HIVE

To create the databases:

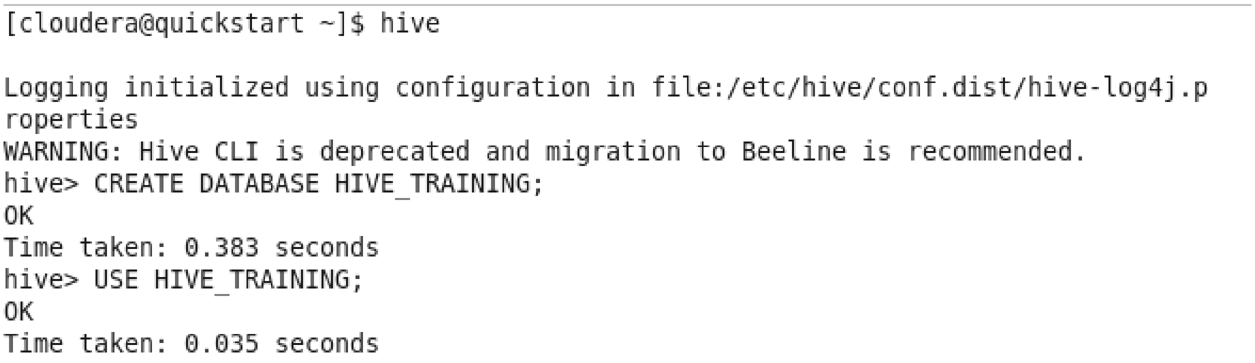
CREATE DATABASE FIRST_TABLE;

In order to use the database:

USE FIRST_TABLE;

To show the databases:

SHOW DB_NAME;



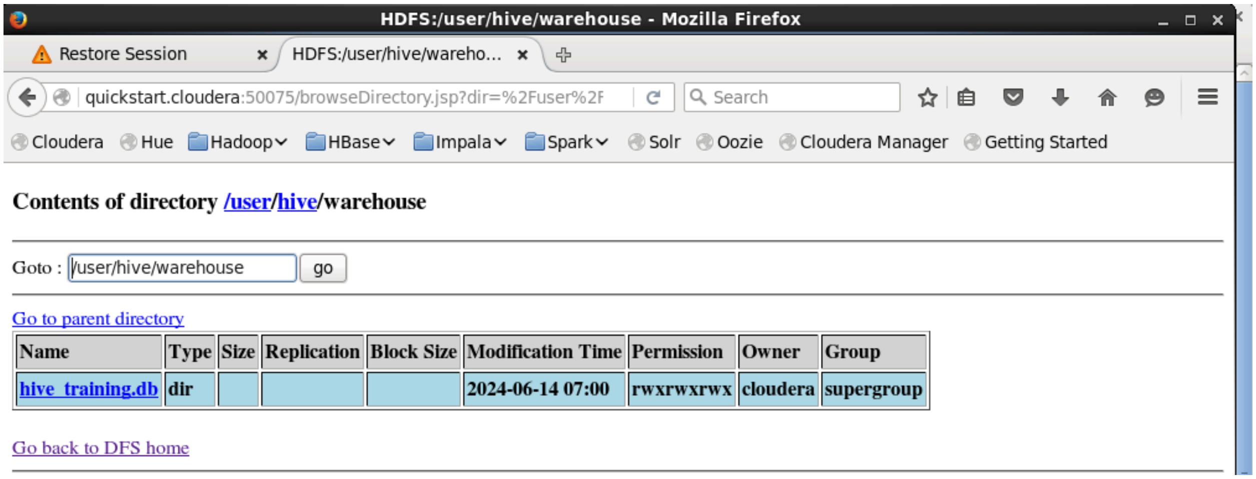
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE DATABASE HIVE_TRAINING;
OK
Time taken: 0.383 seconds
hive> USE HIVE_TRAINING;
OK
Time taken: 0.035 seconds
```

To create the table:

```
CREATE TABLE FIRST_TABLE(
   KEY INT,
   VALUE STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE TABLE FIRST_TABLE(
    >     KEY INT,
    >     VALUE STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 2.309 seconds
```

—---


HDFS:/user/hive/warehouse - Mozilla Firefox
Restore Session
HDFS:/user/hive/wareho...
quickstart.cloudera:50075/browseDirectory.jsp?dir=%2Fuser%2F
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started

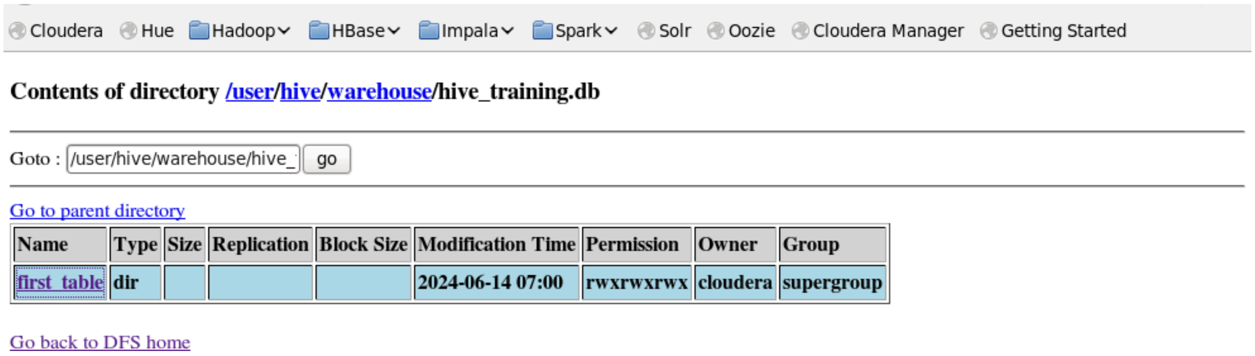
Contents of directory /user/hive/warehouse

Goto : /user/hive/warehouse go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
hive_training.db	dir				2024-06-14 07:00	rwxrwxrwx	cloudera	supergroup

Go back to DFS home



Contents of directory /user/hive/warehouse/hive_training.db

Goto : /user/hive/warehouse/hive_ go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
first_table	dir				2024-06-14 07:00	rwxrwxrwx	cloudera	supergroup

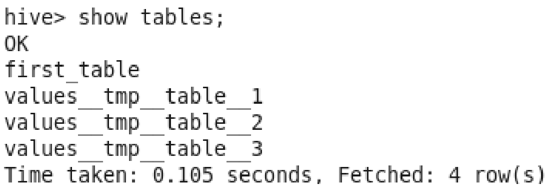
Go back to DFS home

Here you can observe that the database is considered as directory and the tables are considered as files.

—---

To show the list of tables:

show tables;

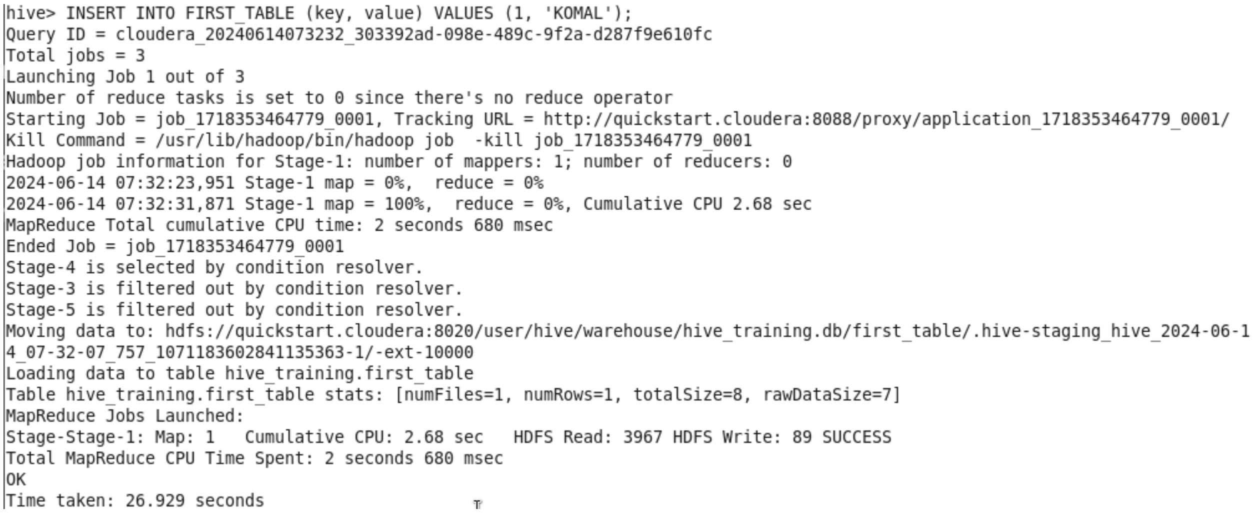


```
hive> show tables;
OK
first_table
values__tmp__table__1
values__tmp__table__2
values__tmp__table__3
Time taken: 0.105 seconds, Fetched: 4 row(s)
```

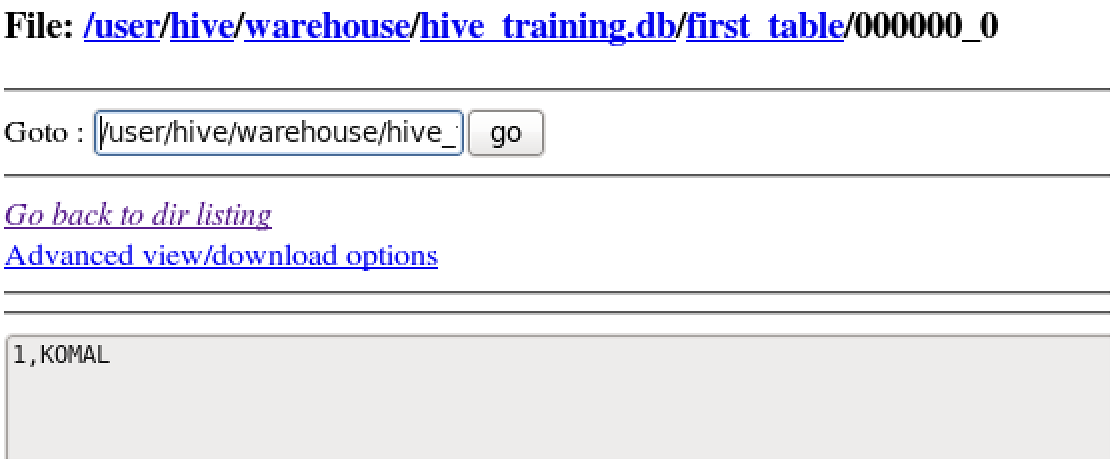
—---

To Insert the values in a table:

INSERT INTO FIRST_TABLE (key, values) VALUES (1, 'KOMAL');



```
hive> INSERT INTO FIRST_TABLE (key, value) VALUES (1, 'KOMAL');
Query ID = cloudera_20240614073232_303392ad-098e-489c-9f2a-d287f9e610fc
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_1718353464779_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718353464779_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718353464779_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-14 07:32:23,951 Stage-1 map = 0%,  reduce = 0%
2024-06-14 07:32:31,871 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.68 sec
MapReduce Total cumulative CPU time: 2 seconds 680 msec
Ended Job = job_1718353464779_0001
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/first_table/.hive-staging_hive_2024-06-1
4_07-32-07_757_1071183602841135363-1/-ext-10000
Loading data to table hive_training.first_table
Table hive_training.first_table stats: [numFiles=1, numRows=1, totalSize=8, rawDataSize=7]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.68 sec   HDFS Read: 3967 HDFS Write: 89 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 680 msec
OK
Time taken: 26.929 seconds
```



File: /user/hive/warehouse/hive_training.db/first_table/000000_0

Goto : /user/hive/warehouse/hive_ go

Go back to dir listing

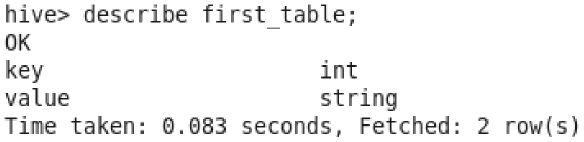
Advanced view/download options

```
1,KOMAL
```

—---

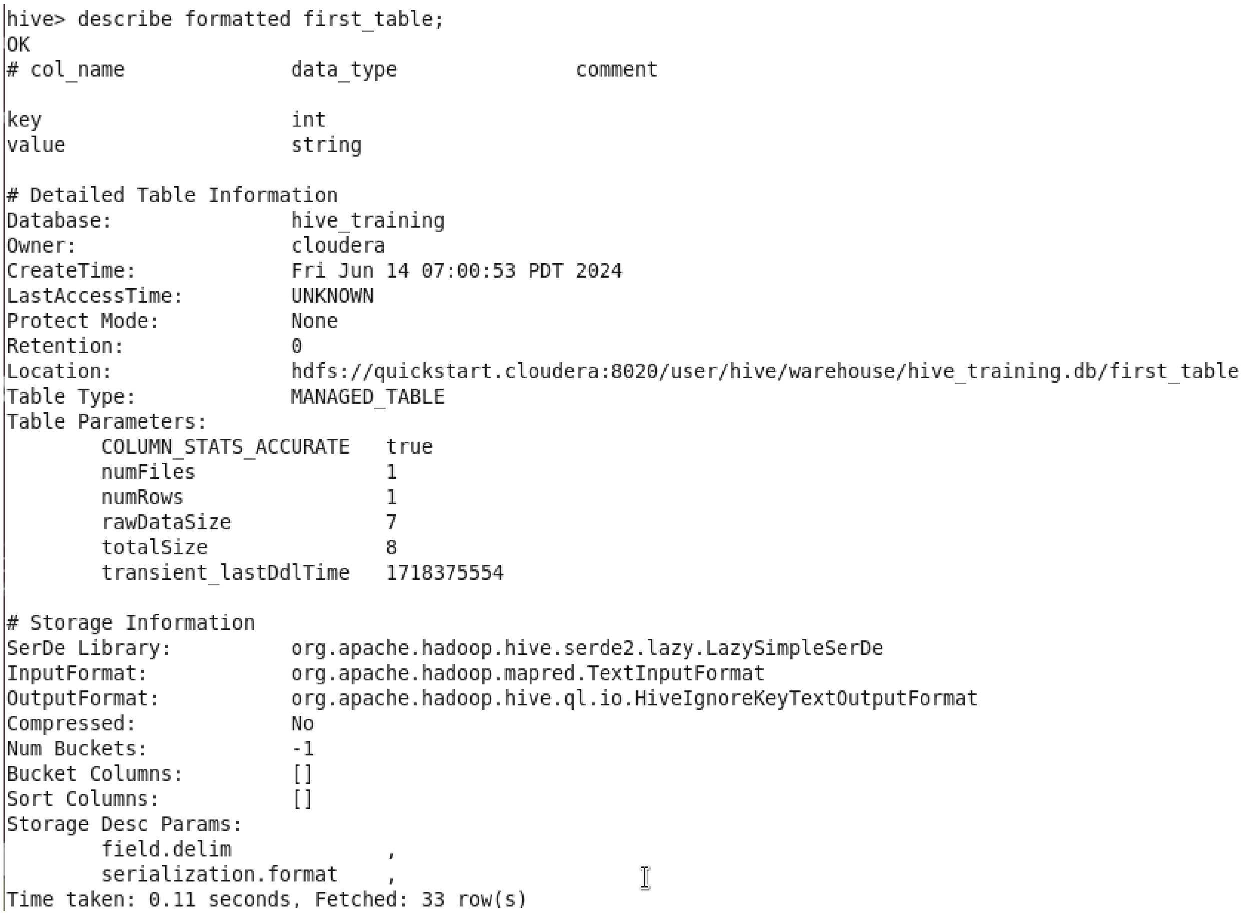
To describe the table structure: (schema information)

describe first_table;



```
hive> describe first_table;
OK
key                     int
value                   string
Time taken: 0.083 seconds, Fetched: 2 row(s)
```

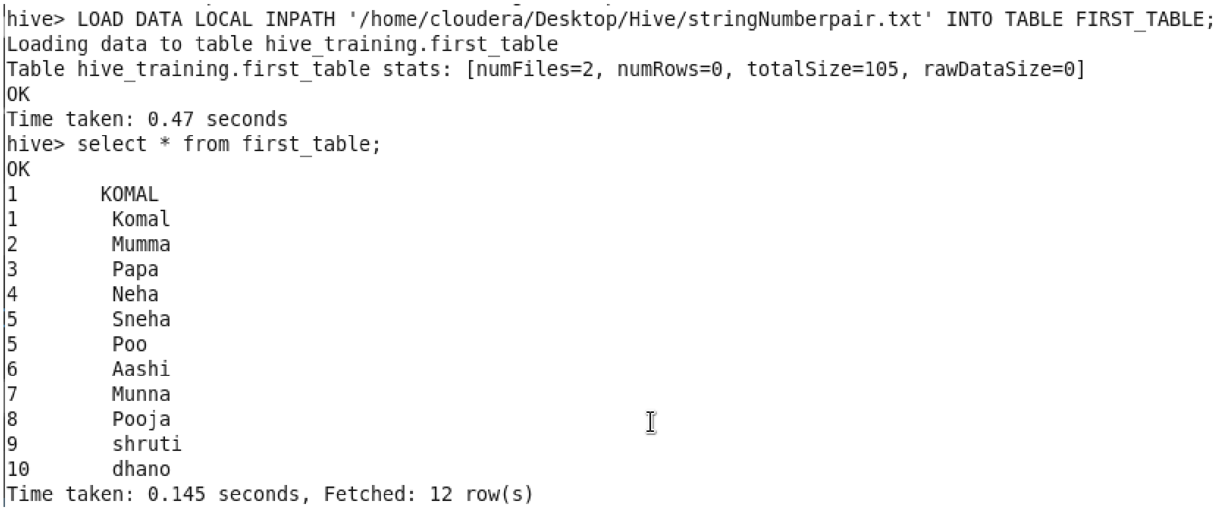
describe formatted first_table;



```
hive> describe formatted first_table;
OK
# col_name              data_type               comment

key                     int
value                   string

# Detailed Table Information
Database:               hive_training
Owner:                  cloudera
CreateTime:             Fri Jun 14 07:00:53 PDT 2024
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/first_table
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   true
        numFiles                1
        numRows                 1
        rawDataSize             7
        totalSize               8
        transient_lastDdlTime   1718375554

# 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: 0.11 seconds, Fetched: 33 row(s)
```

To Insert the bulk load data from local file system to Hive:

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/stringNumberpair.txt' INTO TABLE FIRST_TABLE;



```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/stringNumberpair.txt' INTO TABLE FIRST_TABLE;
Loading data to table hive_training.first_table
Table hive_training.first_table stats: [numFiles=2, numRows=0, totalSize=105, rawDataSize=0]
OK
Time taken: 0.47 seconds
hive> select * from first_table;
OK
1       KOMAL
1        Komal
2        Mumma
3        Papa
4        Neha
5        Sneha
5        Poo
6        Aashi
7        Munna
8        Pooja
9        shruti
10       dhano
Time taken: 0.145 seconds, Fetched: 12 row(s)
```

To insert the bulk data from HDFS to Hive:

LOAD DATA INPATH ‘/hive_training/stringNumberpair.txt’ into table stringNumber

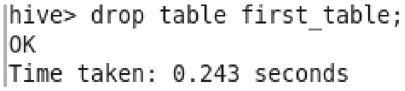
—--

Drop the table:

If we drop the table of type ‘managed table’, then it will delete

- Data
- Table name
- Table shema

drop table first_table;



```
hive> drop table first_table;
OK
Time taken: 0.243 seconds
```

—--

PROBLEM STATEMENT : RETAIL_DATABASE

```
CREATE DATABASE IF NOT EXIST retail_db;

USE retail_db;
```

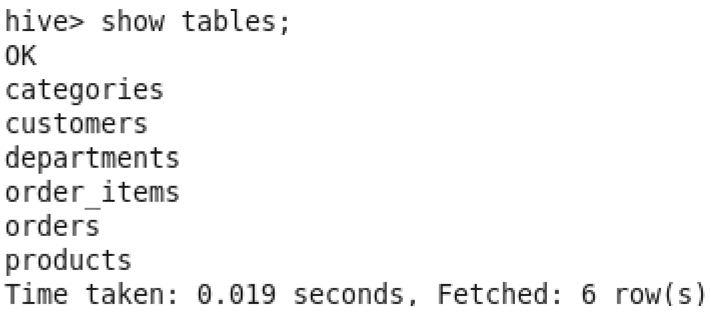
—--

Create tables:

```
CREATE TABLE categories (
    category_id INT,
    category_department_id INT,
    category_name STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;

CREATE TABLE customers (
customer_id int,
customer_fname string,
customer_lname string,
customer_email string,
customer_password string,
customer_street string,
customer_city string,
customer_state string,
customer_zipcode string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;

CREATE TABLE orders (
order_id int,
order_date string,
order_customer_id int,
order_status string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
```

```
STORED AS TEXTFILE
TBLPROPERTIES ('skip.header.line.count'='1');




CREATE TABLE departments (
department_id int,
department_name string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;

CREATE TABLE order_items (
order_item_id int,
order_item_order_id int,
order_item_order_date string,
order_item_product_id int,
order_item_quantity smallint,
order_item_subtotal float,
order_item_product_price float
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;

CREATE TABLE products (
product_id int,
product_category_id int,
product_name string,
product_description string,
product_price float,
product_image string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;
```

```
CREATE TABLE shippers (
   ShipperID INT,
   ShipperName STRING,
   Phone STRING,
);
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1')
STORED AS TEXTFILE;
```



