Lab 10

Aim: Install, Deploy & configure Apache Spark Cluster

Theory:

Apache Spark is an open-source, distributed computing system for big data processing. It's fast, scalable, and supports various tasks like batch processing, real-time streaming, machine learning, and SQL queries. Spark processes data in memory for high performance and works with multiple programming languages (Python, Scala, Java, R). It's widely used for big data analytics and real-time data processing.

Key Features of Apache Spark:

1. **Speed**: Spark processes data in memory, which significantly increases performance for iterative algorithms and real-time processing compared to disk-based systems.
2. **Ease of Use**: It provides high-level APIs in multiple programming languages, including Python, Java, Scala, and R, as well as an interactive mode with tools like PySpark (Python) and Spark Shell (Scala).
3. **Versatility**: It supports a variety of applications, including:
    - **Batch processing** with Spark Core
    - **Stream processing** with Spark Streaming
    - **Machine learning** with MLlib
    - **Graph processing** with GraphX
    - **SQL processing** with Spark SQL
4. **Unified Platform**: Spark integrates seamlessly with other big data tools and frameworks like Hadoop, Kafka, and Hive, enabling a cohesive ecosystem.
5. **Scalability**: It can scale from a single machine to thousands of machines, processing petabytes of data.

Core Components of Apache Spark:

1. **Spark Core**: The foundation of Spark, providing basic functionalities like task scheduling, memory management, and fault recovery.
2. **Spark SQL**: Allows querying of structured and semi-structured data using SQL and DataFrames.
3. **Spark Streaming**: Enables processing of real-time data streams.

4. **MLlib**: A machine learning library with algorithms for classification, regression, clustering, and collaborative filtering.
5. **GraphX**: A library for graph and graph-parallel computation.

Prerequisites:

- **Operating System:** Linux (e.g., Ubuntu), Windows, or macOS.
- **Java:** Java Development Kit (JDK) 8 or later installed.
- **Python (Optional):** For PySpark.
- **Hadoop (Optional):** If using HDFS as the data source.
- **Internet Connection:** For downloading Spark.

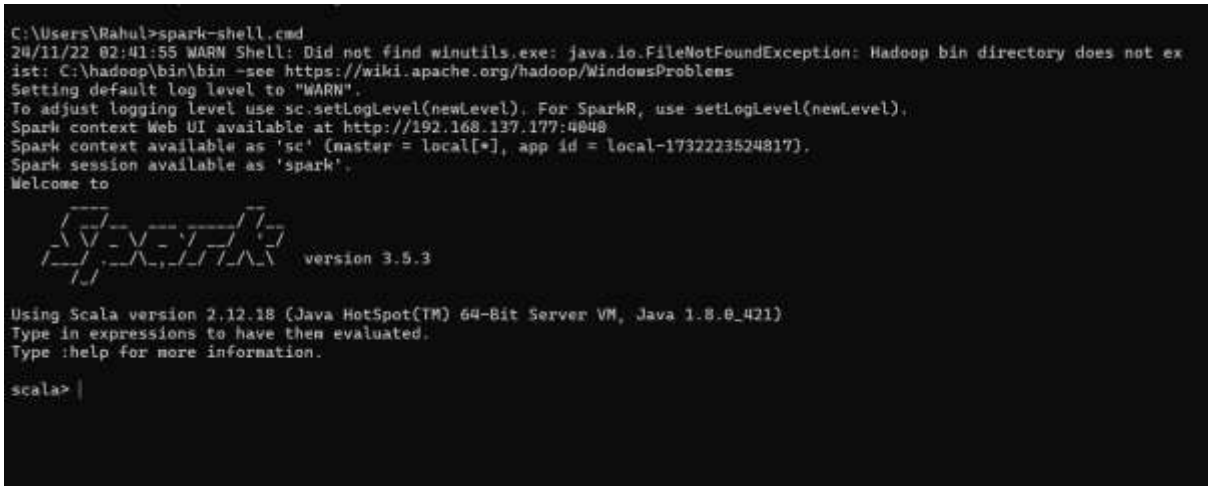
Procedure:

1. Visit the Apache Spark website.
2. Select a Spark version compatible with your system and download the pre-built binary.
3. Extract the downloaded file.

Configure Environment Variables

4. Edit the ~/.bashrc file:
    1. nano ~/.bashrc
5. Add the following lines
    1. export SPARK_HOME=/opt/spark export PATH=$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH
6. Start the Spark master:
    1. start-master.sh
7. Start a Spark worker:
    1. start-slave.sh spark://<master-IP>:<master-port>
8. Running a Simple Spark Application

Output:



Conclusion:

Apache Spark is a powerful, versatile, and efficient tool for big data processing. Its in-memory computation, ease of use, and ability to handle diverse workloads—like batch processing, streaming, and machine learning make it a go-to platform for scalable and real-time data analytics in modern data-driven applications.