International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 301-304, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1001

Development of HOTS Integrated Problem Based Learning (PBL) Chemistry Learning Module on Buffer Solution Material at SMA Negeri 1 Purba

Freddy Tua Musa Panggabean
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Grestina Winda Sari Munthe
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Pasar Maulim Silitonga
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Anna Juniar
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Rini Selly
Chemistry Education Study Program
Universitas Negeri Medan
Medan, Indonesia

Abstract: This research was conducted to design a Problem Based Learning (PBL) based learning module with the ADDIE model. The problems studied are 1) How is the feasibility (validity) and learning effectiveness of the Problem Based Learning (PBL)-based module developed on the buffer solution material to measure students' Higher Order Thinking Skills (HOTS) and 2) What are the students' chemistry learning outcomes using the module-based Problem Based Learning (PBL) is higher than the KKM score. The subjects of this study were 3 expert validators (expert lecturers), and 1 high school chemistry teacher to evaluate the learning to be developed and class XI MIA students of SMA Negeri 1 Purba who were determined by random sampling using an experimental class totaling 30 students. The results of the study showed that 1) the problem-based learning module for chemistry-based learning buffer solution material that was developed was feasible to be used for learning materials to improve students' critical thinking skills. The module is declared effective as seen from the increase in student learning outcomes, namely the average pretest value of 33.5 and posttest 80.13 in the experimental class, 2) The learning outcomes of chemistry using Problem Based Learning (PBL)-based modules developed on buffer solution material are higher than the value KKM (75) with an average score of 80.13.

Keywords: Problem Based Learning, HOTS, module, learning outcomes, BSNP

1. INTRODUCTION

Education is a process of interaction that occurs between teachers and students to assist students in developing their potential. Education is also interpreted as conscious guidance by educators on the spiritual and physical development of students towards the formation of the main personality and skill development through practice so that they are able to reach maturity little by little [1]. Education is always related to the curriculum. According to the KBBI, the curriculum is a set of subjects given to lesson participants. This curriculum has a function as a guide in the implementation of the teaching and learning process in schools for related parties such as teachers, principals, supervisors, parents, the community and the students themselves [2]. The 2013 curriculum that is applied now is a curriculum that prioritizes skills, understanding and character, where students are required to master the material, be active during the teaching and learning process [3] The 2013 curriculum emphasizes educators to have skills in compiling Higher Order Thinking Skill (HOTS) assessment instruments, namely evaluation tools that can train students' critical and creative thinking processes [4]. HOTS is a learning designed to prepare the 21st century generation to have competencies and skills which include: critical thinking and problem solving, creative competence, communication skills and the ability to work together [5].

Currently, the problem with education in Indonesia is related to the quality of education, such as limited facilities in schools, an unequal number of teachers, and the quality of the teachers themselves, which are considered to be lacking in the level of learning materials and approaches used in the learning process used in the learning process. which is less precise and effective [6]. In the teaching and learning process, it is expected that educators can convey the material being taught and provide facilities for learning, while students can understand the material being taught. So that the learning process can run as expected. Maximum learning activities are very important for everyone to understand or gain useful knowledge [7].

The selection and use of good teaching materials is an important factor in the quality of education. Teaching materials that can be used by students as independent learning resources have an important role in improving and developing higher-order thinking skills [8]. Learning modules are books that are written and then printed with the aim that they can be studied independently by students without any guidance from educators (teachers) [9]. It has been equipped with instructions for self-study so that it is referred to as a medium for independent learning. The difference between modules and textbooks is that the module focuses on one material, while the book consists of several materials, so that the use of the module is more effective and efficient [10].

Problem Based Learning (PBL) is one of the innovative learning models that can provide active learning conditions for students. The PBL model prepares lessons for critical and analytical thinking, as well as for finding and using learning resources [11]. The Problem Based Learning (PBL) learning model has been studied by several previous researchers and has been proven to improve student learning outcomes [12], including: concluded that 1) there is an increase in students' creative thinking skills using the PBL model, 2) there is an increase in students' creative thinking skills. student learning outcomes using the PBL model, 3) students' creative thinking skills using the PBL model are better than using the conventional model [13].

Learning by using modules is not only for conveying information but also provides opportunities for students to learn on their own according to their abilities. students can learn on their own to solve problems and understand lessons without being too dependent on the teacher. In addition, learning using modules is expected to change students' study habits for independent study and can help students understand theory in depth through learning experiences [14].

One of the learning activities taught in high school is chemistry learning. Chemistry is a material that is considered difficult by students. The difficulties experienced by students are usually caused by the existence of concepts that must be understood, the relationship between one concept and another, besides that there are also many mathematical calculations [15]. Buffer solution is the main material in chemistry lessons in class XI IPA High School 2nd semester (even). This material is closely related to everyday life, thus making it easier for students to understand the material by connecting it with daily activities and not focusing on theory alone. In studying the buffer solution material, students are required to have good mastery of concepts and mathematical abilities. Because a buffer solution is included in the concept of a solution, it requires an initial understanding of equilibrium, stoichiometry and the concept of acid base [16].

2. METHOD

The type of research used in this research is Research and Development (R&D) with the ADDIE development model (Analysis, Design, Development, Implementation, Evaluation) which aims to produce and develop certain products, and test the effectiveness of these products. The results of this study used Problem Based Learning (PBL) and instruments to measure the HOTS of students in the buffer solution material [17].

The development model used is the ADDIE model. The ADDIE model is a learning model design that is systematically arranged and consists of 5 steps, namely analysis, design, development, implementation, evaluation which includes the design of the entire learning process in a systematic way [18]**.** 1) Analysis, the first stage that must be done is research and information gathering. The research and data collection phase consisted of literature studies and field studies. 2) Design, carried out to identify goals and create a design for learning media that will be developed. 3) Development. The development of the module must be based on several aspects such as the criteria for a good module and the adjustment of the module to the learning material. Furthermore, the module will be validated by an expert validator. 4) Implementation, the application of HOTS-based integrated problem-based learning chemistry learning. 5) Evaluation Phase, Conducted on development products such as content/materials, developed learning media and evaluation of the effectiveness and success of the developed media.

The research also aims to determine the effectiveness of the module in improving student learning outcomes and also the responses of teachers and students regarding the practicality of the module. Student learning outcomes are assessed using pretest and posttest questions [19].

The study began by analyzing the high school chemistry textbook material for buffer solutions. Then it was designed and developed into a problem based learning chemistry module and standardized using a National Education Standards Agency questionnaire conducted by chemistry lecturers and high school chemistry teachers who have experience in teaching chemistry. The final stage is testing the module on students to determine students' chemistry learning outcomes using the HOTS integrated problem based learning model to measure students' higher order thinking skills.

In this study, one group pretest-posttest design was used in implementing the module. This design was chosen because the researcher only used one class as an experimental class, there was no control class as a comparison class. Determination of this design is also adjusted to the actions taken during the study, namely giving an initial test (pretest) then followed by giving treatment for a certain period of time and ending with giving a final test (posttest).

3. RESEARCH RESULT

Research on the development of HOTS integrated problem based learning chemistry learning modules on buffer solution material has been carried out at SMA Negeri 1 Purba through several stages starting from analyzing chemistry books, developing modules with problem based learning syntax which begins with 1) Problem orientation to students 2) Organizing students to learn, 3) Helping investigations and groups, 4) Developing and presenting work and 5) Analyzing and evaluating problem solving processes, module validation by chemistry teachers and chemistry lecturers and implementation to students [20].

3.1 HOTS Integrated Problem Based Learning-Based Chemistry Learning Module Validation

The analysis of the module that uses standardized testing based on the modified BSNP includes 4 aspects: 1) Content feasibility: 2) Language feasibility; 3) Presentation feasibility and; 4) Graphical Feasibility. Validation was given to 2 chemistry lecturers at the State University of Medan and 1 teacher at SMA Negeri 1 Purba. The assessments of the expert validators are presented in Table 3.1 as follows.

Table 3.1 Results of Validation of High School Chemistry Learning Module Based on Problem Based Learning Based on BSNP by Lecturers and Teachers

No	Assesment	Average Score	Validation Criteria
1.	Content eligibility	3,75	Very valid and does not need revision
2.	Language Eligibility	3,79	Very valid and does not need revision
3.	Serving Eligibility	3,75	Very valid and does not need revision
4.	Graphic Eligibility	3,80	Very valid and does not need revision
BSNP Eligibility Average Score		3,77	Very valid and does not need revision
5.	With *Problem Based Learning*	3,66	Very valid and does not need revision

with ;

3,50 < M ≤ 4,00	: Very valid and does not need revision
2.,50 < M ≤ 3,50	: Sufficiently valid and does not need revision
1,50 < M ≤ 2,50	: Invalid and doesn't need revision
M < 1,50	: Invalid and does not need revision

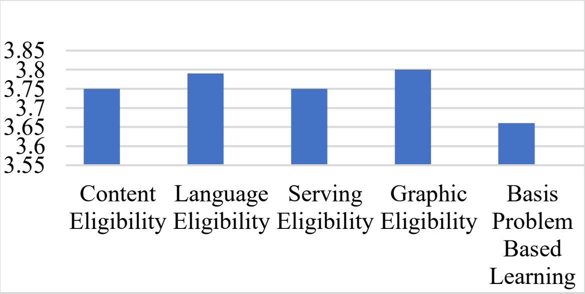


Figure 1. Graph of module feasibility analysis results

Based on the results of the validation of the module development carried out by 2 UNIMED chemistry lecturers. Obtained an average language eligibility score of 3.75; presentation eligibility 3.79; content eligibility 3.75 and graphic eligibility 3.80 which states that the High School chemistry module has met the BSNP criteria. The average score for problem-based learning is 3.66, which means that the module is already based on problem-based learning. Student responses to the developed module also have good criteria with a score on the interest indicator 3.56; 3.42 material indicators and 3.76 language indicators. With an average score of 3.58 which is classified as good criteria.

3.2 The Results of The Trial of The Integrated Problem Based Learning-Based Module HOTS Material for the Buffer Solution

After the media was declared eligible, the stage of validating the test instrument questions was carried out to UNIMED lecturers. Where the results of the test instrument validation there are as many as 10 questions. After being valid, it was then tested on students in class XI MIA 2 and the results were that there were 2 invalid questions and 8 questions declared valid.

Furthermore, at the implementation stage, the problem-based learning chemistry module on the valid buffer solution material was revised according to the suggestions from the lecturer, then tested on students to see if there was an increase in learning outcomes. In this case, the experimental class was first given a pretest to see the students' initial abilities. Where the questions used are test instruments that are validated as many as 8 questions. The average student learning outcomes obtained is 33.5. After obtaining the initial abilities, then learning is carried out using the developed chemistry module. And at the end of the material, a posttest with the same questions was given to find out whether there was an increase during the learning process. And based on the average calculation, the posttest score was 80.13.

Judging from the calculation of statistical data obtained Sig_{count} for pretest data of 0.200 and posttest data of 0.042. From these data it can be concluded that the data is normally distributed as seen from $Sig_{count} > \alpha$, with $\alpha = 0.05$. This shows that the pretest and posttest data are normally distributed, so that the experimental class has students' cognitive abilities that are normally distributed.

Table 3.2 Average Student Learning Outcomes

Class	Pretest	Posttest
Experiment	32,4	80,23

Based on the research that has been done at SMA Negeri 1 Purba, it can be concluded that the chemical module based on problem based learning on the buffer solution material developed is effective and feasible to use for the learning process. With the results of testing the hypothesis that the average student learning outcomes are 80.23. It can be seen that the average learning outcomes obtained are higher than the Minimum Completeness Criteria (KKM) of 75. Where H_a is accepted and H_0 is rejected.

4. CONCLUSIONS

Based on the results of material validation, it shows that the 1) problem-based learning chemistry module of the developed buffer solution material is suitable for use as learning material to improve students' critical thinking skills. The module was declared effective as seen from the increase in student learning outcomes, namely the average pretest score of 33.5 and posttest 80.13 in the experimental class. 2) The results of learning chemistry using a Problem Based Learning (PBL)-based module developed on the buffer solution material is higher than the KKM value (75) with an average value of 80.13.

5. REFERENCES

[1] H. Y. P. Sibuea, “Pembaruan Sistem Pendidikan Di Indonesia : Perkembangan Dan Tantangan,” *J. Kaji.*, vol. 22, no. 2, pp. 151–162, 2017, doi: 10.22212/kajian.v22i2.1520.

[2] N. Rizkia, S. Sabarni, A. Azhar, E. Elita, and R. D. Fitri, “Analisis Evaluasi Kurikulum 2013 Revisi 2018 Terhadap Pembelajaran Kimia Sma,” *Lantanida J.*, vol. 8, no. 2, p. 168, 2021, doi: 10.22373/lj.v8i2.8119.

[3] S. Yudha, O. A. Saputra, R. Purwanto, and A. W.

Nugraha, “Analysis of Chemical Teaching Materials for Class X SMA / MA on The Discussion of The Role of Chemistry in Daily Life,” *J. Pendidik. dan Pembelajaran Kim.*, vol. 10, no. 3, pp. 109–117, 2021, doi: 10.23960/jppk.v10.i3.2021.11.

[4] Z. Rifka, I. Khaldun, and A. Ismayani, “Analisis Pelaksanaan Penilaian Autentik Kurikulum 2013 Oleh Guru Kimia Di SMA Negeri Banda Aceh Tahun Pelajaran 2016 / 2017 Pendahuluan,” *J. Ilm. Mhs. Pendidik. Kim. Vol.2.*, vol. 2, no. 3, pp. 248–255, 2017, [Online]. Available: http://jim.unsyiah.ac.id/pendidikan-kimia/article/view/4929

[5] J. Purba, A. Sutiani, F. T. M. Panggabean, M. Isnaini, and H. D. Hutahaean, “Implementasi Bahan Ajar Kimia Umum Online Terintegrasi Media Dalam Meningkatkan HOTS Ditinjau Dari Kemampuan Awal Mahasiswa,” *J. TIK dan Pendidik.*, vol. 9, no. 1, pp. 52–59, 2022, doi: 10.24114/jtikp.v9i1.35481.

[6] D. Rahman, A. Adlim, and M. Mustanir, “Analisis Kendala Dan Alternatif Solusi Terhadap Pelaksanaan Praktikum Kimia pada SLTA Negeri Kabupaten Aceh Besar,” *J. Pendidik. Sains Indones.*, vol. 3, no. 2, pp. 1–13, 2015, [Online]. Available: http://jurnal.unsyiah.ac.id/jpsi

[7] N. S. Herawati and A. Muhtadi, “Pengembangan modul elektronik (e-modul) interaktif pada mata pelajaran Kimia kelas XI SMA,” *J. Inov. Teknol. Pendidik.*, vol. 5, no. 2, pp. 180–191, 2018, doi: 10.21831/jitp.v5i2.15424.

[8] F. T. M. Panggabean, J. Purba, A. Sutiani, and M. A. Panggabean, “Analisis Hubungan Antara Kemampuan Matematika dan Analisis Kimia Terhadap Hasil Belajar Kimia Materi Kesetimbangan Kimia,” *J. Inov. Pembelajaran Kim.*, vol. 4, no. 1, p. 18, 2022, doi: 10.24114/jipk.v4i1.32904.

[9] K. Dwiningsih, Nf. Sukarmin, Nf. Muchlis, and P. T. Rahma, “Pengembangan Media Pembelajaran Kimia Menggunakan Media Laboratorium Virtual Berdasarkan Paradigma Pembelajaran Di Era Global,” *Kwangsan J. Teknol. Pendidik.*, vol. 6, no. 2, pp. 156–176, 2018, doi: 10.31800/jtp.kw.v6n2.p156--176.

[10] R. Imanda, I. Khaldun, and A. Azhar, “Pengembangan Modul Pembelajaran Kimia Sma Kelas Xi Pada Materi Konsep Dan Reaksi-Reaksi Dalam Larutan Asam Basa,” *J. Pendidik. Sains Indones.*, vol. 5, no. 2, pp. 42–49, 2018, doi: 10.24815/jpsi.v5i2.9816.

[11] F. T. M. Panggabean, P. M. Silitonga, and M. Sinaga, “Development of CBT Integrated E-Module to Improve Student Literacy HOTS,” *Int. J. Comput. Appl. Technol. Res.*, vol. 11, no. 05, pp. 160–164, 2022, doi: 10.7753/ijcatr1105.1002.

[12] I. P. P. A. Antara, “Model Problem Based Learning untuk Meningkatkan Hasil Belajar Kimia Pada Pokok Bahasan Termokimia,” *J. Educ. Action Res.*, vol. 6, no. 1, pp. 15–21, 2022, doi: 10.23887/jear.v6i1.44292.

[13] R. Silaban, F. T. M. Panggabean, F. M. Hutapea, E. Hutahaean, and I. J. Alexander, “Implementasi Problem Based-Learning (PBL) dan Pendekatan Ilmiah Menggunakan Media Kartu Untuk Meningkatkan Hasil Belajar Peserta Didik Tentang Mengajar Ikatan Kimia,” *J. Ilmu Pendidik. Indones.*, vol. 8, no. 2, pp. 69–76, 2020, doi: 10.31957/jipi.v8i2.1234.

[14] M. Y. Soleh, S. Santosa, and M. Indrowati, “Studi Komparasi Penerapan Model Pembelajaran Problem Based Learning dan Inkuiri Terbimbing terhadap Keterampilan Proses Sains Siswa Kelas X SMA Negeri 3 Boyolali Tahun Pelajaran 2013/2014,” *Bio-Pedagogi*, vol. 3, no. 2, p. 1, 2014, doi: 10.20961/bio-pedagogi.v3i2.5328.

[15] E. N. U. Cholifah, S. Yamtinah, and E. Susanti VH, “Hubungan Kemampuan Analisis dan Matematika dengan Prestasi Belajar Siswa pada Materi Larutan Penyangga Kelas XI SMA Negeri 4 Surakarta,” *J. Pendidik. Kim.*, vol. 8, no. 2, p. 179, 2019, doi: 10.20961/jpkim.v8i2.25340.

[16] M. Gultom, D. Fitriyani, M. Paristiowati, Moersilah, Yusmaniar, and Y. Rahmawati, “Analisis Miskonsepsi pada Materi Larutan Penyangga Menggunakan Two-Tier Diagnostic Test,” *JRPK J. Ris. Pendidik. Kim.*, vol. 9, no. 2, pp. 58–66, 2019, doi: 10.21009/jrpk.092.01.

[17] N. Rohmiyati, A. Ashadi, and S. B. Utomo, “Pengembangan modul kimia berbasis inkuiri terbimbing pada materi reaksi oksidasi – reduksi,” *J. Inov. Pendidik. IPA*, vol. 2, no. 2, p. 223, 2016, doi: 10.21831/jipi.v2i2.4869.

[18] J. Purba, F. T. M. Panggabean, A. Widarma, and A. Sutiani, “Development of Online General Chemistry Teaching Materials Integrated with HOTS-Based Media Using the ADDIE Model,” *Int. J. Comput. Appl. Technol. Res.*, vol. 11, no. 05, pp. 155–159, 2022, doi: 10.7753/IJCATR1105.1001.

[19] H. Sulistiani, W. Sumarni, and T. A. Pribadi, “Pengembangan Modul Ipa Terpadu Pada Model Pembelajaran Berbasis Masalah–Pertanyaan Socratik (Mpbm-Ps) Tema Carbon Cycle Untuk Siswa Smp Kelas Vii,” *Unnes Sci. Educ. J.*, vol. 4, no. 2, pp. 905–911, 2015, doi: 10.15294/USEJ.V4I2.7941.

[20] S. Nur, I. P. Pujiastuti, and S. R. Rahman, “Efektivitas Model Problem Based Learning (PBL) terhadap Hasil Belajar Mahasiswa Prodi Pendidikan Biologi Universitas Sulawesi Barat,” *Saintifik*, vol. 2, no. 2, pp. 133–141, 2016, doi: 10.31605/saintifik.v2i2.105.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 305-307, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1002

Salesforce Contact Validation using LinkedIn Sales Navigator for CRM

Sameer F
Computer Science and Engineering
RV College of Engineering
Bangalore, India

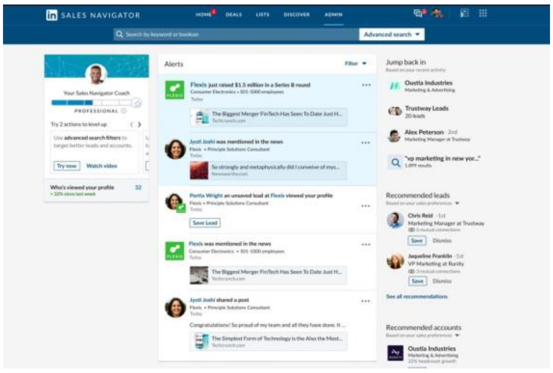
Dr. Chethana R Murthy
Computer Science and Engineering
RV College of Engineering
Bangalore, India

Abstract: Customer relationship management is a process in which a business or other organization administers its interactions with customers, typically using data analysis to study large amounts of information. Keeping the CRM contacts up-to date and in sync is an important task as it is the main pillar to any Customer Relation. This paper presents the use of the LinkedIn Sales Navigator for validating the out-of-sync contacts in CRM. LinkedIn Sales Navigator is a sales intelligence platform that enables virtual selling by allowing sales professionals to build and maintain relationships with their buyers at scale. This paper makes use of Selenium, Java, JSON, and automation to validate the Salesforce Contacts.

Keywords: Salesforce, LinkedIn sales navigator, CRM, Automation, Selenium

I. INTRODUCTION

LinkedIn Sales Navigator is LinkedIn's flagship product for sales teams, enabling reps, managers, and ops leaders to inform their approaches and strategies by taking advantage of the full breadth of LinkedIn's expansive data, insight, and relationship-building tools. Sales Navigator is a sales intelligence platform that enables virtual selling by allowing sales professionals to build and maintain relationships with their buyers at scale. Sales Navigator is designed to be a centerpiece and fixture for modern B2B sales teams, integrating with other sales technologies (such as CRM) to provide a foundation of trusted, reliable, real-time data.



