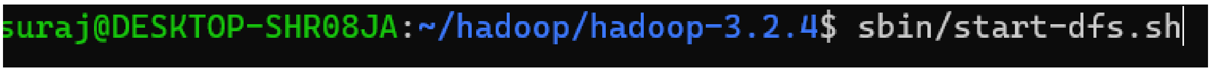
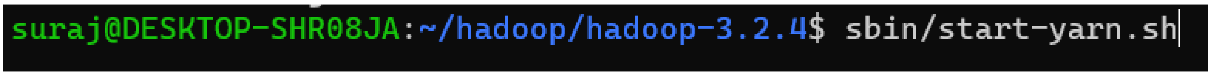
Analysing sample Superstore data-set by creating dynamic partitions using hives

Hadoop command

Go to hadoop directory
Start dfs , then yarn

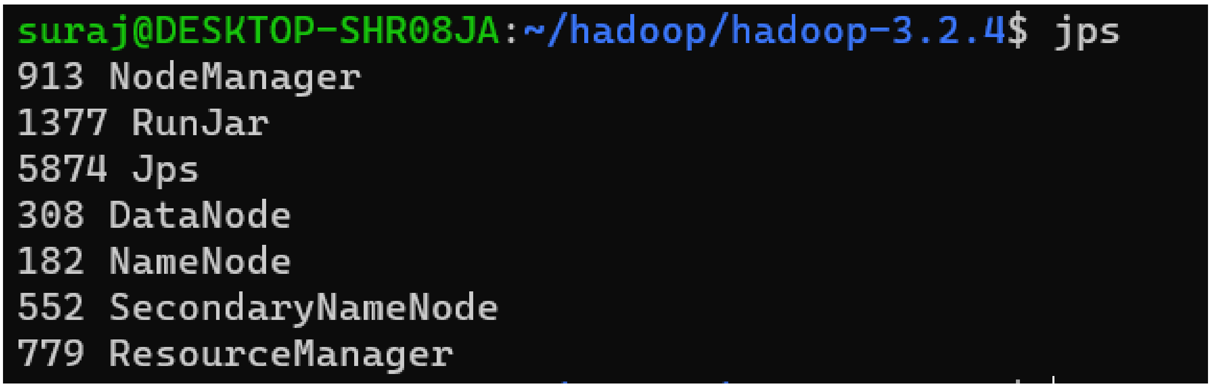


```
suraj@DESKTOP-SHR08JA:~/hadoop/hadoop-3.2.4$ sbin/start-dfs.sh
```



```
suraj@DESKTOP-SHR08JA:~/hadoop/hadoop-3.2.4$ sbin/start-yarn.sh
```

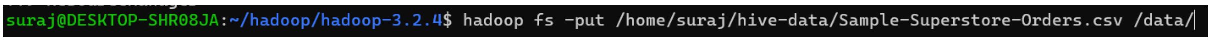
Do check whether the all resources are running or not by doing jps



```
suraj@DESKTOP-SHR08JA:~/hadoop/hadoop-3.2.4$ jps
913 NodeManager
1377 RunJar
5874 Jps
308 DataNode
182 NameNode
552 SecondaryNameNode
779 ResourceManager
```

Now put the data Sample-Superstore-Orders.csv on to HDFS

Please check the code sample provided

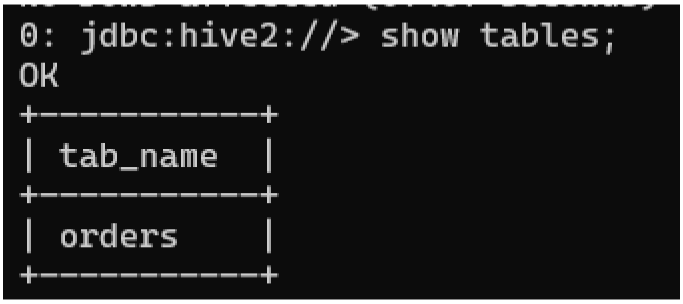


```
suraj@DESKTOP-SHR08JA:~/hadoop/hadoop-3.2.4$ hadoop fs -put /home/suraj/hive-data/Sample-Superstore-Orders.csv /data/
```

