International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 8, pp 65-70, August-2021

https://doi.org/10.47001/IRJIET/2021.508011

Development of a Knowledge Discovery System in Big Data Mining Environment

[1]Ihekeremma A. U. Ejimofor, [2]Obi O.R. Okonkwo

[1,2]Department of Computer Science, NnamdiAzikiwe University, Awka, Nigeria

Corresponding Author Email: iaejims2@yahoo.com

***Abstract* - This paper focused on the development of knowledge discovery system in big data mining environment. In order to carry out the aim of the work, the paper developed a knowledge discovery system in Big Data Mining Environment that could sift through large amounts of data to find previously hidden patterns, discover valuable new insights and make decisions; apply the dynamics involved in big data technologies and use of distributed data storage and analysis architecture of Hadoop MapReduce; conduct performance benchmarking on Relational Database Management System (RDBMS) and Hadoop cluster, create value in several ways and improve performances. The analytic environment provided a powerful in database algorithms and open source algorithms to enable predictive analytics, data mining, statistical analysis, advanced numerical computations and interactive graphics. Automated analysis of historical data were performed by employing Knowledge Discovery and Data mining (KDD) using Map Reduce Methodology and Predictive Analytic Methodology. The Euclidean distance and the pseudo F-statistic validated Hadoop's high scalability and performance in the real time applications domain, minimized data movement thereby ensuring inherent security and better performance. The result showed that a model for big data mining environment was realized which provided an open source framework for cloud computing and distributed file system for fast data loading.**

Keywords: Big Data, Clustering, Hadoop, MapReduce, Knowledge Discovery and Data Mining.

I. INTRODUCTION

The amount of data in today's world is ever increasing, and has even led to new terms describing this as Big Data [1]. Big Data is a term for data sets that are so large or complex that current methodologies and traditional data processing applications are inadequate [2]. This Big data is hard to process using conventional technologies and calls for massive parallel processing[3].Big data open source technologies have gained quite a bit of traction due to the demonstrated ability to parallelly process large amounts of data. Both parallel processing and technique of bringing computation to data has made it possible to process large datasets at high speed.

Knowledge Discovery is defined as 'Any solution that supports the identification of meaningful historical and potential future data sets for the purposes of predicting future events and assessing the attractiveness of various courses of action [4]. Therefore, this paper concentrates on the development of a knowledge discovery system in a big data mining environment by applying the dynamics involved in big data technologies using Apache Hadoop Map Reduce.

Apache Hadoop is an open source framework that is implemented in Java[5]. It allows data to be stored and processed in large data sets in parallel and distributed fashion. Hadoop Data File System (HDFS) allows data to be distributed over different machines and these machines are interconnected on which data is getting distributed and in Hadoop's terms it is called a Hadoop Cluster. In order to process Big Data, MapReduce is used [6] and it is the programming unit of Hadoop, which allows a parallel and distributed processing of data that is lying across the Hadoop Clusters. Every machine in the cluster is processing the data it has and this is known as Map[7]. Finally, the intermediary inputs are combined in order to provide the output, known as Reduce and hence Mapreduce[8].

II. THE PROPOSED SYSTEM

The diagram of the proposed system for a power Distribution System is shown in Figure 1. It comprises of the Data Lake, data warehouse, visualization and analytics and on line business application.

The first session is Data Lake – Raw Data Storage and Processing unit which comprise structured and unstructured data, Hadoop-based and Map reduce algorithms., The Data Lake – Raw Data Storage and Processing unit handles sensor data, Blogs/Email, web data, Internet, Mobile, Computer, Documents, Audio and Videos, Cameras etc. Sensors are used commonly to measure a physical quantity and convert it into a readable digital signal for processing (and possibly storing).

