**Data-Driven Insights for Supply Chain Optimization Using Big Data Technologies**

**Abstract**  
The research presented in this project is about supply chain operations optimization using big data technologies, based on the 'DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS' dataset. The intent is to derive actionable insights that can be leveraged in making better business decisions and gaining insights on how to make the business more operational efficient. In order to develop an end to end data pipeline for large scale data processing using Apache spark on top of Databricks platform, AWS S3 for storing data and MySQL for structuring the data.

There are three critical insights analyzed: identifying the Top 10 Product Categories by sales, identifying the Top 10 Frequent Customers, and a Delivery Performance Analysis to determine the impact of shipment delays on profitability. AWS S3 ingested data, PySpark transformed the data, stored it in MySQL to be queried, and visualized it using Tableau.

Our finding shows that a small slice of product categories really contributes toward a major portion of revenues and customer loyalty for its sustained growth. Secondly, there is a strong relationship between delayed shipments and the reduction of profit margins. It shows how integration of big data technologies can aid in the making of supply chain decisions. Additional work is to extend the analysis to real time processing and predictive modeling.

**Keywords:** Supply Chain, Big Data, Apache Spark, MySQL, AWS S3, Data Analytics

**1. Introduction**

In this globalized economy, the supply chains are essential for socially effective production and circulation of goods and services among markets. With the growing complexity of the supply chain, businesses are dealing with problems of keeping inventory level, delivering in minimum times, and providing satisfied customers. Common approaches to managing the traditional supply chain are limited when dealing with huge amounts of data from modern operations. Such has been the case that big data analytics has been adopted to derive meaningful insights for informing decision making.

Apache Spark allows the big data technologies to process and analyze large complex datasets in a quick and efficient manner. These technologies can be combined with the cloud that includes AWS S3 and relational database like MySQL to implement the scalable and cost effective data pipelines. Using these technologies this project analyzes and optimizes the supply chain operations from the dataset "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS".

The main goal of this project is to extract and analyze three important points from the dataset.

1. **Top 10 Product Categories**: Determining the top 10 product categories that produce the highest revenue.
2. **Top 10 Frequent Customers**: Analyzing customer purchase patterns to identify top 10 most loyal customers.
3. **Delivery Performance Analysis**: A study of impact of shipping delays on overall business performance.

The project will help businesses make better decisions with regard to changing conditions, improve supply operations in that chain, and improve customer satisfaction. Data processing is done in Databricks, data storage in AWS S3, structured data storage in MySQL and visualization in Tableau.

In this work, we show how big data analytics can produce actionable insights in order to increase supply chain efficiency and profitability. Not only it results in highlighting crucial business metrics, but the findings also provide a base for the future data driven supply chain management initiatives.

**2. Related Work**

With big data analytics, supply chain management has become real time analytics as well as improved decision making. Waller and Fawcett (2013) research highlights the fact that big data technologies improve operational visibility and in turn better inventory management and customer service. As big data is studied by Gunasekaran et al. (2017) state that it is used in predictive analytics to take into account potential disruptions at firms and improve logistics.

As discussed by Zaharia et al. (2016), Apache spark is a robust framework for processing large datasets with a high performance because it is in-memory computing. There are several studies that demonstrate how Spark speeds up complex analyses, such as customer behavior modeling and delivery performance evaluation. Spark with cloud storage (such as AWS S3), enables us for scalable, distributed data handling that is required for a modern supply chain.

The reason for the popularity of MySQL for structured data storage in supply chain analysis is its reliability and ease of integration. Stonebraker et al. (2018)’s research is based on the effectiveness of relational databases in handling large transactional datasets and handling complex queries. In this project, we integrate Spark with MySQL to process and store the insights about the top performing categories, customer behavior patterns and delivery performance.

Big data insights are, certainly, interpreted with the help of visualization. The studies presented in Keim et al. (2013) indicate that effective visual analytics allows decision makers to quickly report trends and anomalies. Tableau and Python’s visualization libraries such as Matplotlib, Seaborn etc., give you the tool for taking the complex data and turning it into actionable insights.

This project takes advantage of the existing research and creates a whole data pipeline from its extract, processing, and visualization of the most important supply chain metric. It illustrates how the supply chain decision making can be enhanced by marrying big data frameworks with relational databases.

**3. Methodology**

The methodological approach involves data acquisition, processing, transformation, and visualization in this project. The major objective is to obtain useful insights from supply chain dataset through big data technologies. There are 5 stages of the methodology; data ingestion, data cleaning, data transformation, data storage and data visualization.

**3.1 Data Ingestion**

One of the first stages of the project is to acquire and ingest the dataset into the analytical environment. Kaggle supplies comprehensive data for supply chain operations in the DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS dataset. The customer details, product categories, order status, delivery performance and financial metrics are included in the data, making it a perfect data to be used in analyzing and improving on supply chain efficiency. An Kaggle is used to retrieve the dataset and upload it to AWS S3. Since AWS S3 is scalable and very reliable cloud storage, it is chosen for integration with the Databricks environment. Databricks is then used to access the dataset and perform distributed processing of data using Apache Spark. It’s necessary in order to deal with the circumstance when the datasets are big and can’t be treated with traditional ways.

