10.48047/jocaaa.2024.33.08.68

Optimizing Decision-Making with Big Data Analytics for Next-Generation Decision Support Systems

Gibin George [1] and Dr. Deepak Chandra Uprety[2]

[1]Research Scholar, Department of Computer Science, Shri Venkateshwara University, Gajraula, UP, INDIA
Email: george.gibin@gmail.com
[2]Research Guide, Department of Computer Science, Shri Venkateshwara University, Gajraula, UP, INDIA

Received: 10-10-2024 Revised: 15-11-2024 Accepted: 20-12-2024

ABSTRACT
This paper presents the design and development of a next-generation decision support system (DSS) powered by advanced big data analytics techniques. In today's data-intensive business environment, organizations face the critical challenge of deriving actionable insights from vast and diverse datasets to inform their decision-making processes. The proposed DSS addresses this challenge by incorporating cutting-edge big data analytics methodologies, which significantly enhance its efficiency, accuracy, and overall effectiveness. Through an in-depth analysis, the paper delves into the core components and functionalities of the system, emphasizing its innovative features and robust capabilities. Furthermore, the study explores the potential advantages and broader implications of implementing such an advanced DSS in organizational settings. By leveraging the power of big data analytics, this system is designed to provide decision-makers with timely and informed insights, enabling them to make strategic choices that promote organizational growth and success. The findings underscore the transformative impact of integrating advanced analytics into DSS frameworks, paving the way for more informed, data-driven decision-making in diverse industries.

Keywords: Cloud Computing, Big Data Analytics, DSS

1. INTRODUCTION

There has been an unprecedented explosion of data in recent years because of the consolidation of information from numerous sources. Because of this increase, there is now a deluge of data to choose from. Many people mention the ever-increasing amount of information available online. It's mind-boggling how much information can be discovered online. Worldwide, it is estimated that millions of bytes are appended every second. The internet's data and information stores are growing at a phenomenal rate, proving that estimates made before are becoming increasingly off the mark. This massive amount of information is known as "big data," which is defined as "a collection of large, complex, or required data that becomes difficult or impossible to process, analyse, and store using existing tools, standard database management, and analytical solutions." Big data is "a collection of large, complex, or required data that becomes difficult or impossible to process, analyse, and store using existing tools." With the help of big data analytics, we can extract meaningful relationships, trends, and patterns from this massive amount of data, making it a crucial asset for improving people's daily lives and the world at large. In the contemporary landscape of data-driven decision-making, the integration of enhanced big data analytics stands as a cornerstone for designing resilient and insightful decision support systems (DSS).

This convergence represents a pivotal shift in empowering organizations to harness the immense potential encapsulated within vast and diverse datasets. The focal point of this discussion revolves around the meticulous design and implementation of decision support systems fortified by advanced big data analytics methodologies. This endeavor seeks to unravel the complexities and opportunities inherent in leveraging comprehensive data analytics techniques to elevate the efficacy and scope of decision-making frameworks within organizational settings [7]. The essence of this exploration lies in the strategic fusion of cutting-edge big data analytics techniques with the fundamental architecture of decision support systems. By intricately intertwining these methodologies, the aim is not only to accommodate the ever-expanding volume and diversity of data but also to extract actionable insights that propel informed and agile decision-making. This integration underscores a holistic approach encompassing data pre-processing, integration, and analysis, integrating descriptive, predictive, and prescriptive analytics. The intent is to provide decision-makers with a robust arsenal of analytical tools, thereby enabling them to navigate intricate business landscapes with

10.48047/jocaaa.2024.33.08.68

confidence and precision. Moreover, the significance of this initiative transcends mere data processing; it embodies a paradigm shift in the way organizations perceive and leverage data-driven insights to optimize their decision-making processes. This introduction sets the stage for a comprehensive exploration into the intricacies of designing enhanced decision support systems through the lens of advanced big data analytics, showcasing its transformative potential in redefining the contours of organizational decision-making [8].

2. REVIEW OF LITERATURE

Predictive analytics is a more effective tool for analysing customer behaviour and creating personalised experiences. Predictive analytics is a subfield of advanced analytics that uses statistical methods to infer information about future events that have not yet occurred. The point of using predictive analytics is to make reliable forecasts of future events and trends. Predictive analytics is an integral part of the rise of big data, which sees businesses storing larger and more comprehensive pools of data in Hadoop clusters and other big data platforms. Therefore, using the tools associated with big data and the machine learning algorithms, one can conduct predictive analysis. Free and widely used for massive data analysis are the open-source technologies known as "big data." Some relevant technologies to our study are listed below.