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 5, Issue 8, pp 65-70, August-2021
https://doi.org/10.47001/IRJIET/2021.508011

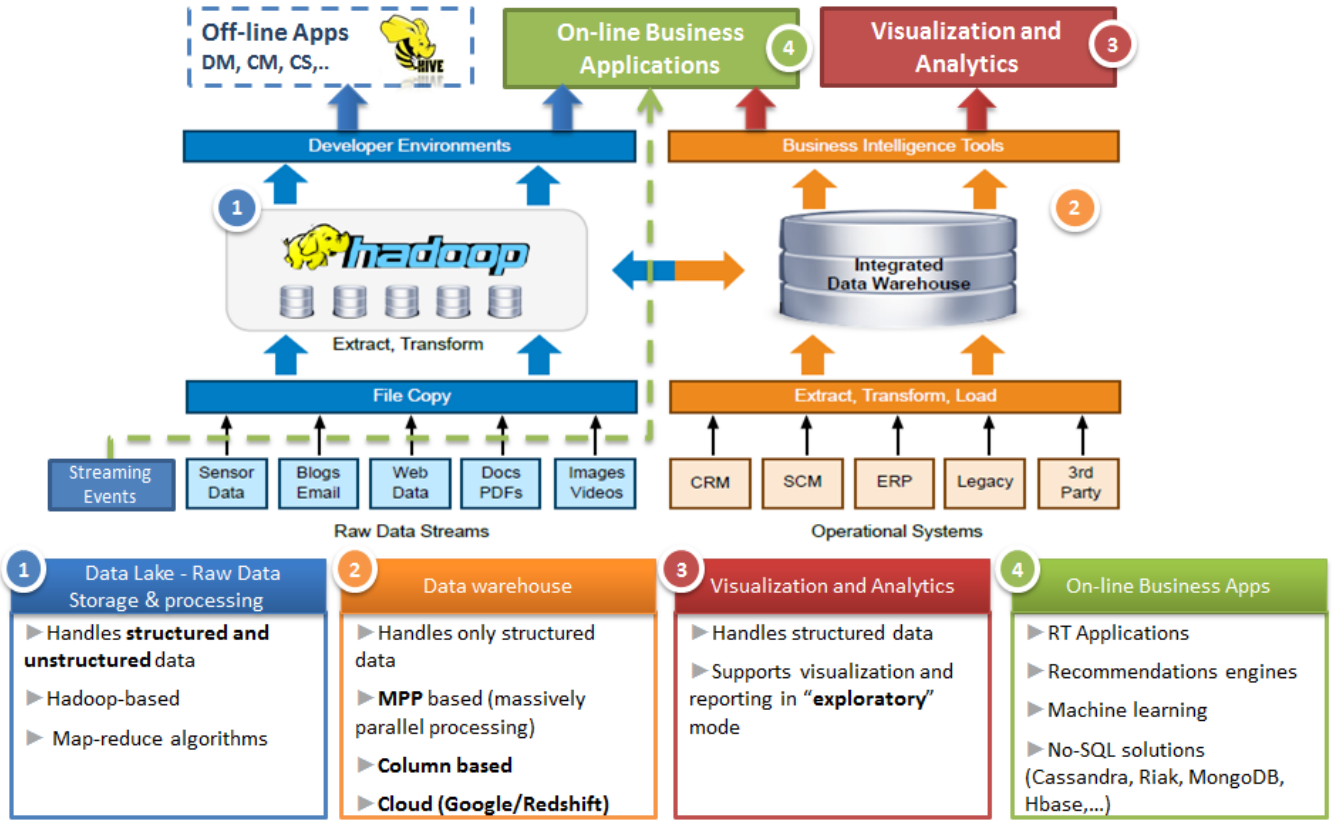

Figure 1: Diagram of the Proposed System

The Data Ware house handles only the structured data, massively parallel processing, column based and cloud data. The Visualization and Analytics section is concerned with structured data and supports visualization and reporting in exploratory mode. The last session which is the On-line application deals with real time applications and machine learning.

2.1 Materials and Method

Materials

The materials used in this paper include Apache Hadoop v3.0, Yet Another Resource Negotiator (YARN), Hadoop Distributed File System (HDFS), Hyper Text Mark Up Language (HTML)/Cascading Style Sheets (CSS), Hypertext Preprocessor 8 (PHP 8), Cloud Watch.

