International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Implements and Integrates Data-Centric Applications based on Big Data Analysis Service on Big Data Platforms

Prabu Ravichandran,
Sr. Data Architect, AWS, Amazon, Raleigh, NC, USA
Prabu.ravichandran07@gmail.com

ABSTRACT
Big data is now an integral part of many fields of research due to the rapid advancement of digital tools for data processing. With the system's help, the business analysis process may be quickly developed and implemented with flexibility, allowing for faster contact and reaction times. It also helps with the management and development of smart data analysis programmes, which are constantly evolving. The suggested method enables users to do batch and streaming computations concurrently by combining online-streaming analysis with offline-batch models. Big data analysis using cloud capabilities and user-customizable workflow techniques are also available. Cloud workflow system modelling, application customisation, dynamic construction, and scheduling are all included in this study. To improve efficiency, it is suggested to use a chain workflow foundation mechanism that combines multiple analytic components into one. To ensure the system's analytical competence, four real-world application examples are given. The results of the experiments demonstrate that the suggested system makes good use of data analysis methods and can handle numerous users logging in at the same time. Network operators have used the suggested SaaS workflow solution, and it has proven successful for them.

Keywords: Big data intelligent analysis; cloud workflow; SaaS; cloud computing

1. INTRODUCTION

A survey on business process integration (BDI) solution implementations was carried out by the Accenture corporation. A whopping 92% of managers are satisfied with the outcomes achieved by BDI solutions, and an overwhelming 89% think that big data analytics and integration is crucial to their company strategy for taking advantage of competition. Computers have been 33% cheaper, storage has become 38% cheaper, and bandwidth has become 27% cheaper annually, according to the Internet Trends Report from KPCB's Mary Meeker. This trend has persisted for the last twenty years. Data selection, collection, storage, communication, searching, visualisation, privacy and security assurance, and overall data handling are the main obstacles to BDI processing. Effective decision making is driven by the efficiency with which big data is handled. Investment costs have been decreased, and the big data management and analytics area has been enhanced, thanks to advancements in computer infrastructures, algorithms, and novel technologies. This has allowed businesses to get the best value for their investment.

1.1 Motivation and significance of BDI study

Big data would be quite valuable, according to the business experts. Businesses are improving the customer experience and cutting expenses through digital transformation. Customers would prefer to be able to view their own data and do transactions while on the move. The information can be made reliable, standardised, and actionable by online processing of bigdata employing analytical tools in organisations. With the use of big data insights, businesses are able to create more BDI apps, make better business decisions, and increase productivity. The possibility of cyberattacks has grown alongside the revolution in computing and digitalization. There is a growing and increasingly complicated cyber danger from hackers. Intelligent and secure business process automation systems have been designed using ML and DL methodologies. Since 2019, more money has been going towards machine learning initiatives than into all of AI's other endeavours put together. In order to get business insight and make choices in real-time, Walmart organisation has used BDI technologies. When planning their marketing strategy, several prominent fast food chains are turning to BDI solutions to uncover emerging business trends. These companies include McDonald's, KFC, and Pizza Hut. In addition to other businesses, casinos have found success in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

recent years by implementing BDI solutions to boost income and encourage repeat business. In order to anticipate guests' actions, preferences, and dietary needs, the hospitality sector employs BDI software. Modern tourists also rely on digital technologies to gather information about any and all tourist-related topics. In order to provide high-quality healthcare services while reducing time and money wasted, the healthcare industry has used BDI. Public services in smart cities are being developed by governments using BDI. Thanks to BDI's data insights and analytical reports, e-commerce sectors like Amazon, Flipkart, etc. have been empowered. Meteorologists were able to make more accurate weather predictions by integrating AI, BDI, and visualisation technologies. Modern agriculture has effectively implemented BDI solutions. The use of digital marketing has been made possible by BDI solutions, which is crucial for the success of any organisation.

The need for powerful data analysis tools has skyrocketed in tandem with the proliferation of both the variety and volume of available data. Data processing systems [1] collect and process data items in order to extract valuable insights from them. Businesses can benefit from data analysis in a number of ways, including better customer experience, more informed business development, and better market judgements. Many older data analysis systems, like SPSS, operate independently. Data processing systems in the big data age typically use distributed computing and the cloud. Data mining, querying, ETL, and online analytical processing (OLAP) are just a few of the data analysis methodologies that have been utilised in a distributed computing fashion to manage large data sets. This research focuses on improving the big data analysis system's performance by integrating workflow technologies.

