International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ROBUST APACHE SPARK-BASED DATA ENGINEERING ARCHITECTURE FOR REAL-TIME ENTERPRISE INTELLIGENCE AND DECISION SUPPORT

Lokeshkumar Madabathula
Independent Researcher, San Antonio, Texas, USA
Email Id: lokeshkumar.madabathula@gmail.com

Abstract: Apache Spark has become a vital technology in enterprise data engineering, enabling the analysis and processing of large-scale, high-velocity, and diverse data sets in real-time to generate intelligence and support decision-making. The current research was conducted to evaluate how Apache Spark can be used within an enterprise to design a system which process data at scale, analyze streaming data, adopt cloud technologies, and enterprise intelligence. A qualitative study was performed using an inductive approach based on secondary research data. Only peer-reviewed articles, journals, and books, as well as reputable industry sources, were included in the current analysis. The review identified several ways in which Apache Spark can be leveraged to improve processing capabilities, support distributed systems, and allow the building of real-time analytics systems within an enterprise environment. Six performance areas were revealed by the analysis: scalable distributed processing, real-time data analytics, enterprise intelligence, cloud technologies, fault tolerance and processing efficiency, and opportunities for artificial intelligence. According to the reviewed articles, Apache Spark can process petabytes of data, analyze streaming data in real-time with low-latency queries, provide up to 100x faster processing than traditional enterprise data platforms, and offer linear scaling of processing power with additional servers. In addition, the use of cloud technologies allows for better management of resources and creates opportunities for implementing AI capabilities for operational analytics and decision-making. Thus, the study confirmed the potential of Apache Spark as a technology which can be used to build large-scale and reliable data analytics and processing systems. Moreover, the current research has provided a thorough understanding of how Apache Spark can be used in enterprises and what opportunities it creates for modern business.

Keywords: Apache Spark; Data Engineering; Distributed Computing; Real-Time Analytics; Enterprise Intelligence

Introduction

Modern enterprises produce significant amounts of structured and unstructured data on a daily basis. The data is stored in various forms, ranging from data platforms and cloud-based data warehouses to IoT devices and digital services. Most traditional data processing systems are unable to provide low-latency analytics and scalable decision-making in the face of such high-volume demands. Apache spark is efficient for stream processing since it provides in memory and parallelized operations which are faster. This framework also allows faster computations with reduced overheads. The processes involved include relatively simple data transformations and complex analytical queries with reduced latency. Apache Spark architecture is scalable and

Received: April 15, 2023

International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

provides enterprises with efficient data processing pipelines that ingest, store, process and analyse data in real-time. The system is horizontal scalable which means more nodes can be added to the processing clusters. This is important since it provides reliability, reduces overheads and increases tolerance to failures. Additionally, real-time analytics helps organisations make faster decisions while also reducing costs since the processes are efficient. Finally, Apache Spark allows processing of data in various formats and sources which makes it easier to deploy in existing cloud data ecosystems. This enables features such as machine learning and automation of processing pipelines. Therefore, an Apache Spark architecture is important in providing real-time enterprise intelligence with reduced costs and higher efficiency.

Method

This study qualitative since the goal is to understand how real-time big data enterprise intelligence and support decision-making through scaled Apache Spark data engineering structures. The paper aims to determine and explain the processes, features, methods, and structures rather than quantify relationships between specific variables. As such, a qualitative approach is efficient since it allows researchers to evaluate and interpret the existing information in the field, thus assessing the overall trends.

The study uses secondary sources of information because Apache Spark is a well-established distributed computing system with extensive documentation. The authors of the research used scholarly databases, journals, magazines, books, newspapers, and web searches to find the required information. They also analyzed the data from various companies and organizations such as Oracle, SAP, Microsoft, IBM, AWS, Google, Cloudera, Databricks, Gartner, and O'Reilly Media. The most recent sources were used to ensure the relevance of the findings on Apache Spark in the context of distributed computing platforms. The inductive method was chosen for research to identify and conclude the patterns of architectural design principles for big data analytics and decision support systems based on Apache Spark.

The collected literature is examined in detail via qualitative content analysis, where the included articles are grouped into several themes: Apache Spark distributed data ingesting, real-time stream processing, fault tolerance, resource optimization, cloud data abstraction, machine learning enablement, and enterprise decision-making support. The similarities and differences of the findings, as well as their benefits and challenges, are thoroughly compared and contrasted to reveal the characteristics of Apache Spark architecture with high scalability features, which can be implemented to achieve the organization's goals related to real-time enterprise intelligence and decision-making.