Use advanced search filters to target better leads and accounts.

Who's viewed your profile 32

Flexis just raised $1.5 million in a Series B round
Consumer Electronics • 501-1000 employees
Today
The Biggest Merger FinTech Has Seen To Date Just H...
Techcrunch.com

Jyoti Joshi was mentioned in the news
Flexis • Principle Solutions Consultant
Today
So strongly and metaphysically did I conceive of mys...

Portia Wright an unsaved lead at Flexis viewed your profile
Flexis • Principle Solutions Consultant
Today

Flexis was mentioned in the news
Consumer Electronics • 501-1000 employees
Today
The Biggest Merger FinTech Has Seen To Date Just H...
Techcrunch.com

Jyoti Joshi shared a post
Flexis • Principle Solutions Consultant
Today
Congratulations! So proud of my team and all they have done. It ...
The Simplest Form of Technology is the Also the Most...
Techcrunch.com

Oustia Industries
Marketing & Advertising
Trustway Leads
20 leads
Alex Peterson · 2nd
Marketing Manager at Trustway
"vp marketing in new yor..."

Chris Reid · 1st
Marketing Manager at Trustway
Jaqueline Franklin · 1st
VP Marketing at Runity

Oustia Industries
Marketing & Advertising

Figure 1: LinkedIn Sales Navigator

CRM (Customer Relationship Management) is a kind of software that stores customer contact information like name, address, age, phone number. It also keeps tracks of customer activity like website visits, numbers of outgoing and incoming phone calls, email, and more.

Salesforce is a cloud-based Customer Relationship Management (CRM) software for managing customer relationships and integration with other systems. This SaaS tool helps to create custom solutions for marketing, sales, services and ecommerce as per business requirements.

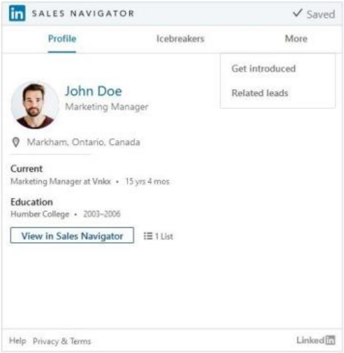
Test automation is the use of software separate from the software being tested to control the execution of tests and the comparison of actual outcomes with predicted outcomes. Test automation can automate some repetitive but necessary tasks in a formalized testing process already in place or perform additional testing that would be difficult to do manually.

Selenium is a free (open-source) automated testing framework used to validate web applications across different browsers and platforms. You can use multiple programming languages like Java, C#, Python etc. to create Selenium Test Scripts.

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. JSON is an open standard file format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays.

2. AIM

The aim of this is paper is to validate contacts in Salesforce CRM using LinkedIn sales navigator to understand if contact is currently with the company or not. The contact validation doesn't have any value unless we understand where contact has moved to. Currently LinkedIn contact data validation feature is limited to only providing information about if contact is with company or not and doesn't provide details to contact's current details like Company Name, Address, Title etc.



John Doe
Marketing Manager
Markham, Ontario, Canada
Current
Marketing Manager at Vnlox • 15 yrs 4 mos
Education
Humber College • 2003–2006

Figure 2: Sales Navigator – Contact Details

Sales Navigator helps to perform three of the most critical functions:

- **Target**: Quickly identify and learn about people and companies.
- **Understand**: Track key developments at target accounts, such as decision makers changing jobs or indicators of buying intent, to act on opportunities as they arise.
- **Engage**: Connect and converse with prospects within a ready-to-do-business environment while tapping into the full extent of LinkedIn's messaging and content-sharing capabilities.

With the help of Sales Navigator, the contact's current details like Company Name, Address, Title, Location and so on can be acquired and validated with CRM data which helps to sync the contacts and keep the CRM data up-to date.

3. RELATED WORK

In this section, we look into the automation part of Contact validation, the tools and programming language that can be used to automate Salesforce CRM and LinkedIn Sales Navigator are discussed.

In [1], the paper addresses the challenge of Automating the testing process for application landscape frameworks that is hard due to the complexity caused by the variety of customer application landscape configurations, used tools and platforms. Although the paper doesn't cover wide range of application landmark frameworks.

In [2], author focuses to solve harder problems such as (1) what are start-up and tear-down activities of tests, (2) what are interesting test data, (3) what expected behavior to check in assertions. They use Meta-heuristic to find test inputs, Symbolic execution, TTCN-3 for test control. The paper doesn't have practical implementation of the solution it provides.

[3] Paper compares Postman, JMeter, and Robot Framework and constructs a test and evaluation system based on different data environments. The paper uses PM model and page object model for developing the test scripts.

The aim of [4] is evaluation and comparison of Katalon Studio and Selenium, used for web application testing. They followed Page object model (POM). The drawback of this paper is paper doesn't talk about test data automation.

[5] presents a description of automated software testing tools based on the type of the test and specifies which tools are the best and more efficient. The paper doesn't conduct experiment to compare the different testing tools and paper lacks a practical demonstration of tools.

4. DESIGN

LinkedIn Sales Navigator has an integration with Salesforce which provides a capability known as Contact Data Validation. It compares the Contacts and their Account associations in CRM with LinkedIn Data and determines which of the contacts are outdated, i.e., they are associated to incorrect account. It populates a field LID__No_longer_at_Company__c on CRM contacts which tells us if the contact has moved to a new company. LinkedIn doesn't have any APIs which will provide this information. Only way is to use the LinkedIn Widget on the Contact Layout, which pulls in the latest contact information from LinkedIn and shows what is their new company name and update contact manually in CRM. Doing this manually for 120k outdated contacts was not feasible so feasible option is to automate this process.

We have used Selenium with Java for automating the process of Contact Data validation.

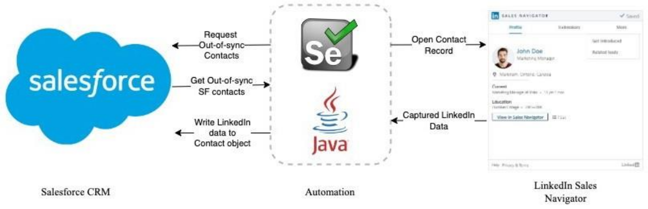


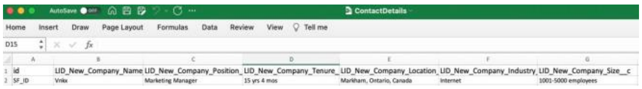
Figure 3: Flow Diagram

The steps in the automation process are as follows:

1. Query Salesforce for out of sync contacts
   - Establish connection with Salesforce CRM
   - Query the out-of-sync contacts
   - Get the CSV output and save it in disk.
2. Get the Contact details from LinkedIn Sales Navigator
   - Read the CSV using OpenCSV java library
   - Open the Contact's record page
   - Capture LinkedIn data for the Contact
   - Write back the captured data to CSV
3. Update the CRM data
   - Read the CSV using OpenCSV java library
   - Create JSON objects of Contact details from CSV
   - Establish connection with Salesforce CRM
   - Update the data to CRM

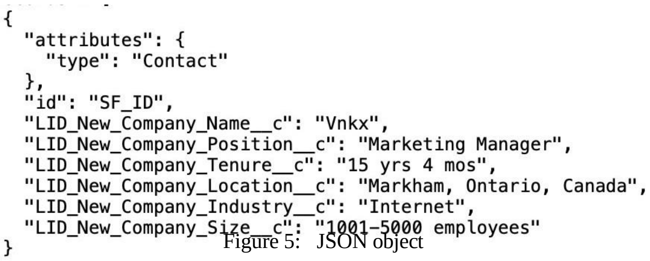
5. RESULTS AND ANALYSIS

We successfully validated the out-of-sync CRM contacts using LinkedIn Sales Navigator. The LinkedIn Sales Navigator provided the contact details such as the current Company, position, Company location, tenure, industry and Company size. These contact details from Sales Navigator helped us for Contact validation.


AutoSave
ContactDetails
Home Insert Draw Page Layout Formulas Data Review View Tell me
D15
fx

	A	B	C	D	E	F	G
1	id	LID_New_Company_Name	LID_New_Company_Position_	LID_New_Company_Tenure_	LID_New_Company_Location_	LID_New_Company_Industry_	LID_New_Company_Size__c
2	SF_ID	Vnkx	Marketing Manager	15 yrs 4 mos	Markham, Ontario, Canada	Internet	1001-5000 employees

Figure 4: CSV with Contact Details



```
{
  "attributes": {
    "type": "Contact"
  },
  "id": "SF_ID",
  "LID_New_Company_Name__c": "Vnkx",
  "LID_New_Company_Position__c": "Marketing Manager",
  "LID_New_Company_Tenure__c": "15 yrs 4 mos",
  "LID_New_Company_Location__c": "Markham, Ontario, Canada",
  "LID_New_Company_Industry__c": "Internet",
  "LID_New_Company_Size__c": "1001-5000 employees"
}
```

Figure 5: JSON object

The Selenium automates the process of querying for the out- of-sync contacts, storing that data to CSV, opening the Contact's record in LinkedIn Sales Navigator and collecting data from Sales Navigator and writing that to the CSV. For the updating data on to CRM, a Selenium script reads all the data from CSV and creates JSOM objects for each contact and updates the contact data to CRM through API calls.

6. CONCLUSION

The Customer Relationship management plays a vital role in any organization. Keeping the CRM contacts up-to date and in sync is an important task. The paper proposes a way to validate the out-of-sync contacts of CRM using LinkedIn Sales Navigator. The validation process is automated using Selenium with Java.

The validation takes some time, as the automation script opens each contact's record in Sales Navigator and then collects the data, this time can be reduced by executing the automation script in parallel.

7. ACKNOWLEDGEMENT

I would like to express my gratitude to Dr. Chethana R Murthy and RV College of Engineering for all their assistance and support.

8. REFERENCES

1. N. Wild, H. Lichter and P. Kehren, "Test Automation Challenges for Application Landscape Frameworks," 2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2020, pp. 330-333, doi: 10.1109/ICSTW50294.2020.00059.
2. Y. Labiche, "Test Automation - Automation of What?," 2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2018, pp. 116-117, doi: 10.1109/ICSTW.2018.00037.
3. C. -H. Hsieh et al., "Evaluation System for Software Testing Tools in Complex Data Environment," 2021 4th International Conference on Information Communication and Signal Processing (ICICSP), 2021, pp. 604-609, doi: 10.1109/ICICSP54369.2021.9611846.
4. E. Pelivani and B. Cico, "A comparative study of automation testing tools for web applications," 2021 10th Mediterranean Conference on Embedded Computing (MECO), 2021, pp. 1-6, doi: 10.1109/MECO52532.2021.9460242.
5. S. K. Alferidah and S. Ahmed, "Automated Software Testing Tools," 2020 International Conference on Computing and Information Technology (ICCIT-1441), 2020, pp. 1-4, doi: 10.1109/ICCIT-144147971.2020.9213735.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 308-314, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1003

The Implementation of the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) in the Decision Support System for Determining Village Development Planning

Oktaviani Ika Wijayanti
Faculty of Economics and Business
Brawijaya University
Malang, Indonesia

Khusnul Ashar
Faculty of Economics and Business
Brawijaya University
Malang, Indonesia

Nurul Badriyah
Faculty of Economics and Business
Brawijaya University
Malang, Indonesia

Abstract: The research was conducted to respond to the existence of personal or group interests in determining village development planning, so to minimize personal and group interests, a decision support system is needed with the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) approaches that have good performance. The purpose of the study was to determine the performance of the method, the quality of the application based on usability. The research was conducted by distributing questionnaires to respondents or participants as expiry users. Where the questionnaire is divided into two parts, namely: the first is a questionnaire to measure the performance of the ANP and SAW methods. The second questionnaire uses the USE qussenere to measure the quality of the application system. The results showed that the performance of the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) methods on the decision support system obtained a Pk% result of 70% which was included in the feasible category. On the quality of the application system based on the usability test value of 75.67%. So that the system is included in the category of decent or good quality.

Keywords: Performance, Decision Support System, Analytical Network Process (ANP), Simple Additive Weighting (SAW)

1. INTRODUCTION

This research was conducted to respond to the existence of personal and group interests in determining village development planning. Based on the regulation of the Minister of Villages/Kelurahan, Development of underdeveloped areas, and transmigration Number 5 of 2015 concerning the determination of priorities for the use of Village/Urban funds in 2015. Village/Kelurahan Funds are funds sourced from the State Revenue and Expenditure Budget designated for transferred Villages/Kelurahan through the Regency/City Regional Revenue and Expenditure Budget and used to fund governance, development implementation, community development, and community empowerment [1]. Problems that often occur in villages are that the development stage in the village must consider the priority scale and elements of justice, as well as the absence of a Decision Support System for Determining Development Priorities in the village and the system currently being used is not maximally computerized.

Ideally, in determining village development, it starts from the process of planning program activities and making decisions that are free from personal and group interests. However, in reality the planning and decision-making processes do not work as they should. This is because the government's role in implementation is still centralized with top-down planning, so that decision-making is dominated by village elites, and is an annual formal routine. Meanwhile, the results of interviews with the Village Head stated that in the planning process many personal or group interests were involved in proposing village development plans. Therefore, a decision support system (DSS) is needed with an approach that can handle multiple criteria and non-structural problems, thereby minimizing personal and group interests.

The decision support system that produces recommendations for development planning priorities uses the criteria set by the government (Permendagri No. 66, 2007)[2] which can be adjusted based on needs. In addition to these criteria, we need a method that can solve semi-structured and non-structured problems that can minimize the existence of personal and group interests in community proposals. The method used in the DSS for determining village development planning must have good performance. The performance of a method can be measured by using a questionnaire containing the appropriate and accommodative variables. While the quality of a system can be measured through Usability Testing by using a USE Questionnaire that contains several variables, namely usability, easy to use, easy to learn and community satisfaction as consumers regarding perceptions of the quality of an application or product.

Several studies on decision support systems that do not or involve the role of public participation have been carried out. Like the research that has been done by Nababan & Tuti on determining the feasibility of operating a house for poor families using the Weighted Product (WP) method [3] which also still has weaknesses in using this method, namely it does not have costs and benefits for the criteria so that it affects the priority weight level. Meanwhile, Aziz, Febriani, Sopandi, & Gustian's research on the decision support system for determining development priorities using the Analytical Hierrachy Process (AHP) [4] still has weaknesses, because the weighting of the method is still of interest and subjective.

Although there has been research on DSS related to development planning, this research implements the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) methods in the decision support system for determining village building planning. The ANP method is used to determine the priority weight of the criteria, where the method can solve problems with many criteria (multi-criteria) and is able to accommodate the relationship of influence between criteria, so that it will eliminate subjectivity in weighting criteria. Meanwhile, the Simple Additive Weighting (SAW) method is used to get village development planning priorities, because of its ability to make a more precise assessment based on the predetermined cost and benefit criteria.

Based on the description above, the focus of this research is to develop a decision support system for determining village development planning using the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) methods to assist decision makers in determining development planning priorities in accordance with the needs of the village community.

1.1 Decision Support System

The concept of a Decision Support System was first introduced in the early 1970s by Michael S. Scott Morton with the term Management Decision System (Sari, 2018). The concept of decision support is characterized by a computer-based interactive system that helps decision makers utilize data and models to solve unstructured problems. Basically DSS is designed to support all stages of decision-making starting from identifying problems, selecting relevant data, determining the approach used in the decision-making process, to evaluating alternative choices [5].

The definition of a Decision Support System (DSS) itself is a flexible, interactive and adaptable computer-based information system developed to support solutions to unstructured specific management problems. Decision Support Systems use data, provide an easy user interface and can incorporate decision-making thinking [6]. Kusrini in his book entitled Concepts and applications of decision support systems defines an information system that provides information, modeling and manipulating data [7]. Meanwhile, Hafiz & Ma'mur define a computer-based information system that provides interactive supporting information between other stakeholders during decision making. So from some definitions Decision Support System can be said as a computer system that helps in managing data into information that can solve problems and provide the right decisions [8].

1.2 ANP dan SAW

Decision making, usually more often used a hierarchical method consisting of goals, criteria, and alternatives. The use of a hierarchy is to make it easier for decision makers. However, there are times when decision making does not only pay attention to the hierarchical structure, but also the network or the dependence and feedback between elements in the cluster (inner dependence) and between clusters (outer dependence). According to Rusydiana & Devi, feedback is able to properly capture the influence of interactions, especially when decision makers are faced with risks and uncertainties in a complex business environment [9]. ANP uses a system of pairwise comparisons to measure the weight of structural components, and in turn makes a ranking of the best alternative choices that must be taken.

The Analytic Network Process (ANP) is a multi-criteria assessment method for decision structuring and analysis that has the ability to measure the consistency of assessment and flexibility in choices at the sub-criteria level. Meanwhile, Saaty defines ANP as a relative measurement method used to derive the composite priority ratio from the individual ratio scale that reflects the relative measurement of the influence of interacting elements with respect to control criteria [10]. ANP is able to accommodate linkages between criteria or alternatives, and allows interaction and feedback from elements within the cluster and between clusters.

According to Nofriansyah, the Simple Additive Weighting (SAW) is often also known as the weighted addition method [11]. The basic concept of the SAW method is to find the weighted sum of the performance ratings for each alternative on all criteria [12]. The SAW method requires the process of normalizing the decision matrix (X) to a scale that can be compared with all existing alternative ratings. The SAW method recognizes the existence of 2 (two) attributes, namely the benefit criteria and the cost criteria. The basic difference between these two criteria is in the selection of criteria when making decisions.

1.3 Development Planning

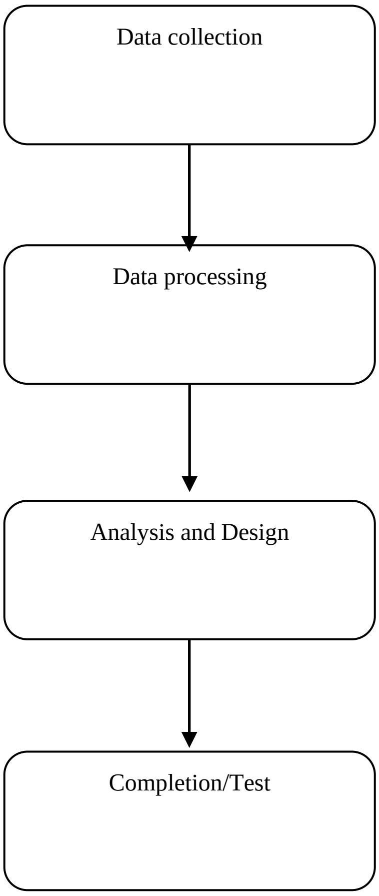
Development is a multi-dimensional process involving important changes in a structure, socio-economic system, public attitudes and national institutions and accelerating economic growth, unemployment inequality and eradicating absolute poverty [13]. In addition to the above understanding, experts provide various definitions of development, but in general there is an agreement that development is a process to make changes to the village, both physical and non-physical. Kartasasmita provides a simpler understanding of the development of a process of change for the better through planned efforts [14]. Although the definition of development varies widely, in general, development can be interpreted as a process of change from one national condition to a national condition that is considered better or continuous progress towards the improvement of an established human life.

Village community development is a village change to create an independent and innovative village [15]. Meanwhile, according to Tjokrowinoto, village development can be carried out based on three principles, namely the principle of integral development, the principle of own strength, and the principle of mutual consensus. The principle of integral development is balanced development from all aspects of the village community [16][17]. The principle of self-strength is that each effort must first be based on its own strength (Soekanto, 2013)[18], the principle of mutual agreement is that development must be carried out correctly to become the needs of the village community and the decision to implement the project is not on the priority of superiors but is a joint decision of the members. society [19]. In the end, village development is the development of village independence starting from a good village planning process, followed by good program management.

Effective village development is not solely due to opportunities but is the result of determining activity priority options, not the result of trial and error, but the result of good planning, because development needs are greater than available resources. Through planning, we want to formulate development activities that efficiently and effectively can provide optimal results in utilizing available resources and developing existing potential.

2. Meotode Penelitian

Broadly speaking, this research uses a descriptive quantitative approach, with the research stages divided into four stages, namely data collection, data processing, analysis and design and completion/testing which is shown in Figure 1 research flow chart..



Gamabar 1. Diagram alur Penelitian

2.1 Data collection

The supporting data in this study is the criteria data that affect the determination of village development planning priorities. These data include urgency (K), public interest (KU), Availability of potential (KP), perceived by many people (DBO), Barriers to income (MP), Cost (B). These data are obtained based on the results of a literature study that can be used for making a decision support system using the ANP method. In addition to the data mentioned in this study, it also requires some supporting data in conducting usability testing, namely, ease of use, ease of use, easy to learn and satisfaction. These data are used to assess the level of quality of the decision support system that has been made and implemented.

2.2 Data processing

The data and information that has been obtained will be used in data processing which includes several activities such as the following:

➢ Criteria analysis and selection

At this stage the aim is to determine what criteria will be used to assess whether the proposed development planning priority program is appropriate or not. The criteria are the results of the author's analysis of journal references, village laws, village planning technical guidelines, books and articles which are then described as a network model.

➢ Determination of Influence Relationship

Determination of the influence relationship is used to determine whether there is an influence relationship between the criteria/clusters with each other. The results of these determinations are used to build the network structure of the ANP method.

➢ Criteria priority weighting and ANP calculation

The priority weighting of the criteria is used to determine how big the relationship is between one criterion and another to the criteria that are affected in determining the choice of priority programs. While the calculation with the ANP method by calculating the priority weight of each criterion. Then create a super matrix which includes a weightless super matrix, a weighted super matrix and a limit matrix. The result of this matrix limit calculation is a list of criteria weight values that have been sorted based on the largest calculated value.