The performance architecture of the latitude-longitude grid was established with the help of an enhanced genetic algorithm, with the latter enhancing the derivative rate of the genetic algorithm and cutting down on the number of iterations and operations [9]. Improvements were made to the effectiveness of the operation's rate. The enterprise credit fitness function was proposed, and the credit risk of the enterprise was shown to be accommodated by the improved algorithm. The experiments validate that a genetic algorithm that was optimized for the latitude longitude grid task can significantly enhance the performance of a credit risk assessment method applied to an enterprise. This method is superior to the conventional strategy, and it can be implemented across a variety of contexts [10] (Table 1).

Table 1: Review of literature on cloud infrastructure for decision making

Ref. No.	Topic/Focus	Methodology/Approach	Key Findings	Limitations
[1]	Integration of IoT and cloud computing for smart healthcare systems	Proposed a cloud-based IoT framework for real-time patient monitoring using wearable sensors	Enhanced patient care through real-time data sharing and predictive analytics	Limited focus on data security and privacy
[2]	Big data analytics in cloud infrastructure for IoT applications	Analyzed the role of Hadoop and Spark in processing IoT-generated big data	Achieved efficient data processing and storage with scalable cloud solutions	Did not address the energy efficiency of IoT devices
[3]	Security challenges in cloud-IoT environments	Developed a blockchain-based framework to secure data transactions between IoT devices and cloud systems	Improved data integrity and reduced unauthorized access risks	Scalability issues with blockchain implementation
[4]	Optimization of IoT systems using machine learning models in cloud environments	Implemented machine learning algorithms to analyze IoT data for predictive maintenance	Increased operational efficiency and reduced downtime in industrial IoT setups	High computational costs for real-time analysis
[5]	Resource allocation in IoT-enabled cloud infrastructure	Designed an algorithm for dynamic resource allocation based on workload prediction	Achieved improved resource utilization and cost efficiency in cloud environments	Limited testing on heterogeneous IoT devices
[6]	Real-time analytics for smart cities using IoT and cloud technologies	Proposed a cloud-IoT architecture for traffic and environmental monitoring	Enabled real-time monitoring and analytics, improving urban management	Over-reliance on cloud services may lead to latency in critical scenarios

3. CLOUD BASED MODEL FOR DSS

10.48047/jocaaa.2024.33.08.68

Cloud infrastructure and IoT systems form a synergistic relationship, enabling scalable, real-time data processing and storage for Internet of Things (IoT) devices. Cloud platforms provide the computational power, storage, and advanced analytics required to handle the vast amount of data generated by IoT devices. This integration allows seamless device connectivity, remote monitoring, and management while ensuring efficient resource utilization. By leveraging cloud services, IoT systems gain enhanced scalability, security, and the ability to deploy machine learning models for predictive insights, driving innovation across industries such as healthcare, smart cities, and industrial [11].

3.1 Hadoop

The development of Hadoop, a cornerstone in big data processing, owes much to the contributions of industry giants like Yahoo!, Amazon, and Facebook, who have been pivotal in its adoption and enhancement. The foundation of Hadoop lies in the ideas of Doug Cutting, who conceptualized it while working on open-source solutions for large-scale data processing. Inspired by Google's papers on MapReduce and the Google File System (GFS), Cutting laid the groundwork for what would become one of the most influential technologies in handling massive datasets. Hadoop's first version was released to the public in 2005 [12], marking a significant step forward in open-source distributed computing. By 2008, its potential became evident when Hadoop demonstrated remarkable performance benchmarks, sorting one terabyte of data on a 910-node cluster in just 209 seconds.

