PRACTICAL-5

AIM: Perform the following MapReduce Programs in CloudxLab and Hadoop on windows.

1.	Implement/run the MapReduce program to count the frequency of each words. (perform on CLOUDXLAB Web-console and CLOUDXLAB Hadoop environment)
2.	Configure the Hadoop on windows and Implement/run the MapReduce program to count the frequency of each numbers.
3.	Find the average age of the people (both male and female) who died in the titanic tragedy using Hadoop MapReduce in cludxlab.

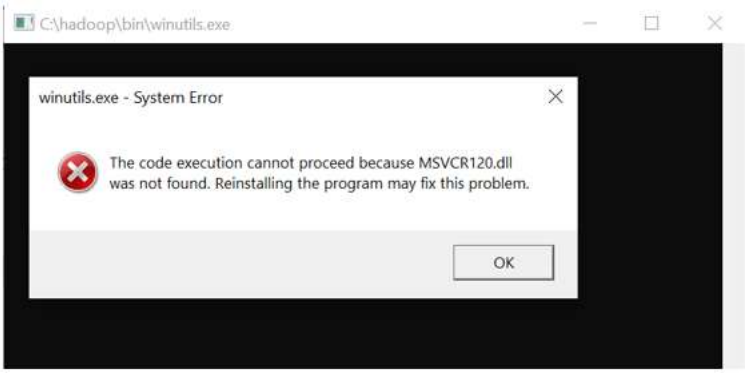
Perform following:

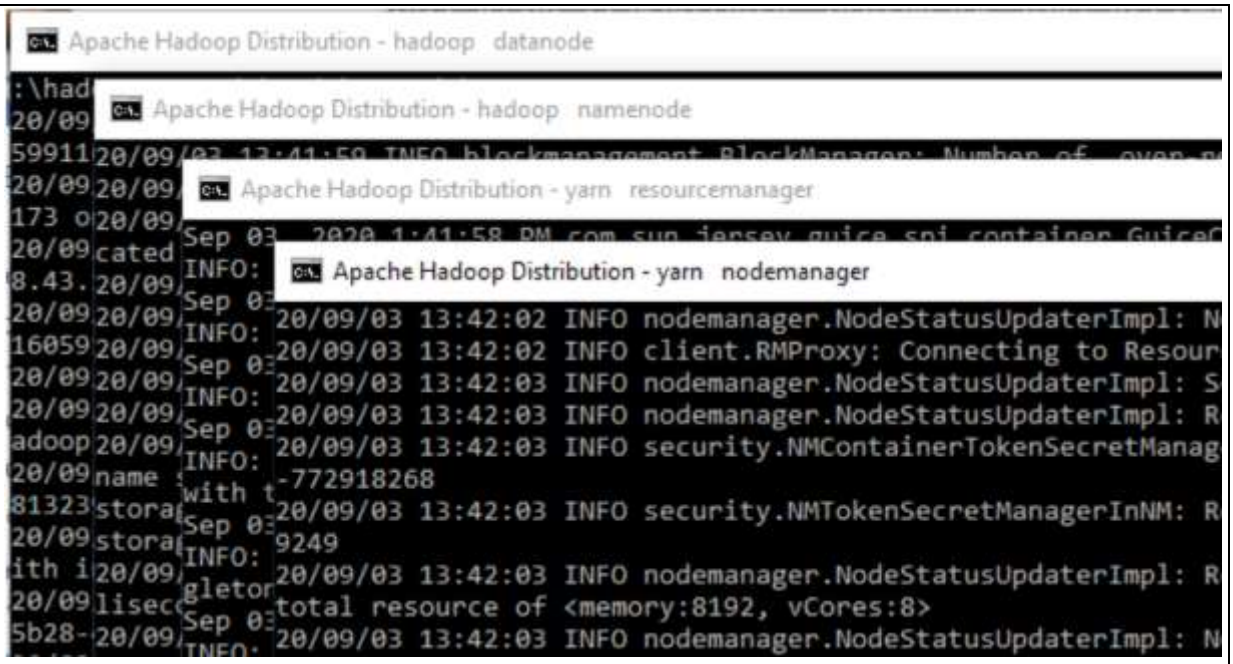
Install WinSCP on windows (It is required to copy the file from your local system to cloudxLab (Web console)

