**Retail Data Analysis using Hive**

**Retail Transactions**

* 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.

**Volume of transactions**

* 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.

**Velocity of transactions**

* In peak hours, 1000’s of transactions will happen in any given second.
* 1000’s of trasactions/sec across all stores contribute to high velocity.

**Variety of transactions**

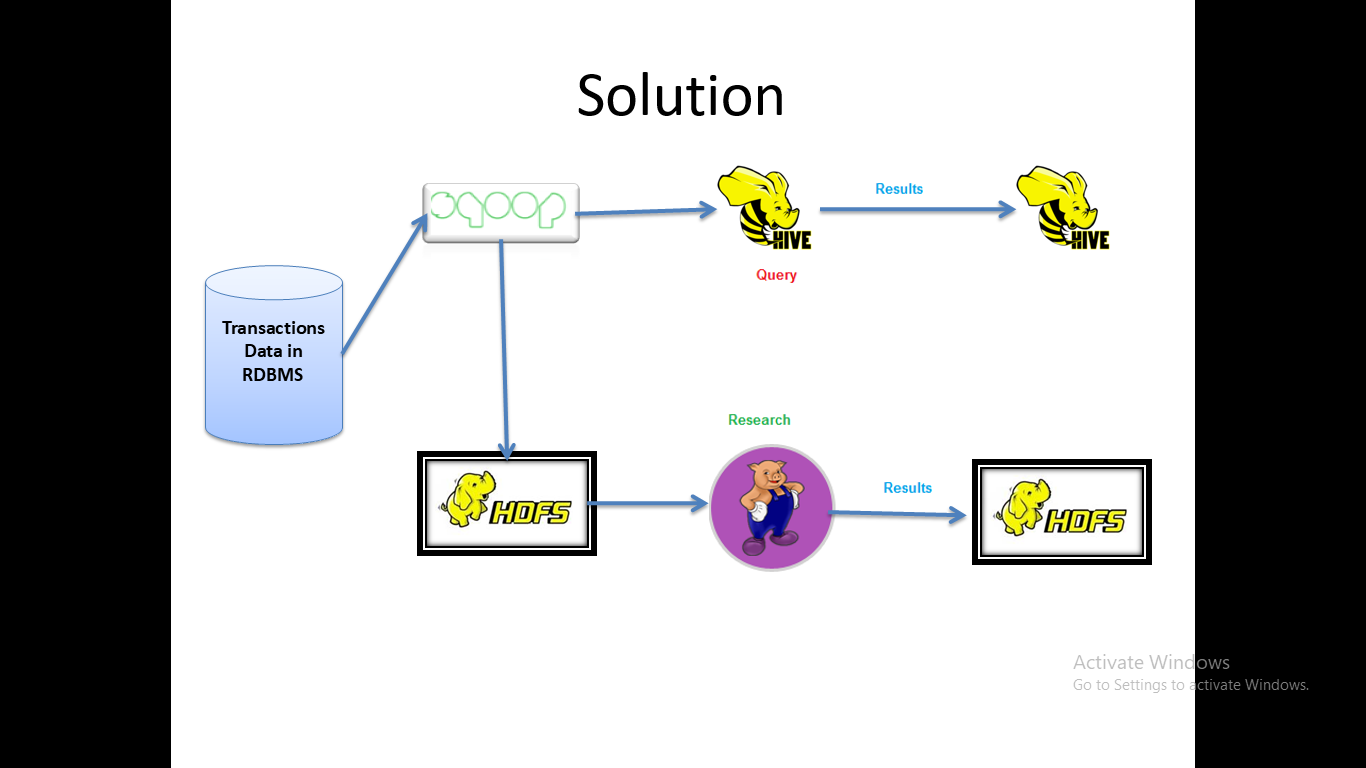
* The most widely known varities of data generated by transactions:
  + - Json Format
    - Xml Format
    - Csv Format

**Its Bigggg Data**

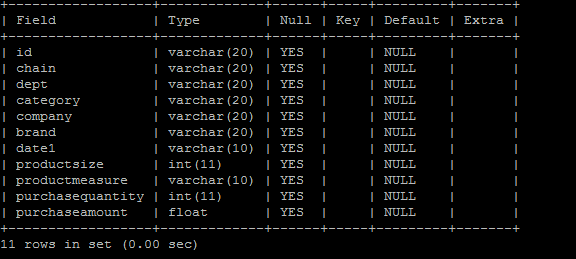
* Having Huge Volume, High velocity and variety will make this data Big Data.
* Challenges:
  + Storage
  + Scalability
  + Processing
  + Sharing

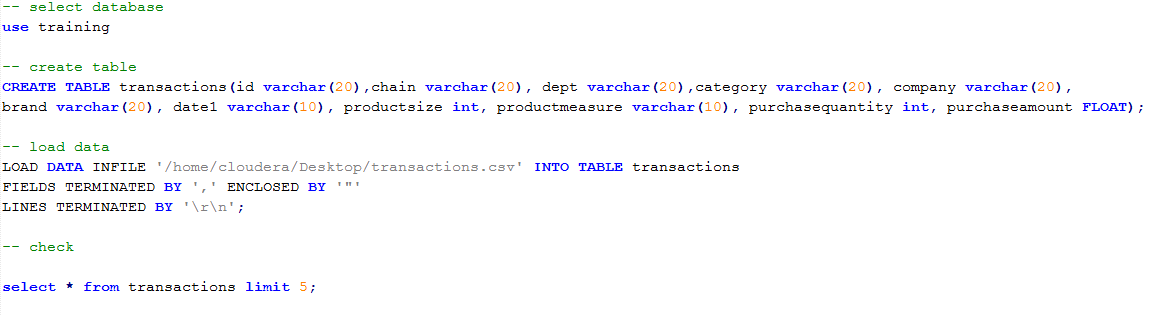
**Goal**

* Process these logs over night as a batch to understand:
  + Demand of a given product.
  + Trend and seasonality of sales.
  + Understand performance of chain.
  + Loyal Customer identification.



