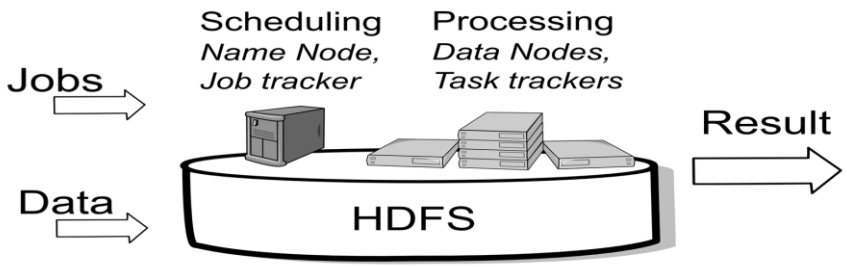
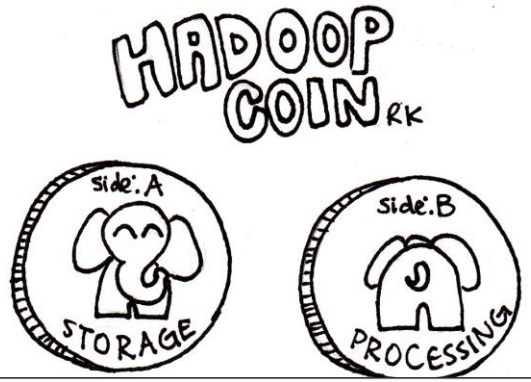
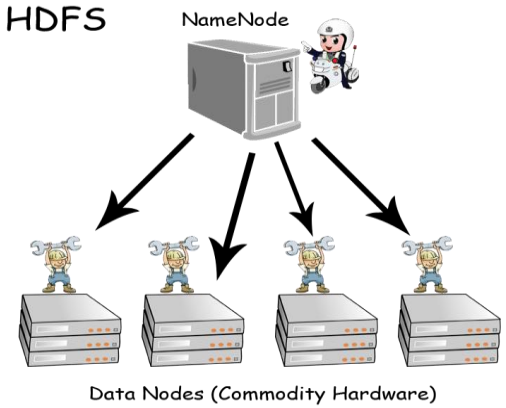
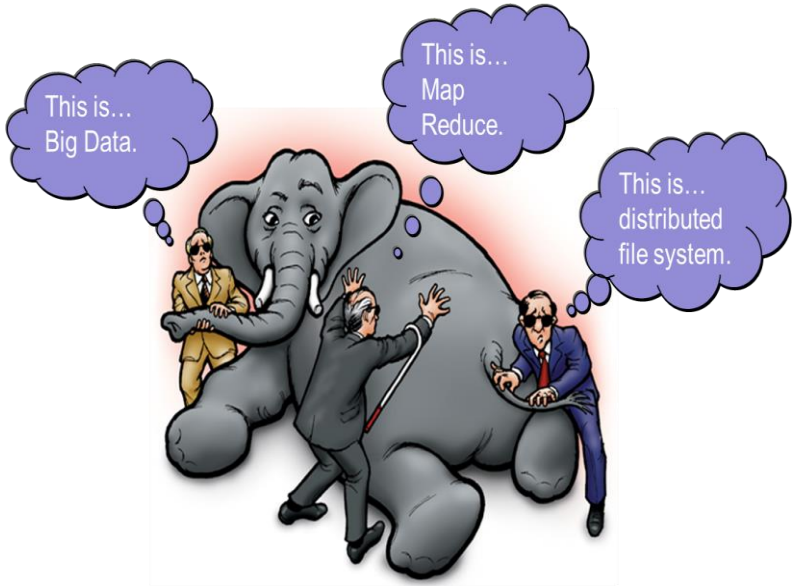





**Vipin Gupta**
BE,RHCSS,RHCE,CEH,CCNA,MCSE,MCSA
vipin2411@gmail.com
Mobile: 93563-10379
www.linuxexpert.in

# Hive

# Hive Interface

```
[root@quickstart ~]# hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
Time taken: 10.73 seconds, Fetched: 1 row(s)
hive> create database record;
OK
Time taken: 0.538 seconds
hive> show databases;
OK
default
record
Time taken: 0.044 seconds, Fetched: 2 row(s)
hive> use record;
OK
Time taken: 0.034 seconds
```