The Apache v3.0 is used to process the big data while YARN spreads the data across the cluster. The Hadoop Distributed File System (HDFS) is a distributed database where the data is saved to become accessible to YARN and Hadoop. Ubuntu Server v16 is used to host the program.HTML/CSS is used to present mined/analyzed data in the frontend.PHP 8, used to transfer mined/analyzed data from the server to the client. R, language for statistical analysis and reporting and. Cloud Watch provides the ability for monitoring condition on specific metrics or log files from different services and sends alert according to condition. Almost all of the services applied on this project will be connected with Cloud Watch to provide emergency alert.

Method

The method used in this paper is the Hadoop MapReduce methodology.

In the proposed system, data comes from multiple sources and are also available in multiple formats, our first start is to integrate and store all the data (2015 - 2017 data) in one single location – a cluster of computers on the cloud. The computers are 3 Google engine computers, each with 128GB Ram and 1TB storage space. The data itself is 4TB.

The dataset cannot be opened on any single computer, and so to integrate it, a java program was ran across the cluster to break down the files into ordered segments, stored across the entire computer. So, while the first computer might contain data1.tsv, data2.tsv ... data30.tsv; the second would contain data 31.tsv - data59... and so on. The Hadoop summary job also attached then reads this data across the cluster reduces all the payments made by each customer into one lump payment, thereby significantly reducing the size of the file by a factor of 36.

This is summarized in the flowchart of Figure 2.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 5, Issue 8, pp 65-70, August-2021
https://doi.org/10.47001/IRJIET/2021.508011

