Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Exploration of Big Data Pipeline Solutions for Business Analysis: A Comprehensive Survey

Pallavi G B[1], Latha N R[2], Shyamala G[3], Kalyana Kiran B. S. Goli[4], D Revanth[5], Gamana Yeluri R[5], Harika N[5], Keerthi P Reddy[5]

1, 2, 3 Associate Professor, Department of CSE B. M. S. College of Engineering, Bengaluru, India-560019
4 Senior Engineering Manager, Oracle India Private Limited, Bengaluru, India-560103
5 Undergraduate Student, Department of CSE B. M. S. College of Engineering, Bengaluru, India-560019

ARTICLE INFO

Received: 24 Dec 2024
Revised: 12 Feb 2025
Accepted: 26 Feb 2025

ABSTRACT

The sudden burst of data has resulted in the emergence of many big data frameworks such as Hadoop, Flink, and cloud-native platforms including Azure, AWS, and Google Cloud. Although these technologies facilitate efficient processing, storage, and analytics for business analysis, organizations are faced with the dilemma of selecting the appropriate framework because of differences in scalability, automation, and performance. Managed cloud platforms focus on smooth integration and operational efficiency, but companies receive no direct guidance on how to select the optimal pipeline for a given workload, especially when working with real-world, heterogeneous datasets such as Yelp. This research delves into the challenges of big data processing, examining primary inefficiencies and architectural trade-offs to offer insights into workflow optimization for data, business analysis, and decision-making. Furthermore, this work not only compared the platforms but also offers some guidance on how to choose the best processing pipeline specific to a complex business dataset like Yelp.

Keywords: big data, cloud computing, machine learning, data analytics.

INTRODUCTION

The swift proliferation of data between sectors has made scalable big data platforms a necessity. Two such widely used ones are Azure and Amazon Web Services (AWS), and both of them possess varying strengths: Azure provides Synapse Analytics, Data Lake, and Databricks for natively cloud-integrated and hybrid implementation, whereas AWS provides Redshift, S3, and EMR with the focus being on cloud-native scaling and automation. Yet, organizations cannot figure out which platform suits their company's business data processing requirements because of differences in performance, price, and scalability. Most research compares technical metrics like latency and cost, but few compare how these platforms handle real-world business datasets with structured (business information), semi-structured (user comments), and unstructured text (customer opinions). To overcome this limitation, this research compares Azure and AWS using the Yelp dataset, comparing key factors like scalability under different workloads, cost-effectiveness, and processing efficiency. Apart from its business applicability, this research also benefits academic research in business intelligence, data science, and cloud computing by providing a structured evaluation framework for cloud- based data analytics. The findings serve as a practical guide for firms while offering a reproducible framework for future research on large-scale business datasets.

LITERATURE SURVEY

A. Big Data Frameworks: A Comparative Study

A number of comparisons have been conducted between conventional and contemporary big data frameworks to evaluate their efficiency in processing large data. Hadoop, one of the most popular big data processing frameworks, is inefficient in processing geospatial data, states Eldahshan et al. [1]. This has given rise to specialized frameworks like SpatialHadoop and GeoSpark that improve the efficiency of Hadoop in processing spatial data. Lenka et al. [2] show that GeoSpark is more runtime efficient in big clusters and is hence highly appropriate for applications like disaster management, city planning, and geospatial health systems.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Studies by Manikandan and Ravi [5] have also been concerned with the efficiency of Hadoop's ecosystem, i.e., the Hadoop Distributed File System (HDFS) and MapReduce, in handling big data. Zhou et al. [6] cite that the combination of Apache Hadoop and Apache Spark in web- based tools such as BDViewer has demonstrated a 50% increase in performance, validating the efficiency of hybrid frameworks in big data processing. Sharma and Kaur [7] cite the ongoing battle between Hadoop's MapReduce and Apache Spark's in-memory processing, citing that Spark is better suited for iterative queries and real-time analysis, whereas MapReduce is better suited for batch-processing tasks.

