Review Of Big Data Process Lifecycle And Technologies In Digital Ecosystem

Ragimova N., Mustafayeva S., Abdullayev V., Bagirli M., Abuzarova V., Khalilov M.

Abstract

Organizations now face new issues in handling and utilizing the enormous amounts of data being generated as a result of the advent of big data. A organized strategy, known as the Big Data Process Lifecycle, is required to overcome these problems. Data Ingestion, Data Storage, Data Processing, Data Analysis, and Data Visualization are among the stages involved in this process. Organizations must comprehend the Big Data Process Lifecycle and related technologies in order to manage Big Data effectively and acquire insightful information for making decisions.

Keywords: big data, Big Data Process Lifecycle, Data Ingestion, Data Storage, Data Processing, Data Analysis, Data Visualization, Apache Kafka, Hadoop Distributed File System (HDFS), Apache Spark, Tableau.

Big Data is often used to describe data sets that are so enormous and intricate that they are difficult to handle or analyze using conventional data processing techniques. These data sets frequently originate from a number of different sources, including social media, sensors, and other sources that produce a lot of data. In order to process, manage, and analyze huge data sets quickly and effectively, organizations require new tools and methodologies. Because of this, new Big Data-focused technologies and methodologies, such Hadoop, Spark, and NoSQL databases, have been created.

Global data production has increased dramatically in recent years, and this growth is anticipated to continue. The ubiquitous usage of social media, the Internet of Things (IoT), and the growing digitization of business operations are just a few of the causes that are fueling this expansion. Organizations increasingly have to manage enormous amounts of data that are too complex and massive to be handled by means of conventional data processing techniques. Organizations must use new tools and technology created expressly to handle Big Data in order to manage and analyze this data effectively. The study and interpretation of Big Data using machine learning, artificial intelligence, and other methods has given rise to new fields like data science. Effective Big Data management necessitates a thorough knowledge of the Big Data Process Lifecycle and the technologies utilized at each level. The following steps are often included in the big data process lifecycle:

1. Data Ingestion: Raw data is gathered from multiple sources and stored in a data repository during the data ingestion stage. The data is often unstructured, therefore processing is necessary before analysis.
2. Data Storage: The data is kept in a data warehouse or a distributed file system, like Hadoop Distributed File System (HDFS), at this point. To save space, the data is typically saved in a compressed and optimized format.
3. Data processing: The data is analysed at this step to draw conclusions and discover patterns. Complex processing operations like data transformation, filtering, and aggregation are involved in this level. At this point, technologies like Apache Spark, Hadoop MapReduce, and Apache Flink are used.
4. Data analysis: The processed data is examined in this step in order to draw conclusions and make decisions. Various data analysis methods, including statistical analysis, machine learning, and data visualization, are used during this step. During this phase, technologies like Apache Hive, Apache Impala, and Apache Drill are used.
5. Data visualization: The insights discovered through data analysis are displayed at this step using visual tools like charts, graphs, and dashboards. This phase aids stakeholders in comprehending and interpreting data analysis findings. During this phase, technologies like Tableau, Power BI, and QlikView are used.

Organizations may construct efficient Big Data systems that can handle massive volumes of data, extract insights, and improve decisions by knowing the Big Data Process Lifecycle and the technologies involved

in each stage.[1, 21]
Big Data describes extraordinarily vast, intricate, and varied data sets that are difficult to process with conventional data management and processing techniques. Big Data is distinguished by its quantity, speed, and variety. The sheer amount of data, which can range from terabytes to petabytes or more, is referred to as the volume of big data. Big Data velocity is the rate at which data is created, processed, and analyzed; this rate might be in real-time or very close to real-time. Big Data is diverse because it generates a wide range of data kinds, including structured, semi-structured, and unstructured data.
Here are some illustrations of big data and the difficulties in managing it:
• Social Media Data. Large volumes of data are produced by social media sites in the form of posts, comments, likes, and shares. The difficulty lies in drawing insights from this frequently unstructured data in order to comprehend client mood, behavior, and preferences. Social media data can often be noisy and untrustworthy, making analysis challenging.
•IoT Data.Large amounts of data are produced by linked devices including sensors, wearables, and smart home appliances as part of the Internet of Things (IoT). Since this data is frequently real-time, it needs to be processed quickly in order to find trends, abnormalities, and opportunities. Managing the sheer amount and variety of data while maintaining data security and quality is difficult with IoT data.
• Healthcare Data. Data generated by the healthcare industry comes from electronic health records, medical equipment, and clinical studies. Due to the high level of sensitivity, data privacy laws must be strictly followed. Additionally, it can be challenging to integrate and evaluate healthcare data because it is frequently complicated, unstructured, and fragmented.
• Financial Data
Trading systems, risk management systems, and client interactions all provide a lot of data for the financial sector. High standards of security and compliance are necessary for this data due to its strict regulation. Integrating and analyzing data from various sources to get a comprehensive picture of risks, opportunities, and customer behavior is difficult when dealing with financial data.
[3, 4]Managing Big Data presents a number of significant difficulties, including:
1. Data Integration: Big Data is challenging to integrate and evaluate since it frequently arrives from several sources and in different formats.
2. Data Quality: Big Data needs to be cleaned and transformed because it is frequently insufficient, erroneous, or inconsistent.
3. Scalability: To manage massive volumes of data, big data requires highly scalable storage and processing systems.
4. Security: Big Data is a target for cyberattacks and data breaches since it is highly valuable and sensitive.
5. Talent Shortage: Due to the increased demand for people with knowledge of big data, there is a talent gap that makes it challenging for businesses to efficiently manage and analyze big data.[21, 23]
A framework called the Big Data Process Lifecycle assists organizations in controlling the entire process of working with big data. It consists of a number of stages, each involving certain tasks and tools. The steps of the Big Data Process Lifecycle are as follows:[5, 22, 24]
• Data ingestion. The process of importing, gathering, and putting substantial amounts of data from diverse sources into a centralized repository or data lake for analysis is known as data ingestion. Because it lays the groundwork for all upcoming stages, this stage is essential to the Big Data Process Lifecycle.
Data ingestion is crucial for a number of reasons:
1. Data sources: Data for organizations is kept in a variety of places, including databases, logs, social media, and IoT devices. Organizations can gather data from many sources and combine it in a single spot for analysis thanks to data intake.
2. Data Volume: Manual data management is challenging given the exponential expansion of data. Data ingestion enables businesses to swiftly acquire and process massive volumes of data through automation.

