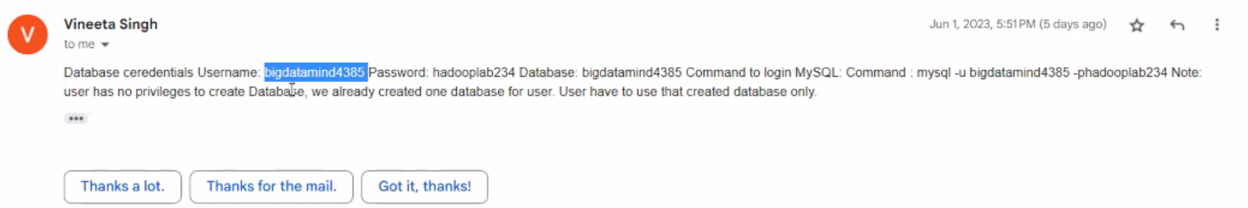
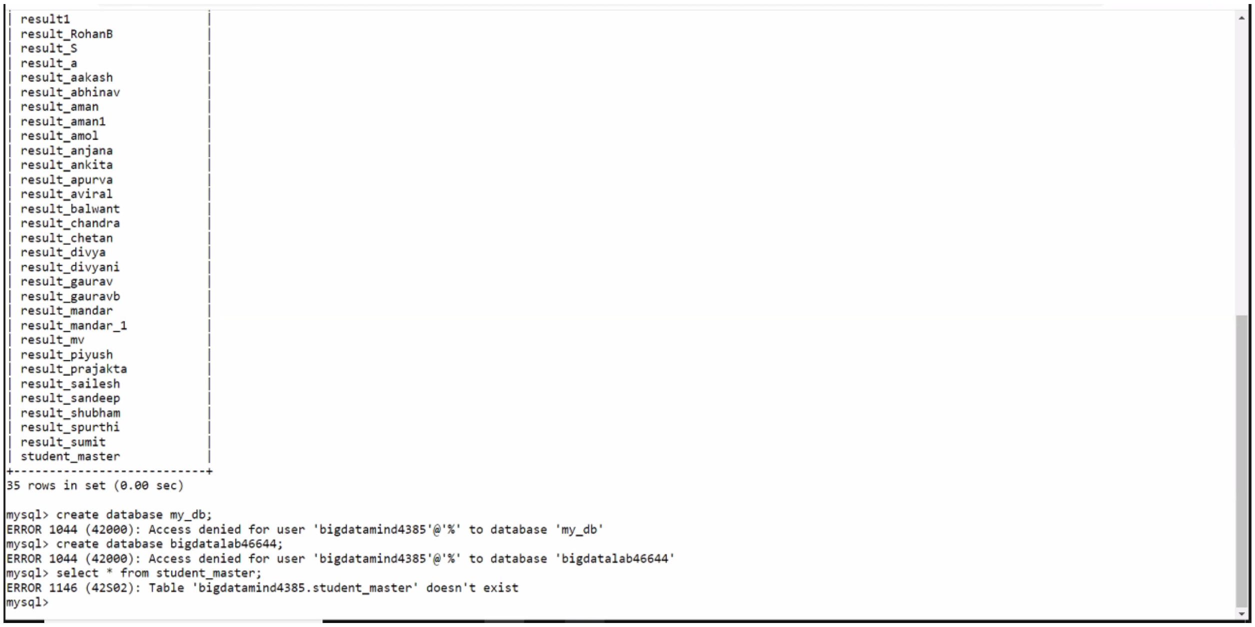
→ Spark with SQL



