Utilizing Google Cloud Platform for Big Data Analytics in E-Commerce: An HDFS and Hive Approach

Samuel Verghese (3120522)

Arnold Joseph (3120369)

Karan Zaveri (3121037)

Harshetha Murthy (3121278)

Nikhil R. Gadekar (3121291)

SRH Hochschule Berlin

Index

Introduction

The exponential growth in data generation necessitates advanced tools and frameworks to manage, process, and extract meaningful insights from large datasets. In this research, we explore the use of Hadoop Distributed File System (HDFS) and Apache Hive to handle and analyze big data. Specifically, we use a dataset of e-commerce transactions to demonstrate how big data technologies can be leveraged to gain valuable insights, which can inform business decisions and strategies.

Problem Statement

E-commerce companies generate massive amounts of data daily, including transaction details, customer information, product specifics, and more. Managing and analyzing this data to extract actionable insights is challenging due to its volume, variety, and velocity. This research aims to utilize HDFS for distributed storage and Hive for efficient querying and analysis to address these challenges. The objectives are:

1. Efficiently store and manage large volumes of e-commerce data using HDFS.
2. Create a Hive table to structure the data for easy querying.
3. Analyze the data to extract insights such as total product values, freight costs, and review scores.
4. Visualize the results to facilitate understanding and decision-making.

Importance of Big Data in E-Commerce

The way data is generated exponentially in different sectors calls for the adoption of sophisticated tools and frameworks to manage, process, and extract meaningful insights from big datasets, especially in e-commerce. Several big data technologies provide essential benefits in dealing with such issues, including Hadoop Distributed File System (HDFS) and Apache Hive. The following are some of the key reasons why big data is irreplaceable:

Ability to Work with Large Data Volumes:

E-commerce platforms produce gigantic amounts of data daily regarding transaction details, customer information, product specifics, etc. Traditional data processing systems find it very difficult to manage such a scale. HDFS enables efficient handling and storage of big data distributed across a large number of nodes in a way that is scalable and fault-tolerant.

Efficient Data Processing:

Huge volumes of structured and unstructured data require potent processing frameworks. Hive enables efficient querying and big-data analysis through a SQL-like interface that further makes the task easy for analysts and data scientists. This is so that enterprises can run complex queries and generate actionable insights at a fast pace.

Improved Decision-Making:

The availability of big data analytics helps companies gain valuable insights into customer behavior, market trends, and operational efficiencies. For instance, purchase behavior patterns, regional sales differences, and customer preferences are revealed through e-commerce data. Such insights will go a long way toward optimizing marketing strategies, inventory management, and customer service for informed decision-making.

Enhanced Customer Experience:

Every e-commerce business needs to understand its customers' needs and preferences. Big data analytics can serve to segment customers, personalize recommendations, and target marketing campaigns on a more explicit basis. In the same way, analysis of review scores and feedback will help improve the products or services of a business for an enhanced level of customer satisfaction.

Cost Efficiency and Resource Optimization:

The infrastructure for big data is expensive to establish, but in the long run, it helps reduce costs by optimizing operations. For example, freight and logistics data can be mined to identify more economical routes and methods of shipping, with less transport cost and improved delivery times.

Scalability and Flexibility:

Big data technologies like HDFS and Hive are designed for horizontal scaling. For example, as the volume of data increases, more nodes can be added to their clusters with only minor modifications to the existing infrastructure. This flexibility ensures that, with each step of the increase in data load, the system processes its processing work efficiently.

Insight into Case Study:

The project "Leveraging HDFS and Hive for Big Data Analytics in E-Commerce" demonstrates the use of these technologies. The current research, indicating how e-commerce data can be processed and analyzed, applies HDFS for distributed storage, data structuring, and querying in Hive. The insights gained into the values of products, costs of freight, and customer review scores are visualized for the end user so they can be better understood and decisions made. The presented case exemplifies the following: through big data, it is possible to make e-commerce operations work more efficiently and optimize strategic business decisions.

Methodology

The research methodology involves several steps, including setting up the HDFS environment, creating Hive tables, loading and transforming data using PySpark, and visualizing the results. Below are the detailed steps:

HDFS Setup

1. **Update and install Google Cloud SDK**:

```
sudo apt-get update

sudo apt-get install google-cloud-sdk
```

2. **Authenticate and copy dataset to cluster**:

```
gcloud auth login

gsutil cp gs://ecom-buck2806/ecom-dataset.csv .
```

3. **Create HDFS directory and upload dataset**:

```
hdfs dfs -mkdir -p /user/ecom-project

hdfs dfs -put ecom-dataset.csv /user/ecom-project/
```

Hive Setup

1. **Create and use a Hive database**:
   ```
   CREATE DATABASE ecom_project;
   USE ecom_project;
   ```

