ENHANCING THE USER STORIES IN AGILE USING BIG DATA METHODOLOGIES

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

SURINDER RANI

Registration number

11616310

Supervisor

RITIKA MAHAJAN

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November (2017)

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0320

Supervisor Name : Ritika Mahajan **UID :** 18370 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Surinder Rani	11616310	2016	K1637	9592879502

SPECIALIZATION AREA : Software Engineering **Supervisor Signature:** _____

PROPOSED TOPIC : Enhancing the User stories in agile using Big Data methodologies

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.43
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.86
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.14
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.14
5	Social Applicability: Project work intends to solve a practical problem.	6.57
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.57

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): Yes
PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Enhancing the User stories in agile using Big Data methodologies**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

11/27/2017 12:46:33 PM

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "ENHANCING THE USER STORIES IN AGILE USING BIG DATA METHODOLOGIES" in partial fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs. Ritika Mahajan. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Surinder Rani

R.No.11616310

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled "**ENHANCING THE USER STORIES IN AGILE USING BIG DATA METHODOLOGIES"**, submitted by **Surinder Rani** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Ritika Mahajan

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____
2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my supervisor, **Mrs. Ritika Mahajan**, Assistant Professor, Computer Science and Engineering Department, **Lovely Professional University**, **Phagwara**, for her invaluable help and guidance during the course of thesis. I am highly indebted to her for constantly encouraging me by giving her critics on my work. I am grateful to her for giving me the support and confidence that helped me a lot in carrying out the research work in the present form. And for me, it's an honor to work under her. I also take the opportunity to thank **Mr. Dalwinder Singh, HOD**, Computer Science and Engineering Department, **Lovely Professional University, Phagwara,** for providing us with the adequate infrastructure in carrying the research work. I would also like to thank my parents and friends for their inspiration and ever encouraging moral support, which went a long way in successful partial completion of my thesis. Above all, I would like to thank the almighty God for His blessings and for driving me with faith, hope and courage in the thinnest of the times.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The big data is the data which is large in quantity and attributes do not have any relation. The term “big data” is relatively new in IT and business. The Big data is a term used where the large volume of data is difficult to process, store and analyze by using traditional existing database technologies. As the nature of big data is indistinct so, there is need to involves considerable processes to identify and translate the data into new insights. There are number of definitions of big data some researchers also define big data as a large volume of scientific data for visualization. Other researchers define big data as “the amount of data just beyond technology's capability to store, manage, and process efficiently. The user story is the story which is designed at the time of effort estimation. The user story is used to estimate the efforts for the software development. The authors have proposed various techniques to improve user story which directly reduce MRE for the effort estimation. In this research work, technique will be designed to improve user story to reduce MRE value of the effort estimation

KEYWORDS

Big Data

User story

Effort estimation

MRE

CHAPTER 1
INTRODUCTION

1.1 Big Data

In today era term data is everywhere and there is need to store, process and manage it since from the beginning of human civilization and human societies. The amount and type of data has been stored, processed and managed depends on different factors such as necessity of human, available tools, technologies, effort, cost, ability to gain insight into the data. The term “big data” is relatively new in IT and business. The Big data is a term used where the large volume of data is difficult to process, store and analyze by using traditional existing database technologies. As the nature of big data is indistinct so, there is need to involves considerable processes to identify and translate the data into new insights. There are number of definitions of big data some researchers also define big data as a large volume of scientific data for visualization. Other researchers define big data as “the amount of data just beyond technology's capability to store, manage, and process efficiently [1].”

In past, human beings used carving on stones, metal sheets, wood, etc like primitive ways of storing and capturing the data. Then they have started capturing data on paper, cloth, etc then human have started using USB Drives, Compact Discs, Hard Drives, etc. Now data is present everywhere to deal with that huge amount of data a new concept has been added i.e., Big Data.

N statistical science a key challenge is Big Data due to which number of researchers has started working on it for getting better context by using different algorithms and implementing it for new framework theoretical implications [2]. The massive or large amount of data comes under Big Data that also contains streams of data that are heterogeneous to each other.

1.1.1 Characteristics of Big Data

- **Volume:** The size of Big Data is huge and value out of data is very much depends on size of data. The volume of data will depend that this data will be considered as Big Data or not. Hence when dealing with Big Data, volume is one characteristic which is need to be considered [4].

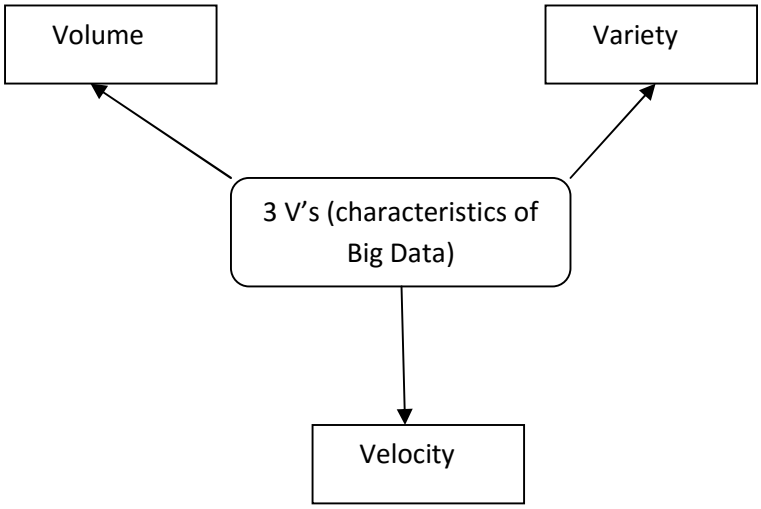

Fig. 1: Characteristics of Big Data

- **Variety:** Variety is the next aspect of Big Data that refers to heterogeneous sources and data nature for both structured and unstructured. In past, the source of data available was datasheets and spreadsheets that have been considered in almost all applications. Now data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc has been considered in different applications. In storage, mining and analysis of data some issues were created by variety of unstructured data.
- **Velocity:** The speed of generated data refers to velocity. In data real potential is determined using how fast data is generated and processed to meet demands.