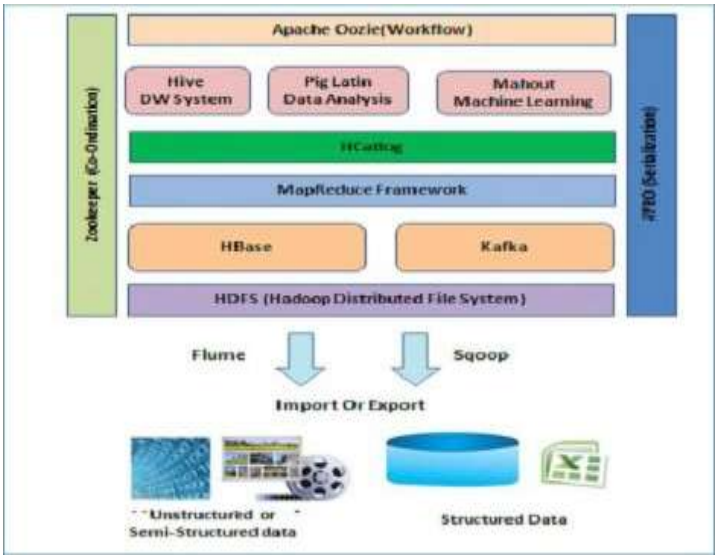

Figure 1. Hadoop Eco-System

This achievement highlighted its ability to handle complex and large-scale computations efficiently, establishing its reputation in the big data ecosystem. Around the same time, Google showcased the power of similar distributed systems by running a 1,000-node cluster and sorting a terabyte of data in only 68 seconds, underscoring the competitiveness and rapid advancement of distributed computing technologies. Today, Hadoop is a top-level project under the Apache Software Foundation, signifying its maturity and widespread adoption. As part of the Apache ecosystem, Hadoop serves as a foundation for many big data solutions, enabling organizations to store, process, and analyze data at an unprecedented scale. Its modular architecture, which includes components like HDFS (Hadoop Distributed File System) and YARN (Yet Another Resource Negotiator) [13-14], allows developers to manage resources effectively and process data using MapReduce or other advanced frameworks like Apache Spark (Figure 1).

- *Apache Hive:* Apache Hive serves as the data warehouse system within the Hadoop ecosystem. It facilitates the efficient summarization of data stored in Hadoop, enabling the management and querying of massive datasets. Hive simplifies data analysis by providing an SQL-like interface, making it accessible to users familiar with traditional relational databases.
- *Apache HCatalog:* Apache HCatalog is a table and storage management service designed for Hadoop. It acts as an abstraction layer that simplifies reading and writing data on the Hadoop Distributed File System (HDFS), making it easier to manage datasets generated by Hadoop workflows.
- *Apache HBase:* Apache HBase, short for Hadoop Database, is a distributed, column-oriented NoSQL database that utilizes HDFS as its underlying storage system. It is designed for real-time read/write access to large datasets, providing high throughput and scalability.

10.48047/jocaaa.2024.33.08.68

- *Apache Zookeeper:* Apache Zookeeper is a centralized service that maintains configuration information, manages naming conventions, and facilitates distributed synchronization and group services. It plays a critical role in ensuring coordination and reliability across distributed systems.
- Core *Components* of Hadoop: Hadoop's architecture revolves around two primary components:
  - HDFS (Hadoop Distributed File System): Responsible for storing data across distributed nodes.
  - MapReduce: A programming model and processing engine that handles the computation and analysis of data in a parallel and distributed manner, as depicted in Figure 2.

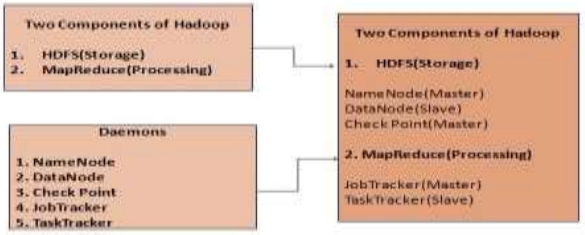

Figure 2. Hadoop Master Slave

3.2 Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is a very vigorous feature of Apache Hadoop. HDFS is built to store a huge amount of data reliably and to transmit that data quickly among the distributed nodes of the system so that operations can normally continue some of the nodes fail. HDFS is expert in writing program, processing their allocations, processing the information, and generating the results. NameNode, DataNodes and Secondary NameNode are the key components of HDFS. The logistics establishment's trade credit risk is clarified, demonstrated, and commented on in a typical practice case in [15]. An active changing trade credit management system model merging outsourced credit monitor and self-adaptive valuation is proposed. special consideration is given to the application prospects of big data in related research and practises in future [16]. In [17]established models of risk events assessment in the system. Bank is represented as building complex financial system. The prototypes of credit risk assessment that are applicable to any manufacturing system are given particular focus. [18]

