**Hive\_Challenge\_Task\_3**

Scenario based questions solutions:

**1. Will the reducer work or not if you use “Limit 1” in any HiveQL query?**

= Yes reducer will work if we use "Limit 1" in any Hive Query.

**2. Suppose I have installed Apache Hive on top of my Hadoop cluster using default metastore configuration.Then, what will happen if we have multiple clients trying to access Hive at the same time?**

= When we install Apache hive on top of Hadoop cluster using default metastore configuration it doesn't allow multiple connection at the same time because in default hive uses Derby database to store its metadata and this derby is RDBMS which allows only one connection at a time.

**3. Suppose, I create a table that contains details of all the transactions done by the customers:**

**CREATE TABLE transaction\_details (cust\_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;**

**Now, after inserting 50,000 records in this table, I want to know the total revenue generated for each month.**

**But, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that I will be taking in order to do so?**

= In order to reduce execution time of any given query we should follow following steps:

\* First of all we can create internal table and feed the data into it because internal table takes less time then any external table.

\* We should create ORC table which works more efficiently in hive.

\* We can increase the no. of reducers assigned to run any perticular query.

\* We can create partition to avoid the extra querying of unwanted data in the table for any perticular query.

\* We can create bucketing also in order to reduce the execution time.

**4. How can you add a new partition for the month December in the above partitioned table?**

= In order to add new partition into the already partitioned table we have to use "alter" function along with "CASCADE".

**5. I am inserting data into a table based on partitions dynamically. But, I received an error – FAILED ERROR IN SEMANTIC ANALYSIS: Dynamic partition strict mode requires at least one static partition column. How will you remove this error?**

= To remove this error we should use:

hive> set hive.exec.dynamic.partition.mode=non-strict;

**6. Suppose, I have a CSV file – ‘sample.csv’ present in ‘/temp’ directory with the following entries:**

**id first\_name last\_name email gender ip\_address**

**How will you consume this CSV file into the Hive warehouse using built-in SerDe?**

= hive> create table sample\_csv\_serde

> (

> id int, first\_name string,

> last\_name string, email string, gender string, ip\_address string

> )

> row format serde 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

> with serdeproperties

> (

> "separatorChar" = ","

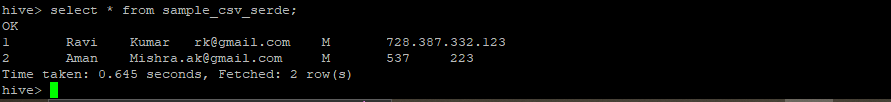
> )

> stored as textfile

> tblproperties("skip.header.line.count"="1");

hive> load data local inpath '/home/cloudera/tmp/sample.csv' into table sample\_csv\_serde;

hive> select \* from sample\_csv\_serde;



**7. Suppose, I have a lot of small CSV files present in the input directory in HDFS and I want to create a single Hive table corresponding to these files.**

**The data in these files are in the format: {id, name, e-mail, country}. Now, as we know, Hadoop performance degrades when we use lots of small files.**

**So, how will you solve this problem where we want to create a single Hive table for lots of small files without degrading the performance of the system?**

**=** create a temporary table:

hive> create table temp\_tbl

(ind int, name string, e\_mail string, country string)

row format delimited

fields terminated by ','

stroed as textfile;

**load data into these file using below command :**

hive> load data inpath '/input\_directory/...' into table temp\_tbl;

**create a table that will store data in sequenceFile format:**

hive> create table seq\_tbl(

ind int, name string, e\_mail string, country string)

row format delimited

fields terminated by ','

stroed as sequencefile;

**transfer data from temporary table into this table:**

hive> Insert overwrite table seq\_tbl select \* from temp\_tbl;

**8. LOAD DATA LOCAL INPATH ‘Home/country/state/’**

**OVERWRITE INTO TABLE address;**

**The following statement failed to execute. What can be the cause?**

= The following statement is failed to execute because overwrite statement works when the file data is already available in any table from their it is used to overwrite the same file into other new table using overwrite statement. And if we have to load the file data we have to use load data local inpath 'Home/country/state/file\_name' into table table\_name;

**9. Is it possible to add 100 nodes when we already have 100 nodes in Hive? If yes, how?**

= Yes it is possible to add 100 more nodes when we have already 100 nodes.

