732\_Saksham Shetkar

Big Data technology

MapReduce CharCount

**Practical : 3**

**Aim : To Implement CharCount problem using Hadoop MapReduce in Eclips.**

Step 1 : Run your cloudera system. Open Eclips.

Step 2 : Click on File > New > java project. Give Project Name (“CharCount”).

Step 3 : Click on Libraries tab. Then click on Add External JARs… Tab To add Hadoop Libraries.

Step 4 : Follow this steps :

Click on File System -> usr -> lib -> hadoop (Select all the libraries (JAR files) -> Click OK.

Again Click on Add External JARs… -> client -> select all jar files -> ok -> Finish.

Step 5 : Now you will see project name “CharCount” on sidebar.

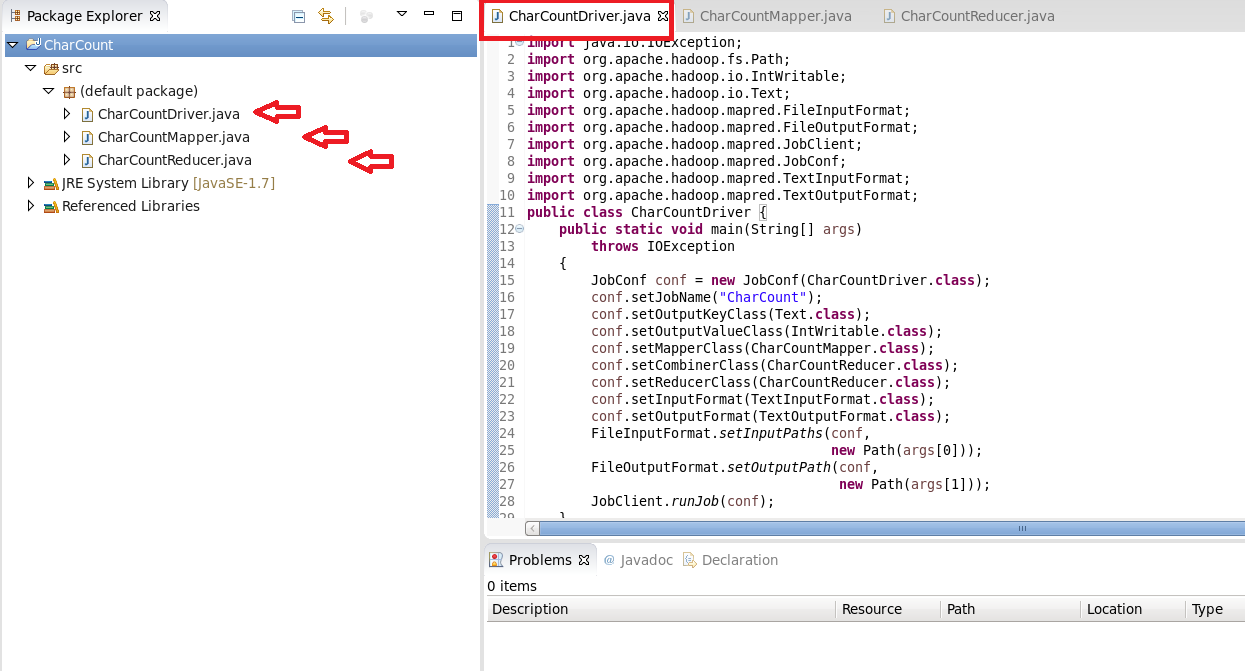
Right Click on Project name “CharCount” -> New -> class.

