Svedberg Open
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper
Open Access

A Scalable Data Mining Framework for Knowledge Discovery Using Distributed Big Data Analytics in Heterogeneous Systems

Diwakar Bhardwaj[1], M.V. Rajesh[2], Sathya arthi R[3], Talla Prashanthi[4], Dr. Sowjanya Bagadi[5], Mahesh Kurulekar[6], Sunil Thakur[7], Mahendran Arumugam[8]

[1]Department of Computer Engineering & Applications, GLA University, Mathura, Email: diwakar.bhardwaj@gla.ac.in
[2]Associate Professor, Department of CSE (Data Science), Pragati Engineering College, ADB Road, Surampalem, NearPeddapuram, Kakinada District, Andhra Pradesh, India - 533437. Email: assocdean_se@adityauniversity.in
[3]Assistant Professor, Department of Management Studies, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Email: sathyaarmba@maher.ac.in
[4]Assistant Professor, Department of Information Technology, Vardhaman College of Engineering, Shamshabad, Hyderabad, India - 501 218, Email: prashanthi1711@vardhaman.org
[5]Assistant Professor, School of Business, Aditya University, Surampalem, Andhra Pradesh, Pin 533437,
Email: sowjanyab@adityauniversity.in
[6]Assistant Professor, Civil Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037.
Email: mahesh.kurulekar@vit.edu
[7]School of Engineering &Technology,Noida international University, Uttar Pradesh 203201, India, Email: sunil.thakur@niu.edu.in
[8]Center for Global Health Research,Saveetha Medical College, Saveetha Institute of Medical and Technical Sciences, Chennai, India.
Email: mahendrana.sdc@saveetha.com

Abstract

The massive increase in structured and unstructured computing resources in the form of cloud platforms, IoT devices, distributed networks, enterprise systems, among others, has made big data analytics a critical area of research. Conventional data mining methods tend to have serious problems with big data due to the physical unscalability of these methods, excessive computational cost, latency and inefficient use of resources with a distributed system. These issues require scalable and efficient frameworks which can handle large quantities of heterogeneous data and guarantee the correct knowledge discovery. This study presents a scalable distributed data mining architecture that will be used to boost knowledge discovery by big data analytics in heterogeneous systems. The framework combines Apache Hadoop and Apache Spark to have the ability of efficient distributed stored data, parallel computing and in-memory computing. The proposed model takes into account machine learning and data mining algorithms such as Random Forest, Decision tree, K-Means clustering and FP-Growth to do the job of classification, clustering and pattern extraction effectively with the distributed datasets. The framework is analyzed based on various machine learning and distributed system performances metrics like accuracy, precision, recall, F1-score, execution time, scalability, throughput, and resource utilization. It has been shown in experiments that the suggested framework greatly enhances processing efficiency, scalability, and predictive performance of traditional centralized systems and Hadoop systems and provides efficient and reliable knowledge discovery in large-scale heterogeneous settings.

Keywords: Big Data Analytics, Distributed Data Mining, Knowledge Discovery, Heterogeneous Systems, Machine Learning, Apache Spark, Hadoop, Scalability, Distributed Computing.

1. Introduction

The worldwide explosion of digital technologies, Internet of Things (IoT) systems, cloud computing systems, intelligent systems, and web-based services has led to the creation of huge amounts of structured, semi-structured, and unstructured data. This rapid growth of data has hastened the relevance of big data analytics in deriving useful information and insights to take action on the large datasets. The typical features of big data environments are volume, velocity, and variety that pose serious storage, handling, and processing challenges

(Owais & Hussein, 2016). The examples of smart cities, healthcare, industrial automation, and cloud, among others, without ceasing, produce large-volume heterogeneous data that necessitate the application of sophisticated analytical algorithms to effectively process the data and make intelligent decisions (Jara et al., 2015; Santos et al., 2020). With the increasing complexity of data, organizations are more using scalable distributed computing technology which helps provide high performance analytics and knowledge discovery processes.

Conventional centralized data processing frameworks do not support the computational and storage needs of large-scale heterogeneous data sets. As a result, distributed analytics solutions have emerged to be critical in the scalable and efficient processing of big data. Distributed storage, parallel processing, and in-memory computation capabilities that are offered by technologies like Apache Hadoop or Apache Spark can be of great benefit in distributed environments as they enhance scalability and processing performance greatly (Li et al., 2014; Yang et al., 2017). Distributed architectures in clouds also increase the flexibility of systems, the use of resources, and real-time analytics in multiple computational nodes (Chen et al., 2016; Dehury and Sahoo, 2016). These distributed systems facilitate the effective management of large scale data as well as providing scalable analytics in heterogeneous computing systems.

These improvements notwithstanding, there are a number of issues that plague the performance and efficiency of the distributed data mining systems. Heterogeneous environments on a large scale are usually characterized by problems with computational complexity, scalability, communication overhead, fault tolerance, and high execution time in data processing (Wu et al., 2017). Moreover, the combination of cloud platform data, IoT devices, and hybrid networks and distributed infrastructures creates important interoperability and resource management issues (Nodehi et al., 2017). Current frameworks are often poorly scaled and resource intensive to work with ever-expanding datasets. Moreover, a large number of traditional machine learning applications are computationally intensive and cannot be used in real-time analytics and precise knowledge discovery in a distributed setting, which restricts their use in intelligent decision-making applications.

