UDC 004.6

ANALYSIS OF EFFICIENCY IN STORING AND PROCESSING UNSTRUCTURED DATA IN BIG DATA ENVIRONMENTS

Aliyev D.
bachelor's degree, Azerbaijan State University of Economics (Baku, Azerbaijan)

АНАЛИЗ ЭФФЕКТИВНОСТИ ХРАНЕНИЯ И ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ В СРЕДЕ BIG DATA

Алиев Д.В.
бакалавр, Азербайджанский государственный экономический университет (Баку, Азербайджан)

Abstract

The article examines key technologies for storing and processing unstructured data within Big Data environments, including Hadoop, NoSQL databases, Apache Spark, and Elasticsearch. Key advantages and limitations of each approach, along with their impact on infrastructure performance and scalability, are analyzed. An example is provided using Apache Kafka for data streaming and PySpark for preprocessing, highlighting the significance of these technologies in handling large volumes of information. Recommendations are given for selecting suitable technologies for different business scenarios. The research demonstrates that the integration of Big Data technologies into business processes enhances flexibility and reduces costs associated with data storage and processing.

Keywords: unstructured data, big data, containerization, data analysis, Apache Spark, Elasticsearch.

Аннотация

В статье рассмотрены основные технологии для хранения и обработки неструктурированных данных в среде больших данных (Big Data), включая Hadoop, NoSQL базы данных, Apache Spark и Elasticsearch. Проанализированы ключевые преимущества и ограничения каждого подхода, а также их влияние на производительность и масштабируемость инфраструктуры. Приведен пример использования Apache Kafka для потоковой передачи данных и PySpark для предобработки, что подчеркивает важность этих технологий в управлении большими объемами информации. Даны рекомендации по выбору подходящих технологий для различных бизнес-сценариев. Исследование показывает, что интеграция Big Data технологий в бизнес-процессы повышает гибкость и снижает затраты на хранение и обработку данных.

Ключевые слова: неструктурированные данные, большие данные, контейнеризация, анализ данных, Apache Spark, Elasticsearch.

Introduction

With the development of data processing technologies and the increase in data volume, the need for efficient solutions for data storage and analysis is growing. One of the key challenges modern organizations faces is the need to work with unstructured data, which constitutes a significant portion of information coming from various sources. Unstructured data, such as text documents, images, videos, and social media data, cannot be easily processed by traditional relational databases. Consequently, there is a need to develop technologies and methodologies capable of effectively

storing and analyzing such data. The goal of this article is to explore existing approaches to the storage and processing of unstructured data in Big Data environments.

One of the main challenges of working with unstructured data is its complexity and diversity. Unlike structured data, which is easily systematized, unstructured data requires more flexible storage and processing approaches. In recent years, technologies such as Hadoop and NoSQL databases have been actively used in Big Data environments, providing distributed data storage and enabling data processing under high loads. This article presents a comparative analysis of technologies used for the storage and processing of unstructured data and evaluates their effectiveness based on key parameters such as processing speed, scalability, and reliability.

In addition to technical aspects, special attention is given to optimizing processes for working with unstructured data and reducing processing costs. The introduction of technologies such as Apache Spark and Elasticsearch provides opportunities for fast data processing and analysis, which is particularly relevant in the context of modern digital business. This article examines the prospects and limitations of these technologies depending on the data characteristics and business needs and offers recommendations for selecting suitable solutions for different usage scenarios.

Main part

One of the primary methods for storing unstructured data is the use of distributed file systems, such as the Hadoop Distributed File System (HDFS). HDFS enables storing large volumes of data by dividing it into blocks and distributing them across multiple nodes, which enhances storage reliability and protects data from loss [1]. For processing unstructured data based on HDFS, Apache Spark is often used – a framework that enables parallel computations and speeds up the analysis process. Unlike traditional MapReduce, Spark offers higher performance by utilizing memory, making it suitable for tasks requiring fast data processing.

NoSQL databases, such as MongoDB and Cassandra, have also become popular tools for storing and processing unstructured data. Unlike relational databases, NoSQL databases can handle various data types, such as JSON documents and graphs, making them flexible for Big Data projects [2]. MongoDB, for instance, allows storing data in a document format, which simplifies working with text information and enables quick search and filtering operations. On the other hand, Cassandra ensures high availability and scalability, making it suitable for applications with heavy loads [3].