```
hive> show tables;
OK
categories
customers
departments
order_items
orders
products
Time taken: 0.019 seconds, Fetched: 6 row(s)
```

—--

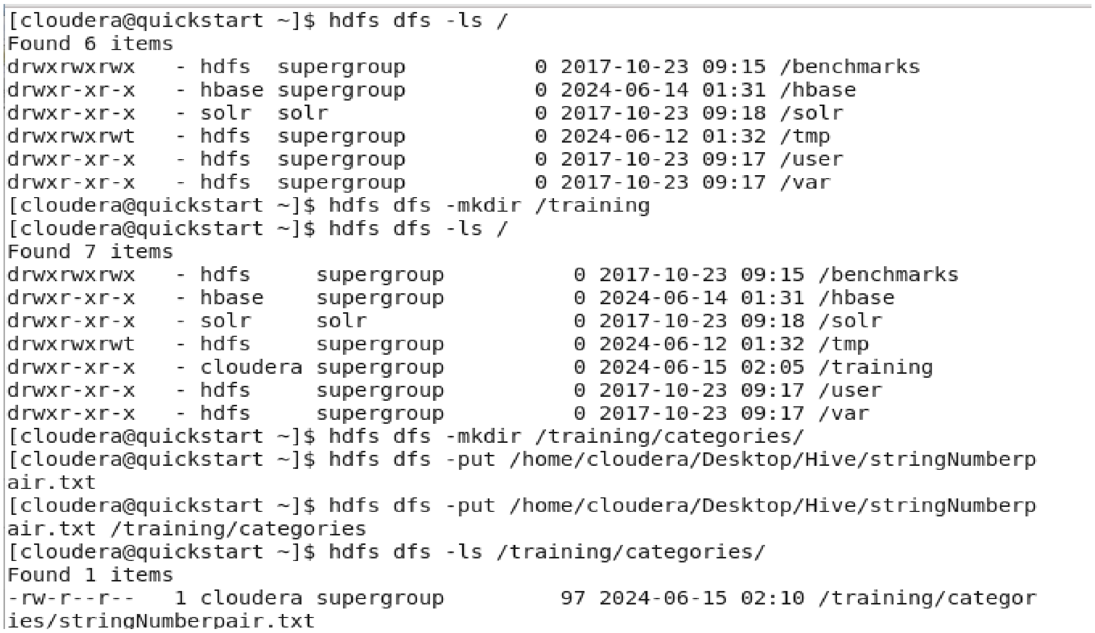
****

MANAGED TABLE USING LOCATION

****

You can load data from HDFS path location

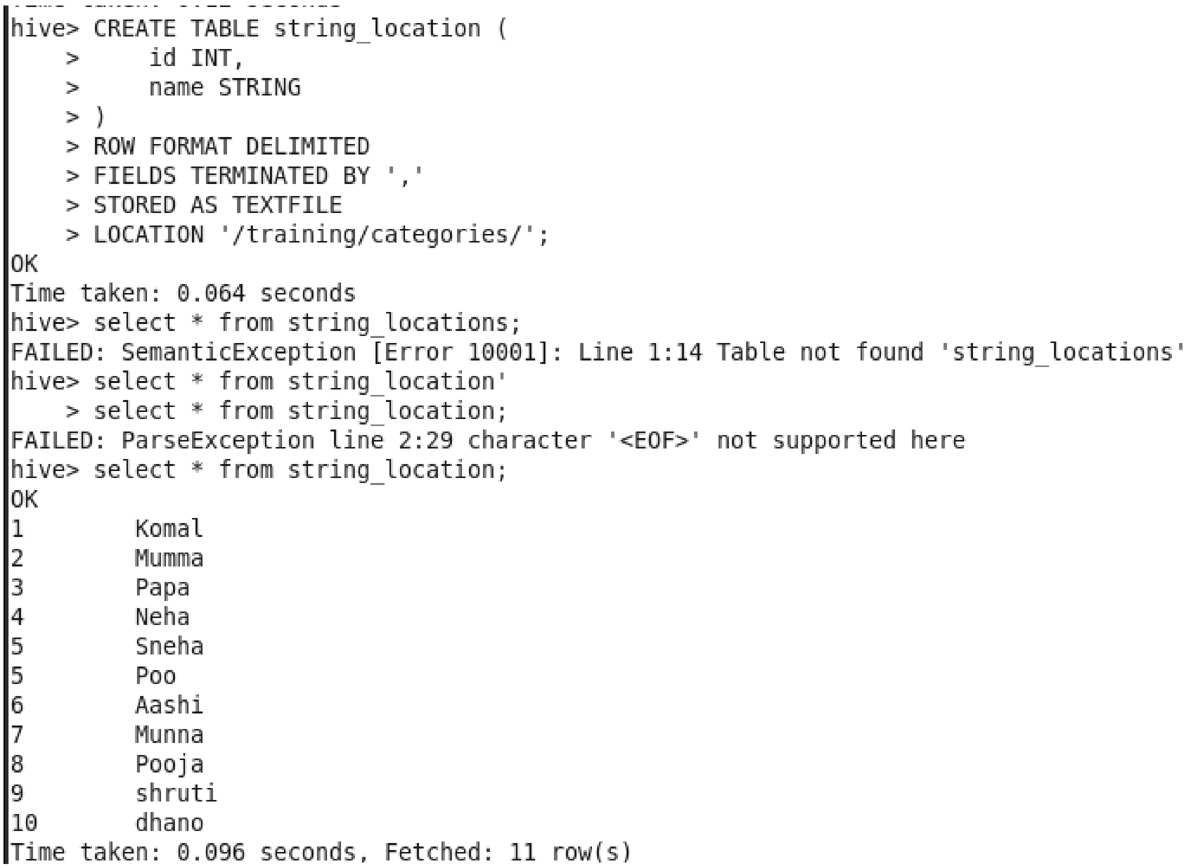
Before that 👍



```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2024-06-14 01:31 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2024-06-12 01:32 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /training
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 7 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2024-06-14 01:31 /hbase
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs     supergroup          0 2024-06-12 01:32 /tmp
drwxr-xr-x   - cloudera supergroup          0 2024-06-15 02:05 /training
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -mkdir /training/categories/
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Desktop/Hive/stringNumberp
air.txt
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Desktop/Hive/stringNumberp
air.txt /training/categories
[cloudera@quickstart ~]$ hdfs dfs -ls /training/categories/
Found 1 items
-rw-r--r--   1 cloudera supergroup         97 2024-06-15 02:10 /training/categor
ies/stringNumberpair.txt
```

Now,

```
CREATE TABLE string_location (
   id INT,
   name STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/training/categories/';
```



```
hive> CREATE TABLE string_location (
    >     id INT,
    >     name STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/training/categories/';
OK
Time taken: 0.064 seconds
hive> select * from string_locations;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'string_locations'
hive> select * from string_location'
    > select * from string_location;
FAILED: ParseException line 2:29 character '<EOF>' not supported here
hive> select * from string_location;
OK
1       Komal
2       Mumma
3       Papa
4       Neha
5       Sneha
5       Poo
6       Aashi
7       Munna
8       Pooja
9       shruti
10      dhano
Time taken: 0.096 seconds, Fetched: 11 row(s)
```

It will not create a new directory in hive whereas it will store in the same HDFS directory as it was previous.

Data will not be move and if we drop the table it will remove the data from the HDFS as well.

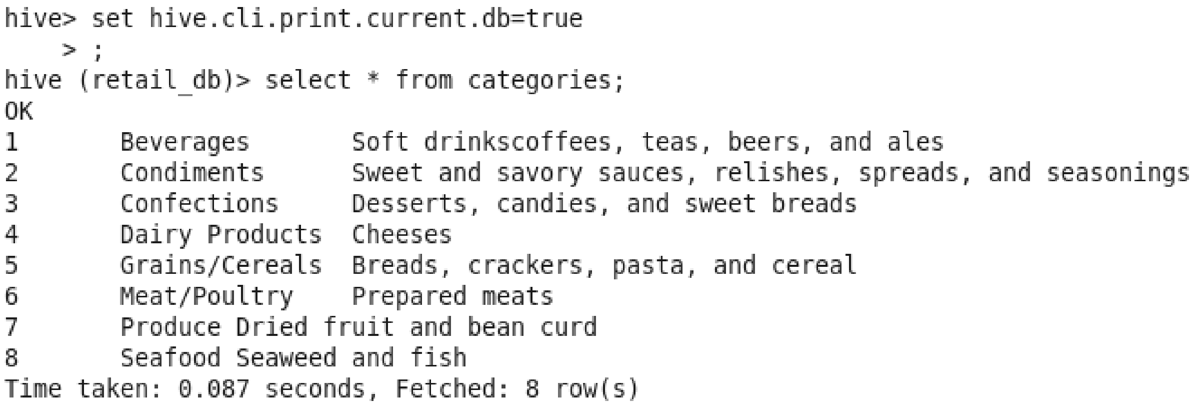
—--

You can connect with Mysql:

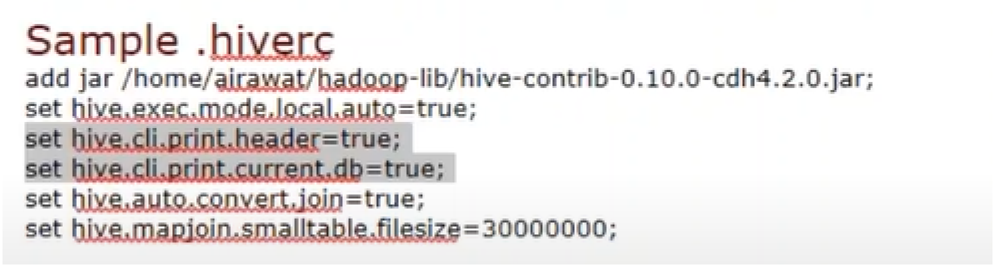
mysql -u hadoop1 -p
—---

To how the current DB name in CLI:

set hive.cli.print.current.db=true



```
hive> set hive.cli.print.current.db=true
    > ;
hive (retail_db)> select * from categories;
OK
1       Beverages       Soft drinkscoffees, teas, beers, and ales
2       Condiments      Sweet and savory sauces, relishes, spreads, and seasonings
3       Confections     Desserts, candies, and sweet breads
4       Dairy Products  Cheeses
5       Grains/Cereals  Breads, crackers, pasta, and cereal
6       Meat/Poultry    Prepared meats
7       Produce Dried fruit and bean curd
8       Seafood Seaweed and fish
Time taken: 0.087 seconds, Fetched: 8 row(s)
```

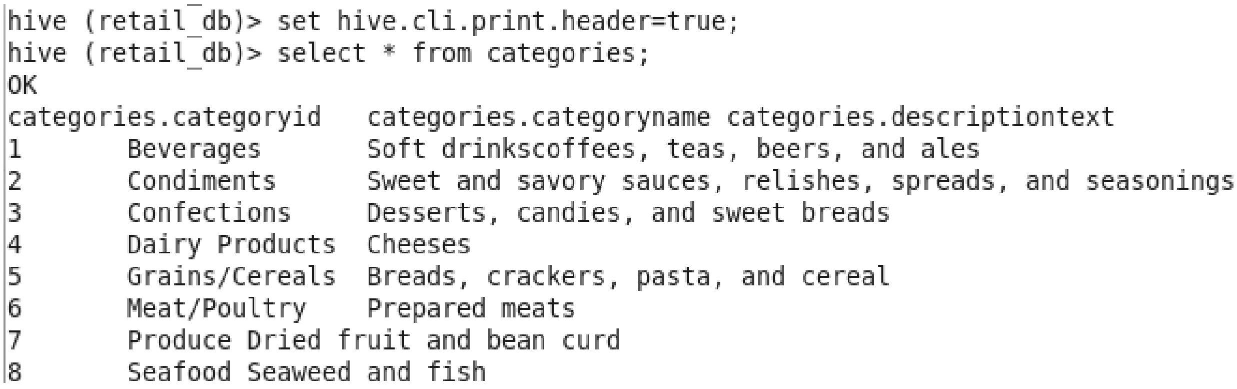


Sample .hiverc

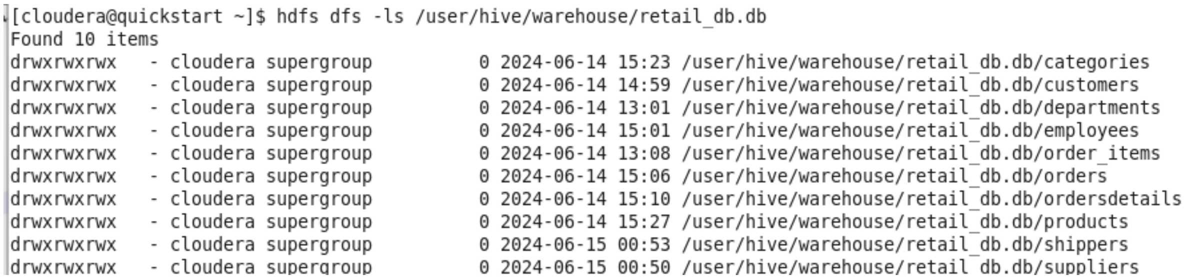
```
add jar /home/airawat/hadoop-lib/hive-contrib-0.10.0-cdh4.2.0.jar;
set hive.exec.mode.local.auto=true;
set hive.cli.print.header=true;
set hive.cli.print.current.db=true;
set hive.auto.convert.join=true;
set hive.mapjoin.smalltable.filesize=30000000;
```

To how the header in CLI:

set hive.cli.print.header=true;



```
hive (retail_db)> set hive.cli.print.header=true;
hive (retail_db)> select * from categories;
OK
categories.categoryid   categories.categoryname categories.descriptiontext
1       Beverages       Soft drinkscoffees, teas, beers, and ales
2       Condiments      Sweet and savory sauces, relishes, spreads, and seasonings
3       Confections     Desserts, candies, and sweet breads
4       Dairy Products  Cheeses
5       Grains/Cereals  Breads, crackers, pasta, and cereal
6       Meat/Poultry    Prepared meats
7       Produce Dried fruit and bean curd
8       Seafood Seaweed and fish
```
—--



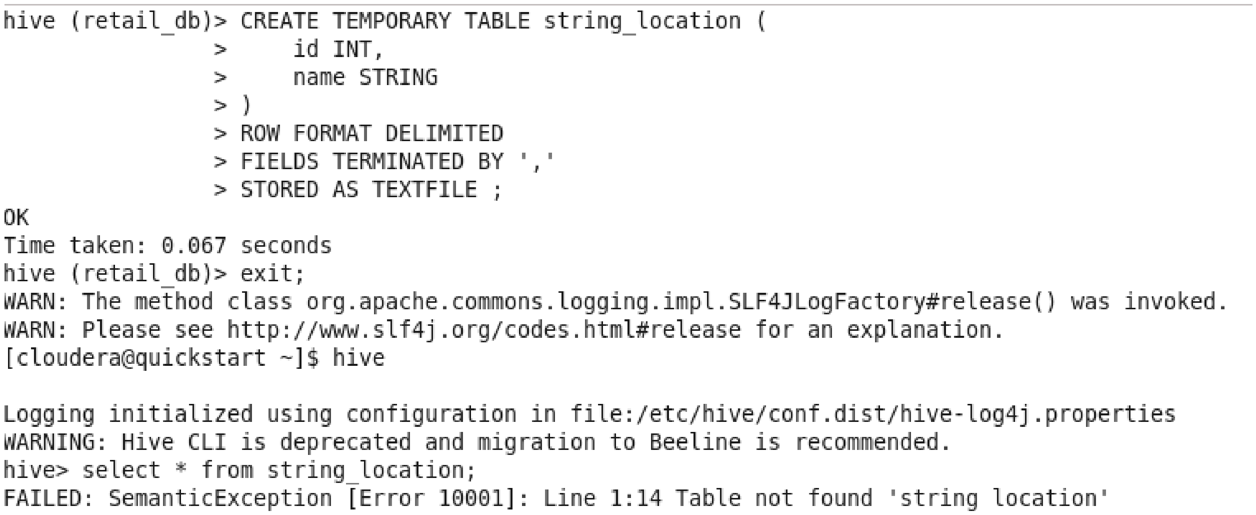
```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/hive/warehouse/retail_db.db
Found 10 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 15:23 /user/hive/warehouse/retail_db.db/categories
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 14:59 /user/hive/warehouse/retail_db.db/customers
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 13:01 /user/hive/warehouse/retail_db.db/departments
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 15:01 /user/hive/warehouse/retail_db.db/employees
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 13:08 /user/hive/warehouse/retail_db.db/order_items
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 15:06 /user/hive/warehouse/retail_db.db/orders
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 15:10 /user/hive/warehouse/retail_db.db/ordersdetails
drwxrwxrwx   - cloudera supergroup          0 2024-06-14 15:27 /user/hive/warehouse/retail_db.db/products
drwxrwxrwx   - cloudera supergroup          0 2024-06-15 00:53 /user/hive/warehouse/retail_db.db/shippers
drwxrwxrwx   - cloudera supergroup          0 2024-06-15 00:50 /user/hive/warehouse/retail_db.db/suppliers
```

—--

TEMPORARY TABLE:

It is available for that hive session only.

```
CREATE TEMPORARY TABLE string_location (
   id INT,
   name STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE ;
```



```
hive (retail_db)> CREATE TEMPORARY TABLE string_location (
                >     id INT,
                >     name STRING
                > )
                > ROW FORMAT DELIMITED
                > FIELDS TERMINATED BY ','
                > STORED AS TEXTFILE ;
OK
Time taken: 0.067 seconds
hive (retail_db)> exit;
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> select * from string_location;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'string_location'
```

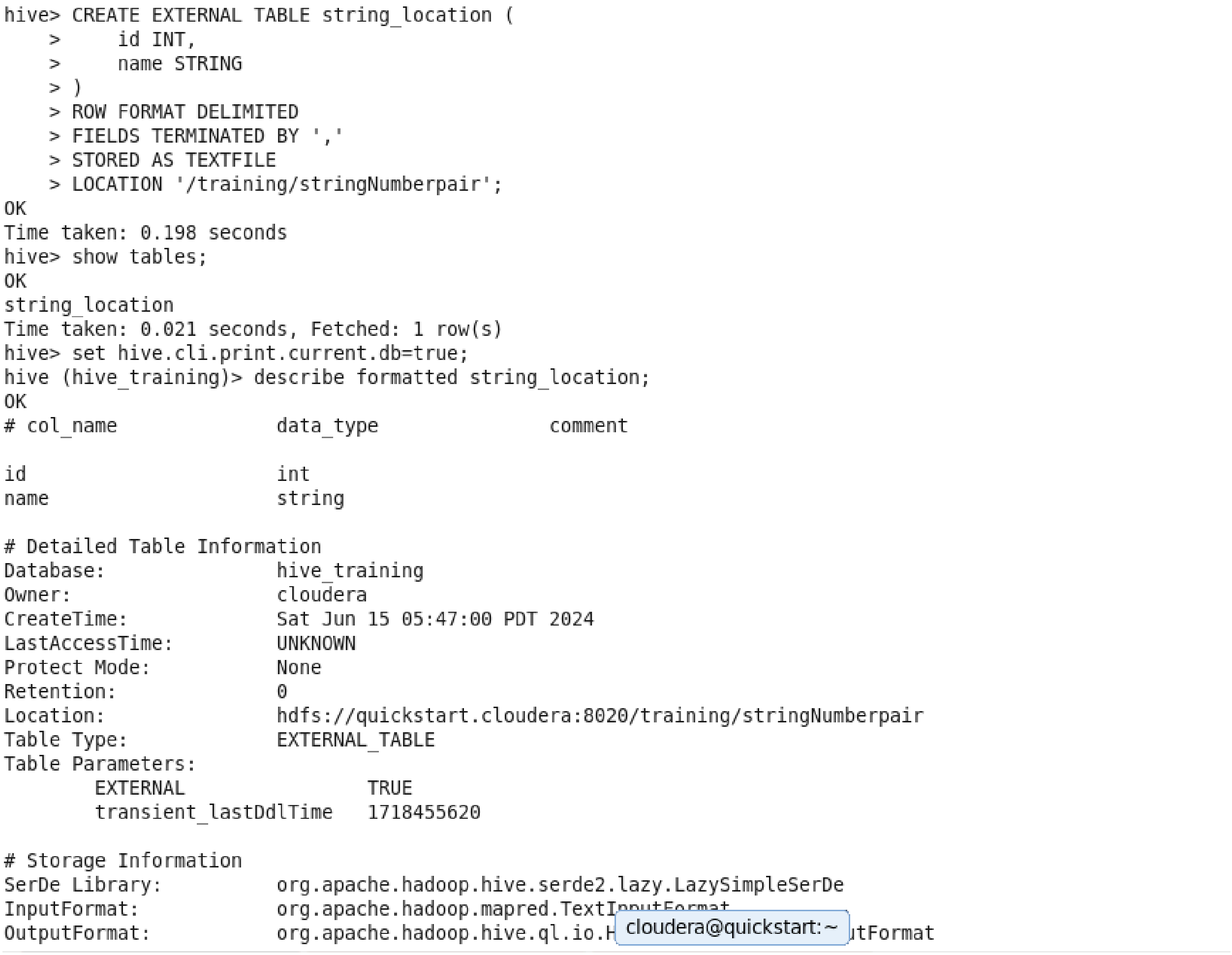
—--

****

EXTERNAL TABLE USING LOCATION

****

CREATE EXTERNAL TABLE string_location (
   id INT,
   name STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/training/';



```
hive> CREATE EXTERNAL TABLE string_location (
    >     id INT,
    >     name STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/training/stringNumberpair';
OK
Time taken: 0.198 seconds
hive> show tables;
OK
string_location
Time taken: 0.021 seconds, Fetched: 1 row(s)
hive> set hive.cli.print.current.db=true;
hive (hive_training)> describe formatted string_location;
OK
# col_name              data_type               comment

id                      int
name                    string

# Detailed Table Information
Database:               hive_training
Owner:                  cloudera
CreateTime:             Sat Jun 15 05:47:00 PDT 2024
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/training/stringNumberpair
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        transient_lastDdlTime   1718455620

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.H                    utFormat
```

Even we try to delete this table, it wont get deleted.

—---

Recap:

A. Managed Table

    - By default, whenever we create the table, its managed/ internal table.
    - It is always created under the location called (/user/hive/warehouse/)
    - Whenever you drop the Managed table, then the table gets dropped also your underlying HDFS directory gets deleted holding the data and schema.

B. External Table

- To create the external table in hive we need to use/write external keyword explicitly while creating the table.
- If you dont specify the location as an argument, again the directory would be created under (/user/hive/warehouse)
- If you want to create the directory for Hive tablein any other location then you need to use the location argument with path.
- Whenever you drop the external table, the table gets dropped but the underlying HDFS data is still available.

****

HIVE PARTITIONING

****

****

STATIC PARTITIONING

****

Scenario 1: Client sending 3 files

Create the Data:

emp_ind.txt-
id,name,city,age
100,Komal,Mumbai, 26
101,Komi,Airoli,25
102,Komu,NaviMumbai,26
103,Koma,Airoli,26

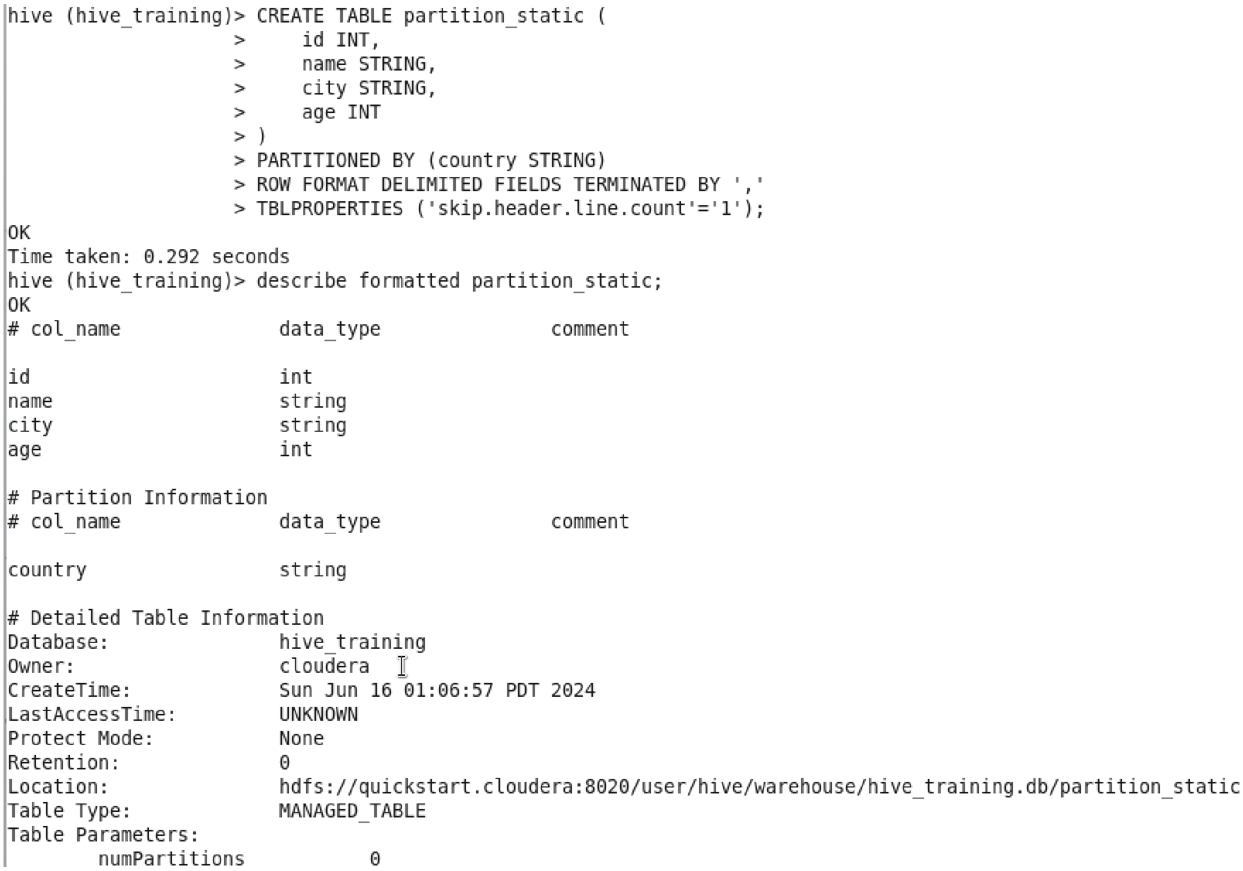
emp_us.txt-
id,name,city,age
200,Hari,CA,40
429,Ram,Texas,39

404,King,dallas,52

emp_uk.txt-
id,name,city,age
300,John,London,40
301,King,London,33
302,Samuel,Edenburg,52

Create static Partition Table:

```
CREATE TABLE partition_static (
   id INT,
   name STRING,
   city STRING,
   age INT
)
PARTITIONED BY (country STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');
```



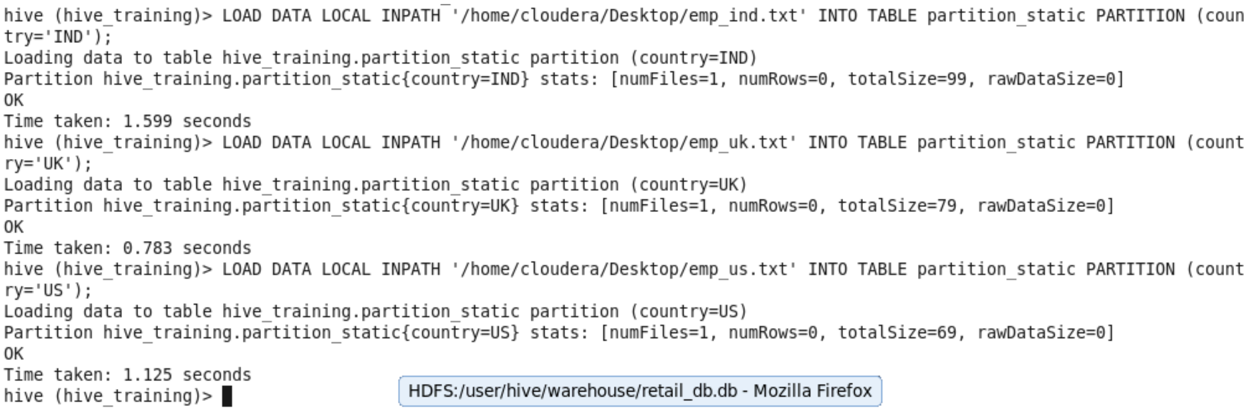
```
hive (hive_training)> CREATE TABLE partition_static (
                    >     id INT,
                    >     name STRING,
                    >     city STRING,
                    >     age INT
                    > )
                    > PARTITIONED BY (country STRING)
                    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
                    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.292 seconds
hive (hive_training)> describe formatted partition_static;
OK
# col_name              data_type               comment

id                      int
name                    string
city                    string
age                     int

# Partition Information
# col_name              data_type               comment

country                 string

# Detailed Table Information
Database:               hive_training
Owner:                  cloudera
CreateTime:             Sun Jun 16 01:06:57 PDT 2024
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_static
Table Type:             MANAGED_TABLE
Table Parameters:
        numPartitions           0
```

Load data:

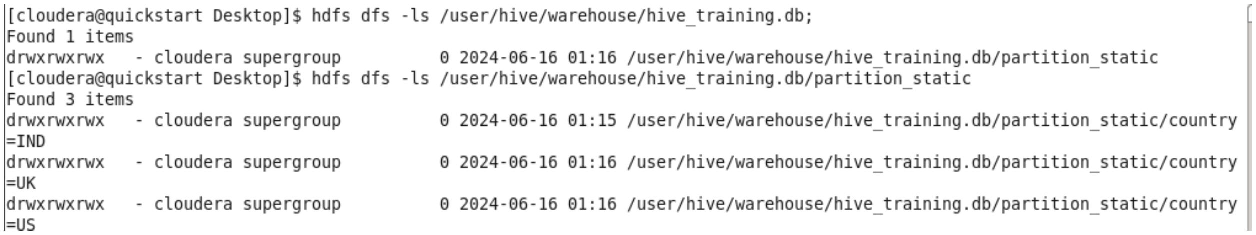
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_ind.txt' INTO TABLE partition_static PARTITION (country='IND');

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_uk.txt' INTO TABLE partition_static PARTITION (country='UK');

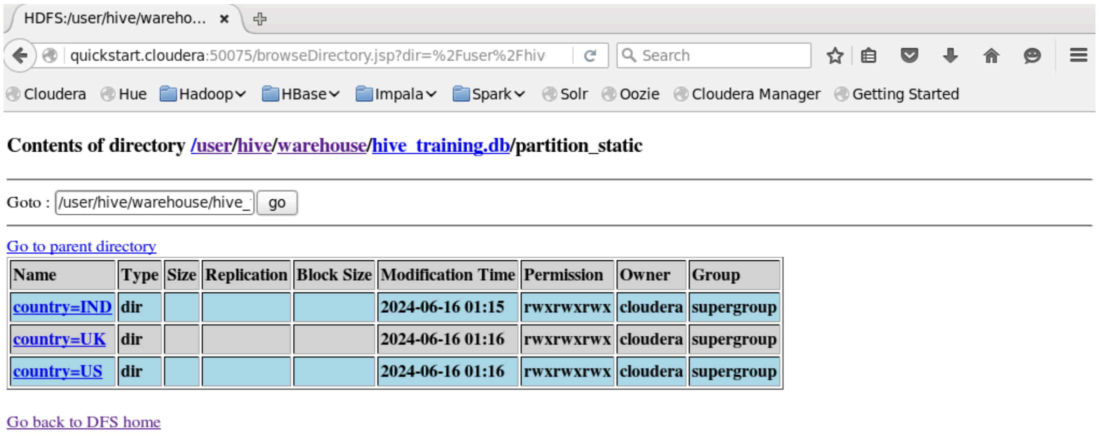
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_us.txt' INTO TABLE partition_static PARTITION (country='US');



```
hive (hive_training)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_ind.txt' INTO TABLE partition_static PARTITION (coun
try='IND');
Loading data to table hive_training.partition_static partition (country=IND)
Partition hive_training.partition_static{country=IND} stats: [numFiles=1, numRows=0, totalSize=99, rawDataSize=0]
OK
Time taken: 1.599 seconds
hive (hive_training)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_uk.txt' INTO TABLE partition_static PARTITION (count
ry='UK');
Loading data to table hive_training.partition_static partition (country=UK)
Partition hive_training.partition_static{country=UK} stats: [numFiles=1, numRows=0, totalSize=79, rawDataSize=0]
OK
Time taken: 0.783 seconds
hive (hive_training)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_us.txt' INTO TABLE partition_static PARTITION (count
ry='US');
Loading data to table hive_training.partition_static partition (country=US)
Partition hive_training.partition_static{country=US} stats: [numFiles=1, numRows=0, totalSize=69, rawDataSize=0]
OK
Time taken: 1.125 seconds
hive (hive_training)>
```



```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /user/hive/warehouse/hive_training.db;
Found 1 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:16 /user/hive/warehouse/hive_training.db/partition_static
[cloudera@quickstart Desktop]$ hdfs dfs -ls /user/hive/warehouse/hive_training.db/partition_static
Found 3 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:15 /user/hive/warehouse/hive_training.db/partition_static/country
=IND
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:16 /user/hive/warehouse/hive_training.db/partition_static/country
=UK
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:16 /user/hive/warehouse/hive_training.db/partition_static/country
=US
```



HDFS:/user/hive/wareho...
quickstart.cloudera:50075/browseDirectory.jsp?dir=%2Fuser%2Fhiv
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started

Contents of directory /user/hive/warehouse/hive_training.db/partition_static

Goto : /user/hive/warehouse/hive_ go

Go to parent directory

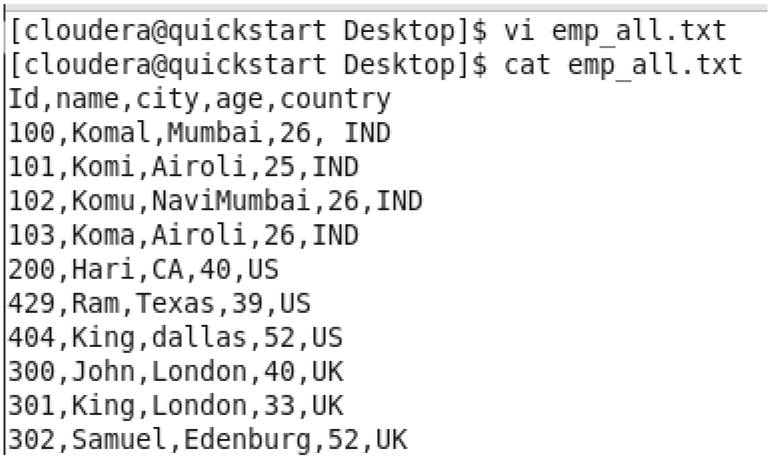
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
country=IND	dir				2024-06-16 01:15	rwxrwxrwx	cloudera	supergroup
country=UK	dir				2024-06-16 01:16	rwxrwxrwx	cloudera	supergroup
country=US	dir				2024-06-16 01:16	rwxrwxrwx	cloudera	supergroup

Go back to DFS home

Scenario 2: client sends 1 file

Data: (emp_all.txt)

Id,name,city,age,country
100,Komal,Mumbai,26, IND
101,Komi,Airoli,25,IND
102,Komu,NaviMumbai,26,IND
103,Koma,Airoli,26,IND
200,Hari,CA,40,US
429,Ram,Texas,39,US
404,King,dallas,52,US
300,John,London,40,UK
301,King,London,33,UK
302,Samuel,Edenburg,52,UK



```
[cloudera@quickstart Desktop]$ vi emp_all.txt
[cloudera@quickstart Desktop]$ cat emp_all.txt
Id,name,city,age,country
100,Komal,Mumbai,26, IND
101,Komi,Airoli,25,IND
102,Komu,NaviMumbai,26,IND
103,Koma,Airoli,26,IND
200,Hari,CA,40,US
429,Ram,Texas,39,US
404,King,dallas,52,US
300,John,London,40,UK
301,King,London,33,UK
302,Samuel,Edenburg,52,UK
```

👍

+++
If the Partition is a part of Data
+++

Create the Partition table 🙂

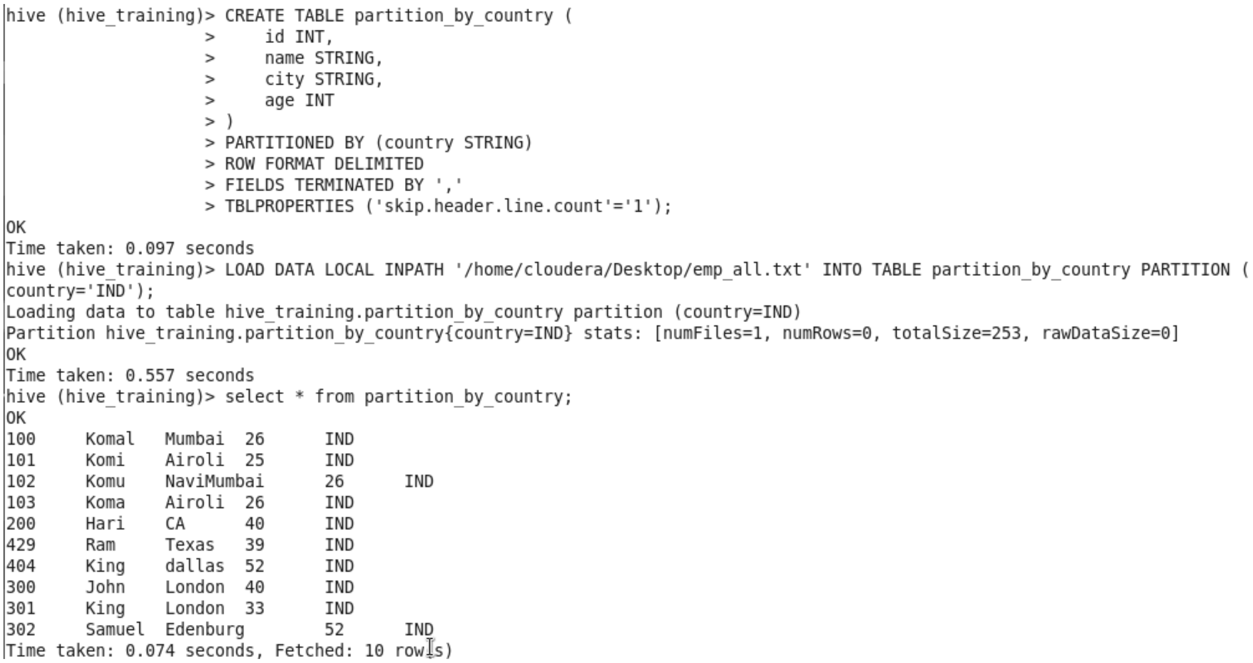
In static partition, the column in which the partition is made should not be present in create table query.

```
CREATE TABLE partition_by_country (
    id INT,
    name STRING,
    city STRING,
    age INT
```

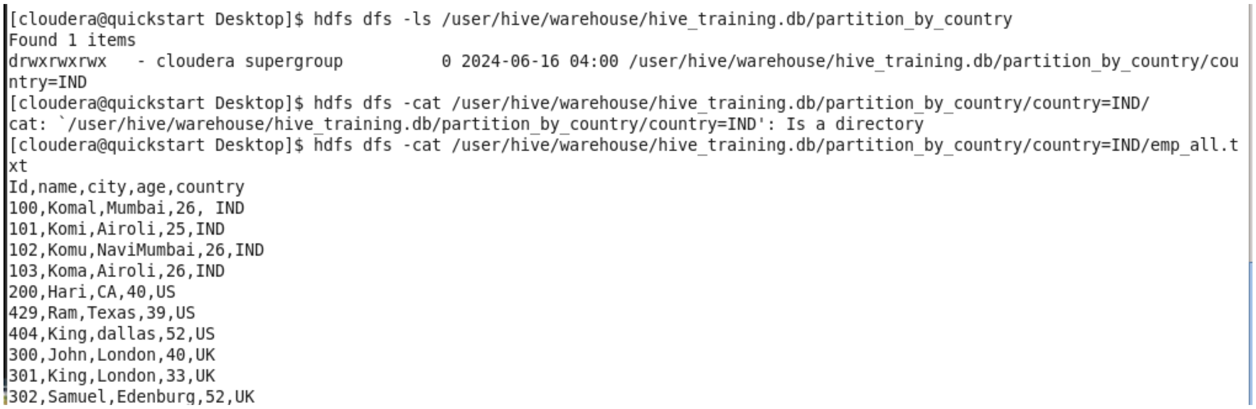
)
PARTITIONED BY (country STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

Load the data 👍

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE partition_by_country PARTITION (country='IND');



```
hive (hive_training)> CREATE TABLE partition_by_country (
                    >     id INT,
                    >     name STRING,
                    >     city STRING,
                    >     age INT
                    > )
                    > PARTITIONED BY (country STRING)
                    > ROW FORMAT DELIMITED
                    > FIELDS TERMINATED BY ','
                    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.097 seconds
hive (hive_training)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE partition_by_country PARTITION (
country='IND');
Loading data to table hive_training.partition_by_country partition (country=IND)
Partition hive_training.partition_by_country{country=IND} stats: [numFiles=1, numRows=0, totalSize=253, rawDataSize=0]
OK
Time taken: 0.557 seconds
hive (hive_training)> select * from partition_by_country;
OK
100     Komal   Mumbai  26      IND
101     Komi    Airoli  25      IND
102     Komu    NaviMumbai      26      IND
103     Koma    Airoli  26      IND
200     Hari    CA      40      IND
429     Ram     Texas   39      IND
404     King    dallas  52      IND
300     John    London  40      IND
301     King    London  33      IND
302     Samuel  Edenburg        52      IND
Time taken: 0.074 seconds, Fetched: 10 rows)
```



```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /user/hive/warehouse/hive_training.db/partition_by_country
Found 1 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 04:00 /user/hive/warehouse/hive_training.db/partition_by_country/cou
ntry=IND
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=IND/
cat: `/user/hive/warehouse/hive_training.db/partition_by_country/country=IND': Is a directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=IND/emp_all.t
xt
Id,name,city,age,country
100,Komal,Mumbai,26, IND
101,Komi,Airoli,25,IND
102,Komu,NaviMumbai,26,IND
103,Koma,Airoli,26,IND
200,Hari,CA,40,US
429,Ram,Texas,39,US
404,King,dallas,52,US
300,John,London,40,UK
301,King,London,33,UK
302,Samuel,Edenburg,52,UK
```

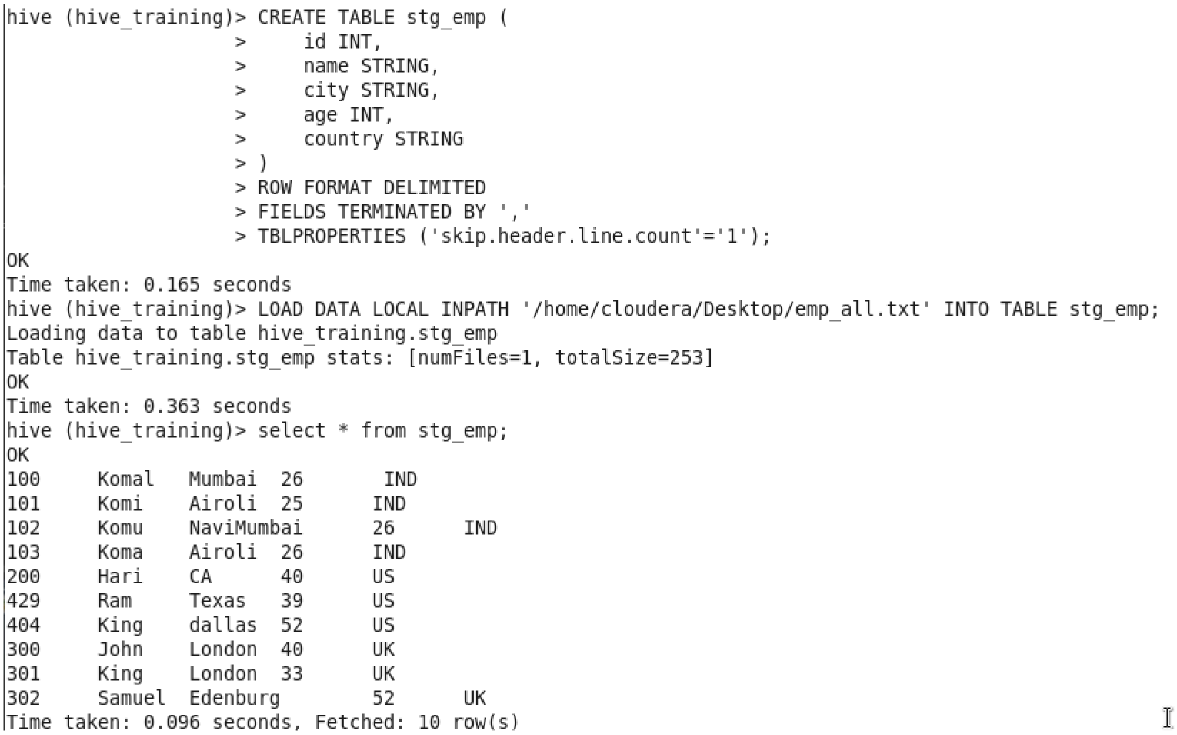
You can see that the partition table and the actual table is different.

So, create the intermediate table 👏

```
CREATE TABLE stg_emp (
   id INT,
   name STRING,
   city STRING,
   age INT,
   country STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');
```

—---

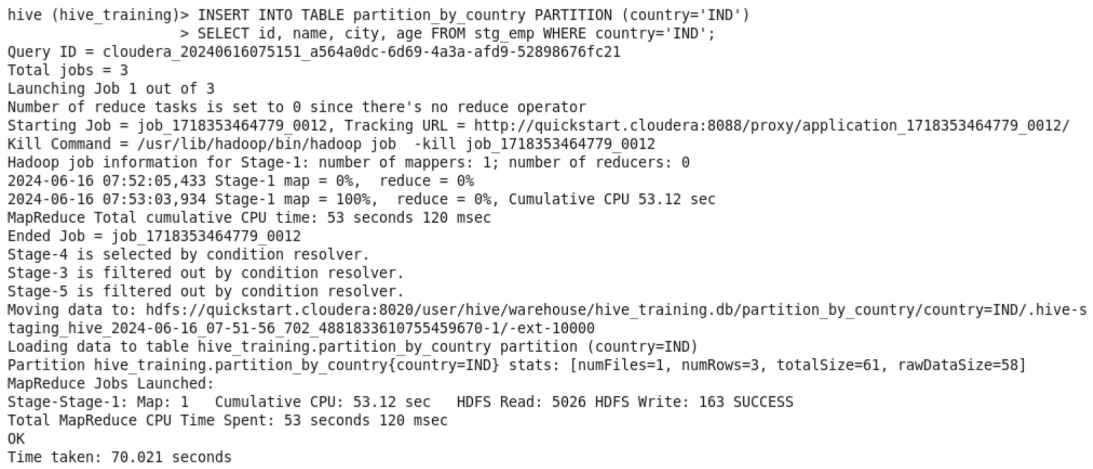
```
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE stg_emp;
```



```
hive (hive_training)> CREATE TABLE stg_emp (
                    >     id INT,
                    >     name STRING,
                    >     city STRING,
                    >     age INT,
                    >     country STRING
                    > )
                    > ROW FORMAT DELIMITED
                    > FIELDS TERMINATED BY ','
                    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.165 seconds
hive (hive_training)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE stg_emp;
Loading data to table hive_training.stg_emp
Table hive_training.stg_emp stats: [numFiles=1, totalSize=253]
OK
Time taken: 0.363 seconds
hive (hive_training)> select * from stg_emp;
OK
100     Komal   Mumbai  26       IND
101     Komi    Airoli  25      IND
102     Komu    NaviMumbai      26      IND
103     Koma    Airoli  26      IND
200     Hari    CA      40      US
429     Ram     Texas   39      US
404     King    dallas  52      US
300     John    London  40      UK
301     King    London  33      UK
302     Samuel  Edenburg        52      UK
Time taken: 0.096 seconds, Fetched: 10 row(s)
```

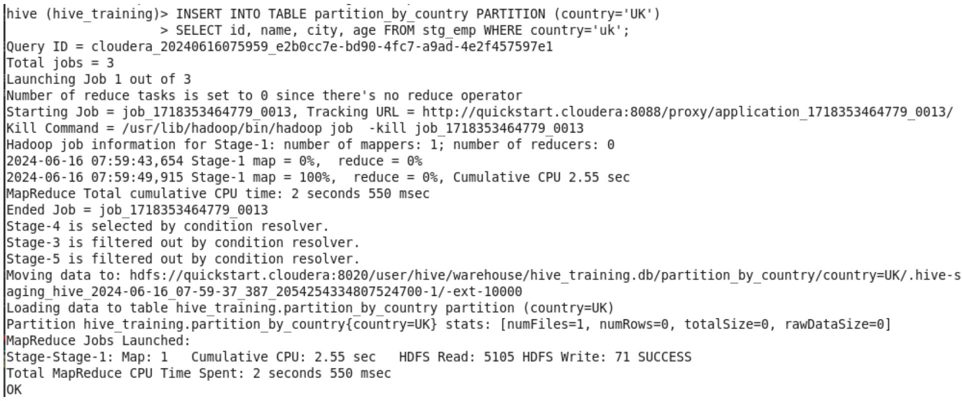
```
truncate table partition_by_country;
```