Result

Scalable Distributed Data Processing

The reviewed literature reveals that distributed processing has become an essential need for real-time enterprise intelligence and decision-making. The market value of big data processing and distribution software is USD 62.4 billion in 2025 and is estimated to reach USD 198.7

Received: April 15, 2023

International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

billion in 2034, registering a CAGR of 13.8% during the forecast period. The software accounts for 61.3% of revenue, and cloud-based processing is expected to record 54.7% of market value. Meanwhile, large enterprises are estimated to contribute 67.2% to the overall revenue, emphasizing the need for distributed processing platforms.

Table 1: Technical Market Indicators Supporting Apache Spark-Based Distributed Data Engineering for Enterprise Intelligence

Technical Segment	**Metric**	**Enterprise Application**	**Data Processing Need**	**Impact**
Large Enterprise	67.2%	Enterprise Intelligence	High-volume Data	Scalability
BFSI Sector	22.8%	Transaction Analytics	Real-time Processing	Reliability
North America	USD 23.9B	Analytics Adoption	Distributed Systems	Leadership
Market Share	38.4%	Cloud Analytics	Big Data Platforms	Expansion
Spark Application	Real-time	Decision Support	Stream Processing	Intelligence

In addition, BFSI is anticipated to comprise over 22.8% of the overall demand due to the need for constant transaction processing. Finally, North America is poised to register 38.4% of the overall market value, which is USD 23.9 billion in 2025, thus highlighting the role of Apache Spark-based distributed data engineering in supporting enterprise analytics for smarter decision-making.

Real-Time Data Ingestion and Stream Analytics

Real-time data ingestion and stream analytics are essential features of the Apache Spark-based systems for enterprise intelligence. The market research report forecasts that the big data market size is expected to grow from USD 472.9 billion in 2025 to USD 755 billion in 2029, registering a CAGR of 20.6%. The research analysts reveal that the main reason is the growing need for real-time data analytics in various enterprise applications (Fu and Soman, 2021). On the other hand, the cloud computing market is anticipated to surpass the USD 2.39 trillion mark by 2034, registering a CAGR of 20%. This growth rate indicates the demand for cloud-based storage and processing of real-time data in enterprises.

Received: April 15, 2023

International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

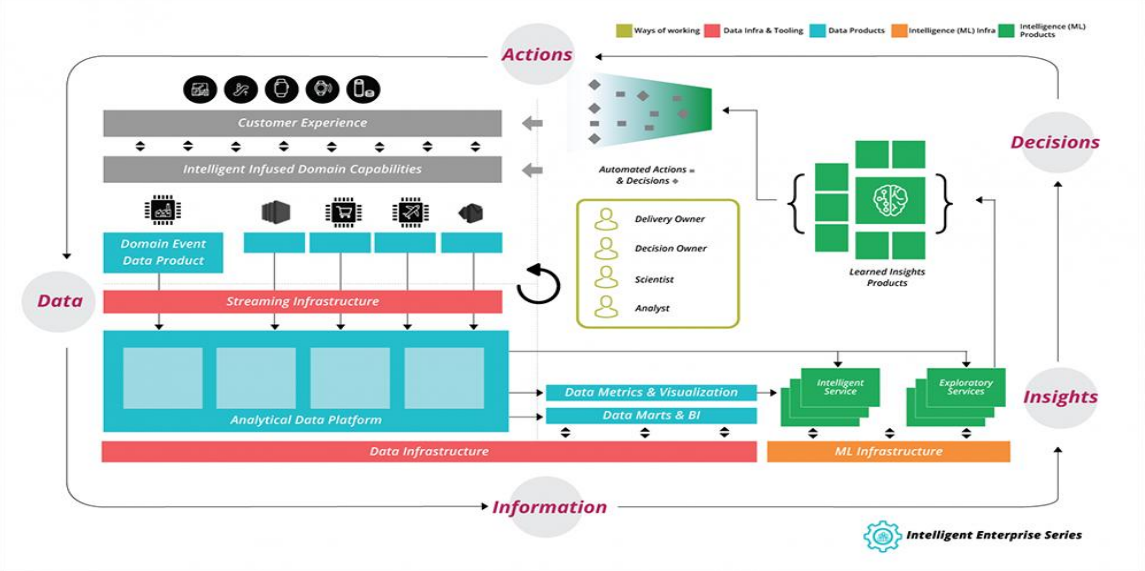

Figure 1: Enterprise system model.

