Implementation of MapReduce in windows

Pre-requisite:

- **Java Installation:** Download Java from this link: https://www.oracle.com/java/technologies/downloads/#java8.
  - Install Java
  - **Set Java environment variable:**
    - Go to Start->Edit the System environment variable->Environment variable.
    - Then Click new and enter variable name as "JAVA_HOME".
    - In the value field, Enter the Java path such as "C:\Java\jdk1.8.0_351".
    - Go to path and click edit. Then add new and type "%JAVA_HOME%\bin".
  - **To check java version:** Open **cmd** and type command "java –version"

Command Prompt

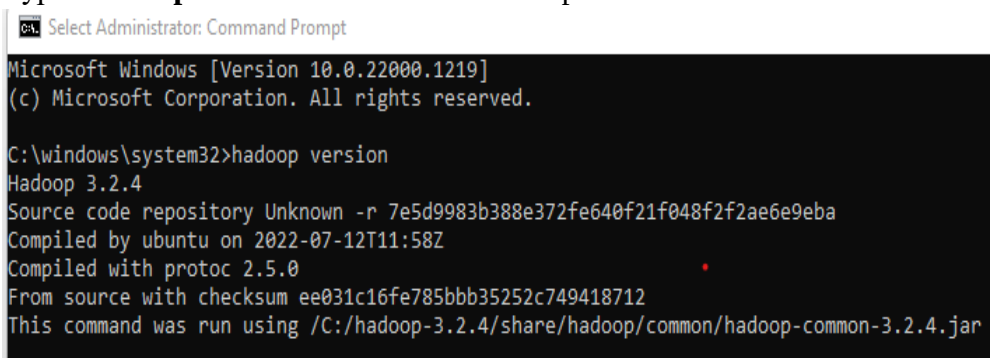
```
Microsoft Windows [Version 10.0.22000.1219]
(c) Microsoft Corporation. All rights reserved.

C:\Users\HP>java -version
java version "1.8.0_351"
Java(TM) SE Runtime Environment (build 1.8.0_351-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.351-b10, mixed mode)
```

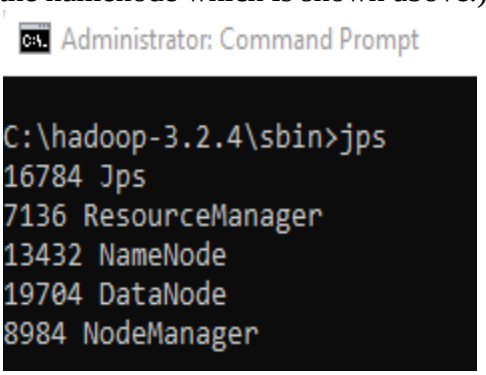
- **Hadoop Installation:** Download Hadoop from this link: https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.4/hadoop-3.2.4.tar.gz
  - Extract the tar file and stored the extracted file in the C drive. "**C:\hadoop-3.2.4**".
  - **Set Hadoop environment variable:**
    - Go to **Start→Edit the System environment variable→Environment variable**.
    - Then Click new and enter variable name as "HADOOP_HOME".
    - In the value field, Enter the Hadoop path such as "C:\hadoop-3.2.4".
    - Go to path and click edit. Then add new and type "%HADOOP_HOME%\bin". Also add "%HADOOP_HOME%\sbin".
  - Go to **C:\hadoop-3.2.4\etc\hadoop\...** folder and edit the below xml files.
    - ✓ **core-site.xml**
    - ✓ **mapred-site.xml**
    - ✓ **hdfs-site.xml**
    - ✓ **yarn-site.xml**

- ✓ Attached the codes in **Hadoop_4_important_xml_files.txt**

- Create folder "data" under "C:\hadoop-3.2.4"
  - ✓ Create folder "datanode" under "C:\hadoop-3.2.4\data"
  - ✓ Create folder "namenode" under "C:\hadoop-3.2.4\data"
- Edit the file **hadoop-env.cmd** from "C:\hadoop-3.2.4\etc\hadoop" by closing the command line "JAVA_HOME=%JAVA_HOME%" instead of set "JAVA_HOME= C:\Java\jdk1.8.0_351" (if your java file in Program Files the instead of give Progra~1 otherwise you will get JAVA_HOME incorrectly set error)
- Replace the **bin** file under "C:\hadoop-3.2.4" with the **bin** file as attached in **Bin.zip**
- Open **cmd** and type command "hdfs namenode –format". You will see through command prompt which tasks are processing, after completation you will get a message like '**namenode format succesfully and shutdown**'.
- **<u>Test hadoop installation:</u>** Open **cmd** as **administrator.**
  - → Type "**hadoop version**" to check the hadoop version.



