Right on Time

Smart Air Travels

By: Le Gu, Jacky Zhang

Initiatives

According to a study conducted by UC Berkeley researchers, flight delays put a $32.9 billion hole in the U.S. economy in 2007, and more than half of that cost is borne by us, the passengers.

This project aims to better understand air travel experiences by analyzing flight on-time performance statistics. By exploring flight delays, the project hope to reveal delay trends by airport, airline, aircraft models and reasons for delay etc.

Analysis

1. Which airlines have the most delays?
2. When is the worst time to travel, in terms of delays expected?
3. Which weather conditions cause the most delays?
4. Which models of plane have the most delays?
5. Which are the busiest airports?

Predictive model:

We seek to build a regression model API that gives the users an estimate of expected delay given information about their flight.

Data Sources

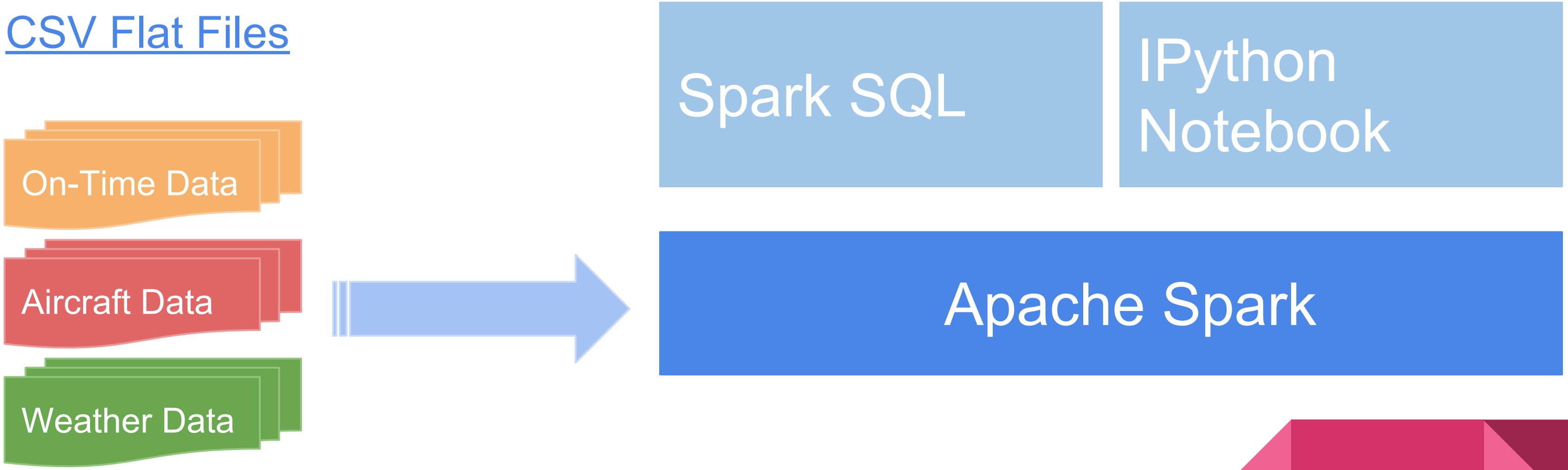
- The United States Department of Transportation tracks on-Time performances of major US carriers. The data set contains flights departure, arrival details for US domestic flights since 1987. This project would take the most recent 10 years' data (20 GB) for analysis.

  On-Time performance

- Statistical Computing Organization provides aircraft-related information such as manufacturer, issue year, model type etc.

  Plane Data

Data Architecture



CSV Flat Files

On-Time Data

Aircraft Data

Weather Data

Spark SQL

IPython Notebook

Apache Spark

Milestones

- Week 1: Data Collection
- Week 2: Data pipeline setup
- Week 3: Spark SQL Analysis
- Week 4: PySpark Analysis, model building
- Week 5: Project wrap-up

Potential Challenges

1. Automating data collection process
2. Dynamically updating prediction model
3. Handling queries on large data sets