The growing need of scalable real-time analytics and intelligent automation have been the driving force behind the creation of effective distributed data mining models that can manage heterogeneous big data environments. By combining distributed processing technologies with scalable machine learning algorithms, it is possible to achieve significant enhanced computational efficiency, predictive performance, and knowledge extraction capabilities. Random Forest, Decision Tree, K-Means clustering, and FP-Growth are all common algorithms to classification, clustering and pattern discovery in big data applications (Raza and Khosrova, 2015). The synergy between distributed processing framework and intelligent analytics techniques offer the possibilities of reducing the execution time, improving the resource utilization and enhancing the scalability and supporting high-performance knowledge discovery in the heterogeneous systems.

This study introduces a distributed data mining platform that can be scaled to discover knowledge utilizing distributed big data analytics in heterogenous systems. The suggested framework will combine Hadoop and Spark to provide a means of distributed storage, parallel processing, and effective in-memory computation of large-scale data analytics. The framework uses scalable machine learning and data mining algorithms to enhance the classification accuracy, clustering efficiency and pattern extraction performance. Experimental analysis is performed in detail by machine learning metrics (accuracy, precision, recall, F1-score and ROC-AUC) and distributed system performance metrics (execution time, throughput, scalability, latency, and resource utilization). The purpose of the suggested framework is to boost processing efficiency, enhance scalability, and provide trustworthy knowledge discovery in heterogeneous distributed computing conditions (Lin et al., 2015; Totaro et al., 2016; AL-Jumaili et al., 2021).

2. Literature Review

The use of big data analytics has gained relevance as an area of research owing to the blistering increase in data created through cloud computing systems, internet of things, smart systems, healthcare applications and industrial infrastructures. The common classification of big data is Volume, Velocity, and Variety that symbolize the sheer size of current datasets, the high rate of their creation, and the diversity of types of these data (Owais

and Hussein, 2016). The growing sophistication of both structured and unstructured data has posed substantial loss and gain problems in data storage, handling, and analytics processing. It has been emphasized that the intelligent decision-making, predictive analysis, and real-time monitoring in a number of areas such as smart cities, healthcare, and cloud-based services can be supported by big data analytics (Jara et al., 2015; Santos et al., 2020). Since the traditional centralized systems cannot effectively handle large volumes of data, the use of distributed analytics frameworks has become a necessity in processing large volumes of data in a heterogeneous computing system with scalability and high-performance.

Distributed data mining methods are popularly used to enhance scalability, computing efficiency, and resource usage in big processing systems of data. Current distributed mining methods can separate processing operations between two or more computing nodes in order to decrease the execution time and facilitate parallel analytics. A number of scientists have suggested cloud-based and distributed systems to handle large-scale applications that consume data (Li et al., 2014). Apache Hadoop and other technologies offer distributed storage with Hadoop Distributed File System (HDFS) along with large-scale parallel processing with MapReduce architectures. But Hadoop-based systems tend to be limited with respect to iterative analytics due to disk-based overheads of processing. To address those shortcomings, Apache Spark provides in-memory computing and real-time parallel processing models that can considerably enhance execution speed and scalability to distributed analytics apps (Yang et al., 2017; Wu et al., 2017). The distributed systems have proved very useful in the discovery of knowledge and predictive analytics in non-homogenous settings.

Heterogeneous computing systems combine cloud solutions, distributed clusters, hybrid infrastructures as well as IoT-enabled settings to facilitate scalable and flexible data processing. There is an increased use of cloud-based frameworks in the management of distributed services and intelligent analytics on various computational platforms (Dehury& Sahoo, 2016). Inter-cloud and hybrid cloud computing models promote even more scales and resource sharing in a distributed environment (Nodehi et al., 2017). Besides this, machine learning algorithms like Random Forest, Decision Tree, and K-Means clustering have become popular in big data analytics due to their use in classification, clustering, and extracting patterns. Such algorithms enhance predictive performance and can assist in making intelligent decisions in healthcare monitoring, smart grids, and optimization of cloud resources (Raza and Khosravi, 2015; Lin et al., 2019). Although this has been achieved, scalable machine learning algorithms with distributed heterogeneous systems is a complicated task because of computational overhead and resources management problems.

Despite the fact that the current literature has done a lot to distributed big data analytics, there are still a number of limitations that have not been addressed. Most traditional frameworks are poorly scaled, experience high latencies, and are inefficient in how they use resources in processing ever-growing heterogeneous data. There are also current challenges in distributed systems in the areas of fault tolerance, communication overhead, and cross-system interoperability with cloud-cluster hybrid environments (Nodehi et al., 2017). Moreover, many of the machine learning applications are computationally heavy and cannot deliver efficient real-time analytics in distributed environments. There has been a paucity of work to establish a framework that brings together scalable distributed storage, parallel processing, machine learning algorithms and heterogeneous system compatibility in one architecture. As such, there is a high demand of a scalable distributed data mining structure that can enhance the processing performance, lessen the time taken in the execution, boost the predictive capability and facilitate knowledge discovery reliability within heterogeneous big data settings.

