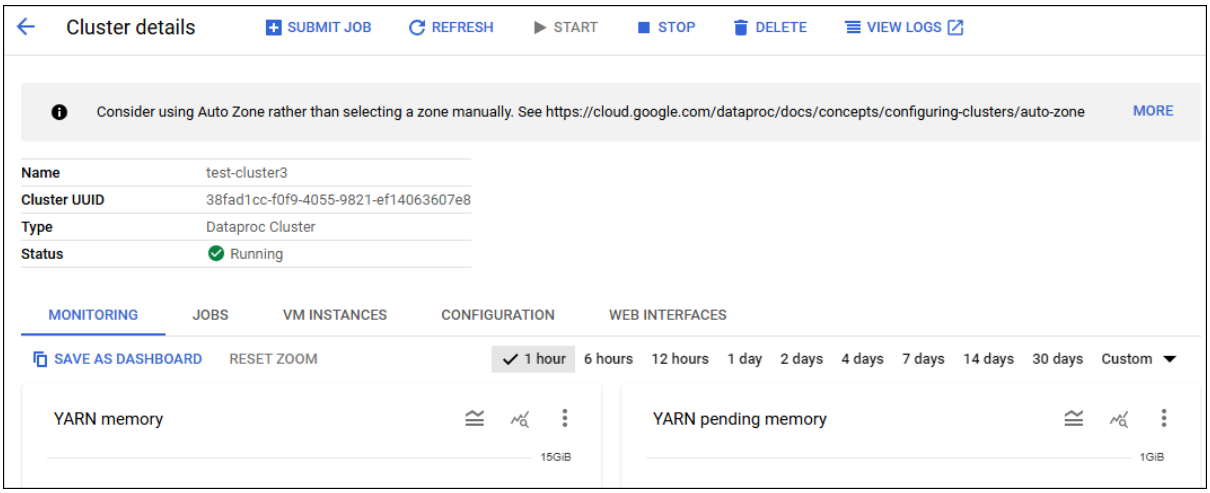
Hadoop Assignment

Task 1: Data Ingestion into HDFS

1. The Cluster is created with 1 master node and 2 worker nodes



2. The given dataset (transactions.csv) is uploaded to HDFS.

Command : hdfs dfs -put Transactions.csv /user/mounika-de

3. The data is available in HDFS and it is accessible

i. hdfs dfs -ls /user/mounika

```
loka_mounika@test-cluster3-m:~$ hdfs dfs -ls /user/mounika
Found 7 items
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-16 05:45 /user/mounika/ExportedData
-rw-r--r--   2 loka_mounika hadoop     216974 2025-02-11 21:54 /user/mounika/Transactions.csv
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-15 00:54 /user/mounika/avro
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-14 19:19 /user/mounika/csv
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-14 19:48 /user/mounika/externaltable
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-15 00:57 /user/mounika/orc
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-15 00:53 /user/mounika/parquet
```

ii. hadoop fs -head /user/mounika-de/Transactions.csv

```
loka_mounika@test-cluster3-m:~$ hadoop fs -head /user/mounika/Transactions.csv
transaction_id,user_id,amount,transaction_date,transaction_type
TXN000001,U0440,651.22,2025-01-06,deposit
TXN000002,U0367,727.29,2025-01-23,withdrawal
TXN000003,U0233,262.94,2025-01-07,refund
TXN000004,U0139,269.5,2025-02-04,refund
TXN000005,U0086,552.44,2025-02-02,purchase
TXN000006,U0200,363.37,2025-01-06,withdrawal
TXN000007,U0017,238.59,2025-02-05,refund
TXN000008,U0226,135.88,2025-01-13,refund
TXN000009,U0090,196.56,2025-01-08,subscription
TXN000010,U0407,449.26,2025-01-28,refund
TXN000011,U0229,363.51,2025-02-02,withdrawal
TXN000012,U0349,133.6,2025-01-19,deposit
TXN000013,U0290,612.98,2025-01-11,withdrawal
TXN000014,U0152,916.83,2025-01-28,purchase
TXN000015,U0087,417.06,2025-01-24,purchase
TXN000016,U0150,626.93,2025-01-13,purchase
TXN000017,U0063,941.1,2025-01-19,withdrawal
TXN000018,U0342,289.85,2025-01-11,subscription
TXN000019,U0453,14.31,2025-01-01,deposit
TXN000020,U0339,314.36,2025-01-05,deposit
TXN000021,U0396,734.01,2025-02-04,refund
TXN000022,U0423,559.32,2025-01-11,deposit
```