  - → Then change the directory by typing "**cd C:/hadoop-3.2.4/sbin**"
  - → Type "**start-all.cmd**" to start all the hadoop daemons.
  - → Type "**jps**" and you will see all the namenode, datanode, resourcemanager and nodemanager has started. (If any of the below didn't started, then first you have to go to "**C:\hadoop-3.2.4\data\datanode**" folder and delete all files, again go to "**C:\hadoop-3.2.4\data\namenode**" folder and delete all files. Then format the namenode which is shown above.)



  - → Open: http://localhost:8088/cluster in any browser to view Nodes of the cluster

→ Open: http://localhost:9870/dfshealth.html#tab-overview in any browser to get overview

→ Now hadoop has succesfully installed in your System.

→

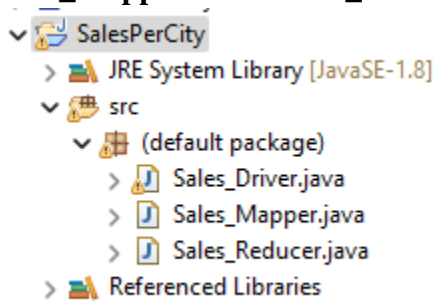
- **Eclipse Installation:** For Java Programming you have to install **Eclipse IDE**. Download Eclipse from this link: https://cutt.ly/HM6clSh

MapReduce:

To implement MapReduce, I have downloaded a dataset from https://www.kaggle.com/datasets/gaurang0405/item-sales, which is "**Sales.csv**". It contains information of item sales of a company of United States in different cities. The goal is to **find out Number of Items Sold in Each City of United States.**

- **Java Programming:**

→ In Eclipse I have create new java project as "**SalesPerCity**" and create 3 classes ("**Sales_Driver**", "**Sales_Mapper**" and "**Sales_Reducer**") in it.



→ To write the MapReduce code, first I have to add external JARs. To do that, right click on **SalesPerCity**, then click

Build Path→Configure Build Path→Libraries→Add External JAR

and select all the JAR files from (**C:\hadoop-3.2.4\share\hadoop\common** and **C:\hadoop-3.2.4\share\hadoop\mapreduce**) folder and then click "**Apply and Close**" button.

→ Then write the code, compile it and export all the classes in JAR file as **SalesPerCity.jar** where Main-class is "**Sales-Driver**".

→ I kept the jar file and csv file in C drive.

- **Start Hadoop:**

→ Open **cmd** as **administrator**, to start the hadoop, type "**start-all.cmd**".

→ By using "**jps**" command, ensure that, hadoop nodes are running.

→ To create an input directory, type "**hadoop fs –mkdir /SalesInput**", where **SalesInput** is input directory.

→ To put csv file as input in hadoop, first I have gave access of csv file by using command: "**hadoop fs -chmod 777 C:/Sales.csv**" and then type "**hadoop fs -put C:/Sales.csv /SalesInput**".

→ To ensure whether input file is successfully imported, type "**hadoop fs -ls /SalesInput/**".

```
Administrator: Command Prompt

C:\hadoop-3.2.4\sbin>hadoop fs -put C:/Sales.csv /SalesInput

C:\hadoop-3.2.4\sbin>hadoop fs -ls /SalesInput/
Found 1 items
-rw-r--r--   3 HP supergroup    2326291 2022-11-24 07:29 /SalesInput/Sales.csv
```

→ To view the content of the file, type "**hadoop dfs -cat /SalesInput/Sales.csv**".

