International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I8P109

Volume 72 Issue 8, 60-69, August 2024

Original Article

Big Data Engineering Using Hadoop and Cloud (GCP/AZURE) Technologies

Shrikaa Jadiga

Independent Researcher, USA.

Corresponding Author : shrikaa99@gmail.com

Received: 12 June 2024 Revised: 17 July 2024 Accepted: 07 August 2024 Published: 29 August 2024

Abstract *- Big Data Engineering is crucial in today's data-driven society, where managing vast amounts of data is key to business success. This paper explores the integration of Hadoop and cloud technologies, specifically Google Cloud Platform (GCP) and Microsoft Azure, to address Big Data challenges. With its components, such as HDFS, MapReduce, and YARN, Hadoop provides a robust framework for distributed storage and processing large datasets. Cloud platforms like GCP and Azure offer scalability, cost-effectiveness, and flexibility, making them ideal for Big Data applications. They support various Big Data tools and provide secure, compliant environments for data processing. By leveraging these technologies, organizations can enhance their data processing capabilities, achieve better resource management, and gain valuable insights from their data. This integration not only optimizes performance but also ensures efficient handling of Big Data, paving the way for innovative solutions and competitive advantages.*

Keywords *- Big data, Hadoop ecosystem, Cloud technologies, Scalability and flexibility, BigQuery.*

1. Introduction

Big Data is defined by its high volume, high velocity, and high variety, encapsulated in the three Vs: volume (the quantity of data), velocity (the rate of data generation and processing), and variety (the different types of data, including formatted, semi-formatted, and unformatted). The primary goal of Big Data analytics is to uncover hidden patterns, unknown correlations, and meaningful insights from large datasets. However, the ever-increasing volume and complexity of Big Data necessitate using more advanced and efficient approaches and tools for storage, processing, and analysis (Gandomi & Haider, 2015). Despite significant advancements, a critical research gap remains in effectively managing and utilizing Big Data for actionable insights. This gap is particularly pronounced in the integration and optimization of Big Data tools and technologies to enhance business processes, customer experiences, and competitive advantage. The role of Big Data Engineering, which involves creating and maintaining structures that support Big Data processes and analyses, is pivotal in addressing this gap (Kambatla et al., 2014). However, there is limited research on the best practices and methodologies for Big Data Engineering, particularly in the context of rapidly evolving cloud technologies. Hadoop, an Apache project, offers an integrated environment for storing and processing Big Data across clusters of computers using a simple programming model. Its capability to scale from a few servers to thousands, each with computation and storage capabilities, makes it a cornerstone of Big Data solutions (White, 2015). However, deploying Big Data applications on cloud computing platforms such as Google Cloud Platform (GCP) and Microsoft Azure, which offer dynamic and versatile resources, introduces additional complexities and opportunities that are not yet fully explored. This research aims to address the existing gaps by examining the integration of Hadoop with cloud technologies, specifically GCP and Azure, to develop optimized approaches for Big Data Engineering. By exploring these areas, this study seeks to contribute to the field by providing insights and recommendations for businesses looking to leverage Big Data for strategic advantage.

1.1. Big Data Engineering Concepts

Big Data Engineering is a crucial field dedicated to developing and maintaining systems to efficiently manage large-scale data. Traditional relational databases struggle with the scalability and performance demands of Big Data, necessitating the use of distributed file systems like the Hadoop Distributed File System (HDFS), which divides and replicates data across multiple nodes to enhance reliability and fault tolerance (Wang et al., 2022). Processing frameworks such as MapReduce and Apache Spark are essential for managing the computational load by breaking down tasks and enabling concurrent processing. Spark's in-memory computing further accelerates data handling, making it a preferred choice for real-time analytics (White, 2020). Data analysis tools like Apache Hive and Pig facilitate querying and

processing large datasets, providing user-friendly interfaces for data manipulation (Guller, 2019). Additionally, ensuring data quality and integrity across distributed systems is a significant challenge, addressed through data cleansing, preprocessing, and conflict resolution mechanisms (Chen & Zhang, 2020). This integrated approach to storage, processing, and analysis underscores the importance of scalable and efficient systems in the field of Big Data Engineering, highlighting ongoing research and development efforts to optimize these technologies. Data storage is among the core competencies of Big Data Engineering since it refers to the foundational step that precedes data ingestion. In this case, efficiency and effectiveness in storing huge volumes of data to enhance their security can hardly be overemphasized. The disadvantage of traditional relational databases is that they cannot scale well to huge data and have poor performance when dealing with large data loads. Thus, distributed file systems have become necessary, including the Hadoop Distributed File System (or HDFS, as it will shortly be referred to). HDFS is the highly available system and distributed file system in Hadoop that stores large data sets on multiple machines. Its structure permits the storage to be in blocks and to achieve data redundancy by storing data blocks on different nodes, which can benefit in case some nodes' hardware fails (Wang et al., 2022).

Besides storage, processing frameworks are other important pillars essential in Big Data Engineering. These frameworks allow for handling large amounts of data by dividing the computational work between multiple nodes or processors. The most popular software framework for processing is MapReduce, associated with Apache Hadoop. MapReduce breaks a problem down into more sub-problems to make it easier to solve and more efficient since it takes less time to work on them concurrently. Another trending framework is Apache Spark, which has been associated with in-memory computing in processing highly complemented data sets compared to disk I/O systems like MapReduce. The reality is that Spark can simultaneously deal with complex tasks, both in the batch data processing sphere and in the real-time stream processing one or machine learning and graph processing, making it indispensable in today's big data sphere (White, 2020). Data analysis tools are a component that also belongs to the Big Data Engineering category. These tools help the data scientist or analyst gain insights from humongous data. Apache Hive, for instance, comes with a data warehouse infrastructure that lies on Hadoop, where users can perform queries on large datasets with a language similar to SQL called HiveQL. This makes it easily understandable for users with SQL experience to easily transition as it combines the essence of the DBMS and the big data platforms. Also, the pig can be referred to as a tool for issuing a high-level script and processing data since it contains a scripting language to perform data transformation and analysis (Guller, 2019). Big Data Engineering systems are composed of one or more of the sources of big data, which are usually distributed systems for the sake of scalability and reliability. Distributed systems are further built in such a way that the loads imposed on a certain node are distributed among different nodes, thereby decreasing the probability of occurrence for bottlenecks and failure points. This means that the architecture is horizontally scalable, which implies that one can scale out by having numerous nodes in the system to accommodate more data or computing. Furthermore, the recovered system in the distributed system involves the replication of the data across the nodes so that even if the number of nodes in systems reduces the capability of the systems, some amount of fault tolerance remains (Wang et al., 2022).