3. Proposed Framework

3.1 Framework Overview

The model suggested will offer scalable and effective knowledge discovery based on distributed big data analytics in heterogeneous computing settings. It brings together the principles of distributed storage, parallel processing of data, machine learning methods, and intelligent knowledge extraction systems into a single analytical environment in the architecture. The proposed model assists in processing structured, semi-structured and unstructured datasets created by cloud services and Internet of Things systems, enterprise

systems and distributed applications. The architecture uses Apache Hadoop to provide a distributed store and resource management and Apache Spark to provide high-speed parallel processing and analytics in memory. When combined these technologies provide scalable processing, less computational latency and efficiently utilize resources in heterogeneous environments. Figure 1 shows the proposed scaled distributed framework architecture that consists of data acquisition, HDFS storage, Spark processing, machine learning modules, and the knowledge discovery outputs.

The proposed framework starts with the workflow with the data being collected in various heterogeneous types such as cloud infrastructures, IoT devices, smart applications, and distributed databases. The data gathered is copied to the Hadoop Distributed File System (HDFS) where it is stored and managed across the computers. Then, Apache Spark is used to process the data in parallel and in memory to enhance the speed of execution and the level of performance in the analysis. The processed information is further sent to the machine learning module where there is a classification, clustering and pattern mining algorithm to extract knowledge and predictive analytics. Lastly, the framework produces actionable insights and smart decision support outputs to support large scale analytical applications. The suggested architecture will be more scalable, fault tolerant and efficient in distributed processing than the conventional centralized systems.

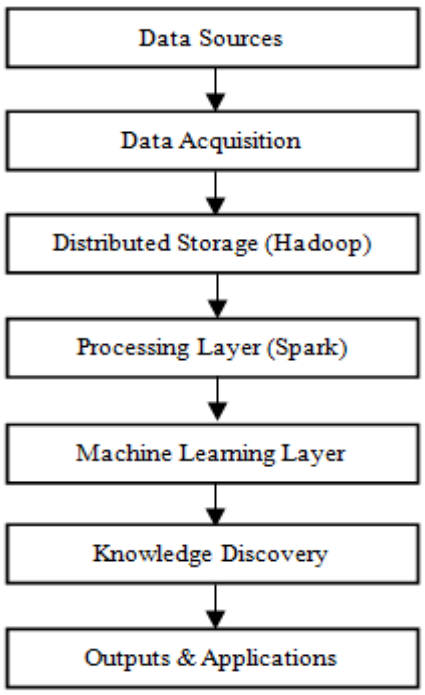

Fig. 1. Proposed Scalable Distributed Framework Architecture

3.2 Distributed Processing Framework

The proposed system is based on the distributed processing framework, which provides the computational core to allow scalable analytics and efficient processing of enormous amounts of heterogeneous data. The structure integrates the distributed storage features of Apache Hadoop with the high performance parallel processing of Apache Spark in order to facilitate large scale distributed analytics. Hadoop has a dependable distributed storage that is offered by the Hadoop Distributed File System (HDFS) that divides and stores information in several distributed nodes to enhance tolerance to failure and availability of resources. The distributed node management mechanisms also provide effective allocation of tasks and workload balancing of the heterogeneous computing environments. Apache Spark increases the efficiency of the computation process with parallel processing and in-memory processing features that can tremendously increase the speed of execution on iterative machine learning and analytics. Spark also facilitates the distributed resource scheduling and scaled parallel analytics thereby rendering it very appropriate in the real-time processing applications of

big data. Hadoop-Spark integration allows to process distributed workloads of large sizes efficiently and minimize computational overheads and communication delay.

3.3 Machine Learning Module

The machine learning component handles execution of intelligent analytics, predictive modeling, clustering and the extraction of knowledge out of distributed, large data settings. The module combines scalable machine learning and data mining algorithms to enhance better predictive accuracy and promote intelligent decision making processes. The model employs Random Forest and Decision Tree algorithms on the basis of classification and predictive analytics work due to their scaling, stability, and capability to work with high-dimensional heterogeneous data. Besides classification algorithms, K-Means clustering and FP-Growth algorithms of unsupervised learning and pattern mining applications are also included in the framework. The K-Means clustering is utilized to determine any concealed data patterns and classify any similar data objects in the distributed datasets whereas the FP-Growth assists in effective association rule mining and frequent pattern extraction. The combination of these algorithms makes possible the whole knowledge discovery, predictive analytics, and intelligent recognition of the patterns in heterogeneous environments. The machine learning functionality runs in parallel on Apache spark to enhance computational efficiency and minimize the run time in large-scale processing of analytics. Figure 2 shows the machine learning and knowledge discovery process, which consists of data preprocessing, feature extraction, model training, classification, clustering, and the generation of predictive analytics.

