NITTE EDUCATION TRUST | NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

ACADEMIC YEAR 2020-2021

KNOWLEDGE • CHARACTER • UNITY

BIGDATA LABORATORY

Report on,
Learning Activity II-Programming Assignment

Submitted by,
Aayesha Nomani (1NT18IS003)

Submitted to,
Mrs. Disha D N,
Assistant Professor,
Department of Information Science and Engineering NITTE Meenakshi Institute of Technology Bangalore-064

Mr. Mahesh Kumar,
Assistant Professor,
Department of Information Science and Engineering NITTE Meenakshi Institute of Technology Bangalore-064

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(An autonomous institution with A+ Grade by NAAC /UGC, Affiliated to Visvesvaraya Technological University, Belgaum, Approved by UGC/AICTE/Govt. of Karnataka)
Yelahanka, Bengaluru-560064

Table of Contents

LIST OF FIGURES

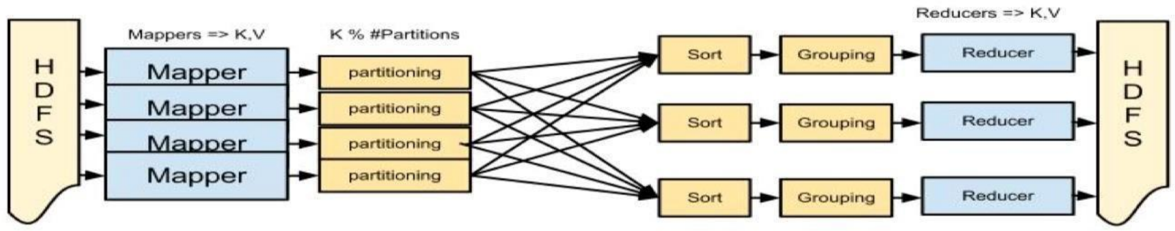
Brief note on Hadoop and Map Reduce

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput.



The MapReduce Pipeline

A mapper receives (Key, Value) & outputs (Key, Value)
A reducer receives (Key, Iterable[Value]) and outputs (Key, Value)
Partitioning / Sorting / Grouping provides the Iterable[Value] & Scaling

Hadoop Map-reduce Problem statement

Exercise-I

Create a dataset in excel as .csv file and it should contain the following fields with at least 20 sample datasets in it.

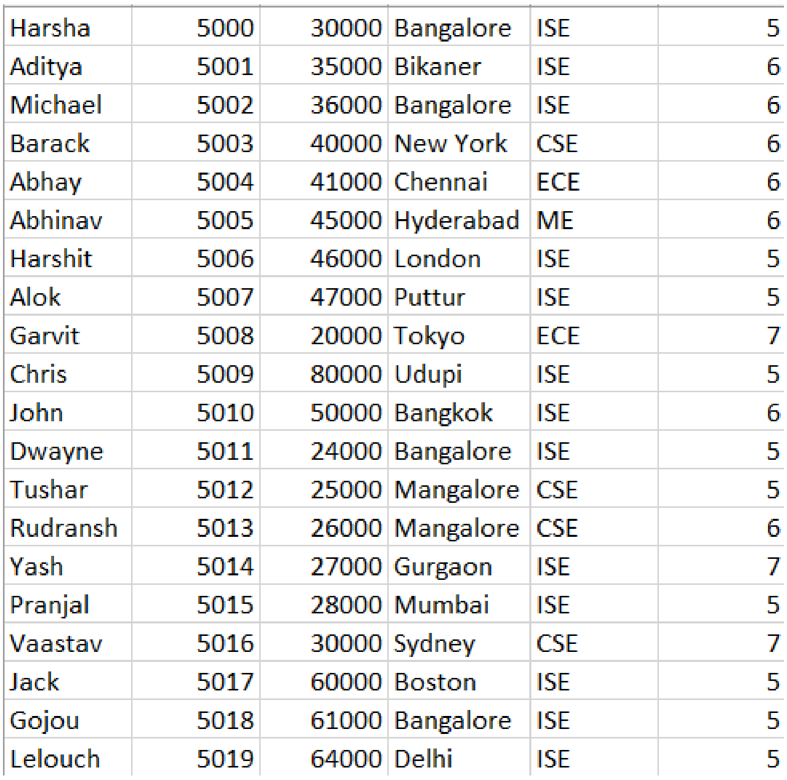
Name	SSN	Salary	Address	Dname	Experience
Harsha	5000	30000	Bangalore	ISE	5

Use the Hadoop MapReduce programming framework to come up with a Program which will take the data from this .csv file and computes the following.

1. Total number of employees who work in ISE department
2. Total number of employees with experience=5 years
3. Count the number of employees who lives in Bangalore.

Dataset Description

LA2.csv



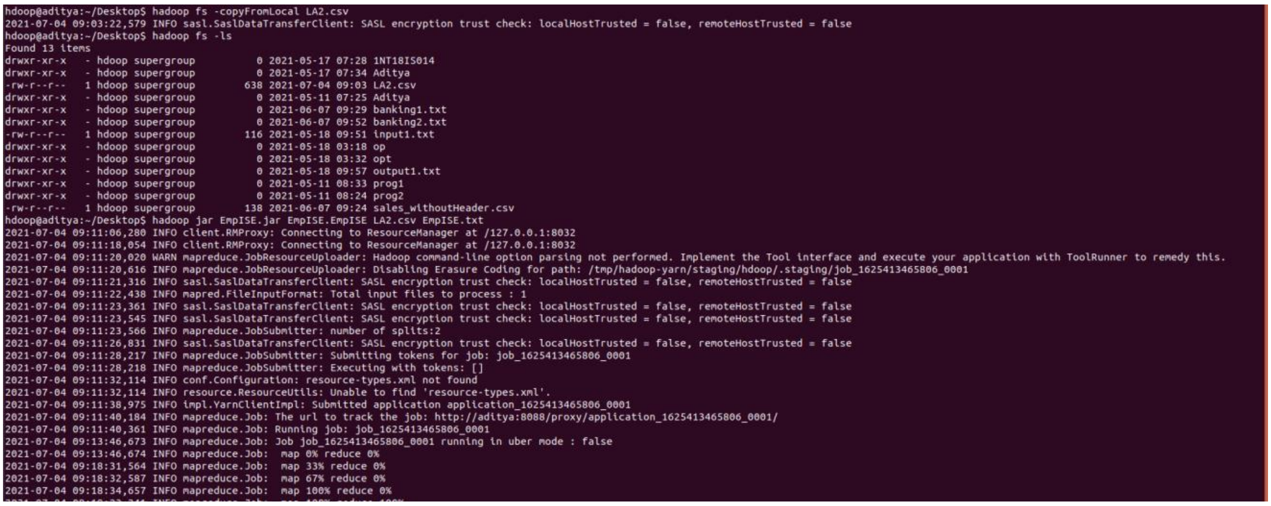
Harsha	5000	30000	Bangalore	ISE	5
Aditya	5001	35000	Bikaner	ISE	6
Michael	5002	36000	Bangalore	ISE	6
Barack	5003	40000	New York	CSE	6
Abhay	5004	41000	Chennai	ECE	6
Abhinav	5005	45000	Hyderabad	ME	6
Harshit	5006	46000	London	ISE	5
Alok	5007	47000	Puttur	ISE	5
Garvit	5008	20000	Tokyo	ECE	7
Chris	5009	80000	Udupi	ISE	5
John	5010	50000	Bangkok	ISE	6
Dwayne	5011	24000	Bangalore	ISE	5
Tushar	5012	25000	Mangalore	CSE	5
Rudransh	5013	26000	Mangalore	CSE	6
Yash	5014	27000	Gurgaon	ISE	7
Pranjal	5015	28000	Mumbai	ISE	5
Vaastav	5016	30000	Sydney	CSE	7
Jack	5017	60000	Boston	ISE	5
Gojou	5018	61000	Bangalore	ISE	5
Lelouch	5019	64000	Delhi	ISE	5

Source Code

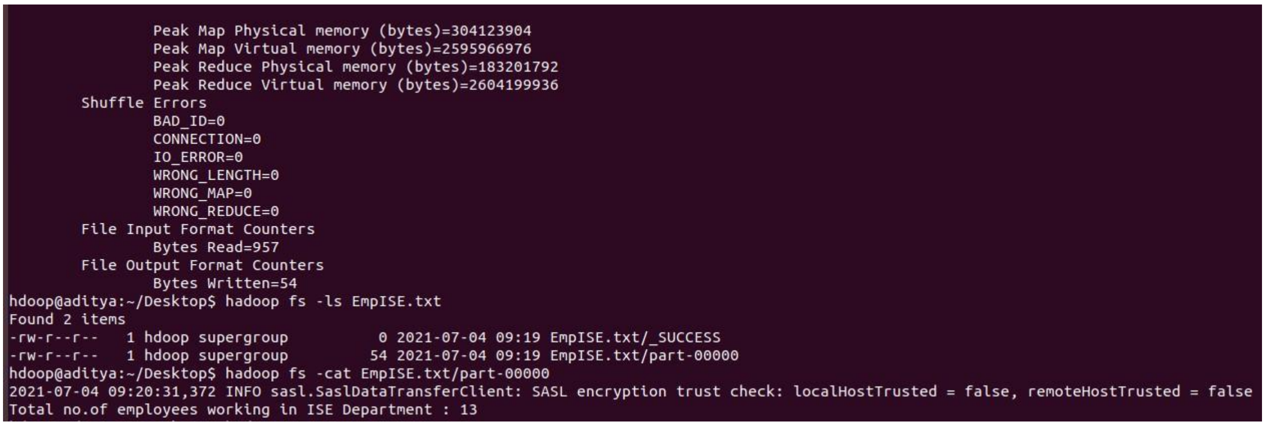
https://github.com/aayeshanomani/1NT18IS003_aayesha_A_bdLab/tree/master/BD%20LA%202

Results and Snapshot (Hadoop Map-reduce Programming)

