Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

Apache Hadoop Ecosystem

Raullien Roos Geingos
Department of Computer Science And Application
Sharda University
Greater Noida, Uttar Pradesh, India

Joyline Bildad
Department of Computer Science And Application Sharda University
Greater Noida, Uttar Pradesh, India

Ravi Prakash Chaturvedi
Department of Computer Science And Application
Sharda University
Greater Noida, Uttar Pradesh, India

Annu Mishra
Department of Computer Science And Application
Sharda University
Greater Noida, Uttar Pradesh, India

Rajneesh Kumar Singh
Department of Computer Science And Application
Sharda University
Greater Noida, Uttar Pradesh, India

Abstract: **The proliferation of user-generated data brought about by the quick development of social media sites like Instagram has created both opportunities and difficulties for data-driven insights. Such large-scale datasets are difficult for traditional data processing techniques to effectively handle in terms of volume, velocity, and variety. In order to handle and analyse Instagram engagement data at scale, this study uses the Apache Hadoop ecosystem to construct a distributed data analytics system.**

The study uses a five-node Hadoop cluster with HDFS (Hadoop Distributed File System) for scalable storage and PySpark (Apache Spark for Python) for effective data processing. The study looks at important engagement indicators like likes, comments, post popularity, and sentiment analysis while concentrating on data ingestion, preprocessing, feature extraction, and advanced analytics. Deeper insights into user interaction patterns and the best content strategies are made possible by the system's usage of distributed computing, which allows for both batch and real-time analysis of big datasets. The results show that Hadoop and PySpark provide a scalable and effective social media analytics solution, which makes them appropriate for uses like content performance optimisation, influencer marketing, and tailored recommendations. By demonstrating how open-source distributed computing frameworks may be used to extract valuable insights from enormous social media datasets, this study advances the area of big data analytics.

Keywords: **Big Data Analytics, Apache Hadoop Ecosystem, PySpark, HDFS (Hadoop Distributed File System), Distributed Computing, Social Media Analytics, Instagram Data Analysis, Sentiment Analysis, Real-time Data Processing, Batch Processing.**

I. INTRODUCTION

Social media sites like Instagram produce enormous volumes of user interaction data every day in the big data era. Likes, comments, shares, and following activity are examples of interactions that provide important information on audience behaviour, content popularity, and user engagement trends. However, there are many obstacles to overcome in terms of processing, storage, and real-time analytics when analysing such vast amounts of data. Because of their computing constraints and scalability issues, traditional data processing technologies find it difficult to effectively address these issues. In order to tackle these problems, this study uses the Apache Hadoop ecosystem to create a distributed data processing system that makes use of PySpark (Apache Spark for Python) for effective data processing and HDFS (Hadoop Distributed File System) for storage. In order to provide profound insights into post performance, user interaction patterns, and the best content tactics, the objective is to analyse Instagram engagement metrics across a five-node Hadoop cluster.

In order to ensure scalable and effective management of massive datasets, this study focusses on data import, preprocessing, feature extraction, and analytics using PySpark. With the goal of offering practical insights for social media strategy optimisation, the study investigates important performance metrics such likes per post, sentiment analysis of comments, and engagement trends.

This study shows how well the Hadoop ecosystem works for social media analytics by putting in place a fully distributed environment that provides a scalable solution for real-time insights. This study makes a significant contribution to both academic and industrial applications since its conclusions can be used in personalised recommendation systems, influencer marketing, and content strategy optimisation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

2. LITERATURE SURVEY

1. Big Data Analytics in Social Media

Traditional data processing techniques are insufficient due to the massive amount of user interaction data generated by social media sites such as Instagram. Big data analytics tools are becoming crucial for deriving valuable insights from unstructured social media data, allowing researchers and organisations to efficiently examine user behaviour and engagement trends (Gandomi and Haider, 2015)[1]. Scalable data processing in social media research has been made easier by the inclusion of distributed computing frameworks like Hadoop and Spark (Hashem et al., 2015)[2].

2. The Apache Hadoop Ecosystem for Large-Scale Data Processing

Because of its distributed architecture, scalability, and fault tolerance, Apache Hadoop has been widely used for big data processing and storage (White, 2012)[3]. MapReduce offers a programming style for parallel processing, whereas Hadoop's HDFS (Hadoop Distributed File System) allows for the dependable storing of big datasets over numerous nodes (Dean & Ghemawat, 2008)[4]. MapReduce is less appropriate for real-time data analysis due to its high latency.