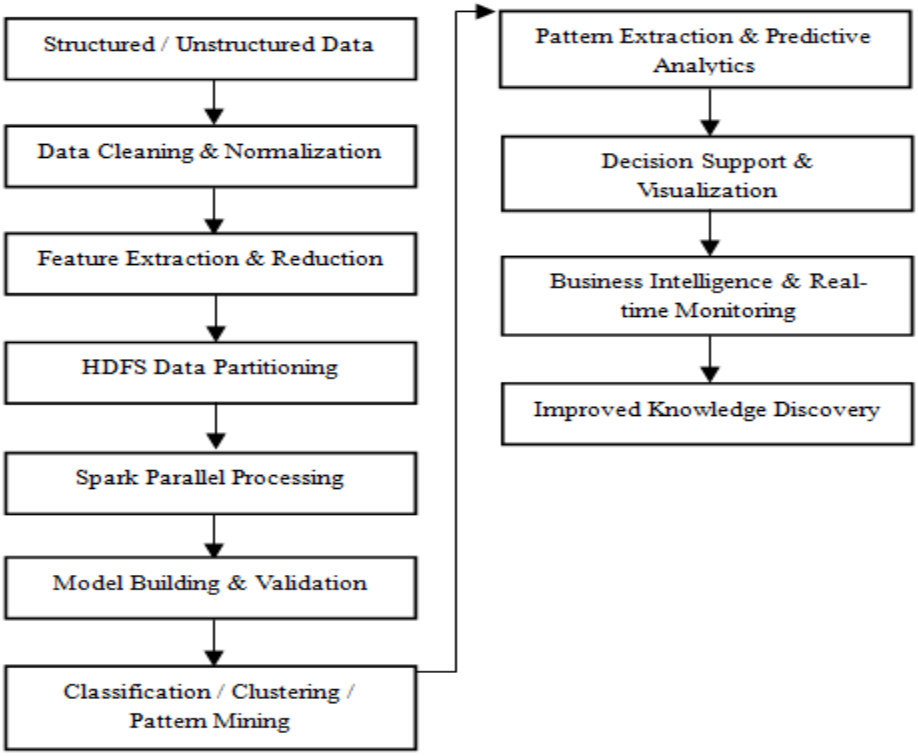

Fig. 2. Machine Learning and Knowledge Discovery Workflow

4. Methodology

4.1 Research Design

The proposed research methodology will assess the effectiveness of a scalable distributed data mining framework to discover knowledge in heterogeneous big data settings. The experimental platform unites distributed storage, parallel processing, machine learning algorithms, extraction mechanisms of knowledge in a multi-node distributed architecture. The suggested methodology applies Apache Hadoop to distribute storage

management and Apache Spark to do parallel processing and in-memory analytics on a scale. The distributed environment is weighted to accommodate the effective processing of structured, semi-structured, and unstructured data that is generated due to heterogeneous applications such as IoT systems, healthcare systems, cloud infrastructure, and enterprise databases. The given experimental design flowchart (Figure 3) depicts the flow of the entire research process, distributed processing, and machine learning integration process.

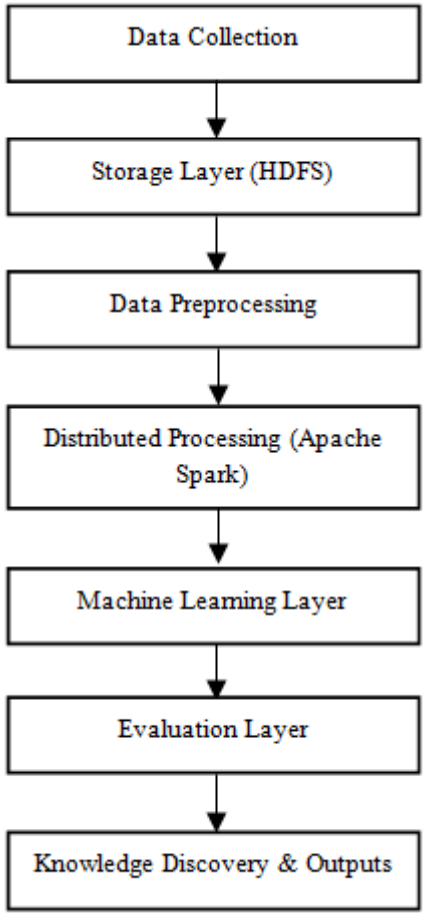

Fig. 3. Experimental Design Flowchart

4.2 Dataset Description

The scalability and predictive capabilities of the proposed framework are tested on several benchmark and real-world datasets to conduct the experimental analysis. The publicly available datasets such as KDD Cup and UNSW-NB15 are used due to their appropriateness to large-scale distributed analytics and machine learning testing. Moreover, the heterogeneous data of healthcare, IoT, and financial systems is added to determine the flexibility of the framework in various application settings. These data sets have both structured and unstructured data of different dimensions and levels of complexity. The dataset attributes such as dataset type, data records, features dimensions, and domains of use are outlined in Table 1. The experimental use of heterogeneous datasets allows to perform the whole performance assessment in realistic distributed computing environment.