Task 2: Creating Hive Tables

1. External hive table

create external table extern_txns (transaction_id string, user_id string, amount int, transaction_date int, transaction_type string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/mounika-de/tabledata/' tblproperties("skip.header.line.count"="1");

2. Internal hive table:

creating table:

create table internal_txns(transaction_id string, user_id string, amount float, transaction_date DATE, transaction_type string) row format delimited fields terminated by ',' tblproperties("skip.header.line.count"="1");

Loading data from local to hadoop default location:

hadoop fs -put /home/loka_mounika/Transactions_New.csv /user/hive/warehouse/mounika_de.db/internal_txns/

Verifying the data:

hadoop fs -ls /user/hive/warehouse/mounika_de.db/internal_txns

```
loka_mounika@test-cluster3-m:~$ hadoop fs -ls /user/hive/warehouse/mounika_de.db/internal_txns
Found 1 items
-rw-r--r--   2 loka_mounika hadoop     216974 2025-02-15 00:09 /user/hive/warehouse/mounika_de.db/internal_txns/Transactions_New.csv
```

--->Observation:

1. External table will create on top of the path we have specified where as internal table creates on hive default location.

2. external table will read the data automatically from the location

Task 3: Data Analysis Using Hive

a. On Managed Table

1. **Count the total number of transactions**.

Query:

select count(t.transaction_id) from internal_txns t;

```
hive> select count(t.transaction_id) from internal_txns t;
Query ID = loka_mounika_20250216214952_2c72b866-0669-4689-b445-84a721d71653
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.85 s
--------------------------------------------------------------------------------
OK
5000
Time taken: 11.981 seconds, Fetched: 1 row(s)
```

There are total 5000 transactions.

Time taken: 7.85 seconds

2. **Find the total transaction amount grouped by transaction type**.

Query:

select sum(internal_txns.amount) as Total_Amount, internal_txns.transaction_type from internal_txns Group By internal_txns.transaction_type;

```
hive>
    >
    > select sum(internal_txns.amount) as Total_Amount, internal_txns.transaction_type from internal_txns Group By internal_txns.transactio
n_type;
Query ID = loka_mounika_20250216215049_d0fb9a85-840d-4d2d-9af4-094410a017a0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.93 s
----------------------------------------------------------------------------------------------
OK
521239.1999999995       deposit
513125.04999999964      purchase
498754.4599999998       refund
501381.7699999997       subscription
475506.7099999996       withdrawal
Time taken: 8.401 seconds, Fetched: 5 row(s)
```

3. **Identify the user who has spent the highest amount**.

Query:

SELECT t.transaction_id, SUM(t.amount) AS TotalSpent FROM internal_txns t GROUP BY

t.transaction_id ORDER BY TotalSpent DESC LIMIT 1;

```
hive> SELECT t.transaction_id, SUM(t.amount) AS TotalSpent FROM internal_txns t GROUP BY
    >  t.transaction_id ORDER BY TotalSpent DESC LIMIT 1;
Query ID = loka_mounika_20250216215145_fea0bcad-05c3-4436-af7a-846d3ecab097
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.65 s
----------------------------------------------------------------------------------------------
OK
TXN004453       1000.0
Time taken: 9.158 seconds, Fetched: 1 row(s)
```

Here, the user with the transaction id TXN004453 has spent the maximum amount which is 1000.

Time taken 8.433 sec

4. **Retrieve all transactions from the last 7 days**.

Query:

select * from internal_txns t where to_date(t.transaction_date) >= date_sub(current_date, 7) and to_date(t.transaction_date) < current_date;

<no records found for recent 7 days, so gave 11 days to fetch some records>

```
hive> select * from internal_txns t where to_date(t.transaction_date) >= date_sub(current_date, 11) and to_date(t.transaction_date) < curre
nt_date;
Query ID = loka_mounika_20250216215316_0d067138-c5f2-4125-bacc-b8569986dafb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 8.11 s
--------------------------------------------------------------------------------
OK
TXN000007       U0017   238.59  2025-02-05      refund
TXN000060       U0424   253.31  2025-02-05      subscription
TXN000150       U0043   335.06  2025-02-05      deposit
TXN000161       U0174   572.96  2025-02-05      refund
```

There are total 144 transactions happened in the last 11 days

b. On External Hive Table:

1. **Count the total number of transactions**.