For that we have to shut down the cluster and add 100 nodes using horizontal scaling of cluster.

Each node must have pre specified roles and should be linked with the correct ip address.

Make sure the status of the cluster after booting the cluster again.

Hive Practicle Questions:

hive> create table customers

> (

> Customer\_Id int, Name string, Age int, Address string, Salary int

> )

> row format delimited

> fields terminated by ','

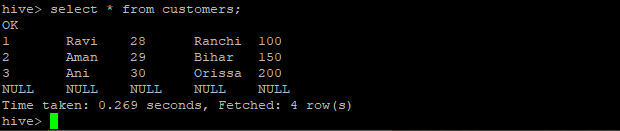
> tblproperties("skip.header.line.count"="1");

OK

Time taken: 1.151 seconds

hive> load data local inpath '/home/cloudera/customers.csv' into table customers;

hive> select \* from customers;



hive> create table order

> (

> O\_Id int, Date date, Customer\_Id int, Amount int

> )

> row format delimited

> fields terminated by ','

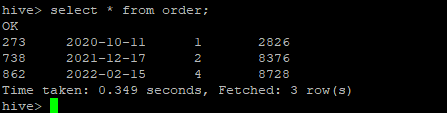
> tblproperties("skip.header.line.count"="1");

OK

Time taken: 0.258 seconds

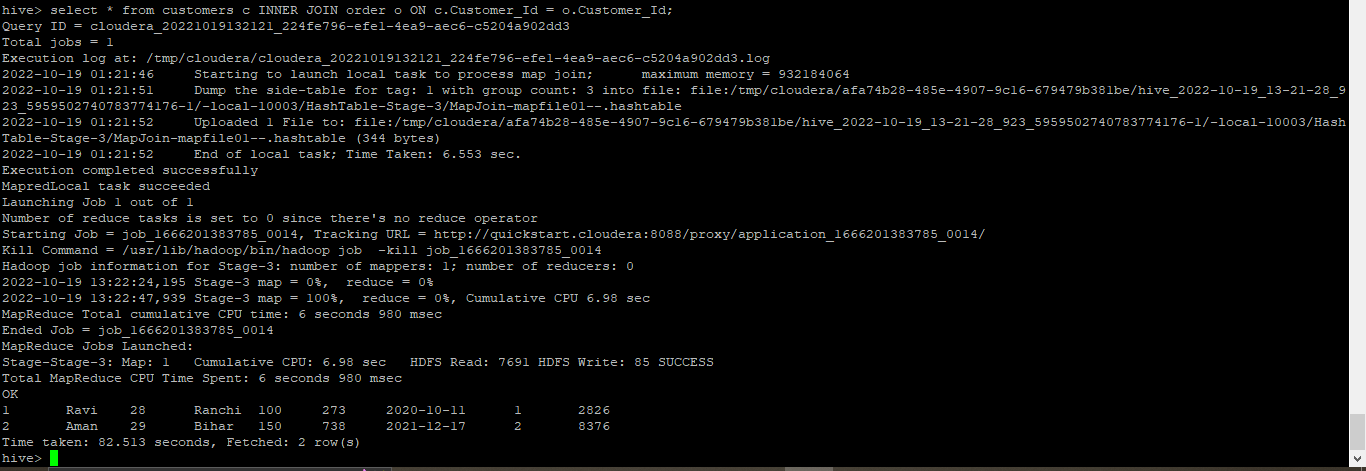
hive> load data local inpath '/home/cloudera/order.csv' into table order;

hive> select \* from order;