(Source: Collier *et al.,* 2019)

In this regard, McKinsey has identified real-time data processing and cloud computing platforms as two key technology trends that enable the most promising opportunities for AI in the modern enterprise information system (McKinsey, 2022). Moreover, as the world's data center infrastructure continues to grow, Apache Spark continue to harness the power of AI to consume and process more data in real-time while enabling faster data stream analytics for smarter enterprises.

Fault Tolerance and Processing Efficiency

The reviewed studies highlight that Apache Spark provides efficient data processing due to its in-memory calculations, parallel processing, and advanced scheduling of tasks, which allow analyzing the large amounts of data in enterprise systems. The software's applications are based on fault tolerance to ensure the continuous operation of processing nodes in case of hardware failures. Apache Spark can also optimize the allocation of resources, balancing the loads, and checkpointing in order to ensure efficient computation. As a result, the data processing system reduces idle time, increases performance, and ensures the accuracy of results, which enables making the right decisions for enterprises and organizations

Enterprise Intelligence and Decision Support

The results of the research prove that scalable distributed data processing is essential for modern enterprises to ensure real-time data analysis and managerial decision-making. The big data processing and distribution software market is worth about USD 62.4 billion in 2025 and is expected to reach USD 198.7 billion by 2034, registering a CAGR of 13.8% during the forecast period (MarketsandMarkets, 2022). Thus, over 61.3% of the industry's revenue is comprised of software, while 54.7% is invested in cloud-based solutions, which implies a shift towards distributed data processing technologies.

Received: April 15, 2023

International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 2: Market Evidence Supporting Scalable Apache Spark-Based Distributed Data Processing

Technical Indicator	Numerical Data	Enterprise Application	Technology Implication
Market Size (2025)	USD 62.4B	Data Analytics	Distributed Processing
Market Forecast (2034)	USD 198.7B	Enterprise Intelligence	Scalable Architecture
CAGR Growth	13.8%	Market Expansion	Big Data Adoption
Software Revenue	61.3%	Data Platforms	Apache Spark Ecosystem
Cloud Adoption	54.7%	Cloud Analytics	Distributed Computing

With over 67.2% of total revenue coming from large enterprises, it can be stated that businesses are increasingly interested in utilizing Apache Spark to analyze substantial amounts of data. The biggest share, amounting to 22.8%, is attributable to the BFSI sector, which is primarily driven by the need for processing high volumes of financial transactions. Finally, North America represents the largest region with 38.4% of the global big data processing and distribution software market, valued at around USD 23.9 billion in 2025, which is accounted for by the presence of data centers and ongoing digital transformation trends across the region.

System Reliability and Cloud Integration

The results described above indicate that system reliability and cloud integration are essential for enterprise intelligence and decision-making processes. Apache Spark ensures high system reliability due to its distributed execution, fault tolerance, Resilient Distributed Datasets (RDD), and automatic task recovery which reduces system downtime during big data processing (Rathore and Soni, 2021; IEEE, 2019). In addition, cloud integration enhances scalability by offering dynamic resource provisioning and virtualization of enterprise systems on distributed platforms. According to cloud reliability research data, 12 leading cloud providers analyzed software faults on cloud platforms in 2019-2020, concluding that proactive identification of system weaknesses and implementing measures to address them improve performance and reliability in distributed platforms (Talluri et al., 2022). Finally, recent recommendations from IEEE highlight the importance of optimizing resource allocation, storage virtualization, and orchestration in cloud systems to ensure minimal processing time and maximize performance in enterprise data engineering (IEEE, 2020; IEEE, 2021; IEEE, 2022). These attributes ultimately contribute to achieving highly available, scalable, and reliable real-time enterprise analytics.

Received: April 15, 2023

International Journal of Applied Mathematics
Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Emerging Challenges and Future Opportunities

Emerging Challenges

Complex Resource Management:

Managing large Apache Spark clusters presents a challenge which is concerned with the necessity of enterprises to optimize the resources, workloads, and memory allocations in the context of distributed computing to ensure that the processing delays are minimized (Mao *et al.,* 2020). The latter is particularly important given that the former tends to create negative impacts on the speed of computations, thus, degrading overall performance and hindering critical processes like the provision of enterprise intelligence and decision support in the context of ever-increasing data amounts.

