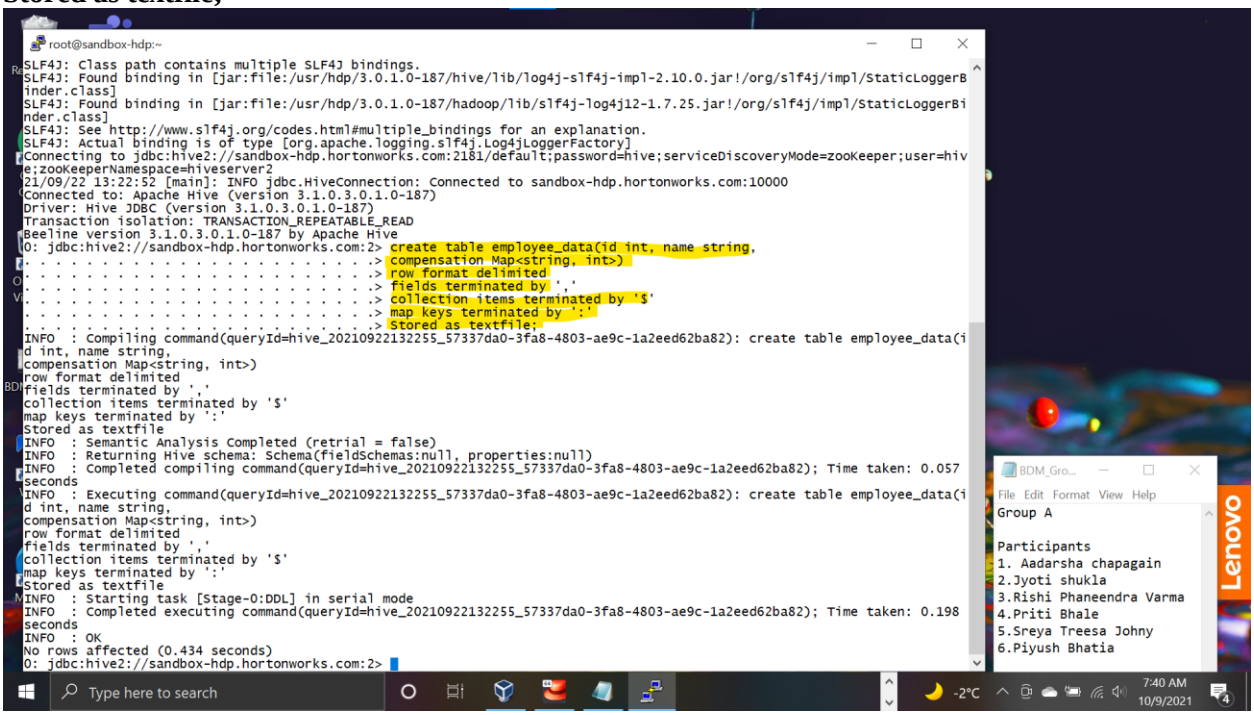
Group A
Lab 2c
Hive part 2 and file format.

Participants:
1. Aadarsha chapagain
2.Jyoti shukla
3.Rishi Phaneendra Varma
4.Priti Bhale
5.Sreya Treesa Johny
6.Piyush Bhatia

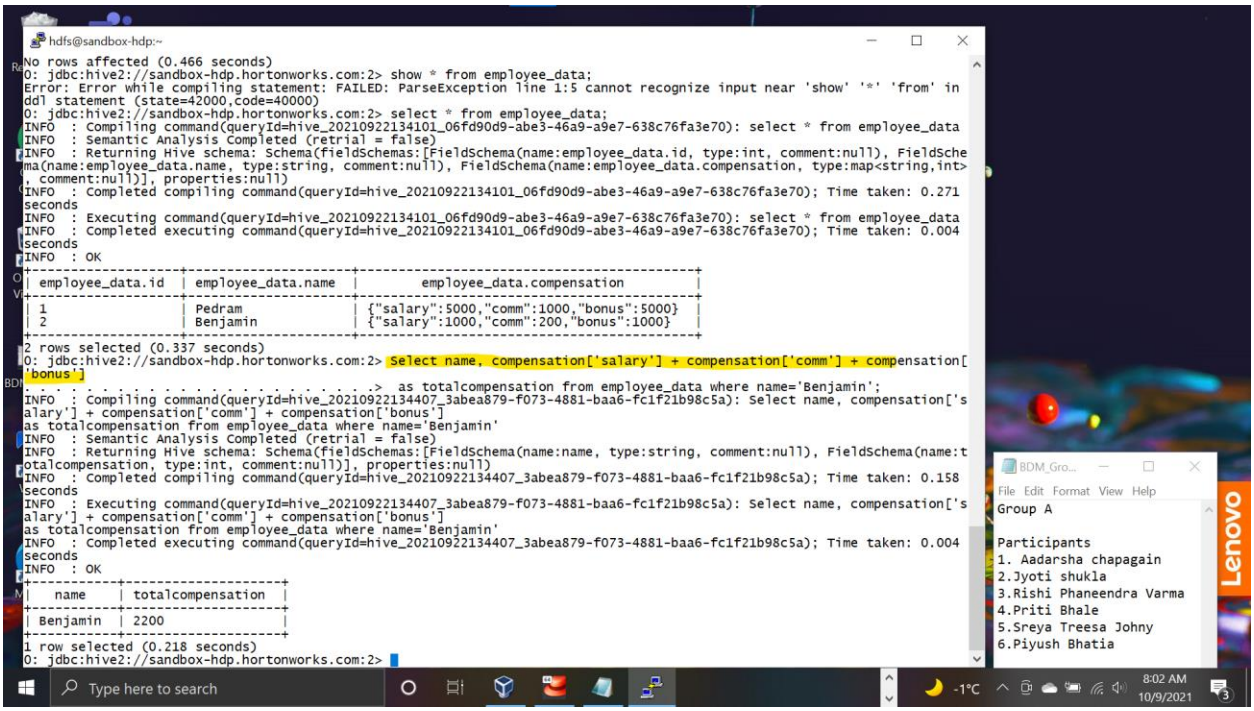
Slide 2:
create table employee_data(id int, name string,
compensation Map<string, int>)
row format delimited
fields terminated by ','
collection items terminated by '$'
map keys terminated by ':'
Stored as textfile;