1.1.2 Applications of Big data

Here are some examples of Big Data applications [5]:

- **Smart Grid case:** it is crucial to monitor and manage the smart grids operations and the national electronic power consumption in real time. To achieve the objective above mentioned there is need to make multiple connections among smart meters, sensors, control centers and other infrastructures. To detect the abnormal behaviors of the connected devices and to identify at-risk transformers Big Data analytics need to use. With the help of Big data Grid Utilities can choose the best treatment or action. The real-time analysis of the generated Big Data allow to model incident scenarios.

- **E-health:** To personalize health users are already using services connected health platforms (e.g., CISCO solution). Big Data is generated from different heterogeneous sources (e.g., laboratory and clinical data, patients symptoms uploaded from distant sensors, hospitals operations, and pharmaceutical data). There are number of beneficial applications for using advanced analysis of medical data sets. It enables to personalize health services (e.g., doctors can monitor online patients symptoms in order to adjust prescription); to adapt public health plans according to population symptoms, disease evolution and other parameters. To decrease the cost expenditure and optimize the operations of hospital the big data has been used.
- **Internet of Things (IoT):** IoT represents one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT are continuously evolving. Nowadays, there are various Big Data applications supporting for logistic enterprises. With the help of sensors, wireless adapters and GPS it becomes possible to detect the position of vehicles. Thus, such data driven applications enable companies not only to supervise and manage employees but also to optimize delivery routes.
- **Public utilities:** In complex water supply network the sensors have been placed in pipelines to monitor the flow of water. In order to detect illegal connections, leakages and controlling value remotely a real time monitoring systems has been implemented by Bangalore water supply and sewage board that has been disclosed in press report. This helps in ensuring an equitable water supply to each and every remote areas or city [6]. In order to fix leaked water pipes there is need to identify it in proper time that reduces the need for valve operators to continuously monitor it manually.
- **Transportation and logistics:** The RFID (Radiofrequency Identification) and GPS have been used by many public road transport companies to track buses and explore interesting data to improve their services. We are able to optimize bus routes and the frequency of trips by collecting data about the number of passengers using the buses in different routes. Various real-time systems has been implemented not only to provide passengers with recommendations but also to offer valuable information on when to expect the next bus which will take him to the desired destination. By predicting the demand about public or private networks the by using mining of Big Data helps in improving the travelling

business. Making predictions from such data is a complicated issue because it depends on several factors such as weekends, festivals, night train, starting or intermediate station. By using the machine learning algorithms, it is possible to mine and apply advanced analytics on past and new big data collection. In fact advanced analytics can ensure high accuracy of results regarding many issues.

- **Political services and government monitoring:** In order to analyze the sentiments of population and monitoring the political trends a United States and India government is using data mining. The personal interviews, social network communications and voter compositions like data sources has been combine in many applications. This helps in detecting local as well as national issues and the valuable resources or utilities has been used by governments after being optimize by Big Data systems. In case of large networks the flow of water can be monitored by placing sensors in the pipeline of water supply chains. The leakages, illegal connections and valves are controlled or monitored in real time by many countries sitting in remote area that ensure the continuous supply of water in different areas of city.

1.2 Testing:

In order to check that system will be able to satisfy all requirements it has been tested along with its all components. The actual requirements has been gathered and then compared in with the results gathered by testing to identify any gaps, errors or mismatch with the requirements. A testing is a analyzing software process that has been defined by ANSI/IEEE 1059 standard that helps in detecting required and existing conditions differences. The software items features have been evaluated that helps in detecting defects, errors or bugs.

The application and software program should meet all the requirements has been verified and validate by software testing.

- Meets the requirements of business and technical that guided its design and development
- Works as per the expectation

The important application code defects, flaws or errors are also identified and fixed by software testing. The requirements and design documents are reviewed in order to decide the important defect comes under test planning [7]. The defect that affects the usability or functionality of the

application is the most important defect. The purpose of software testing is to detect software failures so that defects may be discovered and corrected. This is a non-trivial pursuit. Testing establishes that in what conditions the product will not function properly but it cannot establish that it will work properly under all conditions [8].

1.2.1 Main purpose of Software testing: There are mainly three purposes of software testing given below:

- **Verification:** This confirms that technical specifications are meet using software. A description of functions in terms of measured value of output for particular input has been described by specification that comes under specific fixed conditions. In 3 seconds of submission it is possible to return the eight ordered fields of month against summary table of multi-month account. The data has been retrieved from simple given specification along the SQL query retrieving data line.
- **Validation:** The requirements of business have been meets using software has been confirmed by this process. One of the simple examples of business requirement will be information about customer manager of branch after choosing the name of office branch. In that window information summary and identification of manager has been presented for each customer manager. The summarized, formatted and displayed data details provision are other requirements.
- **Defect finding:** It is the difference between actual and expected answer. In phase of design, specification or coding development introduced fault are introduced by defects in source.

The different aspects of codes like does it will work in the way that it supposed to do and does it will fulfill the requirements has been examined by testing that code using software testing and then executing that code in different conditions or environments. The development and testing organization team is different in most of the current software development culture. The team member testing can be achieved using different rules and the process of developed software has been changed to correct it by derived information from software testing [9].

1.3 Types of testing: Testing can be mainly of two types given below:

Manual Testing: The testing of software is done manually without using any tool or script. The unexpected behavior or bug has been identified by testing software and end user role has been played by tester itself. The unit, integrating, system and user acceptance testing are the different stages of manual testing.

The testing completeness has been ensured by performing different test plans, test cases or test scenarios by tester. The errors are identified by exploring that software explored by testers that also includes exploratory testing.

Automation Testing: The products are tested using software on the basis of scripts written by tester and it is also known as Test automation. The manual process is automated and then test scenarios are again re-run in short time. It is a repeated process that don't need manpower to process it. In point of view of stress, load and performance the application has been tested using automation testing that is different from regression testing. In comparison to manual testing it is prove to be efficient in terms of accuracy, time, test coverage and money.