As data volumes grow, the following problems arise in conventional, straightforward data processing systems: (1) massive volumes of data: the amount of data that needs processing is growing exponentially; (2) data complexity: Structured and semi-structured data are among the new forms of information that are appearing; (3) diversity of mixed loads: as businesses grow and data becomes more diverse, many new self-analysis requirements are being proposed.

2. LITERATURE REVIEW

There is a new way to solve old data analysis challenges with cloud computing [2,3]. It is a way of running computers that makes use of the Internet to deliver resources that are both scalable and virtualized. Instead of using local or remote servers, this computing method uses parallel processors to distribute computing duties. Enterprise data centres will operate in a manner analogous to the Internet, urging companies to allocate their scarce storage and processing resources primarily to applications with mission-critical importance. The cloud offers a variety of services, including infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). The growth of distributed computing technology is another benefit of big data.

The huge data batch processing has been the subject of several research articles [4,5,6]. The MapReduce programming approach was proposed by Google, and its open-source Hadoop has achieved outstanding results both theoretically and in practice [7,8,9,1,11]. Businesses are interested in both batch computing and streaming data processing in real time. A couple of examples include Twitter's Storm streaming computing system and Yahoo's S4 distributed stream computing system. In contrast to Microsoft's TimeStream technology, Facebook employs Data Motorway and Puma for real-time data management.

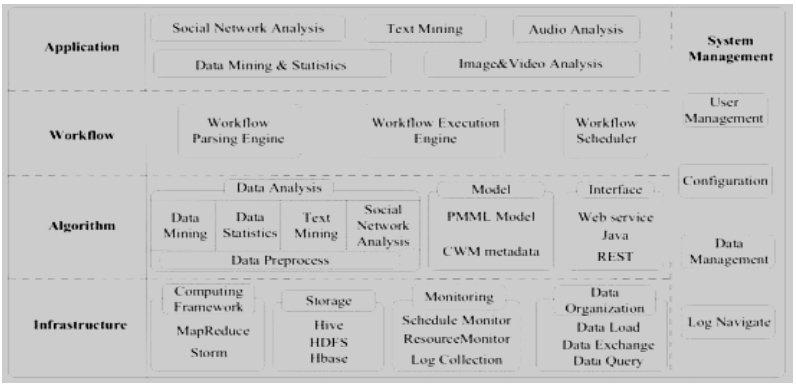

Figure 1. Framework for DDM, or distributed data mining. Predictive Model Markup Language (PMML), Hadoop Distributed File System (HDFS), and Representational State Transfer (REST).

Distributed data mining (DDM) [12-14], a typical SaaS, is our proposed distributed large data analytic system to tackle these difficulties. Infrastructure, algorithms, workflows, applications, and system management are all part of the DDM architecture, as shown in Figure 1. Thanks to its distributed storage capacity and parallel processing capabilities, DDM is able to process

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

and store massive volumes of data.[15-20] Over the Internet, this analysis system offers cloud services for ETL, data mining, OLAP, and reporting.[21-27] In DDM, the workflow system is the central component. A suite of big data analysis apps that users can modify to their liking. By translating user needs into workflow procedures, DDM makes it possible to assign tasks in the cloud in a parallel fashion.

3. BIG DATA ANALYSIS ORIENTED CLOUD WORKFLOW SYSTEM

It defines a cloud workflow system that is aimed at big data research with the purpose of managing and investigating business processes and facilitating shared workflows. As depicted in Figure 2, the system topology primarily consists of a web server, a workflow engine (which can run on a cluster of parallel computers or high-performance servers), and a decentralised cloud service (which offers resources for offline data mining, real-time analysis, and distributed storage). The web server offers user-friendly interfaces (UIs) for application design.

