Group 28:

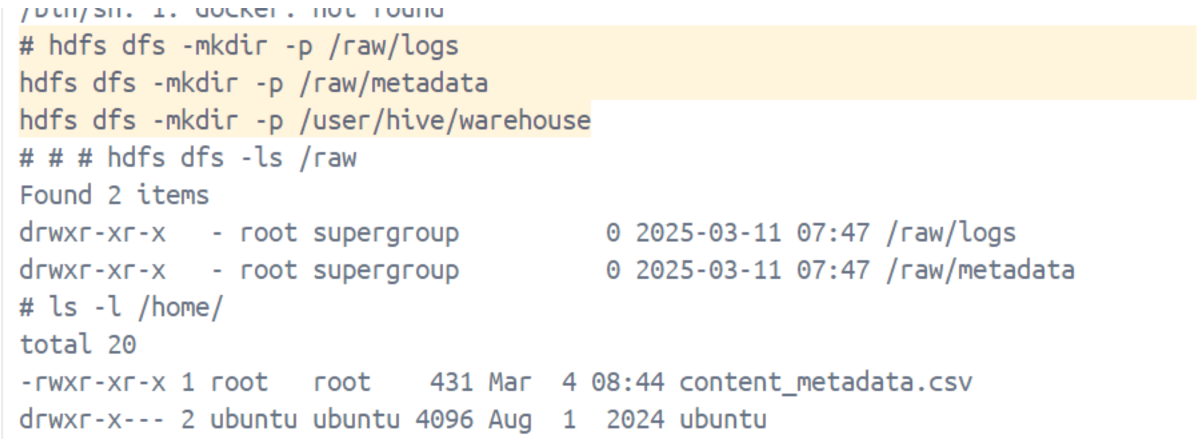
Student ID: 24280041 24280046

This document contains:

1. **Script and Screen Shot of Outputs for:**
   - Step 1: Ingestion Script
   - Step 2: Raw Tables in Hive
   - Step 3: Star Schema
   - Step 4: Transformation
   - Step 5: Queries
2. **Hive DDL from Docker Terminal With all scripts working and outputs in the same order as above**

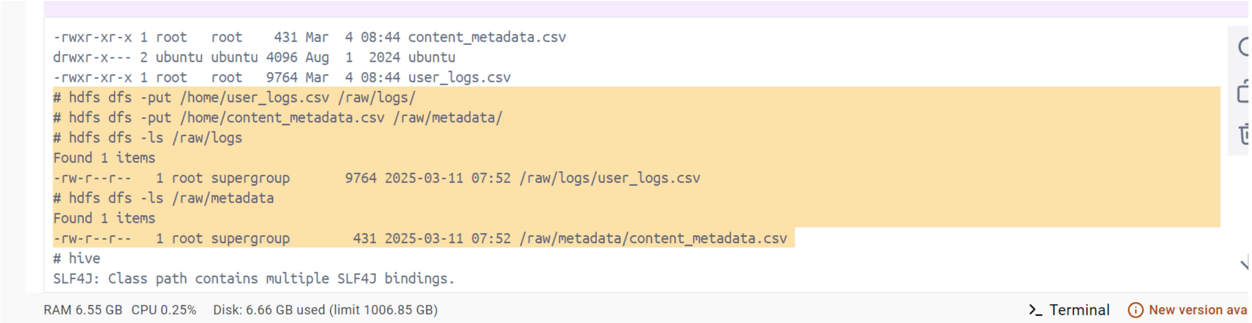
Step 1: Used these commands to ingest data manually first before making it into .sh file and enabling it.

Used these commands to make directory in HDFS



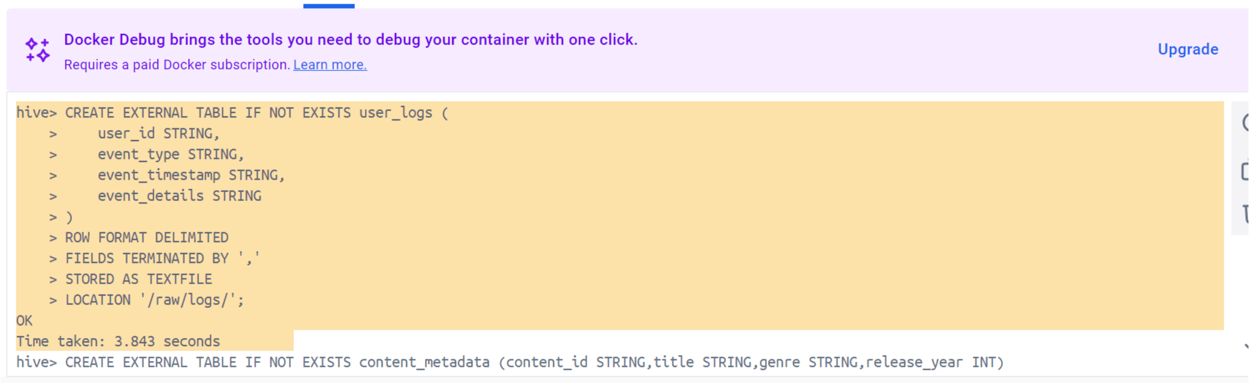
```
/bin/sh: 1: docker: not found
# hdfs dfs -mkdir -p /raw/logs
hdfs dfs -mkdir -p /raw/metadata
hdfs dfs -mkdir -p /user/hive/warehouse
# # # hdfs dfs -ls /raw
Found 2 items
drwxr-xr-x   - root supergroup          0 2025-03-11 07:47 /raw/logs
drwxr-xr-x   - root supergroup          0 2025-03-11 07:47 /raw/metadata
# ls -l /home/
total 20
-rwxr-xr-x 1 root   root    431 Mar  4 08:44 content_metadata.csv
drwxr-x--- 2 ubuntu ubuntu 4096 Aug  1  2024 ubuntu
```