Query optimization and reducing data processing time are essential aspects of working with unstructured data. Elasticsearch, a search system built on Apache Lucene, is used for indexing and searching through large volumes of unstructured information. Elasticsearch allows creating indexes for text data and provides high search speed, which is particularly important in real-time data analysis. Table 1 presents the main characteristics of Hadoop, MongoDB, and Elasticsearch, including their applicability to different data types and performance.

Table 1

Comparison of technologies for unstructured data storage and processing

Technology	**Data type**	**Advantages**	**Limitations**	**Common application areas**
Hadoop	All types	Scalability, reliability	Requires significant resources	Big Data analysis, long-term storage
MongoDB	Documents (JSON)	Flexibility, easy integration	Limited ACID support	Document storage, data analytics
Elasticsearch	Text data	High search speed	Dependent on indexing	Search systems, text analytics
Cassandra	All types	High availability, scalability	Complex setup, resource-intensive	Analytics, high-load systems

Apache Spark	All types	High performance, memory support	Memory-dependent	Data analysis, fast processing of large volumes

To improve the efficiency of unstructured data processing, technologies such as Apache Kafka play a key role, providing real-time data transfer between different system components [4]. Kafka serves as a distributed streaming service that transmits data from sources to processors, allowing rapid processing and analysis of unstructured data. This approach is used for real-time analytics, including monitoring and event tracking, where minimal latency is essential.

An important aspect of unstructured data processing is ensuring data security and integrity. Most technologies, such as MongoDB and Elasticsearch, support built-in data protection mechanisms, including access control and encryption. However, for Big Data containing unstructured information, additional measures, such as access rights distribution and visibility restrictions, are often required, especially in corporate and government projects [5].

Data quality control is also essential, as unstructured data may contain errors and incorrect values. Data quality analysis tools, such as Apache Griffin, help identify anomalies and ensure data meets quality standards. This is especially important when integrating data from various sources, where each error can impact the final analysis results.

Applying Big Data technologies in unstructured data management

Several key technologies are actively applied in Big Data environments for working with unstructured data, each fulfilling a specific role in the data processing and analysis process. These technologies include distributed data storage systems such as Hadoop and NoSQL databases, data streaming tools like Apache Kafka, and search systems such as Elasticsearch.

The first stage of managing unstructured data is collecting and integrating it from various sources [6]. In this task, solutions like Apache Kafka, which processes data in real time and transmits it to storage and analytical systems, become indispensable. Kafka supports low-latency data transmission, which is crucial for applications where data analysis must be immediate. For streaming data processing during analysis stages, Kafka can be integrated with Apache Spark, providing high performance.

Once data is collected, it must be stored and prepared for further processing. Hadoop Distributed File System (HDFS) offers a reliable and scalable solution for storing large volumes of unstructured data. Data is distributed across multiple nodes, ensuring high fault tolerance and data security. HDFS is widely used for long-term data storage and allows processing data using tools such as MapReduce and Spark [7].

NoSQL databases, such as MongoDB and Cassandra, are widely used for effectively storing textual and semi-structured information. These databases allow data to be stored in flexible formats, such as JSON, which is especially useful for document and text information processing. MongoDB provides fast data search and filtering, while Cassandra is optimized for distributed systems and can handle low-latency queries, making it suitable for high-load applications.

When data is ready for analysis, tools like Apache Spark and Elasticsearch can be used for processing. Spark is a high-performance framework for Big Data processing that supports parallel computations in clusters, allowing data analysis at high speed. In turn, Elasticsearch allows creating indexes for text information, speeding up the search process and making it effective even with large data volumes [8]. Elasticsearch is widely used for text search and real-time data analysis, making it suitable for applications that require quick response.

