**ACKNOWLEDGEMENT**

We express our humble gratitude to **Dr C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support. We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr T.V.Gopal**, for his invaluable support.

We wish to thank **Dr Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work. We are incredibly grateful to our Head of the Department**, Dr M. Lakshmi** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work. We want to convey our thanks to our program coordinators **Dr. Vadivu G**, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for their inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor **Dr. Shantha Kumari K**, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr SWARNALAKSHMI** Assistant Professor Department of Data Science and Business Systems SRM Institute of Science and Technology, for providing me with an opportunity to pursue my project under his/her/their mentorship. He provided me with the freedom and support to explore the research topics of my interest. Her/His/Their passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement

SHAIK YASIR TAWFIQ MUGESH RAJ GODFREY ASHWANTH

DATASET

**This is a Product Sales Dataset scraped from the Amazon website** which is download from [Kaggle](https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset?select=Amazon-Products.csv).

* Its product data are separated by 142 categories in csv format, along with the full dataset name **Amazon-Products.csv**.
* Each csv files are consisting of 7 columns and each row has products details accordingly

**Features**

| **name** | **description** |
| --- | --- |
| *name* | The name of the product |
| *main\_category* | The main category of the product belong |
| *sub\_category* | The main category of the product belong |
| *ratings* | The ratings given by amazon customers of the product |
| *no of ratings* | The number of ratings given to this product in amazon shopping |
| *discount\_price* | The discount prices of the product |
| *actual\_price* | The actual MRP of the product |

**Inspiration**

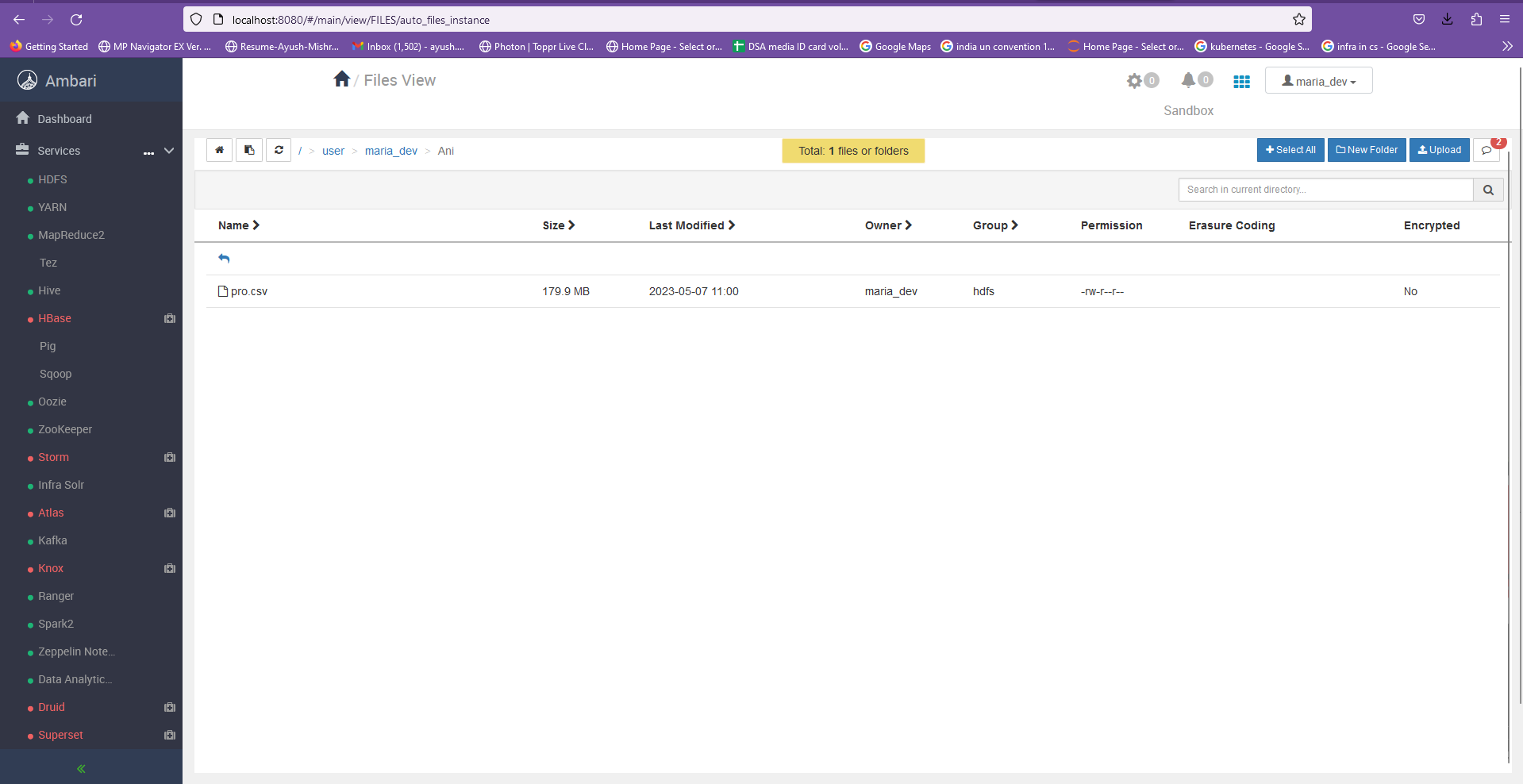
Amazon is an American Tech Multi-National Company whose business interests include E-commerce, where they buy and store the inventory, and take care of everything from shipping and pricing to customer service and returns. I've created this dataset so that people can play with this dataset and do a lot of things as mentioned below

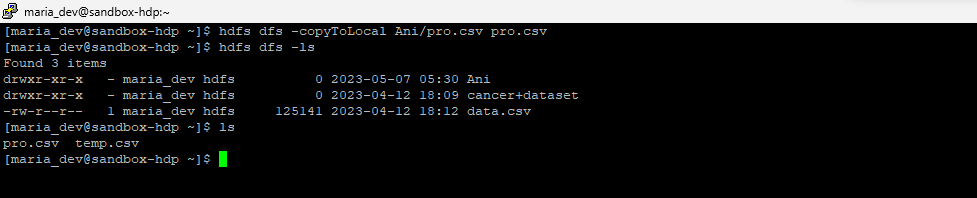
* Dataset Walkthrough
* Data Preprocessing
* Understanding Dataset Hierarchy
* Exploratory Data Analysis
* Data Visualization *Using matlabplot and searborn*
* Data Visualization *Using BI Tools such as****Tableau, PowerBI ,Kibana ,Grafana, Splunk***
* **Making Recommendation System**
* **Webscraping content of each products in detail**

This is a list of some of that things that you can do on this dataset. It's not definitely limited to the one that is mentioned there but a lot more other things can also be done.