Create 3 new classes named :

**CharCountDriver**(having the main function),

**CharCountMapper**,

**CharCountReducer.**

****

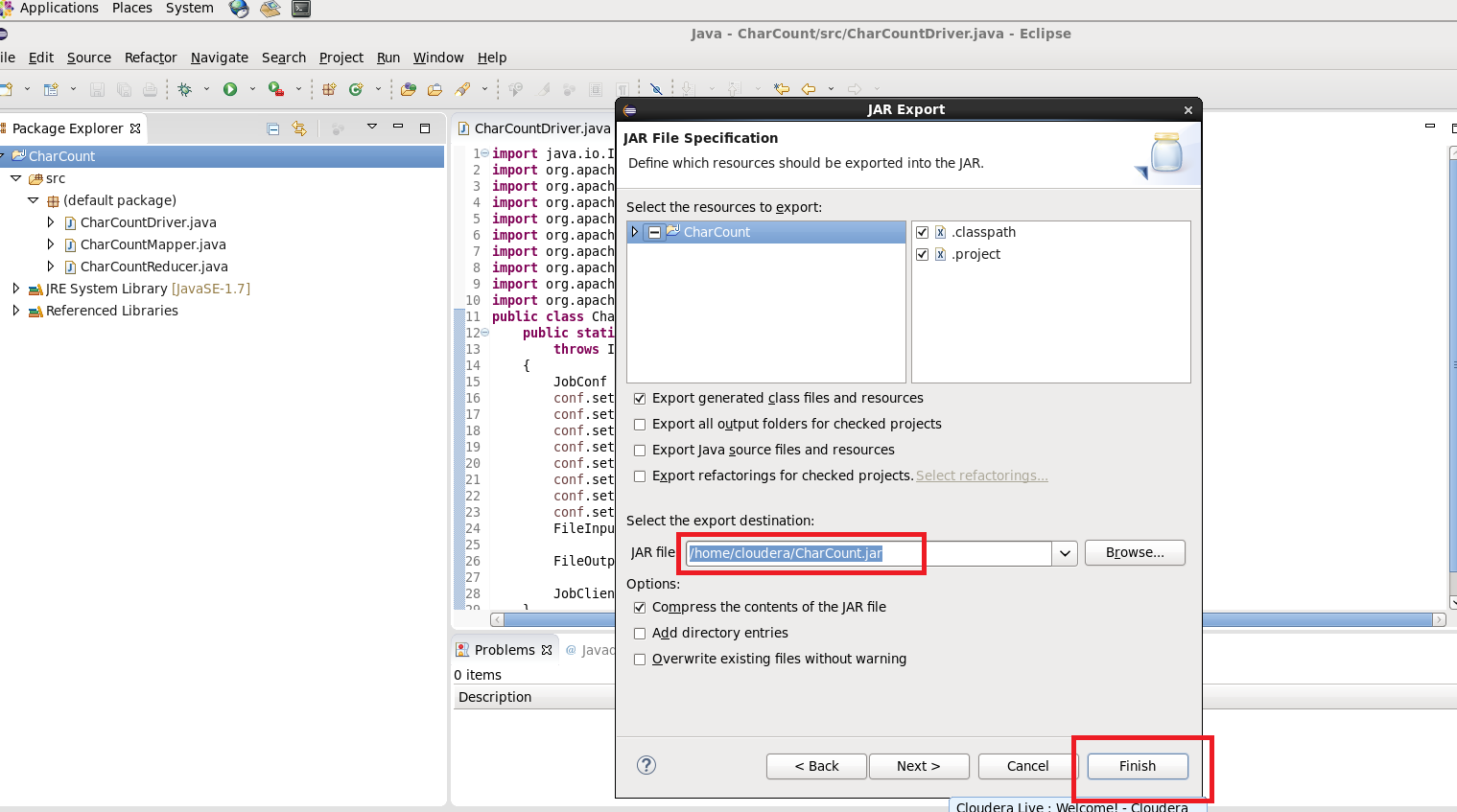
Step 6 : Write your code in java windows.

Step 7 : Right Click on the project name

CharCount -> Export -> Java -> JAR File -> Next ->

For select the export destination for JAR file:

browse -> Name : CharCount.jar -> save in folder -> cloudera -> Finish -> OK



Step 8 : Now open terminal . follow this commands :

