DATA ANALYTICS Naresh kumar V

ASSIGNMENT 1 210701323

HADOOP

INTRODUCTION:

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets.

HISTORY:

2003	Google released the paper, Google File System (GFS).
2004	Google released a white paper on Map Reduce.
2006	○ Hadoop introduced. ○ Hadoop 0.1.0 released. ○ Yahoo deploys 300 machines and within this year reaches 600 machines.
2007	○ Yahoo runs 2 clusters of 1000 machines. ○ Hadoop includes HBase.
2008	○ YARN JIRA opened ○ Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds. ○ Yahoo clusters loaded with 10 terabytes per day. ○ Cloudera was founded as a Hadoop distributor.
2009	○ Yahoo runs 17 clusters of 24,000 machines. ○ Hadoop becomes capable enough to sort a petabyte. ○ MapReduce and HDFS become separate subproject.

Hardware Requirements

2010	o Hadoop added the support for Kerberos. o Hadoop operates 4,000 nodes with 40 petabytes.
	o Apache Hive and Pig released.
2011	o Apache Zookeeper released. o Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage.
2012	Apache Hadoop 1.0 version released.
2013	Apache Hadoop 2.2 version released.
2014	Apache Hadoop 2.6 version released.
2015	Apache Hadoop 2.7 version released.
2017	Apache Hadoop 3.0 version released.
2018	Apache Hadoop 3.1 version released.

1. **Memory:**
    - At least 8 GB of RAM per machine (16 GB or more is recommended for production environments).
2. **Storage:**
    - At least 500 GB of disk space per machine.
    - Use high-speed disks (SSD) for better performance.
3. **CPU:** o Multi-core processors are recommended. At least 4 cores per machine.
4. **Network:** o High bandwidth (1 Gbps or higher) network connection between nodes.

Software Requirements

1. **Operating System:**
    - Linux-based OS (e.g., CentOS, Ubuntu, Debian). ○ Some versions of Hadoop support Windows, but Linux is preferred for production.
2. **Java:**
    - Oracle JDK 8 or OpenJDK 8 (Java 8).
    - Some versions of Hadoop may support Java 11, but it's crucial to verify compatibility.
3. **SSH:**
    - Password-less SSH (Secure Shell) setup for communication between nodes.
4. **Hadoop Distribution:** ○ Latest stable version of Hadoop. You can download it from the Apache Hadoop website.
5. **Additional Software:**
    - Python (optional, but recommended for certain Hadoop ecosystem tools).
    - Various Hadoop ecosystem components (e.g., HDFS, YARN, MapReduce, Hive, HBase, etc.) as required by your specific use case.

Configuration Considerations

1. **Cluster Management:**
    - Use tools like Apache Ambari, Cloudera Manager, or other cluster management tools for easier setup and maintenance.
2. **Resource Management:**
    - Properly configure YARN for resource allocation.
    - Set appropriate heap sizes for NameNode and DataNode based on available memory.
3. **Replication Factor:** ○ Set the HDFS replication factor based on data redundancy needs (default is 3).
4. **Network Configuration:**
    - Ensure proper network configuration and DNS settings.
    - Optimize network settings for Hadoop traffic.

INSTALLATION STEPS

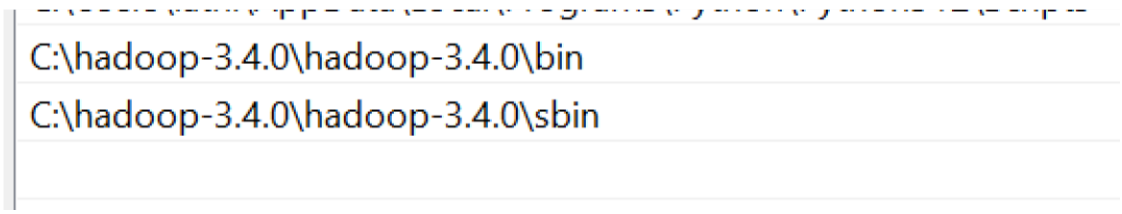
```
C:\Windows\System32>java -version
java version "1.8.0_421"
Java(TM) SE Runtime Environment (build 1.8.0_421-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.421-b09, mixed mode)
```