➢ Construction plan weighting and Simple Additive Weighting (SAW) calculations

The weighting is done by providing an assessment of the proposed program based on each of the predetermined criteria. This Simple Additive Weighting (SAW) method requires the decision system to determine the weight for each attribute. The total score for the proposed alternative program is obtained by adding up all the results of the multiplication between the rating and the weight of each criterion, the result of the calculation is a list of alternative weight values that have been sorted based on the largest calculated value.

2.3 System Analysis and Design

The purpose of this stage is to analyze the system that will be developed according to the needs of the participants. Furthermore, the specification of user needs will be known and who will use the system (User).

➢ Database Design and Development

The database is used as a storage medium for input and output data from the decision support system. The database on this system uses MySQL which is an open source database management system.

➢ System Design and Development

The system developed will be based on a website so that it can be accessed by users from anywhere and anytime. With this system, it is hoped that it can accommodate the process of village community proposals (participants) and decision makers in determining priorities for village development planning programs that are considered the most appropriate using the ANP and SAW methods.

2.4 Testing

➢ Research Instruments

The research instrument used to test the performance of the method and usability test is a series of questionnaires that can process data related to suitability, effectiveness, efficiency, satisfaction with the use of a decision support system. The thing that underlies the use of questionnaires is that questionnaires can provide convenience for respondents to understand and answer the questions asked properly. In addition, the questionnaire makes respondents more comfortable and flexible in answering questions (Munir, 2010). The complete form of the questionnaire package and the Likert measurement scale are shown in Tables 1 and 2..

Table 1. Usability and Method Performance Questionnaire

Variable	No.	Indicator
Usefulness	1	This system helps me to be more effective
	2	This system helps me to be more productive
	3	This system is very useful
	4	This system saves me time when using it
Ease of Use	5	This system is easy to use
	6	This system is practical to use
	7	This system is User Friendly
	8	I don't see any inconsistencies while using this system
	9	Errors that occur in this system are easy to recover quickly and easily
Ease of Learning	10	I learned to use the system quickly
	11	I can easily remember how to use this system
Satisfication	12	I am satisfied with this system
	13	Using this system is a lot of fun
	14	This system works exactly what I want
	15	This system is very comfortable when used
Performa Metode	16	Is this system in accordance with the priority level of the proposed activity program that you want?
	17	Has the proposal you put forward through this system been accommodated?

Table 2. Criteria for measuring the Likert Scale

Score	Answer Criteria
1	1=Strongly Disagree(STS)
2	2=Disagree (TS)
3	3=Sufficiently Agree (CS)
4	4=Agree (S)
5	5=Strongly Agree (SS)

➢ System Feasibility Test and Questionnaire

System and questionnaire feasibility tests need to be carried out to ensure that the results of the system and questionnaire data collection are suitable for analysis. A system and questionnaire that will be used in research must have valid and reliable properties so that it is feasible to be used as a research instrument.

The feasibility test of the system was carried out using the validity of the system by comparing the similarity of the results of the ANP and SAW calculations on the system to the results of manual calculations. If the results of the system are the same as the manual, then it is said to be valid, but if it is not the same then it is invalid, and analysis and improvement must be carried out on the system until it is valid.

The questionnaire feasibility test was carried out using two methods, namely validity and reliability tests. Validity test is used to determine the feasibility of the items in a question. The validity test used is Pearson's corellate bivariate (product moment correlation) and the r table is significant with 5%. While the reliability test was conducted to determine the consistency and reliability of the measuring instrument. In this study, the reliability test was carried out using the Cronbach's Alpha measure. To determine the level of reliability of the instrument used the categories shown in table 3.

Table 3 Reliability Level of Cronbach's Alpha

Reliability Interval	Categori
$0{,}80 < r_{11} \leq 1{,}00$	Very high reliability
$0{,}60 < r_{11} \leq 0{,}80$	High reliability
$0{,}40 < r_{11} \leq 0{,}60$	Medium reliability
$0{,}20 < r_{11} \leq 0{,}40$	Low reliability
$0{,}00 < r_{11} \leq 0{,}20$	Unreliable

➢ Test Method Performance and Usability

Measurement of performance and usability by calculating the percentage of answers from respondents using the formula stated in (1).

$$PK(\%) = \frac{Sekor\ yang\ diobsevasi}{Sekor\ yang\ diharapkan} x\ 100\ \% \qquad (1)$$

The data obtained is then converted based on the table of eligibility categories as shown in table 4.

Table 4. Eligibility Category

Score	Category
<21	Very unworthy
21-40	not feasible
41-60	Enough
61-80	Worthy
81-100	Very worth it

➢ Analysis and Processing of Questionnaire Results

Analysis of the results of the questionnaire was carried out after processing the data first. Data processing is carried out after getting the results of the validity and reliability tests in accordance with the provisions. This data processing aims to measure the percentage value of the feasibility of the method performance and the quality of the application system in the USE questionnaire.

3. RESULT AND DISCUSSION

3.1 System Validity

Decision support system is a product of information systems / information technology that uses a mathematical calculation method approach. As a product with a mathematical approach, validation should be carried out to determine compliance with the calculation rules based on that method. Validation of calculations on the application system is very important to do, this is because it is closely related to the priority weights of criteria and proposals that have an impact on the ranking of criteria and proposals. Validation is done by giving the same input data to the application system and the manual, then the output from the system is compared to the similarity to the manual results. If the system output to the manual has the same value, it is said to be valid, if it is not the same, it is not said to be valid, and an analysis of improvements to the system must be carried out until it is valid. The results of the criteria can be seen in Table 5.

Table 5. Output criteria on the system and manual

Criteria	ANP	
	system	Manual
Inhibiting Income (II)	7,988	7,988
Urgency (U)	7,101	7,101
Felt by Many People FMP)	3,742	3,742
Public Interest (PI)	5,733	5,733
Potential Availability (PA)	5,352	5,352
Cost (C)	4,400	4,400

Table 5. Output criteria for the system and manual, showing the criteria with II code in the system calculation using the ANP method has a value of 7.988, U criteria of 7.101, FMP criteria of 3.742, PI of 5.733, PA of 5.352 and criteria C of 4.400. Based on these data, there is no difference in the results found in the results of calculations and system testing against the manual. This shows that the value of the criteria with the ANP method approach on the decision support system is valid.

Table 6. Alternative Outputs on System and Manual

Alaternative	SAW	
	System	Manual
Road repair	30,852	30,852
Water tunnel	28,705	28,705
Entrepreneurship training	29,821	29,821

While table 6 alternative outputs on the system and manual, with the proposed road improvement in the calculation of the system with the SAW method has a value of 30,852, the culvert program has a weight of 28,705 and the entrepreneurship training program has a weight of 29,821. Based on these data, the results contained in the calculation of the system to the manual are the same. This shows that the decision support system with the SAW method is valid.

3.2 Questionnaire Validity

The validity test of the method performance and usability questionnaire was carried out to find out how much validity the measuring instrument used was with validity analysis using the bivariate correlation product moment method with the help of the SPSS 16 for Windows program. The following are the results of the instrument validity test, application quality and method performance.

Table 7. Results of the Validity Test of Questionnaire Question Items

Indicator	r count	r table	Information
Q1	0,502	0,344	Valid
Q2	0,473	0,344	Valid
Q3	0,697	0,344	Valid
Q4	0,439	0,344	Valid
Q5	0,473	0,344	Valid
Q6	0,535	0,344	Valid
Q7	0,479	0,344	Valid
Q8	0,534	0,344	Valid
Q9	0,469	0,344	Valid
Q10	0,405	0,344	Valid
Q11	0,431	0,344	Valid
Q12	0,555	0,344	Valid
Q13	0,489	0,344	Valid
Q14	0,439	0,344	Valid
Q15	0,445	0,344	Valid
Q16	0,867	0,344	Valid
Q17	0,503	0,344	Valid

The results of testing the validity of the questionnaire in table 5.12 which consists of 17 questions that have been filled out by 33 respondents indicate that all items are valid. That is, based on the comparison of the calculated r value greater than r table = 0.344 and has a positive value, the question item is declared valid.

3.3 Reliability

The reliability test was used to measure the reliability of the usability questionnaire and the performance of the method in research. The reliability test in this study used the Cronbach's Alpha method with the help of SPSS 16 for Windows statistics. The results of the reliability test data can be seen in the following table.

Table 8. Reliability test results Questionnaire questions

Reliability Statistics	
Cronbach's Alpha	N of Items
.789	17

According to the data from the reliability test results in table 8, it is known that there are 17 usability and method performance questions with a Cronbach's Alpha value of 0.789. Based on the conversion of Cronbach's Alpha coefficient value to table 3 the level of reliability is in the high category.

3.4 Method Performance by Aspect

The performance of the method in this study is the accuracy of the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) methods which are applied to website-based information communication technology in processing community proposals to become the priority level of activity programs in village development planning. In the context of method performance, accuracy can be seen from the suitability of the priority level of the proposal based on the wishes of the participants and the decision maker's accommodation of the participant's (community) activity program. The results of processing the performance questionnaire data on the application of the ANP and SAW methods to the application system on the aspects of suitability and accommodation are shown in Figure 2.

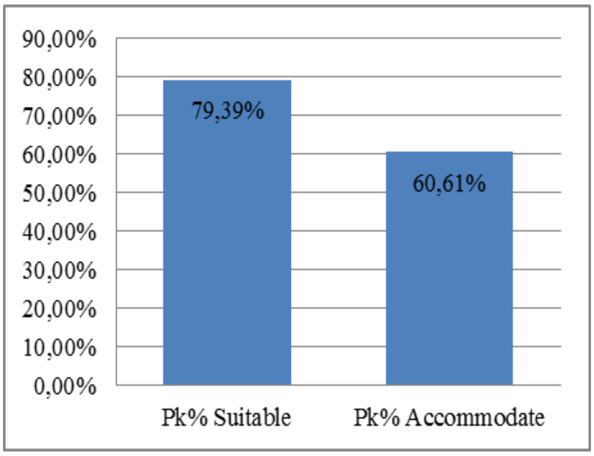


Figure 2. Graph of Method Performance by Aspect

Figure 2 is the result of the performance questionnaire on the application of the ANP and SAW methods to the application system in the aspects of suitability and accommodation. Based on table 4, the system feasibility standard on the suitability aspect shows that the Pk% value of 79.39% is included in the feasible category. Meanwhile, in the accommodative aspect, the Pk% value of 60.61% is in a good category.

3.5 Usability Berdasarkan Aspek

This study uses the USE Questionnaire as a parameter for measuring usability. The questionnaire consists of usability, ease to use, easy to learn, and satisfaction. It is hoped that it can provide information and empirical evidence that the use of the application system is following the needs and can provide convenience for users or participants of West Waru Village. The results of several aspects used to observe the quality or not of an application system are shown in Figure 3.

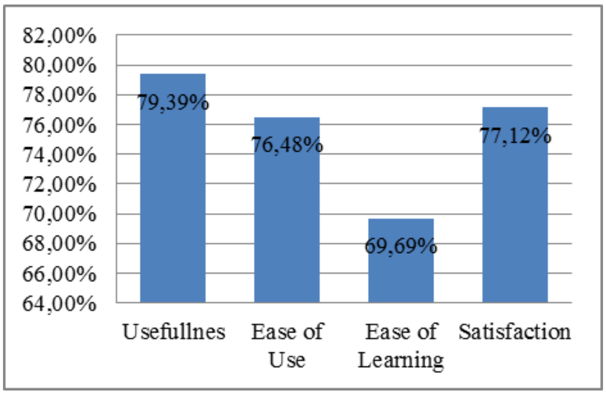


Figure 3. Usability Graph by Aspect

Based on Figure 3 on the aspect of usability (usefulness), which includes effective, productive, useful, and efficient, it is necessary to measure the extent to which the product enables users to achieve their goals. Where the level of usability in the usability aspect using the USE Questionnaire on the application system is 79.39%. Meanwhile, the ease of use aspect which includes easy to use, simple, user-friendly, consistent, and easy to recover is needed to measure how far the ease of use for users is, the Pk% value is 76.48%.

In the easy-to-learn aspect, which includes being fast to learn and easy to remember, it is necessary to measure how far the ease of learning for application system users is. Based on Figure 2, the usability level of the ease of learning in the application system is 69.69%. Finally, the aspect of satisfaction (satisfaction) includes satisfied, pleasant, as desired, and comfortable with Pk% of 77.12%.

Based on each of these values, the PK% level on all aspects of usability is included in the feasible category. It can be seen in table 4 of the system's feasibility standards that the value of 61-80 is included in the appropriate category for application system users in proposing program activities according to their needs..

3.6 Method Performance and Usability

The performance of the method is needed to measure the accuracy of the ANP and SAW methods in processing the priority level of the proposal in the application system. While usability is used to measure how easy the application system is in carrying out its duties. The results of method performance and usability are shown in Figure 4.

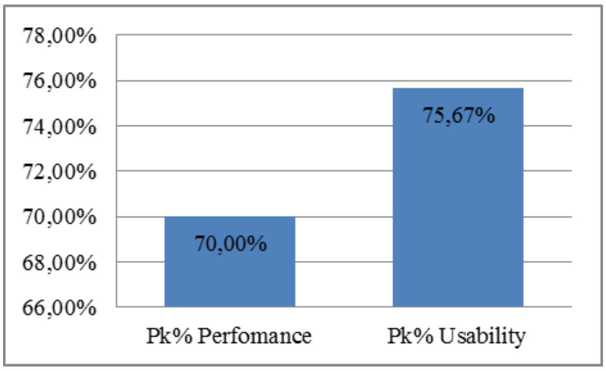


Figure 4. Performance and Usability Test

Figure 4 graphs the method performance and usability, the Pk% value of the method performance using the USE Questionnaire on a decision support system is 70%, while the Pk% usability value is 75.67%. Based on table 4, the feasibility standard for the method performance system on the decision support system and usability is in a good category.

4. CONCLUSION

Based on the results and analysis that has been done, the following conclusions can be formulated in this study, namely:

In terms of suitability, accommodation and overall performance of the Analytical Network Process (ANP) and Simple Additive Weighting (SAW) methods on the decision support system are in the decent or good category.

The quality of the application system is based on usability tests on the aspects of usability (usefulness), ease of use (ease of use), ease of learning (ease of learning), aspects of satisfaction (satisfaction) and overall in the category of decent or good quality.

5. ACKNOWLEDGMENTS

The author would like to thank all those who have helped, especially the fathers and mothers who have provided funding for this research.

6. REFERENCES

[1] Peraturan Menteri Desa, Pembangunan Daerah Tertinggal, Dan Transmigrasi Nomor 5 Tahun 2015 Tentang Penetapan Prioritas Penggunaan Dana Desa.(http://ensiklo.com/2015/03/16/permendes-no-5-tahun-2015-tentang-penetapan-prioritaspenggunaan-dana-desa/ diakses tanggal 12 Juli 2022)

[2] Peraturan Menteri Dalam Negeri No. 66 Tahun 2007 Tentang Perencanaan Pembangunan Desa.

[3] Nababan L., & Tuti E. 2018. Determination Feasibility of Poor Household Surgery By Using Weighted Product Method. International Conference on Cyber and IT Service Management (CITSM), Parapat, IEEE..

[4] Aziz, R. M. A., Febriani, D., Sopandi, A. S., & Gustian, D. 2021. Sistem Pendukung Keputusan Penentuan Prioritas Pembangunan Dengan Analytical Hierrachy Process. Seminar Nasional Pengaplikasian Telematika (SINAPTIKA). Vol 1 No 1. hal. 178-187.

[5] Magdalena, H. 2012. Sistem Pendukung Keputusan untuk Menentukan Mahasiswa Lulusan Terbaik di Perguruan Tinggi (STUDI KASUS STMIK ATMA LUHUR PANGKALPINANG). Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA). Yogyakarta. hal. 49-56. ISSN : 2089-9815.

[6] Turban, E., Sharda, R., & Delen, D. 2011. Decision Support and Business Intelligence Systems. 9th Editon. Pearson Education Inc. India.

[7] Kusrini. 2007. Konsep dan Aplikasi Sistem Pendukung Keputusan. Penerbit Andi. Yogyakarta.

[8] Hafiz, A., & Ma'mur, M. 2018. Sistem Pendukung Keputusan Pemilihan Karyawan Terbaik Dengan Pendekatan Weighted Product (Studi Kasus:PT. Telkom Cab. Lampung). Jurnal Cendikia. Vol. XV. 23-28.

[9] Rusydiana, A. S., & Devi, A. 2013. Analytic Network Process: Pengantar Teori & Aplikasi, Cetakan Pertama. SMART Publishing. Bogor.

[10] Saaty, T. L. 1999. Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World. RWS Publications. Pittsburgh.

[11] Nofriansyah, D., & Defit , S. 2017. Multi criteria decision making (MCDM) pada sistem penduung keputusan, Edisi Pertama. Deepublish. Yogyakarta.

[12] Kusumadewi, S., Hartati, S., Harjoko, A., & Wardoyo, R. 2006. Fuzzy Multi-Attribute Decision Making (Fuzzy MADM). Graha Ilmu. Yogyakarta.

[13] Todaro, Michael. 1977. Pembangunan Ekonomi Di Dunia Ketiga. Erlangga. Jakarta.

[14] Kartasasmita, G. 1997. *Administrasi Pembangunan.* LP3ES. Jakarta.

[15] Kessa, W. 2015. Perencanaan Pembangunan Desa. Cetakan Pertama. Kementerian Desa, Pembangunan Daerah Tertinggal, Dan Transmigrasi Republik Indonesia. Jakarta.

[16] Tjokroamidjojo, B. 1996. Perencanaan Pembangunan. Gunung Agung. Jakarta.

[17] Gai, A. M., Witjaksono, A., & Narjun, M. L. 2020. Identification of the Most Influential Infrastructure for the Development of Disadvantaged Villages in Sumberpetung Village, Malang Regency, Indonesia. International Research Journal of Advanced Engineering and Science. Volume 5, Issue 2, pp. 44-48.

[18] Soekanto, S. 2013. Sosiologi Suatu Pengantar. Rajawali Pers. Jakarta

[19] Achmad, M. 2018. Manajemen Dan Tata Kelola Pemerintahan Desa Perspektif Regulatif Dan Aplikatif. Cetakan Pertama. Balai Puataka. Jakarta

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 315-319, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1004

Developing an ETL Pipeline for Data Analysis

A S Prajwal Babu
Computer Science and Engineering
RV College of Engineering
Bangalore, India

Prof. Suma B
Computer Science and Engineering
RV College of Engineering
Bangalore, India

Abstract:The world's most valuable resource these days is the expanding data. Large organisations continuously produce data about their clients, consumers, and employees in real time. This data cannot be easily interpreted in its raw form, but after being processed and changed, it can be widely used for analytics. This improves a number of the aforementioned business entity's existential traits, including organisational management, market capabilities, and consumer feedback.Given the volume of data that a corporation generates, it is obvious that it will need a significant investment of money, time, talent, and resources to achieve the goal of in-house data processing, calibration, and storage. The goal is to overcome the obstacles businesses present for data-pipelining technology and get processed data directly at the conclusion of the data sync cycle. One sync cycle is the continuous fetching of data created or altered over the course of a given time frame, such as a fortnight or a month.

Keywords: Data pipeline,ETL pipeline Cloud, Data Warehouse, Data Analytics

1. INTRODUCTION

An ETL pipeline is a group of procedures used to transfer data from one or more sources into a database, such as a data warehouse. The three interdependent data integration processes called "extract, transform, and load," or ETL, are used to take data out of one database and transport it to another. Once loaded, data can be used for reporting, analysis, and the creation of useful business insights.

The relevance of utilising such data in analytics, data science, and machine learning programmes to gain business insights develops along with the amount of data, data sources, and data types at organisations. Since turning the raw, unclean data into clean, new, trustworthy data is a crucial step before these projects can be undertaken, the requirement to prioritise these activities puts growing pressure on the data engineering teams. ETL, or extract, transform, and load, is a method used by data engineers to gather data from various sources, transform it into a reliable and useable resource, and then load it into the systems that end users may access and utilise later to address business-related issues.

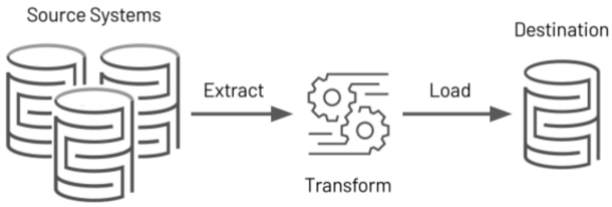


Fig 1: ETL process

Extract

Data extraction from the target sources—which are typically heterogeneous and include business systems, APIs, sensor data, marketing tools, transaction databases, and others—is the initial step of this process. As you can see, while some of these data types are likely to be semi-structured JSON server logs, others are likely the structured outputs of commonly used systems. The extraction can be done in a variety of ways: Three techniques for data extraction:

Partial Extraction - If the primary system alerts you when any data has changed, that is the simplest way to retrieve the data.

With Update Notification of Partial Extraction - Not all systems can send out notifications when an update occurs, but they can still identify the entries that have changed and send out an extract of those records.

Full extract - Some systems are unable to determine which data has been altered at all. In this situation, the only way to obtain the data from the system is through a full extract. For this technique to work, you must have a duplicate of the previous extract in the same format so you can track down the modifications that were performed.

Transform

This stage entails converting the unformatted raw data that has been gleaned from a source into a form that can be accessed by various applications. In order to meet operational requirements, data is cleaned, mapped, and converted during this stage, frequently to a particular schema. This procedure involves many sorts of transformation to guarantee the accuracy and reliability of the data. Instead of loading data straight into the ultimate data source, data is usually placed into a staging database. This procedure guarantees a speedy rollback in the event that things does not proceed as expected. You have the option to create audit reports for legal compliance at this point, as well as identify and fix any data problems.