All the tables which we are going to create can be seen by going to localhost:9870 from your search engine then go to utilities->browse the file system

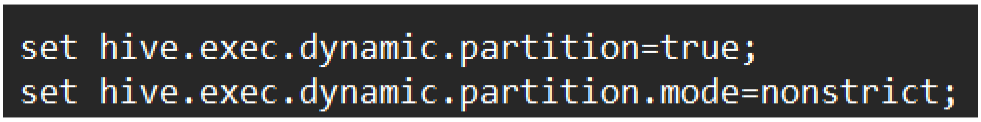
As i have created this tables inside the default directory of hive which is user/hive/warehouse and inside the database /salesdataanalysis.db

I have given the code for creating the first non-partitioned table orders with the help of this table we are going to create the rest of the tables



```
0: jdbc:hive2://> show tables;
OK
+-----------+
| tab_name  |
+-----------+
| orders    |
+-----------+
```

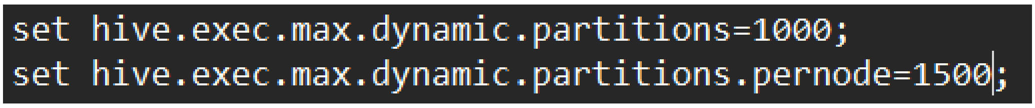
NOTE: by default dynamic partition is not allowed in hive. So in order to create dynamic partition type the below command in your hive terminal.



```
set hive.exec.dynamic.partition=true;
set hive.exec.dynamic.partition.mode=nonstrict;
```

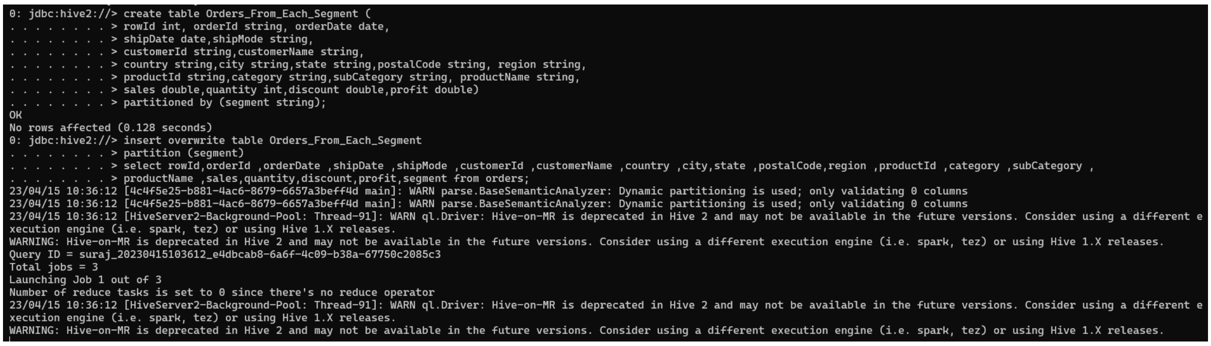
Optional:

If you find any difficulties while creating dynamic partitioning table
Then drop the table and try the below code, after that try creating table



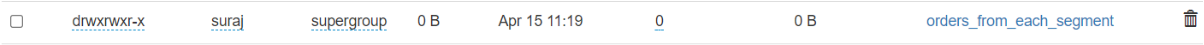
```
set hive.exec.max.dynamic.partitions=1000;
set hive.exec.max.dynamic.partitions.pernode=1500;
```

1. Creating the table orders_from_each_segment (basically i am partitioning the based on different segment present in the original table orders)