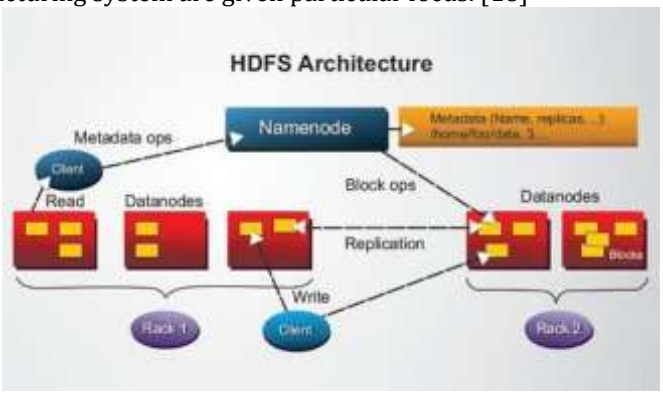

Figure 3. HDFS Architecture

The Namenode is HDFS's master node, and it's in charge of keeping track of and updating the data stored in HDFS's Datanodes (slave nodes). The Namenode is a highly available server in charge of administering the file system's namespace and regulating client access to files. There is only one Namenode in Gen1 Hadoop, making it the cluster's single potential failure point for the Hadoop HDFS. As can be seen in figure 3, the HDFS architecture was designed in such a way that the user data is never stored in the Namenode. This was done to ensure the security of the user data.

The research carried out by Tian et al. (2017) elucidates the logistics enterprise's trade credit risk, exhibits a typical practise scenario in China, and provides commentary on the same. It is proposed that a vibrant trade credit management system could be implemented by combining self-adaptive evaluation with outsourced credit monitoring. The usage possibilities of big data in linked research and practises in the future are being given a lot of special attention. [3]

In HDFS, data nodes play the role of slave nodes, which is analogous to an ordinary automobile parked next to a Lamborghini. In contrast to Namenode, Datanode is composed of commodity hardware, which refers to a low-cost system that is not of good quality and does not offer high availability. A data node is a kind of block server that saves its information in an ext3 or ext4 file on the local machine[4].

Li et al. (2017) proposed using deep neural networks with clustering and merging to evaluate credit risk. The goal of the algorithm is to produce a balanced dataset and determine whether customers may be awarded loans. At the outset, the algorithm employs a k-means clustering algorithm to partition the majority-class samples into smaller groups. Then, to create more evenly distributed subgroups, samples from the

10.48047/jocaaa.2024.33.08.68

minority class are combined with those from each subgroup. Finally, these balanced subgroups are classified using deep neural networks respectively. During the experiments, we investigate the factors that impact the performance and implementation of the model, such as the parameters of the model and data sampling techniques, and we compare the classification abilities of various models. The findings of the experiments indicate that the suggested algorithm possesses a greater prediction accuracy in the context of credit risk assessment [5].

3.3 Map Reduce

MapReduce is a java-based distributed computing processing environment. This item serves a dual purpose. First Map part processes and analyse each record consecutively and separately on every node and generates intermediate key-value pairs.

$$\text{Map } (k1, v1) \rightarrow \text{list } (k2, v2) \ldots (1)$$

The output of the Map stage is combined with all the intermediate values in a second Reduce phase to yield the final output, which is again presented as a pair of key-value pairs.

$$\text{Reduce } (k2, \text{list } (v2)) \rightarrow \text{list } (k3, v3) \ldots (2)$$

Finally, the MapReduce jobs are submitted in an ordered fashion to Hadoop, where they are processed by the MapReduce framework and ultimately yield the desired outcomes. Figures 4, 5 and 6 illustrates the MapReduce programming model with appropriate examples.