Vineeta Singh
to me

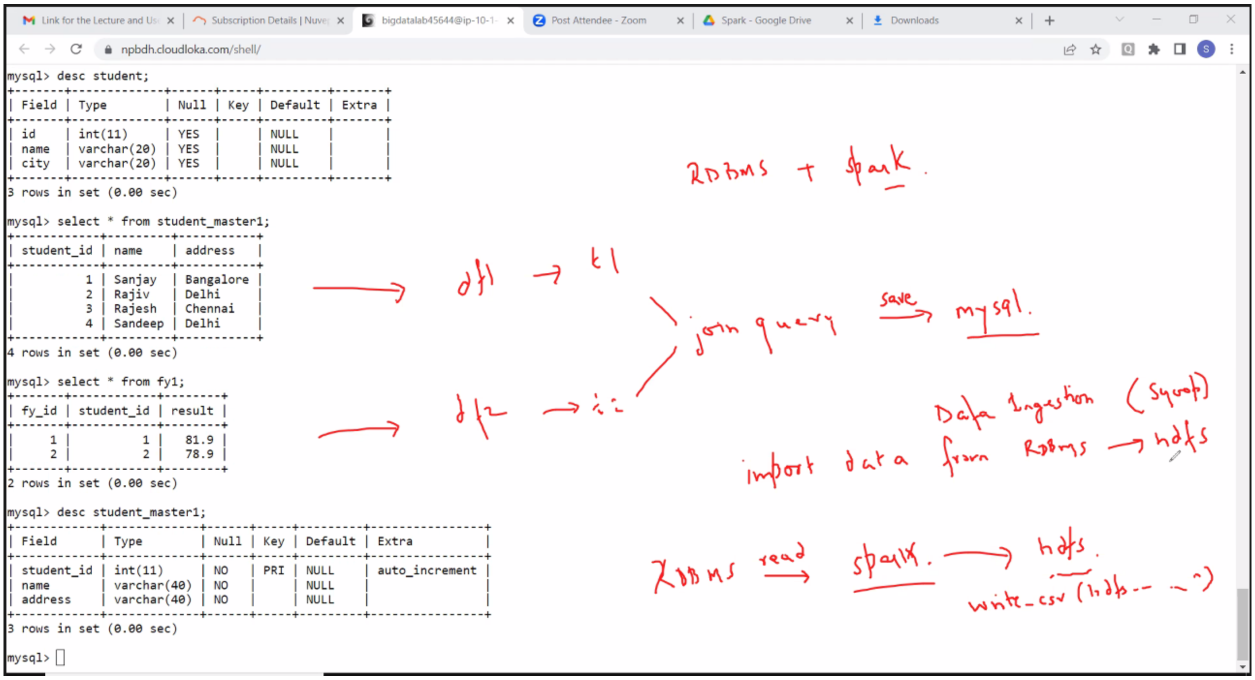
Jun 1, 2023, 5:51 PM (5 days ago)

Database ceredentials Username: bigdatamind4385 Password: hadooplab234 Database: bigdatamind4385 Command to login MySQL: Command : mysql -u bigdatamind4385 -phadooplab234 Note: user has no privileges to create Database, we already created one database for user. User have to use that created database only.

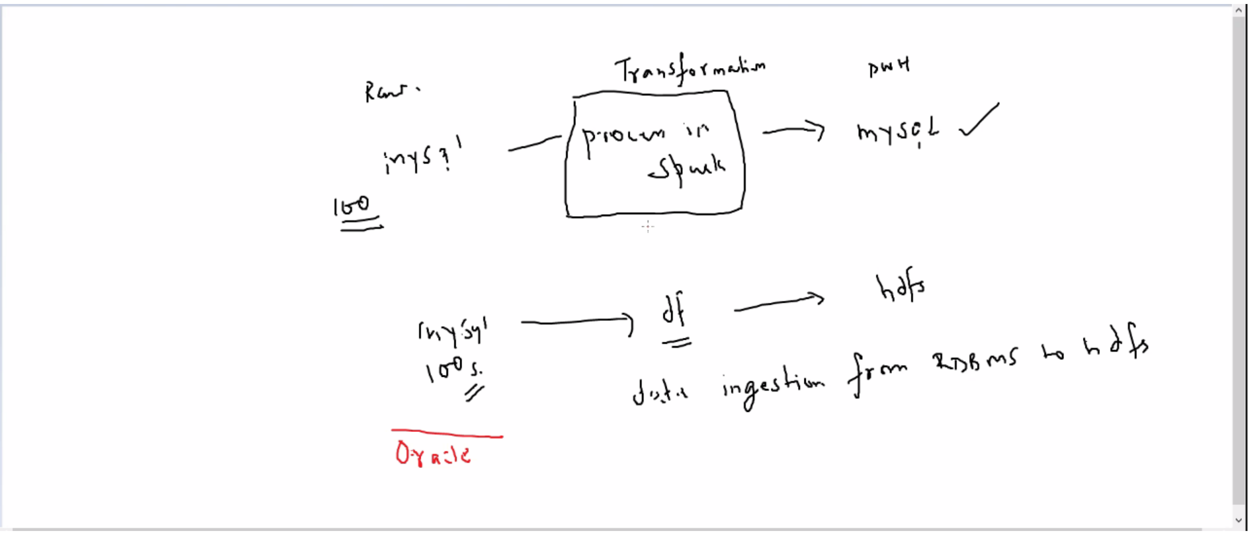


```
| result1                  |
| result_RohanB            |
| result_S                 |
| result_a                 |
| result_aakash            |
| result_abhinav           |
| result_aman              |
| result_aman1             |
| result_amol              |
| result_anjana            |
| result_ankita            |
| result_apurva            |
| result_aviral            |
| result_balwant           |
| result_chandra           |
| result_chetan            |
| result_divya             |
| result_divyani           |
| result_gaurav            |
| result_gauravb           |
| result_mandar            |
| result_mandar_1          |
| result_mv                |
| result_piyush            |
| result_prajakta          |
| result_sailesh           |
| result_sandeep           |
| result_shubham           |
| result_spurthi           |
| result_sumit             |
| student_master           |
+--------------------------+
35 rows in set (0.00 sec)

mysql> create database my_db;
ERROR 1044 (42000): Access denied for user 'bigdatamind4385'@'%' to database 'my_db'
mysql> create database bigdatalab46644;
ERROR 1044 (42000): Access denied for user 'bigdatamind4385'@'%' to database 'bigdatalab46644'
mysql> select * from student_master;
ERROR 1146 (42S02): Table 'bigdatamind4385.student_master' doesn't exist
mysql>
```