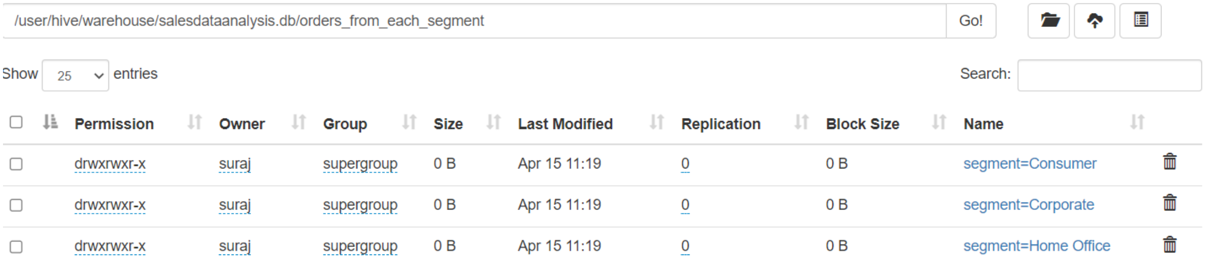
```
0: jdbc:hive2://> create table Orders_From_Each_Segment (
. . . . . . . . > rowId int, orderId string, orderDate date,
. . . . . . . . > shipDate date,shipMode string,
. . . . . . . . > customerId string,customerName string,
. . . . . . . . > country string,city string,state string,postalCode string, region string,
. . . . . . . . > productId string,category string,subCategory string, productName string,
. . . . . . . . > sales double,quantity int,discount double,profit double)
. . . . . . . . > partitioned by (segment string);
OK
No rows affected (0.128 seconds)
0: jdbc:hive2://> insert overwrite table Orders_From_Each_Segment
. . . . . . . . > partition (segment)
. . . . . . . . > select rowId,orderId ,orderDate ,shipDate ,shipMode ,customerId ,customerName ,country ,city,state ,postalCode,region ,productId ,category ,subCategory ,
. . . . . . . . > productName ,sales,quantity,discount,profit,segment from orders;
23/04/15 10:36:12 [4c4f5e25-b881-4ac6-8679-6657a3beff4d main]: WARN parse.BaseSemanticAnalyzer: Dynamic partitioning is used; only validating 0 columns
23/04/15 10:36:12 [4c4f5e25-b881-4ac6-8679-6657a3beff4d main]: WARN parse.BaseSemanticAnalyzer: Dynamic partitioning is used; only validating 0 columns
23/04/15 10:36:12 [HiveServer2-Background-Pool: Thread-91]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different e
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = suraj_20230415103612_e4dbcab8-6a6f-4c09-b38a-67750c2085c3
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
23/04/15 10:36:12 [HiveServer2-Background-Pool: Thread-91]: WARN ql.Driver: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different e
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
```

When you click on the below table present in the hdfs



□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:19	0	0 B	orders_from_each_segment	🗑

You will see that the partition is automatically created because only these 3 segments are there in the table it got created
Note: here we are using insert overwrite command to create dynamic partitioning.

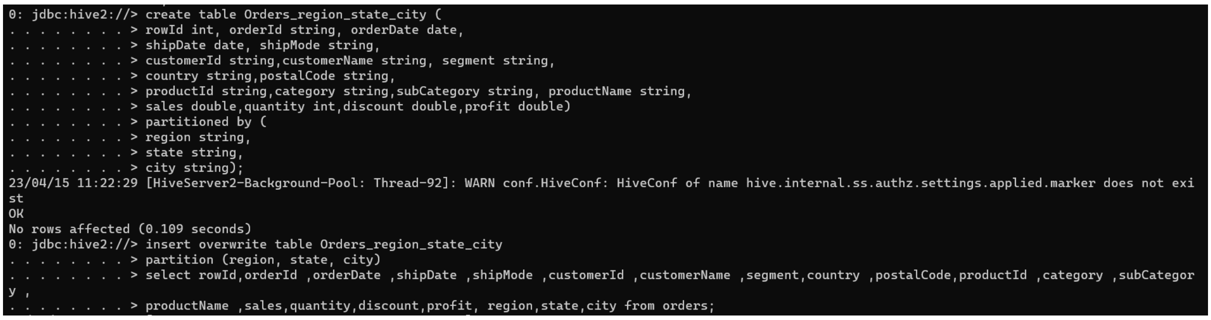


/user/hive/warehouse/salesdataanalysis.db/orders_from_each_segment

Show 25 entries Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:19	0	0 B	segment=Consumer
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:19	0	0 B	segment=Corporate
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:19	0	0 B	segment=Home Office

2. Creating the table orders_region_state_city (basically i am partitioning the table based on region than state and then city present in the original table orders)

