Hadoop-Based Political and Ideological Big Data Platform Architecture and Pattern Mining in Higher Education

Aysar Alrababah

Dept of Computer Science
Universiti Sains Malaysia (USM), Malaysia

ABSTRACT

There is a growing opportunity for self-improvement within the realm of higher education, represented by the combination of big data technologies and higher education. The Education Big Data Processing Framework (EBDPF) is the name of the platform that is the subject of this article, which focuses on the development of an advanced big data processing platform. In order to process huge datasets in an effective manner, the platform makes use of Hadoop's big data storage architecture, in addition to Hive, Flume, and Sqoop for data collection and synchronization, respectively. In this study, the implementation of standard data mining techniques is investigated through the use of MapReduce programming. More specifically, the study investigates the execution efficiency and scalability of these algorithms within the Hadoop platform. The data clustering task in data mining is chosen as a representative problem in this study so that the effectiveness of the EBDPF may be evaluated accordingly. An implementation of the clustering job that is based on MapReduce is created and tested on the Hadoop platform. Extensive experiments are conducted with a range of cluster sizes and data sizes. The findings provide more evidence that Hadoop distributed systems are more efficient and successful than traditional methods when it comes to managing data mining applications. The extended performance analysis of computing power not only provides evidence that the platform is currently effective, but it also provides a signal that it has a significant amount of potential that has not yet been utilized. At the convergence of big data technology and higher education, the EBDPF emerges as a strong and promising framework, demonstrating considerable steps in accelerating data processing within educational contexts. In conclusion, the EBDPF is a framework that demonstrates significant progress.

Keywords: Higher Education; Big Data Technologies; Hive; Flume; Sqoop; Data Collection; Data Synchronization; Hadoop Platform; Pattern Mining Applications; Educational Contexts.

1. Introduction

The implementation of big data technology has brought about a revolution in a variety of fields, including education, in this era of digital transformation. Large-scale data analytics are becoming increasingly utilized by educational institutions of higher learning in order to improve the learning experiences of students, boost the efficiency of the institutions, and make decisions based on the data [1]. As a result of this paradigm shift toward data-driven decision-making in higher education, novel platforms and systems have been developed. These platforms and systems are able to process and analyze huge volumes of data in order to extract significant

https://doi.org/10.70023/piqm24125

insights. This study proposes a cutting-edge project that intends to deliver first-rate big data risk control services in the field of college ideology and politics [2].

The University Ideological and Political Big Data Platform is the concept that is being discussed in this study. This platform provides a comprehensive solution for gathering, processing, and analyzing diverse datasets in order to assist risk control measures in the higher education sector [3]. This is in response to the expansion of digital data sources relating to students across a variety of fields. The University Ideological and Political Big Data Platform is built on a comprehensive design that includes important modules such as data collecting, processing and storage, data synchronization, and system scheduling [4]. This architecture serves as the platform's foundation. Each of these modules collaborates with the others to guarantee a smooth flow of data, effective processing, and rapid analysis of vital information concerning the philosophy and politics of the college that is being discussed [5].

The platform provides a solid architecture for managing large-scale data operations with precision and scalability. This is accomplished through the integration of modern technologies like as Hadoop, Flume, Sqoop, Oozie, and Hue [6]. The mining of regular sequence behavior patterns among college students is one of the key goals of the platform, which is one of the primary aims of the platform. Through the utilization of sophisticated algorithms and data mining techniques, the platform is able to unearth significant information into the sequential behavior of students who access online resources [7].

For the purpose of forecasting future student behaviors, providing targeted interventions, and improving overall student success rates, this information serves as an essential knowledge foundation [8]. In addition, the platform's commitment to user-centric design and operational efficiency is highlighted by the fact that it places a strong emphasis on visualization, ease of use, and maintenance. Through the utilization of technologies such as HUE and Oozie, users are able to create workflows in an understandable manner, monitor progress, and resolve issues through the use of a graphical user interface. This user-friendly approach not only improves the accessibility of the platform, but it also gives administrators and developers the ability to streamline their jobs and improve the efficiency of the system [9].

For the purpose of data collection, the platform utilizes a two-layer Flume agent structure, which is in accordance with the best practices that are prevalent in the market [10]. This structure guarantees the smooth transmission of data from monitoring directories to the Hadoop Distributed File System (HDFS). This architecture not only improves the dependability of data and the efficiency with which it is stored, but it also creates the framework for future scalability and extension as the platform evolves to meet evolving requirements. It is impossible to overestimate the relevance of the platform's architecture taking into account the processing and storage of data [11]. With a particular emphasis on high-efficiency offline processing of huge datasets and robust storage dependability, the platform makes use of the capabilities offered by the Hadoop framework in order to fulfill these essential needs.

The platform ensures a solid basis for data processing, analysis, and storage by adopting Hadoop as the fundamental storage and computing system [12]. Additionally, it addresses the challenges of sustaining complex MapReduce programs by solving these challenges. In addition, the platform's scalability and performance are bolstered by its distributed system architecture. This is demonstrated by comparison studies that demonstrate the system's effectiveness in processing data at a variety of scales. The capability of the platform to grow linearly with the volume of data and the number of processing nodes demonstrates its promise for managing large-scale data activities with the highest possible efficiency and the least amount of performance loss. The most important contributions are discussed in this article.