For example, let’s add a Python code snippet that shows how to use the PySpark library (Python API for Apache Spark) and Elasticsearch for unstructured data processing and analysis. In this code, we assume we have text data, which will first be processed with PySpark and then indexed and stored in Elasticsearch for fast search and analysis.

```
# Importing PySpark and Elasticsearch libraries
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, lower, regexp_replace
from elasticsearch import Elasticsearch, helpers
```

```
# Initializing a Spark session
spark = SparkSession.builder \
    .appName("Big Data Processing Example") \
    .getOrCreate()

# Reading unstructured data from a text file
data_path = "path/to/your/textfile.txt"
df = spark.read.text(data_path)

# Data preprocessing: cleaning text and converting to lowercase
df_cleaned = df.withColumn("value", lower(col("value")))
df_cleaned = df_cleaned.withColumn("value", regexp_replace(col("value"), "[^a-zA-Z0-9\\s]", ""))

# Transforming data to a structure suitable for analysis
df_transformed = df_cleaned.withColumnRenamed("value", "processed_text")

# Converting DataFrame to list for loading into Elasticsearch
processed_data = df_transformed.rdd.map(lambda row: row["processed_text"]).collect()

# Initializing Elasticsearch connection
es = Elasticsearch("http://localhost:9200")

# Function to prepare data in Elasticsearch format
def es_data_generator(data, index_name="processed_data_index"):
    for line in data:
        yield {
            "_index": index_name,
            "_source": {
                "text": line
            }
        }

# Loading data into Elasticsearch
helpers.bulk(es, es_data_generator(processed_data))

print("Data successfully processed and indexed in Elasticsearch.")

# Stopping the Spark session
spark.stop()
```

This example demonstrates the process of unstructured data processing with PySpark and its subsequent loading into Elasticsearch. First, a Spark session is created for loading and cleaning text data. After preprocessing, which includes removing special characters and converting text to lowercase, the data is transformed into a structure suitable for analysis. Then it is converted to a format convenient for loading into Elasticsearch. In Elasticsearch, the data is indexed, allowing for fast search and real-time text information analysis.

Conclusion

Containerization and Big Data technologies represent a significant breakthrough in managing and processing unstructured data, allowing organizations to handle the increasing volumes of information and meet modern business demands. The use of Hadoop, Apache Spark, NoSQL databases, and Elasticsearch substantially enhances the flexibility and performance of infrastructure

for storing and analyzing data. These technologies have proven effective under high loads, providing capabilities for fast processing and real-time data analysis.

Furthermore, the use of data streaming tools such as Apache Kafka enables flexibility and timeliness in processing incoming real-time data. Although additional measures are required for ensuring data security and quality control, the application of containerization and Big Data technologies allows organizations to reduce operational costs and improve the reliability of information systems.

Future developments are expected to further advance these technologies and integrate them with new machine learning methods, establishing a foundation for enhanced capabilities in intelligent analysis and processing of unstructured information. This ongoing progress will open up new possibilities for more sophisticated data handling, ensuring that businesses can leverage unstructured data as a valuable resource in achieving their goals and improving service quality.

References

1. Gordienko E.P., Panenko N.S. Modern technologies for processing and analyzing Big Data in scientific research // Current Problems of Railway Transport. 2018. P. 44-48.
2. Mamedova G.A., Zeynalova L.A., Melikova R.T. Big Data technologies in e-education // Open Education. 2017. Vol. 21. No.6. P. 41-48.
3. Stepanova D.I. The use of the BIG DATA system to improve the efficiency of utilities // World Economy: Security Issues. 2019. No.2. P. 70-78.
4. Fedorova L.A., Hu G., Huang S., Zemlyakova S.A. The application of Big Data technologies in modern enterprises // Bulletin of the Altai Academy of Economics and Law. 2020. No.9-2. P. 322-329.
5. Kulikova O.M., Tropynina N.E. Challenges of using BIG DATA technology in modern market conditions // Innovative Economy: Development Prospects and Improvement. 2022. No.7 (65). P. 16-21.
6. Radchenko I.A., Nikolaev I.N. Big Data technologies and infrastructure // St. Petersburg: ITMO University. 2018. Vol. 52.
7. Alekseev K.A. The use of Big Data in international business // Proceedings of the Institute for System Programming RAS. 2020. Vol. 32. No.4. P. 7-20.
8. Gladchenko V.A. The use of global «Big Data» technologies as an effective tool for risk management in customs authorities // Economics. Law. Innovation. 2019. No.2. P. 36-41.