Further it can of many types given below:

- **Alpha Testing:** This is the most common used testing type in software industry and its objective is that before releasing product to market or users all the possible defects and issues should be identified [10]. This is carried out at the end of software development phase but before Beta testing i.e., it is conducted at developer end.
- **Acceptance Testing:** There are some requirements of business so a test has been performed by client to check the end to end system flow. The software will only be accepted by client when the functionality of all functions is same as expected. This comes under the last phase of the testing.
- **Ad-hoc Testing:** This testing is performed on the basis of ad-hoc that does not put any reference model to test case and not even use any plan or documentation. They check the defects and break in the application by executing a random functionality.
- **Accessibility Testing**: This is used to check that is software or application is accessible by disabled people or not. The deaf, color blindness, mentally disabled, blind, old age are comes under disabilities [11].

- **Beta Testing:** This is formal type of software testing and it is carried out at customer end. Before releasing any product to the market for its actual end user Beta testing has been performed in real environment that ensures any major failure in product.
- **Back-end Testing:** This is used to test the data stores in database by front end application. SQL Server, MySQL, and Oracle etc are the different types of database. It will test the table structure, data structure, scheme and etc. In this testers are directly connected to the database no GUI is involved and data is verified by running few queries on the database [12]. This testing will help in finding data loss, deadlock, data corruption etc types of issues.
- **Browser Compatibility Testing:** The testing team performed this type of testing and it is performed for web applications that ensure the working of software in different browser and operating system [13].
- **Backward Compatibility Testing:** This ensures the working compatibility of newly developed and update software with older version of environment.

1.4 How testing define in Big Data

The software products are not individually tested rather than that the data processing has been verified suing Big Data testing application. The testing of performance and functions becomes important part in case of testing Big Data.

The terabytes of data has been verified by QA engineers of Big data testing that uses commodity cluster and other supportive components. The processing is very fast that require a high level of testing skills and the processing can be of three types given below:

- Batch
- Real time
- Interactive

In Big data testing an important part is also played by quality of data. So, there is need to check the quality of data before testing the application and it comes under the part of database testing

[15]. The various characteristics like **conformity, accuracy, duplication, consistency, validity, data completeness,** etc are come under it.

Now days it testing a big data has become a very big challenge faced by organization due to lack of knowledge on what to test and how much data have to test. The difficulties have been faced by organization in defining the strategy of testing for structured and unstructured data validation. The Hadoop system has been used to processed Big Data in which first step is to load data into HDFS involves in extracting data from different source systems and then back loading them into HDFS. Then Map Reduce operations have been performed on it and final output results has been extracted from HDFS.

1.5 Software Effort Estimation

In industries it is very important to perform accurate estimation of efforts and there should not be overestimation of efforts nor underestimation. As the overestimation will lead to threaten the customers and under will leads to breakdown a project. In order to avoid it the researchers have started focusing on developing the accurate method for software effort estimation rather than using human expert judgment [16]. In system development life cycle an important role is played by software effort estimation as the software project success is affected in case a design of project has been estimated inaccurately. This method can be of two types in the first one a equations and mathematical models has been used to make a algorithm models that has been used in data set have enough to train a model. The second one is prediction system models it is used in the case have available training data set is not efficient to train an algorithm model.

There is need of reliable estimation for control and planning of proper project and the existing software industries doesn't estimate a project in proper way as they don't use estimation appropriately. So, there is need to focus on efforts that improve the situation otherwise user has to suffer a lot. In software engineering effort is used to denote measure of use of workforce and is defined as total time that takes members of a development team to perform a given task. The cost or total time required to produce software projects which is depend on the man-day, man-month, and man-year units [17]. The reason behind estimation is not fixed it vary as most frequent one is project approval.

The problems faced by project designers in controlling and managing software projects are overrun of effort estimate. With inaccurate effort estimates, it surely affects project designers to make correct decisions and leading to the failure of the entire software project development.

1.5.1 User stories:

When a wanted feature is written on a card, expressed in everyday language and written in some case are comes under user story. It is composed of three aspects given below:

- While planning iteration a written description of story has been used.
- All the information that is written on the cards are uncompleted which are further completed by discussion about details that can take place number of times during the project.
- The agreement on the deliverance by team and customer is also very important that's why an acceptance test has been used after the completion of story.

The name given to these three aspects are Card, Conversation and Confirmation and these stories can be either is written on a paper or by software. The use of paper for writing purpose is prove to be more advantageous as it is simple, interactive and discussion are encouraged using it that further can be placed, stored and carried around [18]. The main advantage of using paper for writing purpose is its low tech nature that works as reminder which shows stories are imprecise.

It is important to perform estimation on user stories as cost of developing a user story has been estimated and it is important to estimate each one accurately. The estimations of all the features in one time is not necessary the only thing important is to estimate each new features that can further be selected for inclusion in future. The time is allocated for incorporating all the desired features that is not sufficient for it so there is need to prioritize the user stories development. The financial value of the story, the cost involved in developing the Story, the amount and significance of the knowledge are the factors that are utilized while putting priorities to user stories.

1.6 Techniques used to improve user story in Software development

1.6.1 INVEST Grid: This method was used [19] in Agile requirement management. There are some requirements of it that has been represented by the acronym INVES. The user story has

been met using the criteria of INVEST Grid for this each user story has been analyzed in order to determine if it is independent from each other that helps in moving to another Sprint without interfacing software deployment. In case when a high level of authority is given to client in order to negotiate a change in user story before it becomes a full requirement. The aim is that a provider should allow the change but getting stuck into a problem or running out of time constraints. The first aim of this approach is to introduce threshold level and the user story considered to be independent when there is no need to split it into more sub requirements. This helps in reducing the subjectivity and this approach has been initially developed to evaluate a set of measures and KPI in a business process model. The thresholds can be referred to as a profile.

1.6.2 An Enhanced XP Process Model: The existing model was not efficient that's why a new enhanced XP process model has been proposed. The main phases of XP model are:

- Planning
- Design
- Coding
- Testing

The Project Planning, Analysis or Risk Management, Design or Development and Testing are the main phases of adaptive model. This method is mainly implemented in medium and large scale a project that needs to be continuously evaluated due to changes occurs in requirements of customers [20]. By communicating to the customer during project planning phase a specification of the project has been documented which is composed of feasibility report that is created to prepare a cost benefits analysis (CBA) sheet. The technical, economic and operational feasibilities combined to report feasibility based on request of client.