- ✓ To create a powerful big data processing platform using Hadoop, Flume, Sqoop, and other tools to analyze enormous collegiate philosophy and politics datasets.
- ✓ To implement an EBDPF module to collect different data from sources and easily transfer it to Hadoop utilizing Flume and Sqoop for synchronization and storage.
- ✓ To establish EBDPF data storage and processing for high-volume data processing to ensure dependability and optimize Hadoop storage.
- ✓ To analyze sequential behavior data to identify patterns, predict future behaviors, and enhance risk control measures among college students.

A summary of the research is provided below. In Section 2, the current literature and study techniques are thoroughly examined. The research strategy, methodology and processing procedures are detailed in 3rd Section. The results analysis is covered in 4th Section. Part 5 explores the main conclusion and Future work.

2. Research Methodology

Popescu et al. [13] investigated the application of large data to sentiment analysis, with a focus on ML algorithms such as Support Vector Machines and Naive Bayes. A Hadoop-based architecture allowed for efficient data handling in the research, which made use of a big collection of social media posts. Across all scenarios, the results demonstrated an 89% success rate in predicting future public opinion trends and a 92% success rate in sentiment classification. The study sheds light on the practicality and effectiveness of sentiment analysis on a large scale, offering useful information for companies, politicians, and social scientists.

Sedek et al. [14] provided with a residential dashboard is currently being developed by the Student Affairs Division of UiTM Perlis in order to address the deficiency of complete residential data that exists inside the e-Kolej system itself. A more informed decision-making process and more effective management of student housing will be facilitated by the dashboard. Apache Hive, HiveQL, and Microsoft Power BI are currently being utilized in the project. Positive comments were received on the dashboard, which obtained an average score of 4.71 rating.

Gattoju and Nagalakshmi [15] offered with Hadoop is a vast software framework that is created for the purpose of managing large amounts of data, which may include real-time data, social networking, and money laundering. Unauthorized access, on the other hand, is a significant problem. A revolutionary ChaApache framework is proposed by the author in order to achieve the goals of securing the Hadoop application, reducing error rates, and improving processing times. A framework that is implemented in Python and encrypts 512 bits of data is compared to existing replicas in terms of the amount of time it takes to compute, the amount of resources it uses, the rate at which it shares data, and the speed at which it encrypts data.

Wang et al. [16] suggested an Early Warning Method for College Students' Abnormal Behavior Using Multimodal Fusion and an Improved Decision Tree (EWMABCS-MFIDT). In addition to addressing concerns with discreteness and sparseness in campus big data, it seeks to enhance the precision of early warnings from the beginning. Due to the fact that it improves educational work that is targeted, personalized, and predictive, the EWMABCS-MFIDT approach is an extremely useful instrument for tackling the difficulties that are associated with campus big data

Lv [17] suggested the use of big data analytics applications is increasing the value of healthcare, which is contributing to the growing trend of informatization in the medical and communication fields. The open-source framework known as Apache Hadoop, which is managed by the Apache Foundation, is well-known for its fault tolerance and its ability to provide access to an unlimited amount of computer resources. The study makes use of qualitative analysis to explore the advantages and disadvantages of cloud computing in the field of medical construction and data mining, as well as the achievements of Hadoop and the significance of association rule mining algorithms in the field of medical cloud computing and clinical diagnosis.

Wang [18] proposed an Smart piano gloves are new. The gloves' multiinertial sensors assess piano players' motions in real time, enabling piano learners to identify conventional gestures and improve learning efficiency and interest while cutting costs. Piano movements are unique in their diversity, speed, dynamism, and time-varying nature compared to other applications. We present a non-contact gesture recognition system design strategy. Combining infrared and ambient light sensors Si1143 with capacitive touch-sensitive microcontrollers C8051F700 and C8051F800 enables non-contact gesture identification for multiple activities, gesture detection, and target object distance calibration.

He et al. [19] recommended that a cloud-based platform for high-quality teaching resources of higher vocational sports be developed in order to address the issue of high packet loss rates in data transmission. In order to construct a shared information hierarchy model, the platform makes use of blockchain technology. Additionally, it eliminates intermediate linkages and set up a trustworthy storage mechanism. In order to achieve the lowest possible packet loss rate and the highest possible application value, the method incorporates data into a central database.

Li et al. [20] offered the healthcare sector is currently facing an increase in the number of patients, which calls for improved hospital administration and more health monitoring systems. The ever-increasing upkeep of the dataset makes it harder to maintain, which brings to the necessity of implementing big data solutions. Due to the fact that it generates a significant amount of revenue, the industry is confronted with a large potential threat from attackers. In this paper, the challenges that are faced by the big data industry in terms of security are discussed, and a blockchain-based method is proposed as a solution to these problems.