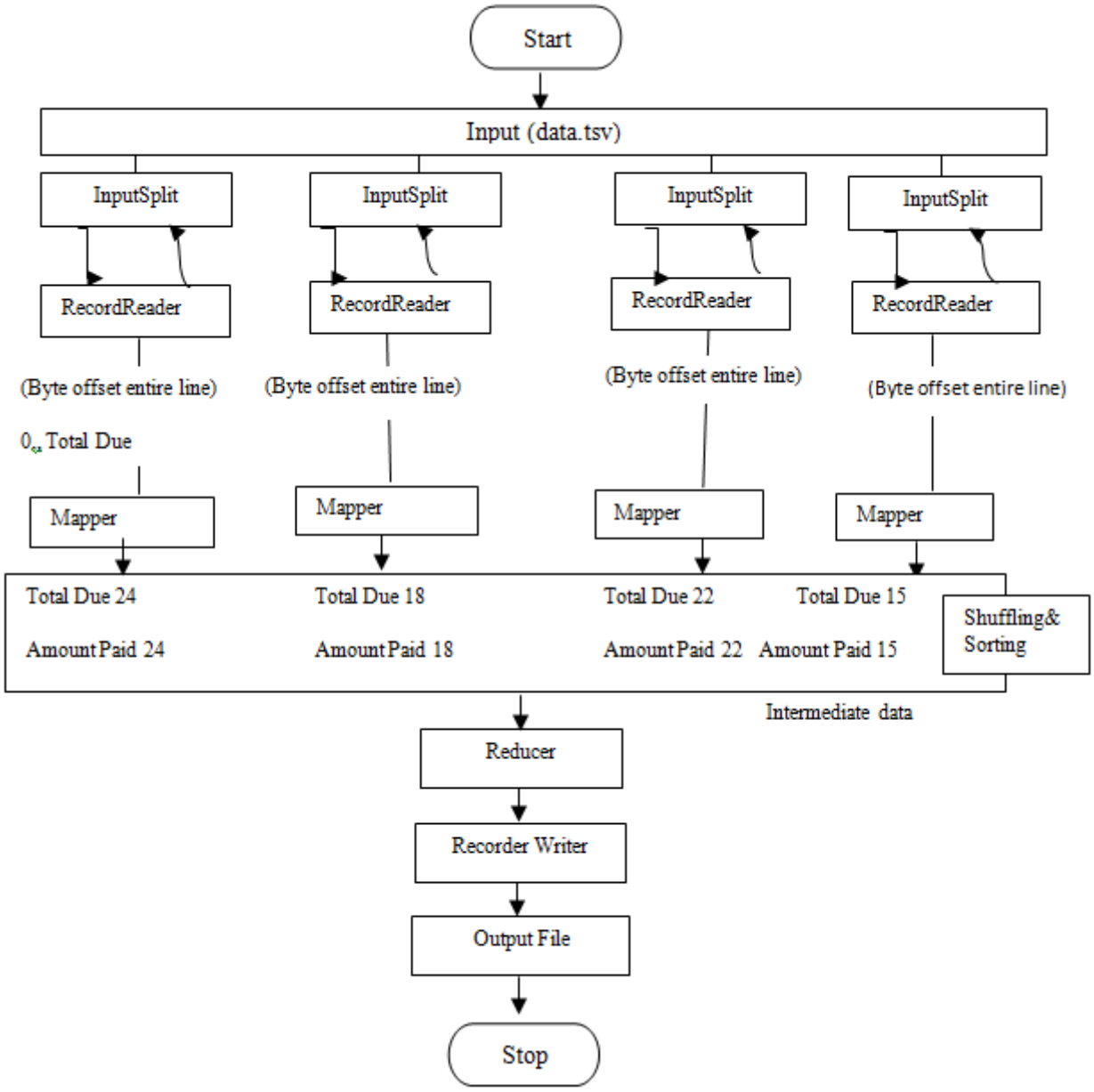

Figure 2: System flow chart for the MapReduce project

Figure 2 shows the system flow chart for the hadoop MapReduce. The file (data.tsv) already stored in HDFS are split into different files depending on the size and each file will have the same number of mappers and predefined interface called RecordReader to read each record/file into key and value pairs, this is done by default.

The RecordReader reads the files based on the type of the format of the file. There are four different file formats: TexInputFormat, KeyvalueTextInputFormat, SequenceFileInputFormat and SequenceAsInputFormat.

It reads each record to know the format and then converts into key value pairs back and forth as byte Offset, Entire line. It starts from 0 and reads the entire line of the record and then sends the data to the mapper, depending on the logic of the program.

The mapper sends out the file into a set of key and value pairs which are sorted and shuffled using the shuffling and sorting interface. The sorted job is then sent to the Reducer which are read by the Record Writer and finally sent out as an output file and you see success or logs directly after your job has been completed successfully.

III. RESULTS AND DISCUSSION

The data for the Days on which payments are made, Control Chart of Units Consumed, Control Chart of Date of

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 5, Issue 8, pp 65-70, August-2021

https://doi.org/10.47001/IRJIET/2021.508011

Payment, Unusual activities detected in the System are presented in Figure3, Figure 4, Figure 5, and Figure 6.

From Figure 3, payments were made mostly on Saturdays. From this, we can deduce that target system does not need as many staff as they currently employ to receive payments during weekdays.

From the control chart given in Figure 4, we see that most points lie outside the control limits. This suggests that the situation is unstable with regard to Amount Paid by the customers. In Figure 5 with the aid of the proposed system, it is shown that the control chart given in Figure 5 had all the points lie within the control limits. This suggests that the situation is stable with regard to the date of payments as defaulters are usually penalized. Also, the figure shows that the average unit consumed by the customers is 304.9 KWH with a deviation of 64.70 KWH.

In Figure 6, the unusual activities recorded in the system are shown and it also shows that 37% of the customers are in the R-2 Metering system. The system developed identified potential fraudulent behaviour especially when a meter is tampered or "frozen." Once the program notices that payments have not been made on a meter for a long time based on some indices like when the owner last made a payment, and how much was made, the payment pattern for the past 2 years; it sends an alert on the screen for prompt action.