1. Total number of employees who work in ISE department

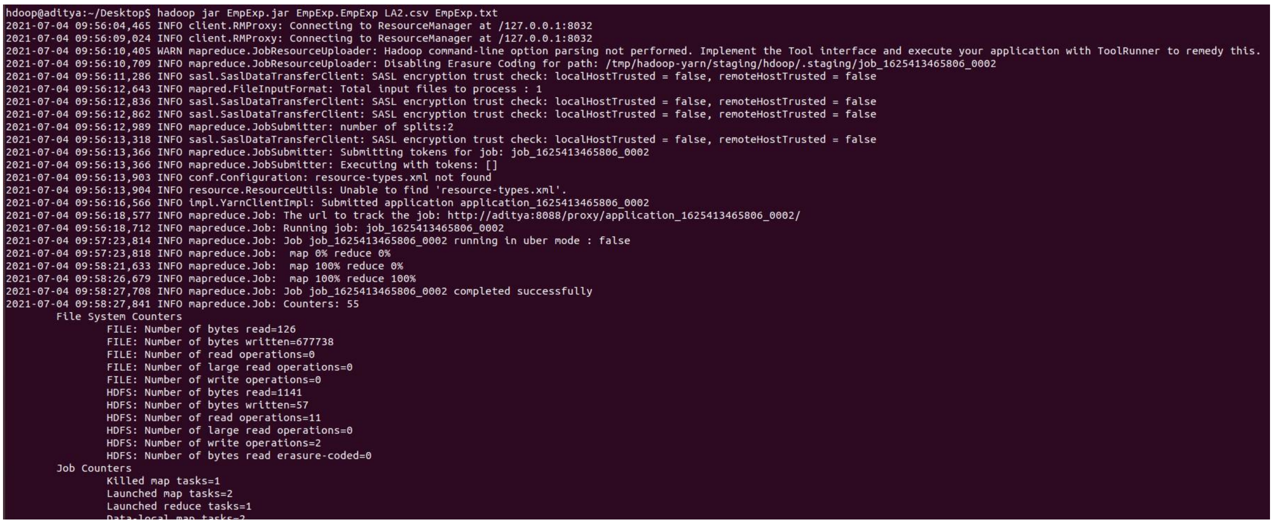


```
hdoop@aditya:~/Desktop$ hadoop fs -copyFromLocal LA2.csv
2021-07-04 09:03:22,579 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hdoop@aditya:~/Desktop$ hadoop fs -ls
Found 13 items
drwxr-xr-x   - hdoop supergroup          0 2021-05-17 07:28 1NT18IS014
drwxr-xr-x   - hdoop supergroup          0 2021-05-17 07:34 Aditya
-rw-r--r--   1 hdoop supergroup        638 2021-07-04 09:03 LA2.csv
drwxr-xr-x   - hdoop supergroup          0 2021-05-11 07:25 Aditya
drwxr-xr-x   - hdoop supergroup          0 2021-06-07 09:29 banking1.txt
drwxr-xr-x   - hdoop supergroup          0 2021-06-07 09:52 banking2.txt
-rw-r--r--   1 hdoop supergroup        116 2021-05-18 09:51 input1.txt
drwxr-xr-x   - hdoop supergroup          0 2021-05-18 03:18 op
drwxr-xr-x   - hdoop supergroup          0 2021-05-18 03:32 opt
drwxr-xr-x   - hdoop supergroup          0 2021-05-18 09:57 output1.txt
drwxr-xr-x   - hdoop supergroup          0 2021-05-11 08:33 prog1
drwxr-xr-x   - hdoop supergroup          0 2021-05-11 08:24 prog2
-rw-r--r--   1 hdoop supergroup        138 2021-06-07 09:24 sales_withoutHeader.csv
hdoop@aditya:~/Desktop$ hadoop jar EmpISE.jar EmpISE.EmpISE LA2.csv EmpISE.txt
2021-07-04 09:11:06,280 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 09:11:18,054 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 09:11:20,020 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-07-04 09:11:20,616 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1625413465806_0001
2021-07-04 09:11:21,316 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:11:22,438 INFO mapred.FileInputFormat: Total input files to process : 1
2021-07-04 09:11:23,361 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:11:23,545 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:11:23,566 INFO mapreduce.JobSubmitter: number of splits:2
2021-07-04 09:11:26,831 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:11:28,217 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1625413465806_0001
2021-07-04 09:11:28,218 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-07-04 09:11:32,114 INFO conf.Configuration: resource-types.xml not found
2021-07-04 09:11:32,114 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-07-04 09:11:38,975 INFO impl.YarnClientImpl: Submitted application application_1625413465806_0001
2021-07-04 09:11:40,184 INFO mapreduce.Job: The url to track the job: http://aditya:8088/proxy/application_1625413465806_0001/
2021-07-04 09:11:40,361 INFO mapreduce.Job: Running job: job_1625413465806_0001
2021-07-04 09:13:46,673 INFO mapreduce.Job: Job job_1625413465806_0001 running in uber mode : false
2021-07-04 09:13:46,674 INFO mapreduce.Job:  map 0% reduce 0%
2021-07-04 09:18:31,564 INFO mapreduce.Job:  map 33% reduce 0%
2021-07-04 09:18:32,587 INFO mapreduce.Job:  map 67% reduce 0%
2021-07-04 09:18:34,657 INFO mapreduce.Job:  map 100% reduce 0%
```

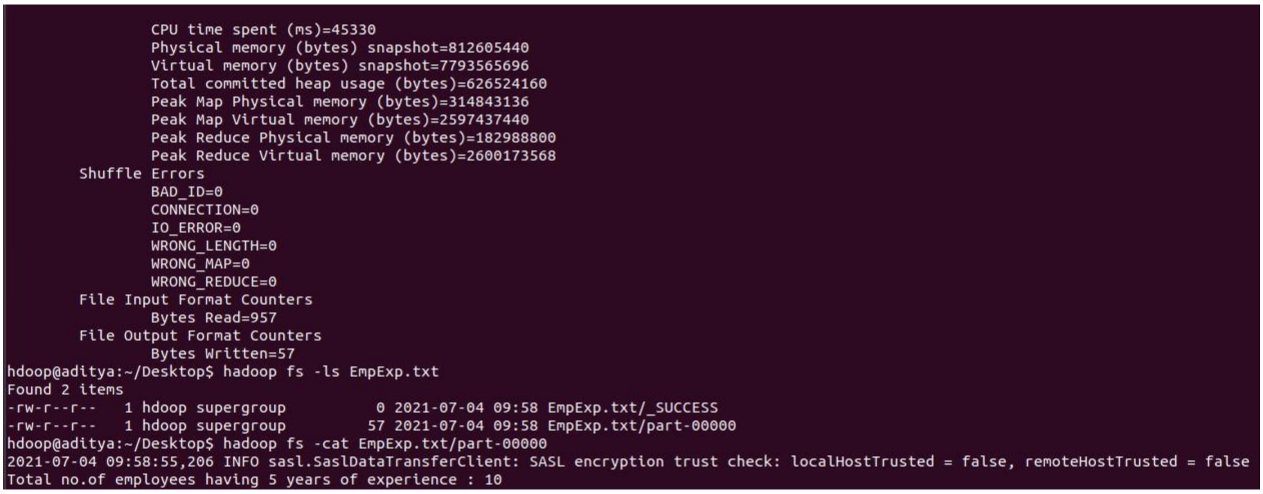


```
                Peak Map Physical memory (bytes)=304123904
                Peak Map Virtual memory (bytes)=2595966976
                Peak Reduce Physical memory (bytes)=183201792
                Peak Reduce Virtual memory (bytes)=2604199936
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=957
        File Output Format Counters
                Bytes Written=54
hdoop@aditya:~/Desktop$ hadoop fs -ls EmpISE.txt
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2021-07-04 09:19 EmpISE.txt/_SUCCESS
-rw-r--r--   1 hdoop supergroup         54 2021-07-04 09:19 EmpISE.txt/part-00000
hdoop@aditya:~/Desktop$ hadoop fs -cat EmpISE.txt/part-00000
2021-07-04 09:20:31,372 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total no.of employees working in ISE Department : 13
```

2. Total number of employees with experience=5 years

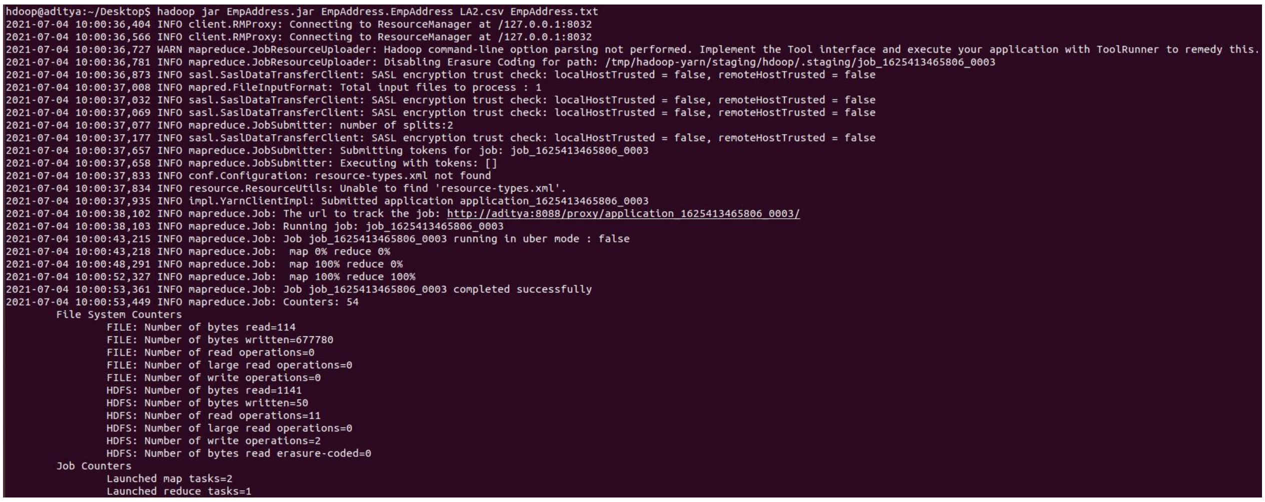


```
hdoop@aditya:~/Desktop$ hadoop jar EmpExp.jar EmpExp.EmpExp LA2.csv EmpExp.txt
2021-07-04 09:56:04,465 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 09:56:09,024 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 09:56:10,405 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-07-04 09:56:10,709 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1625413465806_0002
2021-07-04 09:56:11,286 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:56:12,643 INFO mapred.FileInputFormat: Total input files to process : 1
2021-07-04 09:56:12,836 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:56:12,862 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:56:12,989 INFO mapreduce.JobSubmitter: number of splits:2
2021-07-04 09:56:13,318 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 09:56:13,366 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1625413465806_0002
2021-07-04 09:56:13,366 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-07-04 09:56:13,903 INFO conf.Configuration: resource-types.xml not found
2021-07-04 09:56:13,904 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-07-04 09:56:16,566 INFO impl.YarnClientImpl: Submitted application application_1625413465806_0002
2021-07-04 09:56:18,577 INFO mapreduce.Job: The url to track the job: http://aditya:8088/proxy/application_1625413465806_0002/
2021-07-04 09:56:18,712 INFO mapreduce.Job: Running job: job_1625413465806_0002
2021-07-04 09:57:23,814 INFO mapreduce.Job: Job job_1625413465806_0002 running in uber mode : false
2021-07-04 09:57:23,818 INFO mapreduce.Job:  map 0% reduce 0%
2021-07-04 09:58:21,633 INFO mapreduce.Job:  map 100% reduce 0%
2021-07-04 09:58:26,679 INFO mapreduce.Job:  map 100% reduce 100%
2021-07-04 09:58:27,708 INFO mapreduce.Job: Job job_1625413465806_0002 completed successfully
2021-07-04 09:58:27,841 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=126
                FILE: Number of bytes written=677738
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1141
                HDFS: Number of bytes written=57
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
```