Monsma [13] showcased the Rocket League, a wildly successful esports game that has amassed a devoted fanbase of more than 90 million individuals. Professional teams must continuously strive for performance improvement if they want to remain relevant. To provide the groundwork for data-driven performance in Rocket League, a framework has been built to apply best practices in notational analysis. The goal, components, and success criteria of a notational analysis system are defined by the framework. Through discussions with seasoned researchers in esports performance analysis, it has been tested and refined, and areas of attention have been highlighted for future revisions.

Kovalchik and Stephanie [14] discussed how the explosion of player tracking data has changed the game in a number of sports by opening the door to new ways of measuring performance and assigning points for individual plays. Sports may expand analytical methods in numerous fields, as shown in this study, which emphasizes major contributions from statistical and machine learning methodologies. But it also talks about the methodological problems with monitoring data use in sports, which are still there. Sports have the ability to be a testing ground for new analytical methods, as this study has shown.

Ke et al. [15] displayed the standardised process of building machine learning models for basketball rosters(MLM-BR) ; the NBA and the WNBA have collaborated to create a new framework. The system classifies players and builds a model of the squad using principle components analysis (PCA) and a basic neural network. With it comes a rating system that takes into account things like player performance patterns and their ability to execute under duress.

Using various weighting techniques, the framework found ten player groups, four of which were deemed elite.

Ziyi et al. [16] suggested a technique of comparison analysis for the purpose of assessing the trajectories of several agents in ball games. A trajectory analysis with neural network (TA-NN) strategy that is based on attention mechanism is used in order to identify unique segments in the trajectories of classes that have been provided. Through the highlighting of segmented trajectories and the identification of variables that are connected with specific labels, the approach makes it possible to comprehend the distinctions that exist across classes. By comparing baselines with effective/ineffective attack labels with NBA datasets, the efficacy of the system was validated and confirmed.

Bialecki et al. [17] suggested that esports offers a great amount of data availability, which makes it available to the scientific community. Now available is a dataset consisting of 17930 game-state files that was processed into 55 "replaypacks" from StarCraft II competitions. For statistical and machine learning modeling tasks, as well as for comparisons to laboratory measurements, these data are absolutely necessary. There is potential for the dataset to be used in artificial intelligence, machine learning, psychology, and sports-related research.

Chessa et al. [18] demonstrated a weighted complex network approach(WCNA) to the problem of identifying basketball player communities on the basis of their performances. A sparsification method is also used in order to get rid of weak edges. By calculating the ideal community structure of the "giant component" at each edge removal, modularity and compactness are maximized to their full potential. The normalized mutual information serves as a confirmation of the sparsification transition, which ultimately leads to the optimal number of communities and the most optimal distribution of nodes. In order to facilitate data-driven decision-making in basketball, the concept is used to the regular season of the NBA.

Kleinman and Erica Michelle [19] provided a summary of In high-impact fields like as education, training, and health, complex games that include several right techniques and uncertain consequences are becoming more popular. Because of their high learning curves, however, they are inaccessible to a large number of players, which results in a reduction in diversity among professional circles. Through the use of the Cyclical Phase Model of Self-Regulated Learning, this study investigates how players acquire the skills necessary to play and become proficient in increasingly difficult games. Moreover, it offers empirical insights for the creation of computational support tools for learning, as well as an expanded knowledge of learning and mastery in complicated gaming. By contributing to the development of more accessible and effective high-impact technologies, this study helps to make complicated games more accessible.

Mountifield et al. [20] provided a summary of the field of sport analytics is expanding rapidly within the sporting sector. It provides teams and organizations with the ability to enhance their performance and make well-informed choices on their performance. A dearth of algorithms and statistical approaches that are in the public domain has arisen as a consequence of the increasing competitive goals of sport analytics. It is becoming increasingly popular and a sphere of common knowledge, despite the fact that there are many who say that its use is based on the self-interest of organizations. A literature overview on statistical aspects, performance optimization, theoretical frameworks, and applications of sport analytics is presented in this chapter. The purpose of this study is to demonstrate that sport analytics is more than simply alchemy in the entertainment sector.

3. Proposed methodology

For the purpose of efficiently collecting, processing, and analyzing massive datasets within the sphere of higher education, the methodology that has been presented for the Education Big Data Processing Framework (EBDPF) incorporates a comprehensive approach that merges cutting-edge technologies and powerful data processing algorithms. The following essential elements are what make up the methodology:

a. EBDPF Architecture

Higher education institutions can benefit from the robust platform known as the Education Big Data Processing Framework (EBDPF), which was developed for the purpose of processing and analyzing massive datasets. Components that are networked and help facilitate data gathering, processing, analysis, and visualization are included in its infrastructure. Flume and Sqoop are the technologies that the EBDPF employs to acquire data from a wide variety of sources. This ensures that the data is gathered and synchronized in an effective manner. For the purpose of transmitting data from monitoring directories to the Hadoop Distributed File System (HDFS), the data processing and storage layer employs a two-layer Flume agent structure. This structure ensures high-efficiency offline cleaning processing while also providing data dependability and scalability for storage.