Then moved files manually to the new made directories in HDFS



```
-rwxr-xr-x 1 root   root    431 Mar  4 08:44 content_metadata.csv
drwxr-x--- 2 ubuntu ubuntu 4096 Aug  1  2024 ubuntu
-rwxr-xr-x 1 root   root   9764 Mar  4 08:44 user_logs.csv
# hdfs dfs -put /home/user_logs.csv /raw/logs/
# hdfs dfs -put /home/content_metadata.csv /raw/metadata/
# hdfs dfs -ls /raw/logs
Found 1 items
-rw-r--r--   1 root supergroup       9764 2025-03-11 07:52 /raw/logs/user_logs.csv
# hdfs dfs -ls /raw/metadata
Found 1 items
-rw-r--r--   1 root supergroup        431 2025-03-11 07:52 /raw/metadata/content_metadata.csv
# hive
SLF4J: Class path contains multiple SLF4J bindings.
```

Created tables in hive to test



Docker Debug brings the tools you need to debug your container with one click.
Requires a paid Docker subscription. Learn more.

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS user_logs (
    >     user_id STRING,
    >     event_type STRING,
    >     event_timestamp STRING,
    >     event_details STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/raw/logs/';
OK
Time taken: 3.843 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS content_metadata (content_id STRING,title STRING,genre STRING,release_year INT)
```

Step 2:

External Table for user_logs (Partitioned)

Since the logs table should be partitioned by (year, month, day), run:

CREATE EXTERNAL TABLE IF NOT EXISTS raw_user_logs (

  user_id INT,

  content_id INT,

  action STRING,

  timestamp STRING,

  device STRING,

  region STRING,

  session_id STRING

```
)
PARTITIONED BY (year INT, month INT, day INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/raw/logs';
```

Add Partitions

Since Hive doesn’t automatically detect partitions for external tables, we need to add them manually:

```
ALTER TABLE raw_user_logs ADD PARTITION (year=2025, month=03, day=11) LOCATION '/raw/logs/2025/03/11/';
```

External Table for content_metadata

```
CREATE EXTERNAL TABLE IF NOT EXISTS raw_content_metadata (
   content_id INT,
   title STRING,
   category STRING,
   length INT,
   artist STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/raw/metadata';
```

Step 3:

SQL For Tables and Star Schema. (Hive DDL given at the end)

Star schema as follows:

- **Fact Table**: fact_user_activity (user interactions with content)
- **Dimension Tables**:
    - dim_users (user-related details)
    - dim_content (content metadata)
    - dim_sessions (session-related details)

dim_users Table

```
CREATE TABLE dim_users STORED AS PARQUET AS
SELECT DISTINCT user_id FROM raw_user_logs;
```

dim_content Table (Ignoring NULL columns)

sql

CopyEdit

```
CREATE TABLE dim_content STORED AS PARQUET AS
SELECT DISTINCT content_id, title, category, artist
FROM raw_content_metadata
WHERE content_id IS NOT NULL;
```

dim_sessions Table

sql

CopyEdit

```
CREATE TABLE dim_sessions STORED AS PARQUET AS
SELECT DISTINCT session_id, device, region
FROM raw_user_logs;
```

Step 4:

Fact Table & Load Data Using INSERT OVERWRITE

sql

CopyEdit