2. **Create a Hive table for the dataset**:
   ```
   CREATE TABLE olist_data (
     Id INT,
     order_status STRING,
     order_products_value FLOAT,
     order_freight_value FLOAT,
     order_items_qty INT,
     customer_city STRING,
     customer_state STRING,
     customer_zip_code_prefix INT,
     product_name_length INT,
     product_description_length INT,
     product_photos_qty INT,
     review_score INT,
     order_purchase_timestamp STRING,
     order_approved_at STRING,
     order_delivered_customer_date STRING
   )
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   STORED AS TEXTFILE;
   ```

3. **Load data into the Hive table**:
   ```
   LOAD DATA INPATH '/user/ecom-project/ecom-dataset.csv' INTO TABLE olist_data;
   ```

Data Processing with PySpark

1. **Define schema and load data into DataFrame**:

```
from pyspark.sql.types import *
from pyspark.sql.functions import col, to_timestamp

schema = StructType([
   StructField("Id", IntegerType(), True),
   ...
   StructField("order_delivered_customer_date", StringType(), True)
])

df = spark.read.format("csv").option("sep", ",").option("header",
"true").schema(schema).load("hdfs:///user/ecom-project/ecom-dataset.csv")
```

2. **Convert string columns to timestamp**:

```
df = df.withColumn("order_purchase_timestamp",
to_timestamp(col("order_purchase_timestamp"), "dd-MM-yy HH:mm")) \
      .withColumn("order_approved_at", to_timestamp(col("order_approved_at"), "dd-MM-yy
HH:mm")) \
      .withColumn("order_delivered_customer_date",
to_timestamp(col("order_delivered_customer_date"), "dd-MM-yy HH:mm"))
```

3. **Add day and week columns**:

```
from pyspark.sql import functions as F

df = df.withColumn("day", F.dayofmonth("order_purchase_timestamp"))
df = df.withColumn("week", F.weekofyear("order_purchase_timestamp"))
```

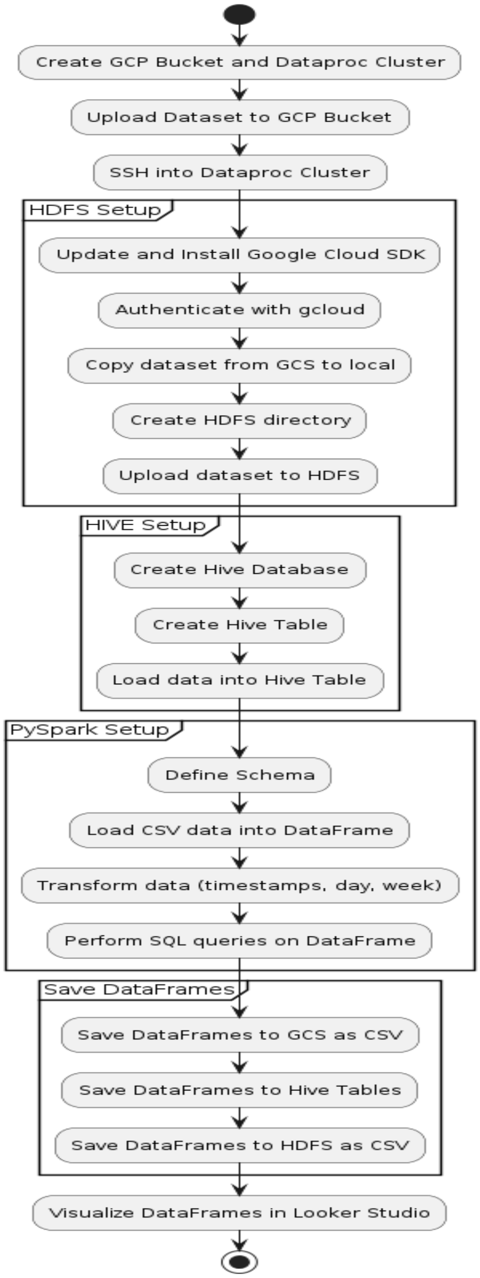
4. **Create aggregated DataFrames**:

   ```
   df1 = spark.sql("""
   SELECT
      id,
      order_status,
      SUM(order_products_value) AS total_products_value,
      SUM(order_freight_value) AS total_freight_value,
      customer_city,
      day
   FROM
      solution
   GROUP BY
      id, order_status, customer_city, day
   """)
   ```

5. **Save results to Google Cloud Storage and Hive**:

   ```
   df1.coalesce(1).write.mode("overwrite").option("header", "true").csv("gs://ecom-buck2806/processed_data1")
   ...
   spark.sql("USE ecom_project")
   df1.write.mode("overwrite").saveAsTable("ecom_project.df1")
   ```