Another critical aspect of Big Data Engineering is solving issues with the input data quality. As the volume and variety of data increases, it is critical for organizations to ensure the quality of input data for precise analysis of business outcomes. The data cleansing process removes duplicity to avoid or minimize the occurrence of duplicated data. In contrast, the data preprocessing takes care of the missing data and, where necessary, eradicates all the errors. Also, achieving data coherency across the distributed systems may be complex since many nodes may update data simultaneously. Issues like these are dealt with in ways that include distributed transactions and conflict resolution mechanisms to ensure data integrity (Chen & Zhang, 2020). The other key concern raised is in the area of data security and confidentiality. As there is an escalating volume of information disclosed and archived, it is paramount to deploy effective security measures to prevent threats from compromising data privacy and integrity. Many error control techniques are employed in data transactions in computer systems and data communication, including data encryption. It is achieved, for example, by setting up access controls and proper authentication procedures that allow authorized users to access the data only. However, meeting the requirements of legal frameworks relating to data protection, like the GDPR, may prevent legal and financial consequences (Hashem et al., 2020).

Data integration and a range of Big Data sources- such as structured, semi-structured, and unstructured data- are essential features of Big Data Engineering. The other type of big data is structured data with a well-defined structure such as in the case of relational databases, and this can be integrated in traditional ETL fashion. However, there is a need for better techniques to integrate and analyze semi-structured data, such as JSON and XML files, and unstructured data, including text and multimedia. Applications such as Apache NiFi and Apache Kafka help collect and combine heterogeneous data, supporting real-time data streaming and processing (García et al., 2022). BD Engineering is an important part of the process of handling and processing large information flows. Big data refers to large datasets that require specific systems to store, process and analyze the data. The different components of these systems include Data storage, Data processing frameworks and Data analytics tools.

A distributed systems architecture guarantees scalability and dependability, thus solving other difficulties encountered regarding data quality, data security and integration. Growing at an increasing rate, big data engineering is an emerging field that witnesses technology enhancements and developments well into the future.

2. Hadoop Ecosystem

As an open-source framework, Hadoop stands in the middle of big data and serves as the core platform for processing large amounts of data. Derived from the Apache software foundation, Hadoop is a framework of tools and solutions that enables efficient storage of large amounts of data, processing the data, and analysis. Some prominent ones include the Hadoop Distributed File System to store data, MapReduce to process the data, and Yet Another Resource Negotiator for managing the resources (White, 2020). HDFS is a system created for storing Big Data in many machines and, at the same time provides the necessary data access rates and fail-safe solutions. In another concept, it divides files into great chunks and then disseminates these chunks across nodes in a cluster and replicates every chunk to render it effective. This replication mechanism keeps HDFS functioning in the presence of hardware failure; thus, it can be considered an efficient solution for storing extremely large amounts of data (White, 2020).

Another component of the Hadoop framework is MapReduce, which works as a processing component in parallel computations on large datasets. It splits work into sub-working units performed in parallel on different nodes, the map functions. The results from the map functions are shuffled and sorted based on specific keys to be read by the reduced functions that combine the results to offer the final result. This mode of processing enables Hadoop to process big data with a speed of petabytes due to this distributed processing model (Guller, 2019). YARN, tackling the resource management layer in Hadoop, improves the system's scalability and resource usage. YARN splits the resource management task from the job scheduling task; this enables multiple data processing languages such as MapReduce, Apache Spark, Apache Flink and many others to run on one Hadoop cluster. Thus, the division provides better resource utilization and helps the system perform various tasks efficiently (White, 2020). Apart from the four parts, several modules are considered a part of the Hadoop ecosystem. There is Apache Hive, which is a data warehousing tool being developed on Hadoop the main purpose of which is to offer a SQL-like language known as HiveQL. Hive will allow users to carry out tasks using standard SQL, such as queries on big data, and it will help bridge the gap between conventional relational data and the Hadoop ecosystem (Guller, 2019). Apache Pig is also an important component in the Hadoop ecosystem, a high-level scripting language for data transformation and analysis. Pig Latin, the language used by Pig, is a scripting language that enables users to write data processing scripts, which are compiled and converted to MapReduce jobs. This makes writing the more complicated data transformation functions easier and handling large data sets easier for the developer (Guller, 2019). HBase is a type of database system that falls under the category of NoSQL as it stores data in real-time mode with read/write access to big data stored in Hadoop's HDFS. It features billions of rows and millions of columns, which are useful to applications that need random and fast access to huge amounts of data. Some operations that can be performed are range scans, operations involving a single row, read/write, and complex queries, providing a versatile and highly efficient way for big data management (Guller, 2019). Apache Spark is another subproject of Hadoop, an in-memory processing framework offering fast and generalized data processing. While MapReduce involves storing intermediate data on the disk, Spark stores intermediate data in the memory, making it faster. Thus, Spark can help with various tasks, including batch processing, real-time stream processing, machine learning, graph processing, etc., making it effective for Big Data applications (Zaharia et al., 2020).