```
CREATE TABLE fact_user_activity (
  user_id INT,
  content_id INT,
  session_id STRING,
  action STRING,
  event_timestamp STRING,
  year INT,
  month INT,
  day INT
) STORED AS PARQUET;
INSERT OVERWRITE TABLE fact_user_activity
SELECT user_id, content_id, session_id, action, `timestamp`, year, month, day
FROM raw_user_logs;
```

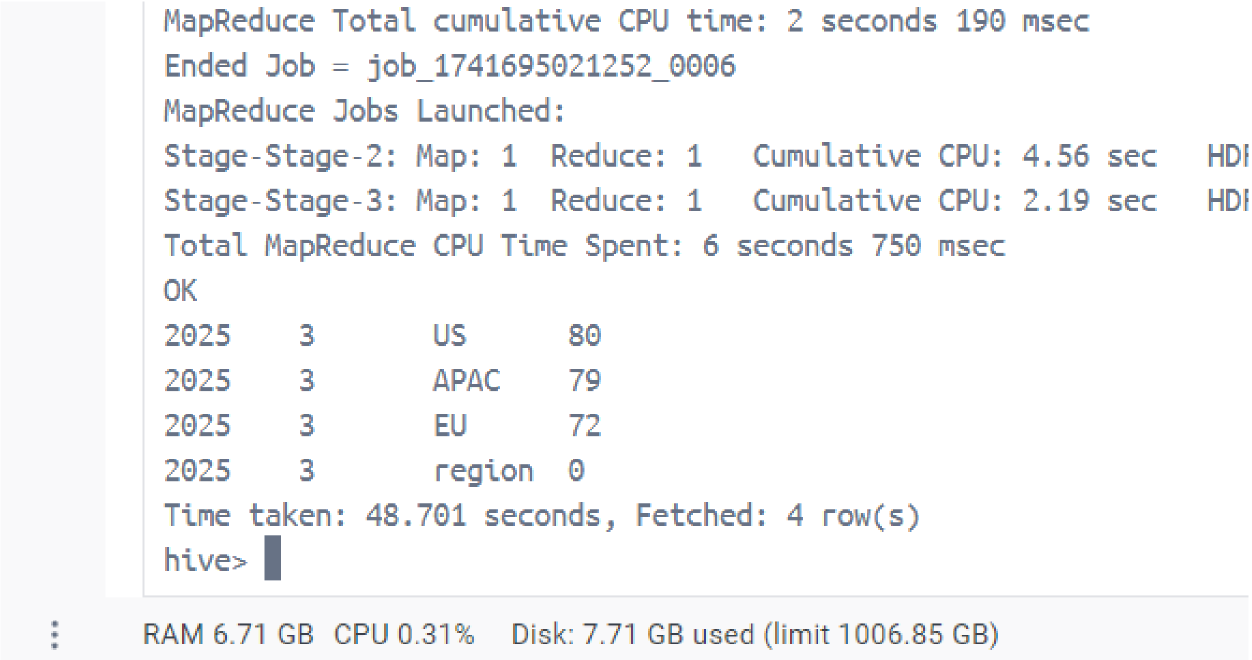
Step 5:

Query 1:

Counts distinct users per region, per month.

```
SELECT
  f.year,
  f.month,
  s.region,
  COUNT(DISTINCT f.user_id) AS monthly_active_users
FROM fact_user_activity f
JOIN dim_sessions s ON f.session_id = s.session_id
WHERE f.year = 2025  -- Example filter
GROUP BY f.year, f.month, s.region
```

ORDER BY f.year, f.month, monthly_active_users DESC;



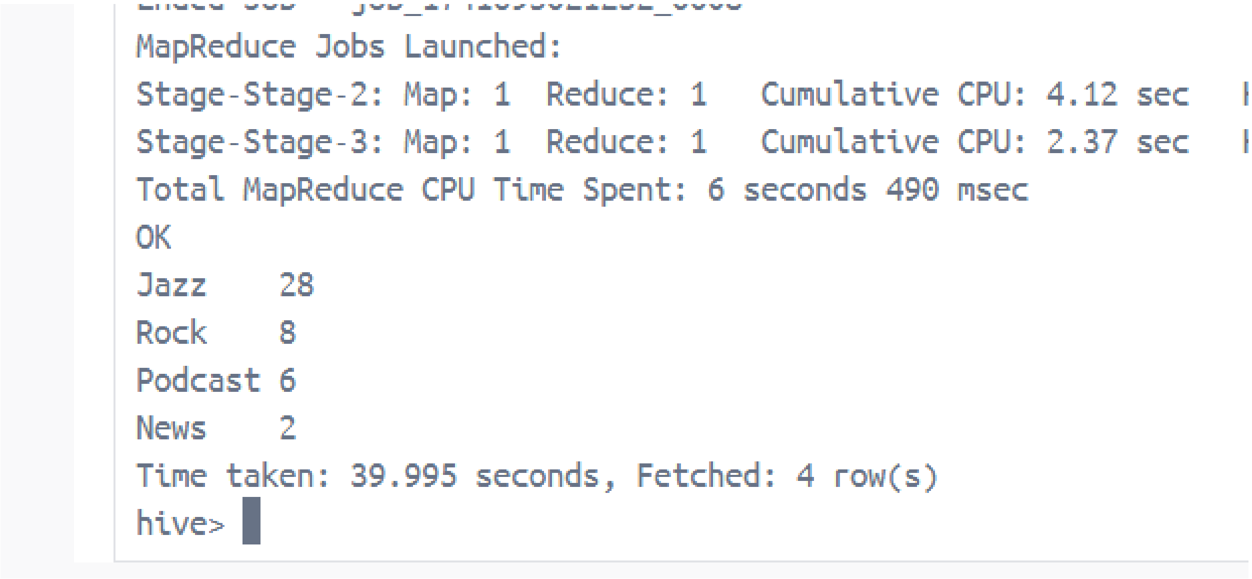
```
MapReduce Total cumulative CPU time: 2 seconds 190 msec
Ended Job = job_1741695021252_0006
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.56 sec   HDI
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 2.19 sec   HDI
Total MapReduce CPU Time Spent: 6 seconds 750 msec
OK
2025    3       US      80
2025    3       APAC    79
2025    3       EU      72
2025    3       region  0
Time taken: 48.701 seconds, Fetched: 4 row(s)
hive>
```

RAM 6.71 GB CPU 0.31%   Disk: 7.71 GB used (limit 1006.85 GB)

Query 2:

Finds the most-played content categories.