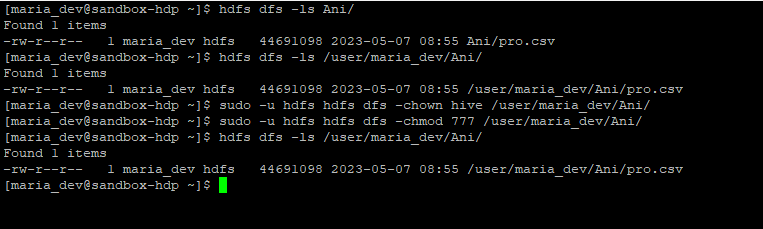
Importing File into HDFS

Uploading File into HDFS using Apache Ambari



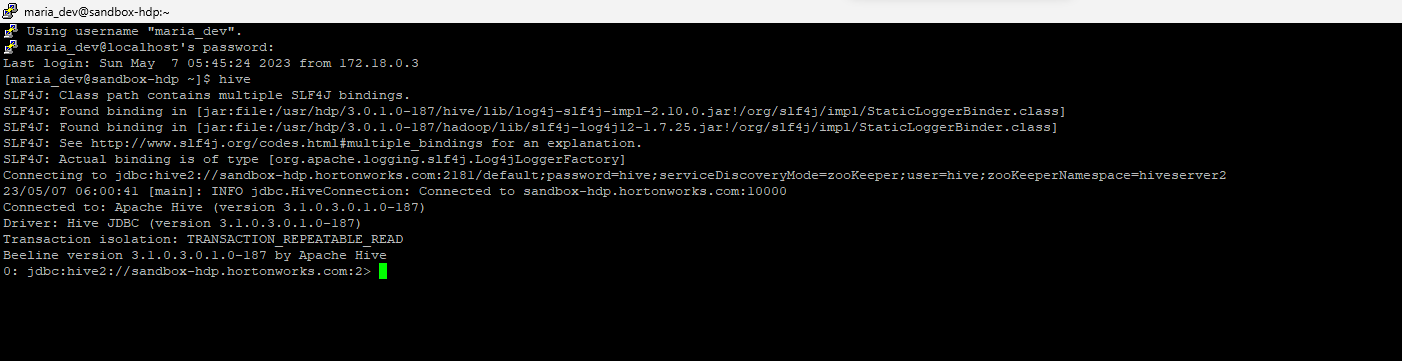


Changing ownership and permissions of data file



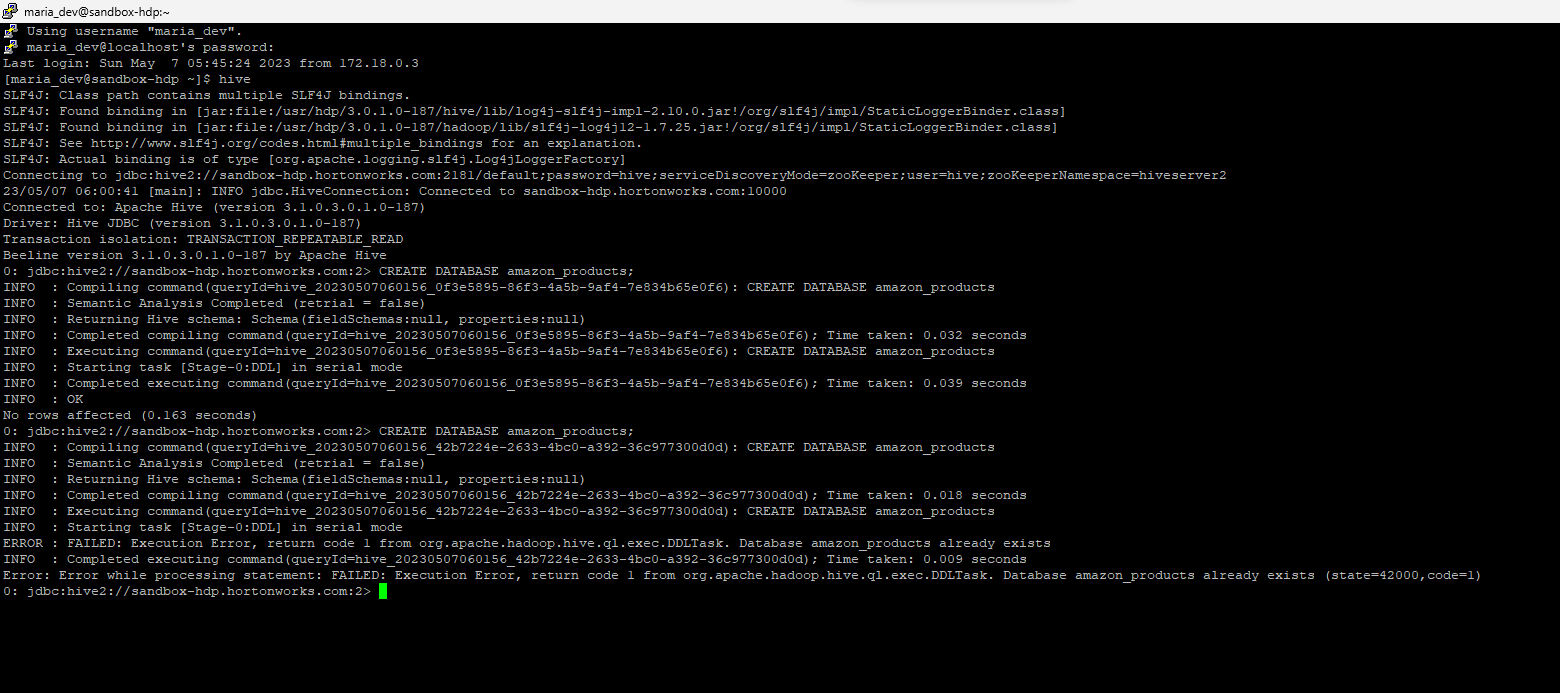
Hive Commands

Launching Hive from HortonWorks

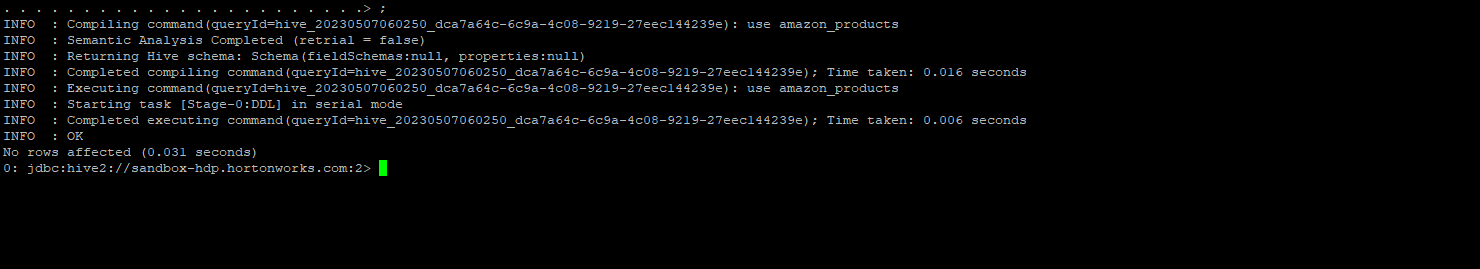


Creating a Database

* *create database* amazon\_products;



