Name: Aditya Wandhekar **SRN: 202201449**
Roll no. 34 **Div : A(A2)**

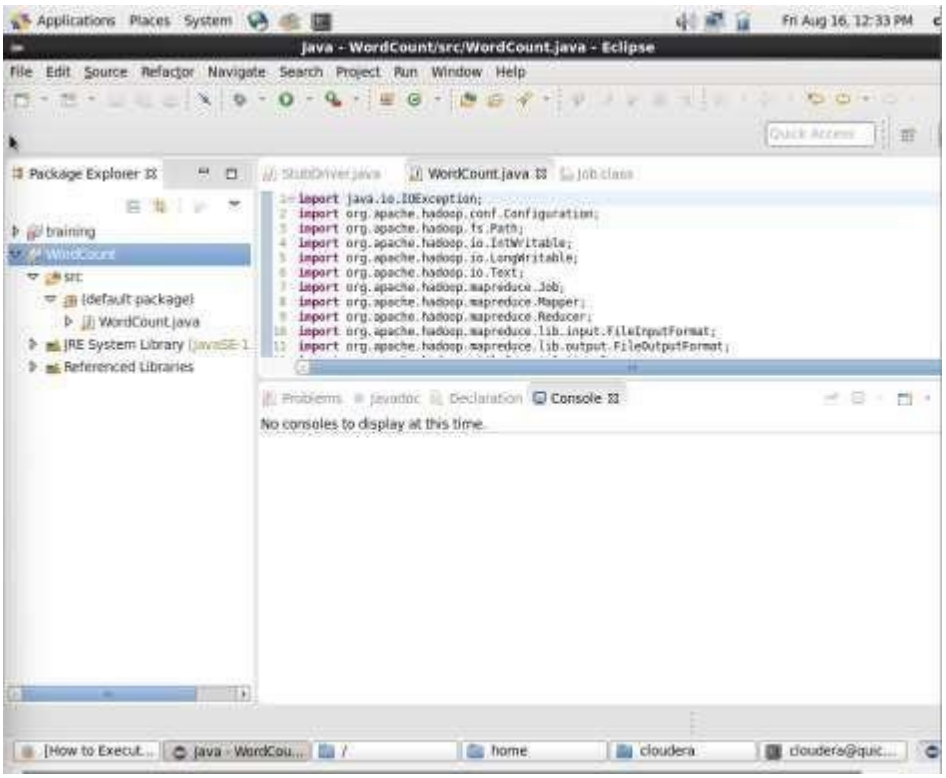
BDA Assignment 3

PS: Word count using map reduce

Solution:

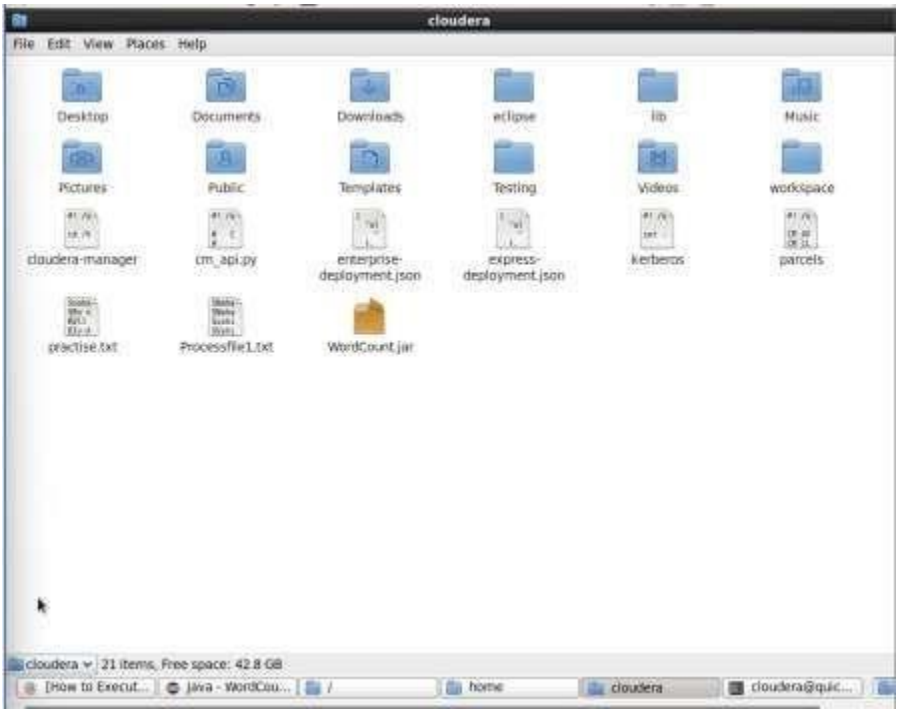
Step1: Make a WordCount.java file in eclipse and paste the code of the WordCount in that by making a new class named WordCount

