S.Y. M.C.A
SEMESTER III

SIDDHI S.KOTRE
ROLL NO: 50

MCA Sem III – Div A

Big Data Analytics and Visualization

Subject Teacher – Dr. Ashwini Renavikar

Lab 4: Hive

Aim: Perform various operations using Hive

Theory:

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive allows users to query large datasets stored in Hadoop's HDFS (Hadoop Distributed File System) using a language called HiveQL, which is similar to SQL.

Key Components of Hive

- **Database**: In Hive, databases are used to organize tables. A database can have multiple tables related to different datasets.
- **Tables**: Hive supports two types of tables:
    - **Managed Tables**: Hive manages both the data and the schema. When a managed table is dropped, the underlying data is also deleted.
    - **External Tables**: Only the schema is managed by Hive, but the data is stored externally. Dropping an external table does not remove the data.
- **Partitions**: Hive tables can be partitioned, which allows Hive to divide a table into multiple segments. This improves query performance by scanning only necessary partitions instead of the entire table.
- **Buckets**: Tables can be further subdivided into buckets based on the hash of a column value. Bucketing is another method for optimizing query performance.

SIDDHI S.KOTRE
ROLL NO: 50

1.Create a Database with the name College

```
hive> create database timscdr;
OK
Time taken: 0.287 seconds
```

2. Create a Hive Managed tables, employee, and dept under database college

Employee:
empcode INT,
empfname STRING,
emplname STRING,
job STRING,
manager STRING,
hiredate STRING,
salary INT,
commission INT,
deptcode INT
Dept table:
deptcode INT,
deptname STRING,
location STRING

```
hive> CREATE TABLE IF NOT EXISTS employee (
    >     empcode INT,
    >     empfname STRING,
    >     emplname STRING,
    >     job STRING,
    >     manager STRING,
    >     hiredate STRING,
    >     salary INT,
    >     commission INT,
    >     deptcode INT
    > ) row format delimited fields terminated by ',
    > ;
OK
Time taken: 0.051 seconds
```

SIDDHI S.KOTRE
ROLL NO: 50

3. Load data in the tables employee and dept using .csv files

```
[cloudera@quickstart ~]$ nano dept_data.txt
[cloudera@quickstart ~]$ nano dept_data.txt
[cloudera@quickstart ~]$ nano employee_data.txt
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/employee_data.txt /user/h
ive/warehouse/timscdr/
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/dept_data.txt /user/hive/
warehouse/timscdr/
```

```
hive> LOAD DATA INPATH '/user/hive/warehouse/timscdr/employee_data.txt' INTO TAB
LE employee;
Loading data to table timscdr.employee
Table timscdr.employee stats: [numFiles=1, totalSize=870]
OK
Time taken: 0.208 seconds
```

4. List all the records in a table

```
hive> LOAD DATA INPATH '/user/hive/warehouse/timscdr/employee_data.txt' INTO TAB
LE employee;
Loading data to table timscdr.employee
Table timscdr.employee stats: [numFiles=1, totalSize=872]
OK
Time taken: 0.215 seconds
hive> SELECT * FROM employee;
OK
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    (
20
9499    TIM     ADOLF   SALESMAN        7698    1981-02-20      1600    300     3
0
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20
9654    SAM     MILES   SALESMAN        7698    1981-09-28      1250    1400    3
0
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       1
0
```

```
hive> LOAD DATA INPATH '/user/hive/warehouse/timscdr/dept_data.txt' INTO TABLE dep
Loading data to table timscdr.dept
Table timscdr.dept stats: [numFiles=1, totalSize=107]
OK
Time taken: 0.112 seconds
hive> select* from dept;
OK
10      FINANCE EDINBURGH
20      SOFTWARE        PADDINGTON
30      SALES   MAIDSTONE
40      MARKETING       DARLINGTON
50      ADMIN   BIRMINGHAM
Time taken: 0.027 seconds, Fetched: 5 row(s)
```

SIDDHI S.KOTRE
ROLL NO: 50

5. Get the table structure

```
hive> describe employee;
OK
empcode                 int
empfname                string
emplname                string
job                     string
manager                 string
hiredate                string
salary                  int
commission              int
deptcode                int
Time taken: 0.038 seconds, Fetched: 9 row(
hive> describe dept;
OK
deptcode                int
deptname                string
location                string
Time taken: 0.039 seconds, Fetched: 3 row(
```