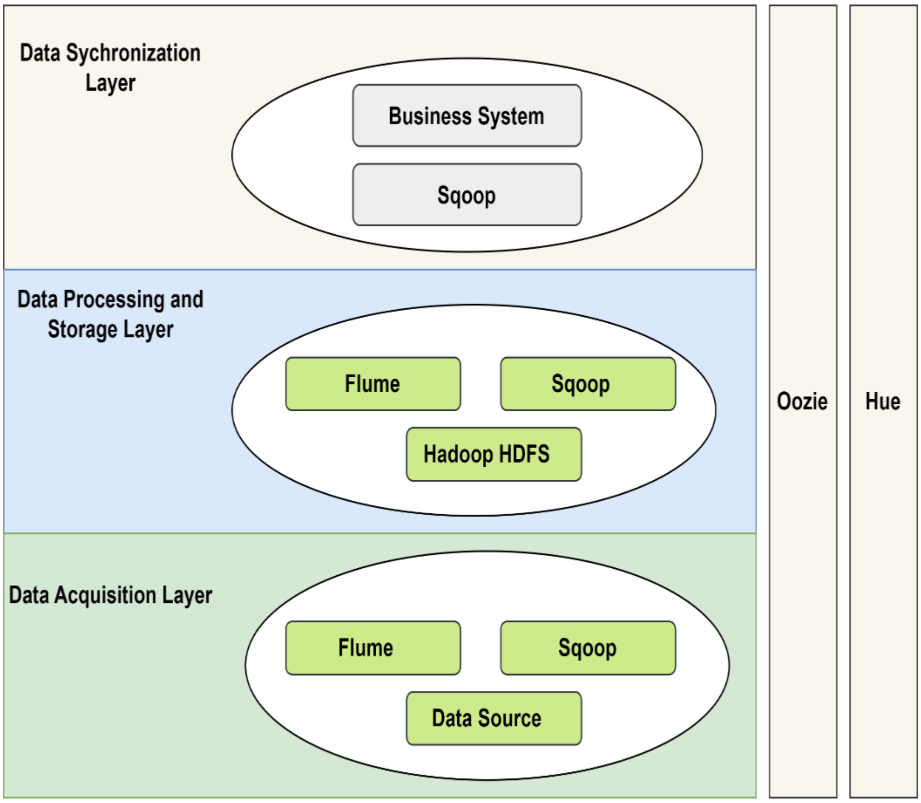

Figure 1. EBDPF Architecture

In figure 1 shows that The layers of EBDPF Architecture, the layer is responsible for managing the synchronization of data that has been collected from a variety of sources. subsequently ensures that the data is integrated and updated in a seamless manner for the purposes of analysis and processing. Integrating with a wide variety of business systems enables the EBDPF to simplify the flow and processing of data in a smooth manner. Enhancing the user experience and making system maintenance easier are two benefits that may be achieved through the utilization of visualization tools and user-friendly interfaces. The EBDPF places a high priority on scalability and performance optimization, which enables the platform to manage large-scale data operations and adapt to different data volumes and processing nodes. Therefore, this guarantees that the framework is capable of processing and analyzing data in educational contexts in an efficient and reliable manner while retaining a high level of efficiency

b. Data Acquisition Layer

In order to enable additional processing and analysis within the platform, the Data Acquisition Layer is vital in the Education Big Data Processing Framework (EBDPF). It is responsible for gathering and synchronizing data from various sources. In order to integrate different data streams efficiently and simplify data collecting, this layer makes use of modern tools like Sqoop and Flume. The Data Acquisition Layer is described in depth below:

Making Use of the Flume Tool: Flume is a distributed, dependable, and strong tool that may efficiently gather, aggregate, and transfer massive volumes of data from several sources to a single data repository. The EBDPF makes use of Flume to collect data from a variety of sources, one of which is text files kept on crawler clusters. When dealing with data from a variety of sources and formats, Flume is the way to go because of its scalability and versatility.

Integrating the Sqoop Tool: Sqoop is a tool that was made to efficiently move large amounts of data from structured data storage like relational databases to Apache Hadoop. To facilitate the smooth movement of data to the Hadoop cluster for additional processing, the EBDPF makes use of Sqoop to synchronize data acquired from various sources. An important part of the Data Acquisition Layer, Sqoop may manage duties related to data synchronization.

Efficient Data Gathering: The EBDPF's Data Acquisition Layer is responsible for making sure that data from different sources, such as text files on crawler clusters, is gathered efficiently. The framework is able to efficiently gather data in either real-time or batch mode via the use of Sqoop and Flume, according to the needs of the data processing jobs. Delays are minimized and data is instantly available for analysis thanks to this efficient data collection technique.

The Data Acquisition Layer is in charge of syncing data from several streams, which might be either structured or unstructured. The EBDPF can unify data from various sources and formats by including Flume and Sqoop, which allows for easy integration into the platform for additional processing. Consistent and up-to-date data is ensured by this synchronization, which improves the accuracy and reliability of subsequent studies.