3. Apache Spark for Efficient Data Processing

Apache Spark was presented as a quick, in-memory distributed computing platform to overcome MapReduce's performance issues (Zaharia et al., 2010)[5]. Spark is perfect for machine learning and large-scale data analytics because of its efficient iterative computation capabilities (Meng et al., 2016)[6]. Researchers have discovered that PySpark (Spark's Python API) improves data processing capabilities in the context of social media analytics, allowing for both batch and real-time monitoring of engagement metrics (Lamba & Garg, 2017)[7].

4. Instagram Data Analysis and User Engagement

In order to find patterns in user engagement, sentiment analysis, and influencer impact, recent research have looked at Instagram data. The impact of Instagram likes, comments, and hashtags on content exposure and engagement rates was investigated by Rahman et al. in 2021[8]. Similar methods were used by Tifrea et al. (2020)[9] to categorise Instagram material according to user interactions using machine learning techniques. These studies emphasise how crucial scalable data analytics tools are to extracting useful information from Instagram data.

5. Applications of Hadoop & Spark in Social Media Analytics

Data from social media may be processed efficiently at scale using Hadoop and Spark, according to research. Tiwari et al. (2019)[10] showed how Hadoop can be used for sentiment analysis on Twitter and came to the conclusion that distributed computing is far more efficient than standard relational databases. In a similar vein, Dhyani & Patel (2021)[11] used Spark-based data processing to examine Facebook user interactions, leading to better scalability and quicker query execution times. The efficacy of the Hadoop ecosystem for extensive social media analytics is confirmed by these results.

6. Research Gap and Contribution

Although big data analytics for social media has been studied in the past, most of these studies have been on Twitter and Facebook, with little research on Instagram data analysis utilising Hadoop and Spark. By using a five-node Hadoop cluster to analyse Instagram engagement metrics, this study seeks to close that gap. This study advances the fields of social media analytics, big data processing, and influencer marketing insights by utilising HDFS for scalable storage and PySpark for effective processing.

2.1. SYSTEM ARCHITECTURE AND COMPONENTS

In order to create a scalable Instagram data analytics system, this research project makes use of the Apache Hadoop ecosystem. Data sources, distributed storage, big data processing frameworks, and visualisation tools are some of the components that make up the architecture. Every element is essential to the effective processing, storing, and analysis of massive amounts of Instagram engagement data.

1. Data Source: Instagram Engagement Data

- Instagram engagement metrics, such as likes, comments, hashtags, timestamps, and user interactions, make up the main dataset. Web scraping, third-party data extraction tools, and the Instagram API are some of the methods used to gather this data. Usually saved in CSV or JSON format, the dataset needs to be preprocessed before analysis can begin.

2. Hadoop Distributed File System (HDFS)

- The main storage system for managing massive amounts of Instagram data is HDFS. It distributes data among several servers for scalability and fault tolerance, running on a five-node Hadoop cluster. By distributing data blocks among several nodes, HDFS reduces the possibility of data loss and guarantees data reliability.

3. YARN (Yet Another Resource Negotiator)

- As the Hadoop cluster's resource management, YARN distributes computational resources in a dynamic manner. It is essential for maximising system performance since it makes sure that data processing jobs are completed effectively and distributes the workload evenly among cluster nodes.

3. PySpark for Big Data Processing

- The main processing engine for Instagram data analysis is Apache Spark, more especially its Python API (PySpark). PySpark enables in-memory processing, which greatly increases computational efficiency in contrast to standard Hadoop MapReduce's high latency. PySpark is in charge of a number of data analysis and transformation activities, such as:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

· *Data Ingestion:* Loading raw Instagram data from HDFS.
· Data Cleaning: Handling missing values, duplicate records, and timestamp conversion.
· *Feature Extraction:* Calculating engagement metrics such as likes per post, comment frequency, peak activity hours, and sentiment scores.
· *Data Aggregation:* Summarizing user interactions over time to identify trends.

5. Spark SQL for Structured Query Processing

- Structured data querying using SQL-like commands is made possible by the integration of Spark SQL into the system. It makes working with structured datasets easy and enables sophisticated analytical queries on Instagram engagement numbers. Additionally, Spark SQL improves data interchange by supporting many formats, including CSV, JSON, and Parquet.

