# Week 5 FAQs

W5:1 Can we partition the external table?same question for buckets, can we create buckets on the external table?

Ans : Yes

W5:2 In Dynamic Partitioning, is there a way to create only partitioned tables and load data directly without creating an additional non-partitioned table?

Ans :Nope.It's a 2 step process

W5:3 I am getting below issue in cloudxlab for SMB Join
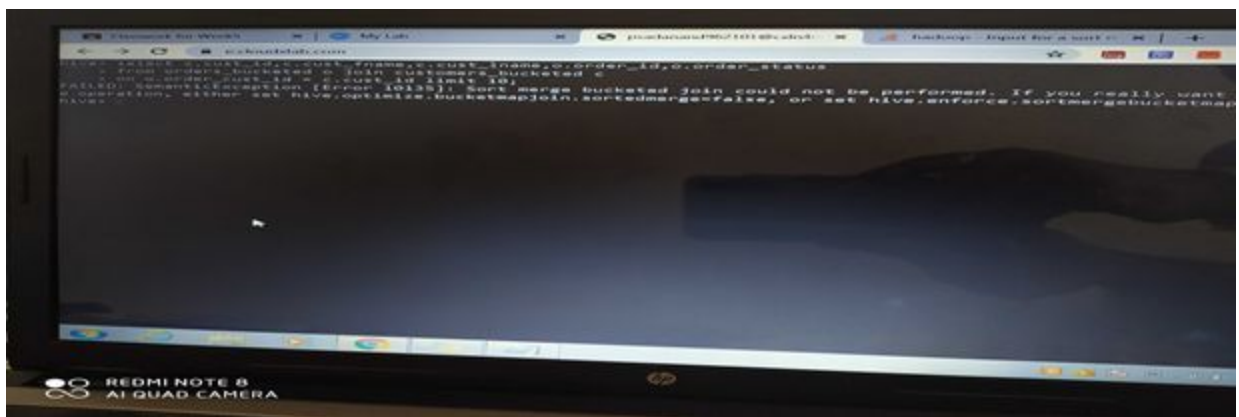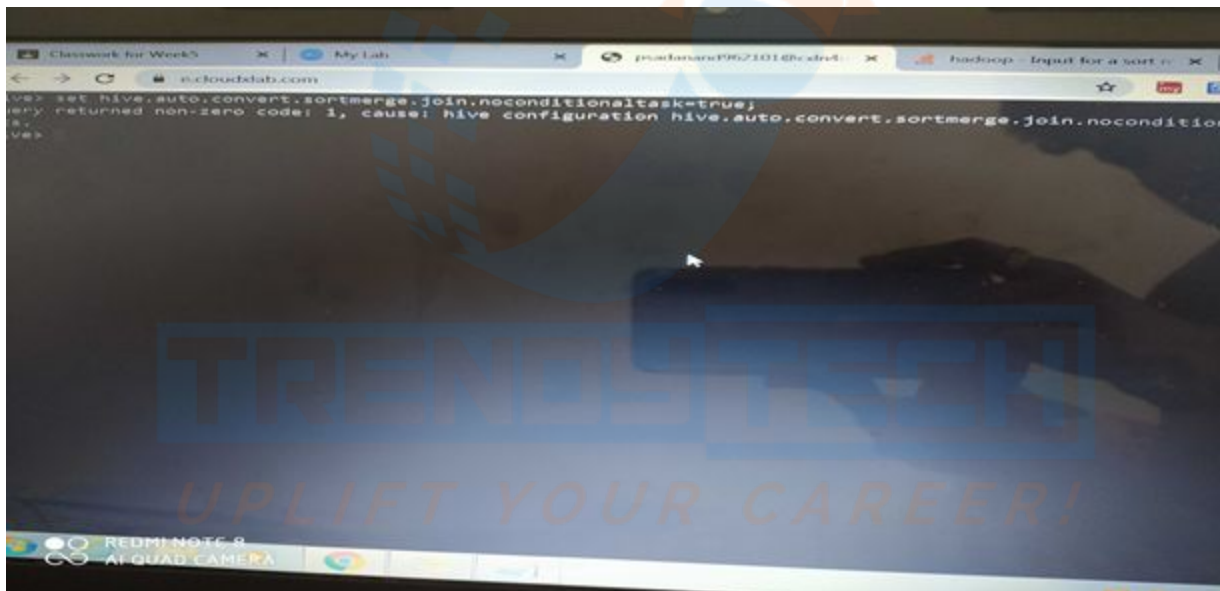
Ans :Join query worked as SMB Join now after setting property
hive.enforce.sortmergebucketmapjoin=false

W5:4 I have a small doubt. Sqoop import and export is working fine with encrypted password .(password alias). but it is throwing errors when i keep the same in sqoop job.