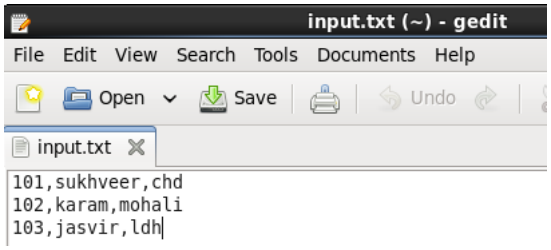
# Creating Table

```
hive> create table student(rno int,name string,city string)row format delimited fields terminated by ',';
OK
Time taken: 0.911 seconds
hive> show tables;
OK
student
Time taken: 0.067 seconds, Fetched: 1 row(s)
```

```
[root@quickstart ~]# hadoop fs -ls /user/hive/warehouse/record.db
Found 1 items
drwxrwxrwx   - root supergroup          0 2019-07-20 19:31 /user/hive/warehouse/record.db/student
[root@quickstart ~]#
```

# Loading Data from Local Shell into Hive Table

```
hive> LOAD DATA LOCAL INPATH '/root/input.txt' INTO TABLE student;
Loading data to table record.student
Table record.student stats: [numFiles=1, totalSize=49]
OK
Time taken: 1.579 seconds
hive>
    > select * from student;
OK
101     sukhveer        chd
102     karam   mohali
103     jasvir  ldh
Time taken: 0.194 seconds, Fetched: 3 row(s)
```



# Describe Table

```
hive> describe student;
OK
rno                     int
name                    string
city                    string
Time taken: 0.301 seconds, Fetched: 3 row(s)
hive> describe extended student;
OK
rno                     int
name                    string
city                    string

Detailed Table Information      Table(tableName:student, dbName:record, owner:root, createTime:1563675836,
 lastAccessTime:0, retention:0, sd:StorageDescriptor(cols:[FieldSchema(name:rno, type:int, comment:null),
FieldSchema(name:name, type:string, comment:null), FieldSchema(name:city, type:string, comment:null)], loc
ation:hdfs://quickstart.cloudera:8020/user/hive/warehouse/record.db/student, inputFormat:org.apache.hadoop
.mapred.TextInputFormat, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat, compress
ed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.serde2.laz
y.LazySimpleSerDe, parameters:{serialization.format=,, field.delim=,}), bucketCols:[], sortCols:[], parame
ters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), stor
edAsSubDirectories:false), partitionKeys:[], parameters:{numFiles=1, transient_lastDdlTime=1563676299, COL
UMN_STATS_ACCURATE=true, totalSize=49}, viewOriginalText:null, viewExpandedText:null, tableType:MANAGED_TA
BLE)
```

# Creating External Table

```
[root@quickstart ~]# hadoop fs -mkdir /hive_demo/
[root@quickstart ~]# hadoop fs -put external-input.txt /hive_demo/
```

```
[root@quickstart ~]# hadoop fs -ls /hive_demo/
Found 1 items
-rw-r--r--   1 root supergroup         44 2019-07-20 20:30 /hive_demo/external-input.txt
[root@quickstart ~]# hadoop fs -cat /hive_demo/external-input.txt
1,priya,jalandhar
2,riya,amritsar
3,ram,ldh
```

```
hive> create external table student_external(rno int,name string,city string)row format delimited fields t
erminated by ',' LOCATION '/hive_demo/';
OK
Time taken: 0.127 seconds
```

```
hive> select * from student_external;
OK
1       priya   jalandhar
2       riya    amritsar
3       ram     ldh
Time taken: 0.129 seconds, Fetched: 3 row(s)
```

```
[root@quickstart ~]# hadoop fs -put input2.txt /hive_demo/
[root@quickstart ~]# hadoop fs -put input.txt /hive_demo/
[root@quickstart ~]# hadoop fs -ls /hive_demo/
Found 4 items
-rw-r--r--   1 root supergroup         44 2019-07-20 20:30 /hive_demo/external-input.txt
-rw-r--r--   1 root supergroup         49 2019-07-20 22:42 /hive_demo/input.txt
-rw-r--r--   1 root supergroup         38 2019-07-20 22:35 /hive_demo/input1.txt
-rw-r--r--   1 root supergroup         35 2019-07-20 22:41 /hive_demo/input2.txt
[root@quickstart ~]# cat input.txt
101,sukhveer,chd
102,karam,mohali
103,jasvir,ldh
[root@quickstart ~]# cat input1.txt
104, v,panchkula
105,jagseer,moga
[root@quickstart ~]# cat input2.txt
104, karan,panchkula
105,jiya,moga
[root@quickstart ~]# vi input1.txt
[root@quickstart ~]# vi external-input.txt
[root@quickstart ~]# vi input.txt
[root@quickstart ~]# hadoop fs -rm /hive_demo/input1.txt
Deleted /hive_demo/input1.txt
[root@quickstart ~]# cp input2.txt input3.txt
[root@quickstart ~]# vi input3.txt
[root@quickstart ~]# hadoop fs -put input3.txt /hive_demo/
```

```
hive> select * from student_external;
OK
1       priya   jalandhar
2       riya    amritsar
3       ram     ldh
101     sukhveer        chd
102     karam   mohali
103     jasvir  ldh
104      karan  panchkula
105     jiya    moga
104     raman   panchkula
105     aman    moga
Time taken: 0.113 seconds, Fetched: 10 row(s)
hive> select * from student_external where rno=104;
OK
104      karan  panchkula
104     raman   panchkula
```