**Transactions dataset**



**Load data into MySQL**

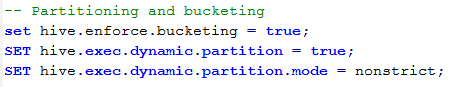
**Retail Data Analysis with Hive**

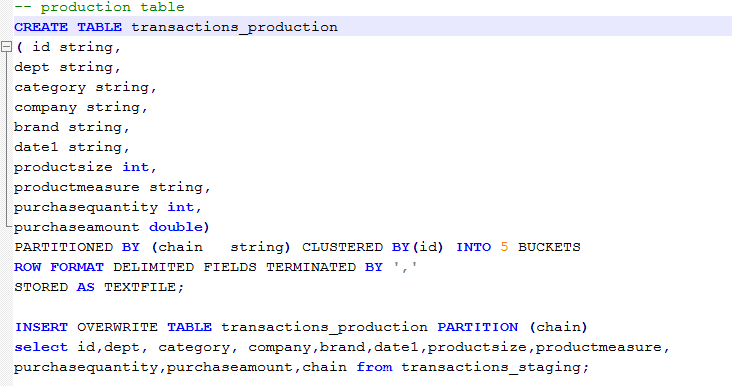
****

**Partioning and bucketing**

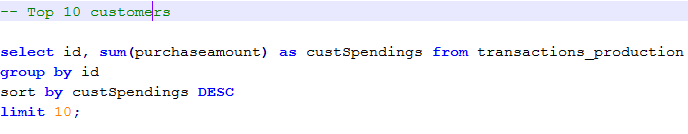
* Partitioning and bucketing in hive will let you do faster querying.
* For dynamic partitioning, load the data in to staging table which is already done.
* Now create a production table, and insert data.

**Open hive and set below properties**

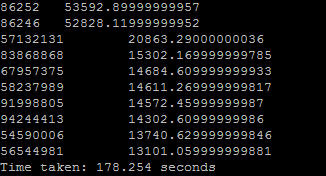
****

**Create production table and load data**

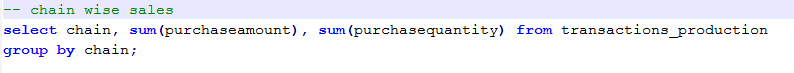
**Top 10 customers**

****

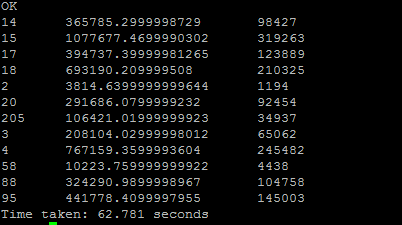
**Snapshot of output**

****

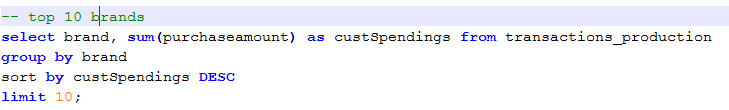
**Chain wise sales**

****

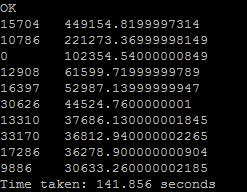
**Snapshot of output**

****

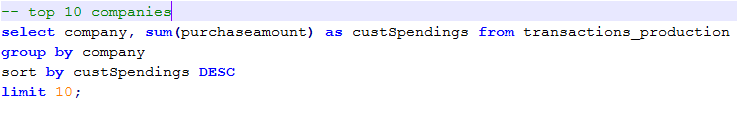
**Top 10 brands**

****

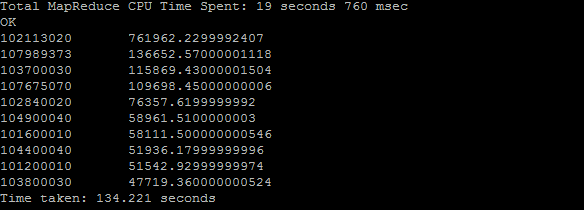
**Snapshot of output**

****

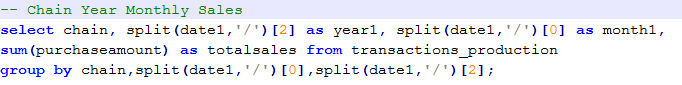
**Top 10 companies**

****

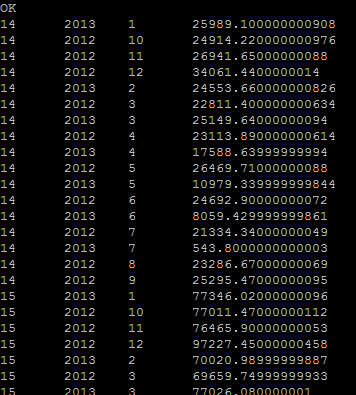
**Snapshot of output**

****

**Chain Year Monthly Sales**

****

**Snapshot of output**

****

**Storing results**

* To store results

