International Journal of Computer Applications Technology and Research
Volume 6–Issue 01, 01-08, 2017, ISSN:-2319–8656

# Business Intelligence for Big Data Analytics

Tomas Ruzgas
Department of Applied Mathematics
Kaunas University of Technology
Kaunas, Lithuania

Jurgita Dabulytė-Bagdonavičienė
Department of Applied Mathematics
Kaunas University of Technology
Kaunas, Lithuania

**Abstract**: This article introduces methods and tools which are designed for analyzing Big Data. In the present research, the most popular software tool opportunities have been compared and the differences and advantages have been identified for Business Intelligence (BI) analytics according to the dominant market requirements of BI. The article also presents the technologies of fast calculation processing, including architecture of *in-memory* and *grid* computing.

**Keywords**: big data, business intelligence, grid computing

## 1. INTRODUCTION

Since time immemorial, mankind has been collecting and analyzing particular data. In the course of time, the necessity of fast and reliable findings has been increasing. Digital Universe Study of International market research and analysis company International Data Corporation (IDC) has revealed that the amount of created and replicated data encompassed 2.8 zettabytes in 2012. IDC predicts that digital space will have expanded to 40 zettabytes (it will be 50 times larger than it was 10 years ago) by 2020. New data is generated so quickly that a graphic data chart will represent ideal exponent. Consultation company Gartner, Inc. has reported that business increases its data from 40% to 60% per annum. This type of growth is influenced by mobile technologies and databases associated with customers and their behavior in supermarkets (such data is accumulated by trade networks). In addition to financial institutions, research data of medical and human genome is not falling behind the trend. Especially data in social networks is generated very quickly. This is the most difficult processed and unstructured multimedia data: free-form text, images, sounds and video clips. Nowadays, the data generated by devices comprises 30% of all data; therefore, it is predicted that this figure will have reached 42% by 2020. A considerable amount of data is created every day, but it is not information. In order to obtain the information from data, it is necessary to process particular data. Data Science is described as data analysis using scientific methods. Strategically important, as well as irrelevant information can be hidden in a large amount of data. The search for important information in a massive amount of data has encouraged the emergence of tools for data analysis, high quality application packages or programming tools that help to orientate in a substantial amount of information. Increase in data and information brings new requirements for information processing by computer systems.

Data mining is extraction of useful information from accumulated data. It is remarkable that technologies are able to transform factual data into useful information and knowledge, which are necessary for performance management, market analysis and the decision-making process (Han et al., 2012). Data mining is considered to be a multifaceted concept: it can be defined as identifying structures (models, connections, statistical models or templates) in databases (Fayyad et al., 1993), as well as the application of statistics for data analysis and predictive modelling in order to discover new patterns and trends in big data sets. It may also be described as big data exploration and analysis by automated or semi-automated means with the purpose to find useful patterns and rules (Berry & Linoff, 2008).

Data mining is used for knowledge discovery in databases. During this process, new information is searched for in large amounts of data sets, that could help to gain knowledge of analyzing data and make suitable decisions (Cios et al., 2007). Data mining method helps to find rules for searching tasks and to solve problems of prediction, classification, clustering and interconnectivity; therefore, it is important to have systems, providing various methods for solving tasks of data mining (Dunham, 2002).

The main purposes of this article are to evaluate the tools for big data analytics, to conduct a comparative analysis of the most popular data mining software tools for business intelligence, to identify the differences and similarities of various opportunities and to describe the technologies of fast calculation processing.

## 2. BUSINESS INTELLIGENCE AND ANALYTICS

Traditional BI market share leaders are disrupted by platforms that expand access to analytics and deliver higher business value. BI leaders should track how traditionalists translate their forward-looking product investments into a renewed momentum and improved customer experience.

The BI and analytics platform market are undergoing a fundamental shift. During the past ten years, BI platform investments have largely been in IT-led consolidation and standardization projects for large-scale systems-of-record reporting. These have tended to be highly governed and centralized, where IT-authored production reports were pushed out to inform a broad array of information consumers and analysts. Now, a wider range of business users are demanding access to interactive styles of analysis and insights from advanced analytics, without requiring them to have IT or data science skills. As the demand from business users for pervasive access to data discovery capabilities is growing, IT sector wants to deliver on this requirement without sacrificing governance.

While the need for system-of-record reporting to run businesses remains, there is a significant change in how companies are satisfying these and new business-user-driven requirements. They are increasingly shifting from using the installed base, i.e. traditional and IT-centric platforms that are the enterprise standard, to more decentralized data discovery deployments that are now spreading across enterprises. There is the transition to platforms that can be rapidly implemented and can be used either by analysts and business users in order to find insights quickly, or by IT to quickly build analytics content in order to meet business requirements and to deliver more timely business benefits. Gartner estimates that more than a half of net new purchasing is data-discovery-driven (Sommer et al., 2014). This shift to a decentralized model, empowering more business users, also drives the need for a governed data discovery approach.

This is a continuation of a six-year trend, where the installed-base, IT-centric platforms are being complemented, and in 2014, they were increasingly displaced for new deployments and projects with business-user-driven data discovery and interactive analysis techniques. This is also increasing IT's concerns and requirements around governance as deployments grow. Making analytics more accessible and pervasive to a broader range of users and use cases is the primary goal of organizations, making this transition.

Traditional BI platform vendors have tried very hard to meet the needs of the current market by delivering their own business-user-driven data discovery capabilities and enticing adoption through bundling and integration with the rest of their stack. However, their offerings have been pale imitations of the successful data discovery specialists (the gold standard being Tableau) and, as a result, have had limited adoption to date. Their investments in next-generation data discovery capabilities have the potential to differentiate them and spur adoption, but these offerings are works in progress (for example, SAP Lumira and IBM Watson Analytics).

Also, in support of wider user adoption, companies and independent software vendors are increasingly embedding traditional reporting, dashboards and interactive analysis into business processes or applications. They are also incorporating more advanced and prescriptive analytics built from statistical functions and algorithms available within the BI platform into analytics applications. This will deliver insights to a broader range of analytics users that lack advanced analytics skills.

As companies implement a more decentralized and bimodal governed data discovery approach to BI, business users and analysts also demand access to self-service capabilities beyond data discovery and interactive visualization of IT-curated data sources. This includes access to sophisticated, yet business-user-accessible, data preparation tools. Business users also look for easier and faster ways to discover relevant patterns and insights in data. In response, BI and analytics vendors introduce self-service data preparation (along with a number of startups such as ClearStory Data, Paxata, Trifacta and Tamr), and smart data discovery and pattern detection capabilities (an area for startups such as BeyondCore and DataRPM) to address these emerging requirements and to create differentiation in the market. The intent is to expand the use of analytics, particularly insight from advanced analytics, to a broad range of consumers and non-traditional BI users, increasingly on mobile devices and deployed in the cloud.

Interest in cloud BI declined slightly during 2015, to 42% compared with last year's 45% — of customer survey respondents reporting they either are (28%) or are planning to deploy (14%) BI in some form of private, public or hybrid cloud. The interest continued to lean toward private cloud and comes primarily from those lines of business (LOBs) where data for analysis is already in the cloud. As data gravity shifts to the cloud and interest in deploying BI in the cloud expands, new market entrants such as Salesforce Analytics Cloud, cloud BI startups and cloud BI offerings from on-premises vendors are emerging to meet this demand and offer more options to buyers of BI and analytics platforms. While most BI vendors now have a cloud strategy, many leaders of BI and analytics initiatives do not have a strategy on how to combine and integrate cloud services with their on-premises capabilities.

Moreover, companies are increasingly building analytics applications, leveraging a range of new multistructured data sources that are both internal and external to the enterprise and stored in the cloud and on-premises to conduct new types of analysis, such as location analytics, sentiment and graph analytics. The demand for native access to multistructured and streaming data combined with interactive visualization and exploration capabilities comes mostly from early adopters, but are becoming increasingly important platform features.

As a result of the market dynamics discussed above, for this Magic Quadrant, Gartner defines BI and analytics as a software platform that delivers 13 critical capabilities across three categories (i.e. to enable, produce and consume) in support of four use cases for BI and analytics. These capabilities support building an analytics portfolio that maps to shifting requirements from IT to the business. From delivery of insights to the analytics consumer, through an information portal often deployed centrally by IT, to an analytics workbench used by analysts requiring interactive and smart data exploration (Tapadinhas, 2014), these capabilities enable BI leaders to support a range of functions and use cases from system-of-record reporting and analytic applications to decentralized self-service data discovery. A data science lab would be an additional component of an analytics portfolio. Predictive and prescriptive analytics platform capabilities and vendors are covered in Fig. 1.

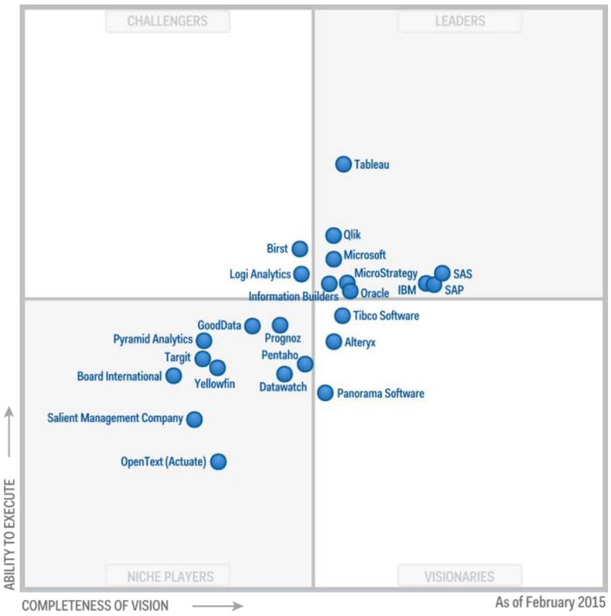


Figure. 1 Magic Quadrant for Business Intelligence and Analytics. Source: Gartner

Vendors are assessed for their support of four main use cases:

- centralized BI provisioning: supports a workflow from data to IT-delivered-and-managed content;
- decentralized analytics: supports a workflow from data to self-service analytics;
- governed data discovery: supports a workflow from data to self-service analytics to systems-of-record, IT-managed content with governance, reusability and promotability;
- OEM/embedded BI: supports a workflow from data to embedded BI content in a process or application.

Vendors are also assessed according to the following 13 critical capabilities: business user data mashup and modelling, internal platform integration, BI platform administration, metadata management, cloud deployment, development and integration, free-form interactive exploration, analytic dashboards and content, IT-developed reporting and dashboards, traditional styles of analysis, mobile, collaboration and social integration and embedded BI (Sallam et al., 2015).

Fig. 1 presents a global view of Gartner's opinion of the main software vendors that should be considered by organizations, seeking to use BI and analytics platforms to develop BI applications. Buyers should evaluate vendors in all four quadrants without assuming that only the Leaders can deliver successful BI implementations. Year-over-year comparisons of vendors' positions are not particularly useful, given the market dynamics (such as emerging competitors, new product road maps and new buying centers); also, clients' concerns have changed. It is also important to avoid the natural tendency to ascribe personal definitions. For the purposes of evaluation in this Magic Quadrant, the measures are very specific and likely to be broader than the axis titles may imply at first glance.

According to the study of Gartner, Inc. (world's leading research and consultancy company of information technology), which was conducted in 2015, SAS and the Tableau were recognized as the world's greatest leaders in the field of business intelligence and analytics platforms. The results of evaluation are presented in (see Fig. 1) (Note: the best position is at the top right corner of the figure).

SAS Institute Inc. offers a vast array of integrated components within its Business Intelligence and Analytics suite that combines deep expertise in statistics and predictive modelling with innovative visualization enabled by powerful in-memory processing capabilities. SAS Visual Analytics is the flagship product in the suite for delivering interactive and self-service analytic capabilities at an enterprise level, i.e. extending the reach of SAS beyond its traditional user base of power users, data scientists and IT developers within organizations. SAS also leverages its range of platform components and expertise in various industries to offer a wide range of vertical- and domain-specific analytic applications.

SAS is again a leader this year as it continues to build momentum with SAS Visual Analytics, which was released in 2012 and has gained some traction in the market against the data discovery leaders through product differentiation and a more accessible pricing model (with a lower entry point than initially offered). SAS also continues to demonstrate very strong vision in many areas such as the expansion of both smart data discovery capabilities and embedded advanced analytics within SAS Visual Analytics, seamless navigation between SAS Visual Analytics and SAS Visual Statistics and integration across other core analytic components of the platform in order to address enterprise requirements for governed data discovery.

- Strengths
  - SAS was rated slightly higher for market understanding (by references) than the average for this Magic Quadrant; this is a composite measure combining ease of use, complexity of analysis and breadth of use. Support for complex analytic use cases is an obvious strength for SAS, but the fact that eight other vendors ranked higher for complexity of analysis may indicate that in many cases the primary product being used is Enterprise BI, which offers more traditional styles of reporting, and that penetration and adoption of Visual Analytics to address more complex use cases is a work-in-progress within SAS's BI customer base. The portfolio of products reaches a broader range of users leveraging the platform to support use cases spanning the full analytic spectrum, which is positive for SAS and a differentiator for its platform.
  - The main reasons why reference customers choose SAS are functionality and product quality, which are clear strengths. SAS delivers a full range of functionality through integrated BI and analytic platform components such as SAS Visual Analytics, SAS Office Analytics and SAS BI/Enterprise BI Server (EBI) as well as complementary products used for data integration, data management, data mining and predictive modelling, all built with a focus on product quality for which SAS was rated just above the overall average.
  - The SAS BI and analytics platform can be deployed to meet the needs of a diverse set of use cases, as indicated by reference organizations that ranked SAS third for frequency of deployment in both centralized and decentralized BI use cases. This diversity positions SAS favourably to differentiate itself from other vendors in the market with a platform that is able to meet both the enterprise IT needs and business self-service needs.
  - Nearly 15% of survey references report using the integrated self-service data preparation capabilities offered by SAS to allow business users and analysts to access, integrate and transform data in preparation for analysis. The availability of integrated business-user data preparation capabilities is a differentiator for SAS compared with other data discovery vendors; particularly Tableau, which relies on third-party integration with vendors such as Alteryx, Paxata and Trifacta to deliver this capability to its customers.
- Weaknesses
  - License cost was again a concern for SAS customers in 2014 and was cited as a barrier to wider deployments by 46% of the reference organizations who responded to the survey, higher than all but one other vendor in the Magic Quadrant. It is expected that this will improve in the next year's survey as customers benefit from the fact that

SAS revamped its Visual Analytics pricing structure in September 2014 to address this concern and offer its customers a per user price point that more closely aligns with competitive data discovery products in the market. With this change, SAS has also made Visual Analytics more accessible to the SMB market with a lower point of entry, i.e. four-core server license priced at $8,000, which can support up to five power users. Under the new pricing structure, the per-user license cost of Visual Analytics is more comparable to leading data discovery offerings, which is critical to SAS's goal of extending the reach of analytics more broadly within its customer base and to win net new customers.

- Customers reported significant difficulty in migrating to the latest release of the SAS platform components that they have deployed, as indicated by its being given the fourth-highest migration difficulty rating. While the migration difficulty rating is high (compared to other Magic Quadrant vendors included in the survey), it should be noted that the score corresponds to a rating between "straightforward" and "somewhat complex," according to the scale used in the survey. It is also likely that the complexity reported by some customers is related to platform-level migrations rather than version updates to individual products.

- Support for complex use cases is platform strength, but SAS references rate both overall ease of use and business benefits delivered as below the overall average. This could be because the adoption of Visual Analytics, while higher than other traditional market share leaders, is still early and has yet to have its full impact on the perceived ease of use; also, the most recent release of EBI, which offers usability improvements, has not yet been widely deployed. Other data discovery platforms are currently doing a better job of executing on the vision of making hard things easy and being accessible to a broader range of users, but SAS Visual Analytics is gaining awareness and traction in the market and has the potential to close the gap.

Tableau's intuitive and visual-based data discovery capabilities have transformed business users' expectations about what they can discover in data and share without extensive skills or training with a BI platform. Tableau's revenue growth during the past few years has very rapidly passed through the $100 million, $200 million and $300 million revenue thresholds at an extraordinary rate compared with other software and technology companies.

Tableau has a strong position on the Ability to Execute axis of the Leaders quadrant, because of the company's successful "land and expand" strategy that has driven much of its growth momentum. Many of Gartner's BI and analytics clients are seeing Tableau usage expand in their organizations and have had to adapt their strategy. They have had to adjust to incorporate the requirements that new users/usage of Tableau bring into the existing deployment and information governance models and information infrastructures. Despite its exceptional growth, which can cause growing pains, Tableau has continued to deliver stellar customer experience and business value. It is expected that Tableau will continue to rapidly expand its partner network and to improve international presence during the coming years.

- Strengths
  - Tableau has clearly defined the market in terms of data discovery, with a focus on "helping people see and understand their data." Currently, it is the perceived market leader with most vendors viewing Tableau as the competitor that they most want to be like and to beat. At a minimum, they want to stop the encroachment of Tableau into their customer accounts.
  - Tableau rates among the top five vendors for aggregate product score, with particular strengths in the decentralized and governed data discovery use cases. In particular, analytic dashboards, free-form exploration, business-user data mashup and cloud deployment are platform strengths. Tableau's direct query access to a broad range of SQL and MDX data sources, as well as a number of Hadoop distributions, native support for Google BigQuery, Salesforce and Google Analytics has been a strength of the platform since the product's inception and often increased its appeal to IT versus in-memory-only options. As a result, customers report having slightly below-average deployment sizes in terms of users, but among the highest data volumes (in this Magic Quadrant).
  - Tableau has managed its growth and momentum well. The company has been able to grow and scale without a significant impact on discounts extended (that is, these are very limited) or customer experience. Most technology companies struggle to manage this balance between growth and execution.
  - Tableau customers report among the highest scores in terms of breadth and ease of use along with high business benefits realized. Gartner inquiries and customer conversations reveal that Tableau users report enthusiasm for the product as a result of being able to rapidly leverage insights from Tableau that have a significant impact on their business. Customers also report faster-than-average report development times.
  - Tableau is an R&D-driven company. It continues to invest in R&D at a higher pace (in terms of revenue percentage, it was 29% in 2014) than most other BI vendors.
- Weaknesses
  - Tableau has a limited product line focused on data discovery. Organizations like buying and managing fewer software assets and vendors. At some point, many of the new generation of visualization and discovery tools that are bundled with other (competitor) applications may gain traction, particularly as they roll out smart data discovery and self-service data preparation differentiators.
  - IT-developed reports and dashboards, traditional styles of analysis, metadata management, development and integration, BI platform administration, embedded BI and collaboration are rated as weaker capabilities of the platform, making it less well suited for centralized and embedded use

cases. When Tableau customers have advanced data preparation, production reporting, advanced analytics, distribution and alerting as requirements, they have to turn to third-party products and partner capabilities. This may also limit its ability for large-scale displacements, but not for large scale surrounding and marginalizing of IT and report-centric incumbents.

- Tableau is the competitive target of most other vendors in this market. It faces competitive threats from every other vendor in the market that is also focused on delivering self-service data discovery and visualization capabilities, in an attempt to slow down Tableau's momentum.
- Tableau offers limited advanced analytics capabilities. R integration has been recently added and is a major improvement for users, needing more statistical and advanced capabilities. Other vendors, such as SAS, SAP and Tibco, have more advanced native capabilities.

Tableau's enterprise features around data modelling and reuse, scalability and embeddability, which enable companies to use the platform in a more pervasive and governed way, are evolving with each release, but are still more limited than IT-centric system-of-record platforms.

# 3. ANALYTICS: BUSINESS VISUALIZATION

Regardless of size or industry sector, organizations collect all types and amounts of data. Unfortunately, traditional architectures and existing infrastructures are not designed to deliver the fast analytical processing needed for rapid insights. As a result, IT is swamped with constant requests for ad hoc analyses and one-off reports. Any delay can frustrate decision makers because it takes too long (or it may be impossible) to get the information needed to answer their questions quickly. Increasingly, decision makers, analysts and other business users want to share reports via email or mobile devices. To help one make sense of the growing data within organization, SAS Institute Inc. product Visual Analytics provides an interactive user experience that combines advanced data visualization, an easy-to-use interface and powerful in-memory technology. This lets a wide variety of users visually explore data, execute analytics and understand what data means. Then they can create and deliver reports wherever needed via the web, mobile devices or Microsoft Office applications.

Data visualization helps explore and make sense of data (Tagarden, 1999). Adding analytics to visualizations helps uncover insights buried in data. Analytics visualization helps discover trends within your business and the market that affect the bottom line. One can quickly recognize outliers that may affect product quality or customer churn. One can also easily recognize parameters in data that are highly correlated. Some of these correlations will be obvious, but others will not. In identifying these relationships, one is able to focus on the areas most likely to influence highest-priority goals. By combining dashboards, reporting, BI and analytics, analytic tools provide both data visualization and analytic visualization. No matter how deep one wants to dive into data, analytic tools provide the capabilities and visualization techniques to take the user there. SAS Visual Analytics lets one go directly from reporting to exploration in the same user experience. With support for data management, report creation, collaboration through SAS Mobile BI apps and Microsoft Office integration, SAS Visual Analytics helps unlock insights and improve efficiency throughout the organization. SAS Visual Analytics reduces the number of tools that should be used and the number of systems that IT must maintain. SAS Visual Analytics combines powerful in-memory technologies with an extremely easy-to-use exploration interface and drag-and-drop analytics capabilities. No coding is required. Report creators, business analysts and even traditional consumers of BI reports can create and share visualizations to gain new insights from their data. SAS Visual Analytics is designed to handle big data, with in-memory processing designed to meet the demands of today and tomorrow. Flexible deployment options let the user easily scale system as data and analytics needs grow. SAS Visual Analytics integrates with Microsoft Office, helping share interactive and self-service reports directly within familiar Microsoft Office applications. These are more than static reports. SAS Visual Analytics allows to build reports that enable collaborative and engaging discussions that can drive deeper insights and better decisions.

The SAS LASR Analytic Server is the in-memory analytics engine for SAS Visual Analytics. In-memory analytics allows quickly determine relationships across hundreds of parameters in billions of rows of data. After all, speed and accuracy are critical to effective analytics. With social media data and freeform text documents becoming part of data ecosystem, the question is often “What valuable information is in all this data?” Data from the social media world, including Twitter streams, Google Analytics and Facebook, as well as call center logs, online comments and other text-based documents can be analyzed to determine much more than the frequency of common terms and phrases. The sentiment around topics, terms and entire text documents can also determined. Through the combination of text sentiment analysis and data visualization techniques, documents can be filtered by topic and sentiment; therefore, areas that need attention may be isolated.

With web-based exploratory analysis and other easy-to-use features, even users without analytical expertise can use predictive analytics to gain precise insights (Matthew et al., 2006). Nontechnical users can create and change queries simply by selecting items from a sidebar or dynamically filtering and grouping data items. Autocharting selects the visualization that best suits the type of data chosen. “What does it mean” pop-up boxes provide explanations of analytical techniques, helping everyone understand the data and what the analysis means. Analytically savvy users can use visualization techniques to spot trends and derive deep intelligence quickly and easily. This eliminates much of the everyday trial-and-error process currently used to identify areas that need further analysis.

How do customers navigate website of organisation? What is the customer journey through organisation support structure? The data accumulated from operational systems provides information to paint a clear picture of how transactions move within those systems. Path analysis with SAS Visual Analytics allows to see those flow patterns and recognize trends, such as where customers enter the website, where they navigate and where they exit. With SAS Visual Analytics, successful flow patterns and isolate flows that failed to deliver the desired action can be identified. This level of analytics visualization provides decision makers with the information required to pinpoint opportunities for improvement. Analytic features are tailored for ease of use; therefore, everyone can create analytic visualizations on their own without learning new skills or engaging IT. Self-service autoloading allows the users to load their own data from Excel spreadsheets and other sources for analysis.

Growing volumes and varieties of big data make it difficult to visualize and understand valuable relationships in data and obtain the analytically based answers, which require to take the best actions. Traditional IT infrastructures are just not designed for rapid and iterative analytical processing and on-the-fly changes to predictive models. It is hard for statisticians, data scientists and business analysts to build the number of models that are needed. They cannot easily experiment with segments or groups, or quickly refine their models to find the best one. SAS Visual Statistics solves these issues. As an add-on to SAS Visual Analytics, it combines interactive data exploration and discovery with the ability to easily build and adjust huge numbers of predictive models. It is really very easy as no coding is required. The in-memory engine reads data into memory once, putting an end to constant and expensive data shuffling.

SAS Visual Statistics provides an interactive, intuitive, drag-and-drop, web-browser interface for creating descriptive and predictive models on data of any size rapidly. It takes advantage of LASR Analytic Server to persist and analyze data in memory and deliver near instantaneous results. When combined with SAS Visual Analytics, it provides a fast and single environment for interactive data exploration and model development. SAS Visual Statistics is designed for statisticians, data scientists and business analysts who want to visually and instantly interact with and analyze complex data. The easy-to-use, drag-and-drop interface provides nonprogramming access to powerful SAS statistical modeling and machine-learning techniques. These techniques are used to predict outcomes that result in better and more targeted actions.

SAS Visual Statistics is an add-on to SAS Visual Analytics Explorer. The common SAS Visual Analytics Explorer environment provides interactive data exploration and analytical modeling capabilities. It can quickly identify predictive drivers among multiple exploratory variables, and interactively discover outliers and data discrepancies. Then, this information may be used to populate interactive environment for sophisticated predictive modelling. The web browser interface makes it a simple drag-and-drop process to create powerful descriptive and predictive models. Multiple sers can easily collaborate to build and refine the best models. Interactive processing is very fast; thus, users can quickly and easily experiment with different techniques.

## 4. GRID: FASTER PROCESSING

These days, IT budgets are typically limited in most organizations, which makes meeting the computing demands of today's business environment a constant challenge. Buying the latest and greatest servers (i.e., scaling up) to meet peak-demand computing loads is one solution, but it can be both costly and inefficient. Organizations' use of business analytics grows, as well as the need for a flexible IT infrastructure that can scale cost-effectively while meeting peak demands and managing growing and increasingly diverse user workloads. Grid enables organizations to create a managed, shared grid computing environment for processing large volumes of data and analytic programmes. The solution provides critical capabilities for meeting an organization's business analytics needs, including workload balancing, job prioritization, high availability, parallel processing, resource assignment and monitoring.

Grid gives IT greater flexibility to meet service level commitments by easily reassigning computing resources to meet peak workloads or changing business demands (Smith et al., 2002). The solution provides a central point of control for administering policies, programmes, queues and job prioritization across multiple types of users and applications to achieve business goals under a given set of constraints. Having multiple servers in a grid computing environment enables jobs to run on the best available resource. If a server fails, its jobs can be transitioned seamlessly to another server, providing high availability. In addition, IT staff can perform maintenance on specific servers without interrupting analytics jobs, as well as introduce additional computing resources without disrupting the business. Multiprocessing capabilities let divide individual jobs into subtasks that are run in parallel

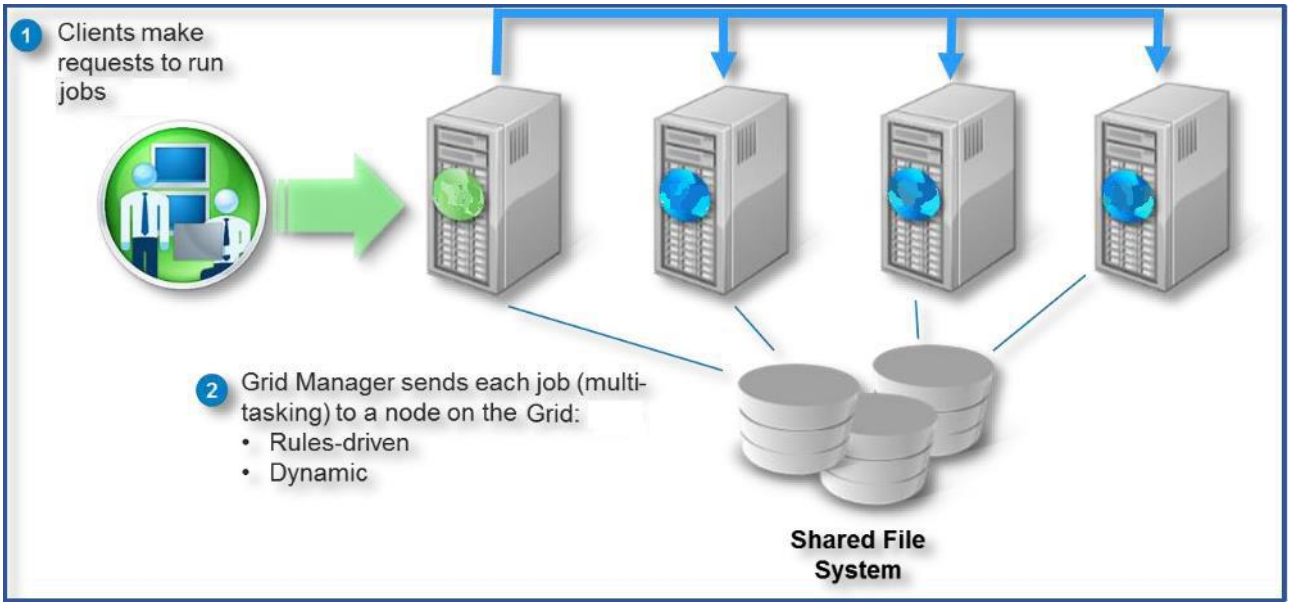

1 Clients make requests to run jobs
2 Grid Manager sends each job (multi-tasking) to a node on the Grid:
• Rules-driven
• Dynamic
Shared File System

Figure 2. Grid Computing Architecture.

on the best available hardware resource. The programmes best-suited for parallel processing are those with large data sets and long run times, as well as those with replicate runs of independent tasks running against large data sets. Processing data integration, reporting and analytical jobs accelerate decision making across the enterprise. Grid lets fully utilize all available computing resources now and cost-effectively scale out as needed, adding capacity in single-processing units to keep IT spending in check (Joseph, 2004). As it can add low-cost commodity hardware resources incrementally, there is no need to size today's environment.

SAS Grid Manager's patented technology uses industry-leading grid computing middleware from Platform Computing to get maximum availability from business analytics environment. The solution gives a competitive advantage by enabling to balance user and application workloads among available computing resources; consequently, it is possible to obtain results much more quickly. IT can add computing resources in the form of lower-cost commodity hardware incrementally, eliminating the need to size today's environment for tomorrow's demands.

SAS data integration and analytical products are automatically tailored for parallel processing in a grid computing environment. To achieve maximum processing efficiency with minimum user intervention, these programs detect the grid environment at the time of execution. The grid-enabled logic, that is produced, can be saved as stored processes for the use by other reporting clients to generate results for more users as cost-effectively as possible. Other SAS solutions, including SAS Enterprise Guide and SAS Risk Dimensions, can automatically submit jobs to a grid of shared computing resources. All programmes can take advantage of grid computing environment with the addition of programming syntax and a structure that allows the submission of entire programmes to the grid or the parallel execution of programme steps (subtasks).

A wide variety of SAS jobs can be scheduled across grid environments for optimal resource utilization and faster processing. Individual jobs can be divided into subtasks that are then executed in parallel to accelerate processing and increase workload throughput. In today's international organizations, nightly batch-processing windows no longer exist. As a result, data is available 24/7 and can be quickly loaded and analyzed.

## 5. CONCLUSIONS

The need for platforms to scale and perform for larger amounts of diverse data will also continue to dominate BI market requirements. At the same time, the ability to bridge decentralized business-user-led analytics deployments with those centralized to serve the enterprise will be a crucial ongoing challenge for IT and BI vendors. With the added complexities introduced by new data sources (such as the cloud, real-time streaming events and sensors and multistructured data) and new types of analysis (such as link/network and sentiment analysis, and new algorithms for machine learning), new challenges and opportunities will emerge to integrate, govern and leverage these new sources to build business value. Leaders of BI initiatives will be under pressure to identify and optimize these opportunities and to deliver results faster than ever before.

In-memory analytical processing build models faster (Zaharia et al., 2012). With the LASR Analytic Server, there is no need to write data to disk or perform data shuffling. SAS Visual Statistics loads all data into memory once and interacts with the data without reloading it each time when a new task is performed. This means the impact of changes to models (e.g., adding new variables or removing outliers) is instantly visible. Because it is designed for concurrent processing, many users can create and run complex models simultaneously. Data and analytic workloads are performed in a distributed form across multiple server nodes, and are multithreaded on each node for blazingly fast speeds.

Because SAS has made grid computing an automatic capability within multiple applications, processing times are greatly reduced. As a result, one can integrate, cleanse and analyze larger volumes of data more quickly.

## 6. REFERENCES

[1] Berry M. J. A. and Linoff G. S. 2008. Mastering Data Mining: The Art and Science of Customer Relationship Management, Wiley, p. 512.

[2] Cios K.J., Pedrycz W., Swiniarski R.W., and Kurgan L. 2007. Data Mining: A Knowledge Discovery Approach, Springer, p. 606.

[3] Dunham M.H. 2002 Data Mining: Introductory and Advanced Topics, Pearson, p. 315.

[4] Fayyad U., Chaudhuri S., Bradley P. 1993. Data Mining and its Role in Database Systems, vol. 5, no. 6, 914–925.

[5] Han J., Kamber M., and Pei J. 2012. Data Mining: Concepts and Techniques – 3rd edition, Elsevier, p. 740.

[6] Joseph J. 2004. Evolution of Grid Computing Architecture and Grid Adoption Models, IBM System Journal, vol. 43, iss. 4, 624-645.

[7] Matthew K.O.L., Christy M.K.C, Kai H.L., Choon L. 2006. Understanding Customer Knowledge Sharing in Web-based Discussion Boards: An Exploratory Study, Internet Research, vol. 16 iss. 3, 289 – 303.

[8] Sallman R. L., Hostmann B., Schlegel K., Tapadinhas J., Parenteau J., and Oestreich T. W. 2015. Magic Quadrant for Business Intelligence and Analytics Platforms.

[9] Smith J., Gounaris A., Watson P., Paton N.W., Fernandes A.A.A., Sakellariou R. 2002. Distributed Query Processing on the Grid, Springer.

[10] Sommer D., Buytendijk F., Schlegel K. 2014. Market Trends: Business Intelligence Tipping PointsHerald a New Era of Analytics.

[11] Tagarden D.P. 1999. Business information visualization, Communications of the AIS, vol. 1, iss. 1, article 4.

[12] Tapadinhas J. 2014. How to Architect the BI and Analytics Platform.

[13] Zaharia M., Chowdhury M., Das T., Dave A., Ma J., McCauley M., Franklin M.J., Shenker S., Stoica I. 2012. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing, Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, April 25-27

[14] Gartner Inc. <http://www.gartner.com>

[15] International Data Corporation < https://www.idc.com>

[16] SAS Institute Inc. <http://www.sas.com>

[17] Tableau <http://www.tableau.com>

# Analysis of Modeling Performance and Simulation Tools for Wireless Sensor Networks

A. MOUIZ
EEA&TI laboratory, Hassan II
University of Casablanca
Faculty of Sciences and
Techniques (FSTM)
Mohammedia, Morocco

A. BADRI
EEA&TI laboratory, Hassan II
University of Casablanca
Faculty of Sciences and
Techniques (FSTM)
Mohammedia, Morocco

A. BAGHDAD
EEA&TI laboratory, Hassan II
University of Casablanca
Faculty of Sciences and
Techniques (FSTM)
Mohammedia, Morocco

A. SAHEL
EEA&TI laboratory, Hassan II
University of Casablanca
Faculty of Sciences and
Techniques (FSTM)
Mohammedia, Morocco

**Abstract**: Wireless sensor networks rapidly invaded several application domains. The nodes that make up these networks are electronic devices designed and sized to meet the needs of surveillance, data collection and transport, communication, etc. However, Operation and design of the wireless sensor network systems require reliable and efficient simulation tools before actual implementation of the application. In this paper an analysis of the performance of the modeling phase of the network and simulation tools was presented as a synthesis to evaluate the performance of the systems of a wireless sensor network.

**Keywords**: Wireless sensor networks; modeling; simulation; routing; energy consumption; OPNNET; OMNeT $^{++}$; NS2.

## 1. INTRODUCTION

Wireless sensor networks have been increasingly successful in scientific and industrial research communities. Their general function is the detection of information in the most hostile environments. Wireless sensor networks provide many solutions for the detection and monitoring of our environment, this allows many possibilities for new applications. System modeling and network simulation are two important steps to reduce the cost and duration of the deployment process.

The design and realization of wireless sensor networks are influenced by several parameters, such as radio module status, limited resources, limited bandwidth, dynamic topology, scalability, etc. These factors serve as guidelines for the development of complex algorithms and protocols that are dedicated to wireless sensor networks [1]. However, modeling techniques and simulations were used to evaluate the performance of wireless sensor network systems.