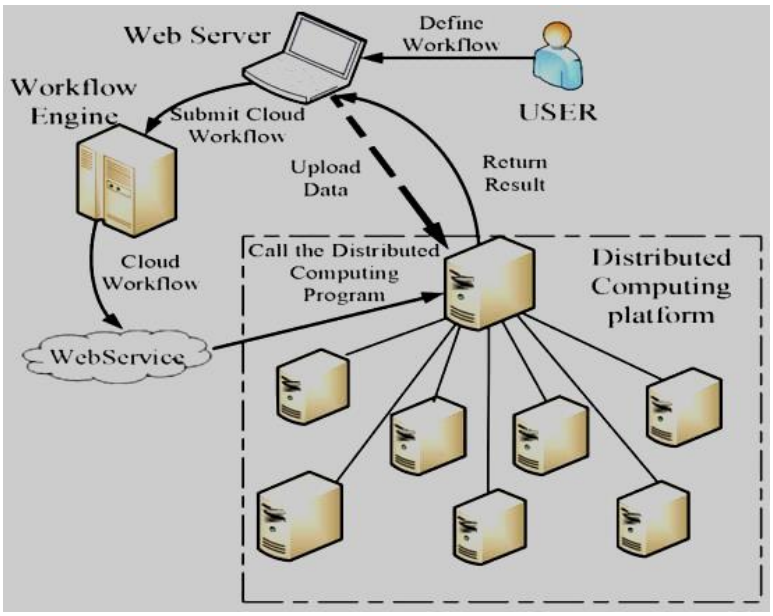

Figure 2. System topology.

In addition to including algorithms for big data analysis, the parallel cloud platform also needs to think about how to interact with processes. The workflow technique is shown in Figure 3, and the detailed steps are as follows:

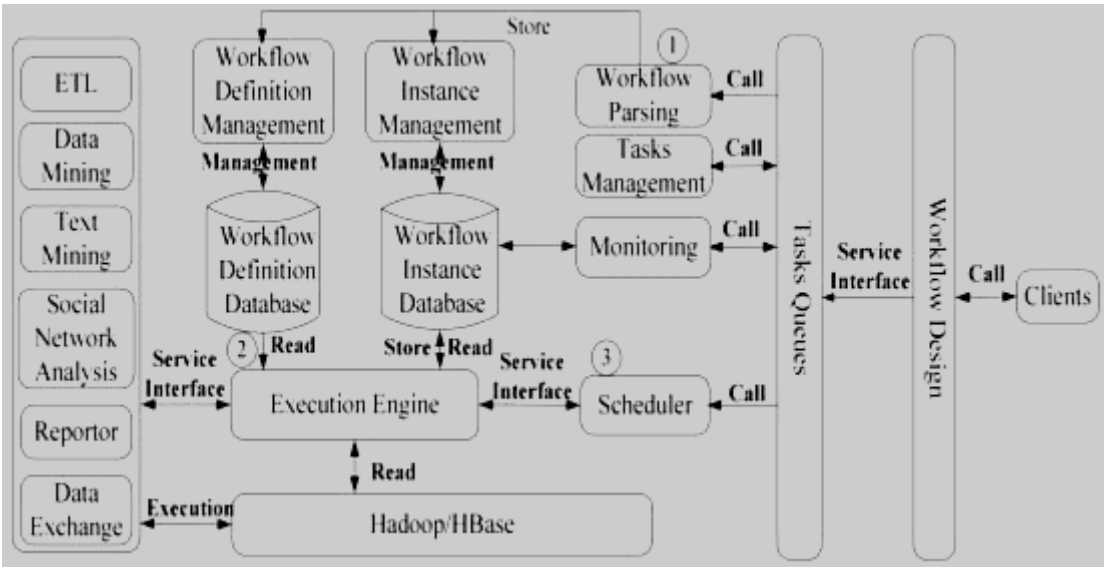

Figure 3. Standard operating procedure. Extract, transform, and load (ETL).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Step 1: The web-based user interface (UI) allows users to simply drag-and-drop components to build workflows for large data analysis applications. Web services distinguish between cloud services (like MapReduce, Hive, and Storm) and traditional business processes (such data interchange, result visualisation, and reports) while processing supplied workflows. It is possible to use the web server to handle routine tasks, and the distributed cloud platform for cloud-related tasks. The workflow definition database stores the stated workflows, which the workflow parsing engine will receive. The workflow configuration settings, along with the parsed XML description documents and DAGs, are stored in the workflow instance database once the user has specified them.
Step 2: After the specifics of the task are uploaded, the workflow execution engine will react to requests to execute the process. The cloud platform initiates the daemon process by requesting workflow XML definition documents and DAGs. Through the use of control and data streams, the daemon executes the tasks sequentially.
Step 3: Hadoop, which is open source and includes MapReduce and the Google File System, as well as Storm, are used to build a platform for distributed computing. For big data analysis, our solution integrates the best features of MapReduce, Storm, HDFS, and HBase from Hadoop. Users will receive the analysis results and, depending on their settings, analysis data will be stored.
Step 4: Users can retrieve the results using the web-based user interface, and the workflow execution engine will conduct the full process according to the workflow description.