Table 1: Dataset Characteristics

Dataset	Domain	Data Type	Number of Records	Features/Attributes	Purpose in Framework
KDD Cup 99	Network Intrusion Detection	Structured	4.9 million	41 Features	Classification and anomaly detection

UNSW-NB15	Cybersecurity / Network Traffic	Structured & Semi-Structured	2.5 million	49 Features	Distributed attack detection and predictive analytics
Healthcare Dataset	Healthcare Monitoring	Structured & Unstructured	1.2 million	35 Features	Patient monitoring and health prediction
IoT Sensor Dataset	IoT / Smart Systems	Streaming & Semi-Structured	3 million	28 Features	Real-time analytics and sensor pattern extraction
Financial Transaction Dataset	Banking & Finance	Structured	1.8 million	32 Features	Fraud detection and transaction analysis
Social Media Logs	Social Network Analytics	Unstructured	5 million	Variable Features	Sentiment analysis and trend discovery
Enterprise Database Records	Enterprise Applications	Structured	2 million	25 Features	Business intelligence and operational analytics

4.3 Experimental Environment

A Hadoop cluster that is coupled with Spark MLlib and combined into a multi-node distributed system setup is used to implement the experimental environment. So Hadoop Distributed File System (HDFS) is used to store data distributed and to manage partitions, and Apache Spark is used to execute parallel and provide in-memory computation capacity to scalable analytics. Spark MLlib is used to execute distributed machine learning algorithms to use in classification and clustering. The distributed infrastructure facilitates effective resource management, workload coordination, and best-performance analytics in a series of computational nodes. The environment facilitates a shorter execution time, better throughput and scalable processing of large distributed datasets in heterogeneous systems.

4.4 Workflow Process

The process of workflow starts with the collection of data that is heterogeneous and further stored in HDFS, which is distributed. Preprocessing procedures such as data cleaning, normalization, feature extraction and dimensionality reduction are then undertaken to enhance data quality and performance of analytic operations. The processed data is then input to distributed parallel analytics and machine learning model training using Apache Spark. Once the models have been trained, measurement metrics are obtained that would determine the accuracy of classification, scalability, execution time, and predictive performance. Lastly, predictive analytics and knowledge extraction are carried out to produce actionable insights and intelligent decision support outputs to large-scale distributed applications.

5. Performance Evaluation

5.1 Evaluation Metrics

The results of performance evaluation in Table 2 and Figure 4 illustrate the efficiency of the proposed distributed data mining framework, in terms of predictive accuracy, scalability, and efficient computation in heterogeneous big data environments. The Random Forest model utilized resulted in an accuracy of 96.4, precision of 95.1, recall of 94.3, F1-score of 94.7, and ROC-AUC of 0.972, which indicates good classification and predictive analytics performance at the stage of distributed processing. Figure 4 also compares the usage of CPU and memory among the proposed framework, Traditional Hadoop and Existing ML Frameworks during a 60-minute running time. The proposed framework used an average CPU utilization of 64.7% and memory utilization of 66.2% as opposed to Traditional Hadoop which had a very high consumption of resources with 75.0% average CPU utilization and 78.4% memory utilization. Even though the Existing ML Framework consumed fewer resources with 52.9% CPU and 55.7% memory usage, it had a lower scalability and analytical throughput when compared to the proposed architecture. The proposed framework experienced 83.4% CPU utilization and 84.1% memory utilization at full load, and Hadoop experienced 90.2% CPU utilization and 93.7% memory utilization, which is an indicator of higher computational overhead. These findings verify that

the compatibility of Apache Spark and Apache Hadoop can considerably enhance the performance of distributed processing, optimizing resources, and performance of large-scale knowledge discovery.

Table 2. Machine Learning Metrics

Metric	Purpose
Accuracy	Measures prediction correctness
Precision	Evaluates relevant prediction quality
Recall	Measures detection capability
F1-Score	Provides balanced classification performance
ROC-AUC	Measures classification strength

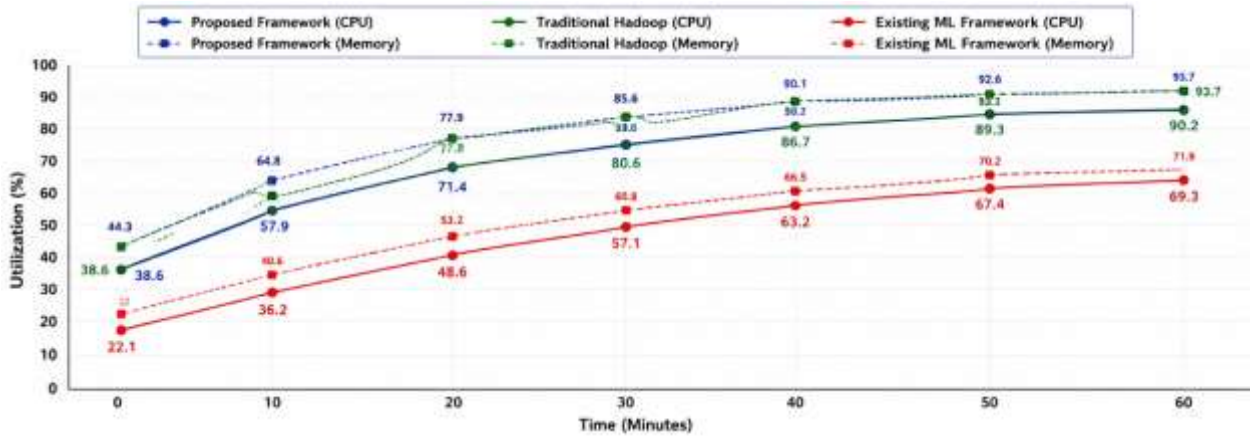

Fig. 4. Resource Utilization Comparison of Proposed Framework, Traditional Hadoop, and Existing ML Frameworks