One similarity to big data in the Hadoop ecosystem is based on the fact that the ecosystem comprises a set of tools and technologies that work in harmony and can be integrated easily, thus offering a one-stop shop solution for big data processing. For instance, Hive and Piglet are data analysis languages. Pig provides users with the ability to manipulate large datasets through the use of high-level data processing languages. Similarly, HBase provides users with fast and random access to large datasets. Spark refines the ecosystem with an in-memory processing capacity that empowers applications with boosted data processing efficiency. Overall, due to the flexibility and extensibility of the Hadoop ecosystem, it can be concluded that it is widely adopted by organizations that encounter big data processing and analysis. Due to the flexibility of its operation and ability to process both batch and real-time workloads, Big Data Engineering cannot do without Apache Spark. In addition, it is evident that Hadoop is an open source, which makes it possible for the tool to grow from strength to strength as more individuals engage in its development (White, 2020). Another advantage that is usually associated with the Hadoop ecosystem is its affordability. As a result, by utilizing standard and non-proprietary IT resources, an organization can develop cost-efficient and stable Big Data architectures. This increases the accessibility of organizations to enhance data processing power, which democratizes this power to even organizations that are not as large as other multinationals that possibly may afford it. Also, the ability to communicate with other tech solutions is flexible in the context of the Hadoop ecosystem. For instance, it can synchronize with such cloud conduits as Google Cloud Platform (GCP) and Microsoft Azure so that, based on cloud computing, organizations can evolve to realize better scalability and flexibility. It gives an excellent synergy between the on-premises and Cloud for Big Data processing (Haridal et al., 2018).

Hadoop environment is core to Big Data handling and encompasses a large component of the framework necessary to process such data. HDFS, MapReduce, and YARN, which help deliver distributed storage and processing, represent its core elements. Other related tools such as Hive, Pig, HBase, and Spark enrich its features, making it a comprehensive platform for Big Data Engineering. Thus, as the Big Data field keeps expanding, the Hadoop ecosystem will remain an essential element that shapes the future and fosters development in the data processing domain (White, 2020; Guller, 2019; Zaharia et al., 2020).

Table 1. Hadoop ecosystem components

Component	Description
HDFS	Distributed file system
MapReduce	Distributed processing
YARN	Resource management
Hive	SQL-like querying
Pig	Scripting language
HBase	NoSQL database
Spark	In-memory processing

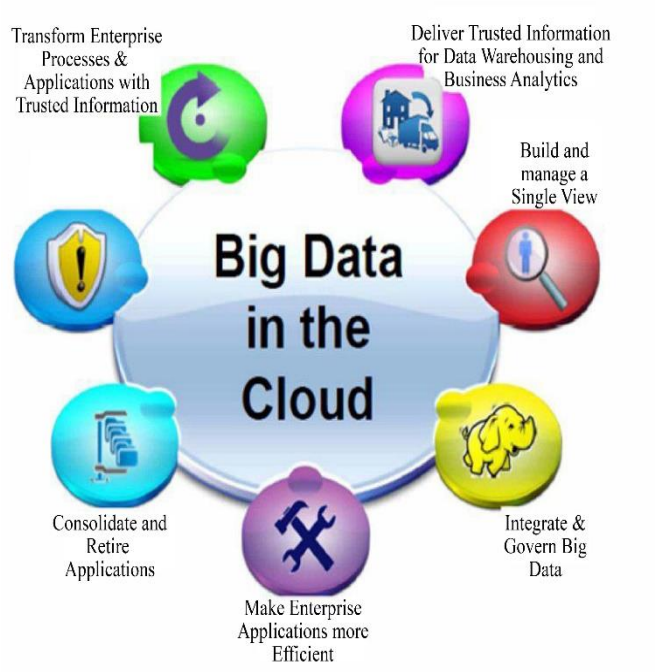

Fig. 1 Cloud technologies for big data

3. Cloud Technologies for Big Data

Cloud computing for Big Data processing allows resource scaling, which is considered one of the greatest assets with benefits such as cost efficiency, flexibility, and simplicity of use (Hashem et al., 2020). Google cloud platform and Microsoft Azure are some huge Cloud providers with vast services for Big Data, specifically an application area (Buyya et al., 2019). Understanding the major elements of the cloud technologies for Big Data and the major clouds, Google Cloud Platform and Microsoft Azure will be insightful.

3.1. Scalability and Flexibility

Another important advantage of cloud platforms is that services can be scaled almost to infinity, which means that depending on the needs of an organization's work, one or another amount of computing resources can be used. This elasticity makes it possible for businesses to easily address difficult workloads without necessarily having to invest heavily in hardware. For instance, GCP can auto-scale features that enable automated adjustments of virtual machines depending on the load's complexity and subsequently save time and costs (Vavilapalli et al., 2020). Likewise, Azure storage and compute resources can be smoothly scaled up or down following the requirements of BIG DATA applications (Li et al., 2021).

3.2. Cost-Effectiveness

Among the key advantages of Big Data applications in the cloud environment, cost reduction is considered to be the main one. On-premise infrastructure is a capital-intensive model wherein a significant investment is required to acquire hardware, especially for EU countries that witnessed cutthroat competition and deteriorating business confidence, leading to lower spending on IT infrastructure. On the other hand, cloud services operate on the usage-based model, meaning they only charge organizations for the services they have utilized. It saves money by not requiring a heavy one-time investment and is cheaper compared to constant daily running expenses. GCP and Azure provide differential, flexible prices and quotas, ensuring that companies at any stage of evolution can use high-technology Big Data (Hashem et al., 2020).

3.3. Ease of Use and Accessibility

Cloud platforms have interfaces and tools that make deploying Big Data applications fairly easy and manageable. For instance, Big Query is one of the services provided by GCP that enables users to execute queries in the form of SQL on enormous data without needing to comprehend the system's deeper levels (Bifet et al., 2019). Azure's Synapse Analytics is a big data analytics workspace that has combined data ingestion, management, and serving capabilities with data warehousing to improve data loading, transformation, and analyzing to meet real-time, day-one business intelligence and machine learning application requirements, according to Armbrust et al., 2019.