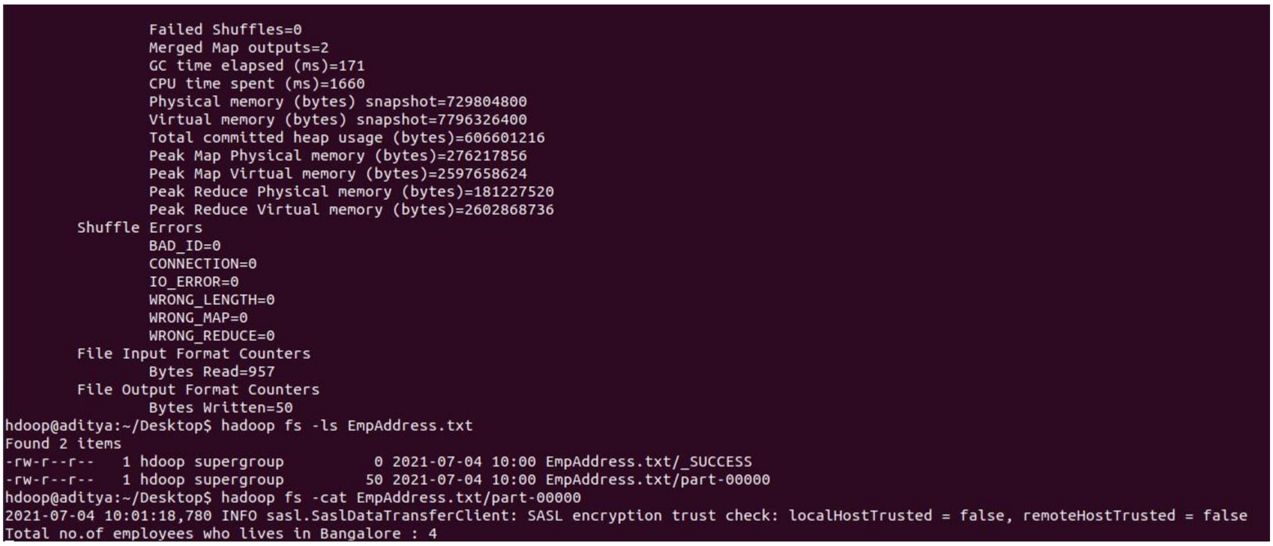


```
                CPU time spent (ms)=45330
                Physical memory (bytes) snapshot=812605440
                Virtual memory (bytes) snapshot=7793565696
                Total committed heap usage (bytes)=626524160
                Peak Map Physical memory (bytes)=314843136
                Peak Map Virtual memory (bytes)=2597437440
                Peak Reduce Physical memory (bytes)=182988800
                Peak Reduce Virtual memory (bytes)=2600173568
        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=957
        File Output Format Counters
                Bytes Written=57
hdoop@aditya:~/Desktop$ hadoop fs -ls EmpExp.txt
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2021-07-04 09:58 EmpExp.txt/_SUCCESS
-rw-r--r--   1 hdoop supergroup         57 2021-07-04 09:58 EmpExp.txt/part-00000
hdoop@aditya:~/Desktop$ hadoop fs -cat EmpExp.txt/part-00000
2021-07-04 09:58:55,206 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total no.of employees having 5 years of experience : 10
```

3. Count the number of employees who lives in Bangalore.



```
hdoop@aditya:~/Desktop$ hadoop jar EmpAddress.jar EmpAddress.EmpAddress LA2.csv EmpAddress.txt
2021-07-04 10:00:36,404 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 10:00:36,566 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-07-04 10:00:36,727 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2021-07-04 10:00:36,781 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1625413465806_0003
2021-07-04 10:00:36,873 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 10:00:37,008 INFO mapred.FileInputFormat: Total input files to process : 1
2021-07-04 10:00:37,032 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 10:00:37,069 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 10:00:37,077 INFO mapreduce.JobSubmitter: number of splits:2
2021-07-04 10:00:37,177 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-07-04 10:00:37,657 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1625413465806_0003
2021-07-04 10:00:37,658 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-07-04 10:00:37,833 INFO conf.Configuration: resource-types.xml not found
2021-07-04 10:00:37,834 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-07-04 10:00:37,935 INFO impl.YarnClientImpl: Submitted application application_1625413465806_0003
2021-07-04 10:00:38,102 INFO mapreduce.Job: The url to track the job: http://aditya:8088/proxy/application_1625413465806_0003/
2021-07-04 10:00:38,103 INFO mapreduce.Job: Running job: job_1625413465806_0003
2021-07-04 10:00:43,215 INFO mapreduce.Job: Job job_1625413465806_0003 running in uber mode : false
2021-07-04 10:00:43,218 INFO mapreduce.Job:  map 0% reduce 0%
2021-07-04 10:00:48,291 INFO mapreduce.Job:  map 100% reduce 0%
2021-07-04 10:00:52,327 INFO mapreduce.Job:  map 100% reduce 100%
2021-07-04 10:00:53,361 INFO mapreduce.Job: Job job_1625413465806_0003 completed successfully
2021-07-04 10:00:53,449 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=114
                FILE: Number of bytes written=677780
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1141
                HDFS: Number of bytes written=50
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
```



```
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=171
                CPU time spent (ms)=1660
                Physical memory (bytes) snapshot=729804800
                Virtual memory (bytes) snapshot=7796326400
                Total committed heap usage (bytes)=606601216
                Peak Map Physical memory (bytes)=276217856
                Peak Map Virtual memory (bytes)=2597658624
                Peak Reduce Physical memory (bytes)=181227520
                Peak Reduce Virtual memory (bytes)=2602868736
        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=957
        File Output Format Counters
                Bytes Written=50
hdoop@aditya:~/Desktop$ hadoop fs -ls EmpAddress.txt
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2021-07-04 10:00 EmpAddress.txt/_SUCCESS
-rw-r--r--   1 hdoop supergroup         50 2021-07-04 10:00 EmpAddress.txt/part-00000
hdoop@aditya:~/Desktop$ hadoop fs -cat EmpAddress.txt/part-00000
2021-07-04 10:01:18,780 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total no.of employees who lives in Bangalore : 4
```

HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

Hive is not

a. A relational database

b. A design for Online Transaction Processing (OLTP)

c. A language for real-time queries and row-level updates

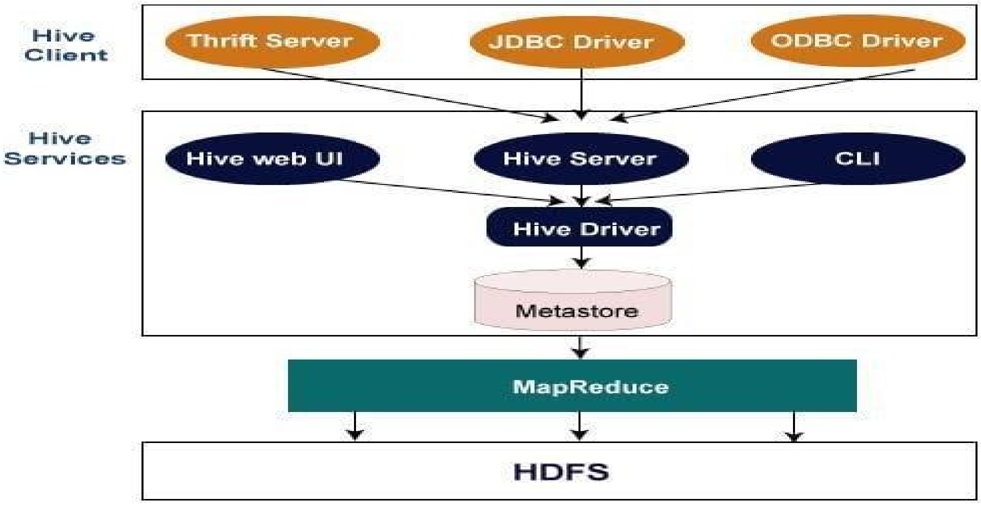
Features of Hive

d. It stores schema in a database and processed data into HDFS.

e. It is designed for OLAP.

f. It provides SQL type language for querying called HiveQL or HQL.

g. It is familiar, fast, scalable, and extensible.



Hive Problem Statement

Exercise-II

Use the above dataset in .csv file and create a database called as EmployeeDB. Create a table under the database called as Employee using HIVEQL. The table fields are same, that is,

Name	SSN	Salary	Address	Dname	Experience
Harsha	5000	30000	Bangalore	ISE	5

Use the HiveQL language to perform the following Query based Map-reduce operations,

1. Insert 5 records using INSERT command.

2. Demonstrate the Alter command for the following cases,

a. Rename the table name to “Emp”.

b. Rename the column name “Dname” to “Dept_name”.

3. Retrieve all the employees whose salary is not less than 50000.

4. Extract all employees who live in Bangalore but having less than 5 years of experience

5. Create separate view containing Name, Dept_name of employees

6. Display Name and SSN and use group by SSN and order by Name

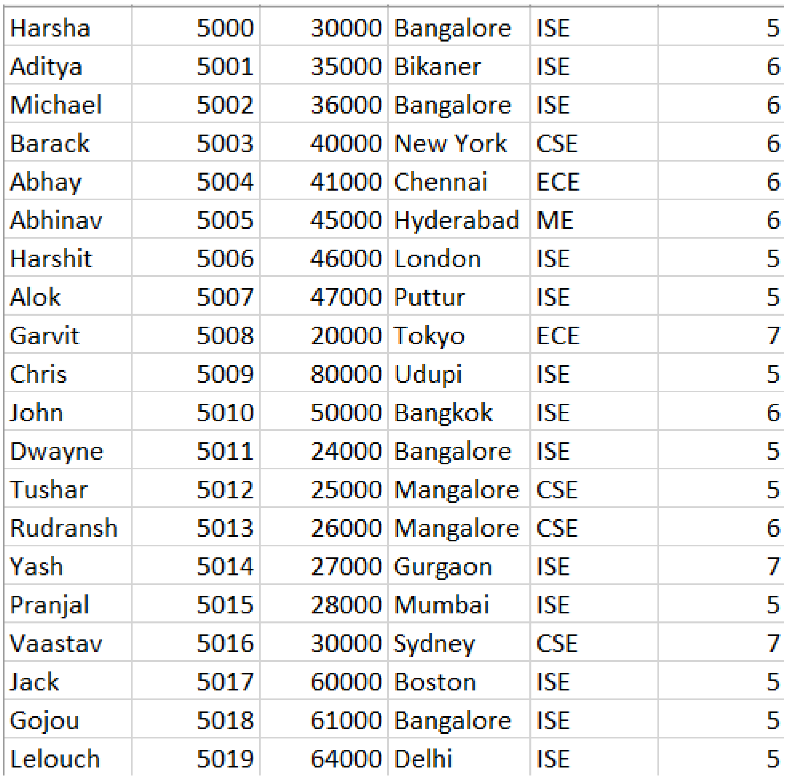
7. Retrieve Maximum salary, minimum salary and Average salary of the employees

8. Create Another table called Department with the following fields (Dname = Dept_name and perform the following joins (outer, left outer, right outer) over Dname