```
SELECT
   c.category,
   COUNT(*) AS play_count
FROM fact_user_activity f
JOIN dim_content c ON f.content_id = c.content_id
WHERE f.action = 'play'
AND f.year = 2025 AND f.month = 3  -- Example filter
GROUP BY c.category
ORDER BY play_count DESC
LIMIT 10;
```



```
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.12 sec
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 2.37 sec
Total MapReduce CPU Time Spent: 6 seconds 490 msec
OK
Jazz    28
Rock    8
Podcast 6
News    2
Time taken: 39.995 seconds, Fetched: 4 row(s)
hive>
```

HIVE DDL For Raw Tables and External Tables and Partitioning in HIVE

+DDL For Star Schema + DDL For Transformation + Querry 1 HIVE DDL + Query 2 DDL + Query 3 DDL

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS raw_user_logs (
  >   user_id INT,
  >   content_id INT,
  >   action STRING,
  >   timestamp STRING,
  >   device STRING,
  >   region STRING,
  >   session_id STRING
  > )
  > PARTITIONED BY (year INT, month INT, day INT)
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
```

```
   > STORED AS TEXTFILE
   > LOCATION '/raw/logs';
NoViableAltException(287@[])
      at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeOrPKOrFK(HiveParser.java:33341)
      at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeOrPKOrFKList(HiveParser.java:29513)
      at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:6175)
      at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:3808)
      at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2382)
      at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1333)
      at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
      at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
      at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
      at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
      at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
      at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
      at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
      at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
      at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
      at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
      at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
      at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
      at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
      at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
      at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
      at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
      at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
      at java.lang.reflect.Method.invoke(Method.java:498)
      at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
      at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 5:4 cannot recognize input near 'timestamp' 'STRING' ',' in column name or primary key or foreign key
hive> # hive
```

```
SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]


Logging initialized using configuration in jar:file:/opt/hive/lib/hive-common-2.3.10.jar!/hive-log4j2.properties Async: true

Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

hive> CREATE EXTERNAL TABLE IF NOT EXISTS raw_user_logs (

  >    user_id INT,

  >    content_id INT,

  >    action STRING,

  >    `timestamp` STRING,   -- Use backticks to avoid conflicts

  >    device STRING,

  >    region STRING,

  >    session_id STRING

  > )

  > PARTITIONED BY (year INT, month INT, day INT)

  > ROW FORMAT DELIMITED

  > FIELDS TERMINATED BY ','

  > STORED AS TEXTFILE

  > LOCATION '/raw/logs';

OK

Time taken: 2.912 seconds

hive> ALTER TABLE raw_user_logs ADD PARTITION (year=2025, month=03, day=11) LOCATION '/raw/logs/2025/03/11/';

OK

Time taken: 0.2 seconds

hive> CREATE EXTERNAL TABLE IF NOT EXISTS raw_content_metadata (

  >    content_id INT,

  >    title STRING,
  >    category STRING,
  >    length INT,
  >    artist STRING
  > )
  > ROW FORMAT DELIMITED
  > FIELDS TERMINATED BY ','
  > STORED AS TEXTFILE
  > LOCATION '/raw/metadata';
OK
Time taken: 0.059 seconds
hive> SHOW TABLES;
OK
raw_content_metadata
raw_user_logs
Time taken: 0.072 seconds, Fetched: 2 row(s)
hive> DESCRIBE FORMATTED raw_user_logs;
OK
# col_name            data_type             comment

user_id             int
content_id          int
action              string
timestamp           string
device              string
region              string
session_id          string

# Partition Information
# col_name            data_type             comment

year                int
```

```
month           int
day             int

# Detailed Table Information
Database:           default
Owner:              root
CreateTime:         Tue Mar 11 12:31:37 UTC 2025
LastAccessTime:     UNKNOWN
Retention:          0
Location:           hdfs://hive:9820/raw/logs
Table Type:         EXTERNAL_TABLE
Table Parameters:
    EXTERNAL            TRUE
    numFiles            1
    numPartitions       1
    numRows             0
    rawDataSize         0
    totalSize           9764
    transient_lastDdlTime   1741696297

# Storage Information
SerDe Library:      org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:        org.apache.hadoop.mapred.TextInputFormat