3.4. Integration with Big Data Tools

Most Big Data tools and technologies are compatible with cloud platforms. Apache Hadoop, Apache Spark, TensorFlow, and the like are fully supported on the Google Cloud platform, making it easy to establish complex data processing and Machine learning pipelines.

3.4.1. Dataproc

Dataproc is a fully managed Hadoop and sparks service that is available on the GCP platform, which enables users to create a cluster and job management without the need to deal

with the physical environment of the infrastructure (Vavilapalli et al., 2020). Azure also strongly supports Big Data frameworks for HDInsight as a managed service for running the Hadoop, Spark, and Kafka analytics services in the Cloud (Isard et al., 2020).

3.5. Security and Compliance

This recognizes data security and compliance as paramount in Big Data applications. Cloud providers provide different security features, such as data encryption, control of access rights, and network security to protect sensitive data. GCP and Azure also adhere to different standards and regulations, including GDPR, HIPAA, and SOC 2; this gives business entities confidence that their data is safe and meets the required standards (Hashem et al., 2020).

3.6. Case Studies and Real-World Applications

Many examples exist of how various organizations have effectively adopted and implemented cloud technologies to address Big Data challenges. For example, in place of handling and storing large amounts of data themselves, Spotify relies on GCP for big data real-time analysis for user behaviour to enhance their recommendation feature. It has also enabled Spotify to integrate large-scale data processing and data analysis at a fast pace and at a relatively low cost (Bifet et al., 2019).

Likewise, e-commerce companies utilize Microsoft Azure to evaluate customers' information and use the results to enhance their marketing solutions, which is enabled by Azure's broad portfolio of analytics tools (Armbrust et al., 2019). Therefore, cloud technologies offer significant advantages for Big Data processing, including scalability, cost-effectiveness, ease of use, and robust security. Platforms like GCP and Azure provide comprehensive services and tools that enable organizations to harness the power of Big Data efficiently. By leveraging these cloud technologies, businesses can gain valuable insights, drive innovation, and stay competitive in the data-driven world (Hashem et al., 2020; Buyya et al., 2019).

3.6.1. Google Cloud Platform (GCP)

Big Data on Google Cloud Platform (GCP) Services are all-encompassing tools that can be used for Big Data processing within an organization's operation. Some of the key services are data warehousing using BigQuery, stream and batch data processing using Dataflow, and Apache Hadoop and Spark job execution using Dataproc (Vavilapalli et al., 2020).

3.6.2. BigQuery

BigQuery is a data warehousing service that runs on the Google Cloud Platform. It is a serverless product that allows one to query large datasets using SQL. It is intended to deal with petabytes of data and is aimed to include real-time analytics and reporting. The storage and the computation are managed as two distinct services, which enables each of them to be scaled independently in BigQuery. This flexibility allows the services to be charged by using resources in a way that makes the solution affordable for Big Data applications.

4. Dataflow

There is a practical transferring and transformation service named Dataflow, which is aimed at stream and batch data pipelines and makes the process much easier. This is accomplished using Apache Beam, an open-source programming model for defining data processing pipelines. Dataflow includes resource allocation, scale-up and scale-down, and other resource management tasks as a part of the service, so that the developers do not have to bother aboutorry about the architecture of the infrastructure once it is set up. This service is perfect for tasks like processing stream data, data munging, data transforming, and data loading processes, and it fits within ML pipelines.

4.1. Dataproc

Dataproc is an automatically managed solution for running Apache Hadoop and Spark clusters. The functionality lets users easily and rapidly create and manage clusters, owing to GCP's base. Dataproc supports many big data tools such as Hive, Pig, HBase, etc., making Dataproc suitable for many big data solutions. The integration with Hadoop empowers the possibility of the effective and linear treatment of big data and big workload data in the Cloud, satisfying the flexibility demand of many businesses (Bifet et al., 2019). In the diagram above, GCP's big data services are depicted together with emphasis on how the three elements, namely BigQuery, Dataflow and Dataproc, are intertwined. These services collectively provide integrated big data analytics for stream data and batch data analysis services.

4.2. Microsoft Azure

Microsoft provides a set of services for big data processing on Azure, such as HDInsight for managed Hadoop clusters, Databricks for interactive data collaboration, and Synapse Analytics as integrated services (Isard et al., 2020).

4.2.1. HDInsight

HDInsight is a Microsoft Azure service that is easily used to run big data applications like Hadoop, Spark, and other such services. It offers an intelligent and efficient way of handling big data; users can easily build and set up clusters. HDInsight is correctly aligned with Azure storage solutions regarding capability and speed for data processing and analysis.

4.2.2. Databricks

Azure Data Bricks is an analytics tool that uses Apache Spark to process big data. It is a social tool for data engineers, data scientists and business analysts to collaborate on big data initiatives. Databricks enable batch processing, stream processing, and machine learning in data, and consequently, it is a useful tool for big data.