3. Variety in Data: The data gathered may be in unstructured, semi-structured, or structured formats. Organizations can benefit from data ingestion by transforming the data into a standardized format that is simple to evaluate.
4. Real-time data: Some applications call for the continuous collection and processing of data as it is produced in real-time. This is significant for situations where rapid decision-making is essential, such fraud detection.
5. Data Security: To ensure that only authorized individuals can access the data, data ingestion can incorporate security mechanisms like authentication, encryption, and authorization.
• Data Storage. A crucial phase of the Big Data Process Lifecycle is data storage. Data must be stored in a central area so that it may be quickly accessible and evaluated after it has been ingested. The following succinct statement sums up the significance of data storage in the Big Data Process Lifecycle:
1. Scalability: As data volumes increase, conventional storage methods can no longer keep up. There are scalable options for storing massive volumes of data, including Hadoop Distributed File System (HDFS), Amazon S3, and Azure Blob Storage.
2. Cost-effectiveness: Implementing and maintaining traditional storage systems, such as Relational Database Management Systems (RDBMS), can be expensive. The pay-as-you-go concept offered by data storage technologies like HDFS and S3 allows businesses to only pay for the storage they really utilize.
3. Flexibility: Data storage technologies allow for a wide variety of data types to be stored. They make it simpler for businesses to store all of their data in a single spot because they can store structured, semi-structured, and unstructured data.
4. Accessibility: Data analysis is made simple by data storage technologies. The ability to access data using a variety of technologies, including SQL, NoSQL, and Hadoop MapReduce, makes it simpler for data scientists and analysts to draw conclusions from the data.
5. Data Security: To ensure that data is stored securely, data storage systems offer strong security features including encryption, access control, and audit trails.
•Data Analysis
Because it is during the data analysis stage of the big data process lifecycle that big data insights and value are obtained. In order to get insights that can be utilized to spur corporate growth and guide decision-making, it entails the application of statistical and analytical approaches to extract useful information from the data, find patterns and correlations, and gain trends.
The Big Data Process Lifecycle requires data analysis for the following main reasons:
1. Business Insights: Data analysis enables businesses to understand consumer behavior, industry trends, and operational efficiency. Organizations may make educated decisions and gain a competitive advantage by analyzing massive amounts of data to find patterns and correlations that would otherwise be impossible to spot.
2. Efficiency Gained: Data analysis can help firms gain operational efficiency by pointing out areas where processes can be optimized, waste can be reduced, and productivity can be increased. Organizations can develop a comprehensive understanding of their operations through the analysis of data from numerous sources, which enables them to make wise decisions to increase efficiency.
3. Predictive Analytics: Data analysis may help firms employ predictive analytics to foresee future market demand, consumer behavior, and trend. Organizations can create predictive models by evaluating past data, which will enable them to foresee future patterns and take well-informed decisions to seize opportunities.
4. Personalization: Data analysis may help businesses provide customers with tailored experiences. Organizations can customise products and services to cater to specific client needs by analyzing customer data to acquire insights into customer preferences, behavior, and needs.
• Data Visualization
The graphic depiction of data and information is known as data visualization. It entails developing visuals,

such as maps, charts, and graphs, to communicate complex data in a way that is simple to comprehend. Data visualization's objective is to assist users in locating patterns, connections, and trends in massive amounts of data that may not be immediately obvious in raw data formats.

Data visualization aids firms in making data-driven decisions, making it a crucial part of the Big Data process lifecycle. Data visualization enables decision-makers to swiftly find insights and make informed decisions by presenting complex data in an understandable style. Data visualization can be used by a business, for instance, to study consumer behavior and pinpoint areas where its goods or services need to be improved.