```
INSERT INTO TABLE partition_by_country PARTITION (country='IND')
SELECT id, name, city, age FROM stg_emp WHERE country='IND';
```



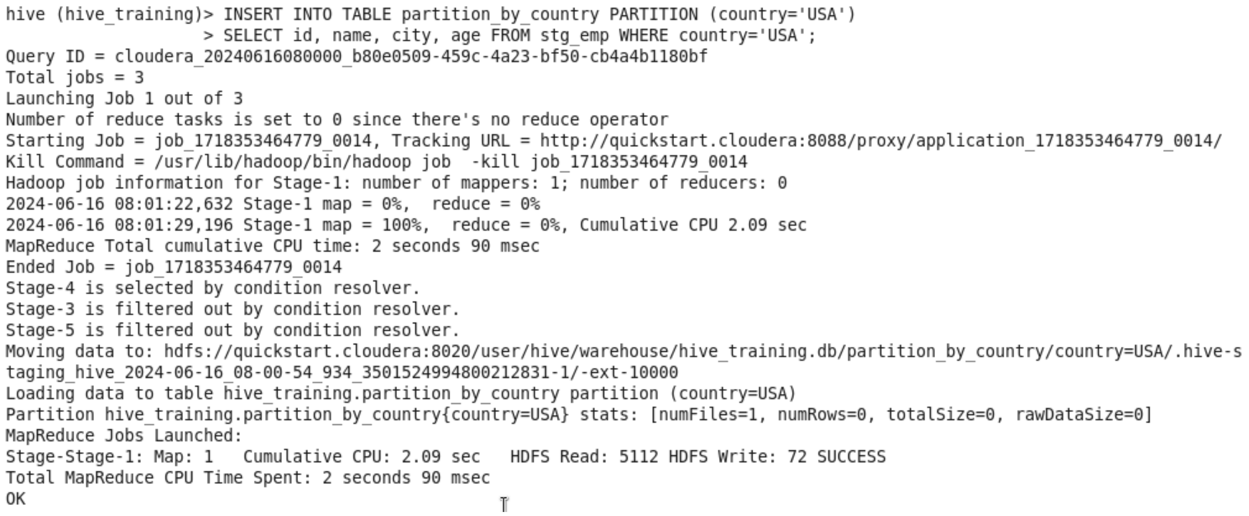
```
hive (hive_training)> INSERT INTO TABLE partition_by_country PARTITION (country='IND')
                    > SELECT id, name, city, age FROM stg_emp WHERE country='IND';
Query ID = cloudera_20240616075151_a564a0dc-6d69-4a3a-afd9-52898676fc21
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1718353464779_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718353464779_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718353464779_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-16 07:52:05,433 Stage-1 map = 0%,  reduce = 0%
2024-06-16 07:53:03,934 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 53.12 sec
MapReduce Total cumulative CPU time: 53 seconds 120 msec
Ended Job = job_1718353464779_0012
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_by_country/country=IND/.hive-s
taging_hive_2024-06-16_07-51-56_702_4881833610755459670-1/-ext-10000
Loading data to table hive_training.partition_by_country partition (country=IND)
Partition hive_training.partition_by_country{country=IND} stats: [numFiles=1, numRows=3, totalSize=61, rawDataSize=58]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 53.12 sec   HDFS Read: 5026 HDFS Write: 163 SUCCESS
Total MapReduce CPU Time Spent: 53 seconds 120 msec
OK
Time taken: 70.021 seconds
```

```
INSERT INTO TABLE partition_by_country PARTITION (country='UK')
SELECT id, name, city, age FROM stg_emp WHERE country='uk';
```



```
hive (hive_training)> INSERT INTO TABLE partition_by_country PARTITION (country='UK')
                    > SELECT id, name, city, age FROM stg_emp WHERE country='uk';
Query ID = cloudera_20240616075959_e2b0cc7e-bd90-4fc7-a9ad-4e2f457597e1
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_1718353464779_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718353464779_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718353464779_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-16 07:59:43,654 Stage-1 map = 0%,  reduce = 0%
2024-06-16 07:59:49,915 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.55 sec
MapReduce Total cumulative CPU time: 2 seconds 550 msec
Ended Job = job_1718353464779_0013
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_by_country/country=UK/.hive-st
aging_hive_2024-06-16_07-59-37_387_2054254334807524700-1/-ext-10000
Loading data to table hive_training.partition_by_country partition (country=UK)
Partition hive_training.partition_by_country{country=UK} stats: [numFiles=1, numRows=0, totalSize=0, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.55 sec   HDFS Read: 5105 HDFS Write: 71 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 550 msec
OK
```

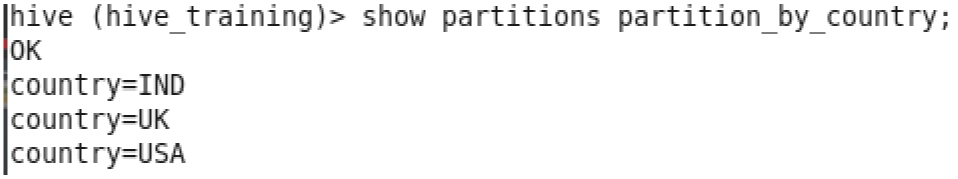
```
INSERT INTO TABLE partition_by_country PARTITION (country='USA')
SELECT id, name, city, age FROM stg_emp WHERE country='USA';
```



```
hive (hive_training)> INSERT INTO TABLE partition_by_country PARTITION (country='USA')
                    > SELECT id, name, city, age FROM stg_emp WHERE country='USA';
Query ID = cloudera_20240616080000_b80e0509-459c-4a23-bf50-cb4a4b1180bf
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_1718353464779_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718353464779_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718353464779_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-16 08:01:22,632 Stage-1 map = 0%,  reduce = 0%
2024-06-16 08:01:29,196 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.09 sec
MapReduce Total cumulative CPU time: 2 seconds 90 msec
Ended Job = job_1718353464779_0014
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_by_country/country=USA/.hive-s
taging_hive_2024-06-16_08-00-54_934_3501524994800212831-1/-ext-10000
Loading data to table hive_training.partition_by_country partition (country=USA)
Partition hive_training.partition_by_country{country=USA} stats: [numFiles=1, numRows=0, totalSize=0, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.09 sec   HDFS Read: 5112 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 90 msec
OK
```

To see the partition:

show partitions partition_by_country;

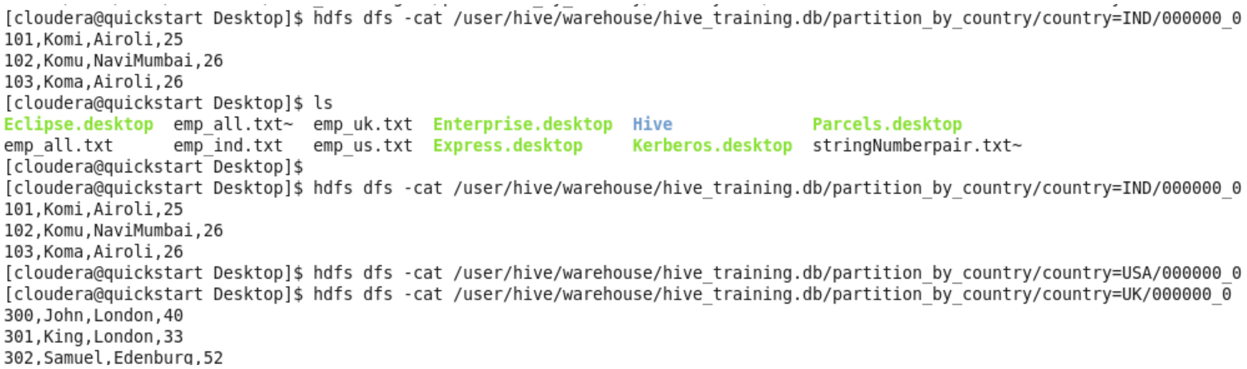


```
hive (hive_training)> show partitions partition_by_country;
OK
country=IND
country=UK
country=USA
```

INSERT OVERWRITE TABLE partition_by_country PARTITION (country='IND')
SELECT id, name, city, age FROM stg_emp WHERE country='IND';

INSERT OVERWRITE TABLE partition_by_country PARTITION (country='UK')
SELECT id, name, city, age FROM stg_emp WHERE country='UK';

INSERT OVERWRITE TABLE partition_by_country PARTITION (country='US')
SELECT id, name, city, age FROM stg_emp WHERE country='US';



```
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=IND/000000_0
101,Komi,Airoli,25
102,Komu,NaviMumbai,26
103,Koma,Airoli,26
[cloudera@quickstart Desktop]$ ls
Eclipse.desktop  emp_all.txt~  emp_uk.txt  Enterprise.desktop  Hive              Parcels.desktop
emp_all.txt      emp_ind.txt   emp_us.txt  Express.desktop     Kerberos.desktop  stringNumberpair.txt~
[cloudera@quickstart Desktop]$
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=IND/000000_0
101,Komi,Airoli,25
102,Komu,NaviMumbai,26
103,Koma,Airoli,26
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=USA/000000_0
[cloudera@quickstart Desktop]$ hdfs dfs -cat /user/hive/warehouse/hive_training.db/partition_by_country/country=UK/000000_0
300,John,London,40
301,King,London,33
302,Samuel,Edenburg,52
```

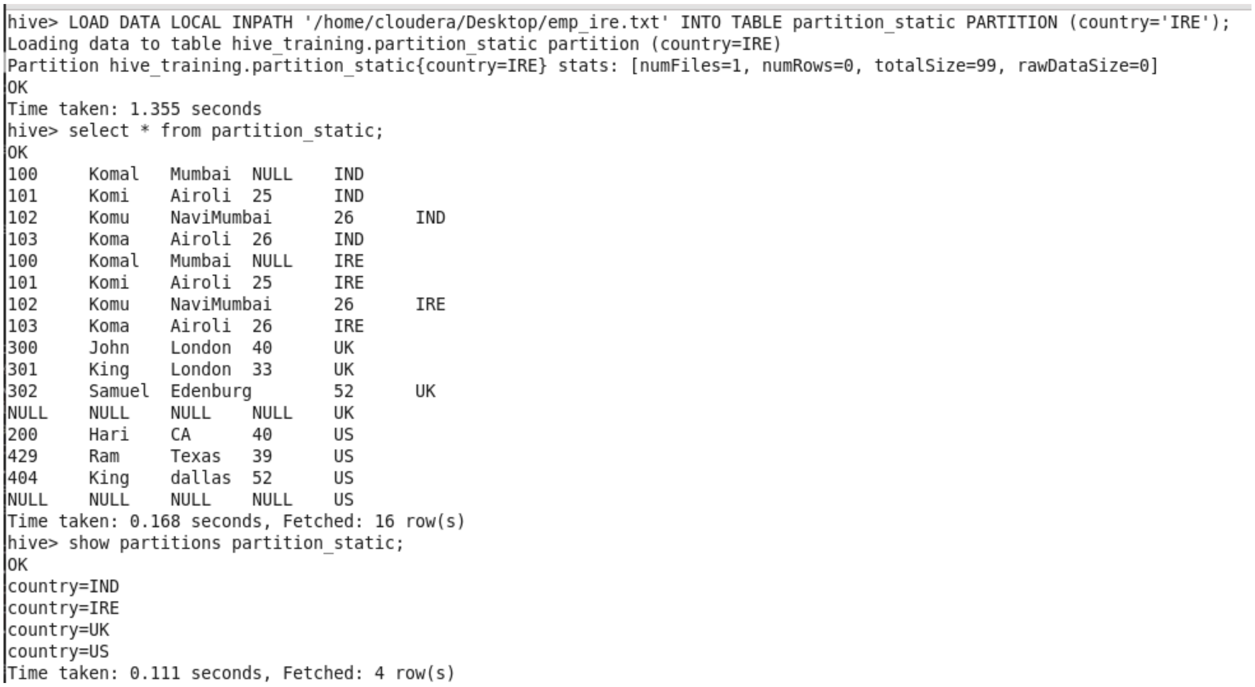
—---

Exercise:

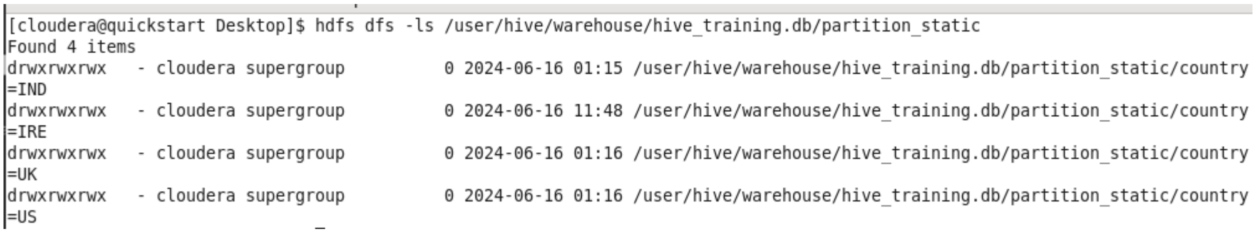
Trying to create the partitions for one more country:

cp emp_ind.txt emp_ire.txt

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_ire.txt' INTO TABLE partition_static PARTITION (country='IRE');



```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_ire.txt' INTO TABLE partition_static PARTITION (country='IRE');
Loading data to table hive_training.partition_static partition (country=IRE)
Partition hive_training.partition_static{country=IRE} stats: [numFiles=1, numRows=0, totalSize=99, rawDataSize=0]
OK
Time taken: 1.355 seconds
hive> select * from partition_static;
OK
100     Komal   Mumbai  NULL    IND
101     Komi    Airoli  25      IND
102     Komu    NaviMumbai      26      IND
103     Koma    Airoli  26      IND
100     Komal   Mumbai  NULL    IRE
101     Komi    Airoli  25      IRE
102     Komu    NaviMumbai      26      IRE
103     Koma    Airoli  26      IRE
300     John    London  40      UK
301     King    London  33      UK
302     Samuel  Edenburg        52      UK
NULL    NULL    NULL    NULL    UK
200     Hari    CA      40      US
429     Ram     Texas   39      US
404     King    dallas  52      US
NULL    NULL    NULL    NULL    US
Time taken: 0.168 seconds, Fetched: 16 row(s)
hive> show partitions partition_static;
OK
country=IND
country=IRE
country=UK
country=US
Time taken: 0.111 seconds, Fetched: 4 row(s)
```

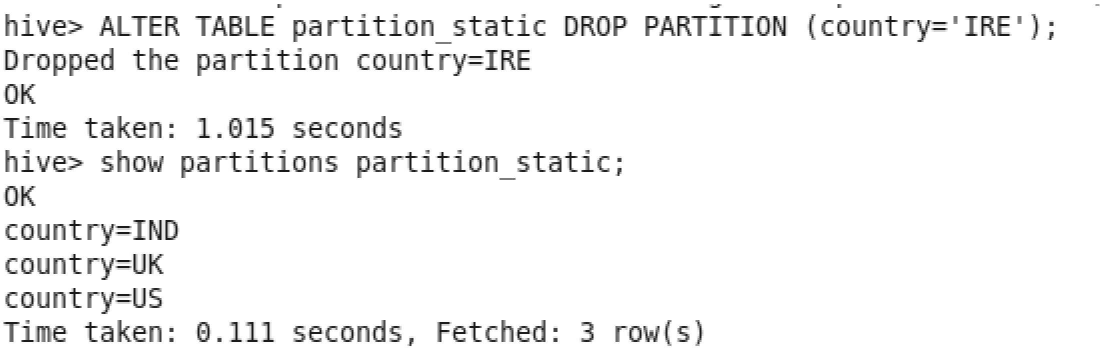


```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /user/hive/warehouse/hive_training.db/partition_static
Found 4 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:15 /user/hive/warehouse/hive_training.db/partition_static/country
=IND
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 11:48 /user/hive/warehouse/hive_training.db/partition_static/country
=IRE
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:16 /user/hive/warehouse/hive_training.db/partition_static/country
=UK
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 01:16 /user/hive/warehouse/hive_training.db/partition_static/country
=US
```

—---

Dropping the Partitions:

ALTER TABLE partition_static DROP PARTITION (country='IRE');



```
hive> ALTER TABLE partition_static DROP PARTITION (country='IRE');
Dropped the partition country=IRE
OK
Time taken: 1.015 seconds
hive> show partitions partition_static;
OK
country=IND
country=UK
country=US
Time taken: 0.111 seconds, Fetched: 3 row(s)
```

—--

RECAP:

Case 1 👏 If the Partition is a part of Data

- Create the temporary table and fetch the records from there.

Case 2 😀If the partition column is not the part of a table

- Load the Data

—---

****

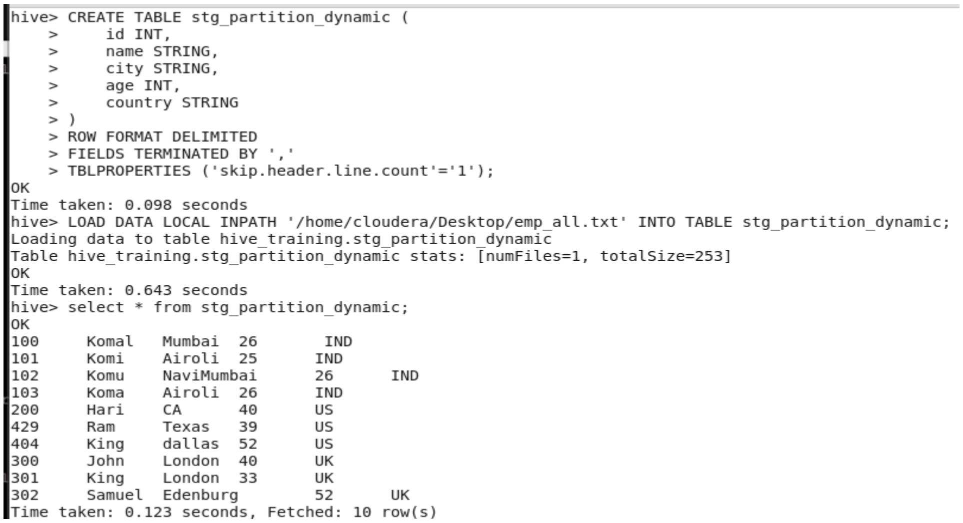
DYNAMIC PARTITIONING

****

```
CREATE TABLE stg_partition_dynamic (
   id INT,
   name STRING,
   city STRING,
   age INT,
   country STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');
```

—---

```
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE
stg_partition_dynamic;
```

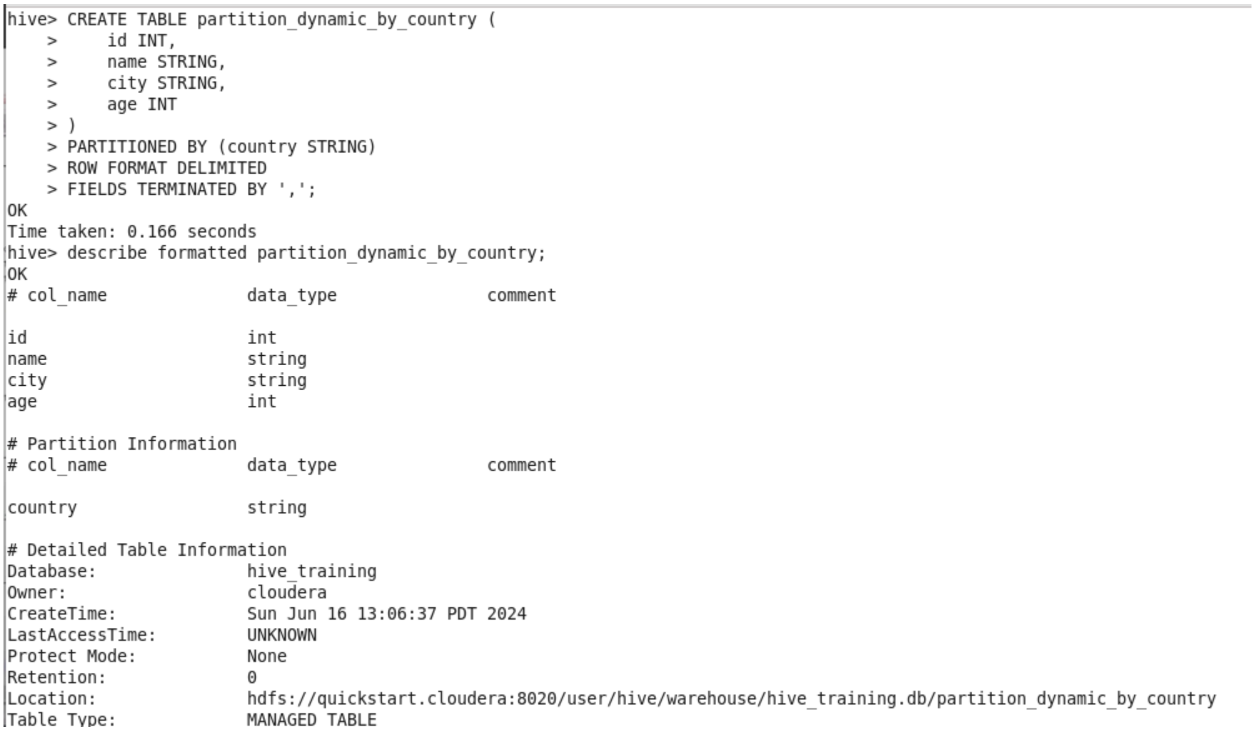


```
hive> CREATE TABLE stg_partition_dynamic (
    >     id INT,
    >     name STRING,
    >     city STRING,
    >     age INT,
    >     country STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.098 seconds
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/emp_all.txt' INTO TABLE stg_partition_dynamic;
Loading data to table hive_training.stg_partition_dynamic
Table hive_training.stg_partition_dynamic stats: [numFiles=1, totalSize=253]
OK
Time taken: 0.643 seconds
hive> select * from stg_partition_dynamic;
OK
100     Komal   Mumbai  26       IND
101     Komi    Airoli  25      IND
102     Komu    NaviMumbai      26      IND
103     Koma    Airoli  26      IND
200     Hari    CA      40      US
429     Ram     Texas   39      US
404     King    dallas  52      US
300     John    London  40      UK
301     King    London  33      UK
302     Samuel  Edenburg        52      UK
Time taken: 0.123 seconds, Fetched: 10 row(s)
```

—---

Now we will create the partition table for country column 😡

```
CREATE TABLE partition_dynamic_by_country (
   id INT,
   name STRING,
   city STRING,
   age INT
)
PARTITIONED BY (country STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```



```
hive> CREATE TABLE partition_dynamic_by_country (
    >     id INT,
    >     name STRING,
    >     city STRING,
    >     age INT
    > )
    > PARTITIONED BY (country STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.166 seconds
hive> describe formatted partition_dynamic_by_country;
OK
# col_name              data_type               comment

id                      int
name                    string
city                    string
age                     int

# Partition Information
# col_name              data_type               comment

country                 string

# Detailed Table Information
Database:               hive_training
Owner:                  cloudera
CreateTime:             Sun Jun 16 13:06:37 PDT 2024
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_dynamic_by_country
Table Type:             MANAGED_TABLE
```

—---

INSERT INTO TABLE partition_dynamic_by_country partition(country) select id, name,city,age,country from stg_partition_dynamic;

—--

Comparing the changes for above Insert command for both static and dynamic 👍

Changes 1 ➕No need to mention the value for the Partition colum (no need of hard coding)

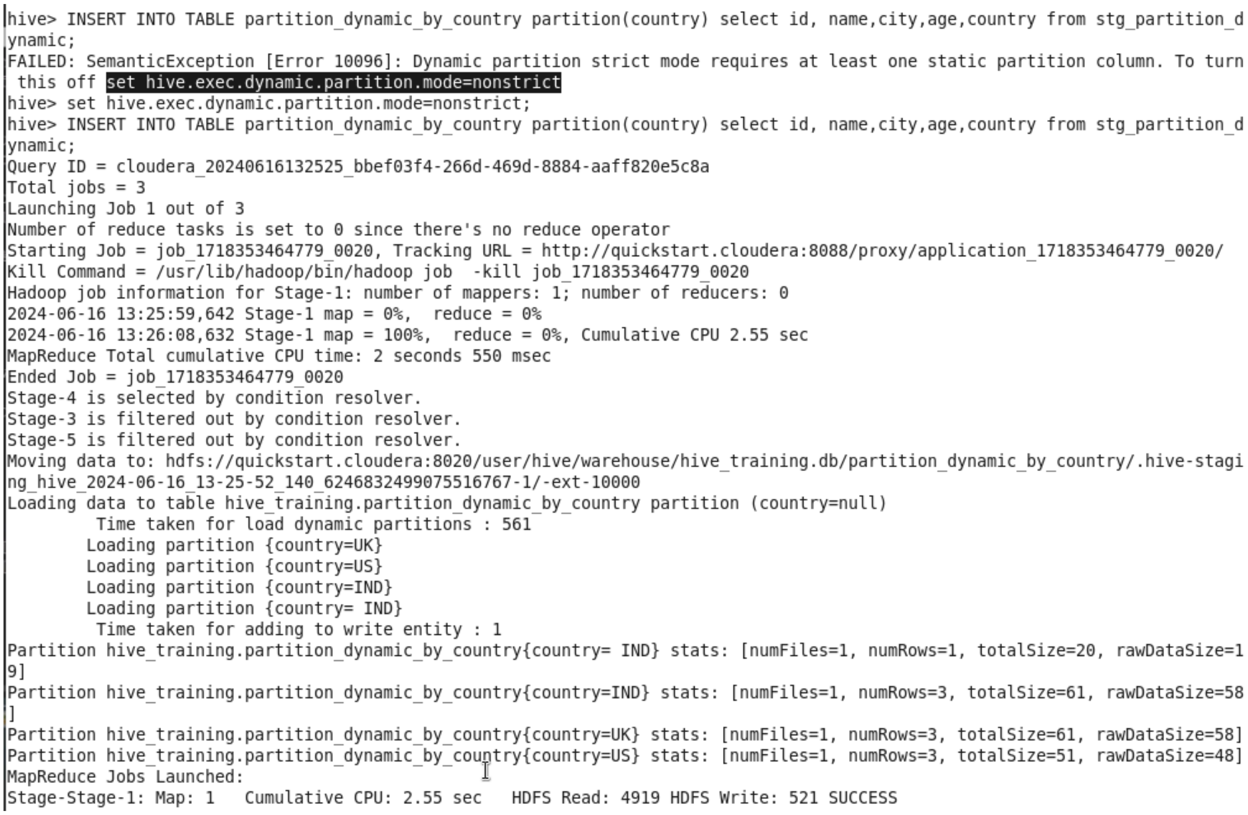
Changes 2 ➕partition column will be present in select clause but as the last column name.