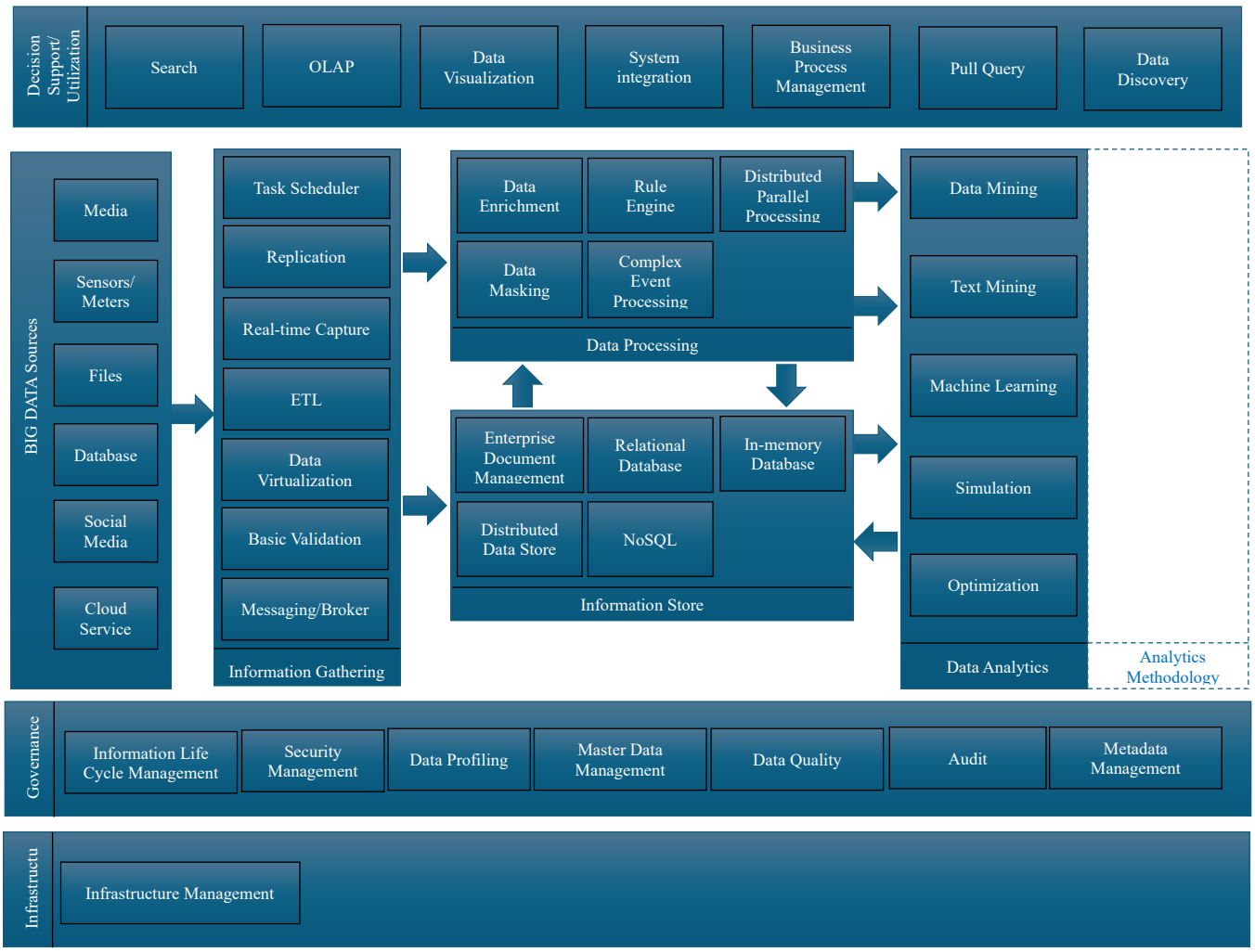

Fig. 2 GCP big data services architecture

The collaboration with Azure services, including Azure Data Lake Storage and Azure SQL Data Warehouse, improves its functions and makes it a multidimensional tool for data processing and analysis (Armbrust et al., 2019).

4.2.3. Synapse Analytics

Azure Synapse Analytics can be described as a unified, integrated service for big data and data warehousing. It allows for end-to-end handling of data from consumption to processing, storage, governance and serving for real-time BI and ML requirements. Synapse Analytics is easily extensible, allowing connection with several Azure services like Power BI and Azure Machine Learning that favor the development of complete analytics solutions. It again includes built-in security and compliance enforcement functionalities that safeguard and make the information compliant. In Figure 1 above, there is a representation of the Azure big data services with details of HDInsight being coupled with Databricks and Synapse Analytics. These services offer ISVs a strong foundation for big data processing, analytics, and machine learning to help organizations make strategic decisions based on the data. As illustrated in the section above, both GCP and Azure have rich and solid solutions for big data processing. Data warehousing can be easily managed through GCP's BigQuery, and stream processing and batch data processing are facilitated through Dataflow and Dataproc from GCP.

At the same time, Microsoft's Azure offers solutions in the form of HDInsight, Databricks, and Synapse Analytics. Through the use of these cloud technologies, large quantities of data could be better managed and analyzed for the purpose of spurring innovation as well as creating competitive advantage in today's big data environment (Vavilapalli et al., 2020; Isard et al., 2020; Armbrust et al., 2019).

5. Comparative Analysis: GCP to Azure

This position shows that performance, scalability, cost, and usability are crucial when comparing GCP and Microsoft Azure for Big Data processing. Both have excellent data handling and analysis services and may be used interchangeably depending on distinct business requirements and situations.

