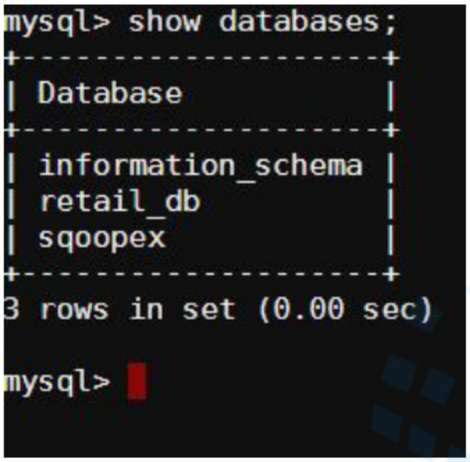
Week 4 FAQs

W4:1 In cloudxlab Mysql I am not able to locate hive metastore DB.



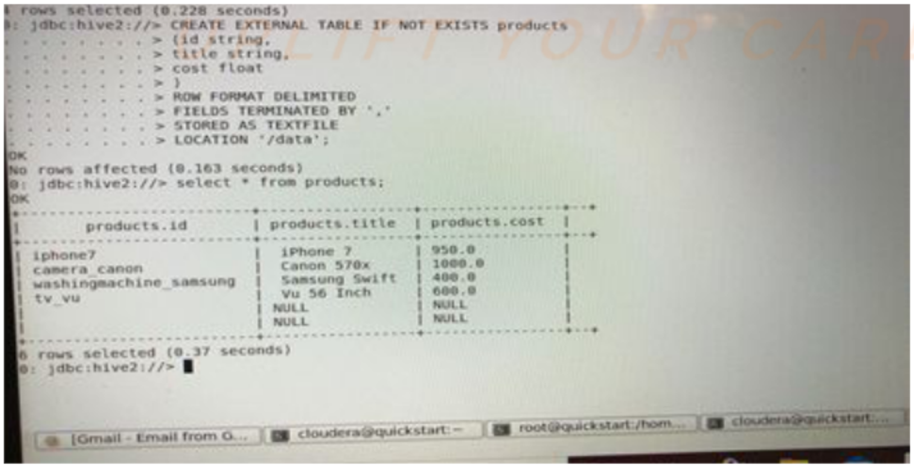
```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| retail_db          |
| sqoopex            |
+--------------------+
3 rows in set (0.00 sec)

mysql> 
```

Ans: yes you cannot see there. it's restricted.. as this is shared env..

W4:2 I am not able to login into HUE in Cloudxlab and not getting any errors.

Ans:hue will be deprecated soon. As that's not the production way of doing things..

W4:3 Why am I seeing nulls in the table?



```
4 rows selected (0.228 seconds)
0: jdbc:hive2://> CREATE EXTERNAL TABLE IF NOT EXISTS products
. . . . . . . . > (id string,
. . . . . . . . > title string,
. . . . . . . . > cost float
. . . . . . . . > )
. . . . . . . . > ROW FORMAT DELIMITED
. . . . . . . . > FIELDS TERMINATED BY ','
. . . . . . . . > STORED AS TEXTFILE
. . . . . . . . > LOCATION '/data';
OK
No rows affected (0.163 seconds)
0: jdbc:hive2://> select * from products;
OK
+--------------------------+-------------------+-------------------+--+
|       products.id        | products.title    | products.cost     |
+--------------------------+-------------------+-------------------+--+
| iphone7                  |  iPhone 7         | 950.0             |
| camera_canon             |  Canon 570x       | 1000.0            |
| washingmachine_samsung   |  Samsung Swift    | 400.0             |
| tv_vu                    |  Vu 56 Inch       | 600.0             |
|                          | NULL              | NULL              |
|                          | NULL              | NULL              |
+--------------------------+-------------------+-------------------+--+
6 rows selected (0.37 seconds)
0: jdbc:hive2://> 
```

Ans: just because the file has some extra empty lines.

W4:4 When we run a select distinct query or any complex query in hive on a table having 4 to 5 rows we see that 1mapper/ 1 reducer is used .. just wanted to confirm if the number of row grows, the number of mapper will also increase, right??

Ans : absolutely.. as there will be more blocks.. and more mappers will be there.

W4:5 Select * from orders where city = 'wa' doesn't seem to invoke any map tasks but why was the answer in the quiz given as map only?

Ans : Since the data was very small hive did some internal optimization so skip mapreduce altogether. If the data is big.. Then treat this like a filter. And for filter a map operation is required. So it's a map only job in a broader sense.

W4:6 if we create an external table , table data will be stored in HDFS. If we create an internal table, where data will store?

Ans Data will stored in hive warehouse directory /user/hive/warehouse in cloudera
FQ And /user/hive/warehouse is the default hive directory in HDFS.
Even internal table data is also available at the same location?
Ans Yes that's correct And only internal/managed tables data gets stored in hive default location.

W4:7 I am getting an error while connecting to hive through beeline using command: beeline -u jdbc:hive2:// I am using itversity lab. Isn't it the right way to connect? Pls help with this

Ans Try this url for connection
jdbc:hive2://nn01.itversity.com:10000
username : scott
password : tiger

W4:8 I was unable to connect to the browser in Cloudera and I don't see any symbol to re-connect to the network. Please help with this?