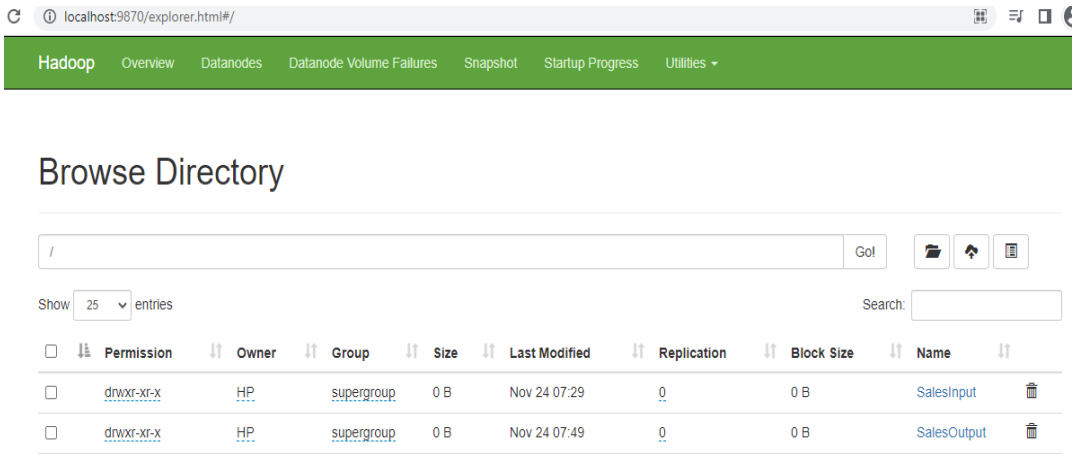
→ Now apply MapReduce program to the input file. To type the command, we have to follow: "**hadoop jar path_of_the_jar_file input_directory output_directory**". So type "**hadoop jar C:/SalesPerCity.jar /SalesInput /SalesOutput**", where **SalesOutput** is output directory. After, applying the jar file you can see the task performed in the MapReduce phase.

→ After completed the MapReduce tasks the output will be stored in the output directory. To see the output, type "**hadoop fs -cat /SalesOutput/***"

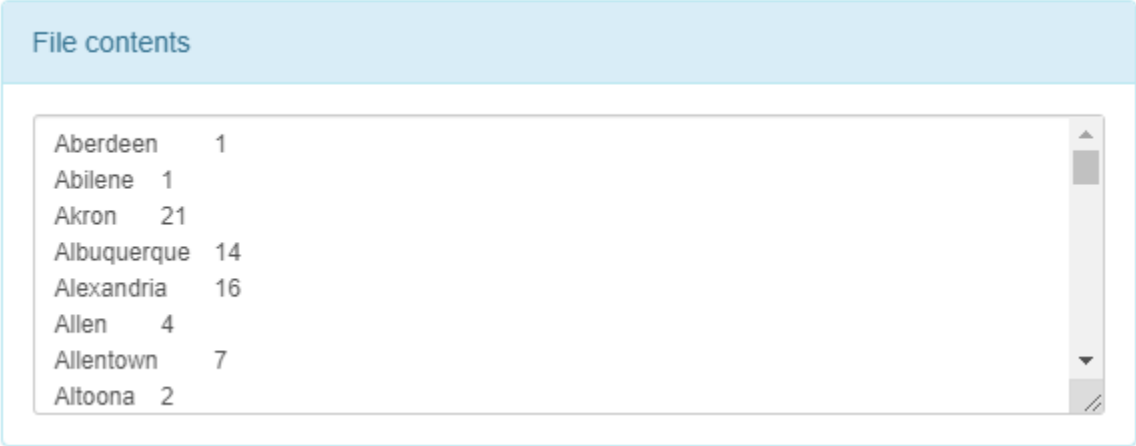
```
Administrator: Command Prompt

C:\hadoop-3.2.4\sbin>hadoop fs -cat /SalesOutput/*
Aberdeen        1
Abilene 1
Akron   21
Albuquerque     14
Alexandria      16
Allen   4
Allentown       7
Altoona 2
Amarillo        10
Anaheim 27
Andover 4
Ann Arbor       5
Antioch 1
Apopka  7
Apple Valley    9
Appleton        2
Arlington       60
```

→ Now, if you want to check the output in localhost, open: http://localhost:9870/explorer.html#/ in any browser.



Here, click **SalesOutput→part-00000→Head the file (first 32K)**



→ To stop the hadoop, type "**stop-all.cmd**"

Now the hadoop single node cluster was installed successfully and the MapReduce program were executed successfully in our windows system.