CHAPTER 2

REVIEW OF LITERATURE

Usharani.K, et.al, (2016), has recommended the use of software effort estimation as number of prediction methods and different algorithms are available in it. In software engineering an important role is played by it that's why number of researchers is working on it to get the improved effort estimation technique. In order to access the efforts an expert judgment are required by industries and further the existing technologies has been used for experts. All the research shows that the existing prediction methods and algorithms are not perfect to be use for software effort estimation. In this paper [21], authors did a survey of total fifteen papers based on software effort estimation prediction and algorithmic methods from different journals. The highly predictive attributes of data set has been considered for estimation of accurate efforts. The backward input selection approach has been used to identify the above mentioned attributes. The authors have concluded different concepts and describe different directions where it can be improved. The existing methods accuracy can be improved using procedures of data processing and different methods can be used in order to evaluate software effort estimation.

Sivakumar D, et.al, (2017), have utilized the COCOMO II post architecture parameters for effort measurement of software project whose representation will not be able to accomplish the obligatory level of exactness. In this paper [22], authors have used a genetic algorithm for improving the measured results precision using above mentioned model. The prediction accuracy has been improved using that methodology that also reduces different uncertainty in used model coefficients A, B, C and D. The model accuracy has been tested in terms of different parameters name as MRE and MMRE and value of change rate is 0.15. The average relative estimation or the best fitness value 3.79 is measured in this analysis and this result is 1.42 smaller than the average relative estimation of 5.99 obtained with the help of the elderly coefficients. These results show that the estimation by using the finest coefficient is much better than the same estimation performed with existing coefficients.

Jianglin Huang, et.al, (2017), has presented a method name as analogy based software that helps in estimating the unseen project cost. This is based on analogies against previous sharing selection features. There are different factors on which validity of the selected features depend

upon and effectiveness is taken as a most crucial factors of different applied data processing techniques on data sets. In this paper [23], authors have proposed a first controlled experiment that helps in studying the class of three-stage data-preprocessing techniques. There are different stages such as data normalization, feature selection for analogy-based effort estimation and missing data imputation are different stages that has been used in above mentioned technique. The ISBSG data has been used for performing investigation on it and its experimental results show that the proposed technique prove to efficient in terms of resultant effort estimation accuracy. In analogy based effort estimation it has prove to effective to use a Z-Score normalization, kNN imputation and mutual information based feature weighting.

Kazunori Iwata, et.al, (2016), have discussed different effects in estimating the amount of efforts by classification that is directly associated with code development. It is very important to estimate the efforts required for new software projects that can also be identified after the completion of project. So, it should be considered while making a model and estimating the required efforts. In this paper [24], authors have presented an embedded software development projects classifications with combination of support vector machine and artificial neural network (ANN). A form of support vector machine, ANN, linear regression has been used to create a effort estimation models after the use of classification. The existing model estimation accuracy has been computed by performing different experiments on it with two criteria one with including classification and one without it. In order to consider statistically significant evidence a test has been performed name as Howell along with variance one way analysis. The results indicate statistically significant differences between certain pairs of models.

Sarwosri, et.al, (2016), have recommended the use of Use Case Point (UCP) in software project estimation that is estimated by performing some calculation son total number of worker prediction based on effort estimation and time required for software development. In 1993, Karner has introduced a UCP that now widely use due to its effectiveness compared to existing approaches. In this paper [25], authors have used UCP and presented different methods to minimize the difference between actual effort and effort estimation by increasing Technical Complexity Factor (TCF) and Environmental Complexity Factor (ECF) in estimation. Then they have used two ways to consider those factors the first one is to add drive cost factor by mapping TFC and ECF in COCOMO II. In second the developing software has been considered for

conducting qualitative research by understanding the problem deeply. The 6.19% estimation deviation has been achived using old UCP on project1 and it is 39.32& on project2. The results for the same parameter using new UCP approach is 5.02% on project1 and 7.94% on project 2. In project1 there is large decrease in deviation and it increases by decrease for smaller scale. In small scale projects the calculation involvement of TF and EF will be effective and it has also been seen that it is better to implement a new UCP in project effort estimation of software development.

Shensi Tong, et.al, (2016), have analyzed that during software development it very challenging and vital to performing effort estimation of software. The past storage data is the factor that creates challenge in number of small and medium sized companies that can be solved averaging the data of cross company for effort estimation that is not an easy task. In this paper [26], authors have proposed a Mixture of Canonical Correlation Analysis and Restricted Boltzmann Machines (MCR) named approach that addresses the issues of data heterogeneity in effort estimation of cross company. In order to represent heterogeneous effort estimation data to present a unified metric is one of the essential ideas in MCR. In effort estimation of heterogeneous cross company there is need to combine restricted Boltzmann Machines method and Canonical Correlation Analysis. In PROMISE repository total 5 public datasets are use to evaluate MCR approach and its evaluation results show that the MCR approach performs better than KNN in terms of estimating it with partial different metrics. The MMRE value is decreased by 0.60 and PRED value is increased by 0.16, in case of MdMRE it decreases by 0.19.

Lixiao Zhou, et.al, (2017), have recommended astronomy as a first area of science to learn from big data and the amount of data is getting increase day by day. In today, most of the research on data requires expensive telescopes and their key basis is data quality. It is very important that software testers will understand big data is about far more than simply data volume. In this paper [27], authors have analyzed characteristics, types, methods, problems, challenges and proposed a new possible software testing solutions for astronomical big data . There is need to explore some novel related principle of systematical methodology that helps in finding solutions to solve test management and efficiency in software testing. The full time and professional software testing engineers are needed to introduce step by step that should pay attention in order to grasp knowledge in astronomy. Software testing engineers should learn to cooperation with astronomer

and data scientist. Professional software testing organizations and normalization in astronomy should be built, especially for astronomical big data and the maturity should be improved on schedule. It has been concluded that it is a valuable and long-time-lasting work to study how to evolve on the base of combining software testing technology with big data, cloud computation and artificial intelligent in astronomy.

