Nanditha.V
160119733133

DATA SCIENCE AND BIG DATA ANALYTICS

EXPERIMENT-8: Installation of Big data technologies and building a Hadoop Cluster

AIM: Installation of Big data technologies and building a Hadoop cluster.

Description:

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Hadoop consists of four main modules:

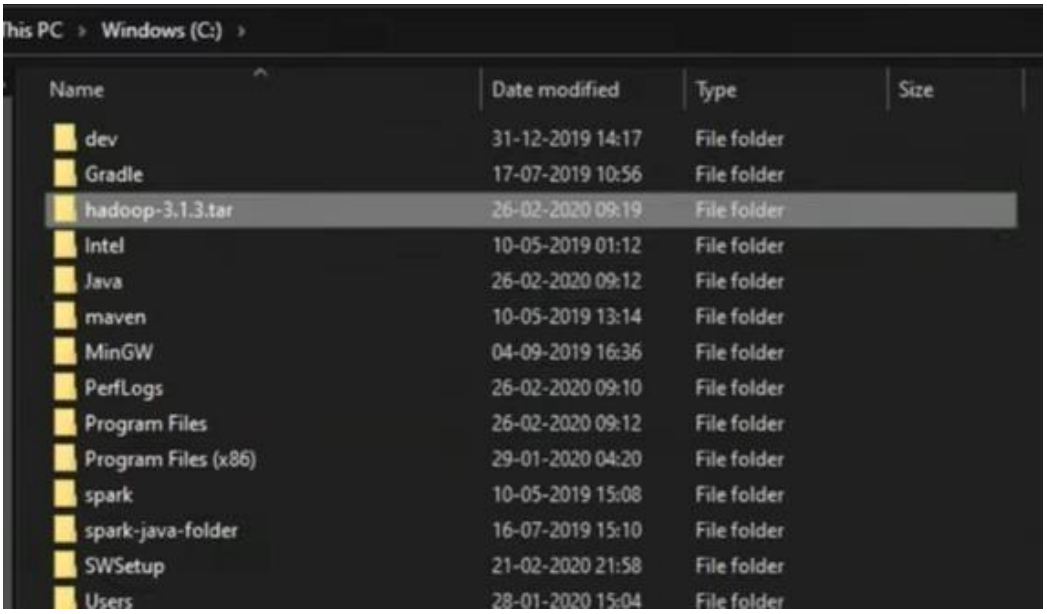
- Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
- Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
- MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
- Hadoop Common – Provides common Java libraries that can be used across all modules.

PROCEDURE:

Step 1: Verify the Java installed
javac -version

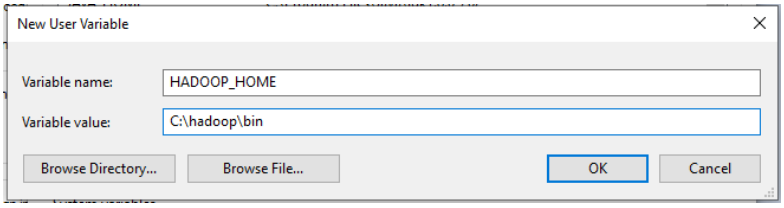
```
PS C:\Users\Geetansh Sahni> javac -version
javac 11.0.4
PS C:\Users\Geetansh Sahni>
```

Step 2: Extract Hadoop at C:\Hadoop



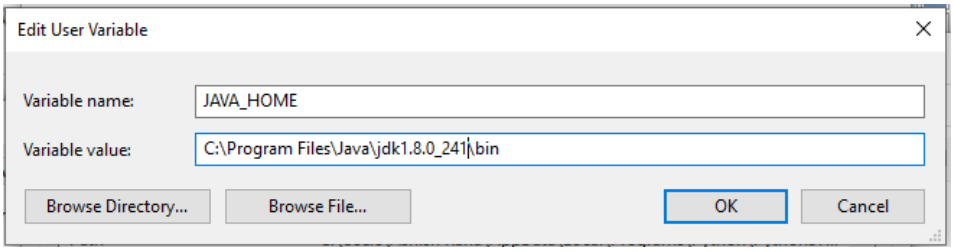
Step 3: Setting up the HADOOP_HOME variable

Use windows environment variable setting for Hadoop Path setting.

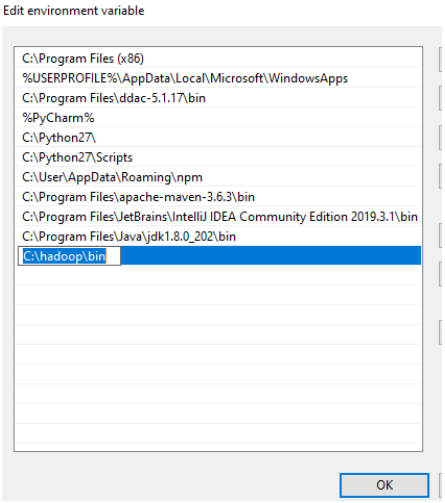


Step 4: Set JAVA_HOME variable

Use windows environment variable setting for Hadoop Path setting.