4. METHODOLOGY

This system also provides data analysis templates for customer churn prediction, social network analysis, self-construct business workflows, and user behaviour and interest analysis. The use of telecom operators' data for service recommendation and CCP is introduced in this section. Demonstrating the system's usefulness and efficiency, two security applications are proposedtwo things: one to identify phishing sites and another to analyse potential threats to a network's security.

4.1. Service Recommendation

Workflow definition: There is a typical pattern of consumer behaviour in the telecommunications industry that operators can use to better propose services and suggest new ones based on user preferences. The two primary components of service suggestion are customer clustering and service recommendation itself. The service recommendation pipeline is illustrated in Figure 4.

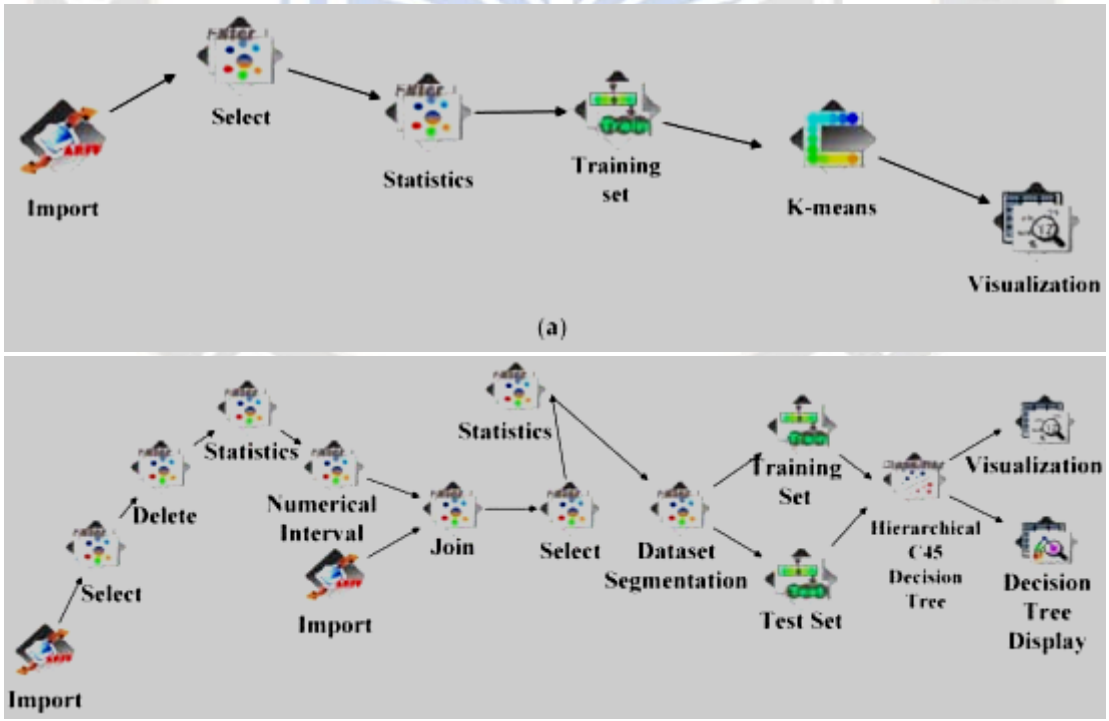

Figure 4. Procedure for making service recommendations: (a) grouping clients; (b) recommending services to those clients.

- Customer Cluster: There are several clusters of customers, with members that exhibit similar patterns of behaviour. The platform receives data using the "Input" component. Because raw data has various properties, the "Select" component chooses relevant ones for applications. The "K-means" component classifies buyers into groups based on their shared habits.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

- Service Recommendation: Based on the results of customer clustering, we apply a classification method to uncover the underlying relationship between client traits and the services they have chosen.In this case, the service numbers match the labels used for categorization. The platform receives data using the "Input" component. Two parts, "Select" and "Delete," allow you to filter features. Due to the unique nature of the service numbers, we classify new clients into various service types using the "C45" component.