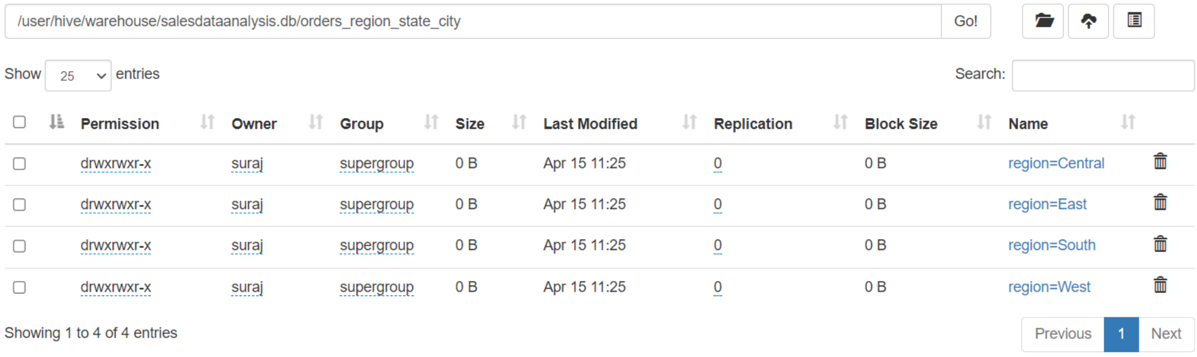


```
0: jdbc:hive2://> create table Orders_region_state_city (
. . . . . . . . > rowId int, orderId string, orderDate date,
. . . . . . . . > shipDate date, shipMode string,
. . . . . . . . > customerId string,customerName string, segment string,
. . . . . . . . > country string,postalCode string,
. . . . . . . . > productId string,category string,subCategory string, productName string,
. . . . . . . . > sales double,quantity int,discount double,profit double)
. . . . . . . . > partitioned by (
. . . . . . . . > region string,
. . . . . . . . > state string,
. . . . . . . . > city string);
23/04/15 11:22:29 [HiveServer2-Background-Pool: Thread-92]: WARN conf.HiveConf: HiveConf of name hive.internal.ss.authz.settings.applied.marker does not exi
st
OK
No rows affected (0.109 seconds)
0: jdbc:hive2://> insert overwrite table Orders_region_state_city
. . . . . . . . > partition (region, state, city)
. . . . . . . . > select rowId,orderId ,orderDate ,shipDate ,shipMode ,customerId ,customerName ,segment,country ,postalCode,productId ,category ,subCategor
y ,
. . . . . . . . > productName ,sales,quantity,discount,profit, region,state,city from orders;
```

When you click on the below table present in the hdfs

☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	orders_region_state_city

You will get this



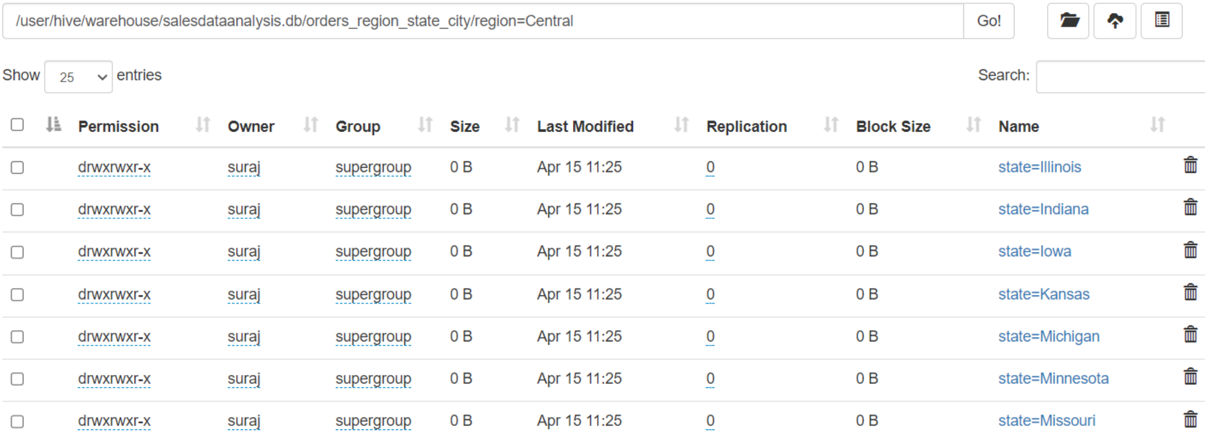
/user/hive/warehouse/salesdataanalysis.db/orders_region_state_city