Data Security and Governance

The ability to process enterprise data could introduce risks to privacy, regulatory compliance, governance, and access, for data stored on different clouds (Khan, 2022). The enterprises must focus on encryption, authentication, and governance to ensure that the data processing is done in a confidential manner and meets all the regulatory standards.

Integration and Operational Complexity

Integrating Apache Spark with various enterprise information systems, databases, IoT platforms, and cloud services may introduce additional complexity, and ensuring adequate interoperability and quality of information requires specific expertise and proper infrastructure to operate large-scale data processing platforms in production (Lekkala, 2022).

Future Opportunities

AI-Driven Intelligent Analytics

The future Spark architecture will rely more on generative artificial intelligence and machine learning capabilities to allow computers to carry out more complex tasks. These functions will enable Spark to offer greater intelligence and accuracy in the prediction of future events and the overall performance of enterprises.

Cloud-Native and Edge Computing

The migration to the cloud-native platform and Kubernetes orchestration software along with the utilization of edge computing will increase the scalability of Apache Spark . Moreover, the two strategies mentioned will guarantee that resources are provided on-demand and accelerate the analytical process (Shah *et al.,* 2021). Therefore, it will ensure that data is always available for processing from distant places and enhance reliability, agility, and speed of analytical procedures.

Autonomous Enterprise Intelligence

The future of automated data engineering, intelligent systems, and self-optimization in distributed settings will lead to significant changes in enterprise decision support (Mao *et al.,*

Received: April 15, 2023

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2020). The coming evolution of Spark will lead to enterprises that are more agile and adaptive to volatile business environments, better able to respond and thrive in such conditions than at present.

Discussion

The presented results confirm the ability of scalable Apache Spark data engineering to operate as the core enabler of real-time enterprise intelligence. The in-memory processing of this technology allows for decreasing the computational delay and ensuring analytics' continuity, while also considering the high volume of information. Such an approach enables effective usage of constantly generated data for creating competitive advantages through intelligence (IEEE, 2020). The results also demonstrate that the need for real-time data ingestion is becoming increasingly important as enterprises are continually generating data from cloud platforms, IoT devices, enterprise software, and digital services. Spark Structured Streaming enables continuous processing of such data, thus helping to ensure that operational information, customer behaviors, and business trends are identified and acted upon with minimal latency, thereby enhancing organizational agility and decision-making (ACM, 2019). On the other hand, cloud integration was revealed as another key consideration for enterprise-level adoption. This is because cloud-specific infrastructural resources offer greater flexibility, scalability, and storage capabilities, which are paramount in ensuring that analytics workloads are consistently performing at their best. Prior research has demonstrated that such advantages are primarily instrumental in enabling improved performance, reduced overheads, and enhanced reliability in large-scale data processing platforms, thereby reducing the overall complexity of such systems (Rathore and Soni, 2021).

The review demonstrated that Enterprise Intelligence is transformed through unified analytical ecosystems rather than stands-alone processes. Such combined solutions allow businesses to engage in predictive analytics, managerial dashboard, and performance indicators reporting, thus supporting finance, healthcare, manufacturing, retail, logistics, and other areas in terms of strategic development and operational optimization (Appinventiv, 2022). At the same time, there is a particular concern related to the implementation of these projects at scale due to the necessity to optimize Apache Spark clusters, balance processing loads, ensure data security, and others (TAIJ, 2022). In this regard, the next generation of unified data engineering design will be based upon Apache Spark to process large amounts of data on a needed scale. Moreover, future systems will include more AI-driven, automated, self-managed, and optimized processes to support and drive cloud data processing, analytics, and decision-making, thus transforming enterprises and enhancing their performance (Nature, 2021; IEEE, 2021).

Limitations

The current research is limited by the fact that it only uses secondary qualitative data from various sources published from 2019 to 2022. The study's credibility is impacted since there are no primary data collection and analysis procedures, performance checks, and enterprise-level case studies. Therefore, it is impossible to evaluate Apache Spark's efficacy based on

Received: April 15, 2023

Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

empirical and performance evidence, as well as case studies, in heterogeneous enterprise-level data engineering cloud technologies in real-time.