Changes 3 ➕No need of WHERE clause.

—---

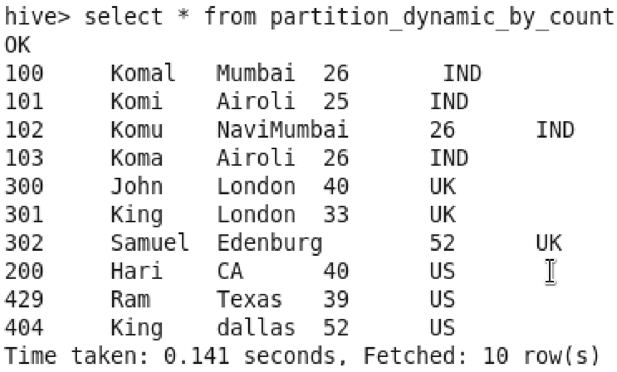
NOTE 🎉

set hive.exec.dynamic.partition=true; (It enables the dynamic partition)
set hive.exec.dynamic.partition.mode=nonstrict; (it allows the dynamic partition)
set hive.exec.max.dynamic.partition.mode=100;
set hive.exec.max.dynamic.partitions.pernode=100;
—--

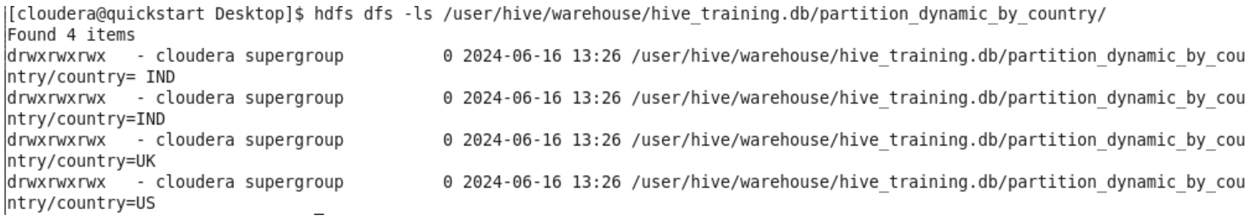


```
hive> INSERT INTO TABLE partition_dynamic_by_country partition(country) select id, name,city,age,country from stg_partition_d
ynamic;
FAILED: SemanticException [Error 10096]: Dynamic partition strict mode requires at least one static partition column. To turn
 this off set hive.exec.dynamic.partition.mode=nonstrict
hive> set hive.exec.dynamic.partition.mode=nonstrict;
hive> INSERT INTO TABLE partition_dynamic_by_country partition(country) select id, name,city,age,country from stg_partition_d
ynamic;
Query ID = cloudera_20240616132525_bbef03f4-266d-469d-8884-aaff820e5c8a
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_1718353464779_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718353464779_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718353464779_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-16 13:25:59,642 Stage-1 map = 0%,  reduce = 0%
2024-06-16 13:26:08,632 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.55 sec
MapReduce Total cumulative CPU time: 2 seconds 550 msec
Ended Job = job_1718353464779_0020
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_training.db/partition_dynamic_by_country/.hive-stagi
ng_hive_2024-06-16_13-25-52_140_6246832499075516767-1/-ext-10000
Loading data to table hive_training.partition_dynamic_by_country partition (country=null)
         Time taken for load dynamic partitions : 561
        Loading partition {country=UK}
        Loading partition {country=US}
        Loading partition {country=IND}
        Loading partition {country= IND}
         Time taken for adding to write entity : 1
Partition hive_training.partition_dynamic_by_country{country= IND} stats: [numFiles=1, numRows=1, totalSize=20, rawDataSize=1
9]
Partition hive_training.partition_dynamic_by_country{country=IND} stats: [numFiles=1, numRows=3, totalSize=61, rawDataSize=58
]
Partition hive_training.partition_dynamic_by_country{country=UK} stats: [numFiles=1, numRows=3, totalSize=61, rawDataSize=58]
Partition hive_training.partition_dynamic_by_country{country=US} stats: [numFiles=1, numRows=3, totalSize=51, rawDataSize=48]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.55 sec   HDFS Read: 4919 HDFS Write: 521 SUCCESS
```

Data gets loaded:



```
hive> select * from partition_dynamic_by_country;
OK
100     Komal   Mumbai  26       IND
101     Komi    Airoli  25      IND
102     Komu    NaviMumbai      26      IND
103     Koma    Airoli  26      IND
300     John    London  40      UK
301     King    London  33      UK
302     Samuel  Edenburg        52      UK
200     Hari    CA      40      US
429     Ram     Texas   39      US
404     King    dallas  52      US
Time taken: 0.141 seconds, Fetched: 10 row(s)
```



```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /user/hive/warehouse/hive_training.db/partition_dynamic_by_country/
Found 4 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 13:26 /user/hive/warehouse/hive_training.db/partition_dynamic_by_cou
ntry/country= IND
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 13:26 /user/hive/warehouse/hive_training.db/partition_dynamic_by_cou
ntry/country=IND
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 13:26 /user/hive/warehouse/hive_training.db/partition_dynamic_by_cou
ntry/country=UK
drwxrwxrwx   - cloudera supergroup          0 2024-06-16 13:26 /user/hive/warehouse/hive_training.db/partition_dynamic_by_cou
ntry/country=US
```

—---

When to use static partitioning vs dynamic partitioning?

- In static partitioning, we use to load the data multiple times as per partition condition
  While in dynamic, it works with single load statement.

- In case, when we have to extract only one partition condition
  Eg: we have only 1 file with data from various country and assume we only want the data for India
  Here we can hard code the data by applying the partitioning condition.

—---
****

HIVE BUCKETING

****

Data: (emp_bucket.txt)

street,city,zip,state,beds,baths,sq_ft,type,price
3526 HIGH ST,SACREMENTO,95838,CA,2,1,796,RESIDENTIAL,59222
45 TI_LST,LA,97654,LAUS,1,2,798,INDUSTRIAL,49876
456 KALA ST,CA,67890,CALIF,2,1,678,INDUSTRIAL,40000
2 ABBEY ST,DUBLIN,98678,IRE,3,2,898,RESIDENTIAL,98000
2 CORNELL ST,DUBLIN,78907,IRE,2,1,789,RESIDENTIAL,87907

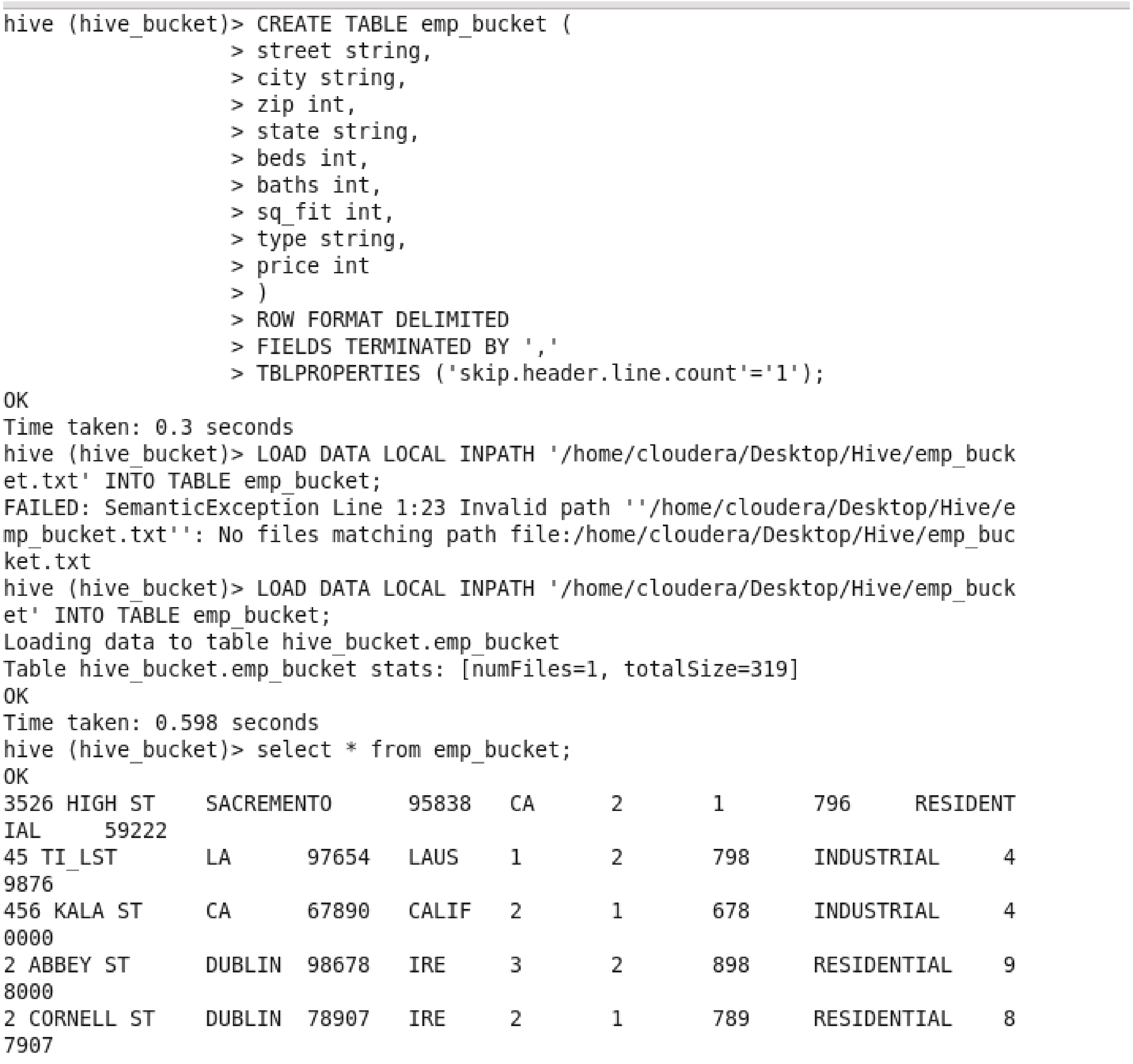
Create a normal table:

```
CREATE TABLE emp_bucket (
street string,
city string,
zip int,
state string,
beds int,
baths int,
sq_fit int,
type string,
price int
)
ROW FORMAT DELIMITED
```

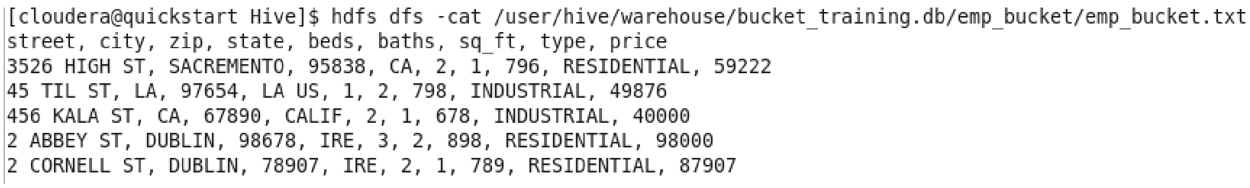
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');
—---

Load data 👍

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/emp_bucket.txt' INTO TABLE emp_bucket;
—--



```
hive (hive_bucket)> CREATE TABLE emp_bucket (
                  > street string,
                  > city string,
                  > zip int,
                  > state string,
                  > beds int,
                  > baths int,
                  > sq_fit int,
                  > type string,
                  > price int
                  > )
                  > ROW FORMAT DELIMITED
                  > FIELDS TERMINATED BY ','
                  > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.3 seconds
hive (hive_bucket)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/emp_buck
et.txt' INTO TABLE emp_bucket;
FAILED: SemanticException Line 1:23 Invalid path ''/home/cloudera/Desktop/Hive/e
mp_bucket.txt'': No files matching path file:/home/cloudera/Desktop/Hive/emp_buc
ket.txt
hive (hive_bucket)> LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/emp_buck
et' INTO TABLE emp_bucket;
Loading data to table hive_bucket.emp_bucket
Table hive_bucket.emp_bucket stats: [numFiles=1, totalSize=319]
OK
Time taken: 0.598 seconds
hive (hive_bucket)> select * from emp_bucket;
OK
3526 HIGH ST    SACREMENTO      95838   CA      2       1       796     RESIDENT
IAL     59222
45 TI_LST       LA      97654   LAUS    1       2       798     INDUSTRIAL     4
9876
456 KALA ST     CA      67890   CALIF   2       1       678     INDUSTRIAL     4
0000
2 ABBEY ST      DUBLIN  98678   IRE     3       2       898     RESIDENTIAL    9
8000
2 CORNELL ST    DUBLIN  78907   IRE     2       1       789     RESIDENTIAL    8
7907
```

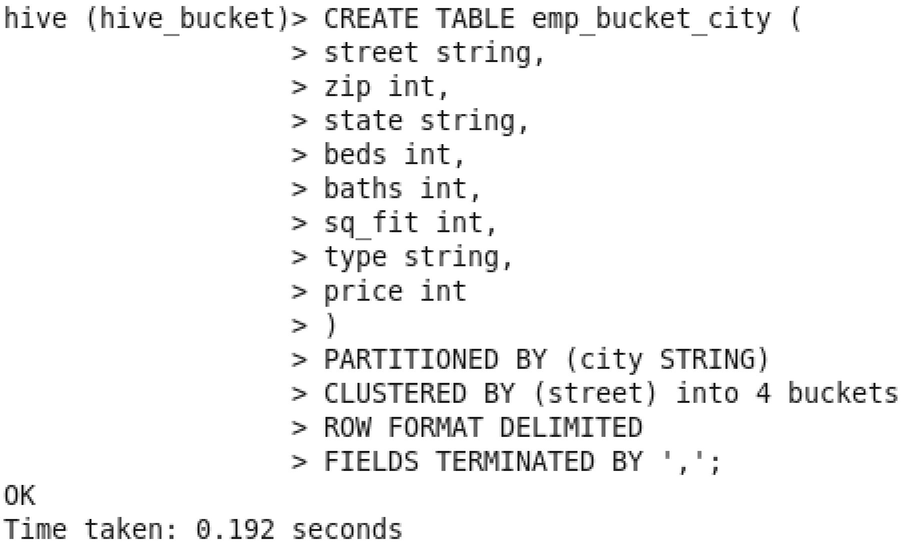


```
[cloudera@quickstart Hive]$ hdfs dfs -cat /user/hive/warehouse/bucket_training.db/emp_bucket/emp_bucket.txt
street, city, zip, state, beds, baths, sq_ft, type, price
3526 HIGH ST, SACREMENTO, 95838, CA, 2, 1, 796, RESIDENTIAL, 59222
45 TIL ST, LA, 97654, LA US, 1, 2, 798, INDUSTRIAL, 49876
456 KALA ST, CA, 67890, CALIF, 2, 1, 678, INDUSTRIAL, 40000
2 ABBEY ST, DUBLIN, 98678, IRE, 3, 2, 898, RESIDENTIAL, 98000
2 CORNELL ST, DUBLIN, 78907, IRE, 2, 1, 789, RESIDENTIAL, 87907
```

—---

Create the partition table:

CREATE TABLE emp_bucket_city (
street string,
zip int,
state string,
beds int,
baths int,
sq_fit int,
type string,
price int
)
PARTITIONED BY (city STRING)
CLUSTERED BY (street) into 4 buckets
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

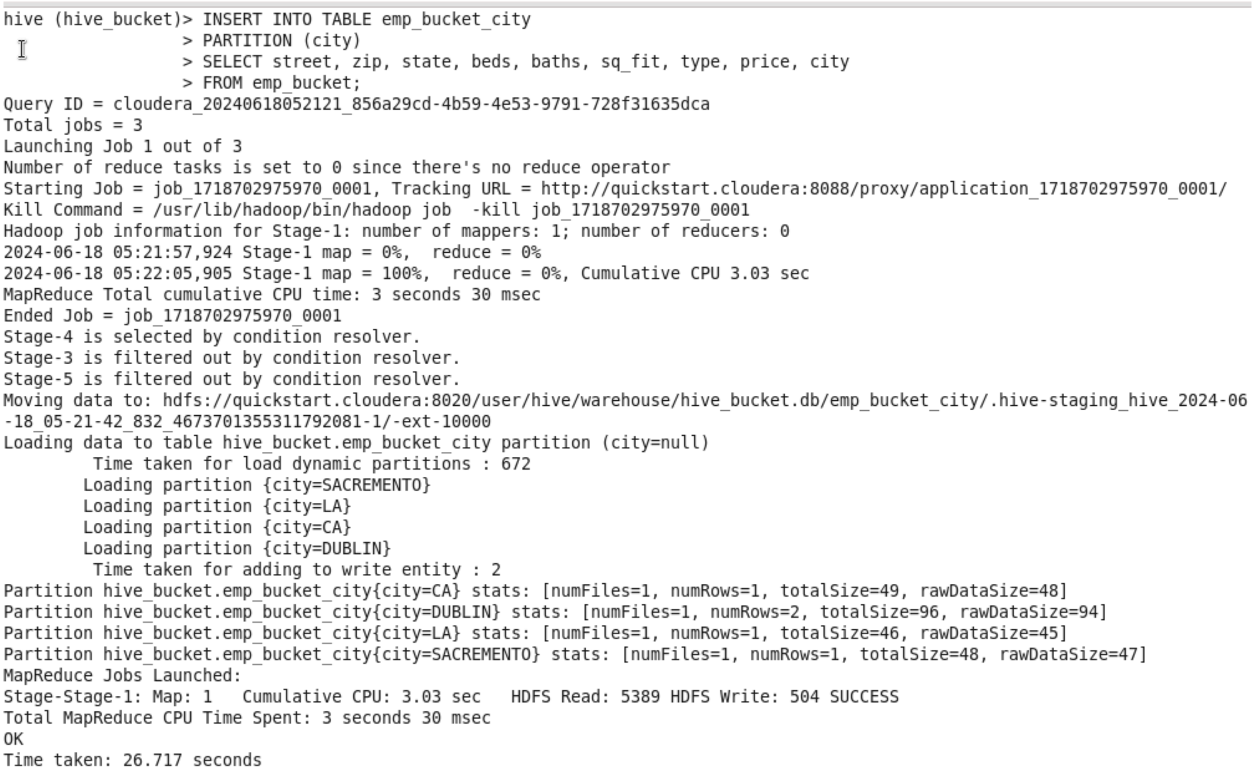


```
hive (hive_bucket)> CREATE TABLE emp_bucket_city (
                  > street string,
                  > zip int,
                  > state string,
                  > beds int,
                  > baths int,
                  > sq_fit int,
                  > type string,
                  > price int
                  > )
                  > PARTITIONED BY (city STRING)
                  > CLUSTERED BY (street) into 4 buckets
                  > ROW FORMAT DELIMITED
                  > FIELDS TERMINATED BY ',';
OK
Time taken: 0.192 seconds
```

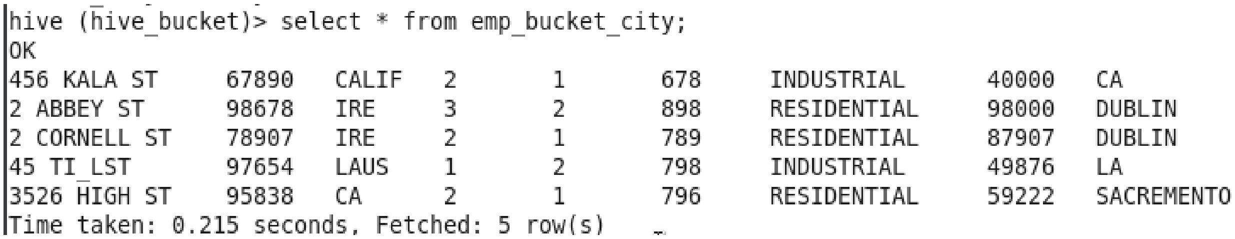
—--

Loading the values into table:

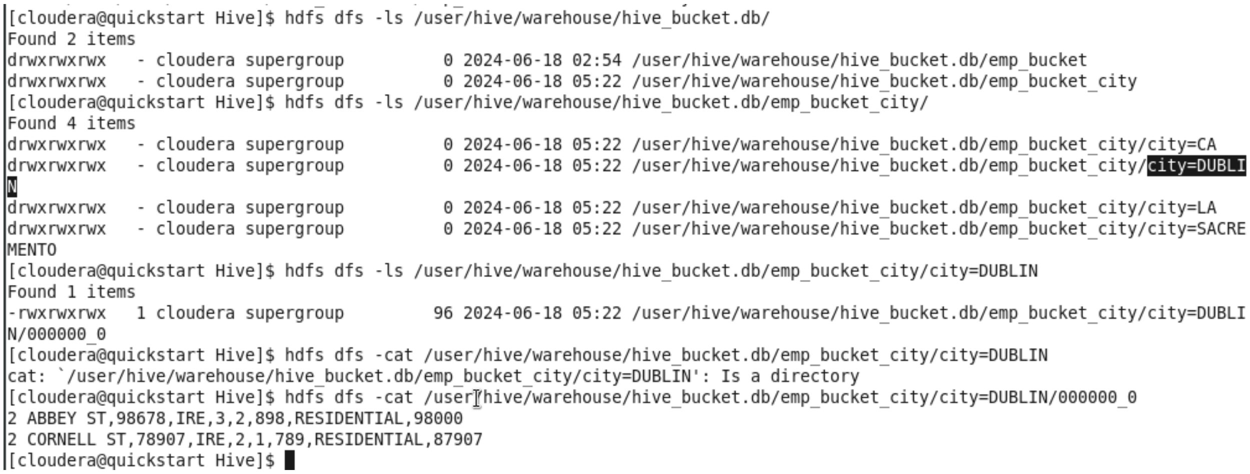
INSERT INTO TABLE emp_bucket_city
PARTITION (city)
SELECT street, zip, state, beds, baths, sq_fit, type, price, city
FROM emp_bucket;



```
hive (hive_bucket)> INSERT INTO TABLE emp_bucket_city
                  > PARTITION (city)
                  > SELECT street, zip, state, beds, baths, sq_fit, type, price, city
                  > FROM emp_bucket;
Query ID = cloudera_20240618052121_856a29cd-4b59-4e53-9791-728f31635dca
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_1718702975970_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718702975970_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718702975970_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-06-18 05:21:57,924 Stage-1 map = 0%,  reduce = 0%
2024-06-18 05:22:05,905 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.03 sec
MapReduce Total cumulative CPU time: 3 seconds 30 msec
Ended Job = job_1718702975970_0001
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: hdfs://quickstart.cloudera:8020/user/hive/warehouse/hive_bucket.db/emp_bucket_city/.hive-staging_hive_2024-06
-18_05-21-42_832_4673701355311792081-1/-ext-10000
Loading data to table hive_bucket.emp_bucket_city partition (city=null)
         Time taken for load dynamic partitions : 672
        Loading partition {city=SACREMENTO}
        Loading partition {city=LA}
        Loading partition {city=CA}
        Loading partition {city=DUBLIN}
         Time taken for adding to write entity : 2
Partition hive_bucket.emp_bucket_city{city=CA} stats: [numFiles=1, numRows=1, totalSize=49, rawDataSize=48]
Partition hive_bucket.emp_bucket_city{city=DUBLIN} stats: [numFiles=1, numRows=2, totalSize=96, rawDataSize=94]
Partition hive_bucket.emp_bucket_city{city=LA} stats: [numFiles=1, numRows=1, totalSize=46, rawDataSize=45]
Partition hive_bucket.emp_bucket_city{city=SACREMENTO} stats: [numFiles=1, numRows=1, totalSize=48, rawDataSize=47]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.03 sec   HDFS Read: 5389 HDFS Write: 504 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 30 msec
OK
Time taken: 26.717 seconds
```



```
hive (hive_bucket)> select * from emp_bucket_city;
OK
456 KALA ST     67890   CALIF   2       1       678     INDUSTRIAL      40000   CA
2 ABBEY ST      98678   IRE     3       2       898     RESIDENTIAL     98000   DUBLIN
2 CORNELL ST    78907   IRE     2       1       789     RESIDENTIAL     87907   DUBLIN
45 TI_LST       97654   LAUS    1       2       798     INDUSTRIAL      49876   LA
3526 HIGH ST    95838   CA      2       1       796     RESIDENTIAL     59222   SACREMENTO
Time taken: 0.215 seconds, Fetched: 5 row(s)
```



```
[cloudera@quickstart Hive]$ hdfs dfs -ls /user/hive/warehouse/hive_bucket.db/
Found 2 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 02:54 /user/hive/warehouse/hive_bucket.db/emp_bucket
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city
[cloudera@quickstart Hive]$ hdfs dfs -ls /user/hive/warehouse/hive_bucket.db/emp_bucket_city/
Found 4 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=CA
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLI
N
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=LA
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=SACRE
MENTO
[cloudera@quickstart Hive]$ hdfs dfs -ls /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLIN
Found 1 items
-rwxrwxrwx   1 cloudera supergroup         96 2024-06-18 05:22 /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLI
N/000000_0
[cloudera@quickstart Hive]$ hdfs dfs -cat /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLIN
cat: `/user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLIN': Is a directory
[cloudera@quickstart Hive]$ hdfs dfs -cat /user/hive/warehouse/hive_bucket.db/emp_bucket_city/city=DUBLIN/000000_0
2 ABBEY ST,98678,IRE,3,2,898,RESIDENTIAL,98000
2 CORNELL ST,78907,IRE,2,1,789,RESIDENTIAL,87907
[cloudera@quickstart Hive]$ 
```

—---

```
CREATE TABLE emp_bucket_state (
street string,
city string,
zip int,
beds int,
baths int,
sq_fit int,
type string,
price int
)
PARTITIONED BY (state STRING)
CLUSTERED BY (city) into 3 buckets
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

—---

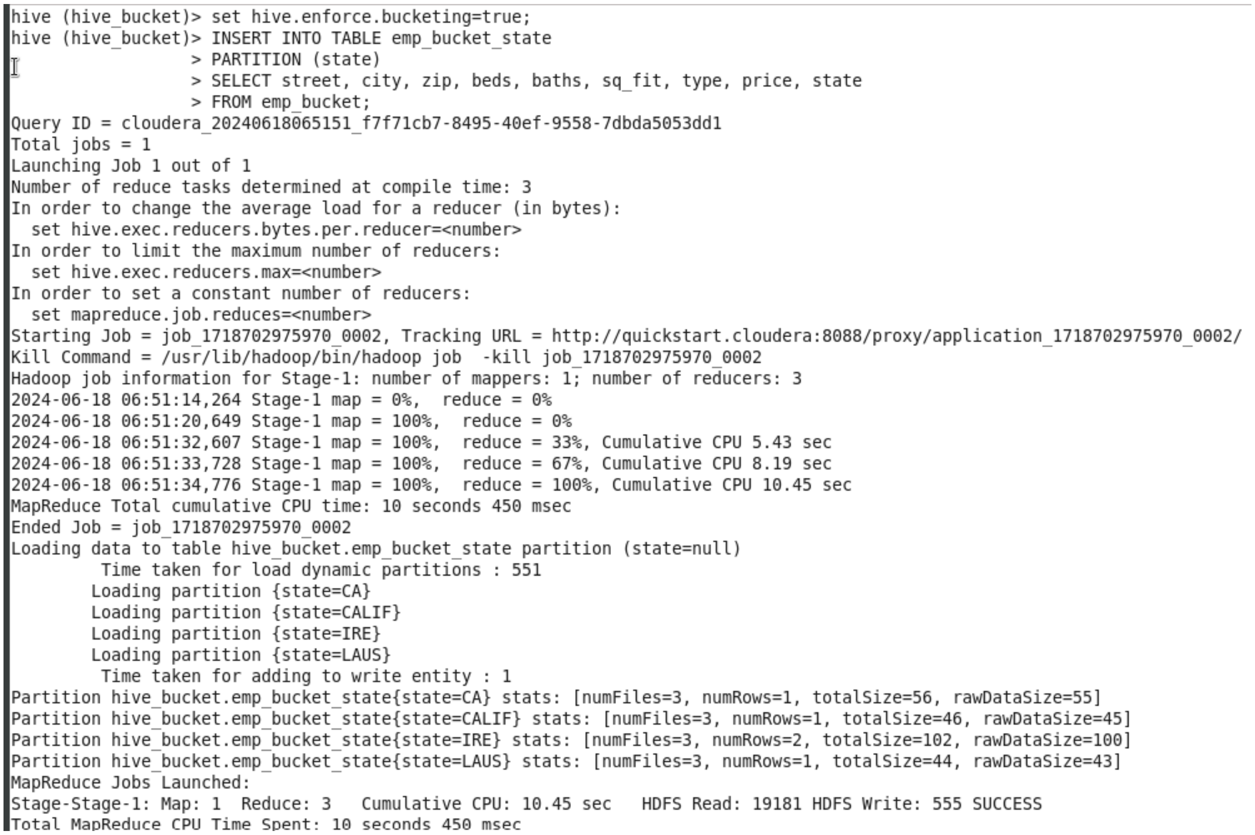
To enforce the Bucketing, we will have to set:

```
set hive.enforce.bucketing=true;
```

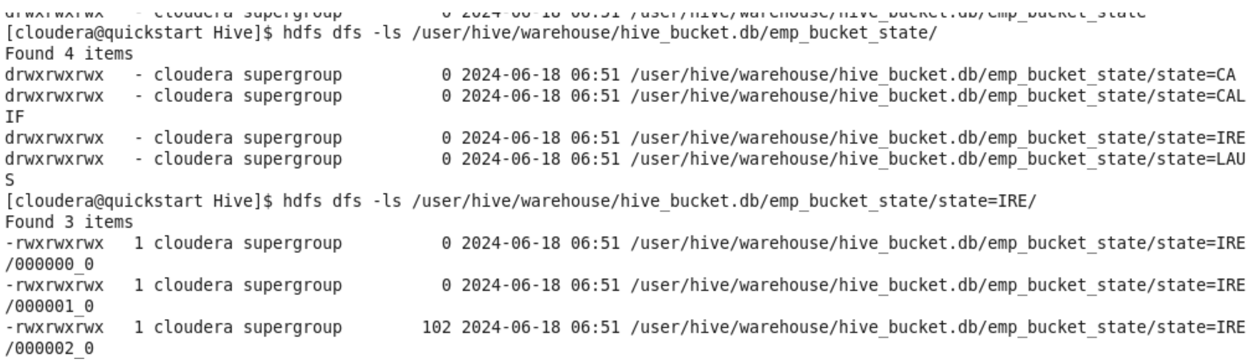
—---

Loading the values into table:

```
INSERT INTO TABLE emp_bucket_state
PARTITION (state)
SELECT street, city, zip, beds, baths, sq_fit, type, price, state
FROM emp_bucket;
```



```
hive (hive_bucket)> set hive.enforce.bucketing=true;
hive (hive_bucket)> INSERT INTO TABLE emp_bucket_state
                  > PARTITION (state)
                  > SELECT street, city, zip, beds, baths, sq_fit, type, price, state
                  > FROM emp_bucket;
Query ID = cloudera_20240618065151_f7f71cb7-8495-40ef-9558-7dbda5053dd1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 3
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_1718702975970_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718702975970_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718702975970_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 3
2024-06-18 06:51:14,264 Stage-1 map = 0%,  reduce = 0%
2024-06-18 06:51:20,649 Stage-1 map = 100%,  reduce = 0%
2024-06-18 06:51:32,607 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 5.43 sec
2024-06-18 06:51:33,728 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 8.19 sec
2024-06-18 06:51:34,776 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.45 sec
MapReduce Total cumulative CPU time: 10 seconds 450 msec
Ended Job = job_1718702975970_0002
Loading data to table hive_bucket.emp_bucket_state partition (state=null)
         Time taken for load dynamic partitions : 551
        Loading partition {state=CA}
        Loading partition {state=CALIF}
        Loading partition {state=IRE}
        Loading partition {state=LAUS}
         Time taken for adding to write entity : 1
Partition hive_bucket.emp_bucket_state{state=CA} stats: [numFiles=3, numRows=1, totalSize=56, rawDataSize=55]
Partition hive_bucket.emp_bucket_state{state=CALIF} stats: [numFiles=3, numRows=1, totalSize=46, rawDataSize=45]
Partition hive_bucket.emp_bucket_state{state=IRE} stats: [numFiles=3, numRows=2, totalSize=102, rawDataSize=100]
Partition hive_bucket.emp_bucket_state{state=LAUS} stats: [numFiles=3, numRows=1, totalSize=44, rawDataSize=43]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 3   Cumulative CPU: 10.45 sec   HDFS Read: 19181 HDFS Write: 555 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 450 msec
```



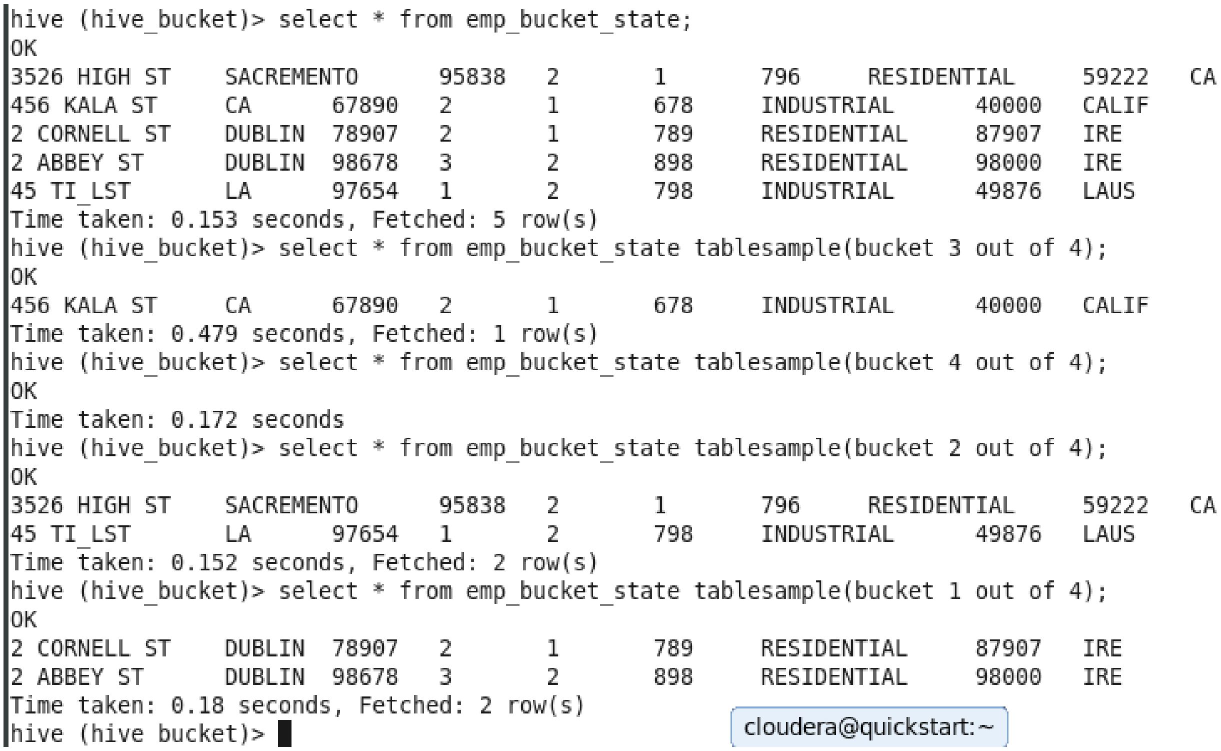
```
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state
[cloudera@quickstart Hive]$ hdfs dfs -ls /user/hive/warehouse/hive_bucket.db/emp_bucket_state/
Found 4 items
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=CA
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=CAL
IF
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=IRE
drwxrwxrwx   - cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=LAU
S
[cloudera@quickstart Hive]$ hdfs dfs -ls /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=IRE/
Found 3 items
-rwxrwxrwx   1 cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=IRE
/000000_0
-rwxrwxrwx   1 cloudera supergroup          0 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=IRE
/000001_0
-rwxrwxrwx   1 cloudera supergroup        102 2024-06-18 06:51 /user/hive/warehouse/hive_bucket.db/emp_bucket_state/state=IRE
/000002_0
```

****

BUCKET TABLE SAMPLING

****

select * from emp_bucket_state tablesample(bucket 2 out of 3)



```
hive (hive_bucket)> select * from emp_bucket_state;
OK
3526 HIGH ST    SACREMENTO      95838   2       1       796     RESIDENTIAL     59222   CA
456 KALA ST     CA      67890   2       1       678     INDUSTRIAL      40000   CALIF
2 CORNELL ST    DUBLIN  78907   2       1       789     RESIDENTIAL     87907   IRE
2 ABBEY ST      DUBLIN  98678   3       2       898     RESIDENTIAL     98000   IRE
45 TI_LST       LA      97654   1       2       798     INDUSTRIAL      49876   LAUS
Time taken: 0.153 seconds, Fetched: 5 row(s)
hive (hive_bucket)> select * from emp_bucket_state tablesample(bucket 3 out of 4);
OK
456 KALA ST     CA      67890   2       1       678     INDUSTRIAL      40000   CALIF
Time taken: 0.479 seconds, Fetched: 1 row(s)
hive (hive_bucket)> select * from emp_bucket_state tablesample(bucket 4 out of 4);
OK
Time taken: 0.172 seconds
hive (hive_bucket)> select * from emp_bucket_state tablesample(bucket 2 out of 4);
OK
3526 HIGH ST    SACREMENTO      95838   2       1       796     RESIDENTIAL     59222   CA
45 TI_LST       LA      97654   1       2       798     INDUSTRIAL      49876   LAUS
Time taken: 0.152 seconds, Fetched: 2 row(s)
hive (hive_bucket)> select * from emp_bucket_state tablesample(bucket 1 out of 4);
OK
2 CORNELL ST    DUBLIN  78907   2       1       789     RESIDENTIAL     87907   IRE
2 ABBEY ST      DUBLIN  98678   3       2       898     RESIDENTIAL     98000   IRE
Time taken: 0.18 seconds, Fetched: 2 row(s)
hive (hive bucket)> 
```

—--

****

JOINS IN HIVE

****

Data:

CUSTOMERS.TXT:

Id,Name,Age,Address,Salary
1,Ross,32,Ahmedabad,2000
2,Rachel,25,Delhi,1500
3,chandler,23,Kota,2000
4,Monika,25,Mumbai,6500
5,Mike,27,Bhopal,8500
6,Phoebe,22,MP,4500
7,Joey,24,Indore,10000

ORDERS.TXT:

```
OID,Date,Customer_ID,Amount
102,2016-10-08 00:00:00,3,3000
100,2016-10-08 00:00:00,3,1500
101,2016-11-20 00:00:00,2,1560
103,2015-05-20 00:00:00,4,2060
```

ORDER_ITEMS.TXT:

```
oid,ord,date,items,amount
102,2016-10-08 00:00:00,Pizza,3000
102,2016-10-08 00:00:00,Juice,3000
100,2016-10-08 00:00:00,Biryani,3000
101,2016-11-20 00:00:00,Paneer,3000
103,2015-05-20 00:00:00,Momos,3000
```

Create Table:

```
CREATE TABLE customers (
ID INT,
NAME STRING,
AGE INT,
ADDRESS STRING,
SALARY INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/onlineshop/customers.txt' INTO
TABLE customers;

CREATE TABLE orders(
OID INT,
DATE STRING,
CUSTOMER_ID INT,
AMOUNT INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/onlineshop/orders.txt' INTO TABLE
orders;
```

```
CREATE TABLE order_items (
OID INT,
ORD_DATE STRING,
ITEMS STRING,
AMOUNT INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/onlineshop/order_items.txt' INTO
TABLE order_items;
```

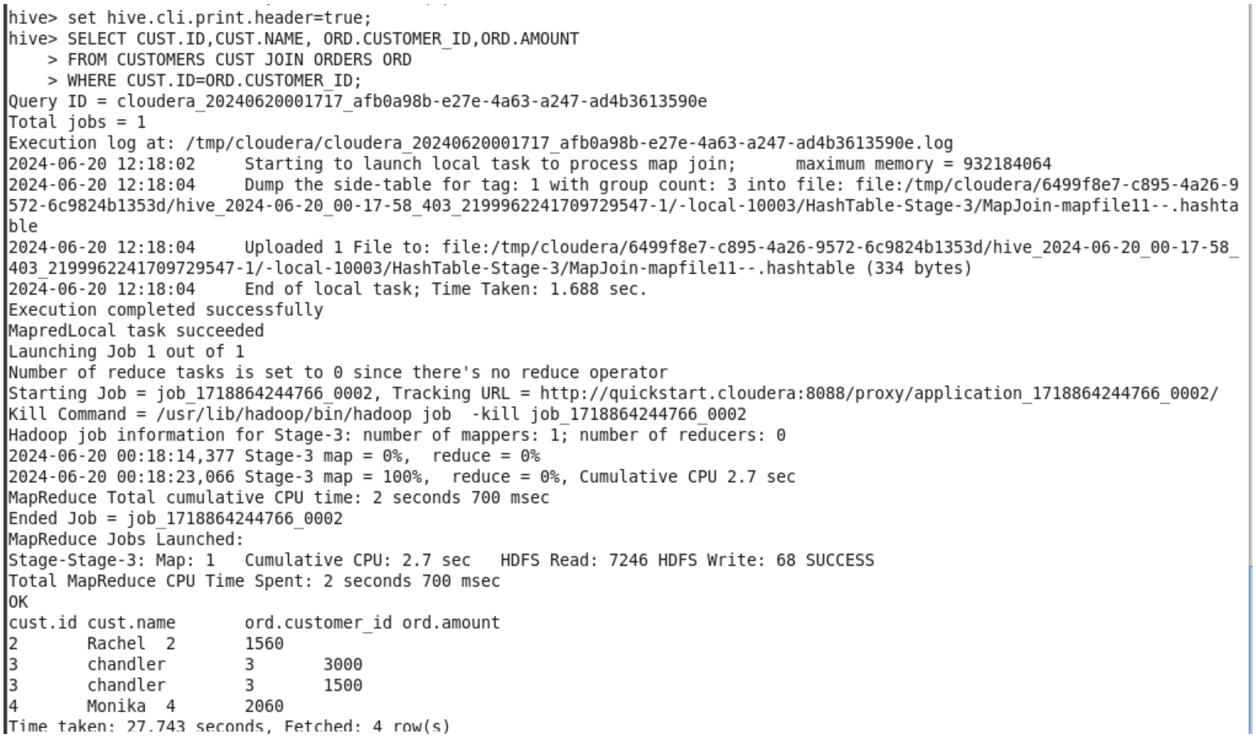
—-----------------------------------

NORMAL JOIN (ANCI 89):

```
SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
FROM CUSTOMERS CUST, ORDERS ORD
WHERE CUST.ID=ORD.CUSTOMER_ID;
```

NORMAL JOIN (ANCI 92):

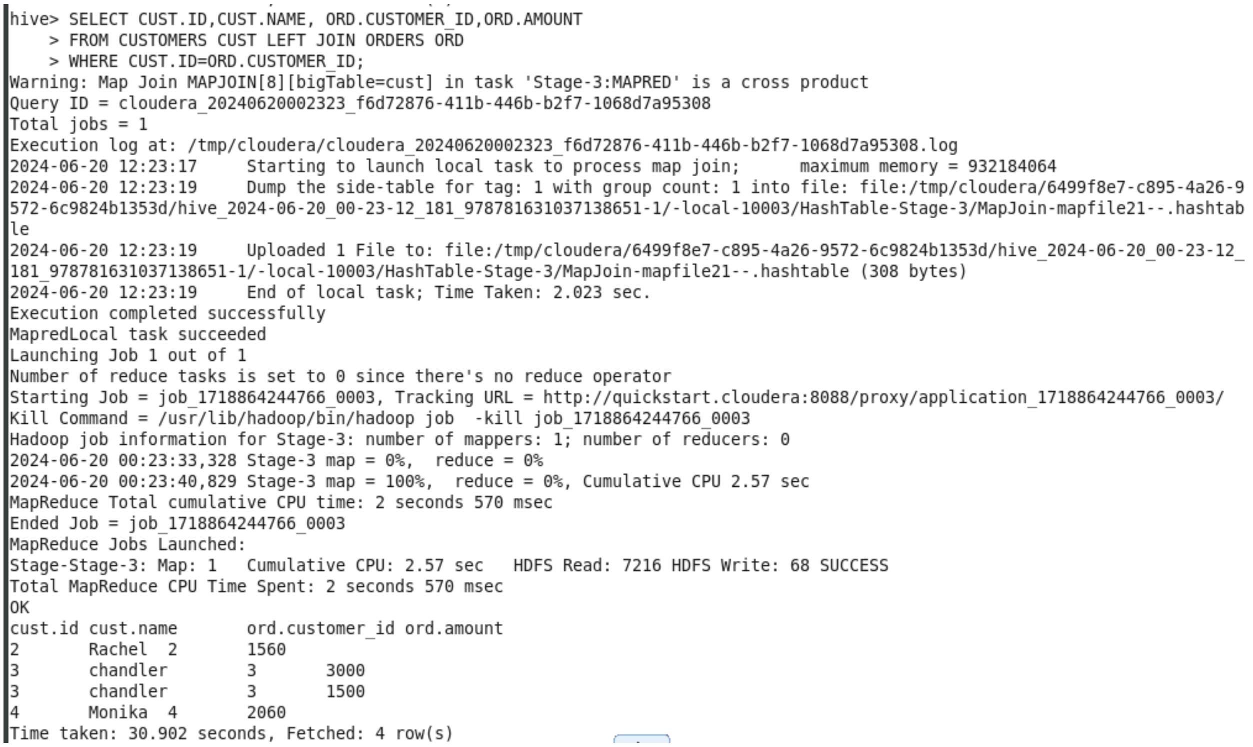
```
SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
FROM CUSTOMERS CUST JOIN ORDERS ORD
WHERE CUST.ID=ORD.CUSTOMER_ID;
```