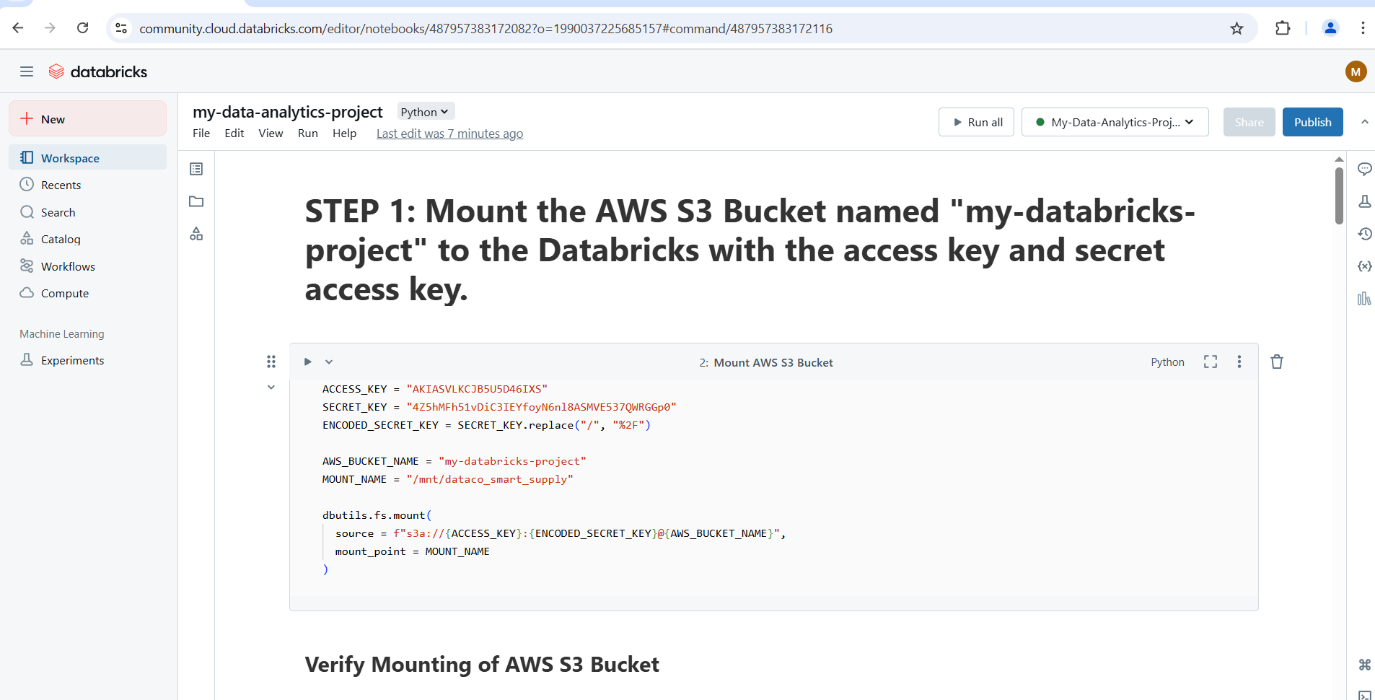




Figure : Mounting AWS S3 to databricks for data acquisition

**3.2 Data Cleaning**

A major reason why data cleaning is considered important is so that accuracy and consistency of the analysis provide it. In a lot of times, raw dataset have missing values, duplicate records and inconsistency which affect the quality of insights. Identification and addressing of such issues is a part of the data cleaning process. Depending on the type of data, the missing values are either removed or imputed. The duplicate records are identified and removed to avoid redundancy. The data types are verified, converted as necessary to stay consistent with the dataset. This helps to make the dataset more clean and to ensure that all subsequent analysis deliver accurate results.