Show 25 entries Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	region=Central
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	region=East
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	region=South
☐	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	region=West

Showing 1 to 4 of 4 entries

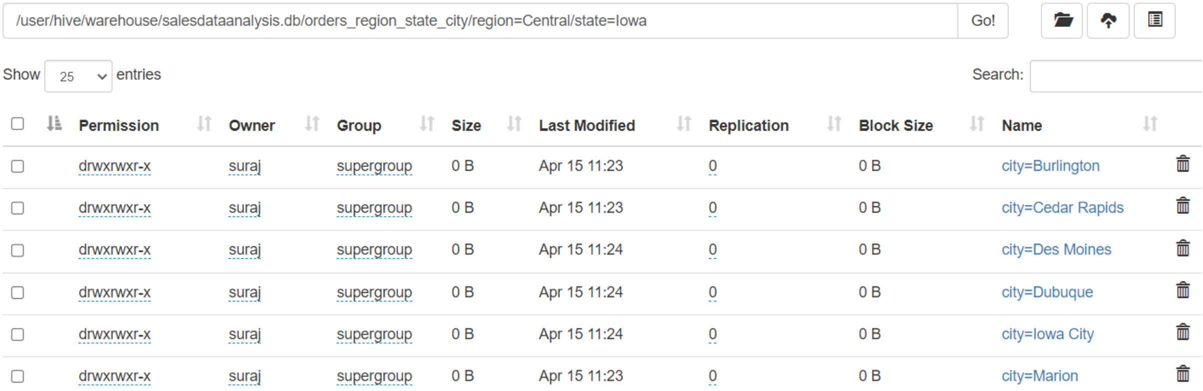
Here there four regions present in the table (so we have the table partitioned by region) Now when you click on any region let say Central , you will find the partitions by state inside the region



/user/hive/warehouse/salesdataanalysis.db/orders_region_state_city/region=Central
Go!
Show 25 entries
Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Illinois	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Indiana	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Iowa	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Kansas	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Michigan	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Minnesota	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	state=Missouri	🗑

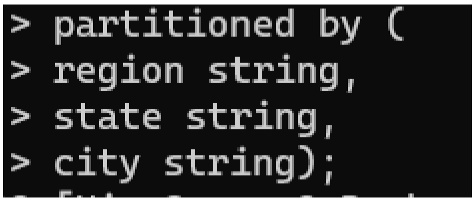
Now when you click on any of the state let say Iowa you will find the city partitions inside that state



/user/hive/warehouse/salesdataanalysis.db/orders_region_state_city/region=Central/state=Iowa
Go!
Show 25 entries
Search:

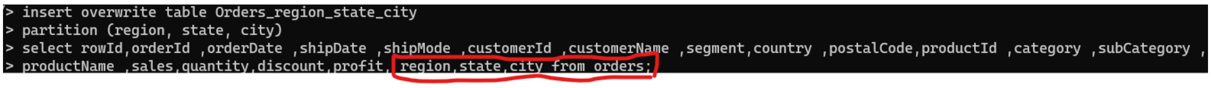
□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:23	0	0 B	city=Burlington	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:23	0	0 B	city=Cedar Rapids	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:24	0	0 B	city=Des Moines	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:24	0	0 B	city=Dubuque	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:24	0	0 B	city=Iowa City	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:23	0	0 B	city=Marion	🗑

Note: while creating partition inside a partition the order of columns putting in the partition is really important
region->state->city