# Partition

- Partition means dividing a table into a coarse grained parts based on the value of a partition column such as a date. This make it faster to do queries on slice of the data.
- It define how data is stored in HDFS.
- Grouping data bases on some column
- **Buckets**
  - Partitions divided further into buckets bases on some other column

# Partition

```
[cloudera@quickstart ~]$ sudo hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database university
    > ;
OK
Time taken: 4.339 seconds
hive> use university;
OK
Time taken: 0.118 seconds
hive> create table students(rno int,name string,branch string,section string)row format delimited fields terminated by ',';
OK
Time taken: 0.971 seconds
hive> describe students;
OK
rno                     int
name                    string
branch                  string
section                 string
Time taken: 0.56 seconds, Fetched: 4 row(s)
```

# Partition

```
[cloudera@quickstart ~]$ sudo hadoop fs -ls /user/hive/warehouse/
Found 3 items
drwxrwxrwx   - root supergroup          0 2019-07-21 03:15 /user/hive/warehouse/myretail.db
drwxrwxrwx   - root supergroup          0 2019-07-20 20:38 /user/hive/warehouse/record.db
drwxrwxrwx   - root supergroup          0 2019-07-22 07:34 /user/hive/warehouse/university.db
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo hadoop fs -ls /user/hive/warehouse/university.db
Found 1 items
drwxrwxrwx   - root supergroup          0 2019-07-22 07:34 /user/hive/warehouse/university.db/students
[cloudera@quickstart ~]$ gedit input-data.txt
[cloudera@quickstart ~]$ cat input-data.txt
1,priya,cse,c1
2,riya,cse,c1
3,aman,cse,c2
4,raman,cse,c2
5,ram,ece,e1
6,sham,ece,e1
7,abhay,ece,e2
8,ajay,ece,e2
9,ritika,me,m1
10,karan,me,m1
11,himansu,me,m2
12,abhishek,me,m2
```

# Partition

```
hive> load data local inpath '/home/cloudera/input-data.txt' into table students;
Loading data to table university.students
Table university.students stats: [numFiles=1, totalSize=180]
OK
Time taken: 2.3 seconds
hive> select * from students;
OK
1       priya   cse     c1
2       riya    cse     c1
3       aman    cse     c2
4       raman   cse     c2
5       ram     ece     e1
6       sham    ece     e1
7       abhay   ece     e2
8       ajay    ece     e2
9       ritika  me      m1
10      karan   me      m1
11      himansu me      m2
12      abhishek        me      m2
```

# Creating Partition and Buckets

```
hive> create table studentsbybranch(rno int,name string,section string)partitioned by (branch string)clustered by (section) into 2 buckets row form
ed fields terminated by ',';
OK
Time taken: 0.137 seconds
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> set hive.exec.dynamic.partition=true;
hive> set hive.enforce.bucketing=true;
```

create table studentsbybranch(rno int,name string,section string)partitioned by (branch string)clustered by (section) into 2 buckets row format delimited fields terminated by ',';