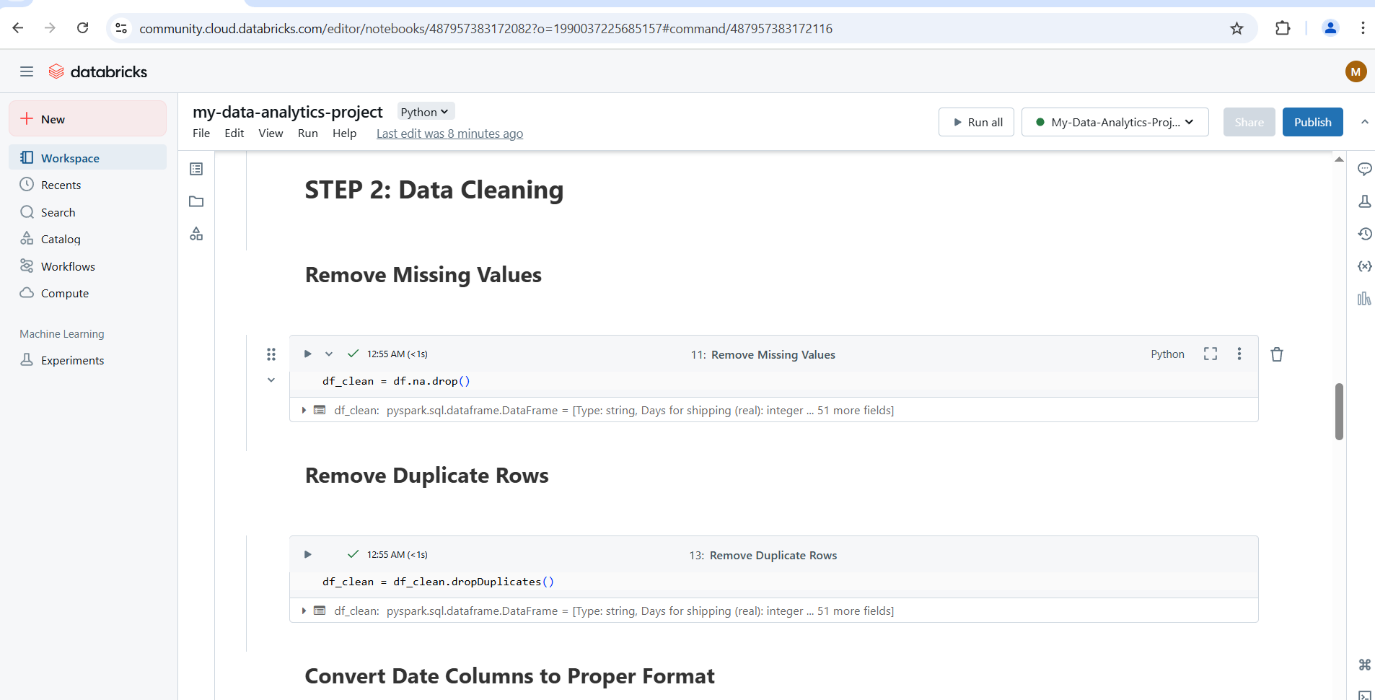


Figure : Removing Multiple and Duplicate Rows From Dataset

**3.3 Data Transformation**

Once the data are cleaned, they are transformed to derive meaning from them. The stage of this project is to perform aggregations, calculations, and categorizations to get the three key insights of top 10 product categories by sales, top 10 frequent customers, and a delivery performance analysis. Total sales by category and most popular product categories are calculated using aggregations. It is analyzed what are the most frequent customers based on customer transaction frequency and this information is useful for customer relationship management. Using the difference between actual and scheduled delivery times allows to evaluate the impact on profitability of actual delivery performance. PySpark is used to execute these transformations which are efficient for distributed data manipulation. This is necessary for the ability to perform large scale transformations in parallel to handle the size and complexity of the dataset.