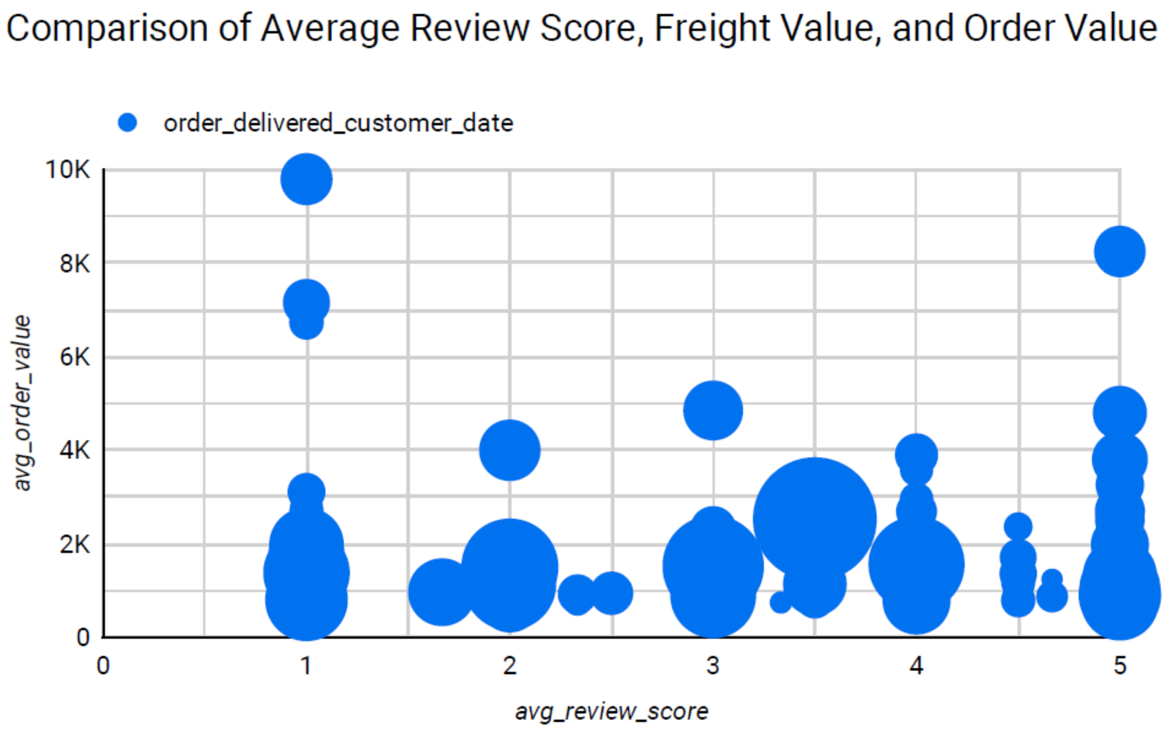
Flow Chart and Explanation


Create GCP Bucket and Dataproc Cluster
Upload Dataset to GCP Bucket
SSH into Dataproc Cluster
HDFS Setup
Update and Install Google Cloud SDK
Authenticate with gcloud
Copy dataset from GCS to local
Create HDFS directory
Upload dataset to HDFS
HIVE Setup
Create Hive Database
Create Hive Table
Load data into Hive Table
PySpark Setup
Define Schema
Load CSV data into DataFrame
Transform data (timestamps, day, week)
Perform SQL queries on DataFrame
Save DataFrames
Save DataFrames to GCS as CSV
Save DataFrames to Hive Tables
Save DataFrames to HDFS as CSV
Visualize DataFrames in Looker Studio

The flowchart illustrates the data processing pipeline on the Google Cloud Platform. It begins with data upload to a Google Cloud Storage bucket, followed by data processing in a Dataproc cluster using HDFS and Hive, and ends with data analysis and visualization using PySpark and Google Cloud Storage for output storage.