Ans : sqoop job
-Dhadoop.security.credential.provider.path=jceks://hdfs/user/cloudera/mysql.week5.jceks
--create job_assignment -- import ....
I kept the -Dhadoop command after the sqoop job. Now it is working fine.

W5:5 In real Time how we get to know what all columns we need to choose to partitions the data as a Developer BA has to inform us or we will get those details with Data Modeling Document

Ans : We can choose a column with low cardinality as,partition column in hive.If we have column on which queries are run,very frequently,then we choose that as partition column

FQ:In real Time how we get to know what all columns we need to choose to partitions the data and how u get to know one which column u will run the query frequenty ,prior making a table how can u query the table

Ans : if you don't have any idea of data then we cant think of any optimization. having a clarity on your use case and clarity of data will help you to structure your tables in right way

FQ: if we have 200 columns in a file so to Perform Design Table optimization in that case we need to analyse each and every column and which column is well suited as a partitioned column means which has a LOW CARNALITY and used FREQUENTLY in QUERY in that scenario checking each and every column approach is good to follow?
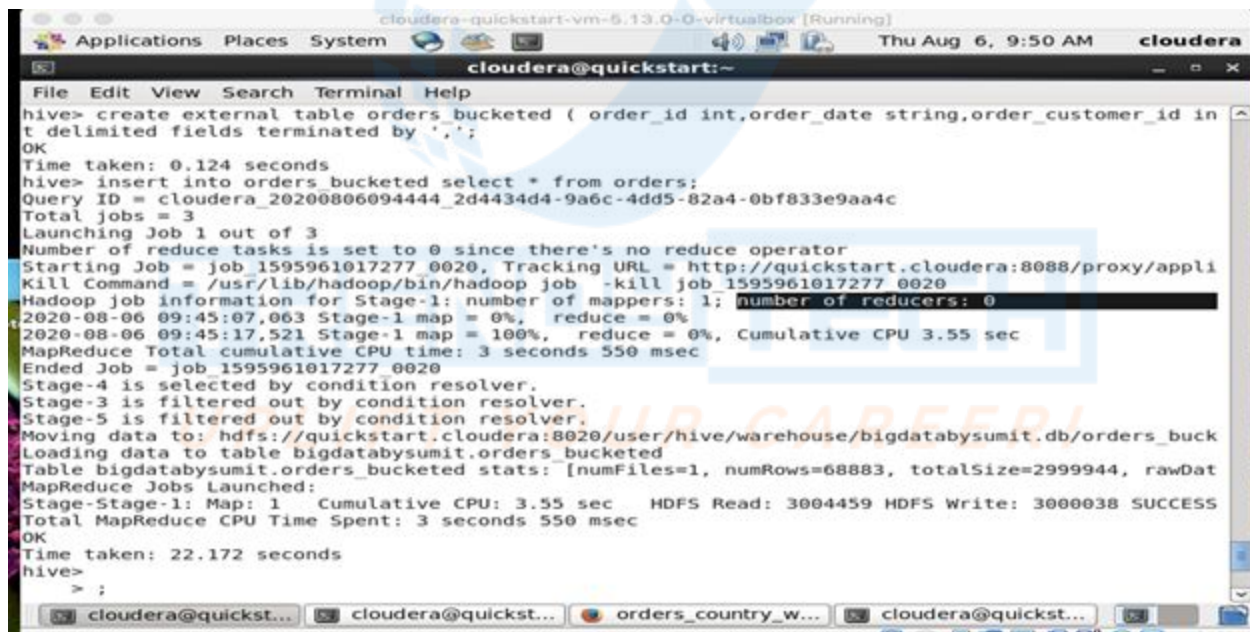
Ans : Ok if we don't have an idea about our data. What cash we do.. Yes of course even if we have 1000 columns having ideas about our data is necessary. Else you can't think of partitioning and bucketing

W5:6 In dynamic partitioning and Bucketing, we require a normal table to data load further into the partitioned or bucketed table and loading to these runs MapReduce (Insert into TableA_Partition select * from TableB) which takes lots of time. With the increased data volume, is there a way to speed up the insert into these tables?