Load

Last but not least, the load function involves copying converted data from a staging region to a target database, which may or may not have existed before. The complexity of this process will vary depending on the requirements of the application. You can use ETL tools or custom code to complete each of these processes.

2. RELATED WORK

The idea of data pipelines is relatively new, and recent innovations in cloud architecture and cloud storage have advanced this particular field. These are the only new developments in the related area of data pipelining.

The following idea served as the foundation for a study on ETL technology that was carried out in 2009 [1]. Extraction-Transformation-Loading (ETL) forms are the earliest computer algorithms that promote initial stacking and sporadic warehouse refreshing. There were some limitations to this; information extraction is still a challenge, largely due to the closed nature of the sources; there are also challenges with streamlining and resume; and the absence of a baseline prevents further research.

Real-time ETL Data Warehousing was then researched in 2012 [2]. The goal was to achieve real-time data warehousing, which is heavily reliant on the selection of an extraction, transformation, and loading (ETL) method in data warehousing technologies (ETL).

In 2013, synchronous research [3] was being conducted in the field of ELT using an information distribution center's ability to directly input unprocessed, raw data while deferring information update and cleaning until needed by pending reports.

ETL was being accepted for a few applications later in 2016 [4, including the healthcare field]. While maintaining its integrity, this information must be appropriately deleted, modified, and packed into the warehouse. It gave the extract, transform, and load (ETL) procedure its seal of approval for correctness, populating the clinical research database as a result.

At the same time, Amazon's S3 [5] service may be useful because it offers bulk storage that doesn't require packing or cleaning. The Simple Storage Service (S3), a cheap capacity utility, had been introduced by Amazon.com. S3 intends to offer storage as a low-effort, widely available assistance with a simple "pay more only as costs emerge" payment approach.

With the introduction of Data lakes after one year in 2017, the adoption of Extract-Load-Transform grew more quickly [6]. The simplest assumption of an information lake is to mash up each piece of data provided by an organisation to produce increasingly important information at finer granularities.

2018 saw the incorporation of a defined methodology to create an R-based platform leveraging SQL. create a framework for R that influences SQL and is predictable and piping-able such that repeatable research on medium-sized data is a simple reality. Therefore, it had scaling issues based on data volume, and algorithms weren't instantaneous for medium data, which increased latency. Another implementation was made later that year to compile scientific data for analysis. To handle scientific data aggregation, transformation, and improvement for scientific data discovery and retrieval, a distributed extract-transform-load system that is horizontally scalable [8].

The improvement of privacy for ETL operations, particularly with biomedical data, was the subject of research in 2019 [9]. Data from many sources can be combined at clinical and translational distribution centres to create the requisite enormous datasets. This was accepted since anonymization was not supported by current ETL tools. Furthermore, at that moment, basic anonymization tools cannot be incorporated in ETL work processes.

Another work procedure related to the widely used On-Demand ETL system was being studied that same year. The Extract Transform Load process (ETL), which is the primary bottleneck in BI arrangements, is addressed creatively by DOD-ETL [10], an instrument that provides it in almost real-time. The main difficulty was to manage several information sources while also providing little latency for real-time responses.

Use in the banking industry was also being investigated later that year. Our new idea (RDD4OLAP) cubes consumed by Spark SQL or Spark Core fundamentals will replace the standard information combination and investigation process. It will do this by utilising Extract-Transform-Load (ETL) concepts, big data processing techniques, and oriented containers clustering architecture [11]. But also provide for very little delay so that you can react instantly.

3. EXISTING FAME WORK

To create a contemporary ETL system, open source frameworks like Apache Airflow might be employed. There are fantastic possibilities to contribute to the open source community that we pretty much rely on when the project is still in the development stage. As a result, we have chosen to release the project as open source under the Apache license.

Below are some of the procedures that Airflow powers:

Data warehousing: prepare, arrange, evaluate the quality of the data, and add information to our expanding data warehouse.

Calculate metrics for both host and visitor for engagement and growth accounting using growth analytics.

Experimentation: Calculate the logic and aggregates of our A/B testing experimentation framework.

Search: Calculate metrics relating to search ranking.

Email targeting: Applying rules to email targeting allows us to target and engage users.

Sessionization: generate datasets for clickstream and time spent

Data infrastructure maintenance: Application of data retention policies, folder cleanup, and database scraping are all examples of data infrastructure maintenance.

Airflow Principles:

- Scalable
- Dynamic
- Extensive
- Elegant

Airflow Features:

- Pure Python
- Useful UI
- Robust Integrations
- Open Source

Architecture

Python has solidified itself as the language of data, much the way English is being used for professional business. Python-like Python was used from the ground up to create Airflow. The code base has extensive unit test coverage, is expandable, well-documented, consistent, and limited.

Python is also used for pipeline creation, making it simple to generate dynamic pipelines from configuration files or other sources of metadata. We adhere to the idea of "configuration as code" for this. Although any language could be used to construct Airflow pipelines using yaml or Json task setup, we thought that some fluidity was lost in translation. It is quite valuable to be able to meta-program, subclass and use import libraries while writing pipelines in code (Python, IDEs). Remember that as long as you create Python that reads these configurations, you can still author jobs in any language or markup.

Airflow can be used for running in just a few commands, however the full architecture consists of the following elements:

A comprehensive CLI (command line interface) for testing, running, backfilling, describing, and clearing DAG components.

An online tool for exploring the definition, dependencies, status, metadata, and logs of your DAGs. The Flask Python web framework serves as the foundation for the web server, which comes packed with Airflow.

A metadata repository which the Airflow uses to maintain track of tasks and jobs statuses and other permanent data, often a MySQL or Postgres database.

A group of workers that distributes the execution of the task instances for the jobs.

The instances of the tasks that are prepared to run are launched by scheduler processes.

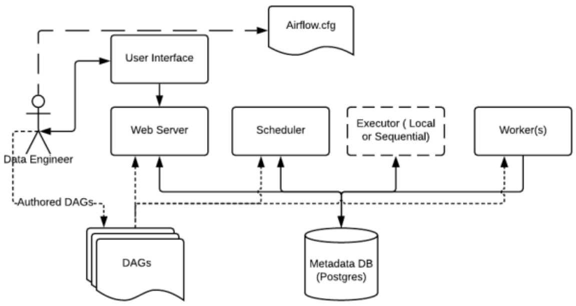


Fig 2: Airflow architecture

There are a few things to consider:
SQL database is used by Airflow to hold metadata about the data pipelines that are being used. This is shown as Postgres in the picture above, which is very popular with Airflow. MySQL is one of the alternative databases that Airflow supports.
- Web Server and Scheduler: The Scheduler and Airflow web server are independent programmes that communicate with the aforementioned database while running locally (in this scenario).
- The Executor is depicted separately above since it is frequently referenced in Airflow and in the documentation, although it actually runs inside the Scheduler and is not a separate process.
- The Worker(s) are independent processes that also communicate with the metadata repository and other elements of the Airflow architecture.
- Airflow.cfg is the configuration file for Airflow, and the Web server, Scheduler, and Workers may all access it.
- DAGs are the Python code-containing DAG files that represent the data pipelines that Airflow will perform. These files must be accessible by the Web Server, Scheduler, and Workers, and their location is specified in the Airflow configuration file.
A DAG defines your process in this manner, but keep in mind that we haven't specified what we actually want to do—A, B, and C could refer to anything. Perhaps A prepares the data that B will use to evaluate it while C emails. It's also possible that A keeps track of your whereabouts so that B can open your garage door and C can turn on your house lights. The DAG's role is to ensure that whatever its constituent activities accomplish occurs at the proper time, in the proper order, or with the proper handling of any unanticipated complications; it is not important what those jobs actually do.

4. PROPOSED FRAMEWORK

The framework which we are discussing in this paper is primarily built using Node JS. The framework is built in such a way that even a person with least programming experience can build an ETL pipeline. Most of the logic which has to be implemented should be done using SQL.

ETL Stack

It is an ETL (Extract/Transform/Load) Stack written in NodeJS. Extract useful data out of raw data, Transform to usable metrics (aggregations) and Load to Enterprise data lake.

ETL Job

An ETL Job (configured as JSON file) is a set of interdependent tasks which run as a single unit of work. It is a logical unit of work - Hourly Viewership metrics, Daily Ad-Analytics metrics.

ETL Task

A Task is a single piece of independent work unit - Compute hourly sessions from Beacon data for example. It can depend on other task(s) to run.

Big data computation

Most tasks work with a source as Data lake, compute on data from lake and put back computed data into data lake. Some cases, they put the data/metrics back to the end-user reporting system.

Tools Used

- Athena

Athena is a partition supported realtime big data crunching system using Facebook's PrestoDB underneath. You can write a SQL query which runs on the data on S3. It can extract, filter, aggregate, group data to create metrics.

- Data lake(S3)

S3 is big data storage system to store objects, logs, records in formats like JSON, Parquet, CSV, Regex parsable text records.

- Postgres

RDBMS database used to store the end-user facing reporting metrics with right indices to fetch data faster. Analytics portal and APIs can use this database to provide reports and visualizations to customers

The framework allows us to create a pipeline which is often referred to as a job. This job will have many interdependent tasks. The tasks are the individual work items which carry out a specific function. All these tasks are joined and interlinked to create a pipeline.

All the jobs in the framework are automated using cron schedules so that without any human intervention all the jobs are running at prescribed time. From extraction of information till loading the useful information into the database is automated. And the data in the database is used to build dashboards, send reports, monitoring and alerting etc.

The picture below Figure 3 shows the basic architecture of the whole process of how the ETL pipeline is working.

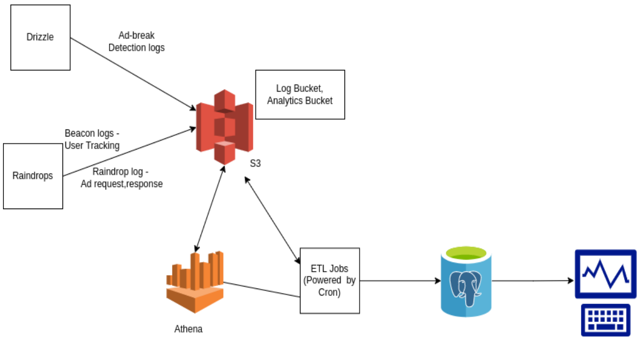

Drizzle
Ad-break Detection logs
Log Bucket, Analytics Bucket
Beacon logs - User Tracking
Raindrops
Raindrop log - Ad request,response
S3
ETL Jobs (Powered by Cron)
Athena

Fig 3: Basic Architecture

The architecture displays how the data is extracted from different data sources or deployments. All this data is stored in Amazon S3, which is the data warehouse. The data which is needed for the pipeline is extracted from the data warehouse and specific transformations are applied to the raw data using Amazon Athena. The transformed data is then loaded into the database. From the database the data is queried and displayed into the dashboard.

5. METHODOLOGY

The following discussion touches upon the methodology of how the pipeline is implemented from the first step of extraction to the last step of displaying the data.

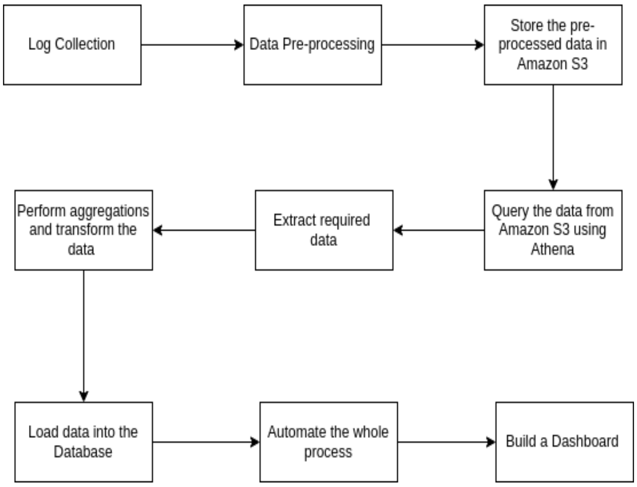

Log Collection
Data Pre-processing
Store the pre-processed data in Amazon S3
Query the data from Amazon S3 using Athena
Extract required data
Perform aggregations and transform the data
Load data into the Database
Automate the whole process
Build a Dashboard

Fig 4: Flow Chart of the methodology

Figure 4 represents the flow of events happening while building a pipeline:

Log Collection

The log data or the raw data is collected from different data sources or deployments and stored in the Data Warehouse.

Data Pre-processing

The raw data or the logs which are collected can be any format, usually the logs will be in a Json format. The data is processed and converted to csv format with the useful data in it. The pre-processed data is again stored in Amazon S3.

Extraction

Using Amazon Athena, few useful pre-processed data is queried and used in the pipeline for getting some insights upon that data.

Perform Aggregation and transformation

The raw data is aggregated and then transformed to get desired output using the query engine which is the Amazon Athena.

Loading

The final transformed data is loaded into the database, and stored over there.

Automation

The whole process is automated using cron schedulers. The process is scheduled at a particular time, depending on the frequency of which the process should be run. Accordingly the process is scheduled and automated.

Dashboard

A dashboard is built upon the data in the postgres to provide business insights to the customers. The dashboard can also be used for monitoring and alerting the development team if there are any data discrepancies.

6. RESULTS AND DISCUSSION

In a perfect world, data analysts have access to all the information they need and don't have to worry about how or where it is stored since analytics just function.

The reality of analytics has been far more convoluted up until recently. Building and maintaining flimsy ETL (Extract, Transform, Load) pipelines that pre-aggregated and filtered data down to a consumable level was required in order to access data because expensive data storage and underpowered data warehouses made this impossible. ETL software providers compete based on how specialised and adaptable their data pipelines were.

We are now getting closer to the analysts' ideal thanks to technology. Fragile ETL pipelines are a thing of the past thanks to practically free cloud data storage and significantly more powerful contemporary columnar cloud data warehouses. Extract and load the raw data into the destination, then transform it after the load, is the modern data architecture. Numerous advantages result from this distinction, including improved adaptability and usability.

7. CHALLENGES

The challenges involved on this framework are:

- Retries for Failures: NO Retries for failed jobs. Manual retries should be done for failed jobs
- No Cross-job Dependency: A task can depend on other task(s) in the same job, but can’t depend on a task of another job.
- Non-parsable Logs: Logs collected from ETL Jobs are not parsable. So debugging requires manual analysis of Logs.
- Too many Slack notifications: Slack notifications are too many - causing more ignorance for failed jobs.

8. CONCLUSION

Technology trends are aware that processing, storage, and bandwidth are now accessible to anyone. The price of computation has decreased over time due to technological advancements. Similarly, the price of a gigabyte has dropped from around $1 million to a few cents in a period of about 35 years. Data warehouses can now hold significantly higher data volume as a result of these drastic cost reductions. It is no longer necessary for organisations to pre-aggregate and, in the process, delete a significant amount of source data. This makes it possible for analysts to do analysis that is both deeper and more thorough than previously. Despite the World Wide Web not existing until 1991, internet transport costs have drastically fallen. It fell from approximately $1,200 per Mpbs to a few cents in less than twenty years. The cloud, or the utilisation of remote, decentralised, web-enabled computer resources, is the result of the convergence of these three cost-reduction developments. A wide variety of cloud-native applications and services have also emerged as a result of cloud technology.

Many firms adopt a manual, ad hoc approach to data integration; in fact, 62 percent utilise spreadsheets like Google Sheets and Excel to combine data from several files and visualise the information. 2 In order to do this, files must be downloaded, values must be manually changed or cleaned, intermediate files must be created, and other similar operations.

Ad hoc data integration has a number of disadvantages, to mention a few:

- only suitable for very modest data amounts
- Slow
- Human error prone
- Not safe enough for critical information
- Often not reproducible

Maintaining the boundaries between distinct data sources' silos while filling in the gaps with "federated" queries, which directly query various source systems and integrate data in real time, is a more long-term solution. Organizations can use SQL query engines like Presto to accomplish this. This federated approach's drawback is that it has a lot of moving pieces and performs poorly with big amounts of data. The truth is that a methodical, repeatable strategy to data integration—a data stack—is necessary for a scalable, sustainable approach to analytics.

9. FUTURE WORK

The framework can be improved in many ways. A few updates which are thought to be added are :

- Run in Pods: Dockerize, build for Kubernetes and Run Jobs and Tasks in Kubernetes. Different ETL tasks require different sized/resourced Pods.
- Parsable Logs: Collect Logs from ETL Jobs and parse them to understand what is failing and reason for failing
- Automated Retries: Retry a failing job for n number of times (already exists for few type of tasks) at job level and various time periods (retry a job after 1 hour if daily job failed)
- ETL Dashboard: ETL Dashboard to show the progress of the job execution, what is failing, one click retries etc.
- Non-NodeJs Tasks: Few tasks are better to run in other languages - Shell script, Python etc. Wrapper for such tasks to embed into NodeJS Job/Task Framework
- Monitoring Support: Monitor Analytics Pipeline also as a first grade component - Raise alerts for failures, handle/recover from failures etc.
- Cost and Stats: Costs of Athena and other operations at fine grain level, building stats dashboard etc.

10. ACKNOWLEDGEMENT

Prof. Suma B, Mr. Venkatesha and Mr. Sai Charan deserve special thanks for their insightful remarks and ideas as well as for giving the authors the chance to examine the development process, artefacts, and records.

11. REFERENCES

[1] Panos Vassiliadis, “A Survey of Extract-Transform-Load Technology.,” International Journal of Data Warehousing and Mining, July 2009

[2] Kamal Kakish, Theresa A Kraft, “ETL Evolution for Real-Time Data Warehousing”, presented at Conference: Proceedings of the Conference on Information Systems Applied Research, At New Orleans Louisiana, USA,2012

[3] Florian Waa, Tobias Freudenreich, Robert Wrembel, Maik Thiele, Christian Koncilia, Pedro Furtado, “On-Demand ELT Architecture for Right-Time BI: Extending the Vision”, International Journal of Data Warehousing and Mining 9(2):21-38 · April 2013

[4] Michael J. Denney, MA,1 Dustin M. Long, PhD,2 Matthew G. Armistead, BS,1 Jamie L. Anderson, RHIT, CHTS-IM,3 and Baqiyyah N. Conway, PhD4, “Validating the Extract, Transform, Load Process Used to Populate a Large Clinical Research Database,” Int. J. Med. Inform., 94, 2016

[5] Valerio Persico, Antonio Montieri, Antonio Pescapè, “On the Network Performance of Amazon S3 Cloud-Storage Service”,5th IEEE International Conference on Cloud Networking (Cloudnet), 2016

[6] Pwint Phyu Khine, Zhao Shun Wang, “Data Lake: A New Ideology in Big Data Era”, 4 th International Conference on Wireless Communication and Sensor Network [WCSN2017], At Wuhan, China, 2017

[7] Benjamin S. Baumer, “A Grammar for Reproducible and Painless Extract-Transform-Load Operations on Medium Data”, arXiv:1708.07073v3 [stat.CO], 23 May 2018

[8] Ibrahim Burak Ozyurt and Jeffrey S Grethe, “Foundry: a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement”, Database (Oxford). 2018.

[9] FabianPrasser, HelmutSpengler, RaffaelBild, JohannaEicher, Klaus A.Kuhn, “Privacy-enhancing ETL-processes for biomedical data”, International Journal of Medical Informatics, Volume 126, June 2019, Pages 72-81

[10] Gustavo V. Machado, Ítalo Cunha, Adriano C. M. Pereira, Leonardo B. Oliveira , “DOD-ETL: distributed on-demand ETL for near real-time business intelligence “, Journal of Internet Services and Applications volume 10, Article number: 21, 2019.

[11] Noussair Fikri, Mohamed Rida, Noureddine Abghour, Khalid Moussaid & Amina El Omri, “An adaptive and real-time based architecture for financial data integration”, Journal of Big Data volume 6, Article number: 97, 2019.

[12] Aiswarya Raj,Jan Bosch,Tian J. Wang,Helena Holmström Olsson, “Modelling Data Pipelines”, at 46th Euromicro Conference on Software Engineering and Advanced Applications (IEEE), 2020.

[13] Marko Jamedžija, Zoran Đurić, “Moonlight: A Push-based API for Tracking Data Lineage in Modern ETL processes”, at 20th International Symposium INFOTEH-JAHORINA(IEEE), 2021.

[14] Noussair Fikri , Mohamed Rida, Noureddine Abghour, Khalid Moussaid, Amina El Omri, “An adaptive and real-time based architecture for financial data integration”, in Springer open, journal of big data, 2019.

[15] Valerio Persico,Antonio Montieri, Antonio Pescapè, “On the Network Performance of Amazon S3 Cloud-storage Service”, at 5th IEEE International Conference on Cloud Networking, 2016.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 320-322, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1005

Online Vehicle Rental System

Aniruddha S
B.E in Computer Science Eng
R V College of Engineering
Bangalore

Shivaraj B Karagera
B.E in Computer Science Eng
R V College of Engineering
Bangalore

Manas M N
Assistant Professor
R V College of Engineering
Bangalore

Abstract: The main objective of the paper is to discuss a customer-centric vehicle rental process with state-of-the-art technology. By digitalizing the manual process of document verification, selecting a vehicle and also the payment process the end user will have a smooth experience of renting a vehicle.Digitalizing the service will give the customer a choice of booking from anywhere and any time.The system is very convenient in case where the customer is new to a city or a place and when customer prefers to checkout himself rather than using agents to avail the benefit of booking and renting a vehicle.The system will also be providing various features like additional driver,protections etc for the customer and thus enriching the rental experience.