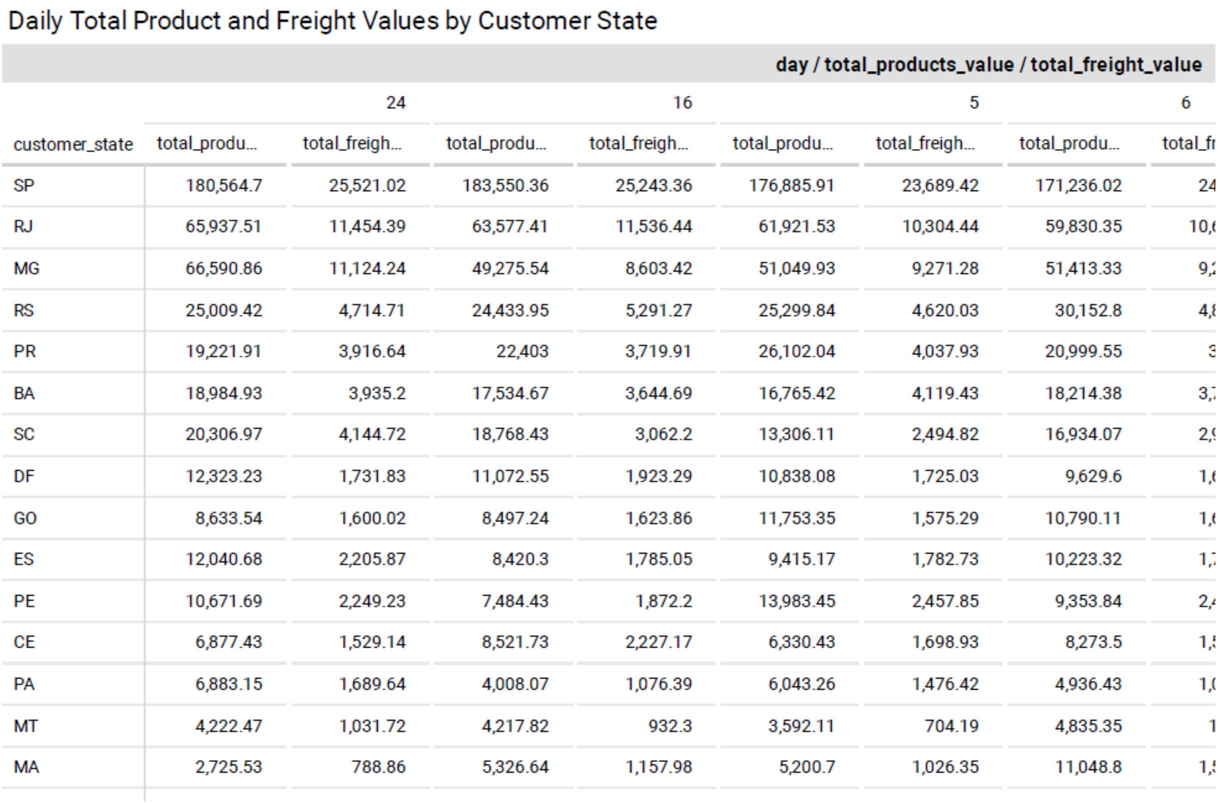
Results and Visualization

The data analysis revealed several key insights, visualized using charts and graphs. These visualizations help in understanding the underlying patterns and trends in the e-commerce data, providing a clear picture of how different factors impact the overall business performance.



This chart compares the average review scores of products with their corresponding freight and order values.

Insights: Higher order values and freight costs might relate to a higher review score, implying customer satisfaction with more premium products or quicker shipping service. Variations in the review scores for different order values and freight costs would further improve product quality or shipping efficiency.