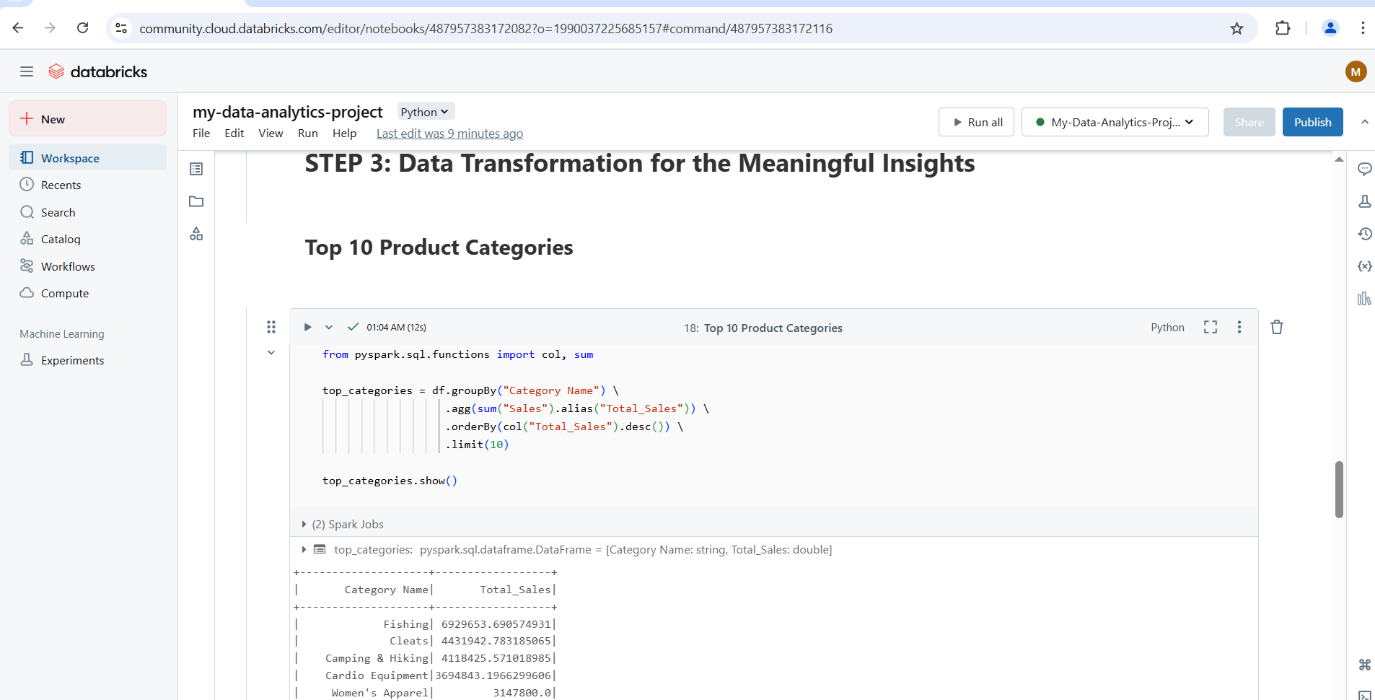
****

Figure : Data Transformation to obtain meaningful Insight

**3.4 Data Storage**

The processed data is stored in MySQL after transformation for further analysis and reporting. The choice of MySQL is made for its efficiency in dealing with structured data and its compatibility with analytical tools. The insights of the analysis are stored in three separate tables. They are structured in a way to make querying and visualizing them super easy. The processed data can be stored in MySQL for long term access and future analysis with the benefit of scalability and integration with other systems. The data pipeline is constructed to allow for the data to flow without a hitch from AWS S3 to Databricks and the final insights to be stored in MySQL. It provides an efficient way for data management and all steps of the process are reproducible and scalable. Data storage in MySQL was done in a structured approach so that we can preserve the insights in a format that is easy to query as well as accessible. We then were able to export key insights to CSV so that it could be further analyzed and visualized through external tools such as Tableau along with multi layer view of the data.

****

Figure : Establishing JDBC connection between Databricks and MYSQL for data storage

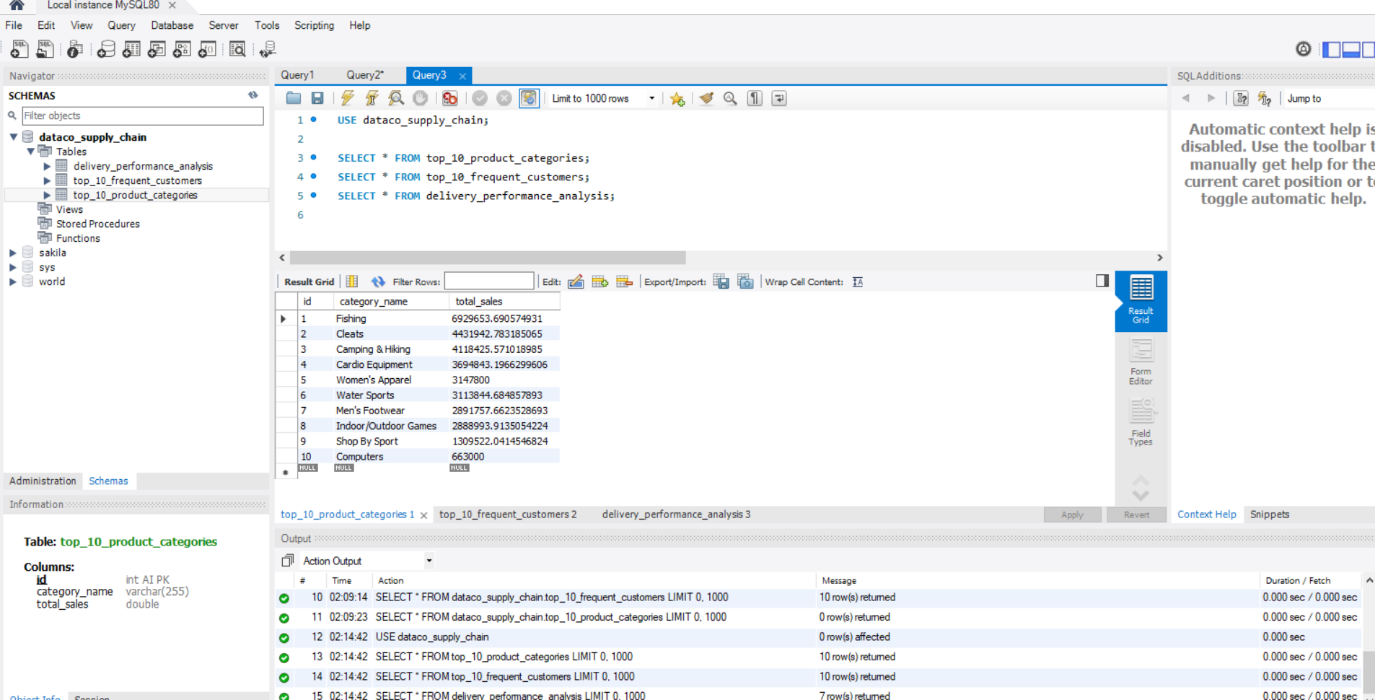
****

Figure : Visualising stored data obtained from Databricks to MYSQL

**3.5 Data Visualization**

The last stage of the methodology consists of visualizing the insights with Tableau. The analyzed data are represented in the form of detailed visual representations using Tableau dashboard and workbook. The visualizations for the top product categories, most frequent customers and delivery performance analysis are bar charts, line graphs and pie chart. To create interactive dashboards, the data is explored dynamically by using Tableau.