B. Big Data Technology and IoT Breakthroughs

Technological innovations in big data technologies have been extensively discussed by Vijayaraj et al. [9], who talk about various frameworks, tools, and techniques, and the distributed computing, in-memory computing, and hybrid cloud models. They highlight the significance of cloud-native big data platforms in maintaining infrastructure costs low and performance high.

In the context of IoT and big data, Ahmad et al. [17] describe how big data analytics can be utilized to handle high-speed data streams that are produced by IoT devices. They describe how real-time processing systems like Apache Kafka, Apache Flink, and Google Cloud Dataflow can be utilized to process large-volume IoT data streams. Faizan and Prehofer [15] also highlight the requirement for effective stream processing tools in order to prevent data loss and provide guaranteed delivery of data in IoT applications.

C. Real-time data streaming and processing

One of the major characteristics of big data analytics is the processing of data in real time. Le Noac'h et al. [11] introduce a discussion on the effectiveness of Apache Kafka in distributed stream computing, with special emphasis on faults, event-time processing, and scalability in high-throughput environments. Tableau and Power BI have been compared in terms of performance in interactive visualization, with Ali et al. [3] observing that although Tableau is renowned for better interactivity and simplicity, its licensing cost is extremely high and can be considered a demerit. Power BI is more versatile to be combined with Microsoft services but is subject to performance problems when it processes extremely large datasets.

D. Big Data Analytics Tools Complete Survey

Finally, a systematic comparison of various big data analytics tools has been provided by Sahu et al. [20], analyzing their performance at the different phases of data management— collection, storage, processing, and visualization. Tools like Tableau, MongoDB, and Hadoop have been identified with their respective strengths, and future research will involve improving their interoperability across industry-specific use cases. Alalawi et al. [21] provide an overview of AWS Cloud development tools and services, providing a comprehensive survey of AWS Cloud tools, and discussing challenges in terms of tool selection based on specific use cases.

E. Data Storage and Management

Data storage and management are fundamental requirements for big data applications. Comparative studies of real-time databases such as MongoDB and Firebase with big data query engines such as Impala and Hive are conducted by Mustafa et al. [8] and conclude that Impala provides better performance in processing large volumes of data. Comparative studies of storage facilities such as HDFS, Amazon S3, and Amazon RDS for real-time big data applications, each with their respective advantages, are conducted by Jamal et al. [18]. Also, the studies on the efficiency of Apache Pig and Apache Cassandra in Hadoop environments have concluded that Cassandra is more suitable for real-time activities because of its distributed nature, while Pig is more suitable in processing large volumes of structured and unstructured data [19].

F. Machine Learning and Data Analytics

Machine learning techniques are becoming more and more popular in big data processing to enhance data processing efficiency. Smith and Zhang [4] propose a hybrid scheme combining machine learning-based data reduction and GPU-based rendering and demonstrates a 40% improvement in rendering performance. Pandey and Bist [12] provide AI-based decision- making and data security frameworks and propose the AI-driven Information Value Chain

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

framework to solve problems such as algorithmic bias, real-time processing constraint, and security threat. Gupta and Kumari [13] conduct comparative research on programming languages, i.e., Python and Scala, for big data processing and conclude that although Scala provides better execution time and scalability, Python is easier for developers to use and used mainly in data science.

Various visualization techniques, including 1D, 2D, and 3D visualizations, have been investigated to mitigate volume, variety, and velocity of large data challenges. Tableau, D3.js, and Plotly have been identified as versatile visualization tools, with AI visualizations poised to revolutionize healthcare, finance, and cybercrime industries [14]. Sahu et al.

G. Big Data Scalability Challenges and Solutions

Big data scalability is still a challenge, especially in distributed computing systems. Pushpaleela et al. [16] make their application modernization strategies in AWS Cloud hinge on the significance of migrating legacy systems to cloud-native architecture. They talk about prominent strategies such as "Lift-and-Shift," "Re-platforming," and "Refactoring" which enable organizations to migrate applications to AWS based on complexity and business needs. Zhao-hong et al. [10] refer to the capability of the Apache Spark and TensorFlow frameworks to process large-scale data without compromising security and privacy. New encryption algorithms for real-time encryption, differential privacy, and secure multiparty computation are likely to improve privacy protection in big data systems.