Effortless Integration for Subsequent Processing: The Data Acquisition Layer's principal goal is to include data from diverse sources into the EBDPF in a smooth manner so that it can undergo subsequent processing and analysis. The framework is able to accomplish its data collection, synchronization, and transfer goals with relative ease thanks to Flume and Sqoop. Data is then analyzed utilizing sophisticated data mining algorithms and analytical techniques on the Hadoop cluster. This tight connectivity improves the platform's overall efficiency and simplifies the data processing processes. All things considered, the EBDPF's Data Acquisition Layer is vital for gathering, coordinating, and integrating data from many sources with the help of Flume and Sqoop. This layer lays the groundwork for successful data processing and analysis within the platform by assuring efficient data gathering and synchronization. Its ultimate goal is to enable informed decision-making and insights production in the higher education domain.

c. *Data Processing and Storage Layer*

The Education Big Data Processing Framework, also known as EBDPF, is an essential component for the implementation of effective data processing and storage. The management of the transfer of data from monitoring directories to the Hadoop Distributed File System (HDFS) is accomplished through the utilization of a Flume agent structure that is composed of two layers. In this figure 2, the first layer agent is responsible for collecting data from the crawler cluster, while the second layer agent is responsible for receiving and sending the data to HDFS for storage and processing. This structure ensures that the transfer of data is both seamless and efficient. Through the use of a two-layer Flume agent structure, data flow can be optimized, data dependability can be improved, and the process of transferring huge datasets to HDFS can be simplified. Because of this, efficient data processing and storage for the purposes of analysis and decision-making are made possible.

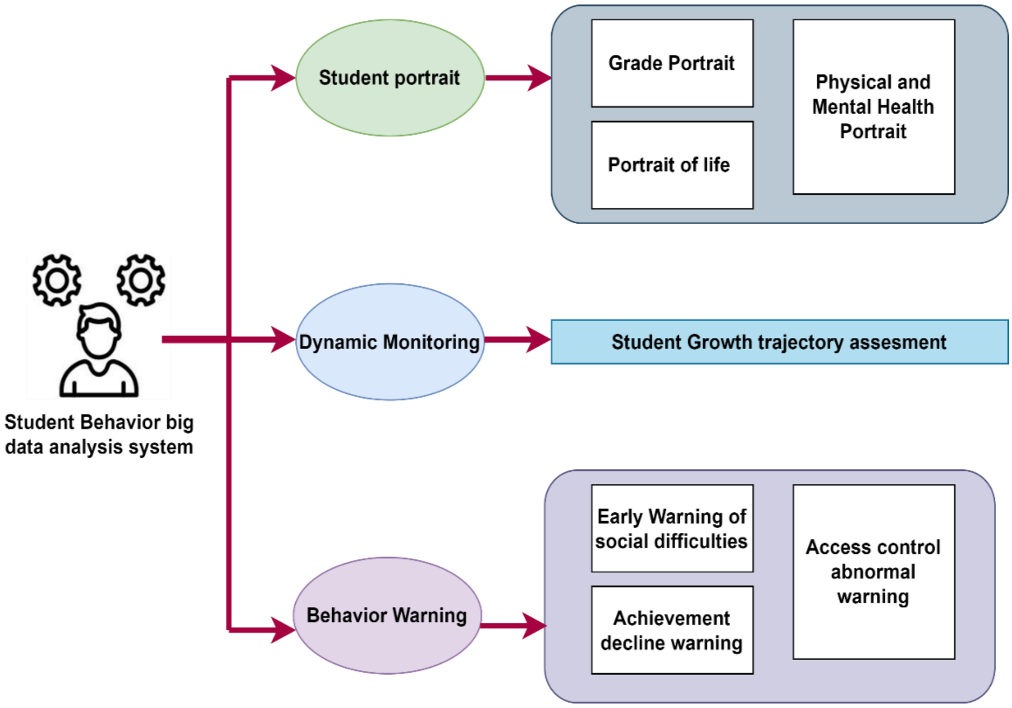
Student portrait
Grade Portrait
Portrait of life
Physical and Mental Health Portrait
Student Behavior big data analysis system
Dynamic Monitoring
Student Growth trajectory assesment
Behavior Warning
Early Warning of social difficulties
Achievement decline warning
Access control abnormal warning

Figure 2. Data Processing and Storage Layer

The EBDPF also places an emphasis on high-efficiency offline cleaning processing of massive datasets. This ensures that data is cleaned, converted, and readied for analysis without affecting the operations that are being performed in real time related to data processing. When it comes to storage, the EBDPF places a high priority on data dependability and scalability, and the primary storage system that it uses is HDFS. The danger of data loss or corruption is reduced as a result of this, which guarantees that the data is stored in a dependable and error-tolerant manner. Due to the fact that HDFS is scalable, the platform is able to accommodate increasing amounts of data without being negatively affected by performance issues. There is a significant contribution made by the Data Processing and Storage Layer to the seamless integration of data within the EBDPF. This layer makes it possible to do advanced analytics, data mining, and visualization operations. Users are given the ability to draw important insights and make educated decisions based on processed data as a result of this seamless connection, which improves the overall efficiency and efficacy of data processing inside the platform.

d. Data Synchronization Layer

An essential part of the Education Big Data Processing Framework (EBDPF) is the Data Synchronization Layer, which is responsible for coordinating the synchronization of data obtained from different sources. It makes sure that the platform for processing and analysis is always up-to-date and integrated. To provide a complete dataset useful for universities, the layer compiles information from several sources, including text files, databases, web servers, and other data repositories. Also, this layer is in charge of data updates, so new data gets added to the dataset without a hitch, old records get updated with the most recent information, and data consistency is maintained across all sources. For analysis and decision-making, users can obtain correct data in real-time.

