BIG DATA LAB

Name: S.L.A.Laisha

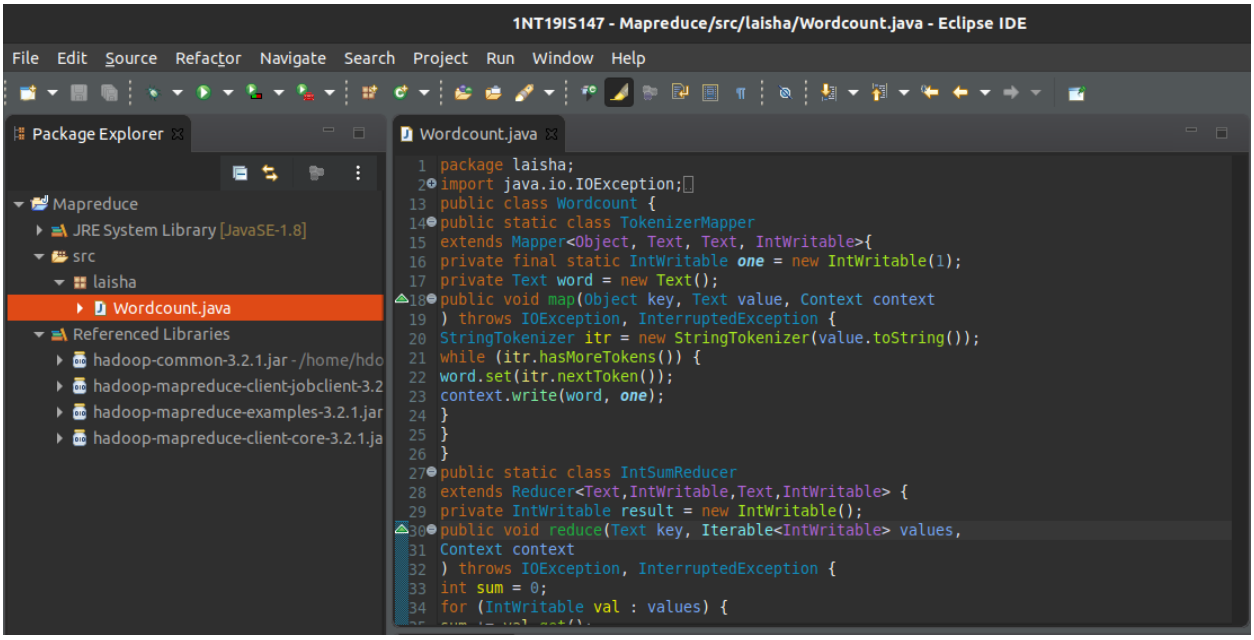
USN: 1NT19IS147

Date:14.06.2022

Exercise-3: MAPREDUCE

Use the Hadoop framework to write a custom MapReduce program to perform word count operation on a custom data set.

Initially create a new project, package and class in eclipse to run a java code.
To install jar files:
Right click on project (Mapreduce)
Click on -> build a path -> add external archives -> Hadoop 3.2.1 -> share
In share 1. Click on common ->open hadoop-common-3.2.1.jar
2. Click on mapreduce -> open hadoop-mapreduce-client-core-3.2.1.jar



Right click on project -> export -> java -> jar file -> next
Browse the address of the java file and save it in desktop/document/downloads and name it (laisha.jar)

IN TERMINAL:

Run the commands:

cd $HADOOP_HOME
cd sbin
jps
start-all.sh

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~$ cd $HADOOP_HOME
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1$ cd sbin
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ jps
4292 org.eclipse.equinox.launcher_1.5.600.v20191014-2022.jar
5096 Jps
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [admin1-HP-280-G4-MT-Business-PC]
Starting resourcemanager
Starting nodemanagers
```

hdfs dfs -mkdir -p ~ /input
hdfs dfs -appendToFile - ~/input/text.txt
Create a file and add content to it. ->(ctrl D two times)

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -mkdir -p ~ /input
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -appendToFile - ~ /input/test.txt
appendToFile: stdin (-) must be the sole input argument when present
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -appendToFile - ~/input/test.txt
My name is Laisha and I am studying in NMIT2022-06-17 09:27:34,714 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = f
alse, remoteHostTrusted = false
```

hadoop jar /home/hdoop/Desktop/laisha.jar ~/input ~/out

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hadoop jar /home/hdoop/Desktop/laisha.jar ~/input ~/out
2022-06-17 09:29:50,969 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-17 09:29:51,230 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
 your application with ToolRunner to remedy this.
2022-06-17 09:29:51,247 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_16554379493
08_0001
2022-06-17 09:29:51,336 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 09:29:51,462 INFO input.FileInputFormat: Total input files to process : 1
2022-06-17 09:29:51,529 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 09:29:51,572 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 09:29:51,588 INFO mapreduce.JobSubmitter: number of splits:1
2022-06-17 09:29:51,705 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2022-06-17 09:29:52,122 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1655437949308_0001
2022-06-17 09:29:52,122 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-17 09:29:52,284 INFO conf.Configuration: resource-types.xml not found
2022-06-17 09:29:52,284 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-17 09:29:52,448 INFO impl.YarnClientImpl: Submitted application application_1655437949308_0001
2022-06-17 09:29:52,477 INFO mapreduce.Job: The url to track the job: http://admin1-HP-280-G4-MT-Business-PC:8088/proxy/application_1655437949308_0001
/
```

hdfs dfs -cat ~/out/part*

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -cat ~/out/part*
2022-06-17 09:33:26,116 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
I       1
Laisha  1
My      1
NMIT    1
am      1
and     1
in      1
is      1
name    1
```

The word count of all the words in the file are 1.

Add more content to the file created.
Again run the above commands

hdfs dfs -appendToFile - ~/input/test.txt

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -appendToFile - ~/input/test.txt

hello NMIT hello laisha2022-06-17 09:35:48,162 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrus
ted = false
```

hadoop jar /home/hdoop/Desktop/laisha.jar

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$  hadoop jar /home/hdoop/Desktop/laisha.jar
Exception in thread "main" java.lang.ArrayIndexOutOfBoundsException: 0
        at laisha.Wordcount.main(Wordcount.java:50)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
```

hadoop jar /home/hdoop/Desktop/laisha.jar ~/input ~/output

// use a new output dir when u append content to existing file

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hadoop jar /home/hdoop/Desktop/laisha.jar ~/input ~/output
2022-06-17 09:36:42,043 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-17 09:36:42,274 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute
 your application with ToolRunner to remedy this.
2022-06-17 09:36:42,298 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_16554379493
08_0002
```

// u can see map and reduce to be 100% in the picture

```
2022-06-17 09:36:43,336 INFO mapreduce.Job: Running job: job_1655437949308_0002
2022-06-17 09:36:47,413 INFO mapreduce.Job: Job job_1655437949308_0002 running in uber mode : false
2022-06-17 09:36:47,414 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-17 09:36:51,513 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-17 09:36:55,545 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-17 09:36:55,563 INFO mapreduce.Job: Job job_1655437949308_0002 completed successfully
2022-06-17 09:36:55,643 INFO mapreduce.Job: Counters: 54
```

hdfs dfs -cat ~/output/part*

U can see the wordcount of the words in file

```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -cat ~/output/part*
2022-06-17 09:37:34,081 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
I       1
Laisha  1
My      1
NMIT    2
am      1
and     1
hello   2
in      1
is      1
laisha  1
name    1
```