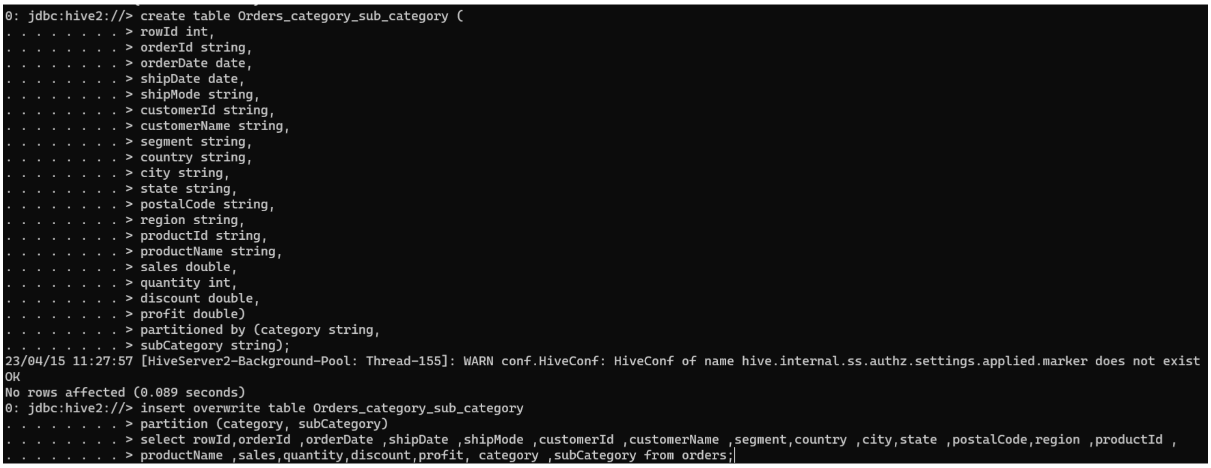
```
> partitioned by (
> region string,
> state string,
> city string);
```

Also note that the column on which you are creating partitions should be listed at the last while inserting the data into the table



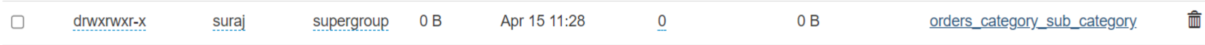
```
> insert overwrite table Orders_region_state_city
> partition (region, state, city)
> select rowId,orderId ,orderDate ,shipDate ,shipMode ,customerId ,customerName ,segment,country ,postalCode,productId ,category ,subCategory ,
> productName ,sales,quantity,discount,profit, region,state,city from orders;
```

3. Creating the table orders_region_state_city (basically i am partitioning the table on category and then on subcategory present in the original table orders)



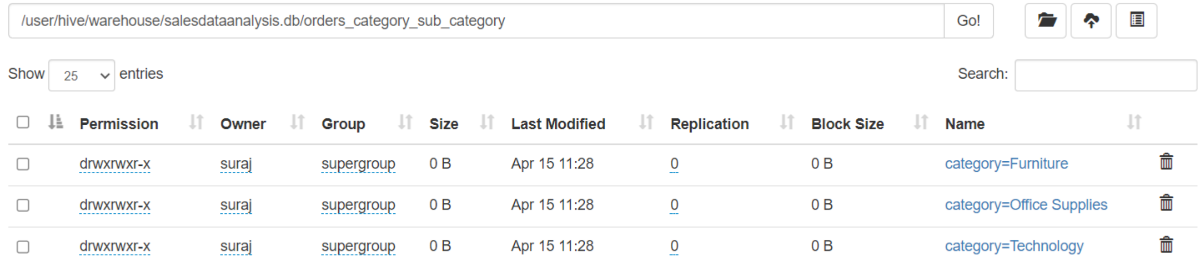
```
0: jdbc:hive2://> create table Orders_category_sub_category (
. . . . . . . . > rowId int,
. . . . . . . . > orderId string,
. . . . . . . . > orderDate date,
. . . . . . . . > shipDate date,
. . . . . . . . > shipMode string,
. . . . . . . . > customerId string,
. . . . . . . . > customerName string,
. . . . . . . . > segment string,
. . . . . . . . > country string,
. . . . . . . . > city string,
. . . . . . . . > state string,
. . . . . . . . > postalCode string,
. . . . . . . . > region string,
. . . . . . . . > productId string,
. . . . . . . . > productName string,
. . . . . . . . > sales double,
. . . . . . . . > quantity int,
. . . . . . . . > discount double,
. . . . . . . . > profit double)
. . . . . . . . > partitioned by (category string,
. . . . . . . . > subCategory string);
23/04/15 11:27:57 [HiveServer2-Background-Pool: Thread-155]: WARN conf.HiveConf: HiveConf of name hive.internal.ss.authz.settings.applied.marker does not exist
OK
No rows affected (0.089 seconds)
0: jdbc:hive2://> insert overwrite table Orders_category_sub_category
. . . . . . . . > partition (category, subCategory)
. . . . . . . . > select rowId,orderId ,orderDate ,shipDate ,shipMode ,customerId ,customerName ,segment,country ,city,state ,postalCode,region ,productId ,
. . . . . . . . > productName ,sales,quantity,discount,profit, category ,subCategory from orders;
```