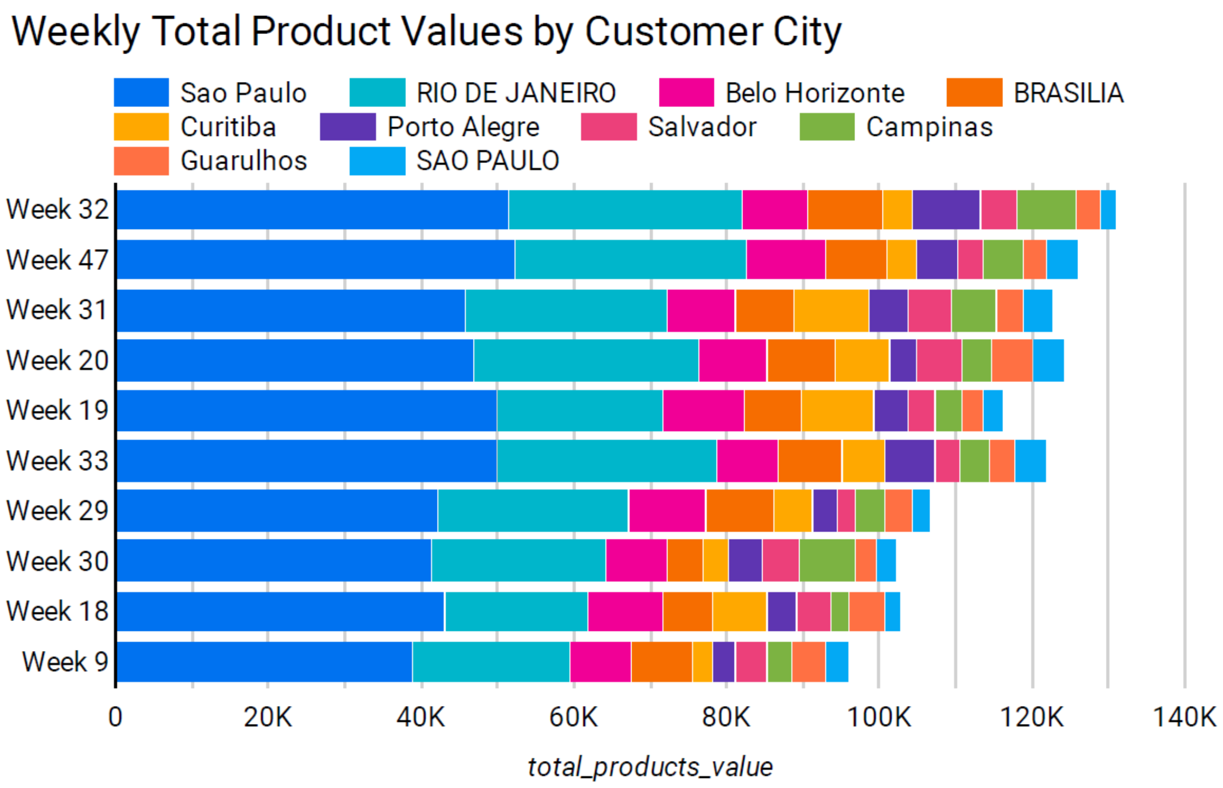


Daily Total Product and Freight Values by Customer State

	day / total_products_value / total_freight_value							
	24		16		5		6	
customer_state	total_produ...	total_freigh...	total_produ...	total_freigh...	total_produ...	total_freigh...	total_produ...	total_fr
SP	180,564.7	25,521.02	183,550.36	25,243.36	176,885.91	23,689.42	171,236.02	24
RJ	65,937.51	11,454.39	63,577.41	11,536.44	61,921.53	10,304.44	59,830.35	10,6
MG	66,590.86	11,124.24	49,275.54	8,603.42	51,049.93	9,271.28	51,413.33	9,2
RS	25,009.42	4,714.71	24,433.95	5,291.27	25,299.84	4,620.03	30,152.8	4,8
PR	19,221.91	3,916.64	22,403	3,719.91	26,102.04	4,037.93	20,999.55	3
BA	18,984.93	3,935.2	17,534.67	3,644.69	16,765.42	4,119.43	18,214.38	3,7
SC	20,306.97	4,144.72	18,768.43	3,062.2	13,306.11	2,494.82	16,934.07	2,9
DF	12,323.23	1,731.83	11,072.55	1,923.29	10,838.08	1,725.03	9,629.6	1,6
GO	8,633.54	1,600.02	8,497.24	1,623.86	11,753.35	1,575.29	10,790.11	1,6
ES	12,040.68	2,205.87	8,420.3	1,785.05	9,415.17	1,782.73	10,223.32	1,7
PE	10,671.69	2,249.23	7,484.43	1,872.2	13,983.45	2,457.85	9,353.84	2,4
CE	6,877.43	1,529.14	8,521.73	2,227.17	6,330.43	1,698.93	8,273.5	1,5
PA	6,883.15	1,689.64	4,008.07	1,076.39	6,043.26	1,476.42	4,936.43	1,0
MT	4,222.47	1,031.72	4,217.82	932.3	3,592.11	704.19	4,835.35	1
MA	2,725.53	788.86	5,326.64	1,157.98	5,200.7	1,026.35	11,048.8	1,5

