Week 6 FAQs

W6:1 What is the importance of surrogate keys in Hive?

Ans: Surrogate key is a column or combination of column set as primary key instead of real ones whenever real ones are not possible due to some reason...thus it helps to implement a primary key for the table

W6:2 in real time which scenario we use distribute by and sort by

Ans :There is no fixed scenario. That purely depends upon the situation and requirements.

W6:3 "Parquet provides good support for deeply nested data structures" Now what is nested data structures here. Can anyone explain with an example ?

Ans : nested data structure means a tree like structure. where some data that is nested or fully contained within something.

W6:4 Like the above code skip headers property is applied. But I wanted to know if we can apply more than one tbl properties while creating a new table.

Ans : Yes more than one table properties can be used while creating a hive table.
Example:tblproperties("skip.header.line.count"="1", "skip.footer.line.count"="1"

W6:5 would like to check how to check the explanation in hive.

Ans :You can use the Apache Hive EXPLAIN command to display the actual execution plan that Hive query engine generates and uses while executing any query in the Hadoop ecosystem. Eg:- EXPLAIN select * from orders. Also EXPLAIN EXTENDED <query> gives extra info about

W6:6 How to create indexes in hive ?

Ans : Below are few commands to work with indexes....CREATE INDEX table01_index ON TABLE table01 (column2) AS 'COMPACT';
SHOW INDEX ON table01;
DROP INDEX table01_index ON table01; But indexes are removed from hive 3.0 version onwards and not recommended to use also (edited)

W6:7 I am trying to do the java project to copy and paste but I don't seem to have the data

Ans could you please clarify what actually you mean by java project here. A screenshot will always be helpful.

W6:8 I have table with 6 columns now I need to create partition table with partitioning on 5 columns someone suggest how I can write the insert command for these I have created a partitioned table

Ans :Have you checked the "Hive Partitioning With 2 Columns" ? Similarly you can proceed with your number of partition columns.

W6:9 Insert into table tablename partition(column names) select How I can write the select query for these

Ans : using multiple AND *operators in where clause*

W6:10 I mean the partition column should be the last column in the select statement ryt but in these cases in which order I can write the partition column names in the select query ?

Ans: the order you want to keep. Suppose partition columns are A - E, then you have to insert data in the same order. Check the partitioning on 2 column topic - same logic will be applied.

W6:11 Row based , column based file formats are different from row based ,column based databases (mysql and hbase respectively) ..?

Ans: files systems and databases are 2 different things… in file formats we talk about various file formats like text file format, parquet, orc.. which can be classified in 2 categories. In databases we have row based and column based. Row based example: mysql, Column based example hbase.

W6:12 If we get the input file as json format how do we handle it in hive ? I have tried creating a normal table on top of the json data and after that inserting that into an orc hive table with json serde, but it does not work. It is still showing data in json format in the hive orc

```
File  Edit  View  Search  Terminal  Help
hive> select * from customers_json limit 2;
OK
{"custno": 1,"firstname": "Dai","lastname": "Noble","gender": 2,"age": 53,"profession": "Human Resources","contactNo": "16441019280 99","emailId": "eu.lacus.Quisque@l
om","city": "Berlin","state": "Berlin","isActive": 1,"createdDate": "2017-03-01","UpdatedDate": "2018-08-25"}    NULL    NULL    NULL    NULL    NULL    NULL    NULL
ULL     NULL    NULL    NULL
{"custno": 2,"firstname": "Kay","lastname": "Wise","gender": 2,"age": 26,"profession": "Public Relations","contactNo": "1674031917099","emailId": "ridiculus@Sednunc.
y": "Oldenzaal","state": "Overijssel","isActive": 2,"createdDate": "2017-04-10","UpdatedDate": "2018-08-30"}    NULL    NULL    NULL    NULL    NULL    NULL    NULL
ULL     NULL    NULL    NULL
Time taken: 0.039 seconds, Fetched: 2 row(s)
hive> select * from customers_json_serde limit 2;
OK
{"custno": 1,"firstname": "Dai","lastname": "Noble","gender": 2,"age": 53,"profession": "Human Resources","contactNo": "16441019280 99","emailId": "eu.lacus.Quisque@l
om","city": "Berlin","state": "Berlin","isActive": 1,"createdDate": "2017-03-01","UpdatedDate": "2018-08-25"}    NULL    NULL    NULL    NULL    NULL    NULL    NULL
ULL     NULL    NULL    NULL
{"custno": 2,"firstname": "Kay","lastname": "Wise","gender": 2,"age": 26,"profession": "Public Relations","contactNo": "1674031917099","emailId": "ridiculus@Sednunc.
y": "Oldenzaal","state": "Overijssel","isActive": 2,"createdDate": "2017-04-10","UpdatedDate": "2018-08-30"}    NULL    NULL    NULL    NULL    NULL    NULL    NULL
ULL     NULL    NULL    NULL
Time taken: 0.043 seconds, Fetched: 2 row(s)
hive> select * from customers_json_parquet limit 2;
OK
{"custno": 1,"firstname": "Dai","lastname": "Noble","gender": 2,"age": 53,"profession": "Human Resources","contactNo": "16441019280 99","emailId": "eu.lacus.Quisque@l
om","city": "Berlin","state": "Berlin","isActive": 1,"createdDate": "2017-03-01","UpdatedDate": "2018-08-25"}    NULL    NULL    NULL    NULL    NULL    NULL    NULL
ULL     NULL    NULL    NULL
{"custno": 2,"firstname": "Kay","lastname": "Wise","gender": 2,"age": 26,"profession": "Public Relations","contactNo": "1674031917099","emailId": "ridiculus@Sednunc.
```

```
create table customers_json
(custno string  ,firstname string  ,lastname string  ,
gender string  ,age string  ,profession string  ,contactno string  ,
emailid string ,city string  ,state string  ,isactive string  ,
createddate string  ,updateddate string );


load data local inpath '/home/cloudera/Desktop/Shared/customers.json' into table customers_json;


create table if not exists customers_json_parquet(
custno string  ,firstname string  ,lastname string  ,
gender string  ,age string  ,profession string  ,contactno string  ,
emailid string ,city string  ,state string  ,isactive string  ,
createddate string  ,updateddate string
)
row format serde 'org.openx.data.jsonserde.JsonSerDe';|;

insert into customers_json_parquet select * from customers_json;
```