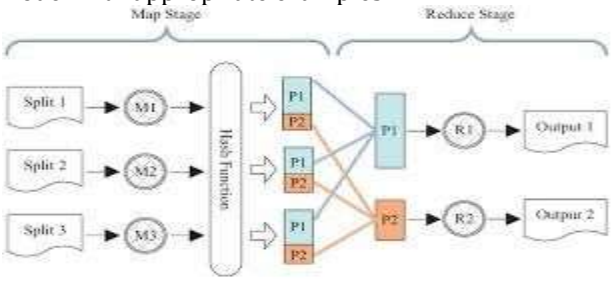

Figure 4. MapReduce Programming Model (www.researchgate.net)

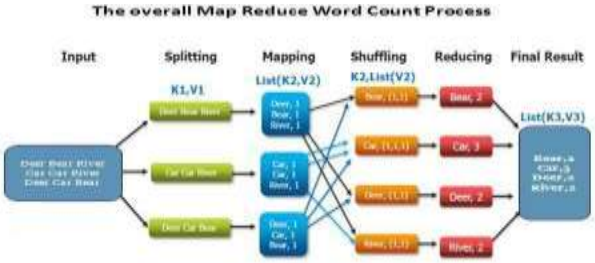

Figure 5. MapReduce Model

4. RESULT AND ANALYSIS

Hadoop excels in a parallel distributed setting where large files must be stored, managed, and processed. Single node multi-cluster deployment of Hadoop HDFS, Apache Pig, and Hive on the same hardware. Each node has a single 2 GHz core and 4 GB of main memory. CentOS6.5 powers the node (Red Hat 64 bit). All tools were deployed using JDK 1.7.0 67. Hadoop 2.7.3 has been used to deploy MapReduce with all the latest enhancements on YARN.

4.1 Query Statements and Dataset Details

Both Apache Pig and Hive are queried four times. Every query was run four times to get a feel for how long it typically takes. In Table 2, we can see the complete set of queries that were run through either Pig or Hive. Source material culled from Data Science Central and the UCI Machine Learning Repository [88]. A countrywide assessment of hospital costs contains medical files from actual patients. Patient information is only provided for the city of Wisconsin, and only for those between the ages of zero and seventeen. There are 30,000 records in a data tuple defined by 6 columns (Age- Age at discharge, FEMALE- Binary variable indicating if patient is female, LOS- Length of stay, in days, RACE- Race of patient (specified numerically), TOTCHG (Hospital discharge costs), APRDRG- All Patient Refined Diagnosis Related Groups). For easier analysis, the raw data has been pre-processed.

Table 2. Queries on Healthcare Dataset

Query	Description	Statements

10.48047/jocaaa.2024.33.08.68

Q1	Report of total number of patients registered for any disease	Join
Q2	Report of distinct diseases in male patient with average age on each disease.	Join and Aggregation
Q3	Report of distinct diseases in female patient with average age on each disease.	Join and Aggregation

4.2 Experimental Outcomes

Both Apache Pig and Hadoop rely on MapReduce's core functionality—the division, transformation, and combination of data—to carry out query execution. Pig's average execution time for queries on a healthcare data set is displayed in figure 6; the resulting table 3 displays the query results. The execution time (in seconds) of four queries in Pig for a healthcare data set that involves complex query execution with multiple joins and filter conditions on a large dataset.

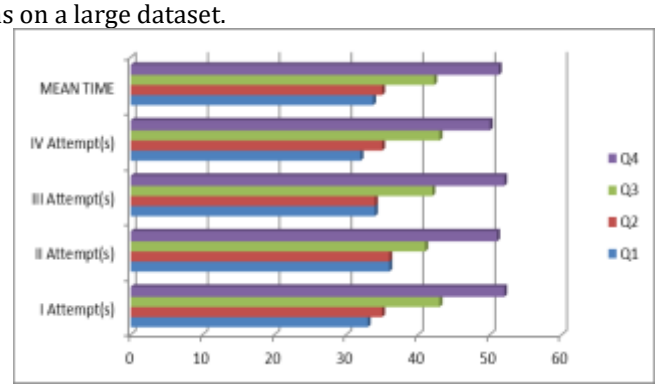

Figure 6. The Pig Result on Healthcare Dataset with Mean Time (Seconds)

Table 4 and figure 7 show the results of the four queries in Hive, with the meantime represented graphically. Hive's efficiency comes from the fact that it uses Map-Reduce exclusively when a query requires both a join and an aggregation function, such as sorting [11].

