Project : Airline Data Analysis

Overview - The Quick Pitch

The objective is to analyze the airlines data to provide the airline on time performance analysis to the end user using PySpark programming.
This project has been done on a legit Hadoop cluster of 13 Nodes availed on subscription from a third party called Itversity. The data that's used is huge volume data, with the size being 30 GB.

The Problem

Airline on time performance refers the service success rate by the airlines based on the schedule. Airline delay is the most important issue in the airline industry, because it will lead to economic crisis in the airline business for the owners. This project analyses the airline data to provide the necessary cleansed data to derive meaningful insights from it.

The Solution

Analytics have been performed on US Flight data to get meaningful insights from it like delay in departure and arrival, ranking airlines on the basis of on-time arrival, ranking on the basis of cancellations etc.
In order to perform analytics on this data, the data should be quality and required format data.
The collected raw data is in form of a CSV file is taken from a website called Kaggle, which provides these datasets for analysis purposes. This data is then uploaded on the Hadoop cluster - HDFS. After this, the data is read from HDFS programmatically through PySpark code and various operations are applied, to get the delay details, rankings etc.
The filtered quality compressed dataset in Parquet format is written into Hive Tables for further analysis.

Software Requirements

- Mac OS
- Jupyter Notebooks
- PyCharm CDE
- PySpark 2.4
- Hive
- HDFS
- Yarn