Big Data Final Project By Matthew Latondresse

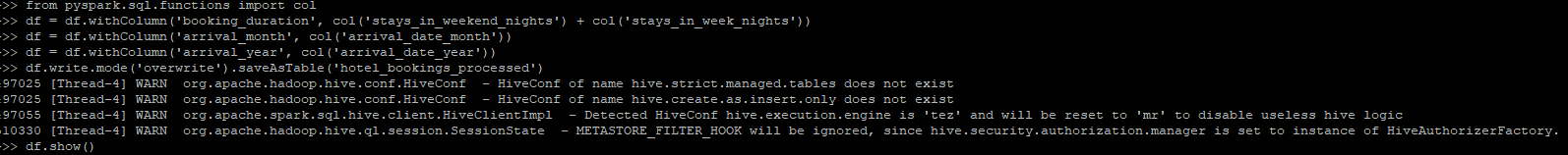
The hospitality industry is heavily influenced by customer behavior, booking trends, and external factors such as seasonality and economic conditions. Understanding these trends can help hotels optimize their operations, improve customer satisfaction, and increase profitability. This project aims to analyze hotel booking data to identify key trends, booking preferences, and cancellation patterns. Specifically, it seeks to address the following questions:

1. How do booking trends vary over the years?
2. What are the patterns in booking cancellations?
3. How does the average booking duration vary by hotel type?

The insights from this analysis are significant in the contemporary context as they can help hotel managers make data-driven decisions to enhance their services and marketing strategies. Understanding booking and cancellation patterns can lead to better inventory management, targeted promotions, and improved customer satisfaction.

I started by uploading the dataset to HDFS to make it accessible to my Spark cluster. HDFS is a distributed file system that provides high-throughput access to application data and is well-suited for large datasets.

Using PySpark, I loaded the dataset from HDFS into a Spark data frame. Spark provides powerful in-memory processing capabilities and can handle large datasets efficiently, making it an ideal choice for data processing and transformation.

I filled missing values to ensure the dataset was complete and ready for analysis. Handling missing values is crucial for maintaining data integrity and ensuring accurate analysis results.

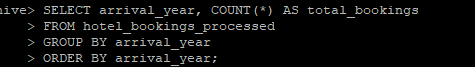
I added new columns to the data frame to facilitate my analysis. These columns included booking\_duration, which combined stays\_in\_weekend\_nights and stays\_in\_week\_nights, and extracted arrival\_month and arrival\_year from the arrival\_date.

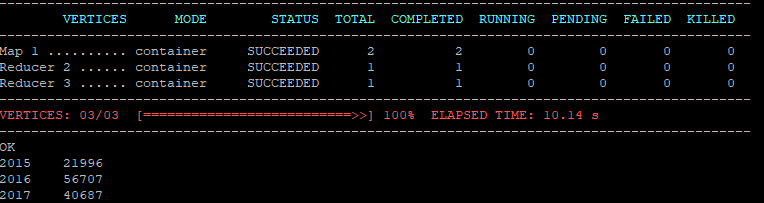
I saved the cleaned and transformed data frame to a Hive table for easy querying. Hive provides a SQL-like interface to query and analyze large datasets stored in Hadoop.

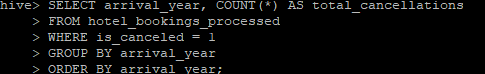


I used Hive to run SQL queries against the processed data. This step allowed me to derive insights and perform aggregations, such as counting total bookings and cancellations by year and calculating average booking durations by hotel type.

This query helped us understand how the number of bookings changed over the years. It could reveal trends such as increasing or decreasing popularity of the hotels.



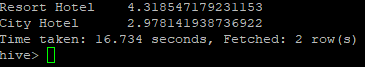
This query showed the cancellation trends over the years, helping identify any patterns or external factors (e.g., economic downturns, pandemics) that might influence cancellations.





This query provided insights into how long guests typically stay at different types of hotels. This information is valuable for marketing strategies, resource allocation, and improving customer experience.





By following these steps, we effectively processed and analyzed the hotel booking data, allowing me to uncover valuable insights about booking trends, customer behavior, and hotel performance. This process is crucial for making data-driven decisions, optimizing operations, and enhancing the overall guest experience in the hospitality industry.

This project demonstrated the practical application of big data technologies, including Spark and Hive, to analyze a large dataset. It showcased the integration of various tools to achieve comprehensive data processing and analysis. The visualizations and findings provided valuable insights into hotel booking trends, cancellations, and customer preferences.