Implement/run the MapReduce program to count the frequency of each words. (Perform on CLOUDXLAB Web-console and CLOUDXLAB Hadoop environment)	
Required Files: mapper.py, reducer.py, numbers.txt, word.txt, MapReduceClient.jar	
MapReduce Program in Python (Run on Cloudxlab Local Web Console)	
1.	Create your own folder and copy the mapper.py, reducer.py, and word.txt in the folder.
2.	Change the permission of both python files. chmod a+x pms/reducer.py chmod a+x pms/mapper.py
3.	Run the mapper.py on word.txt file. cat word.txt | python mapper.py
4.	sort the data cat word.txt | python mapper.py | sort
5.	Run the mapper.py and reducer.py on word.txt cat word.txt | python mapper.py | sort | python reducer.py
MapReduce Program in Python (Run on Cloudxlab Hadoop environment)	
1.	Login into cloudxlab. Open web-console.
1.	Copy the mapper.py and reducer.py file in Local console (Any Folder or in same directory). Copy the word.txt file in Hadoop Console(HDFS) in user directory(Ex-pareshsolanki7426) Any Folder (in our case MapReduce folder)
2.	Change the permission of mapper.py and reducer.py chmod a+x reducer.py chmod a+x mapper.py
3.	Run following to find the location of hadoop-streaming jar file (it is required to use in mapreduce profram) find /usr/hdp -name hadoop-streaming.jar
4.	Perform following command to run MapReduce program on word.txt

	hadoop jar /usr/hdp/2.6.2.0-205/hadoop-mapreduce/hadoop-streaming.jar -input /user/pareshsolanki7426/MapReduce/word.txt -output /user/pareshsolanki7426/MapReduce/Wordcount -mapper "python mapper.py" -file /home/pareshsolanki7426/mapper.py -reducer "python reducer.py" -file /home/pareshsolanki7426/reducer.py
5.	See the final result hadoop fs -cat /user/pareshsolanki7426/MapReduce/Wordcount/part-00000
Configure the Hadoop on windows and Implement/run the MapReduce program to count the frequency of each numbers.	
Required files	
1	jdk-8u351-windows-x64.exe
	LINK: https://www.oracle.com/in/java/technologies/javase/javase8u211-later-archive-downloads.html#license-lightbox
2	hadoop-3.2.4.tar.gz
	LINK: https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.4/hadoop-3.2.4.tar.gz
3	hadoop3_xFixedbin.rar
4	msvcr120.zip
	LINK: https://www.dll-files.com/msvcr120.dll.html
5	VC_redist.x64.exe
	LINK: https://learn.microsoft.com/en-us/cpp/windows/latest-supported-vc-redist?view=msvc-170 and find the link https://aka.ms/vs/17/release/vc_redist.x64.exe
STEPS to Configure the HADOOP on WINDOWS	
1	Create a folder “Java” in C: drive
2	Install java (preferred jdk 8) – During Installation, on the “Destination folder GUI window”, Click on button "change" and select the path “C:\Java” – Go to “C:\program Files\java\jdk1.8.0_351 and cut the folder “jdk1.8.0_351”. Copy this folder in C:\Java\ – Delete the folder c:\program Files\java – Set the Environment User variable JAVA_HOME=C:\Java\jdk1.8.0_351\bin – Set the Environment System variable path=C:\Java\jdk1.8.0_351\bin
3	Download hadoop-3.2.4.tar.gz from apache site. – Run Winrar Application as administrator and Extract it in C:\ – Rename “C:\hadoop-3.2.4.tar.gz” to “C:\hadoop”