Ans:-In a normal table how can you directly load json without a json serde? thats wrong

W6:13 1. Create a table2. Load data using OCR or PARQUET file format.Can we further use partitioning or bucketing on the same table where data is loaded in OCR or PARQUET file format ?

Ans:- yes absolutely..

W6:14 while trying to do week5 Assignment with optimised file formats and compression, I am getting an error in Map join.Illegal Argument Exception- Illegal character in path at index 85 - for state = Cases being reassigned to states.

Ans:- In the query ,You are trying right outer join as map side join here.You can try inner join as map side join to get matches from both tables.And use on clause instead of where n join Columns as state and date.

W6:15 Which is the compatible platform for Avro file format?

Ans:- Avro formats commonly used in kafka and in landing zone where the files are loaded directly from source.

W6:16 When we are using dictionary encoding since we are not storing actual value it will save storage space but while processing we need to decode the data will it not impact the processing time? If not why?

Ans:- the more savings/benefits we are getting in terms of less storage and thus less IO required is compromised the little bit of processing overhead

W6:17 Is it possible to directly load data from a .CSV file stored in HDFS to an optimized table with partitions buckets and orc file format.In video I saw initially we are creating a raw table on top of .CSV and then loading data to another table which is optimised (having orc file format)

Ans :- When we load the data from the HDFS to the hive table using load command then it is just like cutting and pasting the data. Then how would hive know which partition or bucket it is.(except only static partitioning where we explicitly mention which partition this file to go).
-That's why first we need to impose the basic structure on top of the file so that hive knows which column a particular value in a record in that file belongs to so that it can partition or bucket it in the opt table.

W6:18 Parquet is good for handling nested data, what does it mean?

Ans:-https://blog.twitter.com/engineering/en_us/a/2013/dremel-made-simple-with-parquet.html#:~:text=Parquet%20stores%20nested%20data%20structures,the%20Dremel%20paper%20from%20Google.&text=I%2FO%20will%20be%20reduced,required%20to%20read%20the%20input.

W6:19 What is the exact meaning of a best compression ratio? Correct me if i am wrong: as per my understanding it is the good balance between the speed and the compression in terms of storage.

Ans:- compression ratio is nothing but (uncompressed file size / compressed file size). So higher/deeper the compression, more will be time and less operational speed. So we choose the technique which gives good balance between CR and speed

W6:20 suppose the last 16k bytes of orc contains some data of file footer along with the whole of postscript, in this case does it discards the file footer and reads only postscript first?secondly if the postscript is more than 16k bytes how it proceeds in this case?

Ans :- the last byte of the orc file gives/holds the length of postscript in file. and postscript is the only thing in orc which is NOT compressed, so we read the postscript first. postscript also contains the length of footer. So after reading length from ps we read footer. postscript is never more than 256 bytes in length.

W6:21 Partition on State referring to null and the partitions directory's are created with some dynamic numbers

Ans :- While inserting into the partition table use the state column at the end. Then it will partition on the basis of state.

W6:22 All, which approach is more used in the Industry for loading data in a partitioned table(a). using a non-partitioned table to load data into a partitioned table(b) or directly creating a Partitioned table, adding the data to the Partitioned table directory, and running MSCK REPAIR to update metadata so that the Partitioned table shows data?

Ans:Approach (a) Approach (b) is more like static partitioning

W6:23 In case of partitioning on an ORC file, first partitions are created based on the partition column and then ORC further subdivides the data based on STRIPES and ROW GROUPS right?