6. Resilient Distributed Datasets (RDDs) & DataFrames

PySpark processes data using RDDs (Resilient Distributed Datasets) and DataFrames:

- ✓ *RDDs*: Handle unstructured and semi-structured Instagram data, such as raw JSON files containing post details.
- ✓ *DataFrames:* Manage structured engagement data, enabling faster query execution and advanced analytics.

7. Feature Engineering and Sentiment Analysis

The system extracts key engagement metrics and performs sentiment analysis on Instagram comments. Feature extraction includes:

- ✓ *Total Engagement Rate: (Likes + Comments) per post.*
- ✓ *Peak Engagement Time: Identifying optimal posting hours.*
- ✓ *Hashtag Effectiveness: Analyzing hashtag influence on post visibility.*
- ✓ *Sentiment Analysis: Classifying comments as positive, neutral, or negative using Natural Language Processing (NLP) techniques.*

8. Processed Data Storage and Visualization

Once the data is processed, it is stored in multiple formats for further analysis and visualization:

- ✓ *HDFS:* Retains the processed data for scalability and reusability.
- ✓ *External Databases (e.g., PostgreSQL, MongoDB):* Used for structured query-based analysis.
- ✓ *Data Export (CSV/JSON):* Enables integration with visualization tools like Tableau, Matplotlib, or Power BI.

9. Data Visualization for Insights

The extracted insights are visualized using charts, graphs, and reports to provide a clear understanding of Instagram engagement trends. This step helps identify:

- ✓ Most engaging posts based on likes and comments.
- ✓ Optimal posting times for maximum audience interaction.
- ✓ Impact of hashtags on content visibility.
- ✓ Sentiment trends in user comments.

10. Cluster Management and Performance Optimization
To ensure smooth execution, the system incorporates:

- ✓ Resource Optimization: Dynamic allocation of CPU, memory, and disk usage across the cluster.
- ✓ Parallel Processing: Distributing computation tasks across worker nodes for higher efficiency.
- ✓ Fault Tolerance: Ensuring data availability by replicating blocks across HDFS nodes.

2.2. ALGORITHM
The following algorithm outlines the step-by-step process of data collection, preprocessing, analysis, and visualization for Instagram engagement data using the Apache Hadoop ecosystem.

Algorithm Starts:
Hadoop-Based Instagram Data Analytics
Input: Raw Instagram dataset (likes, comments, hashtags, timestamps, follower counts)
Output: Processed insights (engagement trends, sentiment analysis, best posting times)

1. Initialize the Hadoop Ecosystem
- ❖ Set up the HDFS cluster across five nodes.
- ❖ Configure YARN for resource allocation.

2. Data Collection & Storage
- ❖ Extract Instagram engagement data via Instagram API or dataset imports.
- ❖ Store the raw data in HDFS in JSON/CSV format.

3. Data Preprocessing in PySpark
- ❖ Load raw data from HDFS into PySpark.
- ❖ Clean the dataset:
  - ✓ Handle missing values (remove or impute).
  - ✓ Convert timestamps into structured formats.
  - ✓ Normalize text (remove special characters from hashtags).
- ❖ Store the cleaned data back into HDFS.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

4. Feature Extraction & Transformation

- ❖ Compute engagement metrics:
  - ✓ Total Engagement = Likes + Comments per post.
  - ✓ Engagement Rate = (Likes + Comments) / Followers.
  - ✓ Peak Activity Time = Most active hours for user interactions.
- ❖ Perform Sentiment Analysis on comments (optional).

5. Data Storage for Querying

- ❖ Store processed data in HDFS for further computation.
- ❖ Export structured data to external databases (PostgreSQL, MongoDB) for visualization.

6. Data Analysis & Visualization

- ❖ Use PySpark SQL for querying processed data.
- ❖ Generate visual reports using Matplotlib, Tableau, or Power BI.
- ❖ Identify key trends, best posting times, and audience sentiment.

7. Performance Optimization

- ❖ Optimize YARN resource allocation for better parallel processing.
- ❖ Ensure fault tolerance by replicating data across nodes.