[cloudera@quickstart ~]$ ls

abc.txt CharCount cloudera-manager Desktop Downloads

enterprise-deployment.json kerberos Music Pictures Templates

WordCloud.jar

abc.txt~ CharCount.jar cm\_api.py Documents eclipse

express-deployment.json lib parcels Public Videos

workspace

[cloudera@quickstart ~]$ pwd

/home/cloudera

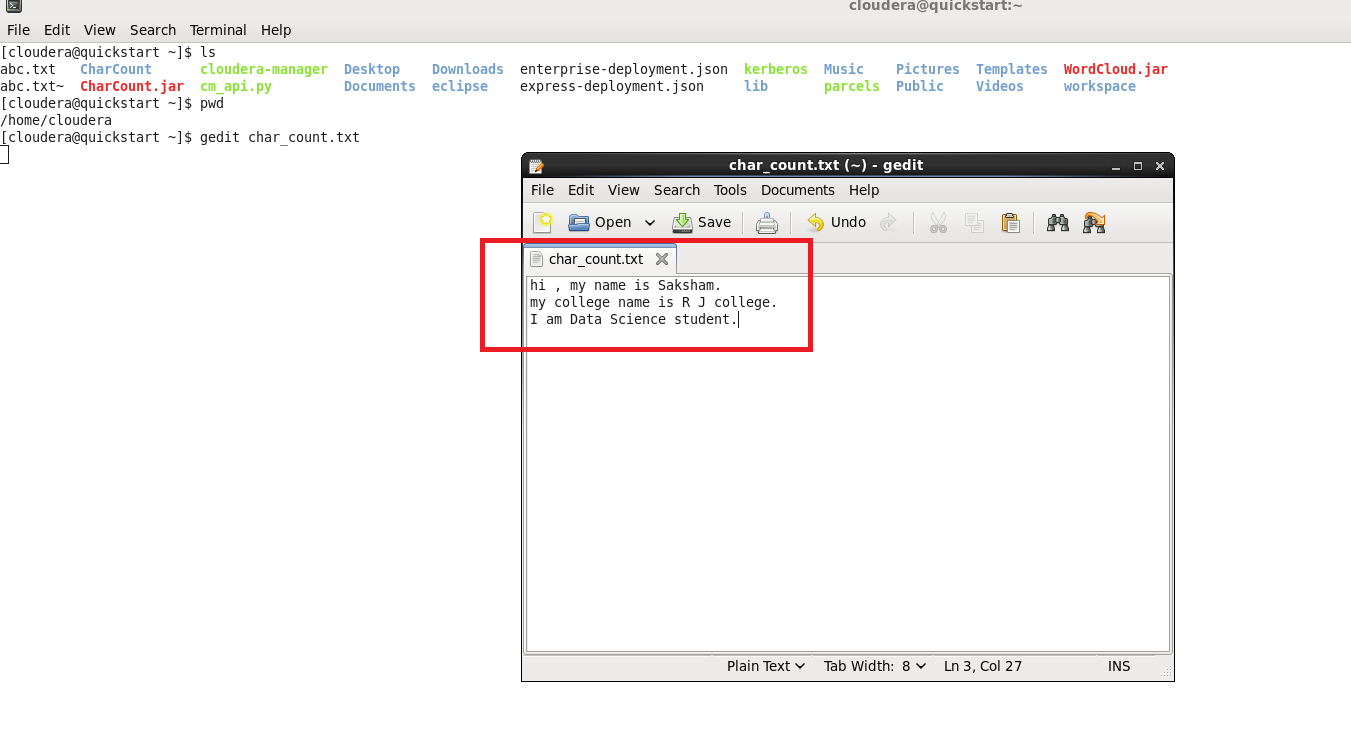
[cloudera@quickstart ~]$ gedit char\_count.txt

[cloudera@quickstart ~]$ cat char\_count.txt

hi , my name is Saksham.

my college name is R J college.

I am Data Science student.