6. Perform following operation on the employee table

a. Prepare a list of all the jobs and their respective number of employees working for that Job.

```
hive> SELECT job, COUNT(*) AS num_employees
    > FROM employee
    > GROUP BY job;
Query ID = cloudera_20240829010404_b06800a4-a1e5-4ea8-b8a9-17b2024daee5
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_1724910377336_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_172491037733
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-08-29 01:04:22,622 Stage-1 map = 0%,  reduce = 0%
2024-08-29 01:04:26,793 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.53 sec
2024-08-29 01:04:31,951 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.3 sec
MapReduce Total cumulative CPU time: 1 seconds 300 msec
Ended Job = job_1724910377336_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.3 sec   HDFS Read: 8450 HDFS Write: 78 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 300 msec
OK
ANALYST 4
MANAGER 3
PRESIDENT       1
SALESMAN        4
SOFTWAREENGINEER        3
TECHNICALLEAD   1
```

SIDDHI S.KOTRE
ROLL NO: 50

b. Display all the MANAGERS in order where the one who served the most appears first.

```
hive> SELECT empfname, emplname, hiredate
    > FROM employee
    > WHERE job = 'MANAGER'
    > ORDER BY hiredate ASC;
Query ID = cloudera_20240829010707_2267d5f4-a4ae-487f-840d-0ff04cd1abd3
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_1724910377336_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-08-29 01:07:22,560 Stage-1 map = 0%,  reduce = 0%
2024-08-29 01:07:26,695 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.75 sec
2024-08-29 01:07:31,815 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.57 sec
MapReduce Total cumulative CPU time: 1 seconds 570 msec
Ended Job = job_1724910377336_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.57 sec   HDFS Read: 8259 HDFS Write: 66 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 570 msec
OK
KIM     JARVIS  1981-04-02
BELLA   SWAN    1981-05-01
KEVIN   HILL    1981-06-09
```

c. Display the FULL NAME and SALARY drawn by an analyst working in dept no 20

```
hive> SELECT CONCAT(empfname, ' ', emplname) AS full_name, salary
    > FROM employee
    > WHERE job = 'ANALYST' AND deptcode = 20;
OK
CONNIE SMITH    3000
MADII HIMBURY   2000
Time taken: 0.058 seconds, Fetched: 2 row(s)
```

SIDDHI S.KOTRE
ROLL NO: 50

7. Perform the following joins between employee and dept table using key deptcode

a. Left Join

```
hive> SELECT e.*, d.deptname, d.location
    > FROM employee e
    > LEFT JOIN dept d
    > ON e.deptcode = d.deptcode;
Query ID = cloudera_20240829012525_09711e45-4174-4d55-8e54-3a47fbd2629e
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240829012525_09711e45-4174-4d55-8e54-3a47fbd2629e.log
2024-08-29 01:25:19     Starting to launch local task to process map join;      maximum memory = 49807360
2024-08-29 01:25:20     Dump the side-table for tag: 1 with group count: 5 into file: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a
605-22ceb4234867/hive_2024-08-29_01-25-17_803_4782246135174257994-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashta
ble
2024-08-29 01:25:20     Uploaded 1 File to: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a605-22ceb4234867/hive_2024-08-29_01-25-17_
803_4782246135174257994-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (447 bytes)
2024-08-29 01:25:20     End of local task; Time Taken: 0.572 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_1724910377336_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0008
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-08-29 01:25:25,089 Stage-3 map = 0%,  reduce = 0%
2024-08-29 01:25:29,202 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.59 sec
MapReduce Total cumulative CPU time: 590 msec
Ended Job = job_1724910377336_0008
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 0.59 sec   HDFS Read: 8502 HDFS Write: 1163 SUCCESS
Total MapReduce CPU Time Spent: 590 msec
OK
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    0       20      SOFTWARE        PADDINGTON
9499    TIM     ADOLF   SALESMAN        7698    1981-02-20      1600    300     30      SALES   MAIDSTONE
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20      SOFTWARE        PADDINGTON
9654    SAM     MILES   SALESMAN        7698    1981-09-28      1250    1400    30      SALES   MAIDSTONE
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10      FINANCE EDINBURGH
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20      SOFTWARE        PADDINGTON
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       10      FINANCE EDINBURGH
9844    PAUL    TIMOTHY SALESMAN        7698    1981-09-08      1500    0       30      SALES   MAIDSTONE
9876    JOHN    ASGHAR  SOFTWAREENGINEER        7788    1983-01-12      3100    0       20      SOFTWARE        PADDINGTON
9900    ROSE    SUMMERS TECHNICALLEAD   7698    1981-12-03      2950    0       20      SOFTWARE        PADDINGTON
9902    ANDREW  FAULKNER        ANALYST 7566    1981-12-03      3000    0       10      FINANCE EDINBURGH
9934    KAREN   MATTHEWS        SOFTWAREENGINEER        7782    1982-01-23      3300    0       20      SOFTWARE        PADDI
NGTON
9591    WENDY   SHAWN   SALESMAN        7698    1981-02-22      500     0       30      SALES   MAIDSTONE
9698    BELLA   SWAN    MANAGER 7839    1981-05-01      3420    0       30      SALES   MAIDSTONE
9777    MADII   HIMBURY ANALYST 7839    1981-05-01      2000    200     20      SOFTWARE        PADDINGTON
```