Ans : Think of it this way, if we have to send California data into the California folder.. this of course requires some work. this work is worth doing because you just do this work once and get forever performance gains.. and normally daily we might be dealing with incremental data which we have to send in tables.. it won't be too huge

W5:7 after creating my bucketed tables e.g "orders_bucketed" and i run "insert into orders_bucketed select * from orders" zero reducers run instead of 8

Ans : it is because you forgot to set the property hive.enforce.bucketing to true

W5:8 I am trying JOIN operation using MAPJOIN hint and I have also set both auto convert and ignore hint properties as false but still the reducer is launched. but ideally the reducer should not be launched. Can you explain this?

Ans: an order by clause is there in your query, it is using a reducer for global ordering of data.Remove the order by and check.As properties are set correctly ,it should work omitting order by

W5:9 In an interview ..interviewer asked me what is the SEMI JOINS in Hive..anyone has any idea about this

Ans:In SMB all joins are possible.refer this cloudera doc.. this is quite good.
https://docs.cloudera.com/HDPDocuments/HDP2/HDP-2.0.0.2/ds_Hive/optimize-joins.html

W5:10 In Sorting what happens when no of reducers are 2, data size from mappers is 1GB and property set hive.exec.reducers.bytes.per.reducer is set to default i.e 250 MB?

Ans: I believe if the number of reducers property is set then it takes precedence over anything else... but you can still verify the same.

FQ: ok but in this case as each reducer can only hold 250MB data and we have 2 reducers in place(due to property) so in total 500MB data can be taken up by reducers, then how will the remaining 500MB data be handled?

Ans :If you set a number of reducers that takes precedence that means if mappers give 4gb then 4gb is divided among 2 reducers.there won't be any role of 250 mb property that's what I meant by saying it takes precedence.and in such cases you might end up getting out of memory error..

W5:11 I'm curious to know how data will get divided if we bucket based on date column, means each bucket will get which type of data?

Ans:1. think like case with string.. internally there will be a hash function based on which it will divide it in buckets
2.:In case of bucketing on Date datatype column,
date is first converted to unix_timestamp which transforms date into int/bigint and then mod is calculated

W5:12 When do we call a dataset as small. Is there any particular size, record count or column count to decide it as a small dataset.

Ans : small means 25 mb or less. this number can be changed. This is covered as part of my video.

W5:13 In hive optimization , there is a small table and big table , as we know that the table which is having size less than 25mb is considered as a small table . My question, Is this size (25 mb for smaller file) fixed for a real production environment ?

Ans: This can be changed. The default size is 25mb

W5:14 I have a file in hive with date fields as varchar, i want them in yyyy-mm-dd format ..how i can achieve this? i tried to use cast function but could not change it

Ans: We can use the to_date(colName, pattern) function. Pattern we can give dd-MM-yyyy

W5:15 Can we create a partition within the partition of a table ??

Ans:When you define multiple column names in partitioned by clause at time of creating table structure.. 1st column will be parent partition(eg.country) and inside that 2nd column partition(eg. State) will get created partitioned by (country, state)

**W5:16 [root@quickstart ~]# sqoop import **

**--connect jdbc:mysql://quickstart.cloudera:3306/retail_db **

**--username retail_dba **

--password cloudera

--table orders

--warehouse-dir /user/cloudera

Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail. Please set $ACCUMULO_HOME to the root of your Accumulo installation. 20/11/13 14:32:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0 20/11/13 14:32:46 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead. 20/11/13 14:32:46 ERROR tool.BaseSqoopTool: Error parsing arguments for import:

Ans: Issue is with '--table' clause There must be a space between table name and '\'. I tried the same command with a space in between 'orders' and '\', and it works

W5:17 i am using cloudxlab and need orders and customers table for join optimization, how can i get these two tables to my cloudxlab given that these two tables are not provided

ans:Use winscp to transfer data to local of the cloudxlab

W5:18 How window function helps in optimization can any one explain ?