[cloudera@quickstart ~]$ hdfs dfs -ls /

Found 8 items

drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks

drwxr-xr-x - hbase supergroup 0 2023-01-19 05:20 /hbase

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:05 /inputdir

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:14 /outputdir

drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr

drwxrwxrwt - hdfs supergroup 0 2023-01-19 05:20 /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 /char\_inputdir

[cloudera@quickstart ~]$ hdfs dfs -ls /

Found 9 items

drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks

drwxr-xr-x - cloudera supergroup 0 2023-01-24 07:32 /char\_inputdir

drwxr-xr-x - hbase supergroup 0 2023-01-19 05:20 /hbase

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:05 /inputdir

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:14 /outputdir

drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr

drwxrwxrwt - hdfs supergroup 0 2023-01-19 05:20 /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 -put /home/cloudera/char\_count.txt /char\_inputdir/

[cloudera@quickstart ~]$ hdfs dfs -ls /char\_inputdir

Found 1 items

-rw-r--r-- 1 cloudera supergroup 84 2023-01-24 07:34 /char\_inputdir/char\_count.txt

[cloudera@quickstart ~]$ hdfs dfs -cat /char\_inputdir/char\_count.txt

hi , my name is Saksham.

my college name is R J college.

I am Data Science student.

[cloudera@quickstart ~]$ hadoop jar /home/cloudera/CharCount.jar CharCountDriver /char\_inputdir/char\_count.txt /char\_outputdir

*-----It will start executing the program.---------*

*File System Counters*

*FILE: Number of bytes read=292*

*FILE: Number of bytes written=431746*

*FILE: Number of read operations=0*

*FILE: Number of large read operations=0*

*FILE: Number of write operations=0*

*HDFS: Number of bytes read=350*

*HDFS: Number of bytes written=100*

*HDFS: Number of read operations=9*

*HDFS: Number of large read operations=0*

*HDFS: Number of write operations=2*

*Job Counters*

*Launched map tasks=2*

*Launched reduce tasks=1*

*Data-local map tasks=2*

*Total time spent by all maps in occupied slots (ms)=38087*

*Total time spent by all reduces in occupied slots (ms)=7773*

*Total time spent by all map tasks (ms)=38087*

*Total time spent by all reduce tasks (ms)=7773*

*Total vcore-milliseconds taken by all map tasks=38087*

*Total vcore-milliseconds taken by all reduce tasks=7773*

*Total megabyte-milliseconds taken by all map tasks=39001088*

*Total megabyte-milliseconds taken by all reduce tasks=7959552*

*Map-Reduce Framework*

*Map input records=3*

*Map output records=84*

*Map output bytes=501*

*Map output materialized bytes=298*

*Input split bytes=224*

*Combine input records=84*

*Combine output records=36*

*Reduce input groups=25*

*Reduce shuffle bytes=298*

*Reduce input records=36*

*Reduce output records=25*

*Spilled Records=72*

*Shuffled Maps =2*

*Failed Shuffles=0*

*Merged Map outputs=2*

*GC time elapsed (ms)=1921*

*CPU time spent (ms)=4060*

*Physical memory (bytes) snapshot=911654912*

*Virtual memory (bytes) snapshot=4693946368*

*Total committed heap usage (bytes)=704643072*

*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=126*

*File Output Format Counters*

*Bytes Written=100*

[cloudera@quickstart ~]$ hdfs dfs -ls /

Found 10 items

drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks

drwxr-xr-x - cloudera supergroup 0 2023-01-24 07:34 /char\_inputdir

drwxr-xr-x - cloudera supergroup 0 2023-01-24 07:45 /char\_outputdir

drwxr-xr-x - hbase supergroup 0 2023-01-19 05:20 /hbase

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:05 /inputdir

drwxr-xr-x - cloudera supergroup 0 2023-01-19 06:14 /outputdir

drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr

drwxrwxrwt - hdfs supergroup 0 2023-01-19 05:20 /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 -ls /char\_outputdir

Found 2 items

-rw-r--r-- 1 cloudera supergroup 0 2023-01-24 07:45 /char\_outputdir/\_SUCCESS

-rw-r--r-- 1 cloudera supergroup 100 2023-01-24 07:45 /char\_outputdir/part-00000

[cloudera@quickstart ~]$ hdfs dfs -cat /char\_outputdir/part-00000

**Output :-**

3

15

, 1

. 3

D 1

I 1

J 1

R 1

S 2

a 7

c 4

d 1

e 9

g 2

h 2

i 4

k 1

l 4

m 6

n 4

o 2

s 4

t 3

u 1

y 2

[cloudera@quickstart ~]$