Query:

select count(e.transaction_id) from external_txns e;

```
hive> select count(e.transaction_id) from external_txns e;
Query ID = loka_mounika_20250216215441_8af9a094-feff-44f4-aded-4b4b3dcae648
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.34 s
--------------------------------------------------------------------------------
OK
5000
Time taken: 7.693 seconds, Fetched: 1 row(s)
```

There are total 5000 transactions

Time taken: 7.693 sec

2. **Find the total transaction amount grouped by transaction type**.

Query:

select sum(e.amount) as Total_Amount, e.transaction_type from external_txns e Group By e.transaction_type;

```
hive> select sum(e.amount) as Total_Amount, e.transaction_type from external_txns e Group By e.transaction_type;
Query ID = loka_mounika_20250216215534_f260e672-38af-4ff1-b3ad-d7f4dee6372a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.20 s
--------------------------------------------------------------------------------
OK
521239.1997451782       deposit
513125.05059432983      purchase
498754.4605484009       refund
501381.7700147629       subscription
475506.70995140076      withdrawal
Time taken: 8.55 seconds, Fetched: 5 row(s)
```

3. **Identify the user who has spent the highest amount.**

Query:

SELECT e.transaction_id, SUM(e.amount) AS TotalSpent FROM external_txns e GROUP BY e.transaction_id ORDER BY TotalSpent DESC LIMIT 1;

```
hive> SELECT e.transaction_id, SUM(e.amount) AS TotalSpent FROM external_txns e GROUP BY e.transaction_id ORDER BY TotalSpent DESC LIMIT 1;

Query ID = loka_mounika_20250216215610_d1774de3-c040-4f6c-aa16-2f9df22fb0ce
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0068)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.99 s
--------------------------------------------------------------------------------
OK
TXN004453       1000.0
Time taken: 8.478 seconds, Fetched: 1 row(s)
```

The user with the transaction id TXN004453 has spent the max amount whih is 1000

4. **Retrieve all transactions from the last 7 days**.

Query: select * from external_txns e where to_date(e.transaction_date) >= date_sub(current_date, 11) and to_date(e.transaction_date) < current_date;

```
hive> select * from external_txns e where to_date(e.transaction_date) >= date_sub(current_date, 11) and to_date(e.transaction_date) < curre
nt_date;
Query ID = loka_mounika_20250216220339_4f32578d-a037-434c-8e82-e2f59a56442a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739210662788_0069)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 8.52 s
--------------------------------------------------------------------------------
OK
TXN000007       U0017   238.59  2025-02-05      refund
TXN000060       U0424   253.31  2025-02-05      subscription
TXN000150       U0043   335.06  2025-02-05      deposit
TXN000161       U0174   572.96  2025-02-05      refund
TXN000227       U0093   358.69  2025-02-05      subscription
TXN000269       U0282   973.84  2025-02-05      refund
```

--->Observation:

- Both external and managed tables are giving the same data but external table performs better as it reads data directly from HDFS, whereas internal table reads data from Hive's internal storage hence, it takes more time.

Task 4: Performance Optimization

1. **Create a partitioned external hive table based on the transaction_type column.**

Query:

create external table partitioned_txns (transaction_id string, user_id string, amount DOUBLE, transaction_date DATE) PARTITIONED BY(transaction_type string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/hive/warehouse/partitioned_txns/' tblproperties("skip.header.line.count"="1");

2. Load data into the partitioned table from non-partitioned table.

Query:

INSERT OVERWRITE TABLE partitioned_txns PARTITION(transaction_type) SELECT transaction_id,user_id,amount,transaction_date,transaction_type FROM mounika_de.external_txns;

```
hive> select * from partitioned_txns limit 5;
OK
TXN000001       U0440   651.219970703125        2025-01-06      deposit
TXN000012       U0349   133.60000610351562      2025-01-19      deposit
TXN000019       U0453   14.3100004196167        2025-01-01      deposit
TXN000020       U0339   314.3599853515625       2025-01-05      deposit
TXN000022       U0423   559.3200073242188       2025-01-11      deposit
Time taken: 0.152 seconds, Fetched: 5 row(s)
```