OutputFormat:       org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:         No
Num Buckets:        -1
Bucket Columns:     []
Sort Columns:       []
Storage Desc Params:
    field.delim         ,
    serialization.format    ,
Time taken: 0.272 seconds, Fetched: 45 row(s)

hive> DESCRIBE FORMATTED raw_content_metadata;

OK

# col_name            data_type             comment


content_id          int

title               string

category            string

length              int

artist              string


# Detailed Table Information

Database:           default

Owner:              root

CreateTime:         Tue Mar 11 12:32:34 UTC 2025

LastAccessTime:     UNKNOWN

Retention:          0

Location:           hdfs://hive:9820/raw/metadata

Table Type:         EXTERNAL_TABLE

Table Parameters:

    EXTERNAL            TRUE

    numFiles            1

    totalSize           431

    transient_lastDdlTime   1741696354


# Storage Information

SerDe Library:      org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

InputFormat:        org.apache.hadoop.mapred.TextInputFormat

OutputFormat:       org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

Compressed:         No

Num Buckets:        -1
```

```
Bucket Columns:         []

Sort Columns:           []

Storage Desc Params:

        field.delim             ,

        serialization.format    ,

Time taken: 0.053 seconds, Fetched: 33 row(s)

hive> SELECT * FROM raw_user_logs WHERE year=2025 AND month=3 AND day=11 LIMIT 10;

OK

NULL    NULL    action  timestamp       device  region  session_id      2025    3       11

133     1004    play    2023-09-01 00:00:00     desktop US      sess_3  2025    3       11

110     1008    forward 2023-09-01 00:00:00     tablet  APAC    sess_24 2025    3       11

125     1004    skip    2023-09-01 00:00:00     mobile  US      sess_23 2025    3       11

110     1003    forward 2023-09-01 00:00:00     mobile  APAC    sess_22 2025    3       11

157     1002    forward 2023-09-01 00:00:00     tablet  EU      sess_4  2025    3       11

122     1005    skip    2023-09-01 00:00:00     tablet  EU      sess_1  2025    3       11

119     1002    forward 2023-09-01 00:00:00     desktop APAC    sess_41 2025    3       11

181     1006    play    2023-09-01 00:00:00     mobile  US      sess_28 2025    3       11

196     1010    pause   2023-09-01 00:00:00     desktop APAC    sess_33 2025    3       11

Time taken: 1.361 seconds, Fetched: 10 row(s)

hive> SELECT * FROM raw_content_metadata LIMIT 10;

OK

NULL    title   category        NULL    artist

1000    Title 1000      Jazz    290     Artist 1

1001    Title 1001      Jazz    149     Artist 7

1002    Title 1002      Jazz    179     Artist 9

1003    Title 1003      Rock    283     Artist 1

1004    Title 1004      Jazz    149     Artist 10

1005    Title 1005      Podcast 192     Artist 10

1006    Title 1006      Rock    148     Artist 5

1007    Title 1007      Jazz    232     Artist 3

1008    Title 1008      News    181     Artist 8
```

```
Time taken: 0.108 seconds, Fetched: 10 row(s)

hive> CREATE TABLE dim_users STORED AS PARQUET AS

    > SELECT DISTINCT user_id FROM raw_user_logs;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = root_20250311133055_c1a40568-a797-45e4-8662-d3ab582d7121

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_1741695021252_0001, Tracking URL = http://hive:8088/proxy/application_1741695021252_0001/

Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0001

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2025-03-11 13:31:04,273 Stage-1 map = 0%,  reduce = 0%

2025-03-11 13:31:09,362 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.86 sec

2025-03-11 13:31:14,519 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.57 sec

MapReduce Total cumulative CPU time: 4 seconds 570 msec

Ended Job = job_1741695021252_0001

Moving data to directory hdfs://hive:9820/user/hive/warehouse/dim_users

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.57 sec   HDFS Read: 18015 HDFS Write: 641 SUCCESS

Total MapReduce CPU Time Spent: 4 seconds 570 msec

OK

Time taken: 20.587 seconds

hive> CREATE TABLE dim_content STORED AS PARQUET AS

    > SELECT DISTINCT content_id, title, category, artist

    > FROM raw_content_metadata
  > WHERE content_id IS NOT NULL;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20250311133138_1611034c-5f9c-4a84-b817-273763027230
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_1741695021252_0002, Tracking URL = http://hive:8088/proxy/application_1741695021252_0002/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-03-11 13:31:43,348 Stage-1 map = 0%,  reduce = 0%
2025-03-11 13:31:47,466 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2025-03-11 13:31:52,589 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.44 sec
MapReduce Total cumulative CPU time: 4 seconds 440 msec
Ended Job = job_1741695021252_0002
Moving data to directory hdfs://hive:9820/user/hive/warehouse/dim_content
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.44 sec   HDFS Read: 9184 HDFS Write: 1023 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 440 msec
OK
Time taken: 15.544 seconds
```