This article aims to present an analysis of the performance of modeling and simulation tools that are dedicated to networks of wireless sensors. In the following sections, we will present the interest and the important requirements of modeling in the field of wireless sensor networks. Then, a typical system model of a wireless sensor network is provided. We then describe some simulation tools and their properties. At the end, we present the results of the comparative study in the form of a synthesis. We draw the conclusion and our prospects in the final section.

## 2. MODELING IN THE FIELD OF WIRELESS SENSOR NETWORKS

In wireless sensor networks, models are the first steps in the realization of a new idea or a new approach. The models created are often simulated in order to estimate their validity by comparing them with other similar models [2]. The modeling of the same node or of the same network may be different according to the level of abstraction considered.

For the modeling of a wireless sensor network system according to the precision of the information wishing to validate, the designer uses in principle different varieties of levels of abstraction. Subsequently, it descends from a level of abstraction to a more detailed level by making the model more subtle and more sophisticated [3]. In other words, abstraction often involves simplifying and replacing a complex and detailed architecture of the device with a comprehensible model by which we can solve a problem. Thus, the model will be general and easily manipulable.

The modeling of the same sensor node may be different depending on the objective assigned to the simulation [4]. For example, if one wishes to optimize the energy consumption in a sensor node or in a global network, there may be a sensor node model or a sensor network model. Consequently, a node can be modeled at several levels of abstraction according to the objective of the modeling but also according to the information available during the design stage for which the model will be used. In this section, the system modeling of a wireless sensor network is studied. Then, a summary of the requirements of the modeling is presented. Finally, a typical system model of a wireless sensor network is provided.

### 2.1 The Requirements of Wireless Sensor Networks Modeling

In order to have credible and sufficiently realistic results thanks to simulations, a correct and well detailed modeling of the real-time properties of the sensor nodes and of the environmental models is conceivable and mandatory [1].

However, there are many requirements for modeling and simulation for wireless sensor networks such as heterogeneity, scalability, power consumption, etc.

Most of the systems in deployed wireless sensor networks are heterogeneous systems. So a modeling of different nodes and the management of the interconnections between them is really necessary. So, we must take into account the support of the heterogeneity in the modelization of a network system of wireless sensors. Due to their high density in the area to be observed, the sensor nodes must be able to adapt their operation in order to maintain the desired topology. In the case of a corrupt or damaged node for an energy reason, the network must be able to take this modification into account while ensuring an equal quality of service. Therefore, modeling must take into account the entry or loss of nodes in the network in order to solve the problem of scalability [5]. In terms of energy consumption, network designers need to obtain accurate power and timing data to adjust their applications before deployment in real-world environments. Because the malfunctioning of a node implies a change in the topology and imposes a reorganization of the network.

## 2.2 A Typical Model of Wireless Sensor Networks System

In order to properly simulate a wireless sensor network system, it is necessary to reproduce the behavior and operation of a wireless sensor in a computer environment. A wireless sensor network consists mainly of three parts: a node system, the network and the physical environment. The figure above represents a typical model of the wireless sensor network system.

The node system in the model shown is composed of two parts: a hardware and other software part. For the hardware platform, it consists of a processing unit which collects data from the capture unit, processes it and decides when and where to send it. It must also carry out programs and different protocols communication. The platform also consists of an RF transceiver, a sensor and a battery whose energy is the most valuable resource in a sensor network, as it directly affects the lifetime of micro- Sensors and therefore of an entire network. Additional components may be added depending on the field of application, such as, for example, a location system such as a GPS, an energy generator or a mobilizer enabling it to move [6]. The software model includes operating system, protocol stack, and application software implementation etc. The nodes are connected among by the wireless network.

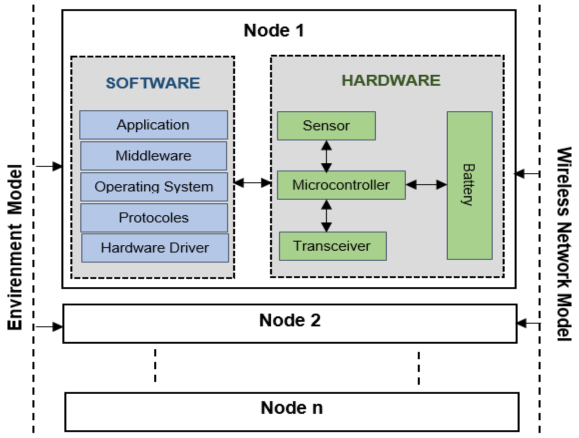


Figure. 1 A Typical Model of Wireless Sensor Networks System

This figure shows a model that contains the topology of the network and transfers the packets between the nodes. It also implements numerous radio frequency channel models. The environment model specifies how physical parameters in the environment vary both spatially and temporally.

Choices on modeling techniques, communication between sensors, had to be made. The development of a simulation environment and a simple example of a sensor network was essential [7]. We needed a modular environment to choose the type of component to be integrated into the simulation in order to make this simulation reliable. We will be able to predict the proper functioning of the network, its performance, its organization, its energy consumption, etc.

## 3. SIMULATION TOOLS AND ENVIRONMENTS FOR WIRELESS SENSSOR NETWORKS

Several simulators are developed for wireless sensor networks. These simulators allow to simulate models of nodes or networks in a virtual environment. The simulations provide a good approximation to verify the different diagrams and applications developed for wireless sensor networks at low cost and in less time.

The most well-known simulators are OPNET [8], OMNeT $^{++}$ [9] and NS2 [10] which are used to simulate the characteristics of networks with specific sensor nodes. In addition, they are mainly used to compare the efficiency of the algorithms used in the network (MAC, routing, etc).

In this section we describe some simulation tools and describe their characteristics based on information published for the study of wireless sensor networks. A comparative study of these simulators, as well as the emulation tools is then carried out. This analysis is also presented as a mapping of the simulators according to the level of abstraction and the design stage of the sensor node to which they are dedicated.

### 3.1 OPNET Modeler

OPNET Modeler is a network simulator developed with the C ++ language and based on an intuitive graphical interface. OPNET is considered the main simulation and modeling tool for commercial networks in the world [11]. It can be used as a research tool and also as a network design / analysis tool, its handling and its use is relatively easy.

The use can build a network model using predefined node models and provided by the OPNET library from equipment and commercially available fixed networking protocols. The simulation under this tool also provides as standard a list of implementations of routers, switch workstations. In wireless networks, the strong point of OPNET lies in the precise modeling of radio transmission, modeling in detail different characteristics.

The OPNET Modeler simulator has three levels of hierarchical structure to define each aspect of the system, from the highest level to the lowest level (network domain, node domain and process domain) [8]. This simulator is free only for universities, it comes with a version for academic use, but with limited capacities.

### 3.2 OMNeT $^{++}$

OMNeT $^{++}$ is an open source simulation environment that uses the C ++ language for simulation models. This simulator offers a robust graphical interface for animation and debugging, and an integrated simulation kernel, so it has graphical tools for real-time simulation of construction and evaluation of results.

The main purpose of OMNeT $^{++}$ is to simulate network communications and also IT systems, it provides the basic implementation of various hardware modules (base radio, CPU) and software (routing schemes Simple) for wireless sensor networks. It adapts mainly to very large network topologies. Indeed, thanks to its flexible basic architecture, it is able to simulate hardware architectures [9]. This platform has become known not only within the scientific community but also in the industrial world. And it is thanks to this modular architecture that it is easier to implement new protocols. The OMNeT $^{++}$ simulator provides both the estimate of the energy consumed in the communication unit and in the processing unit. There may be a single module or a compound module. The first module is a .cc file and an .h file. The second module comprises simple modules or other connected modules connected among. The parameters, sub-modules and ports of each module are specified in a .ned file.

The most recent general purpose simulation environment for sensor networks is called Castalia, it is modular and extensible [11]. The problem with OMNeT $^{++}$ is the lack of a library of modules specific to wireless sensor networks, but many research groups are working on this point to add additional modules specific to wireless sensor networks.

### 3.3 NS2

Network Simulator NS2 is an open source simulation environment that uses the C ++ language for sensor networks. Its main axes are IP networks and its use consists of a good knowledge of Tool Command Language. It is a very popular discrete event simulator for research purposes [10]. He is often involved in the study of multipoint or unipoint routing algorithms, transport protocols, and session protocols. NS2 can be a good choice due to its large community.

The NS2 simulator is more suited to sensor networks because it includes a basic energy modeling, it also allows to model very well the physical layer of the OSI model with different transmission systems, wired or not. The simulator can organize a simple mobility simulations. Moreover, thanks to the .nam (Network Animator) extension, one can see the results of a simulation once completed [11]. The latter mainly considers the energy consumed in the communication unit. This energy is relative to the different internal states of its radio module (transceiver).

Among the problems of NS2 is that it has natively no graphical interface and its object-oriented design that introduces an unnecessary interdependence between the modules. Also it does not have the ability to model the execution time of the application code or the operating system in real time [12]. Thus, all simulations are carried out on the command line.

### 3.4 Emulation Tools

In addition to these three simulation tools, there are also other emulation tools for wireless sensor networks. The emulator is a sub-set of simulator that allows to analyze the codes intended to be embedded in the targeted platform. It allows to emulate realistically the behavior of the embedded software.

Emulation [13] can combine software and hardware implementation. The general principle of these emulators is to run on computer the applications intended to be implemented in the node. Thus, some emulators allow to estimate the energy consumption which corresponds to the details of the behavior of the node. But they are limited to the operating systems used.

There are several types of emulators and they are often dedicated to specific platforms or software. For example, those that simulate nodes that have their own operating system, such as TOSSIM for nodes that use TINYOS, or COOJA for nodes that use CONTIKI [13].

The following table summarizes and compares the three simulation tools studied according to their properties and characteristics. This comparison was made according to published information for the study of wireless sensor networks.

**Table 1. Analysis and comparison of simulation tools for wireless sensor networks**

| | **OPNET** | **OMNeT $^{++}$** | **NS2** |
|---|---|---|---|
| **License** | Commercial | Commercial, academic | Open source |
| **Language supported** | C/C++/Java | C/C++ | C/C++/OTCL |
| **General / Specific** | Specifically designed for WSN | General Simulator | General Simulator |
| **Scalability n > 100** | Excellent | Good | Fair |
| **Mobile network simulation** | ✓ | ✓ | ✓ |
| **Communication with other modules** | ✓ | - | - |
| **3D Radio modelling** | ✓ | - | - |

## 4. DISCUSSION AND ANALYSIS

The target node model must take into account the different activities of the node in the network. It will have to be able to distinguish the impact in energy when adding, replacing or removing a function in the node and it must not depend initially on the technologies, since the results of simulation must give information to the designers from the first stages of node design. The modeling must start at first with a high level of abstraction. This makes it possible to take into account the phases of specification and functional design very early in the progress of the design process. However, low-level abstraction modeling is often designed to simulate node systems after design steps, and is not suited to describing the different activities of nodes in the network.

Several simulation tools are developed for wireless sensor networks, three of these simulators have been analyzed, compared and presented on table 1 according to their performances. The node models in NS2 and OMNeT $^{++}$ are made up of different modules, such as the PHY module, the MAC module, etc. These two simulators provide the energy consumption according to the configuration of the nodes. Moreover, the two node models that exist in NS2 and OMNeT $^{++}$ do not take into account the energy consumption in the acquisition device. Indeed, in node models, although the user can propose applications in each module of the model, the user can not implement communicating, dependent applications that must run simultaneously in different modules. No module supports data processing such as data compression, decision making, etc. However, in order to have better and more accurate simulation for a wireless sensor network, both OMNeT $^{++}$ and NS2 software must be well extended or modified.

For better modeling, the OPNET simulation tool offers better support, better maintenance and proven simulation models. In addition, this simulator offers excellent scalability to large networks. It is capable of communicating with other modules, so it is the only one of the other two simulators that offers 3D radio modeling. It is also equipped with several MAC standard and routing support [12] (802.11, 802.16. 8, UMTS, SMART MAC, GRP, OLSR, and TORA).

Some wireless sensor network simulators do not depend on existing architectures or technologies, but offer users the ability to imagine and choose the appropriate hardware characteristics for simulation. Note that they are based on specific nodes. They often use predefined hardware / software models.

For a better simulation of a wireless sensor network model, the selected simulation tool must have a parallelism allowing the execution of several tasks simultaneously. The simulator must also have an event triggering property in order to synchronize several functions or trigger a change of state of a function. It must also have the notion of a pause which makes it possible to stop a running process and to restart it at the desired moment, as well as the time management allowing to generate a loop period duration, as an internal clock, a function. The results of the chosen simulation tool must provide the evolution over time of the variation in power consumption. These results would help the designer in choosing the hardware and software to be used in the early design stages of the node. In order to consider the energy impact of the different functions to be implemented in the node model, it is necessary to model the direct link between each of the functions and their respective consumptions.

## 5. CONCLUSION

Wireless sensor networks are paving the way for a variety of applications in many fields and are of considerable interest and a new stage in the evolution of information and communication technologies.

In this paper, an analysis of the performance of modeling and simulation tools for sensor networks has been done. We have seen that simulators offer an accurate approximation at low cost and often in less time to check the different applications developed for wireless sensor networks. In addition, the simulation phase also provides an easy-to-use validation environment and a better understanding of network behaviors. However, the current state of the art on node models and simulators does not meet all design methodology requirements.

## 6. REFERENCES

[1] A. Mouiz, A. Badri, A. Baghdad, A. Ballouk, and H. Lebbar. “Analysing Study of a Contribution to Minimize the Energy Consumption and Improve the Performance of Wireless Sensor Networks”, International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol.6, No1. Jan-Feb 2016.

[2] H. Unterassinger, M. Dielacher, M. Flatsher, S. Gruber, G. Kowalczyk, J. Prainsack, T. Herndl, J. Schweighofer, and W. Pribyl. “A Power Management Unit for Ultra-Low Power Wireless Sensor Networks”, IEEE Africon, the Falls Resort and Conference Centre, ISSN: 2153-0033, Livingstone, Zambia, 13 - 15 Sept 2011.

[3] W. Du, D. Navarro, F. Mieyeville, and I. O'connor. “IDEA1: A Validated SystemC-Based Simulator for Wireless Sensor Networks”, Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on, ISSN: 2155-6806, 15 Nov 2011.

[4] B. Kan, L. Cai, L. Zhao, and Y. Xu. “Energy Efficient Design of WSN Based on an Accurate Power Consumption Model”, Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on, ISSN: 2161-9646, 21-25 Sept 2007.

[5] M. Korkalainen, M. Sallinen. “A Survey of RF-Propagation Simulation Tools for Wireless Sensor Networks”, Sensor Technologies and Applications (SENSORCOMM), 2010 Fourth International Conference on, ISBN: 978-1-4244-7537-7, 18-25 July 2010.

[6] G. V. Merrett, N. M. White, N. R. Harris, and B. M. Al-Hashimi. “Energy aware simulation for wireless sensor networks”, in Proc. of the 6th Annual IEEE communications society conference on Sensor, Mesh and Ad Hoc Communications and Networks, ser. SECON’09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 64–71.

[7] G. N. Bravos, A. G. Kanatas, A. Kalis. “Power Control Techniques for Energy Efficient Wireless Sensor Networks”, Wireless Conference 2006 - Enabling Technologies for Wireless Multimedia Communications (European Wireless), 12th European, ISBN: 978-3-8007-2961-6, 1 June 2011.

[8] X. Chang, “Network simulations with OPNET”, Simulation Conference Proceedings, 1999 winter. ISBN: 0-7803-5780-9. 06 Aug 2002.

[9] X. Xian, W. Shi, and H. Huang, “Comparison of OMNET++ and other simulator for WSN simulation”, Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on. ISSN: 2156-2318. 1 August 2008.

[10] A. R. Khan, S. M. Bilal, and M. Othman. “A performance comparison of open source network simulators for wireless networks”, IEEE International Conference on Control System Computing and Engineering, Nov 2012.

[11] M. Korkalainen, M. Sallinen, N. Kärkkäinen, P. Tukeva. “Survey of Wireless Sensor Networks Simulation Tools for Demanding Applications”, Networking and Services, 2009. ICNS '09. Fifth International Conference on. ISBN: 978-1-4244-3688-0. 26 May 2009.

[12] S. A. Madani, J. Kazmi, S. Mahlknecht. “Wireless sensor networks: modeling and simulation”, World's largest Science, Technology & Medicine Open Access book publisher. Chapter from the book Discrete Event Simulations.

[13] M. Imran, A. M. Said, and H. Hasbullah, “A Survey of Simulators, Emulators and Testbeds for Wireless Sensor Networks”, Information Techonology (ITSim), International Symposium in, pp. 897 – 902. ISBN: 978-1-4244-6715-0. Jun 2010.

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 13-18, 2017, ISSN:-2319–8656

# Software Performance Workload Modelling

Vijay Datla

**Abstract:** The debate between performance engineers and business stakeholders over non-functional requirements is probably as old as the performance discipline itself. 'What set of transactions is enough to represent my system?', 'Why do we not load test every transaction?' , 'Our volumes are much higher than what the targets show' are some of the common questions that need to be answered. From a technical perspective, benefits from load testing every transaction are not enough to justify the effort involved in the exercise. However, for a business, even a small risk of one untested low volume transaction affecting the others or bringing down the entire system is high enough to raise a flag. This paper is an attempt to balance these concerns by discussing how to create workload models that are closer representations of the real world enterprise applications. It answers common requirement gathering questions like where to look for information, on what basis to include and exclude use cases from workloads and how to derive a complete and convincing workload model. This paper highlights the risks associated with selective modelling and the possible mitigations. It also brings to the table tips and tricks of the trade, some lessons learnt the hard way.

**Keywords**: Performance, Modelling, Vijay Datla, Vijay, Datla

## 1. REQUIREMENT ANALYSIS:

Just like any Software Development Lifecycle (SDLC), a Performance lifecycle also begins with Requirements Analysis with the difference that the requirements are purely non-functional in nature. Non-functional requirement is a requirement that specifies the criteria that can be used to judge the operation of a system rather than a functional behavior. There are several kinds of non-functional requirements like Security, Maintainability, Usability and so on but the specifics that we are interested in are Performance, Scalability and to a certain extent Availability. Requirements gathering forms the foundation for all future performance engineering activities on a project. Mistakes made in understanding the business requirements translate into setting of wrong goals and takes all the performance efforts into the wrong direction. Requirements gathering is therefore the key to a successful Performance Engineering project. But even before getting into requirements, it is important to understand the objectives. It is a common misconception that performance can only be done to measure the response time of the system. In literal terms, measuring performance of a system is purely Performance Testing which is part of a larger discipline called Performance Engineering. Performance testing is a means; an enabler in achieving the Performance engineering objectives. So what are these objectives?

•Measuring and improving Performance of an application
•Meeting the non-functional requirement targets
•Improving user experience
•Benchmarking the application and hardware
•Validating Hardware Sizing Once the objectives are clear, the next step is to define the scope at a high level, meaning which modules or what part of the solution will need to be tested as part of the performance exercise. To go deeper into the objectives and scope of performance, it is essential to have a thorough understanding of the system. This understanding can come not just by studying the application but also by studying the business.

### 1.1. Asking the right questions:

- Customer base
- Growth rate
- Concurrency
- Volume centircs vs user centrics
- Most common transactions
- Response time requirements
- User arrival pattern

Gathering requirements for performance testing is the most challenging task given that there is no one place with consolidated information and most sources are external. Readily available non-functional performance requirements and statistics is a rare occurrence. However, it not the lack of availability that adds to the challenge, it is the process of gathering and consolidating data from various sources that's a cumbersome task. More than getting the right answers, it is about asking the right questions. Since the process involves dealing with business, its important to frame questions more comprehendible to a business mind. Instead of asking what is the concurrency or throughput target, try asking what is the customer base of the business? How many of these customers will be accessing the system at any given time? The following should give an idea:
• What is the expected business growth rate?
• Is the system volume centric or user centric?
• What response time is the system required to serve in case of web based OLTP transactions?

• What are the most common use cases? or transactions that happen on the system most frequently?
• Do all users arrive into the system over a small window or are they spread across the day?
• What are the peak periods of access to the system?
• Are there periodic tasks that the system is designed to accomplish?
E.g.
• End of Month/Quarter reports?
• Close of business?
• Seasonal sales?
• Year End closing? And so on

### 1.2. Picking the right resources:

Common sources like Business Analytics, RFP, Business volume reports, Audit reports, Inputs from legacy system, Capacity sizing document, Webserver access logs, Data ware house, google analytics.On enterprise level projects there can be several sources of information when gathering the non-functional requirements.

**•Business Analysts (BAs)**

•BAs are always the first source of information for non-functional requirements. They may or may not have all the information required, but they will be able to make the connection to the right business contacts.

**•RFPs**

•RFPs usually contain a non-functional requirements section. The requirements specific to Performance may be few and non-elaborated but will still contain response times, customer base, transaction volumes etc.

**•Business Reports**

•There are several reports that the business maintains like Volume reports, Accounting, auditing reports that can provide insight into business statistics

**•Legacy Systems**

•In case of legacy modernization projects, there already is a system, maybe a mainframe that is still serving the business. Running simple select queries on this system can help in studying the real world transaction volumes and load patterns

**•Hardware Sizing Documents**

•In the initial stages of SDLC, enterprise projects go through the process of determining the hardware required to support the solution. This sizing is based on the throughput that the system is expected to achieve. So either on a high level or in detail, some study is already done at this stage that can often be used as opposed to reinventing the wheel.

**•Google Analytics**

•For enterprise applications with already existing websites, Google Analytics is a web-analytics solution that provides detailed insights into the website traffic. It reports traffic patterns, sources of incoming load, navigation patterns, detailed load patterns over a period of time and much more.

**•Data Warehouse**

•Most enterprise projects maintain data warehouses for storing archived information that can be accessed to obtain non-functional details

**•Domain Research**

•In most cases there is existing research in the market that has been done on various kind of applications catering to several domains. If there is absolutely no information available in house then these researches can be a good place to start from.

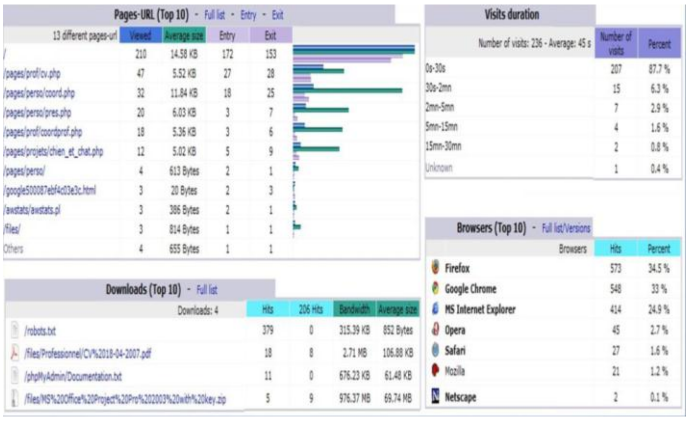
**•Log Parsers**

•In case of implementations with an existing system in place, server access logs are excellent sources of real time information. There are several tools in the market that parse access logs into comprehendible, meaningful information. There are several log parsing tools in the market that can produce meaningful data from Web Server logs. AWStats is one such open source log parsing tool that is being used here as an example. This parser extracts data from a web server access log and converts it into meaningful server statistics. It is much like a web analytics tool, only that it works offline. It produces graphs that provide insight into load patterns in terms of user visits, page visits bandwidth etc.

The tool lists the most commonly accessed pages which helps in determining the high volume transactions. It also indicates the browser most commonly used to access the application website. With the advancement of browsers features and variety in the market, this information is useful in deciding what browser to use when simulating load on the application.

The most important use of the tool is in studying the user arrival and load pattern. The hourly graphs outline the user arrival pattern and the
weekly, monthly and yearly graphs help in determining the peak periods.

The below charts show the website usage patterns in terms of top accessed URLs, top downloads, average user visit durations and distribution of browsers for incoming requests.



Pages-URL (Top 10) - Full list - Entry - Exit

| 13 different pages-url | Viewed | Average size | Entry | Exit |
|---|---|---|---|---|
| / | 210 | 14.58 KB | 172 | 153 |
| /pages/prof/cv.php | 47 | 5.52 KB | 27 | 28 |
| /pages/perso/coord.php | 32 | 11.84 KB | 18 | 25 |
| /pages/perso/pres.php | 20 | 6.03 KB | 3 | 7 |
| /pages/prof/coordprof.php | 18 | 5.36 KB | 3 | 6 |
| /pages/projets/chien_et_chat.php | 12 | 5.02 KB | 5 | 9 |
| /pages/perso/ | 4 | 613 Bytes | 2 | 1 |
| /google500087ebf4c03e3c.html | 3 | 20 Bytes | 2 | 3 |
| /awstats/awstats.pl | 3 | 386 Bytes | 2 | 1 |
| /files/ | 3 | 814 Bytes | 1 | 1 |
| Others | 4 | 655 Bytes | 1 | 1 |

Visits duration

| Number of visits: 236 - Average: 45 s | Number of visits | Percent |
|---|---|---|
| 0s-30s | 207 | 87.7 % |
| 30s-2mn | 15 | 6.3 % |
| 2mn-5mn | 7 | 2.9 % |
| 5mn-15mn | 4 | 1.6 % |
| 15mn-30mn | 2 | 0.8 % |
| Unknown | 1 | 0.4 % |

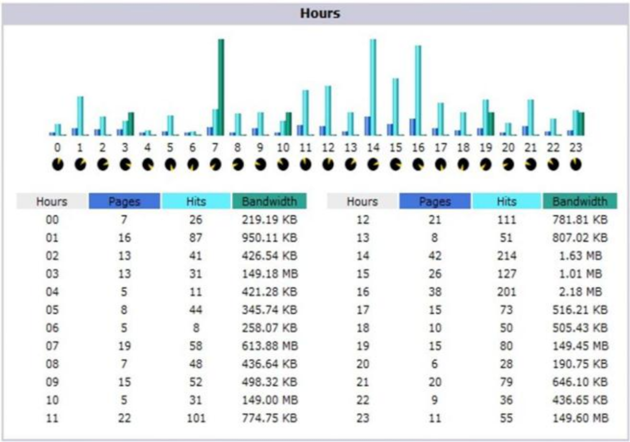
Downloads (Top 10) - Full list

| Downloads: 4 | Hits | 206 Hits | Bandwidth | Average size |
|---|---|---|---|---|
| /robots.txt | 379 | 0 | 315.39 KB | 852 Bytes |
| /files/Professionnel/CV%2018-04-2007.pdf | 18 | 8 | 2.71 MB | 106.88 KB |
| /phpMyAdmin/Documentation.txt | 11 | 0 | 676.23 KB | 61.48 KB |
| /files/MS%20Office%20Project%20Pro%202003%20with%20key.zip | 5 | 9 | 976.37 MB | 69.74 MB |

Browsers (Top 10) - Full list/Versions

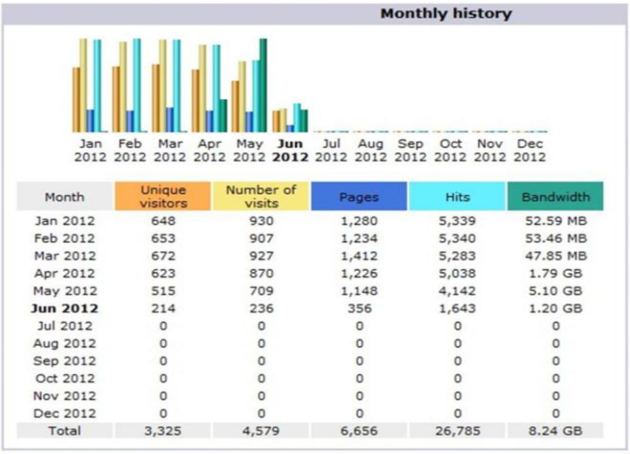
| Browsers | Hits | Percent |
|---|---|---|
| Firefox | 573 | 34.5 % |
| Google Chrome | 548 | 33 % |
| MS Internet Explorer | 414 | 24.9 % |
| Opera | 45 | 2.7 % |
| Safari | 27 | 1.6 % |
| Mozilla | 21 | 1.2 % |
| Netscape | 2 | 0.1 % |

The below graph shows the hourly distribution of load. This kind of information helps determine the peak hours of the day and the % increase in load during the peak hour. Having access to this information also helps in deciding off peak windows for scheduling batch and cron jobs during the day.



| Hours | Pages | Hits | Bandwidth | Hours | Pages | Hits | Bandwidth |
|---|---|---|---|---|---|---|---|
| 00 | 7 | 26 | 219.19 KB | 12 | 21 | 111 | 781.81 KB |
| 01 | 16 | 87 | 950.11 KB | 13 | 8 | 51 | 807.02 KB |
| 02 | 13 | 41 | 426.54 KB | 14 | 42 | 214 | 1.63 MB |
| 03 | 13 | 31 | 149.18 MB | 15 | 26 | 127 | 1.01 MB |
| 04 | 5 | 11 | 421.28 KB | 16 | 38 | 201 | 2.18 MB |
| 05 | 8 | 44 | 345.74 KB | 17 | 15 | 73 | 516.21 KB |
| 06 | 5 | 8 | 258.07 KB | 18 | 10 | 50 | 505.43 KB |
| 07 | 19 | 58 | 613.88 MB | 19 | 15 | 80 | 149.45 MB |
| 08 | 7 | 48 | 436.64 KB | 20 | 6 | 28 | 190.75 KB |
| 09 | 15 | 52 | 498.32 KB | 21 | 20 | 79 | 646.10 KB |
| 10 | 5 | 31 | 149.00 MB | 22 | 9 | 36 | 436.65 KB |
| 11 | 22 | 101 | 774.75 KB | 23 | 11 | 55 | 149.60 MB |

The below graph shows the distribution of load over a year. This information helps in understanding seasonal workloads if any experienced by the business and in turn the application.



| Month | Unique visitors | Number of visits | Pages | Hits | Bandwidth |
|---|---|---|---|---|---|
| Jan 2012 | 648 | 930 | 1,280 | 5,339 | 52.59 MB |
| Feb 2012 | 653 | 907 | 1,234 | 5,340 | 53.46 MB |
| Mar 2012 | 672 | 927 | 1,412 | 5,283 | 47.85 MB |
| Apr 2012 | 623 | 870 | 1,226 | 5,038 | 1.79 GB |
| May 2012 | 515 | 709 | 1,148 | 4,142 | 5.10 GB |
| **Jun 2012** | 214 | 236 | 356 | 1,643 | 1.20 GB |
| Jul 2012 | 0 | 0 | 0 | 0 | 0 |
| Aug 2012 | 0 | 0 | 0 | 0 | 0 |
| Sep 2012 | 0 | 0 | 0 | 0 | 0 |
| Oct 2012 | 0 | 0 | 0 | 0 | 0 |
| Nov 2012 | 0 | 0 | 0 | 0 | 0 |
| Dec 2012 | 0 | 0 | 0 | 0 | 0 |
| Total | 3,325 | 4,579 | 6,656 | 26,785 | 8.24 GB |

## 2. DEFINING SCOPE:

Consider high volume, Complex design, Business Impact, Resource Intensive, Seasonal peaks. With this gathered Information define the categorization and target for combined use case volumes in each category and test the high volume use case in each category.

### 1.3. Categorization

•One other premise that can go a long way in maximizing the code coverage of performance testing efforts and de-risking the system is categorization.

•Several enterprise transactions can be classified as variations or flavors of one base transaction. Even though there will be slight variation in input parameters, the backend tables and the data access objects will be the same. For instance, a customer updating his phone number vs. updating his address in the profile. Even though both transactions start out differently, they essentially perform an UPDATE on the profile table, and one can be termed as a flavor of the other.

•Along similar lines, the transaction could be the same but coming in from different sources. E.g. a request for account creation could come in from the web, from an agency or from customer service agents over the phone. However different the sources, the execution flow in all cases would involve a call to the same WebService and would end in an INSERT in the accounts and related reference tables.

•Once you have identified sets of similar transactions, combine the volumes of each; select the transaction with the highest volume to represent the set; and load test it to the combined volume.

•This approach covers wider grounds while limiting the effort involved in preparing and maintaining test frameworks for each transaction.

Most complex enterprise applications today are heavily data dependent. A simple example of such a transaction would be funds transfer in a bank account. To complete this transaction, there is a pre-requisite of having enough funds in a source account. If we keep executing this transaction over a set of accounts, the data will need to be refreshed either by using a different set of accounts or by changing the available balance on existing accounts.

To make it more complex, there are systems like Service Request Management Systems that are designed around flow of data from one stage to the other. Performance testing such systems becomes a nightmare because one successful execution of tests requires useful data to be created at each stage and the entire cycle repeated for the next run.

This added complexity introduces another factor which is the Return on Investment i.e. whether the effort involved in preparing for and maintaining a test case from one run to the other is worth the benefit from testing it.

In essence, it cannot be just one factor that can sufficiently determine the transaction set but it has to be a combination of all. Whatever the selection process, the choices are influenced by aggressive delivery schedules and there is always a trade-off.

## 3. CREATING WORKLOAD MODEL:

Factors to be considered are growth rates, transactional distribution, complex transactions.

When defining targets it is important to account for growth rate. Non-functional requirements are usually defined in the initial stages of the project. By the time the solution goes to production business volumes grow considerably. The targets defined for performance testing should be raised by the growth rate factor up to the roll out dates.

For a simple system where most transactions take the same amount of time to complete, the conversion from throughput to concurrency and vice-versa can be generalized to a simple formula:

$T=C/(tt+rt)$

Where T is the throughput (tps) in page views per second

C is the concurrency

tt is the think time between pages in seconds

And rt is the Response time of each page in seconds

However, in case of complex longer transactions the workload model has to be worked out differently.

Let us take an example of a generic Core Banking Application. A core banking solution will comprise of several modules that cater to Teller Banking, Online/Net Banking, Tele Banking, Mobile Banking, Customer Service etc.

All these modules function as different entry points into the system. Despite the different interfaces and web layers, they will all access the same backend services, data objects and database tables. So if we were to define scope and create a workload model for this application, we will have to look at the architecture on a whole by considering requirements of individual modules and how they interact with each other and the external interfaces.

Unlike functional testing, performance testing efforts have to be limited to only a select number of transactions. Before deriving a workload model, we have to first select transactions that form the scope of performance testing within each module.

For simplicity, let us work with three modules of our core banking application- Teller Banking, Online Banking and Phone Banking.

Functionally, there are a total of 22 use cases arising from these modules as listed in the table above. For a business, the ideal risk-free scenario is to performance test all 22 use cases. However, the effort involved in creating a load test framework for 22 Use cases and maintaining it across builds and releases can be a very challenging and time consuming activity. Projects seldom have the resources and the time to support the ask. Moreover, the benefit from load testing every use case is usually not worth the effort involved.

So we need to draw a line at a certain throughput, i.e. define a threshold below which a use case will not be considered for load testing. Use cases highlighted with green in the table above are transactions chosen on account of their high volumes.

Now that we have defined the scope, we will derive a workload using the requirements and data available. In most enterprise applications, the requirements are a combination of volume-centric and user-centric targets, i.e. module level concurrency and business volumes targets for every use case. For instance, in our example of the Banking application, its easy to know how many bank tellers will be using the core banking application, how many customer service agents will be working on the customer service module and so on. Assuming that we have statistics on transaction volumes from say the previous year, using simple mathematical logics, we can derive a workload model. But first lets define some variables:

Total application concurrency – C

Concurrency of a module y – Cy

Total number of modules in the application – m

Therefore, C = C1 + C2 + ….. + Cm. For sake of simplicity, lets represent it by SUM[C1:Cm]

Now lets get into distribution within a module. Lets say the total number of transactions in the module y is n. Consider a transaction x in the module y. Lets say the target volume of x is Vx per hour and the length of x is Lx.

Its important to note that the target transaction volumes should be of the time of the rollout. So, if the requirements were defined in 2016, the application goes live in 2017 and the growth rate is 10% then the target volumes for performance testing should be 120% of the 2016 volumes.

Since each transaction has its own length, i.e. a different number of pages, it is important to first translate business volumes into page views and then go over distribution. Hence, the target page views per second, i.e. Tx = Vx * Lx

User distribution i.e. the distribution of the module level concurrency amongst its transactions or use cases will be a function of the target page views Tx.

Therefore, concurrency of a transaction x in a module y i.e.

Cx = ROUND ( Tx / SUM[T1:Tn] ) * Cy

The excel above is a sample workload model for our example. Please note that the values are mere assumptions and in no way represent the actual volumes of bank.

The information at hand was the distribution of a concurrency of 604 users across the three modules, Teller Banking, Online Banking and Phone Banking. Also known were the target volumes for each of the shortlisted use cases. A study of the use case navigation and call flow helped determine the length (number of pages) of each use case. Applying the above formulae over the given information, targeted throughout (Page Views per second) per use case and a concurrency distribution within each module was calculated.

Once created, it is important to get a sign-off on a workload model before starting execution. This ensures that the requirements set forth for the Performance testing exercise are correct and validates the assumptions made.

Since a workload model relates more to the business, it is important to represent the information well. A pictorial representation of information is more likely to be well noticed and understood when compared to an excel containing a whole lot of numbers.

