Big Data Analytics
Experiment No. 03

Title : To install and configure Hadoop.

Theory :

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Hadoop was originally designed for computer clusters built from commodity hardware, which is still in common use. It has since also found use on clusters of higher-end hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel.

This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

Components of Hadoop :

Hadoop is a framework that uses distributed storage and parallel processing to store and manage Big Data. It is the most commonly used software to handle Big Data. There are three components of Hadoop.

1. Hadoop HDFS - Hadoop Distributed File System (HDFS) is the storage unit of Hadoop.
2. Hadoop MapReduce - Hadoop MapReduce is the processing unit of Hadoop.
3. Hadoop YARN - Hadoop YARN is a resource management unit of Hadoop.

STEPS TO INSTALL HADOOP

- **Installation and Configuration of Hadoop:**

1. . Install or download Java 1.8.0

   https://www.oracle.com/java/technologies/javase/javase8-archive-downloads.html

2. Download the latest stable release from the Apache Hadoop download websites

   https://hadoop.apache.org/docs/r2.8.0/

3. Install Java JDK 1.8.0 in "C:\JAVA“
4. Extract file Hadoop 2.8.0.zip and place under "C:\Hadoop-2.8.0".
5. Set the path HADOOP_HOME Environment variable on windows 10
6. Set the path JAVA_HOME Environment variable on windows 10.
7. Next we set the Hadoop bin directory path and JAVA bin directorypath.

- **Configuration:**

   1. Edit file C:/Hadoop-2.8.0/etc/hadoop/core-site.xml, paste below xml paragraph and save this file.

      ```
      <configuration>

      <name>fs.default</name>

      <value>hdfs://localhost:9000</value>

      </property>
      ```

```
</configuration>
```

2. Rename "mapred-site.xml.template" to "mapred-site.xml" and edit this file C:/Hadoop-2.8.0/etc/hadoop/mapred-site.xml, paste below xml paragraph and save this file.

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

3. Create folder "data" under "C:\Hadoop-2.8.0". Create folder "datanode" under "C:\Hadoop-2.8.0\data". Create folder "namenode" under "C:\Hadoop-2.8.0\data" data.
4. Edit file C:\Hadoop-2.8.0/etc/hadoop/hdfs-site.xml, paste below xml paragraph and save this file.2.8.0/etc/hadoop/mapred-site.xml, paste below xml paragraph and save this file.

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>C:\hadoop-2.8.0\data\namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>C:\hadoop-2.8.0\data\datanode</value>
</property>
</configuration>
```

5. Edit file C:/Hadoop-2.8.0/etc/hadoop/yarn-site.xml, paste below xml paragraph and save this file.

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

6. Edit file C:/Hadoop-2.8.0/etc/hadoop/hadoop-env.cmd by closing the command line "JAVA_HOME=%JAVA_HOME%" instead of set "JAVA_HOME=C:\Java" (On C:\java this is path to file jdk.18.0).
7. Download file Hadoop Configuration.zip (Link:https://raw.githubusercontent.com/MuhammadBilalYar/Hadoop-On-Window/master/Hadoop%20Configuration.zip).
8. Delete file bin on C:\Hadoop-2.8.0\bin, replaced by file bin on file just download (from Hadoop Configuration.zip).
9. Open cmd and typing command "hdfs namenode –format".
10. Open cmd and change directory to "C:\Hadoop-2.8.0\sbin" and type "start-all.cmd" to start apache.
11. Open: http://localhost:8088 to check if hadoop is running