8. Output Insights & Decision Making

- ❖ Generate final reports on:
- ✓ Most engaging Instagram posts.
- ✓ Best-performing hashtags.
- ✓ Optimal posting schedule.
- ✓ Sentiment trends from user comments.
- ✓ End of Algorithm

FLOWCHART:

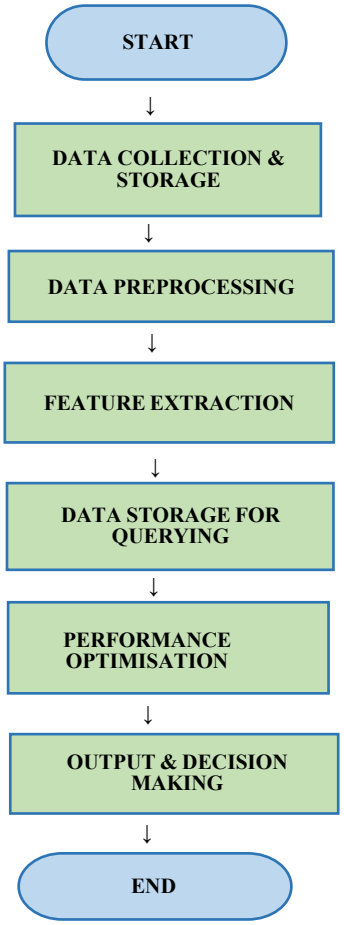


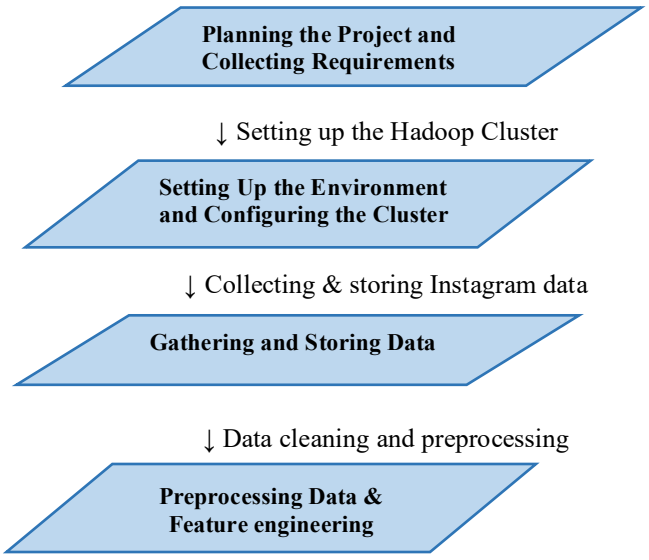
Fig. 1 Flowchart of Instagram Data Analytic using Hadoop & PySpark

Flowchart Explanation:

1. Start: Initialising the Hadoop ecosystem is the first step in the process.
2. Data Collection & Storage: Data about Instagram engagement is gathered and kept in HDFS.
3. Data Preprocessing: PySpark is used for data transformation, cleansing, and organising.
4. Feature Extraction: Sentiment analysis and key engagement metrics are carried out.
5. Data Storage for Querying: Both external databases and HDFS store processed data.
6. Data Analysis & Visualization: Visualisation tools and queries are used to generate insights.
7. Performance Optimization: Through the use of YARN and parallel processing, the system guarantees efficiency.
8. Output & Decision Making: Trends and final reports are produced.