Check for Hadoop version

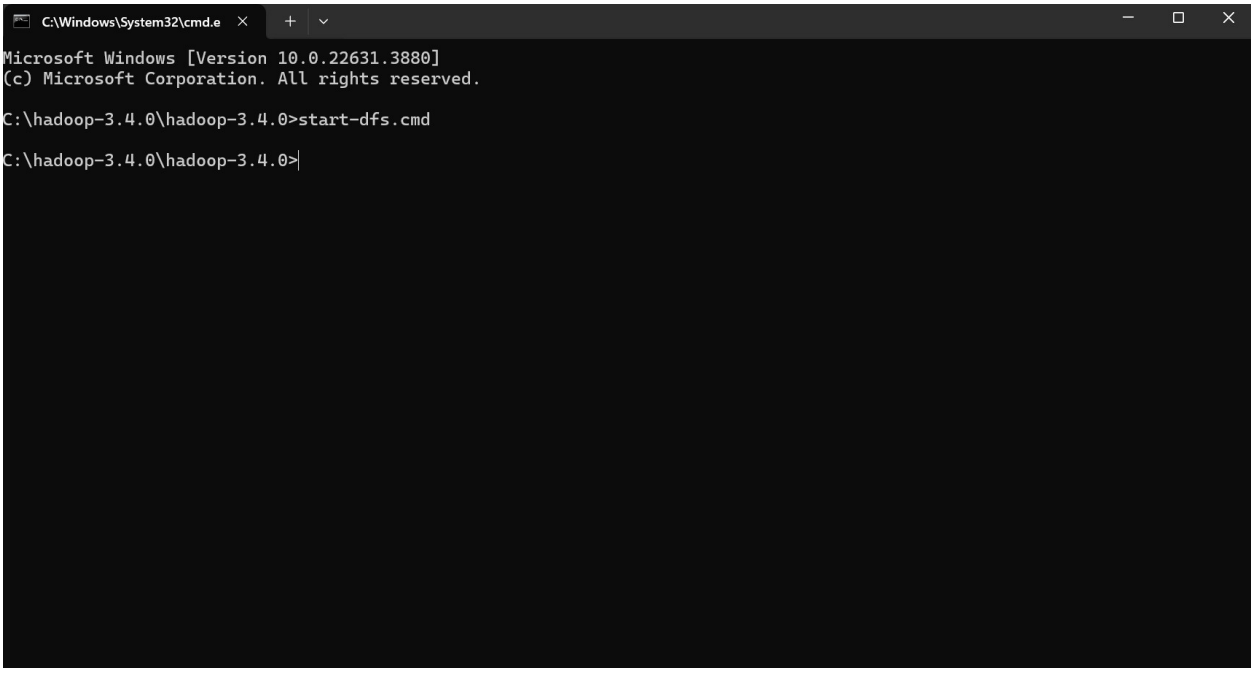
```
C:\Windows\System32>hadoop
Usage: hadoop [--config confdir] [--loglevel loglevel] COMMAND
where COMMAND is one of:
  fs                   run a generic filesystem user client
  version              print the version
  jar <jar>            run a jar file
                       note: please use "yarn jar" to launch
                             YARN applications, not this command.
  checknative [-a|-h]  check native hadoop and compression libraries availability
  conftest             validate configuration XML files
  distch path:owner:group:permisson
                       distributed metadata changer
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath            prints the class path needed to get the
                       Hadoop jar and the required libraries
  credential           interact with credential providers
  jnipath              prints the java.library.path
  kerbname             show auth_to_local principal conversion
  kdiag                diagnose kerberos problems
  key                  manage keys via the KeyProvider
  trace                view and modify Hadoop tracing settings
  daemonlog            get/set the log level for each daemon
 or
  CLASSNAME            run the class named CLASSNAME

Most commands print help when invoked w/o parameters.
```