EXISTING WORK

Miryala et al. [22] compare the capabilities of AWS and Azure, focusing on important features such as compute power, storage, networking, database management, and pricing models. They highlight that while AWS has superior global scalability, Azure has superior enterprise integration, particularly with Microsoft-based systems. According to their findings, AWS's EC2 and Lambda offer superior scalability over Azure's Virtual Machines and Functions, but Azure's hybrid cloud architecture provides greater flexibility for enterprise applications. Daniel et al. [23] elaborate on the above by comparing big data analytics capabilities in AWS, Azure, and Google Cloud, emphasizing the capabilities of services such as Hadoop, Spark, AI- based analytics, and real-time processing. Again, they fail to offer a comparative cost analysis of AWS tools such as Glue, EMR, and Redshift with Apache-based solutions. Borra et al. [24] shift the focus to serverless computing, comparing its impact on scalability, cost-effectiveness, and operational efficiency in AWS, Azure, and GCP. While they make interesting observations, they fail to compare the impact of serverless architectures on processing big data workloads with large-scale data, such as those based on Apache Spark and Hadoop. Similarly, Oladimeji [25] compares the efficiency of data pipelines between AWS Glue and Azure Data Factory, evaluating processing speed, scalability, fault tolerance, and cost-effectiveness. While the study provides interesting benchmarks, it offers limited discussion of data transformation efficiency. All these studies identify the comparative strengths and weaknesses of cloud platforms in various areas, but they leave a gap in the evaluation of the entire gamut of big data processing workflows, particularly in terms of cost, scalability of performance, and end-to-end data transformation.

PROPOSED SOLUTION

The vast amount of research on big data frameworks points to the increasing use of cloud- based and hybrid solutions. Azure, AWS, and Google Cloud are some of the platforms that offer varied tools for storage, processing, and analytics, supporting different scalability and deployment requirements. Amongst them, Azure and AWS stand out due to their extensive portfolio of services and ability to address complex big data challenges. Azure enables hybrid deployments through offerings like Synapse Analytics, Data Lake, and Databricks, whereas AWS emphasizes cloud-native automation through offerings like EMR, Kinesis, and Redshift.

The special feature of this study is the application of these platforms to the Yelp dataset in real life, which is a sophisticated business dataset with structured data (business data), semi- structured data (tips and check-ins), and unstructured text (reviews). This method not only measures technical performance but also assesses whose big data pipeline—Azure's or

AWS's—is more appropriate for business analytics. Though both platforms have become a mainstay, an end-to-end comparison of all steps in a big data pipeline is still lacking. Trade- offs in performance, cost, and enterprise

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

integration are essential to understand when selecting a cloud solution. Azure's hybrid approach and AWS's cloud-native automation dictate choice according to business needs.

To fill this gap, the current research establishes a systematic methodology for a direct side-by- side comparison of both platforms. The suggested architecture—Data Source, Processing, and Visualization Layers—guarantees simplicity and modularity, allowing for unbiased comparisons in the same conditions. A decision framework is also provided to assist practitioners in choosing the most appropriate processing pipeline according to dataset properties and workload needs. By using this systematic framework for the Yelp dataset, the paper presents actionable insights to businesses for maximizing their big data processes and selecting the most appropriate platform for business analytics.

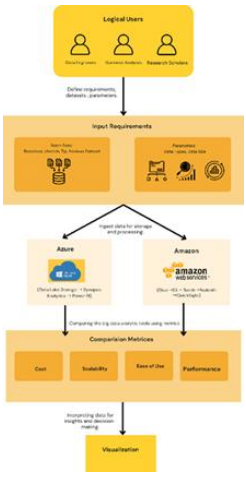

Figure. 1: System Architecture

The architecture of the system determines major stakeholders such as data engineers, business analysts, and researchers who specify the requirements, datasets, and metrics for evaluation. It then splits into two concurrent streams: one using Azure's cloud services (e.g., Azure Data Lake, Synapse Analytics, Databricks, Power BI) and the other using AWS cloud-native offerings (e.g., S3, Sage Maker, EMR, Quick Sight). This two-pipeline strategy allows for a systematic comparison of performance, scalability, and cost under the same workloads and actual conditions.