Mr. Kunal Sharma, et.al, (2016), have analyzed that (ETL) is a process of data warehousing to transfer data from basis database by performed some alteration policy on extract information. Then on target base that transformed data has been loaded back and with necessity of Bid Data different organization has started moving toward big data technology. They have started using Hadoop, Hive and HBase to store their information In order to migrate data from RDBMS to Hadoop, organization has started using process of data migration and then these data has been utilized to perform various analyses. There is issue of discrepancy in the task of data migration that occurs due to different reasons that results in inaccurate analysis of data. In this paper [28], authors have given a review on comprehensive facts legalization framework between RDBMS and Hadoop. Then they have planned general testing framework for information justification after ETL method. new product obtain during testing of framework on special data set shows to the point obligatory to test data increases linearly with boost in number of account. The main restraint of the framework is that present realization does not have burn test kind of ability where in tester can observe how greatly information get varied after movement process.

Krishna Kumar Singh, et.al, (2016), has analyzed that from last two decades scientist are working on Big Data and its analytics. The researchers have gained lost of advancement in all verticals of Big Data but very less has worked on cloud based software testing in it. The 5 V's are the base of Big Data analytics and sometimes unwanted data has been generated by analytical results for financial forecasting. In analytical scenario an important role is played by Cloud computing that supports everything as service like IAAS, SAAS, PAAS, TAAS etc. So, in this environment there is need of testing as well as validation tool in the same environment. Testing as a Service (TaaS) is being offered by many players through cloud. Dearth literature availability and wide application of testing tools in financial market cloud computing Big Data prompted us to work on the area of cloud based automated validation and testing tool model. In this paper [29], authors have addressed the real challenges of online cloud based automated testing not a as

a service (TaaS). They have also introduced a new model that includes mandatory tool applicable in the financial market computing. This has been analyzed bu testing that the proposed model is applicable during testing and validation of the desired data for financial forecasting.

Adiba Abidin, et.al, (2016), has presented that now days large statistics is a big matter of conversation. Their uses have been seen in paper to practical magazine from public media to journal. The word large statistics refers to intricate facts sets whose mass is beyond the skill of conventional processing technique within a pet period of instance. Large statistics consists of huge size that might be peta bytes or Exabyte's of statistics consisting of billions to trillions of account of millions of public. Testing of vast amount of facts is a huge dare. With the appearance of public medium, cloud and smart phones industry have to contract with the huge amount of statistics. While full-size statistics provide solution to compound business troubles like analysis of huge data serves as a starting point for quicker and better choice make, new products and services are being progress for the clients. [30], authors have alert on the different techniques that have been implemented. The most vital object that tester has to stay in mind is the active scenery of the statistics and other different act block issue linked with large information. The planned technique can be used for testing large data.

Jérôme Lacaille, et.al, (2016), have recommended the use of Snecma's test benches for testing new engine configuration or engine parts. Through each test awake to two thousand sensors detain all bit of data generate by the train or the worktable cell itself. It is very hard to manually examine all this information. Numeric account are analyzed and coded as series of labels each representing a diverse fleeting or stabilize phase. The labels are issue from categorization of local arithmetic model parameter with (and without) their first situation and are store in a circulated file allowing similar search by classic map/reduce format. Then it becomes a set easier to gaze for a specific guide in a set of tens of years of numerical account. Parallel distance is built to balance labels or brand sequences [31]. Their road-map is increasingly increasing the distributed database of labels and topics (with links to original documents and numeric records). The first pace was to recognize the diverse phases extracted from little subsets of sequential capacity and build local models for known patterns. In parallel, our facts base is classified into topics and a example of query system is implement. This incremental procedure allows us to

construct our database increasingly, adding novel patterns when expert are asking for them with no any require for a revamp of the scheme.

Dawit G. Tesfagiorgish, et.al, (2015), have analyzed that during transformation of this much large volume of data, some cases of data disparity, blunder and/or loss of positive facts will take place that leads to an ineffective facts alteration. Testing plays an important role in those cases to check occurrence of such possible errors. The obtainable class testing methods are also variable, revisit unfair results, fail to give answers pro facts difference or have some confines which do not treat every and every piece of the facts into the course. [32], authors have future a novel loom of large facts transformation testing to is based on the idea of data reverse engineering. It is a full approach that reverses the whole change process and does a link testing on both and every door of the facts if the new source facts can be construct back from the end data, once successful ETL course is ended. Despite ETL's intricacy processes they also need to be confirmed and validated to make sure that a winning and precise execution is done that yields a concrete and robust data system. The new proposed system for big data alteration validation testing and facts quality pledge will contribute to conquer the challenge faces in statistics accuracy and facts semantics fluctuation of data relocation process.

Zakia Asad, et.al, (2015), has concluded that pressure is increased on the data centers network due to movement of massive volume of data in cloud. So, in this paper the authors have used the mixing technique, spate coding along with software defined network control to propose a new scheme to dynamically reduce the volume of communication. For this purpose they have introduced a novel spate coding algorithm which helps in achieving the real world use cases for networks of data centers. Further by performing a proof of concept implementation on the proposed system they continuously bridge the gap between theory and practice. The experiment results of proposed coding based scheme is compared with vanilla Hadoop implementation, Combiner-N-Code and an in network combiner and shows the better performance. The results are compared in terms of communication volume which is up to 62% better than existing schemes, in terms of good put it is improved by 76%, disk utilization is by 38% and in terms of number of bits that can be transmitted per joule of energy is up to 200%. The results show that the proposed scheme is advantageous from the existing techniques in terms of different parameters [33].

Zakia Asad, et.al, (2015), has proposed a network coding technique CodHoop employ by system for the same purpose mentioned in previous paper. In this paper, authors have used a network middle box service and specifically index coding for controlling the dynamically reduction in communication volume. Further they have presented the motivating use case for this class of applications and used Hadoop as a representative. The results of the proposed scheme are compared with the Hadoop in terms of number of parameters. A result shows that the proposed scheme is in average 31% better than use case translates depending vanilla Hadoop implementation. This shows that there is 31% less utilization of equipment energy in Hadoop scheme and in proposed scheme 31% jobs can run simultaneously or can say job completion time is reduced by 31%. The coding based scheme used in this requires a XORing of packets whose operations are computationally very fast. In this case the given memory has larger bandwidth by which authors are able to process closer to link rate. Even in the worst case this coder has 809 Mbps of throughput on a 1 Gbps link [34].