* *use amazon\_products*



Creating Table

*CREATE TABLE product\_sales (*

*name STRING,*

*main\_category STRING,*

*sub\_category STRING,*

*ratings FLOAT,*

*no\_of\_ratings INT,*

*discount\_price FLOAT,*

*actual\_price FLOAT*

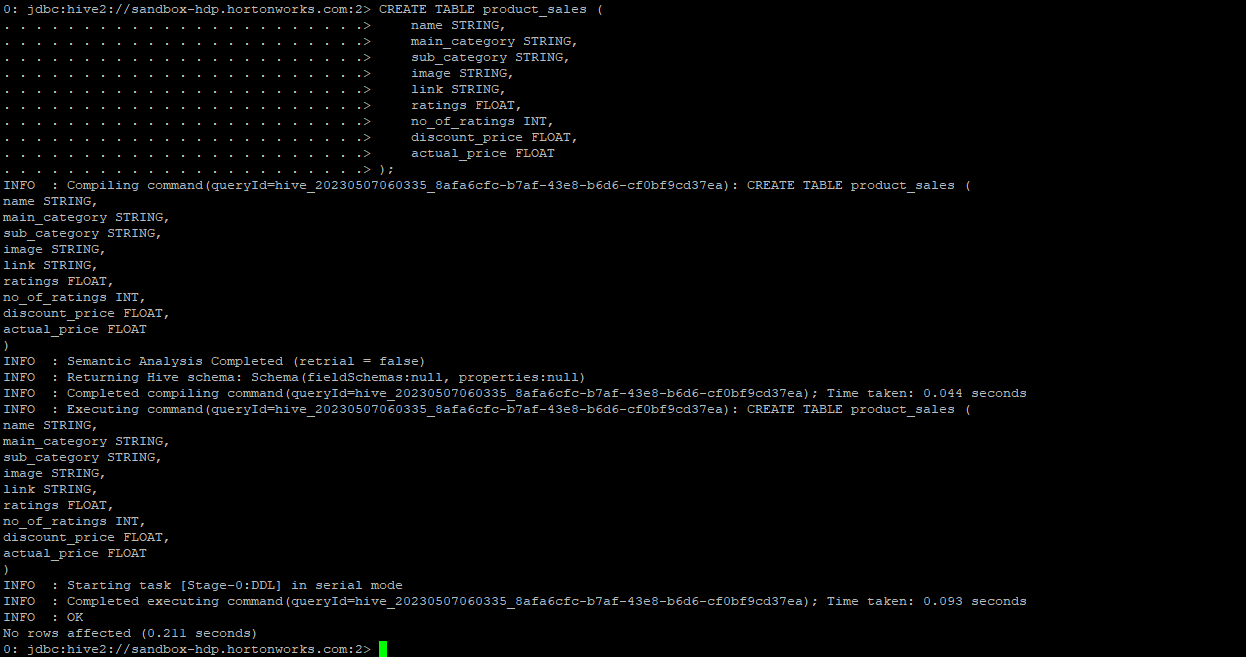
*)*

*ROW FORMAT DELIMITED*

*FIELDS TERMINATED BY ','*

*STORED AS TEXTFILE*

*TBLPROPERTIES("skip.header.line.count"="1");*



Loading CSV data into HIVE

*LOAD DATA INPATH '/path/to/csv/file.csv' OVERWRITE INTO TABLE product\_sales;*

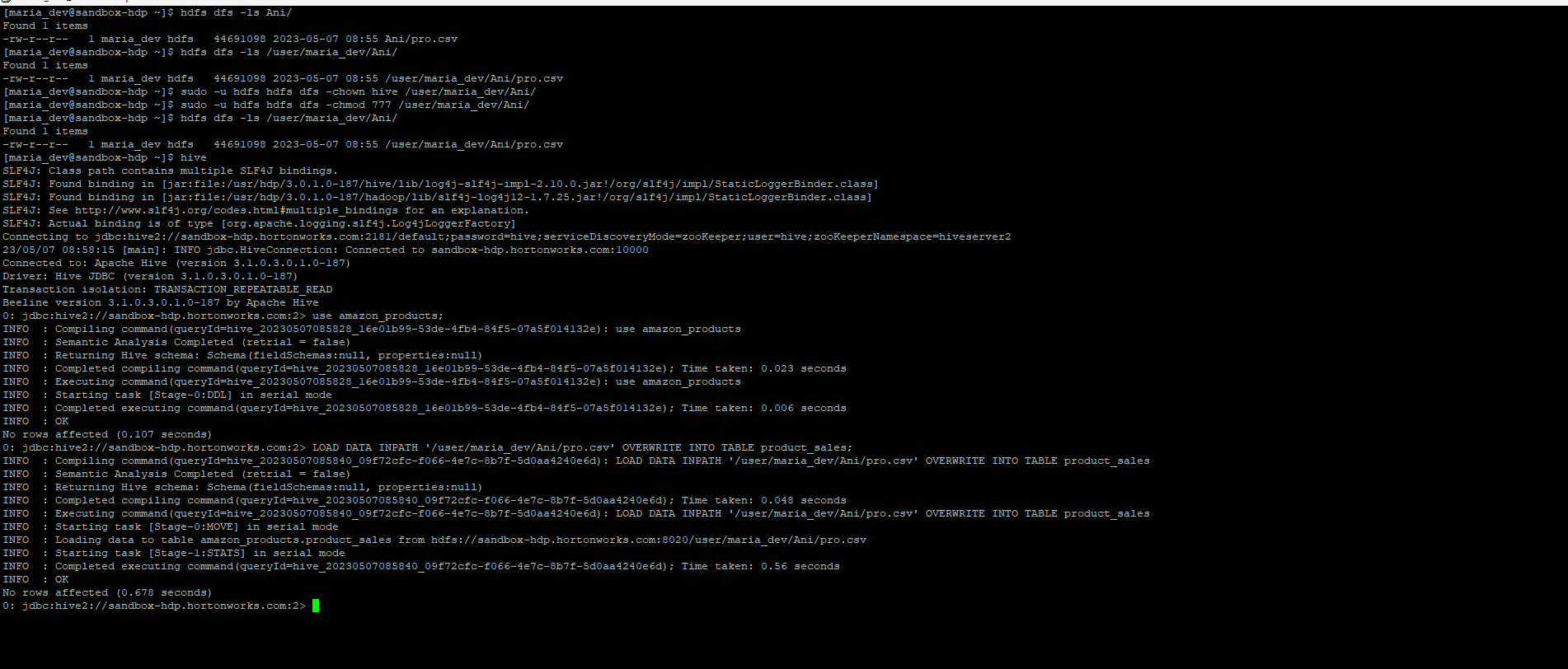
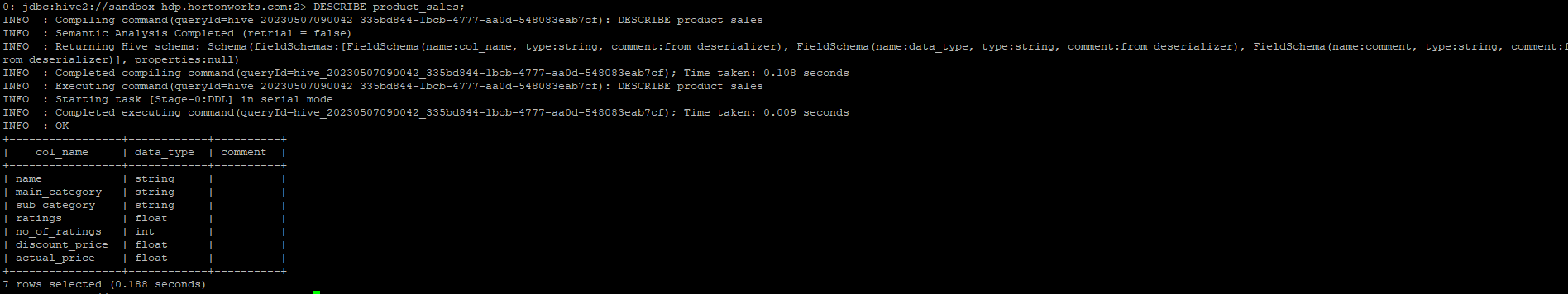