- Also while making WordCount.java add external libraries as well from 2 locations
- Librarires->Add external jars->File system->user->lib->Hadoop (Slect all the jar file then add)
- Librarires->Add external jars->File system->user->->client->(Select all the jar files then add)



Step2: Export this .java file as .jar file by doing following steps:

-Export->Java->jarfile->browse(i.e the place where you want to extract the jar file/home/cloudera)



Step 3: Execute the following commands which include

-Making text file in local system by writing text in it

- Making a directory in the hdfs enviorment

-In the directory in the hdfs copying the text from local system to hdfs system

```
[cloudera@quickstart ~]$ hdfs dfs -ls
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - cloudera supergroup          0 2024-08-14 08:44 /data
drwxr-xr-x   - cloudera supergroup          0 2024-08-14 08:54 /data_copy
drwxr-xr-x   - hbase    supergroup          0 2024-08-16 11:47 /hbase
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs     supergroup          0 2024-08-05 21:06 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /inputfolder
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Processfile1.txt /inputfol
der/
[cloudera@quickstart ~]$ hdfs dfs -cat /inputfolder/Processfile.txt
cat: `/inputfolder/Processfile.txt': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cat /inputfolder/Processfile1.txt
SNeha Thakkar
SNeha Thakkar
Sushil THakkar
SUshil THakkar
CLoudera
Cloudera
Cloddera
Big Data
Big Data
CLoudera
Sneha Thakkar
```

-Execution of how does it calculate the word count using the jar file:

```
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount /inpu
tfolder/Processfile1.txt /out1
24/08/16 12:25:54 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
24/08/16 12:25:56 INFO input.FileInputFormat: Total input paths to process : 1
24/08/16 12:25:56 INFO mapreduce.JobSubmitter: number of splits:1
24/08/16 12:25:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_17
23834051029_0001
24/08/16 12:25:59 INFO impl.YarnClientImpl: Submitted application application_17
23834051029_0001
24/08/16 12:25:59 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1723834051029_0001/
24/08/16 12:25:59 INFO mapreduce.Job: Running job: job_1723834051029_0001
24/08/16 12:26:39 INFO mapreduce.Job: Job job_1723834051029_0001 running in uber
 mode : false
24/08/16 12:26:39 INFO mapreduce.Job:  map 0% reduce 0%
24/08/16 12:27:06 INFO mapreduce.Job:  map 100% reduce 0%
24/08/16 12:27:18 INFO mapreduce.Job:  map 100% reduce 100%
24/08/16 12:27:20 INFO mapreduce.Job: Job job_1723834051029_0001 completed successfully
24/08/16 12:27:21 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=240
                FILE: Number of bytes written=287187
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=251
                HDFS: Number of bytes written=62
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=22932
                Total time spent by all reduces in occupied slots (ms)=10948
```

```
        Total time spent by all map tasks (ms)=22932
        Total time spent by all reduce tasks (ms)=10948
        Total vcore-milliseconds taken by all map tasks=22932
        Total vcore-milliseconds taken by all reduce tasks=10948
        Total megabyte-milliseconds taken by all map tasks=23482368
        Total megabyte-milliseconds taken by all reduce tasks=11210752
    Map-Reduce Framework
        Map input records=11
        Map output records=18
        Map output bytes=198
        Map output materialized bytes=240
        Input split bytes=125
        Combine input records=0
        Combine output records=0
        Reduce input groups=7
        Reduce shuffle bytes=240
        Reduce input records=18
        Reduce output records=7
        Spilled Records=36
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=383
        CPU time spent (ms)=1890
        Physical memory (bytes) snapshot=401895424
        Virtual memory (bytes) snapshot=3125907456
        Total committed heap usage (bytes)=287834112
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=120
```

```
    File Output Format Counters
            Bytes Written=62
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2024-08-16 12:27 /out1/_SUCCESS
-rw-r--r--   1 cloudera supergroup         62 2024-08-16 12:27 /out1/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
BIG     2
CLODDERA        1
CLOUDERA        3
DATA    2
SNEHA   3
SUSHIL  2
THAKKAR 5
[cloudera@quickstart ~]$ cat /home/cloudera/Processfile1.txt
SNeha Thakkar
SNeha Thakkar
Sushil THakkar
SUshil THakkar
CLoudera
Cloudera
Cloddera
Big Data
Big Data
CLoudera
Sneha Thakkar
[cloudera@quickstart ~]$
```

OUTPUT:

```
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
BIG     2
CLODDERA        1
CLOUDERA        3
DATA    2
SNEHA   3
SUSHIL  2
THAKKAR 5
[cloudera@quickstart ~]$ cat /home/cloudera/Processfile1.txt
SNeha Thakkar
SNeha Thakkar
Sushil THakkar
SUshil THakkar
CLoudera
Cloudera
Cloddera
Big Data
Big Data
CLoudera
Sneha Thakkar
```