## 4. WORKLOAD VALIDATION:

In order to redesign the complete workload model it is recommended to do an early validation such as reverse calculation, Think time between pages(TT), Avg response time for each page(RT), time to complete execution x-, Achieved throughput

Requirement analysis, market research and solution design are based on a series of assumptions and it is important to ensure that the assumptions are correct by validating that the goals are achievable. This validation can be done without having to execute the load tests, just by doing some reverse calculations.

For example, lets assume that the average Think time between pages i.e. TT is set at an average of 10Secs and the Response Time target for each web page i.e. RT is 4Secs.

Hence the throughput of a business transaction x that can be achieved by the derived Concurrency Cx is

Vx = Cx * 3600 / (Lx * (TT+RT))

where Lx is the length of x i.e. number of pages. If the achieved Vx is in line with the targeted business transaction volumes then it is safe to say that the assumption of think times and the response time requirements are correct.

Validation can also be done post-execution at either the front-end or the back-end. At the front end, there are load generation tools that report counts of execution of transactions under test. Lets take the example of the IBM Rational Performance Tester load test tool. In the test report as one of the metrics, you can see the number of hits made to each page in the test suite. This number is a count of how many transactions were successfully completed on the system.

From the back end, post every test run a simple query on the database can give a count of volumes achieved during a test run.

## 5. THE BIRDS EYE VIEW:

For Performance Testing to reveal accurate characteristics of a system, the workload model should be a close representation of real world production load pattern. For complex enterprise applications user interface is just one entry point into the system. There are several other interfaces, WebServices scheduled jobs etc that share the system resources. To simulate a real world production load pattern it is essential to look at the complete picture and account for at least incoming load from all possible sources.

With the increasing complexity of business models and interdependence on business partners and service providers, interaction with external subsystems through interfaces and exposed WebServices and messaging interfaces is one primary source of incoming load. Other sources are inter-module communications between modules under test and those that are out of scope of the Performance test exercise.

Another activity to account for is the daily Batch jobs and schedulers that run during the regular business hours. For those that run during off-peak hours, its important to test and ensure that the execution of all scheduled batch jobs complete during the designated window and do not overflow into the regular business hours. Along similar lines, there are regular backup and archival activities that need to be allocated resources.

One other consideration that needs to go into a completing a workload is the recurring business activities that take place over and above the regular tasks. For example Close-of-Business, End-of-Month reporting, Quarterly reports etc.

## 6. SEASONAL WORKLOAD MODELS:

These models are business critical.There are a few domains that every so often, experience a substantial variation in their load pattern. These are called seasonal workloads. For applications that cater to these domains, ensuring performance and stability during such seasonal workloads also becomes the responsibility of the performance test exercise. Some examples of such seasonal workloads are:

•eCommerce Applications for Retailers: End of Season Sales, Holidays like Christmas and Thanksgiving

•Banking and Financial Applications: End of Year Closing

•Job Portals: Graduation Period

•Human Resource Management Systems: Appraisals etc

## 7. SELECTIVE MODELLING – RISK ANALYSIS:

There is always some amount of risk involved with selective modeling. Some transaction, some piece of code, SQL, stored procedure etc always rolls out without being performance tested.

An untested transaction can consume excessive system resources, starving other transactions of computational resources and causing a delay in overall system responses, or in the worst case scenario, crash the system.

However small, this risk associated with selective modeling can raise several flags if it has the potential to cause loss of revenue for the business. Because it is highly impractical to load test every transaction, a mitigation strategy needs to be defined.

There is no one thing that can be done to ensure that the system is risk free from performance problems. Several efforts have to run in parallel to cover maximum ground.

•Use Functional tests, UAT and System tests to detect bad transactions

•Monitor servers during UAT and Functional tests

•Load the test environments with near-production volume data

•Analyze offline reports from test servers for any abnormal system usage

•Plan one round of Performance testing with UAT or Functional tests running in parallel on the same environment

**Tips:**

These tips are some lessons that have been learnt from requirement gathering processes with several customer and hence are generic and applicable to all domains like banking, insurance, retail, telecommunication etc:

•Make sure you have a complete understanding of how the business that is being served by the application. What major functionalities does it cater to and what external systems does it interact with. Try to relate that to the solution design

•If and when possible, visit the business on-site to understand the system usage and study the load patterns

•Always set targets at peaks and not the average volumes

•Account for growth rates by targeting the volumes projected for the rollout timeline

•For a new system with no existing data, derive the data volumes. During execution, load test with databases holding at least near-production volume of data

•Ensure that there is room for server maintenance activities at average loads

•Last but the most important, get a sign-off on the requirements set forth for performance before starting execution

## CONCLUSION:

In this session we have gone over the process of gathering and defining requirements for performance testing of enterprise applications. We have seen how workload models can be derived for simple as well as complex use cases using the data available from various sources on projects.

We listed some factors that can help in defining the scope of performance testing activities, the risks involved and possible mitigations for addressing business concerns arising from not performance testing all transactions.

In conclusion, there is no one defined method for creating a comprehensive workload model. The selection process has to be a factor of business priorities, application complexity and project timelines. While there is always some amount of risk involved with performance testing over selective modeling, a lot can be done to mitigate or minimize the possible impact on business.

## REFERENCES:

- **AWStats**
- **Google Analytics**

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 19-22, 2017, ISSN:-2319–8656

# Performance Lifecycle in Banking Domain

Vijay Datla

**Abstract**: Performance assurance and testing plays a key role in complex applications and is an essential element of the application development life cycle. This case study is about integrating performance at a large national bank. Learn how custom monitoring, Six Sigma techniques, performance testing, and daily production reports played an important role in identifying production issues. This paper illustrates and examines the challenges and successes of performance planning, testing, analysis, and optimization after the release of X Bank's CRM application.

**Keywords**: Performance, Lifecycle, Banking, Vijay Datla, Vijay, Datla

## 1. INTRODUCTION

Software Performance Lifecycle (SPL) is an approach that can be applied to all types of technologies and industries. This solution allows the true possibilities of the system and software under test to be examined. It allows for precise planning and budgeting. The SPL approach can begin at any stage of the Software Development Life Cycle (SDLC) and will mature as the wheel turns (or the life cycle progresses). However, the earlier performance is evaluated, the sooner design and architectural flaws can be addressed and the faster and cheaper the software development life cycle becomes. The wheel in this case is the development life cycle as a whole, not just one application release but all releases from the start of the application. The SPL steps include planning, testing, monitoring, analysis, tuning and optimization. These steps will be discussed in conjunction with this case study. The idea is to begin the SPL approach at any point on the wheel (or development life cycle). As the wheel turns the SPL approach will position itself to start earlier and earlier in the development life cycle for future release levels. In the example explained below, SPL started at the end of the first release of the application to be tested. Due to the late introduction, we ran into different issues and problems, but we jumped on and started the performance lifecycle. The introduction of the SPL approach will save significant time and money, while ensuring end user satisfaction. Our organization used these set of techniques and procedures and called it Software Performance Lifecycle (SPL)

## 2. CASE STUDY

This performance case study involves a major national bank with over 2,000 branches, fictionally named X Bank for this study. The bank was facing performance issues with various portions of their CRIVI application. They were experiencing high response times, degraded throughput, poor scaling properties, and other issues. This caused un-acceptance from their end users and customer base.

During the first round of implementing performance at X Bank, our responsibilities as consultants was to help elevate their performance issues. We were in charge of managing and executing the Software Performance Life Cycle, which included items such as planning, scripting, testing, analyzing, tuning, and managing the performance lab. As the application grew in size, so did the team. lt started with two Senior Performance Engineers and evolved to a Senior Performance Engineer, a Senior Application Developer, two Scripting Resources, Environment Team Resource, and a part-time Database Admin.

The production environment consisted of five ACS (Application Combined Servers) and one NT database server. The rollout plan for X Bank called for 500 branches every 6 months until reaching the goal of 2,000 branches.

The CRM application technologies consisted of an ASP front end on IIS Web servers, C++ middle tier on MO Series and an Oracle database.

## 2.1 Planning & Setup Phase

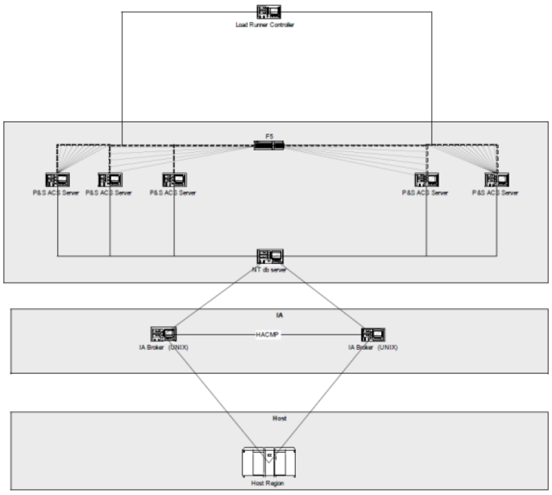
The first step in the SPL process is planning, this entails planning for the entire process, creating a performance test plan, and setting up the performance environment. The initial responsibilities of the two Performance Engineers was to interact with the bank resources, business analysts, developers, system administrators, database administrators, application engineers, and others to gather enough information to devise a performance test plan. To help create the performance test plan, we needed to fully understand the application behavior at X Bank. First, we sat down with business analysts to understand the major pain points and learn the application usage at the bank. We also made trips out to different branches and spoke to actual end users of the application to analyze their user experience and performance concerns.

Next, we needed to get a better understanding of the database volumes in production to allow us to properly populate the perfom1ance test lab database. To do this we received database row counts from the production database administrator for the previous three months, and we used that information in conjunction with the growth projections to appropriately populate the performance test lab database.

All the information gathered during the planning phase enabled us to get a better understanding and positioned us to create a performance test plan. The test plan included actual use case! business process steps, SLA goals, database sizing information, performance lab specifications, and exit criteria for this round.

Next we set out to create an onsite performance test lab. The test lab included load testing servers, application servers, a database server, an Integrated Architecture (IA) server, and a host system. The perfom1ance lab environment mimicked production in terms of the application servers but lacked in terms of number of CPUs on the IA Server. The load testing software of choice was LoadRunner and Win Runner. We utilized one load testing controller, two load generators, and two end user workstations. The workstations were used in conjunction with WinRunner to truly understand end user experience under load. Lastly a custom dashboard application was written to monitor application transaction response times at the web and application tiers and to provide server statistics information

The last step was to install the application on the servers and load the appropriate data volumes into the performance database server. We used Perl scripts to generate the appropriate volumes of test data, which gave us the flexibility to increase volume as needed.



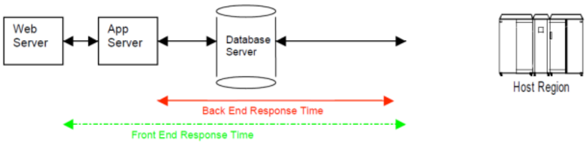
Above picture shows Performance Setup

## 2.2 Performance Testing & Monitoring Phase

At this point, we had devised a performance test plan, had an understanding of the performance concerns, and created a performance test lab. Now the next step was to begin performance testing and application system monitoring.

The first step was to create test scripts for all the outlined business process in our test plan using LoadRunner. The scripting process included recording the script, enhancing the script and running configuration audits. Enhancing the scripts included items such as parameterization, correlation, verification, logic, extra coding, and anything else that was required to mimic a real end user. Recording, enhancing, and single playbacks were all performed under the LoadRunner Vugen utility. The Vugen utility is LoadRunner's scripting engine. The next step was to run configuration audits, with multiple users, using the LoadRunner Controller for all defined test scripts. The LoadRunner Controller is the utility utilized to generate virtual user traffic. The reason for running configuration audits was to make sure issues, such as data dependencies and concurrency problems, did not arise in multiple user mode. After we completed the configuration audits, we executed load tests separately against each of the five AC8 servers. The separate tests were conducted to make sure all servers were behaving exactly the same in terms of response times, throughput, and utilization. Next we ramped up to two servers where we calculated BP (Business Process) throughput and compared them to our goals. Other testing goals, used for comparison, included concurrent virtual user mark, business process throughput for other business processes, and transaction response times. After the two AC5 server tests, we scaled up to five AC8 servers, and were not able to meet all our throughput goals.

The logs showed us which application transaction was being executed, along with specific information of each transaction. lt showed the back-end response time which is depicted as a solid line in the diagram below. The next entry in the log file was the front-end time which does not include GUI rendering time (shown as dotted line below). As the front end time includes the back-end times, the difference provided us with just the web server response time, giving us another data point for our analysis. The last two entries provided the exact request size and response size of each transaction, thus allowing us to verify the correct data sizes for the appropriate business processes.



## 2.3 Tuning & Optimization Phase

The next steps in the SPL include tuning and optimization of the application. Through an iterative testing cycle, which included testing, database resets, monitoring, and analysis we made significant changes to the application. These changes included items such as the login cache mechanism, which originally cached unnecessary content, and changes to a third-party party DLL that was affecting database queries. We also changed the application configuration file to start the correct number of application server instances, which in turn maximized the server resources.

## 2.4 Results

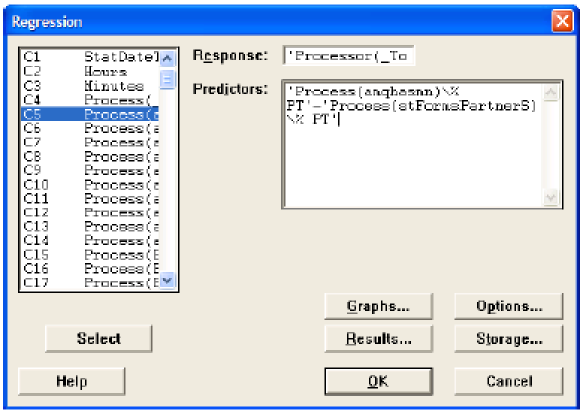
We measured the round trip application response times a several different ways. As we ran our tests we utilized LoadRunner to give us end-to-end response time, which did not include GUI rendering. So we utilized WinRunner to get true end-to-end response times which included GUI rendering time. The way we approached this was by running a full load with LoadRunner and placing two workstations on a emulated branch circuit running WinRunner. The WinRunner statistics provided response times that included the WAN emulation Of a remote branch circuit, as well as GUI rendering. This in tum provided us with a true user experience and end-to-end response times. Also to allow the business users to understand the feel of the application under load we had them walk through the application while we were running a full load test. The business users performed the business process steps from a lab that emulated remote branches. Lastly the business users provided response times from the branch circuit lab when no load was emulated on the system. This provided a true picture, meaning the best we can expect from our end user response times on system. The changes made allowed us to drastically reduce response times for most transactions while increasing the throughput of the application

Before we started SPL process at bank, they are facing performance issues at hundreds of branches.

Response times were high for key transactions, and marketing days were not acceptable by end users. X Bank did not truly understand the application scaling properties, capacity planning, hardware sizing, nor was the bank able to identify the performance issues they were having. But after the first pass through of the SPL process of planning, testing, monitoring, analysis, tuning, and optimization the bank was meeting SLAs, had proper hardware sizing in place, decreased costs, and gain confidence on marketing days.

## 3. PRODUCTION MONITORING

After we rolled all the changes into production and completed the first iteration of SPL at X Bank, we began monitoring production on a daily basis. The application production team provided daily server statistics for all of production. We used six sigma techniques such as regression analysis, processor capabilities charts, Xbar-S charts and boxplots to aid in our production monitoring. First, we performed daily regression analysis on server processes to isolate any top consuming processes. The regression analysis consists of gathering process information from all servers, which was supplemented by Perfmon, Minitab, and Perl. Perform is a windows monitoring solution, Minitab is a statistical computing system, and Perl, a scripting language, was used to parse and format Perfmon data to fit Minitab. Next, the formatted data was imported into Minitab and a Minitab worksheet was created. After the Minitab worksheet was created, a regression analysis was performed to find the top consuming processes [M|Nll]3]. For this analysis, Processor Total was the 'Resp ' processes, to be analyzed, were the 'Predictors' (x variable). See picture below.



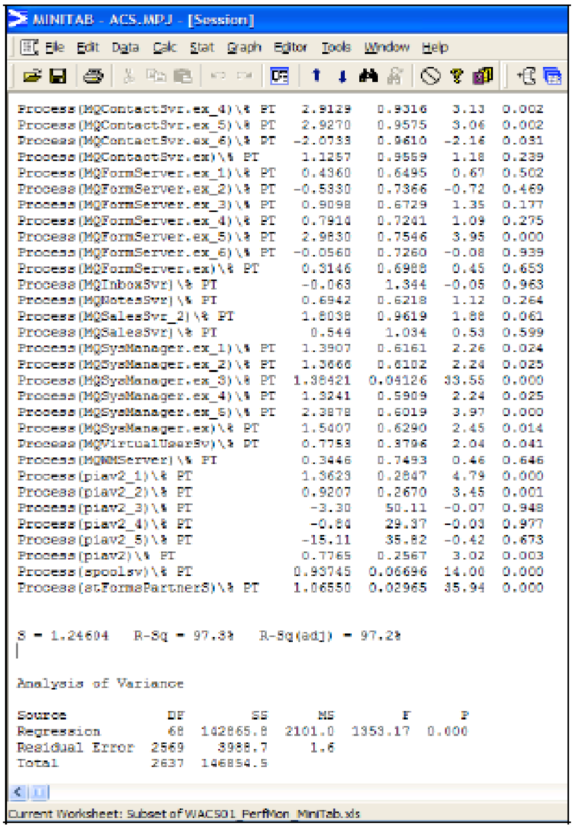
```
C1      StatDateT
C2      Hours
C3      Minutes
C4      Process(
C5      Process(a
C6      Process(a
C7      Process(a
C8      Process(a
C9      Process(a
C10     Process(a
C11     Process(a
C12     Process(a
C13     Process(a
C14     Process(a
C15     Process(B
C16     Process(B
C17     Process(B
```

Response: 'Processor(_To

Predictors: 'Process(anqhasnn)\%
PT'-'Process(stFormsPartnerS)
\% PT'

After indicating the response and predictors in Figure 1 the next step was to execute the regression analysis. The Output created from the regression analysis is shown in Figure 2. Figure 2 provides a value called R-Sq, the Higher the value of R-Sq the more relevant this data set is to our analysis. ln our example the R-Sq value was 9i'.2%, indicating our data having a really good fit to the model.

Below Picture shows Regression output, Processor vs Processes



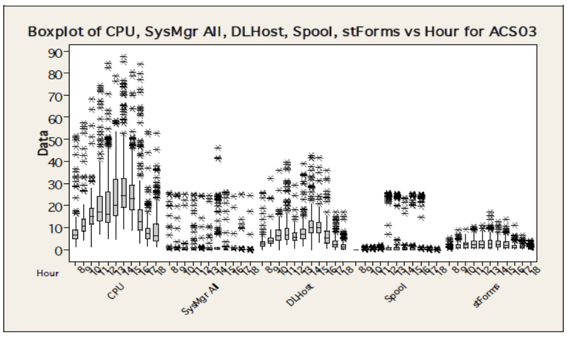
```
Process(MQContactSvr.ex_4)\% PT    2.9129   0.9316   3.13  0.002
Process(MQContactSvr.ex_5)\% PT    2.9270   0.9575   3.06  0.002
Process(MQContactSvr.ex_6)\% PT   -2.0733   0.9610  -2.16  0.031
Process(MQContactSvr.ex)\% PT      1.1257   0.9559   1.18  0.239
Process(MQFormServer.ex_1)\% PT    0.4360   0.6495   0.67  0.502
Process(MQFormServer.ex_2)\% PT   -0.5330   0.7366  -0.72  0.469
Process(MQFormServer.ex_3)\% PT    0.9098   0.6729   1.35  0.177
Process(MQFormServer.ex_4)\% PT    0.7914   0.7241   1.09  0.275
Process(MQFormServer.ex_5)\% PT    2.9830   0.7546   3.95  0.000
Process(MQFormServer.ex_6)\% PT   -0.0560   0.7260  -0.08  0.939
Process(MQFormServer.ex)\% PT      0.3146   0.6988   0.45  0.653
Process(MQInboxSvr)\% PT           -0.063    1.344  -0.05  0.963
Process(MQNotesSvr)\% PT           0.6942   0.6218   1.12  0.264
Process(MQSalesSvr_2)\% PT         1.8038   0.9619   1.88  0.061
Process(MQSalesSvr)\% PT            0.544    1.034   0.53  0.599
Process(MQSysManager.ex_1)\% PT    1.3907   0.6161   2.26  0.024
Process(MQSysManager.ex_2)\% PT    1.3668   0.6102   2.24  0.025
Process(MQSysManager.ex_3)\% PT   1.38421  0.04126  33.55  0.000
Process(MQSysManager.ex_4)\% PT    1.3241   0.5909   2.24  0.025
Process(MQSysManager.ex_5)\% PT    2.3878   0.6019   3.97  0.000
Process(MQSysManager.ex)\% PT      1.5407   0.6290   2.45  0.014
Process(MQVirtualUserSv)\% PT      0.7753   0.3796   2.04  0.041
Process(MQWMServer)\% PT           0.3446   0.7493   0.46  0.646
Process(piav2_1)\% PT              1.3623   0.2847   4.79  0.000
Process(piav2_2)\% PT              0.9207   0.2670   3.45  0.001
Process(piav2_3)\% PT               -3.30    50.11  -0.07  0.948
Process(piav2_4)\% PT               -0.84    29.37  -0.03  0.977
Process(piav2_5)\% PT              -15.11    35.82  -0.42  0.673
Process(piav2)\% PT                0.7765   0.2567   3.02  0.003
Process(spoolsv)\% PT             0.93745  0.06696  14.00  0.000
Process(stFormsPartnerS)\% PT     1.06550  0.02965  35.94  0.000


S = 1.24604   R-Sq = 97.3%   R-Sq(adj) = 97.2%


Analysis of Variance

Source            DF        SS      MS        F      P
Regression        68  142865.8  2101.0  1353.17  0.000
Residual Error  2569    3988.7     1.6
Total           2637  146854.5
```

Current Worksheet: Subset of WACS01_PerfMon_MiniTab.xls

After performing the above steps for all the production servers, the analysis provided us with a list of top processes of each production server. Below picture shows the high consuming processes. The higher T value the higher controlling factor the predictor, or the process.

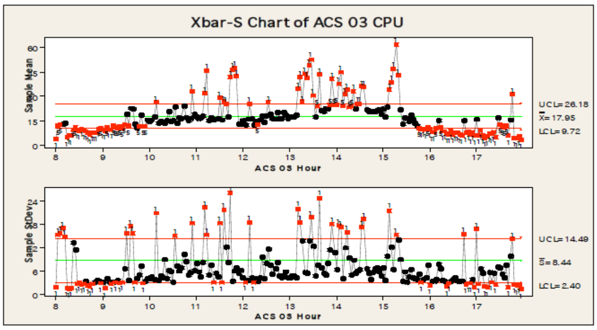
| ACS01 | | |
|---|---|---|
| **Predictor** | **T** | **P** |
| DLLHOST_1)\% PT | 89.62 | 0 |
| MQSysManager.ex_1)\% PT | 84.13 | 0 |
| MQSysManager.ex)\% PT | 73.96 | 0 |
| MQSysManager.ex_2)\% PT | 64.52 | 0 |
| MQSysManager.ex_3)\% PT | 63.25 | 0 |
| MQSysManager.ex_5)\% PT | 50.08 | 0 |
| stFormsPartnerS)\% PT | 38.54 | 0 |
| MQSysManager.ex_4)\% PT | 29.16 | 0 |
| MQBLServer)\% PT | 28.35 | 0 |
| cqmghost)\% PT | 11.53 | 0 |
| spoolsv)\% PT | 9.95 | 0 |
| Inetinfo)\% PT | 8.55 | 0 |
| LSASS)\% PT | 7.82 | 0 |
| piav2)\% PT | 7.8 | 0 |
| amqzlaa0)\% PT | 7.48 | 0 |
| piav2_2)\% PT | 7 | 0 |
| piav2_1)\% PT | 6.99 | 0 |

Below table shows list of top CPU consuming processes :

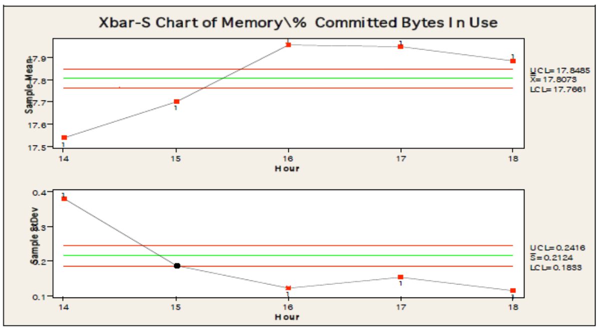


The CPU and memory graphs below displays a control chart for subgroup means (an X chart) and a control chart for subgroup standard deviations (an S chart) in the same graph window. The X chart is drawn in the upper half of the screen; the S chart in the lower half. Seeing both charts together allows you to track both the process level and process variation at the same time [M|Nll]3]. The x-axis of the graphs represents duration (time) while the y-axis represents sample mean or sample standard deviation. We primarily used this information to detect trends overtime. Minitab draws the average (center line), the upper control limit (UCL), and the lower control limit (LCL) lines by default.

CPU Processor control chart:

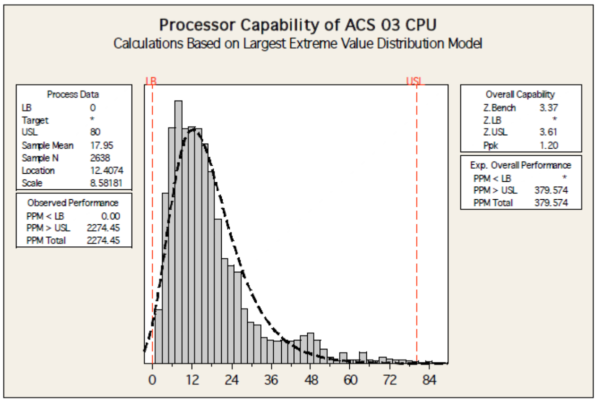


Memory Usage during peak



The processor capability graph [MINI03] below shows the processor distribution model around the CPU utilization
for the server on a given day. We used this information to see how many times (or parts per million, PPM) the data exceed our upper specifications limit (USL). In our case any data point outside the 80% USL mark is considered defective because the application degraded after the CPU utilization hit 80%. In this graph there are a few things to keep in mind, Left Boundary (LB), Upper Specification limit (USL), and parts per million (PPM). Ln Graph 1, PPM > USL is 2274, indicating that for every 1 million data points of CPU utilization we are following outside our acceptable range 2274 times.

Below Picture shows Processor capability:



| Process Data | |
|---|---|
| LB | 0 |
| Target | * |
| USL | 80 |
| Sample Mean | 17.95 |
| Sample N | 2638 |
| Location | 12.4074 |
| Scale | 8.58181 |

| Observed Performance | |
|---|---|
| PPM < LB | 0.00 |
| PPM > USL | 2274.45 |
| PPM Total | 2274.45 |

| Overall Capability | |
|---|---|
| Z.Bench | 3.37 |
| Z.LB | * |
| Z.USL | 3.61 |
| Ppk | 1.20 |

| Exp. Overall Performance | |
|---|---|
| PPM < LB | * |
| PPM > USL | 379.574 |
| PPM Total | 379.574 |

We used the above tables and graphs to create daily production reports. From the daily reports we created weekly and monthly trends to isolate any long-term problems. Specifically, it helped us isolate a few high processes that were behaving differently in production than in our test lab. The reason some the processes were behaving differently in production was due the fact that we could not test all business processes during this round of testing and only limited our testing to the business processes that produced 80% of the volume. The 80% was used based on the rule of thumb that20% of the transactions produce 80% of the volume. This allowed us to get a good indication of production volume without spending months scripting many different business processes. It seemed that transactions that were a part of the untested business processes were the top consuming processes to show up in production, and not in the performance lab. Due to our daily monitoring and reports, we were able to resolve these types of issues prior to any production downtime. ln conjunction, we utilized a custom dashboard application that provided response time information at the web and application layer for every transaction. It also provided real-time server and network statistics for all production servers. The dashboard alerted the application support team if any transaction or server network statistics breached the SLA thresholds.

## 3.1 Conclusion

Start the SPL approach at any stage of the SDLC. The bank faced performance issues in production, which caused un-acceptance from end users, bank personal, and the possible loss of future product upgrades. This could have cost millions of dollars in product revenue and maintenance licenses. But it did not, even though we started the SPL process of planning, testing, monitoring, analysis, tuning, and optimization after release l was in production. Since we were already on the wheel, we were able to include SPL earlier and earlier in the Software Development Life Cycle as the product grew, or as the wheel turned. The SPL approach saved significant time and money, ensured production readiness, improved performance and scalability, and built confidence.

## 4. REFERENCES

[1] MINITAB Statistical Software

# Effect of Information and Communication Technology-induced Multitasking on Academic Performance of University Students in Uganda

Peter Jegrace Jehopio
Department of Planning and Applied Statisticshhh
Makerere University
Uganda

Ronald Wesonga
Department of Planning and Applied Statistics
Makerere University
Uganda

Douglas Andabati Candia
Department of Planning and Applied Statistics
Makerere University
Uganda

**Abstract**: Numerous researches on information and communication technology (ICT)-induced multitasking among students document a number of unfavourable consequences, such as heightened distraction and less attention, hampered learning and hindered productivity at the expense of better academic performance. This study focused on the effect of information and communication technology induced multitasking on academic performance of university students in Uganda. To this end, primary data were collected during the month of May 2016 using stratified cluster sample design. A self-reported questionnaire was used to collect data from 312 students of Makerere University who participated in the study. Through structural equation modelling (SEM), it was demonstrated that ICT-induced multitasking does not affect academic performance directly but through self-regulation, attention span, emotional control and productivity focus. Nonetheless, multitasking does not always have negative consequences. To a majority of students, multitasking provides emotional satisfaction and enjoyment, which do correlate positively with good academic performance. Indeed, multitasking can be an effective use of time when well-regulated and an efficient tool in problem solving. Multitasking may only be indicative of the changing nature of norms. Traditionally, one was expected to give and receive undivided attention when talking in a face-to-face conversation with another; yet new norms are evolving for the networked society, such as responding to text messages promptly. To buffer the negative effect of ICT-induced multitasking on academic performance, one needs a facility with a good degree of self-regulation, attention span, emotional control and productivity focus.

**Keywords:** ICT; SEM; multitasking; academic performance; technology-induced; Uganda

## 1. INTRODUCTION

Students tend to multitask very often during learning activities[1]. Common multitasking activities during learning are social networking, surfing, chatting, texting, tweeting, downloading music and movies, listening to music, studying another lesson, e-mailing, video gaming, note-taking, eating, and drinking [2]. Research on information and communication technology (ICT)-induced multitasking among students documents a number of unlikeable outcomes, such as heightened distraction and less attention, hampered learning and hindered productivity at the expense of better academic performance [3-7] [8] [9, 10] [11, 12] [9, 13]. Nonetheless, other recent studies suggest that multitasking does not always have negative outcomes and may even have beneficial cognitive outcomes [14] [15].

The effect of information and communication technology induced multitasking on academic performance of university students in Uganda was investigated. To this end, through stratified cluster sample design, a self-reported questionnaire was used to collect data from 312 students of Makerere University. Moving structural equation modelling (SEM), it was demonstrated that ICT-induced multitasking does not affect academic performance directly but through self-regulation, attention span, emotional control and productivity focus.

### 1.1 Literature Review

Technology-induced multitasking and its damaging influence on academic performance have been widely studied [16] [17] [9] [13]. Further, research on information and communication technology (ICT)-induced multitasking among students documents a number of distasteful consequences, such as heightened distraction and less attention, hampered learning and hindered productivity at the expense of better academic performance [3] [4] [5] [6-9] [10] [9, 13]. With the ubiquity of cellular connection, text messaging, social media and the Internet, the modern multitasker is consistently engaged and always “on” at previously unimagined levels [18]. Studies show that a multitasking mind is one which is highly compromised: it juggles, divides, and sacrifices key mental faculties, often at the expense of proper information processing and encoding [19] [20]. Multitasking is known to impair attention [21]. Additionally, multitasking is often characterized by staying up late at night [22], which often positively correlates with lower levels of academic success [23].

Multitasking can be defined as being exposed to different information sources and switching between different media [24], which may be either sequential or concurrent based on the time spent on each task before switching to another. If the switching between the tasks is very short in duration (say, from attending a lecture being delivered to taking notes on the lecture), then that is concurrent multitasking. However, if the switches occur in longer durations (say, from attending a lecture being delivered to surfing the Internet), then that is sequential multitasking. It has been registered that individuals "engage in multitasking behaviour despite their metacognitive judgment about the performance costs [25] [26]. In contrast, [21]found that "self-regulated students were more likely to sustain their attention on classroom learning, and therefore less likely to text-message during class," i.e. self-regulated students are unlikely to multitask.

Students tend to multitask very often during learning activities [1]. Common multitasking activities during learning are social networking, surfing, chatting, texting, tweeting, downloading music and movies, listening to music, studying another lesson, e-mailing, video gaming, note-taking, eating, and drinking [2]. In related studies, [27] found that students switch tasks an average of 27 times per hour. [11] reported that students multitask 42 percent of class time. [28] found that 84 percent of college students engage in non-learning related media multitasking behaviours during lecture. Besides, [9] found that students seated near multitasking peers were consistently distracted and performed worse on retention measures compared to those sitting near students who were not multitasking.

Findings suggest that students' technology use is highly attributed to their anxiety without technology and dependency on technology, rather than any actual preference for multitasking [5] [29]. Apparently, the driving force behind multitasking is emotional rewards gained even at the cost of learning [26] [30] [26]. To this point, numerous studies have examined the relationship between anxiety and media multitasking [30]. Considering the documented value of social connection and social capital, this neurological dynamic may explain common research findings in which socially focused forms of multitasking and distraction, such as Facebook and Twitter, are often the most pervasive multitasking endeavour [31]. [32] noted that compulsive texting shares features with their compulsive Internet use given that both enable social interactions and have similar reasons for use, such as allowing for rapid text-based communication that promotes multitasking. An important conclusion from the study was that females would endorse greater frequency of texting compared to males. Indeed, [33]found that females do handle multitasking better than males. Also, [34] found that females were more susceptible to multitask compared to males and the female that engage in multitasking are more likely to have difficulties with academics [35].

Technology-induced multitasking resides within the construct of attentional control [13], and within the broader framework of self-regulation [36]. Attentional control is the ability to sustain deep and focused cognitive attention [10]. Even when students did not actively engage in multitasking, they reported that other students' laptops used in class were perceived as a distraction [37] [9]. Multitaskers are likely to give less attention to immediate, face-to-face communication because they are also thinking about their social network. A related concept is that of *poly-consciousness*, in which people's access to communication technologies can divide consciousness between immediate ("here and now") interaction settings and more distant settings, which undermines the immediate interaction conversation [38]. The implication of the foregoing discourse is that multi-tasking, divided attention, and the presence of a cell phone may interfere with one's ability to become acquainted with another.

At the same time, several recent studies suggest that multitasking does not always have negative outcomes and may even have beneficial cognitive outcomes [15]. For example, [39]found no significant correlation between media multitasking and a range of psychosocial well-being factors, including emotional positivity, sociability, and impulsivity. In other studies, even positive effects of media multitasking on well-being have been suggested. For example, interacting with family members while viewing television enhanced children's prosocial behaviour, and media multitasking was positively correlated with university students' emotional satisfaction, albeit at the cost of cognitive performance [40]. To be fair, multitasking is necessary for certain professions and is an indisputable phenomenon in education and life [13]. For example, [14] demonstrated that listening to a pleasant music while performing an academic test helped students to overcome stress, to devote more time to more stressful and more complicated task and the grades were higher. Multitasking can be an efficient use of time; a relatively manageable endeavour when necessary; or, when well monitored or well-regulated and, an effective tool in problem solving [41]. For example, multi-tasking can effectively provide a necessary avenue to interact with multiple others all at once in order to accomplish various goals [42]. In addition, certain people prefer to switch between multiple tasks within the same time block, and such "*polychronic-oriented*" individuals can be more satisfied with work that involves multi-tasking[43]. Furthermore, people who are hyper-connected generally report that they do not have problems attending to everyday tasks and inter-personal relationships [44]. It may be a question of changing nature of norms, traditionally people were expected to give and receive undivided attention when talking in face-to-face conversation with another, yet new norms are being developed for the networked society, such as responding to text messages promptly [45]. [46]concluded that students who multitask perform better academically.

In effect, contemporary students are described as digital natives (*homo zappiens*) and effective multitaskers. Digital natives [47] are individuals who are surrounded by digital technologies [48]. The ability to multitask across various multimedia environments is regarded as a significant characteristic of digital natives [49]. Other common features include effective communication, self-directed learning, and digital thinking [50] [47] [49]. Furthermore, some believe that the brains and cognitive capacity of those engaged in frequent multitasking will expand and adapt as a result of the behaviour, which may help them become ''nimble, quick-acting multitaskers''[51], who are able to manage signals from multiple sources at a time and are well prepared for careers in the information industry using technology.