Step 5: Set Hadoop and Java bin directory path



Step 6: Hadoop Configuration :
For Hadoop Configuration we need to modify Six files that are listed below-

1. Core-site.xml

2. Mapred-site.xml

3. Hdfs-site.xml

4. Yarn-site.xml

5. Hadoop-env.cmd

6. Create two folders datanode and namenode

Step 6.1: Core-site.xml configuration

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

Step 6.2: Mapred-site.xml configuration

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

Step 6.3: Hdfs-site.xml configuration

```
<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>
```

Step 6.4: Yarn-site.xml configuration

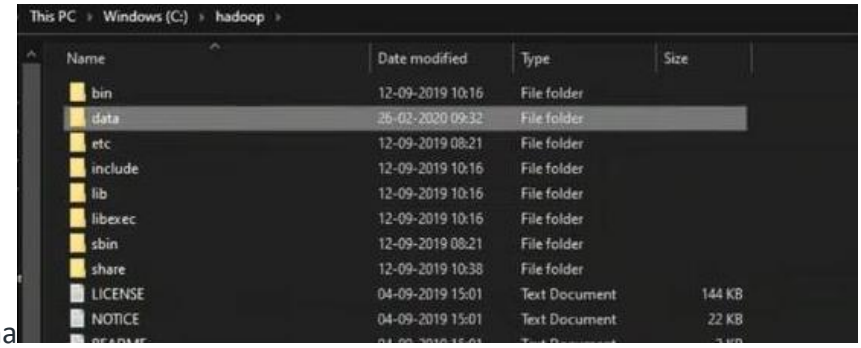
```
<configuration>
  <property>
     <name>yarn.nodemanager.aux-services</name>
     <value>mapreduce_shuffle</value>
  </property>
  <property><name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
   <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

Step 6.5: Hadoop-env.cmd configuration
Set "JAVA_HOME=C:\Java" (On C:\java this is path to file jdk.18.0)

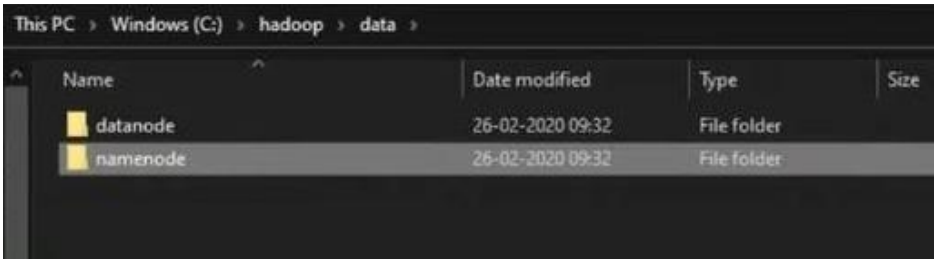
```
16