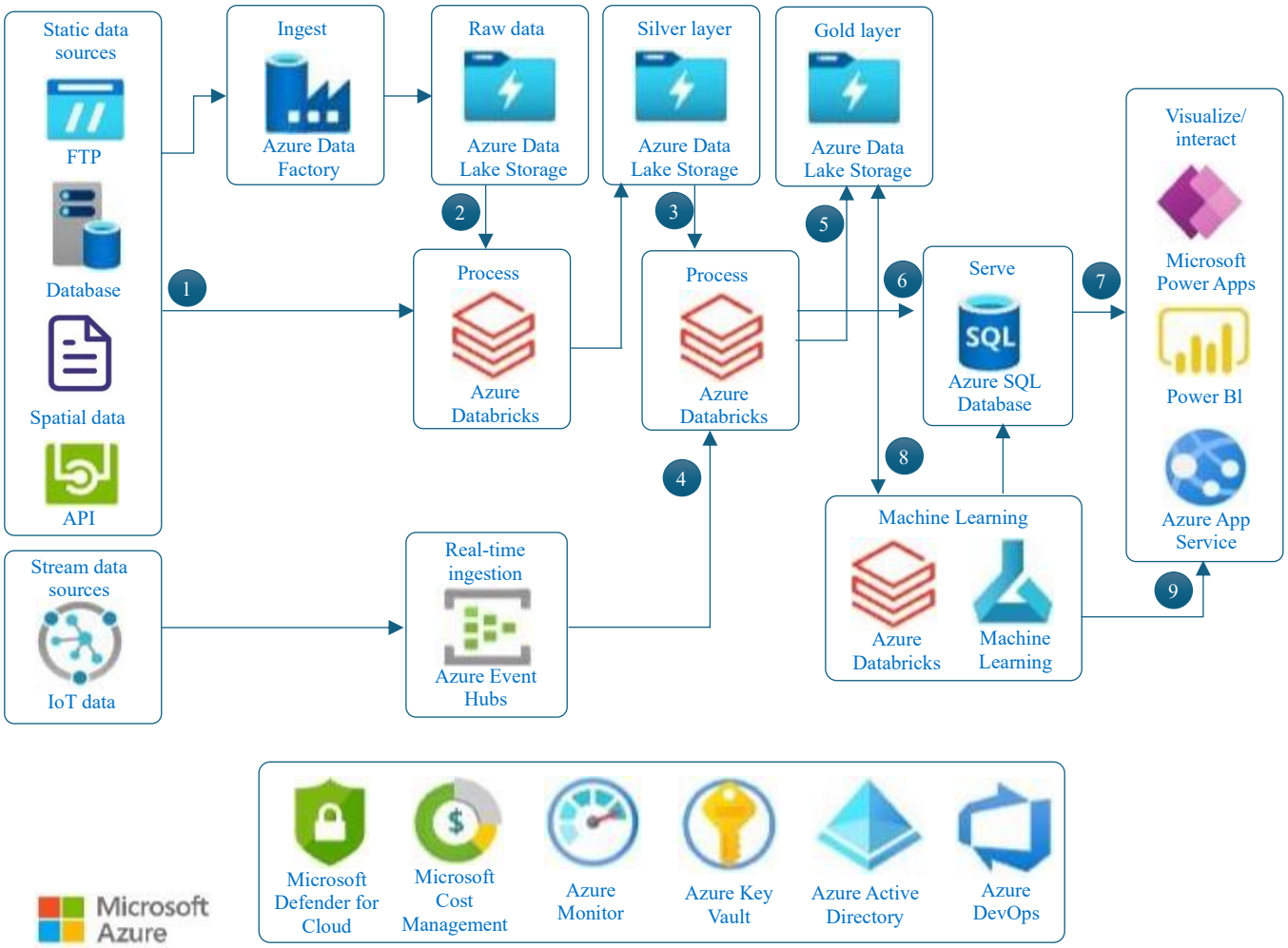

Fig. 3 Azure big data services architecture

5.1. Performance

It is specifically important to note that BigQuery, which is being used by BGCP, is reported to be extremely fast. Being a serverless data warehouse, BigQuery enables users to perform a variety of computations with the help of complex SQL queries on large datasets with latency. The system is built in such a way that storage and computing are physically separated from each other, which means that queries are answered quickly, making the system preferred for real-time analytics and interactive querying (Li et al., 2021). On the other hand, Azure's Synapse Analytics also offers comparable performance, primarily when fused with Microsoft's other services. Synapse Analytics represents the integration of big data and the data processing tools of a data warehouse. That deep integration with Microsoft tools such as Power BI and Azure Machine Learning improves its capability of performing well in end-to-end data analytics workflows (Zaharia et al., 2020).

5.2. Scalability

Regarding scalability, both the GCP and Azure do well. This enables a simple scale-up or scale-down of resources depending on the load of work that an application might have to deal with. For example, BigQuery can support petabytes of data without requiring any tuning from the business side, thus enabling business operations to scale up or down their data processing needs easily (Vavilapalli et al., 2020). Azure Synapse Analytics also has great scalability, which is in common with other products of the sapient company. It has the added-on demand elephant scaling both in computing and storage, meaning you can scale up or down inputs and outputs depending on the user's need. This flexibility makes sure that Azure is capable and ready to handle big data and data analytics projects without being a burden or slowing down large projects (Li et al., 2021).

5.3. Cost

This means that cost is one of the objects that can vary and lead to choosing either GCP or Azure. Since both platforms work on the use of resources, users are charged the number of resources they have used. This model can greatly impact cost savings compared to the on-premises IT infrastructure since it relieves a customer from large initial investments in hardware. Regarding this, pricing by GCP on

BigQuery is mostly done according to the amount of data stored in the services and the number of queries made. It can be particularly economical for organizations where the amount of query work varies, and the used money must be adaptable. Pricing-wise, Azure Synapse Analytics has different tiers with different available options for businesses to choose from depending on how much they are willing to spend and how much they will be using it. The two interfaces provide a pricing estimating tool and tools to control, track, and minimize the cost (Hashem et al., 2020).

5.4. Ease of Use

Another aspect of comparison between GCP and Azure based on comparative analysis is ease of use. GCP provides simple and smooth operability, and the complex elements are self-explanatory, making it easy to deploy and use big data applications. One of the ways through which data analysts interact with BigQuery is through an SQL-like query language, which not only makes it easy for the analysts to make queries since they do not need to learn a new language if they already know SQL (Bifet et al., 2019). Another widely used and relatively easy-to-understand tool is Azure Synapse Analytics, although it is closely connected to Microsoft. It makes integrating different tools and services, such as Power BI and Azure Machine Learning, in data analytics. This is because the use of Synapse Analytics gives the user a single interface that is convenient to work with when it comes to data and analytics operations, hence increasing efficiency and collaboration among users (Armbrust et al., 2019).