Time management skill is an important aspect of behaviour for self-regulation, which involves setting goals, prioritizing, time estimation and problem solving [6] and as an intervening variable may explain the influence of multitasking on academic performance. If an individual has a good plan of what to do, he may not be distracted by other media activities. In addition, time management could buffer the negative effect of media multitasking. [52] found that many college students report that they were unable to go more than 10 minutes without checking their laptop, smartphone, tablet or e-reader. Many students pause in their learning activity to read and reply immediately to incoming text messages, or browse online while preparing homework [53, 54] . Research has demonstrated that students who use a laptop computer in the classroom report occasional email checking and frequent instant message sending and receiving. These students judged themselves to be less attentive during the lecture and to attain lower academic performance levels than other students [55{Golub, 2010 #23, 56]. It was also found that the self-assessments of students on failure to complete homework correlated significantly with their high usage of instant messaging software and specific types of multi-tasking activities [57] [31] [58]. Moreover, these behaviours interfered with schoolwork and was negatively related to overall college grade point average (GPA) performance [18] [59].

The debate regarding the effect of multitasking on academic performance has not yet come to a consensus [34] [60]. For instance, [30] observed 185 undergraduate students in three experimental conditions where learners were distracted with varying numbers of text messages. Findings showed that learning success decreased as the amount of texting increased. Another experimental study found that using mobile phones during lectures interfered with the learning gains of undergraduate students regardless of the degree of texting [61]. On the other hand, [62] designed a similar experiment with 120 university students where receiving instant messages or texting during video lectures did not have any effect on performance. Other studies have revealed a negative association between the frequency of multitasking in learning settings and GPA indicative of academic performance [63][[64][[11, 31, 65].

## 1.2 Conceptual Framework

Literature portrays that ICT-induced multitasking is replete among students [5] [29] [1] [2]. It has been pointed out that technology-induced multitasking resides within the construct of attentional control [13], and within the broader framework of self-regulation [36]. Time management skill is an important aspect of behaviour for self-regulation [6] [66] and, as an intervening variable, may explain the influence of multitasking on academic performance [67]. If an individual has a good plan of what to do, he may not be distracted by other media activities. In addition, time management could buffer the negative effect of media multitasking [67]. Further, ICT-induced multitasking among students documents a number of distasteful consequences including hindered productivity at the expense of better academic performance [3] [4] [5] [6].

It is therefore apparent that ICT-induced multitasking influences academic performance among university students; through self-regulation, attention span, emotional control and productivity focus, as depicted in the Figure 1 which follows.

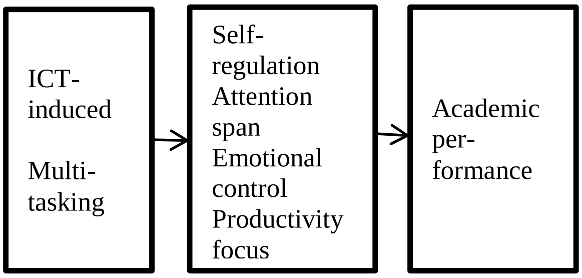


Figure 1: Conceptual framework for the effect of ICT-induced multitasking on student academic performance.

## 1.3 Study Objectives

The main objective of the study was to determine the effect that information and communication technology induced multitasking has on academic performance of university students in Uganda. Specifically, the study sought to investigate the following: the effect of ICT-induced multitasking on student attention; the influence of ICT-induced multitasking on student self-regulation; the consequence of ICT-induced multitasking on student productivity; how often students multitask during study session (say, during a one-hour lecture); whether ICT-induced multitasking is contagious among students; if student characteristics influence ICT-induced multitasking; whether ICT-induced multitasking impairs face-to-face interaction with others; and if ICT-induced multitasking is emotional-reward driven at the expense of better academic performance.

## 2. METHODS

To achieve the objectives of this study primary data were collected in May 2016 using stratified cluster sample design, through a self-reported questionnaire, from 312 Makerere University students. Students offering arts, sciences, male and female were targeted. Data were collected on the various characteristics under the constructs presented in the Figure 1 contained, with academic performance transformed into a binary outcome (good or poor).

## 3. RESULTS

Findings from the study are presented beginning with the characteristics of respondents, then the model for ICT-induced multitasking on academic performance.

### 3.1 Characteristics of Students who Engage in ICT-induced Multitasking

In the Table 1, a description of the characteristics of respondents of the study is made. From the Table 1 while

attending lecture, 62 percent of students multitask. Slightly, fewer female students (48%) multitask compared to their male counterparts (52%). Sciences-based majors (74%) do adversely multitask in comparison to their arts-based counterparts (27%). In the middle of working on an assignment, 70 percent of students multitask. Students who have ever stayed up late to multitask were 60 percent while those who have ever woken up early to multitask were 43 percent. Students who multitask and have more friends online than face-to-face were 52 percent. Up to 89 percent of students engage in multitasking upon seeing fellow students so doing.
On self-regulation, 66 percent of students report that they possess good time management skills. Seventy (70) percent report that they have a clear idea of what they want to accomplish during each upcoming week, but only 47 percent do make a list of what they have to do each day. While 74 percent of students often desist from multitasking so as to allow themselves focus on academic work, only 62 percent of the students have enough time to complete their assignments as thoroughly as they would like to. Up to 55 percent of students sometimes multitask without a specific goal.

Regarding emotional control: 50 percent of students have ever spent time even when advisable not to. Seventy three (73%) engage in multitasking to escape boredom and a similar percentage (74%) believe that multitasking provides them enjoyment. However, 45 percent of students have ever felt apprehensive about the much time that they spend multitasking.

With respects to productivity, only 49 percent of students reported that multitasking helps them to be more productive in their study time and only 62 percent of students who multitask have enough time to complete their assignments as flawlessly as they would like to.

Furthermore, during a one-hour lecture, on average, students multitask (switch on-and-off tasks) five (5) times with the longest attention span on the lecture being 41 minutes. The average number of minutes a student spends multitasking during a one-hour (60 minute) lecture is 13.

**Table 1: Characteristics of university students who engage in ICT-induced multitasking**

| **General Characteristics** | **Percentage** |
|---|---|
| 1. Students who multitask while attending lecture. | 62.1 |
| 2. Female students who multitask while attending lecture. | 47.8 |
| 3. Male students who multitask while attending lecture. | 52.2 |
| 4. Students of sciences-based major who multitask while attending lecture | 73.5 |
| 5. Students of arts-based major who multitask while attending lecture. | 26.5 |
| 6. Students who multitask in the middle of working on an assignment. | 70.3 |
| 7. Students who have ever stayed up late to multitask. | 59.9 |
| 8. Students who have ever woken up early to multitask. | 43.2 |
| 9. Students who report that multitasking distracts them from academic work. | 45.7 |
| 10. Students who multitask and have more friends online than face-to-face. | 51.7 |
| 11. Students who find themselves engaging in multitasking upon seeing fellow students so doing. | 88.7 |
| **Self-regulation** | |
| 12. Students who reported that they possess good time management skills. | 65.7 |
| 13. Students who have a clear idea of what they want to accomplish during each upcoming week. | 70.2 |
| 14. Students who make a list of things they have to do each day. | 47.0 |
| 15. Students who have enough time to complete their assignments as thoroughly as they would like to. | 62.2 |
| 16. Students who often desist from multitasking so as to allow themselves focus on academic work. | 74.1 |
| 17. Students who find it hard to resist multitasking. | 45.9 |
| 18. Students who consider multitasking to be to be a good study tool. | 52.1 |
| 19. Students who sometimes multitask without a specific goal. | 54.5 |
| **Emotional Control** | |

| | |
|---|---|
| 20. Students who have ever spent time multitasking even when advisable not to. | 49.7 |
| 21. Students who sometimes engage in multitasking to escape boredom. | 72.9 |
| 22. Students who believe multitasking provides them enjoyment. | 73.7 |
| 23. Students who become frustrated when conditions do not permit multitasking. | 55.0 |
| 24. Students who become irritable when conditions do not permit multitasking. | 49.3 |
| 25. Students who have ever engaged in multitasking even when they feel not to. | 54.1 |
| 26. Students who have ever felt apprehensive about the much time they spend multitasking. | 45.2 |
| **Productivity** | |
| 27. Students who report that multitasking helps them to be more productive in their study time. | 48.8 |
| 28. Students who have enough time to complete their assignments as thoroughly as they would like to. | 62.2 |
| **Attention span** | |
| 29. The number of times students multitask (switch on-and-off tasks) during a one-hour lecture. | 5.4 |
| 30. On average, the longest duration (in minutes) during a one-hour (60 minutes) lecture that a student can go without multitasking. | 40.7 |
| 31. Average number of minutes a student spends multitasking during a one-hour (60 minute) lecture. | 12.5 |

Consequent to the conceptual framework presented in the Figure 1, structural equation modelling was moved in order to concurrently study the indirect effect of ICT-induced multitasking (independent variable) on academic performance (dependent variable) through one’s self-regulation, productivity focus and attention span (intermediate variables). The model equations were then:

$$self_{reg} = \alpha_1 + \beta_1 multi + e_{self_{reg}}$$

$$atten_{span} = \alpha_2 + \beta_2 multi + e_{atten_{span}}$$

$$Emotn_{contol} = \alpha_3 + \beta_3 multi + e_{emotn_{control}}$$

$$prod = \alpha_4 + \beta_4 multi + e_{prod}$$

$$\begin{aligned} acad_{perf} = \alpha_5 + \beta_5 self_{reg} + \beta_6 atten_{span} \\ + \beta_7 emotn_{control} + \beta_8 prod \\ + e_{acad_{perf}} \end{aligned}$$

## 3.2 Structural Equation Model for Predictors of Academic Performance

Following from the conceptual framework presented in Figure 1, results of structural equation modelling are presented in Table 2.

**Table 2: Structural equation model of ICT-induced multitasking on academic performance**

| Potential Factors | | Coefficients | Odds Ratio | P>\|z\| |
|---|---|---|---|---|
| Self-regulation | <- | | | |
| | Multitasking | -0.853 | 0.426 | 0.004 |
| Attention span | <- | | | |
| | Multitasking | -7.112 | 0.001 | 0.006 |
| Productivity | <- | | | |
| | Multitasking | -8.717 | 0.001 | 0.005 |
| Emotional control | <- | | | |
| | Multitasking | -1.098 | 0.334 | 0.000 |
| Academic Performance | <- | | | |
| | Self-regulation | 0.342 | 1.408 | 0.213 |
| | Attention span | -0.004 | 0.996 | 0.514 |
| | Productivity | 0.002 | 1.002 | 0.772 |
| | Emotional control | 0.052 | 1.054 | 0.647 |

From the Table 2, ICT-induced multitasking is seen to negatively significantly ($p<0.05$) affect academic performance through self-regulation, attention span, productivity and emotional control.

On self-regulation, 0.426 decrease in the log-odds of self-regulation is expected for students who multitask during academic engagement compared to those who do not. Holding all other independent variables constant, an increase in academic performance of students who possess a higher degree of self-regulation is expected.

With regards to attention span, 0.001 decrease in the log-odds of attention span is expected for students who multitask compared to those who do not; holding all other independent variables constant consequently, resulting into a decrease in academic performance for students who multitask.

On productivity, notable 0.001 decrease in the log-odds of productivity is expected for students who multitask compared to those who do not. Holding all other independent variables constant, an increase in academic performance for students who focus on productivity but not multitasking is expected.

Regarding emotional control, 0.334 decrease in the log-odds of emotional control is expected for students who multitask compared to those who do not. Holding all other independent variables constant consequently there will be an increase in academic performance for students as their level of emotional control increases.

## 4. DISCUSSION

This study focused on the effect of information and communication technology induced multitasking on academic performance of university students in Uganda. Indeed, ICT-induced multitasking is replete among students [5] [29] [1] [2]. Findings of this study show that 62 percent of university students multitask while attending lectures. Regarding multitasking, on average, students switch tasks five (5) times during a one-hour lecture. To switch tasks five (5) times while attending a one-hour lecture is indeed to do so often; which is in tandem with observation by [1], that Students tend to multitask very often during learning activities. The study by [27] which demonstrated that 'students switch tasks an average of 27 times per hour,' does not specifically focus on the particular type of multitasking. This study, however, specifically focuses on ICT-induced multitasking while attending lecture. Also, observed in this study is that slightly fewer female students (48%) multitask compared to their male counterparts (52%). This may not necessarily mean that female students multitask less but may only be in the case of agreement with [33], who found that females do handle multitasking better than males.

Sciences-based majors (74%) do adversely multitask in comparison to their arts-based counterparts (27%). The large disparity may be because science-based majors are more apt to grow into digital natives in comparison to arts-based majors because the content of what sciences-based majors study is closely or is directly and practically related to ICT.

It was also noted that up to 89 percent of students engage in multitasking upon seeing fellow students so doing, which likely implies that ICT-induced multitasking is contagious. Furthermore, [59] found that students seated near multitasking peers were consistently distracted and performed worse on retention measures compared to those sitting near students who were not multitasking. Therefore, being physically close to a multitasking peer is likely to negatively affect ones academic performance.

Although 70 percent of students reported that they have a clear idea of what they want to accomplish during each upcoming week, only 47 percent do make a list of what they have to do each day, which likely implies that a number of students lack attentional control [13] and, therefore, self-regulation [36] which in turn negatively correlates with poor academic performance. Moreover, this study found that up to 55 percent of students sometimes multitask without a specific goal.

With respect to self-regulation, ICT-induced multitasking was observed to negatively affect academic performance, yet self-regulation involves setting goals, prioritizing, time estimation and problem solving [6] [66], which are significant for good academic performance. With regards to attention span, this study found out that, through attention span, multitasking negatively influences academic performance which also [21] observed. Furthermore, multitasking is often characterized by staying up late at night [22], which often positively correlates with lower levels of academic success [23]. With so much mentioned, obviously multitasking lowers productivity.

Nonetheless, multitasking does not always have negative consequences and may even have beneficial cognitive outcomes [15]. Indeed, this study found out that up to 74 of university students report that multitasking provides them enjoyment. Multitasking was noted to positively be correlated with university students' emotional satisfaction, albeit at the cost of cognitive performance [40]. To be fair, multitasking is necessary for certain professions and is an indisputable phenomenon in education and life [13]. Multitasking can be an efficient use of time; a relatively manageable endeavour when necessary; or, when well monitored or well-regulated, an effective tool in problem solving [41] [68]. Furthermore, people who are hyper-connected generally report that they do not have problems attending to everyday tasks and inter-personal relationships [44]. It may be a question of changing nature of norms, traditionally people were expected to give and receive undivided attention when talking in face-to-face conversation with another, yet new norms are being developed for the networked society, such as responding to text messages promptly [45, 69].

Up to 62 percent of university students multitask while attending lectures. On average, students switch tasks five (5) times during a one-hour lecture. Slightly fewer female students (48%) were noted to multitask during study time compared to their male counterparts (52%). Sciences-based majors (74%) do multitask more compared to their arts-based counterparts (27%). Moreover, up to 89 percent of students engage in multitasking upon seeing fellow students so doing which likely implies that ICT-induced multitasking is contagious.

Nonetheless, multitasking does not always have negative consequences and may even have beneficial cognitive outcomes. Indeed, this study found out that up to 74 of university students report that multitasking provides them enjoyment; which positively correlates with university students' emotional satisfaction, which is normally healthy for better academic performance. Indeed, multitasking can be an efficient use of time when well regulated and an effective tool in problem solving. Multitasking may only be indicative of the changing nature of norms. Traditionally people were expected to give and receive undivided attention when talking in face-to-face conversation with another; yet new norms are being developed for the networked society, such as responding to text messages promptly. To buffer the negative effect of ICT-induced multitasking on academic performance, institutions of higher learning need to provide an environment where students are monitored for a good degree of self-regulation, attention span, emotional control and productivity focus.

## 5. CONCLUSION

This study noted that ICT-induced multitasking does not affect academic performance directly but through self-regulation, attention span, emotional control and productivity focus.

Noting that ICT-induced multitasking affects academic performance through self-regulation, attention span, emotional control and productivity focus, to buffer the negative effect of ICT-induced multitasking on academic performance university students need be facilitated to possess a high degree of self-regulation, attention span, emotional control and productivity focus. Multitasking during lectures should specifically be discouraged, since ICT-induced multitasking tends to be contagious. Since multitasking does not always have a negative consequence, it may not be completely discouraged, given that many do derive emotional satisfaction through it. Provided it is properly regulated, multitasking is beneficial for better academic performance. Therefore, the future scope of the studies should focus on developing model frameworks that supports integration of ICT-induced multitasking that directly supports students' better academic performance.

## 6. REFERENCES

[1] C. Calderwood, P. L. Ackerman, and E. M. Conklin, "What else college students “do” while studying? an investigation of multitasking," *Computers in Human Behavior*, vol. 75, 2014.

[2] J. H. Kuznekoff and S. Titsworth, "The impact of mobile phone usage on student learning," *Communication Education*, vol. 62, pp. 233–252, 2013.

[3] A. Salomon and Y. B. Kolikant, "High-school students' perceptions of the effects of non-academic usage of ICT on their academic achievements," *Computers in Human Behavior*, vol. 64, p. 143e151, 2016.

[4] A. Kononova, E. Joo, and S. Yuan, "If I choose when to switch: Heavy multitaskers remember online content better than light multitaskers when they have the freedom to multitask," *Computers in Human Behavior*, vol. 65, p. 567e575, 2016.

[5] C. A. Terry, P. Mishra, and C. J. Roseth, "Preference for multitasking, technological dependency, student metacognition, & pervasive technology use: An experimental intervention," *Computers in Human Behavior* vol. 65, p. 24e251, 2016.

[6] Y. Zhang and P. P. Rau, "An Exploratory Study to Measure Excessive Involvement in Multitasking Interaction with Smart Devices," *Cyberpsychology, Behavior, and Social Networking*, vol. 19, 2016.

[7] A. Lepp, J. E. Barkley, and A. C. Karpinski, "The relationship between cell phone use, academic performance, anxiety, and satisfaction with life in college students," *Computers in Human Behavior*, vol. 31, 2014.

[8] T. Judd, "Making sense of multitasking," *The role of Facebook. Computers & Education*, vol. 70, pp. 194–202, 2014.

[9] F. Sana, T. Weston, and N. J. Cepeda, "Laptop multitasking hinders classroom learning for both users and nearby peers," *Computers & Education*, vol. 62, 2013.

[10] N. Unsworth, B. D. McMillan, G. Brewer, and G. J. Spillers, "Everyday attention failures: An individual differences investigation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 38, p. 1765e1772, 2012.

[11] J. M. Kraushaar and D. C. Novak, "Examining the affects of student multitasking with laptops during the lecture," *Journal of Information Systems Education*, vol. 21, p. 241e251, 2010.

[12] L. D. Rosen, A. F. Lim, L. M. Carrier, and N. A. Cheever, "An empirical examination of the educational impact of text message-Induced task switching in the classroom: Educational implications and strategies to enhance learning," *Psicologia Educativa*, vol. 17, p. 163e177, 2011.

[13] E. Wood and L. Zivcakova, "Understanding multimedia multitasking in educational settings," *In L. D. Rosen, N. A. Cheever, & L. Mark Carrier (Eds.), Wiley handbook of psychology, technology, and society*, p. 404e453, 2015.

[14] A. Cabanac, L. Perlovsky, and M. Bonniot-Cabanac, "Music and academic performance," *Behavioural Brain Research*, vol. 256, pp. 257–260, 2013.

[15] J. Winter, D. Cotton, J. Gavin, and J. D. Yorke, "Effective e-learning? Multi-tasking, distractions and boundary management by graduate students in an online environment," 2010.

[16] H. Kauffman, "A review of predictive factors of student success in and satisfaction with online learning," *Research in Learning Technology,* vol. 2015, p. 26507, 2015.

[17] H. Hembrooke and G. Gay, "The laptop and the lecture: The effects of multitasking in learning environments," *Journal of computing in higher education,* vol. 15, p. 46e64, 2003.

[18] S. Bellur, K. L. Nowak, and K. S. Hull, "Make it our time: In class multitaskers have lower academic performance," *Computers in Human Behavior,* vol. 53, 2015.

[19] G. Bozeday, "Media multitasking and the student brain," *School Specialty,* 2013.

[20] T. F. Heatherton and D. D. Wagner, "Cognitive neuroscience of self-regulation failure," *Trends in Cognitive Sciences,* vol. 15, p. 132e139, 2011.

[21] F. Y. F. Wei, Y. K. Wang, and M. Klausner, "Rethinking college students' self-regulation and sustained attention: Does text messaging during class influence cognitive learning?," *Communication Education,* vol. 61, p. 185e204, 2012.

[22] W. A. Austin and M. W. Totaro, "Gender differences in the effects of Internet usage on high school absenteeism," *The Journal of Socio-Economics,* vol. 40, 2011.

[23] H. C. Tsai and S. H. Liu, "Relationships between time-management skills, Facebook interpersonal skills and academic achievement among junior high school students," *Social Psychology of Education,* p. 1e14, 2015.

[24] E. Ophir, C. Nass, and A. D. Wagner, "Cognitive control in media multitaskers," *PNAS,* vol. 1e5, 2009.

[25] J. R. Finley, A. S. Benjamin, and J. S. McCarley, "Metacognition of multitasking: How well do we predict the costs of divided attention? ," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. Applied, 20, p. 158e165, 2014.

[26] X. Wang, "Excelling in multitasking and enjoying the distraction: Predicting intentions to send or read text messages while driving," *Computers in Human Behavior,* vol. 64, p. 584e590, 2016.

[27] C. Marci, "A (biometric) day in the life: Engaging across media," *Paper presented at Re: Think 2012, New York, NY,* 2012, March.

[28] Y. Fan, S. Gong, Y. Wang, and Z. Wang, "Advances in Psychology," vol. 6, pp. 914-922, 2016.

[29] K. Sim, N.; and S. Stein, "Reaching the unreached: de-mystifying the role of ICT in the process of doctoral research," *Research in Learning Technology* vol. 2016, p. 30717 2016.

[30] L. D. Rosen, M. L. Carrier, and N. A. Cheever, "Facebook and texting made me do it: Media-induced task-switching while studying," *Computers in Human Behavior,* vol. 29, p. 948e958, 2013.

[31] R. Junco and S. R. Cotten, "No A 4 U: the relationship between multitasking and academic performance," *Computers & Education,* vol. 59, p. 505e514, 2012.

[32] K. M. Lister-Landman, S. E. Domoff, and E. F. Dubow, "The Role of Compulsive Texting in Adolescents' Academic Functioning," *Psychology of Popular Media Culture. Advance online publication,* 2015.

[33] G. Stoet, D. B. O'Connor, M. Conner, and K. R. Laws, "Are women better than men at multi-tasking? ," *BMC Psychology,* p. 1:18, 2013.

[34] S. Xu, Z. Wang, and P. David, "Media multitasking and well-being of university students," *Computers in Human Behavior,* vol. 55, p. 242e250, 2016.

[35] F. Y. Hong, S. I. Chiu, and D. H. Hong, "A model of the relationship between psychological characteristics, mobile phone addiction and use of mobile phones by Taiwanese university female students," *Computers in Human Behavior,* vol. 28, pp. 2152–2159, 2012.

[36] M. S. Hagger, C. Wood, C. Stiff, and N. L. D. Chatzisarantis, "Ego depletion and the strength model of self-control," *A meta-analysis. Psychological Bulletin,* vol. 136, p. 495e525, 2010.

[37] C. B. Fried, "In-class laptop use and its effects on student learning," *Computers & Education,* vol. 50, p. 906e914, 2008.

[38] S. Misra, L. Cheng, J. Genevie, and M. Yuan, "The iPhone effect: the quality of in-person social interactions in the presence of mobile devices " *Environment and Behavior. Advance online publication,* 2014.

[39] S.-I. Shih, "A null relationship between media multitasking and well-being," *PLoS One,* vol. 8, p. e64508, 2013.

[40] Z. Wang and J. M. Tchernev, "The "myth" of media multitasking: Reciprocal dynamics of media multitasking, personal needs, and gratifications," *Journal of Communication,* vol. 62, p. 493e513, 2012.

[41] S. A. Brasel and J. Gips, "Media multitasking behavior: Concurrent television and computer usage," *Cyberpsychology, Behavior and Social Networking,* vol. 14, p. 527e534, 2011.

[42] P. David, L. Xu, J. Srivastava, and J. Kim, "Media multitasking between two conversational tasks," *Computers in Human Behavior,* vol. 29, 2013.

[43] A. Arndt, T. J. Arnold, and T. D. Landry, "The effects of polychromic-orientation upon retail employee satisfaction and turnover," *Journal of Retailing,* vol. 82, 2006.

[44] A. Smith, "The best (and worst) of mobile connectivity," *Washington, DC: Pew Research Center,* 2012.

[45] D. K. Forgays, I. Hyman, and J. Schreiber, "Texting everywhere for everything: gender and age differences in cellphone etiquette and use " *Computers in Human Behavior,* vol. 31, 2014.

[46] J. L. Badge, N. F. W. Saunders, and A. J. Cann, "Beyond marks: new tools to visualise student engagement via social networks," *Research in Learning Technology,* vol. 20, 2012.

[47] M. Prensky, "Digital natives, digital immigrants, part 1," *On The Horizon,* vol. 9, 2001.

[48] T. Cochrane, L. Antonczak, H. Keegan, and V. Narayan, "Riding the wave of BYOD: developing a framework for creative pedagogies," *Research in Learning Technology,* vol. 2014, p. 24637 2014.

[49] W. Veen and B. Vrakking, "Homo zappiens: Growing up in a digital age," *London, UK: Network Continuum Education,* 2006.

[50] C. Cronin, T. Cochrane, and A. Gordon, "Nurturing global collaboration and networked learning in higher education," *Research in Learning Technology,* vol. 2016, p. 26497, 2016.

[51] Q. J. Anderson and L. Rainie, "Millennials will benefit and suffer due to their hyperconnected lives," *Pew Internet & American Life Project,* 2012.

[52] S. Kessler, "38% of college students can't go 10 minutes without tech [STATS]," *Mashable Tech,* 2011.

[53] W. Barrat, M. Hendrickson, A. Stephens, and J. Torres, "The Facebook: Computer mediated social networking," *Student Affairs Online,* vol. 6, p. 1e5, 2005.

[54] L. E. Levine, B. M. Waite, and L. L. Bowman, "Electronic media use, reading, and academic distractibility in college youth," *Cyberpsychology & Behavior,* vol. 10, p. 560e566, 2007.

[55] M. Duggan and L. Rainie, "Cell phone activities 2012," *Pew Internet & American Life Project,* 2012.

[56] T. L. Golub and M. Miloloza, "Facebook, academic performance, multitasking and self-esteem," *In 10th special focus symposium on ICESKS: Information, communication and economic sciences in the knowledge society,* 2010.

[57] M. A. Moreno, L. Jelenchick, R. Koff, J. Eikoff, C. Diermyer, and D. A. Christakis, "Internet use and multitasking among older adolescents: An experience sampling approach," *Computers in Human Behavior*, vol. 28, p. 1097e1102, 2012.

[58] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educational Psychologist,* vol. 38, p. 43e52, 2003.

[59] D. K. Wentworth and J. H. Middleton, "Technology use and academic performance," *Computers & Education,* vol. 78, p. 306e311, 2014.

[60] M. L. Courage, "Translational science and multitasking: Lessons from the lab for the everyday world," *Developmental Review* vol. 35 pp. 1–4, 2016.

[61] Y. Ellis, B. Daniels, and A. Jauregui, "The effect of multitasking on the grade performance of business students," *Research in Higher Education Journal*, vol. 8, 2010.

[62] D. R. Lawson, "The effects of text messaging on memory recall in college Learning and Instruction students," *North Carolina: Western Carolina University. Unpublished PhD dissertation,* vol. 41, p. 94e105, 2013.

[63] S. M. Ravizza, D. Z. Hambrick, and K. M. Fenn, "Non-academic internet use in the classroom is negatively related to classroom learning regardless of intellectual ability," *Computers & Education,* vol. 78, 2014.

[64] L. Burak, "Multitasking in the university classroom," *International Journal for the Scholarship of Teaching And Learning,* vol. 6, 2012.

[65] W. C. Jacobsen and R. Forste, "The wired generation: academic and social outcomes of electronic media use among university students," *Cyberpsychology, Behavior, and Social Networking,* vol. 14, p. 275e280, 2011.

[66] P. R. Pintrich, "Multiple goals, multiple pathways: The role of goal orientation in learning and achievement," *Journal of Educational Psychology,* vol. 92, p. 544, 2000.

[67] X. Yang, X. Xu, and L. Zhu, "Media multitasking and psychological wellbeing in Chinese adolescents: Time management as a moderator," *Computers in Human Behavior,* vol. 53, pp. 216–222, 2015.

[68] L. Lin, "Breadth-biased versus focused cognitive control in media multitasking behaviors," *Proceedings of the National Academy of Sciences*, vol. 106, p. 15521e15522, 2009.

[69] L. Rainie and B. Wellman, "Networked: The new operating system," *The MIT Press,* 2012.

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 32-37, 2017, ISSN:-2319–8656

# Performance Forecast of DB Disk Space

Srikanth Kumar Tippabhotla
CTS

**Abstract**: ln the absence of special purpose monitoring and/or modeling software designed specifically for forecasting database disk space requirements, a solution was developed using general purpose database facilities and office suite products. The results achieved were (1) an understanding of the heretofore unknown trends and patterns in the use of disk space by individual databases (2) the ability to accurately and proactively forecast the additional disk space needed for individual databases, and (3) the ability to reclaim the forecast unused disk space, all based upon linear regression analyses.

**Keywords**: Performance; DB

## 1. INTRODUCTION

As the sheer number, size, and complexity of databases deployed in organizations continues to climb each year, the effective management of those instances becomes a challenge for the IT staff, and in particular for those charged with maintaining database integrity, availability, performance, and recoverability. Practices rooted in reactive and frequently eleventh-hour individual heroics no longer make the grade in today's business environment. Rather, one needs to create and adopt repeatable, proactive methodologies in order to provide appropriate IT service delivery.

The database management systems (DBMS) utilized consisted of:

- Oracle Server (Oracle)
- Sybase Adaptive Sen/er Enterprise (Sybase)
- IBM DB2/UDB (UDB)
- Microsoft SQL Server (SQL Server)

Operating system (OS) environments included:

- UNIX -- IBM AIX
- UNIX -- Sun Microsystems Solaris
- Microsoft Windows Server.

As is the case with most operational support groups an on-call pager rotation schedule placed a team member in the "hot seat" 24 hours a day for a one- week period. During this tour of duty, the on-call person would respond to pages that were programmatically generated by the combination of the BMC Patrol Monitor for Sybase/Oracle/UDB product (hereafter referred to as Patrol) and the Tivoli Enterprise Manager suite. Additional, manually-generated pages were also issued by IT staff members at the round-the-clock computer operations center.

## 2. METHODS

Before getting into the details of the system, let's lay the foundation for that discussion with some background information.

**Definition of Terms and Database Concepts:**

**DBMS:** Software package that allows you to use a computer to create a database; add, change, and delete data in the database; sort the data in the database; retrieve data in the database; and create forms and reports using the data in the database.

**Instance:** A database instance consists of the running operating environment which allows users to access and use a database. A database (as a generic structured store of data) becomes an instance when instantiated as a system and made available via its database management system. Specific database providers can define database instances in terms of the precise hardware and software resources required to make them available: thus the Oracle database requires allocated system memory and at least one background process before the database counts as an instance.

**Database:** A database is a collection of information stored in a computer in a systematic way, such that a computer program can consult it to answer questions. The software used to manage and query a database is known as a database management system (DBMS). The properties of database systems are studied in information science.

in the case of Oracle and UDB there is a one-to-one relationship between an instance and a database; e.g. there is one and only one database associated with each instance.

Sybase and SQL Server, on the other hand, have a one-to-many relationship between an instance and its databases; e.g., one instance can have multiple databases defined and managed within it.

**The Physical Representation of Database Content on Disk**

All of the "stuff" that's stored and managed by a database (tables, indices, procedures, packages, rules, constraints) has to ultimately reside on disk. Below is an overview of the different approaches taken by the various DBMS architectures.

Oracle and UDB use the concept of a tablespace as the metaphor for holding the contents of a database. There is a one-to-many relationship between an instance (database) and its tablespaces. That is, an instance can and usually does have a number of tablespaces associated with it. Those tablespaces, however are not shared among other, unrelated instances.

A tablespace, in turn, consists of one or more "datafi|es" (Oracle) or "containers" (UDB) which are the actual physical files on disk that are visible to the hosting OS.

Sybase, on the other hand, use the concept of a database device to hold the contents of a database. A database device is somewhat akin to a tablespace. Database devices, in turn, consist of physical files on disk which are visible to the OS. While a database device can be used by multiple databases within an instance, the practice at the author's location is to associate only one database to any particular database device. Therefore, for purposes of trending and analysis, disk utilization metrics at the internal database level were gathered and used for forecasting, and not at the database device level.

ln order to define a common terminology for the tablespace (Oracle, UDB) and database (Sybase) constructs across the various DBMSs, the author coined the term “data holder”. When an instance or database experiences a near or complete shortage of disk space, it experiences that shortage at its “data holder’ level. That condition is manifested in DBMS error messages to that effect. Similary, from the OS‘s perspective, the files which hold the databases content are stored just like any other file; i.e., inside an OS file system. While there are differences between the way the UNIX and Windows Server file systems work internally, logically they can be viewed as a pre-defined amount of disk space for holding files. Analogically, a data holder is to an RDBMS database as a file is to an OS file system -- both are layers of abstraction in the path to the final representation of database content on disk.

Regardless of how file systems are instantiated to a particular OS image (SAN, NAS, arrays, internal disk, others), they all share the common attribute of having been assigned a finite size that meets the anticipated needs of that file system. An image of an OS would typically have many file systems defined to it, each with a different size.

if a database instance has a 10GB file system defined for its use, that file system can hold any number of files as long as the sum of their sizes is <= 10GB. Any attempt to increase the size of an existing file or create a new file that would bring that sum over 10GB would be met with an OS error message and a denial of that attempt.

## **Statement of the Problem**

No matter which DBMS is involved, all databases operate within the constraint of having to house all of their content within a set of data holders, each of which is pre-defined to be of a certain size. When any such data holder is first defined to the database, it will appear to the OS to be a file or set of files which occupies the full size of the defined data holder. For example, creating a 5GB tablespace in Oracle will result in a file or set of files whose sum of file system disk occupancy, as seen by its hosting OS, will be 5GB. However, from the DBMS's perspective at this point in time, the tablespace is empty or 0% full, and has no database content in it yet. It shows up as an empty tablespace with the capacity to hold 5GB worth of database objects. If the hosting OS file system were defined at 10GB, it would see the file system now as 50% full.

As database objects (tables, indices, etc.) are defined and subsequently populated using that data holder, it will present itself to the DBMS as housing n bytes. As n encroaches on 5GB it will come up against the 100% full internal DBMS mark and the DBMS will not be able to add any more content to that data holder until it's is made larger, or content is deleted. At that point the DBMS will return error codes to any database operation which would result in the need for more disk space in the effected data holder; e.g., SQL INSERT requests and certain types of SQL UPDATE requests.) Note that there would be no OS error messages since no attempt has been made to increase the size of the underlying files.

Such a condition would be seen as a loss of availability to parts or all of the application using that database. In order to get the application running again, a short-term quick fix for this situation would be to increase the size of the data holder, providing that the hosting file system had unused space in it for such an increase. lf the file system were full, other IT players would have to be contacted to see if alternative solutions could be tried; e.g., the applications group might see if there was any data that could be deleted from the database, or the host system administration group would see if they could increase the size of the effected file systems. While these options were being explored, the application would be out-of-sen/ice. Were this to occur in the off-hours, an even greater delay in restoration to normal service would be expected as on-call people as contacted to remedy this basic disk space problem from remote locations. The magnitude of this issue became apparent to the author when he saw that that there were over 2,800 data holders that made up the Sybase, Oracle, and UDB instances. These 2,800 data holders in turn consist of over 5,100 individual data files. Managing 2,800 data holders and 5,100 files in a reactive, pager event-driven basis was simply not working. A proactive, quantitatively based forecasting approach was needed.

## **Statement of the Solution**

ln order to prevent these types of database disk space problems from occurring, or at least to greatly reduce their likelihood of occurrence, what was needed was information that characterized the usage patterns of each of the data holders in the enterprise over time. Having that data would allow one to extract the underlying trends and patterns exhibited in the data holders over an extended period, and to forecast what the future needs were of each data holder.