17  @rem Set Hadoop-specific environment variables here.
18
19  @rem The only required environment variable is JAVA_HOME.  All others are
20  @rem optional.  When running a distributed configuration it is best to
21  @rem set JAVA_HOME in this file, so that it is correctly defined on
22  @rem remote nodes.
23
24  @rem The java implementation to use.  Required.
25  set JAVA_HOME=%JAVA_HOME%
```

Step 6.6: Create datanode and namenode folders
1. Create folder "data" under "C:\Hadoop-2.8.0"

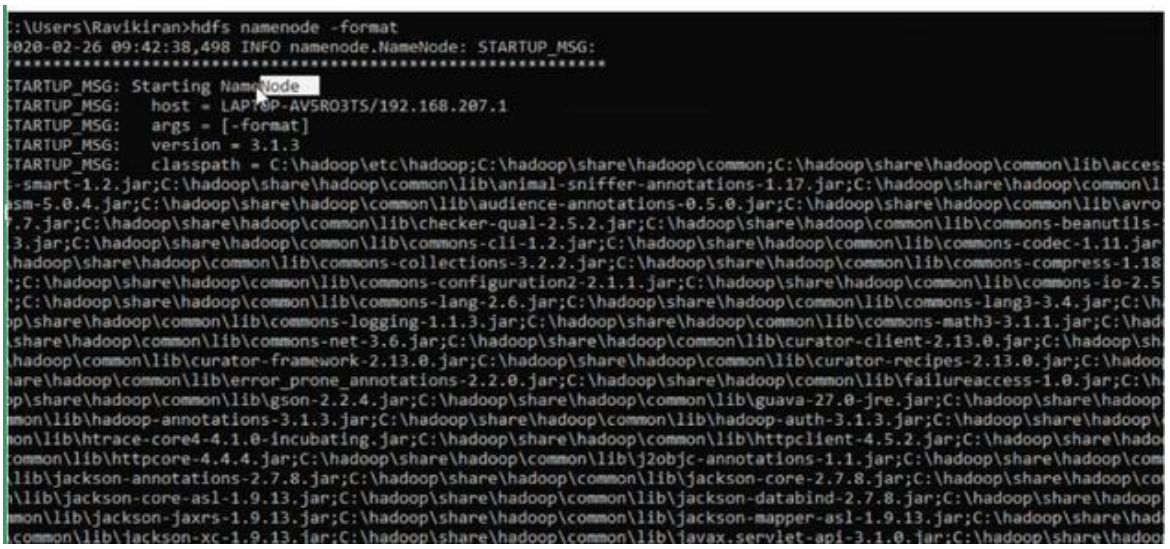
2. Create folder "datanode" under "C:\Hadoop-2.8.0\data"



3. Create folder "namenode" under "C:\Hadoop-2.8.0\data"



Step 7: Format the namenode folder
Open command window (cmd) and typing command "hdfs namenode –format"



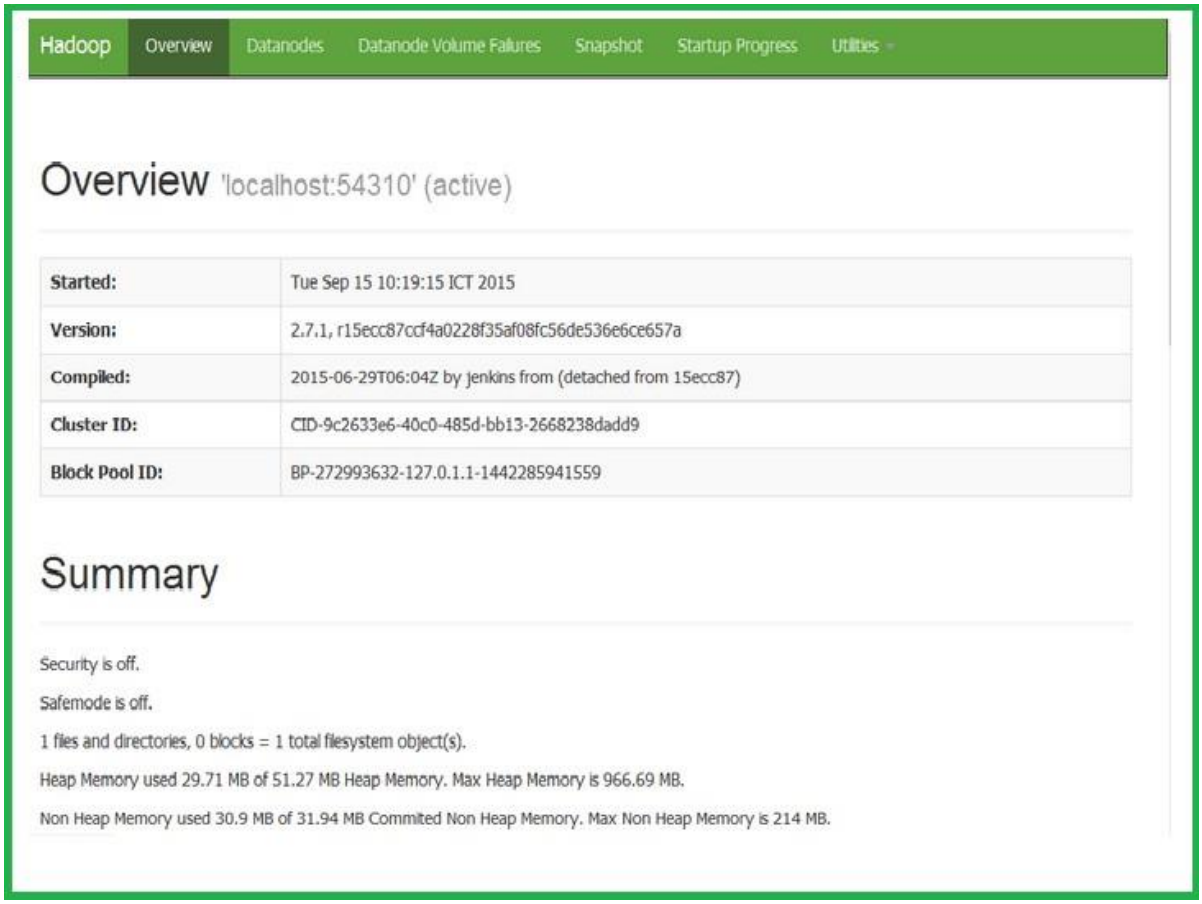
Step 8: Testing the setup
Open command window (cmd) and typing command "start-all.cmd"

Step 8.1: Testing the setup:
Ensure that namenode, datanode, and Resource manager are running

Step 9: Open: http://localhost:8088

Step 10:

Open: http://localhost:50070



ANALYSIS:

Here, we have successfully installed the Hadoop.