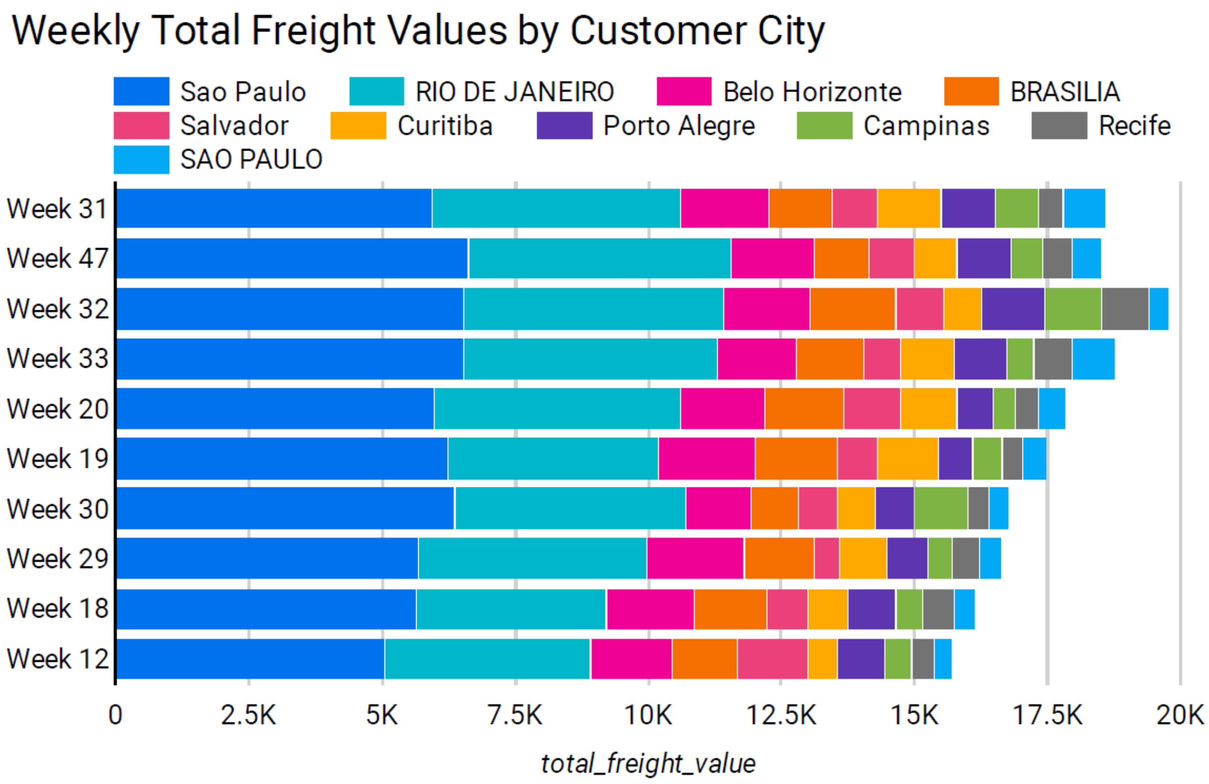
This chart represents daily total values of products sold and freight costs incurred, disaggregated by the customer state.

Insights: States that are showing consistently high values concerning products and freight can be recommended for marketing, with logistics improvements. Knowing states with changing values might be the reason behind seasonal trends or probable delivery inefficiencies.



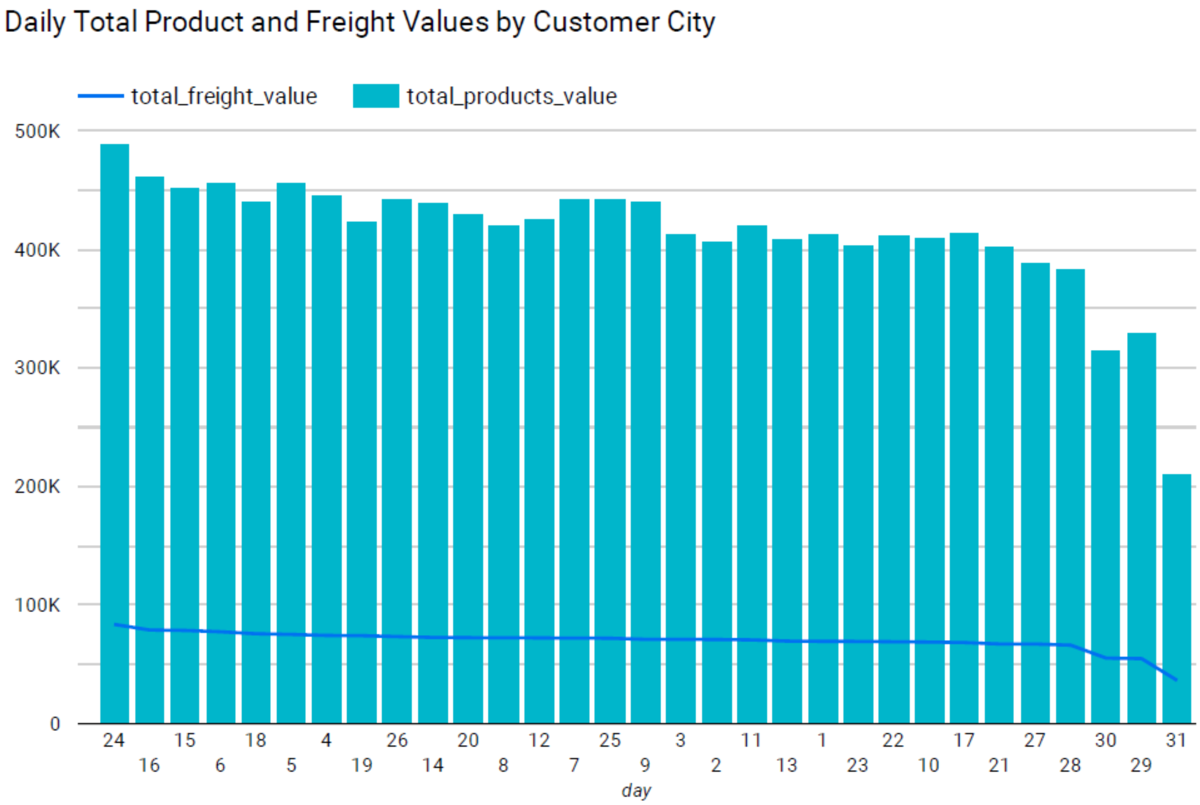
This graph compares the weekly products' sum values of each customer city.