Ans:Windows function helps us to write complex queries in one query instead of sub/nested queries.. And, it also helps in reducing the dataset to query by applying a proper window (like duration of day, month)

W5:19 While creating this table in MySQL i am getting an error. Tried a lot but could not find out what is the issue.

```
hive> create table IF NOT EXISTS State_Testing
    > (seq INT NOT NULL PRIMARY KEY,
    > date VARCHAR (30),
    > state VARCHAR(50) NOT NULL,
    > total_samples INT,
    > negative INT,
    > positive INT);
MismatchedTokenException(171!=301)
        at org.antlr.runtime.BaseRecognizer.recoverFromMismatchedToken(BaseRecognizer.java:617)
        at org.antlr.runtime.BaseRecognizer.match(BaseRecognizer.java:115)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:5063)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:2557)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:1589)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1065)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:201)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:166)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:522)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1356)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1473)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1285)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1275)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:226)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:175)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:389)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:781)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:699)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:634)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 2:9 mismatched input 'NOT' expecting ) near 'INT' in create table statement
hive>
```

Ans : its mysql syntax but you are trying to execute on hive..string data type should be used instead of varchar2 in case of hive

W5:20 We set the property for hive.enforce.bucketing as true. This means by default in that session all the partitioned tables created will have this bucketing enabled. What if we quit the session and connect back do we need to set this property again or is it applicable only for a particular session?

Ans : We set the property for hive.enforce.bucketing as true. This means by default in that session all the partitioned tables created will have this bucketing enabled? Yes.
If you quit and reconnect the session you need to set the property again because the default value is false.

W5:21 Can we alter no. of buckets ?

Ans :We can change the number of buckets in a table using alter table command.
But this won't restructure the existing data.Change will be reflected only for new entries

W5:22 a. from products_no_buckets insert into table products_w_buckets select id,name,cost,category;

b. insert into table products_w_buckets select id,name,cost,category from products_no_buckets;

Ans : I just went through Hive Bucketing Practical video where we have used 1st query ... so both of them are fine, just the syntax is different.

W5:23 For Partitioning , can we consider the date column? , as in rdbms , i.e daywise , month wise etc.. ?

Ans : we can do partition on the date column as there would be distinct 365 partitions for a year. So if we have a table lets say having 4-5 years of data then we can do partition based on the date column.

W5:24 In Hive - if we need to do partitioning on the column "title" , do we need to change the DDL for the table and order in the source file?? ie: title column should come at last and same in source fileAs I tried to give the "title" as partition , it did partition on the cost. Please explain is this feasible in practical scenario.??create table freshproducts(id string ,title string,cost float)

Ans:Create a table freshproducts as given above and load from the file. Create a another

table freshproducts_partition as below

(Id string,

Cost float)

Partition by (title string).

Now insert data from fresh products to freshprodutcs_partition

Insert overwrite table into freshproducts_partition partition(title)

Select id,title, cost from freshproducts;

W5:25 select /*+MAPJOIN(o)*/
c.customer_id,c.customer_fname,c.customer_lname,o.order_id,o.order_date
from orders o left outer join customers c on
(o.order_customer_id=c.customer_id) limit 10;

"When left table is small & right table is bigger then left outer join cannot be treated like map side join" -- here left table is order which is bigger then customer , so how can the statement is to be interpreted ?

Ans : There is a property hive.mapjoin.smalltable.filesize which is set to 25 mb by default. Which means if your table size is less than 25 mb that would be considered as a small table for the system. Now even though orders are bigger than customers but it is still smaller than this property. Hence orders are considered as small tables. And if the table is small then only it will be loaded in memory..

W5:26 in Join optimizations, especially in Map side joins, there is a local task that runs even before the actual MR job starts. Where does this local task run ? on data nodes which have small tables residing, or on Namenode ?

Ans : It will be on locally where small table resides

Dated Till :-19th December 2020