When you click on the below table



□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	orders_category_sub_category	🗑

You will find partition on category

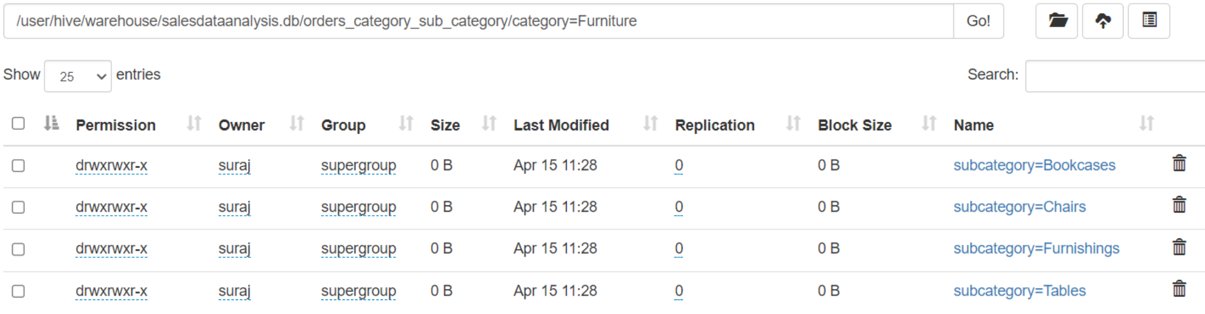


/user/hive/warehouse/salesdataanalysis.db/orders_category_sub_category Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	category=Furniture	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	category=Office Supplies	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	category=Technology	🗑

And when you click on any of the above partition let say furniture you will find subcategory based partition of furniture partition

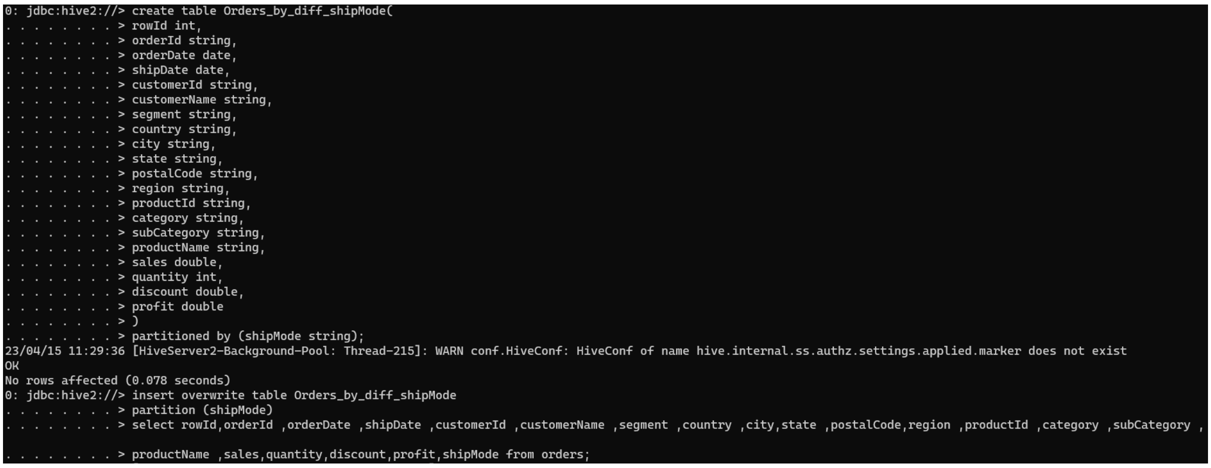


/user/hive/warehouse/salesdataanalysis.db/orders_category_sub_category/category=Furniture Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	subcategory=Bookcases	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	subcategory=Chairs	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	subcategory=Furnishings	🗑
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	subcategory=Tables	🗑