Slide 7:

Select name, compensation['salary'] + compensation['comm'] + compensation['bonus']

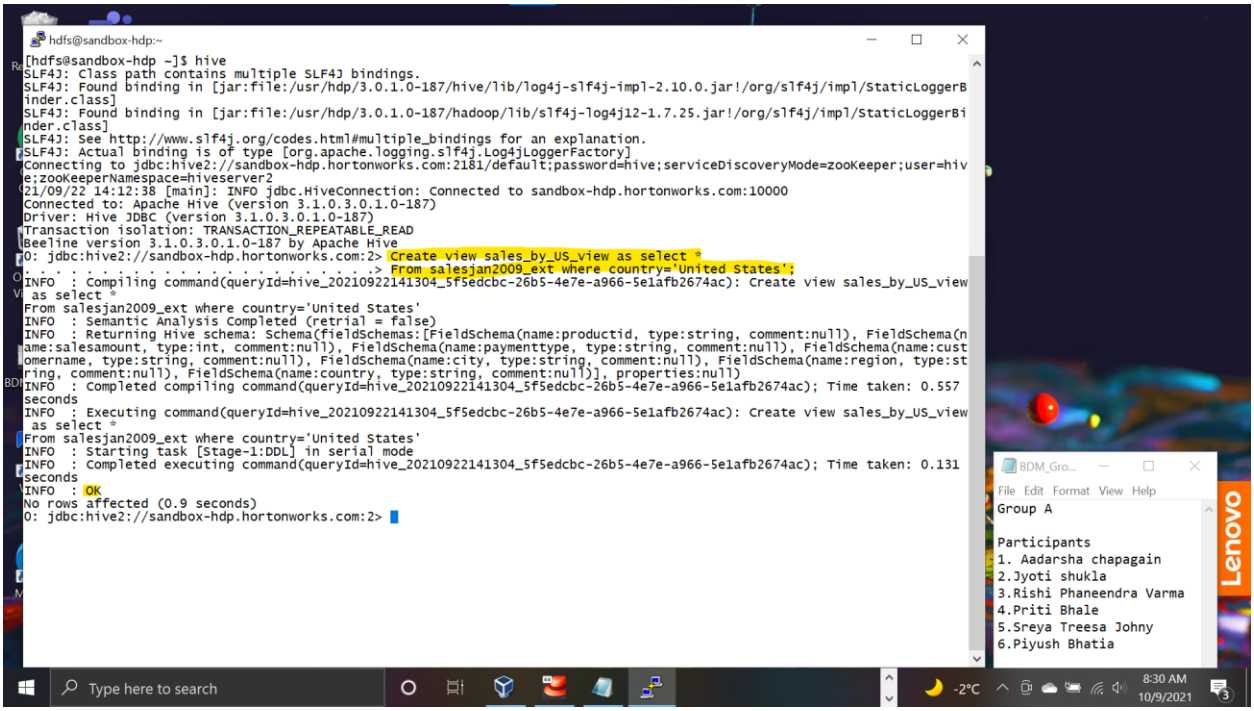
as totalcompensation from employee_data where name='Benjamin';



Slide 8:

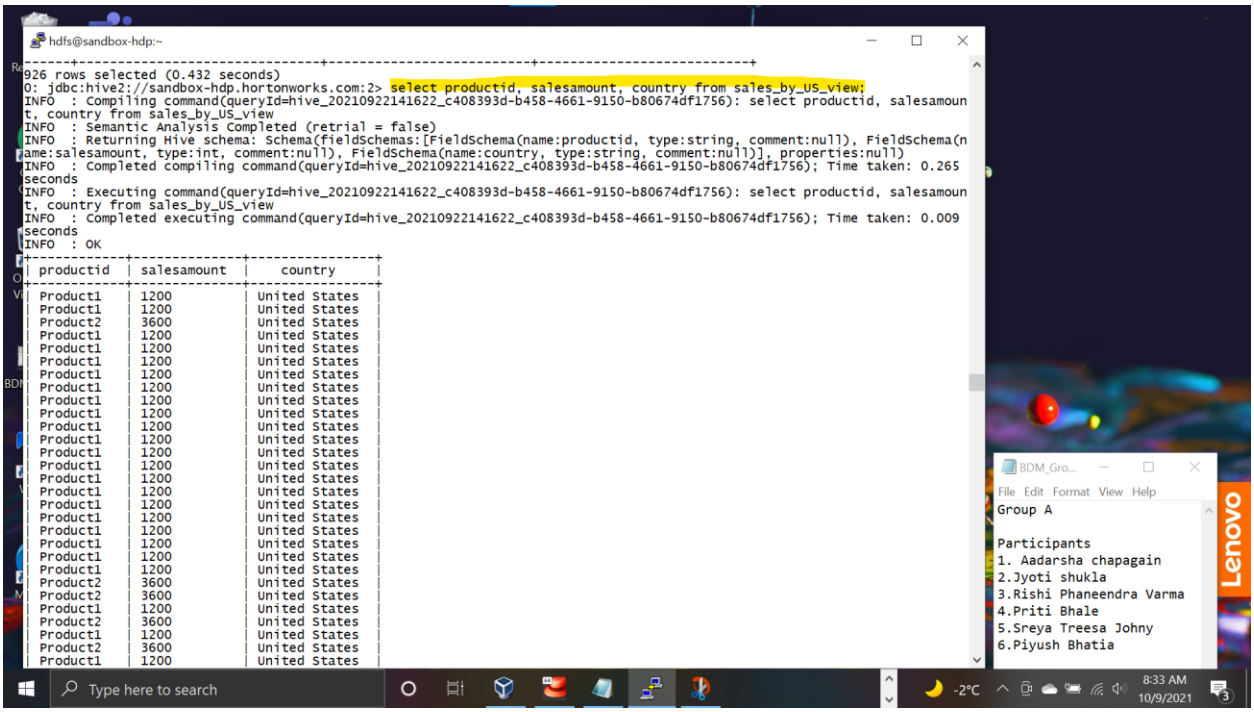
Create view sales_by_US_view as select *

From salesjan2009_ext where country='United States';



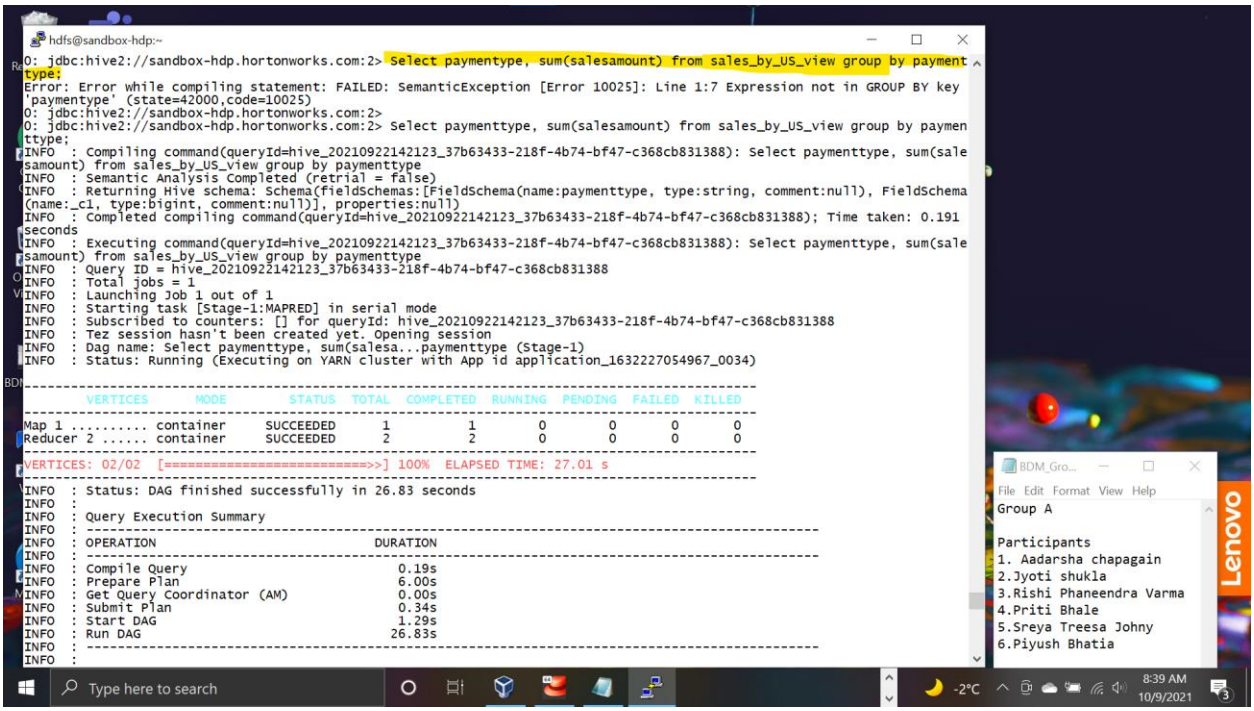
Slide 9:

Select productid, salesamount, country from sales_by_US_view.



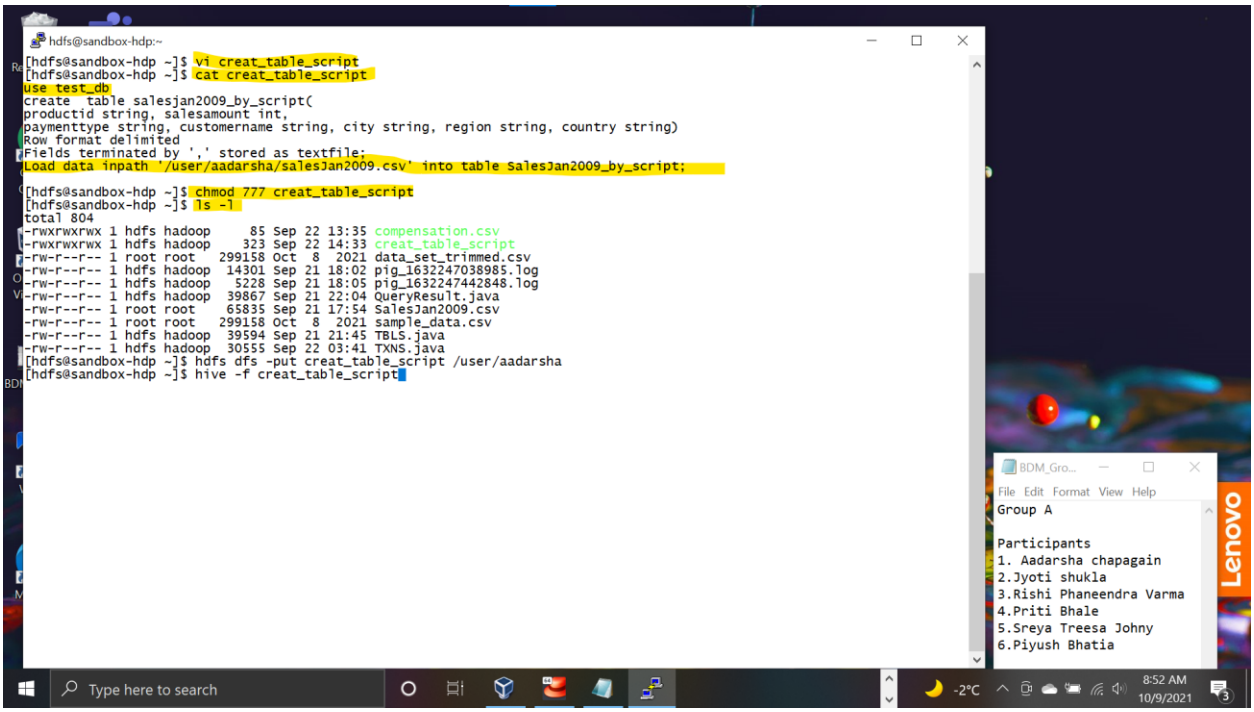
Slide 10:

Select paymenttype, sum(salesamount) from sales_by_US_view group by paymenttype



Slide 11:

Create a script to create a table and change permission and run via hive -f



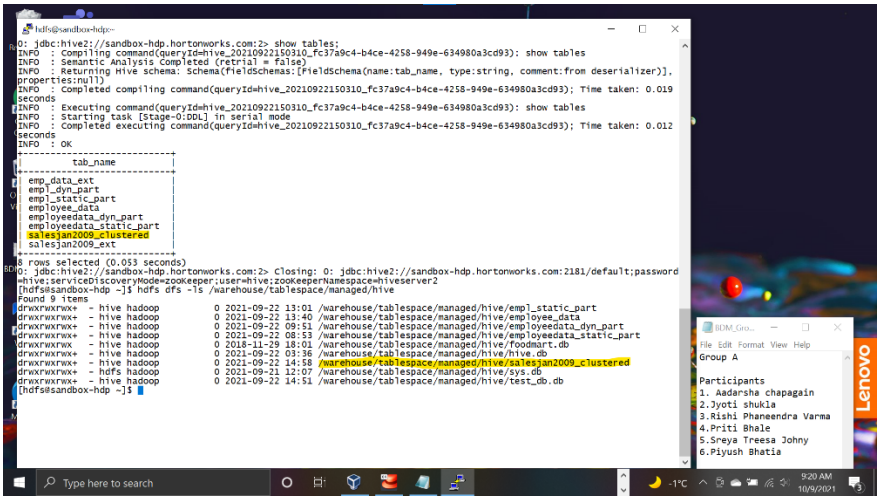
Slide 26:

create table salesjan2009_clustered(

productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)

clustered by (country) into 10 buckets

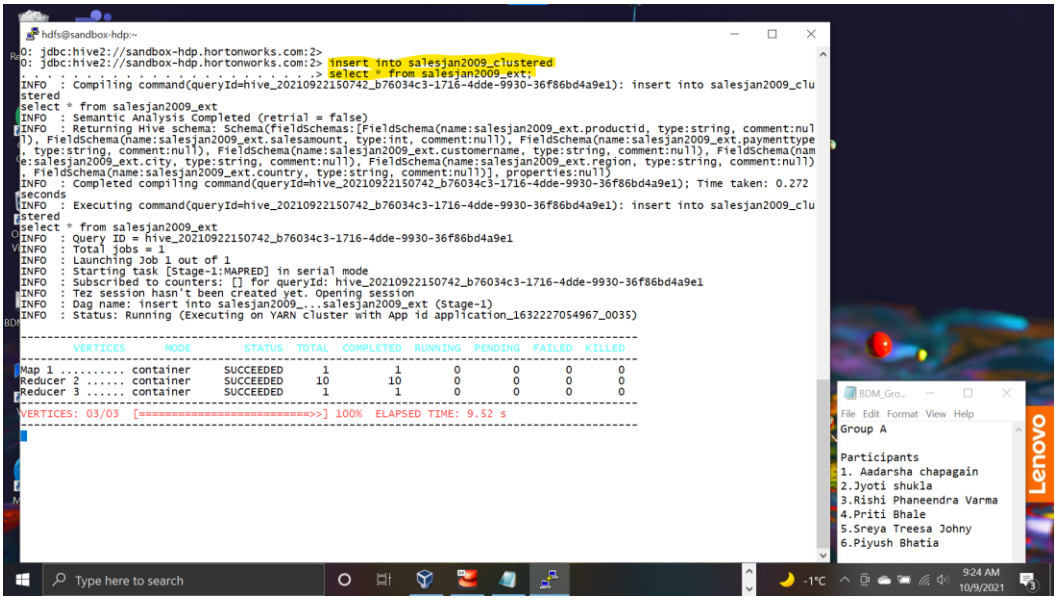
stored as ORCFILE



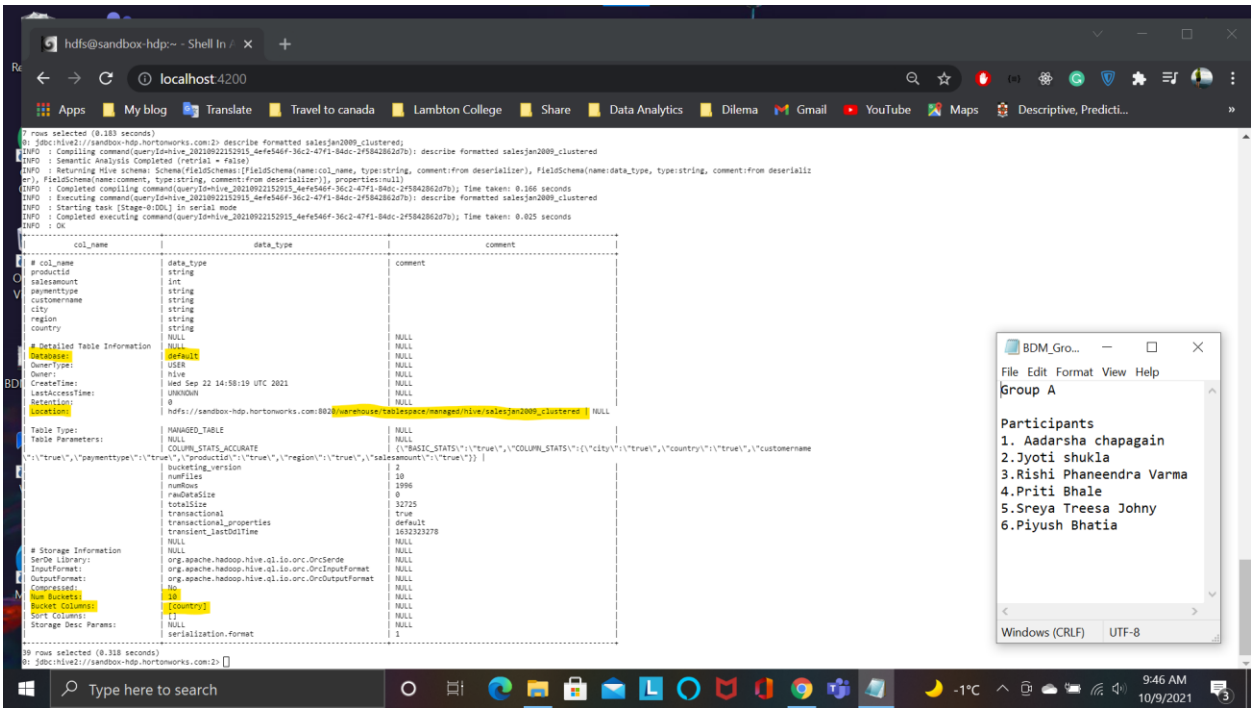
Slide 27:

insert into salesjan2009_clustered

select * from salesjan2009_ext;

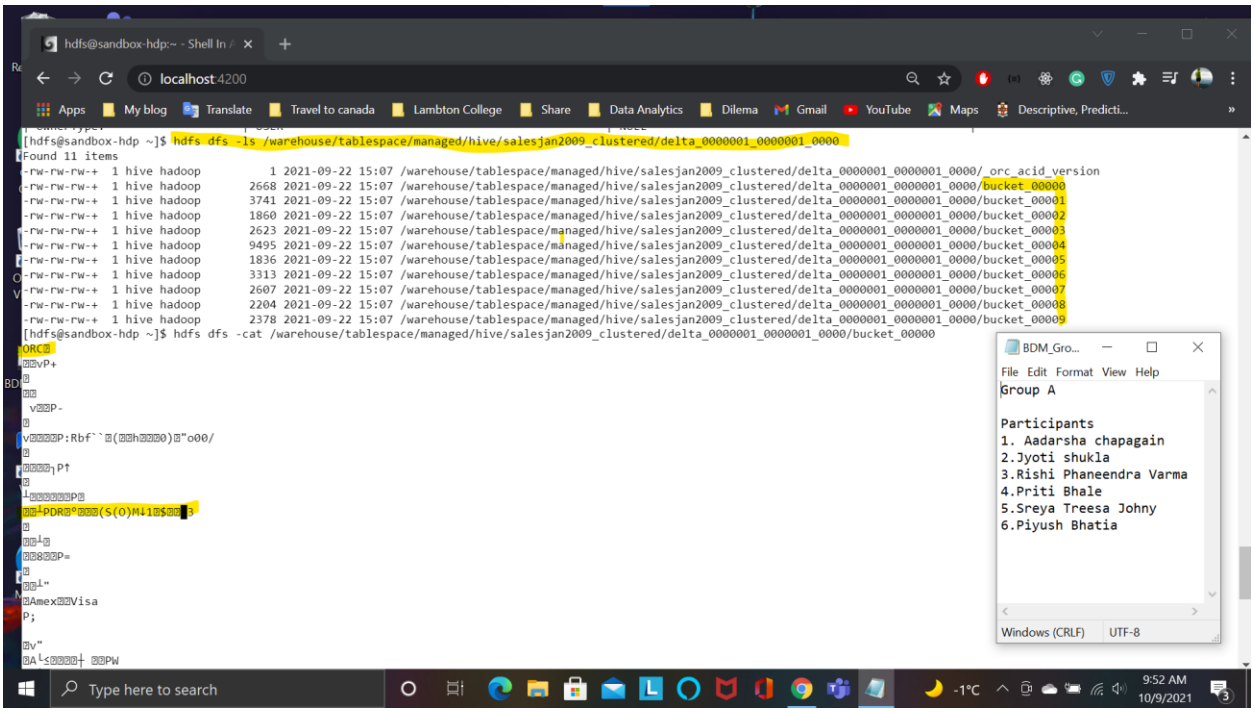


Slide 28:



Slide 29:

10 buckets are created and orc file is not in human readable form.