Workflow instance: In order to analyse the class data, we must first sort the 300 million rows of raw data into 20 columns. In these columns, you will find information on customers, charges, voice communication (including calling, called, local, roaming, long distance, and the number of calls and call duration), and messages. The application allows for the definition of six client clusters: advanced, low-end, midrange value-added, advanced call, and midrange value-added. In order to train the categorization model, which can forecast customer behaviour and propose appropriate services to new consumers, we use customer clusters and the quantity of services.

5. RESULTS AND DISCUSSION

Figure 5 depicts a data analysis procedure that is used as an example. By contrasting three distinct execution forms—NoneWF, which is the absence of a workflow system and the manual execution of tasks, CommonWF, and ChainWF— The proposed cloud-based workflow system's effectiveness was evaluated.

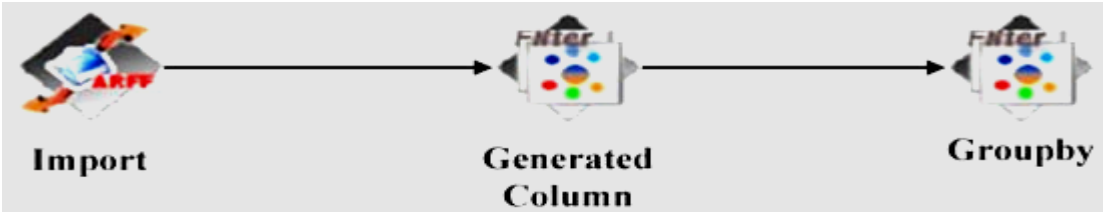

Figure 5. Data analysis workflow.

Two cloud companies, "Column generation" and "Groupby," and one regular company make up the example workflow's "Input" component. By combining the "Generated Column" and "Groupby" components into a single chain, ChainWF ensures that processed data remains in memory until all processing is finished, rather than being sent to the hard disc. Users can choose between HDFS or HBase as the destination for the analysis findings. All forty-three files, which included WAP website addresses, timestamps, and phone numbers, were authentic Wireless Application Protocol (WAP) logs used in the tests. Figure 6 displays the test findings. The horizontal axis shows the test data scale, which ranges from 50 GB to 230 GB, while the vertical axis shows the run time.

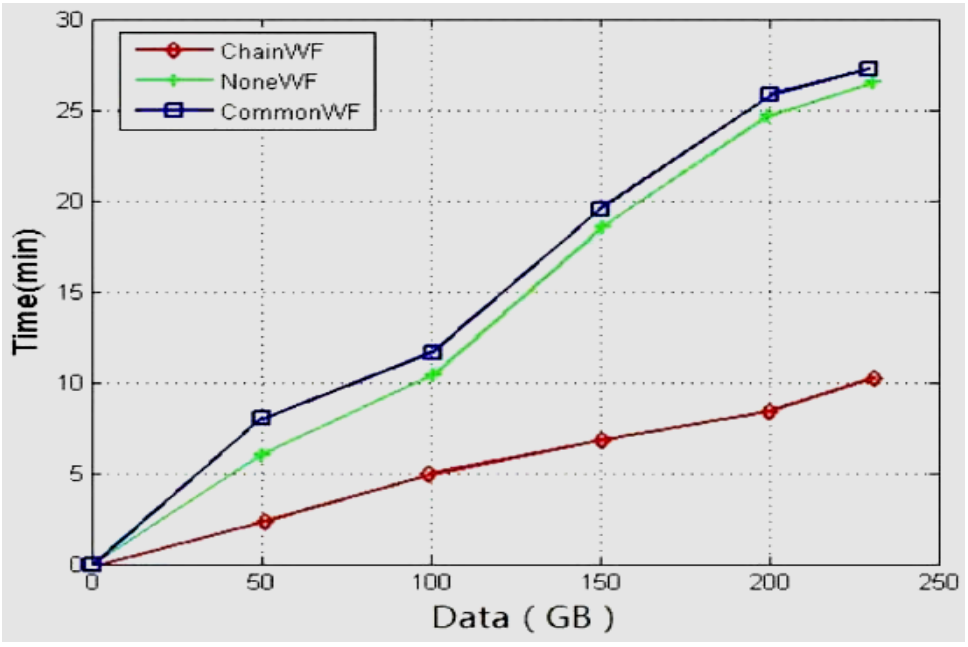

Figure 6. Various execution forms' performance.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