In order to conduct rapid and accurate analyses, real-time data synchronization is crucial. In order to improve responsiveness and relevance, the EBDPF uses techniques to synchronize data continuously or at predetermined intervals. Data validation tests, error management procedures, and data cleansing processes are all part of data quality assurance, which aims to keep data accurate and consistent. The Data Synchronization Layer's principal goal is to facilitate the platform-wide integration of data for processing and analysis. Users are given a unified and up-to-date dataset for sophisticated analytics, data mining, visualization, and reporting tasks by the framework's effective management of data synchronization. Overall, data-driven decision-making in higher education is made more efficient and successful by this seamless integration.4.

4. Results and Discussion

The Education Big Data Processing Framework (EBDPF) demonstrated efficient resource utilization, scalability, and overall effectiveness in processing large datasets for data analysis in higher education. Efficiency metrics showed a system efficiency of 0.7 to 0.8, indicating efficient resource utilization and task completion within the Hadoop environment. Scalability analysis showed that as dataset size increased, efficiency improved, indicating the system's ability to handle larger datasets without significant performance decrease. Overall, the EBDPF demonstrated good scalability and efficiency in handling complex data mining tasks within a distributed computing environment. To emphasize the proposed model's advantages and enhancements, contrast its performance with those of other cutting-edge models. The existing algorithms, such as SVM [13], Naïve Bayes [13], and HiveQL [14], are taken for comparison study based on the metrics.

a. Scalability Analysis of Pattern Mining Tasks

A system's capacity to effectively manage changing amounts of data is evaluated through the process of scalability analysis. One example of this is a data mining framework that operates within a Hadoop environment. In this particular scenario, the analysis is centered on determining how the efficiency of the system shifts as the amount of data that is input gradually grows. It is essential to have a good understanding of how well the system scales with larger datasets in order to effectively manage the increasing demands of data processing jobs. The goal is to have this understanding.

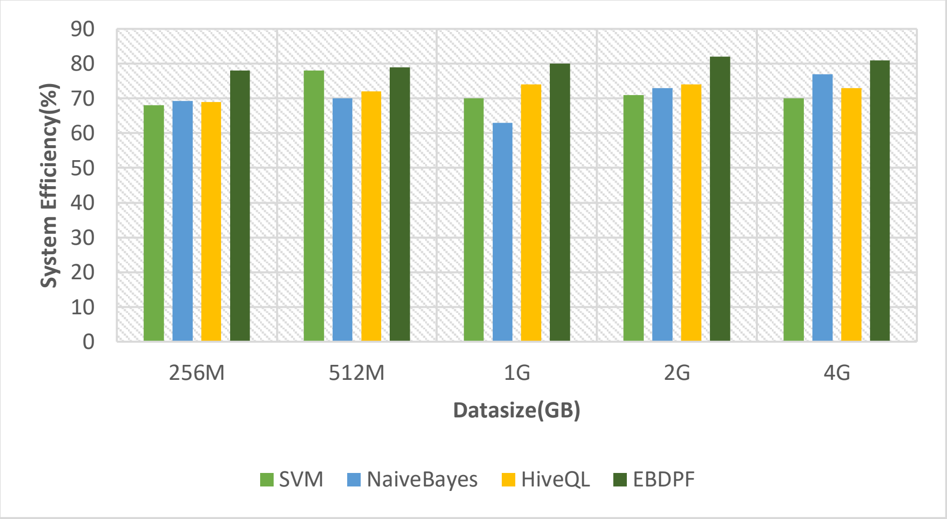

Figure 3. Scalability Analysis

In figure 3,The scalability analysis illustrates a positive impact, demonstrating that the effectiveness of the distributed system improves in proportion to the amount of the datasets. With regard to scalability, this influence implies that the system is able to properly manage larger datasets without seeing a significant drop in performance. The fact that the amount of data and the number of nodes have been shown to have an effect on the efficiency of the system highlights the capability of the system to scale its resources and processing power in accordance with the workload.

b. Efficiency Analysis of Distributed System

Evaluation of the performance of a distributed system, such as Hadoop, based on the usage of computational nodes is accomplished through the utilization of efficiency metrics. Within the context of this scenario, the number of compute nodes is altered, and the efficiency ratio is evaluated. By analyzing the efficiency ratio, one can gain an understanding of how efficiently the system makes use of its resources, regardless of the number of compute nodes that are present. The execution of this analysis is necessary in order to optimize the setup of the distributed system and guarantee the highest possible level of performance.

