DATA ANALYTICS
ASSIGNMENT-1

Apache Hadoop:

Apache Hadoop is an open-source framework that allows for the distributed processing of large datasets across clusters of computers using simple programming models. It is designed to scale up from a single server to thousands of machines, each offering local computation and storage. Hadoop is widely used for big data analytics and supports a range of use cases, from data warehousing and ETL (Extract, Transform, Load) operations to machine learning and data mining.

1.1 History of Hadoop:

Hadoop originated from projects at Yahoo! and Apache Nutch, which were inspired by Google's papers on the Google File System (GFS) and MapReduce.

- **2003**: Google published the GFS paper, which described a distributed file system designed to handle large data sets using commodity hardware.
- **2004**: Google released the MapReduce paper, describing a programming model and associated implementation for processing and generating large datasets.
- **2006**: Doug Cutting and Mike Cafarella created Hadoop to support the Nutch search engine project.
- **2008**: Hadoop became a top-level Apache project, gaining wider adoption and further development.
- Since then, Hadoop has become a cornerstone of big data, with ongoing contributions and improvements from the open-source community and industry.

1.2 Versions of Hadoop:

Hadoop has evolved through several major versions, each offering performance improvements, feature enhancements, and expanded support.

- **Hadoop 1.x**: The initial version, featuring MapReduce as the primary processing engine and HDFS for storage.
- **Hadoop 2.x**: Introduced YARN (Yet Another Resource Negotiator) which allowed for better resource management and enabled multiple processing models other than MapReduce to work on Hadoop clusters (like Apache Tez, Apache Spark).
- **Hadoop 3.x**: Introduced features such as erasure coding for more efficient storage, support for more than 2 Name Nodes, and containerization with Docker.

1.3 System Requirements for Hadoop (all OS):

Hadoop can be installed on various operating systems like Windows, macOS, and Linux. Below are general system requirements:

- **RAM**: Minimum 4 GB RAM (8 GB or more recommended).
- **CPU**: Multi-core processor (quad-core or higher recommended for faster processing).
- **Disk Space**: At least 20 GB of free disk space for storage (more for large datasets and better performance).
- **Java**: Requires Java Development Kit (JDK) 8 or higher.
- **Operating Systems**:
    - Linux (preferred OS for Hadoop clusters due to its stability and performance in distributed environments).
    - macOS (for development purposes). ○ Windows (not typically used for production clusters but supported in standalone mode).

1.4 Installation Steps on Macos:

Step 1: Install Java 1.8 version

```
mohamedhussainshahulhameed@Mohameds-Laptop ~ % java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)
```

Step 2: Enable SSH for local host

```
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/id_rsa.pub
ssh localhost
```

Step 3: Download Hadoop and modify z-profile

```
source ~/.zprofile
```

Step 4: Configure Hadoop

```
sudo code $HADOOP_HOME/etc/hadoop/hadoop-env.sh
/usr/libexec/java_home
export
JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_333.jdk
/Contents/Home
```

Step 5: Edit core-site.xml

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

Step 6: Edit hdfs-site.xml

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

Step 7: Edit mapred-site.xml

```
<configuration>
 <property>
  <name>yarn.app.mapreduce.am.env</name>
  <value>HADOOP_MAPRED_HOME=/opt/homebrew/opt/hadoop</value>
 </property>
 <property>
  <name>mapreduce.map.env</name>
  <value>HADOOP_MAPRED_HOME=/opt/homebrew/opt/hadoop</value>
 </property>
 <property>
  <name>mapreduce.reduce.env</name>
  <value>HADOOP_MAPRED_HOME=/opt/homebrew/opt/hadoop</value>
 </property>
</configuration>
```