SIDDHI S.KOTRE
ROLL NO: 50

b. Right Join

```
hive> SELECT e.*, d.deptname, d.location
    > FROM employee e
    > RIGHT JOIN dept d
    > ON e.deptcode = d.deptcode;
Query ID = cloudera_20240829012626_2c745285-ba4f-4b67-b81f-cd2116051e42
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240829012626_2c745285-ba4f-4b67-b81f-cd2116051e42.log
2024-08-29 01:27:01     Starting to launch local task to process map join;      maximum memory = 49807360
2024-08-29 01:27:01     Dump the side-table for tag: 0 with group count: 4 into file: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a
605-22ceb4234867/hive_2024-08-29_01-26-59_438_3362757624584638640-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashta
ble
2024-08-29 01:27:01     Uploaded 1 File to: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a605-22ceb4234867/hive_2024-08-29_01-26-59_
438_3362757624584638640-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable (1140 bytes)
2024-08-29 01:27:01     End of local task; Time Taken: 0.59 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_1724910377336_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0009
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-08-29 01:27:06,212 Stage-3 map = 0%,  reduce = 0%
2024-08-29 01:27:10,306 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.66 sec
MapReduce Total cumulative CPU time: 660 msec
Ended Job = job_1724910377336_0009
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 0.66 sec   HDFS Read: 7718 HDFS Write: 1211 SUCCESS
Total MapReduce CPU Time Spent: 660 msec
OK
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10      FINANCE EDINBURGH
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       10      FINANCE EDINBURGH
9902    ANDREW  FAULKNER        ANALYST 7566    1981-12-03      3000    0       10      FINANCE EDINBURGH
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    0       20      SOFTWARE        PADDINGTON
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20      SOFTWARE        PADDINGTON
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20      SOFTWARE        PADDINGTON
9876    JOHN    ASGHAR  SOFTWAREENGINEER        7788    1983-01-12      3100    0       20      SOFTWARE        PADDINGTON
9900    ROSE    SUMMERS TECHNICALLEAD   7698    1981-12-03      2950    0       20      SOFTWARE        PADDINGTON
9934    KAREN   MATTHEWS        SOFTWAREENGINEER        7782    1982-01-23      3300    0       20      SOFTWARE        PADDI
NGTON
9777    MARII   HIMPHRY ANALYST 7839    1981-05-01      3000    200     20      SOFTWARE        PADDINGTON
```

SIDDHI S.KOTRE
ROLL NO: 50

c. Full Join

```
hive> SELECT e.*, d.deptname, d.location
    > FROM employee e
    > FULL OUTER JOIN dept d
    > ON e.deptcode = d.deptcode;
Query ID = cloudera_20240829012828_1e2ec380-457e-4294-a1b0-2acd38a10fbb
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_1724910377336_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0010
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2024-08-29 01:28:22,974 Stage-1 map = 0%,  reduce = 0%
2024-08-29 01:28:29,165 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.01 sec
2024-08-29 01:28:34,323 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.81 sec
MapReduce Total cumulative CPU time: 1 seconds 810 msec
Ended Job = job_1724910377336_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 1.81 sec   HDFS Read: 16042 HDFS Write: 1211 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 810 msec
OK
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10      FINANCE EDINBURGH
9902    ANDREW  FAULKNER        ANALYST 7566    1981-12-03      3000    0       10      FINANCE EDINBURGH
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       10      FINANCE EDINBURGH
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20      SOFTWARE        PADDINGTON
9777    MADII   HIMBURY ANALYST 7839    1981-05-01      2000    200     20      SOFTWARE        PADDINGTON
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20      SOFTWARE        PADDINGTON
9934    KAREN   MATTHEWS        SOFTWAREENGINEER        7782    1982-01-23      3300    0       20      SOFTWARE        PADDI
NGTON
9900    ROSE    SUMMERS TECHNICALLEAD   7698    1981-12-03      2950    0       20      SOFTWARE        PADDINGTON
9876    JOHN    ASGHAR  SOFTWAREENGINEER        7788    1983-01-12      3100    0       20      SOFTWARE        PADDINGTON
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    0       20      SOFTWARE        PADDINGTON
9844    PAUL    TIMOTHY SALESMAN        7698    1981-09-08      1500    0       30      SALES   MAIDSTONE
9698    BELLA   SWAN    MANAGER 7839    1981-05-01      3420    0       30      SALES   MAIDSTONE
9591    WENDY   SHAWN   SALESMAN        7698    1981-02-22      500     0       30      SALES   MAIDSTONE
9654    SAM     MILES   SALESMAN        7698    1981-09-28      1250    1400    30      SALES   MAIDSTONE
```