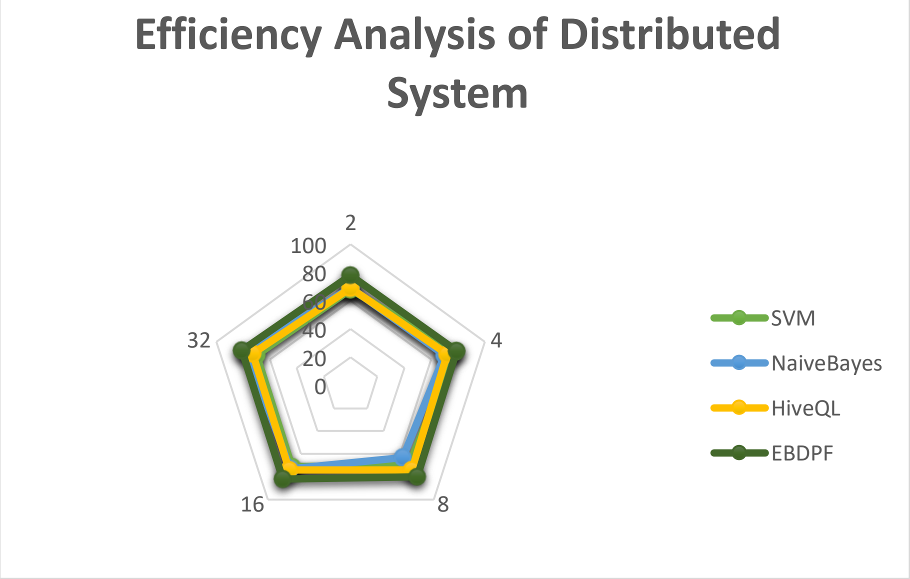

Figure 4. Efficiency Analysis

In figure 4,The efficiency metrics, which range from 70% to 80% in terms of system efficiency, serve as a comprehensive evaluation of the overall performance of the distributed system located within the Education Big Data Processing Framework (EBDPF) for the purpose of data mining jobs. Efficiency values that are higher indicate that resources are being utilized effectively and that jobs are being completed in an efficient manner inside the Hadoop environment.The efficiency ratio, determined by multiplying the execution time on the wall of the distributed system by the number of processing nodes, offers significant insights into the efficiency of resources and the completion of tasks.

c. *Impact of Data Size and Nodes*

The impact of the size of the data and the number of nodes is described as follows: The impact of data size and nodes seeks to investigate the ways in which the effectiveness of a system is affected by the quantity of data that is input as well as the number of computing nodes that are present. Through the examination of a variety of data sizes and the modification of the number of nodes, this analysis offers a comprehensive perspective on the dynamic relationship that exists between these two elements. It is helpful to have an understanding of the impact in order to make educated decisions on the setup of the system and the distribution of resources in order to achieve maximum efficiency in a distributed computing environment that is based on Hadoop.

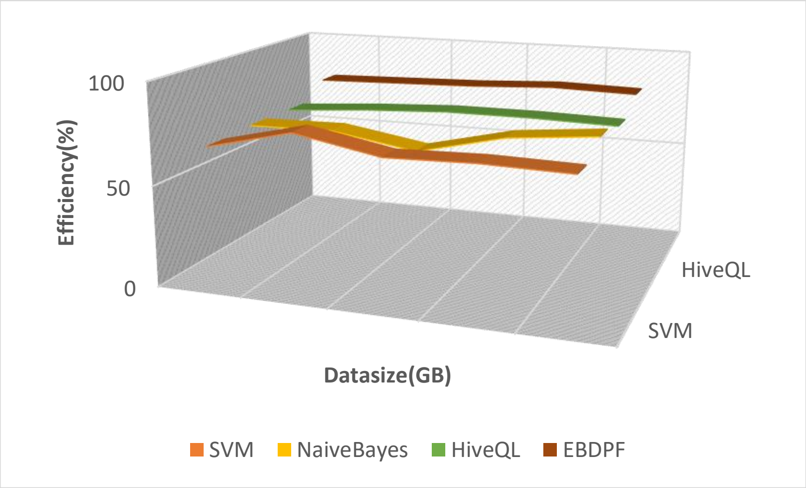

Figure 5. Impact of Data Size and Nodes

In figure 5,The experiment demonstrates that the EBDPF, which is driven by the Hadoop distributed computing environment, demonstrates strong scalability and efficiency in processing data mining jobs on artificial datasets of varied sizes.The experiment also demonstrates that the EBDPF is able to handle large amounts of data. The impact on performance is observable in the system's ability to process and analyze huge datasets in an effective manner, which contributes to the implementation of data-driven decision-making in higher education institutions. In addition to highlighting the framework's effectiveness in managing complicated data mining activities within a distributed computing environment, the results demonstrate the computational advantages and development possibilities of the framework.