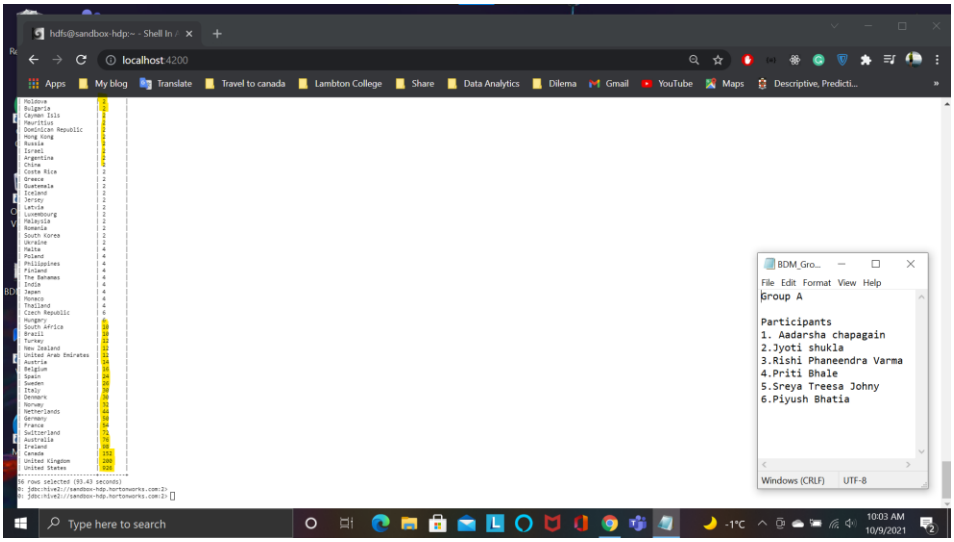
Slide 32:

select country, count(*) as count from SalesJan2009_ext

group by (country)

order by count;

Here we can see that the data is skewed and do not have normal distribution.



Slide 33: create table salesjan2009_skewed(

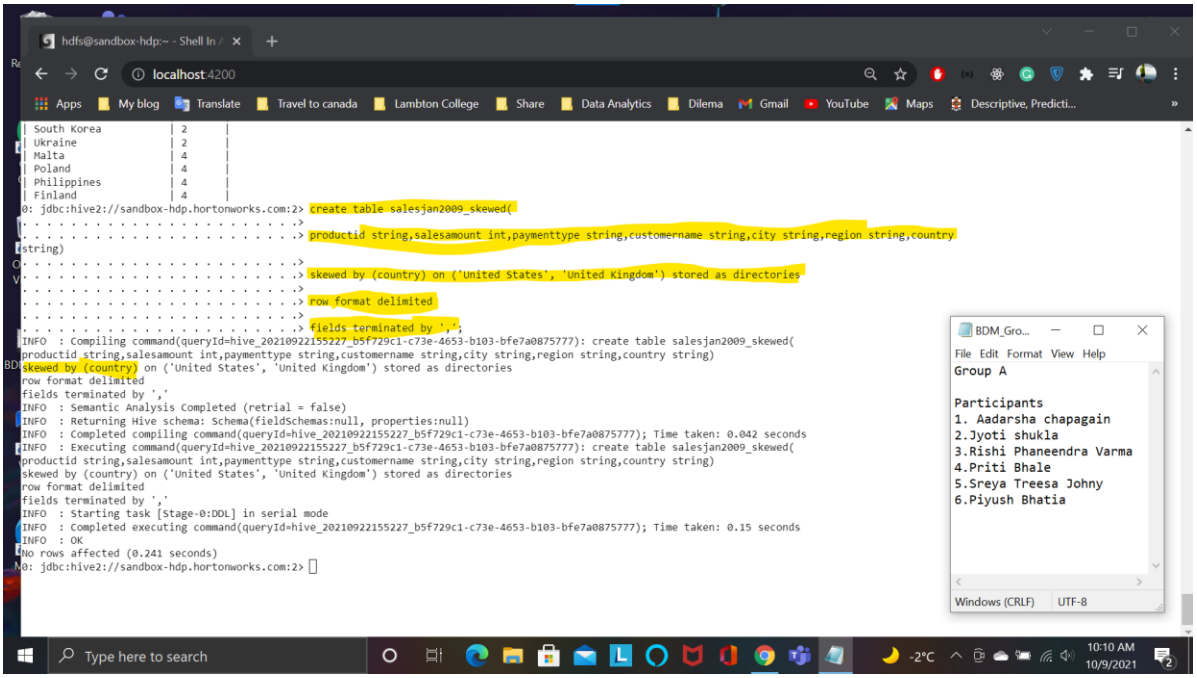
productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)

skewed by (country) on ('United States', 'United Kingdom') stored as directories

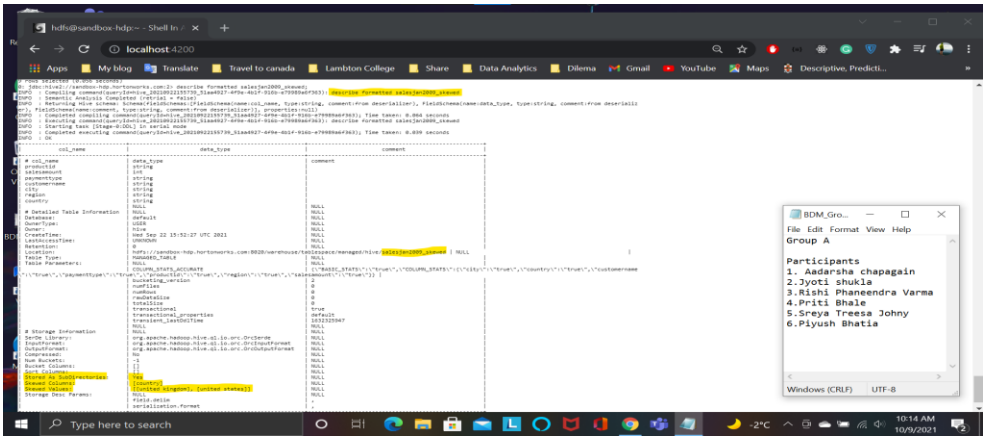
row format delimited

fields terminated by ',';