Insights: High weekly sale cities can be targeted for promotional campaigns and maintaining inventory. This is going to enable us to get sales trends across cities and, therefore, allocate resources and schedule marketing efforts more accurately.



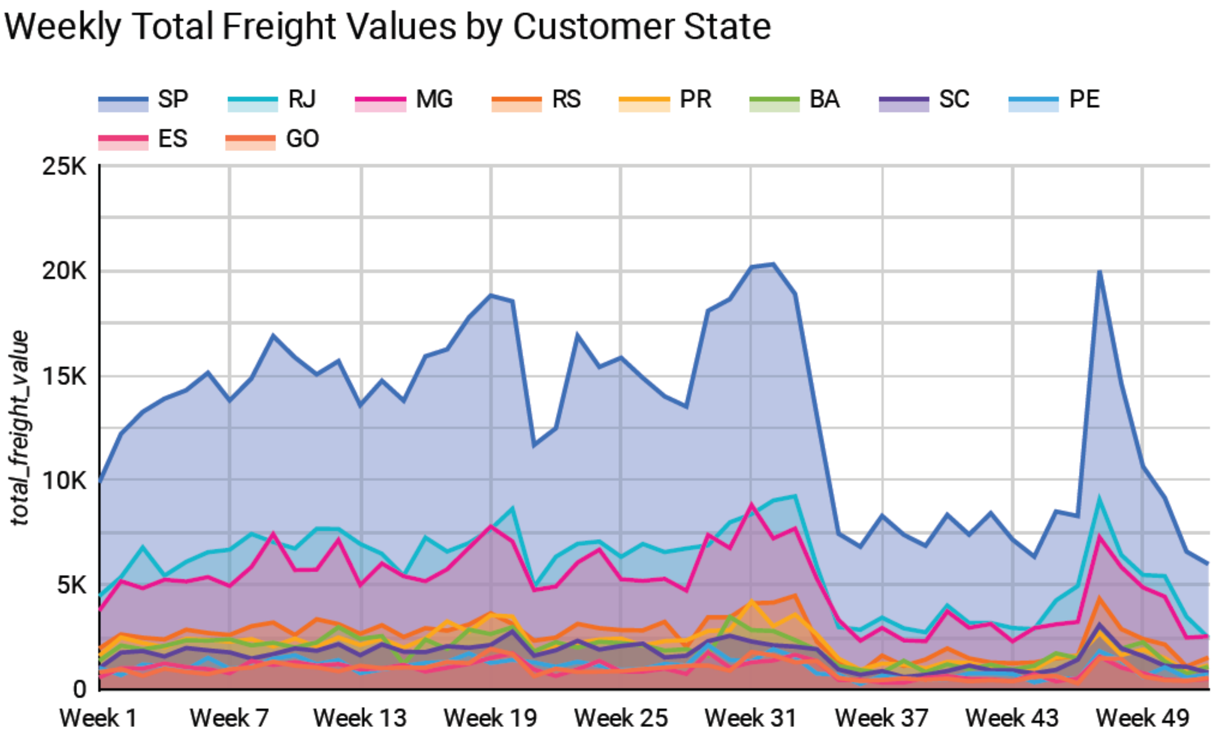
Total freight values every week concerning each customer city are illustrated.

Insights: A city with a cost at the higher end might be working through optimized shipping routes or using some logistic strategy for such reduced costs. These results will allow an overall examination and even the possibility of diminishing the general shipping expenses.