Ans : Absolutely

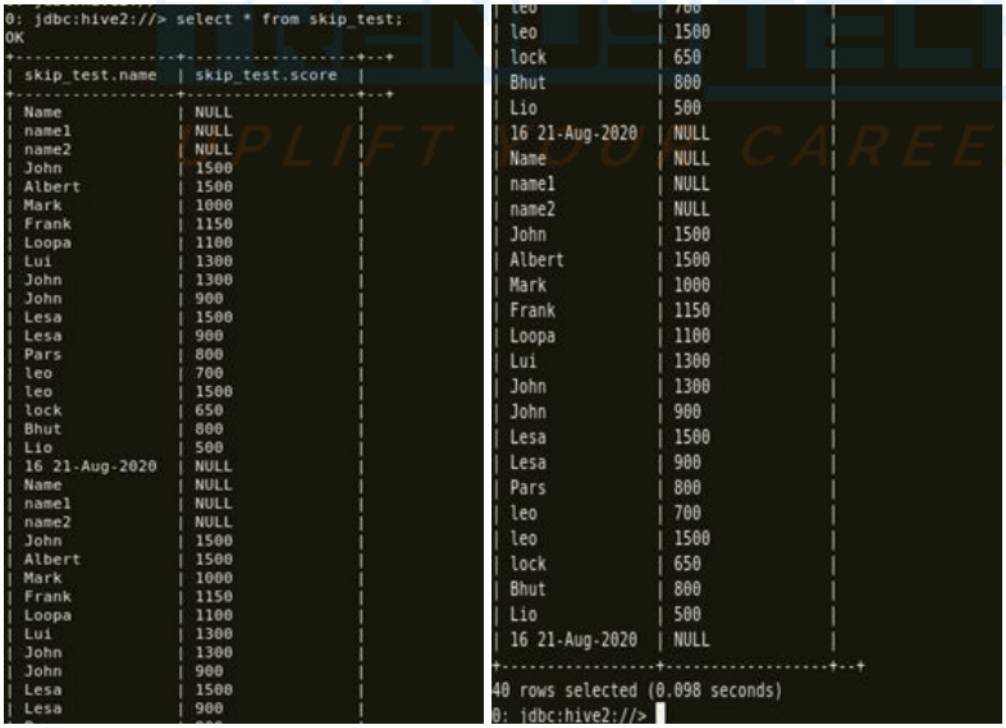
W6:24 With Immutability we can insert only once in a table however, I used load data local inpath command twice and it appended the data rather than overwriting it. Can you please explain this behavior?

Ans:Loading can be done both in append mode and overwrite mode.
Ex1: load data local inpath '/path/to/file' into table <table_name>;
The above command will append the table if you use twice but..
Ex2: load data local inpath '/path/to/file' **overwrite** into table <table_name>;
This command will overwrite the table.

W6:25 Created an Immutable table :

```
create table skip_test
(
  name string,
  score int
)
row format delimited
fields terminated by ','
tblproperties("immutable"="true");
load data local inpath '/home/cloudera/Hive/skip_dataset.csv' into table skip_test
```

This will load 20 records into the table
Now, if I run the load data command again, it appends 20 more records to skip_test table whereas it should give an error because of violating Immutability.



Ans: https://www.docs4dev.com/docs/en/apache-hive/3.1.1/reference/LanguageManual_DML.html

W6: why the number of Buckets is equal to the number of Reducers as in real-time total reducers count might be significantly less compared to total mappers and it will slow down the data insert due to no of Reducers launched?

Ans:Refer this Conversation for understanding

FQ: Counter question - How many output files get created?

Ans:output files are created based on the number of buckets but why can't these be generated by mapper?
Before putting data in buckets, there is a need to sort the hash function ouput so that the bucket column with the same hash function output goes to same bucket and for this purpose reducer is required
eg :
if we bucketed on Id column with 12 values from 1 to 12 then
bucket 0 -> {4,8,12}
bucket 1 -> {1,5,9}
bucket 2 -> {2,6,10}
bucket 3 -> {3,7,11}
This result is very similar to the shuffle and sort process which is part of reducer that's why we require reducers here.
And each reducer is launched to work on a respective hash function output and group all the similar output in one file

W6:26All, even after setting the below required properties for bucket map join i am not able to see bucket map join in the EXPLAIN EXTENDED query.

set hive.optimize.bucketmapjoin=true;
SET hive.enforce.bucketing=true;
select c.customer_id ,c.customer_fname ,c.customer_lname ,o.order_id ,o.order_date
from orders o join customers c on (o.order_customer_id = c.customer_id)
limit 10;

Ans : SET hive.auto.convert.join = true- to apply map join
SET hive.enforce.bucketing=true -- to enforce bucketing
SET hive.optimize.bucketmapjoin=true-- to apply bucket map join
In the query no hints are used.So we can:

set hive.ignore.mapjoin.hint=truePlz try if this works..also the tables have to be both bucketed on join columns n no.of buckets in one table should be an integral multiple of no of buckets in other table

---

Dated Till 18th December 2020