HIVE DDL For Star Schema:

```
hive> CREATE TABLE dim_users STORED AS PARQUET AS
  > SELECT DISTINCT user_id FROM raw_user_logs;
```

```
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20250311133055_c1a40568-a797-45e4-8662-d3ab582d7121
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_1741695021252_0001, Tracking URL = http://hive:8088/proxy/application_1741695021252_0001/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-03-11 13:31:04,273 Stage-1 map = 0%,  reduce = 0%
2025-03-11 13:31:09,362 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.86 sec
2025-03-11 13:31:14,519 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.57 sec
MapReduce Total cumulative CPU time: 4 seconds 570 msec
Ended Job = job_1741695021252_0001
Moving data to directory hdfs://hive:9820/user/hive/warehouse/dim_users
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.57 sec   HDFS Read: 18015 HDFS Write: 641 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 570 msec
OK
Time taken: 20.587 seconds
hive> CREATE TABLE dim_content STORED AS PARQUET AS
   > SELECT DISTINCT content_id, title, category, artist
   > FROM raw_content_metadata
   > WHERE content_id IS NOT NULL;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
```

```
Query ID = root_20250311133138_1611034c-5f9c-4a84-b817-273763027230
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_1741695021252_0002, Tracking URL = http://hive:8088/proxy/application_1741695021252_0002/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-03-11 13:31:43,348 Stage-1 map = 0%,  reduce = 0%
2025-03-11 13:31:47,466 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2025-03-11 13:31:52,589 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.44 sec
MapReduce Total cumulative CPU time: 4 seconds 440 msec
Ended Job = job_1741695021252_0002
Moving data to directory hdfs://hive:9820/user/hive/warehouse/dim_content
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.44 sec   HDFS Read: 9184 HDFS Write: 1023 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 440 msec
OK
Time taken: 15.544 seconds
hive> CREATE TABLE dim_sessions STORED AS PARQUET AS
   > SELECT DISTINCT session_id, device, region
   > FROM raw_user_logs;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20250311133211_bd71f6da-c762-467c-9b07-fb89e1b6f5c2
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_1741695021252_0003, Tracking URL = http://hive:8088/proxy/application_1741695021252_0003/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-03-11 13:32:15,178 Stage-1 map = 0%,  reduce = 0%
2025-03-11 13:32:19,280 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.4 sec
2025-03-11 13:32:24,401 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.43 sec
MapReduce Total cumulative CPU time: 3 seconds 430 msec
Ended Job = job_1741695021252_0003
Moving data to directory hdfs://hive:9820/user/hive/warehouse/dim_sessions
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.43 sec   HDFS Read: 19039 HDFS Write: 1366 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 430 msec
OK
Time taken: 14.523 seconds
hive> CREATE TABLE fact_user_activity (
    >     user_id INT,
    >     content_id INT,
    >     session_id STRING,
    >     action STRING,
    >     event_timestamp STRING,
    >     year INT,
    >     month INT,
    >     day INT
    > ) STORED AS PARQUET;
```

```
OK

Time taken: 0.05 seconds

hive> INSERT OVERWRITE TABLE fact_user_activity

    > SELECT user_id, content_id, session_id, action, `timestamp`, year, month, day

    > FROM raw_user_logs;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = root_20250311133247_87b21063-ce00-4bcf-9bf4-0c458bda021c

Total jobs = 3

Launching Job 1 out of 3

Number of reduce tasks is set to 0 since there's no reduce operator

Starting Job = job_1741695021252_0004, Tracking URL = http://hive:8088/proxy/application_1741695021252_0004/

Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0004

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0

2025-03-11 13:32:52,308 Stage-1 map = 0%,  reduce = 0%

2025-03-11 13:32:56,406 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.92 sec

MapReduce Total cumulative CPU time: 1 seconds 920 msec

Ended Job = job_1741695021252_0004

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://hive:9820/user/hive/warehouse/fact_user_activity/.hive-staging_hive_2025-03-11_13-32-47_265_2347698501231064156-1/-ext-10000

Loading data to table default.fact_user_activity

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1   Cumulative CPU: 1.92 sec   HDFS Read: 15079 HDFS Write: 2912 SUCCESS

Total MapReduce CPU Time Spent: 1 seconds 920 msec

OK

Time taken: 10.483 seconds

hive> SELECT * FROM fact_user_activity LIMIT 10;

OK

SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
```

```
SLF4J: Defaulting to no-operation (NOP) logger implementation

SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.

NULL    NULL    session_id      action  timestamp       2025    3       11

133     1004    sess_3  play    2023-09-01 00:00:00     2025    3       11

110     1008    sess_24 forward 2023-09-01 00:00:00     2025    3       11

125     1004    sess_23 skip    2023-09-01 00:00:00     2025    3       11

110     1003    sess_22 forward 2023-09-01 00:00:00     2025    3       11

157     1002    sess_4  forward 2023-09-01 00:00:00     2025    3       11

122     1005    sess_1  skip    2023-09-01 00:00:00     2025    3       11

119     1002    sess_41 forward 2023-09-01 00:00:00     2025    3       11

181     1006    sess_28 play    2023-09-01 00:00:00     2025    3       11

196     1010    sess_33 pause   2023-09-01 00:00:00     2025    3       11

Time taken: 0.105 seconds, Fetched: 10 row(s)

hive> SELECT * FROM dim_users LIMIT 10;

OK

NULL

100

103

104

105

106

107

108

109

110

Time taken: 0.069 seconds, Fetched: 10 row(s)

hive> SELECT * FROM dim_content LIMIT 10;

OK

1000    Title 1000      Jazz    Artist 1

1001    Title 1001      Jazz    Artist 7

1002    Title 1002      Jazz    Artist 9
```

```
1003    Title 1003    Rock    Artist 1
1004    Title 1004    Jazz    Artist 10
1005    Title 1005    Podcast Artist 10
1006    Title 1006    Rock    Artist 5
1007    Title 1007    Jazz    Artist 3
1008    Title 1008    News    Artist 8
1009    Title 1009    Jazz    Artist 9
Time taken: 0.082 seconds, Fetched: 10 row(s)
hive> SELECT * FROM dim_sessions LIMIT 10;
OK
sess_10 desktop APAC
sess_11 desktop APAC
sess_14 desktop APAC
sess_16 desktop APAC
sess_2  desktop APAC
sess_21 desktop APAC
sess_22 desktop APAC
sess_3  desktop APAC
sess_30 desktop APAC
sess_33 desktop APAC
Time taken: 0.072 seconds, Fetched: 10 row(s)
```