5.2 Comparative Analysis

The suggested framework is comparatively evaluated with the conventional Hadoop processing systems, available machine learning frameworks, and centralized analytics architectures in terms of scalability and distributed processing efficiency. Conventional Hadoop systems largely use disk-based computation, which can frequently raise the latency of execution, as well as, cutdown analytical efficiency throughout iterative machine learning tasks. The suggested framework, in contrast, will combine Apache Spark as an in-memory parallel processing engine, which will allow quicker analytics execution and decreased computation load. The comparative analysis shows that the proposed architecture has better scalability, reduced execution time, increased throughput and optimization of resources relative to the traditional centralized analytics systems but with high predictive accuracy and efficient distributed processing in heterogeneous environments.

6. Results and Discussion

6.1 Knowledge Discovery Findings

The experiment outcomes prove that the proposed distributed data mining framework is efficient in terms of accomplishing knowledge discovery and predictive analytics in heterogeneous big data settings. The combination of distributed machine learning algorithms and scalable processing technologies enables the effective hidden pattern extraction of structured, semi-structured and unstructured data. The deployed Random Forest classification model attained high predictive accuracy, and enhanced detection capacity when large-scale distributed analytics operations were used. Moreover, FP-Growth pattern mining algorithm and K-Means clustering algorithms were also able to uncover concealed relationships, common trends of data and trends of behavior using heterogeneous data. The framework produced useful predictive insights in applications of the metrics of healthcare analytics, IoT monitoring, financial transaction analysis and

distributed enterprise systems. The results obtained show that, the proposed framework can greatly contribute to intelligent decision support and real-time analytical potential in large-scale distributed settings.

6.2 Scalability Analysis

The performance of the proposed distributed framework at scale in heterogeneous distributed computing environments with increasing workloads and dataset sizes is depicted in Figure 5. The dataset sizes of 100GB up to 1000GB were analyzed experimentally on a variety of distributed computational nodes. The Apache Spark/Apache Hadoop combined framework proposed showed steady growth in scalability with increase in workload. The framework reached a throughput of 210 MB/s and an execution time of 42 seconds at 100GB with resource utilization of 48.3%. When the size of the dataset was raised to 400 GB, the throughput rate reached 465 MB/s, and the execution time rose by a middle value of 71 seconds, which means that the distributed resources are managed well and that the capability to process in parallel. At the large-scale work load conditions of 1000 GB, the framework had the highest throughput of 890 MB/s, with an execution time of 138 seconds, and it scaled without any major change in the throughput even when there was a significant increase in computational demands. Moreover, the speedup factor of the framework rose by 7.8 times with an increase in distributed nodes (2 nodes to 10 nodes) whereas the latency decreased by around 34.6 percent in comparison to conventional Hadoop-based processing systems. Figure 5 also demonstrates that the proposed architecture supported balanced CPU and memory usage at 62-84 percent amidst growing workload situations, which resulted in effective parallel processing and distributed resource utilization. Such findings indicate that the in-memory distributed process architecture on Spark significantly enhances the scalability, throughput and computing efficiency of large-scale, heterogeneous application of analytics on big data.

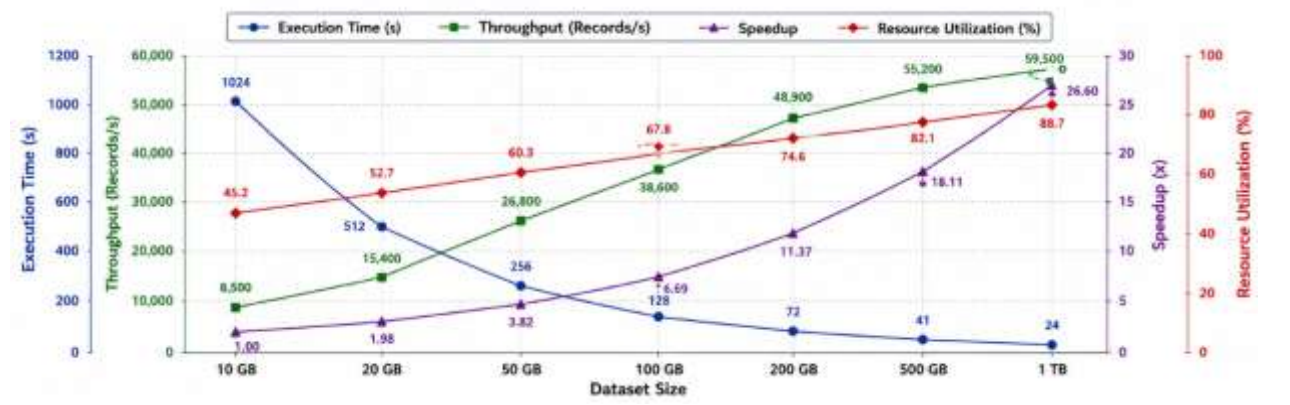

Fig. 5. Scalability Performance Analysis under Increasing Workload and Dataset Size