For local linux based
User: root
Password: password


Link for the Lecture and Us
Subscription Details | Nuve
bigdatalab45644@ip-10-1-
Post Attendee - Zoom
Spark - Google Drive
Downloads
npbdh.cloudloka.com/shell/

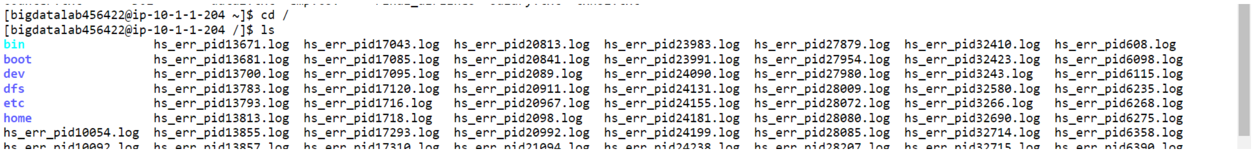
```
mysql> desc student;
+-------+-------------+------+-----+---------+-------+
| Field | Type        | Null | Key | Default | Extra |
+-------+-------------+------+-----+---------+-------+
| id    | int(11)     | YES  |     | NULL    |       |
| name  | varchar(20) | YES  |     | NULL    |       |
| city  | varchar(20) | YES  |     | NULL    |       |
+-------+-------------+------+-----+---------+-------+
3 rows in set (0.00 sec)

mysql> select * from student_master1;
+------------+---------+-----------+
| student_id | name    | address   |
+------------+---------+-----------+
|          1 | Sanjay  | Bangalore |
|          2 | Rajiv   | Delhi     |
|          3 | Rajesh  | Chennai   |
|          4 | Sandeep | Delhi     |
+------------+---------+-----------+
4 rows in set (0.00 sec)

mysql> select * from fy1;
+-------+------------+--------+
| fy_id | student_id | result |
+-------+------------+--------+
|     1 |          1 |   81.9 |
|     2 |          2 |   78.9 |
+-------+------------+--------+
2 rows in set (0.00 sec)

mysql> desc student_master1;
+------------+-------------+------+-----+---------+----------------+
| Field      | Type        | Null | Key | Default | Extra          |
+------------+-------------+------+-----+---------+----------------+
| student_id | int(11)     | NO   | PRI | NULL    | auto_increment |
| name       | varchar(40) | NO   |     | NULL    |                |
| address    | varchar(40) | NO   |     | NULL    |                |
+------------+-------------+------+-----+---------+----------------+
3 rows in set (0.00 sec)

mysql> 
```