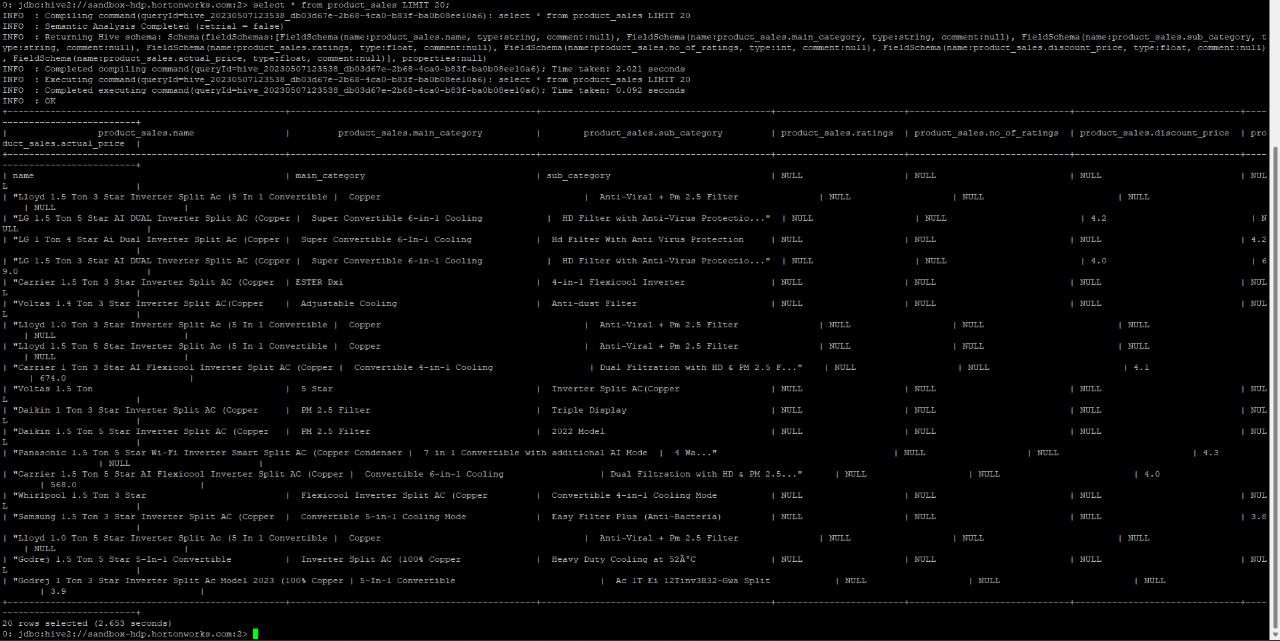
Table Schema

*describe product\_sales;*



Viewing Data using HiveQL

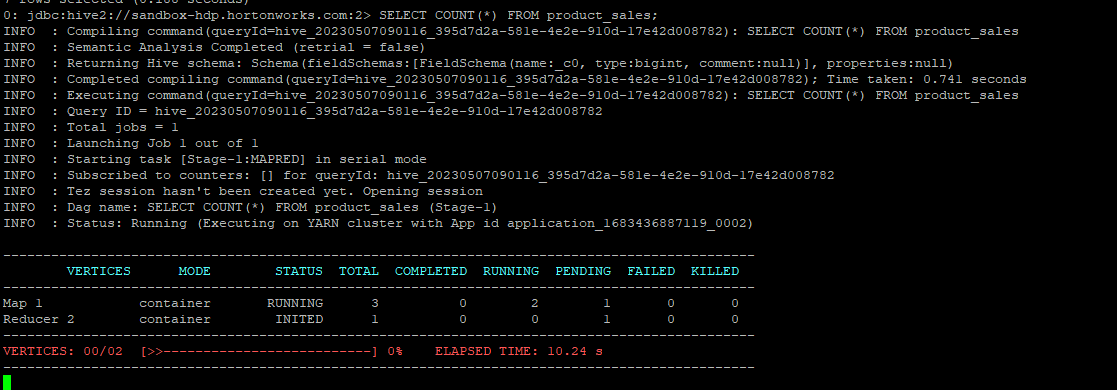
*Select \* from product\_sales Limit 20*