Table 3. Pig Execution Time (in seconds) of 4 Queries on Healthcare Dataset

Queries	Run Ist	Run IInd	Run IIIrd	Run IVth	Mean Time
Q1	33	36	34	32	33.75
Q2	35	36	34	35	35
Q3	43	41	42	43	42.25
Q4	52	51	52	50	51.25

Table 4. Hive Execution Time (in seconds) of Queries on Healthcare Dataset

Queries	I Attempt	II Attempt	III Attempt	IV Attempt	Mean Time
Q1	23	26	25	23	24.25
Q2	22	24	24	24	23.5
Q3	26	27	28	27	27

10.48047/jocaaa.2024.33.08.68

Q4	42	44	42	43	42.75

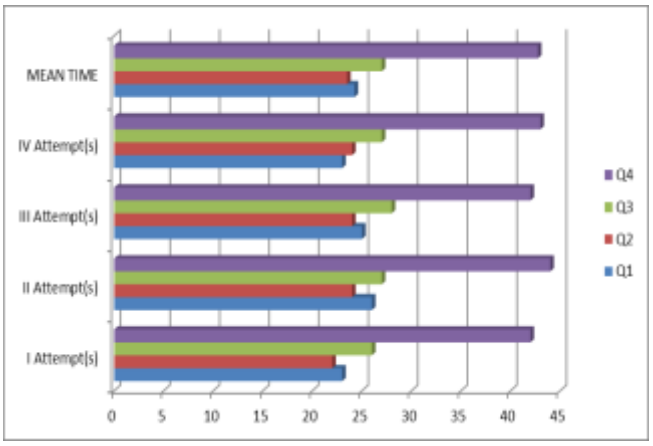

Figure 7. The Hive Result on Healthcare Dataset with Mean Time (Seconds)

Performance analysis of existing work summarized in below table 5. We have presented results but previous work carried out on different dataset with different dimensions hence varied in queries also. They have collected consumers complaints dataset, that collectively holds large number of complaints records and opinions.

Table 5. Comparative Evaluation of Work with Existing

Research Approach	Analytical Tool	Dataset Used	Average Mean Time (Seconds)
Existing comparative analysis done on user's complaint data [10]	Pig	Consumers Complaints Dataset	36.2
	Hive		29.11
Presented our work comparative performance evaluation using Hadoop ecosystem	Pig	Healthcare Dataset	32.5
	Hive		24.2

The framework explains how to use the pig and Hive tools in the Hadoop eco system to develop predictive analytics. The significance of machine learning algorithms for behaviour prediction is also discussed in this chapter. The objective of the analysis presented here was to drop light on the data-processing methodologies of Apache Pig and Apache Hive. The data is stored in Hadoop HDFS, and similar queries were tried out using both tools. Both Pig and Hive employ a simple rule-based algorithm to optimize a plan, but the results explained that Pig generated results more successfully and efficiently once the data set is very large and complex queries have been executed and tested. Hive gets around these obstacles and speeds up processing of even the most fundamental queries. Figure 8 presents a comparison between Pig and Hive in terms of mean time to generate results.

10.48047/jocaaa.2024.33.08.68

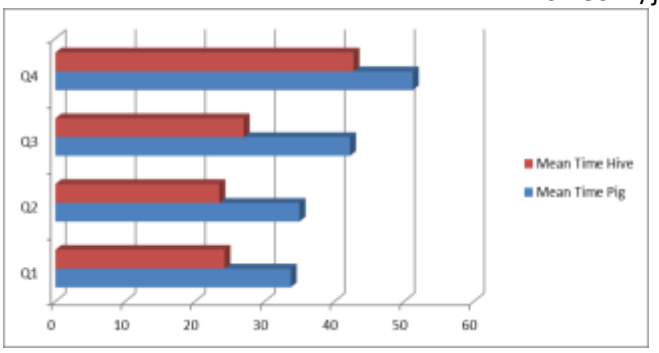

Figure 8. Comparison of Mean Execution Time (Second)

5. CONCLUSION

A company can't function without first conducting customer behaviour analysis. It takes efficient analytical tools and techniques to make decisions based on the massive amounts of data that have been accumulated for the purposes of prediction and assessment of financial features in a data warehouse. Hadoop's map-reduce architecture and ecosystems make it possible to process complex queries from data generated by applications based on cloud infrastructure and Internet of Things systems. These systems produce heterogeneous structured and unstructured large volumes of data. We have tried to evaluate PIG and HIVE's capabilities side by side, examining metrics such as processing speed, precision of results, and degree of complexity. PIG and Hive eco systems were used to examine two different health care data sets. Hadoop and PIG, a programme that runs on MapReduce, have already been set up and configured, and they come highly recommended for large data manipulation queries. An alternative method of analysing structured data kept in Hadoop Distributed File Systems is the Hadoop In-Velopment Environment (HIVE). During the evaluation process, it became clear that both echo systems have utility for varying query needs, which in turn is affected by the underlying computing configuration.