Ans check the network settings in virtual box manager , whether its showing Network -

W4:9 Can someone plz assist in connecting beeline via itversity plz? I'm using !connect jdbc:hive2://nn01.itversity.com:10000 Error: !connect: event not found. Even if I remove !connect, it still fails

Ans use: beeline -u jdbc:hive2://

W4:10 If I create an External table with location /data/ and if I load a huge dataset into that table. So how is that huge data going to get stored? Like the entire huge data get stored in same location available? what happens if that location doesn't have enough space to store huge data? Is it going to split data to different nodes? I understand that replicas of data get stored in different nodes, does the same apply if there comes space issues?

Ans If the case of the large files also, data gets stored in different nodes in the form of blocks. This concept is independent of the file path/location. If the table is external then it's just loading data. How will you face space issues?

W4:11 In which scenario we can use a managed hive table and external hive table?

Ans: if data is used by multiple teams or by multiple technologies e.g. the same data is used by both hive and spark then we can't create the managed tables since if we delete the table then both data + metadata will be deleted so others won't be able to use it. Also when data lands in data lake we have to create the external tables only on them to read. If we want hive to manage both data + metadata (including deletion) we go for managed tables.

W4:12 After loading the file in HIVE, I can see NULLs. Is it possible to remove them?

Ans: we can use a different table and do a insert overwrite to overwrite the null data with the actual values..provided the file from which we are overwriting have correct values

W4:13 Why are we keeping the data in HDFS where there is less scope of modifications as generally we made the changes or update the DATA rather than Schema (we very rarely add new columns)?

Ans Big data is used for analytical processing and not meant to be used as a transaction processing system like DBMS's. Hence updation of data is also not frequent operation. For e.g if you are analyzing weather forecast data of the last 5 years, we are unlikely to update the values which were captured years ago.

W4:14 We create permanent functions to be used as UDFs. I wanted to know where we give the path of the jar, which path does it have to be in a multi node cluster.

Ans :Copy the JAR file to the host on which Hive Server is running. Save the JARs to any directory you choose, give the hive user read, write, and execute access to this directory, and make a note of the path so that you can give while creating a permanent function. Hope it's clear

W4:15 I have a doubt regarding the external table in Hive when we specify the location '/data/' and if suppose we have 2 files inside the data directory in that, will it not be supported that means we can only give one file /data/products.csv. Please clarify.

Ans: Data of all the files under the data directory will be visible on hive external table
If all the files are not with same delimiter then
1. The files with proper delimiter will be loaded properly
2. The files with different delimiter entire data will be displayed in first column n rest all columns will have NULL value

W4:16 I am getting an error while connecting to hive through beeline using command: beeline -u jdbc:hive2:// I am using Itversity lab. Isn't it the right way to connect?

Ans: Try this url for connection
jdbc:hive2://nn01.itversity.com:10000/
username : scott
password : tiger
OR
was able to resolve the issue as follows:
1.Just type beeline
2.Once u are inside beeline,
type:
!connect jdbc:hive2://gw02.itversity.com:10000/
Username: scott
Password: tiger

W4:17 In mapreduce example (i.e restaurants/movies) we have seen that for distinct conditions, there was no function of reducer job but in hive query example if we run a simple distinct query it will run a reducer, why?

Ans: while getting distinct values, though there is no work of reducer, but the shuffle and sorting are required as the o/p of mappers are not the final o/p. so the shuffle and sort take place only if the reducer is there. That's why the reducer can't be ignored. In the MR example also we were not ignoring the reducer due to this.

W4:18 I have created a database in Hive named as swaroopDB and also one table with some data. But I can’t see this in hdfs, I mean if i give command as hadoop fa -ls /user/hive/warehouse/swaroopDB its saying no such file or directory. Any idea folks what can be done here? I am using Itversity labs.

Ans: It should be hdfs dfs -ls /apps/hive/warehouse/swaroopDB.In itversity labs the hive directory path resides in apps/hive/warehouse.

W4:19 “derby it allows only one connection at a time” above line can any one please elaborate.

Ans: suppose you have 2 or more hive jobs to run.Then only one job can run at a time. The other jobs has to wait until the one which is running is finished

W4:20 When we load data in hive the metadata(schema) is stored in a database of rdbms. But what if the table is created using data from hive tables only so where the metadata for this table will be stored?

Ans: metadata(schema) is always stored in RDBM irrespective of whether it's a Managed table or External as per HIVE's architecture

W4:20 Why does HDFS have high latency / why can we not read from HDFS very quickly? Can anyone please help to understand the reason behind this.

Ans: There are several reasons for HDFS high latency:

1. MapReduce based execution model - which generally involves lots of intermediate operations
2. HDFS design model - design to hold huge volumes of data with support for multiple structure and file formats.
3. Huge disk IO Operations - as it follows MapReduce framework, which involves high IO operations causing latency.
4. Disk persistence - most data are persisted in disk to support Job recovery.
5. Multiple node involvement - as hdfs is based on distributed architecture the access request 1st goes to name node then to data nodes causing delay in receiving data

W4:21 Whether Hive triggers a Mapreduce job for simple read commands like displaying all the records of a really huge table.

Example: Let's say we have a 4-node cluster and our table orders are spanned across all the nodes. Now, If I just want to fire some simple query like select * from orders; will Hive trigger any Mapreduce job? If not, how can Hive retrieve data from all the blocks in all 4 nodes?

Ans : There is something called a fetch task in hive. So when we run simple commands in hive like select * , limit then it won't launch a map reduce job rather it launches fetch task. It can just read the file and dump it to you. But if we have a query like select col1 from table then MR job will be launched as it has to parse each line to get the exact column. We can consider it roughly like cat operation. But thing is it has some threshold size which means if your table size is larger than the threshold then it has to launch a map reduce job even for simple select * as well.

Dated Till :- 15th December 2020