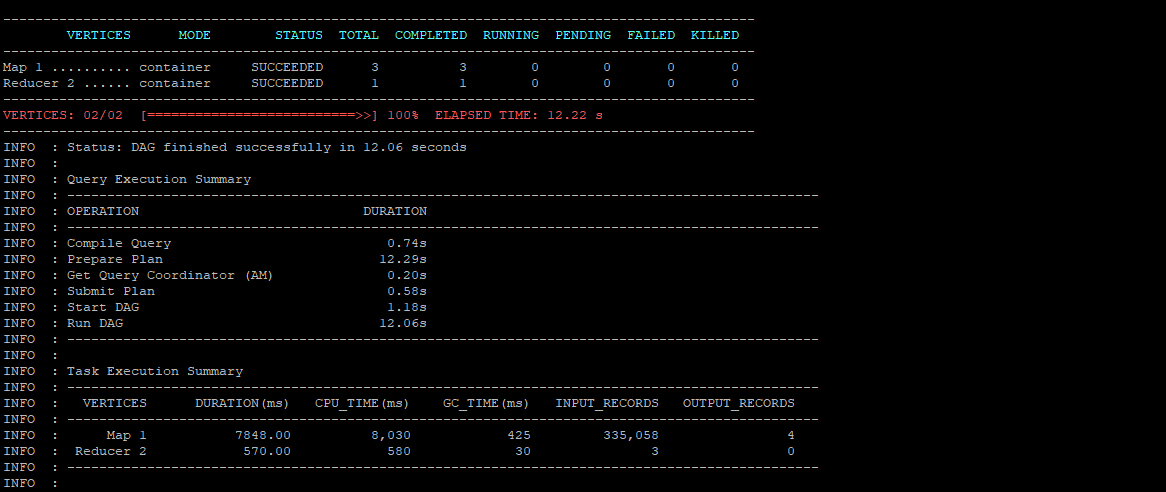


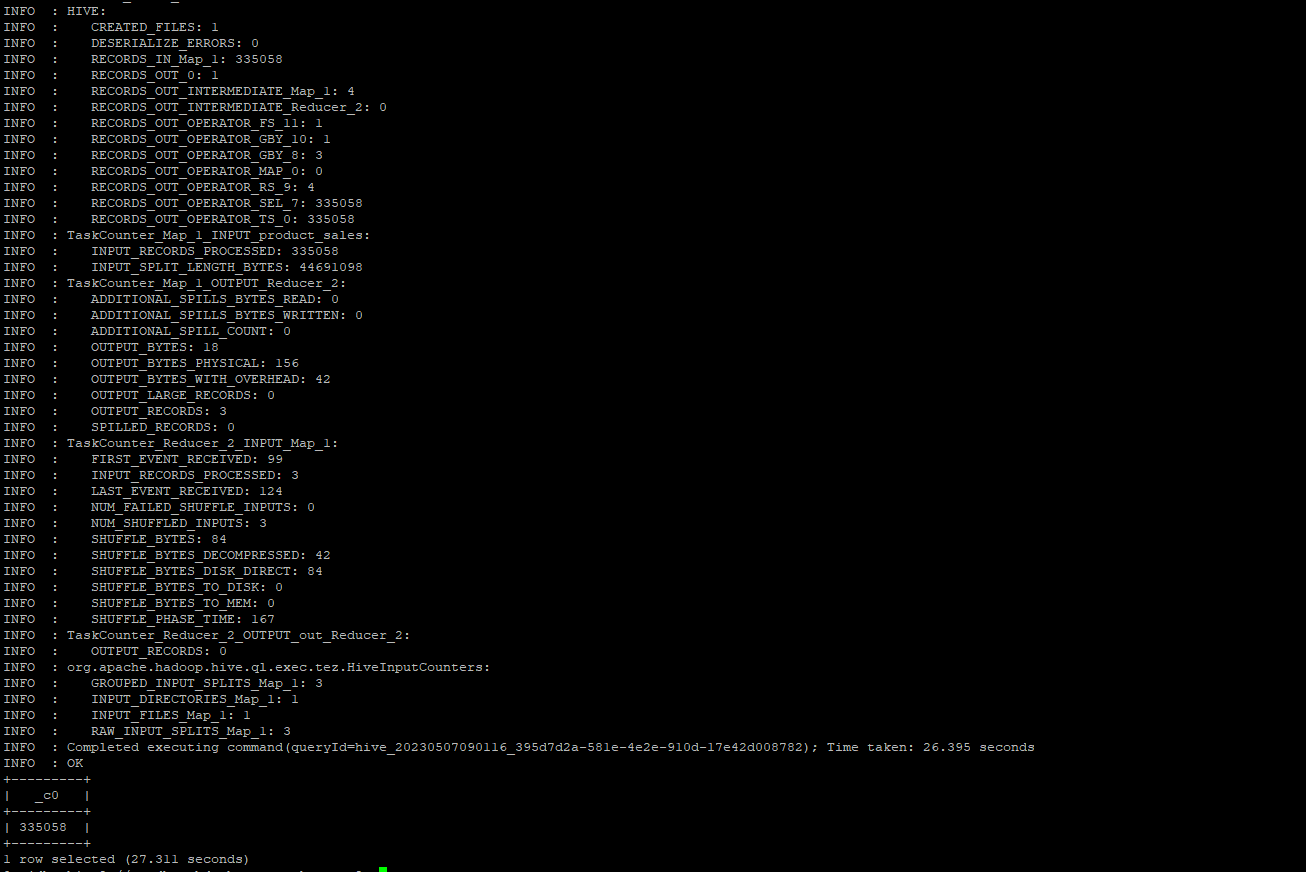
Data Analysis

Viewing Number of rows

*Select count(\*) from product\_sales;*