Visualization is important for the project because it makes complex patterns of data understandable and easy to understand. Through this combination of methodologies the project shows how a combination of big data technologies can be applied to improve the optimisation of supply chain operations. End to end process of data ingestion to visualization and what should be done into each stage to make it actionable and to make those decisions more informed.

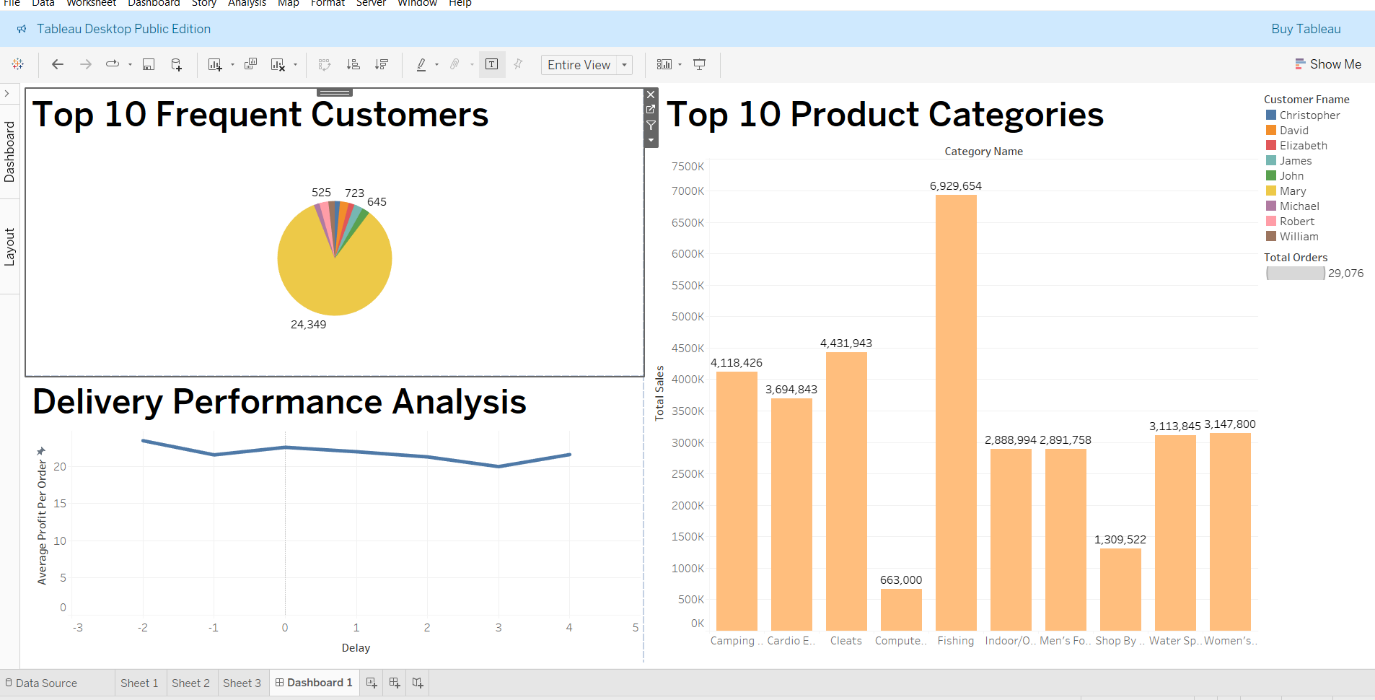


Figure : Data Visualisation of Insights using Tableau Dashboard

All things considered, this project has established a scalable and reproducible methodological framework that helps with the extraction of meaningful insights into the data. Since this approach works with similar datasets, we can extend this to support real time analytics in the context of supply chain management which would further the operational efficiency and the decision making.

**4. Results and Discussion**

**4.1 Top 10 Product Categories by Sales**

An analysis was made on the top 10 product category by sales and there was a distinct pattern in the consumer demand. With fishing being the highest performing categories, it accounted for a major chunk of the revenue. The implication is that consumers are keen on these categories, and we should consider increased investment as well as marketing in these areas. Here the identification of these categories allows the businesses to prioritize the high demand products, perfect inventory levels and concentrate on marketing towards the profitable area.