Step 8: Edit yarn-site.xml

```
<configuration>
```

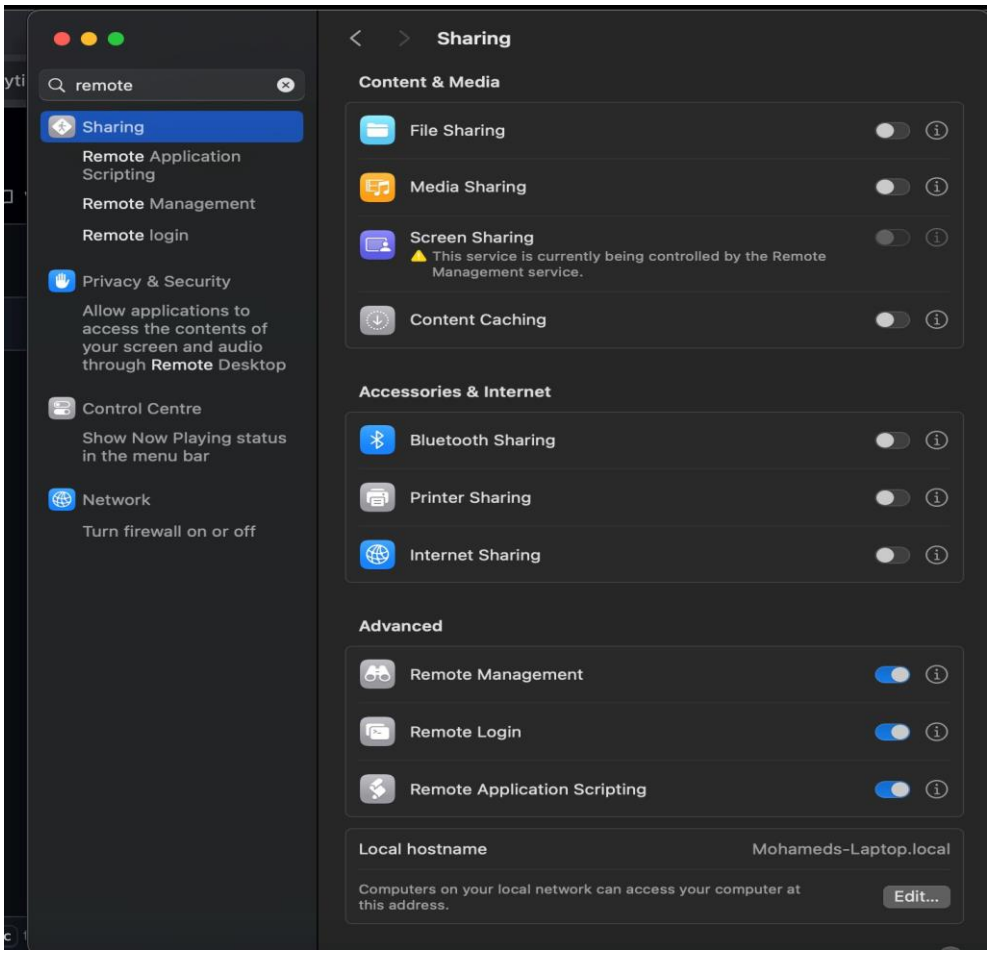
```
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

Step 9: Start Hadoop

Start-all.sh

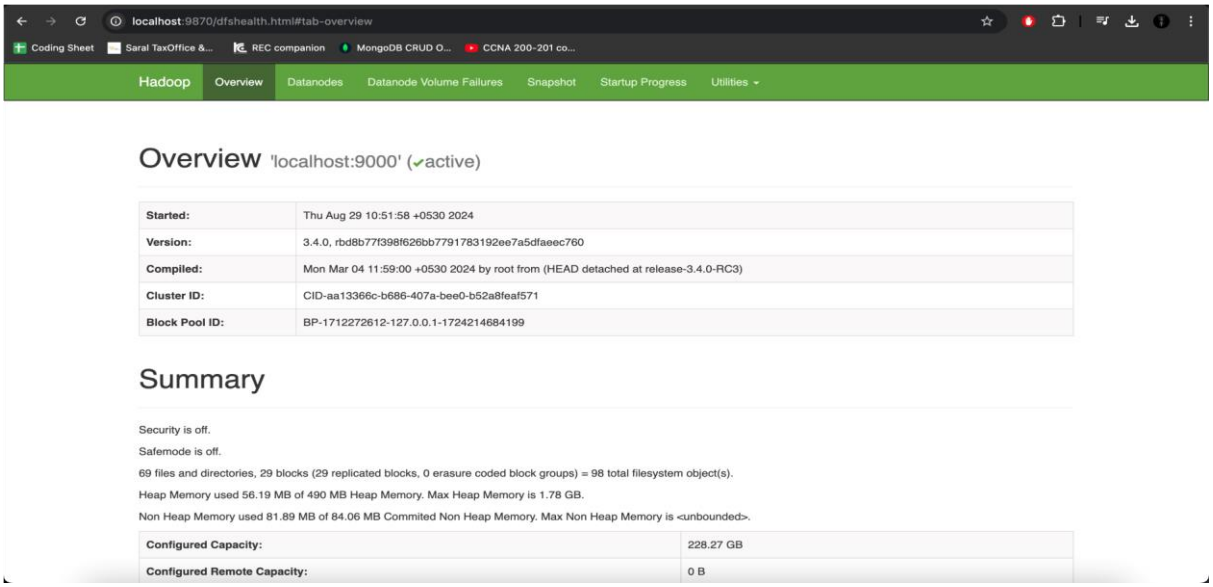
1.5 Installation Screenshots

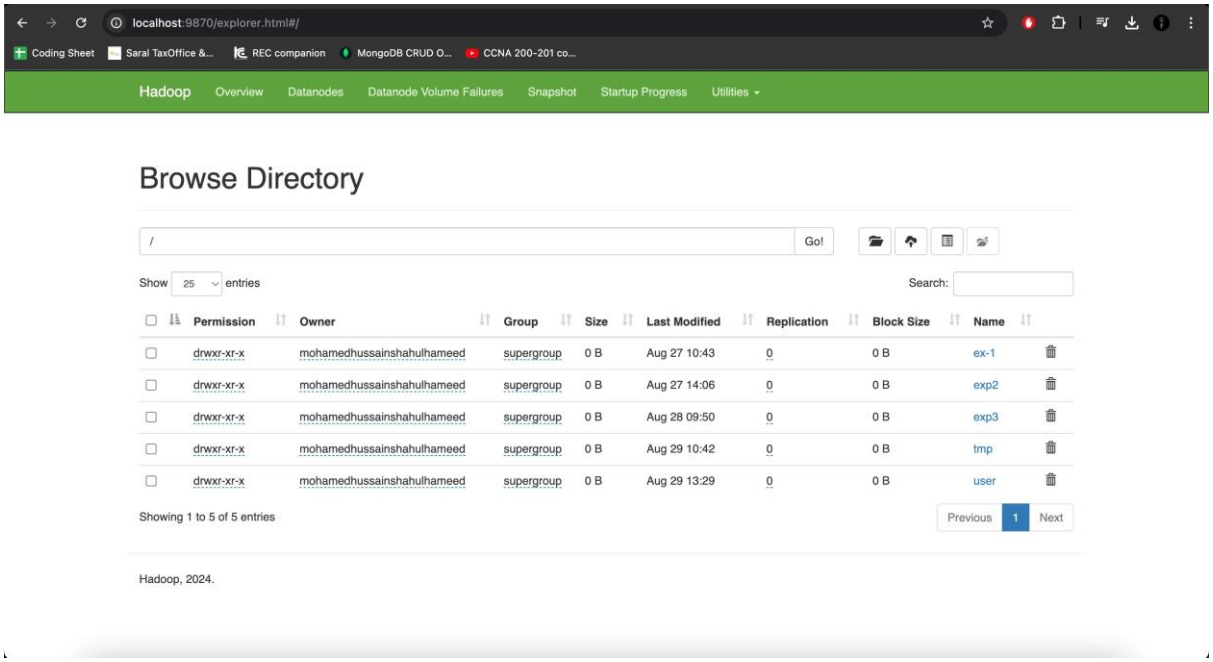
```
mohamedhussainshahulhameed@Mohameds-Laptop ~ % echo $JAVA_HOME
/Library/Java/JavaVirtualMachines/jdk1.8.0_202.jdk/Contents/Home
mohamedhussainshahulhameed@Mohameds-Laptop ~ %
```