d. Inner Join

```
hive> SELECT e.*, d.deptname, d.location
    > FROM employee e
    > INNER JOIN dept d
    > ON e.deptcode = d.deptcode;
Query ID = cloudera_20240829012929_e3645afe-8e46-446e-958a-4a73f3ebbc4f
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240829012929_e3645afe-8e46-446e-958a-4a73f3ebbc4f.log
2024-08-29 01:29:43     Starting to launch local task to process map join;      maximum memory = 49807360
2024-08-29 01:29:43     Dump the side-table for tag: 1 with group count: 5 into file: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a
605-22ceb4234867/hive_2024-08-29_01-29-41_275_8282792907026899814-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashta
ble
2024-08-29 01:29:43     Uploaded 1 File to: file:/tmp/cloudera/f4b5c355-58ba-4f1e-a605-22ceb4234867/hive_2024-08-29_01-29-41_
275_8282792907026899814-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashtable (447 bytes)
2024-08-29 01:29:43     End of local task; Time Taken: 0.797 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_1724910377336_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0011
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-08-29 01:29:49,734 Stage-3 map = 0%,  reduce = 0%
2024-08-29 01:29:53,820 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.86 sec
MapReduce Total cumulative CPU time: 860 msec
Ended Job = job_1724910377336_0011
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 0.86 sec   HDFS Read: 8722 HDFS Write: 1163 SUCCESS
Total MapReduce CPU Time Spent: 860 msec
OK
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    0       20      SOFTWARE        PADDINGTON
9499    TIM     ADOLF   SALESMAN        7698    1981-02-20      1600    300     30      SALES   MAIDSTONE
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20      SOFTWARE        PADDINGTON
9654    SAM     MILES   SALESMAN        7698    1981-09-28      1250    1400    30      SALES   MAIDSTONE
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10      FINANCE EDINBURGH
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20      SOFTWARE        PADDINGTON
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       10      FINANCE EDINBURGH
9844    PAUL    TIMOTHY SALESMAN        7698    1981-09-08      1500    0       30      SALES   MAIDSTONE
9876    JOHN    ASGHAR  SOFTWAREENGINEER        7788    1983-01-12      3100    0       20      SOFTWARE        PADDINGTON
9900    ROSE    SUMMERS TECHNICALLEAD   7698    1981-12-03      2950    0       20      SOFTWARE        PADDINGTON
9902    ANDREW  FAULKNER        ANALYST 7566    1981-12-03      3000    0       10      FINANCE EDINBURGH
9934    KAREN   MATTHEWS        SOFTWAREENGINEER        7782    1982-01-23      3300    0       20      SOFTWARE        PADDI
```

SIDDHI S.KOTRE
ROLL NO: 50

8. Create a view on the employee table for employees whose salary is greater than 2000 and then drop that view.

```
hive> CREATE VIEW high_salary_employees AS
    > SELECT *
    > FROM employee
    > WHERE salary > 2000;
OK
Time taken: 0.099 seconds
hive> SELECT * FROM high_salary_employees;
OK
9369    TONY    STARK   SOFTWAREENGINEER        7902    1980-12-17      2800    0       20
9566    KIM     JARVIS  MANAGER 7839    1981-04-02      3570    0       20
9782    KEVIN   HILL    MANAGER 7839    1981-06-09      2940    0       10
9788    CONNIE  SMITH   ANALYST 7566    1982-12-09      3000    0       20
9839    ALFRED  KINSLEY PRESIDENT       7566    1981-11-17      5000    0       10
9876    JOHN    ASGHAR  SOFTWAREENGINEER        7788    1983-01-12      3100    0       20
9900    ROSE    SUMMERS TECHNICALLEAD   7698    1981-12-03      2950    0       20
9902    ANDREW  FAULKNER        ANALYST 7566    1981-12-03      3000    0       10
9934    KAREN   MATTHEWS        SOFTWAREENGINEER        7782    1982-01-23      3300    0       20
9698    BELLA   SWAN    MANAGER 7839    1981-05-01      3420    0       30
9860    ATHENA  WILSON  ANALYST 7839    1992-06-21      7000    100     50
Time taken: 0.052 seconds, Fetched: 11 row(s)
hive> DROP VIEW high_salary_employees;
OK
Time taken: 0.045 seconds
```

