if  we create  2TB table in HDFS and  then delete it using

hadoop fs -rmr location

is the  2 TB  data  still be occupying the space out there.?

Will it be retrievable?

or is  that space will be overwritten by any new data written to hdfs ?

how can we proof that 2TB data is there after deletion ?

I am asking the question from different angle, but one of you answer may cover all those

yes the data will be recoverable automatically by the hdfs since the namenode will have the metadata and the data block copies exists in other location based on its replication factor. so the data will be maintained by the hdfs even the data deelted manually. it won't be overwrittened by any new data.

1. How to see external table structure?

2. If the external table structure is deleted, how the data can stay? I thought the opposite is true, i.e. one can remove the data without removing the structure!!

3. How to export the outputs of multiple hive queries (either from a script file or on the hive prompt) to in HDFS in one shot (i.e. not creating external table for each query)

e need to look in to the physical location to retrieve data from external table

The key difference between external and managed table in Hive is that data in the external table is not managed by Hive. When you create external table you define HDFS directory for that table and Hive is simply "looking" in it and can get data from it but Hive can't delete or change data in that folder. When you drop external table Hive only deletes metadata from its metastore and data in HDFS remains unchanged.

Managed table basically is a directory in HDFS and it's created and managed by Hive. Even more - all operations for removing/changing partitions/raw data/table in that table MUST be done by Hive otherwise metadata in Hive metastore may become incorrect (e.g. you manually delete partition from HDFS but Hive metastore contains info that partition exists).

In Hadoop definative guide I think author meant that it is a common practice to write MR-job that produces some raw data and keeps it in some folder. Than you create Hive external table which will look into that folder. And than safelly run queries without the risk to drop table etc. In other words - you can do MR job that produces some generic data and than use Hive external table as a source of data for insert into managed tables. It helps you to avoid creating boring similar MR jobs and delegate this task to Hive queries - you create query that takes data from external table, aggregates/processes it how you want and puts the result into managed tables.

Another goal of external table is to use as a source data from remote servers, e.g. in csv format.

one TT communicates with only one Data Node bcoz files are stored as blocks in different data nodes.

**Question 1)** JT assigns task(t1) to Task tracker(TT1). Clear till this point. Now, can JT assign task(t2) to the same Task Traker(TT1) while TT1 still needs to send acknowledge to JT on task(t1)

**Question 2**) If Question 1 is Yes. Doesn't this contradict our statement that one TT Communicates with only one datanode.

                           because, if Question 1 is yes, Tast tracker(TT1) communicates with some datanode(d1) for task(t1) also at the same time Task tracker(TT1) communicates with datanode(d2) for task(t2).

How to you load huge data in to cluster? And their commands

Explain Map-Reduce.

Map-Reduce Architecture

Can we able to set number of maps in Map-Reduce?

Can we able to set number of reduce in Map-Reduce?

How you set number of maps in Sqoop?

What is partitioners? Explain where to use it and how to use it

How shuffle, sorting happens

Difference between Text and String?

Is String Mutable?

What is writable? And it uses

Explain Load and Dump in Pig programming?

How do you set maps and reduce in Pig programming?

What are the User Defined Functions in Pig?

How to you load huge data in to cluster? And their commands

Explain Map-Reduce.

Map-Reduce Architecture

Can we able to set number of maps in Map-Reduce?

Can we able to set number of reduce in Map-Reduce?

How you set number of maps in Sqoop?

What is partitioners? Explain where to use it and how to use it

How shuffle, sorting happens

Difference between Text and String?

Is String Mutable?

What is writable? And it uses

Explain Load and Dump in Pig programming?

How do you set maps and reduce in Pig programming?

What are the User Defined Functions in Pig?

 1. Explain Project, what is your role & responsibilities and what you have done?

2. Explain Map-Reduce program average of column in algorithmic process

3. Explain Map-Reduce in real time scenario

4. Explain Map-Reduce architecture

5. what are various datatypes? (Map-reduce, hive, pig, hbase)

6. what datatype you will store datetime in PIG?

7. Is there any map input format?

8. describes the relationship between MapReduce and Pig?

9. Remaining Questions related to BI tools and analytics