Viewing Min ,Max, Avg sales of each department grouped by categories

*SELECT*

*category,*

*COUNT(\*) AS num\_sales,*

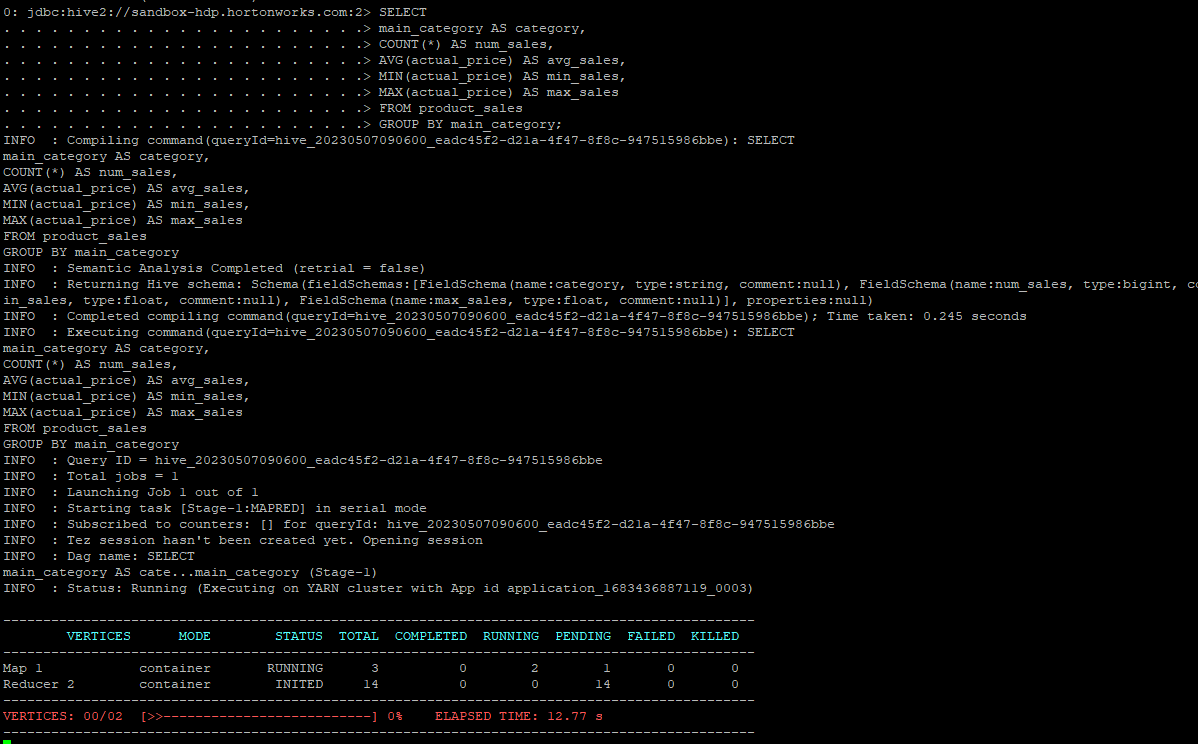
*AVG(sales) AS avg\_sales,*

