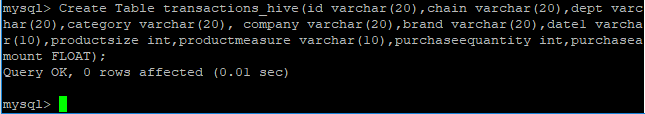
**Project 1**

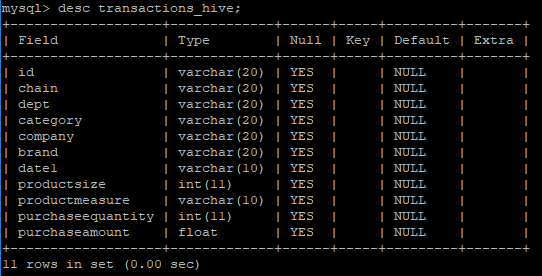
**Retail Data Analysis with Apache Hive**

* Retail stores daily generate millions of transactions logs.
* Analyzing these logs would generate beautiful insights and improve business.
* Storing these logs on traditional databases would be costly and scalability will be a big challenge
* Stores like walmart are spread across different locations.
* Daily millions of customers visit these stores and generate billions of logs.
* This billions of logs contribute to huge volume of data.
* Having Huge Volume, High velocity and variety will make this data Big Data.
* Challenges:
* Storage
* Scalability
* Processing
* Sharing
* What is our Goal
  + We will process data and find following.
    - Demand of a given product
    - Trend and seasonality of sales
    - Understand performance of chain
    - Loyal Customer identification
* We have a Transactional data which we will process using hive
* First, we have to load data into mysql

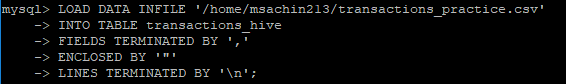
Step 1: Create table in mysql



Here We have created Table transactions\_hive



Step 2: Load Data into table



Step 3: Now Use Apache Sqoop for import data from Mysql to Into Hive



sqoop import --connect jdbc:mysql://101.53.130.146/msachin213 --username msachin213 -password 25Msj1s43Ctgz3a --table transactions\_hive --hive-import --hive-table transactions\_hive -m 1

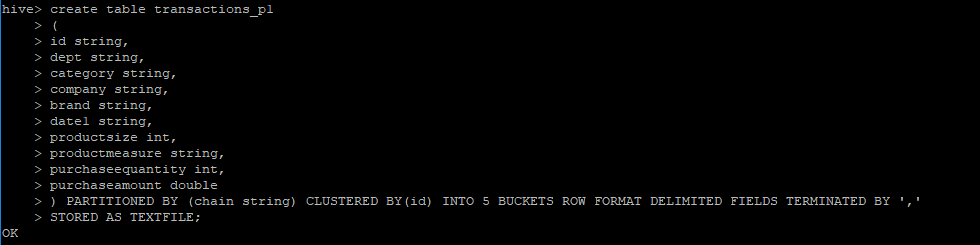
**This is the command for importing dat from mysql to hive using sqoop tool**

**Step 4: Partitioning and Bucketing**

* In normal sql database we have only one structure which is row and column factor in apache hive we can categories data by using partitioning and bucketing that’s why hive can access data faster than normal sql database.
* Partitioning and bucketing in hive will let you do faster querying.
* For dynamic partitioning, load the data in to staging table which is already done.
* Now create a production table, and insert data
* For enabling hive bucketing and partitioning have to write some command
* set hive.enforce.bucketing =true ;
* set hive.exec.dynamic.partition = true;
* set hive.exce.dynamic.partition.mode = nonstrict;

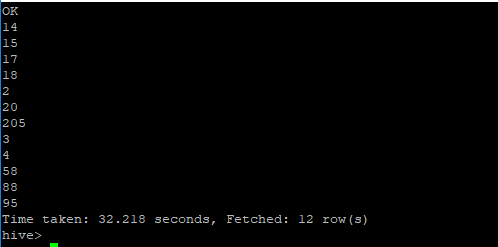
**Step 5: Create table with Partition and Bucket**

* When We imported data from mysql to hive using sqoop, schema of mysql and hive was same now we will create another table in hive with partition and bucket and then append data of first hive table.



* Here we created table with partition name of chain and with five buckets.
* Here we can see how many partitions has been created





**Step 6: Now Load Data Into transactions\_p1 from 1st hive table**

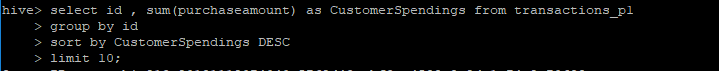


INSERT OVERWRITE TABLE transactions\_p1 PARTITION (chain)

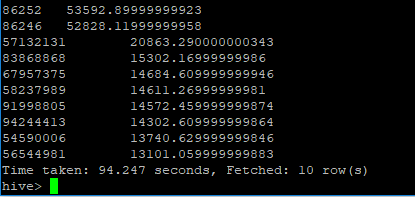
select id,dept,category,company,brand,date1,productsize,productmeasure,purchaseequantity,purchaseamount,chain from transactions\_hive;

**Step 7: Now all the things are set we have data into hive table in proper manner now we will fire query and get results from table**

* **Top 10 Customers**
* This is query for getting top 10 customers



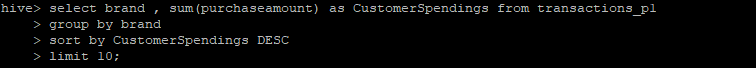
* Result



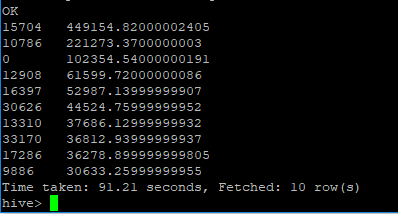
* **Chain(Partition) wise sales**
* This is query for getting Chain(Partition) wise sales



* Result
* Top 10 Brands
* This is query for getting Chain(Partition) wise sales

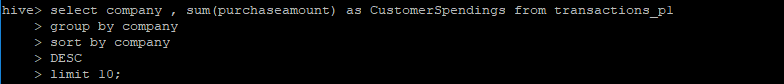


* Result

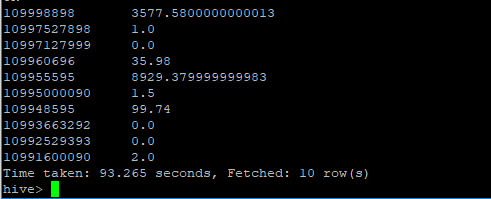


* Top 10 Company

* This is query for getting top 10 company



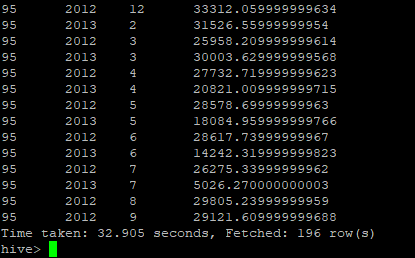
* Result



* Chain Year Monthly Sales
* This is query for getting Chain Year Monthly Sales



* Result



* We can store result in table

create table top10companies11

As

select company,sum(purchaseamount) as CustomerSpendings from transactions\_p1

group by company

sort by CustomerSpendings DESC ;