Dno	Dname
6	ISE

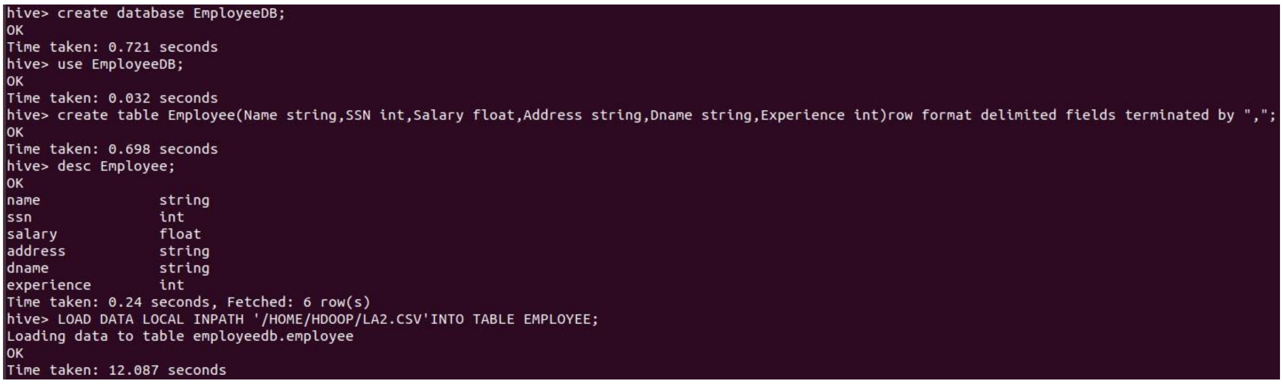
Dataset Description

LA2.csv

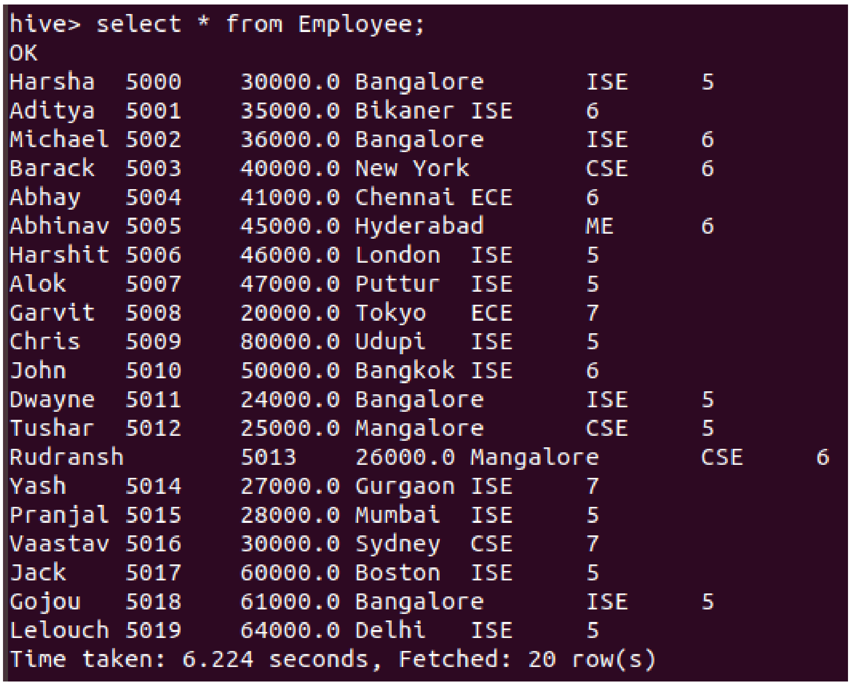


Harsha	5000	30000	Bangalore	ISE	5
Aditya	5001	35000	Bikaner	ISE	6
Michael	5002	36000	Bangalore	ISE	6
Barack	5003	40000	New York	CSE	6
Abhay	5004	41000	Chennai	ECE	6
Abhinav	5005	45000	Hyderabad	ME	6
Harshit	5006	46000	London	ISE	5
Alok	5007	47000	Puttur	ISE	5
Garvit	5008	20000	Tokyo	ECE	7
Chris	5009	80000	Udupi	ISE	5
John	5010	50000	Bangkok	ISE	6
Dwayne	5011	24000	Bangalore	ISE	5
Tushar	5012	25000	Mangalore	CSE	5
Rudransh	5013	26000	Mangalore	CSE	6
Yash	5014	27000	Gurgaon	ISE	7
Pranjal	5015	28000	Mumbai	ISE	5
Vaastav	5016	30000	Sydney	CSE	7
Jack	5017	60000	Boston	ISE	5
Gojou	5018	61000	Bangalore	ISE	5
Lelouch	5019	64000	Delhi	ISE	5

Results and Snapshots



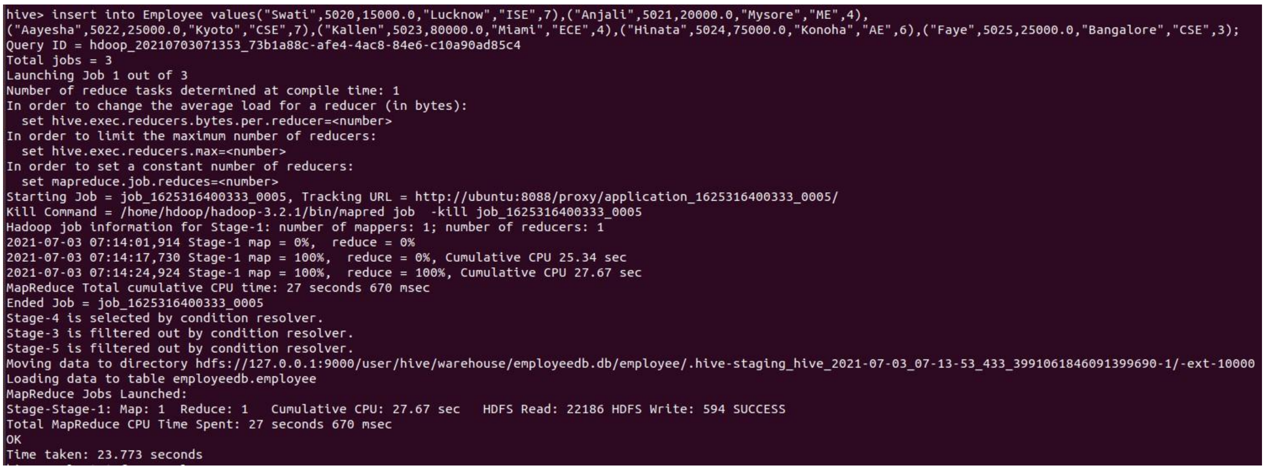
```
hive> create database EmployeeDB;
OK
Time taken: 0.721 seconds
hive> use EmployeeDB;
OK
Time taken: 0.032 seconds
hive> create table Employee(Name string,SSN int,Salary float,Address string,Dname string,Experience int)row format delimited fields terminated by ",";
OK
Time taken: 0.698 seconds
hive> desc Employee;
OK
name                    string
ssn                     int
salary                  float
address                 string
dname                   string
experience              int
Time taken: 0.24 seconds, Fetched: 6 row(s)
hive> LOAD DATA LOCAL INPATH '/HOME/HDOOP/LA2.CSV'INTO TABLE EMPLOYEE;
Loading data to table employeedb.employee
OK
Time taken: 12.087 seconds
```



```
hive> select * from Employee;
OK
Harsha  5000    30000.0 Bangalore       ISE     5
Aditya  5001    35000.0 Bikaner ISE     6
Michael 5002    36000.0 Bangalore       ISE     6
Barack  5003    40000.0 New York        CSE     6
Abhay   5004    41000.0 Chennai ECE     6
Abhinav 5005    45000.0 Hyderabad       ME      6
Harshit 5006    46000.0 London  ISE     5
Alok    5007    47000.0 Puttur  ISE     5
Garvit  5008    20000.0 Tokyo   ECE     7
Chris   5009    80000.0 Udupi   ISE     5
John    5010    50000.0 Bangkok ISE     6
Dwayne  5011    24000.0 Bangalore       ISE     5
Tushar  5012    25000.0 Mangalore       CSE     5
Rudransh        5013    26000.0 Mangalore       CSE     6
Yash    5014    27000.0 Gurgaon ISE     7
Pranjal 5015    28000.0 Mumbai  ISE     5
Vaastav 5016    30000.0 Sydney  CSE     7
Jack    5017    60000.0 Boston  ISE     5
Gojou   5018    61000.0 Bangalore       ISE     5
Lelouch 5019    64000.0 Delhi   ISE     5
Time taken: 6.224 seconds, Fetched: 20 row(s)
```

Query 1

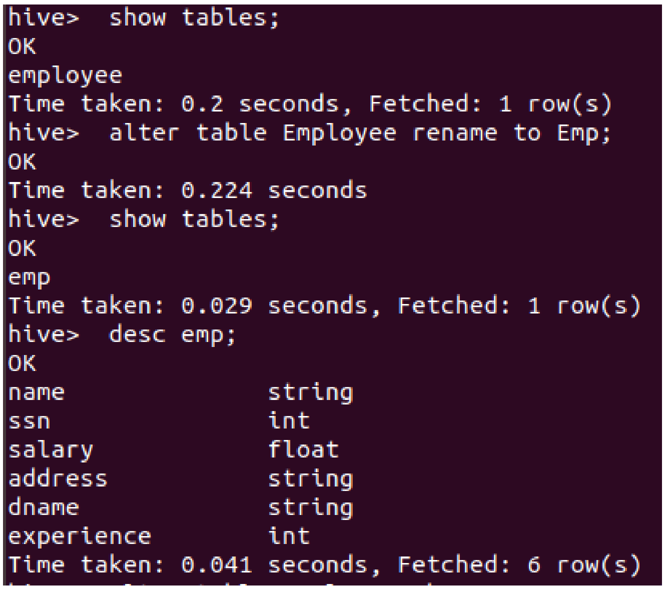
Insert 5 records using the INSERT command.



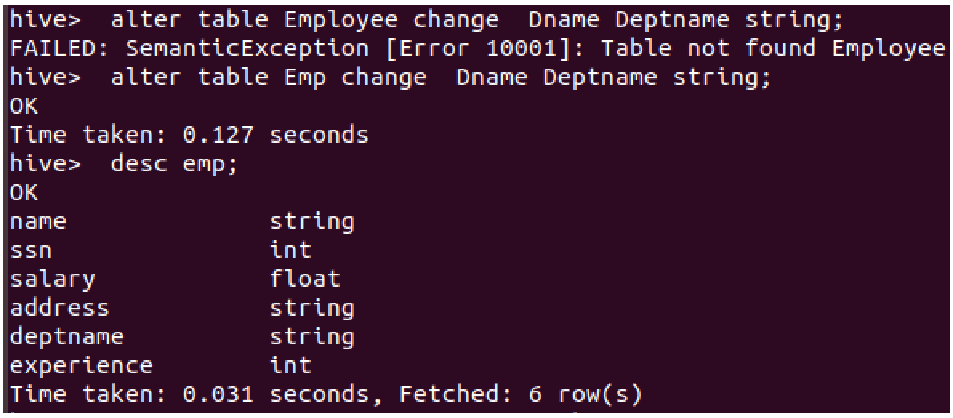
```
hive> insert into Employee values("Swati",5020,15000.0,"Lucknow","ISE",7),("Anjali",5021,20000.0,"Mysore","ME",4),
("Aayesha",5022,25000.0,"Kyoto","CSE",7),("Kallen",5023,80000.0,"Miami","ECE",4),("Hinata",5024,75000.0,"Konoha","AE",6),("Faye",5025,25000.0,"Bangalore","CSE",3);
Query ID = hdoop_20210703071353_73b1a88c-afe4-4ac8-84e6-c10a90ad85c4
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 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_1625316400333_0005, Tracking URL = http://ubuntu:8088/proxy/application_1625316400333_0005/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625316400333_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-03 07:14:01,914 Stage-1 map = 0%,  reduce = 0%
2021-07-03 07:14:17,730 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 25.34 sec
2021-07-03 07:14:24,924 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.67 sec
MapReduce Total cumulative CPU time: 27 seconds 670 msec
Ended Job = job_1625316400333_0005
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/employeedb.db/employee/.hive-staging_hive_2021-07-03_07-13-53_433_3991061846091399690-1/-ext-10000
Loading data to table employeedb.employee
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 27.67 sec   HDFS Read: 22186 HDFS Write: 594 SUCCESS
Total MapReduce CPU Time Spent: 27 seconds 670 msec
OK
Time taken: 23.773 seconds
```

Query 2

Demonstrate the Alter command for the following cases,
a. Rename the table name to “Emp”.
b. Rename the column name “Dname” to “Dept_name”.