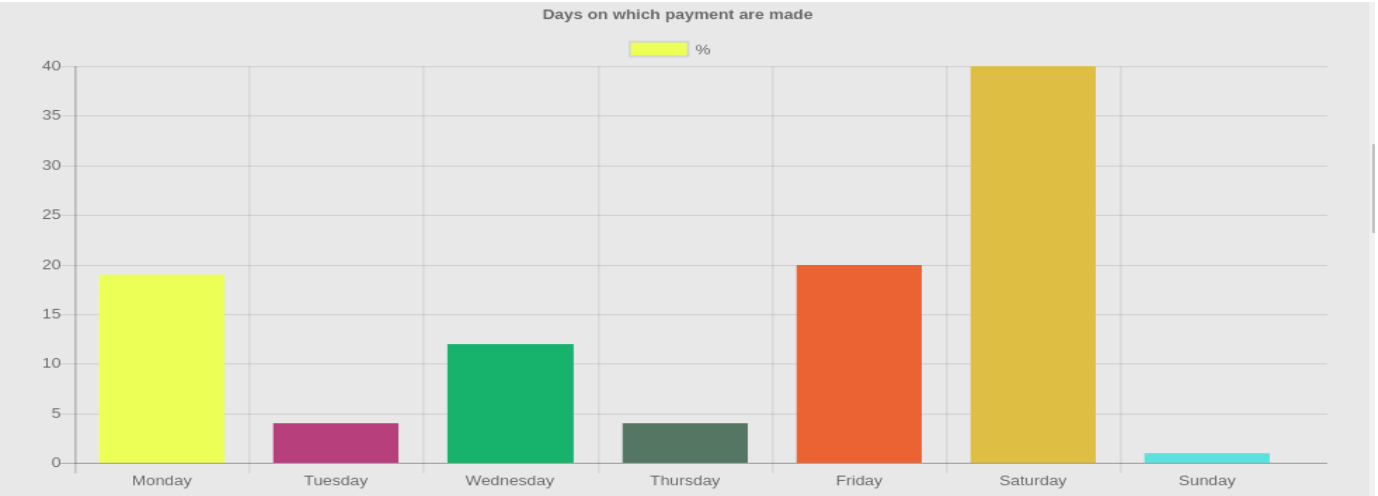

Figure 3: Days on which payments are made

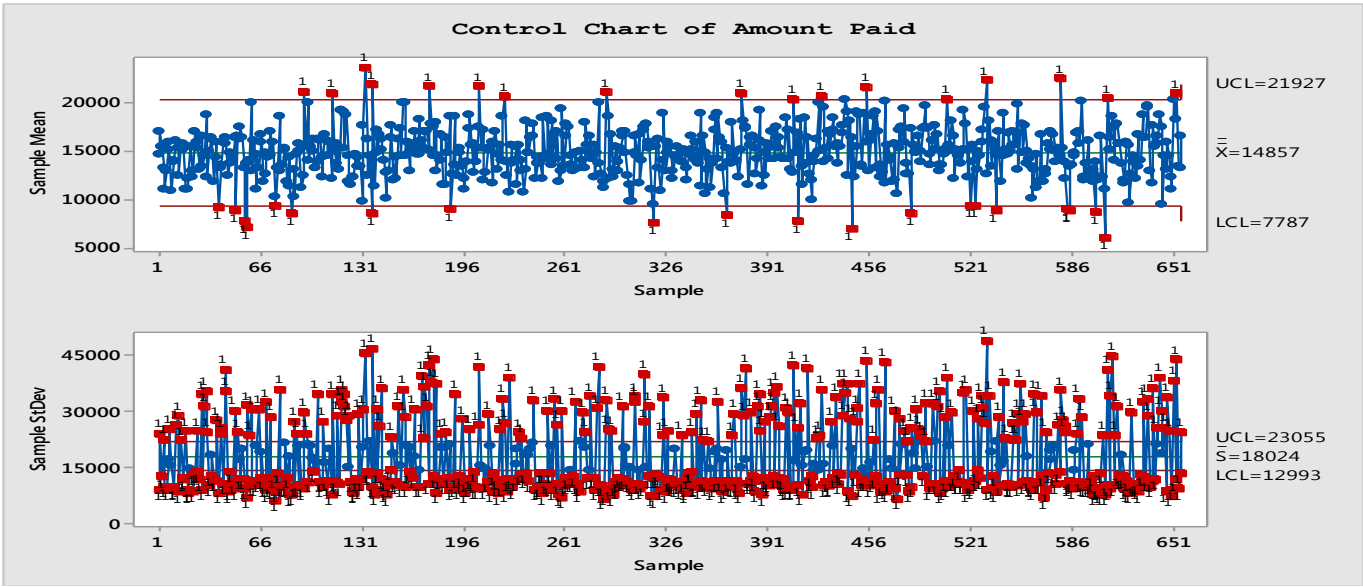

Figure 4: Control Chart of Units Consumed

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 5, Issue 8, pp 65-70, August-2021
https://doi.org/10.47001/IRJIET/2021.508011

