US Airline On-Time Analytics system

Overview

US Airline On-Time dataet, which consists of flight arrival and departure from 1987 to 2008 details are available at Harvard Dataverse Repository.

With this project, we want to analyze this dataset and answer questions, such as when is the best time of day/day of week/time of year to fly to minimum delays, or do older planes suffer more delays.

Scope

More specifically, we will build a system which will:

1. Aquire data from source and save to a data lake
2. Build an ETL pipeline to extract data from data lake, transformate it and load it to the target system, a multi-dimensional data warehouse

Then, we will use tools to analyze the data in data warehouse and answers the questions.

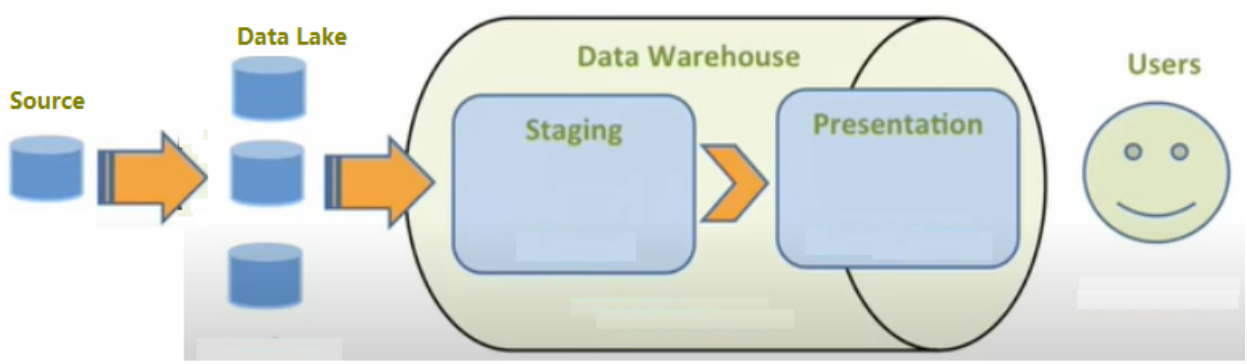
Data sources

Source data can acquired here
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

This dataset is static. It includes 22 compress csv files, one for each year. It has nearly 120 million records, takes up 1.6 gb space compressed or 12 gb when uncompressed.

In addition to the on-time dataset, airport, plane and carrier dataset are also provided as csv files.

Architecture/Data flow



1. Source data will be acquired and saved to data lake at local or cloud, in form of data files
2. Data files selected by certain criterias will be loaded from the data lake to the staging area for cleansing and transformation.
3. Cleansed and transformed data will be loaded to data warehouse for the end users's use.

Major steps:

We split the project into different tasks and stages:

1. Acquire the data from source and store it to data lake, on local or cloud
2. Load data from data lake to staging area for exploration
    a. Determine the data cleansing and transformation strategy. Decide which rows , as well as which fields will be loaded to target system
    b. Perform data warehousel data modeling and system design. Decide on dimension tables, fact tables, their attributes, data mapping, data loading sequence,
3. Development/testing
    - Writing scripts that will perform the data extract, data cleansing/transformation, data loading.
    - Running end-to-end integration test for small set of data
4. Scale up/performance

    Load testing. Optimize the solution for bid data volume

5. Deployment
    1) Run one time load to acquire data from source to data lake
    2) Based on the year range selected, execute the ETL pipeline end-to-end to move the data from data lake to data warehouse
6. Data analysis

Using SQL and other tools to analyze the data that has been loaded to data warehouse and answer the questions

Technology

Data warehouse strategy

We will use Kimballs' approach ： start from the questions and use star-schema for data modeling.

Storage:

Location	Solution
Data lake	Local or cloud (AWS S3 or Azure Blob or HDFS) Data file format: parque (prefered), csv
Staging area	RDBMs tables/parque files/
Data warehouse	RDBMS

Tools:

Process	**Tools**
Acquiring data	Python, API modules (such as requests)
Data exploring	Jupyter notebook, Python, Spark, Pandas, SQL
Data Transformation/ cleansing	Python, Spark, Pandas, SQL
Data loading	Python, SQL, bulk load tools (to be decided) Spark
ETL orchestration	To be decided
Reporting/Query/Analysis	Jupyter Notebook Python (pandas, Matplotlib module,) SQL