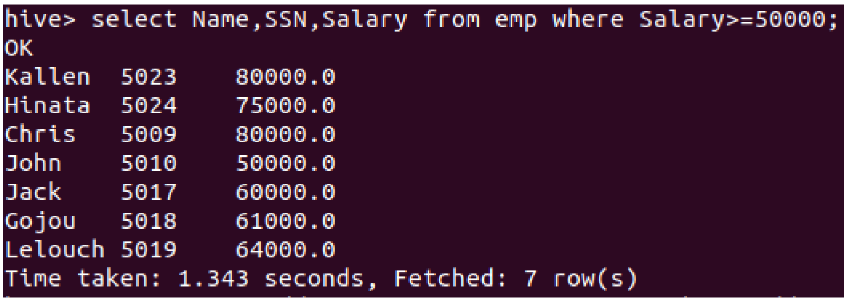
```
hive>  show tables;
OK
employee
Time taken: 0.2 seconds, Fetched: 1 row(s)
hive>  alter table Employee rename to Emp;
OK
Time taken: 0.224 seconds
hive>  show tables;
OK
emp
Time taken: 0.029 seconds, Fetched: 1 row(s)
hive>  desc emp;
OK
name                  string
ssn                   int
salary                float
address               string
dname                 string
experience            int
Time taken: 0.041 seconds, Fetched: 6 row(s)
```



```
hive>  alter table Employee change  Dname Deptname string;
FAILED: SemanticException [Error 10001]: Table not found Employee
hive>  alter table Emp change  Dname Deptname string;
OK
Time taken: 0.127 seconds
hive>  desc emp;
OK
name                  string
ssn                   int
salary                float
address               string
deptname              string
experience            int
Time taken: 0.031 seconds, Fetched: 6 row(s)
```

Query 3

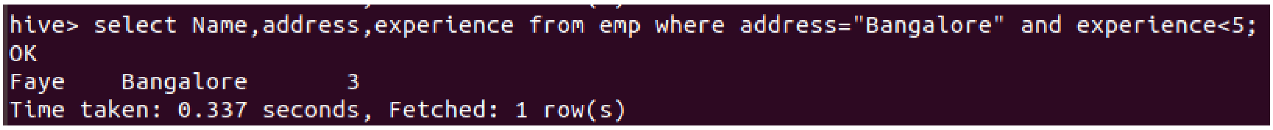
Retrieve all the employees whose salary is not less than 50000.



```
hive> select Name,SSN,Salary from emp where Salary>=50000;
OK
Kallen  5023    80000.0
Hinata  5024    75000.0
Chris   5009    80000.0
John    5010    50000.0
Jack    5017    60000.0
Gojou   5018    61000.0
Lelouch 5019    64000.0
Time taken: 1.343 seconds, Fetched: 7 row(s)
```

Query 4

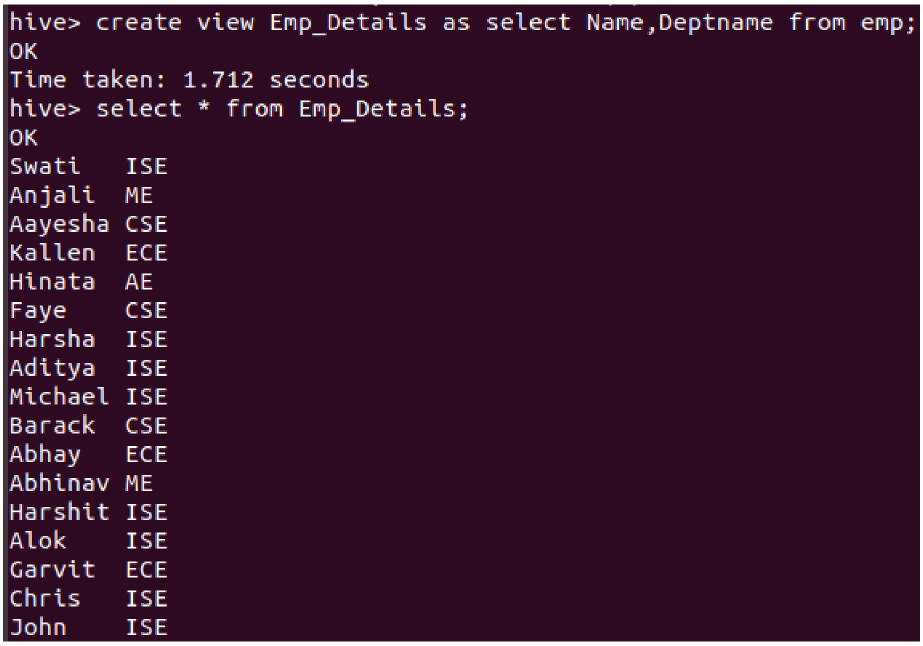
Extract all employees who live in Bangalore but having less than 5 years of experience.



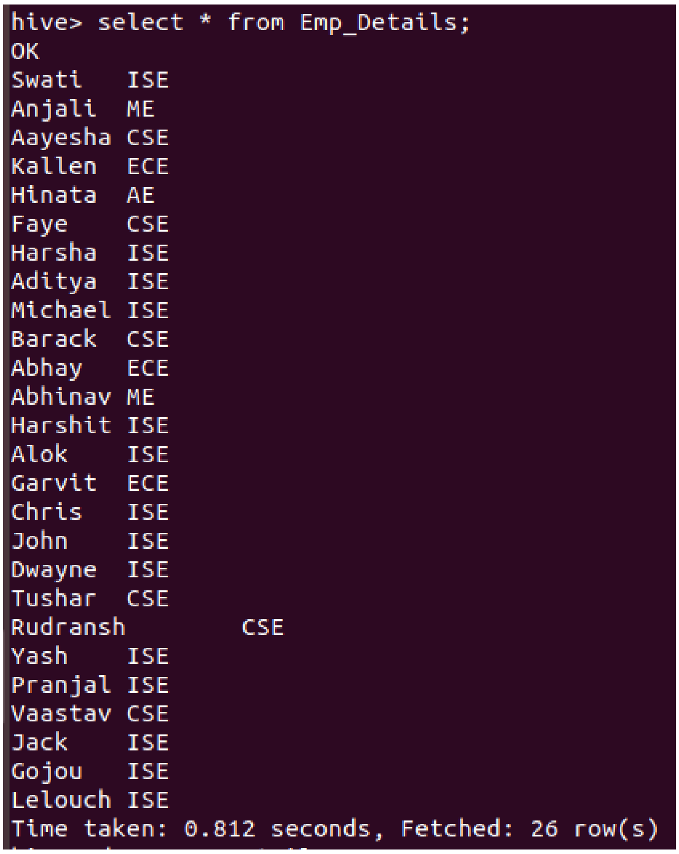
```
hive> select Name,address,experience from emp where address="Bangalore" and experience<5;
OK
Faye    Bangalore       3
Time taken: 0.337 seconds, Fetched: 1 row(s)
```

Query 5

Create separate view containing Name, Dept_name of employees



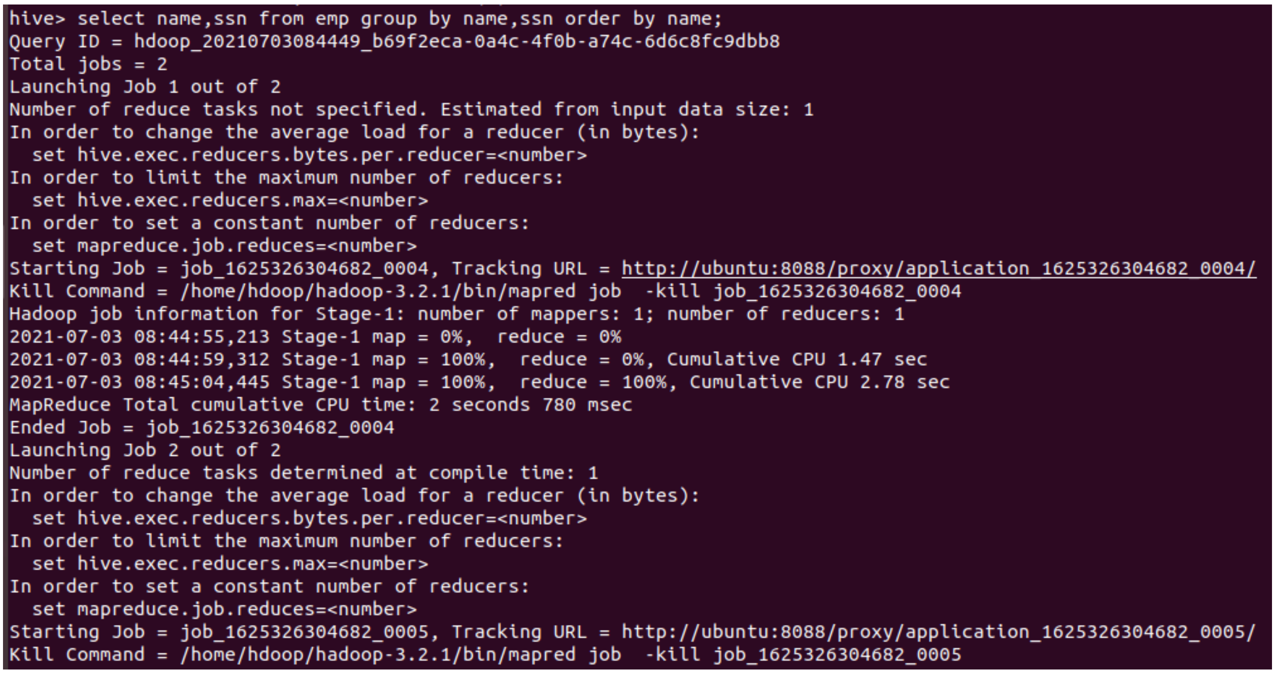
```
hive> create view Emp_Details as select Name,Deptname from emp;
OK
Time taken: 1.712 seconds
hive> select * from Emp_Details;
OK
Swati   ISE
Anjali  ME
Aayesha CSE
Kallen  ECE
Hinata  AE
Faye    CSE
Harsha  ISE
Aditya  ISE
Michael ISE
Barack  CSE
Abhay   ECE
Abhinav ME
Harshit ISE
Alok    ISE
Garvit  ECE
Chris   ISE
John    ISE
```



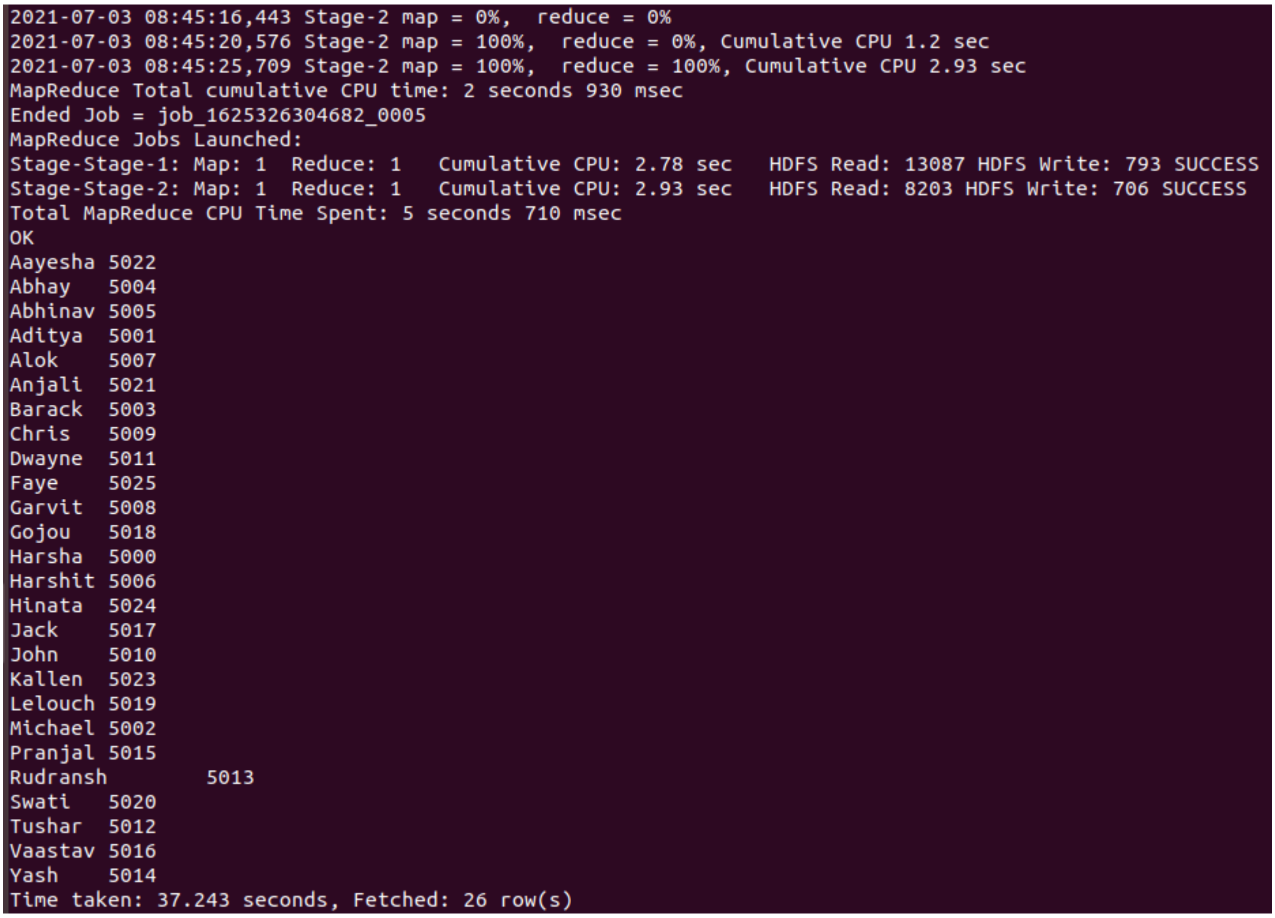
```
hive> select * from Emp_Details;
OK
Swati   ISE
Anjali  ME
Aayesha CSE
Kallen  ECE
Hinata  AE
Faye    CSE
Harsha  ISE
Aditya  ISE
Michael ISE
Barack  CSE
Abhay   ECE
Abhinav ME
Harshit ISE
Alok    ISE
Garvit  ECE
Chris   ISE
John    ISE
Dwayne  ISE
Tushar  CSE
Rudransh        CSE
Yash    ISE
Pranjal ISE
Vaastav CSE
Jack    ISE
Gojou   ISE
Lelouch ISE
Time taken: 0.812 seconds, Fetched: 26 row(s)
```

