International Journal of Scientific Engineering and Science
Volume 8, Issue 6, pp. 11-14, 2024. ISSN (Online): 2456-7361

Integration of Big Data and Data Engineering in Modern Organizations

Sultan Yerbulatov
Lead Data & Analytics Engineer, Chevron Eurasia Business Unit LLP Tengizchevroil, Atyrau, Republic of Kazakhstan
sultan.yerbulatov@gmail.com

Abstract**—** *In the modern world, organizations are faced with large amounts of data that require effective processing and analysis to make informed decisions. In this regard, the integration of big data and data engineering (Data Engineering) play an important role. Big data integration is a complex process of combining, storing and processing data from various sources, including structured and unstructured data. The strategic use of big data requires competent data management at all stages of its life cycle. Data Engineering, on the other hand, focuses on the development and maintenance of data architectures that ensure the efficient flow of data from its sources to end users. This process includes data collection, transformation and loading (ETL), as well as the creation of an infrastructure for data storage and management. In this context, the integration of big data and Data Engineering are becoming key components of successful data analysis and strategic decision-making in modern organizations. The purpose of the work is to consider the process of integrating big data and Data Engineering in modern organizations. The methodological foundations were scientific works, specialized literature and opinions of experts in this field.*

Keywords**—** *Data Engineering, big data, modern technologies, organizations, companies, firms, integration of modern capabilities, data analysis.*

I. INTRODUCTION

Data engineering is a process that helps manage volumes of data: process, analyze and extract valuable information. It is based on the following principles:

1. Capturing and processing information: Data scientists create and implement systems to collect information from a variety of sources such as databases, web resources, sensors and others. They then process the information into the required format and structure.
2. Data storage and management: Databases and repositories are developed and maintained where information is stored in an organized form. Data management processes are being optimized to ensure high-speed access and reliability.
3. Data processing and analysis: develop data processing processes, including data cleaning, transformation and aggregation. Creating data pipelines allows you to extract valuable insights from vast amounts of information.

If we talk about the role of data engineering in the development of technology, then it plays a key role in the evolution of modern technologies and the business world as a whole [1]:

1. Innovation and Foresight: Data scientists help companies identify new opportunities and create innovative products and services based on data analysis. They use machine learning and analytics techniques to predict trends, improve efficiency, and optimize business processes.
2. Make informed decisions: Data engineers provide access to up-to-date and accurate information, giving leaders the confidence to make informed strategic and operational decisions.
3. Improved Productivity and Efficiency: With the use of data engineering, companies can optimize their business processes and increase productivity. Data analytics and process automation help improve efficiency, reduce costs and improve the quality of products and services.
4. Development of personalized solutions: Data engineers help create personalized products and services, customizing them to the needs of each client. Analyzing data about customers, their preferences and behavior makes it possible to provide personalized recommendations and improve user experience [2].

1. General characteristics of big data and Data Engineering

Data engineering is a sequence of activities aimed at making data accessible and usable to Data Scientists, data analysts, business intelligence developers and other specialists in the organization. Specialists are required to develop and create systems for large-scale data collection and storage, as well as prepare them for subsequent analysis.

An organization often uses a variety of operations management software (e.g., ERP, CRM, production systems, etc.) containing databases of various information. In addition, data can be stored as individual files or received in real time from external sources (for example, IoT devices). Due to the variety of data storage formats, an organization faces difficulty in obtaining a clear picture of the business's health and ongoing analytics.

Data engineering solves this problem step by step.

The data engineering process consists of performing a sequence of tasks that transform a large volume of raw data into a practical product that meets the needs of analysts, data scientists, machine learning engineers and other specialists. Typically the entire process consists of the following steps.

Fig. 1. Simplified structure of the data engineering process

International Journal of Scientific Engineering and Science
Volume 8, Issue 6, pp. 11-14, 2024. ISSN (Online): 2456-7361

When data is consumed (Data ingestion), it moves from numerous sources—SQL and NoSQL databases, IoT devices, websites, streaming services, and so on—to the end system for the purpose of transformation for subsequent analysis. Data comes in a variety of forms and can be either structured or unstructured.

During the Data transformation stage, fragmented data is adapted to the needs of end users. This stage includes eliminating errors and duplicate data, standardizing them and bringing them into the required format.

Data serving ensures the transfer of processed data to end users - a business intelligence platform, a dashboard or a data science team.

Data flow orchestration ensures that all tasks are completed successfully by giving visibility into the data engineering process. This process coordinates and continually monitors data processes to identify and resolve data quality and accuracy issues.

A data pipeline is the technique used to automate the data engineering process's steps of ingestion, transformation, and transmission.

Big Data engineering. When considering data engineering, it is impossible to ignore the concept of Big Data, which stands out for its characteristics: volume, speed of receipt, variety and reliability. This is usually important in large-scale technology corporations like YouTube, Amazon or Instagram. Big Data engineering is the process of building vast data warehouses and distributed systems with outstanding scalability and reliability under fault conditions.