To that end, the author devised the simple collector mentioned earlier for all of the Oracle databases. The collector itself is a single SELECT statement that leverages several PL/SQL features. This statement is run just once a day on a single Oracle instance. That single instance has database links set up for all of the other Oracle instances in the enterprise. This allows the SQL to gather the information from all other instances in the complex via database links, and to gather all of the information for all data holders on all instances into a single database table for analysis and longer term storage The PL/SQL loops through the dba_db_|inks table, generates the SQL needed for each instance, and then executes the generated SQL. The execution is serial, gathering the needed information about all tablespaces in any one instance and then going on to the next instance until information from all instances is placed into the central repository table. The information gathered from each instance was described above and is repeated here in its database format in Table below:

| Column Name | Data type |
|---|---|
| Batch_date | DATE |
| Instance | VARCHAR2 255 |
| RDBMS type | VARCHAR2 6 |
| Data_holder_name | VARCHAR2 30 |
| Allocated_bypes | NUMBER |
| Free_bytes | NUMBER |

Note: “batch_date" is the date and time when the data was collected. ln order to provide a consistent value for all of the tablespaces in all of the instances at data collection time, the current date and time at the start of the collection process is stored as a constant. It's then reused in all of the data extracted from each instance, even though the actual extraction times might be a minute or two offset from that value. Given the longitudinal nature of the data used in the analyses, this difference in time is not a problem. Having a consistent date and time for the all values collected that day allows grouping by date and time in subsequent analyses.

As noted earlier, shortly after the Oracle collectors were created and put into place, another team member created collectors for Sybase and UDB. These collectors gather the same six fields as shown in Table 1 since the data holder concept, by intent and design, is extensible to all DBMSs. Due to architectural dissimilarities between Oracle, Sybase, and UDB, the actual means of collecting the information is unique to each DBMS. The extracted data, however, has the same meaning and is not sensitive to any particular DBMS collection context. Once this information was gathered to cover a reasonable period of time, it was possible to subject the data and its derivatives to a number of analyses, described below.

**Results:**

Forecastinq Analvses Performed on the Data:

In order to use the collected data, it first had to be placed on a common platform that would provide the analytics needed for forecasting, descriptive statistics, etc. While the Oracle collector accumulated all of its obsen/ations about each Oracle instance into a single table on the central Oracle collector instance, the Sybase and UDB collectors used a different approach. They created individual flat files in comma separated value (CSV) format on each Sybase and UDB instance's host. What was needed was a means to gather the content of these three disparate data sources and put it in one spot. The initial solution chosen for this forecasting system was to use Microsoft's OLAP Services, a component of MS SQL Server versions 7.0 and 2000. The CSV files from each Sybase and UDB instance were automatically gathered together each day and sent via file transfer protocol (FTP) to an MS SQL Server instance. There they were loaded into a common table (t_data_hoIder) via Data Transformation Services. Similarly, the Oracle data for each day was automatically extracted out of its table and loaded into the same location. That table's layout is identical to the one shown in Table 1. All columns have the "NOT NULL" attribute. The intent of designing the data structure in this way was to provide a means of performing multi-dimensional analyses along variables of interest. While the main focus was to forecast when each individual data holder would run out of space, the presence of the instance, DBMS, and data holder columns allowed one to dice-and-slice the data along those values. These are discussed later in the paper.

The following derived measures were created for each data holder:

- Bytes_used: (bytes_allocated — bytes_free)
- Percent Used: (Bytes_used/bytes_allocated)*1O0
- Percent Free: (bytes_free/bytes_allocated)*100

With the table in place on MS SQL Server, the author defined a multi-dimensional data structure and set up analyses in MS OLAP Services. That structure and its analyses vetted (serial) time against bytes_used for each unique combination of instance name and data holder name. This analysis used the most recent 365 day's of daily data points for each data holder as input and yielded the slope, intercept, and Pearson product moment correlation coefficient (squared) for each data holder on each instance. Numerous other descriptive statistics were also calculated for each data holder such as the mean, median, mode, variance, standard deviation, and number of observations. The output of these OLAP Services analyses was in the form of a table which contained a row for each of the 2,800 data holders. The columns in each row of this table were the instance name, the DBMS type, the data holder name, slope, the intercept, the R2, the number of observations, the mean, the median, the mode, the variance, and the standard deviation for that "set".

Additional columns in this table, by way of programming done by the author within OLAP Sen/cies, were the most current values for bytes_used bytes_allocated, and bytes_free, along with the value of the date of the most recent observation in the past year's set of data for this instance/DBMS/data_holder set. Lastly, OLAP Sen/ices was further programmed to take these values and forecast what the expected shortfall or surplus would be, in bytes, for each data holder six months from the current date, using the method described immediately below:

Using the simple linear equation "y = mx + b", the author solved for "y" in order to see how many bytes would be in use (bytes_used) at time where "X" was displaced 180 days fon/vard from the current date. Applying the slope "m" calculated for each data holder we arrived at the projected bytes_used six months from now.

Further programming in OLAP Services yielded the difference between the most current bytes_a||ocated number and the six-month forecast bytes_used value If the difference between the current bytes_a|located and the forecast bytes_used was positive, that indicated that we would have a surplus of that exact magnitude six months from now for that particular data holder in that particular instance. lf, on the other hand, that result was negative, we would have a shortfall of that size in six months, and would need to take corrective action now so as to forestall an on-call coverage pager event in the future. The above calculations were all predicated upon filling the data holder to 100% of its allocated capacity at the six month target date, since we were forecasting bytes_used against the most current bytes_a||ocated. Based upon practical experience, and given the variance observed over time in each data holder on each instance, or collectively across the DBMS or instance dimensions, it was evident that allowing a data holder to approach 100% full was a dangerous practice. It did not take into account the periodic, seasonal, and random ebbs and flows that were observed in the data holder's behavior with respect to bytes_used. Not all data holders grew at a linear rate; rather, they exhibited troughs and crests in bytes_used over the course of time. The consensus within the author's workgroup, after having looked at the data for all instances and data holders, was that a best practice would be to forecast data holders to reach their 70% full "saturation" point. That being agreed, bytes_used was adjusted in the equation by dividing it by 0.70. The values of the surplus/shortfall column for all 2,800 data holders were then examined for the largest negative values. This was done by importing the OLAP cube into Excel. The conditional formatting feature in Excel was used to show those data holders with a projected shortfall to have their numbers literally "in the red". In some oases there was sufficient space in the underlying file system(s) to satisfy the forecast shortfall, and the data holder was increased in size by the amount calculated by one of the database staff. However, in other cases, there was not sufficient free space in the underlying file system(s) and a formal request was created for the UNIX disk management group to add the needed number of bytes. This proactive, forecasting approach allowed such requests to be fulfilled well in advance of the six month projected shortfall. The forecast numbers, in and of themselves, were not used blindly. Examination of the R2 values for each data holder was used to assess how well the observed data points resonated to the march of time. If the R2 value was below 0.80, visual inspection of the plotted bytes used data points was undertaken to help understand the pattern, if any, in the data. If the data was “all over the place" for a particular data holder, the database team would make a best estimate of what to do with that individual data holder, and manually monitor it more closely in the coming six months. Since R2 is the percent of the observed variance that’s accounted for by the independent variable (time in this case), 0.80 was arbitrarily used as a line in the sand against which all forecasts were evaluated for usefulness.

These analyses were run on a regular and automatic basis every three months. The results were examined by the database group to see which data holders needed an adjustment to accommodate their projected growth (or decline) in the next six months.

## 3. BEYOND FORECASTING

### 3.1 Additional insights Provided by the data:

Having a year’s worth of data in the database now allowed the author to pose specific queries about the nature of all the databases in the organization. It was only by having this data and exploiting its emergent properties that these insights were possible Heretofore, such questions had been impossible to answer quantitatively due to there being no historical data. Best guesses were made, current real-time data was used, and the collective anecdotal experience of the workgroup was combined to produce SWAG answers.

Now, with the actual data at hand, a number of questions could be and were answered:

- Q. Which data holders exhibit has the highest or lowest mi of growth?

- A. By sorting the OLAP cube on the slope value, we display all data holders ordered by their rate of growth, from positive to negative.

- ' Q. Which data holders exhibit has the highest data content occupancy?

~ A. By sorting the OLAP cube on the mean bytes used value, we display all data holders ordered by amount of information they store.

- Q. Which data holders exhibit has the highest disk occupancy?

- A. By sorting the OLAP cube on the mean bytes_al|ocated value, we display all data holders ordered by amount of disk space they take up.

- Q. Which data holders are the most over- or under-utilized?

- A. By sorting the OLAP cube on their Percent Used or Percent Free values, we can display all data holders ordered by where they stand in the O-100% data holder full category.

Perhaps the most valuable insight was provided by creating a pivot table/chart in Excel, which was published to the corporate intranet for use by managers, developers, business analysts, and others. This allowed IT staff to visualize the trends and other characteristics of the data in an interactive manner. Since Excel has an internal limit of 65K rows in a worksheet, and we had 2,800 data holders with 365 observations each, or 1,022,000 rows, the raw data could not be put into Excel. instead, the author elected to only use the data from the production instances, and to further aggregate that into its weekly mean values. This was done by writing a trivial SQL statement to find the weekly means for bytes_used, bytes_free, and bytes_a||ocated for each unique combination of instance name, DBMS type , data holder name, and week number within the year. (Since the data was on MS SQL Server, the datepart “week” value was used in the GROUP BY clause. The datepart "year" was used in the expression to order the data appropriately, since we had multiple years in the database.)

The data for the pivot tables and charts consisted of the elements shown in below Table:

| |
|---|
| dbms_prd_name (Sybase, Oracle, UDB) |
| db_instance (instance name) |
| data_holder_name |
| dbs_year (YYYY) |
| dbs_week (1-52) |
| Bytes_used (by that data holder) |
| Percent_Full (for that data holder) |

Two pivot tables and two pivot charts were created from this data: one that showed the (absolute) bytes_used values, and one that showed the (relative) percent full values. The bytes_used pivot table and chart had the following characteristics:

- Its x-axis was time, expressed as the past 12 months, using the year and the week within the year. Since the past 12 months would span a year in all cases but the beginning of a new year, two pivot select buttons appear on that axis: one for year and one for week within year. For the most part these buttons were unused, and the entire 12 months of data was viewed.
- its y-axis plotted bytes_used.
- Pivot buttons were provided for:
  - ➢ dbms_prd_name
  - ➢ db instace
  - ➢ data holder name

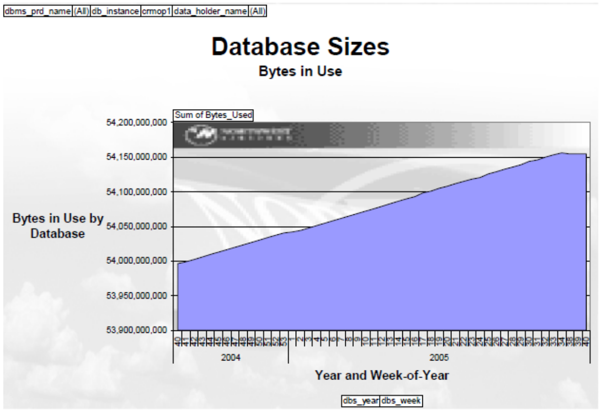
This allows the views to dice-and-slice the data along any of these dimensions. For example, these questions were answered in the pivot chart:

- Whats the pattern of bytes_used over the past year for:
  - ➢ All Oracle instances?
  - ➢ All Sybase instances?
  - ➢ All UBD instances?
  - ➢ Oracle and Sybase combined?
  - ➢ Oracle and UDB combined?

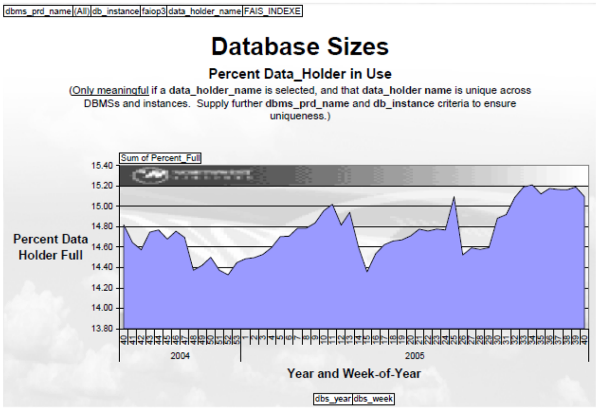
- Sybase and UDB combined?
- Sybase and Oracle and UDB combined?

- What's the pattern of bytes_used over the past year for:
  - Any individual instance?
  - Any combination of instances? (Note this also permits any combination of instances of interest. regardless of the DBMS that's hosting them.)
- What's the pattern of bytes used over the past year for:
  - Any individual data holder? (Note that one must enter an instance name for this to be meaningful. Otherwise it would show the total value for all data holders that have that name, regardless of the instance name.)
  - Any combination of data holders?

An example of this pivot chart is shows in Figure below:



The percent used pivot table and chart had the same setup as the bytes used pivot table. However, since the percent calculation was performed at the data holder level in the raw data, it would not be valid to do any roll-ups on the DBMS or instance dimensions. Therefore, a warning message was written to appear on the pivot chart that read Only meaningful if a data holder name is selected, and that data holder name is unique across DBMSs and instances. Supply further dbms_prd_name and db_instance criteria to ensure uniquenessm') Below Figure shows the pivot chart. By using this second pivot chart, people in the IS infrastructure could see which data holders are close to their 100% full limits. Also, the pivot table content can be copy/pasted into a new spreadsheet and then sorted on its percent full value to show all data holder's in the enterprise in order by their percent full values. Using excel Filtering, these queries can further be refined into DBMS type, instance name, and data holder name.



The Excel spreadsheet was created such that one can refresh the raw data from the source database on MS SQL Server with a single click. Therefore, each month it's possible to completely update the pivot tables and charts with virtually no effort.

## 4. DISCUSSION

By measuring and storing just these two, simple metrics every day for each data holder on each instance (bytes_allocated, bytes_free), the organization was able to evolve from its previous reactive mode to a more proactive and methodical process.

### Benefits

Some of the benefits that accrued through this shift in focus and the use of applied mathematics were:

- With this data now published on a regular monthly basis to the intranet, the consumers of it have gained considerable insights into the seasonal and other variations in their data usage patterns.

- The work group responsible for acquiring disk space forthe entire IS organization can now set realistic budget values for next years disk space requirements, based upon the higher level rollups of the bytes_used data.

- Pager call reduction: the 1,041 pages that were previously issued per year for database disk space problems dropped to only a handful.

- The rates of growth of the various applications or business systems at the organization were now quantified and published. This allowed the IT organization to compare those rates between applications, year-over-year, etc.

- The organization can now identify any anomalousrates that might indicate that an application change (intended or not) or business driver variation was having a significant impact on the rate at which data was being accrued in a database.

- Descriptive statistics can be compared between data holders to better understand their central tendencies and dispersion characteristics.

## 5. CONCLUSION

lt’s frequently amazing how just a tiny set of data observations can form the basis of uncovering the underlying substrates of an IT infrastructure. Automatically gathering just two values per day from each data holder in the enterprise allowed the IT staff to quantitatively and visually depict the patterns that had had been there all the time -- they had just not been measured. Applying that knowledge freed up considerable staff time which had previously been consumed by unnecessary, reactive paging events and by daily disk space monitoring. Now that these analyses have given us a glimpse into the nature of the organization's database environment, additional next steps can be considered:

- One could correlate (orjoin, in “database-eze") OS file system statistics with the DBMS data above so as to automatically determine if there is enough space in a file system to address the forecast deficit.

- We could explore non-linear regression relationships, any number of classical transforms (log, power, exponential, Nmorder polynomials, squares, cubes, inverses, etc.) of the dependent and independent variables could be performed to determine if any combination of those yields higher R values.

The very first DBMSs, which appeared in the early hunter-gather phase of IT, required a tremendous amount of staff time just to keep them running. Knowledge about and experience with them was scarce, often acquired as folklore from the tribal elders around the corporate campfires. Secret handshakes and magical amulets abounded. Mystical robes were frequently donned in order to exorcise the demons that plagued that software. However, over time, the DBMS vendors as a group added more and more self-managing capabilities to those systems, which made them less and less labor intensive. Even with those enhancements, the pesky problems associated with database disk space persisted. Some vendors did add features that self-managed the data holders by automatically extending them into their host file systems on an as needed basis, following business rules set up by the database administrators. Vendors are even putting in features that retract over-allocated data holders into disk space footprints that more closely resemble their normal usage patterns As those features become the accepted best practices in the workplace, the clerical tedium of managing database disk space will eventually become a faded memory, much like the punch card.

The system described in this paper is an attempt to provide a stepping stone to bridge the gap between the present state of DBMS capabilities and the future, and to do so with a methodology firmly rooted in quantitative analysis.

## 6. REFERENCES

[1] Vijay Datla, “Software Performance Tuning”, IJARCST, vol. 4, issue 4, 2016.

[2] http://www.scolumbiasd.k12.pa.us/hsf/business/gengler/compapp/accintro.htm

[3] Vijay Datla "Performance of Ecommerce Implementation", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, issue 12, 2016.

[4] http://en.wikipedia.org/wiki/Database_instance

[5] Vijay Datla “Software Performance Workload Modelling”, International Journal of Computer Applications Technology and Research (IJCATR), Volume 6-Issue 1, 2017. doi:10.7753/IJCATR0601.1003

[6] http://en.wikipedia.org/wiki/Database_instance

[7] Vijay Datla “Performance Lifecycle in Banking Domain”, International Journal of Computer Applications Technology and Research (IJCATR), Volume 6-Issue 1, 2017. doi:10.7753/IJCATR0601.1004

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 32-37, 2017, ISSN:-2319–8656

# Performance Tuning of Data Warehouse

Srikanth Kumar Tippabhotla
CTS

**Abstract**: There are many data warehouse performance aspects. and knowing all the different performance aspects. issues and considerations for building your data warehouse can be a tremendous task. Learn how to build your data warehouse for maximum SOL access performance while maintaining the key values for partitioning and parallelism considerations. Through this article you will learn many design options and alternatives to maximize performance for your high performance data warehouse design.

**Keywords**: Performance, Data warehouse

## 1. INTRODUCTION

There are many data warehouse performance aspects, and knowing all the different performance aspects, issues and considerations for building your data warehouse can be a tremendous task. DB2 offers many performance aspects missing from other DBMSs that can make a huge performance difference. This article will highlight the performance advantages of DB2 and how to handle all the design issues, alternatives and considerations experienced while building a large high performance data warehouse.

This article also explores how to build your data warehouse for maximum SOL access performance while building or maintaining the key values for partitioning and parallelism considerations. Since no data warehouse should be design as an island of information, OLAP tool performance aspects for large numbers of concurrent users are also critical design points that will be discussed. Through this article you will be exposed to many design options and alternatives to maximizing performance for your data warehouse system.

## 2. DATA SOURCES

After analyzing and choosing the best data sources, the first task is to design and build the extract, transformation, load (ETL) and data maintenance processes. Parallelism needs to be designed into these processes to reduce the time windows for ETL, maintenance and maximize end-user access. Parallel processing should be considered and weighed against each design alternative to ensure that the best overall design is achieved.

Many corporations only focus on the data sources that can be gathered internally and not information that is available for purchase from outside vendors. This can be a major mistake as the needed information or extra value may only exist from an external source. Outside data should be considered but it should have analyzed to determine the efforts and processes for data standardization and possible complex data conversions.

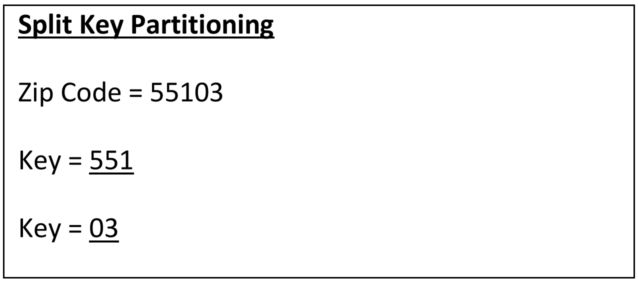
Any data brought into the data warehouse should be as clean, consistent and carefree as possible. Data transformations processes can be very expensive by adding a large number of additional l/O’s to the overall processing. This additional l/O’s can be expensive for large amounts of daily or weekly ETL processes. These ETL cleansing or transformation processes also need to be consistent in business rules and context across multiple business rules and context across multiple diverse source systems. Data consistency and standard domains and ranges are vital for developing clean and usable data that can easily be grouped, rolled up, and associated appropriately for useful business decisions.

Different types of business decision points and business intelligence analysis require different types of data to make useful and insightful decisions from the warehouse data. The input data used and its transformation into the warehouse needs to be reflective of the type of decisions and the overall project objectives set by management for the data warehouse.

Evaluating different input source systems available for your data warehouse, a comparison criterion of how much standardization and transformation each source requires needs to be analyzed. This analysis should determine the additional number of I/O’s for standardization and transformation required by each source Considering these additional amounts of processing can quickly add up especially when additional validation and dimensional key translation IIO activity is included also.

## 3. DATA STRUCTURES AND PARTITIONING

Common data warehouse designs define fact and dimension data table objects. Properly designing these tables to minimize their size, the number of required indexes, size and the clustering method of the data is very Important for overall performance. DB2 partitioning scheme is used to minimize object size and separate out data locking Issues.



**Split Key Partitioning**

Zip Code = 55103

Key = 551

Key = 03

One of the ways to partition DB2 database tablespaces on the z/ OS environment is to split a primary key data into two or more columns shown in the above figure. Any of these partial keys can be used to achieve different partitioning clustering orders and grouping opportunities to leverage a desired data warehouse analysis point; This redefinition must be done carefully to make sure end-user access or ETL process performance does not suffer and is enhanced by the splitting the key and potentially the processing. Also by splitting the data along the separated column values. the data can be grouped to a particular machine, partition or department. This method can also be very effective for segregating the data to different physical devices that can provide an extra performance factor by separating the it’s to different physical devices. Keys that are randomly generated from a formula or algorithm can also be used for partitioning. This method can also be very effective for randomizing the data across the database structures. Using a secondary column as the data partitioning limit key the data might be able to be spread evenly across 100 segments of the database tablespace. These 100 different partitions could then be placed across several different machines or DASD volumes to help parallelism, processing performance or backup and recovery issues. Care must be taken when splitting a key into multiple columns because of Implications of end-user SQL or OLAP tools but in most cases analysis shows these negative or out weighted by the tremendous performance improvements offered by the clustering and grouping opportunities.

Another way to partition database tablespace is through the use of results keys as shown in the below picture

**Result Key Partitioning**

ID = 96808755, Divide by 40

Result = 2420218

Remainder = 35

Sometimes a key is developed from a formula or business practice that uses a derived key. This can be done for a variety of reasons such as security, new business practice or for randomizing the key values. Result keys or hashing keys are usually a secondary key, calculated and used for a particular analysis type or grouping purpose. A simple remainder formula or a complicated calculation can be embedded in a source system that can be leveraged into the design of the partitioning of the warehouse. Documenting this type of key generation is very important because a transformation process will probably need to duplicate the formula or hashing process to generate new key values. Generating keys through formulas or hashing algorithms is good for precisely distributing the data across a range of partitions or machines as desired. The distribution can be done evenly or through a processing that maximizes data density. Manipulating the formulas or hashing routines can also be used to target the data to certain locations while maintenance is done on other portions of the database. The difficulty with using formulas or hashing routines for partitioning is that the end-user can rarely use this type of key for querying the database unless properly defined and accessible through a user-defined function (UDF}. So using this type of index should be researched and justified carefully but its flexibility can be tremendous performance asset.

In some DBMS limit keys are defined in an index structure provides the mechanism for separating the data across partitions or nodes. These limit keys can be customized to reflect the data value distributions to provide an even distribution. The distribution can be controlled via e single column value or multiple column values. Composite keys are much more common than partial keys for indexing (Below Figure). Composite keys are built from many different columns to make the entire key unique or to include numerous columns in the index structure.

**Composite Key Partitioning**

Region Code = 01

Zip Code = 10015

Complete key = 0110099

CIT Sub-partition = 0110050

Sub-partition2 = 0110099

Sub-partltion3 = 0120050

Sub-partition x.= 0540099

For example, a product department will not distinctly identify an item but add SKU number, color and size and it will uniquely identify an item. Each of the columns used to form the composite index provide a partitioning option along with the options of combining columns to form additional partitioning alternatives. During dimension definition sometimes it is best to minimize the number of objects. if the dimension's keys have common domains sometimes it is convenient to combine the entities and combine their index into a composite key. Combining objects should be studied deeply to make sure relationships and integrity are maintained but it can sometimes be a great performance enhancer by eliminating the extra IO going to the extra dimension table. Partitioning with composite keys can also help spread the data appropriately when using secondary column as partitioning limit keys.

Index structures, configuration parameter and partitioning definitions can all be used to separate indexes away from data information across the DASD configuration the physical nodes or the database partitions. Based on a key or a rule the data is stored in a particular allocation of DASD. Separating the data away from its index structure makes it easier to reorganize, maintain, backup or recover in the event of a hardware failure. The data separation is also very important because of the potential to evenly distribute the IOs of any parallel processing. Distributing the lO and data allows multiple processing entry points so that no single process or device becomes saturated with I10 requests. Eliminating and minimizing lO can make or break performance objectives. A prototype design should be put together and typical queries estimated. SQL Traces on the system during the user testing can help point out what tables are most popular or which ones might become bottlenecks or system issues. Determining the optimum number of partitions or nodes depends on a variety of factors. The physical CPU configuration and the HO hardware can be a big performance factor. By matching the hardware to the design, the designer can determine how many partitions and parallel processing streams your CPU, IO and network can support.

## 4. MULTI-DIMENSIONAL CLUSTERING

Another new feature in DB2 Ver8 is new patent pending clustering technique MDC- Multi dimensional clustering. This feature does exactly as the name Implies; it clusters the data against multiple dimension keys. This unique MDC clustering is achieved by managing data row placement into a brand new page extent blocks space management scheme based on their dimensional key values. The new space management, placement and access are facilitated through a new Version 8 Block Index object type. This new Block Index object type is created for each of the dimensions, is similar in structure to a normal index but cross references rows to a larger dimensional data block instead of an individual row. The new extent data page block sizes are chosen at Multi-Dimensional Clustering definition time and if additional space is needed consecutive block extents are defined. Since the rows are managed to a data block, the cross-referencing Block index information needed is smaller, resulting in a smaller index structure. With consecutive pages and the Block index only referencing data blocks, Block index reorganization will rarely or not is needed as often as a regular indexes referencing individual rows. Taking data placement management one- step further than partitioning, Multi-Dimensional Clustering groups the rows to the various dimensional key blocks ideally organizing the rows for data warehousing and OLAP application access. End-user SOL can reference the Multi-Dimensional Clustering Block indexes individually, and or combined with regular indexes and utilized in all the intra and inter parallelism optimizer access methods to quickly retrieve large amounts of data. Since the Multi-Dimensional Clustering blocks are defined and extended in consecutive page blocks, similar data is contained in consecutive pages making caching, pre- fetching, RID lists and accessing data that much quicker. Clustering along multi-dimensional keys also has tremendous value for regular insert, update activities also. With a regular table, the data is placed via a single lustering value and becomes un-clustered with more insert and update activity. Multi-Dimensional Clustering tables maintain their clustering continuously over time because the clustering is based on multiple clustering keys that point to large data blocks instead of individual rows. Multi-Dimensional Clustering —MDC tables are the next step in database table design evolution. With all of its advantages this patent pending unique DB2 clustering technique gives new performance advantages and flexibility to data warehouse, OLAP and even OITP applications database designs.

## 4.1 Specialized Tables

DB2's Materialized Query Tables (MQTS) formerly known as Automatic Summary Tables and summary tables that total departments or product lines also called horizontal aggregation can greatly enhance end-user access performance by minimizing the amount of data accessed. For example, by using a MOT or summary table of monthly totals, sales-to-date figures could be developed with fewer I/O than referencing all the detail sales data. Tracking query activity can sometimes point to special data requirements or queries that happen on a frequent basis. These queries may be totaling particular products or regions that could be developed and optimized through a MQT or horizontal aggregate. Analysis must be done to justify the definition of an MOT to make sure it is used enough. Like all aggregates, MQTs and horizontal aggregates if used enough can eliminate IO and conserve CPU resources. MQTs and Horizontal aggregates work from a particular dimension key that is can be easily separated from the rest of the data.

Another method for creating specialized tables is through the use of Global Temporary Tables. Care needs to be taken to include the GTTs information in all OLAP or end-user tool information so it can be evaluated and possibly utilized for end user query result sets. Sometimes GTTs can also be used to limit the data accessed and can provide extra security against departments looking at other department's data. The GTTs or horizontal aggregate security technique is very effective and also maximizes query performance by minimizing access to only the data needed for the analysis. Another method of speeding analysis is through the use of Materialized Views as aggregate data stores that specialize in a limited dimensional data. These are good for taking complicated join predicates and stabilizing the access path for end-users. Data warehouse data can also be summarized into MVs to provide standard comparisons for standard accounting periods or management reports. Aggregate functions work best when defined to existing end-user comparison points, for example a department, product code or time data. These aggregates can be used extensively for functions and formulas because of their totaled data and definite criteria for gathering the information.

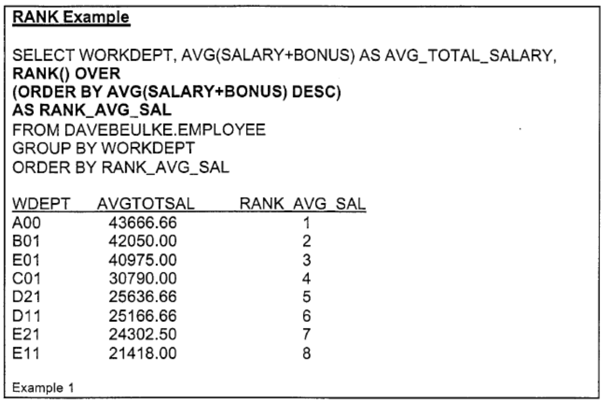
Documentation and meta-data about aggregates must be well published and completely understood by all end-users and their tools. Any end-user or OLAP tools should be aggregate aware and be able to include the different appropriate aggregates in their optimization costing and end-user processing. These data aggregates can save a tremendous amount of |l'Os and CPU. Make sure the aggregates and summaries are monitored to demonstrate and justify their creation.

## 5. UNIQUE DB2 SQL OLAP FEATURES

The SQL OLAP functions performed inside DB2 provide the answers much more efficiently then manipulating the data in a program. Like join activities and other data manipulation that DB2 can do directly, the SQL OLAF' functions can greatly reduce overall IEO and CPU utilization. These OLAP functions are particularly good for getting the top number of data rows that match a criterion.

## 5.1 OLAP Rank Function Example 01

This function can be used to order and prioritize your data according to your specific criteria. RANK orders your SQL query data by your criteria and assigns successive sequential numbers to the rows returned. The SQL query ranked rows can be individual data rows or groups of data rows.



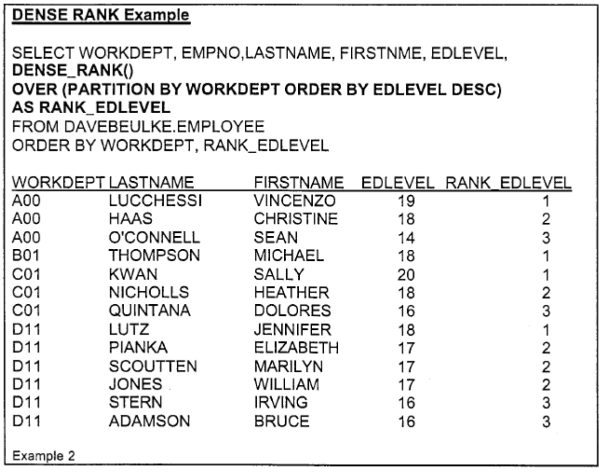
**RANK Example**

```
SELECT WORKDEPT, AVG(SALARY+BONUS) AS AVG_TOTAL_SALARY,
RANK() OVER
(ORDER BY AVG(SALARY+BONUS) DESC)
AS RANK_AVG_SAL
FROM DAVEBEULKE.EMPLOYEE
GROUP BY WORKDEPT
ORDER BY RANK_AVG_SAL

WDEPT    AVGTOTSAL      RANK_AVG_SAL
A00       43666.66            1
B01       42050.00            2
E01       40975.00            3
C01       30790.00            4
D21       25636.66            5
D11       25166.66            6
E21       24302.50            7
E11       21418.00            8
```

Example 1

## 5.2 DENSERANK Function Example 02

This function can also be used to order and prioritize your data according to your specific criteria. DENSE_RANK orders your data and assigns successive sequential numbers based on the Over Partition data values found. DENSE_RANK differs from RANK because common values or ties are assigned the same number.



**DENSE RANK Example**

```
SELECT WORKDEPT, EMPNO,LASTNAME, FIRSTNME, EDLEVEL,
DENSE_RANK()
OVER (PARTITION BY WORKDEPT ORDER BY EDLEVEL DESC)
AS RANK_EDLEVEL
FROM DAVEBEULKE.EMPLOYEE
ORDER BY WORKDEPT, RANK_EDLEVEL
```

```
WORKDEPT LASTNAME      FIRSTNAME   EDLEVEL  RANK_EDLEVEL
A00      LUCCHESSI     VINCENZO      19             1
A00      HAAS          CHRISTINE     18             2
A00      O'CONNELL     SEAN          14             3
B01      THOMPSON      MICHAEL       18             1
C01      KWAN          SALLY         20             1
C01      NICHOLLS      HEATHER       18             2
C01      QUINTANA      DOLORES       16             3
D11      LUTZ          JENNIFER      18             1
D11      PIANKA        ELIZABETH     17             2
D11      SCOUTTEN      MARILYN       17             2
D11      JONES         WILLIAM       17             2
D11      STERN         IRVING        16             3
D11      ADAMSON       BRUCE         16             3
```

Example 2

There are many more considerations that effect data ware house system performance. Nut taking advantage of the tips and techniques discussed with in the ETL processes, parallelism, partitioning and OLAP functions can greatly Improve overall performance.

## 6. REFERENCES

[1] Vijay Datla, “Software Performance Tuning”, IJARCST, vol. 4, issue 4, 2016.

[2] "DB2 UDB for OSISQO V6 Performance Topics" IBM

Redbook SG24-5351

[3] "DB2 UDB for AIX V3.1 Administration Guide" IBM

[4] "DB2 UDB for AIX V8.1 SOL Reference" IBM

[5] Vijay Datla "Performance of Ecommerce Implementation", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, issue 12, 2016.

[6] "Approaches and Methodologies for Capacity Planning for Business Intelligence Applications" IBM Redbook SG24 -5689

[7] Vijay Datla “Software Performance Workload Modelling”, International Journal of Computer Applications Technology and Research (IJCATR), Volume 6-Issue 1, 2017. doi:10.7753/IJCATR0601.1003

[8] "DB2 UDB for ZIOS V8 Administration Guide“IBM

[9] "DB2 UDB for Unix, Linux and Windows V8 Administration Guide“ IBM

[10] Vijay Datla “Performance Lifecycle in Banking Domain”, International Journal of Computer Applications Technology and Research (IJCATR), Volume 6-Issue 1, 2017. doi:10.7753/IJCATR0601.1004

[11] Bob Lyle "DB2 OLAP Functions" International DB2 Users Group European Conference (2001)

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 42-50, 2017, ISSN:-2319–8656

# The Current State of Phishing Attacks against Saudi Arabia University Students

Bushra Mohamed Elamin Elnaim
Department of Computer Science and Information
Sattam bin Abdulaziz University
Al Sulail, Kingdom of Saudi Arabia

Hayder Abood S.Wsmi.Al-Lami
Department of Computer Science and Information
Sattam bin Abdulaziz University
Al Sulail, Kingdom of Saudi Arabia

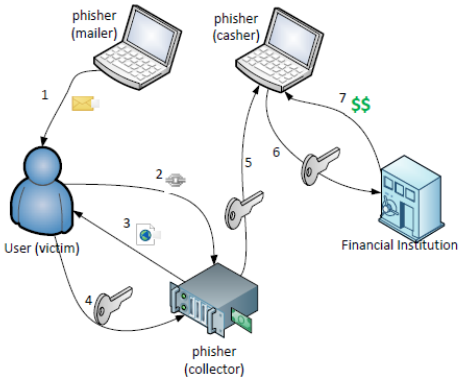
**Abstract**:

Research into phishing and social engineering is a very interesting area since a significant number of attacks are conducted with the help of social engineering and phishing as the main vector to either obtain credentials or trick the user into executing a malware infected file. The goal of our research was to examine the students' familiarity with threats in the form of phishing attacks conducted via the Internet. A questionnaire was conducted to determine the students' ability to recognize phishing attacks and if they know how to protect themselves. The motivation behind this research is to explore the Saudian Student population's self assessment in regard to phishing attacks and to assess their capability on a limited data set for purpose of obtaining a baseline for future research.

**Keywords**: phishing attack; Saudi Arabia student; social engineering; Phishing Attacks in KSA universities; Types of phishing attacks.