Data visualization also makes complex data analysis results more understandable to stakeholders who might lack the technical expertise needed to comprehend the underlying data. Organizations can communicate insights to non-technical stakeholders and entice them to engage in the decision-making process by visualizing data.

The amount, pace, variety, and authenticity of data are only a few of the difficulties associated with managing big data. Organizations must have the appropriate infrastructure and tools to collect, store, and handle the massive amount of data that is generated every day. The velocity of data is the rate at which data is produced, and for companies that are not equipped to handle it, this rate can be overwhelming. The complexity of managing Big Data is also increased by the range of data sources and formats. The correctness and dependability of the data, also known as veracity, is essential for enterprises to make wise judgments. For enterprises to properly manage and use Big Data, a thorough understanding of the Big Data Process Lifecycle and related technologies is essential. Organizations can choose the appropriate tools and technology for each stage of the lifecycle to manage Big Data effectively by using an organized approach. Data ingestion tools like Apache Kafka, Flume, and AWS Kinesis can aid enterprises in efficiently gathering and storing data. Organizations may store and manage huge volumes of data using tools like Hadoop Distributed File System (HDFS), Apache Cassandra, and Amazon S3. Organizations can use data processing tools like Hadoop MapReduce, Apache Spark, and Apache Flink to clean, transform, and analyze their data. Organizations can use data analysis tools like Apache Hive, Apache Impala, and Apache Drill to find patterns and insights in the data. Finally, firms can use data visualization tools like Tableau, Power BI, and QlikView to build visual representations of their data for decision-making.

Conclusion

In summary, handling big data is a difficult and complex undertaking. By using an organized method to manage Big Data effectively, organizations can overcome these difficulties by having a solid understanding of the Big Data Process Lifecycle and related technologies. Organizations can acquire important insights and take well-informed decisions to achieve a competitive advantage by utilizing the appropriate tools and technology for each step of the lifecycle.

References

[1] Viktor Mayer-Schönberger and Kenneth Cukier, "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Houghton Mifflin Harcourt, 2013.

[2] Paul Zikopoulos, Chris Eaton, David Corrigan, Tom Deutsch, and Krzysztof Czarnecki, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill Education, 2011.

[3] Bill Schmarzo, "Big Data: Understanding How Data Powers Big Business", John Wiley & Sons, 2013.

[4] Foster Provost and Tom Fawcett, "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking", O'Reilly Media, Inc, 2013.

[5] David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann Publishers, 2013.

[6] Tom White, "Hadoop: The Definitive Guide", O'Reilly Media, Inc, 2020.

[7] Bill Chambers and Matei Zaharia, "Spark: The Definitive Guide", O'Reilly Media, Inc, 2020.

[8] Ellen Friedman and Kostas Tzoumas, "Apache Flink: A Practical Guide to Building Scalable, Real-Time Data Processing Applications", O'Reilly Media, Inc, 2019.
[9] Neha Narkhede, Gwen Shapira, Todd Palino, "Kafka: The Definitive Guide", O'Reilly Media, Inc, 2021.
[10] Hari Shreedharan, "Apache Flume: Distributed Log Collection for Hadoop", O'Reilly Media, Inc, 2013.
[11] Madhurima Mukhopadhyay, "Learning AWS Streaming Data Solutions", Packt Publishing, 2018.
[12] Jeff Carpenter and Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilly Media, Inc, 2020.
[13] Naoya Hashimoto, "Amazon S3 Cookbook", Packt Publishing, 2015.
[14] Thilina Gunarathne, Srinath Perera, Inturu Sriram, "Hadoop MapReduce v2 Cookbook Second Edition", Packt Publishing, 2015.
[15] Edward Capriolo, Dean Wampler, Jason Rutherglen, "Programming Hive", O'Reilly Media, Inc, 2012.
[16] Ravi Magham and Venkata Giri, "Apache Impala (incubating) - Analyzing Big Data in Real Time", Packt Publishing, 2017.
[17] Charles Givre and Paul Rogers, "Learning Apache Drill", Packt Publishing, 2015.
[18] Daniel G. Murray, "Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software", Wiley, 2019.
[19] Matthew Adams, "Power BI: The Ultimate Step-by-Step Guide for Beginners to Expert", Independently published, 2020.
[20] Barry Harmsen and Miguel García, "QlikView 11 for Developers", Packt Publishing, 2012.
[21] Wikipedia, Big data, https://en.wikipedia.org/wiki/Big_data
[22] Tutorialspoint, Big Data Analytics - Data Life Cycle, https://www.tutorialspoint.com/big_data_analytics/big_data_analytics_lifecycle.htm
[23] Oracle, What is Big Data?, https://www.oracle.com/big-data/what-is-big-data/
[24] Redacción Tokio, Big Data life cycle: understand all Big Data phases, https://www.tokioschool.com/en/news/big-data-life-cycle/
[25] David Taylor, Top 15 Big Data Tools and Software (Open Source) 2023, https://www.guru99.com/big-data-tools.html