3. verifying the data :

hadoop fs -ls /user/hive/warehouse/partitioned_txns**/**

```
loka_mounika@test-cluster3-m:~$ hdfs dfs -ls /user/hive/warehouse/partitioned_txns
Found 5 items
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-13 09:08 /user/hive/warehouse/partitioned_txns/transaction_type=deposit
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-13 09:08 /user/hive/warehouse/partitioned_txns/transaction_type=purchase
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-13 09:08 /user/hive/warehouse/partitioned_txns/transaction_type=refund
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-13 09:08 /user/hive/warehouse/partitioned_txns/transaction_type=subscription
drwxr-xr-x   - loka_mounika hadoop          0 2025-02-13 09:08 /user/hive/warehouse/partitioned_txns/transaction_type=withdrawal
```

4. Test the query performance on partitioned and non-partitoned tables.

i. non-partitoned table

Query:

SELECT * from internal_txns WHERE transaction_type='deposit';

Time taken: 8.749 seconds, Fetched: 1030 row(s)

ii. partitoned table

Query:

SELECT * from partitioned_txns WHERE transaction_type='deposit';

Time taken: 0.205 seconds, Fetched: 1030 row(s)

--->Observation:

Partitioned tables are significantly faster due to its ability to scan purticular row based on the query and whereas non-partitioned tables are slower as it scans entire table to fetch the data.

Task 5: Data Storage Format Comparison

1. external Hive tables that store the same raw data but in different formats:

- Avro Format:

create external table external_avro(transaction_id string, user_id string, amount DOUBLE, transaction_date DATE, transaction_type string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS AVRO LOCATION '/user/mounika/avro/';

Loading Data to Table: insert overwrite table external_avro select * from internal_txns;

```
hive> select * from external_avro limit 5;
OK
TXN000001       U0440   651.22  2025-01-06      deposit
TXN000002       U0367   727.29  2025-01-23      withdrawal
TXN000003       U0233   262.94  2025-01-07      refund
TXN000004       U0139   269.5   2025-02-04      refund
TXN000005       U0086   552.44  2025-02-02      purchase
Time taken: 0.215 seconds, Fetched: 5 row(s)
```

- ORC Format:

Command: create external table external_orc(transaction_id string, user_id string, amount DOUBLE, transaction_date DATE, transaction_type string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS orc LOCATION '/user/mounika/orc/';

Loading Data to Table:

Command: insert overwrite table external_orc select * from internal_txns;

```
hive> select * from external_orc limit 5;
OK
TXN000001       U0440   651.22  2025-01-06      deposit
TXN000002       U0367   727.29  2025-01-23      withdrawal
TXN000003       U0233   262.94  2025-01-07      refund
TXN000004       U0139   269.5   2025-02-04      refund
TXN000005       U0086   552.44  2025-02-02      purchase
Time taken: 0.14 seconds, Fetched: 5 row(s)
```

- Parquet Format:

Command:

create external table external_parquet(transaction_id string, user_id string, amount DOUBLE, transaction_date DATE, transaction_type string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS parquet LOCATION '/user/mounika/parquet/';

Loading Data to Table:

Command:

insert overwrite table external_parquet select * from internal_txns;

```
hive> select * from external_parquet limit 5;
OK
TXN000001       U0440   651.22  2025-01-06      deposit
TXN000002       U0367   727.29  2025-01-23      withdrawal
TXN000003       U0233   262.94  2025-01-07      refund
TXN000004       U0139   269.5   2025-02-04      refund
TXN000005       U0086   552.44  2025-02-02      purchase
Time taken: 0.145 seconds, Fetched: 5 row(s)
```

2. Run a simple GROUP BY query on each table create above

- **AVRO Table**:

Query: select sum(external_avro.amount) as Total_Amount,external_avro.transaction_type from external_avro Group By external_avro.transaction_type;

```
hive> select sum(external_avro.amount) as Total_Amount,external_avro.transaction_type from external_avro Group By external_avro.transaction
_type;
Query ID = loka_mounika_20250216222158_26799b8a-71f5-4560-9f33-bf96fc3f08c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0070)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.65 s
----------------------------------------------------------------------------------------------
OK
521239.1997451782       deposit
513125.05059432983      purchase
498754.4605484009       refund
501381.7700147629       subscription
475506.70995140076      withdrawal
Time taken: 9.0 seconds, Fetched: 5 row(s)
```

- ORC Table:

Query: select sum(external_orc.amount) as Total_Amount,external_orc.transaction_type from external_orc Group By external_orc.transaction_type;

```
hive> select sum(external_orc.amount) as Total_Amount,external_orc.transaction_type from external_orc Group By external_orc.transaction_typ
e;
Query ID = loka_mounika_20250216222234_f544dcfd-4a9b-42fa-be99-9e3d03e49cd3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0070)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.48 s
--------------------------------------------------------------------------------
OK
521239.1997451782       deposit
513125.05059432983      purchase
498754.4605484009       refund
501381.7700147629       subscription
475506.70995140076      withdrawal
Time taken: 7.811 seconds, Fetched: 5 row(s)
```

- Parquet Table:

Query: select sum(external_parquet.amount) as Total_Amount,external_parquet.transaction_type from external_parquet Group By external_parquet.transaction_type;

```
hive> select sum(external_parquet.amount) as Total_Amount,external_parquet.transaction_type from external_parquet Group By external_parquet
.transaction_type;
Query ID = loka_mounika_20250216222307_69b6edd3-ed35-4794-bedb-690d8aca6f64
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1739210662788_0070)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.50 s
--------------------------------------------------------------------------------
OK
521239.1997451782       deposit
513125.05059432983      purchase
498754.4605484009       refund
501381.7700147629       subscription
475506.70995140076      withdrawal
Time taken: 8.814 seconds, Fetched: 5 row(s)
```

3. Compare query execution times across different data formats.

Query:

- AVRO: Time taken: 9 sec, fetched : 5 row(s)

- ORC: Time taken: 7.811 sec, fetched : 5 row(s)

- Parquet: Time taken: 8.814 sec, fetched : 5 row(s)

- **CSV**: Time taken: 8.34 sec, fetched : 5 row(s)

4. Analyze and document the differences in performance.

- Parquet: Time taken: 0.662 sec

- Paquet is a columnar format in which all the values of single columns are stored together.

- Columnar file formats offer high performance by enabling better compression and faster retrieval of data.
- Parquet is more useful for query-intensive workloads.
- Parquet is ideal for large-scale analytics and has efficient compression capabilities.

- AVRO: Time taken: 0.771 sec

- Row-oriented storage
- Supports multiple compression techniques but may not be as efficient as parquet.
- Avro provides faster writes and slower reads

- ORC: Time taken: 8.093 sec

- ORC is column-oriented and suitable for write intensive tasks.
- This is similar to Parquet but have less community support and this might be very challenging to get assistance when issues arise.

- **CSV**: Time taken: 8.34 sec

- Easy to understand and editable in any text editors.
- This is not suitable for partitioned data storage and can not store large datasets.

Overall, the performance of parquet is best than other data formats followed by ORC, AVRO and finally CSV.

--->Observation:

Bonus Task (Optional)

Export the processed data from Hive to HDFS for further analysis

Step #1: Create a temporary table in Hive

```
create table export_temp(transaction_id string, user_id string, amount DOUBLE, transaction_date DATE, transaction_type string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE;
```

Step #2: Load exported table data to temporary table in Hive

```
INSERT OVERWRITE TABLE export_temp SELECT * FROM external_avro;
```

Step #3: Create a directory in HDFS

Command: hdfs dfs -mkdir /user/mounika/ExportedData

Step #4: Export Hive temporary table data to csv in HDFS - execute below command in HDFS

hdfs dfs -cp /user/hive/warehouse/mounika_de.db/export_temp/* /user/mounika/ExportedData/

```
loka_mounika@test-cluster3-m:~$ hdfs dfs -ls /user/mounika/ExportedData/
Found 1 items
-rw-r--r--   2 loka_mounika hadoop     270785 2025-02-16 05:45 /user/mounika/ExportedData/000000_0
```

--->Observation:

- Data is suvccessfully transferred to HDFS from hive and can be used for further analysis.