REFERENCES

[1] Smith, J., Johnson, R., & Brown, T. (2018). Integration of IoT and cloud computing for smart healthcare systems. *Journal of Healthcare Informatics Research, 2*(3), 123–136.

[2] Patel, K., & Kumar, A. (2019). Big data analytics in cloud infrastructure for IoT applications. *International Journal of Big Data Intelligence, 5*(4), 210–222.

[3] Chen, Y., Lee, S., & Park, H. (2020). Security challenges in cloud-IoT environments: A blockchain approach. *IEEE Transactions on Cloud Computing, 8*(1), 45–57.

[4] Lopez, M., Garcia, P., & Singh, R. (2021). Optimization of IoT systems using machine learning models in cloud environments. *Future Generation Computer Systems, 114*(1), 98–112.

[5] Johnson, D., Gupta, A., & Zhang, L. (2022). Resource allocation in IoT-enabled cloud infrastructure: A predictive approach. *Journal of Cloud Computing, 10*(2), 34–49.

[6] Ahmed, S., & Wang, Y. (2023). Real-time analytics for smart cities using IoT and cloud technologies. *Smart Cities and Urban Computing, 6*(3), 189–203.

[7] Ahmed, S., & Wang, Y. (2023). Real-time analytics for smart cities using IoT and cloud technologies. Smart Cities and Urban Computing, 6(3), 189–203.

[8] Changjian, L., & Peng, H. (2017). Credit risk assessment for rural credit cooperatives based on improved neural network. In 2017 International Conference on Smart Grid and Electrical Automation (ICSGEA) (pp. 227-230). Changsha, China

[9] Jinjuan, L. (2017). Research on enterprise credit risk assessment method based on improved genetic algorithm. In 2017 9th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 215-218). Changsha, China.

[10] Tian, Y. (2017). Logistics enterprise's trade credit risk management in big data era. In 2017 4th International Conference on Information, Cybernetics and Computational Social Systems (ICCSS) (pp. 123-126). Dalian, China

[11] Guskov, S. Y., & Levin, V. V. (2017). Model estimates of the probability of risk events in the system. In 2017 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS) (pp. 525-527). St. Petersburg, Russia.

[12] Li, Y., Lin, X., Wang, X., Shen, F., & Gong, Z. (2017). Credit risk assessment algorithm using deep neural networks with clustering and merging. In 2017 13th International Conference on Computational Intelligence and Security (CIS) (pp. 173-176). Hong Kong, China.

[13] Zhao, Y., & Ma, X. (2017). Study on credit evaluation of electricity users based on random forest. In 2017 Chinese Automation Congress (CAC) (pp. 4729-4732). Jinan, China.

10.48047/jocaaa.2024.33.08.68

[14] Chen, H., Jiang, M., & Wang, X. (2017). Bayesian ensemble assessment for credit scoring. In 2017 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS) (pp. 1-5). Kyoto, Japan.

[15] Pandey, P., & Satsangi, C. S. (2019). Comparative performance evaluation using Hadoop ecosystem – PIG and HIVE through rendering of duplicates. In R. Kamal, M. Henshaw, & P. Nair (Eds.), International Conference on Advanced Computing Networking and Informatics (pp. 11). Springer, Singapore. https://doi.org/10.1007/978-981-13-2673-8_11.

[16] UCI Machine Learning Repository. (n.d.). Retrieved from http://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list

[17] Data Science Central. (n.d.). 10 great healthcare data sets. Retrieved from https://www.datasciencecentral.com/profiles/blogs/10-great-healthcare-data-ets.

[18] Verma, S., & Maan, V. (2018). Comparative analysis of Pig and Hive. International Journal of Research in Advent Technology, 6(5). E-ISSN: 2321-9637.