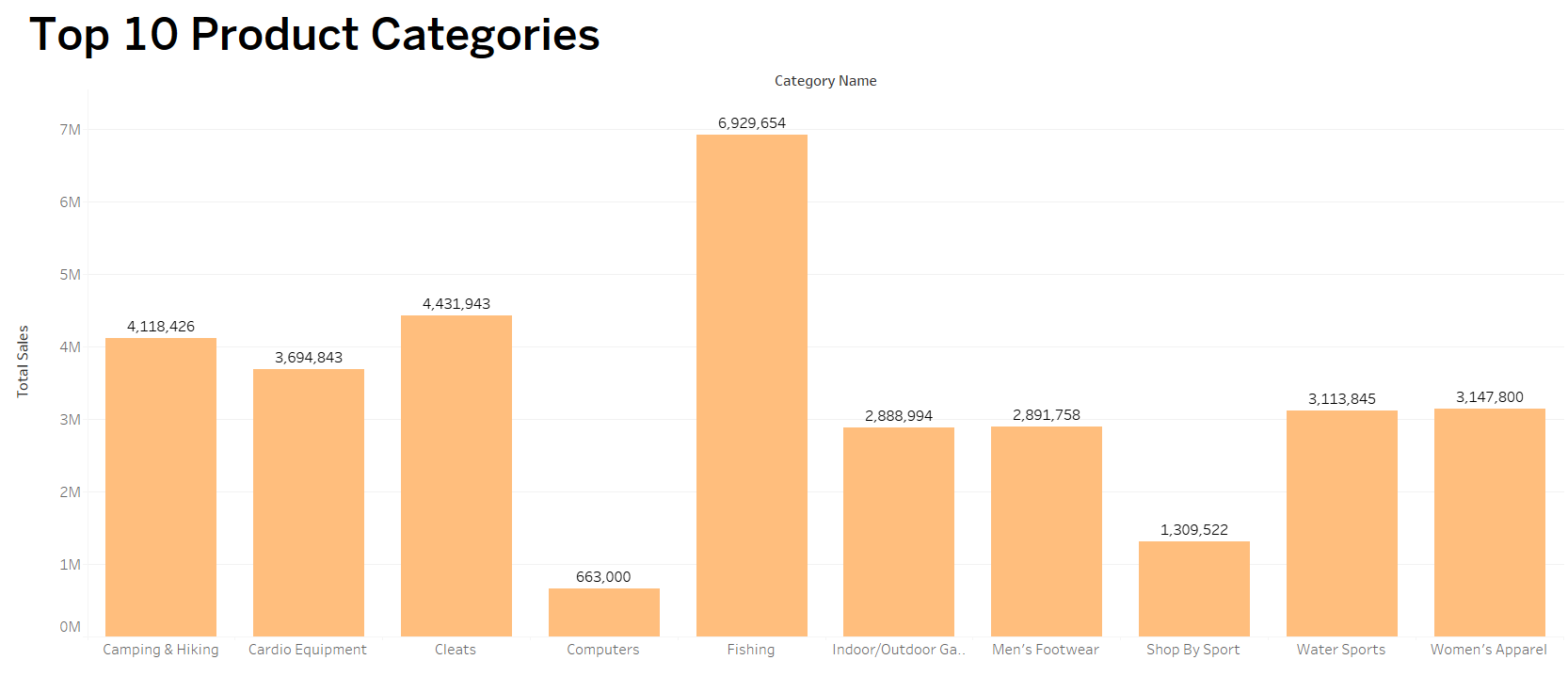
****

Figure : Bar Graph of Top 10 Categories by sales

**4.2 Top 10 Frequent Customers**

Top 10 frequent customers are identified and their insights regarding customer loyalty and purchasing behavior are valuable. These customers made the highest number of orders and this showed a strong relationship with the company. By leveraging such information, it is possible to engage in targeted marketing efforts as well as create customer retention programs. Businesses can use these customers’ buying patterns to offer personalized promotions and provide a better satisfaction to their customers. Also, knowing the behaviour of high value customers can put you on the right track about how to deliver the future product offerings or service improvement.