Query 6

Display Name and SSN and use group by SSN and order by Name.



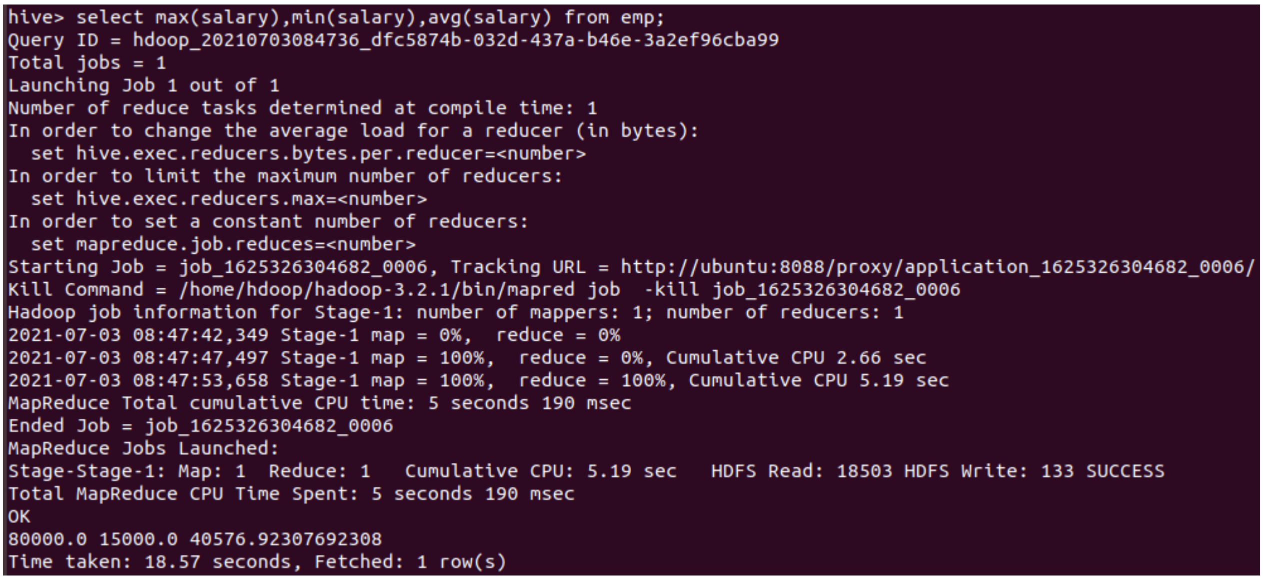
```
hive> select name,ssn from emp group by name,ssn order by name;
Query ID = hdoop_20210703084449_b69f2eca-0a4c-4f0b-a74c-6d6c8fc9dbb8
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in 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_1625326304682_0004, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0004/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-03 08:44:55,213 Stage-1 map = 0%,  reduce = 0%
2021-07-03 08:44:59,312 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.47 sec
2021-07-03 08:45:04,445 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.78 sec
MapReduce Total cumulative CPU time: 2 seconds 780 msec
Ended Job = job_1625326304682_0004
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_1625326304682_0005, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0005/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0005
```



```
2021-07-03 08:45:16,443 Stage-2 map = 0%,  reduce = 0%
2021-07-03 08:45:20,576 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.2 sec
2021-07-03 08:45:25,709 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.93 sec
MapReduce Total cumulative CPU time: 2 seconds 930 msec
Ended Job = job_1625326304682_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.78 sec   HDFS Read: 13087 HDFS Write: 793 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.93 sec   HDFS Read: 8203 HDFS Write: 706 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 710 msec
OK
Aayesha 5022
Abhay   5004
Abhinav 5005
Aditya  5001
Alok    5007
Anjali  5021
Barack  5003
Chris   5009
Dwayne  5011
Faye    5025
Garvit  5008
Gojou   5018
Harsha  5000
Harshit 5006
Hinata  5024
Jack    5017
John    5010
Kallen  5023
Lelouch 5019
Michael 5002
Pranjal 5015
Rudransh        5013
Swati   5020
Tushar  5012
Vaastav 5016
Yash    5014
Time taken: 37.243 seconds, Fetched: 26 row(s)
```

Query 7

Retrieve Maximum salary, minimum salary and Average salary of the employees

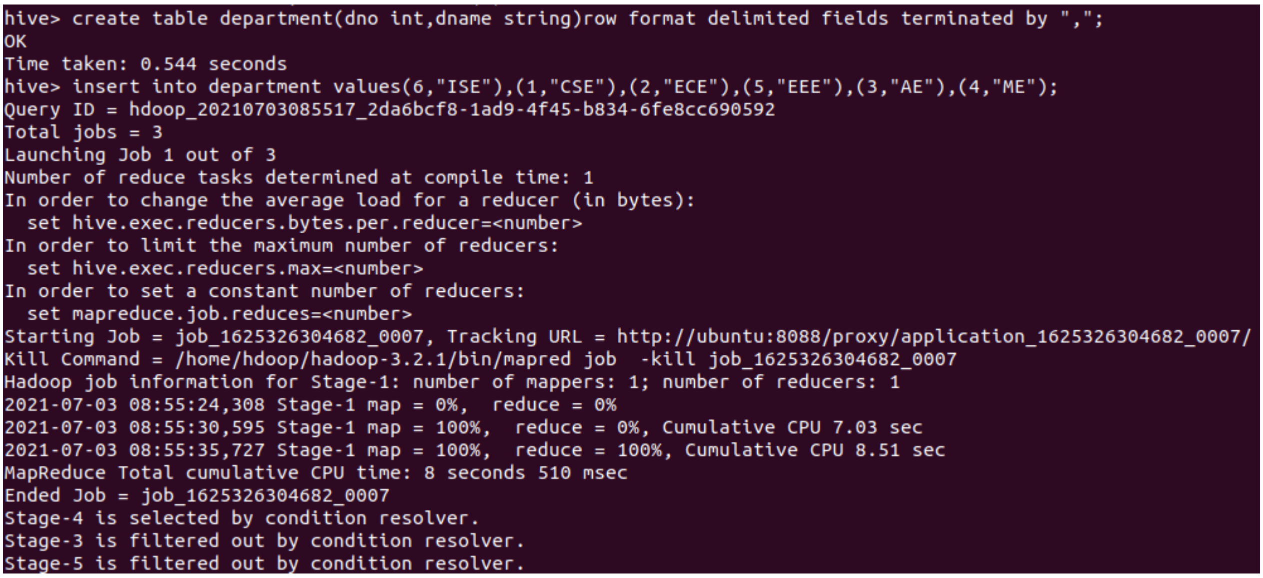


```
hive> select max(salary),min(salary),avg(salary) from emp;
Query ID = hdoop_20210703084736_dfc5874b-032d-437a-b46e-3a2ef96cba99
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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_1625326304682_0006, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0006/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-03 08:47:42,349 Stage-1 map = 0%,  reduce = 0%
2021-07-03 08:47:47,497 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.66 sec
2021-07-03 08:47:53,658 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.19 sec
MapReduce Total cumulative CPU time: 5 seconds 190 msec
Ended Job = job_1625326304682_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.19 sec   HDFS Read: 18503 HDFS Write: 133 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 190 msec
OK
80000.0 15000.0 40576.92307692308
Time taken: 18.57 seconds, Fetched: 1 row(s)
```

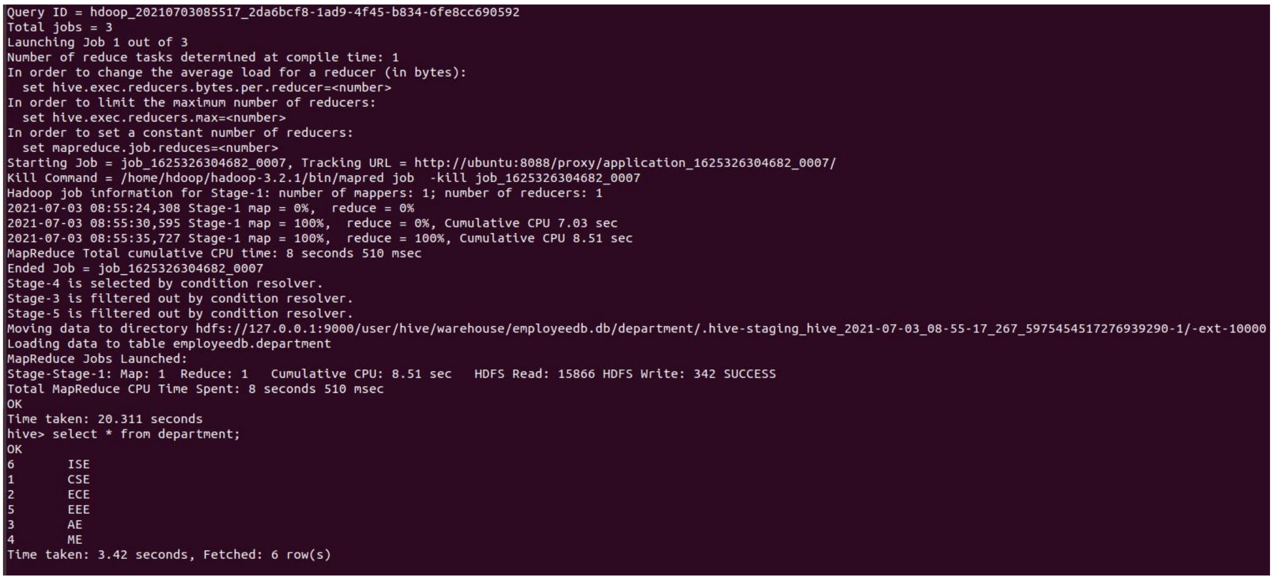
Query 8

Create Another table called Department with the following fields (Dname = Dept_name and perform the following joins (outer, left outer, right outer) over Dname.