We need to specify by which column and value for skewed table;

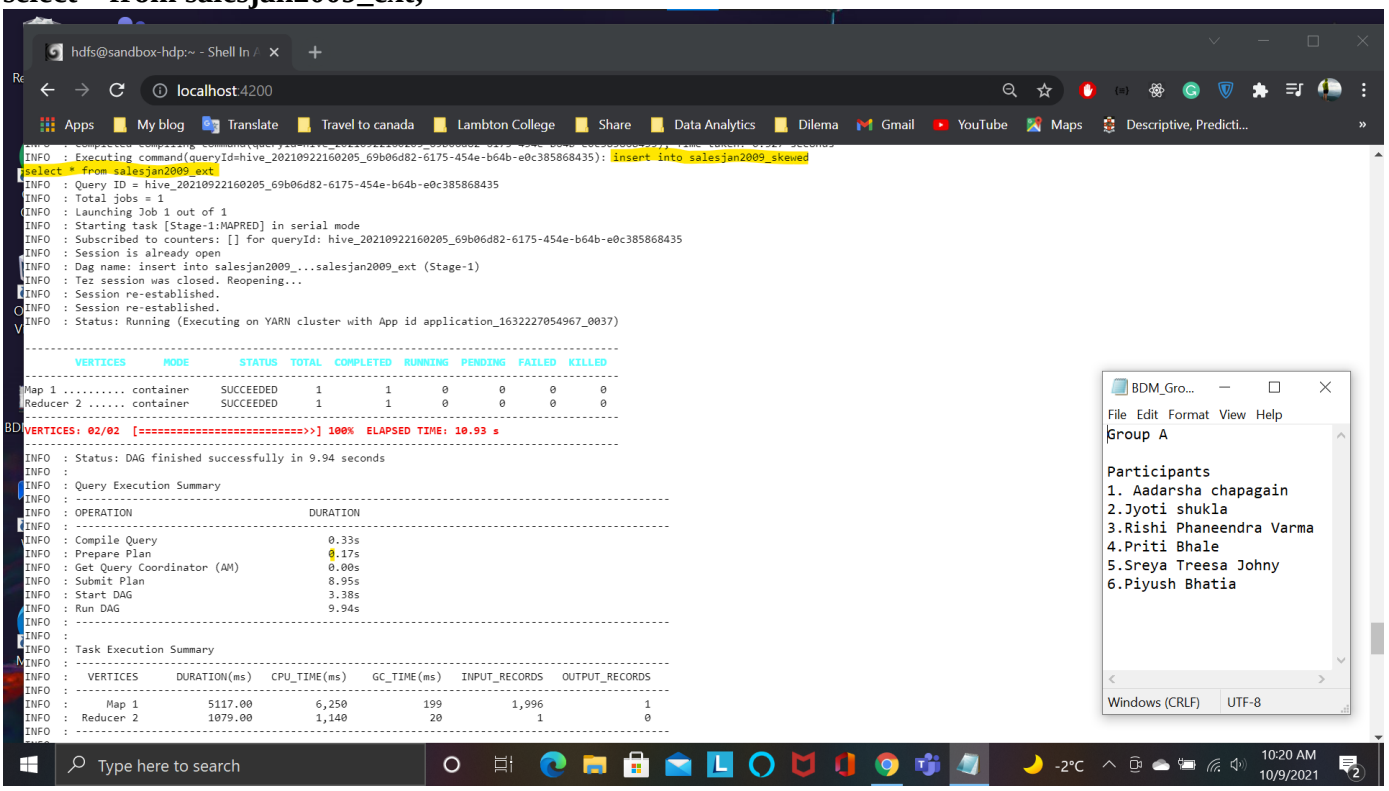


Slide 34: Skewed values are stored as sub directories



Slide 35:

insert into salesjan2009_skewed
select * from salesjan2009_ext;



Slide 89: create table salesjan2009_parquet1(

productid string,salesamount int,paymenttype string,customername string,city string,region string,country string)

stored as PARQUET

TBLPROPERTIES ("parquet.compress"="snappy");

Storing file as Parquet format. It is columnar file format and data are partitioned by columns.