5.5. Case Studies and Real-world Applications

The pros and cons of the chosen platforms are described through the analysis of the multiple cases. For instance, Spotify uses GCP's BigQuery to generate real-time analytics by utilizing the service's speed and scalability for behavioural analysis of the user to enhance the application's recommendation system (Zaharia et al., 2020). On the other hand, e-commerce companies use Azure Synapse Analytics to connect customer data and improve marketing techniques, hence improving the consumers' experience through its highly effective analytical functionalities, as Li et al. (2021) postulated. Both GCP and Azure provide the kind of solution that could be termed as 'Big Data friendly,' meaning it is powerful and cost-effective in handling the said Big Data sets. BigQuery is the best tool of GCP for real-time analytics, and the interactive query feature makes it preferable. Azure Synapse Analytics is developed to work best with other Azure services and other Microsoft technologies to complement its end-to-end data analytics solutions. Which of these platforms will be the choice will, therefore, depend on the business requirements, existing architecture, and stack, as well as the architecture's integration and user interface preferences.

6. Future Trends in Big Data Engineering

The subject area of Big Data Engineering is equally dynamic since new technologies and new methods of data processing frameworks allied to cloud services are continuously on the agenda. It could be said that such changes are expected to revolutionize the ways how organizations gather, store, transform and analyze information to develop and adopt more evolved approaches to managing data. The main trends to be expected in Big Data Engineering are the enhanced connection between Big Data and machine learning/ Artificial Intelligence, the emergence of edge computing, further developments in the data processing framework, and the building up of cloud services.

Table 2. GCP vs. Azure comparison

Feature	GCP (BigQuery)	Azure (Synapse Analytics)
Performance	High	High
Scalability	Excellent	Excellent
Cost	Variable	Variable
Ease of Use	User friendly	Integrated

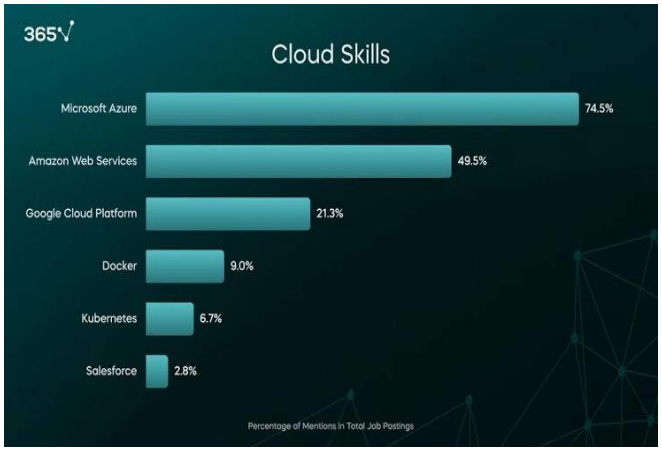

Fig. 4 Cloud skills

7. Integration of Machine Learning and AI

ML and AI are now at the core of Big Data Engineering. These technologies help organizations derive more value from data by automating many analytical processes and discovering patterns that would otherwise be impossible to identify manually. Implementing machine learning approaches enables the processing of real-time and large databases. It performs predictive analysis, detection of deviation from the norm, recommendation services, and factors that create business value (García et al., 2022). AI and ML are now being incorporated into the existing big data platforms to improve the system's function. For example, when it comes to applications, tools and frameworks such as Google Cloud AI and Azure Machine Learning are equipped to make the process of applying machine learning at a large scale relatively easier. Integration of these services is relatively easier due to their pre-built models and/or APIs meant for integrating them with other data pipelines, which helps fast-track the development of AI-based applications (Dhar, 2023). Big data engineering will increasingly invest in ML and AI as methods to improve intelligent data processing and analytics.

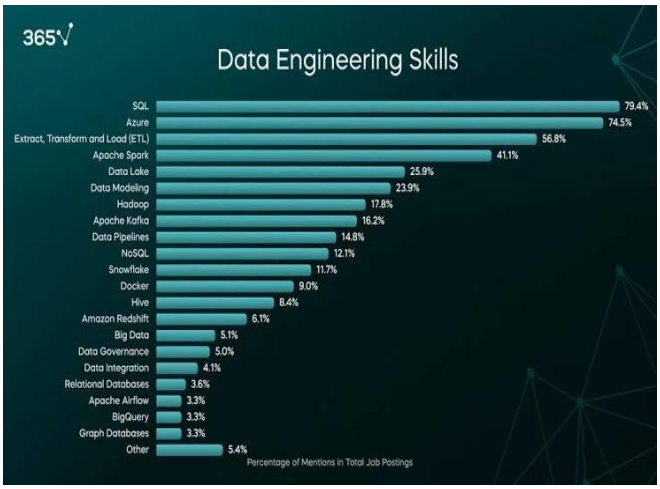

Fig. 5 Data engineering skills

8. Rise of Edge Computing

Big Data Engineering is another trend expected to cause a shift in the near future momentarily in edge computing. Edge computing can be defined as processing near the data source at the network periphery and not in a data hub. This approach effectively minimizes delay and amount of bandwidth consumed, which means it is suitable for real-time processing such as streaming and real-time applications that require quick responses (Shi et al., 2022). The rising IoT devices and ever-growing data processed at the edges call for new data processing frameworks. Machine learning can be processed at the network's edge by lowering the requirements from central servers, and edge devices can preprocess and filter data before forwarding it to the cloud. This way, the load can be shifted from central data centres, and only necessary data is communicated, enhancing system efficiency (García et al., 2022).

9. Advancements in Data Processing Frameworks

Data processing frameworks are not static entities as there is a tendency for newer ones to come up to fit in the increasing needs of Big Data Engineering. As for the innovations, these are aimed at increasing the levels of performance, scalability and working functionality. For example, Apache Spark has become one of the leaders in the field of big data processing owing to the in-memory data processing and numerous capabilities for the work in various modes such as batch, stream, and machine learning modes (Zaharia et al., 2020). There are also thus new frameworks arising in a bid to deal with certain issues in big data processing. For example, Apache Flink is powerful in stream processing and can serve as an efficient tool for real-time data analysis and event-driven systems. Flink's stateful stream processing engine delivers good all throughput low latency requirements. It is well-suited for applications such as fraud detection, predictive maintenance and real-time recommendation systems (as cited in Dhar, 2023).