# Load Data from original Table into Partition Table

```
  > from students s INSERT OVERWRITE TABLE studentsbybranch PARTITION(branch) select s.rno,s.name,s.section,s.branch DISTRIBUTE BY branch;
Query ID = root_20190722080000_0624fda7-588d-4ce1-80e9-83d425784fa6
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1563805409766_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1563805409766_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1563805409766_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-07-22 08:00:43,206 Stage-1 map = 0%,  reduce = 0%
2019-07-22 08:01:02,418 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.29 sec
2019-07-22 08:01:17,799 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.72 sec
MapReduce Total cumulative CPU time: 6 seconds 720 msec
Ended Job = job_1563805409766_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
```

# Partition in HDFS

```
[cloudera@quickstart ~]$ sudo hadoop fs -ls /user/hive/warehouse/university.db
Found 2 items
drwxrwxrwx   - root supergroup          0 2019-07-22 07:47 /user/hive/warehouse/university.db/students
drwxrwxrwx   - root supergroup          0 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo hadoop fs -ls /user/hive/warehouse/university.db/studentsbybranch
Found 4 items
drwxrwxrwx   - root supergroup          0 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=__HIVE_DEFAULT_PARTITION__
drwxrwxrwx   - root supergroup          0 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=cse
drwxrwxrwx   - root supergroup          0 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=ece
drwxrwxrwx   - root supergroup          0 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=me
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo hadoop fs -ls /user/hive/warehouse/university.db/studentsbybranch/branch=cse
Found 2 items
-rwxrwxrwx   1 root supergroup         21 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=cse/000000_0
-rwxrwxrwx   1 root supergroup         21 2019-07-22 08:02 /user/hive/warehouse/university.db/studentsbybranch/branch=cse/000001_0
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo hadoop fs -cat /user/hive/warehouse/university.db/studentsbybranch/branch=cse/000000_0
2,riya,c1
1,priya,c1
[cloudera@quickstart ~]$ sudo hadoop fs -cat /user/hive/warehouse/university.db/studentsbybranch/branch=cse/000001_0
3,aman,c2
4,raman,c2
```

```
hive> select * from studentsbybranch limit 6;
OK
NULL    NULL    NULL    __HIVE_DEFAULT_PARTITION__
2       riya    c1      cse
1       priya   c1      cse
3       aman    c2      cse
4       raman   c2      cse
5       ram     e1      ece
Time taken: 0.283 seconds, Fetched: 6 row(s)
```

# Joins

```
hive> create table employee(name string,salary float,city string) row format delimited fields terminated by ',';
OK
Time taken: 0.234 seconds
hive> load data local inpath '/home/cloudera/emp.txt' into table employee;
Loading data to table university.employee
Table university.employee stats: [numFiles=1, totalSize=64]
OK
Time taken: 0.409 seconds
hive> create table mailid(name string,email string) row format delimited fields terminated by ',';
OK
Time taken: 0.098 seconds
hive>
    > load data local inpath '/home/cloudera/email.txt' into table mailid;
Loading data to table university.mailid
Table university.mailid stats: [numFiles=1, totalSize=68]
OK
Time taken: 0.549 seconds
hive> select * from employee;
OK
abhishek        50000.0 chandigarh
karan   40000.0 ludhiana
aman    60000.0 delhi
Time taken: 0.159 seconds, Fetched: 3 row(s)
hive> select * from mailid;
OK
abhishek        abhishek@gmail.com
aman    aman@gmail.com
riya    riya@gmail.com
Time taken: 0.106 seconds, Fetched: 3 row(s)
```

```
[cloudera@quickstart ~]$ cat emp.txt
abhishek,50000,chandigarh
karan,40000,ludhiana
aman,60000,delhi
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ cat email.txt
abhishek,abhishek@gmail.com
aman,aman@gmail.com
riya,riya@gmail.com
```

