BDA Lab
Assignment-1

Experiment 1: Install apache Hadoop and show the steps inviolved.

Software Requirements

- Ubuntu Linux (Virtual Machine)
- Java (OpenJDK 11)
- Apache Hadoop 3.x
- SSH

Description

Apache Hadoop is an open-source framework used for distributed storage and processing of large datasets using the HDFS and MapReduce programming model. Hadoop follows a master–slave architecture consisting of NameNode, DataNode, and Secondary NameNode.

Steps Involved

Step 1: Install Java

```
sudo apt update
sudo apt install openjdk-11-jdk -y
```

Verify:

```
java -version
```

Step 2: Download and Install Hadoop

```
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
tar -xvzf hadoop-3.3.6.tar.gz
sudo mv hadoop-3.3.6 /usr/local/hadoop
sudo chown -R hadoop:hadoop /usr/local/hadoop
```

Step 3: Set Environment Variables

Edit .bashrc:

```
nano ~/.bashrc
```

Add:

```
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Apply:

```
source ~/.bashrc
```

Step 4: Configure JAVA_HOME

```
nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

Set:

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Step 5: Configure Hadoop Files

core-site.xml

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

hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

Step 6: Enable SSH

```
ssh-keygen
cat ~/.ssh/id_*.pub >> ~/.ssh/authorized_keys
chmod 600 ~/.ssh/authorized_keys
```

Step 7: Format and Start Hadoop

```
hdfs namenode -format
start-dfs.sh
```

Verify:

```
jps
```

Result

Apache Hadoop was successfully installed and verified by running HDFS services and checking active daemons using the jps command.

Experiment 2: Study To study the various freely available frameworks to execute big data analytics scripts. do the study and analysis of at least 5 tools/frameworks.

1. Apache Hadoop

- Type: Batch Processing
- Core Components: HDFS, MapReduce, YARN
- Advantages:
    - Highly scalable
    - Fault tolerant
- Limitations:
    - High latency
    - Disk-based processing

2. Apache Spark

- Type: In-memory processing
- Components: Spark Core, Spark SQL, MLlib, Spark Streaming
- Advantages:
    - Faster than Hadoop
    - Supports Python, Java, Scala
- Use Case:
    - Machine learning and real-time analytics

3. Apache Flink

- Type: Stream + Batch Processing
- Features:
    - Low latency
    - Event-time processing

- Use Case:
    - Real-time analytics and monitoring

4. Apache Hive

- Type: SQL-based analytics
- Feature:
    - Converts SQL queries into MapReduce/Spark jobs
- Use Case:
    - Data warehousing and reporting
- Advantage:
    - Easy for SQL users

5. Apache Pig

- Type: Scripting framework
- Language: Pig Latin
- Use Case:
    - Data preprocessing and ETL
- Advantage:
    - Fewer lines of code than MapReduce

Comparative Analysis

Framework	Processing Type	Language Support	Best Use Case
Hadoop	Batch	Java	Large offline data
Spark	Batch + Streaming	Python, Java	Fast analytics
Flink	Real-time	Java, Scala	Streaming data
Hive	SQL Analytics	SQL	Data warehouse
Pig	Data Flow	Pig Latin	ETL processing

Result

The study of various freely available Big Data frameworks was carried out successfully. Each framework has distinct features and is suitable for different Big Data analytics scenarios.