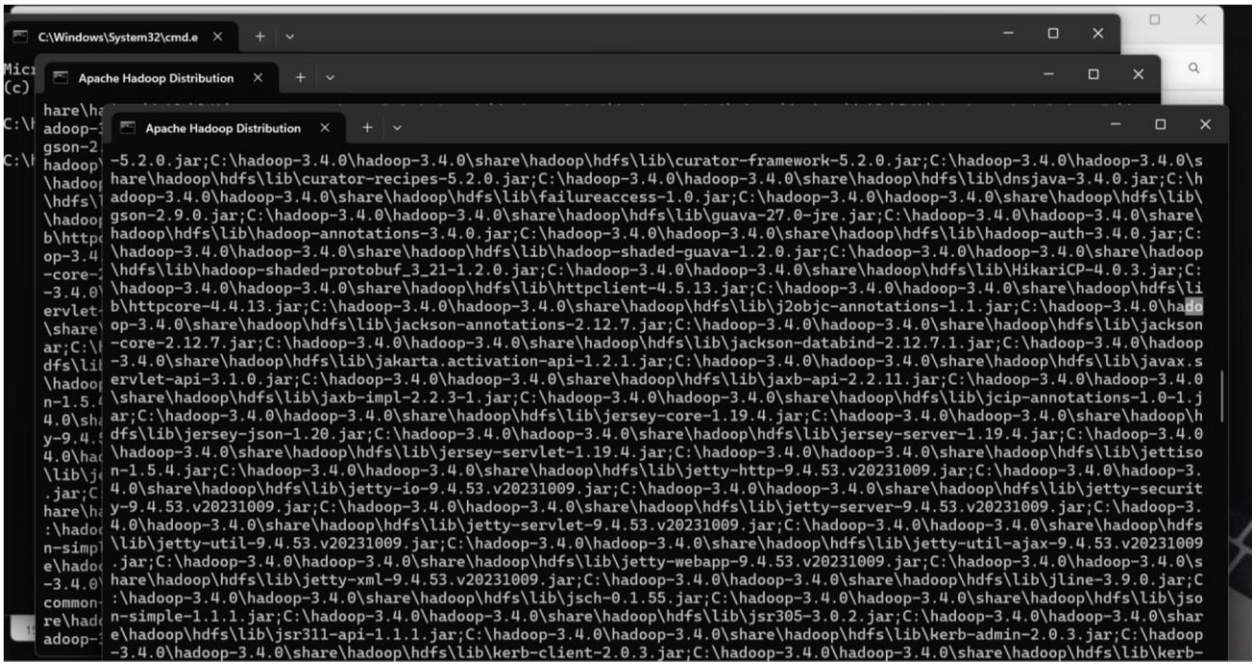
add path variables for java and Hadoop

HADOOP_HOME	C:\hadoop-3.4.0\hadoop-3.4.0\bin
JAVA_HOME	C:\java\jdk-1.8

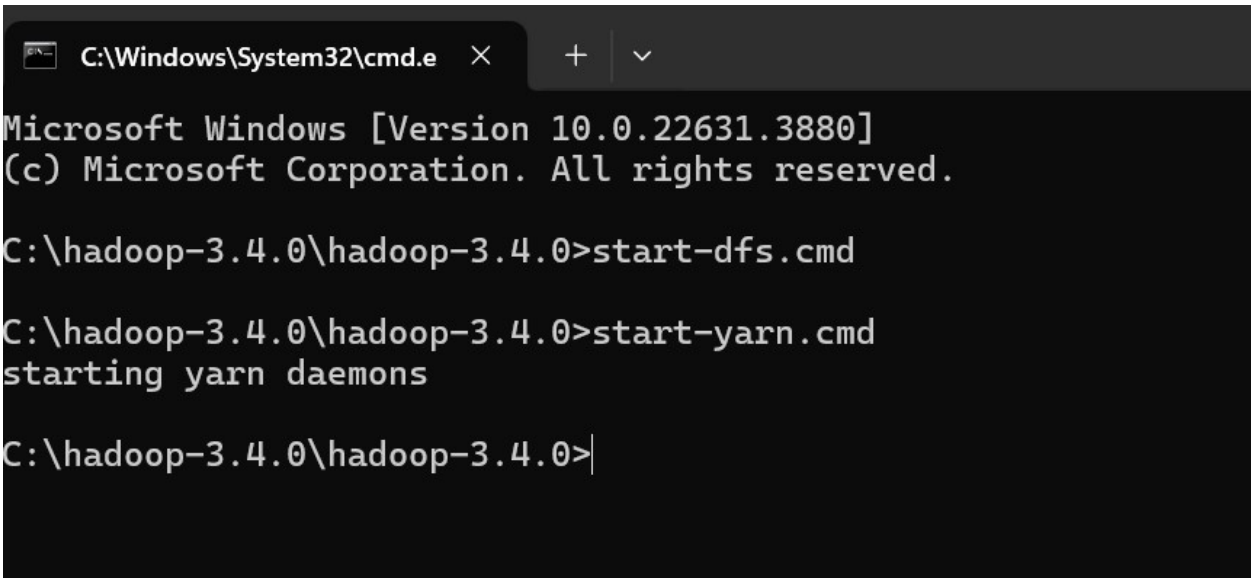