9. Create an index on the employee table then drop that view.

```
hive> CREATE INDEX idx_salary
    > ON TABLE employee (salary)
    > AS 'COMPACT'
    > WITH DEFERRED REBUILD;
OK
Time taken: 0.222 seconds
hive> ALTER INDEX idx_salary ON employee REBUILD;
Query ID = cloudera_20240829013535_a0c26686-f3ca-4400-975e-2304346153e0
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_1724910377336_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1724910377336_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1724910377336_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-08-29 01:35:42,449 Stage-1 map = 0%,  reduce = 0%
2024-08-29 01:35:47,552 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.52 sec
2024-08-29 01:35:52,683 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.31 sec
MapReduce Total cumulative CPU time: 1 seconds 310 msec
Ended Job = job_1724910377336_0012
Loading data to table timscdr.timscdr__employee_idx_salary__
Table timscdr.timscdr__employee_idx_salary__ stats: [numFiles=1, numRows=15, totalSize=1485, rawDataSize=1470]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 1.31 sec   HDFS Read: 9795 HDFS Write: 1581 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 310 msec
OK
Time taken: 15.25 seconds
hive> DROP INDEX idx_salary ON TABLE employee;
```

SIDDHI S.KOTRE
ROLL NO: 50

10. Hive Partitioning
Student table
roll INT,
stfname STRING,
stlname STRING,
course STRING
Create a partitioned table Student_course based on course and Demonstrate static and dynamic partitioning collection

```
[cloudera@quickstart ~]$ nano student_data_all.txt
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/student_data_all.txt /use
r/hive/warehouse/timscdr/
[cloudera@quickstart ~]$ hdfs dfs -cat /user/hive/warehouse/timscdr/student_data
_all.txt
1,TONY,STARK,JAVA
2,TIM,ADOLF,HTML
3,KIM,JARVIS,EXCEL
4,SAM,MILES,HTML
5,KEVIN,HILL,EXCEL
6,CONNIE,SMITH,HADOOP
7,ALFRED,KINSLEY,PYTHON
8,PAUL,TIMOTHY,HTML
9,JOHN,ASGHAR,JAVA
10,ROSE,SUMMERS,DBA
11,ANDREW,FAULKNER,HADOOP
12,KAREN,MATTHEWS,JAVA
13,WENDY,SHAWN,HTML
14,BELLA,SWAN,EXCEL
15,MADII,HIMBURY,HADOOP
16,ATHENA,WILSON,HADOOP
hive> CREATE TABLE Student (
    >     roll INT,
    >     stfname STRING,
    >     stlname STRING,
    >     course STRING
    > ) row format delimited fields terminated by ',';
OK
Time taken: 0.035 seconds
```

SIDDHI S.KOTRE
ROLL NO: 50

```
hive> LOAD DATA INPATH '/user/hive/warehouse/timscdr/student_data_all.txt' INTO TABLE Student;
Loading data to table timscdr.student
Table timscdr.student stats: [numFiles=1, totalSize=333]
OK
Time taken: 0.211 seconds
hive> SELECT * FROM Student;
OK
1       TONY    STARK   JAVA
2       TIM     ADOLF   HTML
3       KIM     JARVIS  EXCEL
4       SAM     MILES   HTML
5       KEVIN   HILL    EXCEL
6       CONNIE  SMITH   HADOOP
7       ALFRED  KINSLEY PYTHON
8       PAUL    TIMOTHY HTML
9       JOHN    ASGHAR  JAVA
10      ROSE    SUMMERS DBA
11      ANDREW  FAULKNER        HADOOP
12      KAREN   MATTHEWS        JAVA
13      WENDY   SHAWN   HTML
14      BELLA   SWAN    EXCEL
15      MADII   HIMBURY HADOOP
16      ATHENA  WILSON  HADOOP
NULL    NULL    NULL    NULL
Time taken: 0.048 seconds, Fetched: 17 row(s)
hive> SET hive.exec.dynamic.partition = true;
hive> SET hive.exec.dynamic.partition.mode = nonstrict;
```