Xuelian Lin, et.al, (2012), has recommended that for job analysis and optimization of MapReduce the accurate performance model is required. The numbers of steps are needed to be perform in case MapReduce that make it a challenging task. In MapReduce the number of steps is directly proportional to complexity, with increase in number of steps complexity increased at steady rate. In this paper to measure the MapReduce task complexity, authors have used a new concept which helps in analyzing the detail composition. The concept of SP, CEF and RCC is also defined in this paper to accurately measure the cost of Map or reduce function. To calculate the cost of each item they have decomposed the major cost items and make a new cost model based on vector, equation. The result of model is verified on a several clusters of Hadoop and it shows the effectiveness of proposed model. In this proposed model, authors have not considered the combine operation and serialization cost. By improving the proposed scheme the results can be improve in terms of serialization. In case of resource contentions in the cluster, proposed model will not be able to accurately predict the execution time of task [35].

Chang Liu, et.al, (2013), has recommended that for data intensive computation in application of big data, a low cost and high efficiency can be achieved by an environment of cloud computing. The cloud computing is cost effective and very flexible but it will restrict the control of user on their own data which results in data security problem. In this paper, the authors have proposed a

Cloud Background Hierarchical Key Exchange (CBHKE) novel hierarchical scheme. This key exchange scheme will help in achieving the secure and efficient scheduling for cloud computing environment. They have designed a layer by layer iterative key exchange strategy to achieve a more efficient Authentication Key Exchange (AKE) even without compromising the data security. The experimental and theoretical results of proposed scheme Cloud Background Hierarchical Key Exchange (CCBKE) and Internet Key Exchange (IKE) is superior in terms of efficiency. The proposed scheme CBHKE key exchange scheme help in improving the efficiency but at the same time they become slow in case of large datasets [36].

CHAPTER 3

PROBLEM DEFINITION

Previously in companies many customers were calling and they demands of requirements these calls record could not be kept because data was very large to store or tackle but with big data every user detail is kept safe because it is very important for companies. They can analysis it and use it in future to grow their business or marketing etc. so every data related to customers or users is very important for companies and with the invention of big data technique it became easy for us to store large amount of data that otherwise was not possible in data warehouses or data mining. At first if we want to see the detail of some customer or user we have to go through every detail of the person but with big data we can access the detail very easily of about our requirements. The data which is used is of big data. The hadoop is used for the implementation. The user story is the designed at the time of information gathering to develop the software. The efficient design of the user story can improve the software effort estimation. In this research, work technique will be designed which can identify the functions which are required and which are not required in the software. This leads to reduction software development cost and time .

The big data is the type of data which store any type of information and remove the constraints of type data and relational data bases. The data which is stored in the database is highly sensitive due to which various attacks need to breach security of the database. The big data has very dynamic nature due to its non relational nature. The cloud computing has the architecture in which virtual servers are involved to store the data. The data which is stored on the virtual servers is the big data.

The Map reduce is the technique will is applied to analyze, perform operations on the big data. In the Map reduce, the HDFC file system is used which is hierarchal distributed file system. In this file system, the data is divided into small chucks and each chuck is treated individually. The chuck is assigned to each virtual machine on which different process is executed. In this research work, the HDFC file system will be used to analyze cloud data, the data is divided in such a way that load in the network can be balanced. The proposed improvement leads to increase efficiency in terms of various parameters.

The big data is the type data which does not have any relation with the other type of data. In this research work, the map reduce technique is applied which can test the big data applications. The performance of testing techniques can be analyzed in terms of fault detection rate and execution time. Which is increased the user stories requirements efficiently

CHAPTER 4

SCOPE OF STUDY

The big data is the type data which does not have any relation with the other type of data. The big data has various techniques to analyze the input data. The map reduce is one of the techniques to analyze the big data. The map is the transformation function and reduce is the data aggregation function. The testing is the technique of software engineering which can test the application and detect faults from the application. In this research work, the map reduce technique is applied which can test the big data applications. The technique will be designed to improve the user story for the reduction in software development cost and time.

In this research work, the map reduce technique is applied which can test the big data applications. The performance of testing techniques can be analyzed in terms of fault detection rate and execution time. It is a cost reduction, faster and better decision making technique, new product and services, perform risk analysis and makes or provide better customer or user experiences.

Hadoop is a framework that allows us to store data and process large data sets in parallel and distributed fashion in which HDFS (storage) allows us to dump any kind of data across the cluster and MapReduce (processing) allows parallel processing of the data stored in HDFS. It can be used to reduce the storage space by detecting the faulty functions and the user stories could be enhanced. It could be used by big enterprises as big data will be used as a sensor that detects the faulty function from the required and non required functions because the data set that we will use will be of big data in this research work.

CHAPTER 5

OBJECTIVES OF STUDY

Following are the various objectives of this research work:

1. To study and analyze various big data testing techniques in data mining.

2. To propose improvement in map reduces technique to improve user story for software development.

3. The proposed improvement will be based on the classification technique to identify required and non-required functions.

4. This will help to manage the productivity to many folds.

5. This will focus on the enhancement of the agile product backlog to the efficient management of the project.

6. To propose improvement in map reduces technique for the fault detection in the big data application.

7. The proposed improvement will be based on automated slicing for the defect prediction in the software.

8. To remove the faulty functions after detecting them so that storage space could be reduced.

9. To enhance the user stories after detection and removal of the faulty functions.

10. To enlighten the Enterprise by providing the good and user friendly interface between the users and the employees.

11. To minimize or reduce the communication gap between the customers and the software developers.

12. To be helpful to keep or maintain the record of customers demands and requirements.

CHAPTER 6

RESEARCH METHODOLOGY

This research is based on detecting faults from the big data applications. In the base paper, the map reduce technique is applied which can transform and aggregate data to detect faults from the application. In this work, technique of automated slicing will be applied which will traverse functional association and generate percentage of fault prediction. To implement proposed technique, SVM classifier will be applied which can classify the faulty and non-faulty functions. It is the main objective of SVM to determine the best function by maximizing the margin between the two classes. This is due to the fact that there are many such linear hyperplanes. The amount of space or distance amongst two classes is known as hyperplane. The shortest between the closes data points to a point on the hyperplane is known as margin. This can further help us in defining the way to extend the margin which can help in selecting only a few hyperplanes for the solution to SVM even when so many hyperplanes are available. . In this research work the functions which required and which are not required are identify. The functions which are not required will be removed from the projects due to which its storage will be increased. The data which is used is of large amount. The hadoop is used for the implementation .Hadoop is a very huge amount of distributed class processing infrastructure which is divided the large amount of data into small pieces of information which user may choose this information as a requirements . The gap of traditional approaches will dilute out.