# Equal Join

```
hive> select e.name,e.city,e.salary,m.email from employee e join mailid m on e.name=m.name;
Query ID = root_20190722082929_6ca64b04-3848-4832-bf4b-2e761ba27f25
Total jobs = 1
Execution log at: /tmp/root/root_20190722082929_6ca64b04-3848-4832-bf4b-2e761ba27f25.log
2019-07-22 08:29:57     Starting to launch local task to process map join;      maximum memory = 1013645312
2019-07-22 08:29:59     Dump the side-table for tag: 0 with group count: 3 into file: file:/tmp/root/3c70ae99-a8db-4c97-aadd-0ec419775e28/hive_20
19-07-22_08-29-40_720_5877268676930452613-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.hashtable
2019-07-22 08:30:00     Uploaded 1 File to: file:/tmp/root/3c70ae99-a8db-4c97-aadd-0ec419775e28/hive_2019-07-22_08-29-40_720_587726867693045261 3-
1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.hashtable (372 bytes)
2019-07-22 08:30:00     End of local task; Time Taken: 2.967 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1563805409766_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1563805409766_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1563805409766_0003
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2019-07-22 08:30:17,585 Stage-3 map = 0%,  reduce = 0%
2019-07-22 08:30:33,708 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
MapReduce Total cumulative CPU time: 2 seconds 890 msec
Ended Job = job_1563805409766_0003
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.89 sec   HDFS Read: 6641 HDFS Write: 81 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 890 msec
OK
abhishek        chandigarh      50000.0 abhishek@gmail.com
aman    delhi   60000.0 aman@gmail.com
Time taken: 54.134 seconds, Fetched: 2 row(s)
```

# Left Outer Join

```
hive> select e.name,e.city,e.salary,m.email from employee e left outer join mailid m on e.name=m.name;
Query ID = root_20190722083232_9e8bae24-fbdd-4d82-aab3-e6d6ff0323b5
Total jobs = 1
Execution log at: /tmp/root/root_20190722083232_9e8bae24-fbdd-4d82-aab3-e6d6ff0323b5.log
2019-07-22 08:32:58     Starting to launch local task to process map join;      maximum memory = 1013645312
2019-07-22 08:33:00     Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/root/3c70ae99-a8db-4c97-aadd-0ec419775e28/hive_20
19-07-22_08-32-45_968_1274500128372120373-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile11--.hashtable
2019-07-22 08:33:00     Uploaded 1 File to: file:/tmp/root/3c70ae99-a8db-4c97-aadd-0ec419775e28/hive_2019-07-22_08-32-45_968_1274500128372120373-
1/-local-10003/HashTable-Stage-3/MapJoin-mapfile11--.hashtable (382 bytes)
2019-07-22 08:33:00     End of local task; Time Taken: 2.452 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1563805409766_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1563805409766_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1563805409766_0004
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2019-07-22 08:33:17,751 Stage-3 map = 0%,  reduce = 0%
2019-07-22 08:33:30,424 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.44 sec
MapReduce Total cumulative CPU time: 2 seconds 440 msec
Ended Job = job_1563805409766_0004
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.44 sec   HDFS Read: 6608 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 440 msec
OK
abhishek        chandigarh      50000.0 abhishek@gmail.com
karan   ludhiana        40000.0 NULL
aman    delhi   60000.0 aman@gmail.com
Time taken: 46.75 seconds, Fetched: 3 row(s)
```

# Full Outer Join

```
    > select e.name,e.city,e.salary,m.email from employee e full outer join mailid m on e.name=m.name;
Query ID = root_20190722083434_d77cfd5a-1beb-4afd-8462-db57f60b8046
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1563805409766_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1563805409766_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1563805409766_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-07-22 08:35:00,712 Stage-1 map = 0%,  reduce = 0%
2019-07-22 08:35:32,164 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 4.78 sec
2019-07-22 08:35:33,291 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.21 sec
2019-07-22 08:35:49,919 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.49 sec
MapReduce Total cumulative CPU time: 12 seconds 490 msec
Ended Job = job_1563805409766_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 12.49 sec   HDFS Read: 13436 HDFS Write: 131 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 490 msec
OK
abhishek        chandigarh      50000.0 abhishek@gmail.com
aman    delhi   60000.0 aman@gmail.com
karan   ludhiana        40000.0 NULL
NULL    NULL    NULL    riya@gmail.com
Time taken: 67.799 seconds, Fetched: 4 row(s)
```