## 1. INTRODUCTION

Phishing is a form of social engineering in which an attacker, also known as a phisher, attempts to fraudulently retrieve legitimate users' confidential or sensitive credentials by mimicking electronic communications from a trustworthy or public organization in an automated fashion [1]. The word "phishing" appeared around 1995, when Internet scammers were using email lures to "fish" for passwords and financial information from the sea of Internet users; "ph" is a common hacker replacement of "f", which comes from the original form of hacking, "phreaking" on telephone switches during 1960s [2]. Early phishers copied the code from the AOL website and crafted pages that looked like they were a part of AOL, and sent spoofed emails or instant messages with a link to this fake web page, asking potential victims to reveal their passwords [3]. A complete phishing attack involves three roles of phishers. Firstly, mailers send out a large number of fraudulent emails (usually through botnets), which direct users to fraudulent websites. Secondly, collectors set up fraudulent websites (usually hosted on compromised machines), which actively prompt users to provide confidential information. Finally, cashers use the confidential information to achieve a pay-out. Monetary exchanges often occur between those phishers. Show Figure(1):



Figure(1): Phishing Information Flow

## 2. LITERATURE REVIEW

## 2.1 Types of Phishing Attacks

Numerous different types of phishing attacks have now been identified. Some of the more prevalent are listed below:

### 2.1.1 Deceptive Phishing

The term "phishing" originally referred to account theft using instant messaging but the most common broadcast method today is a deceptive email message. Messages about the need to verify account information, system failure requiring users to re-enter their information, fictitious account charges, undesirable account changes, new free services requiring quick action, and many other scams are broadcast to a wide group of recipients with the hope that the unwary will respond by clicking a link to or signing onto a bogus site where their confidential information can be collected.[4]

### 2.1.2 **Malware Phishing**

Phishing scams involving malware require it to be run on the user's computer. The malware is usually attached to the email sent to the user by the phishers. Once you click on the link, the malware will start functioning. Sometimes, the malware may also be attached to downloadable files.

Phishers take advantage of the vulnerability of web security services to gain sensitive information which is used for fraudulent purposes. This is why it's always a good idea to learn about the various phishing techniques, including phishing with Trojans and Spyware.[5]

### 2.1.2 Key loggers and Screen loggers

Key loggers and screen loggers are varieties of malware that track input from the keyboard and send relevant information to the hacker via the Internet. They can embed themselves into the user's browsers as small utility programs.[6]

### 2.1.4 Session hijacking attack

The Session Hijacking attack consists of the exploitation of the web session control mechanism, which is normally managed for a session token.

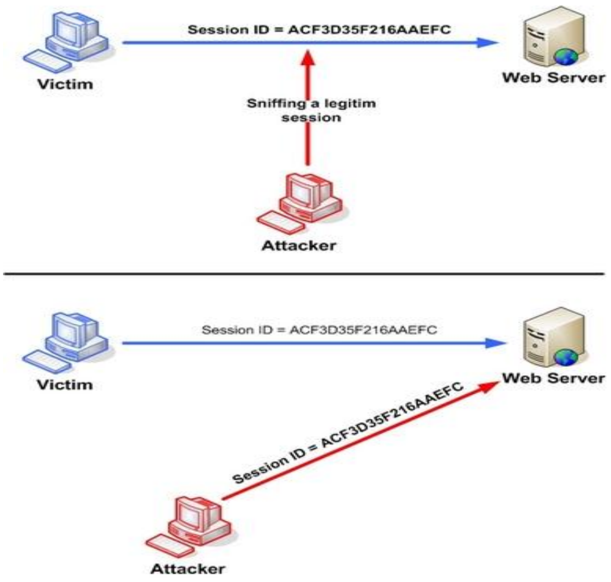
Because http communication uses many different TCP connections, the web server needs a method to recognize every user's connections. The most useful method depends on a token that the Web Server sends to the client browser after a successful client authentication. A session token is normally composed of a string of variable width and it could be used in different ways, like in the URL, in the header of the http requisition as a cookie, in other parts of the header of the http request, or yet in the body of the http requisition.

The Session Hijacking attack compromises the session token by stealing or predicting a valid session token to gain unauthorized access to the Web Server.

The session token could be compromised in different ways; the most common are:[7]

- Predictable session token;
- Session Sniffing;
- Client-side attacks (XSS, malicious JavaScript Codes, Trojans, etc);

As an example the following figure:


Session ID = ACF3D35F216AAEFC
Victim
Web Server
Sniffing a legitim session
Attacker
Session ID = ACF3D35F216AAEFC
Victim
Web Server
Session ID = ACF3D35F216AAEFC
Attacker

Figure(2): Session Sniffing

In figure(2) , as we can see, first the attacker uses a sniffer to capture a valid token session called "Session ID", then he uses the valid token session to gain unauthorized access to the Web Server.

### 2.1.5 Trojan Virus

A Trojan horse or Trojan is a type of malware that is often disguised as legitimate software. Trojans can be employed by cyber-thieves and hackers trying to gain access to users' systems. Users are typically tricked by some form of social engineering into loading and executing Trojans on their systems. Once activated, Trojans can enable cyber-criminals to spy on you, steal your sensitive data, and gain backdoor access to your system. These actions can include:

- Deleting data
- Blocking data
- Modifying data
- Copying data
- Disrupting the performance of computers or computer networks

Unlike computer viruses and worms , Trojans are not able to self-replicate.[8]

### 2.1.6 DNS Poisoning

DNS poisoning is a method which gives the impression that hackers took control of some known sites or not. DNS is the protocol which connects domain name and IP address for any site in the world has one or more IP addresses. When we write in the browser "google.com" our computer has three options for finding IP address or addresses for "google.com".
1.Prima option-hosts file in C :/ windows / system32 / drivers

/ etc / hosts
2.A second option - private DNS (server, router)
3.A third option - public DNS servers (OpenDNS, Google DNS).

Wherever you find the IP address for "google.com" our computer stops and no longer see the other variants. For example, if the IP address found for "google.com" in the hosts file, it does not go and that public private DNS to confirm the validity of those addresses.
Thus we can to fool the PC, we can tell him anything, he will believe anything found in the hosts file.[9]

### 2.1.7 System Reconfiguration Attacks

Modify settings on a user's PC for malicious purposes. For example: URLs in a favorites file might be modified to direct users to look alike websites. For example: a bank website URL may be changed from "bankofabc.com" to "bancofabc.com".**[10]**

### 2.1.8 Data Theft

Data theft is the act of stealing computer-based information from an unknowing victim with the intent of compromising privacy or obtaining confidential information. Data theft is increasingly a problem for individual computer users, as well as big corporate firms.

There is more than one way to steal data. Some popular methods are listed below:[11]

- E-commerce: You should make sure that your data is safe from prying eyes when you sell or buy things on the Web. Carelessness can lead to leaking your private account information.
- Password cracking: Intruders can access your machine and get valuable data if it is not password-protected or its password can be easily decoded (weak password).
- Eavesdropping: Data sent on insecure lines can be wiretapped and recorded. If no encryption mechanism is used, there is a good chance of losing your password and other private information to the eavesdropper.
- Laptop theft: Increasingly incidents of laptop theft from corporate firms occur with the valuable information stored in the laptop being sold to competitors. Carelessness and lack of laptop data encryption can lead to major losses for the firm**.**

### 2.1.8 DNS-Based Phishing Attacks

Domain Name System (DNS)-based **phishing** or hosts file modification is called Pharming. The requests for URLs or name service return a bogus address and subsequent communications are directed to a fake site when the hackers tamper a company’s host files or domain name. As a result, users remain unaware about the fraud website controlled by hackers.[12]

### 2.1.9 Man-**in-the-Middle Phishing**:

Abbreviated as **MITMA**, a **man-in-the-middle attack** is an attack where a user gets between the sender and receiver of information and sniffs any information being sent. In some cases, users may be sending unencrypted data, which means the man-in-the-middle (MITM) can obtain any unencrypted information. In other cases, a user may be able to obtain information from the attack, but have to unencrypted the information before it can be read. In the picture below is an example of how a man-in-the-middle attack works. The attacker intercepts some or all traffic coming from the computer, collects the data, and then forwards it to the destination the user was originally intending to visit. Shown in figure (3):[13]

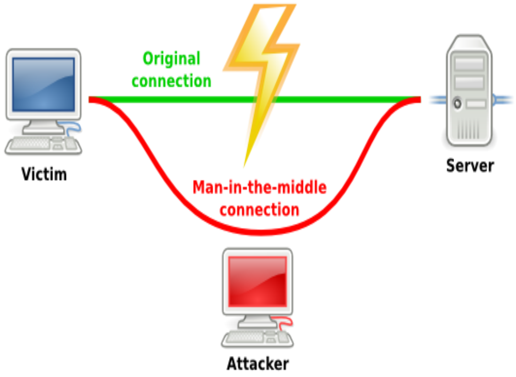


Figure (3): shows Man in Middle Attack

### 2.1.10 SEO poisoning (search poisoning)

SEO poisoning, also known as search poisoning, is an attack method in which cybercriminals create malicious websites and use search engine optimization tactics to make them show up prominently in search results. The sites are associated with terms that large numbers of people are likely to be using in searches at any given time, such as phrases related to holidays, news items and viral videos. According to Web sense Security Labs, up to a quarter of the first page of search results for trending topics are linked to malicious websites.

The attackers create websites with names and descriptions associated with popular or trending topics. For example, in the weeks leading up to Halloween, the attackers might launch sites offering free templates for Halloween costumes; in the weeks or months leading up to Christmas, they might launch holiday recipe sites. The sites might be devoid of relevant content or might feature content stolen from valid sites. The real purpose, however, is to infect visitors with malware or fraudulently access sensitive information to be used for identity theft. Malware on the site may coopt the visitor's computer for a botnet or install a Trojan horse to steal login information. Another ploy is to present the user with a product that they think they are purchasing to access their credit card details. [14]

## 2.2 Related Works

There are different studies which focus on Phishing attacks in KSA:

In [15] research was conducted to introduce and analyze a high level (a country-based) anti-phishing countermeasure implemented in Saudi Arabia. An investigation was carried out to examine whether this countermeasure is effective against phishing scenarios. The countermeasure is considered effective when Phishing websites are reached by users who surf the Internet inside Saudi Arabia whereas it is ineffective when the websites are reached by users who surf the Internet from outside Saudi Arabia.

In [16] study, discuss and propose the phishing attack stages and types, technologies for detection of phishing web pages, and conclude with some important recommendations for preventing phishing for both consumer and company that can be taken to reduce vulnerabilities to phishing attacks.

In [17] the purpose of this paper is to report the findings of the study commissioned by the Communications and Information Technology Commission to ascertain the magnitude of spam in the Kingdom of Saudi Arabia and formulate a comprehensive multi-pronged solution for handling spam in Saudi Arabia based upon best international practices, current situation and national requirements. It is only focus on determining the current state of spam in KSA, focusing on obtaining a good understanding of the nature and prevalence of spam within Saudi Arabia. This information will then form the basis upon which the anti-spam national strategy framework will be based.

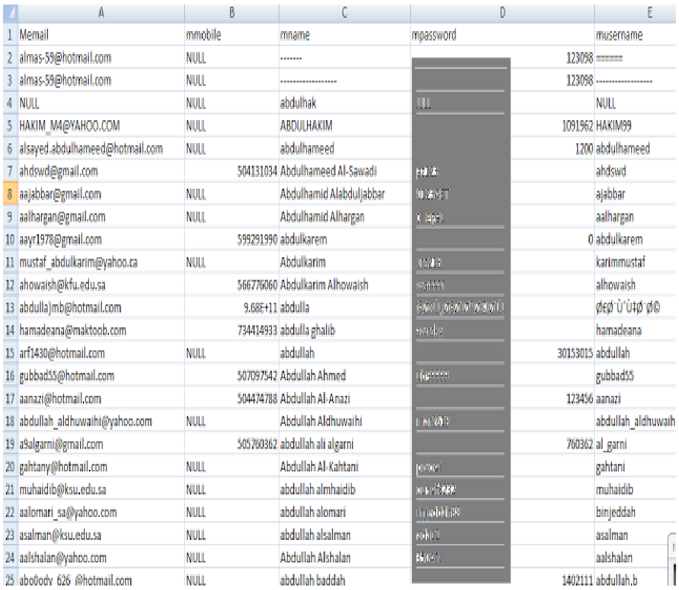
In [18] This paper discusses the stand of Saudi Arabian government against cyber crime and its IT act. It analyzes the cybercrime in the Kingdom and the anti-cyber crime law. It shows that Saudi Arabia was ranked first as the most vulnerable of the Gulf countries to fall victim to cyber-crimes, such as website hacking. It shows that most of the people in KSA know about cyber crime but very less is aware of the associated legislation to combat these crimes. Therefore, in KSA it has been clear how computer crimes can affect people live especially for those financial crimes. Although, the information security is increased but also the unauthorized access for example were dramatically increased. It also concludes that knowing the laws of computer crimes should be considered the first solution to reduce them.

## 2.3 Examples of Phishing Attacks in KSA Universities

Due to the frequent use by Saudi students and professors of computer networks for learning and teaching, universities have a large degree of exposure to cyber attacks.

“At the end of May 2015, a hacker in Saudi Arabia claimed to have hacked and stolen information from a Saudi university’s network,” he explains, “including the personal details, academic results and schedules of 4,000 university students.”[19]

In 2012, The Official Website of **King Saud University (KSU)** , is a public university located in Riyadh, Saudi Arabia hacked by some unknown Hacker. Database of 812 Users hacked from http://printpress.ksu.edu.sa/ and dumped on Internet by Hacker on a file sharing site including Mail address list, mobile phones and passwords.[20]. See figure (4):



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Memail | mmobile | mname | mpassword | musername |
| 2 | almas-59@hotmail.com | NULL | ------- | [redacted] 123098 | ====== |
| 3 | almas-59@hotmail.com | NULL | ----------------- | [redacted] 123098 | ----------------- |
| 4 | NULL | NULL | abdulhak | [redacted] | NULL |
| 5 | HAKIM_M4@YAHOO.COM | NULL | ABDULHAKIM | [redacted] 1091962 | HAKIM99 |
| 6 | alsayed.abdulhameed@hotmail.com | NULL | abdulhameed | [redacted] 1200 | abdulhameed |
| 7 | ahdswd@gmail.com | 504131034 | Abdulhameed Al-Sawadi | [redacted] | ahdswd |
| 8 | aajabbar@gmail.com | NULL | Abdulhamid Alabduljabbar | [redacted] | ajabbar |
| 9 | aalhargan@gmail.com | NULL | Abdulhamid Alhargan | [redacted] | aalhargan |
| 10 | aayr1978@gmail.com | 599291990 | abdulkarem | [redacted] 0 | abdulkarem |
| 11 | mustaf_abdulkarim@yahoo.ca | NULL | Abdulkarim | [redacted] | karimmustaf |
| 12 | ahowaish@kfu.edu.sa | 566776060 | Abdulkarim Alhowaish | [redacted] | alhowaish |
| 13 | abdullajmb@hotmail.com | 9.68E+11 | abdulla | [redacted] | ØεØ`Ù`Ù‡Ø`Ø© |
| 14 | hamadeana@maktoob.com | 734414933 | abdulla ghalib | [redacted] | hamadeana |
| 15 | arf1430@hotmail.com | NULL | abdullah | [redacted] 30153015 | abdullah |
| 16 | gubbad55@hotmail.com | 507097542 | Abdullah Ahmed | [redacted] | gubbad55 |
| 17 | aanazi@hotmail.com | 504474788 | Abdullah Al-Anazi | [redacted] 123456 | aanazi |
| 18 | abdullah_aldhuwaihi@yahoo.com | NULL | Abdullah Aldhuwaihi | [redacted] | abdullah_aldhuwaih |
| 19 | a9algarni@gmail.com | 505760362 | abdullah ali algarni | [redacted] 760362 | al_garni |
| 20 | gahtany@hotmail.com | NULL | Abdullah Al-Kahtani | [redacted] | gahtani |
| 21 | muhaidib@ksu.edu.sa | NULL | abdullah almhaidib | [redacted] | muhaidib |
| 22 | aalomari_sa@yahoo.com | NULL | abdullah alomari | [redacted] | binjeddah |
| 23 | asalman@ksu.edu.sa | NULL | abdullah alsalman | [redacted] | asalman |
| 24 | aalshalan@yahoo.com | NULL | Abdullah Alshalan | [redacted] | aalshalan |
| 25 | abo0ody_626_@hotmail.com | NULL | abdullah baddah | [redacted] 1402111 | abdullah.b |

Figure (4): King Saud University database hacked

## 3. MAIN STUDY

### 3.1 Participants

The Questionnaire was administrated to fifty (50) participants, who were undergraduate students, from Prince Sattam Bin A bdulaziz University. Age ranged from 18 to 22 with the gender 40 male and 10 female. Each participants took part in the survey on a fully voluntary basis. A summary of the demographics of the participants in the main study is shown in table ( 1 ).

Table (1): Participant Demographics in The Main Study.

| Characteristics | Total |
|---|---|
| ***Sample Size*** | 50 |
| ***Gender*** | |
| *Male* | 40 |
| *Female* | 10 |
| ***Age Range ( From 18 To 22)*** | 50 |
| ***Average Hours Per Week on The Internet*** | |
| *0—5* | 3 |
| *6—10* | 5 |
| *11—15* | 5 |
| *16—20* | 9 |
| *20 +* | 28 |

### 3.2 Procedure

The questionnaire was handled out to participants in-person by the researcher. First, the nature of the research was explained to each participant individually. They were told that they were free to withdraw from the study at any time without having to give a reason for withdrawing. Then participants were asked to complete the questionnaire, they were asked whether or not they know what the term " Phishing Attack" means, and if they know how to protect against phishing attack, they were asked also if they know the difference between http and https protocols, if they were familiar with the term " Social Engineering", if they were check the URL after the opening new web site link. The individual participciant was given 15 min to complete the questionnaire. After completing the questionnaire, participants were thanked for their valuable time and effort in taking part in the study.

### 3.3 Results

We collected 50 valid responses and our sample consisted of students from five different scientific fields of study: Computer science department, business management, Arabic, Islamic studies, and mathematic department. As shown in table (2) and figure (5) which were part of the Prince Sattam bin Abdulaziz University.

Table (2): Show five different scientific fields of study

| | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|
| **Computer** | 13 | 26.0 | 26.0 | 26.0 |
| **Math.** | 7 | 14.0 | 14.0 | 40.0 |
| **Management** | 10 | 20.0 | 20.0 | 60.0 |
| **Arabic** | 10 | 20.0 | 20.0 | 80.0 |
| **Islamic** | 10 | 20.0 | 20.0 | 100.0 |
| **Total** | 50 | 100.0 | 100.0 | |

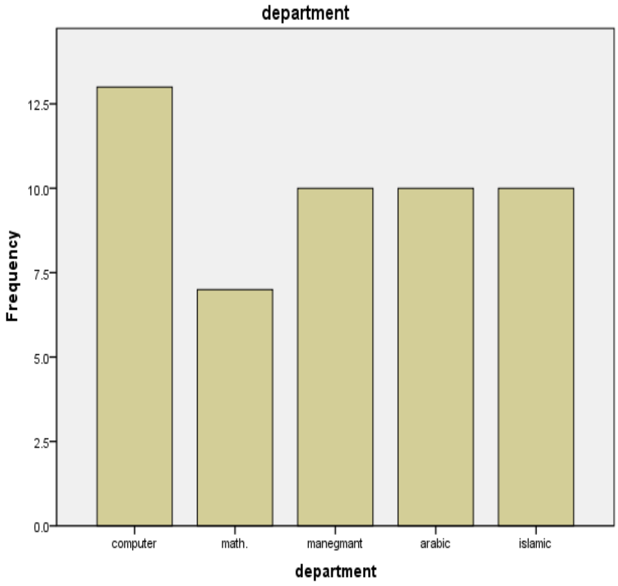


Figure (5) : Frequency of five different department of study

Most of our test subjects, 58.0% stated that they were took 21+ hours per week of internet usage, while 42% were took ≤ 20 hours per week of internet usage as shown in Table (3) and figure (6).

Table (3): The average hours per week of internet experience

| *Average Hours* | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|
| 0-5 | 3 | 6.0 | 6.0 | 6.0 |
| 6-10 | 3 | 6.0 | 6.0 | 12.0 |
| 11-15 | 4 | 8.0 | 8.0 | 20.0 |
| 16-20 | 11 | 22.0 | 22.0 | 42.0 |
| +21 | 29 | 58.0 | 58.0 | 100.0 |
| Total | 50 | 100.0 | 100.0 | |

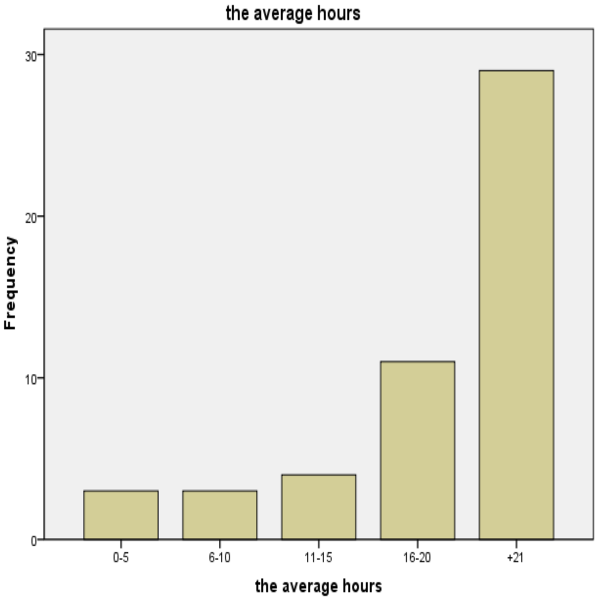


Figure (6): The Average hours per week of internet experience

Table (4) and figure (7) shows the students answer that if they were familiar with the term "Phishing Attack" or not, 50% answered that they were familiar with the term phishing attack and most of them in computer science department, while 50% were not familiar with this term.

Table (4): Student's familiarity with the term Phishing Attack

| *Answer* | | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|---|
| Valid | yes | 25 | 50.0 | 50.0 | 50.0 |
| | no | 25 | 50.0 | 50.0 | 100.0 |
| | Total | 50 | 100.0 | 100.0 | |

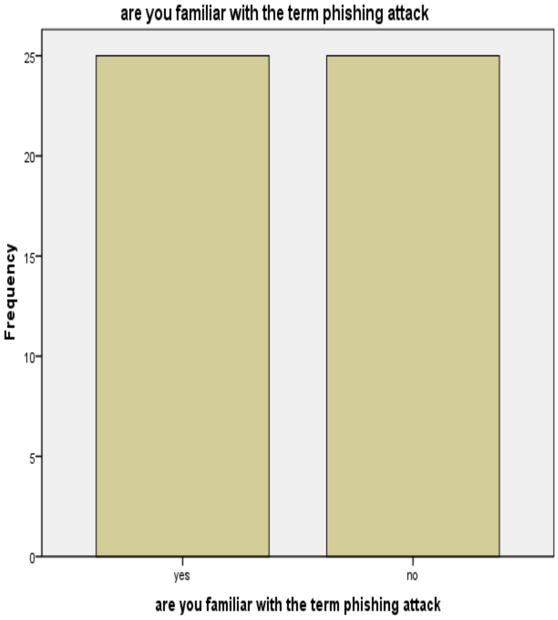


Figure (7): Student's familiarity with the term Phishing Attack

Table (5) and figure (8) shows the students answer that if they know how to protect themselves against phishing attack, 62% were not know how to protect themselves against phishing attacks, while 38% were know.

Table (5): students that were know how to protect themselves against phishing attack

| *Answer* | | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|---|
| Valid | yes | 19 | 38.0 | 38.0 | 38.0 |
| | no | 31 | 62.0 | 62.0 | 100.0 |
| | Total | 50 | 100.0 | 100.0 | |

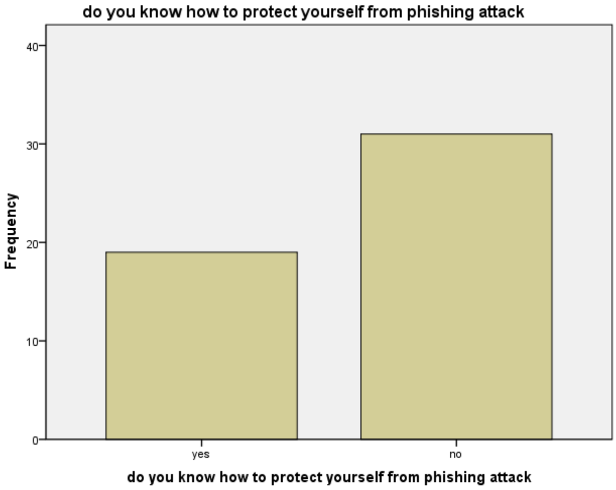


Figure (8): students that were know how to protect themselves against phishing attack

Table (6) and figure (9) shows the students answer that if they were know the difference between http and https protocols, 78% were not know the difference between http and https protocols , while 22% were know.

Table (6): students that were know the difference between http and https protocols

| Answer | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| Yes | 11 | 22.0 | 22.0 | 22.0 |
| No | 39 | 78.0 | 78.0 | 100.0 |
| Total | 50 | 100.0 | 100.0 | |

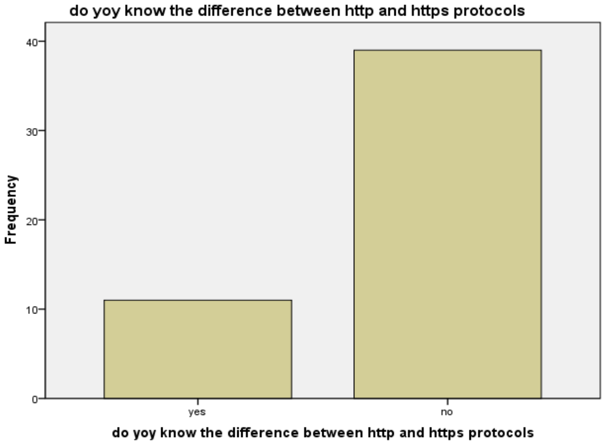


Figure (9): students that were know the difference between http and https protocols

Table (7) and figure (10) shows the students answer that if they were familiar with the term " Social Engineering" or not, 72% were not familiar with the term "social engineering", while 28% were familiar with the term" social engineering".

Table (7): students that were familiar with the term social engineering

| *Answer* | | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|---|
| Valid | yes | 14 | 28.0 | 28.0 | 28.0 |
| | no | 36 | 72.0 | 72.0 | 100.0 |
| | Total | 50 | 100.0 | 100.0 | |

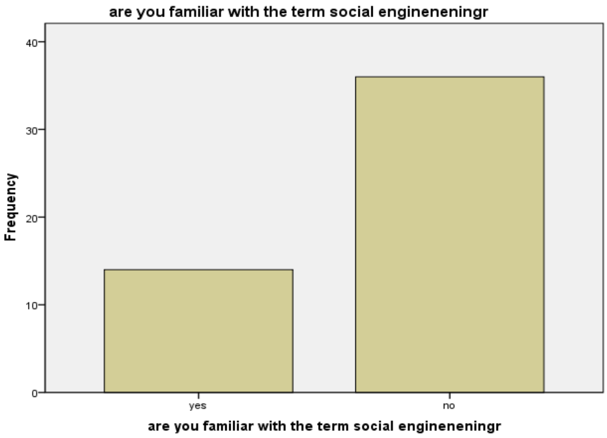


Figure (10): students that were familiar with the term social engineering

Table (8) and figure (11) shows the students answer that if they check the URL after the opening new website link or not, 82% were not check the URL after the opening new website link, 18% check the URL after the opening new website link.

Table (8): students that were check the URL after the opening new website link

| *Answer* | | *Frequency* | *Percent* | *Valid Percent* | *Cumulative Percent* |
|---|---|---|---|---|---|
| Valid | yes | 9 | 18.0 | 18.0 | 18.0 |
| | no | 41 | 82.0 | 82.0 | 100.0 |
| | Total | 50 | 100.0 | 100.0 | |

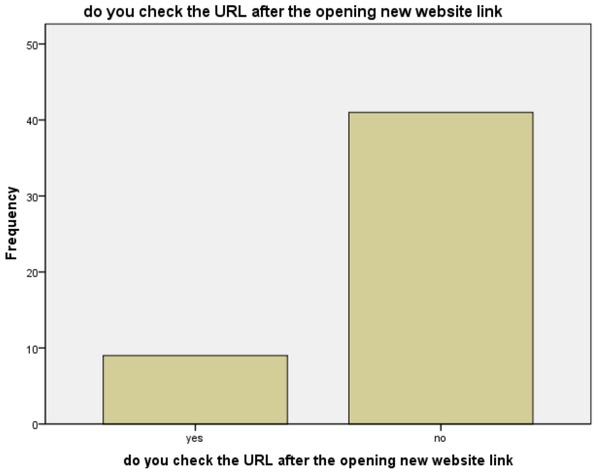


Figure (11): students that were check the URL after the opening new website link

## 4. CONCLUSION

Some of the interesting results our researches are:

- 58.0% stated that they were took 21+ hours per week of internet usage, while 42% were took ≤ 20 hours per week of internet usage.
- 50% answered that they were familiar with the term phishing attack and most of them in computer science department, while 50% were not familiar with this term.
- 62% were not know how to protect themselves against phishing attacks, while 38% were know.
- 78% were not know the difference between http and https protocols , while 22% were know.
- 72% were not familiar with the term "social engineering", while 28% were familiar with the term" social engineering".
- 82% were not check the URL after the opening new website link, 18% check the URL after the opening new website link.

## 5. LIMITATIONS

- The study was only conducted on the College of Science and Humanity Studies At Sulail.
- The study was conducted on a student between 18 and 22 years old.
- Students were not given training on phishing attack.

## 6. RECOMMENDATIONS

Educating saudian students on phishing techniques is an important aspect of internet security, if students were educated about phishing, it can help them understand the methods used to differentiate whether a website is a illegitimate site or a phishing site.

## 7. ACKNOWLEDGMENTS

Our thanks to the Staff of the College of Science and Humanity Studies At Sulail, Prince Sattam bin abdulaziz university for their contributions and comments.

## 8. REFERENCES

[1] Markus Jakobsson and Steven Myers. Phishing and countermeasures: understanding the increasing problem of electronic identity theft. John Wiley & Sons, Inc., 2007.

[2] Anti Phishing Working Group. Origins of the word "phishing". Available at: http://www.antiphishing.org/word_phish.htm accessed on 24/11/2016.

[3] Phishing - word spy. Available at: http://www.wordspy.com/words/phishing.asp, accessed on 27/11/2016.

[4] Available at: http://www.pcworld.com/article/135293/article.html, Accessed on 13/12/2016.

[5] Available at: http://www.phishing.org/phishing-techniques, Accessed on 13/12/2016.

[6] Available at: http://www.innovateus.net/science/what-are-different-types-phishing-attacks, Accessed on 15/12/2016.

[7] Available at: https://www.owasp.org/index.php/Session_hijacking_attack, Accessed on 16/12/2016.

[8] Available at: https://usa.kaspersky.com/internet-security-center/threats/trojans#.WF5Xhn1S6Aw, Accessed on 17/12/2016.

[9] Available at: http://en.videotutorial.ro/otravirea-dns-ului-metoda-folosita-frecvent-de-hackeri, Accessed on 17/12/2016.

[10] Available at: http://www.pcworld.com/article/135293/article.html, Accessed on 17/12/2016.

[11] Available at : https://cybercrime.org.za/data-theft, Accessed on 19/12/2016.

[12] Available at: http://www.innovateus.net/science/what-are-different-types-phishing-attacks, Accessed on 21/12/2016.

[13] Available at: http://www.computerhope.com/jargon/m/mitma.htm, accessed on: 23/12/2016.

[14] Available at: http://whatis.techtarget.com/definition/search-poisoning, Accessed on: 20/12/2016.

[15] Abdullah M.Alnajim, High Level Anti-Phishing Countermeasure: A case Study, IEEE Computer Society, 2011.

[16] Wajeb Gharibi, Some Recommended Protection Technologies for Cyber Crime on Social Engineering Techniques – Phishing, Journal of Communication and Computer, USA, Vol.8 No.7, 2011.

[17] Mishaal Abdullah Al-Kadhi, Assessment of the status of spam in the Kingdom of Saudi Arabia, Communications and Information Technology Commission, Saudi Arabia, 2011.

[18] Bushra M. Elamin, Cyber Crime in Kingdom of Saudi Arabia: The Threat Today and the Expected Future, Information and Knowledge Management, Vol.3 No12, 2013.

[19] Available at: http://www.al-fanarmedia.org/2015/12/arab-universities-are-vulnerable-to-cyber-attacks-experts-say/, Accessed on: 24/12/2016.

[20] Available at: http://thehackernews.com/2012/01/saudi-arabias-king-saud-university.html, Accessed on: 26/12/2016.

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 51-55, 2017, ISSN:-2319–8656

# Student Feedback Mining System Using Sentiment Analysis

R.Menaha
IT Department
Dr.Mahalingam College of Engineering and Technology
Pollachi, Tamilnadu, India

R.Dhanaranjani, T.Rajalakshmi, R.Yogarubini
IT Department
Dr. Mahalingam College of Engineering and Technology

**Abstract-** Academic industries used to collect feedback from the students on the main aspects of course such as preparations, contents, delivery methods, punctual, skills, appreciation, and learning experience. The feedback is collected in terms of both qualitative and quantitative scores. Recent approaches for feedback mining use manual methods and it focus mostly on the quantitative comments. So the evaluation cannot be made through deeper analysis. In this paper, we develop a student feedback mining system (SFMS) which applies text analytics and sentiment analysis approach to provide instructors a quantified and deeper analysis of the qualitative feedback from students that will improve the students learning experience. We have collected feedback from the students and then text processing is done to clean the data. Features or topics are extracted from the pre-processed document. Feedback comments about each topic are collected and made as a cluster. Classify the comments using sentiment classifier and apply the visualization techniques to represent the views of students. This proposed system is an efficient approach for providing qualitative feedback for the instructor that enriches the students learning.

**Keywords**- Students Feedback, Text processing, Clustering, Topic extraction, Sentiment analysis.

## 1. INTRODUCTION

Students provide feedback in quantitative ratings and qualitative comments related to preparation, contents, delivery methods, punctual, skills, appreciation, and learning experience. The delivery methods and preparation component refers to instructor’s interaction, delivery style, ability to motivate students, out of class support, etc. The content refers to course details such as concepts, lecture notes, labs, exams, projects, etc. The preparation refers to student’s learning experience such as understanding concepts, developing skills, applying acquired skills, etc. The paper correction refers to correction of mistakes and providing solutions to overcome it. The punctual refers to the class timing and assignment or record submission. The appreciation refers to the comments given when something is done perfectly. Analyzing and evaluating this qualitative data helps us to make better sense of student feedback on instruction and curriculum.

Recent methods for analyzing student course evaluations are manual and it mainly focuses on the quantitative feedback. It does not support for deeper analysis. This paper focus on providing qualitative and quantitative feedback to analyze and provide better teaching to improve the student’s performance.

The paper will be structured as follows: Section 2 will review the techniques used in text processing and sentiment analysis approach in the background. Section 3 will describe the related works of the current research about the student feedback mining system. Section 4 will provide the proposed system of this paper. Section 5 will have experiments and future works to be implemented and finally we concluded in section 6.

## 2. BACKGROUND

Text mining approach is useful in the sentiment analysis process. In this section, we provide a brief description about the methods that we adopted to extract the keywords from the students feedback document.

### 2.1 Tokenization

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded [4].

### 2.2 Stop word removal

Stop words are words which are filtered out before or after processing of natural language data. These words are removed to extract only the meaningful information [4]. The list of stop words may be 'the, is, at, which, on, who, where, how, hi, before, after' etc.

### 2.3 Clustering

Clustering is the process of making a group of abstract objects into classes of similar objects [4]. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

### 2.4 Classification

Data classification is the process of organizing data into categories for its most effective and efficient use [4]. A well-planned data classification system makes essential data easy to find and retrieve. This can be of particular importance for risk management, legal discovery, and compliance.

### 2.5 Sentiment Analysis

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied in review and social media for a variety of applications, ranging from marketing to customer service. Sentiment analysis aims in determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication.

## 3. RELATED WORKS

Francis F. Balahadia; Ma. Corazon G. Fernando; Irish C. Juanatas [7], developed the teacher's performance evaluation tool using opinion mining with sentiment analysis. They collected the feedback from the students and identified the strength and weakness of the particular teacher. They evaluated the qualitative and quantitative data and provided sentiment score of the teacher in a school.

Nabeela Altrabsheh; Mihaela Cocea; Sanaz Fallahkhair [13], reduced the stress and time consuming of analyzing the feedbacks while teaching. To overcome it, they processed automatically using sentiment analysis. They used Support Vector Machine(SVM) to provide a higher level of pre-processing.

Alok Kumar; Renu Jain Feedback [1], proposed an automatic evaluation system based on sentiment analysis. Feedback is collected in the form of running text and sentiment analysis is performed to identify important aspects using supervised and semi supervised machine learning techniques.