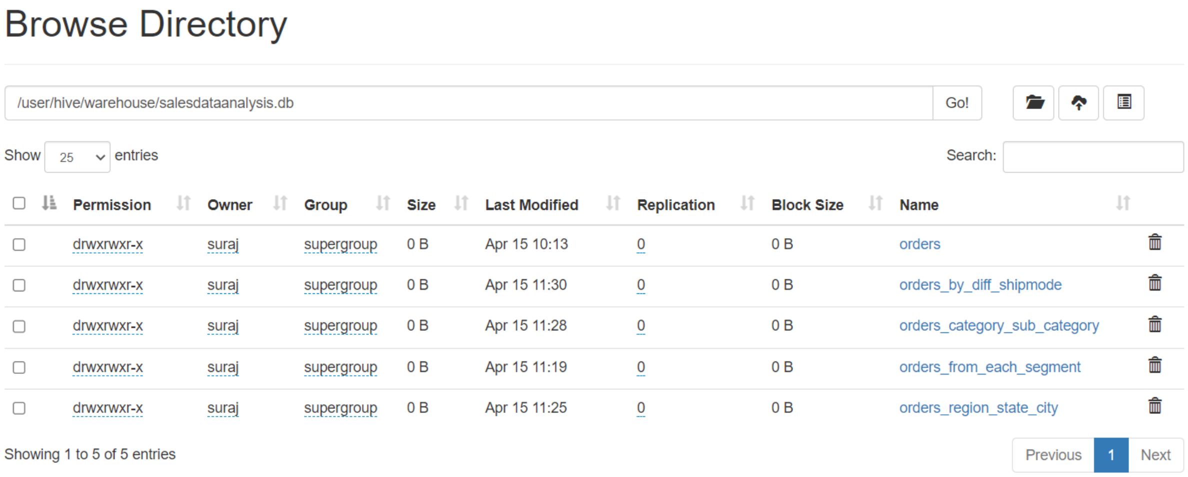
4. Creating the table orders_region_state_city (basically i am partitioning the table based on type of shipping mode present in the original table orders)



```
0: jdbc:hive2://> create table Orders_by_diff_shipMode(
. . . . . . . . > rowId int,
. . . . . . . . > orderId string,
. . . . . . . . > orderDate date,
. . . . . . . . > shipDate date,
. . . . . . . . > customerId string,
. . . . . . . . > customerName string,
. . . . . . . . > segment string,
. . . . . . . . > country string,
. . . . . . . . > city string,
. . . . . . . . > state string,
. . . . . . . . > postalCode string,
. . . . . . . . > region string,
. . . . . . . . > productId string,
. . . . . . . . > category string,
. . . . . . . . > subCategory string,
. . . . . . . . > productName string,
. . . . . . . . > sales double,
. . . . . . . . > quantity int,
. . . . . . . . > discount double,
. . . . . . . . > profit double
. . . . . . . . > )
. . . . . . . . > partitioned by (shipMode string);
23/04/15 11:29:36 [HiveServer2-Background-Pool: Thread-215]: WARN conf.HiveConf: HiveConf of name hive.internal.ss.authz.settings.applied.marker does not exist
OK
No rows affected (0.078 seconds)
0: jdbc:hive2://> insert overwrite table Orders_by_diff_shipMode
. . . . . . . . > partition (shipMode)
. . . . . . . . > select rowId,orderId ,orderDate ,shipDate ,customerId ,customerName ,segment ,country ,city,state ,postalCode,region ,productId ,category ,subCategory ,

. . . . . . . . > productName ,sales,quantity,discount,profit,shipMode from orders;
```

Final view of tables



Browse Directory

/user/hive/warehouse/salesdataanalysis.db

Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 10:13	0	0 B	orders
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:30	0	0 B	orders_by_diff_shipmode
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:28	0	0 B	orders_category_sub_category
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:19	0	0 B	orders_from_each_segment
□	drwxrwxr-x	suraj	supergroup	0 B	Apr 15 11:25	0	0 B	orders_region_state_city

Showing 1 to 5 of 5 entries