4	– Open “hadoop-env.cmd” file from “C:\hadoop\etc\hadoop\” in edit mode (in Notepad). – Find the location of JAVA_HOME in file and Paste the path “C:\Java\jdk1.8.0_351” in place of %JAVA_HOME%
5	– Go to Environment user variable. Click on NEW. Give variable name as “HADOOP_HOME” and value as “C:\hadoop\bin” – Go to Environment System variable. Double Click on “path”. Click on “new” and add the path “C:\hadoop\bin” – Go to Environment System variable. Double Click on “path”. Click on “new” and add the path “C:\hadoop\sbin”
6	Go to “C:\hadoop\etc\hadoop\” and open “core-site.xml” in edit mode and add the following code in between tag <configuration> </configuration>: <property> <name>fs.defaultFS</name> <value>hdfs://localhost:9000</value> </property>
7	– Go to “C:\hadoop” and Create folder named "data". – Inside “data” folder, Create folder "namenode" and "datanode" – Note: These folder paths are used in Next step as value
8	Go to “C:\hadoop\etc\hadoop\”. Open “hdfs-site.xml” in edit mode and add the following code in between <configuration> </configuration>: <property> <name>dfs.replication</name> <value>1</value> </property><property> <name>dfs.namenode.name.dir</name> <value>C:\hadoop\data\namenode</value> </property><property> <name>dfs.datanode.data.dir</name> <value>C:\hadoop\data\datanode</value> </property>
9	Go to “C:\hadoop\etc\hadoop\”. Open “mapred-site.xml” in edit mode and add the following code in between <configuration> </configuration>: <property> <name>mapreduce.framework.name</name> <value>yarn</value> </property>
10	Go to “C:\hadoop\etc\hadoop\”. Open “yarn-site.xml” in edit mode and add the following code in between <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>
11	Delete the folder "bin" from "C:\hadoop"
12	Extract "hadoop3_xFixedbin.rar". Copy the "bin" folder and paste it in "C:\hadoop" location.
13	Double click on "C:\hadoop\bin\winutils.exe". If you get error (See below Figure) then follow the following step: – Download "msvcr120.zip" and extract "msvcr120.dll" file. – Go to "C:\Windows\System32" and paste the file "msvcr120.dll" here. – Go to "C:\hadoop\bin". Double click on "winutils.exe" (Note: Run without error message) 
14	Download "msvc-170" (File name is "vc_redist.x64.exe") and run it. Restart the machine.
15	HADOOP SETUP is ready. To run the Hadoop, execute the following commands.
	Open "Command Prompt" (Run as administrator) C:> hdfs namenode –format (Note: Run only once after Hadoop configuration over) C:> cd hadoop C:\hadoop>cd bin C:\hadoop\bin>start-all.cmd (Note: It will start the four components: datanode, namenode, yarn-resource-manager, and yarn-node-manager)



NOTE: Do not close all above components. Keep it in running mode.
If everything is OK, then open the browser and run following:
localhost:9870

Perform the Mapreduce program:	
1.	Copy following files on "C:\" Drive – MapReduceClient.jar – input_file.txt
2	OPEN Command Prompt (Run as Administrator)
3	C:\>cd hadoop C:\hadoop>cd sbin C:\hadoop\sbin>start-all.cmd
4	C:\hadoop\sbin>hadoop fs -mkdir /inputdata C:\hadoop\sbin>hadoop fs -put c:/input_file.txt /inputdata C:\hadoop\sbin>hadoop fs -ls /inputdata/ C:\hadoop\sbin>hadoop dfs -cat /inputdata/input_file.txt
5	C:\hadoop\sbin>hadoop jar C:/MapReduceClient.jar wordcount /inputdata /outputdata
6	C:\hadoop\sbin>hadoop dfs -cat /outputdata/*
7	C:\hadoop\sbin>stop-all.cmd
Find the average age of the people (both male and female) who died in the titanic tragedy using Hadoop MapReduce in cludxlab.	
	Perform it in CloudxLab or Windows