Partitioned table:

```
CREATE EXTERNAL TABLE IF NOT EXISTS user_logs_partitioned (
  user_id STRING,
  event_type STRING,
  event_timestamp STRING,
  event_details STRING
)
PARTITIONED BY (year INT, month INT, day INT)
ROW FORMAT DELIMITED
```

```
FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/raw/logs/';
```

HIVE DDL For Transformation:

```
hive> INSERT OVERWRITE TABLE fact_user_activity
    > SELECT user_id, content_id, session_id, action, `timestamp`, year, month, day
    > FROM raw_user_logs;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20250311133247_87b21063-ce00-4bcf-9bf4-0c458bda021c
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1741695021252_0004, Tracking URL = http://hive:8088/proxy/application_1741695021252_0004/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2025-03-11 13:32:52,308 Stage-1 map = 0%,  reduce = 0%
2025-03-11 13:32:56,406 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.92 sec
MapReduce Total cumulative CPU time: 1 seconds 920 msec
Ended Job = job_1741695021252_0004
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://hive:9820/user/hive/warehouse/fact_user_activity/.hive-staging_hive_2025-03-11_13-32-47_265_2347698501231064156-1/-ext-10000
Loading data to table default.fact_user_activity
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.92 sec   HDFS Read: 15079 HDFS Write: 2912 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 920 msec
OK
Time taken: 10.483 seconds
```