```
mohamedhussainshahulhameed@Mohameds-Laptop ~ % ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/Users/mohamedhussainshahulhameed/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /Users/mohamedhussainshahulhameed/.ssh/id_rsa
Your public key has been saved in /Users/mohamedhussainshahulhameed/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:DQuoib4aqa9qUaMHQRzztUj39rVL2JgLjjhSmHcNE+s mohamedhussainshahulhameed@Mohameds-Laptop.local
The key's randomart image is:
+---[RSA 3072]----+
|o+.. +           |
| o+ +.=          |
|  .o.=.o.  .     |
| oo=. =..+* .    |
|.o*o.E oS=.+     |
|.+o.o o . o .    |
|oooo . . . .     |
|.oo .            |
|O+.              |
+----[SHA256]-----+
```

```
mohamedhussainshahulhameed@Mohameds-Laptop ~ % jps
79619 Jps
75348 ResourceManager
75526 NodeManager
75014 DataNode
75151 SecondaryNameNode
74831 NameNode
52430 RunJar
```




localhost:9870/explorer.html#/
Coding Sheet
Saral TaxOffice &...
REC companion
MongoDB CRUD O...
CCNA 200-201 co...
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x mohamedhussainshahulhameed supergroup 0 B Aug 27 10:43 0 0 B ex-1
drwxr-xr-x mohamedhussainshahulhameed supergroup 0 B Aug 27 14:06 0 0 B exp2
drwxr-xr-x mohamedhussainshahulhameed supergroup 0 B Aug 28 09:50 0 0 B exp3
drwxr-xr-x mohamedhussainshahulhameed supergroup 0 B Aug 29 10:42 0 0 B tmp
drwxr-xr-x mohamedhussainshahulhameed supergroup 0 B Aug 29 13:29 0 0 B user
Showing 1 to 5 of 5 entries
Previous
1
Next
Hadoop, 2024.