Keywords: kafka; OTP; microservice; java; spring boot

1. INTRODUCTION

We certainly consider online shopping,online banking etc.Similarly,the project aims in digitalizing the rental system thus adding to the digitalization of world.Currently there are various vehicle rental giants who offer online services of vehicle booking and renting,but those either involve manual interaction with agents or consumes a lot of time.Even though there are public services available in almost all places the joy of having own vehicle will give high degree of freedom to move.Considering users who are new to places, its difficult to cope up with public transportation and timings associated with it.This would greatly restrict the freedom of user.Providing users with a service which can help them overcome these various challenges is need of an hour.The proposed system is designed to address these issues faced by users and give them a rich experience.Here user need not own a vehicle instead user can rent a vehicle on demand by booking it. Presently all the steps are manual,hence time consuming and it's really difficult to track each vehicle. Logging and searching is difficult in manual records.

The proposed system digitalizes the manual process of document verification,vehicle selection ,payment and renting a vehicle thus improving existing systems and significant decrease in the time consumed.The system is different from existing ones,here once the booking is done the customer need not wait in a queue for his turn to verify the booking by agent before handing over the vehicle.The user can verify all the details online using proposed system and directly go to the key booth,give the OTP and start renting.This even has major impact on tourism industry.Tourists need not worry about transportation in new places,they can rent the vehicle wherever they want.This eliminates a major constraints on tourism

2. LITERATURE REVIEW

The proposed model depends on the Vehicle Renting System, which is a service we inspected the present working circumstance of the renting technique. Considering the vehicle rental industry and proposing a self drive online car renting system,the procedure of booking a vehicle for the rental reason is manually done[1]. At current renting, users are dependent on a manual system which provides deals to them as a human resource. Nowadays we find Cab Services very easy to book, pay, or drop as they have formed their structures into helpful applications similarly as locales. So there is a need to change the arrangement of the vehicle Renting Service. But the vehicle rental system still works on older systems ,which includes manual interaction, waiting in queue, and waiting for confirmation,allotment of vehicle at a particular location etc. Creating a system where customers can book their automobile for rental and request services across the world[3]. Our system reflects on these problems and brings out a solution of self-driving cars where the whole system would be digitized.

3. METHODOLOGY

The Microservice architecture is being followed where instead of the system being monolithic,we decompose it to independent pieces called services and those are being called on demand, it also is highly maintainable,easily testable,loosely coupled,independently deployable and can be owned by small teams thus making it much easier for debugging and feature changes if any that might arise in future, thus the whole system is being designed.

The project is built using Java and Spring Boot. Considering the features of spring-boot for the microservice architecture it's been used in the development process,Spring-Boot is of great help to create a new service as it gives us features like auto code generation , third party libraries and minimal configuration.It also helps to create a stand alone application.

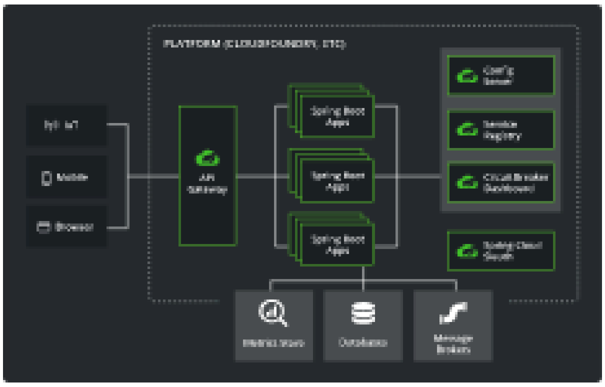


fig.1 Spring-boot architecture[6]

The microservices in the proposed system are booking management service, checkout management service, customer management service and configuration management service.

3.1 Booking

This service will take care of the user making a booking for a vehicle,this phase would be usually done in advance, this includes creating a user trip details so that when the user comes for renting. Booking details will have beginning and ending details of the trip, preference for vehicle, extra protections or features needed that are essential for the trip , and the payment for the trip which will be paid as an advance for booking.

3.2 Checkout

This service is the continuation of the booking service, where he will enter the flow when he would be starting his rental journey, Here users would still be given a freedom to modify some of the booking details. The actual vehicle selection would be happening here, because in the booking, only a group of vehicle identifiers (acriss code) will be considered ,now depending on the identifier and the vehicle availability at that point of time and location, the vehicles would be shown to the user,who will select his choice.The final payment needs to be done here along with refundable cash deposit for security purposes . Once he completes the flow , he can collect the vehicle from the garage , where he would need to show the one time password (OTP) to the agent to collect his vehicle.

3.3 Configuration Management service

As we are considering the microservice architecture and there are many things that will be changing over time , the features would be set configurable. To maintain the configuration the configuration management service is used , which would give all the necessary configs needed,like details of a car,details of a branch etc.

4. CUSTOMER MANAGEMENT SERVICE

This service would collect all the booking details for a user and be used for future trips. Once the user is logged in all the details would be automatically synced and help him rent a vehicle faster and easier. It would also act as a customer care when the user rents a car. Any help in case of emergency needed during the rental will be taken care of by the service.

5. KAFKA

It is used to integrate all the above mentioned services because of its ease of solving the problem of scaling and reliability issues.Its a publish-subscribe model where events are sent by various services when they are available and listened by other services asynchronously.

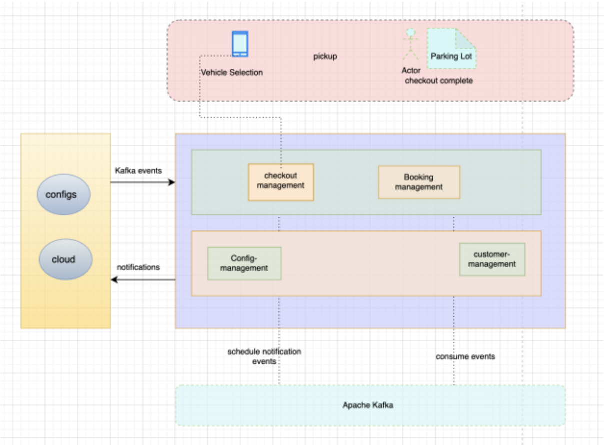


fig.2 System architecture of proposed model

The above diagram describes the work done in a detailed manner.Various services used are depicted. Users will go through the booking and checkout management system which are using other services. The services communicate with each other using kafka events

6. CONCLUSION

The developed system was able to achieve most of the objectives that were considered to be solved, the user is having a smooth flow and there is no middle-man involved which eradicates the communication barrier in different geographical locations.It's time saving since all steps starting from documentation and ending on getting vehicle keys can be done online and gives the freedom of booking/renting a vehicle anytime and anywhere in the world.The system is a self-service model hence customer is the main actor which gives him all power in rental-experience.In coming days this system can also be integrated with new services like delivery and collection service,which would deliver the vehicle to user to preferred location and pickup the same,thus giving user a good rental experience.

7. REFERENCES

[1] Suryadev Singh Rathore, Mahik Chaudhary, 'Analysis of Self Drive Rental Cars Industry in India'. IJMBS Vol. 8, Issue 4, Oct - Dec 2018.

[2] Prof. B.A. Jadhawar. and Komal A. Bhosale, "Research Paper on Java Interactional Development Environment Programming Tool".

[3] Amey Thakur Department of Computer Engineering. University of Mumbai, Mumbai, Maharashtra, India. Car Rental System.International Journal for Research in Applied.Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 9 Issue VII July 2021

[4] Online Car Rental System by Rahul Kulkarni, Chaitanya R, Pratibha, Pooja A Pati, Nikeeta Biradar August 2021|IJIRT | Volume 8 Issue 3 | ISSN: 2349-6002.

[5] Conor Muldoon, Levent Gorg u, John J. O'Sullivan, Wim G. Meijer, and Gregory M. P. O'Hare "Engineering testable and maintainable software with Spring Boot and React".

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullende

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 323-336, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1006

Technostress and Its Determinants: A Psycho–Physiological Complications Among Workforce In Nigeria

Abasiama G. Akpan
Department of Computer Science and Mathematics
Evangel University, Akaeze, Nigeria

Victoria N. Ezeano
Department of Computer Science and Mathematics
Evangel University, Akaeze, Nigeria

Udeme Offiong
Department of Psychology
Chukwuemeka Odumegu University, Igbariam
Nigeria

Abstract
The goal of Information and Communication Technology (ICT) was to make our lives easier (i.e., by providing faster communications around the globe, efficacy in work processes, and so on). If modern technology was designed to empower us, to set us free, and to leave us satisfied, why do we often feel (techno-) stressed due to the use of this technology? This paper unravels a modern disease caused by inability to cope with computer technology in a healthy manner called Technostress. Technostress scale, a twenty – two item Likert-type, two sub-scale questionnaires designed for the study was administered on two hundred and one samples drawn from five faculties in National Open University of Nigeria (NOUN) and five Commercial banks in Port Harcourt metropolis, Nigeria. The data were analyzed using t-test, Correlation and ANOVA statistics. The results revealed that academic staff manifested higher levels of technostress than the employees from the banking sector, a positive correlation was observed between computer hassles and stress reaction. In conclusion, ICT training and stress management were highlighted as solutions for technostress in the two human industries.

Keywords: Techno-overload, Techno-complexity, Techno-insecurity, Techno-uncertainty, Technostress

1.0 Introduction

The world continues to be an information-driven arena where efficiency and competition are measured by the fast pace of information accessibility via Information and Communication Technology (ICT). With the innovations of new features in technologies and its capabilities to provide various services and transactions, many organisations now compete for efficiency and improved productivity. Thus, the rapid advances and changes in new technology have caused institutions and organisations to continuously introduce employees to updated technology and software packages to stay technologically current, and many a times abreast in their area of focus. However, the rapid introduction of technology in the workplace may cause individuals in organizations to suffer from a combination of technology fatigue and aversion. Thus, the presence of Information Communication Technology contributing to workload efficiency, effectiveness and good performance has created a new phenomenon called technostress; which is defined as a modern maladaptation resulting from the failure to cope with ICT and changing requirements related to the use of ICT [1, 2]. The main aim of this study is, therefore, to investigate the impact which the incidence of technostress has among education administrators and those in the banking sector where computer technology has been deployed as the main tool for their daily work. This is with the belief that the population under study could be more vulnerable to information overload and fatigue caused by IT.

1.1 *Technostress and its Components*

Brod [1] defined technostress as a modern disease of adaptation caused by inability to cope with the new computer technologies in a healthy manner. According to him, it manifests itself in two distinct but related ways; that is in the struggle to accept computer technology, and in the more specialised form of over-identification with computer technology. He identified the symptoms of technostress to include irritability, headaches, nightmares, resistance to learning about the computer or the outright rejection of the technology while Tarafdar, Tu, Ragu-Nathan and Ragu-Nathan [3] identified five components of technostress to be:

- ***Techno-overload*****:** A situation where ICT users are forced to work fasterand longer
- ***Techno-invasion****:* A situation where ICT users feel that they can be reached anytime or constantly "connected" which caused a blurring between work-related and personal contexts

- ***Techno-complexity:*** A situation where ICT users feel that their skills are inadequate due to the complexity related to ICT. Consequently, they are forced to spend time and effort to learn and understand the various aspects of ICT
- ***Techno-insecurity*****:** A situation where ICT users feel threatened that they who are better in ICT compared to them
- ***Techno-uncertainty:*** A situation where ICT users feel uncertain and unsettled since ICT is continuously changing and need upgrading.

Kupersmith [4] aver that technostress has only one form of stress which interacts with other forms of stress. He pointed out five related but distinct components of technostress:

- Performance anxiety, which refers to the tendency of an individual to engage in negative thoughts and statements.
- Information overload, which is tension as a result of too much information which exceeds a person's apprehension capability.
- Role conflict which describe the friction between different functions.
- Self-definitions
- Organisational factors such as colleagues, facilities, policies, culture and management.

Tams *et al*. [5] demonstrated the strengths of a multi – method approach in technostress research. In the author’s experiment, participants performed a computer – based task (a memory game) while instant messages frequently interrupted them. Messaging the resulting stress on a psychological level (using self- report measures) and a physiological level (using measures of stress hormone excretion), the authors explained a higher degree of the variance in task performance than with each method alone.

With the advent of ICT in virtually every work setting, workers are all saddled with one form of psycho-physiological complications or another due to constant adjustment to different facets of ICT. It is thus worrisome that such complications are most times overlooked and de-emphasised. The contemporary worker is gradually going blind because of poor computer screen; he or she is continually saddled with overwhelming helplessness associated with massive loss of data due to computer break down. The physiological effect of resultant stress reaction is gradually reducing the quality of life of an average worker today who must depend on ICT for daily result. Since it is almost impossible to shy away from the ICT revolution and its gains, it is very important to recognise the anxiety associated with technostress in order to help the individual involved adjust well to the challenges. It is on this premise that this study is investigating the influence of technostress on the workforce of both the educational and the banking sectors in Nigeria.

1.2 *Objectives of the Study*

The objectives of this study include:

1. To determine the levels of technostress between faculty members and bank officials.
2. To determine gender differences in the manifestation of technostress.
3. To find out the influence of age in manifestations of technostress among participants.
4. To determine the relationship between computer hassles and stress manifestations.

1.3 *Research Hypotheses*

The hypotheses tested in the study are:

1. There is a statistically significant difference between faculty members and bank officials in manifestations of technostress.
2. There is a statistically significant that males will exhibit significantly higher level of technostress than the females.
3. There is a statistically significant the older participants will exhibit significantly higher level of technostress than younger participants.
4. There is a significant and positive correlation between computer hassles and stress manifestations.

2.0 *Literature Review*

Providing insight into the physiology of stress, Arnetz and Berg [6] through their research observed that individuals are most likely to experience higher levels of adrenaline and nor adrenaline during work periods with computers. Adrenaline and nor adrenaline are catecholamines secreted by the adrenal gland. The increased excretion rates of adrenaline and noradrenaline are associated with both under load and overload (stress) stimulation and emotional arousal. Other effects of the increased catecholamine levels, as part of sympathetic nervous responses, are increased heart rate and blood pressure. Increased heart rate and blood pressure have been observed in persons performing a computer task [7]. Other research has shown that there is increased skin conductance level (SCL) while performing a computer task [7]. Skin conductance level is an indicator of increased sympathetic nervous reaction (the more you sweat the better the conductance). Another indirect indicator of being "stressed" by computer use, is an increased jaw muscle electromyography (like clenching your teeth, an index of the user's 'anger') while performing a computer task [8]. In a related study, Charlesworth and Nathan [9], remarked that up to 75% of all visits to physicians are the result of stress-related disorders. Their study concluded that hypertension; coronary heart disease, headaches, asthma, gastrointestinal disorders, and many skin disorders are all related to stress. Same token, same analogy, many factors have been identified to be the causes of technostress. Clute [10] cited the top three reasons that cause technostress as, inexperience with computers, performance anxiety and lack of training/insufficient staffing. As a matter of fact, many other studies also claimed lack of training as one of the main reasons for technostress [11]. Common organisational factors found by Clute [10] to be the sources of technostress include lack of participatory management styles, lack of communication, and lack of involvement. In fact, most studies revealed that those who suffered from technostress were mostly angry because they were forced to accept the technology without being consulted before the implementation of the technology [12] A survey by Masey and Stedman [13] showed that the increase in demands for technology was among the main attributing factors adding to job stress. They pointed out three ways on how stress is inherent in technology as through (a) client expectation; (b) aggressive marketing schemes from software manufacturers; and (c) desire to always be on the cutting edge of technology. Massey and Stedman [13] further conducted a survey of information-technology users concerning feelings about stress in their work environments. The researchers stated that 86% of surveyed workers indicated that their jobs were more stressful now than they were five years ago. They attributed the added stress to being understaffed and having additional responsibilities. They concluded that the nature of information technology "demands a high degree of meticulousness and attracts the type of individuals who are already prone to stress' like Type A personality individuals [14].Ragu-Nathan *et al*. [15] agreed that it is the characteristics of information communication technologies that are creating stress in technology end-users.

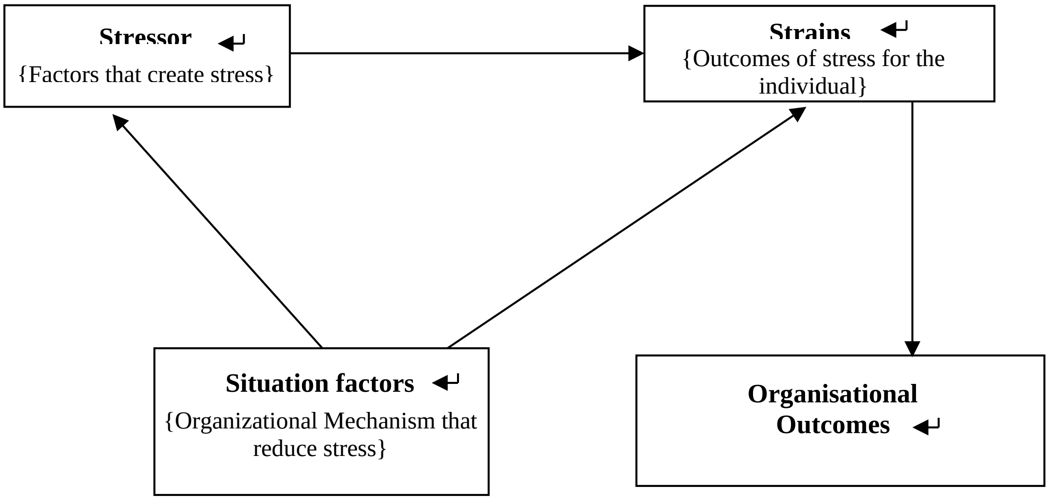

Stressor ↵
{Factors that create stress}
Strains ↵
{Outcomes of stress for the individual}
Situation factors ↵
{Organizational Mechanism that reduce stress}
Organisational Outcomes ↵

Figure 1: Transactional – based Model of Stress [15]

Figure 1 show the stress construct, which was utilized based on organizational behaviour literature.

This provides the theoretical background for understanding stress in the organizational environment.

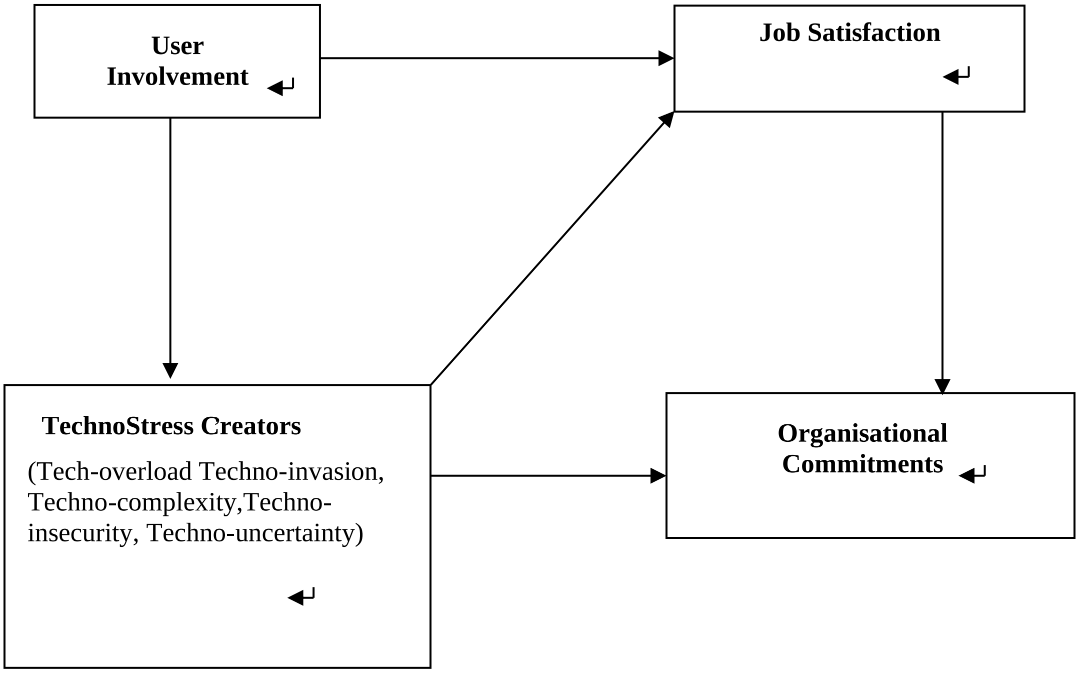

User Involvement
Job Satisfaction
TechnoStress Creators
(Tech-overload Techno-invasion, Techno-complexity,Techno-insecurity, Techno-uncertainty)
Organisational Commitments

Figure 2: Conceptual Model of Technostress [15]

The general model of stress was further developed to link the theoretical concept of organizational stress of ICT usage in the organization; and explain how the use of ICTs can potentially create stress and this negatively impact individual job satisfaction and organizational commitment [15]. Studies reviewed traced to psycho-physiological impact of stress and more specifically, the influence of technostress of job satisfaction and emotional stability. The present study therefore continues this interesting and insightful debate.

3.0 *Methodology*

The research design employed for this study is survey design and specifically, the statistical method used for the analysis is a 2X2X5 ANOVA design. The independent variables are gender (male, female), place of work (university/bank) and age (20 - 29; 30 - 39; 40 - 49; 50 - 59; 60 - 69), while the dependent variables are participants' responses on the technostress scale. The independent variables were tested with mean, SD, t-test, One Way ANOVA and Correlation analysis. The results obtained were linked to the specified hypotheses.

3.1 *Population and Sample*

The population for the study was made up of National Open University of Nigeria (NOUN) where academic seminars and Inaugural lectures delivery system is majorly IT-based and staff of five banks whose services are driven with high level technology within Port Harcourt metropolis. The banks are Zenith bank, First bank, Union Bank, Unity Bank and GT bank. The actual samples for the study are 102 randomly selected respondents among the academics from NOUN and 99 randomly selected respondents among the bank staff from the five purposefully selected banks mentioned above.