In Figure 7, we can see the outcomes of the system's throughput analysis. Throughput varied smoothly over time, with an average of 34,702 records/s. This outcome can fulfil the requirements of analysis in real-time. Data analysis delay was often measured in milliseconds, as illustrated in Figure 8. The data middleware experienced some noticeable delays (a few hundred milliseconds) in transmitting a large quantity of data to the spout module because of task scheduling, which allowed it to access several resources at once. The spout module wouldn't need to gather much data because the platform could process it whenever it chose. Middleware could handle the processing.

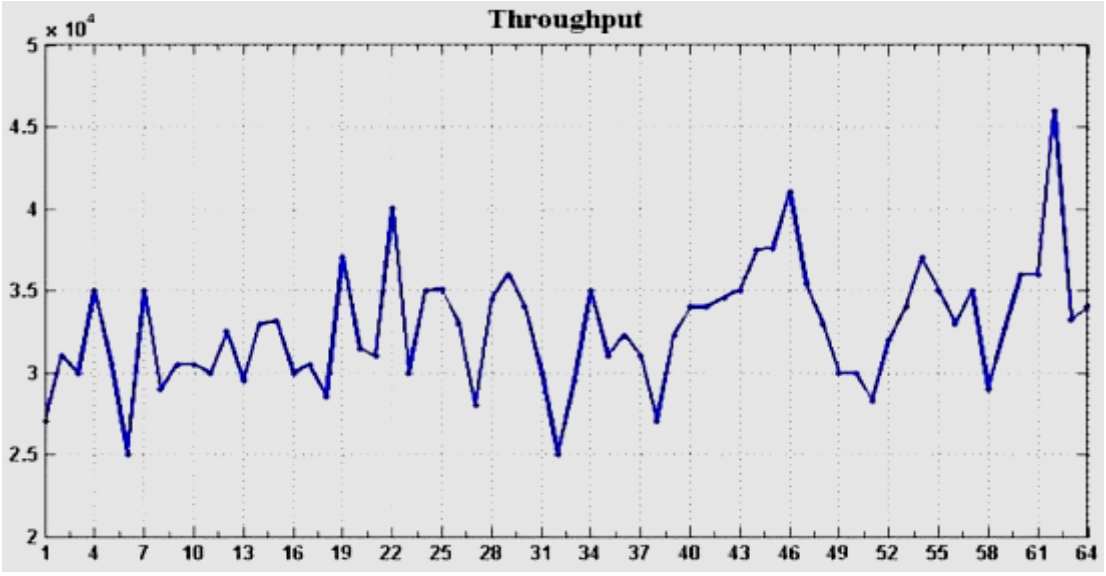

Figure 7. Throughput analysis of the system.