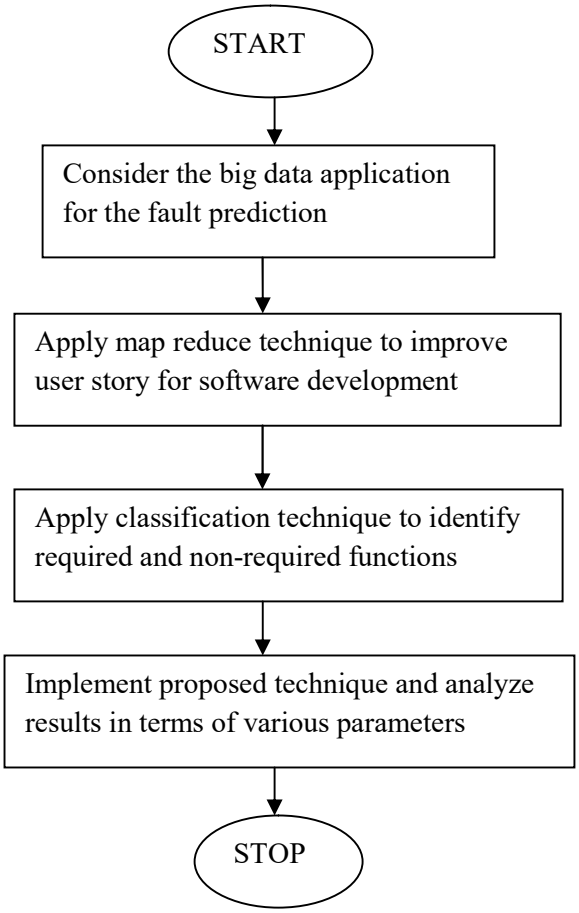

Figure: 2 Proposed Flow chart

Chapter Summary:

In this chapter, the problem formulated within the base paper is explained in the first section. Further, the various objectives to be achieved in this research are defined in the next section. Towards the end, the research methodology which defines the novel approach proposed in this paper is presented along with the proposed flowchart. The flowchart shows the step-wise procedure of proposed algorithm.

CHAPTER 7

EXPECTED OUTCOMES

Following are the various expected outcomes of this research:

1. The map reduce algorithm will be improved to design the user story of the software.

2. The proposed improvement can be compared with the other big data application testing techniques which analyze reliability of the model.

3. The proposed algorithm can reduce development time and cost which improve reliability of the modal.

4. The map reduce algorithm will be improved for the fault detection using automated slicing. This directly increases fault detection value.

5. The user stories record could be kept and maintained very well for the future guidelines that will be a mark to improve the work.

6. It will provide high performance data services.

7. It will provide a platform that could be used for read and right files easily with HDFS and MapReduce.

8. The input from the users could be mapped and traced and output could be generated by aggregating the both.

9 It will be beneficial to use this purposed technique as it could enhance the user friendly environment and interactive interface.

10. This technique could be used by various enterprises to analyze the user's interest and a produce a software product just according to them.

11. This will help to raise the output of the production on organizational level.

CHAPTER 8

SUMMARY AND CONCLUSION

In this research work, it has been concluded that testing is the technique which can test the software applications to predict software defects. The big data is the type of data which does not have relationship between each other. In Big data testing an important part is also played by quality of data. So, there is need to check the quality of data before testing the application and it comes under the part of database testing. The various characteristics like **conformity, accuracy, duplication, consistency, validity, data completeness,** etc are come under it. Now days it testing a big data has become a very big challenge faced by organization due to lack of knowledge on what to test and how much data have to test. The difficulties has been faced by organization in defining the strategy of testing for structured and unstructured data validation. The Hadoop system has been used to processed Big Data in which first step is to load data into HDFS involves in extracting data from different source systems and then back loading them into HDFS.The map reduce is the technique which is applied in the base paper to predict software development time and cost from the big data application. The proposed technique will be based on the classification which classify required and non required functions which reduce software development cost and time .

REFERENCES

[1] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan, "The rise of "big data" on cloud computing: Review and open research issues", Elsevier Information Systems, vol.47, pp.98–115, 2015.

[2] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Springer International Publishing Switzerland 2014, vol.21, pp. 214–227, 2014.

[3] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences, vol.27,pp.1-18, 2017.

[4] A.G. Picciano, "The Evolution of Big Data and Learning Analytics in American Higher", Education Journal of Asynchronous Learning Networks, vol. 3, pp. 9-20, 2012.

[5] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences, vol.27,pp.1-18, 2017.

[6] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE 2013 46th Hawaii International Conference on System Sciences, vol.13, pp.995-1004, 2013.

[7] John E. Bentley, Wachovia Bank, Charlotte NC, "Software Testing Fundamentals—Concepts, Roles, and Terminology, SUGI 30 Planning, Development and Support, vol. 3, pp. 130- 141, 2003.

[8] Kaner, Cem, Falk, Jack, Nguyen, Hung Quoc, "Testing Computer Software", 2nd Ed.. New York, John Wiley and Sons, vol. 1, pp. 1-480, 1999.

[9] Kolawa, Adam; Huizinga, Dorota, "Automated Defect Prevention: Best Practices in Software Management",. Wiley-IEEE Computer Society Press, vol. 3, pp. 41–43, 2007.

[10] Kolawa, Adam; Huizinga, Dorota (2007). Automated Defect Prevention: Best Practices in Software Management43. Wiley-IEEE Computer Society Press. p. 86. ISBN44 047004212545 . 46 .

[11] Section 1.1.2, Certified Tester Foundation Level Syllabus47, International Software Testing Qualifications Board

[12] Kaner, Cem48; James Bach, Bret Pettichord, "Lessons Learned in Software Testing: A Context-Driven Approach", John Wiley & Sons, vol. 2, pp. 4-10, 2001.