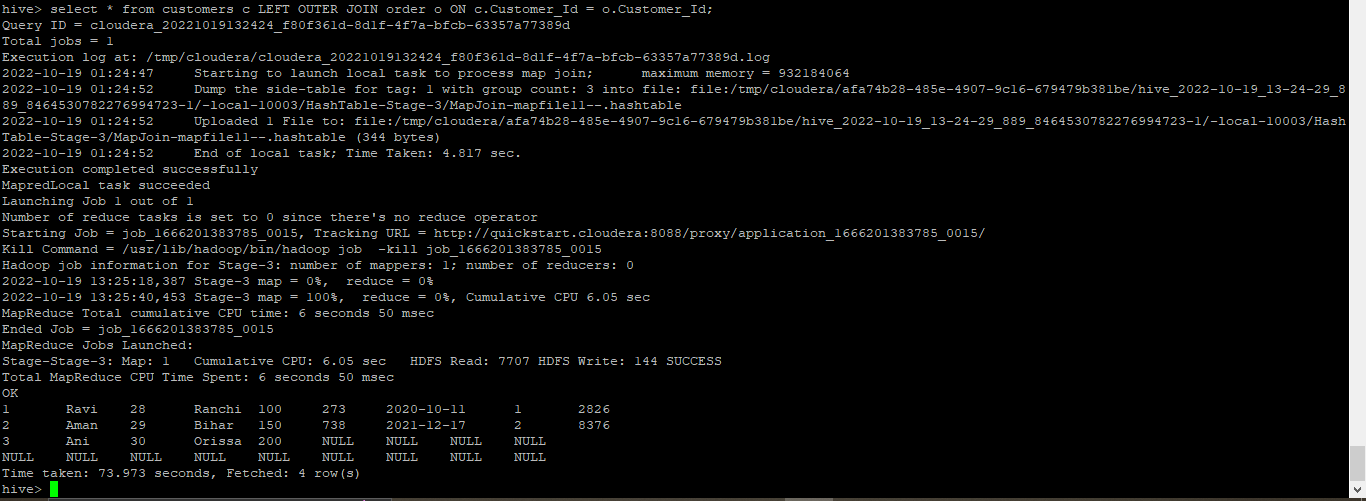
**INNER JOIN**

hive> select \* from customers c INNER JOIN order o ON c.Customer\_Id = o.Customer\_Id;