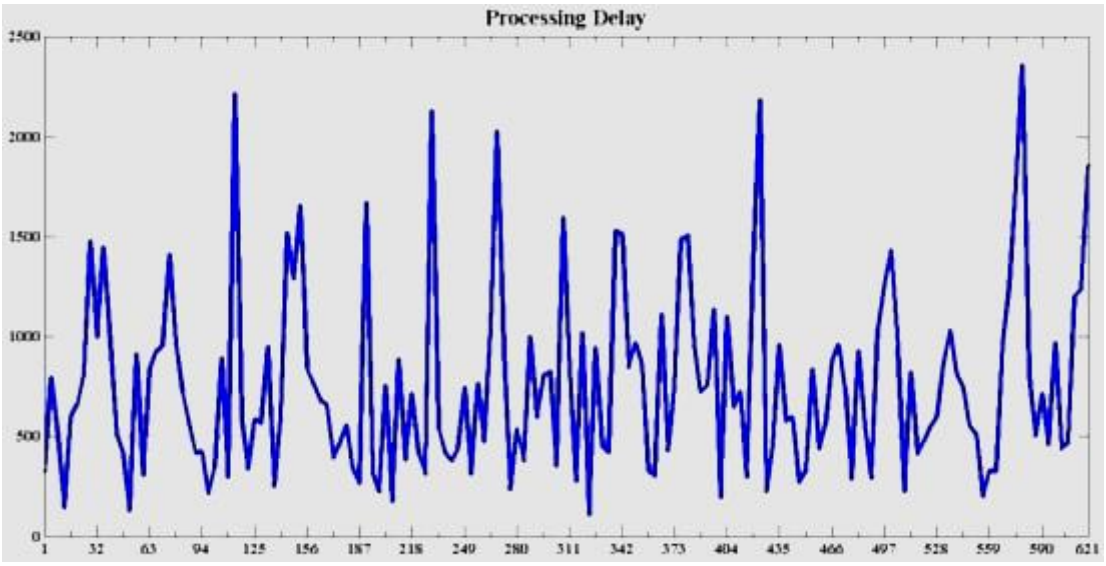

Figure 8. Data analysis latency.

CONCLUSIONS AND FURTHER WORKS

A lightweight cloud workflow system based on RESTful APIs was proposed in this study. The system can parse, execute, store, and schedule workflows, among other things; it also incorporates distributed big data intelligence analytic algorithms and gives users a user interface to set up workflows. Incorporating batch offline and online real-time analyses into the system makes it more resilient and adaptable, which is necessary for handling online streaming data analysis. Afterwards, methods for distributed data analysis may be executed with more flexibility. Additionally, we implemented the ChainWF mechanism, which allows for the simultaneous execution of numerous tasks in a single ChainJob, therefore enhancing the

efficiency of workflow execution. We used four common examples of big data analytics—service suggestion, customer churn prediction, phishing site identification, and network security scenario analysis—to show how cloud workflow parses, executes, and communicates data. We integrated the two kinds of analysis into a single cluster that could manage streaming and batch processing using Lambda architecture. Different workflow processes and the system's concurrency were tested to see which one was more efficient. A number of users were able to access the system at the same time, and the data analysis methods were utilised efficiently, according to the results. We also made sure the real-time and batch analysis models were working properly. It was discovered that the system handled batch and real-time analytics effectively.