The architecture of working with Big Data differs significantly from traditional data processing, since we are dealing with huge flows of information that are rapidly changing and do not fit into conventional data warehouses. In this context, a data lake comes to the rescue [3].

II. KEY DIFFERENCES BETWEEN DATA ENGINEERING AND BIG DATA

TABLE 1. Key differences between data mining and big data

	Data Engineering	**Big Data**
Main activities	They provide meaningful information that helps organizations make informed decisions.	They drive organizations to innovate and ideate and create new opportunities by analyzing complex data.
Key Tools	The main tools are ETL tools, SQL and traditional databases.	The most important tools include Hadoop, Spark, Kafka and NoSQL databases.
Analytics	They provide basic analytics and reporting.	They provide advanced analytics, machine learning and artificial intelligence.
Career path	Entry-level data engineering positions include data engineer, data analyst, and database administrator.	Entry-level big data jobs include big data engineer, data scientist, and machine learning engineer.
Data volume	Process small to moderate amounts of data for analysis.	Process specifically large and complex volumes of data for analysis.
Diversity of data	Deals with structured data.	Includes structured, semi-structured and unstructured data.
Impact on business	Allow organizations to make informed decisions after analyzing data.	They drive innovation, predict future trends and create new opportunities by analyzing big data.
Performance	High performance was required when working with small to moderate amounts of data.	High performance is required when working with large and complex amounts of data [4].

III. HOW DATA INTEGRATION WORKS

There is no one-size-fits-all approach to data integration. At the same time, several typical elements are involved in the integration process. They consist of a unified view of all data provided by a master server, clients, and a variety of other data sources.

A typical scenario involves sending a client request to the primary data server. The server then retrieves the required information from internal and external sources, combining them into a single view, which is then returned to the client.

An integral part of the data integration process is the ETL process, which includes the stages of extraction, transformation and loading. This process extracts data from sources, moves it to intermediate storage, where the data is cleaned and transformed before loading into the final source (usually a data warehouse or data warehouse).

A more modern approach to integrating data into a centralized repository is ELT. ELT modifies the sequence of steps, keeping the same steps as ETL - first, raw data is extracted from sources and loaded into the target source, where transformation occurs as needed. Typically, the target system for ELT is a data lake or cloud data warehouse.

Organizations have different data integrity requirements, resulting in a variety of data integration types. Key types include data consolidation, data virtualization, and data replication. These categories delineate the essential tenets of data integration; let us examine them more thoroughly.

Data consolidation, often thought of as a classic ETL process, is the process of combining data from different sources while removing redundancies and errors. The received data is transferred to a single storage facility, such as a data warehouse. Despite its complexity, this approach is often used in various scenarios. A common destination for data consolidation is a standard data warehouse, which is discussed next.

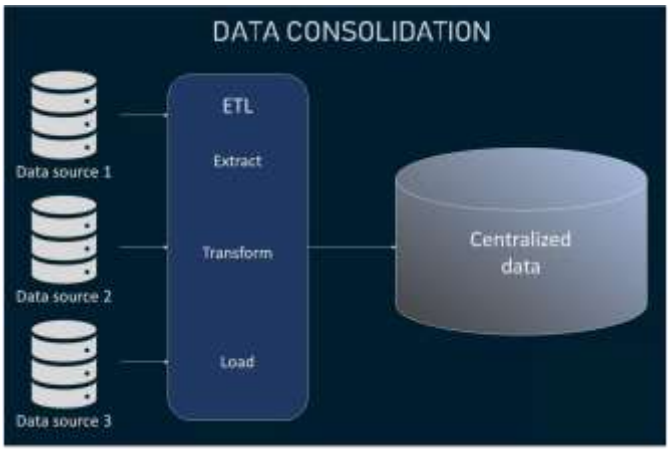

Fig. 2. How Data consolidation works

International Journal of Scientific Engineering and Science
Volume 8, Issue 6, pp. 11-14, 2024. ISSN (Online): 2456-7361

The fundamental concept behind data consolidation is to provide end users with comprehensive information in one place for more detailed analysis and reporting.

Ideal application. Organizations looking to reduce the number of systems on which data is stored and standardize their information resources.

Data virtualization. Data virtualization is a technique that combines data from multiple sources at a virtual level. In this case, there is no data movement or transformation using ETL; they remain physically in their original databases. Instead, integrated virtual (logical) views of the required data are created for querying and on-demand access. Data virtualization can be implemented using data federation, a process that creates a virtual database that does not contain the data but knows the paths to its actual location.