```
hive> set hive.cli.print.header=true;
hive> SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
    > FROM CUSTOMERS CUST JOIN ORDERS ORD
    > WHERE CUST.ID=ORD.CUSTOMER_ID;
Query ID = cloudera_20240620001717_afb0a98b-e27e-4a63-a247-ad4b3613590e
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240620001717_afb0a98b-e27e-4a63-a247-ad4b3613590e.log
2024-06-20 12:18:02     Starting to launch local task to process map join;      maximum memory = 932184064
2024-06-20 12:18:04     Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/cloudera/6499f8e7-c895-4a26-9
572-6c9824b1353d/hive_2024-06-20_00-17-58_403_2199962241709729547-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile11--.hashta
ble
2024-06-20 12:18:04     Uploaded 1 File to: file:/tmp/cloudera/6499f8e7-c895-4a26-9572-6c9824b1353d/hive_2024-06-20_00-17-58_
403_2199962241709729547-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile11--.hashtable (334 bytes)
2024-06-20 12:18:04     End of local task; Time Taken: 1.688 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_1718864244766_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718864244766_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718864244766_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-06-20 00:18:14,377 Stage-3 map = 0%,  reduce = 0%
2024-06-20 00:18:23,066 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.7 sec
MapReduce Total cumulative CPU time: 2 seconds 700 msec
Ended Job = job_1718864244766_0002
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.7 sec   HDFS Read: 7246 HDFS Write: 68 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 700 msec
OK
cust.id cust.name       ord.customer_id ord.amount
2       Rachel  2       1560
3       chandler        3       3000
3       chandler        3       1500
4       Monika  4       2060
Time taken: 27.743 seconds, Fetched: 4 row(s)
```

—---

SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
FROM CUSTOMERS CUST LEFT JOIN ORDERS ORD
WHERE CUST.ID=ORD.CUSTOMER_ID;



```
hive> SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
    > FROM CUSTOMERS CUST LEFT JOIN ORDERS ORD
    > WHERE CUST.ID=ORD.CUSTOMER_ID;
Warning: Map Join MAPJOIN[8][bigTable=cust] in task 'Stage-3:MAPRED' is a cross product
Query ID = cloudera_20240620002323_f6d72876-411b-446b-b2f7-1068d7a95308
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240620002323_f6d72876-411b-446b-b2f7-1068d7a95308.log
2024-06-20 12:23:17     Starting to launch local task to process map join;      maximum memory = 932184064
2024-06-20 12:23:19     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/6499f8e7-c895-4a26-9
572-6c9824b1353d/hive_2024-06-20_00-23-12_181_978781631037138651-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashtab
le
2024-06-20 12:23:19     Uploaded 1 File to: file:/tmp/cloudera/6499f8e7-c895-4a26-9572-6c9824b1353d/hive_2024-06-20_00-23-12_
181_978781631037138651-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile21--.hashtable (308 bytes)
2024-06-20 12:23:19     End of local task; Time Taken: 2.023 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_1718864244766_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718864244766_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718864244766_0003
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-06-20 00:23:33,328 Stage-3 map = 0%,  reduce = 0%
2024-06-20 00:23:40,829 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.57 sec
MapReduce Total cumulative CPU time: 2 seconds 570 msec
Ended Job = job_1718864244766_0003
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.57 sec   HDFS Read: 7216 HDFS Write: 68 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 570 msec
OK
cust.id cust.name       ord.customer_id ord.amount
2       Rachel  2       1560
3       chandler        3       3000
3       chandler        3       1500
4       Monika  4       2060
Time taken: 30.902 seconds, Fetched: 4 row(s)
```

```
SELECT CUST.ID,CUST.NAME, ORD.CUSTOMER_ID,ORD.AMOUNT
FROM CUSTOMERS CUST RIGHT JOIN ORDERS ORD
WHERE CUST.ID=ORD.CUSTOMER_ID;
```

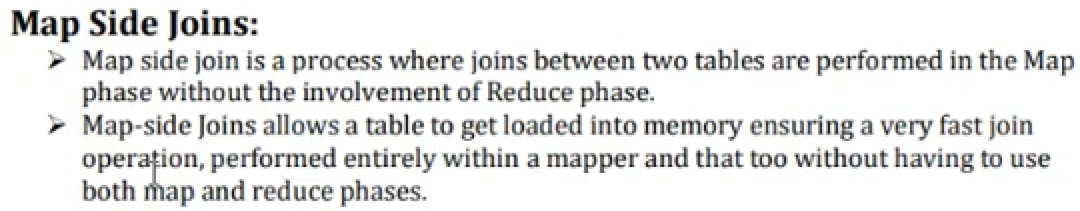
—---

Hive in its own Provides 3 types of joins:

1. Map Side Joins
2. Bucket Joins
3. SORT MERGE BUCKET (SMB) MAP JOIN

—---

++++++++++++++++++++++++++++++
MAP SIDE JOINS
++++++++++++++++++++++++++++++



Map Side Joins:

- ➢ Map side join is a process where joins between two tables are performed in the Map phase without the involvement of Reduce phase.
- ➢ Map-side Joins allows a table to get loaded into memory ensuring a very fast join operation, performed entirely within a mapper and that too without having to use both map and reduce phases.

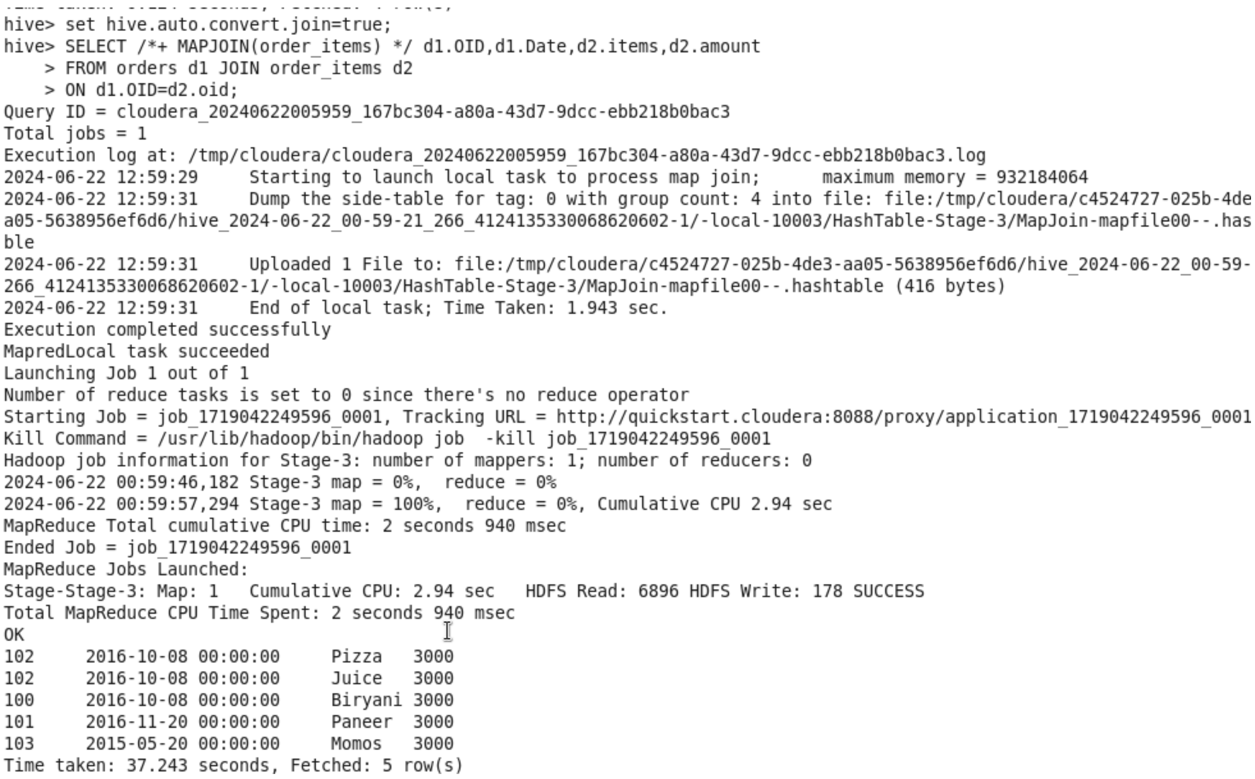
Set the property:

```
set hive.auto.convert.join=true;
```

—--

```
SELECT /*+ MAPJOIN(order_items) */ d1.OID,d1.Date,d2.items,d2.amount
FROM orders d1 JOIN order_items d2
ON d1.OID=d2.oid;
```

Note:
The table which contains less data will be the part of MAPJOIN clause.
Number of reducers are set to 0.



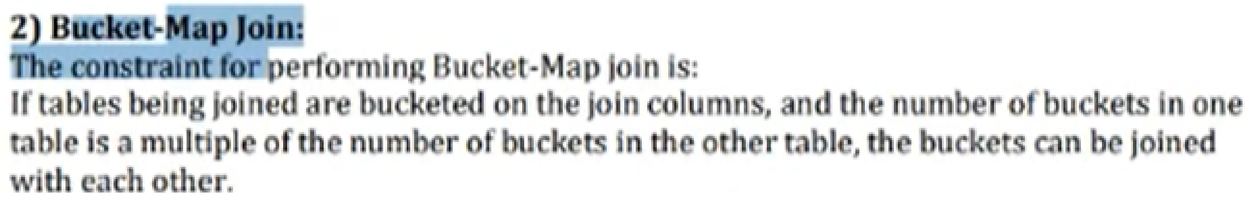
```
hive> set hive.auto.convert.join=true;
hive> SELECT /*+ MAPJOIN(order_items) */ d1.OID,d1.Date,d2.items,d2.amount
    > FROM orders d1 JOIN order_items d2
    > ON d1.OID=d2.oid;
Query ID = cloudera_20240622005959_167bc304-a80a-43d7-9dcc-ebb218b0bac3
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240622005959_167bc304-a80a-43d7-9dcc-ebb218b0bac3.log
2024-06-22 12:59:29     Starting to launch local task to process map join;      maximum memory = 932184064
2024-06-22 12:59:31     Dump the side-table for tag: 0 with group count: 4 into file: file:/tmp/cloudera/c4524727-025b-4de
a05-5638956ef6d6/hive_2024-06-22_00-59-21_266_4124135330068620602-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.has
ble
2024-06-22 12:59:31     Uploaded 1 File to: file:/tmp/cloudera/c4524727-025b-4de3-aa05-5638956ef6d6/hive_2024-06-22_00-59-
266_4124135330068620602-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.hashtable (416 bytes)
2024-06-22 12:59:31     End of local task; Time Taken: 1.943 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_1719042249596_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1719042249596_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1719042249596_0001
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-06-22 00:59:46,182 Stage-3 map = 0%,  reduce = 0%
2024-06-22 00:59:57,294 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.94 sec
MapReduce Total cumulative CPU time: 2 seconds 940 msec
Ended Job = job_1719042249596_0001
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.94 sec   HDFS Read: 6896 HDFS Write: 178 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 940 msec
OK
102     2016-10-08 00:00:00     Pizza   3000
102     2016-10-08 00:00:00     Juice   3000
100     2016-10-08 00:00:00     Biryani 3000
101     2016-11-20 00:00:00     Paneer  3000
103     2015-05-20 00:00:00     Momos   3000
Time taken: 37.243 seconds, Fetched: 5 row(s)
```

—---

++++++++++++++++++++++++++++++++

BUCKET JOINS

++++++++++++++++++++++++++++++++



2) Bucket-Map Join:

The constraint for performing Bucket-Map join is:

If tables being joined are bucketed on the join columns, and the number of buckets in one table is a multiple of the number of buckets in the other table, the buckets can be joined with each other.

For this we need to set the property:

set hive.optimize.bucketmapjoin = true;

—---

CREATE TABLE orders_bucket (
OID INT,
DATE STRING,
CUSTOMER_ID INT,
AMOUNT INT
)
CLUSTERED BY (DATE) INTO 3 buckets
ROW FORMAT DELIMITED

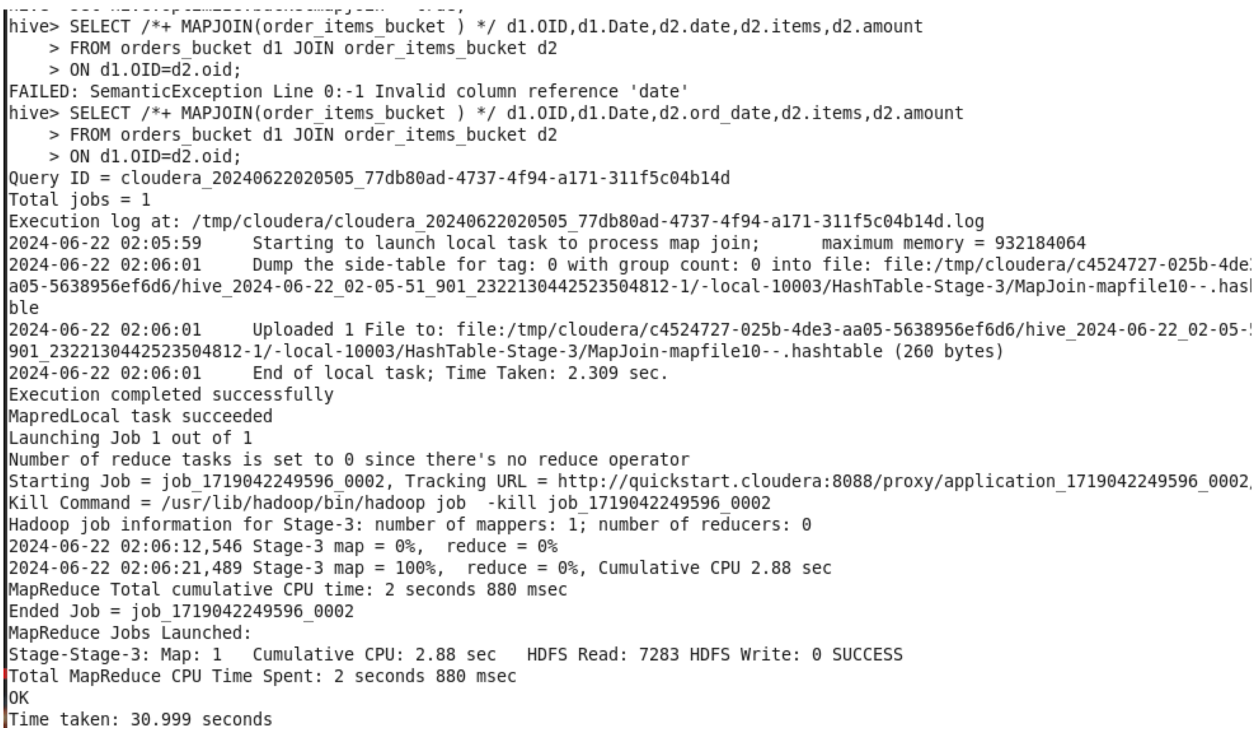
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

CREATE TABLE order_items_bucket (
OID INT,
ORD_DATE STRING,
ITEMS STRING,
AMOUNT INT
)
CLUSTERED BY (ORD_DATE) INTO 3 buckets
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

—---

To Join:

SELECT /*+ MAPJOIN(order_items_bucket) */
d1.OID,d1.Date,d2.ord_date,d2.items,d2.amount
FROM orders_bucket d1 JOIN order_items_bucket d2
ON d1.OID=d2.oid;



```
hive> SELECT /*+ MAPJOIN(order_items_bucket ) */ d1.OID,d1.Date,d2.date,d2.items,d2.amount
    > FROM orders_bucket d1 JOIN order_items_bucket d2
    > ON d1.OID=d2.oid;
FAILED: SemanticException Line 0:-1 Invalid column reference 'date'
hive> SELECT /*+ MAPJOIN(order_items_bucket ) */ d1.OID,d1.Date,d2.ord_date,d2.items,d2.amount
    > FROM orders_bucket d1 JOIN order_items_bucket d2
    > ON d1.OID=d2.oid;
Query ID = cloudera_20240622020505_77db80ad-4737-4f94-a171-311f5c04b14d
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20240622020505_77db80ad-4737-4f94-a171-311f5c04b14d.log
2024-06-22 02:05:59     Starting to launch local task to process map join;      maximum memory = 932184064
2024-06-22 02:06:01     Dump the side-table for tag: 0 with group count: 0 into file: file:/tmp/cloudera/c4524727-025b-4de
a05-5638956ef6d6/hive_2024-06-22_02-05-51_901_2322130442523504812-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.has
ble
2024-06-22 02:06:01     Uploaded 1 File to: file:/tmp/cloudera/c4524727-025b-4de3-aa05-5638956ef6d6/hive_2024-06-22_02-05-
901_2322130442523504812-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable (260 bytes)
2024-06-22 02:06:01     End of local task; Time Taken: 2.309 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_1719042249596_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1719042249596_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1719042249596_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2024-06-22 02:06:12,546 Stage-3 map = 0%,  reduce = 0%
2024-06-22 02:06:21,489 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.88 sec
MapReduce Total cumulative CPU time: 2 seconds 880 msec
Ended Job = job_1719042249596_0002
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.88 sec   HDFS Read: 7283 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 880 msec
OK
Time taken: 30.999 seconds
```

—---

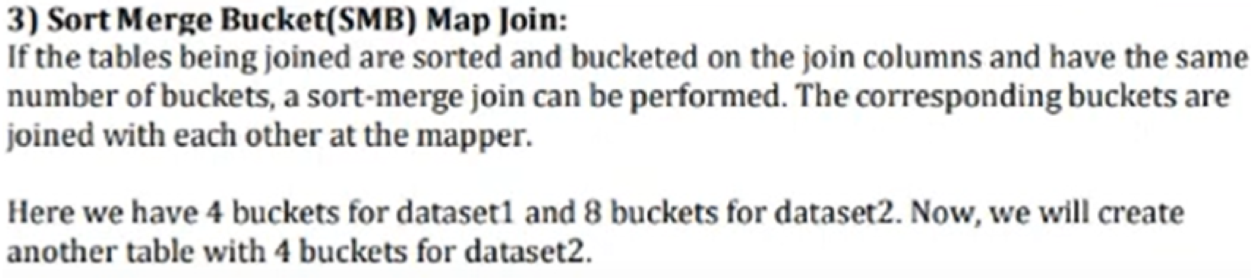
Load the data:

```
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/onlineshop/orders.txt' INTO TABLE orders_bucket;
```

```
LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/onlineshop/order_items.txt' INTO TABLE order_items_bucket;
```

—---

++++++++++++++++++++++++++++++++++++
SORT MERGE BUCKET (SMB) MAP JOIN
++++++++++++++++++++++++++++++++++++



3) Sort Merge Bucket(SMB) Map Join:
If the tables being joined are sorted and bucketed on the join columns and have the same number of buckets, a sort-merge join can be performed. The corresponding buckets are joined with each other at the mapper.

Here we have 4 buckets for dataset1 and 8 buckets for dataset2. Now, we will create another table with 4 buckets for dataset2.

—---

Set Properties:

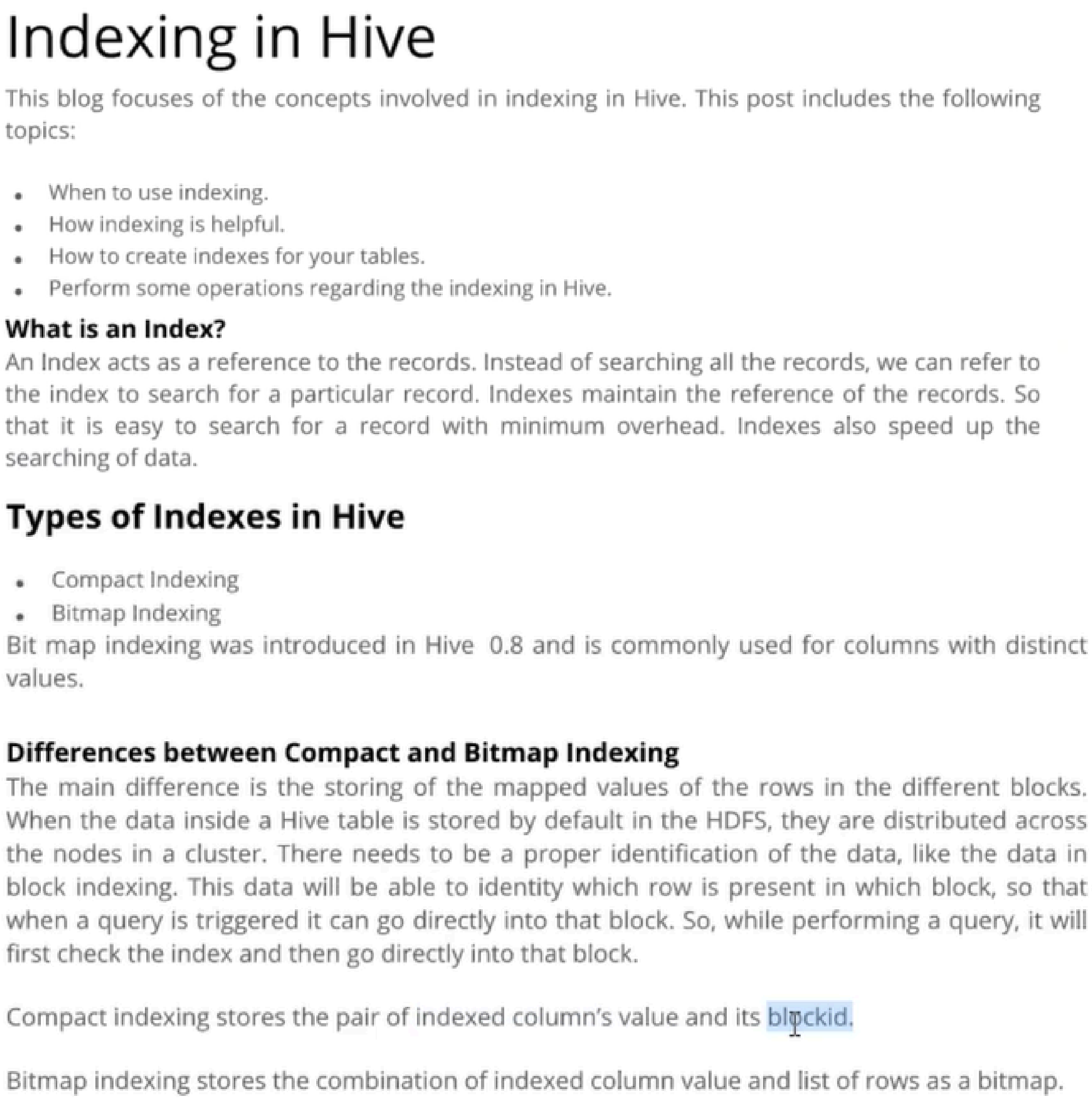
```
set hive.input.format=org.apache.hadoop.hive.ql.io.BucketizedHiveInputFormat;

set hive.optimize.bucketmapjoin=true;