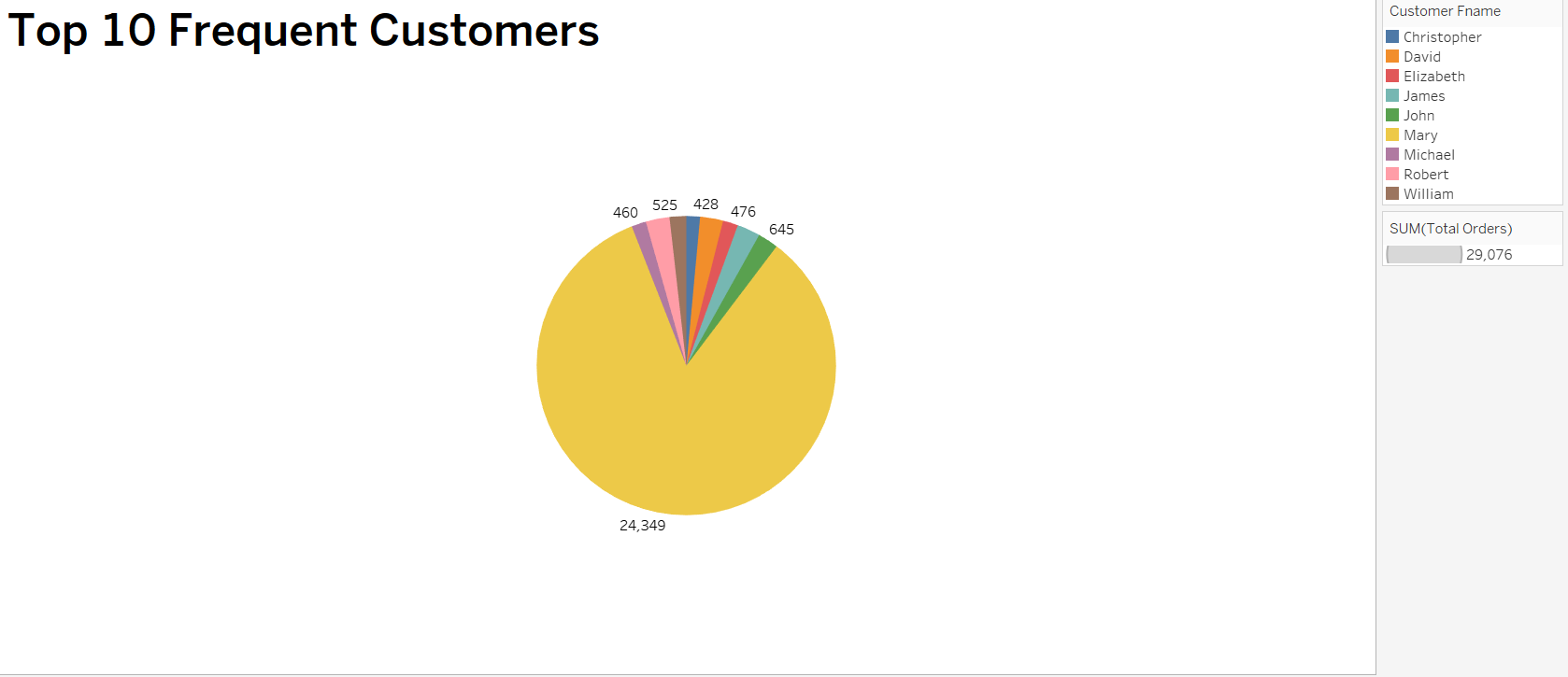


Figure : Pie Chart of Top 10 Most Frequent Customers

**4.3 Delivery Performance Analysis**

The delivery performance analysis was to evaluate the effect of delayed shipment on profitability. It was found that delayed deliveries led to clear correlation with reduced profit margins. Low profitability for longer delayed orders indicates that delivery efficiency is important to maintaining financial performance. It is an insight showing that the supply chain needs to be optimized to reduce delays and enhance customer satisfaction. From this pattern, businesses are able to identify bottlenecks and develop a streamlining of logistics processes.

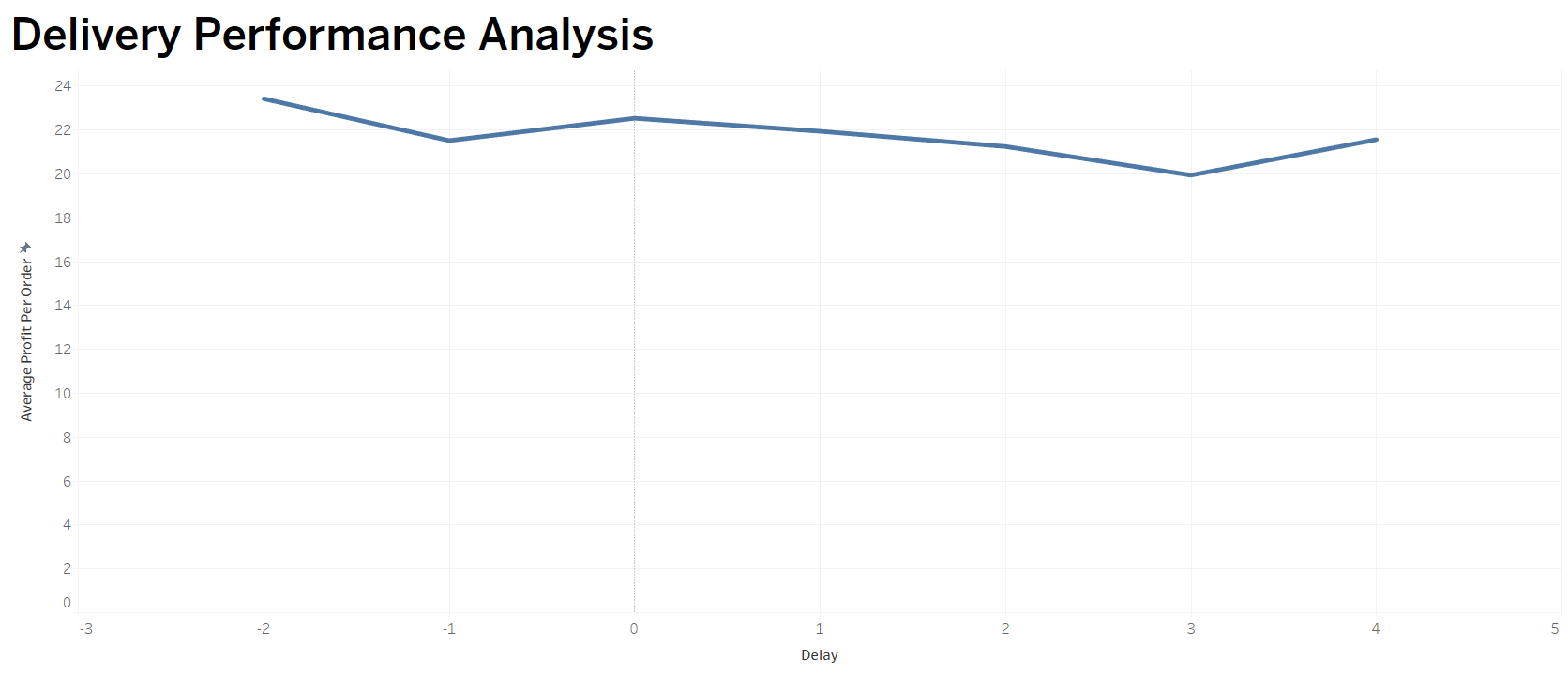


Figure : Line Graph of Delivery Performance Analysis

In combination, these three insights give a clear picture of how the supply chain’s operational efficiency operates. The validity of big data analytics in determining the key performance indicators and improving the business processes and decision making is demonstrated by the findings. This makes businesses poised to develop and adopt data driven growth and competitive strategies for long term.