4. HADOOP WORKING

4.1 Roadmap for Hadoop-based Instagram Data Analytics



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

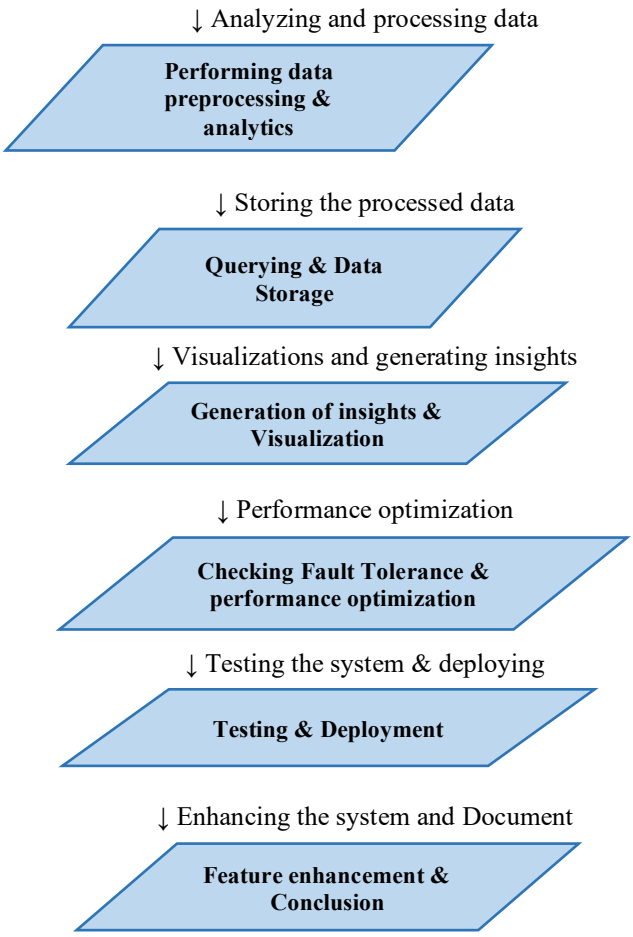


fig. 2 Roadmap

5. HADOOP ENVIRONMENT SETUP

In order to implement a distributed data processing system for Instagram data analysis, it is essential to first set up a Hadoop environment. Hadoop distributes data across multiple nodes in a cluster, allowing for scalable, fault-tolerant, and parallel data processing. This section offers a step-by-step tutorial for configuring a five-node Hadoop cluster, with a MacBook serving as the NameNode (Master) and four physical machines serving as DataNodes (Workers).

1. System Requirements:
Before I install the Hadoop, I made sure all the nodes in the cluster met the following hardware and software requirements:

1.1 Hardware requirements:

✧ **Master node (in this case the MacBook - NameNode)**
✓ CPU: it had a multi-core processor.
✓ RAM: 8GB or more.
✓ Storage: SSD with a least 100GB freeSpace

✧ **Worker node (4 physical machines - DataNodes)**
✓ CPU: must have multi-core processor.
✓ RAM: 4GB or more.
✓ Storage: at least 50GB per node

1.2 Software requirements:

✓ **MacBook (Master Node):** macOS with MacPorts installed
✓ **Worker Nodes:** Any Unix-based OS (in this case macOS)
✓ **Java Development Kit (JDK 8 or higher)**
✓ **SSH (Secure Shell) for password-less communication**
✓ **Hadoop 3.x (latest stable version)**

2. Installation and Configuration:

Step 1: Install Java (JDK or Higher)
✧ Installing Java on my MacBook using MacPorts since Hadoop needs it:

sudo port install openjdk11

Verifying the installation:
java -version

Step 2: Installing Hadoop on the Master node (MacBook)

✧ Using MacPorts, we installed Hadoop:

sudo port install hadoop

Verifying the installation:
hadoop version

Step 3: We configured SSH for password-less Authentication:

✧ Secure password-free SSH access between the nodes is necessary for Hadoop. So we set up SSH keys on the master node.
✧

```
ssh-keygen -t rsa
ssh-copy-id username@datanode1
ssh-copy-id username@datanode2
ssh-copy-id username@datanode3
ssh-copy-id username@datanode4
```

Verifying the connectivity:
ssh datanode1

Step 4: We modified the following Hadoop configuration files:

◆ **hadoop-env.sh (Setting Java home)**

```
export
JAVA_HOME=/Library/Java/JavaVirtualMachines/openjdk
11/Contents/Home
```

◆ **core-site.xml** (Properties of HDFS)

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

```
    </property>
</configuration>
```

◆ **hdfs-site.xml** (Settings for storage replication)

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>3</value>
    </property>
</configuration>
```

◆ **yarn-site.xml** (Settings for resource manager)

```
<configuration>
    <property>
        <name>yarn.resourcemanager.hostname</name>
        <value>master</value>
    </property>
</configuration>
```

◆ **slaves** (Worker nodes list)

```
datanode1
datanode2
datanode3
datanode4
```

Step 5: Formatting the Hadoop Distributed File System (HDFS):

✧ We need to format the name node before starting the hadoop.