[13] McConnell, Steve, "Code Complete (2nd ed.)", Microsoft Press, vol. 2, pp. 950- 960, 2004.

[14] M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem, and P. Helland, "The end of an architectural era: (it's time for a complete rewrite)," in VLDB, vol. 5, pp. 1150–1160, 2003.

[15] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," OSDI, vol. 3, pp. 137–150, 2004.

[16] Simon WU Iok Kuan, "Factors on software effort estimation", International Journal of Software Engineering & Applications (IJSEA), vol. 8, pp. 23-32, 2017.

[17] Jovan Živadinović, Jovan Živadinović, Dragan Maksimović, Aleksandar Damnjanović, Slañana Vujčić, "Methods of effort estimation in software engineering", I International Symposium Engineering Management And Competitiveness 2011 (EMC2011), vol. 7, pp. 417-422, 2011.

[18] Evita Coelho, Anirban Basu, "Effort Estimation in Agile Software Development using Story Points", International Journal of Applied Information Systems (IJAIS), vol. 3, pp. 7-10, 2012.

[19]Luigi Buglione, Alain Abran, "Improving the User Story Agile Technique Using the INVEST Criteria", 2013 Joint Conference of the 23nd International Workshop on Software Measurement (IWSM) and the Eighth International Conference on Software Process and Product Measurement (Mensura), vol. 9, pp. 49-53, 2013.

[20] M. R. J. Qureshi, “An Evaluation of the Improved XP Software Development Process Model”, Sci.Int.(Lahore), vol. 2, pp. 79-82, 2008.

[21] Usharani.K, Vignaraj Ananth.V, Velmurugan.D, “A Survey on Software Effort Estimation”, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), vol. 23, pp. 505-509, 2016.

[22] Sivakumar D, Sureshkumar C, “Effort Estimation of Software Projects With Optimized Coefficients Using Soft Computing Technique”, Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2017), vol. 15, pp. 84-89, 2017.

[23] Jianglin Huang, Yan-Fu Li, Jacky Wai Keung, Y. T. Yu, W. K. Chan, “An Empirical Analysis of Three-stage Data-Preprocessing for Analogy-based Software Effort Estimation on the ISBSG Data”, 2017 IEEE International Conference on Software Quality, Reliability and Security, vol. 21, pp. 442-449, 2017.

[24] Kazunori Iwata, Toyoshiro Nakashima, Yoshiyuki Anan, Naohiro Ishii, “Effort Estimation for Embedded Software Development Projects by Combining Machine Learning with Classification”, 2016 4th Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science & Engineering, vol. 23, pp. 265-270, 2016.

[25] Sarwosri, Muhammad Jabir Al Haiyan, Mujahid Husein, Aditya Putra Ferza, “The Development of Method of The Enhancement of Technical Factor (TF) and Environmental Factor (EF) to The Use Case Point (UCP) to Calculate The Estimation of Software’s Effort”, 2016 International Conference on Information, Communication Technology and System (ICTS), vol. 23, pp. 203-207, 2016.

[26] Shensi Tong, Qing He, Yuting Chen, Ye Yang, Beijun Shen, “Heterogeneous Cross-Company Effort Estimation through Transfer Learning”, 2016 23rd Asia-Pacific Software Engineering Conference, vol. 14, pp. 169-176, 2016.

[27] Lixiao Zhou, Maohai Huang, “Challenges of software testing for astronomical big data”, 2017 IEEE 6th International Congress on Big Data, vol. 8, pp. 529-532, 2017.

[28] Mr. Kunal Sharma, Dr. Vahida Attar, "Generalized Big Data Test Framework for ETL Migration", 2016 International Conference on Computing, Analytics and Security Trends (CAST), vol. 9, pp. 528-532, 2016.

[29] Krishna Kumar Singh, Dr. Priti Dimd, Sachin Rohatgi, "Cloud Testing and Authentication Model in Financial Market Big Data Analytics", IEEE 5th International Conference on System Modeling & Advancement in Research Trends, vol. 13, pp. 242-245, 2016.

[30] Adiba Abidin, Divya Lal, Naveen Garg, Vikas Deep, "Comparative Analysis on Techniques for Big Data Testing", 2016 IEEE international conference on information technology, vol. 6, pp. 219-223, 2016.

[31] Jérôme Lacaille, William Bense, Ion & Stephan Berechet, Cynthia Faure, "Indexation of Numeric Bench Test Records A Big Data Vision", Aerospace Conference, 2016 IEEE, vol. 6, pp. 121-127, 2016.

[32] Dawit G. Tesfagiorgish, Li JunYi, "Big Data Transformation Testing based on Data Reverse Engineering", UIC-ATC-ScalCom-CBDCom-IoP 2015, vol. 7, pp. 649-652, 2015.

[33] Zakia Asad, Mohammad Asad Rehman Chaudhry, David Malone, "Greener Data Exchange in the Cloud: A Coding Based Optimization for Big Data Processing", IEEE Journal on Selected Areas in Communications, vol. 5, pp.1-18, 2015.

[34] Zakia Asad, M. Asad Rehman Chaudhry, D. Malone, "Codhoop: A system for optimizing big data processing", in IEEE International Systems Conference (SysCon), 2015, pp. 295–300.

[35] Xuelian Lin, Zide Meng, Chuan Xu, Meng Wang, "A practical performance model for hadoop mapreduce", in IEEE CLUSTER Workshops, vol.4 pp. 231– 239, 2012.

[36] Chang Liu, Xuyun Zhang, Chengfei Liu, Yun Yang, Rajiv Ranjan, Dimitrios Georgakopoulos, Jinjun Chen, "An Iterative Hierarchical Key Exchange Scheme for Secure Scheduling of Big Data Applications in Cloud Computing", 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, vol. 4, pp. 9-15, 2013.

APPENDIX

USB Universal Serial Bus

PDF Portable Document Format

RFID Radiofrequency Identification

GPS Global Positioning System

ETL Extract Transform and Loading

HDFS Hadoop Distributed File System

IOT Internet of Things

GUI Graphical User Interface

IAAS Internet as a Services

SAAS Software as a Services

PAAS Platform as a Services

TAAS Testing As a Services