3.2 *Data Collection Procedure*

Data for the study were collected using questionnaire. A total of 250 questionnaires were administered on the respondents who were randomly selected among the academics in the five faculties in NOUN and the randomly selected staff of the five chosen commercial banks. Five research assistants; one in each bank was employed for this study and they were given a brief training on the relevant interpersonal skills needed for the study. They were also provided with relevant materials (letter of introduction, writing materials) for the study. Every copy of the questionnaire returned was scrutinised by the researchers. The respondents were expected to respond to all the statements in the questionnaire, in addition to their demographic profile. The statements were 20 in number and were presented on a Likert-type scale. Out of the 250 copies of questionnaire distributed, a total of 201 were found usable. This made up 80.4% of the questionnaire. This figure was thus found enough for the research to proceed with the data analysis.

3.3 *Study Instrument*

A questionnaire was designed for this study. Contents of the questionnaire were adapted from Hudiburg, Computer Hassle Subscale (CHS) and Omoluabi Psycho-physiological Symptom Checklist (PSC). Experts in social and medical sciences helped to validate the contents of the questionnaire. Specifically, the research instrument contained 10 items that assessed computer hassles and 12 items assessing stress reactions. Items for computer hassles include level of IT competence, state of computer screen, incidences of virus attack and loss of data, slow booting and power fluctuation. Others include incessant keyboard error, proficiency in typing, level of competence with computer packages and the ability to use adopted software. The subscale that assesses stress reaction tries to identify the following stress manifestations: backache, headache, pain in the eyes, irritability, poor sleep, dizziness, emotional outburst, poor bowel movement, interpersonal difficulties. The questionnaire was pilot-tested on 20 participants drawn from commercial banks and two study centres of NOUN using test-retest methodology spanning two weeks interval to establish the reliability coefficient of the questionnaire. The reliability coefficient of the questionnaire is 0.71.

Data Analysis: The study employed mean, standard deviation, t-test and One-Way ANOVA for the data analysis. This was done with the aid of statistical package for social science (SPSS), version 15.0.

Results

	N	(%)
Department		
Academic	102	50.75
Financial	99	49.25
Total	**201**	**100**
Gender		
Male	107	53.23
Female	94	46.77
Total	**201**	**100**
Age (Years)		
20-29	55	27.36
30-39	57	28.36
40-49	56	27.86
50-59	27	13,43
60-69	6	2.99
Total	**201**	**100**

Table I: Demographic Profile of the Respondents (N = 201)

The demographic profiles provide information on employees' background, gender and age. A total of 201 participants took part in this study. Out of the total number, 50.75% of them were academic staff while 49.25% were selected from financial institutions. In terms of gender, 53.23% were males while females were 46.77%. An analysis of the age of the participants indicates that 27.36% were between the age group of 20 to 29 years, 28.36%, while those 30 to 39 years were 28.36% of the respondents. In addition, 27.86% were between 40 to 49 years, 13.43% were between 50 to 59 years, while 2.99% of the respondents were in the age range of 50 to 59 years. In order to examine the differences in the symptoms and manifestations of tehnostress among faculty members of National Open University of Nigeria (NOUN) and participants drawn from the banking sector, the mean, standard deviation and t-test of their responses were computed. The result is presented in Table 2.

Table 2: Mean, SD and T-test of Respondents on Technostress Scale

Department		Computer Hassle	Stress Reaction	T	Df

Academic Staff	Mean	29.8627	21.7353		
	N	102	102		
	Std. Deviation	5.36774	4.20025	8.96*	199
Staff of Financial Bank	Mean	23.3333	19.7980		
	N	99	99		
	Std. Deviation	4.96107	3.51667	3.54*	199
	N	201	201		
	Std. Deviation	6.10898	3.98896		
Total	Mean	26.6468	20.7811		

Note: * = Significance, Probability level (P) = <.05, df= 199, Critical t= 1.65

The result in Table 2 shows that academic staff obtained mean scores of 29.86 and 21.74 on computer hassle and stress reaction sub-scales respectively, while employees from the banking sector obtained mean scores of 23.33 and 19.80 on the scales respectively. For a test of significance, the t-test statistic was computed. The result indicates score of 8.96 and 3, 54 respectively for the two groups of respondents with a df of 199. Result further indicates that the observed differences in obtained scores are statistically significant at probability level = <.05, df - 199, Critical t = 1.65. Thus, hypothesis one which states that there will be statistically significant differences between faculty members of open and distance learning and bank officials on manifestations of technostress are hereby confirmed.

To ascertain the influence of gender on manifestations of technostress. The mean, SD and T-test scores of the respondents are analysed in Table 3

Table 3: Mean, SD and T-test of Male and Female Participants on Technostress Scale

Gender		**Computer Hassle**	**Stress reaction**	**t**	**df**
Male	Mean	26.5607	20.4766		
	N	107	107		
	Std. Deviation	5.87636	4.17166	-.211	199
Female	Mean	26.7447	21.1277		
	N	94	94		
	Std. Deviation	6.39377	3.76239	-1.63	199
	N	201	201		
	Std. Deviation	6.10898	3.98896		
Total	Mean	26.6468	20.7811		

Note * = Significance, Probability level (P) = < .05, df = 199, Critical t = 1.65

The result in Table 3 shows that the male respondents obtained mean scores of 26.57 and 20.47 on computer hassle and stress reaction sub-scales, respectively, while females obtained mean scores of 26.74 and 21.13 on the scales respectively. The t-test result indicates score of-.211 and -

1.63 respectively for the two groups of respondents with a df. of 199. Result further indicates that the observed differences in obtained scores are not statistically significant at P. = .05, df = 199, critical t = 1.65. Hypothesis two that states that the males will exhibit significantly higher level of technostress than the females are hereby rejected. To ascertain the influence of age on manifestations of technostress, the mean, SD and ANOVA scores of the respondents are presented in Tables 4 and 5.

Table 4: Mean and SD Scores of Age Influence on Technostress

Age Group		Computer Hassles	Stress Reaction
20-29	Mean	26.7455	19.3818
	N	55	55
	Std. Deviation	5.18266	3.93696
30-39	Mean	25.7193	20.7544
	N	57	57
	Std. Deviation	5.79333	3.66587
40-49	Mean	27.4286	21.2857
	N	56	56
	Std. Deviation	5.81177	4.53958
50-59	Mean	25.4444	21.5556
	N	27	27
	Std. Deviation	7.71279	3.51188
60-69	Mean	35.8333	26.5000
	N	6	6
	Std. Deviation	4.66548	4.18330
Total	Mean	26.7413	20,8060
	N	201	201
	Std. Deviation	6.11169	4.16259

A summary of the result in Table 4 indicates that the highest mean score was obtained among those on the age range of 60 to 69 years on the two sub-scales while the lowest mean score was observed among those aged 20 to 29 years. This result will be further analysed on the discussion section.To further ascertain if the observed differences in Table 4 are statistically significant, the ANOVA statistics was computed and presented in Table 5.

Table 5: ANOVA Summary Table

		Sum of Squares	Df	Mean Square	F
Computer	Between Groups	627.388	4	156.847	4.492*
	Within Groups	6843.159	196	34.914	
	Total	7470.547	200		
Stress	Between Groups	334.294	4	83.574	5.231*
	Within Groups	3131.138	196	15.975	
	Total	3465.433	200		

Note: * Significant, P < .05, df, 4/196, Critical f = 3.03

Result obtained in Table 5 indicates that the statistical observations were significant at P < .05, df, 4/196, critical f = 3.03. Thus, hypothesis 3 that states that older participants will exhibit significantly higher level of technostress than younger participants is confirmed. To further ascertain the relationship between the sub-measures of technostress: computer hassles and stress manifestations, the correlation statistics is presented in Table 6.

Table 6: Correlation Matrix of the Measure

		VAROOOOI	VAR 00002
VAR00001	Pearson	I	.422**
	Sig.		.000
	N	201	201
VAR00002	Pearson	422**	1
	Sig.	.000	
	N	201	201

** Correlation is significant at the 0.01 level

The result indicates that the correlation is significant at the 0.01 level (2 tailed). Hypothesis 4 that states that there will be significant and positive correlation between computer hassles and stress manifestation is thereby confirmed.

For this study, we were also curiously interested in finding out the frequency of stress reactions amongst the two groups, based on the sub-scale that assesses stress manifestations. Table 7 shows the results.

Table 7: Stress Reactions/Manifestations among Respondents

Symptoms of stress manifestastion - Technostress	**Academics**			**Bankers**		
	Mal	Female	%	Male	Female	%
Backache	47	34	79.4	19	12	31.0
Headache	48	41	87.3	22	20	40.0
Pains in the eyes	42	36	76.5	26	21	47.5
Irritability	38	34	71.0	14	14	28.2
Poor sleep	32	26	56.8	21	18	39.4
Dizziness	38	34	71.0	22	14	36.4
Emotional Outburst	26	22	47.1	12	II	23.2
Poor Bowel Movement	31	22	44.2	14	07.	21.2
Interpersonal Difficulties	31	24	53.9	06	07	15.2

The above table shows the percentages of the technostress manifestations by the samples from both the academic community and the banking sector.

4.0 *Discussion*

The study revealed statistically significant differences on the scores of the respondents on technostress exist. Specifically, it was observed that academic staff obtained higher scores on computer hassles and stress reactions (two sub-scales of technostress) than those obtained by employees of commercial banks. The t-test result revealed that the observed differences were statistically significant at: obtained t = 8.6; 3.54, critical t = 1.65, Probability = <.05, df = 199. This is an interesting finding because one would have expected that employees of commercial banks would have presented significantly higher levels of technostress than academic staff considering the impression created by the bank workers and the myth surrounding the time they close for work. However, this study made use of a special category of academics, the open and distance learning faculty members of which job specifications and expectations are largely ICT- based. This is in recognition of the fact that the global arena is currently an e-based arena where Information and Communication Technology is chiefly employed in almost all spheres of life and education sector is not exclusive of this. This, therefore, confirms the fact that educators have adopted ICT as an effective teaching tool if they are to compete effectively and perform efficiently in the global age. The study also revealed no statistically significant differences on symptom and manifestations of technostress among the male and female respondents. In other words, technostress is not gender specific. This finding is in tandem with Raja, Azlina and Siti's [16] study on technostress wherein no significant difference among male and female respondents on technostress exists, because the global challenges are triggering competition and gender role reversal. There is visible gender role reversal in all spheres of working life, thus more women now work in formal organisations than it was in the last century. This study also discovered that older respondents manifested higher levels of technostress than the younger ones. Specifically, it was observed that those aged 60 - 69 years and above presented highest symptoms of technostress than others. This group was closely followed by those aged 40-49 years and 50-59 years respectively. The One-Way ANOVA results indicate that the observed differences were statistically significant at P. = <.05, df = 4/196, obtained f value = 4.492; 5.231, critical f = 3.03. It is not an anomaly to assume that the low scores obtained by the younger respondents is an indication of acceptance and comfortability with the ICT. Those in this age range could be described as digital natives. This is aptly captured in Marc Prensky digital migrant and digital native dichotomy (Raja, Azlina and Siti's [16]. He identified a digital native as a person who understands the value of digital technology and uses this to seek out opportunities for implementing it. A digital migrant on the other hand was described as an individual who was born before the existence of digital technology and adopted it lo some extent later in life. It is however not surprising that older respondent manifesting more technostress than the younger could be as a result of the pressure to adjust and change to the new information communication technology. Also, the correlation statistics showed positive correlation between the subscales of technostress: computer hassles and stress manifestations. This shows that those subjected to high level of technostress are prone to stress-related physiological and psychological complications. Such stress-related complications include lowered immunity, backache, neck ache, tiredness and sleep problems. Others include hypertension, headaches, dizziness, poor appetite, asthma, gastrointestinal disorders, skin rashes, blurred vision, emotional outburst and interpersonal difficulties. This is detrimental to the worker and the institution because it will greatly affect the productivity level.

5.0 *Conclusion*

The contemporary world is a very stressful one with new different series of demands and changes to grapple with. As technology is here to stay, it is crucial we appreciate the emotional and physiological responses to technology as well as fashion out adaptable ways of adjusting to it. Technostress can reduce employee productivity and create dissonance in the work environment, costing employers time and money. Given the trend toward an increasingly faster-paced and more stressful work environment, it seems reasonable to develop effective training and wellness programmes to decrease employees' stress levels and to enhance their sense of technological mastery and personal value. As an antidote to counter the problems that could emanate from technostress and other stress-induced activities in society, the following could be helpful for the individuals and institutions in general.

Implications for the Individual

To reduce the psycho-physiological impact of technostress, individuals should strive to get enough exercise - this is known to reduce stress. They should also learn relaxation techniques - this can help them to sleep better and relieve stress-related physical pains such as stomach pains, headache and backache. In addition, individuals should desist from drinking too much alcohol or caffeine. Instead of helping, these stimulants increase the stress level. They should eat regular rneals and a health meal, control emotional outburst through deep-breathing exercises, practice time management and make sure that the work environment is comfortable. If it is not, they should ask for help from their institutions.

Implications for Institutions

The findings of the study have implications on Distance Learning (DL). Generally, policy measures in DL should strive to sustain an ever-present system of training using effective technologies. DL institutions should strive to adopt user friendly hardware and software with provision for adequate training for the staff. This adduced implication is in tandem with the implication of the result of the study that indicates that DL practitioners manifested higher level of technostress than those in the banking sector. Truly, the two sectors under study rely heavily on ICT for their work schedule and output. It appears that banking sectors are more active in training and exposing their staff to new and effective technologies than those in the education sector. Thus, there is the need for DL institutions and DL administrators to provide improved and adequate ICT facilities for their workforce. This observation stresses the need for adequate training for the staff. In addition to this, DL institutions and administrators should create a better communication channel within the work environment as well as encourage improved level of reassurance, patience and stability within the institution. This is because the practice and expectations of DL work requires full concentration and borderless time which can be tasking on individuals emotionally and physiologically. Also, results of this study have a lot of implications for DL in West Africa. To be able to compete globally, African countries need to bridge the digital divide in learning and development and DL is a viable tool in this area. Thus, concerted effort should be made in providing DL practitioners in the sub-region with up-to-date technology to be able to provide best practices in education.

References

[1] Brod, C. (1984). Technostress: The Human Cost of the Computer Revolution. Reading: Addison-Wesley Publishing Company.

[2] Fuglseth, A. M., and Sørebø, O (2014). The effects of technostress within the context of employee use of ICT. *Comput. Hum. Behav.* 40, 161–170. doi: 10.1016/j.chb.2014.07.040

[3] Ragu-Nathan, T.S., Tarafdar, M., Ragu-Nathan, B. and Tu, Q. (2008). "The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation," Information Systems Research, 19:4

[4] Kupersmith, J. (1992). Technostress and the Reference Librarian. Reference Services Review, Vol. 20, pp. 7-14.

[5] Tams, S., Hill, K., Ortiz de Guinea, A., Thatcher, J. & Grover, V. (2014). Neurols – alternative or Compliment to existing methods? Illustrating the holistic effects of neuroscience and self – reported data in the context of technostress research. Journal of the Association for Information Systems, 15 (10), 723 – 753.

[6] Arnetz, Bengt B.and Berg, Mats (1993). Techno - Stress. Psycho-Physiological sequences of Poor Man-machine Interface. In Michael J. Smith & Gavriel Salvendy (Eds.). Human-Computer Interaction: Applications and Case Studies. Amsterdam: Elsevier, 891- 896.

[7] Muter, P., Furedy, J. J., Vincent, A. and Pelcowitz, T. (1993). User- hostile Systems and Patterns of Psycho physiological Activity. Computers in Human Behaviour, 9, 105-111.

[8] Emurian, Henry H. (1993). Human-computer Interactions: Are there adverse Health Consequences? Computers in Human Behavior, 5, 265 - 275.

[9] Charlesworth, E. & Nathan, R. Stress Management. NY: Ballantine.

[10] Clute, R. (1998). Technostress: A Content Analysis. Kent State University. Kupersmith, J. (1992). Technostress and the Reference Librarian. Reference Services Review, Vol. 20, pp. 7-14.

[11] Al-Fudail, M., &Mellar, H. (2007). Investigating Teacher Stress When Using Technology (Electronic Version). Computers & Education, Vol. 51, No. 3, pp. 1103-10.

[12] Al-Qallaf, C. L. (2006). Librarians and Technology in Academic and Research Libraries in Kuwait: Perceptions and Effects. Libri, Vol. 56, pp. 168-79.

[13] Massey, M. & Stedman, D. (1995). Emotional climate in the information technology organization: Crisis or crossroads? Cause/Effect Magazine, 8 (4), 7-19.

[14] Friedman, M. & Rosenman, R. H. (1974). Type Behaviour and your Heart. NY: Knopf.

[15] Wei, Oiu (2013). Impact of Technostress on Job Satisfaction and Organizational Commitment. M.sc Thesis, Management. Massey University, Auckland, New Zealand.

[16] Raja, I., Azline, A. & Siti, B. (2007). Technostress: A study of academic and non-academic staff. In Dainoff, M. J. (Ed.). Egomomics and Health Aspects. eidelberg: springer-Verlege.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 337-340, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1007

Blockchain as a Solution of Information Security and Data Privacy Issues: Review

Ndung’u Rachael Njeri
Department of Information Technology
Murang’a University of Technology
Kenya

Abstract: The growth of technology has seen development of smart devices that are connected to each other giving rise to device-mesh technology. This has given rise to many owners of these devices sharing data through various web applications such as online marketplaces. The protection of data is paramount for every organization dealing with such data. An evaluation of Blockchain technology as a solution to data privacy is studied. The study concludes that though blockchain is the technology to pursue for securing and protection data, it has numerous challenges and limitations towards data privacy. More research is needed to guarantee an absolute data privacy protection.

Keywords: Blockchain; Information security; Data privacy; Issues of privacy; Personal Identifying Information

1. INTRODUCTION

Today’s technology advancement has evolved tremendously with Artificial Intelligence (AI) taking the lead. In the IEEE computer society’s top ten technology predictions for 2021, machine learning, robotics and industrial Internet of Things (IoT) were seen as the technologies that would highly hold disruptive potential by 2021 going forward, amongst other top technologies. Smart devices evolution brings with it issues of mass data collection of very high magnitude bringing the term ‘big data’ in the fore. The big data has issues with its management, analytics and data privacy issues [1].

[2] when writing for CPO Magazine on data privacy in the era of the Internet of Things noted that new smart home devices like the Amazon Echo and Google Home were raising numerous legal and data privacy issues, primarily because these IoT devices were recording conversations that were held in daily life. Smart toys powered by AI are everywhere playing with kids, giving them company and socializing with kids. But for them to get answers asked by the kids they have to be connected to the internet. Hackers take advantage and use such toys to infiltrate to confidential data by monitoring or illegally spying on children. A major concern with such data sourced from these smart devices is their security and privacy. How these data are used could compromise the data privacy regulations, which can result to compromising the processed information. It’s the responsibility of organizations developing smart devices to take responsibility and ensure that their products are protected from unauthorized access.

Protection of information and data access is paramount for every organization dealing with huge amounts of company and personnel data. Data management is becoming an important frontier in many organizations, which are dealing with data providers, data collectors and data processors. Blockchain technology is among the latest strategies of decentralized data storage that reduces to the minimum unauthorized access of stored data through trust. Some research works [1] [3] have concentrated on bid data privacy challenges. From an information security viewpoint, many systems have been developed aiming on privacy of personal data, such as data anonymization that protect personal identifiable information (PII), differential privacy technique that adds noise to the computational procedure before data sharing, and encryption schemes that allow processing of encrypted data [4]. This paper aims to assess the solutions on data privacy issues using blockchain technology, as a recent emerging technology. Description of data privacy will be assessed and blockchain technology explained. Critical examination of blockchain will be done in light of establishing how it can be used as a solution for the data privacy.

2. DATA PRIVACY

Many literatures have been written trying to understand the term data privacy. There are many viewpoints to evaluate data privacy. No complete and comprehensive description has been given to it since many researchers evaluate privacy issues depending on where they stand. Issues of privacy can be based on the users where their personal private data is disclosed to those who are not merited and without the users’ consent. Organization’s data privacy would mean securing confidential data about the organization from its competitors and once such data is leaked, data privacy is compromised.

Data privacy relates to control of the distribution and use of consumer information including and not limited to statistical features of human populace such as age, income, used to identify individuals, search history and personal profile information [5] [6] [7].

Privacy is multi-dimensional that requires profound address of the issues necessary to fully understand this important issue. Many database handlers provide privacy controls based on how they understand privacy issues. Data privacy taxonomy, can be considered in four technical dimensions; data providers - individuals or organizations that provide data to be stored, data collectors - individuals or organizations that initially collects, uses and stores from the providers, data user - individual or organizations that solicits for acquired data as third parties and data warehouse- the data store itself as key players in matters data privacy. Data privacy is guided by basic tenets such as need to specify the purpose, need to acquire consent, limiting data collection and use to only what is required, data disclosure and its retention is limited as possible, data collected is accurate and security safeguards are assured and that data is open to provider who verifies compliance to these tenets. They identified four dimensions to understand data privacy as purpose, visibility, granularity and retention [8].

The explicit responsibility of data usage, access and providence of secure safeguards that will ultimately secure that data. The data collectors are also guided on how long they can store the data and to whom they can disclose it to through consent given by the data provider. Many at times data sources/owners have no control of how their data, i.e. how it will be used and for how long. Majority of data controllers will share data at their disposal without any regard of privacy regulations. Any technology that give the data owner control of his data, how it will be used by firms and authorities without compromising security and limiting them the capability to personalize services, will go a long way to provide technical solutions to data privacy.

Smart devices that are been employed by many organizations are collecting personal data and storing them in the databases. The IoT objects are connected to each to other and are provided with internet. These objects accumulate a lot of information from their surroundings and share with each other through the internet connection over software system. They produce a lot of information that are used by reliant services such as online marketing. This raised data privacy issues since these objects distribute personal data that would reveal identities of their owners [9].