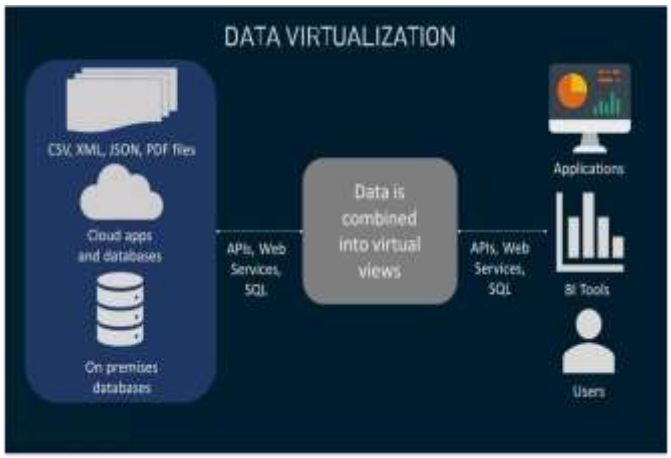

Fig. 3. How Data virtualization works

When a user makes a request, the virtualization layer accesses the source systems and instantly creates a unified view of the requested data. This process can be accomplished using various mechanisms, including APIs.

Information from inventory is transferred to the point of sale database. This method is often used when it is necessary to supplement data in one system with information from another. For example, a sales department interacts with a point-of-sale (POS) system, but needs certain data stored in an inventory database. This way, instead of using multiple applications to re-enter data, the sales team can interface with a core system that contains an exact copy of the data from the inventory system.

The data replication process can be carried out in three ways: • Full table replication - copying all new, updated and existing data from sources to the target storage; • Incremental key-based replication - copying only data that has changed since the last update; • Incremental log-based replication—copies data based on changes in the log files of the source systems [5].

TABLE 2. Advantages and disadvantages of data integration

Advantages	Disadvantages
Data integration provides a holistic view of data that helps companies make informed strategic decisions.	One of the common challenges in data integration is having data in silos. Different departments in an organization may use different systems and databases, creating a fragmented set of data. This makes it difficult to create a common view of the information in the organization. To solve this problem, it is necessary to implement integration solutions that can combine data into a comprehensive and interconnected set.
The integration brings together customer data from multiple sources, providing a comprehensive view to better understand and personalize the experience.	Inaccurate, contradictory or incomplete data pose a serious problem. Data from different sources may contain errors, discrepancies, or formatting issues that affect the quality of analysis and decision making. Data quality management, including cleaning and validation processes, is necessary to ensure the accuracy and reliability of integrated data.
Automating data integration processes reduces the cost of manual entry, processing and reduces the risk of human errors.	Data integration can become technically complex due to the variety of data sources with unique structures and formats. Working with legacy systems or custom applications may require complex technical solutions such as custom code and conversions. Investing in flexible and scalable integration tools can simplify this process.
Data integration allows you to identify trends and opportunities faster, providing a competitive advantage.	As businesses grow, new data integration needs arise. It is important to ensure that your data integration solution is scalable to meet changing data volumes and requirements. Cloud solutions provide the ability to easily scale data integration as your organization grows.
Centralized data control simplifies compliance with data protection regulations and ensures security, enhancing customer trust.	Implementing data integration solutions can require significant financial investment and maintenance costs. For small businesses or startups looking to leverage data integration, this can be a financial barrier. It is important to evaluate these costs against the long-term benefits and savings that can be achieved through data integration in order to make informed decisions about the integration strategy [6].

IV. CONCLUSION

Thus, it can be said that effective data integration requires a comprehensive approach covering goals, careful selection of tools, strong data protection, collaboration and systematic testing with monitoring. By following these best practices, organizations can realize the full potential of their data integration initiatives. This helps you make informed decisions, improve customer experiences, and achieve competitive advantage in today's data-rich business landscape.

REFERENCES

1. Data engineering: Basic principles and role in the development of modern technologies. [Electronic resource] Access mode: https://infozone.pro/data-engineering-basic-principles-and-role-in-the-development-of-modern-technologies/. – (accessed 25.01.2024).
2. The present and future of data engineering. [Electronic resource] Access mode: https://habr.com/ru/companies/vk/articles/661777 /.– (accessed 01/25/2024).
3. Data Engineering: concepts, processes and tools. [Electronic resource] Access mode:https://habr.com/ru/articles/743308 / .– (accessed 25.01.2024).
4. Data Engineering Vs Big Data. [Electronic resource] Access mode:https://digitaldefynd.com/IQ/data-engineering-vs-big-data-complete-guide /.– (accessed 01/25/2024).
5. Data Integration: Approaches, Techniques, Tools, and Best Practices for Implementation . [Electronic resource] Access mode: https://www.altexsoft.com/blog/data-integration /.– (accessed 25.01.2024).
6. Using data integration capabilities in the modern business landscape. [Electronic resource] Access

mode:https://www.coditation.com/blog/harnessing-the-power-of-data-integration-in-the-modern-business-landscape. – (accessed 25.01.2024).