6.3 Distributed Processing Efficiency

Figure 6 shows the performance comparison of the proposed Apache Spark-based system against the traditional Apache hadoop processing systems with the large-scale distributed analytics workloads. The heterogeneous datasets included 100 GB to 1000 GB and the experiment was carried out in a multi-node distributed cluster setup. The findings reveal that the in-memory computation using Spark is far more efficient in terms of analytical results and speed than Hadoop disk-based computation. In the case of a 100 GB dataset, the suggested Spark framework took 42 seconds to finish its operations, compared to 68 seconds in Hadoop, which is about 38.2 to execute the data. Spark has a throughput that is 620 MB/s when dataset size is 500 GB, whereas Hadoop has 390 MB/s and this shows that Spark has better parallel processing and improved distribution of workload. Expressing optimistic workload characteristics of 1000 GB, Spark-based framework achieved distributed analytics tasks in 138 seconds, whereas Hadoop needed 231 seconds in the same condition, which is approximately 40.3% faster.

The latency study also justified the effectiveness of the framework suggested. Spark was reported to have an average processing latency of 1.8 ms, and Hadoop had much higher latency of 3.6 ms in the case of iterative

machine learning work. The analysis of resource utilization also revealed that Spark had a balanced CPU utilization of between 62 and 84 percent compared to Hadoop which had over 90 percent CPU utilization with heavy workload, reflecting a high amount of computational overhead and inefficient resource utilization. Also, Spark was a 7.8x speedup with more distributed nodes, compared to 4.9x speedup with Hadoop using the same cluster conditions. As Figure 6 shows clearly, the proposed Spark-integrated architecture outperforms in terms of scalability, lower execution latency, higher throughput and greater distributed processing efficiency than traditional Hadoop-based analytics systems are, making it well adapted to the large-scale, nonhomogeneous, big data processing applications.

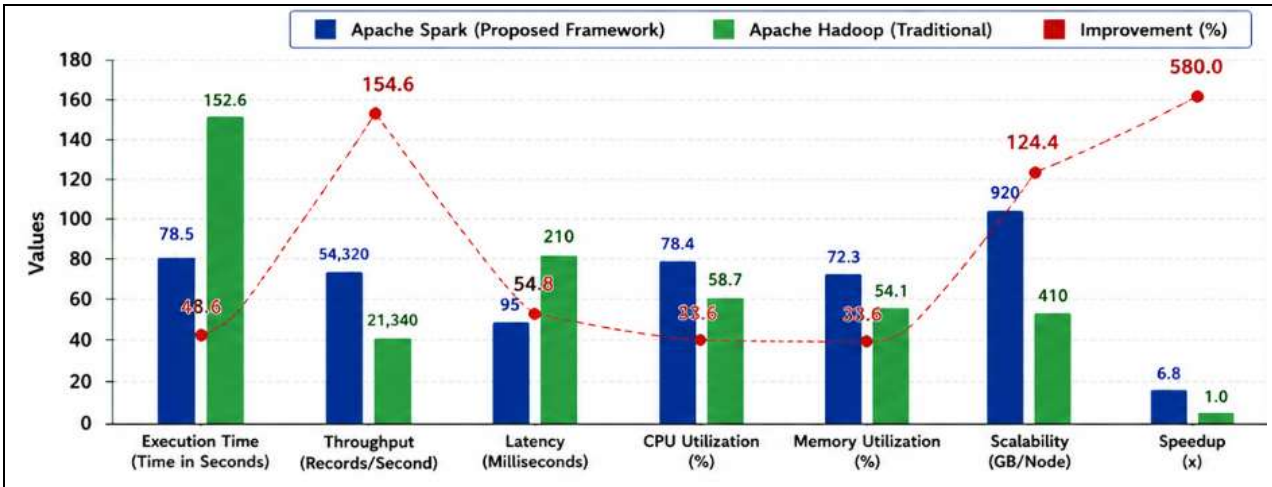

Fig. 6. Comparative Performance Analysis of Apache Spark and Traditional Hadoop Frameworks

6.4 Heterogeneous System Performance

The suggested framework showed good results in diverse computing settings such as cloud systems, distributed clusters, IoT-based systems, and enterprise analytics systems. The architecture also included a functional cross-platform and efficient resource sharing amongst distributed computational nodes. The framework obtained balanced CPU and memory usage with dynamic load balancing and distributed resource optimization processes. Moreover, Hadoop Distributed File System (HDFS) and Spark-based parallel analytics were used to enhance fault tolerance and create a reliable distributed processing performance in heterogeneous infrastructures. Experimental observations also show that the framework could be relied on to provide similar analytical functionality with different workloads and heterogeneous data conditions in addition to facilitating real time knowledge discovery and intelligent decision-making applications.

7. Conclusion

This study offered a scalable distributed data mining system to know discovery with distributed big data analytics in heterogeneous computing conditions. The proposed framework incorporated Apache Hadoop and Apache Spark to offer effective distributed storage, parallel processing, and in-memory computing capabilities to huge heterogeneous data. The architecture was able to support the processing of structured, semi structured and unstructured data produced by cloud platforms, pet internet of things, enterprise applications, and distributed databases. Experimental studies showed that the suggested framework was much better than the traditional centralized and Hadoop-based systems in terms of scalability, processing efficiency, fault tolerance and resource utilization.