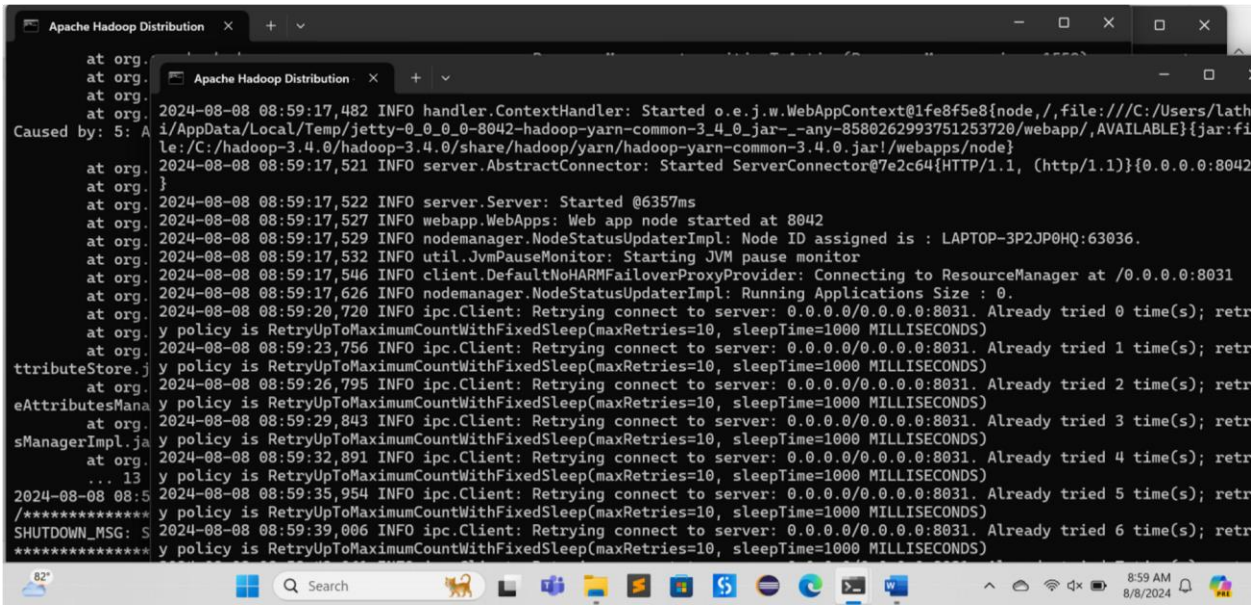
Run start -dfs.cmd



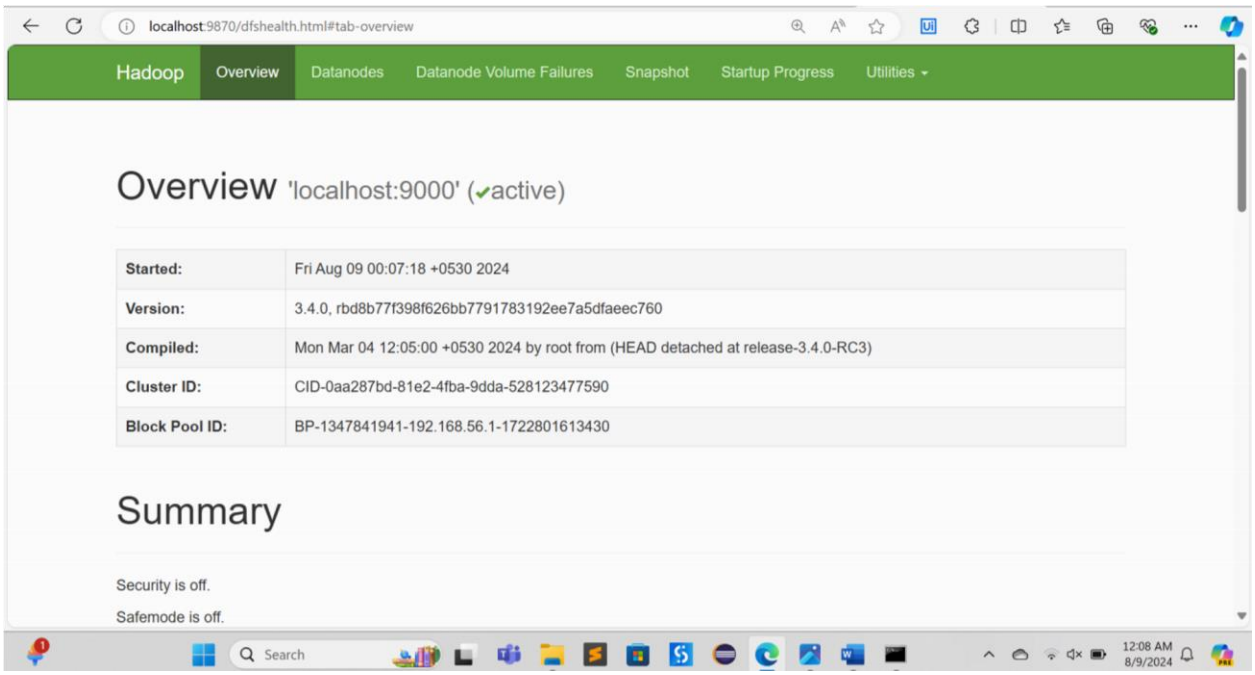


Run start-yarn.cmd





Run in the localhost using localhost:9870



Run using localhost:8088