## 4. PROPOSED SYSTEM

We proposed a system to mine the feedback given by the students and obtain knowledge from that and present that information in qualitative way. Feedback was collected for a course; those feedbacks were pre-processed using text processing techniques. In preprocessing, the feedback files are generated as a flat file. The flat file is tokenized into sentences and the keywords are listed after removing the stop words. We have identified the frequency of each word and extract the topic which has the highest frequency count. Similar comments in each topic are clustered and then the clustered words are classified into positive or negative comments. The classified comments are generated as a chart for easy visualization.

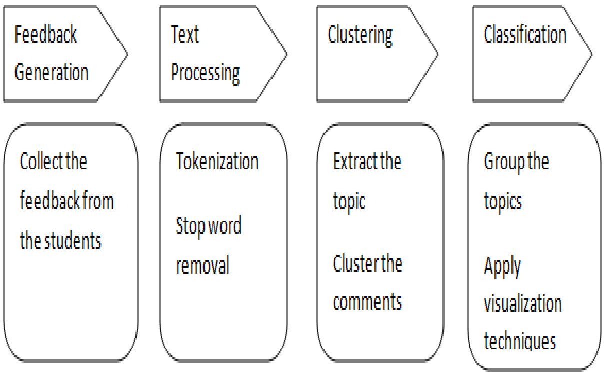


Fig.:1 Block diagram of proposed system

### 4.1 Keyword Extraction

To clean the data, the collected feedback is subjected to tokenization and stop word removal. The following sequence of steps shows that how do we performed preprocessing.

Step 1: Read each student feedback document and append it into a document D.

Step 2: Tokenize D based on. or, to identify the sentence.

Step 3: On each sentence, remove the stop words.

Step 4: Update the document D.

Step 5: Now the document D is removed from stop words.

### Topic Extraction

From the preprocessed document, the parts of a sentence like adjectives, verbs, adverbs, pronouns, nouns, proper noun etc., is removed to identify the topics available in the student feedback. The topic might be teaching, project, communication, interaction, punctuality etc., The following algorithm we have used for the extraction of topics from the feedback document.

```
Input: feedback collection
Output: Topics
A[] ← set of adjectives,verbs,objects,etc.,
D ← preprocessed feedback file
Ch ← empty character set
δ ← Threshold
T ← empty file
//Remove the adjectives,verbs,objects,etc.,
While( D!=EOF)
        Do Tokenization
        Ch= word in D
        If Ch==A then
                Remove  Ch from D.
End While
//Topic Extraction
While(D!=EOF)
    Ch=Word from D
    Count ← The frequency of Ch in D
    If Count > δ
        T ← Append Ch
End While
//T contain list of topics
```

Fig 2.:Topic Extraction Algorithm

We have used a threshold δ to limit the number of topics. The frequency of each word in D is counted by using the equation 1. If a word exceeds the δ then it is identified as topic.

$$\mu(w)= \sum \frac{N(w)}{T}$$

Where µ(w) is the frequency of the word w in document D. N(w) is the number of times the word w appeared in D. And T is the total number of words in D.

### 4.2 Clustering

Each Topic from T is identified as cluster. Read each student feedback document for a course and identify the comments given by the students related to the topic and make it as cluster. We have used pattern matching for comparing the topics with the student feedback document. Semantic similarity can also be used to compute the relatedness between the topic and the feedback comments. The identified topics which are made as cluster by using our approach are Faculty interaction, Punctual, Project, Lab, etc.

### 4.3 Classification

Read each cluster comments and classify those comments as either positive or negative. An array list is maintained for positive and negative comments. The student feedback comments are compared with the list to conclude that comment is either positive or negative on that topic. The following algorithm shows the sequence of steps employed to perform classification.

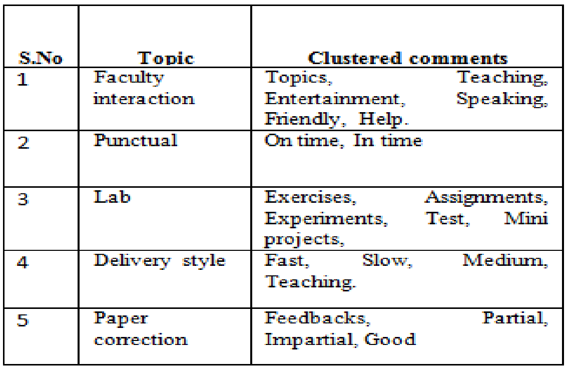
```
Input: Clusters
Output: features
P[] ← set of positive words(good,best,etc.,)
NA[] ← set of negative words(bad,poor,etc.,)
N← No of clusters.
C1 to CN← cluster sets , Ch ← empty
character set
//Classification
Classify(C1 to CN)
While(i<N)
    Read Each Cluster from C1 to CN
    Ch← Read each comment from a cluster
    If Ch contains P then
        Positive_count++
        Else if Ch contains NA then
                Negative_count++
End While
```

Fig.:3 Classification Algorithm

## 5. EXPERIMENTS

In this paper, we build a Student Feedback Mining System to analyze topics and their sentiments from student generated feedback. The feedbacks are collected from the students for a single course for easy evaluation and to improve student's learning file.

We tokenized the feedbacks into sentences and removed the stop words. Topics were extracted from the feedback document. The following table shows the extracted topics and comments related to it, which were made as cluster is shown in the table 1. From each cluster, the comments are classified as positive or negative which is shown in the following table 2.



| S.No | Topic | Clustered comments |
|---|---|---|
| 1 | Faculty interaction | Topics, Teaching, Entertainment, Speaking, Friendly, Help. |
| 2 | Punctual | On time, In time |
| 3 | Lab | Exercises, Assignments, Experiments, Test, Mini projects, |
| 4 | Delivery style | Fast, Slow, Medium, Teaching. |
| 5 | Paper correction | Feedbacks, Partial, Impartial, Good |

Table 1 Clustering

| S.No | Topic | +ve | -ve |
|---|---|---|---|
| 1 | Faculty interaction | 90 | 10 |
| 2 | Punctual | 60 | 40 |
| 3 | Lab | 45 | 55 |
| 4 | Delivery style | 78 | 22 |
| 5 | Paper correction | 60 | 40 |

Table 2 Classification

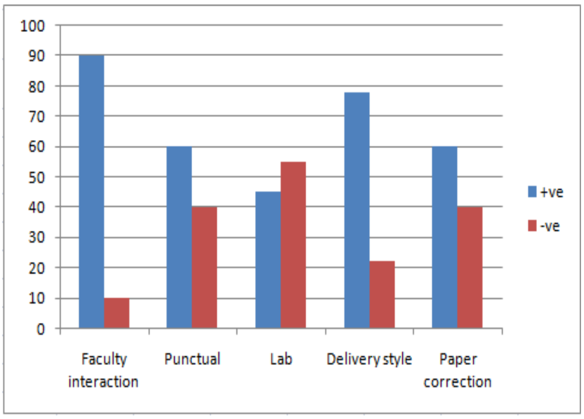


Fig.:4 Chart representation

## 5. DISCUSSIONS AND FUTURE WORKS

As future work, the proposed system can be extended to include semantic similarity for clustering the student feedback .And also different visualization techniques would be used to project the results.

## 6. CONCLUSION

In this paper, a Student Feedback Mining System is build to analyze topics and their sentiments from student generated feedback. This system uses preprocessing, topic extraction, clustering, classification to represent the student views in a graphical way. This system will be useful to improve the students learning and instructor's methods of delivery.

## 7. REFERENCES

1. Alok Kumar; Renu Jain in ***Sentiment analysis and Feedback Evaluation***
2. AnshulMittal,ArpitGoel in***Stock prediction using twitter Sentiment analysis.***
3. Brennan, J. & Williams, R. (2004) ***Collecting and Using Student Feedback***. A Guide to Good Practice (LTSN, York).
4. Dhanalakshmi V in ***Opinion mining from student feedback data using supervised learning algorithms.***
5. Donovan, J., Mader, C. E., & Shinsky, J. (2012). ***Constructive Student Feedback: Online vs. Traditional Course Evaluations.***
6. Elaine Keane & Iain Mac Labhrainn , ***Obtaining Student Feedback on Teaching & Course Quality*** , CELT, April 2005
7. Francis F. Balahadia; Ma. Corazon G. Fernando; Irish C. Juanatas in ***Teacher's performance evaluation tool using opinion mining with sentiment analysis***
8. GokarnIlaNitin ,Asst.Prof.GottipatiSwapna , Prof.VenkyShankararaman in ***Analyzing Educational Comments for Topics and Sentiments: A Text Analytics Approach.***
9. HarshaliP.Patil,MohammadAtiquein ***Sentiment analysis for Social media.***
10. K. P. Mohanan, ***the place of student feedback in teaching evaluation*** http://www.cdtl.nus.edu.sg/publications/studfeedback/StudFeedback_Teach Quality.pdf
11. Mark McGuire,ConstanceKampf Aarhus University in ***Using Social Media Sentiment Analysis to Understand Audiences: A New Skill for Technical Communicators?***
12. M.S.Neethu,R.Rajasree in ***Sentiment analysis in twitter using machine learning techniques.***
13. Nabeela Altrabsheh; Mihaela Cocea; Sanaz Fallahkhair in ***Sentiment Analysis: Towards a Tool for Analysing Real-Time Students Feedback***
14. Tan Li Im, PhangWai San, Chin Kim On Center of Excellence in Semantic Agents Universities Malaysia in ***Rule-based Sentiment Analysis for Financial News***
15. Yao, Y., & Grady, M. L. (2005). ***How do faculty make formative use of student evaluation feedback?***: A multiple case study. *Journal of Personnel Evaluation in Education*, *18*(2), 107-126.
16. Zhao, Y., Karypis, G., & Du, D. Z. (2005). ***Criterion functions for document clustering*** (Doctoral dissertation, University of Minnesota.).
17. Object Refinery Limited. "JFreeChart". JFreeChart. Jan 2009. Web. 17 Apr. 2015 http://www.jfree.org/jfreechart/
18. The Apache Software Foundation. "Apache Poi Project". Apache POI. Jan 2005. Web. 17 Apr. 2015. https://poi.apache.org/

### WEB REFERENCES

1. ***URL:****http://stackoverflow.com*
2. ***URL:****http://javaworld.com*
3. ***URL:****http://programcreek.com*
4. *www.wikipedia.com*

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 56-60, 2017, ISSN:-2319–8656

# Iris Recognition Using Modified Local Line Directional Pattern

S. Nithya
Dr. Mahalingam College of Engineering and Technology
Coimbatore, India

N. Madhuvarshini
Dr. Mahalingam College of Engineering and Technology
Coimbatore, India

V. Nivetha
Dr. Mahalingam College of Engineering and Technology
Coimbatore, India

E. Indira
Dr. Mahalingam College of Engineering and Technology
Coimbatore, India

**Abstract**: In recent years, as one of the emerging biometrics technologies, iris recognition has drawn wide attentions. It has many advantages such as uniqueness, low false recognition rate and so it has broad applications. It mainly uses pattern recognition and image processing methods to describe and match the iris feature of the eyes, and then realizes personal authentication. In image processing field, local image descriptor plays an important role for object detection, image recognition, etc. Till now, a lot of local image descriptors have been proposed. Among all kinds of local image descriptors, it is well-known that LBP is a popular and powerful one, which has been successfully adopted for many different applications such as face recognition, texture classification, object recognition, etc. Currently, a new trend of the research on LBP is to encode the directional information instead of intensity information. LLDP is an LBP-like descriptor that operates in the local line-geometry space. We used modified finite radon transform (MFRAT) to implement the LLDP descriptor for iris recognition and obtained 80.03% accuracy.

**Keywords**: Iris recognition; LLDP; line response; feature extraction; CBIR

## 1. INTRODUCTION

The internet has grown rapidly and due to the improvement of internet, the availability of the number of images has also been increased. Hence, the demand for efficient search and retrieval has increased. Although many researches have been done in the field of image search and retrieval, there are still many challenging problems to be solved [5]. The problem of storage and manipulation of images still exist. To overcome these problems, almost all images are represented in compressed formats like JPEG and MPEG [6]. Early techniques of image retrieval were based on manual textual annotation of images. It is difficult to characterize images by interpretation of what we see. Hence, color, shape, and texture based image retrieval started gaining prominence [7].

In the early 1990's, Content Based Image Retrieval (CBIR) [8], [9] was proposed to overcome the limitations of text based image retrieval. Increase in communication bandwidth, information content and the size of the multimedia datasets has given rise to the concept of CBIR [7]. The images can be retrieved by contents of images like color, texture and shape. This system is called CBIR, which is an intensive and difficult area [6]. This technique enables a user to extract an image based on a query from a dataset containing a large amount of images.

A very fundamental issue in designing a CBIR system is to select the image features that best represents the image contents in a dataset [7]. The effective CBIR system needs efficient extraction of low level features like color, texture and shapes for indexing and fast query image matching with indexed images for the retrieval of similar images. Features are extracted from images in pixel and compressed domains [6]. In CBIR, the features of the image are extracted efficiently and then represented in a particular form to be used effectively in the matching of images [6]. Here, floating point data is largely used. Texture analysis, has been widely employed in applications such as remote sensing, medical image analysis, document analysis, face identification, fingerprint identification, iris recognition [2].

Iris recognition is an emerging personal identification method in biometrics. Iris recognition is using the person's iris to identify or verify who the person is. Iris is a thin, circular structure in the eye with a diameter of only about 10 mm. Iris image acquisition is a key issue in iris recognition, as the quality of the captured image greatly affects the performance of the overall system. The captured image should have high resolution and contrast to show the texture in detail.

## 2. LITERATURE REVIEW

### 2.1 LBP

The LBP was first introduced by Ojala et al [11]. LBP was introduced for texture classification [3]. LBP is a non-parametric method that captures the local structures of images [4]. The Local Binary Pattern method represents textures as the joint distribution of underlying microstructures, modeled via intensity differences in a pixel neighborhood [12].

A binary pattern was obtained by comparing the neighboring pixels ($N$ = 8) in a 3 × 3 window with the pixel in the center of them, as shown in Fig. 1 based on

$$LBP_{P,R} = \sum_{p=0}^{P-1} S(g_p - g_c)2^p \qquad (1)$$

$$\text{where, } S(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases}$$

$g_c$ is the gray value of the center pixel, $g_p$is the gray value of its neighbors, P is the number of neighbors, and R is the radius of the neighborhood.

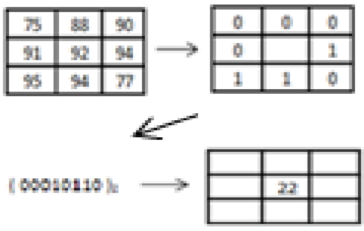


**Figure 1. The process of determining the decimal value for a pixel**

The LBP feature representation has been used in a wide range of texture classification scenarios and has proven to be highly discriminative. However, a restriction of LBP is its sensitivity to affine transformations [12].

## 2.2 LTP

As said by Subrahmanyam Murala, R. P. Maheshwari in [3], LBP was extended to a three-valued code called the LTP. Unlike LBP, LTP does not threshold the pixels into 0 and 1, rather it uses a threshold constant to threshold pixels into three values. It uses three-valued function and the binary LBP code is replaced by a ternary LTP code as:

$$f(x, p_c, T) = \begin{cases} +1, & x \geq p_c + T \\ 0, & |x - p_c| < T \\ -1, & x \leq p_c - T \end{cases} \qquad (2)$$

## 2.3 dLBP

According to Yılmaz Kaya, Omer Faruk Ertugrul and Ramazan Tekin, [2] in Directional Local Binary Pattern (dLBP), the pixels belonging to the same orientation are compared. The orientation may take $0^0$, $45^0$, $90^0$ and $135^0$ through clockwise. After determining the neighbors, the pixel values of them are compared with the value of the center pixel and the other process is carried out as the same as in traditional LBP [2]. The center pixels take the decimal value as its binary value is:

Pc = {S (P0 > P1), S (P1 > P2), S (P2 > P3), S (P3 > P4), S (P4 > P5), S (P5 > P6), S (P6 > P7), S (P7 > P0)} (3)

where, S denotes the comparison.

## 2.4 LLDP

As mentioned in the paper, Local line directional pattern for palmprint recognition by Yue-Tong Luo, Lan-Ying Zhao, Bob Zhang, Wei Jia, Feng Xue, Jing-Ting Lu, Yi-Hai Zhu, Bing-Qing Xu, the feature extraction in LBP structure descriptor is extended from intensity and gradient spaces to line space. This descriptor is very suitable for palmprint recognition. A new feature space i.e., the line feature space, instead of the gradient space or the intensity feature space, is used to compute robust code. Different coding schemes are exploited to produce LLDP descriptor, which is based on line direction numbers and is much better than all other existing LBP structure descriptors. It achieves better recognition performance than that of bit strings.

# 3. PROPOSED SYSTEM

## 3.1 Iris Description

Identification is a one-to-many comparison, which answers the question of "who the person is?" [1]. Iris recognition uses the unique patterns of the human iris to recognize a given individual [13].

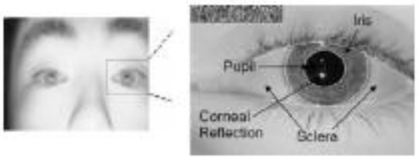


**Figure 2. Human Iris**

As reported by Jin-Suk Kang, image acquisition is a very important process as iris image with bad quality will affect the entire iris recognition process. For optimal performance, it is essential to acquire high-quality iris images.

## 3.2 Overview

In [16], Jia et al, proposed a principal line extraction method based on modified finite radon transform (MFRAT). To extract line responses, we use MFRAT. In this paper, to perform iris recognition, we follow three steps namely, indexing stage, searching stage and recognition stage. In the first step, we extract the features of the image such as mean, energy and entropy. These features are then stored in feature vector. Feature extraction is the process of locating an outstanding part, quality and characteristics in a given image [7]. We compare the features between the query image and the dataset images from the corresponding feature vectors in the searching stage. In recognition stage, to find the similarity between the query image and the dataset images, we used the distance measure. To measure the distance for similarity between the images, we use Manhattan Distance. Manhattan distance, which is computed by the sum of absolute differences between two feature vectors of images, is also called the City block distance metric [6].

**Step 1:** The line responses for four orientations namely $0^0$, $45^0$,$90^0$, $135^0$ are calculated for every non-overlapping region of the query image as well as the database images. From the line responses, senary code is computed.

**Coding strategy:** The index numbers of the minimum line response $r_4$ and the maximum line response $r_1$ are utilized for coding as:

**Senary code =** $r_4 * 6^1 + r_1 * 6^0$

**Step 2:** The next step is to compute the histogram. Histogram is one of the effective summaries of pixel intensity distribution over the image and it also provides a non-parametric statistical summary of information in an image [15].

**Step 3:** Feature extraction follows histogram generation. Here, features like mean, entropy and energy are extracted. Mean is the average of intensity values and describes the brightness of the image [6]. Energy is measured as a texture feature to calculate the uniformity of intensity level

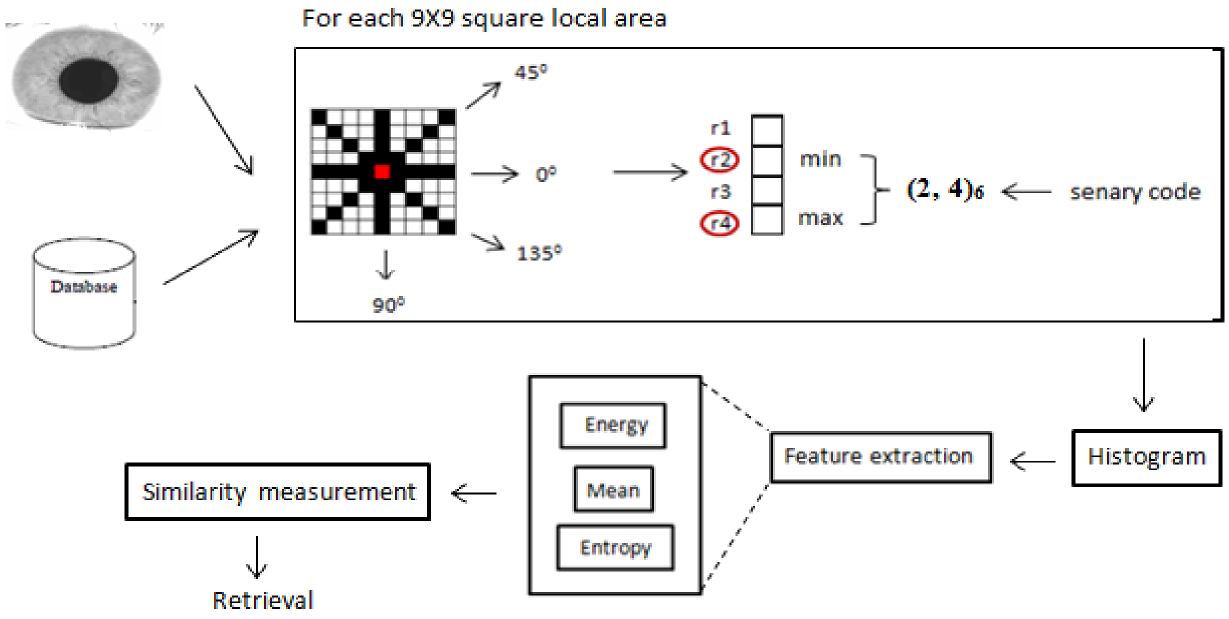


**Figure 3. LLDP for Iris recognition**

distribution [6]. Entropy measures the randomness of the distribution of intensity levels [6].

**Step 4:** To find the perfect match for the query image from the images in the dataset, similarity measure is used. Here, we use Manhattan distance as similarity measure. All the extracted features are considered in measuring similarity.

## 4. RESULTS AND DISCUSSION

As there are many algorithms for palmprint recognition, we tried using the LLDP descriptor with slight modifications for iris recognition. Since the iris lines are not much complicated, we reduced the number of orientations from 12 to 4. By reducing the number of orientations, time complexity gets reduced and hence, less space is enough to store these values. We reduced the size of every non-overlapping square region, thereby utilizing this reduction in time complexity. We converted every pixel values to 16-bit unsigned integer which can store huge values.

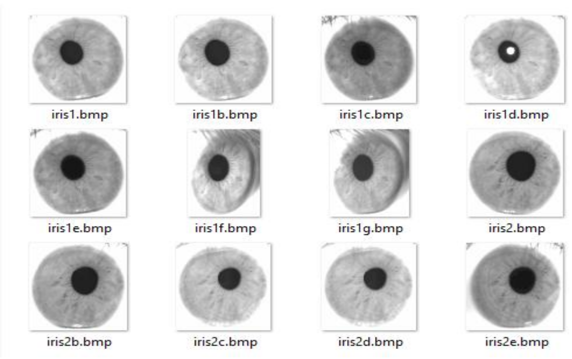


**Figure 4. Iris dataset**

Out of 254 images in the dataset, we took 12 images for testing purpose. We performed effective match for the query image by extracting some features like mean, energy and entropy out of the image. Fazal Malik and Baharum Baharudin say that the energy with high value shows the distribution of intensity values is for a small number of bins of histograms. If the value of entropy is high then the distribution is among greater intensity levels in the image. This measurement is the inverse of energy. The simple image has low entropy while the complex image has high entropy.

Euclidean distance is most commonly used for similarity measurement in image retrieval [6]. We used Manhattan distance because it is high in terms of precision and robust to outliers [6]. Retrieval effectiveness is the aim of most retrieval experiments. To measure the retrieval effectiveness, we use precision and recall. Precision and recall measures have been widely used for evaluating the performance of the CBIR system [7]. Because of simple calculations and easy interpretation of results, we opted for these measures. Precision is the measurement of the retrieved relevant images to the query of the total retrieved images [6]. Recall is the measurement of the retrieved relevant images to the total dataset images [6]. For example, a CBIR method for a query image retrieves totally 15 images with 9 relevant images out of totally 50 relevant images in dataset. Then the precision is 9/15= 60% and recall is 9/50 =18%. Since, precision and recall has shown different results, we used both the measures together to calculate the retrieval effectiveness but not either alone. However, these two measurements cannot be considered as complete accuracy for the effective image retrieval.

**Table 1. Average retrieval rate for various techniques**

| **Average Retrieval Rate** | |
|---|---|
| **Technique** | **Iris recognition** |
| LBP | 77.80 % |
| LTP | 79.86 % |
| LLDP | 79.97 % |
| Proposed system | 80.03 % |

**Table 2. Similarity measure using different distance measures for various techniques**

| Retrieval Accuracy | | |
|---|---|---|
| **Technique** | **Euclidean distance** | **Manhattan Distance** |
| LBP | 78.90 % | 75.64 % |
| LTP | 81.20 % | 82.13 % |
| LLDP | 78.19 % | 79.51 % |
| Proposed system | 79.68 % | 80.03 % |

**Table 3. Accuracy rate for different dimensions in iris recognition**

| Matrix size | Accuracy rate | | | |
|---|---|---|---|---|
| | **LLDP** | **LBP** | **LTP** | **Proposed** |
| 3X3 | 75.71 % | 72.57 % | 74.76 % | 76.13 % |
| 5X5 | 77.35 % | 74.13 % | 76.81 % | 78.34 % |
| 7X7 | 78.84 % | 75.87 % | 77.96 % | 79.1 % |
| 9X9 | 79.97 % | 77.8 % | 79.86 % | 80.03 % |
| 13X13 | 80.27 % | 78.01 % | 80.1 % | 80.29 % |

It has been observed that the accuracy rate is almost similar when the size of the square local area is fixed to be 9 or 13. Hence, we used 9 as the matrix size.

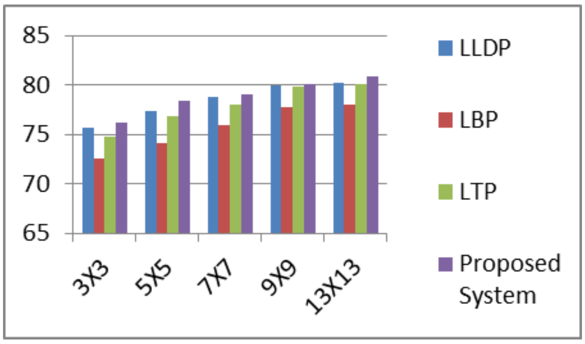


**Figure 5. Various dimensions of different techniques with their accuracy rate**

## 5. CONCLUSION

The LLDP descriptor showed better results for palmprint recognition in 13X13 dimension. On applying the same descriptor for recognizing iris, we obtained 80.29% accuracy. By performing slight modifications in the LLDP descriptor, and applying the same for iris recognition, we obtained similar accuracy 80.03% with 9X9 dimension. 12 directions with $15^0$ interval were considered in palmprint recognition. We achieved better results even by considering 4 directions with an interval of $45^0$ for recognizing iris. This reduction in number of orientations reduces space complexity. Number of computations also gets reduced and hence time complexity gets reduced.

## REFERENCES

[1] Yue-Tong Luo, Lan-Ying Zhao, Bob Zhang, Wei Jia, FengXue, Jing-Ting Lu, Yi-Hai Zhu, Bing-Qing Xu, Local line directional pattern for palm print recognition Science Direct Vol.50 (C) Feb (2016) 26 – 44.

[2] Yılmaz Kaya, Omer Faruk Ertugrul, Ramazan Tekin, Two novel local binary pattern descriptors for texture analysis Science Direct Vol. 34 (C) Sep (2015) 728 – 735.

[3] Subrahmanyam Murala, R. P. Maheshwari, Member, IEEE, and R. Balasubramanian, Member, IEEE Local Tetra Patterns: A New Feature Descriptor for Content-Based Image Retrieval IEEE Vol. 21 (5) May (2012) 2874 – 2884.

[4] Vinh Dinh Nguyen, Dung Duc Nguyen, Thuy Tuong Nguyen, Vinh Quang Dinh, Jae Wook Jeon, Support Local Pattern and Its Application to Disparity Improvement and Texture Classification, IEEE, Vol. 24 (2) Feb (2014) 263 – 277.

[5] Mohsen Zand, Shyamala Doraisamy, Alfian Abdul Halin, Mas Rina Mustaffa, Texture classification and discrimination for region-based image retrieval, Science Direct, Vol. 26 (2015) 305 – 316.

[6] Fazal Malik , Baharum Baharudin, Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain, Science Direct, Vol. 25 (2013) 207 – 218.

[7] Kommineni Jenni, Satria Mandala, Mohd Shahrizal Sunar, Content Based Image Retrieval Using Colour Strings Comparison, ScienceDirect**,** Vol. 50 ( 2015 ) 374 – 379.

[8] Badrinarayan Raghunathan, S.T. Acton, A content based retrieval engine for remotely sensed imagery, IEEE, April (2000) 161 - 165.

[9] Badrinarayan Raghunathan, S.T. Acton, A content based retrieval engine for circuit board inspection, International Conference on Image Processing, Oct (1999) 104 - 108.

[10] G. Deep, L. Kaur, S. Gupta, Directional local ternary quantized extrema pattern: A new descriptor for biomedical image indexing and retrieval, Science Direct, Vol.19 (2016) 1895 -1909.

[11] T. Ojala, M. Pietikainen, T. Maeenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE, Vol. 24 (7) July (2002) 971 – 987.

[12] Sebastian Hegenbart, AndreasUhl, A scale - and orientation – adaptive extension of Local Binary Patterns for texture classification, Science Direct, Vol. 48 (2015) 2633 – 2644.

[13] Jin-Suk Kang, Mobile iris recognition systems: An emerging biometric technology Science Direct 1 (2012) 475-484.

[14] Shahid Latif, Rahat Ullah, Hamid Jan, A Step towards an Easy Interconversion of Various Number Systems, IJECS-IJENS, Vol. 11 No: 02 April (2011) 86 – 91.

[15] Rama Murthy Garimella, Moncef Gabbouj, Iftikhar Ahmad, Image Retrieval: Information and Rough Set Theories, Science Direct, Vol. 54 ( 2015 ) 631 – 637.

[16] D.S. Huang, W. Jia, D. Zhang, Palmprint verification based on principal lines, Science Direct, Vol. 41 (4) (2008) 1316 – 1328.

17] IRIS DATASET: downloaded at
http://www.mae.cuhk.edu.hk/~cvl/main_database.htm

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 61-65, 2017, ISSN:-2319–8656

# Design and Reliability Analysis of Differential Resistance to Current Conversion Circuit for Biomedical Application of Gas Sensing

Zeinab Hijazi[1, 2], Daniele Caviglia[1], Hussein Chible[2], and Maurizio Valle[1]
[1]University of Genova, DITEN, COSMIC Lab, Italy
[2]Lebanese University, EDST, MECRL, Lebanon

**Abstract**: Gas sensing in biomedical applications shows a conversion of the concentration of the exhaled gas to a variation in resistance, so an electronic integrated interface circuit is required to analyse the exhaled gases, which are indications for many diseases. In this paper, a resistance to current conversion circuit based on differential biasing for Electronic nose (E-nose) breath analyser is presented. Over more than 5-decades (500Ω to 100MΩ) input resistance range, a precision, less than 1%, required by novel gas sensing system in portable applications, is preserved. Therefore, the proposed circuit obtains high accuracy under simulation. The outputs of the proposed Resistance to Current (R-to-I) conversion circuit achieve a percentage error below 0.25% under environment corners. The reliability of the proposed circuit is also investigated under the effect of process variations. In order to assess the correctness of the proposed architecture, the circuit was compared to similar solutions presented in literature where the proposed architecture attains a worst-case percentage error of 0.05%.

**Keywords**: resistive gas sensors; E-nose; portable applications; R-to-I conversion; process variation; environmental corners;

## 1. INTRODUCTION

Nowadays, breath analysis has become a patient friendly technique that allows rapid disease diagnosis and permits the early detection of impairment organs and/or other illness. Indeed, human breath contains thousands of volatile organic compounds (VOCs) that may be used as predictive biomarkers of various diseases [1]. Currently, exhaled breath analysis is proposed as a novel effective and alternative technique to blood or urine tests. This technique, which is applicable on all patients, allows early, real time status monitoring and diagnoses plenty of diseases without the need for any medical screening and with no limitations in supply [2]. Portable machines for continuous monitoring, and personal health care is a need. The electronic nose (E-nose) used as portable breath analyzer was developed to copy the human olfactory system. The latter is one of the human five senses. Inspired from biology, E-nose has the ability to differentiate and classify various chemical odours and mainly use gas sensors to generate chemical changes on odour molecules, and conduct further analysis on the gases. Fig. 1 shows the E-nose system versus the human smelling system.

In [3], an electronic nose based on conductive polymer sensors was designed and fabricated. The electronic interface circuit was able detect three hazardous gases that lead to liver damage. However, the resistance of the sensor ranged between 1kΩ and 1.5MΩ. In [4] a breath analyser prototype for detecting ammonia gas in exhaled human breath is described. Such prototype was specialized for ammonia gas detection where the sensor resistance ranges from 15MΩ to 65MΩ. In [5]

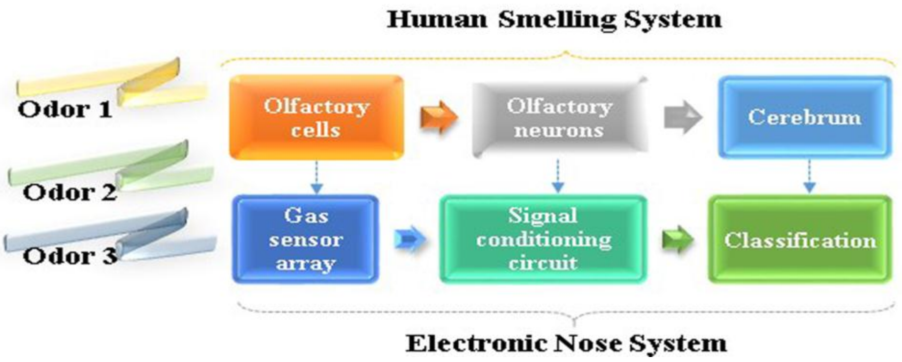


Fig. 1. E-nose system verses human smelling system

ethanol-sensing properties for alcohol breath analysis using metal oxide (MOX) resistive gas sensors was presented. This study verified that increasing the sensors resistance range allows to measure low gas concentrations. This concludes that wider resistance range allows identifying gases with lower concentration.

This work aims to design of a wide range reliable electronic solution providing high accuracy in measuring the resistance value for exhaled breath proper identification. The wide range readout circuit targets to condition wide resistance range resulting in detecting low gas concentrations. The objective of this paper is to realize the R-to-I conversion circuit for a wide resistance range extending from 500Ω to 100MΩ while preserving a precision better than 1% to satisfy the requirements of novel portable gas sensing monitoring [6]. Wider resistance range results in detecting more VOCs, thus, identifying more diseases [5]. The proposed interface takes advantage from the enhanced oscillator approach resistance to time conversion architecture used for environmental monitoring proposed in [7] and [8]. In order to ensure the correctness of the proposed architecture, the R-to-I conversion circuits in [7], [8] and [9] for gas sensing in environmental monitoring, were designed. The proposed R-to-I conversion circuit achieved a maximum percentage error of 0.05%.

## 2. READOUT ELECTRONIC INTERFACE

MOX gas sensors are characterized by their small sizes, economical cost, high sensitivities in detecting trace concentrations of chemical compounds, possibility of on-line operation and possible bench production. However, MOX gas sensors respond in a similar way to different types of oxidizing or reducing species, hence, these sensors are not selective. Integration of heterogeneous nano-structured microarray gas sensors having different sensing materials was proposed in [10] to overcome this limitation.

MOX gas sensors operate on the principle of chemo resistive sensing. Alteration of the nano-structured oxide thin films occurs in the presence of targeted analyte. Therefore, the transducer of the sensor converts the chemical information about the concentration of the gas in the atmosphere into an electrical signal, which is a variation of resistance. Such sensors show a wide range performance since the baseline resistance depends on several chemical and physical parameters, fabrication process, technology, and the sensor operating conditions [11].

Several possible readout circuits for wide range resistive gas sensors are proposed in literature. Expensive Pico ammeters or high-resolution analog-to-digital converters (ADCs) with a programmable gain amplifier (PGA) or a scaling factor system can be used [12]. On the other side, logarithmic sub-range detector based resistance to digital conversion architecture presents another solution. In this method, the resistance range is first converted to a voltage and then measured by an ADC. In order to accommodate the wide range resistance, such architecture divides the input resistance into sub ranges and uses logarithmic sub range detection to detect the correct sub range [13]. Quasi-digital electronic interface circuits based on enhanced oscillator approach in which the information about the sensor is converted into duty cycle, frequency, or period represent an effective solution since it is possible to merge the ingrained simplicity to analog devices with the accuracy and noise immunity typical of digital sensors [14]. Therefore, our proposed approach takes advantage of the resistance-to-time conversion technique, which is advisable when wide-range resistances are to be considered [7], [11], and [14]. This paper focuses on the implementation of the resistance to current conversion circuit, which is an essential part to achieve the overall readout conversion accuracy.

