Airline Dataset Analysis using Hadoop, Hive, Pig and Athena

Business Overview:

Airline data analysis involves extracting meaningful insights and patterns from the vast amount of data collected by airlines during their operations. This analysis helps airlines make informed decisions, optimize processes, enhance efficiency, improve customer experiences, and ultimately increase profitability. Here are some key areas where data analysis plays a crucial role in the airline industry:

- **Flight Performance Analysis:** Airlines analyze data related to flight schedules, on-time performance, delays, and cancellations. This analysis helps identify the root causes of delays, assess the efficiency of flight routes, and make adjustments to improve overall performance.
- **Demand Forecasting:** By analyzing historical booking data, airlines can predict passenger demand for different routes and time periods. Accurate demand forecasting helps optimize flight schedules, set ticket prices, and allocate resources efficiently.
- **Route Profitability:** Data analysis allows airlines to assess the profitability of various routes by considering factors such as passenger load, operating costs, and ticket prices. This insight helps airlines make informed decisions about route expansion or reduction.
- **Customer Segmentation:** Airlines analyze passenger data to segment customers based on their travel behavior, preferences, and demographics. This segmentation enables personalized marketing strategies and tailor-made services for different customer groups.
- **Revenue Management:** Airlines use data analysis to implement effective revenue management strategies. By adjusting ticket prices dynamically based on demand and booking patterns, airlines can maximize their revenue and fill available seats more efficiently.
- **Crew Performance and Scheduling:** Data analysis helps evaluate crew performance, track their working hours, and optimize crew scheduling. This ensures compliance with regulations, improves crew satisfaction, and reduces operational inefficiencies.
- **Maintenance Predictions:** Airlines analyze maintenance data to predict potential issues with aircraft components. This proactive approach helps prevent unplanned maintenance and reduces aircraft downtime.
- **Operational Efficiency:** Data analysis is used to assess the efficiency of various operational processes, such as baggage handling, boarding procedures, and turnaround times. Identifying bottlenecks and areas for improvement can lead to smoother operations.
- **Customer Experience Enhancement:** By analyzing customer feedback and survey data, airlines can identify areas where they can enhance the passenger experience. This may include improving in-flight services, airport facilities, and customer service interactions.

- Safety and Security Analysis: Airlines analyze safety and security-related data to identify patterns and trends that may impact operations. This allows airlines to implement measures to enhance safety and mitigate risks effectively.

To perform these analyses, airlines utilize data analytics tools, machine learning algorithms, and other advanced technologies. Data visualization techniques are also commonly used to present the findings in a clear and actionable format for decision-makers. It is essential for airlines to have robust data governance practices in place to ensure data quality, security, and compliance with relevant regulations.

Aim:

The main objective of this project is to showcase the implementation of Airline Data Processing using open source technologies like Hadoop, Hive, Pig, HDFS, and Athena. The project aims to process and analyze large volumes of airline-related data efficiently and effectively. Hadoop and HDFS provide a distributed storage and processing framework, enabling the handling of massive datasets. Hive and Pig serve as query languages and data processing tools that simplify data manipulation tasks. Additionally, Athena offers a serverless query service, facilitating interactive analysis of data stored in Amazon S3. By utilizing these technologies, the project aims to demonstrate how airlines can leverage big data solutions to extract valuable insights, optimize operations, and improve overall performance within the industry.

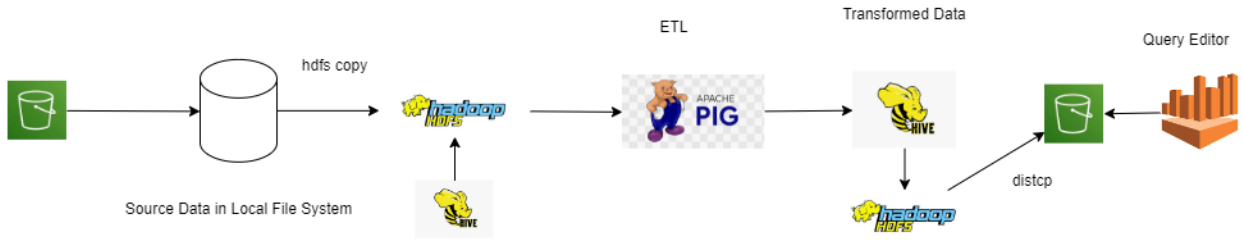
Tech Stack:

Language: Python, SQL
Services: AWS EMR, Apache Hadoop, Apache Hive, Pig, AWS Athena

Key Takeaways:

- Understanding the source data
- Understanding the data models
- Understanding the importance of each service in detail
- Creating AWS EMR cluster
- Connecting to EMR cluster using Putty
- Understanding the Hive Architecture
- Understanding the Hive Data types
- Understanding the Pig Architecture
- Understanding the Pig Data types
- Understanding the basic Unix commands
- Difference between Hive-managed tables and Hive-external tables
- Data preprocessing with Pig
- Basic EDA using Hive
- Understanding Hive partitioning and clustering
- Understanding the Pig ETL job
- Perform analysis using Amazon Athena
- Understanding the Job Orchestration of pipeline

Architecture:



hdfs

hive

pig

s3

Athena