set hive.optimize.bucketmapjoin.sortedmerge=true;
```

—---

****

HIVE INDEXING

****



Indexing in Hive

This blog focuses of the concepts involved in indexing in Hive. This post includes the following topics:

- When to use indexing.
- How indexing is helpful.
- How to create indexes for your tables.
- Perform some operations regarding the indexing in Hive.

What is an Index?

An Index acts as a reference to the records. Instead of searching all the records, we can refer to the index to search for a particular record. Indexes maintain the reference of the records. So that it is easy to search for a record with minimum overhead. Indexes also speed up the searching of data.

Types of Indexes in Hive

- Compact Indexing
- Bitmap Indexing

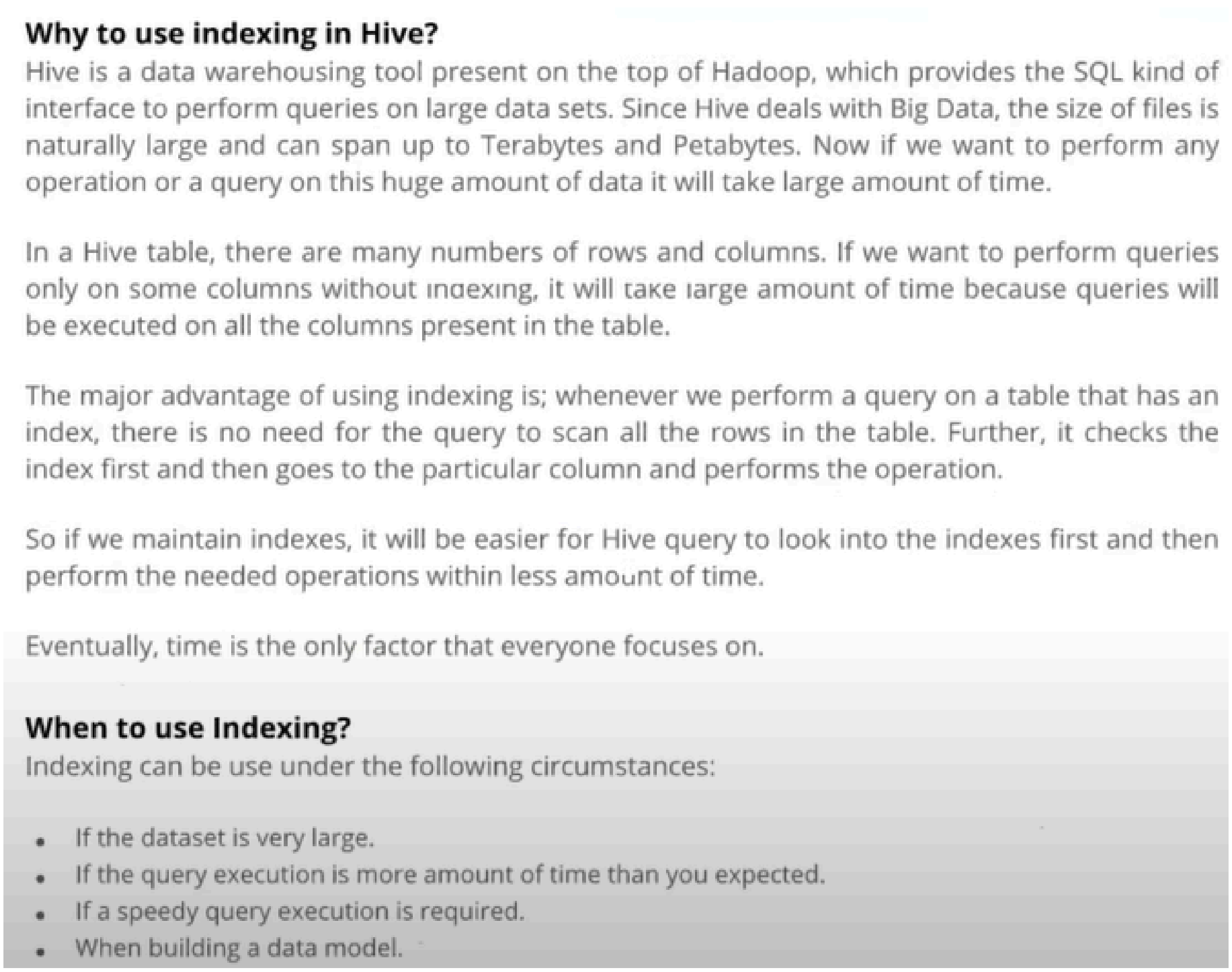
Bit map indexing was introduced in Hive 0.8 and is commonly used for columns with distinct values.

Differences between Compact and Bitmap Indexing

The main difference is the storing of the mapped values of the rows in the different blocks. When the data inside a Hive table is stored by default in the HDFS, they are distributed across the nodes in a cluster. There needs to be a proper identification of the data, like the data in block indexing. This data will be able to identity which row is present in which block, so that when a query is triggered it can go directly into that block. So, while performing a query, it will first check the index and then go directly into that block.

Compact indexing stores the pair of indexed column's value and its blockid.

Bitmap indexing stores the combination of indexed column value and list of rows as a bitmap.



Why to use indexing in Hive?

Hive is a data warehousing tool present on the top of Hadoop, which provides the SQL kind of interface to perform queries on large data sets. Since Hive deals with Big Data, the size of files is naturally large and can span up to Terabytes and Petabytes. Now if we want to perform any operation or a query on this huge amount of data it will take large amount of time.

In a Hive table, there are many numbers of rows and columns. If we want to perform queries only on some columns without indexing, it will take large amount of time because queries will be executed on all the columns present in the table.

The major advantage of using indexing is; whenever we perform a query on a table that has an index, there is no need for the query to scan all the rows in the table. Further, it checks the index first and then goes to the particular column and performs the operation.

So if we maintain indexes, it will be easier for Hive query to look into the indexes first and then perform the needed operations within less amount of time.

Eventually, time is the only factor that everyone focuses on.

When to use Indexing?

Indexing can be use under the following circumstances:

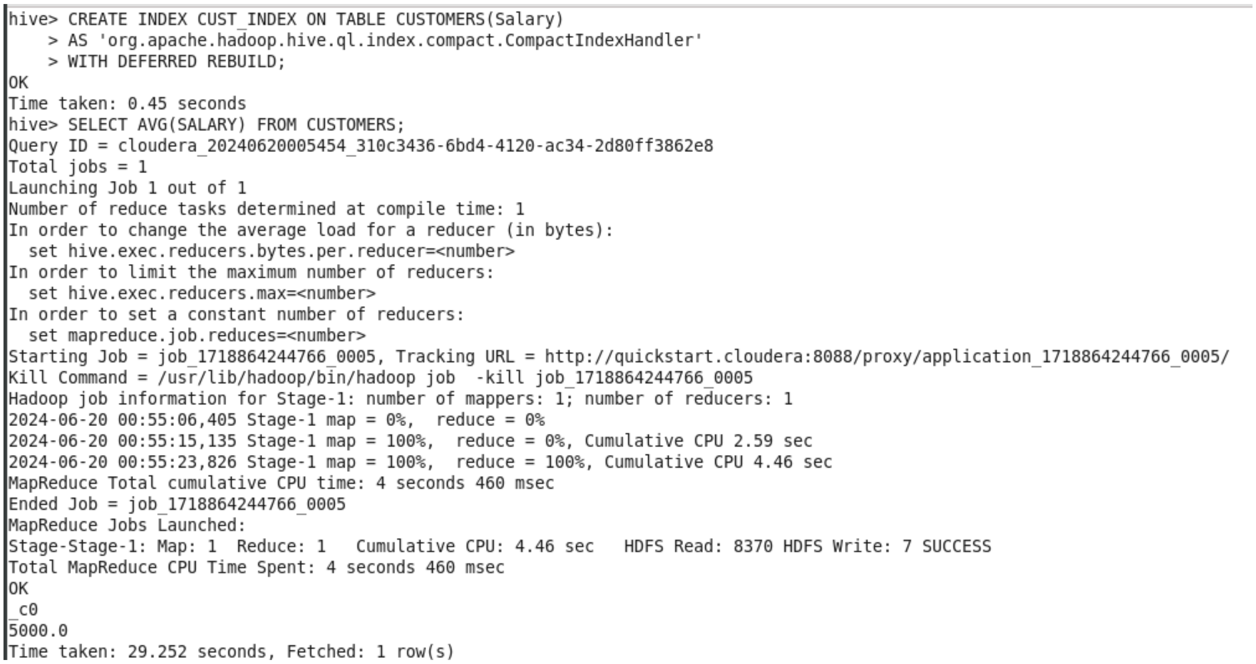
- If the dataset is very large.
- If the query execution is more amount of time than you expected.
- If a speedy query execution is required.
- When building a data model.

2 TYPES OF INDEXING:

- Compact
- bitmap

—---

```
CREATE INDEX CUST_INDEX ON TABLE CUSTOMERS(Salary)
AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
WITH DEFERRED REBUILD;

SELECT AVG(SALARY) FROM CUSTOMERS;
```

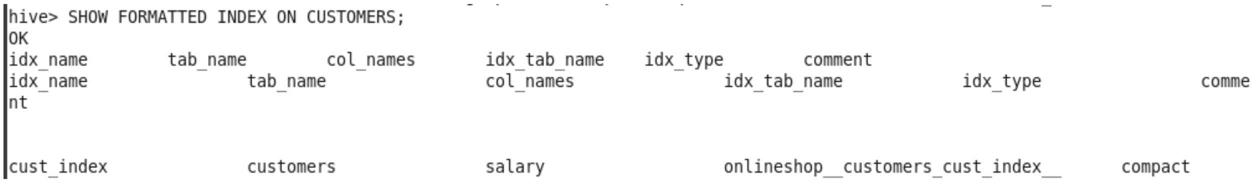


```
hive> CREATE INDEX CUST_INDEX ON TABLE CUSTOMERS(Salary)
    > AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
    > WITH DEFERRED REBUILD;
OK
Time taken: 0.45 seconds
hive> SELECT AVG(SALARY) FROM CUSTOMERS;
Query ID = cloudera_20240620005454_310c3436-6bd4-4120-ac34-2d80ff3862e8
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_1718864244766_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718864244766_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718864244766_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-06-20 00:55:06,405 Stage-1 map = 0%,  reduce = 0%
2024-06-20 00:55:15,135 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.59 sec
2024-06-20 00:55:23,826 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.46 sec
MapReduce Total cumulative CPU time: 4 seconds 460 msec
Ended Job = job_1718864244766_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.46 sec   HDFS Read: 8370 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 460 msec
OK
_c0
5000.0
Time taken: 29.252 seconds, Fetched: 1 row(s)
```

—---

To show the index on the original table:

SHOW FORMATTED INDEX ON CUSTOMERS;

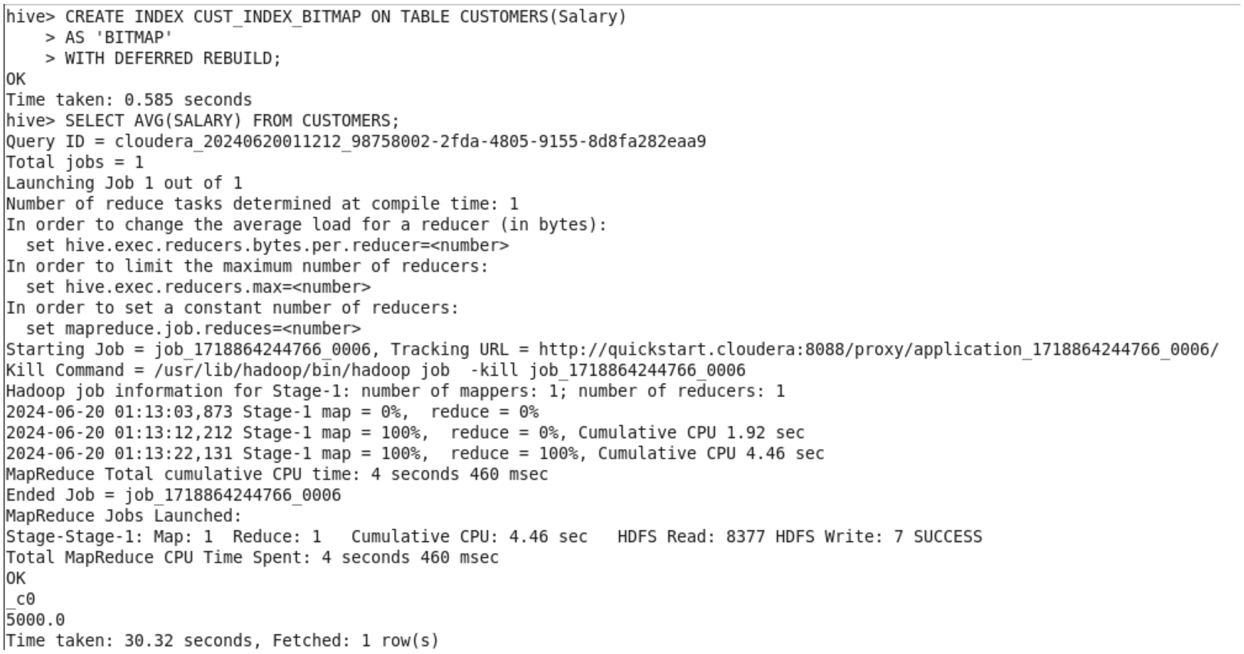


```
hive> SHOW FORMATTED INDEX ON CUSTOMERS;
OK
idx_name        tab_name        col_names       idx_tab_name    idx_type        comment
idx_name                tab_name                col_names               idx_tab_name            idx_type                comme
nt


cust_index              customers               salary                  onlineshop__customers_cust_index__      compact
```

—---

Bitmap Indexing:

CREATE INDEX CUST_INDEX_BITMAP ON TABLE CUSTOMERS(Salary)
AS 'BITMAP'
WITH DEFERRED REBUILD;



```
hive> CREATE INDEX CUST_INDEX_BITMAP ON TABLE CUSTOMERS(Salary)
    > AS 'BITMAP'
    > WITH DEFERRED REBUILD;
OK
Time taken: 0.585 seconds
hive> SELECT AVG(SALARY) FROM CUSTOMERS;
Query ID = cloudera_20240620011212_98758002-2fda-4805-9155-8d8fa282eaa9
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_1718864244766_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718864244766_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718864244766_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-06-20 01:13:03,873 Stage-1 map = 0%,  reduce = 0%
2024-06-20 01:13:12,212 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.92 sec
2024-06-20 01:13:22,131 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.46 sec
MapReduce Total cumulative CPU time: 4 seconds 460 msec
Ended Job = job_1718864244766_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.46 sec   HDFS Read: 8377 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 460 msec
OK
_c0
5000.0
Time taken: 30.32 seconds, Fetched: 1 row(s)
```

—---

Drop Index Tables:

DROP INDEX IF EXISTS CUST_INDEX ON CUSTOMERS;
DROP INDEX IF EXISTS CUST_INDEX_BITMAP ON CUSTOMERS;

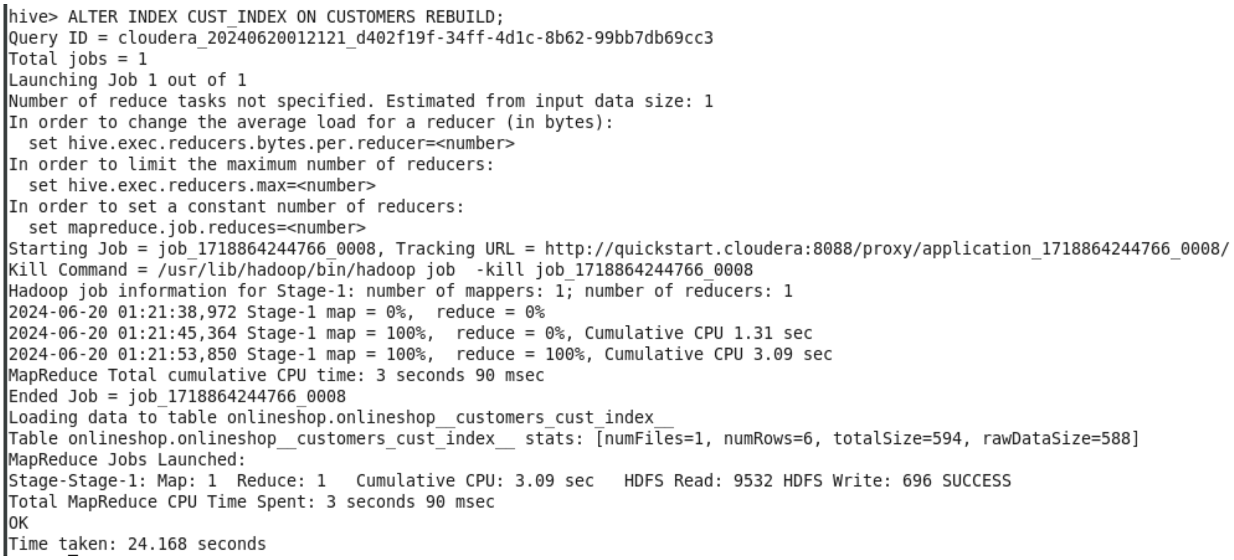
—---

Again create the one of the same index table:

SELECT AVG(SALARY) FROM CUSTOMERS;

Alter Index tables:

ALTER INDEX CUST_INDEX ON CUSTOMERS REBUILD;



```
hive> ALTER INDEX CUST_INDEX ON CUSTOMERS REBUILD;
Query ID = cloudera_20240620012121_d402f19f-34ff-4d1c-8b62-99bb7db69cc3
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_1718864244766_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1718864244766_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1718864244766_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-06-20 01:21:38,972 Stage-1 map = 0%,  reduce = 0%
2024-06-20 01:21:45,364 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.31 sec
2024-06-20 01:21:53,850 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.09 sec
MapReduce Total cumulative CPU time: 3 seconds 90 msec
Ended Job = job_1718864244766_0008
Loading data to table onlineshop.onlineshop__customers_cust_index__
Table onlineshop.onlineshop__customers_cust_index__ stats: [numFiles=1, numRows=6, totalSize=594, rawDataSize=588]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.09 sec   HDFS Read: 9532 HDFS Write: 696 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 90 msec
OK
Time taken: 24.168 seconds
```

ACID TRANSACTIONAL FEATURES IN HIVE

Set the below properties:

```
SET hive.support.concurrency=true;
SET hive.enforce.bucketing=true;
SET hive.exec.dynamic.partition.mode=nonstrict;
SET hive.txn.manager=org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
SET hive.compactor.initiator.on=true;
SET hive.compactor.worker.threads=1;
SET hive.optimize.sort.dynamic.partition=false;
```

Create a file: emp_acid.txt

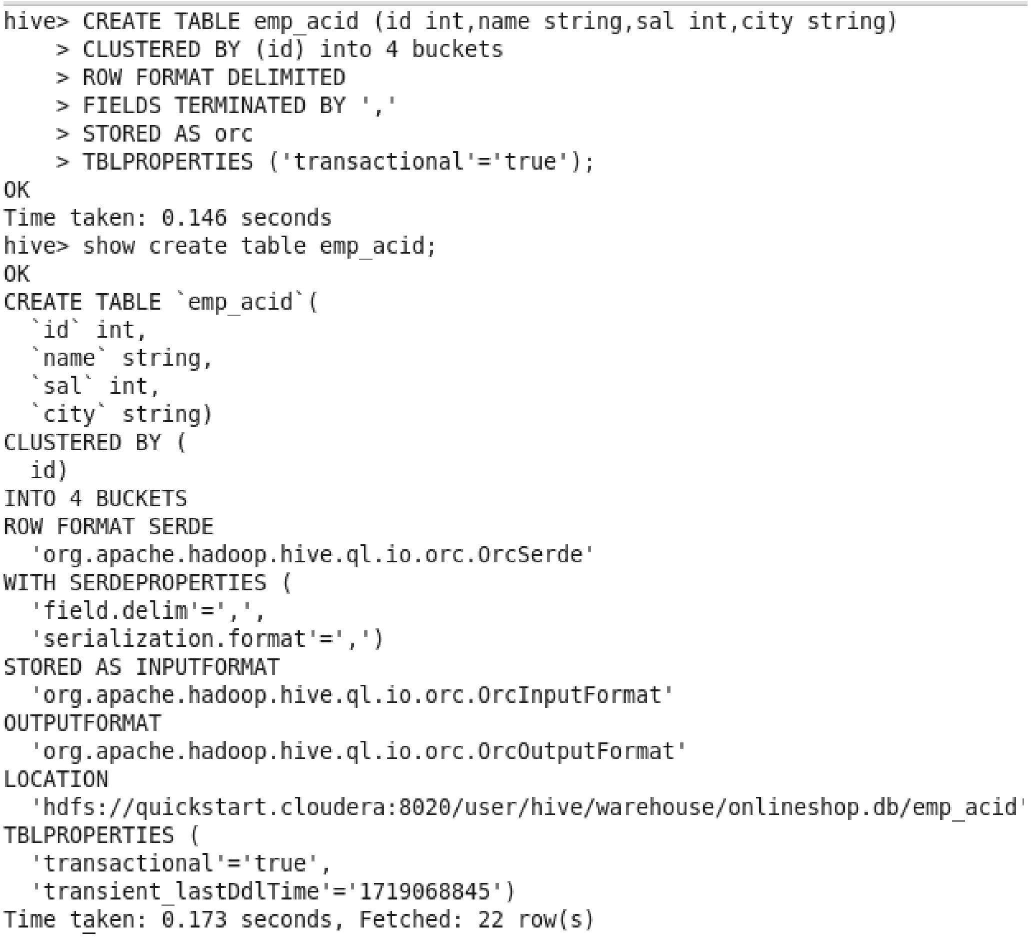
```
Id,name,sal,city
101,saif,100,Mumbai
102,Anup,200,Pune
103,Ram,300,Pune
```

Create a staging Table also called as lock and roll:

CREATE TABLE stg_acid (id int,name string,sal int,city string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Hive/ACID/emp_acid' INTO TABLE stg_acid;

CREATE TABLE emp_acid (id int,name string,sal int,city string)
CLUSTERED BY (id) into 4 buckets
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS orc
TBLPROPERTIES ('transactional'='true');

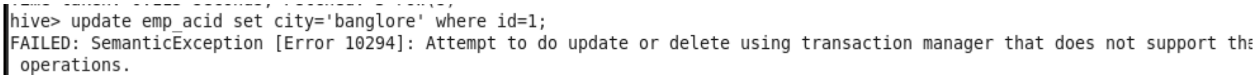


```
hive> CREATE TABLE emp_acid (id int,name string,sal int,city string)
    > CLUSTERED BY (id) into 4 buckets
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS orc
    > TBLPROPERTIES ('transactional'='true');
OK
Time taken: 0.146 seconds
hive> show create table emp_acid;
OK
CREATE TABLE `emp_acid`(
  `id` int,
  `name` string,
  `sal` int,
  `city` string)
CLUSTERED BY (
  id)
INTO 4 BUCKETS
ROW FORMAT SERDE
  'org.apache.hadoop.hive.ql.io.orc.OrcSerde'
WITH SERDEPROPERTIES (
  'field.delim'=',',
  'serialization.format'=',')
STORED AS INPUTFORMAT
  'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'
LOCATION
  'hdfs://quickstart.cloudera:8020/user/hive/warehouse/onlineshop.db/emp_acid'
TBLPROPERTIES (
  'transactional'='true',
  'transient_lastDdlTime'='1719068845')
Time taken: 0.173 seconds, Fetched: 22 row(s)
```

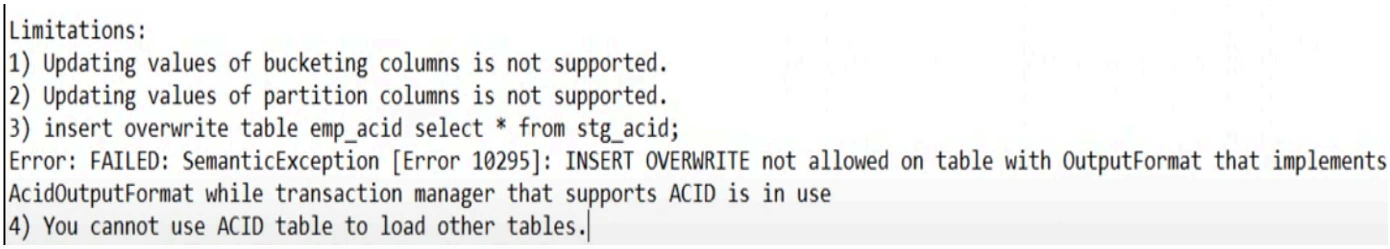
—---

INSERT INTO TABLE emp_acid SELECT * FROM stg_acid;

This below command will give error:
UPDATE emp_acid set city’banglore’ where id=1



```
hive> update emp_acid set city='banglore' where id=1;
FAILED: SemanticException [Error 10294]: Attempt to do update or delete using transaction manager that does not support the
 operations.
```



```
Limitations:
1) Updating values of bucketing columns is not supported.
2) Updating values of partition columns is not supported.
3) insert overwrite table emp_acid select * from stg_acid;
Error: FAILED: SemanticException [Error 10295]: INSERT OVERWRITE not allowed on table with OutputFormat that implements
AcidOutputFormat while transaction manager that supports ACID is in use
4) You cannot use ACID table to load other tables.
```

—---