Query 1 DDL:

```
hive> SELECT

  >    f.year,

  >    f.month,

  >    s.region,

  >    COUNT(DISTINCT f.user_id) AS monthly_active_users

  > FROM fact_user_activity f

  > JOIN dim_sessions s ON f.session_id = s.session_id

  > WHERE f.year = 2025  -- Example filter

  > GROUP BY f.year, f.month, s.region

  > ORDER BY f.year, f.month, monthly_active_users DESC;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = root_20250311142406_3fa5adc6-3b22-4304-a693-73eea3ebe2b7

Total jobs = 2

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.

SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

2025-03-11 14:24:16     Starting to launch local task to process map join;      maximum memory = 477626368

SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".

SLF4J: Defaulting to no-operation (NOP) logger implementation

SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.

2025-03-11 14:24:17     Dump the side-table for tag: 1 with group count: 50 into file: file:/tmp/root/f280238c-aeba-4b03-9fa3-42d8adb61222/hive_2025-03-11_14-24-06_316_3053109984737745539-1/-local-10006/HashTable-Stage-2/MapJoin-mapfile01--.hashtable

2025-03-11 14:24:17     Uploaded 1 File to: file:/tmp/root/f280238c-aeba-4b03-9fa3-42d8adb61222/hive_2025-03-11_14-24-06_316_3053109984737745539-1/-local-10006/HashTable-Stage-2/MapJoin-mapfile01--.hashtable (2585 bytes)
```

```
2025-03-11 14:24:17     End of local task; Time Taken: 1.486 sec.
Execution completed successfully
MapredLocal task succeeded
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_1741695021252_0005, Tracking URL =
http://hive:8088/proxy/application_1741695021252_0005/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2025-03-11 14:24:24,837 Stage-2 map = 0%,  reduce = 0%
2025-03-11 14:24:29,902 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.95 sec
2025-03-11 14:24:35,082 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.56 sec
MapReduce Total cumulative CPU time: 4 seconds 560 msec
Ended Job = job_1741695021252_0005
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_1741695021252_0006, Tracking URL =
http://hive:8088/proxy/application_1741695021252_0006/
```

```
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0006

Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1

2025-03-11 14:24:45,716 Stage-3 map = 0%,  reduce = 0%

2025-03-11 14:24:48,797 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.91 sec

2025-03-11 14:24:52,896 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 2.19 sec

MapReduce Total cumulative CPU time: 2 seconds 190 msec

Ended Job = job_1741695021252_0006

MapReduce Jobs Launched:

Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.56 sec   HDFS Read: 16306 HDFS Write: 190 SUCCESS

Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 2.19 sec   HDFS Read: 6216 HDFS Write: 192 SUCCESS

Total MapReduce CPU Time Spent: 6 seconds 750 msec

OK

2025    3       US      80

2025    3       APAC    79

2025    3       EU      72

2025    3       region  0

Time taken: 48.701 seconds, Fetched: 4 row(s)
```

Query 2 DDL

```
hive> SELECT

    >     c.category,

    >     COUNT(*) AS play_count

    > FROM fact_user_activity f

    > JOIN dim_content c ON f.content_id = c.content_id

    > WHERE f.action = 'play'

    > AND f.year = 2025 AND f.month = 3  -- Example filter

    > GROUP BY c.category

    > ORDER BY play_count DESC

    > LIMIT 10;
```

```
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20250311142721_971cebab-7e1c-406d-86a7-f1598c563ead
Total jobs = 2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.17.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2025-03-11 14:27:26	Starting to launch local task to process map join;	maximum memory = 477626368
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
2025-03-11 14:27:27	Dump the side-table for tag: 1 with group count: 11 into file: file:/tmp/root/f280238c-aeba-4b03-9fa3-42d8adb61222/hive_2025-03-11_14-27-21_806_1616347904249172090-1/-local-10006/HashTable-Stage-2/MapJoin-mapfile11--.hashtable
2025-03-11 14:27:27	Uploaded 1 File to: file:/tmp/root/f280238c-aeba-4b03-9fa3-42d8adb61222/hive_2025-03-11_14-27-21_806_1616347904249172090-1/-local-10006/HashTable-Stage-2/MapJoin-mapfile11--.hashtable (552 bytes)
2025-03-11 14:27:27	End of local task; Time Taken: 1.043 sec.
Execution completed successfully
MapredLocal task succeeded
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_1741695021252_0007, Tracking URL =
http://hive:8088/proxy/application_1741695021252_0007/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0007
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2025-03-11 14:27:33,884 Stage-2 map = 0%,  reduce = 0%
2025-03-11 14:27:37,990 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.7 sec
2025-03-11 14:27:42,089 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.12 sec
MapReduce Total cumulative CPU time: 4 seconds 120 msec
Ended Job = job_1741695021252_0007
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_1741695021252_0008, Tracking URL =
http://hive:8088/proxy/application_1741695021252_0008/
Kill Command = /opt/hadoop/bin/hadoop job  -kill job_1741695021252_0008
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2025-03-11 14:27:52,621 Stage-3 map = 0%,  reduce = 0%
2025-03-11 14:27:56,621 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.99 sec
2025-03-11 14:28:00,737 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 2.37 sec
MapReduce Total cumulative CPU time: 2 seconds 370 msec
Ended Job = job_1741695021252_0008
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.12 sec   HDFS Read: 13890 HDFS Write: 191 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 2.37 sec   HDFS Read: 5751 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 490 msec
```

```
OK
Jazz    28
Rock    8
Podcast 6
News    2
Time taken: 39.995 seconds, Fetched: 4 row(s)
hive>
```

Query 3 DDL: