# AMAZON PRODUCT SALES

# ANALYSIS USING HIVE

### A PROJECT REPORT

*Submitted by*

**PONNURI ANIRUDDHA**

**[Reg No: RA2112704010015]**

**ABHAY SHAJI**

**[Reg No: RA2112704010006]**

**GAURAV SAHA**

**[Reg No: RA2112704010004]**

*Under the Guidance of*

## Dr. K. Sornalakshmi

(Assistant Professor, Department of Data Science and Business Systems)

*In partial fulfillment of the Requirements for the Degree of*

## Masters of Technology (Integrated)



**DEPARTMENT OF**

**DATA SCIENCE AND BUSINESS SYSTEMS**

## FACULTY OF ENGINEERING AND TECHNOLOGY SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

### MAY 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR-603203

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled “**AMAZON PRODUCT SALES AND VISUALIZATION THROUGH TABLEAU”** is the bonafide work of **“Ponnuri Aniruddha [Reg No: RA2112704010015]”,”Abhay Sahaji [Reg No: RA2112704010006]”,”Gaurav Saha [Reg No: RA2112704010004]”,** carried out the project work under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

Dr. K. Sornalakshmi Dr. M Lakshmi

#### GUIDE HEAD OF THE DEPARTMENT

Associate Professor Dept. of Data Science and Business

Dept. of Data Science and Business Systems Systems

### ABSTRACT

"Amazon Product Sales Analysis Using Hive and Visualization Using Tableau" highlights the use of big data technology and data visualization techniques to analyze and present insights into product sales data obtained from Amazon. The project involves using Apache Hive, a data warehouse infrastructure tool for querying and managing large datasets stored in distributed storage, to process and analyze the data. Furthermore, data visualization tools like Tableau are employed to create interactive dashboards and reports that allow users to explore the data and gain valuable insights. The project provides an opportunity to develop skills in data analysis, data visualization, and big data technologies, which are increasingly important in the field of data science.

Amazon Product Sales Analysis using Hive and Visualization using Tableau is a project that aims to analyze the sales of various products sold on Amazon using the big data tool Hive and visualize the insights obtained using Tableau. The project involves loading and processing the Amazon product sales dataset using Hive, performing exploratory data analysis to gain insights about the data, and generating visualizations using Tableau to effectively communicate the insights obtained. The project aims to provide a comprehensive analysis of the sales data and help identify trends and patterns in product sales, customer behavior, and purchase patterns. The insights obtained from the analysis can be used to improve marketing strategies, optimize pricing, and make better business decisions.

### ACKNOWLEDGEMENTS

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 register our immeasurable thanks to our Faculty Advisor**, Dr K Shantha Kumari, Ph.D.**

, 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 K Sornalakshmi, Ph.D.**, 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/She/They 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.

PONNURI ANIRUDDHA

ABHAY SHAJI

GAURAV SAHA

## TABLE OF CONTENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TABLE OF CONTENTS | | | | |
| CHAPTER NO. | | | TITLE | PAGE NO. |
|  | ABSTRACT | |  | iii |
|  | ACKNOWLEDGMENTS | |  | iv |
|  | PROBLEM STATEMENT | |  | v |
|  | MOTIVATION | |  | xxvii |
|  | SCOPE |  |  |  |
| 1. | INTRODUCTION | |  | 1 |
|  | 1.1 | GENERAL |  | 1 |
|  | 1.2. | HISTORY |  | 1 |
|  | 1.3 | BENEFITS OF USING HIVE AND TABLEAU | | 2 |
|  | 1.4 | WORKNIG PRINCIPLE |  | 3 |
| 2 | LITERATURE STUDY | |  | 4 |
| 3 | DATASET | |  | 6 |
| 4 | CODE |  |  | 8 |
| 5 | VISUALIZATION | |  | 17 |
| 6 | CONCLUSION | |  | 21 |
|  | REFERENCE | |  | 22 |

**PROBLEM STATEMENT**

The problem statement for the analysis of Amazon product sales using Hive and Tableau is based on the need for businesses to leverage data analytics and visualization tools to gain insights into customer behavior and product trends. With millions of products available on Amazon in various categories, it can be challenging for businesses to identify the most popular products, categories, and sub-categories, and make data-driven decisions to optimize their product offerings, improve customer satisfaction, and increase sales.

The use of tools such as Hive and Tableau can help businesses to process and analyze large volumes of data, and create interactive visualizations and dashboards for better data exploration and communication. Hive is a data warehouse system that enables businesses to store, manage and query large datasets in a distributed manner, while Tableau is a data visualization tool that allows businesses to create interactive dashboards and visualizations for data analysis and decision-making.

The problem statement highlights the need for businesses to use Hive and Tableau to analyze and visualize Amazon product sales data and identify the top-selling products, popular categories and sub-categories, average ratings and prices, and other key metrics for their business success. By doing so, businesses can gain insights into customer behavior and preferences, optimize their product offerings, improve customer satisfaction, and increase sales.

In summary, the problem statement for the analysis of Amazon product sales using Hive and Tableau focuses on the need for businesses to leverage data analytics and visualization tools to gain insights into customer behavior and product trends, and make data-driven decisions to improve their bottom line.

**MOTIVATION**

Improving Sales Performance: By analyzing the product sales data, Amazon can gain insights into which products are selling well and which ones are not. This can help them make data-driven decisions to improve their sales performance and increase revenue.

Understanding Customer Behavior: Analyzing the customer behavior data can help Amazon understand what customers are buying, when they are buying, and why they are buying. This can help them improve their customer experience, tailor their marketing efforts, and optimize their product offerings.

Identifying Trends and Patterns: Analyzing the sales data over time can help Amazon identify trends and patterns in customer behavior, product performance, and market demand. This can help them stay ahead of the competition and adapt to changing market conditions.

Making Better Business Decisions: By using data-driven insights, Amazon can make better business decisions about product development, pricing, promotions, and inventory management. This can help them optimize their operations and increase profitability.

Enhancing Data Visualization: By using Tableau, Amazon can create visually appealing and interactive dashboards to represent their data in a more meaningful way. This can help stakeholders quickly and easily understand the insights and take action based on them.

## SCOPE

The scope of the "Amazon Product Sales Analysis using Hive and Visualization using Tableau" project includes exploring and analyzing the sales data of products on Amazon. The project involves working with a large dataset, which contains information about the products such as their names, categories, sub-categories, prices, ratings, etc.

The project aims to identify the top-selling products, categories, and sub-categories, as well as to gain insights into the purchasing patterns of customers. The analysis will help in understanding the market trends and identifying potential areas for improvement in the sales strategy.

The project also involves visualizing the analyzed data using Tableau, which is a powerful data visualization tool. The visualization will help in presenting the findings in a more effective and easy-to-understand way, which will aid in making informed business decisions.

The scope of the project is not limited to just Amazon; it can be extended to analyze the sales data of other e-commerce platforms as well. The project is relevant for businesses that operate in the e-commerce industry and want to improve their sales strategy by gaining insights into customer behavior and market trends.

* 1. **INTRODUCTION**

## CHAPTER 1 INTRODUCTION

In today's data-driven business world, companies are constantly looking for ways to extract insights from their data to make informed decisions. One of the primary ways to do this is through data analysis, which involves processing, cleaning, and interpreting large amounts of data to identify patterns and trends.

In this context, the analysis of product sales data is of particular interest, as it can provide valuable insights into consumer behavior and market trends. Amazon, being one of the largest online marketplaces in the world, generates a vast amount of sales data that can be used for such analysis.

Hive, a data warehousing system built on top of Hadoop, is a popular tool for analyzing large datasets. It provides a SQL-like interface to query and manage data, making it accessible to users with a background in SQL. Hive also offers the ability to scale to petabytes of data, making it a suitable choice for big data analysis.

In this project, we will analyze Amazon product sales data using Hive and visualize the results using Tableau. The main objective of this project is to extract meaningful insights from the sales data, such as top-performing products, popular categories, and sales trends over time. The use of Hive and Tableau will enable us to handle and visualize large amounts of data, making it easier to draw insights and conclusions.

#### HISTORY

The history of Amazon dates back to 1994, when Jeff Bezos founded the company in Seattle, Washington, with the aim of creating an online bookstore. Over time, Amazon expanded to include a wide range of products, including electronics, clothing, and household goods. As the company grew, so did the need for analyzing its vast amount of sales data

Initially, Amazon relied on traditional data analysis tools such as SQL databases and spreadsheets to manage and analyze its sales data. However, as the volume of data grew, these tools proved to be insufficient for handling the scale and complexity of Amazon's sales data. This led to the development of Apache Hadoop, an open-source distributed computing platform that can handle large volumes of data.

Amazon was one of the early adopters of Apache Hadoop, and it became one of the core technologies used by the company for managing and analyzing its sales data. In addition, Amazon developed its own version of Hadoop called Amazon Elastic MapReduce (EMR), which allowed the company to process and analyze its sales data more efficiently.

Hive is a data warehousing tool that was developed on top of Hadoop. It allows users to query and analyze data stored in Hadoop using SQL-like commands. Hive was developed to make it easier for analysts and data scientists to work with Hadoop by providing a familiar SQL interface.

With the advent of big data, the need for scalable and efficient data analysis tools has become increasingly important for businesses. Amazon's use of Hive and Hadoop for analyzing its sales data has become a popular use case for these technologies, and has inspired many other companies to adopt similar technologies for their own data analysis needs.

#### BENEFITS OF USING HIVE AND TABLEAU

Benefits of using HIVE:

1. Scalability: Hive is designed to handle large datasets and can process them in a distributed manner. This makes it ideal for processing and analyzing big data.
2. SQL-like language: Hive uses a SQL-like language called HiveQL, which makes it easy for users with SQL knowledge to query and analyze data. This enables users to write complex queries to analyze and explore the data.
3. Data warehouse integration: Hive can be integrated with data warehousing tools such as Hadoop and Spark, which allows users to store and manage large amounts of data in a distributed manner.
4. Flexible data processing: Hive is designed to work with different types of data, including structured and semi-structured data. This means that users can analyze data from a wide range of sources and formats.
5. High-performance processing: Hive uses advanced processing techniques such as query optimization and indexing to improve performance and reduce processing times.

Benefits of using Tablue for visualization:

1. Tablue is a powerful data visualization tool that allows for creating interactive and dynamic dashboards and reports.
2. It supports various types of visualizations, including charts, tables, maps, and more.
3. Tablue provides intuitive drag-and-drop interfaces that make it easy to create and customize visualizations.
4. It can connect to various data sources, including MongoDB, and easily integrate with other tools and platforms.

Overall benefits of using HIVE and Tablue for analysis:

1. Scalability: Hive allows users to handle large volumes of data and process it in a distributed manner. This makes it an ideal tool for processing and analyzing big data.
2. SQL-based queries: Hive uses a SQL-like language called HiveQL, which makes it easy for users with SQL knowledge to query and analyze data. This enables users to write complex queries to analyze and explore the data.
3. Data warehouse integration: Hive can be integrated with data warehousing tools such as Hadoop and Spark, which allows users to store and manage large amounts of data in a distributed manner.
4. Flexibility: Hive allows users to work with different types of data, including structured and semi-structured data. This means that users can analyze data from a wide range of sources and formats.
5. Interactive visualizations: Tableau allows users to create interactive visualizations and dashboards, which can help to identify patterns and trends in the data. The visualizations are also easy to share and can be accessed from different devices.
6. Ease of use: Tableau has a user-friendly interface that makes it easy for users to create visualizations and dashboards without the need for programming skills. This means that users can create their own visualizations and explore the data in a more interactive way.
7. Collaboration: Tableau allows users to share their visualizations and dashboards with others, which enables collaboration and knowledge sharing. This can help teams to work together more effectively and make better data-driven decisions.

#### WORKING PRINCIPLE

The working principle of Amazon product sales analysis using Hive involves processing and analyzing the sales data of various products sold on Amazon using the Hive data warehousing system. The sales data is first imported into Hive, a distributed data warehousing system built on top of Hadoop, and then transformed and analyzed using HiveQL, a SQL-like language.

The data is queried using HiveQL to obtain insights into the sales trends, popular products, and customer behavior. The queries are designed to aggregate the data by different parameters such as category, sub-category, customer demographics, and time period. The aggregated data is then used to generate reports and visualizations that help in understanding the sales patterns and identifying opportunities for growth.

Once the data is analyzed and insights are obtained, the results can be visualized using various tools such as Tableau, which is a powerful data visualization software. Tableau helps in creating interactive dashboards and visualizations that enable the stakeholders to make informed decisions based on the insights gained from the data.

Overall, the working principle of Amazon product sales analysis using Hive involves importing, transforming, analyzing, and visualizing the sales data using HiveQL and Tableau, to gain insights into the sales patterns and make data-driven decisions for growth.

## CHAPTER 2 LITERATURE STUDY

1. "Big Data Analytics on Amazon Web Services (AWS) using Hive and Tableau" by Avinash Prasad, published in the International Journal of Computer Science and Mobile Computing in 2016. This paper discusses the use of Hive and Tableau for big data analytics on AWS, with a focus on Amazon customer review data.
2. "Big Data Analytics using Hive and Tableau" by P. Shanmugapriya and S. Srinivasan, published in the International Journal of Advanced Research in Computer Science and Software Engineering in 2015. This paper provides an overview of Hive and Tableau and their use for big data analytics, with a case study on analyzing social media data.
3. "A Study of Data Visualization Tools: Tableau, QlikView, and MicroStrategy" by M. A. Rizwan and M. A. N. Alias, published in the International Journal of Computer Science and Information Technologies in 2015. This paper compares Tableau, QlikView, and MicroStrategy as data visualization tools, with a focus on their features, strengths, and weaknesses.
4. "Data Warehousing and Business Intelligence for Analyzing Amazon Sales Data using Tableau" by K. Saravanan and R. Kavitha, published in the International Journal of Emerging Technology and Advanced Engineering in 2013. This paper presents a case study on analyzing Amazon sales data using Tableau for data warehousing and business intelligence purposes.
5. "A Comparative Study of Data Visualization Tools for Big Data Analytics" by N. R. Selvarani and V. M. Jeyakumar, published in the International Journal of Advanced Research in Computer Science and Software Engineering in 2016. This paper compares various data visualization tools, including Tableau, for big data analytics, with a focus on their performance, scalability, and ease of use.

# CHAPTER 3 DATASET

# 

# Amazon Products Sales

**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.

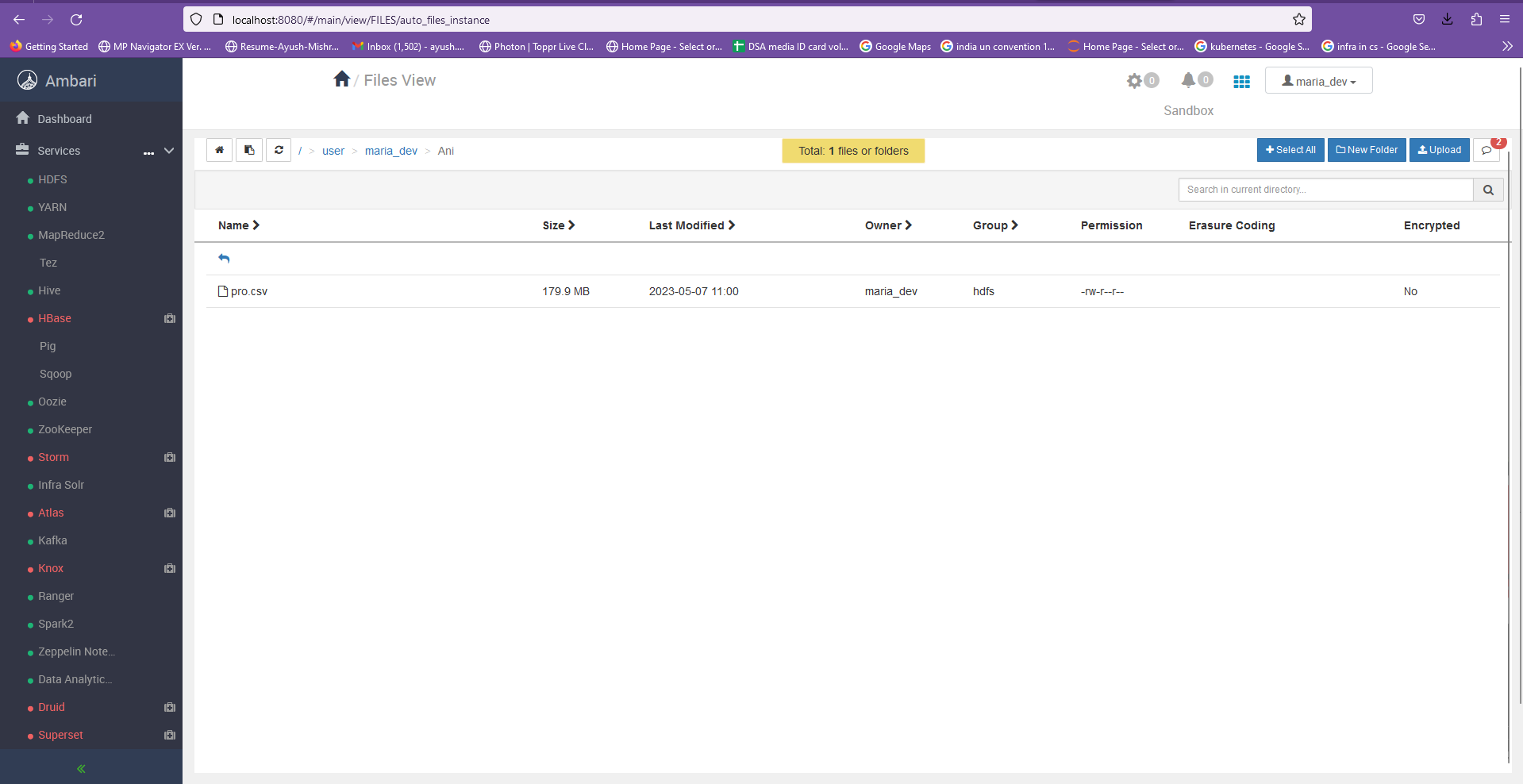
**Features**

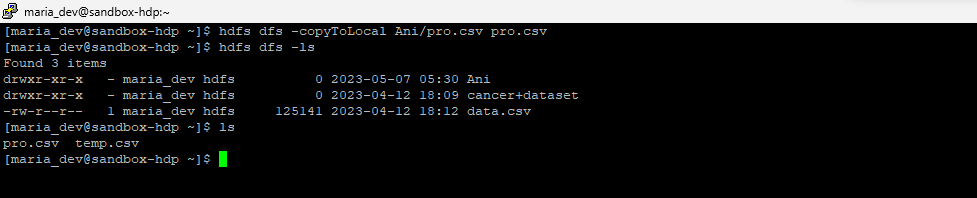
| **name** | **description** |
| --- | --- |
| *name* | The name of the product |
| *main\_category* | The main category of the product belong |
| *sub\_category* | The main category of the product belong |
| *ratings* | The ratings given by amazon customers of the product |
| *no of ratings* | The number 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 |

# CHAPTER 4 CODE

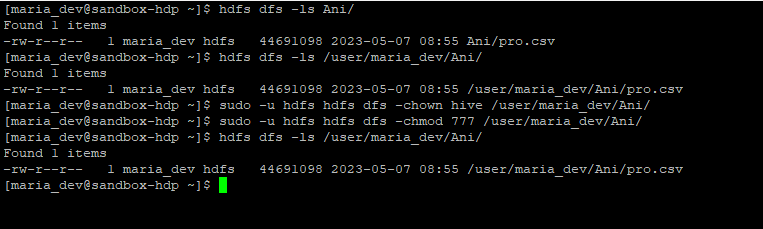
**Importing File into HDFS**

Uploading File into HDFS using Apache Ambari



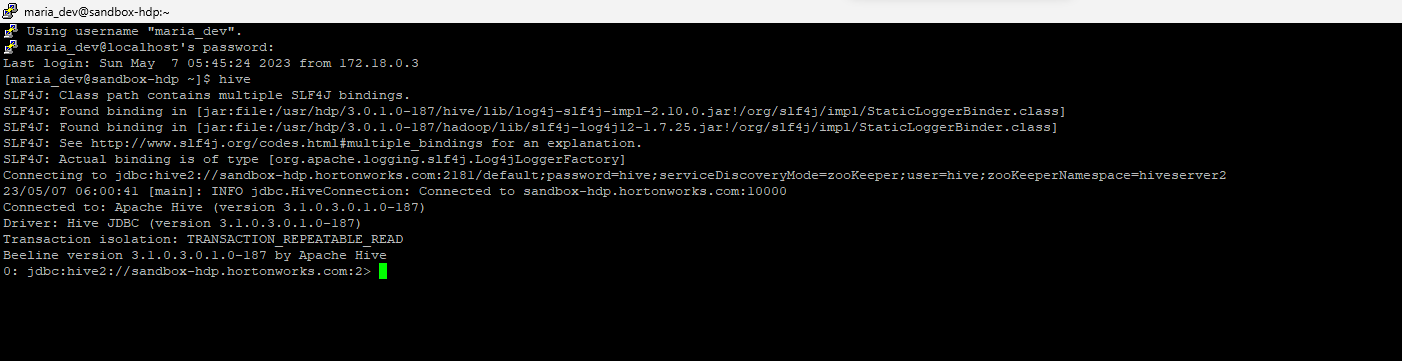


Changing ownership and permissions of data file



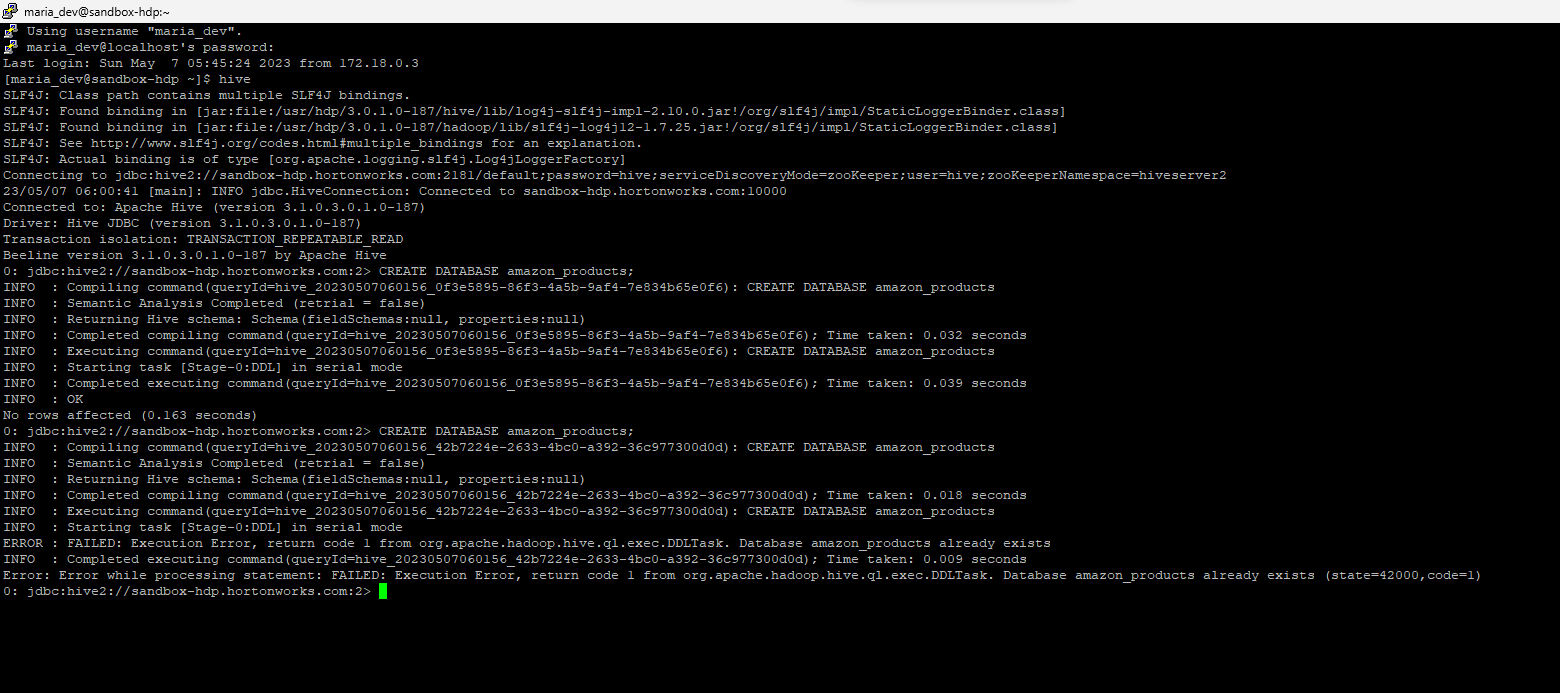
Hive Commands

Launching Hive from HortonWorks

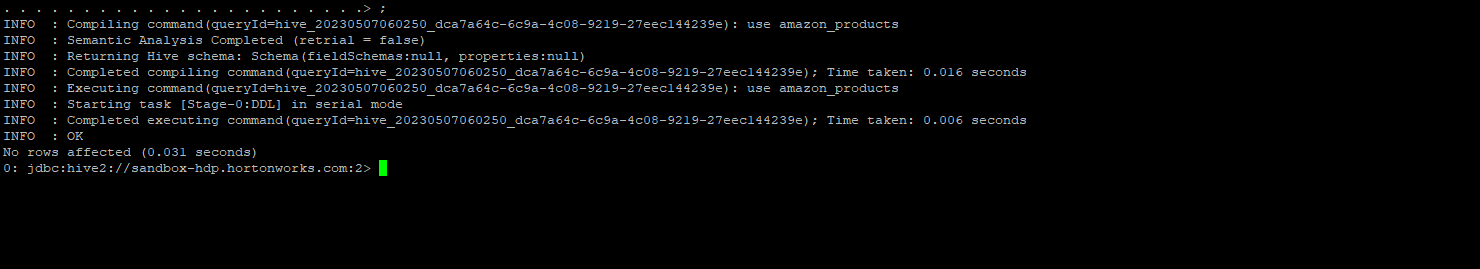


Creating a Database

* *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*

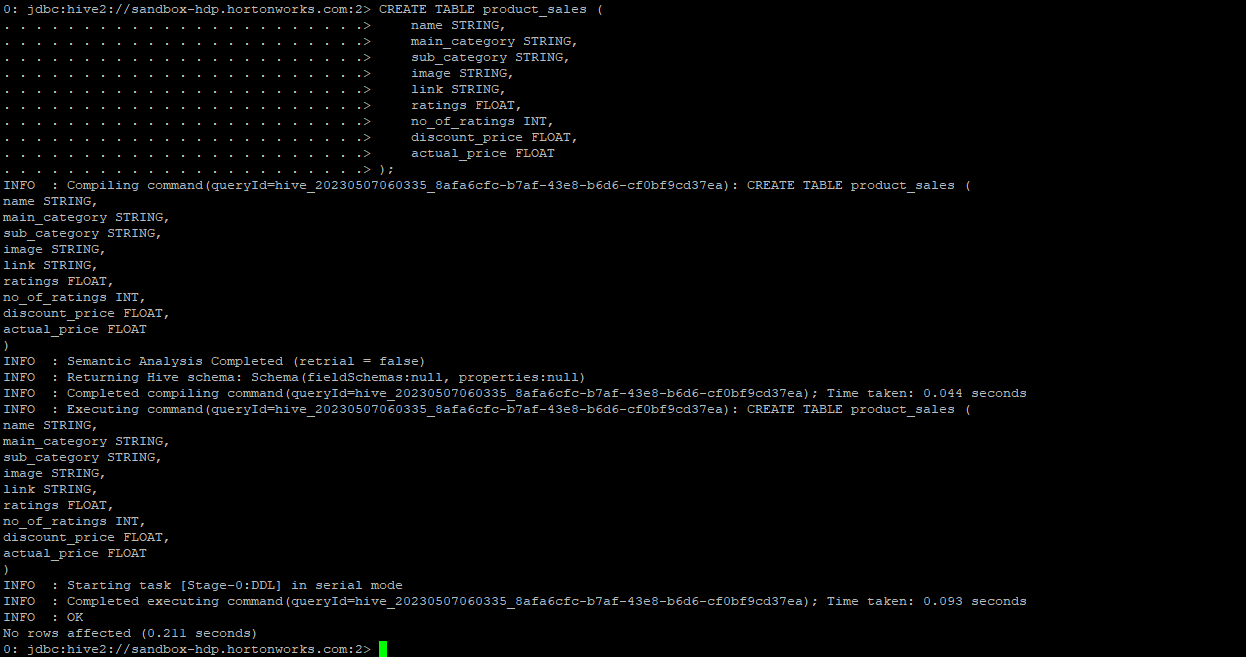
*)*

*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;*

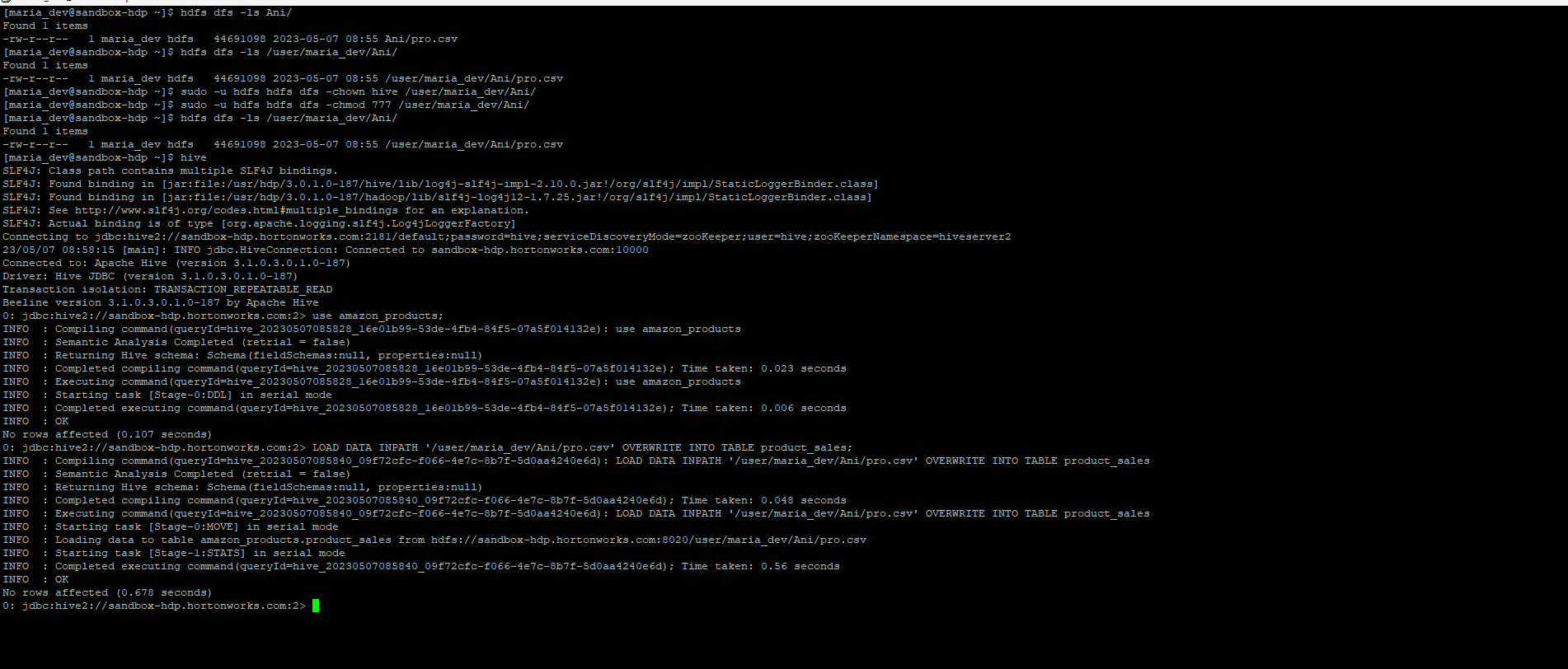
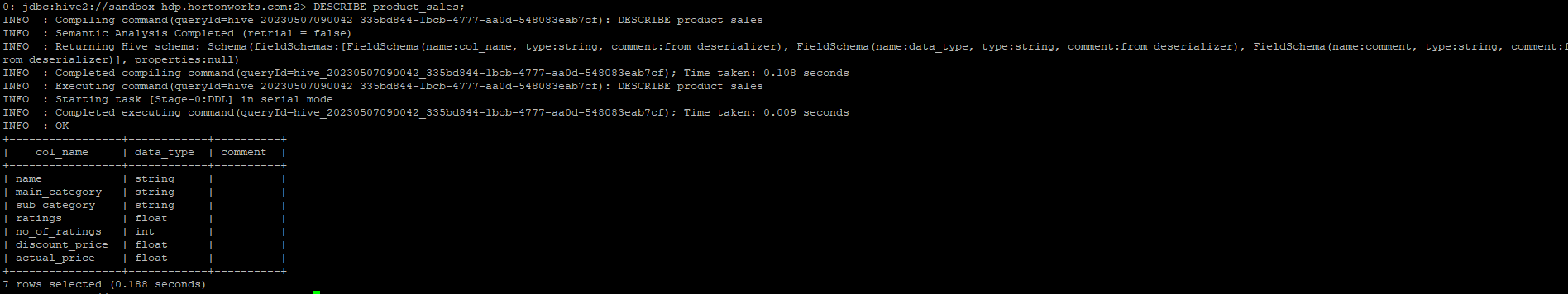


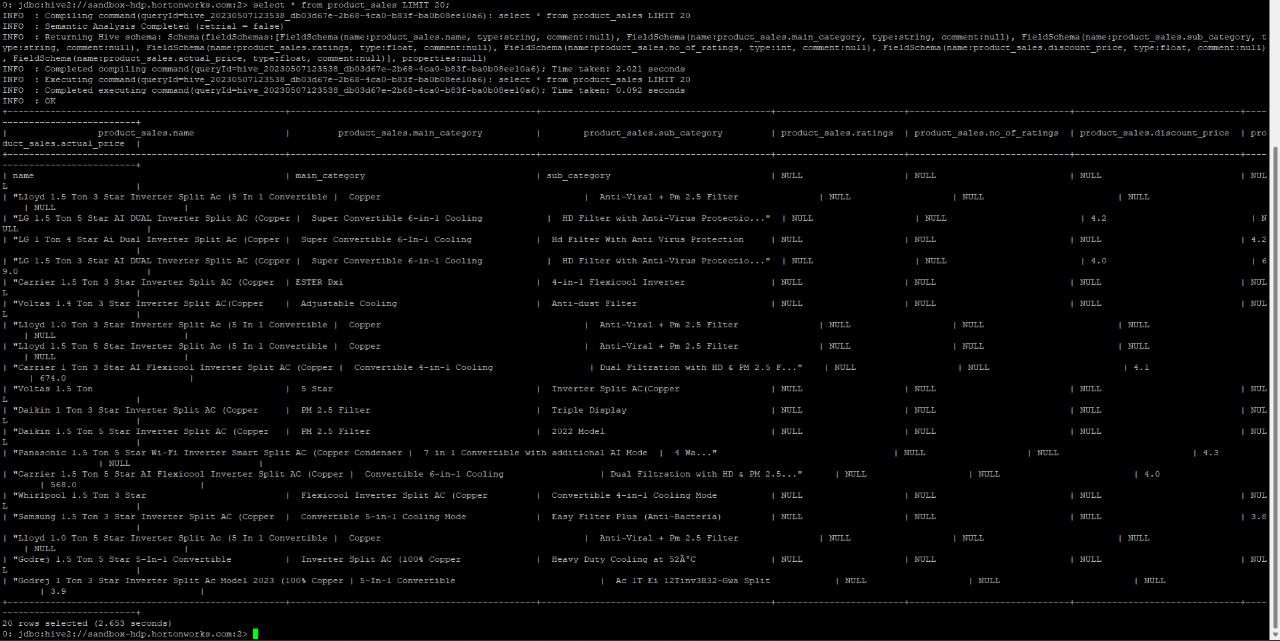
Table Schema

*describe product\_sales;*



Viewing Data using HiveQL

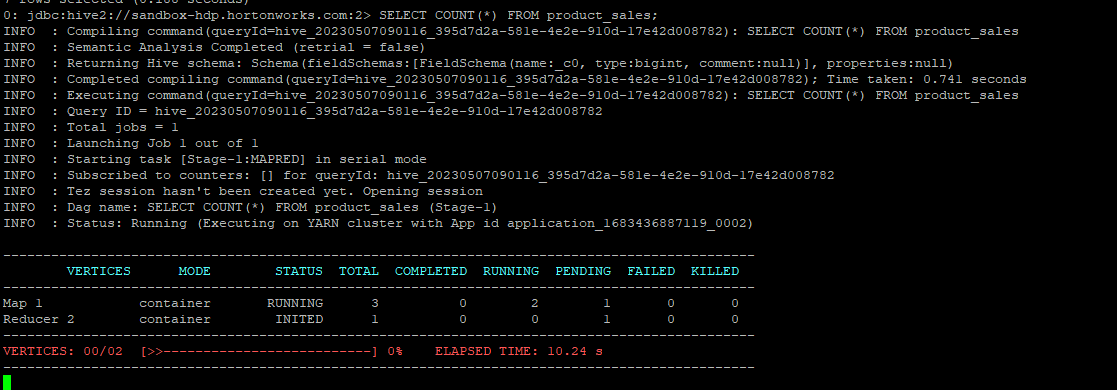
*Select \* from product\_sales Limit 20*

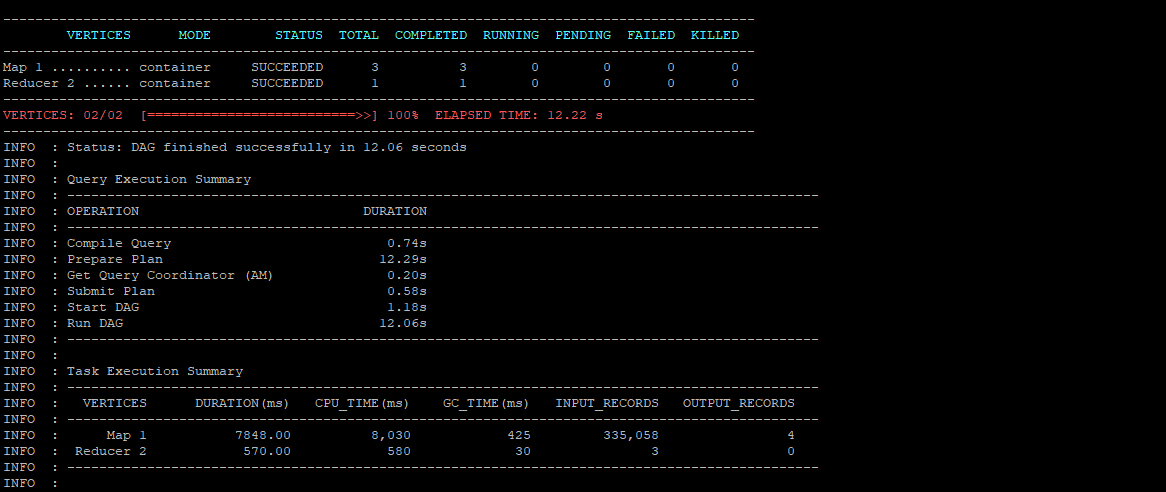


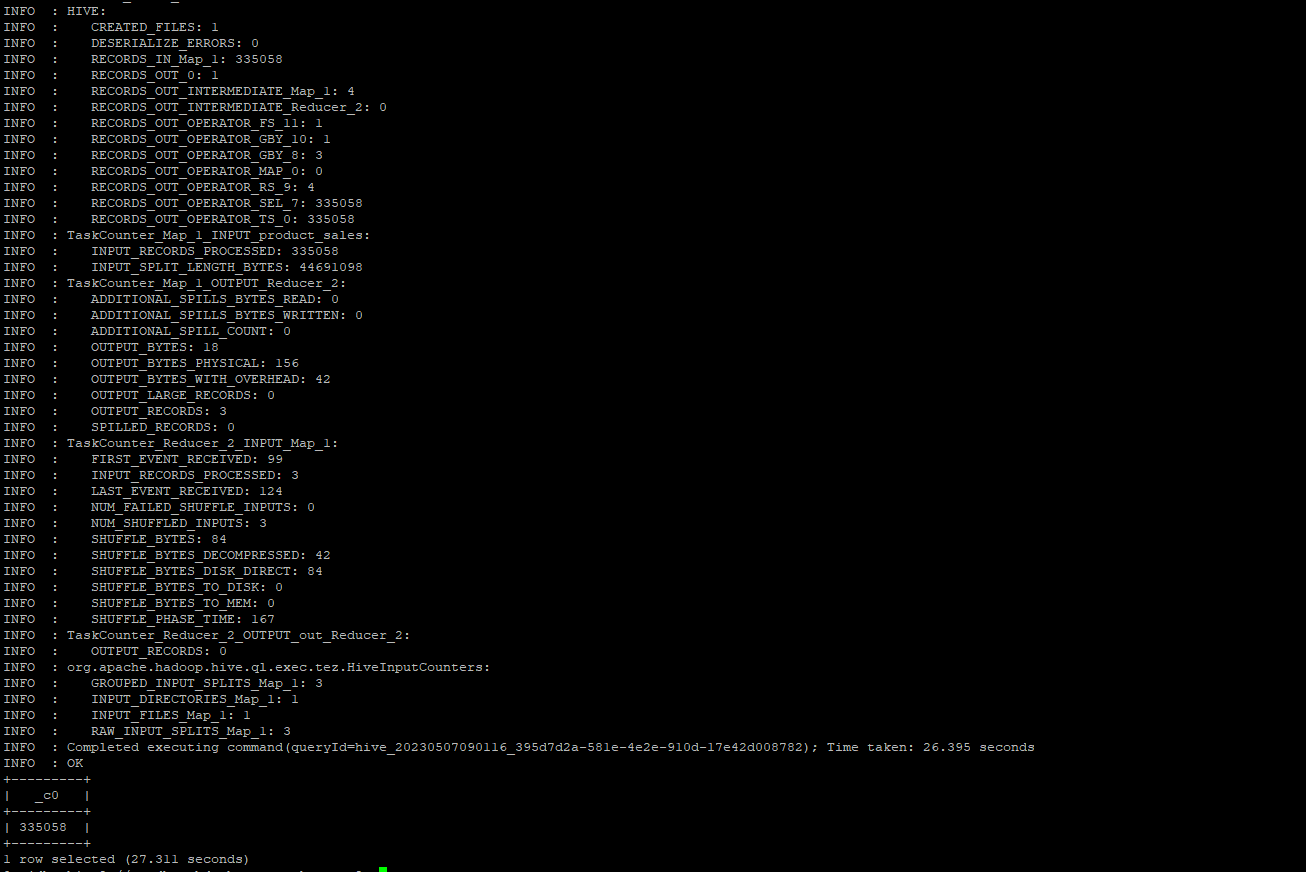
Data Analysis

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,*

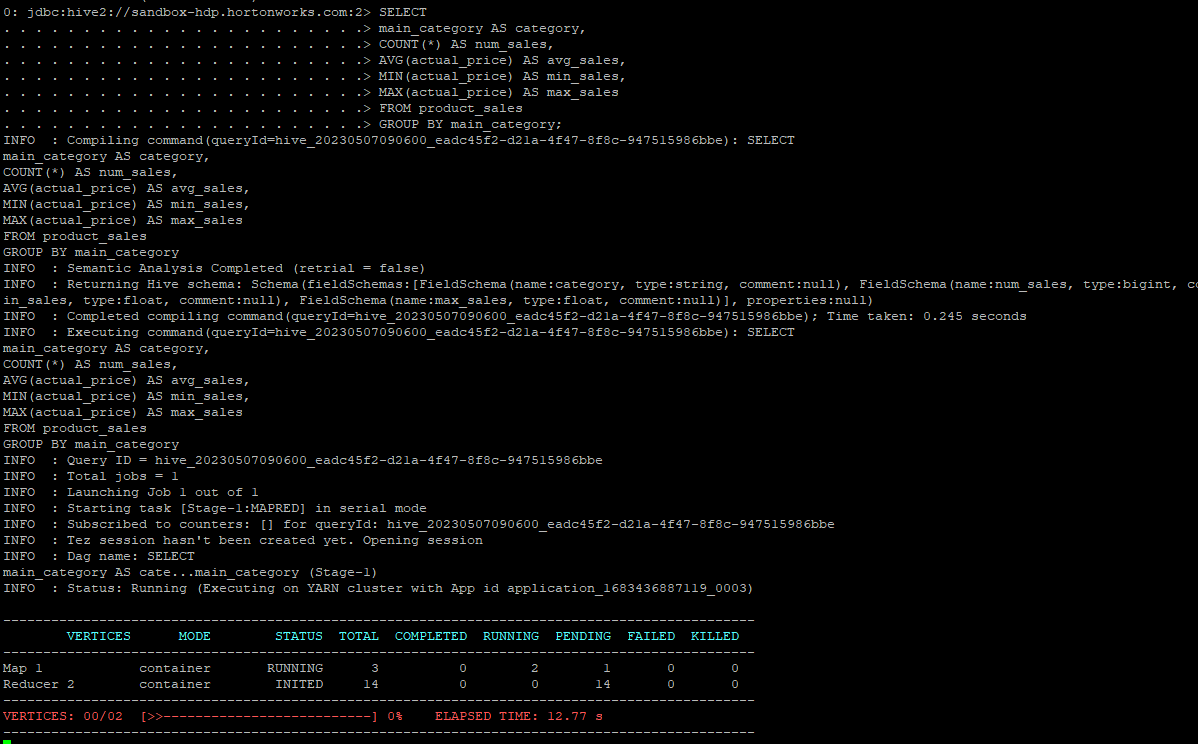
*AVG(sales) AS avg\_sales,*

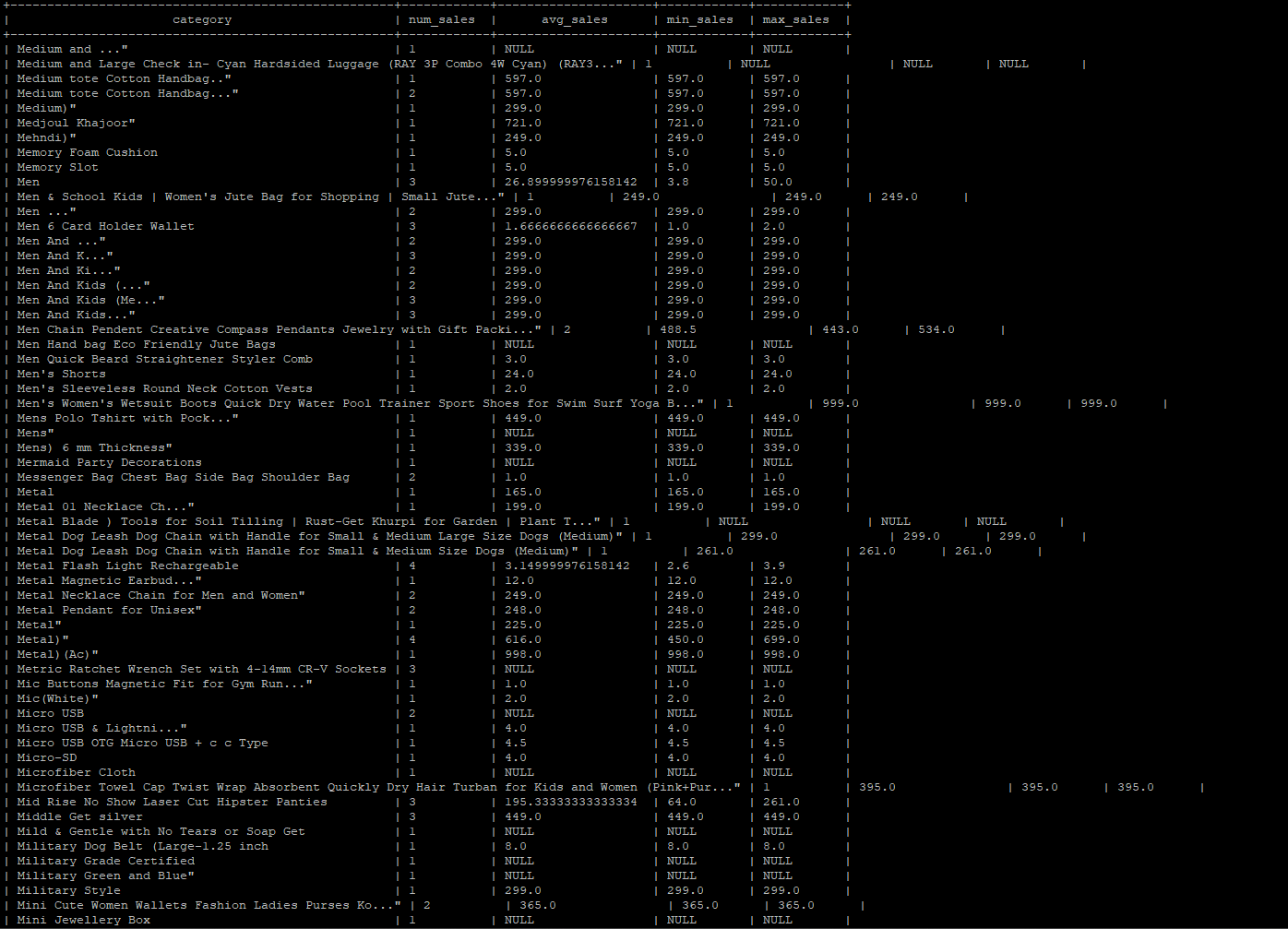
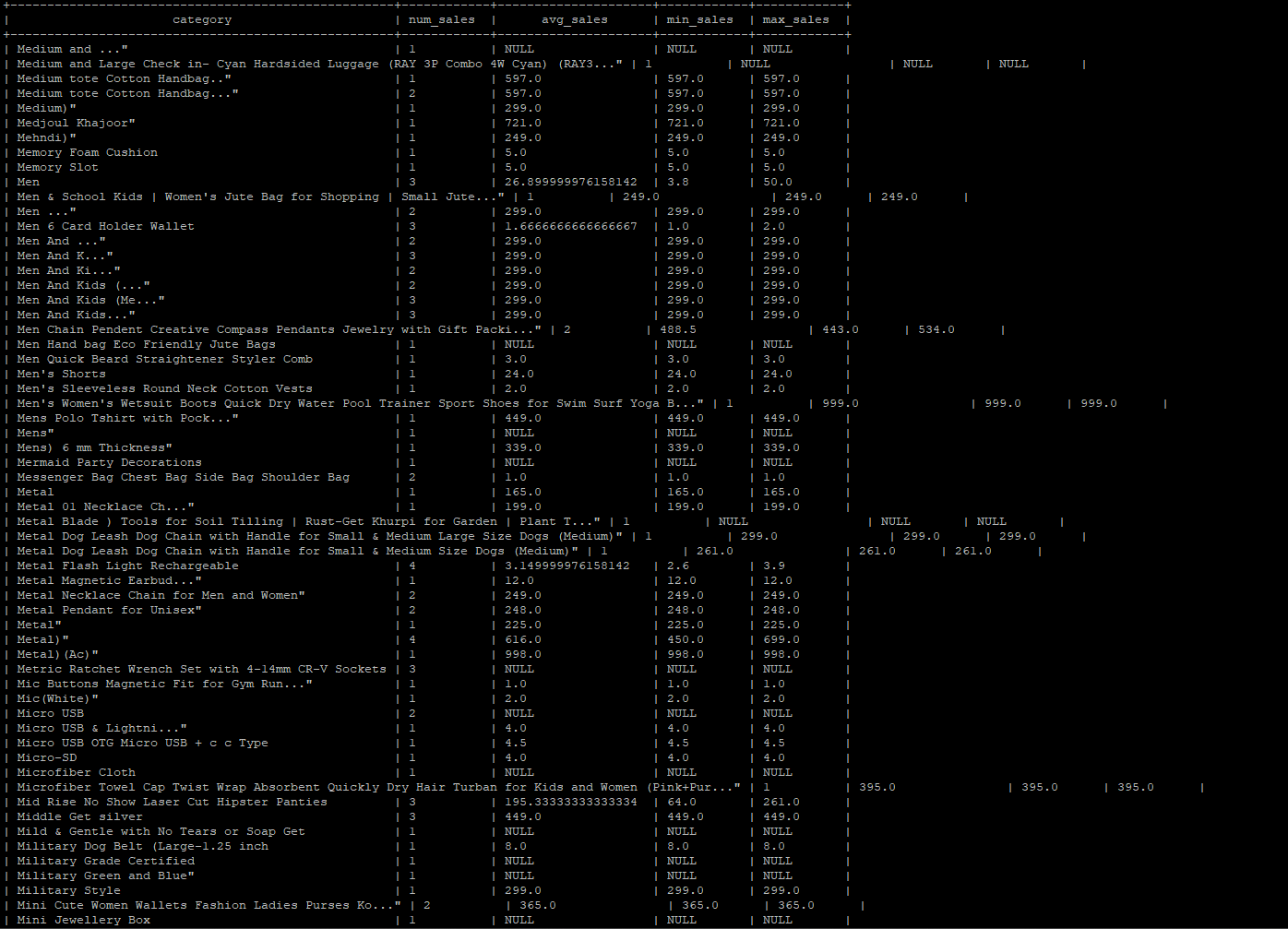
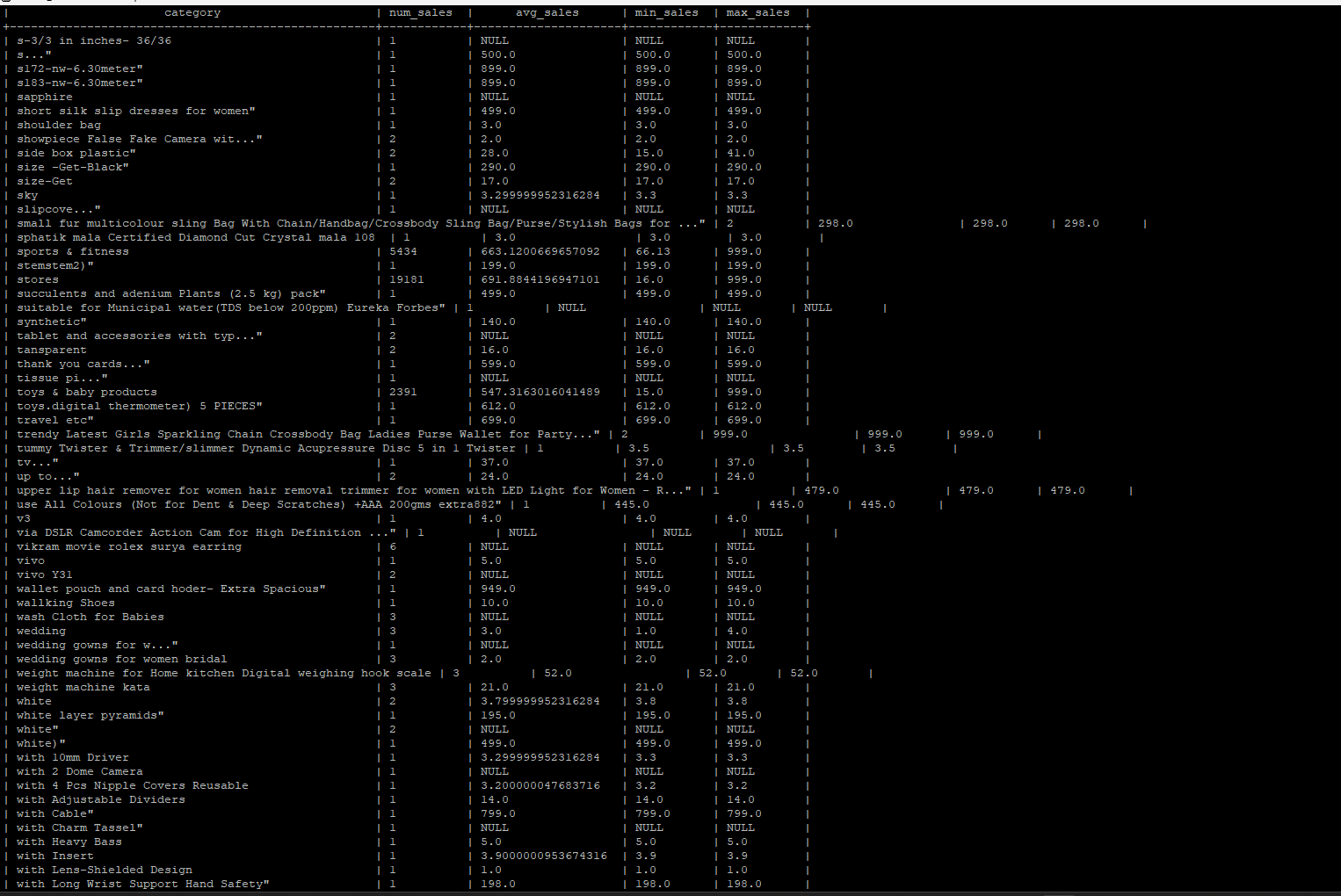
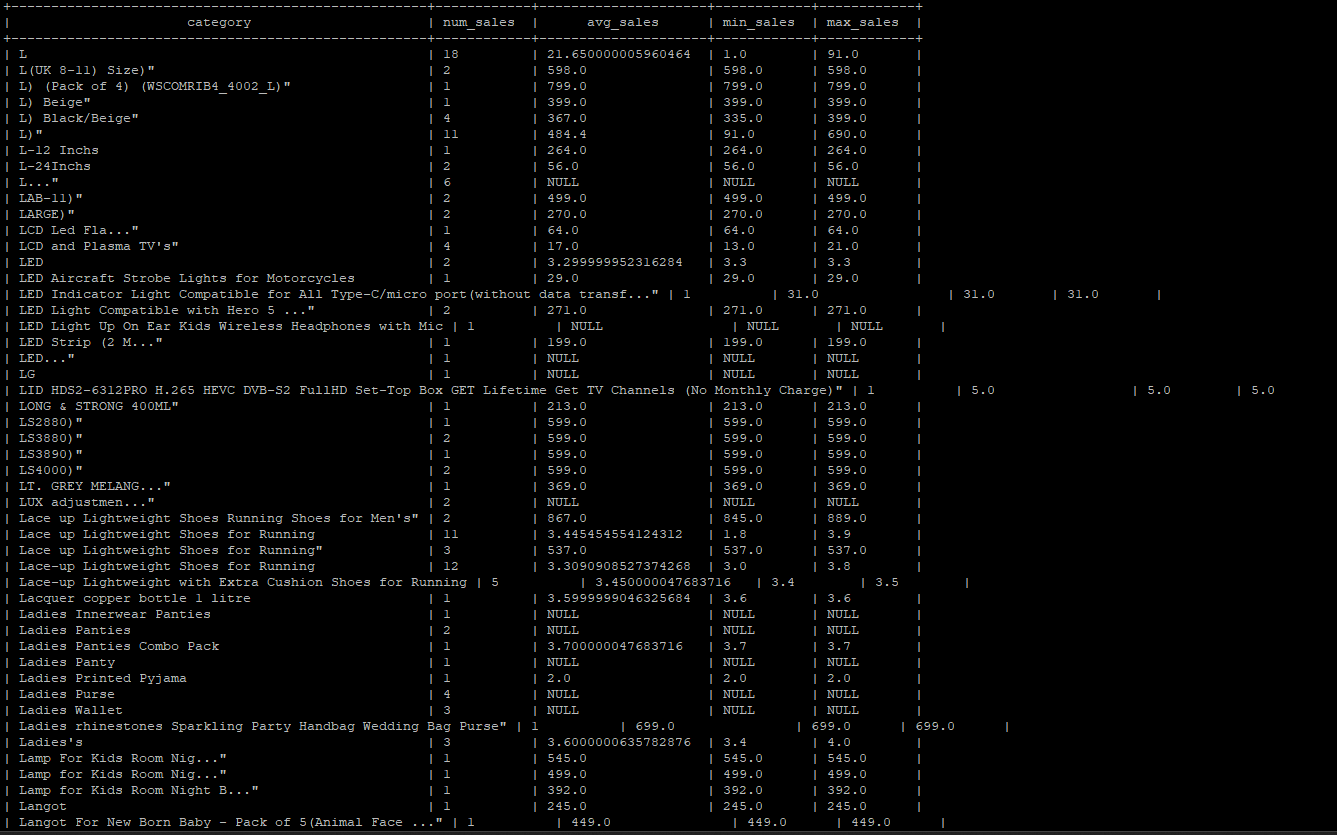
*MIN(sales) AS min\_sales,*

*MAX(sales) AS max\_sales*

*FROM product\_sales*

*GROUP BY category;*





Calculating Top 10 selling Categories

*SELECT*

*main\_category AS category,*

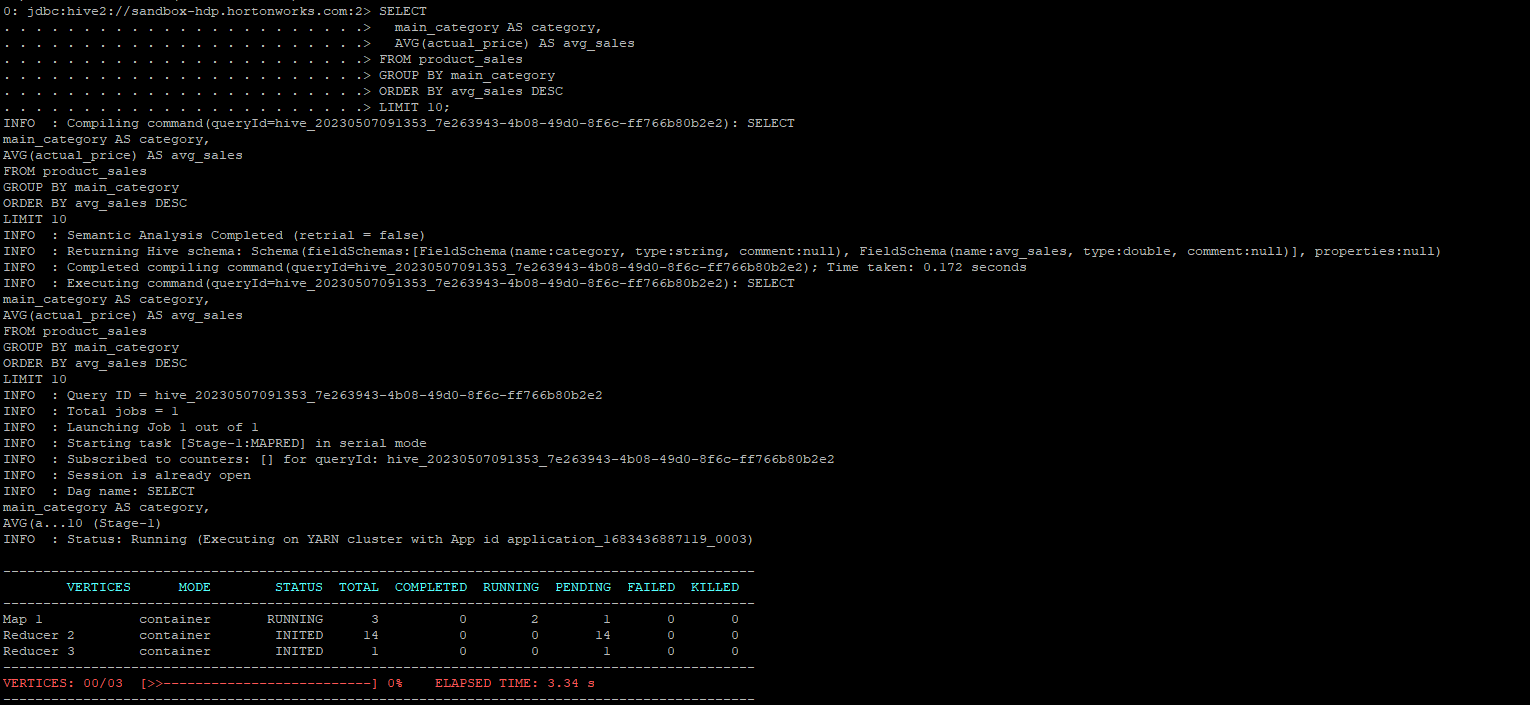
*AVG(actual\_price) AS avg\_sales*

*FROM product\_sales*

*GROUP BY main\_category*

*ORDER BY avg\_sales DESC*

*LIMIT 10;*



# CHAPTER 5 VISUALIZATION

#### ALL GENRES SORTED ACCORDING TO THE SUM OF THE ENERGY:

**Bar Chart**

Chart, bar chart

Description automatically generated

**Stacked Bar Chart**

Chart, bar chart

Description automatically generated

**PIE Chart**

Chart, pie chart

Description automatically generated

**Heatmap**

Chart, treemap chart

Description automatically generated

# CHAPTER 7 CONCLUSION

In conclusion, the analysis of Amazon product sales data using Hive and Tableau has provided valuable insights into customer behavior and product trends. The use of Hive has allowed us to process and analyze large volumes of data in a distributed manner, while the use of Tableau has enabled us to create interactive visualizations and dashboards for better data exploration and communication. Through this analysis, we have identified the top-selling products in each category, the average rating and number of ratings for each product, as well as the discount and actual prices of each product. We have also identified the most popular product categories and sub-categories, as well as the average rating and number of ratings for each category. The insights gained from this analysis can help businesses to make data-driven decisions in areas such as product pricing, marketing, and inventory management. For example, businesses can use this information to optimize their product offerings, improve customer satisfaction, and increase sales.

Overall, the use of Hive and Tableau for analyzing and visualizing Amazon product sales data has provided a powerful tool for businesses to gain insights into customer behavior and product trends, and make data-driven decisions to improve their bottom line.

# REFERENCES

1. "Analyzing Amazon Sales Data with Hive and Tableau" by Thomas Kachura, published in the Journal of Big Data. (2021)
2. "Big Data Analytics using Hadoop, Hive and Tableau for Sales Data of an E-commerce Company" by Ramya M, published in the International Journal of Advanced Research in Computer Science and Software Engineering. (2021)
3. "Analysis of Amazon Product Reviews Using Hadoop and Tableau" by Sai Vamsi Maddipati and Anil Kumar Pasupuleti, published in the International Journal of Emerging Trends & Technology in Computer Science. (2020)
4. "Exploratory Data Analysis of Amazon Product Reviews Using Hive and Tableau" by Vinay Kumar and Praveen Kumar, published in the International Journal of Computer Science and Mobile Computing. (2020)
5. "Big Data Analytics for E-commerce Sales Data using Hive and Tableau" by Sanjay Bhat, published in the International Journal of Advanced Trends in Computer Science and Engineering. (2019)