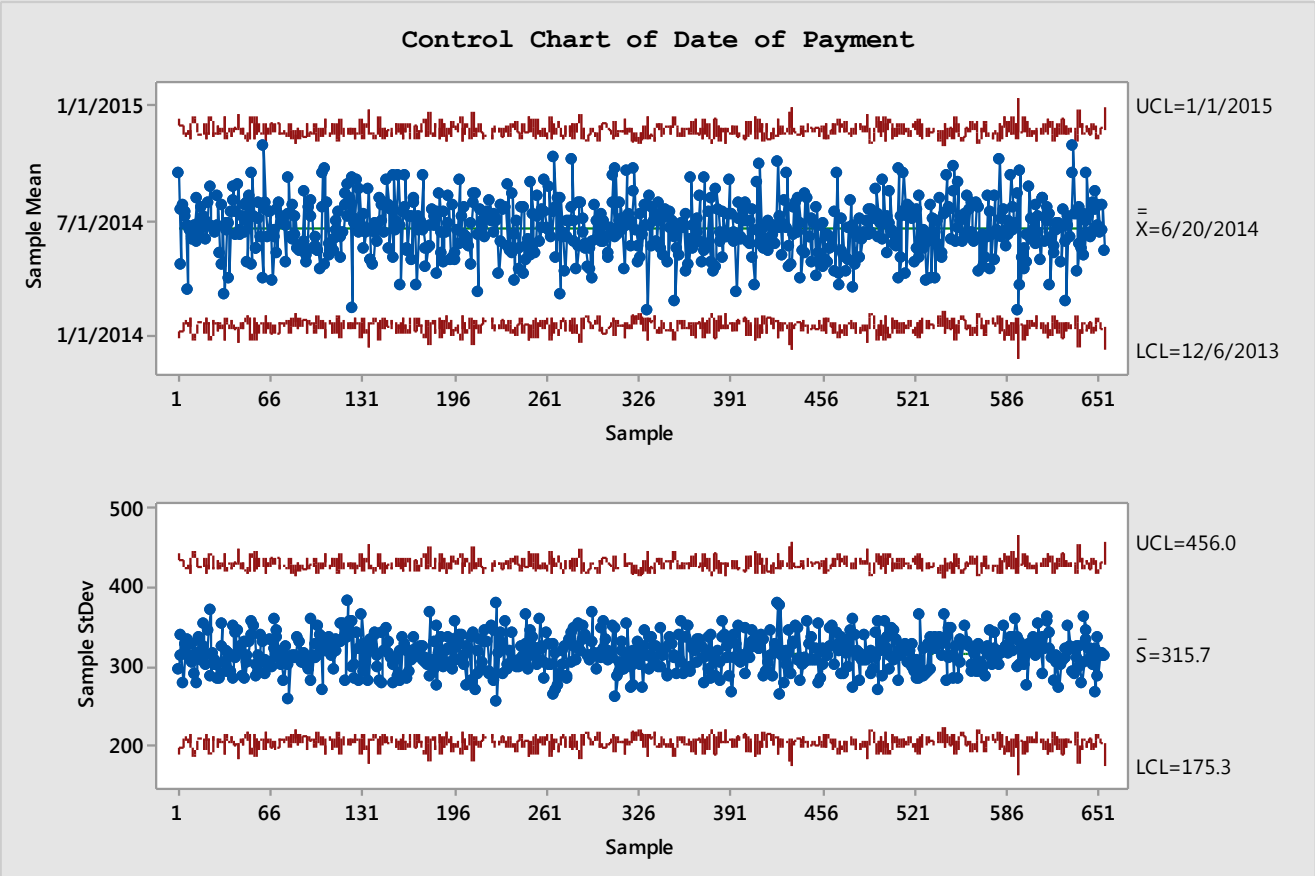

Figure 5: Control Chart of Date of Payment

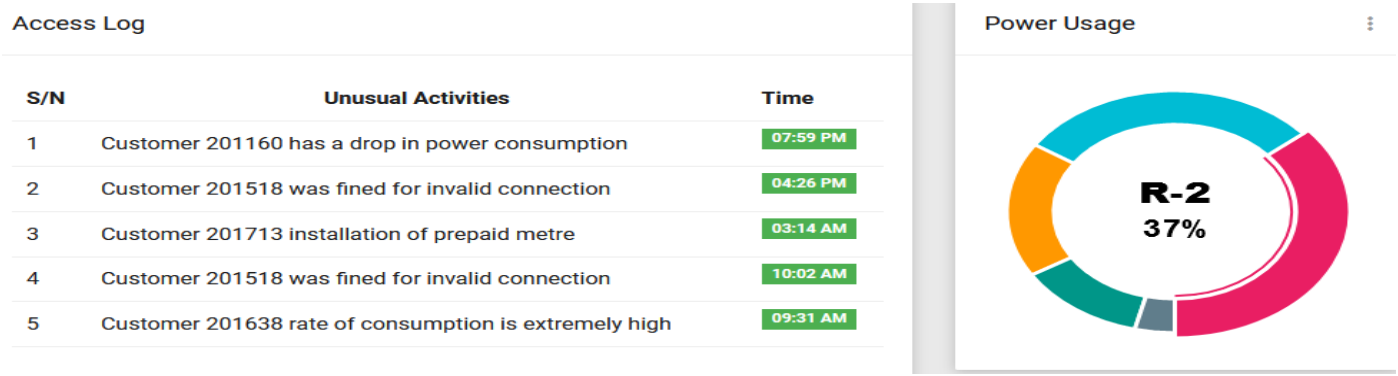

Figure 6: Unusual activities detected in the System

IV. CONCLUSION

This paper has shown the development of a knowledge discovery system in big data mining environment by developing a model that could detect and predict fraudulent or "rare events" in the system. Experimental results show the runtime performance can be improved by more than 80% in Hadoop; thus our mechanism is suitable for multiple types of MapReduce job and can greatly reduce the overall completion time. The predictive analyst has rendered valuable service, by leveraging existing data to enhance knowledge.

REFERENCES

[1] Hand, A., Niu, F., and Ré, C. Hazy, "Making It Easier to Build and Maintain Big Data Analytics." *Communications of the ACM*, vol. 8 no. 5, pp. 40-49, 2013.

[2] Kobielus J., "Parallel database systems: The future of high performance database systems," *Commun. The ACM*, vol. 35, no. 6, pp. 85-98, 2014.

[3] Larose Daniel T. and Larose Chantal D., "Discovering Knowledge in Data: An Introduction to Data Mining," pp. 4-5, 2015.

ISSN (online): 2581-3048
Volume 5, Issue 8, pp 65-70, August-2021