```
hdfs namenode -format
```

Step 6: Starting Hadoop Services:

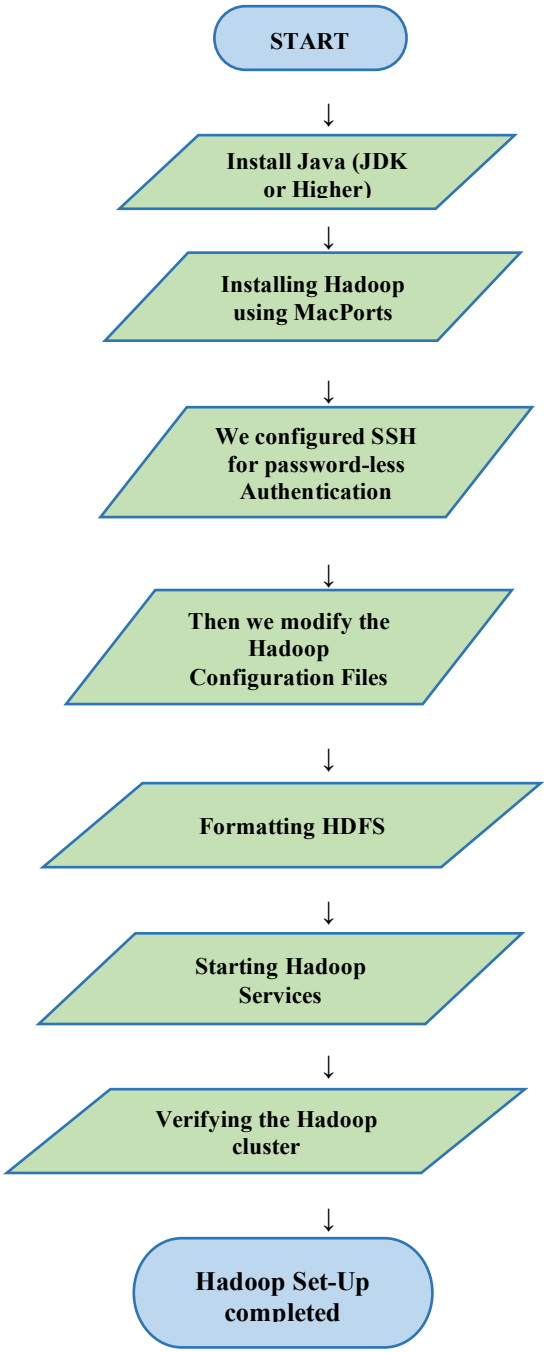
```
start-dfs.sh
start-yarn.sh
```

Verifying the running process:

```
jps
```

Step 7: Verifying the Hadoop cluster

FLOWCHART OF THE ENVIRONMENT SET-UP:



6. fig. 3 hadoop environment setup

6. CONCLUSION

A strong big data analytics pipeline for Instagram data is ensured by the combination of HDFS for scalable storage, PySpark for high-performance computing, and visualisation tools for insight production. This architecture adds to the larger field of big data and influencer marketing analytics by offering a scalable, effective, and economical solution for social media engagement analysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 03, March-2025

8. ACKNOWLEDGMENTS

We would like to extend our deepest gratitude to our guide Dr. Ravi Prakesh for his valuable guidance.

9. REFERENCES

[1] Haider, M., and Gandomi, A. (2015). **Beyond the Hype: Analytics, Methods, and Concepts of Big Data.** 35(2), 137-144, International Journal of Information Management.

[2] Hashem et al. (2015), I. A. T. Review and Open Research Issues Regarding the Emergence of **"Big Data" in Cloud Computing. Systems of Information**, 47, 98-115.

[3] T. White (2012). **Hadoop: The Complete Manual**. Media by O'Reilly.

[4] Ghemawat, S., and Dean, J. (2008). **Simplified Data Processing on Big Clusters using MapReduce.** ACM Communications, 51(1), 107-113.

[5] M. Zaharia and colleagues (2010). Spark: Using Working Sets for Cluster Computing. **Hot Topics in Cloud Computing: Proceedings of the USENIX Conference (HotCloud'10)**.

[6] Meng, X., and associates (2016). **MLlib: Apache Spark Machine Learning.** 17(1), 1235–1241, Journal of Machine Learning Research.

[7] Garg, K., and Lamba, A. (2017). **Apache Spark is used for real-time processing in big data analytics.** IEEE International Cloud Computing Conference, pp. 129–134.

[8] M. S. Rahman et al. (2021). **Data-Driven Analysis of Instagram User Engagement Trends.** Journal of Social Media Analytics, 9(1), 56-72.

[9] C. Tifrea and colleagues (2020). **Classifying Instagram Content with Machine Learning Methods**. Social Media Data Science Workshop Proceedings, 142-150.

[10] R. Tiwari and colleagues (2019). **Analysis of Twitter Sentiment with Hadoop and Spark.** 198–210 in Journal of Big Data Research, 6(3).

[11] Patel, P., and V. Dhyani (2021). **Big Data Analytics for Facebook User Data Processing with Apache Spark.** 7(2), 89-102, International Journal of Data Science.