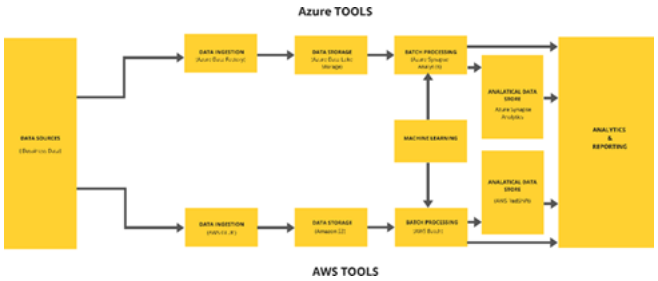

Figure. 2: Design Flow

To make the evaluation platform-independent, the system adheres to an abstract subsystem model that normalizes the data flow through important phases like ingestion, storage, processing, machine learning, analytical data storage, and visualization. The abstraction makes sure that results are not implementation-specific and can be applied across various big data environments. Instead of a vendor-specific approach alone, which would restrain insights within one paradigm, this two-stream architecture allows for an equitable comparison, holding the subtleties of both the cloud-based environment of Azure and managed cloud offerings of AWS.

CONCLUSION

Whereas some studies have made side-by-side comparisons between Azure and AWS pipelines, most assessments compare individual aspects—e.g., single performance benchmarks, cost, or individual pipeline phases—using synthetic or domain-specific data. In contrast, this work suggests a holistic, end-to-end evaluation framework that

addresses all phases of the big data processing pipeline from data ingestion and storage to processing, analysis, and visualization using the real-world Yelp dataset, a complex business dataset. Our suggested methodology provides insights into how organizations can evaluate Azure's offerings (Data Factory, Data Lake, Synapse Analytics, and Databricks) in addition to AWS's offerings (Redshift, S3, and AWS Glue) under similar conditions. It also proposes a decision framework to help practitioners choose the best pipeline setup in accordance with given workload characteristics and dataset needs. By providing such a structured framework, our contribution gives actionable evidence-based advice allowing businesses to coordinate their big data plans with the most efficient and cost-effective cloud platform, closing a large research gap.