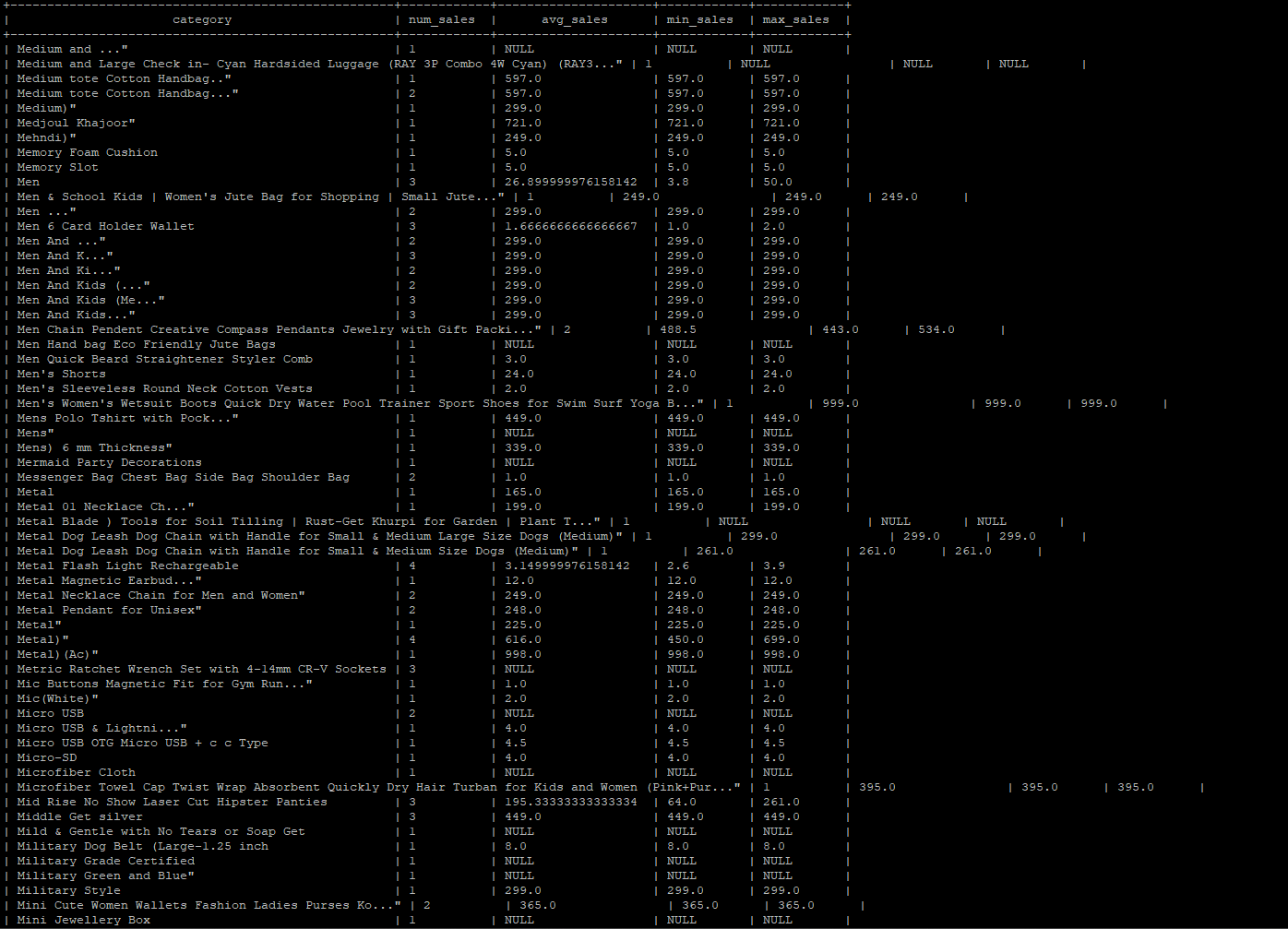
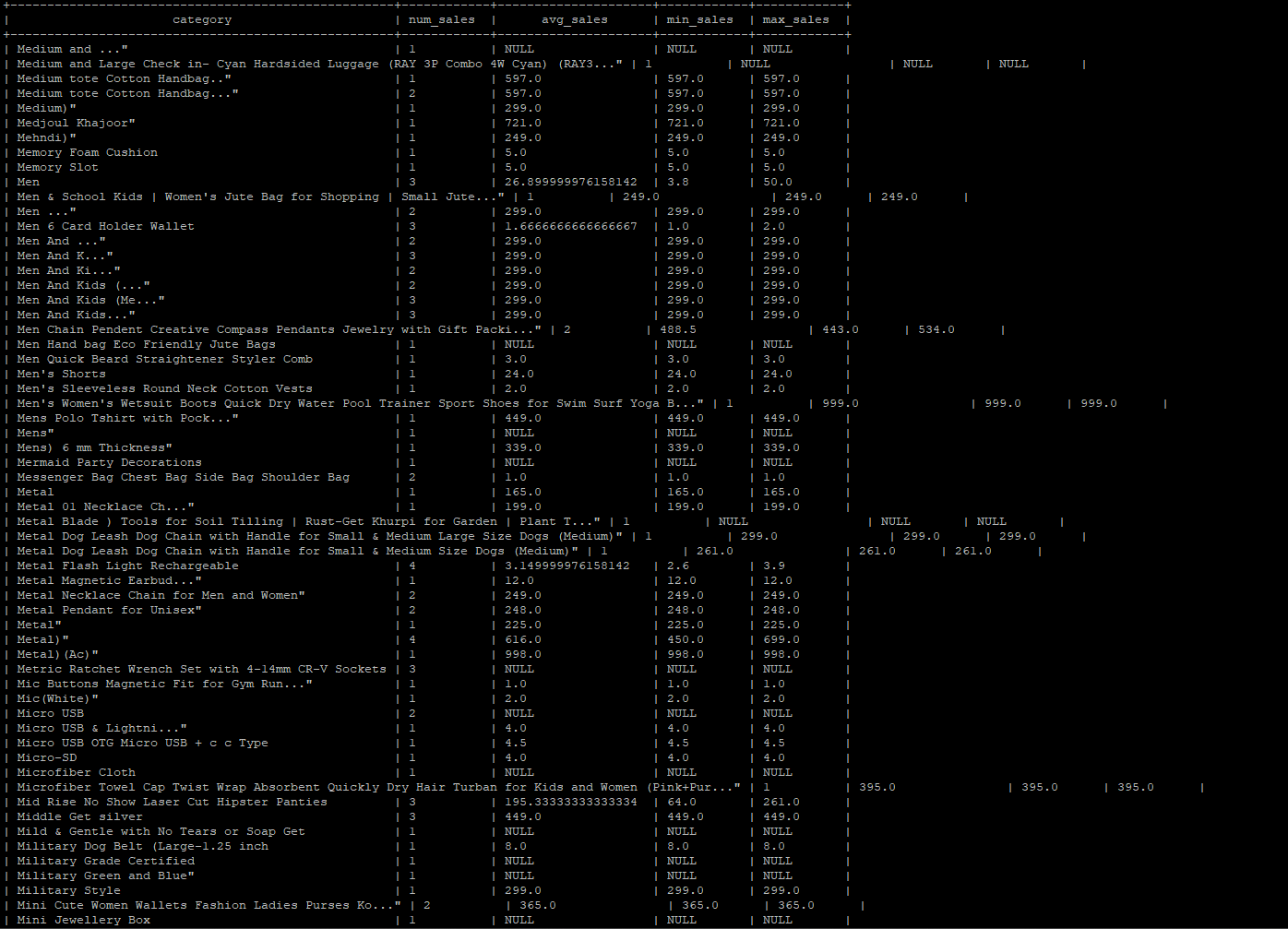
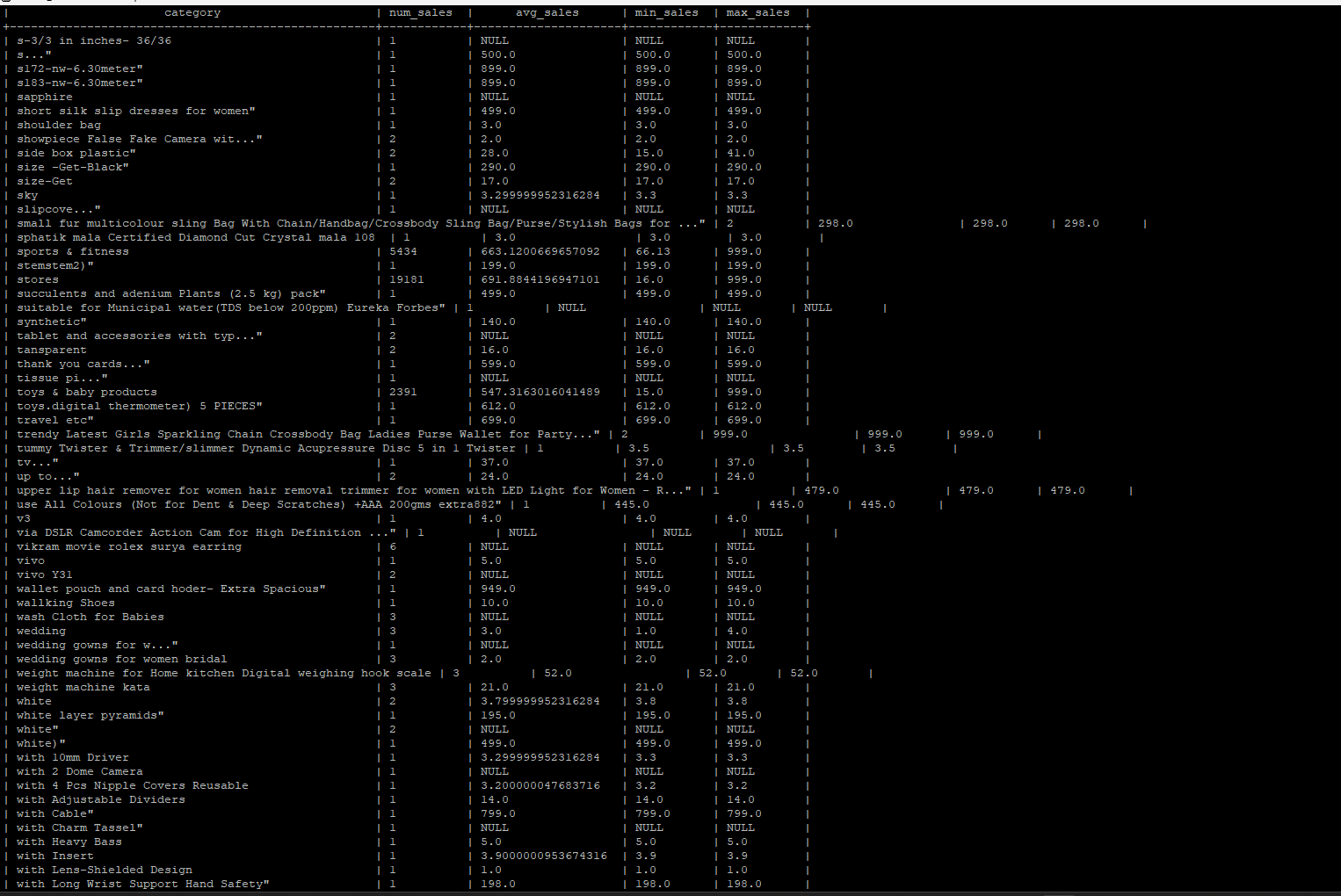
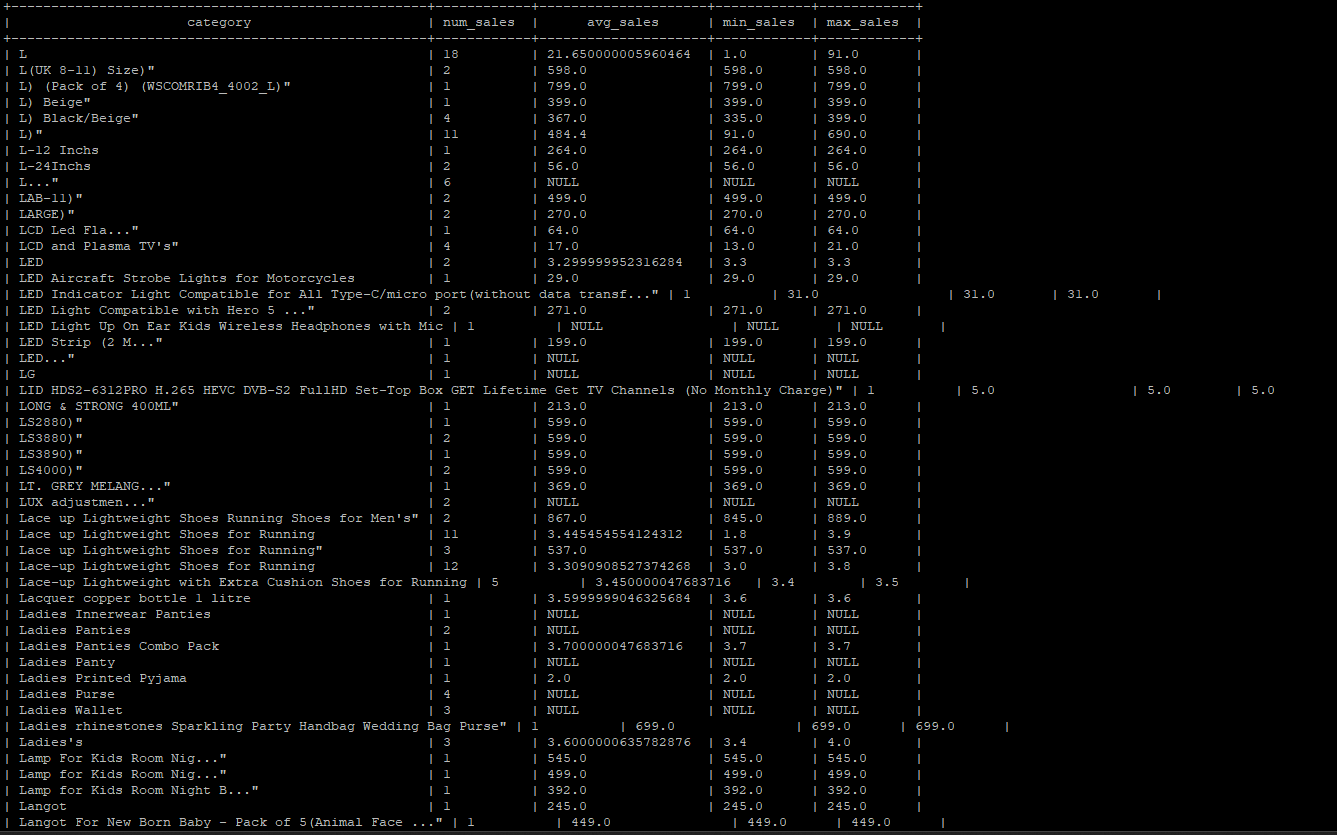
*MIN(sales) AS min\_sales,*