5. Conclusions and Future work

The Education Big Data Processing Framework (EBDPF) is the focal point of this research, which aims to transform data processing in educational settings by investigating

Hadoop-Based Political and Ideological Big Data Platform Architecture and Pattern Mining in Higher Education. Standard data mining approaches implemented using MapReduce programming show that the EBDPF is effective and can manage data quantities and cluster configurations that are different. Hadoop distributed solutions outperform conventional approaches in data mining applications, according to extensive trials. The examination of processing power reveals unrealized possibilities for additional improvements. Improvements to the user interface, security features, scalability testing, algorithmic refinement, and integration of sophisticated technologies are all areas that will be explored further in the EBDPF framework. The EBDPF can be made better at mining massive datasets for useful patterns and insights by algorithmic tweaks. The EBDPF can make better, more nuanced decisions with the use of cutting-edge tech like AI and machine learning, which can improve its analytical capacities and predictive modeling. To make sure the framework can handle complicated educational datasets efficiently and adaptably, scalability testing is conducted. To protect the privacy and integrity of data, it is wise to invest in user interface improvements that increase accessibility and security. The EBDPF can maintain its status as a strong and innovative framework in higher education by taking these factors into account moving forward.

REFERENCES

[1]. Chen, Chengjie, et al. "TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining." Molecular Plant 16.11 (2023): 1733-1742.

[2]. Ageed, Zainab Salih, et al. "Comprehensive survey of big data mining approaches in cloud systems." Qubahan Academic Journal 1.2 (2021): 29-38.

[3]. Jiang, Yang, et al. "Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the National Assessment of Educational Progress (NAEP)." Computers & Education 193 (2023): 104680.

[4]. Li, Chunquan, Yaqiong Chen, and Yuling Shang. "A review of industrial big data for decision making in intelligent manufacturing." Engineering Science and Technology, an International Journal 29 (2022): 101021.

[5]. Bai, Xiaomei, et al. "Educational big data: Predictions, applications and challenges." Big Data Research 26 (2021): 100270.

[6]. Rafique, Azra, Kanwal Ameen, and Alia Arshad. "E-book data mining: real information behavior of university academic community." Library Hi Tech 41.2 (2023): 413-431.

[7]. Caspari-Sadeghi, Sima. "Learning assessment in the age of big data: Learning analytics in higher education." Cogent Education 10.1 (2023): 2162697.

[8]. Huang, Daigen. "Design and Development of University Smart Campus Platform Based on Big Data." Proceedings of the 2nd International Conference on Big Data Economy and Digital Management, BDEDM 2023, January 6-8, 2023, Changsha, China. 2023.

[9]. Khrabatyn, Roman, et al. "Technologies for designing and programming big data in e-learning." Вісник Тернопільського національного технічного університету 109.1 (2023): 72-79.

[10]. Ma, Changxi, Mingxi Zhao, and Yongpeng Zhao. "An overview of Hadoop applications in transportation big data." Journal of traffic and transportation engineering (English edition) (2023).

[11]. Bi, Jiana, Xiangjun Chen, and Shuang Wang. "Development of Network Public Opinion Analysis System in Big Data Environment Based On Hadoop Architecture." Procedia Computer Science 228 (2023): 291-299.

[12]. Yang, Xiaoqing, et al. "Exploring the integration of big data analytics in landscape visualization and interaction design." Soft Computing (2024): 1-18.

[13]. Popescu, Andrei, and Michal Val'ko. "Social Media Sentiment Analysis in the Age of Big Data: Understanding User Behavior and Predicting Trends." International Journal of Business Intelligence and Big Data Analytics 6.1 (2023): 31-39.

[14]. Sedek, Khairul Anwar, et al. "Design and implementation of big data visualization for student housing analysis." Journal of Information and Knowledge Management (JIKM) 2 (2023): 124-142.

[15]. Gattoju, S., & Nagalakshmi, V. (2023). Design of ChaApache framework for securing Hadoop application in big data. Multimedia Tools and Applications, 82(10), 15247-15269.

[16]. Wang, Yubiao, et al. "An early warning method for abnormal behavior of college students based on multimodal fusion and improved decision tree." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-23.

[17]. Lv, Lujun. "Artificial Intelligence Medical Construction and Data Mining Based on Cloud Computing Technology." International Conference on Big Data Analytics for Cyber-Physical System in Smart City. Singapore: Springer Nature Singapore, 2022.

[18]. Wang, Xuyan. "Check for updates Recognition of Piano Play Gesture Based on Infrared Sensor Detection Rod." Proceedings of the 2023 2nd International Conference on Educational Innovation and Multimedia Technology (EIMT 2023). Vol. 8. Springer Nature, 2023.

[19]. He, Xin, and Hongjie Cao. "Cloud Computing-Based Sharing Platform for High-quality Teaching Resources of Higher Vocational Physical Education." International Conference on E-Learning, E-Education, and Online Training. Cham: Springer Nature Switzerland, 2023.

[20]. Kummar, Sohit, et al. "Emplacement Big Data Visualization in Medical Sector." Available at SSRN 4618416 (2023).