https://doi.org/10.47001/IRJIET/2021.508011

[4] Larose Daniel T. and. Larose Chantal D., "Data Mining and Predictive Analysis," pp. 56-67, 2015.

[5] Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., "Big Bench: Towards an Industry Standard Benchmark for Big Data Analytics."*ACM SIGMODInt. Conf.*, pp. 1197- 1208, 2017.

[6] Narang, A., Srivastava, A., and Katta, N. "High Performance Offline and Online Distributed Collaborative Filtering." *The 12th International Conference on Data Mining (ICDM)*, pp. 549-558, 2013.

[7] Dean J. and GhemawatS., "MapReduce: simplified data processing on large clusters," *in Proceedings of the 6th Symposium on Operating Systems Design & Implementation*, vol. 3, pp. 102–111, 2014.

[8] Russom, P. "Managing Big Data." Available Online at: The Data Warehousing Institute, 2013. https://tdwi.org/articles/2013/10/01/executivesummary-managing-big-data.aspx

Citation of this Article:

Ihekeremma A. U. Ejimofor, Obi O.R. Okonkwo, "Development of a Knowledge Discovery System in Big Data Mining Environment" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 5, Issue 8, pp 65-70, August 2021. Article DOI https://doi.org/10.47001/IRJIET/2021.508011

****