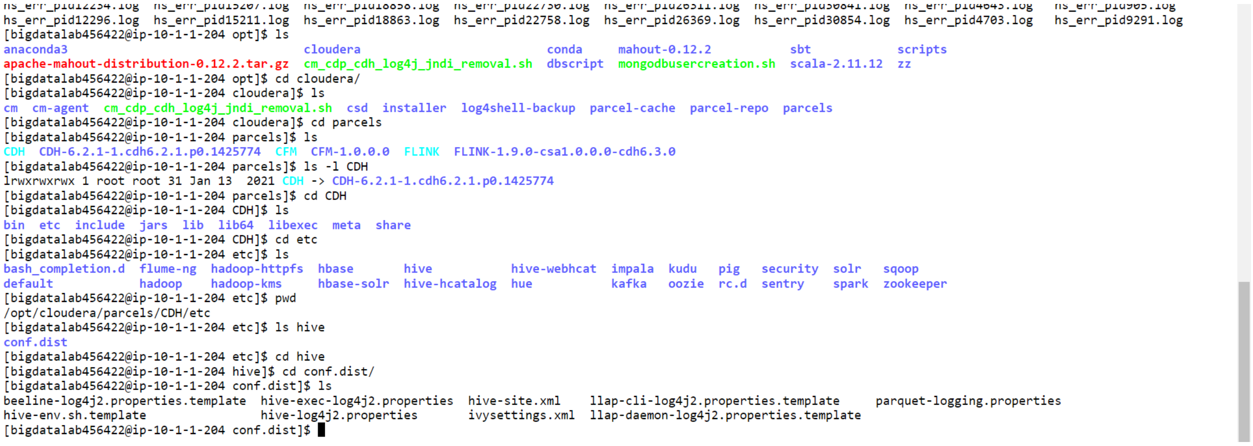
Path for local

`/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/etc/hive/conf.dist/hive-site.xml`

1



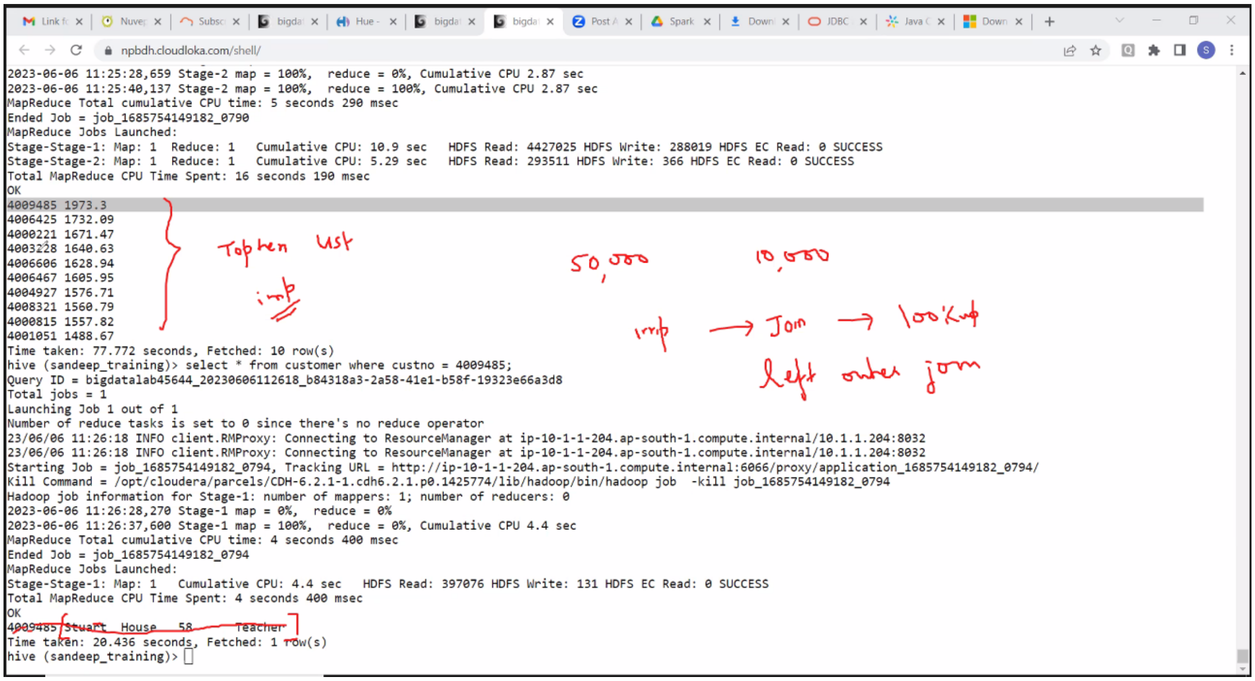
```
[bigdatalab456422@ip-10-1-1-204 ~]$ cd /
[bigdatalab456422@ip-10-1-1-204 /]$ ls
bin                  hs_err_pid13671.log  hs_err_pid17043.log  hs_err_pid20813.log  hs_err_pid23983.log  hs_err_pid27879.log  hs_err_pid32410.log  hs_err_pid608.log
boot                 hs_err_pid13681.log  hs_err_pid17085.log  hs_err_pid20841.log  hs_err_pid23991.log  hs_err_pid27954.log  hs_err_pid32423.log  hs_err_pid6098.log
dev                  hs_err_pid13700.log  hs_err_pid17095.log  hs_err_pid2089.log   hs_err_pid24090.log  hs_err_pid27980.log  hs_err_pid3243.log   hs_err_pid6115.log
dfs                  hs_err_pid13783.log  hs_err_pid17120.log  hs_err_pid20911.log  hs_err_pid24131.log  hs_err_pid28009.log  hs_err_pid32580.log  hs_err_pid6235.log
etc                  hs_err_pid13793.log  hs_err_pid1716.log   hs_err_pid20967.log  hs_err_pid24155.log  hs_err_pid28072.log  hs_err_pid3266.log   hs_err_pid6268.log
home                 hs_err_pid13813.log  hs_err_pid1718.log   hs_err_pid2098.log   hs_err_pid24181.log  hs_err_pid28080.log  hs_err_pid32690.log  hs_err_pid6275.log
hs_err_pid10054.log  hs_err_pid13855.log  hs_err_pid17293.log  hs_err_pid20992.log  hs_err_pid24199.log  hs_err_pid28085.log  hs_err_pid32714.log  hs_err_pid6358.log
hs_err_pid10092.log  hs_err_pid13857.log  hs_err_pid17310.log  hs_err_pid21094.log  hs_err_pid24238.log  hs_err_pid28207.log  hs_err_pid32715.log  hs_err_pid6390.log
```