**Conclusion and Future Work**

**Conclusion**

This project successfully shows the importance of big data analytics in optimization of supply chain management. We extracted, and analyzed with critical insights from the DataCo SMART SUPPLY CHAIN dataset by using Apache Spark (Databricks), AWS S3, MySQL, and Tableau. The three primary insights were Top 10 Product Categories, Top 10 Frequent Customers, and Delivery Performance Analysis, and the actionable findings from them can help improve operational efficiency and profitability.

The Top 10 Product Categories insight informed the businesses which categories brought the highest revenue, which allowed prioritizing these products and best allocating the resources. Like the Top 10 Frequent Customers analysis, the most active customers were also identified in the Top 10 Frequent Customers analysis for businesses to enhance customer relationship, deliver personalized marketing and improving retention strategies. The Delivery Performance Analysis showed us that shipping delays decrease profit per order, and with that demonstrated a clear picture that operational failures could translate to financial regret.

Furthermore, the technical implementation of the project focused on the scalable and efficient data architecture. The data storage was done on AWS S3 to make sure that it is reliable and accessible to large datasets. High speed processing and transformation of the data was made possible with Databricks and Apache Spark, and the data was stored in structured way in MySQL for further analysis. Finally, given the complexity of the data we had to present, Tableau based visualizations proved to be very useful in showing the data in an intuitive and actionable way. Combined, these technologies prove to businesses how they can leverage these technologies to both manage and analyze large datasets in order to accomplish better decision making.

In general, this project showcases the superiorities of data driven decisions making in supply chain management. With historical data as a point of reference, businesses can identify redundancy, enhance customers’ satisfaction and increase their profits. The insights are leveraged to create data informed strategies to better focus product, customer satisfaction and operational efficiencies.

**Future Work**

Although this project captured a great deal of value in its insights, there are a few avenues for future enhancement and exploration. The experience that we are providing is particularly in the area of real time data processing. Businesses may be able to stop and start the supply chain more quickly and take advantage of real time supply chain data through the use of tools such as Apache Kafka. It is this advancement that would increase agility and consequently allow for proactive decision making.

The application of predictive analytics is another promising extension. It became possible to use advanced machine learning models such as regression, time series forecasting, etc., to forecast future sales trends, predict time of delivery, and optimize inventory management. This would provide an edge to these organizations to be ready for market demands and eliminate future uncertainties. Furthermore, advanced customer segmentation is a critical path for the future. By Cluster techniques like K Mean or DBSCAN, deeper insight in the customer purchasing behavior can be achieved. By enabling personalized marketing, they’d be able to drive revenue growth and better customer retention programs since customers are satisfied.

Another area of further analysis that would have significant benefits is supply chain optimization. The overall picture of the supply chain could be provided by simply incorporating such data as inventory levels, supplier performance, logistics tracking. This would allow businesses to have a better idea about bottlenecks, reduce delivery time, and improve overall operations for better efficiency.

Real-time delivery to Business Intelligence (BI) tools like Power BI or Google Data Studio may be automated via integration with these tools to make the delivery of the real time reports and dashboards. This would enable different stakeholders to have quicker, better informed decisions made by them by providing up to date interactive visualizations across different departments.

If these future enhancements are adopted by businesses, the process of supply chain enhancement will be further optimized, and the decisions will also become more effective, and resources will be more effectively managed. With further advancement in technology and data analytics, the project should expand to incorporate these state of the art methodologies to achieve high level of operational efficiency and lead to further increase in profitability of the project over the long period.

**References**

*Gunasekaran, A., Patel, C., & McGaughey, R. E. (2004) 'A framework for supply chain performance measurement,' International Journal of Production Economics, 87(3), pp. 333-347.*

*Forrester, J. W. (1961) Industrial Dynamics, Cambridge, MA: MIT Press.*

*Zaharia, M. et al. (2016) 'Apache Spark: A Unified Engine for Big Data Processing,' Communications of the ACM, 59(11), pp. 56-65.*

*Fisher, M. L. (1997) 'What is the right supply chain for your product?', Harvard Business Review, 75(2), pp. 105-116.*

*Kaggle (2023) DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS Dataset. Available at:* [*https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis*](https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis) *(Accessed: March 2025).*

*Gandomi, A. and Haider, M. (2015) 'Beyond the hype: Big data concepts, methods, and analytics,' International Journal of Information Management, 35(2), pp. 137-144.*