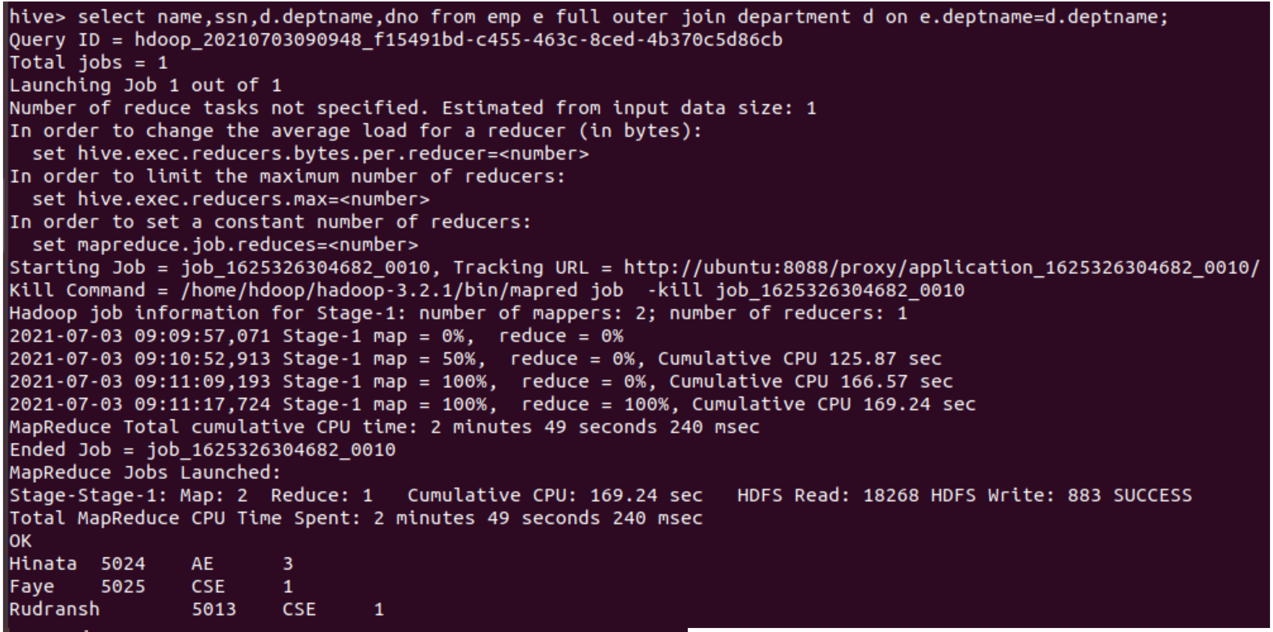
Dno	Dname
6	ISE



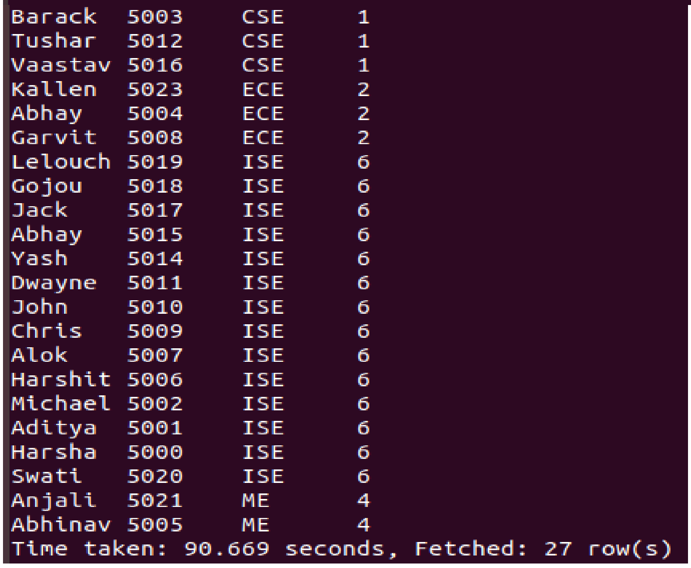
```
hive> create table department(dno int,dname string)row format delimited fields terminated by ",";
OK
Time taken: 0.544 seconds
hive> insert into department values(6,"ISE"),(1,"CSE"),(2,"ECE"),(5,"EEE"),(3,"AE"),(4,"ME");
Query ID = hdoop_20210703085517_2da6bcf8-1ad9-4f45-b834-6fe8cc690592
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 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_1625326304682_0007, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0007/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-03 08:55:24,308 Stage-1 map = 0%,  reduce = 0%
2021-07-03 08:55:30,595 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.03 sec
2021-07-03 08:55:35,727 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.51 sec
MapReduce Total cumulative CPU time: 8 seconds 510 msec
Ended Job = job_1625326304682_0007
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
```



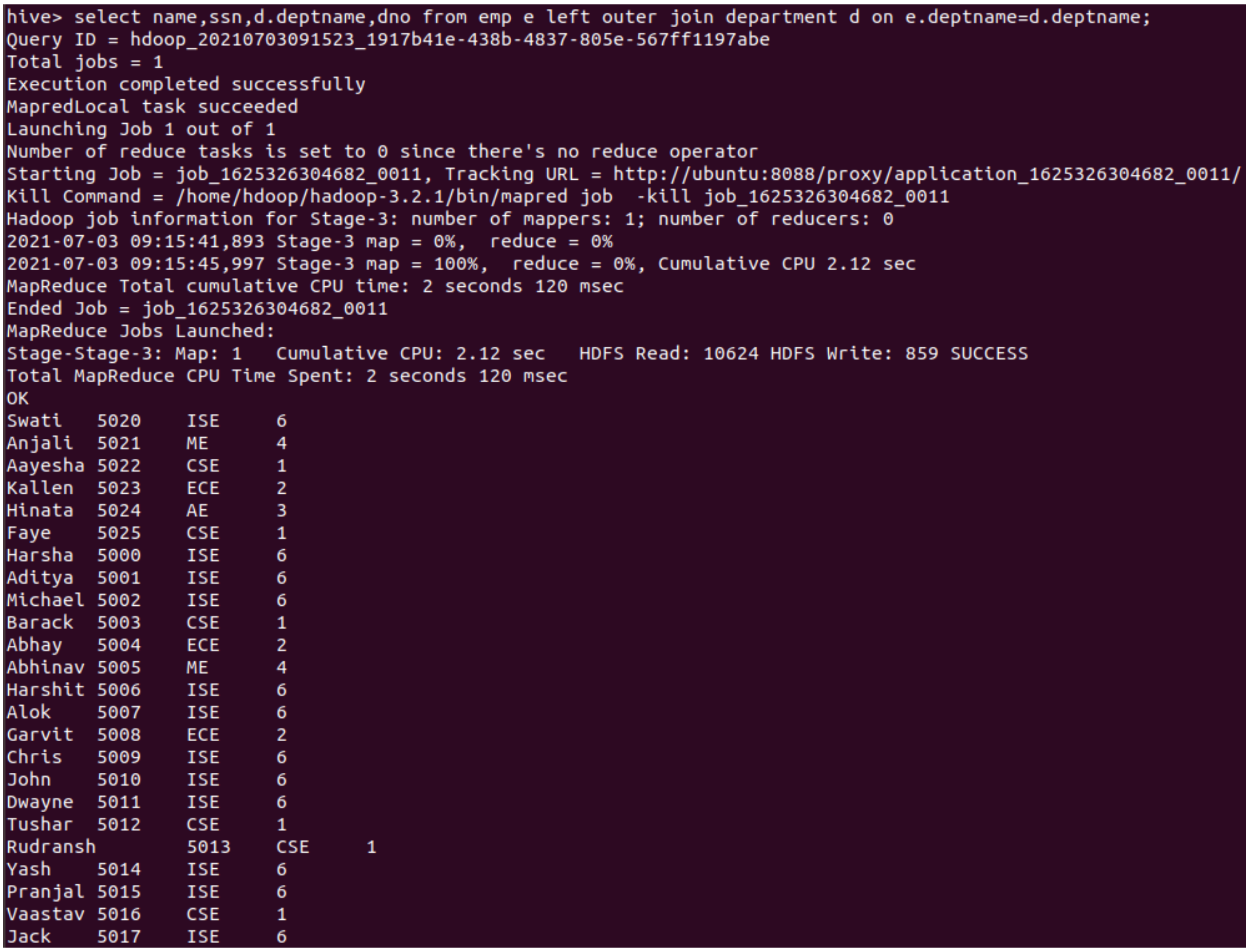
```
Query ID = hdoop_20210703085517_2da6bcf8-1ad9-4f45-b834-6fe8cc690592
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the 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_1625326304682_0007, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0007/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-03 08:55:24,308 Stage-1 map = 0%,  reduce = 0%
2021-07-03 08:55:30,595 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.03 sec
2021-07-03 08:55:35,727 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.51 sec
MapReduce Total cumulative CPU time: 8 seconds 510 msec
Ended Job = job_1625326304682_0007
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://127.0.0.1:9000/user/hive/warehouse/employeedb.db/department/.hive-staging_hive_2021-07-03_08-55-17_267_5975454517276939290-1/-ext-10000
Loading data to table employeedb.department
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.51 sec   HDFS Read: 15866 HDFS Write: 342 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 510 msec
OK
Time taken: 20.311 seconds
hive> select * from department;
OK
6       ISE
1       CSE
2       ECE
5       EEE
3       AE
4       ME
Time taken: 3.42 seconds, Fetched: 6 row(s)
```



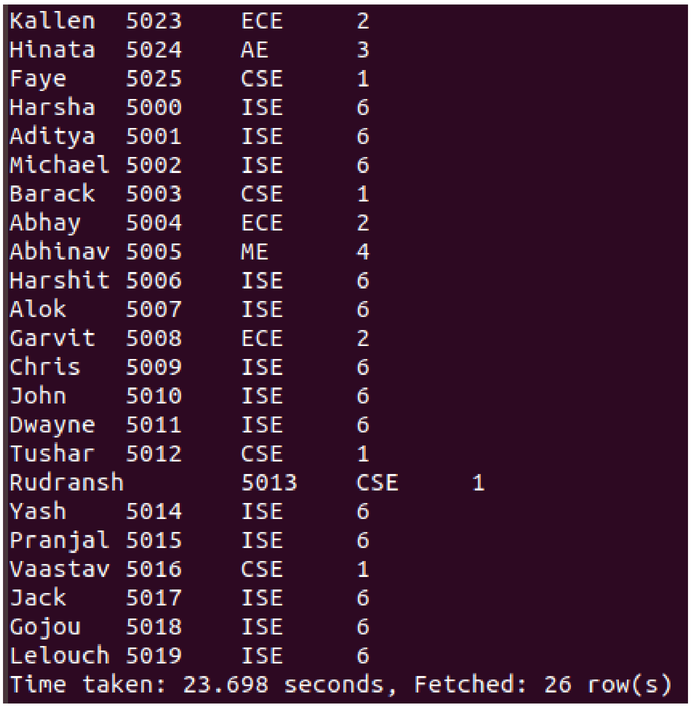
```
hive> select name,ssn,d.deptname,dno from emp e full outer join department d on e.deptname=d.deptname;
Query ID = hdoop_20210703090948_f15491bd-c455-463c-8ced-4b370c5d86cb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change 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_1625326304682_0010, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0010/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0010
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2021-07-03 09:09:57,071 Stage-1 map = 0%,  reduce = 0%
2021-07-03 09:10:52,913 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 125.87 sec
2021-07-03 09:11:09,193 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 166.57 sec
2021-07-03 09:11:17,724 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 169.24 sec
MapReduce Total cumulative CPU time: 2 minutes 49 seconds 240 msec
Ended Job = job_1625326304682_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 169.24 sec   HDFS Read: 18268 HDFS Write: 883 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 49 seconds 240 msec
OK
Hinata  5024    AE      3
Faye    5025    CSE     1
Rudransh        5013    CSE     1
```