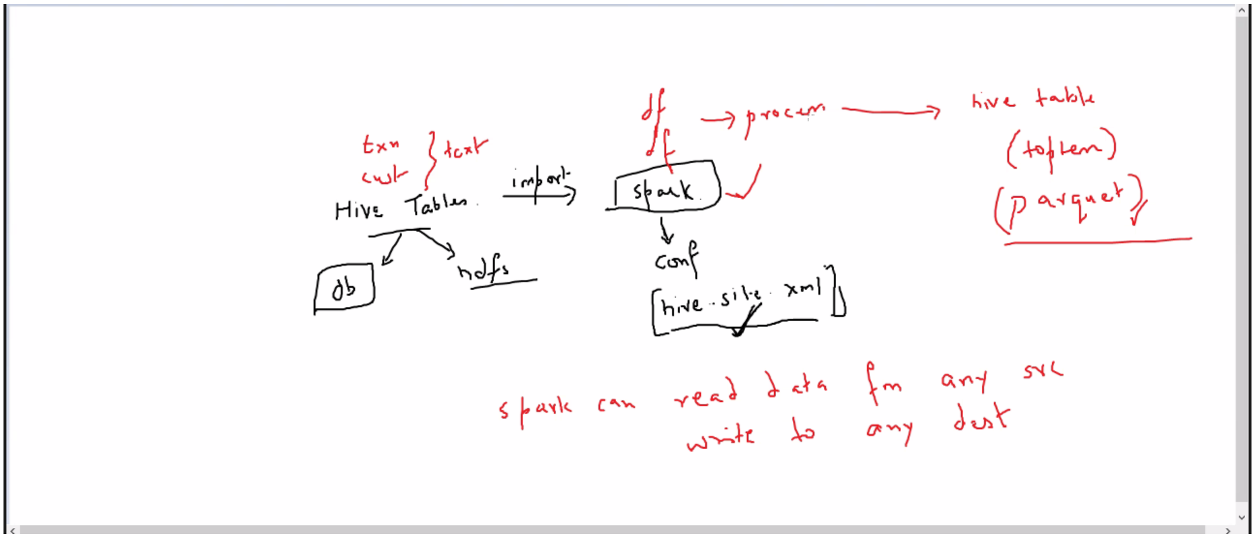
```
hs_err_pid12254.log  hs_err_pid15207.log  hs_err_pid18858.log  hs_err_pid22750.log  hs_err_pid26311.log  hs_err_pid30841.log  hs_err_pid4645.log   hs_err_pid965.log
hs_err_pid12296.log  hs_err_pid15211.log  hs_err_pid18863.log  hs_err_pid22758.log  hs_err_pid26369.log  hs_err_pid30854.log  hs_err_pid4703.log   hs_err_pid9291.log
[bigdatalab456422@ip-10-1-1-204 opt]$ ls
anaconda3                          cloudera                          conda     mahout-0.12.2           sbt           scripts
apache-mahout-distribution-0.12.2.tar.gz  cm_cdp_cdh_log4j_jndi_removal.sh  dbscript  mongodbusercreation.sh  scala-2.11.12  zz
[bigdatalab456422@ip-10-1-1-204 opt]$ cd cloudera/
[bigdatalab456422@ip-10-1-1-204 cloudera]$ ls
cm  cm-agent  cm_cdp_cdh_log4j_jndi_removal.sh  csd  installer  log4shell-backup  parcel-cache  parcel-repo  parcels
[bigdatalab456422@ip-10-1-1-204 cloudera]$ cd parcels
[bigdatalab456422@ip-10-1-1-204 parcels]$ ls
CDH  CDH-6.2.1-1.cdh6.2.1.p0.1425774  CFM  CFM-1.0.0.0  FLINK  FLINK-1.9.0-csa1.0.0.0-cdh6.3.0
[bigdatalab456422@ip-10-1-1-204 parcels]$ ls -l CDH
lrwxrwxrwx 1 root root 31 Jan 13  2021 CDH -> CDH-6.2.1-1.cdh6.2.1.p0.1425774
[bigdatalab456422@ip-10-1-1-204 parcels]$ cd CDH
[bigdatalab456422@ip-10-1-1-204 CDH]$ ls
bin  etc  include  jars  lib  lib64  libexec  meta  share
[bigdatalab456422@ip-10-1-1-204 CDH]$ cd etc
[bigdatalab456422@ip-10-1-1-204 etc]$ ls
bash_completion.d  flume-ng  hadoop-httpfs  hbase       hive            hive-webhcat  impala  kudu   pig   security  solr   sqoop
default            hadoop    hadoop-kms     hbase-solr  hive-hcatalog   hue           kafka   oozie  rc.d  sentry    spark  zookeeper
[bigdatalab456422@ip-10-1-1-204 etc]$ pwd
/opt/cloudera/parcels/CDH/etc
[bigdatalab456422@ip-10-1-1-204 etc]$ ls hive
conf.dist
[bigdatalab456422@ip-10-1-1-204 etc]$ cd hive
[bigdatalab456422@ip-10-1-1-204 hive]$ cd conf.dist/
[bigdatalab456422@ip-10-1-1-204 conf.dist]$ ls
beeline-log4j2.properties.template  hive-exec-log4j2.properties  hive-site.xml    llap-cli-log4j2.properties.template     parquet-logging.properties
hive-env.sh.template                hive-log4j2.properties       ivysettings.xml  llap-daemon-log4j2.properties.template
[bigdatalab456422@ip-10-1-1-204 conf.dist]$ 
```