10. Evolution of Cloud Services

Cloud services are still improving and providing more sophisticated and specific instrumentation to Big Data Engineering. Today's market offers several prominent CSPs, such as Google Cloud Platform, Microsoft Azure, and Amazon Web Services, steadily enhancing their portfolios from concerned data processing services, storage, and analytics. These innovations allow firms to select the most effective solutions depending on the organization's requirements and the integration of appropriate tools in the data processing system. For example, GCP's BigQuery has released new features like materialized views and BI Engine, adding more power to query and analytics. Azure Data Warehouse is a comprehensive big data and data warehousing solution that conforms to customers' demands. AWS currently has various services, such as Redshift for data warehousing, Glue for ETL and Sage Maker for Machine learning (Li et al., 2021). The growth of the cloud service also contributed to the use of hybrid and multi-cloud models. Multiload is fast becoming a strategic norm as organizations look to get the best out of multiple providers regarding performance, costs, and stability. This way, it is possible to use the best qualities of each provider and guarantee business continuity even when providers' services are interrupted (Hashem et al., 2020).

11. Enhanced Data Security and Privacy

Since the amount of information that needs to be stored is constantly increasing, as well as the level of its sensitivity, protection has become an issue of utmost concern. In the future trends of Big Data Engineering, remedies are likely to improve more numbers of security concerns in order to minimize the cases of unauthorized access or theft of data. This involves achievement in areas like encryption, access control measures, and compliance with modern data protection requirements like the GDPR and CCPA (Dhar, 2023). Cloud providers are also following up with new advances in their security business to cater for these problems. Some services that help organizations manage and protect their data include GCP's Cloud Security Command Centre and Azure Security Centre. Further, data anonymization and differential privacy techniques are also emerging to preserve data's privacies continuously, allowing data analysis (Shi et al., 2022).

12. Increased Focus on Data Governance

Proper data management is essential for the quality, validity and reliability of the information used. The future trends of Big Data Engineering focus on compliance with data governance frameworks and standards. This encompasses prioritizing good data governance, adopting data ownership and governance roles, and using data cataloguing and lineage to ensure data lineage and usage are well monitored (García et al., 2022). Businesses will also have to incorporate data quality management tools to identify and correct any flaws in the data. Data cleansing can be integrated with machine learning, where the algorithms take charge of the data cleaning

processes, hence ensuring a reduction in the risk of wrong analysis or decision-making (as cited in Dhar, 2023).

13. Conclusion

This paper focuses on highlighting the significance of Apache Hadoop and Cloud (GCP/Azure) technologies in Big Data Engineering. It has parts like HDFS, MapReduce and YARN, which form the core of the Hadoop ecosystem and are essential for handling big data as they address scalability issues in storage and processing. Supporting them with Hive, Pig, HBase, and Spark tools complements Hadoop for a wide range of big data uses. Google Cloud Platform (GCP) and Microsoft Azure have been aspiring cloud solutions that can work in conjunction with Hadoop's features and advantages, such as scalability and flexibility.

BigQuery and Dataflow from GCP, Dataproc, and Azure's HDInsight, Databricks, and Synapse Analytics are the best, most versatile, and most interactive tools for handling big data. These platforms provide productivity, reliability, flexibility, and cost optimization to meet modern enterprise demands. Comparing the two offers, it is possible to focus on the specifics of each platform to show they can serve different purposes in big data. Some of GCPs' biggest strengths are the speed and the flexibility they provide, while Azure is best for its integration with Microsoft services, providing detailed analytic capabilities. Indeed, as the advancement of technology grows increasingly complex, it is apparent that Hadoop and cloud technologies will retain their value in effectively capturing and analyzing enormous amounts of data to foster new ideas and shape intelligent decision-making.

References

[1] Michael Armbrust et al., "Scaling Spark in the Real World: Performance and Usability," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1840-1843, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nicolas Kourtellis, Gianmarco De Francisci Morales, and Albert Bifet, *Large-Scale Learning from Data Streams with Apache SAMOA*, Learning from Data Streams in Evolving Environments, Springer, Cham, vol. 41, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Rajkumar Buyya, James Broberg, and Andrzej M. Goscinski, *Cloud Computing: Principles and Paradigms*, Wiley, pp. 1-664, 2010. [Google Scholar] [Publisher Link]

[4] Jeffrey Dean, and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, no.1, pp. 107-113, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[5] Vasant Dhar, "The Future of Artificial Intelligence," *Big Data*, vol. 4, no. 1, pp. 1-67, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mohammed Guller, *Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis*, Apress, pp. 1-277, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ibrahim Abaker Targio Hashem et al., "The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues," *Information Systems*, vol. 47, pp. 98-115, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[8] Michael Isard et al., "Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks," *EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems*, pp. 59-72, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[9] Jairo R. Montoya-Torres et al., "Big Data Analytics for Intelligent Transportation Systems," *IFAC-PapersOnline*, vol. 54, no. 2, pp. 216-220, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Blend Berisha, Endrit Mëziu, and Isak Shabani, "Big Data Analytics in Cloud Computing: An Overview," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 11, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Vinod Kumar Vavilapalli et al., "Apache Hadoop YARN: Yet Another Resource Negotiator," *SOCC '13: Proceedings of the 4th Annual Symposium on Cloud Computing*, pp. 1-16, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tom White, *Hadoop: The Definitive Guide*, 3rd ed., O'Reilly Media, pp. 1-688, 2012. [Google Scholar] [Publisher Link]

[13] Matei Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56-65, 2016. [CrossRef] [Google Scholar] [Publisher Link]