Experiment 4

Aim: SQL Queries with Hive: Write and execute SQL queries using Apache Hive to analyze structured data stored in HDFS, and perform operations such as filtering, joining, and aggregating data.

Hive Commands and Output:

```
hive> show databases;
OK
default
madhurima_database
Time taken: 0.037 seconds, Fetched: 2 row(s)
hive> show databases;
OK
default
madhurima_database
Time taken: 0.286 seconds, Fetched: 2 row(s)
hive> use madhurima_database;
OK
Time taken: 0.05 seconds
hive> CREATE TABLE Customers(customer_id INT, name STRING, city STRING) ROW FORM
AT  DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.956 seconds
hive> CREATE TABLE Orders(order_id INT, customer_id INT, amount FLOAT, date  STR
ING) ROW FORMAT  DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.111 seconds
hive> SHOW TABLES;
OK
customers
employees
orders
student
student_records
Time taken: 0.113 seconds, Fetched: 5 row(s)
hive> DESCRIBE madhurima_database.customers;
OK
customer_id             int
name                    string
city                    string
Time taken: 0.107 seconds, Fetched: 3 row(s)
hive> DESCRIBE madhurima_database.orders;
OK
order_id                int
customer_id             int
amount                  float
date                    string
Time taken: 0.08 seconds, Fetched: 4 row(s)
```

Fig 1: Database Commands and Table Creation

```
hive> LOAD DATA LOCAL INPATH 'file:///home/cloudera/Downloads/Customers.txt' INT
O TABLE customers;
Loading data to table madhurima_database.customers
Table madhurima_database.customers stats: [numFiles=1, totalSize=96]
OK
Time taken: 0.693 seconds
```

Fig 2: Loading Data in Customer Table

```
hive> LOAD DATA LOCAL INPATH 'file:///home/cloudera/Downloads/Orders.txt' INTO T
ABLE orders;
Loading data to table madhurima_database.orders
Table madhurima_database.orders stats: [numFiles=1, totalSize=190]
OK
Time taken: 0.352 seconds
```

Fig 2: Loading Data in Orders Table

```
hive> SELECT * FROM customers;
OK
1       Alice   New York
2       Bob     Los Angeles
3       Charlie Chicago
4       Dave    New York
5       Eve     Miami
6       Frank   Boston
Time taken: 0.12 seconds, Fetched: 6 row(s)
hive> SELECT * FROM orders;
OK
101     1       100.5   2024-10-01
102     2       200.0   2024-10-02
103     1       50.75   2024-10-03
104     3       80.0    2024-10-03
105     4       300.0   2024-10-04
106     5       120.0   2024-10-05
107     2       150.0   2024-10-05
108     6       250.0   2024-10-06
Time taken: 0.055 seconds, Fetched: 8 row(s)
```

Fig 3: Displaying Created Tables

```
hive> SELECT o.order_id, o.amount, c.name, c.city
    > FROM orders o
    > JOIN customers c
    > ON o.customer_id = c.customer_id
    > WHERE c.city = 'New York';
Query ID = cloudera_20241011034040_8e1e2542-c655-4d2f-9490-e2e8d7799ebe
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20241011034040_8e1e2542-c655-4d2f-9490-
e2e8d7799ebe.log
2024-10-11 03:40:37     Starting to launch local task to process map join;     m
aximum memory = 1013645312
2024-10-11 03:40:39     Dump the side-table for tag: 1 with group count: 2 into
file: file:/tmp/cloudera/f95ff959-3534-4d4f-8dd3-5b840be8b312/hive_2024-10-11_03
-40-26_450_410633455240172418-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01
--.hashtable
2024-10-11 03:40:39     Uploaded 1 File to: file:/tmp/cloudera/f95ff959-3534-4d4
f-8dd3-5b840be8b312/hive_2024-10-11_03-40-26_450_410633455240172418-1/-local-100
03/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (309 bytes)
2024-10-11 03:40:39     End of local task; Time Taken: 2.427 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_1728640488912_0001, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1728640488912_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1728640488912_0001
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-10-11 03:40:56,655 Stage-3 map = 0%,  reduce = 0%
2024-10-11 03:41:08,828 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.94 se
c
MapReduce Total cumulative CPU time: 1 seconds 940 msec
Ended Job = job_1728640488912_0001
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.94 sec   HDFS Read: 7113 HDFS Write: 7
4 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 940 msec
OK
101     100.5   Alice   New York
103     50.75   Alice   New York
105     300.0   Dave    New York
Time taken: 44.604 seconds, Fetched: 3 row(s)
```

Fig 4: Join Operation on Tables Customers and Orders

```
hive> SELECT c.name, o.order_id, o.amount
    > FROM customers c
    > JOIN orders o
    > ON c.customer_id = o.customer_id;
Query ID = cloudera_20241011034444_6477f33c-b88b-410a-9a32-23450f8220f1
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20241011034444_6477f33c-b88b-410a-9a32-
23450f8220f1.log
2024-10-11 03:44:36     Starting to launch local task to process map join;     m
aximum memory = 1013645312
2024-10-11 03:44:38     Dump the side-table for tag: 0 with group count: 6 into
file: file:/tmp/cloudera/f95ff959-3534-4d4f-8dd3-5b840be8b312/hive_2024-10-11_03
-44-30_570_1060744741217842858-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile2
0--.hashtable
2024-10-11 03:44:38     Uploaded 1 File to: file:/tmp/cloudera/f95ff959-3534-4d4
f-8dd3-5b840be8b312/hive_2024-10-11_03-44-30_570_1060744741217842858-1/-local-10
003/HashTable-Stage-3/MapJoin-mapfile20--.hashtable (407 bytes)
2024-10-11 03:44:38     End of local task; Time Taken: 1.976 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_1728640488912_0003, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1728640488912_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1728640488912_0003
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-10-11 03:44:49,361 Stage-3 map = 0%,  reduce = 0%
2024-10-11 03:44:59,084 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.69 se
c
MapReduce Total cumulative CPU time: 1 seconds 690 msec
Ended Job = job_1728640488912_0003
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.69 sec   HDFS Read: 6988 HDFS Write: 1
22 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 690 msec
OK
Alice   101     100.5
Bob     102     200.0
Alice   103     50.75
Charlie 104     80.0
Dave    105     300.0
Eve     106     120.0
Bob     107     150.0
Frank   108     250.0
Time taken: 30.692 seconds. Fetched: 8 row(s)
```

Fig 5: Join Operation on Tables Customers and Orders

```
hive> SELECT c.name, SUM(o.amount) AS total_spent
    > FROM customers c
    > JOIN orders o
    > ON c.customer_id = o.customer_id
    > GROUP BY c.name;
Query ID = cloudera_20241011034949_bc8b5ef0-d906-494e-9a28-162f20c24f3a
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20241011034949_bc8b5ef0-d906-494e-9a28-
162f20c24f3a.log
2024-10-11 03:49:17     Starting to launch local task to process map join;     m
```

Fig 6: Aggregation Operation on Tables Customers and Orders

```
Total MapReduce CPU Time Spent: 2 seconds 200 msec
OK
Alice   151.25
Bob     350.0
Charlie 80.0
Dave    300.0
Eve     120.0
Frank   250.0
Time taken: 34.643 seconds, Fetched: 6 row(s)
```

Fig 7: Aggregation Operation Output