*MAX(sales) AS max\_sales*

*FROM product\_sales*

*GROUP BY category;*





Calculating Top 10 selling Categories

*SELECT*

*main\_category AS category,*

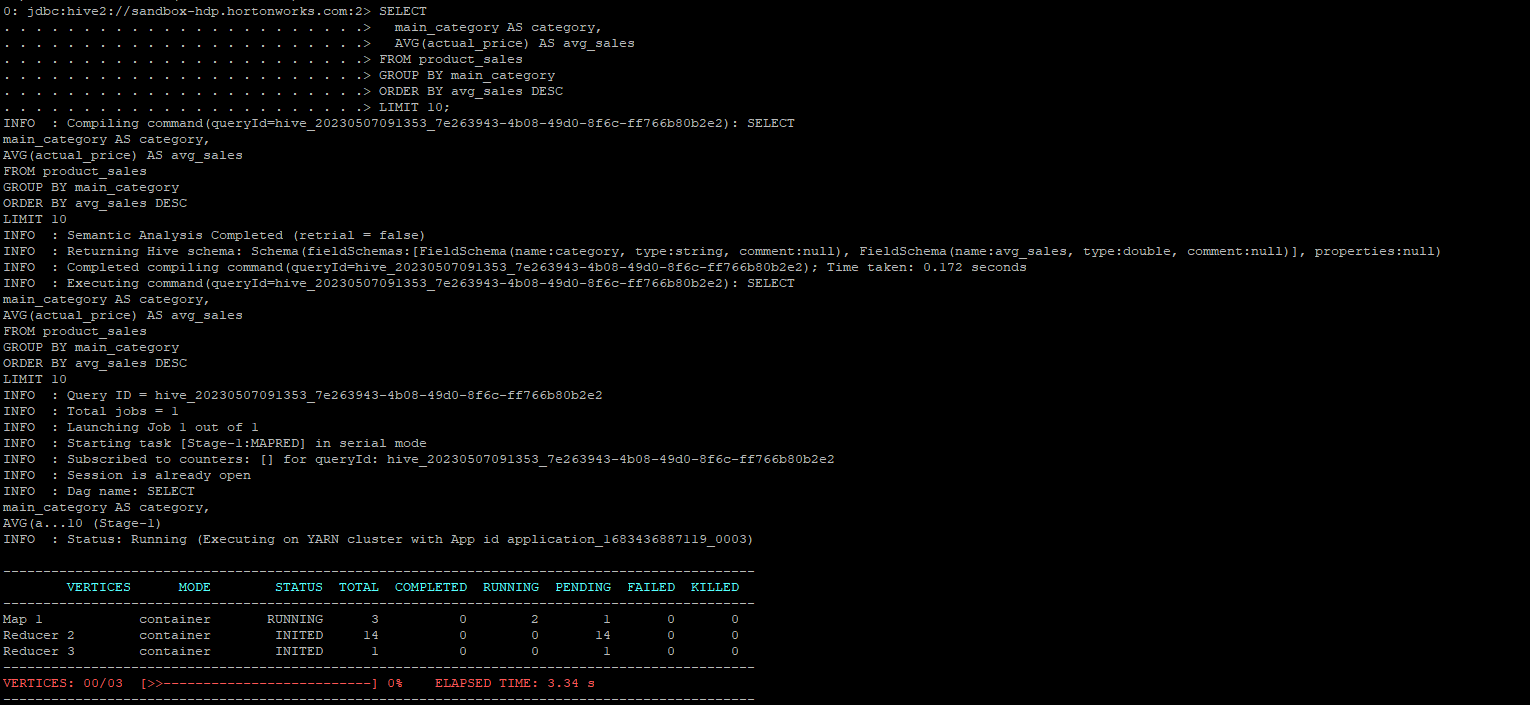
*AVG(actual\_price) AS avg\_sales*

*FROM product\_sales*

*GROUP BY main\_category*

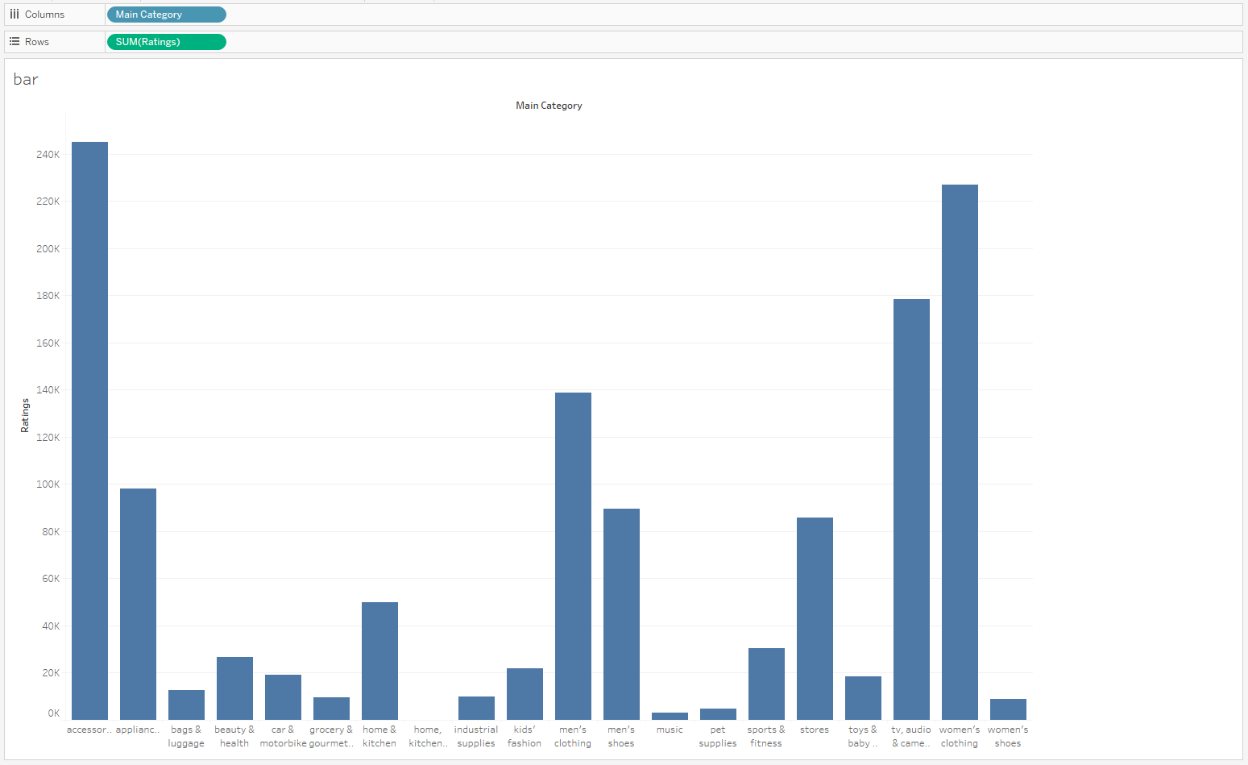
*ORDER BY avg\_sales DESC*

*LIMIT 10;*

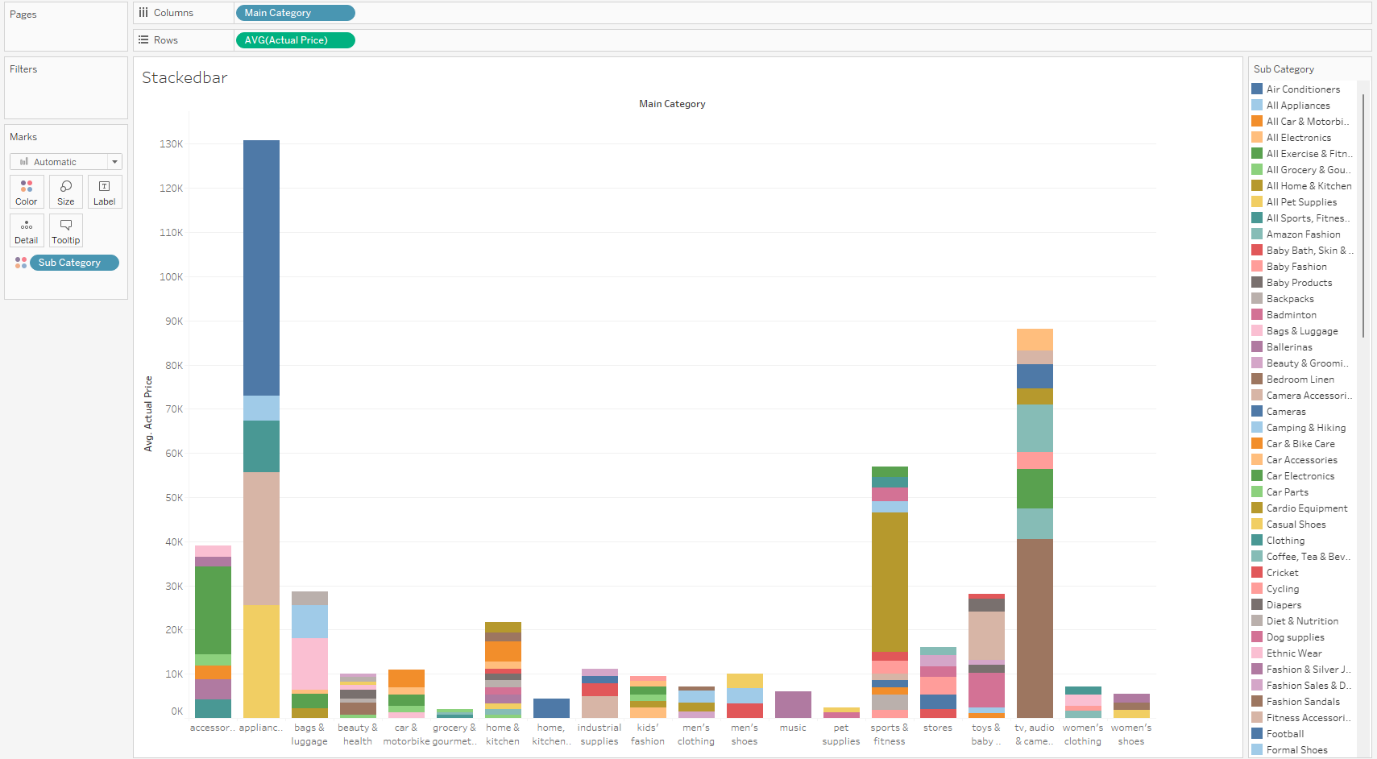


DATA VISUALIZATION

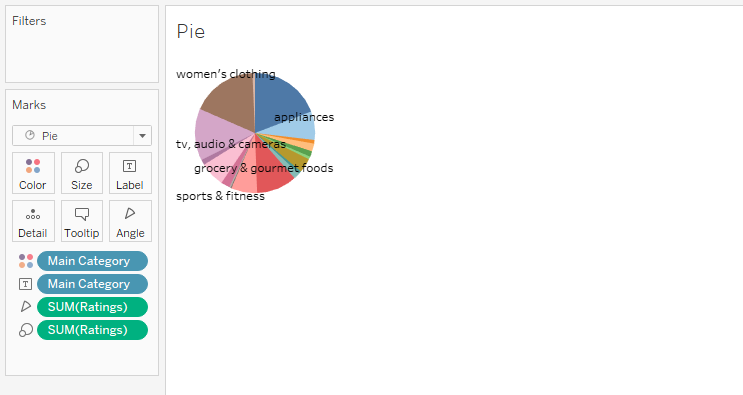
Bar Chart



Stacked Bar Chart



PIE Chart



Heatmap