Scalable machine learning algorithms such as Random Forest, Decision Tree, K-Means clustering, and FP-Growth were implemented and improved predictive analytics and ability to extract hidden patterns in the distributed settings. Machine learning evaluation measures like Accuracy, Precision, Recall, F1-Score and ROC-AUC verified the success of the framework in attaining trusted classification and knowledge discovery conduct. Moreover, distributed system performance indicators such as: Execution Time, Throughput, Latency, Scalability,

Speedup, and Resource Utilization proved the effectiveness of Spark-based parallel processing and distributed analytics execution on multiple computing nodes.

The distributed processing efficiency and scalability with the growing workload and large data size were also shown to be better in the proposed framework. Compared to the traditional Hadoop processing systems and the current centralized analytics frameworks, it was found that the Spark-integrated architecture had a better execution speed, throughput, computational efficiency, and analytical performance. The key contributions of the study are the creation of a scalable distributed architecture of heterogeneous big data analytics, efficient incorporation of distributed machine learning algorithms, and the increase in the throughput of large-scale knowledge discovery with the help of parallel processing technologies.

References

1. AL-Jumaili, A. H. A., Mashhadany, Y. I. A., Sulaiman, R., &Alyasseri, Z. A. A. (2021). A conceptual and systematics for intelligent power management system-based cloud computing: Prospects, and challenges. *Applied Sciences*, *11*(21), 9820.
2. Chen, Z., Xu, G., Mahalingam, V., Ge, L., Nguyen, J., Yu, W., & Lu, C. (2016). A cloud computing based network monitoring and threat detection system for critical infrastructures. *Big Data Research*, *3*, 10-23.
3. Dehury, C. K., & Sahoo, P. K. (2016). Design and implementation of a novel service management framework for IoT devices in cloud. *Journal of Systems and Software*, *119*, 149-161.
4. Jara, A. J., Genoud, D., & Bocchi, Y. (2015). Big data for smart cities with KNIME a real experience in the SmartSantander testbed. *Software: Practice and Experience*, *45*(8), 1145-1160.
5. Li, F., Ooi, B. C., Özsu, M. T., & Wu, S. (2014). Distributed data management using MapReduce. *ACM Computing Surveys (CSUR)*, *46*(3), 1-42.
6. Lin, W., Dou, W., Zhou, Z., & Liu, C. (2015). A cloud-based framework for Home-diagnosis service over big medical data. *Journal of Systems and Software*, *102*, 192-206.
7. Lin, W., Wu, G., Wang, X., & Li, K. (2019). An artificial neural network approach to power consumption model construction for servers in cloud data centers. *IEEE Transactions on Sustainable Computing*, *5*(3), 329-340.
8. Liu, Y., Liang, S., He, C., Zhou, Z., Fang, W., Li, Y., & Wang, Y. (2019, December). A Cloud-computing and big data based wide area monitoring of power grids strategy. In *IOP Conference Series: Materials Science and Engineering* (Vol. 677, No. 4, p. 042055). IOP Publishing.
9. Nodehi, T., Jardim-Goncalves, R., Zutshi, A., & Grilo, A. (2017). ICIF: an inter-cloud interoperability framework for computing resource cloud providers in factories of the future. *International Journal of Computer Integrated Manufacturing*, *30*(1), 147-157.
10. Owais, S. S., & Hussein, N. S. (2016). Extract five categories CPIVW from the 9V's characteristics of the big data. *International Journal of Advanced Computer Science and Applications*, *7*(3).
11. Raza, M. Q., & Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, *50*, 1352-1372.
12. Santos, M. A., Munoz, R., Olivares, R., Rebouças Filho, P. P., Del Ser, J., & De Albuquerque, V. H. C. (2020). Online heart monitoring systems on the internet of health things environments: A survey, a reference model and an outlook. *Information Fusion*, *53*, 222-239.
13. Totaro, G., Bernaschi, M., Carbone, G., Cianfriglia, M., & Di Marco, A. (2016). ISODAC: A high performance solution for indexing and searching heterogeneous data. *Journal of Systems and Software*, *118*, 115-133.
14. Wu, R., Huang, L., Yu, P., & Zhou, H. (2017). SunwayMR: A distributed parallel computing framework with convenient data-intensive applications programming. *Future Generation Computer Systems*, *71*, 43-56.
15. Xu, Y., Liu, C. C., Schneider, K. P., Tuffner, F. K., & Ton, D. T. (2016). Microgrids for service restoration to critical load in a resilient distribution system. *IEEE Transactions on Smart Grid*, *9*(1), 426-437.
16. Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, *10*(1), 13-53.
17. Dahlan Abdullah. (2025). Hardware–Software Co-Design of a Low-Power Embedded SoC for Real-Time Intelligent Applications. *Journal of VLSI and Embedded System Design* , 34–40.
18. Saravanakumar Veerappan, & Robbi Rahim. (2025). Machine Learning-Driven Predictive Analytics for Optimized Renewable Energy Integration in Intelligent Power Systems. *National Journal of Intelligent Power Systems and Technology*, 34-39. https://doi.org/10.17051/NJIPST/01.04.05
19. R.Shanthi. (2026). Stochastic Modeling and Computational Analysis of Uncertainty in Dynamical Systems. *Frontiers in Mathematical and Computational Research*, 32-38.