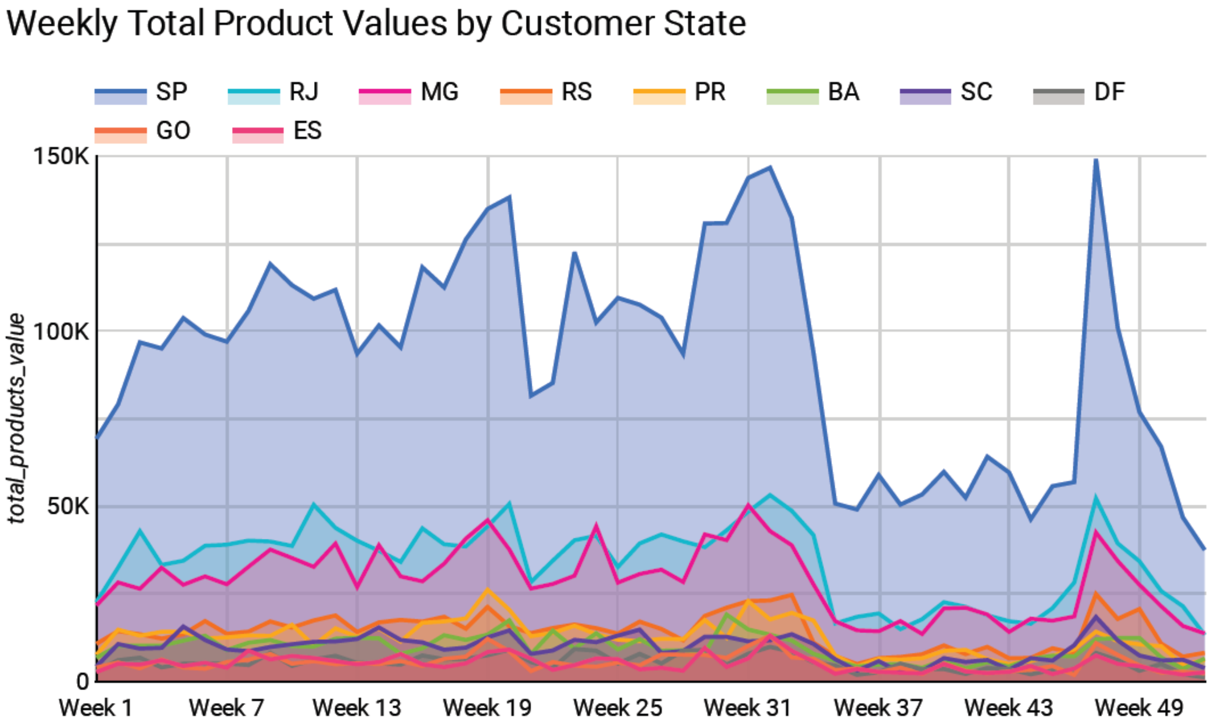
This chart shows the daily sum of sold product values and their corresponding freight costs

Insights: Those cities that have constantly shown high product value and freight cost can be regarded as big markets. You can increase your sales and customer satisfaction by concentrating your marketing and inventory resources in these cities. Cities with expensive costs of shipping about products may gain from optimizing shipping strategies—for example, shipping consolidation or negotiating better rates with carriers. It enables an understanding of peak shopping days, which can be targeted for promotions and inventory planning.



This chart represents the total freight values incurred every week, categorized by customer state.

Insights: The better logistical planning approach for the cost reduction would involve states with higher total weekly freight values. To these ends, maybe looking at local warehousing, delivery route optimization, and other such initiatives would be helpful. The freight values should be checked regularly, and according to that, the efficiency of implemented logistics strategies should be checked over time. The logistics strategies can be specifically developed for the states where the freight values are not constant; this will help to manage the resources properly and improve the delivery timeline.



This graph describes the total number of product values sold and groups them by the customer state it is sold to.

Insights: States with a high product value play a vital role as key markets. Such states should be targeted for inventories, marketing campaigns, and efforts to engage customers. Knowledge about product value trends in weeks enables forecasting demand so that better supply chain and inventory management can be done. In states where the values of products are consistently rising, launches of new product or promotional activities should be initiated and taken up to catch those consumer trends.

Discussion

The analysis provided several insights into the e-commerce transactions dataset. For instance, understanding the distribution of product values and freight costs across different cities and states can help optimize logistics and pricing strategies. The comparison of review scores with order and freight values can inform customer satisfaction and service improvement initiatives. However, the analysis also highlighted some challenges, such as data quality issues and the need for efficient processing frameworks for large datasets.

Merits and Demerits

Merits

1. **Scalability**: HDFS and Hive allow for efficient handling of large datasets.
2. **Flexibility**: PySpark provides powerful data processing capabilities.
3. **Insightful Visualization**: Visual representations aid in a better understanding of data.

Demerits

1. **Complexity**: Setting up and managing the big data infrastructure requires technical expertise.
2. **Resource Intensive**: Big data processing can be resource-intensive and costly.
3. **Data Quality**: Ensuring data quality and consistency can be challenging.

Conclusion

This research demonstrates the effectiveness of using HDFS and Hive in managing and analyzing large-scale e-commerce data. By leveraging these technologies, businesses can gain valuable insights that can drive strategic decisions and improve operational efficiency. The findings also underscore the importance of robust data processing pipelines and the potential benefits and challenges of big data analytics.