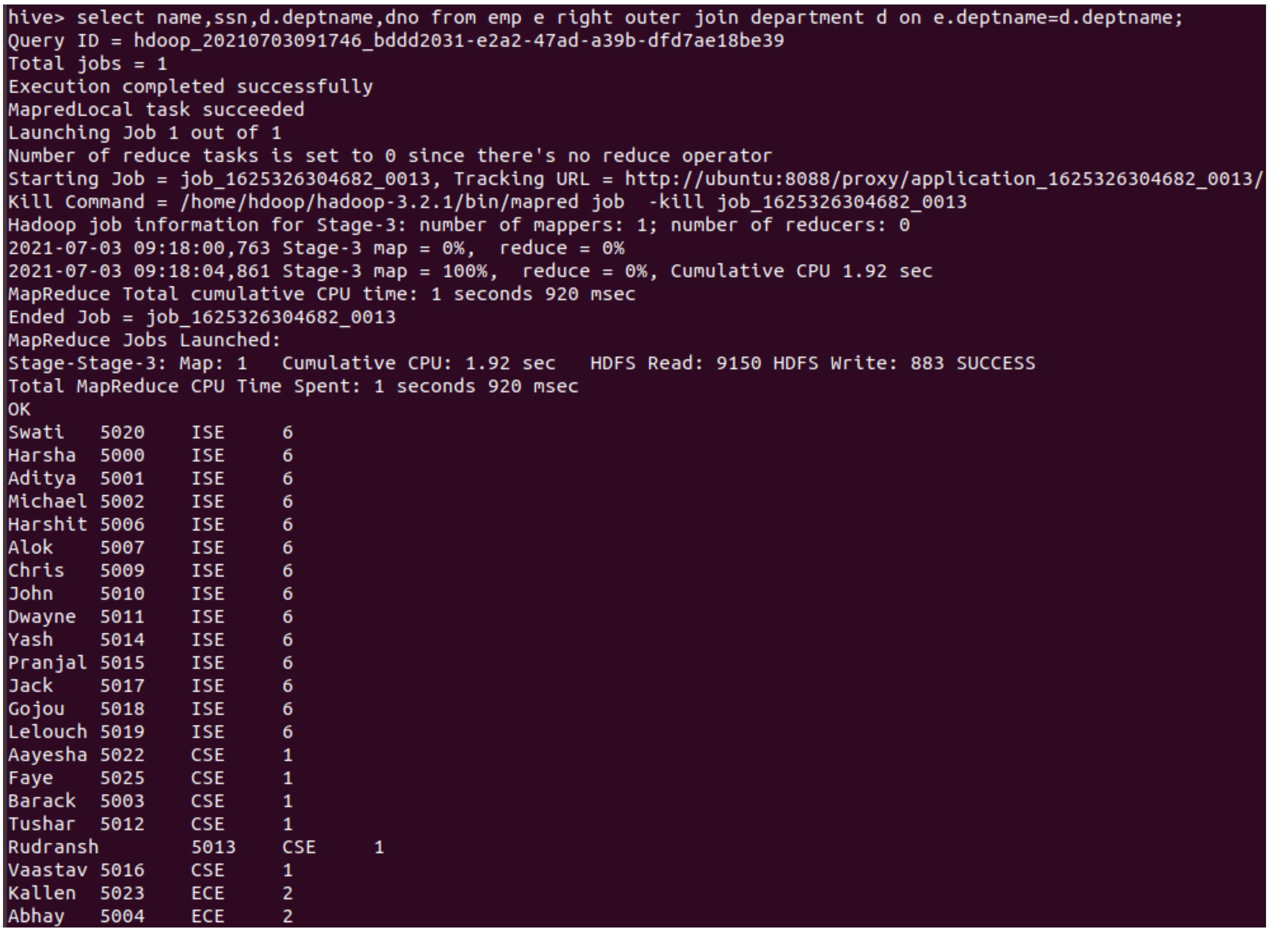
```
Barack  5003    CSE     1
Tushar  5012    CSE     1
Vaastav 5016    CSE     1
Kallen  5023    ECE     2
Abhay   5004    ECE     2
Garvit  5008    ECE     2
Lelouch 5019    ISE     6
Gojou   5018    ISE     6
Jack    5017    ISE     6
Abhay   5015    ISE     6
Yash    5014    ISE     6
Dwayne  5011    ISE     6
John    5010    ISE     6
Chris   5009    ISE     6
Alok    5007    ISE     6
Harshit 5006    ISE     6
Michael 5002    ISE     6
Aditya  5001    ISE     6
Harsha  5000    ISE     6
Swati   5020    ISE     6
Anjali  5021    ME      4
Abhinav 5005    ME      4
Time taken: 90.669 seconds, Fetched: 27 row(s)
```



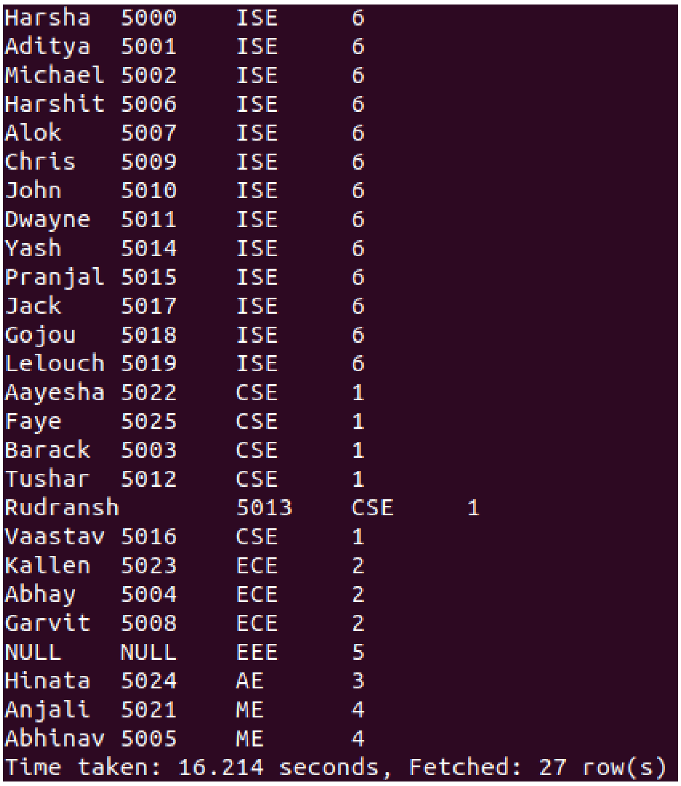
```
hive> select name,ssn,d.deptname,dno from emp e left outer join department d on e.deptname=d.deptname;
Query ID = hdoop_20210703091523_1917b41e-438b-4837-805e-567ff1197abe
Total jobs = 1
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1625326304682_0011, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0011/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0011
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-07-03 09:15:41,893 Stage-3 map = 0%,  reduce = 0%
2021-07-03 09:15:45,997 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.12 sec
MapReduce Total cumulative CPU time: 2 seconds 120 msec
Ended Job = job_1625326304682_0011
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.12 sec   HDFS Read: 10624 HDFS Write: 859 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 120 msec
OK
Swati   5020    ISE     6
Anjali  5021    ME      4
Aayesha 5022    CSE     1
Kallen  5023    ECE     2
Hinata  5024    AE      3
Faye    5025    CSE     1
Harsha  5000    ISE     6
Aditya  5001    ISE     6
Michael 5002    ISE     6
Barack  5003    CSE     1
Abhay   5004    ECE     2
Abhinav 5005    ME      4
Harshit 5006    ISE     6
Alok    5007    ISE     6
Garvit  5008    ECE     2
Chris   5009    ISE     6
John    5010    ISE     6
Dwayne  5011    ISE     6
Tushar  5012    CSE     1
Rudransh        5013    CSE     1
Yash    5014    ISE     6
Pranjal 5015    ISE     6
Vaastav 5016    CSE     1
Jack    5017    ISE     6
```



```
Kallen  5023    ECE     2
Hinata  5024    AE      3
Faye    5025    CSE     1
Harsha  5000    ISE     6
Aditya  5001    ISE     6
Michael 5002    ISE     6
Barack  5003    CSE     1
Abhay   5004    ECE     2
Abhinav 5005    ME      4
Harshit 5006    ISE     6
Alok    5007    ISE     6
Garvit  5008    ECE     2
Chris   5009    ISE     6
John    5010    ISE     6
Dwayne  5011    ISE     6
Tushar  5012    CSE     1
Rudransh        5013    CSE     1
Yash    5014    ISE     6
Pranjal 5015    ISE     6
Vaastav 5016    CSE     1
Jack    5017    ISE     6
Gojou   5018    ISE     6
Lelouch 5019    ISE     6
Time taken: 23.698 seconds, Fetched: 26 row(s)
```



```
hive> select name,ssn,d.deptname,dno from emp e right outer join department d on e.deptname=d.deptname;
Query ID = hdoop_20210703091746_bddd2031-e2a2-47ad-a39b-dfd7ae18be39
Total jobs = 1
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_1625326304682_0013, Tracking URL = http://ubuntu:8088/proxy/application_1625326304682_0013/
Kill Command = /home/hdoop/hadoop-3.2.1/bin/mapred job  -kill job_1625326304682_0013
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2021-07-03 09:18:00,763 Stage-3 map = 0%,  reduce = 0%
2021-07-03 09:18:04,861 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.92 sec
MapReduce Total cumulative CPU time: 1 seconds 920 msec
Ended Job = job_1625326304682_0013
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.92 sec   HDFS Read: 9150 HDFS Write: 883 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 920 msec
OK
Swati   5020    ISE     6
Harsha  5000    ISE     6
Aditya  5001    ISE     6
Michael 5002    ISE     6
Harshit 5006    ISE     6
Alok    5007    ISE     6
Chris   5009    ISE     6
John    5010    ISE     6
Dwayne  5011    ISE     6
Yash    5014    ISE     6
Pranjal 5015    ISE     6
Jack    5017    ISE     6
Gojou   5018    ISE     6
Lelouch 5019    ISE     6
Aayesha 5022    CSE     1
Faye    5025    CSE     1
Barack  5003    CSE     1
Tushar  5012    CSE     1
Rudransh        5013    CSE     1
Vaastav 5016    CSE     1
Kallen  5023    ECE     2
Abhay   5004    ECE     2
```



```
Harsha  5000    ISE     6
Aditya  5001    ISE     6
Michael 5002    ISE     6
Harshit 5006    ISE     6
Alok    5007    ISE     6
Chris   5009    ISE     6
John    5010    ISE     6
Dwayne  5011    ISE     6
Yash    5014    ISE     6
Pranjal 5015    ISE     6
Jack    5017    ISE     6
Gojou   5018    ISE     6
Lelouch 5019    ISE     6
Aayesha 5022    CSE     1
Faye    5025    CSE     1
Barack  5003    CSE     1
Tushar  5012    CSE     1
Rudransh        5013    CSE     1
Vaastav 5016    CSE     1
Kallen  5023    ECE     2
Abhay   5004    ECE     2
Garvit  5008    ECE     2
NULL    NULL    EEE     5
Hinata  5024    AE      3
Anjali  5021    ME      4
Abhinav 5005    ME      4
Time taken: 16.214 seconds, Fetched: 27 row(s)
```

References

Hadoop & Map Reduce:

https://www.youtube.com/watch?v=U3fkWvaqgl

8

https://www.youtube.com/watch?v=K0aDh_sfVrc

Hive:

https://www.youtube.com/watch?v=SAX8b3AN3Uc