## 3. PROPOSED METHOD

### a. Resistance to Time Conversion

Due to their wide input resistance range, designing an electronic interface circuit for resistive sensors is considered as a challenging task [7], [11], and [13]. Over the years, various techniques were proposed to measure electrically a gas change. These studies were either based on the conventional method of adding an analog to digital converter or based on the resistance to frequency or resistance to time period conversion. The latter conversion techniques are advantageous since it is possible to translate the sensed signal into a frequency value, which is subsequently digitized and acquired via a counter. Another advantage of this approach is the robustness of the sensor output signal with respect to noise and disturbances. In addition, the hardware costs in the resistance to frequency based interface circuits are lower than that in conventional ADC methods.

As a result, it is more convenient to transform the resistive information to period of a square wave and perform a time measurement. In resistance to digital conversion, the resistance of the sensor (Rsens) is first converted to a current through a

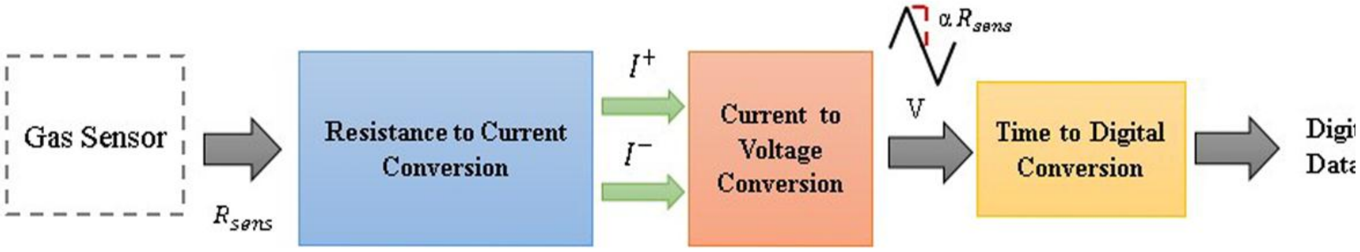

Gas Sensor
$R_{sens}$
Resistance to Current Conversion
$I^+$
$I^-$
Current to Voltage Conversion
$\alpha R_{sens}$
V
Time to Digital Conversion
Digi
Data

Fig.2. Block diagram of the resistance to digital conversion.

voltage buffer, which fixes the voltage across Rsens resulting in a current variation. An accurate current mirror preserves the accuracy while delivering alternate current to an integrator to be converted into a voltage signal is needed. The latter is then converted into period using window comparators and a flip-flop. Finally, the sensor signal will be converted into a digital code through period to code conversion, so to be directly delivered to the digital processing stage of recognition system without requiring an ADC. Such architecture is considered as area efficient [15]. Fig.2 describes the conversions required by the electronic interface for achieving the resistance to digital conversion signal.

## b. R-to-I Conversion Circuit

A novel circuit architecture able to cover a resistance range from 500Ω to 100MΩ (more than 5-decades) into current with high accuracy is proposed. Fig. 3 represents the architecture proposed to obtain such goal. In this architecture, the voltage across the sensors resistance is bounded between two Operational Transconductance Amplifiers OTA1 and OTA2, such that:

$$V_{sens} = V_{ref1} - V_{ref2} \tag{1}$$

In this configuration, the variation of Rsens is converted into a current signal variation (Isens), such that:

$$I_{sens} = \frac{V_{ref1}-V_{ref2}}{R_{sens}} \tag{2}$$

Where, Vref1= 0.5V, Vref2=-0.5V, this results in 1V fixed across the sensor's resistance Rsens. Therefore, for Rsens varying between 500Ω and 100MΩ, Isens will ideally vary between 2mA to 10nA.

A new differential R-to-I conversion circuit with push-pull current mirror architecture required to feed the following integrator with alternate accurate current is proposed. The generated sensed current (Isens) is sourced through transistors Mp1-Mp2 being in parallel with Mp1'-Mp2' and sinked through transistors Mn1-Mn2 being in parallel with Mn1'-Mn2'. The sensed current is then replicated through transistors Mp3-Mp4 with a scaling ratio K: 1 being K= (800µm+800µm) to 800µm i.e. K=2/1 for the PMOS current mirror side. Whereas for the NMOS current mirror, the sensed current is replicated through transistor Mn3-Mn4 with a similar scaling ratio K=2/1, but of different geometric size where, K= (300µm+300µm) to 300µm. The output current of the current mirror (Iout) is equal to Isens divided by the scaling factor K (Isens/K). To improve the accuracy of the current mirror a cascode branch, formed by transistors Mp2, Mp4 from the PMOS side and Mn2, Mn4 from the NMOS side, are added. Scaling the input branch of the current mirror using parallel-connected transistors for both push and pull current mirrors in the proposed architecture allows increasing the transistor's width, thus, improving the accuracy while supporting more current especially when Rsens is low.

## 4. SIMULATUIN RESULTS

The proposed R-to-I conversion circuit was designed in 3.3V- 0.35µm, N-WELL, four metal AMS CMOS technology. The functionality of this circuit is simulated in PSPICE OrCAD CAD tool.

The OTAs used in the architecture are two stage Capacitor Multiplier Miller Compensated whose topology is shown in Fig.3. TABLE I presents the specifications of both OTAs used in the proposed architecture where the OTA1 is biased by Vref1=0.5V and OTA2 is biased by Vref2=-0.5V. The compensation capacitor used is only 0.7pF while the compensation resistor is equal to 60 kΩ.

To measure the accuracy of the converted current, the relative absolute percentage error (% Error ) of both mirrored output currents through the PMOS and NMOS current mirrors (IoutPMOS) and (IoutNMOS), taking the input current across the sensor's resistance (Isens) as a reference, is computed according to (3).

$$\%Error = \frac{|I_{sens}-I_{out}|}{I_{sens}} \times 100 \tag{3}$$

The designed architecture was simulated under environmental corners that are Best, Typical, and Worst cases, to test the reliability of the designed architecture. Where, Best corresponds to high voltage supply i.e. ±1.8V and low temperature $-40^{0}$C, on the contrary, Worst corresponds to low voltage i.e. ±1.5V and high temperature $80^{0}$C, and Typical corresponds to ±1.65V at $27^{0}$C. Fig. 5 shows the absolute percentage error of output currents of the circuit i.e. the output

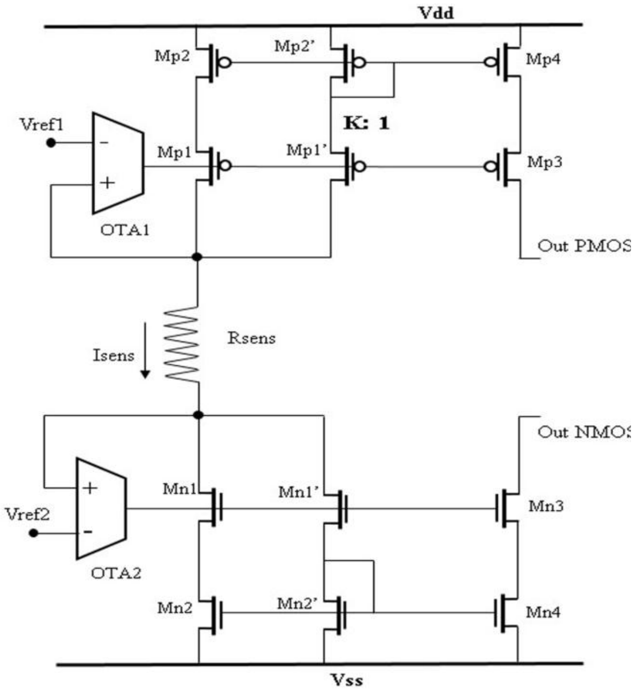


Fig. 3. Proposed resistance to current conversion circuit for resistive gas sensor

TABLE I. SPECIFICATIONS OF THE OTAS USED IN THE PROPOSED R-TO-I CONVERSION ARCHITECTURE

| | **OTA1** | **OTA2** |
|---|---|---|
| **DC Gain** | 58dB | 90.5dB |
| **BW** | 4.2kHz | 103.6Hz |
| **PM** | $112^0$ | $104^0$ |

through the PMOS current mirror (out PMOS), where, the output through the NMOS current mirror (out NMOS) is grounded, Fig.5.a and vice versa Fig.5.b.

The maximum percentage error obtained along the full studied resistance range was calculated, while simulating the circuit under different process variations with typical environmental corners. Fig.6 and Fig.7 illustrate the obtained results along a resistance range from 500Ω to 100MΩ considering both outputs of the circuit (Out PMOS) and (Out NMOS) each at once. As shown in Fig.6 and Fig.7, the resulted accuracy verifies the proper performance of the proposed architecture, and the corner analysis assures the reliability of the implemented circuit. The low values of the achieved % Error validates the requirements of new portable gas sensing systems preserving a precision ≤ 1%.

To ensure the correctness of the proposed architecture the R-to-I conversion circuit presented in [7], [8] were designed. Such architectures are used for gas sensing systems in environmental monitoring where the minimum input resistance value was 1kΩ. Considering the input resistance range proposed for portable breath analysis applications from 500Ω to 100MΩ, the proposed architecture has extended the minimum range to 500 Ω where it achieves an accuracy of 0.05%. Over the remained range, Table II shows the comparison between the proposed architecture and the other architectures over the range 1kΩ to 100 MΩ. The proposed architecture achieved a maximum percentage error better than 0.044% over the compared resistance range, which assure the capability of the proposed architecture to be used for E-nose (breath analyzer) systems.

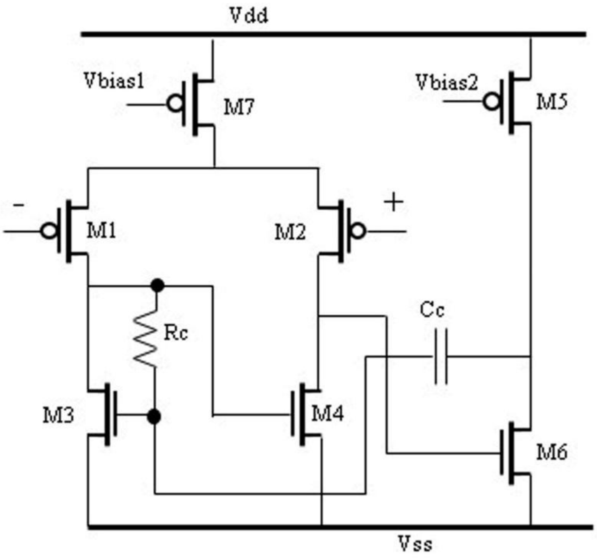


Fig.4. Capacitor Multiplier OTA architecture

To test the non-ideal effects caused by the current mirrors for the sensor current signal transfer in the current mode conversion, the sensitivity of the output current with respect to the input sensor's current is studied. For both outputs i.e. Out PMOS and Out NMOS, the circuit achieves a sensitivity of 0.5003 and 0.499 respectively knowing that the ideal computed sensitivity is 0.5. The sensitivity is defined as:

$$S = \frac{dI_{out}}{dI_{sens}} \tag{4}$$

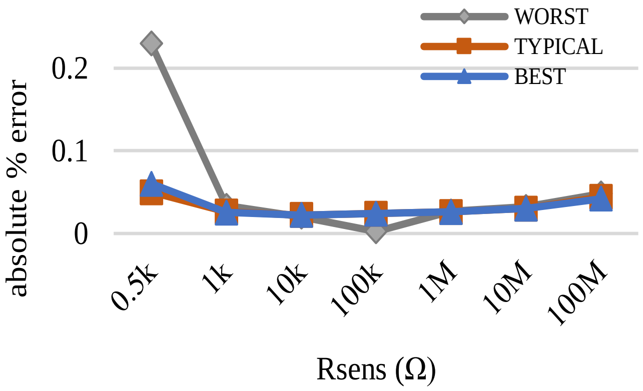


Fig.5.a. Absolute percentage error on the output of the PMOS current mirror for the designed architecture under environmental corners.

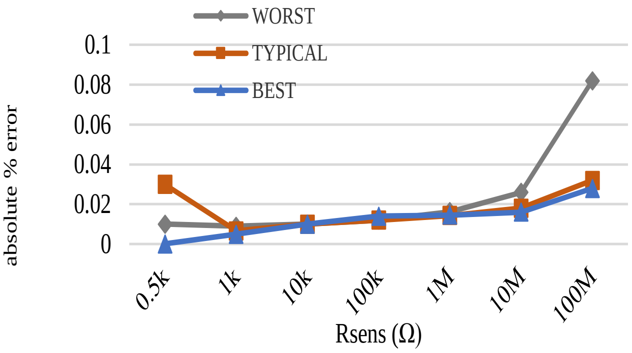


Fig.5.b. Absolute percentage error on the output of the NMOS current mirror for the designed architecture under environmental corners.

FIG. 6. PROCESS CORNER CHECK CONSIDERING THE OUTPUT THROUGH PMOS CURRENT MIRROR

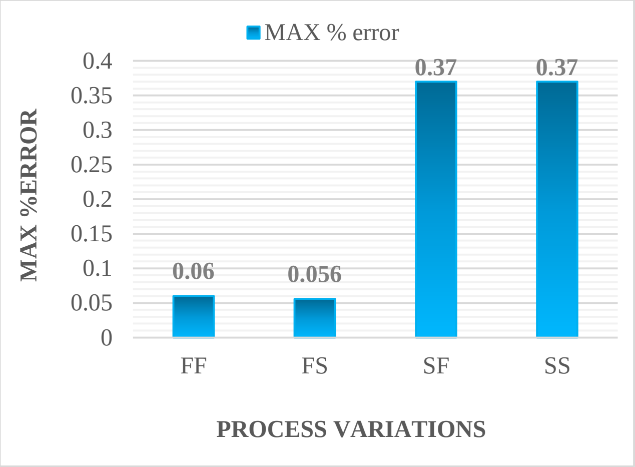


FIG7. PROCESS CORNER CHECK CONSIDERING THE OUTPUT THROUGH THE NMOS CURRENT MIRROR

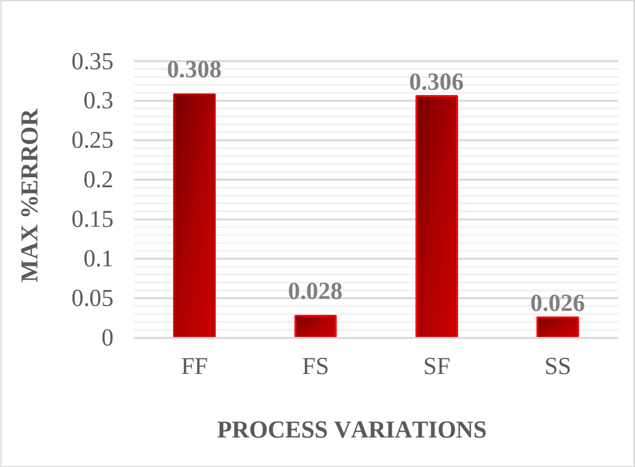


TABLE .II. COMPARISON BETWEEN MAX PERCENTAGE ERROR OBTAINED FROM THREE DIFFERENT R-TO-I ARCHITECTURES

| | [7] | [8] | [9] | Proposed |
|---|---|---|---|---|
| **Max.% Error**<br>**under Typical Case** | 0.29% | 0.24% | 0.25% | 0.044% |

## 5. CONCLUSIONS AND FUTURE WORK

This paper presented the design of R-to-I conversion circuit for gas sensing in biomedical applications. The proposed resistance to current conversion circuit architecture achieved high accuracy and preserved a precision less than 1% required by novel gas sensing system in portable applications over a wider resistance range (500Ω to 100MΩ) compared to the solutions presented in literature. The percentage error of both considered outputs of the circuit Out PMOS and Out NMOS was always below 0.25% and 0.1% respectively under environment corners. The reliability of the proposed circuit was also investigated under the effect of process parameters. The sensitivity of the PMOS and NMOS current mirrors was 0.5003 and 0.499 respectively. Compared to other R-to-I architectures proposed in literature, the proposed architecture achieved a max % Error of 0.05% over the proposed resistance range. The presented work is a step towards the implementation of an electronic readout circuit for resistive gas sensors based on the design and architecture provided herein.

## 6. REFERENCES

[1] S. Dragonieri, et al. “An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD”, 2009.

[2] Y. Fang, J. Lahiri “Breath analysis: biomarkers, nanosensors, and midinfrared absorption sensors”, Chemical Sensors, pp. 3-9, 2013.

[3] K. T. Tang et a.l., "A Low-Power Electronic Nose Signal-Processing Chip for a Portable Artificial Olfaction System," in IEEE Transactions on Biomedical Circuits and Systems, Aug. 2011.

[4] P. Gouma, et al., "Nanosensor and Breath Analyzer for Ammonia Detection in Exhaled Human Breath," in IEEE Sensors Journal, vol. 10, no. 1, pp. 49-53, Jan. 2010.

[5] T. Santhaveesuk, et al. "Enhancement of Ethanol Sensing Properties by Alloying With ZnO Tetrapods," in IEEE Sensors Journal, vol. 10, no. 1, pp. 39-43, Jan. 2010

[6] M. Grassi, P. Malcovati, A. Baschirotto, “Fundamental Limitations in Resistive Wide-Range Gas-Sensor Interface Circuit Design”, in: Sensors and Microsystems - Lecture Notes in Electrical Engineering, vol. 54, 2010, pp.25-40, Springer.

[7] M. Grassi, P. Malcovati, and A. Baschirotto, “A 141-dB dynamic range CMOS gas-sensor interface circuit without calibration with 16- bit digital output word,” IEEE J. Solid-State Circuits, vol. 42, no. 7, Jul. 2007.

[8] F. Conso, M. Grassi, P. Malcovati, A. Baschirotto. “Reconfigurable Integrated Wide-Dynamic-Range Read-Out Circuit for MOX Gas- Sensor Grids Providing Local Temperature Regulation”, proceedings of SENSORS’12, Taipei, Oct. 2012, pp.1822-1825.

[9] Z. Hijazi, D. Caviglia, M. Valle and H. Chible, "Wide range resistance to current conversion circuit for resistive gas sensors applications," 2016 12th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), Lisbon, 2016, pp. 1-4.E.

[10] Barborini, et al. "The influence of nanoscale morphology on the resistivity of cluster-assembled nanostructured metallic thin films," New Journal of Physics, vol. 12, p. 12pp, 2010

[11] A. De Marcellis, et al. "A novel time-controlled interface circuit for resistive sensors," Sensors, 2011 IEEE, Limerick, 2011, pp. 1137-1140.

[12] A. Depari et al., "A New and Fast-Readout Interface for Resistive Chemical Sensors," in IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 5, pp. 1276-1283, May 2010.

[13] M. Choi, et al. , "Wide input range 1.7µW 1.2kS/s resistive sensor interface circuit with 1 cycle/sample logarithmic sub-ranging," 2015 Symposium on VLSI Circuits (VLSI Circuits), Kyoto

[14] C. Azcona, et al., "Low-Power Wide-Range Frequency-Output Temperature Sensor," in IEEE Sensors Journal, vol. 14, no. 5, pp. 1339-1340, May 2014.

[15] C. L. Chang, et al, "An ADC-free adaptive interface circuit of resistive sensor for electronic nose system," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, 25

International Journal of Computer Applications Technology and Research
Volume 6–Issue 1, 66-69, 2017, ISSN:-2319–8656

# Image Encryption and DWT Based Copy-move Image Forgery Detection

Gawtham Srinivasan R
Dr. Mahalingam College of Engineering
&Technology
Pollachi, India

Lokesh K
Dr. Mahalingam College of Engineering
&Technology
Pollachi, India

**Abstract**: In the present world, digital images are one of the main carrier of information. However the digital images are easily getting tampered due to the availability of image editing software such as adobe Photoshop and so on. It is possible to remove important features or information from an image without leaving any traces. Therefore as a solution to the above mentioned problem, we are proposing a unique solution to detect the copy-move image forgery. In proposed scheme the image is encrypted at client side and it will be decrypted at receiver end for authentication purpose. The image is segmented into non-overlapping blocks .Then statistical features are extracted and reduced to facilitate the measurement of similarities. Finally the Euclidian distance has been calculated and duplicate image blocks are identified after post processing. Experiments results demonstrate that our proposed method is able to detect multiple examples of copy-move image forgery and precisely locate the duplicated regions. We are currently working to improve detection in overlapping blocks.

## 1. INTRODUCTION

Digital images are subjected to various kinds of attacks and damages due to the prevalence of various image editing software such as adobe photo shop etc. One of the most important type of image manipulation is copy move forgery. In this type, a portion of the image is copied and placed in a different location. We devised a method to identify the copy move forgery using some statistical features.

## 2. RELATED WORKS

Most methods used in the detection of copy–move forgery can be categorized as either block-based methods or key point based methods. The first such method was proposed by Fridrich using a block matching detection scheme based on discrete cosine transform (DCT). Popescu and Farid proposed a copy–move forgery detection method, which differs in its representation of overlapping image blocks using principal component analysis (PCA) instead of DCT. Some approaches involve the extraction of points of interest using a scale-invariant feature transform (SIFT) capable of detecting and describing clusters of points belonging to cloned areas.

SIFT-based schemes are still limited in their detection performance due to the fact that it is only possible to extract key points from specific locations in an image. In addition, these methods are susceptible to a number of post-processing operations, such as blurring and flipping.

However, some key points of duplicate regions cannot be identified using key point based algorithms and copied regions with little textural structure may be missed entirely.

## 3. COPY MOVE FORGERY DETECTION

The social media, news sources mainly depends on the digital images to represent the truth of the stories; however, digital images processing tools readily available to make the tampered images for malicious reasons. Various methods have been developed to counter tampering and forgery in order to ensure the authenticity of images. Current forgery detection methods can be categorized as active and passive (blind). Most active methods are based on digital signatures and watermarking; however, this requires that data be pre-processed, which can be troublesome. Passive methods are used to analyse images without using a priori information (such as embedded watermarks or signatures), such that a blind decision must be made regarding whether images have been tampered with. Most passive techniques are based on supervised learning through the extraction of specific features to differentiate the original image from tampered versions. The practicality and wide applicability of passive methods have made them a popular topic of research.

Copy–move is the most common form of digital image forgery, in which a portion of an image is copied and pasted into another portion of the same image to conceal something or duplicate elements. The wide availability of image processing software has made it easy to perform copy–move operations. The region altered by copy–move forgery is often almost imperceptible by the human eye; therefore, detecting evidence of these actions is an important issue in image forensics. This paper presents a robust algorithm for the detection of copy–move forgery based on the histogram of oriented gradients (HOG). The performance of the proposed method is compared with existing methods with regard to detection accuracy and computational complexity.

## 4. ENCRYPTION AND DECRYPTION

Nowadays internet is used for faster transmission of large volume of important and valuable data, since internet has many points of attack, it is vulnerable to many kinds of attack, so this information need to be protected from unauthorized access, To protect data from unauthorized access there are many data protection techniques like Nulling Out, Masking Data, Watermarking, Encryption etc. are implemented.
Data Encryption is one of the widely used techniques for data protection. In Data Encryption, data is converted from its

original to other form so that information cannot be accessed from the data without decrypting the data i.e. the reverse process of encryption.

The original data is usually referred as plain data and the converted form is called cipher data. Encryption can be defined as the art of converting data into coded form which can be decoded by intended receiver only who possess knowledge about the decryption of the ciphered data. Encryption can be applied to text, image and video for data protection.

In proposed work we use two simple techniques namely:

- Permutation.
- Substitution.

In Permutation pixels of images are relocated to different location in the image using the chaotic map sequence.

In Substitution pixels of the original image are multiplied with the key image pixels.

We combine both substitution and permutation to form an encrypted image. This encrypted image is used for authentication.

At the receiver end, Receivers are provided with the key and the encrypted image. The receiver decrypts the image and process the image for the identification of copy move forgery.

Decryption is simply the reverse process of encryption.

## 4.1 Chaotic Sequence

Pseudo Random Number Generators (PRNGs) are widely used in many applications, such as numerical analysis, probabilistic algorithms, secure communications, integrated circuit testing, computer games and cryptography. The quality of randomness is usually the main criterion to distinguish the different PRNGs. Besides the quality of randomness, implementation cost and throughput are also important factors to evaluate the effectiveness of the PRNGs in applications, such as modern communications, image encryption, video encryption and sensor networks, and so on.

Chaos has widely been used in cryptography in recent years. Chaotic maps are often used in encrypting images. Chaos is applied to expand the diffusion and confusion in the image. Due to the desirable properties of nonlinear dynamical systems, such as pseudorandom behaviour, sensitivity to initial conditions, unpredictability and periodicity, chaos-based encryption is suggested as a new and efficient way, to deal with the intractable problem of fast and highly secure image encryption.

### 4.1.1 Logistic Map for Image Encryption

For image encryption we use logistic map. It will exhibit the chaotic behavior. Both continuous and discrete chaotic maps are available. In this work discrete map is used, this kind of maps usually takes the form of iterated functions. In this work logistic map is used. The logistic map is a simple one dimensional map and is given in Equation (1),

$$x_{n+1} = rx_n(1 - x_n) \qquad (1)$$

Logistic map is a polynomial mapping of degree 2. In Equation (1) $X_n$ € [0, 1] and is known as the phase space of the logistic map, r is the control parameter that controls the behavior of the map.

- With r between 0 to 1 the map is independent of the initial condition.
- For r between 1 to 2 the trajectory will quickly reach the value, map is independent of the initial condition.
- For r between 2 to 3 the trajectory will reach the value in as specific manner that is it will revolve around the value for some time to reach the value.
- With r between 3 to 3.45 for almost all the initial conditions the population will oscillate between two values and these values are depends on the value of b.
- At r approximately 3.57 is the onset of chaos, at the end of the period-doubling cascade. From almost all initial conditions we can no longer see any oscillations of finite period. Slight variations in the initial population yield dramatically different results over time

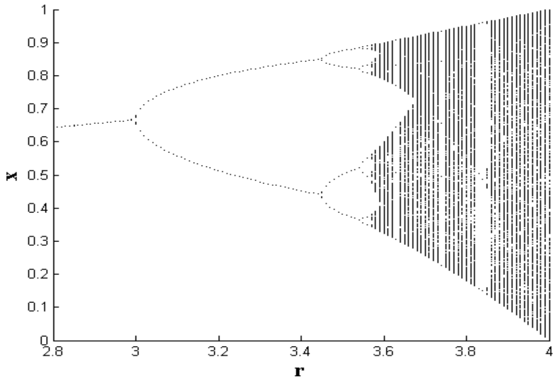


**Figure 3.1.1 Bifurcation diagram of logistic map**

Beyond r = 4, the values eventually leave the interval [0, 1] and diverge for almost all initial values.

Figure 4.1.1 summarizes the above points and the horizontal axis shows the values of the parameter r while the vertical axis shows the values of x. From Figure 4.1.1 it is clear that for the values above r = 3.82 the map exhibits the chaotic behavior. The map used in this work is a discrete one, it is in the form of iterated function. This map is used because of its easy computation and greater complexity.

### 4.1.2 Sine Map for Key Encryption

The sine map is same as that of the logistic map chaotic behavior and is defined by Equation (3),

$$x_{n+1} = r\sin(\pi x_n)/4 \qquad (3)$$

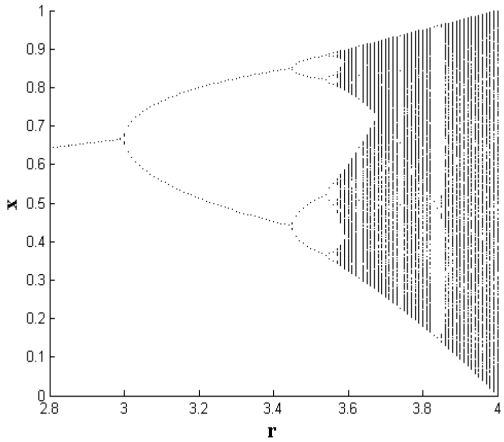


**Figure .1.2 Bifurcation diagram of sine map**

Where parameter 'r' is in the range of (0 to 4). Figure 4.1.2 shows the chaotic behavior of the sine map it is almost same as that of the logistic map. The sine function provides the good randomness behavior for this chaotic function.

The above mentioned are the chaotic behavior of the map we used for the encryption and decryption process.

## 5. COPY MOVE FORGERY DETECTION SCHEME

After the completion of Encryption and Decryption process the image is obtained for the process of identification of copy-move forgery. The image is preprocessed before the identification of forgery takes place.

The preprocessing includes:

- Convert the colored image into grayscale image.
- Resize the image into a fixed size of 256 x 256.
- Divide the images into non-overlapping sub blocks.
- Here we divide the image into 64 sub blocks of 32x32 size.

After the preprocessing, we considered some of the following statistical features for the identification of forgery in the images. The features are: Mean, Median, Mode, Variance, Standard Deviation, min, max, kurtosis, skew, Entropy, Moments.

The statistical features are extracted from the images and the features are normalized for better result.

After the normalization of the features, Distance between the blocks had been calculated. The distance measure applied here is Euclidean distance.

The distance between two determines the match. (i.e.) Less distance implies the best match and farthest distance implies the dissimilarities.

While finding the distance, values with zero are eliminated since it defines the distance between same values (i.e.) x-x. A threshold value is fixed for computing the match between the images. The threshold value is the value below which the matches are occurred. Threshold value is fixed based on the trial and error method. Multiple images are taken and check for multiple images and the threshold values are fixed. Finally we have fixed the threshold value by comparing the multiple images and found out the mean of the value as 0.175.

## *6.* WAVELET TRANSFORM

A wavelet is a mathematical function used to divide a given function or continuous time signal into different scale components. Usually one can assign a frequency range to each scale component. Each scale component can then be studied with a resolution that matches its scale. A wavelet transform is the representation of a function by wavelets. The wavelets are scaled and translated copies (known as "daughter wavelets") of a finite length or fast decaying oscillating waveform (known as the "mother wavelet"). Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for accurately deconstructing and reconstructing finite, non-periodic and/or non-stationary signals.

Wavelet transforms are classified into discrete wavelet transforms (DWTs) and continuous wavelet transforms (CWTs). DWTs use a specific subset of scale and translation values or representation grid. Applications of wavelet transform are transform data, and then encode the transformed data, resulting in effective compression and for communication applications.

## 7. DISCRETE WAVELET TRANSFORM

In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time). Applications for discrete wavelet transform are signal coding, to represent a discrete signal in a more redundant form, often as a preconditioning for data compression, Practical applications can also be found in signal processing of accelerations for gait analysis, in digital communications.

## 8. EXPERIMENT RESULT AND DISCUSSIONS

The plain image is encrypted and a key image with changed pixels is taken to calculate the NPCR and UACI. $C_1(i,j)$ is the encrypted plain image and $C_2(i.j)$ is the changed pixel key image.

### NPCR and UACI

The number of changing pixel rate **(NPCR)** and the unified averaged changed intensity (UACI) are two most common quantities used to evaluate the strength of image encryption algorithms/ciphers. The NPCR and UACI are designed to test the number of changing pixels and the number of averaged changed intensity between cipher text images, respectively, when the difference between plaintext images is subtle (usually a single pixel). Although these two tests are compactly defined and are easy to calculate, test scores are difficult to interpret in the sense of whether the performance is good enough. For example, the upper-bound of the NPCR score is 100%, and thus it is believed that the NPCR score of a secure cipher should be very close to this upper-bound.The attacker may have a slight change (modify one pixel) of the plain image to find some meaningful relationships between the plain image and the encrypted. If one minor change in the plain image causes a significant change in the cipher image, this indicates that the encryption scheme resists differential attacks more efficiently. To test the influence of only one pixel change in the plain image over the whole encrypted image, two common measures are used: Number of Pixels Change Rate (NPCR) and Unified Average Changing Intensity (UACI),

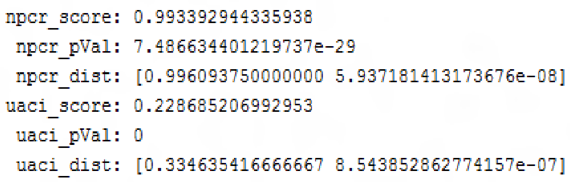
$$NPCR = \frac{\sum_{i,j} D(i,j)}{W \times H} \times 100\%$$

$$UCAI = \frac{1}{W \times H}\left[\sum_{i,j} \frac{\left|C_1(i,j) - C_2(i,j)\right|}{255}\right] \times 100\%$$

Original Image

Encrypted Image

NPCR and UACI results =



```
npcr_score: 0.993392944335938
 npcr_pVal: 7.486634401219737e-29
 npcr_dist: [0.996093750000000 5.937181413173676e-08]
uaci_score: 0.228685206992953
 uaci_pVal: 0
 uaci_dist: [0.334635416666667 8.543852862774157e-07]
```

### Receiver Operating Characteristics (ROC) Analysis

ROC curves are popularly used as performance metrics for classification task. The ROC curve is acquired by applying a threshold value to the classifier predicted score and obtaining a (TP and FP) value for each threshold to generate the curve.

True positive (TP) = the number of cases correctly identified as Forgery
False positive (FP) = the number of cases incorrectly identified as Forgery
True negative (TN) = the number of cases correctly identified as Mismatch
False negative (FN) = the number of cases incorrectly identified as Mismatch.

**Table 8.1 ROC analysis table**

| | FRAUD | NO FRAUD |
|---|---|---|
| **FRAUD** | 16 | 5 |
| **NO FRAUD** | 2 | 57 |

**Accuracy:** The accuracy of a test is its ability to differentiate the forged and not forged cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

**Senstivity:** The sensitivity of a test is its ability to determine the forged cases correctly. To estimate it, we should calculate the proportion of true positive in forged cases. Mathematically, this can be stated as:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

**Specificity:** The specificity of a test is its ability to determine the non forged cases correctly. To estimate it, we should calculate the proportion of true negative in non forged cases. Mathematically, this can be stated as:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

**Table 8.2 ROC Result table**

| **Features** | **Result Obtained (in %)** |
|---|---|
| Accuracy | 91.25 |
| Specificity | 96.61 |
| Sensitivity | 76.19 |
| Precision | 88.88 |

## 9. CONCLUSION AND FUTURE WORK

The detection of forgery in digital images is an interesting topic in forensic science. This paper proposes an effective method for detecting duplicated regions based on the mathematical statistical features in spatial domain. Experiment results demonstrate that the proposed algorithm is able to detect and precisely locate multiple instances of copy–move forgery in a single image. Next our aim is to detect the copy move forgery in wavelet domain using DWT. Also for precise identification of forgery we are planned to use overlapping sub blocks concept.

## 10. REFERENCE

[1] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, G. Serra, A SIFT-based forensic method for copy–move attack detection and transformation recovery, IEEE
Trans. Inform. Forensics Sec. 6 (2011) 1099–1110.

[2] O.M. Al-Qershi, B.E. Khoo, Passive detection of copy–move forgery in digital images: states-of-the-art, Forensic Sci. Int. 206 (1) (2013) 284–295.

[3] S. Bayram, H.T. Husrev, N. Memon, An efficient and robust method for detecting copy–move forgery, in: Proceedings of the IEEE International
Conference on Acoustics, Speech and Signal Processing, 2009, pp. 1053–1056.

[4] A. Costanzo, I. Amerini, R. Caldelli, M. Barni, Forensic analysis of SIFT keypoint removal and injection, IEEE Trans. Inform. Forensics Sec. 9 (9) (2014)
1450–1464.

[5] L. Chen, W. Lu, J. Ni, W. Sun, J. Huang, Region duplication detection based on Harris corner points and step sector statistics, J. Vis. Commun. Image Rep.
24 (2013) 244–254.

[6] CoMoFoD Database. <http://www.vcl.fer.hr/comofod>.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Comput. Vis. Pattern Recognit. (2005) 20–25.

[8] J. Fridrich, D. Soukal, J. Lukas, Detection of copy–move forgery in digital images, in: Proceedings of Digital Forensic Research Workshop, 2003, pp. 19–
23.

[9] J. Fridrich, Methods for tamper detection in digital images, in: Proceedings of the ACM Workshop on Multimedia and Security, 1999, pp. 19–23.

[10] T. Gloe, M. Kirchner, A. Winkler, R. Behme, Can we trust digital image forensics? in: Proceedings of the 15th International Conference on Multimedia,
2007, pp. 78–86.

[11] S. Khan, A. Kulkarni, An efficient method for detection of copy–move forgery using discrete wavelet transform, Int. J. Comput. Sci. Eng. 2 (2010) 1801–
1806.

[12] X. Kang, S. Wei, Identifying tampered regions using singular value decomposition in digital image forensics, in: Proceedings of International
Conference on Computer Science and Software Engineering, 2008, pp. 926–930.

[13] B. Mahdian, S. Saic, Detection of copy–move forgery using a method based on blur moment invariants, Forensic Sci. Int. 171 (2007) 180–189.

[14] D. Lowe, Object recognition from local scale-invariant features, Proc. Int. Conf. Comput. Vis. 2 (1999) 1150–1157.