REFRENCES

[1] ElDahshan, K., Elsayed, E., and Mancy, H., "Comparative Analysis of Tools for Big Data Visualization and Challenges," IAENG International Journal of Computer Science, 2024.
[2] Lenka, R. K., Barik, R. K., Gupta, N., Ali, S. M., Rath, A., Dubey, H., "Comparative Analysis of SpatialHadoop and GeoSpark for Geospatial Big Data Analytics," 2nd International Conference on Contemporary Computing and Informatics (ic3i), pp. 484–487, IEEE, 2016.
[3] Ali, S. M., Gupta, N., Nayak, G. K., Lenka, R. K., "Big Data Visualization: Tools and Challenges," Proceedings of the 2nd International Conference on Contemporary Computing and Informatics (IC3I), 656– 661, IEEE, 2016. DOI: 10.1109/IC3I.2016.7918084.
[4] Smith, J., Zhang, Y., "Big Data Visualisation - An Update until Today," IEEE Transactions on Visualization and Computer Graphics, 29(4), 1234–1248, 2023.
[5] Manikandan, S. G., Ravi, S., "Big Data Analysis using Apache Hadoop," Proceedings of the 2014 IEEE Conference on Big Data, 2014.
[6] Zhou, H., Chen, J., Zhang, Y., "BDViewer - A Web-Based Big Data Processing and Visualization Tool," Proceedings of the International Conference on Big Data and Cloud Computing, 157–165, 2021.
[7] Sharma, M., and Kaur, J., "A Comparative Study of Big Data Processing: Hadoop vs. Spark," 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2019, pp. 1073-1077.
[8] Mustafa, S. M. N., Farooq, M. U., Zehra, S. S., and Noronha, J. P. T., "A Comparative Study of the Performance of Real time databases and Big data Analytics Frameworks," 2023 7th International MultiTopic ICT Conference (IMTIC), Jamshoro, Pakistan, 2023, pp. 1-7, doi: 10.1109/IMTIC58887.2023.10178651.
[9] Vijayaraj, J., Saravanan, R., Victer Paul, P., and Raju, R., "A comprehensive survey on big data analytics tools," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/GET.2016.7916733.
[10] Zhao-hong, Y., Hui-yu, W., Bin, Z., Zhi-he, H., and Wan-lin, L., "A literature review on the key technologies of processing big data," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2018, pp. 202-208, doi: 10.1109/ICCCBDA.2018.8386512.
[11] Le Noac'h, P., Costan, A., and Bougé, L., "A performance evaluation of Apache Kafka in support of big data streaming applications," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 4803-4806, doi: 10.1109/BigData.2017.8258548.
[12] Pandey, M., and Bist, A. S., "A Study of Big Data Analytics: Tools, Applications, and Information Value Chain," 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690840.
[13] Gupta, Y. K., and Kumari, S., "A Study of Big Data Analytics using Apache Spark with Python and Scala," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 471-478, doi: 10.1109/ICISS49785.2020.9315863.
[14] R. S. Raghav, S. Pothula, T. Vengattaraman and D. Ponnurangam, "A survey of data visualization tools for analyzing large volume of data in big data platform," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2016, pp. 1-6, doi: 10.1109/CESYS.2016.7889976.
[15] M. Faizan and C. Prehofer, "Managing Big Data Stream Pipelines Using Graphical Service Mesh Tools," 2021 IEEE Cloud Summit (Cloud Summit), Hempstead, NY, USA, 2021, pp. 35-40, doi: 10.1109/IEEECloudSummit52029.2021.00014.

Research Article

[16] R.C. Pushpaleela, S. Sankar, K. Viswanathan and S.A. Kumar, "Application modernization strategies for AWS cloud," 2022 1st International Conference on Computational Science and Technology (ICCST), 2022.
[17] M. Ahmad, S. Kanwal, M. Cheema and M. A. Habib, "Performance Analysis of ECG Big Data using Apache Hive and Apache Pig," 2019 8th International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, 2019, pp. 2-7, doi: 10.1109/ICICT47744.2019.9001287.
[18] A. Jamal, R. Fleiner and E. Kail, "Performance Comparison between S3, HDFS and RDS storage technologies for real-time big-data applications," 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 2021, pp. 000491-000496, doi: 10.1109/SACI51354.2021.9465594.
[19] Y. K. Gupta and T. Mittal, "Comparative Study of Apache Pig Apache Cassandra in Hadoop Distributed Environment," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1562-1567, doi: 10.1109/ICECA49313.2020.9297532.
[20] S. K. Sahu, M. M. Jacintha and A. P. Singh, "Comparative study of tools for big data analytics: An analytical study," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 37-41, doi: 10.1109/CCAA.2017.8229827.
[21] A. Alalawi, A. Mohsin and A. Jassim, "A survey for AWS cloud development tools and services," IET Conference Proceedings CP777, vol. 2020, no. 6, pp. 17-23, September 2020.
[22] Naresh Kumar Miryala, Cloud Performance A Comparative Study of Aws vs. Azure, International Journal of Computer Engineering and Technology (IJCET), 15(2), 2024, pp. 208-223.
[23] Daniel, S., Brightwood, S. and Oluwaseyi, J., 2024. Cloud-based big data analytics (aws, azure, google cloud).
[24] Borra, P. and Pamidipoola, H.P., 2025. Serverless Computing: The Future of Scalability and Efficiency with AWS, Azure, and GCP. Future, 5(2)
[25] Oladimeji, O. (2023). Enhancing Data Pipeline Efficiency Using Cloud-Based Big Data Technologies: A Comparative Analysis of AWS and Microsoft Azure. Journal of Multidisciplinary Research and Innovation, 2(1), 11-19.