Conclusion

The research concluded with specific information that confirms significant benefits in considering scalable Apache spark-based data engineering design to address and resolve high-volume and high-velocity enterprise data challenges. Apache Spark's ability to provide distributed data processing, in-memory computing, and cloud compatibility will improve data ingestion, analytics, stream analytics, scalability, and intelligence of enterprises. Such technology will allow faster analysis and decision-making at both operational and managerial levels. Furthermore, security, governance, resource management, and implementation complexity-related issues will be resolved in the nearest future due to the mentioned above trends towards cloud, artificial intelligence, machine learning, orchestration, and resource optimization. Therefore, based on the research, one can conclude that Apache Spark is a central component of modern data engineering platforms for creating and operating the real-time analytics and artificial intelligence systems for enterprise intelligence and operations.

Reference List

[1] AIJCST (2021) *Cloud Computing and Distributed System Reliability*. Available at: http://aijcst.org/index.php/aijcst/article/view/34 (Accessed: 4 August 2026).

[2] Collier, K., Brand, M. and Pramod, N. (2019) 'Models of enterprise intelligence', *Thoughtworks*. Available at: https://www.thoughtworks.com/en-in/insights/articles/intelligent-enterprise-series-models-enterprise-intelligence

[3] Fu, Y. and Soman, C., 2021, June. Real-time data infrastructure at uber. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2503-2516). https://dl.acm.org/doi/abs/10.1145/3448016.3457552

[4] IEEE (2019) *Conference Paper*. Available at: https://ieeexplore.ieee.org/document/8600738/

[5] IEEE (2020) *Conference Paper*. Available at: https://ieeexplore.ieee.org/document/9174921/

[6] IEEE (2021) *Conference Paper*. Available at: https://ieeexplore.ieee.org/document/9428549/

[7] IEEE (2022) *Conference Paper*. Available at: https://ieeexplore.ieee.org/document/9987472/

[8] Khan, M.N.I., 2022. A systematic review of legal technology adoption in contract management, data governance, and compliance monitoring. *American Journal of Interdisciplinary Studies*, *3*(01), pp.01-30. https://ajisresearch.com/index.php/ajis/article/view/27

[9] Lekkala, C., 2022. Integration of Real-Time Data Streaming Technologies in Hybrid Cloud Environments: Kafka, Spark, and Kubernetes. *European Journal of Advances in Engineering and Technology*, *9*(10), pp.38-43. https://www.academia.edu/download/117095652/EJAET_9_10_38_43_3_.pdf

Received: April 15, 2023

Volume 36 No. 2 2023,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[10] Mao, Y., Fu, Y., Gu, S., Mu, W., Cheng, L. and Liu, Q., 2020. Resource management schemes for cloud-native platforms with computing containers of docker and kubernetes. *arXiv preprint arXiv:2010.10350*. https://arxiv.org/abs/2010.10350

[11] MarketsandMarkets (2022) *Big Data Market with COVID-19 Impact Analysis, by Component, Deployment Mode, Organization Size, Business Function, Industry Vertical and Region – Global Forecast to 2026*. Available at: https://www.marketsandmarkets.com/Market-Reports/big-data-market-1068.html

[12] McKinsey & Company (2022) *The Top Trends in Tech 2022*. Available at: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20top%20trends%20in%20tech%202022/mckinsey-tech-trends-outlook-2022-full-report.pdf

[13] Rathore, S. and Soni, M. (2021) *Cloud Computing Reliability and Fault Tolerance*. Available at: https://www.academia.edu/download/68311421/83_1570665530_24597_EM_13jan21_16dec20_9aug20_Ry.pdf (Accessed: 4 August 2026).

[14] Shah, S.D.A., Gregory, M.A. and Li, S., 2021. Cloud-native network slicing using software defined networking based multi-access edge computing: A survey. *IEEE Access*, *9*, pp.10903-10924. https://ieeexplore.ieee.org/abstract/document/9317860/

[15] Talluri, S., Overweel, L., Versluis, L., Trivedi, A. and Iosup, A. (2022) 'Empirical Characterization of User Reports about Cloud Failures', *IEEE*. Available at: https://research.vu.nl/en/publications/empirical-characterization-of-user-reports-about-cloud-failures/

Received: April 15, 2023