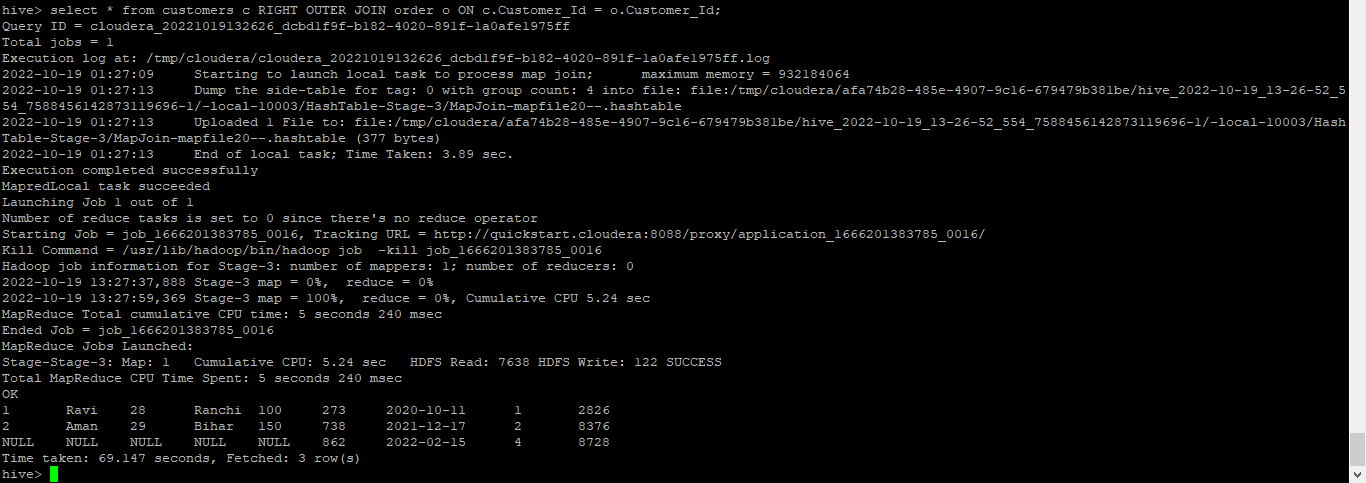
**LEFT OUTER JOIN**

hive> select \* from customers c LEFT OUTER JOIN order o ON c.Customer\_Id = o.Customer\_Id;



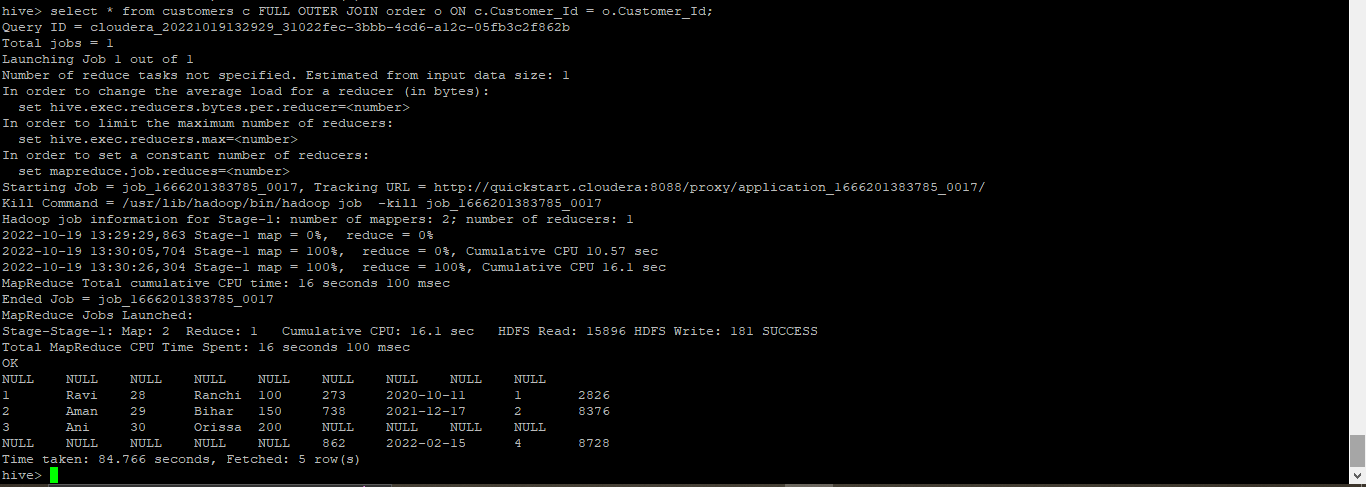
**RIGHT OUTER JOIN**

hive> select \* from customers c RIGHT OUTER JOIN order o ON c.Customer\_Id = o.Customer\_Id;



**FULL OUTER JOIN**

hive> select \* from customers c FULL OUTER JOIN order o ON c.Customer\_Id = o.Customer\_Id;



**BUILD A DATA PIPELINE WITH HIVE**

1. hive> create table zoo

> (

> Animal\_name string, hair string, feathers string, eggs string, milk string, airborne string,

> aquatic string, predator string, toothed string, backbone string, breathes string,

> venomous string, fins string, legs int, tail string, domestic string, catsize string, type int

> )

> row format delimited

> fields terminated by ','

> tblproperties("skip.header.line.count"="1");

OK

hive> load data local inpath '/home/cloudera/zoo\_data.txt' into table zoo;

hive> select \* from zoo;