→

While joining an important data with some other lookup table , always use left outer join so that you don’t lose any data



```
2023-06-06 11:25:28,659 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.87 sec
2023-06-06 11:25:40,137 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.87 sec
MapReduce Total cumulative CPU time: 5 seconds 290 msec
Ended Job = job_1685754149182_0790
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.9 sec   HDFS Read: 4427025 HDFS Write: 288019 HDFS EC Read: 0 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.29 sec   HDFS Read: 293511 HDFS Write: 366 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 190 msec
OK
4009485 1973.3
4006425 1732.09
4000221 1671.47
4003228 1640.63
4006606 1628.94
4006467 1605.95
4004927 1576.71
4008321 1560.79
4000815 1557.82
4001051 1488.67
Time taken: 77.772 seconds, Fetched: 10 row(s)
hive (sandeep_training)> select * from customer where custno = 4009485;
Query ID = bigdatalab45644_20230606112618_b84318a3-2a58-41e1-b58f-19323e66a3d8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
23/06/06 11:26:18 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/06/06 11:26:18 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
Starting Job = job_1685754149182_0794, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1685754149182_0794/
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job  -kill job_1685754149182_0794
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2023-06-06 11:26:28,270 Stage-1 map = 0%,  reduce = 0%
2023-06-06 11:26:37,600 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.4 sec
MapReduce Total cumulative CPU time: 4 seconds 400 msec
Ended Job = job_1685754149182_0794
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.4 sec   HDFS Read: 397076 HDFS Write: 131 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 400 msec
OK
4009485 Stuart  House   58      Teacher
Time taken: 20.436 seconds, Fetched: 1 row(s)
hive (sandeep_training)> 
```



spark can read data frm any src
write to any dest