Organizations on the other hand are collecting information, combining facts from separate sources, merging and swapping them using smart software, and then selling them as merchandise compromising with the very rule of consent from the data providers. This is calling for privacy protection from the governments. Therefore, issues of user personal data disclosure and misuse is threatening the very core of information security in organizations. Many firms are concerned on how stolen data can harm their businesses and how it can be used to compromise how they do business. Wrongly acquired data can harm the reputation of individuals or organizations, which can make them lose self-image and business partners. This would be very critical and getting how to prevent and safely secure such data is among the high priorities that government, firms and individuals are seeking protection for. Though there are many legal, ethical and policy avenues advanced for data privacy protection, they are beyond the scope of this paper. This paper seeks to examine privacy protection through technical projects, and as was mentioned earlier, Blockchain will form the basis of such venture discussed herein.

3. BLOCKCHAIN TECHNOLOGY

The blockchain is a decentralized, digitized, highly distributed public ledger or record that can be accessed by anyone via the Internet, or privately through a restricted network. It works like an enormous virtual accounting book, which records details of every single transaction of data between two parties. It's replicated, shared and coordinated digital data structure that is maintained by consensus protocol and it spreads across many establishments, countries and several sites. It's under peer-to-peer network with no central control. Blockchain transactions uses cryptographic keys where public keys are used as an address to the system and the private key used for signing transactions.

The digital data is packaged into blocks, which are connected together sequentially, in a manner that makes such data immutable once recorded, unless collusion of member nodes on the peer-to-peer network to alter such data. Each block contains a hash pointer as a link to a previous block, a timestamp of when the block was created and the data being transacted. There are numerous blockchains and each blockchain serves a different purpose. Blockchains can be open and public, or they can be privately run by enterprises or even individuals. There many varieties of blockchains thus lacking single set standards, complicating how they are implemented [10] [11]. There are many participants that may qualify as data controllers- those who determine the purpose and manner of processing- for themselves and data processors - those who process on behalf of data controllers must be guided by applicable data privacy laws [12] [17].

3.1 How blockchain and data privacy works

[12] [17] considered the blockchain technology and how it works. He deduced that for blockchain to work effectively the members participating on the chain must be on the same level of knowledge about the blockchain principles. Though many users can upload any kind of data, including personal data, many generic blockchain may not be able to provide data privacy protections thus calls for the users to uphold data protection laws as they upload data onto the blockchain since data transmitted through the chain is visible to every participant though the information cannot be removed from the chain.

The data providers, data controllers, and data processors must be guided by governance agreements among the participants in data privacy. Due to the fact that blockchain deals with data which is distributed in nature, many issues emerge on who is responsible to control what data, thus posing a challenge on which data privacy laws to be followed. To achieve data privacy on blockchain, use of hash personally identifiable information is applied serving as locus link to an off- chain data store. The degree of anonymity that can be provided by blockchain is that of pseudonomity, where virtual identity is required for transactions, while integrity of blockchain largely depends on the complex proof-of-work protocol and largely on honest miners. To manage data privacy, decentralizing data over private- by design systems is way to go. The benefit of adopting such a technology is that of decentralized records that are not controlled centrally and that contains immutable transactions. These transactions are registered and authenticated thus ensuring detection of misuse and tampering of data. This can be achieved by adoption of peer-to-peer systems which blockchain technology employs [9] [12] [17].

Blockchain is a decentralized (not centrally controlled) peer-to-peer system using proof-of-work consensus algorithm that relies upon cooperation among individual nodes to carry out information transmission operations. To add more security and integrity of information while using blockchain, cryptographic public key is used for authentication. This lessens over reliance of third-parties and offer more security to transactions and identity privacy. Like in bitcoin systems, which implements cryptographic Proof of Work (PoW) together with nested chain of hashed addresses to remove the need for third party providing security and privacy even when dealing with unknown nodes. Bit-message used offers anonymity in a trustless network through transmitting encrypted messages in messaging streams. The identities of data owners and their profiles are kept private by using trustless system [13] [11]. Therefore, blockchain can be used to automatically control decisions about collecting, storing and sharing sensitive data by making the ledger act as a legal confirmation for accessing or storing data since its immutable.

Similarly, researchers are working on a protocol that would sit on top of an existing blockchain that promises ‘secrets contracts’ as opposed to ‘smart contracts’ which will make the nodes on the blockchain be able to compute data without ‘seeing’ it. This would allow users to maintain control over personal data through preventing its monetization by online platforms. This would enhance their trustworthiness without having to give individuals access to their specific personal data.

3.2 Blockchain challenges

Blockchain technology is not a technology without issues but why is it deemed to offer security of data for organizations is due to the very nature of the blockchain, anybody wanting to tamper with it would have to make changes to records which are stored on multiple computers, or use a lot of computing power to mine a new branch of the blockchain. Many establishments and researchers are wondering whether blockchain technology would provide data privacy, integrity and anonymity as secure means to deal with data on information systems. When reviewing about blockchain technology, [14] found out that the technology has faults when it comes to offering privacy. Its anonymity element could be traced through tracing some subset of addresses that could be mapped to an IP address by simply observing the transaction relay traffic making the transaction linking possible. This would greatly compromise the data privacy since the data owners could be identified. The proposed technique from that review was transaction mixing technique in order to achieve increased privacy. A mixing transaction technique would allow the participants on the blockchain to move from one user address to another without a clear trace linking between the addresses. Such transactions could act as basic to aid improve anonymity when transaction linking becomes more puzzling [14].

The degree of anonymity that can be provided by blockchain is that of pseudonomity, where virtual identity is required for transactions, while integrity of blockchain largely depends on the complex proof-of-work protocol and largely on honest miners. This pseudonomity means that the block can be traced back to its source thus it doesn’t have an exclusive anonymous status. With PoW, the majority attack, which is given at 51% is common. That is the likelihood of mining a block depends on the work done by the miner. This mechanism will make people want to link together to mine more blocks contributing to mining groups. When such a group holds 51% computing power, it can take control the blockchain. Ostensibly, it can cause security issues [15]. Other technical challenges and limitations of the blockchain technology was that of throughput, latency, size and bandwidth, versioning, hard forks, scalability and multiple chains. Very few literatures were found dealing with these challenges and this portend a rich ground of further research [14].

[16] reviewed literature on blockchain and they identified three algorithms that were designed to maintain how the blockchains could remain verifiable as distributed ledgers, one with pattern clustering algorithm that would cluster nodes into different clusters while using sequence data to represent how nodes behave, and uses similarity metrics and utilizes Euclidean distance. Another algorithm they identified was one using post quantum that ensured security against external threats thus upholding blockchain authentication, where secret keys are generated using lattice basis delegation. The third algorithm discoursed was a multilink integrated factor that would shorten time for validation while improving the dependability of the communication within the blockchain utilizing node link number that would identify those nodes with high capacity of trust and communication. However, all the three algorithms were not without issues, even though they were with high privacy and high secure metrics, malicious nodes were would still exist.

The review of blockchain technology by [14] summarized their findings, which are demonstrated on figure 1 below

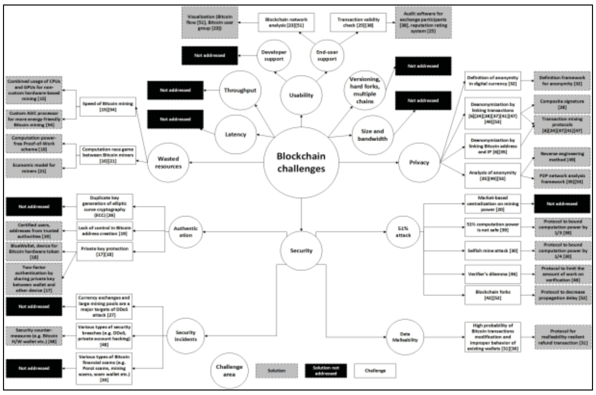

Visualization (Bitcoin flow [51], Bitcoin user group [23])
Blockchain network analysis [23][51]
Transaction validity check [25][38]
Audit software for exchange participants [38], reputation rating system [25]
Not addressed
Developer support
End-user support
Not addressed
Combined usage of CPUs and GPUs for non-custom hardware-based mining [15]
Speed of Bitcoin mining [15][54]
Not addressed
Throughput
Usability
Versioning, hard forks, multiple chains
Not addressed
Definition of anonymity in digital currency [32]
Definition framework for anonymity [32]
Custom ASIC processor for more energy-friendly Bitcoin mining [54]
Not addressed
Latency
Size and bandwidth
Deanonymization by linking transactions [6][24][28][37][41][47][49][53]
Composite signature [28]
Transaction mixing protocols [6][24][37][41][47]
Computation power-free Proof-of-Work scheme [16]
Computation race game between Bitcoin miners [16][21]
Economic model for miners [21]
Wasted resources
Blockchain challenges
Privacy
Deanonymization by linking Bitcoin address and IP [6][35]
Analysis of anonymity [35][49][53]
Reverse engineering method [49]
P2P network analysis framework [35][53]
Not addressed
Duplicate key generation of elliptic curve cryptography (ECC) [26]
Market-based centralization on mining power [20]
Not addressed
Certified users, addresses from trusted authorities [19]
Lack of control in Bitcoin address creation [19]
Authentication
51% attack
51% computation power is not safe [39]
Protocol to bound computation power by 1/3 [39]
BlueWallet, device for Bitcoin hardware token [18]
Private key protection [17][18]
Security
Selfish mine attack [30]
Protocol to bound computation power by 1/4 [30]
Two-factor authentication by sharing private key between wallet and other device [17]
Verifier’s dilemma [46]
Protocol to limit the amount of work on verification [46]
Blockchain forks [42][52]
Protocol to decrease propagation delay [52]
Not addressed
Currency exchanges and large mining pools are a major targets of DDoS attack [27]
Security counter-measures (e.g. Bitcoin H/W wallet etc.) [48]
Various types of security breaches (e.g. DDoS, private account hacking) [48]
Security incidents
Data Malleability
High probability of Bitcoin transactions modification and improper behavior of existing wallets [31][36]
Protocol for malleability-resilient refund transaction [31]
Not addressed
Various types of Bitcoin financial scams (e.g. Ponzi scams, mining scams, scam wallet etc.) [33]
Challenge area
Solution
Solution not addressed
Challenge

Figure 1. Summary of the identified challenges and solutions of Blockchain

3.3 Solutions for Data Anonymity

The implemented blockchain-based methods that improve on the anonymity, a data privacy concept has been applied on technologies such as Bitcoin. Bitcoin is an online payment system, which is publicly viewed as a means of sending monies anonymously. Users may take too lightly the amount of Personally Identifying Information (PII) obviously linked to this digital currency. To conform to Anti- Money Laundering and Know Your Customer regulations based on the online marketplaces, online marketplaces monitor user activity and collect PII from the bank accounts and credit cards used to purchase Bitcoins. A discoverable link between real-world identities and online Bitcoin transactions can be eventually created from that information. To eliminate this anonymity threat, privacy-conscious users rely on Bitcoin mixing services to remove identity-based connections from their coins. Several mixing schemes are been used with the bitcoin transactions such as Bitcoin Mixer, where mixing service’s inclination to reprocess shared pots for storing and redistributing Bitcoins makes it readily identifiable in the blockchain and makes it easier to monitor for addresses depositing to and receiving coins from the service, Bit Launder that complicates analysis by using unpredictable payout timing. The mixing service uses repetitive extracting addresses, which makes it easier to monitor the blockchain for questionable receiving addresses, and to calculate possible originating addresses through balance variances, Bitcoin Blender where timing analysis reliability when tested was reduced due to variable time differences recorded between deposits and withdrawals across trials. However, research showed unique blockchain mixing characteristics can be used to mark each mixing service.

Nevertheless, these techniques have their limitations and therefore they are limited to offer the desired anonymity when transacting using bitcoins. To improve on anonymity on data privacy, such mixing techniques could be improved by eliminating their various limitations and either extending or remodeling to have a technique that can offer better degree of anonymity as a way to conceal personal identity while using blockchain transactions.

3.4 Blockchain Opportunities

Despite the blockchain challenges, there is a sea of opportunities for data privacy protection. Personal data can be collected into personal identifiable information (PII) and stored in distributed shared ledgers. Data privacy is about maintaining data integrity, confidentiality and availability, while blockchain might not enforce confidentiality, it enforces strong integrity and availability owning its decentralized nature in that data within it are transparent to members of the blockchain network. The ledger technology has an opportunity in cybersecurity for securing personal data to transactional data through its encrypted form of data transmission. This technology can be utilized to offer regulation relating to personal identifiable information and data privacy, organizations should embrace the technology and work towards integrating it with data privacy's regulation of the organization or country so as to achieve the best from both worlds.

4. CONCLUSION

Blockchain technology is in the recent past been one of the most reputed disruptive technologies that deemed the ultimate information security and data privacy breach solution. Though in its very early stages of development, many governments and reputed organizations are on research to seek how this technology could help to secure data such as land records, patient's records in the health environments, supply chain markets data and even the data privacy for online platforms. Blockchain is at infancy since its implementation is still very low. This paper however showcases the benefits of blockchain, its various challenges especially on data privacy and could be solutions for those challenges.

The key point of this paper was to explore whether this technology can be used as solution to data privacy. In above literature, it was pointed out that among major blockchain challenges was that of anonymity since the kind provided by blockchain is pseudonomity where one would be traced back through IP address mapping to identify the owner of transaction, yet the anonymity envisaged is that regardless of any trace, owners of transactions cannot be identified.

Several literatures studied herein reports how blockchain was envisaged to work but very few shows how it's been implemented and the success about its implementation. The various blockchain-based techniques used to provide data privacy by enhancing anonymity were without their challenges.

It's the researchers' assertion that as much as blockchain technology is looked upon to be a solution of data privacy, it's still yet to be demonstrated as the ultimate solution of data privacy. The researcher feels that more research is needed to model data privacy enhancing techniques or methods, which can guarantee better data protection and privacy based on blockchain technology. This doesn't mean that blockchain is not worthy technology as far as security of data is concerned. The literature shows how data recorded on blockchain would be difficult to interfere with since it required consensus of participants to change any details and this is achieved through rigorous proof-of-work algorithms thus immutability of data is achieved, which strongly provides data security. Therefore, blockchain is commendable technology as far as data security is concerned though much more research is required to improve on its data privacy provision.

5. REFERENCES

[1] S. Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data," in IEEE Access, vol. 4, pp. 2751-2763, 2016, doi: 10.1109/ACCESS.2016.2577036.

[2] Nicole Lindsey, March 17, 2017 - https://www.cpomagazine.com/data-privacy/data-privacy-era-internet-of-things/

[3] Soria-Comas, J., Domingo-Ferrer, J. Big Data Privacy: Challenges to Privacy Principles and Models. *Data Sci. Eng.* **1,** 21–28 (2016). https://doi.org/10.1007/s41019-015-0001-x

[4] Zyskind, G., & Nathan, O. (2015, May). Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops* (pp. 180-184). IEEE.

[5] Foxman, E. R., & Kilcoyne, P. (1993). Information technology, marketing practice, and consumer privacy: Ethical issues. *Journal of Public Policy & Marketing*, *12*(1), 106-119.

[6] IEEE TRENDS- https://www.computer.org/press-room/2017-news/top-technology-trends-2018

[7] Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. Journal of the Academy of Marketing Science, 45(2), 135-155.

[8] Barker, K., Askari, M., Banerjee, M., Ghazinour, K., Mackas, B., Majedi, M., ... & Williams, A.(2009, July). A data privacy taxonomy. In British National Conference on Databases (pp. 42-54). Springer, Berlin, Heidelberg.

[9] Conoscenti, Marco; Vetrò, Antonio; De Martin, Juan Carlos (2016). Blockchain for the Internet of Things: a Systematic Literature Review. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir (MAR), Nov. 29 2016-Dec. 2 2016. pp. 1-6

[10] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[11] Aitzhan, N. Z., & Svetinovic, D. (2016). Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams. *IEEE Transactions on Dependable and Secure Computing*.

[12] McMyn, A., & Sim, M. (2017). R3 Reports with Hogan Lovells.

[13] J. Warren, "Bitmessage: A peer-to-peer message authentication and delivery system," white paper (27 November 2012), https://bitmessage. org/bitmessage. pdf, 2012.

[14] Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology?—a systematic review. *PloS one*, *11*(10), e0163477.

[15] Lin, I. C., & Liao, T. C. (2017). A survey of blockchain security issues and challenges. *Int. J. Netw. Secur.*, *19*(5), 653-659.

[16] Gorkhali, A., Li, L., & Shrestha, A. (2020). Blockchain: A literature review. *Journal of Management Analytics*, *7*(3), 321-343.

[17] https://www.hlengage.com/_uploads/downloads/5425GuidetoblockchainV9FORWEB.pdf

International Journal of Computer Applications Technology and Research
Volume 11–Issue 08, 341-350, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1108.1008

Model for Enhancing Performance of Network Intrusion Detection based on Hybrid Feature Selection and Unsupervised Learning Techniques

Joseph Mbugua Chahira
Department of Information Science
Garissa University, Kenya

Jane Kinanu Kiruki
Department of Computer Science
Chuka University, Kenya

Abstract

As security threats change and advance in a drastic way, relevant of the organizations implement multiple Network Intrusion Detection Systems (NIDSs) to optimize detection and to provide comprehensive view of intrusion activities. But NIDSs trigger a massive amount of alerts even for a day and overwhelmed security experts as they require high levels of human involvement in creating the system and/or maintaining it. The main goal in this work is to enhances the structural based alert correlation model to improve the quality of alerts and detection capability of NIDS by grouping alerts with common attributes based on unsupervised learning techniques. This work compares four unsupervised learning algorithms namely Self-organizing maps (SOM), K-means, Expectation and Maximization (EM) and Fuzzy C-means (FCM) to select the best cluster algorithm based on Clustering Accuracy Rate (CAR), Clustering Error (CE) and processing time. The result inferred that the proposed model based on hybrid feature selection, PCA and EM is effective in terms of Clustering Accuracy Rate (CAR) and processing time for The NSL-KDD Dataset

Key words: Network Intrusion Detection, unsupervised learning, Clustering, alert correlation,
Structural-based AC.

Introduction

Intrusion detection is a system for detecting intrusions and hence works as the major defensive mechanism in a network environment[1][2]–[4]. Its main goal is to automatically monitor network traffic and classify them as normal or suspicious activities and inform the Security Analyst or response system to take appropriate action before the intrusion compromises the network.

Network monitoring has been used extensively for the purposes of security, forensics and anomaly detection. However, recent advances have created many new obstacles for NIDSs. Firstly, they generate huge volume of low quality evidence and in different format produced by distributed IDS systems (Application, Network and Host based). Unfortunately, most of the alerts generated are either false positive, i.e. benign traffic that has been classified as intrusions, or irrelevant, i.e. attacks that are not successful . This results in slow training and testing correlation processes, higher resource consumption, lower accuracy and higher performance costs.

The unsupervised machine learning algorithms are applied when there is no class to be predicted but when the instances are to be subdivided into natural groups of instances determined by the features available to represent the items into clusters [3], [5]. The algorithms can be trained on unlabeled data or can be applied to the test or evaluation data without training. The trained clustering algorithms build internal representation of unlabeled training data during training which apply to the test data set. The untrained clustering algorithms determines natural differences between subsets of data without prior insight into the data.

In order to improve the quality of alerts for analysis, some research in alert clustering for finding structural correlation have been done. In Structural-based AC (SAC), alerts are correlated based on similarity of attributes. Similarity index or function is used to determine the degree of relationships. Although it can discover known group of alerts or attack steps, research by [6][7]–[9] claimed that it cannot discover the causal relationships among alerts. The major problem in previous techniques is they relied heavily on Security Experts (SE) in developing and maintaining their correlation system. They are based on pre-defined rules or expert knowledge to manage and analyze the intrusion alerts and as a result, rules or knowledge for such systems need to be updated periodically as patterns of attacks change drastically [8], [10].

The aim of this work is to enhance the Structural-based AC model using machine learning technique to improve the quality of alerts and identify attack strategy. An intelligent hybrid clustering model is developed based on normalization, discretization and Improved Unit Range (IUR) technique to preprocess the dataset, EMFFS, Principal Component Analysis (PCA), SAC and proposed Post-Clustering algorithms is implemented to reduce the alerts dimensionality and optimize the performance and unsupervised learning algorithm to aggregate similar alerts and to reduce the number of alerts. In the proposed model the performance of various unsupervised learning techniques like Self-organizing maps (SOM), Expectation Maximization, K-means, hybrid clustering and Fuzzy c-means (FCM) is compared based on four measurements techniques applied are: (1) Clustering Error (CE) is the number of alerts that are wrongly clustered. (2) Error Rate (ER) is the percentage of wrongly clustered alerts, ER = (CE ÷ Total number of alerts observed) x 100, (3) Accuracy Rate (AR) is the percentage of alerts that are accurately clustered as they should be, AR = 100 – ER, and (4) Time is the algorithm processing time in seconds.

Related work

Collection mechanism and reduction of IDS alert framework (CMRAF) [11] was proposed to remove the duplicates IDS alerts and reduce the number of false alerts. They use information gain ratio algorithms to extract the similarities between set of alerts and provide the highest weight to the most effective features based on the class of alerts belonging to the algorithm.

Alert correlation using a novel clustering approach, [12], applied an incremental clustering approach to reduce the amount of alerts generated by IDS. Three attributes, destination IO, signature-id, and timestamp had been extracted and hashed by using MD5. The hash value from the next input tuple is checked against hash value of the existing clusters. The hashing technique is used to speed up the comparison in checking the similarities of alert attributes.

An improved framework for intrusion alert correlation by Elshoush *at el*, (2012), divided alert correlation into ten main components and contained them in the Data Normalization Unit, Filter-based Correlation Unit and Data Reduction Unit. Similar alerts are fused based on seven extracted features, namely Event ID, timesec, SrcIPAddress, DestPort, DestIPAddress, OrigEventName, and SrcPort in order to remove duplicate alerts created by the independent detection of the same attack by different sensors.