REFERENCES

1. Data Process System. Available online: **https://en.wikipedia.org/wiki/Data_processing_system** (accessed on 16 February 2018).
2. Cloud Computing. Cloud Computing [EB/OL]. Available online: **http://en.wikipedia.org/wiki/Cloud_computing** (accessed on 16 February 2018).
3. Li, B.; Zhang, L.; Ren, L.; Chai, X.; Tao, F.; Wang, Y.; Yin, C.; Huang, P.; Zhao, X.; Zhou, Z. Typical characteristics, technologies and applications of manufacturing. *Comput. Integr. Manuf. Syst.* **2012**, *18*, 1345–1356, (In Chinese with English Abstract). [**Google Scholar**]
4. Li, B.D.; Mazur, E.; Diao, Y.L. SCALLA: A platform for scalable one-pass analytics using MapReduce. *ACM Trans. Database Syst.* **2012**, *37*, 1–43. [**Google Scholar**] [**CrossRef**]
5. Morales, G.D.F. SAMOA: A platform for mining big data streams. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; ACM Press: New York, NY, USA, 2013; pp. 777–778. [**Google Scholar**]
6. Lim, L.; Misra, A.; Mo, T.L. Adaptive data acquisition strategies for energy-efficient, smartphone-based, continuous processing of sensor streams. *Distrib. Parallel Databases* **2013**, *31*, 321–351. [**Google Scholar**] [**CrossRef**]
7. White, T. *Hadoop: The Definitive Guide*; O'Reilly Media: Sebastopol, CA, USA, 2012. [**Google Scholar**]
8. Yu, L.; Zheng, J.; Wu, B.; Shen, W.C.; Wang, B.; Qian, L.; Zhang, B.R. BC-PDM: Data mining, social network analysis and text mining system based on cloud computing. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 1496–1499. [**Google Scholar**]
9. Lam, C. *Hadoop in Action*; Manning Publications Co.: Shelter Island, NY, USA, 2010; pp. 265–270. [**Google Scholar**]
10. Duipmans, E.; Pires, L.F. *Business Process Management in the cloud: Business Process as a Service (BPaaS)*; University of Twente: Enschede, The Netherlands, 2012. [**Google Scholar**]
11. Cordys. Google apps + BPM: Power to the business user [EB/OL]. March 2009. Available online: **http://www.cordys.com/cordyscmscn_com/cn_google_apps.php** (accessed on 16 February 2018).
12. Nadella, Geeta Sandeep. Validating the Overall Impact of IS on Educators in U.S. High Schools Using IS-Impact Model – A Quantitative PLS-SEM Approach, DAI-A 85/7(E), Dissertation Abstracts International, Ann Arbor, ISBN 9798381388480, 189, (2023).
13. Gonaygunta, Hari, Factors Influencing the Adoption of Machine Learning Algorithms to Detect Cyber Threats in the Banking Industry, DAI-A 85/7(E), Dissertation Abstracts International, Ann Arbor, United States, ISBN 9798381387865, 142, 2023.
14. Ramya Manikyam, J. Todd McDonald, William R. Mahoney, Todd R. Andel, and Samuel H. Russ. 2016.Comparing the effectiveness of commercial obfuscators against MATE attacks. In Proceedings of the 6th Workshop on Software Security, Protection, and Reverse Engineering (SSPREW'16)
15. R. Manikyam. 2019.Program protection using software based hardware abstraction.Ph.D. Dissertation.University of South Alabama.
16. GPB GRADXS, N RAO, Behaviour Based Credit Card Fraud Detection Design And Analysis By Using Deep Stacked Autoencoder Based Harris Grey Wolf (Hgw) Method, Scandinavian Journal of Information Systems 35 (1), 1-8.
17. R Pulimamidi, GP Buddha, Applications of Artificial Intelligence Based Technologies in The Healthcare Industry, Tuijin Jishu/Journal of Propulsion Technology 44 (3), 4513-4519.
18. R Pulimamidi, GP Buddha, AI-Enabled Health Systems: Transforming Personalized Medicine And Wellness, Tuijin Jishu/Journal of Propulsion Technology 44 (3), 4520-4526.
19. GP Buddha, SP Kumar, CMR Reddy, Electronic system for authorization and use of cross-linked resource instruments, US Patent App. 17/203,879.
20. Hari Gonaygunta (2023) Machine Learning Algorithms for Detection of Cyber Threats using Logistic Regression, 10.47893/ijssan.2023.1229.

21. Hari Gonaygunta, Pawankumar Sharma, (2021) Role of AI in product management automation and effectiveness, https://doi.org/10.2139/ssrn.4637857
22. Sri Charan Yarlagadda, Role of Artificial Intelligence, Automation, and Machine Learning in Sustainable Plastics Packaging markets: Progress, Trends, and Directions, International Journal on Recent and Innovation Trends in Computing and Communication, Vol:11, Issue 9s, Pages: 818–828, 2023.
23. Sri Charan Yarlagadda, The Use of Artificial Intelligence and Machine Learning in Creating a Roadmap Towards a Circular Economy for Plastics, International Journal on Recent and Innovation Trends in Computing and Communication, Vol:11, Issue 9s, Pages: 829-836, 2023.
24. Amol Kulkarni, Amazon Athena Serverless Architecture and Troubleshooting, International Journal of Computer Trends and Technology, Vol, 71, issue, 5, pages 57-61, 2023.
25. Amazon Redshift Performance Tuning and Optimization,International Journal of Computer Trends and Technology, vol, 71, issue, 2, pages, 40-44, 2023.
26. B. Nagaraj, A. Kalaivani, S. B. R, S. Akila, H. K. Sachdev, and S. K. N, "The Emerging Role of Artificial intelligence in STEM Higher Education: A Critical review," International Research Journal of Multidisciplinary Technovation, pp. 1–19, Aug. 2023, doi: 10.54392/irjmt2351.
27. D. Sivabalaselvamani, K. Nanthini, Bharath Kumar Nagaraj, K. H. Gokul Kannan, K. Hariharan, M. Mallingeshwaran, Healthcare Monitoring and Analysis Using ThingSpeak IoT Platform: Capturing and Analyzing Sensor Data for Enhanced Patient Care, IGI Global eEditorial Discovery, Pages: 25, 2024. DOI: 10.4018/979-8-3693-1694-8.ch008.