A probabilistic-based approach proposed [13], correlate and aggregate security alerts by measuring and evaluating the similarities of alert attributes. They use a similarity metric to fuse alerts into meta-alerts to provide a higher-level view of the security state of the system. Alert aggregation and scenario construction are conducted by enhancing or relaxing the similarity requirements in some attribute fields. But similarity correlation is the only way for them to aggregate the alerts. They have to compare all the alert pairs and have to determine lot of thresholds with expert knowledge which lead to their huge volume of computing workload.

Methodology

In this study, the quantitative approach is preferred as the main method due to certain characteristics, such as performance measures, dataset evaluations and the usability of the results. This research has employed a deductive reasoning because it seems to be more appropriate to test the proposed solutions. It addresses the issues of improving the quality of alerts that are generated by multiple NIDSs and recognizing the attack strategy from the unrelated alerts. The identified problems under these issues and the coverage of each objective in this research are solved though these steps

Step (1) Read the pre-processed alerts as inputs.

Alerts that have been processed by Multi-Filter Feature Selection (EMFFS) Method are read from the database as inputs to clustering phase.

Step (2) Reduce the alerts high dimensionality.

All alerts with their attributes are dimensionally reduced using statistical PCA

Step (3) Adopt unsupervised learning algorithm which gives the highest accuracy. Expectation Maximization (EM), (K-means, FCM and SOM. unsupervised learning algorithm are tested and compared.

Step (4) Measure and validate the clustering and post-clustering performances. The performances of the proposed clustering system can be measured using predefined measurements.

Step (5) Save the analysis and experimental results. The analysis and experimental results are recorded and saved in the database. It includes the details on all of the identified clusters attack steps as well as the statistical analysis.

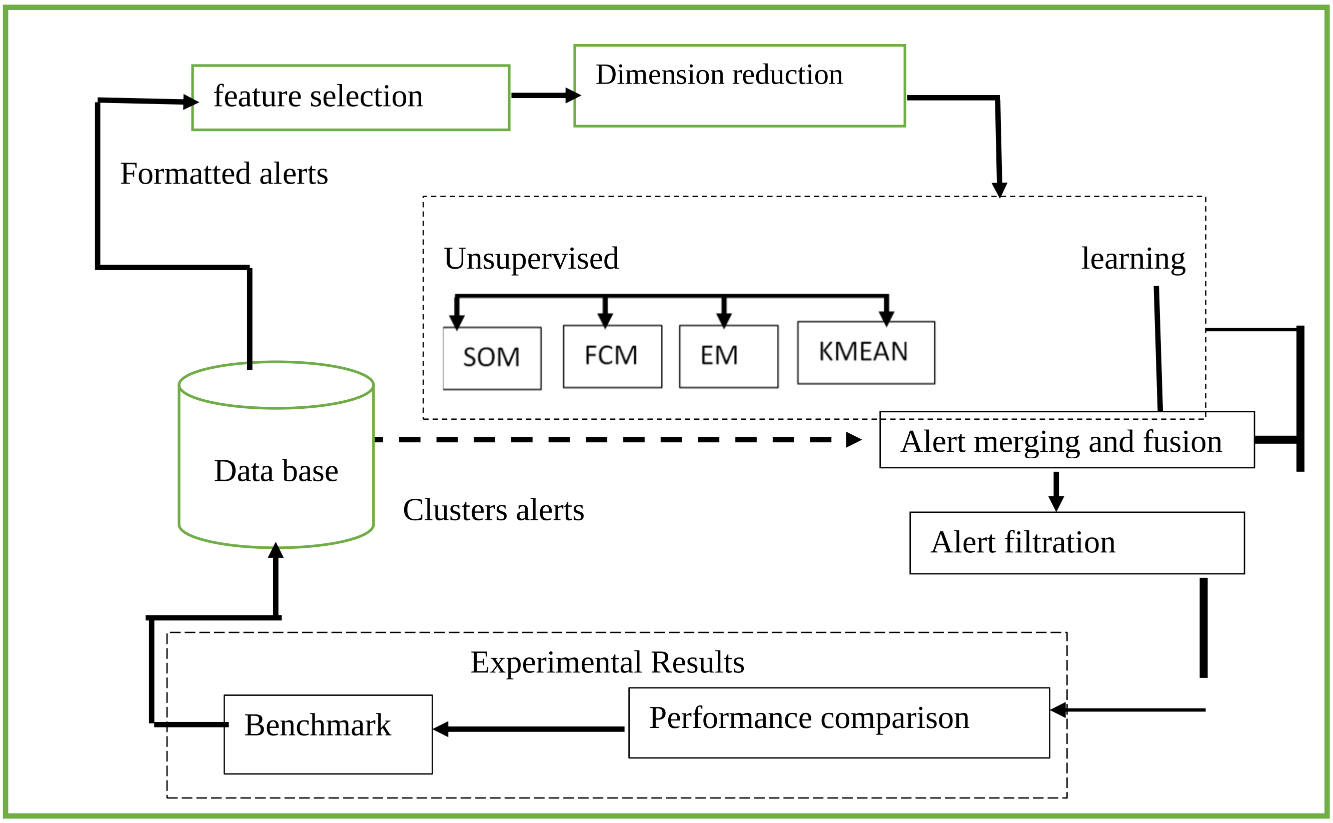


Figure 1: The flowchart of enhanced structural-based alert correlation model

Steps

Ensemble-based Multi-Filter Feature Selection (EMFFS) Method

The main aim of feature selection is to eliminates irrelevant and repetitive features from the dataset to make robust, efficient, accurate and lightweight intrusion detection system To achieve this objective, a model for network intrusion detection system based on Multi-Filter Feature Selection (**EMFFS**) Method is implemented developed by [14] to find the best set of features that are used in this work. The feature selection techniques integrated are Correlation Feature Selection (CFS) based evaluator with Best-first searching method, Information Gain (IG) based Attributes Evaluator with ranker searching method, and Chi Squared and Ranker searching method.

Unsupervised learning techniques

Self Organising Maps

The Self-Organizing Map [15], [16] is a neural network model for analyzing and visualizing high dimensional data and it belongs to the category of competitive learning network. The SOM defines a mapping from high dimensional input data space onto a regular two-dimensional array designed architecture as input vector with six input values and output is realized to two dimension spaces.

The SOM is a neural network trained with a competitive learning rule in an unsupervised manner. A competitive learning rule means that the neurons compete to respond to a stimulus, such as a connection vector (recall that a connection vector describes properties of a network connection, such as the destination port and number of packets sent). The neuron that is most excited by the stimulus, i.e. whose weight vector is most similar to the connection vector, wins the competition. The winning neuron earns the right to respond to that stimulus in future, and the learning rule adjusts its weight vector so that its response to that stimulus in future will be enhanced, i.e. by moving the weight vector closer to the connection vector. This means that the next time that same connection vector is presented, the neuron that won the competition for that same vector last time will be more excited by it. During training, the SOM learns to project connection vectors that are close together in terms of Euclidean distance onto neurons that are close to each in the output grid. In this way, the SOM learns relationships between the connections a vector, expressing them as spatial relationships in the output grid. The training algorithm also ensures that the weight vectors of the neurons area good representation of the connection vectors in the training data. This is achieved by aiming for a low mean quantisation error, where the quantisation error is the distance between a connection vector and the winning neuron ‘s weight vector. The mean quantisation error is the average of this over all connection vectors in the training set

K - MEANS

The K-means algorithm, starts with k arbitrary cluster centers in space, partitions the set of the given objects into k subsets based on a distance metric [17]. The centers of clusters are iteratively updated based on the optimization of an objective function. This method is one of the most popular clustering techniques, which are used widely, since it is easy to be implemented very efficiently with linear time complexity (Biswas, Shah, Tammi, & Chakraborty, 2016). The principle goal of employing the K Means clustering scheme is to separate the collection of normal and attack data that behave similarly into several partitions which is known as K^{th} cluster centroids. In other words, K-Means estimates a fixed number of K, the best cluster centroid representing data with similar behavior. The algorithm initially has empty set of clusters and updates it as proceeds. For each record it computes the Euclidean distance between it and each of the centroids of the clusters. The instance is placed in the cluster from which it has shortest distance. Assume we have fixed metric M, and constant cluster Width W. Let di (C, d) is the distance with metric M, cluster centroid C and instance d where centroid of cluster is the instance from feature vector

Fuzzy c-means (FCM)

Fuzzy c-means (FCM) is an improvement of K-means algorithm has become very important in the application of intrusion detection. In fuzzy C-means is a clustering method that calculates the membership function between each test data instance and each cluster [10], [20]. The test data instance is allocated to the cluster which has higher membership [15], [21]. In fuzzy C-means, the individual data point can belong to several clusters at the same time. Nevertheless, the degree of membership is determined by membership grades which are assigned to each data point. For each x_i in dataset D the fuzzy C-means algorithm assigns membership grade u_{ij} which shows the degree of x_i membership in cluster j ($0 \leq u_{ij} \leq 1$). The membership grades are calculated for each example based on the minimization of an objective function which measures the distance between each data point and the cluster centers. If m is the size of the input dataset and K is the number of clusters, this objective function is calculated as follows:

K membership value to each center. After that, it finds higher membership and assigns the instance to higher membership cluster. In other words, the instance in test dataset will divided into two clusters according to the degree of membership to C_1 and C_2 in this case. In the above equation q is the fuzziness exponent and can be any real value greater than 1 depending on the kind of problem. c_j is the center of j-th cluster and its dimensions are equal to that of input vector x_i. Creating the clusters is done through an iterative optimization process for objective function in which membership grades u_{ij} and cluster centers c_j are updated Once the Fuzzy C-means algorithm obtains the unlabeled dataset of magnitude m as input, it executes the above process and the output are two matrices: The Matrix U which consist of membership grades of each data example in each of the K clusters and matrix C which includes the cluster centers for K clusters then.

To create K disjoint subsets from the dataset based on matrix U, one subset for each individual example in the training dataset is determined based on its maximum membership grade i.e.

for each x_i: if $u_{iw} = \max \{u_{ij}\}$ then $x_i \in D_w$,
where $i = 1, 2, \ldots, m; j = 1, 2, \ldots, K$.

After calculating the subset for all examples, the training dataset is divided to K disjoint subsets D1, D2, . . ., DK. These K subsets are used to train classification techniques like ANN, SVM etc.

Expectation and Maximization Algorithm (EM)

The EM algorithm [22] [17]is a clustering technique in data mining and consists of two repeated steps, Expectation and Maximization. It is based on Gaussian finite mixtures model (GMM) for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables [23]. The EM algorithms alternates between performing an expectation (E) step, which computes the expectation of the log- likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. The model consists of a set of k probability distributions that represent the data of each cluster while the number of iteration and log likelihood difference between two iterations are parameters that defines each of the k distributions. Initially, the algorithm makes guesses for these parameters based on the input data, then determines the probability that a particular data instance belongs to a particular cluster for all data using these parameter guesses. The distribution parameters are revised again and this process is repeated until the resulting clusters have some level of overall cluster 'goodness' or until a maximum number of algorithm iterations are reached.

Mathematically, the algorithm attempts to find the parameters θ, that maximize the probability function, log P (x; θ) of the observed data. It reduces the difficult task of optimizing log P (x; θ) into a sequence of simpler optimization sub problems, whose objective functions have unique global maxima that can often be computed in closed form. These sub problems are chosen in a way that guarantees their corresponding solutions $\varphi^{(1)}$ $\varphi^{(2)}$ ... and will converge to a local optimum of log P (x; θ). The Expectation step (E-step) of the algorithm estimates the clusters of each data instance given the parameters of the finite mixture. During the E-step, the algorithm chooses a function $f(g_t)$, that lower bounds log P (x; θ) everywhere, and for which $f(\varphi^{(1)}) = \log P(x; \varphi^{(t)})$. The Maximization step (M-step) of the algorithm tries to maximize the likelihood of the distributions that make up the finite mixture, given the data. During the M-step, the algorithm moves to a new parameter set $\varphi^{(t+1)}$, that maximizes $f(g_t)$. As the value of the lower-bound $f(g_t)$ matches the objective function at $\varphi^{(t)}$, it follows (9), so the objective function monotonically increases during each of the iterations in EM.

$$\text{Log } P(x; \varphi^{(t)}) = g_t(\varphi^{(t)})\, g_t(\leq \varphi^{(t+1)}) = \log P(x; \varphi^{(t+1)}) \qquad \text{eqn 1}$$

Training data with the results of normalization and discretization techniques enter clustering step. The dataset will be divided into number of clusters in FCM, K-means, and EM to find the optimal results. Similarly, well test the SOM by simultaneously varying the epochs and lattice configuration. Two third of the dataset will be used for training and the rest is for testing.

The NSL-KDD Dataset

The simulated attacks in the NSL-KDD dataset fall in one of the following four categories [24].

i. Denial of service attack (Dos), where attempts are to shut down, suspend services of a network resource remotely making it unavailable to its intended users by overloading the server with too many requests to be handled. e.g. syn flooding. Relevant features include source bytes and percentage of packets with errors. Examples of attacks includes back, land, Neptune, pod, Smurf, teardrop

ii. Probe attacks, where the hacker scans the network of computers or DNS server to find valid IP, active ports, host operating system and known vulnerabilities with the aim discover useful information. Relevant features include duration of connection and source bytes. Examples includes IP sweep, n map, port sweep, Satan

iii. Remote-to-Local (R2L) attacks, where an attacker who does not have an account with the machine tries to gain local access to unauthorized information through sending packets to the victim machine in filtrates files from the machine or modifies in transit to the machine. Relevant features include number of file creations and number of shell prompts invoked. Attacks in this category includes ftp_ write, guess_ passwd, I map, multi hop, phf, spy, warezclient, warezmaster

iv. User-to-Root (U2R) attacks, where an attacker gains root access to the system using his normal user account to exploit vulnerabilities. Relevant features include Network level features – duration of connection and service requested and host level features - number of failed login attempts. Attacks includes buffer overflow, load module, Perl, rootkit

Experimentation, Results and Discussion

In implementation of the model, the researcher used WEKA Software. Three set of experiments were conducted and the results are tabulated in Table 1: In first experiment clustering with data preprocessing based on hybrid feature selection only (i.e., labeled as HFS), the second experiment clustering with PCA only (i.e., labeled as PCA), and the third experiment clustering with HFS and PCA (i.e., labeled as IPCA). The four measurements techniques applied are: (1) Clustering Error (CE) is the number of alerts that are wrongly clustered. (2) Error Rate (ER) is the percentage of wrongly clustered alerts, ER = (CE ÷ Total number of alerts observed) x 100, (3) Accuracy Rate (AR) is the percentage of alerts that are accurately clustered as they should be, AR = 100 – ER, and (4) Time is the algorithm processing time in seconds.

Table1: Clustering Performance based on Self-organizing maps (SOM), Expectation Maximization, K-means and Fuzzy c-means (FCM)

Mode	FCM				K Means				SOM				EM			
	CE	ER	AR	TI	CE	ER	AR	TI	CE	ER	AR	TI	CE	ER	AR	TI
HFS	74	17.5	82.6	1.3	57	13.4	86.6	4.4	135	31.8	68.2	4.2	45	10.6	89.4	1.9

PCA	133	31.3	68.6	3.6	141	33.3	66.2	5.2	170	40.1	60.0	6.5	86	20.3	79.7	2.7
IPCA	67	15.8	84.2	4.8	46	10.9	89.2	6.2	112	26.4	73.6	7.4	41	9.7	90.3	4.6

The number of clusters in FCM, K-Means, and EM were varied to find the optimal results. The SOM was tested by concurrently changing the epochs and lattice configuration. Two third of the dataset were used for training and the rest for testing. The optimum result on SOM (73.58%) was obtained after being trained for 2500 epochs using hexagonal 4 by 6 lattice type and produced 12 clusters. The SOM's best processing time both for training and testing was obtained after 7.4 seconds. Increasing or decreasing the processing changes the results if the dataset, epochs and lattice type are larger (Siraj et al., 2009c). The results of k-means clustering algorithm indicated that the performance depends on the number of clusters which are applied, and increasing or decreasing the cluster beyond the number of data types only lessens the efficiency of the model. Identifying the number of clusters therefore significantly changes to the results.

The research has to determine the number of clusters that are expected in advance in order to obtain good results. In this work several clusters were tested and the optimum results (89.2) were obtained at 22 clusters in a time of 6.2 seconds. However, the challenge of identifying the number of clusters in a dynamic network, is much more difficult since there is no base data to assist in deciding the number of clusters.[26] The best clustering algorithm was EM 90.3% and is arrived at 14 clusters in a time of 4.6 seconds. In respectively cluster, related alerts are clustered together and represent an attack step. The value of CE of FCM, K- Means, SOM and hybrid is larger, and hence a large number of alerts that belong together in one cluster are put into other different clusters. The result inferred that the proposed model based on hybrid feature selection, PCA and EM is effective in terms of clustering accuracy and processing time for this dataset.

Conclusion

The output is a hybrid machine learning approach for automated alert clustering and filtering based on EMFFS, Principal Component Analysis (PCA) and Expectation and maximization techniques that gives optimum results to aggregate similar alerts and to reduce the number of alerts compared to other unsupervised learning algorithms tested. The results are promising in terms of clustering accuracy rate (89.2) and processing time (6.2 sec). The model cannot reveal the memberships of attack stages like that of multi-stages attack which comprise of one/more attack steps.

REFERENCES

[1] L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 446–452, 2015.

[2] K. Kumar, "Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms," vol. 150, no. 12, pp. 1–13, 2016.

[3] S. T. Ikram and A. K. Cherukuri, "Improving Accuracy of Intrusion Detection Model Using PCA and Optimized SVM," vol. 24, no. 2, pp. 133–148, 2016.

[4] D. Perez, M. A. Astor, D. P. Abreu, and E. Scalise, "Intrusion Detection in Computer Networks Using

Hybrid Machine Learning Techniques," 2017.

[5] Y. Kumar, "AI based Hybrid Ensemble Technique for Network Security," no. Icaet, pp. 1–10, 2016.

[6] S. Upadhyay, "A Survey on IDS Alerts Classification Techniques," vol. 105, no. 12, pp. 27–33, 2014.

[7] M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Eng.*, vol. 30, no. 2011, pp. 1–9, 2012.

[8] A. I. Madbouly, A. M. Gody, and T. M. Barakat, "Relevant Feature Selection Model Using Data Mining for Intrusion Detection System," *Int. J. Eng. Trends Technol.*, vol. 9, no. 10, pp. 501–512, 2014.

[9] H. T. Elshoush and I. M. Osman, "An Improved Framework for Intrusion Alert," vol. I, pp. 4–9, 2012.

[10] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," *PLoS One*, vol. 11, no. 11, pp. 1–18, 2016.

[11] S. Chaurasia and A. Jain, "Ensemble Neural Network and K-NN Classifiers for Intrusion Detection," vol. 5, no. 2, pp. 2481–2485, 2014.

[12] C. Science, "A Hybrid Approach to improve the Anomaly Detection Rate Using Data Mining Techniques," no. July, 2015.

[13] F. Valeur, "Real-Time Intrusion Detection Alert Correlation," *Security*, no. June, p. 199, 2006.

[14] J. M. Chahira, "Model for Intrusion Detection Based on Hybrid Feature Selection Techniques," *Int. J. Comput. Appl. Technol. Res.*, vol. 09, no. 03, pp. 115–124, 2020.

[15] V. Sannady and P. Gupta, "Intrusion Detection Model in Data Mining Based on Ensemble Approach," pp. 1654–1658, 2016.

[16] A. Shrivastava, M. Baghel, and H. Gupta, "A Review of Intrusion Detection Technique by Soft Computing and Data Mining Approach," no. 3, pp. 224–228, 2013.

[17] M. M. Siraj, M. A. Maarof, and S. Z. M. Hashim, "Intelligent alert clustering model for network intrusion analysis," *Int. J. Adv. Soft Comput. its Appl.*, vol. 1, no. 1, pp. 33–48, 2009.

[18] B. Subba, S. Biswas, and S. Karmakar, "Enhancing effectiveness of intrusion detection systems : A hybrid approach."

[19] N. A. Biswas, F. M. Shah, W. M. Tammi, and S. Chakraborty, "FP-ANK: An improvised intrusion detection system with hybridization of neural network and K-means clustering over feature selection by PCA," *2015 18th Int. Conf. Comput. Inf. Technol. ICCIT 2015*, pp. 317–322, 2016.

[20] M. Amini, "Effective Intrusion Detection with a Neural Network Ensemble Using Fuzzy Clustering and Stacking Combination Method," vol. 1, no. 4, pp. 293–305, 2014.

[21] B. S. Harish and S. V. A. Kumar, "Anomaly based Intrusion Detection using Modified Fuzzy Clustering," vol. 4, pp. 54–59, 2017.

[22] M. M. Siraj, M. A. Maarof, and S. Z. M. Hashim, "A Hybrid Intelligent Approach for Automated Alert Clustering and Filtering in Intrusion Alert Analysis," *Int. J. Comput. Theory Eng.*, vol. 1, no. 5, pp. 539–545, 2009.

[23] Y. Wahba, E. ElSalamouny, and G. ElTaweel, "Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction," *Ijcsi*, vol. 12, no. 3, pp. 255–262, 2015.

[24] M. R. Parsaei, S. M. Rostami, and R. Javidan, "A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset," vol. 7, no. 6, pp. 20–25, 2016.

[25] M. M. Siraj, M. A. Maarof, and S. Z. M. Hashim, "Intelligent clustering with PCA and unsupervised

learning algorithm in intrusion alert correlation," *5th Int. Conf. Inf. Assur. Secur. IAS 2009*, vol. 1, pp. 679–682, 2009.

[26] S. Duque and M. Nizam, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," *Procedia - Procedia Comput. Sci.*, vol. 61, pp. 46–51, 2015.