Team D-Hawks

John Bever, Karunakar Kotha, Leo Salemann, Shiva Vuppala, Wenfan Xu

Weather & Transportation

Streaming the Data, Finding Correlations

July 19, 2017

Overview

Our Summer 2017 BIGDATA 230B Project will focus on looking for correlations between weather and transportation data. We have identified Data for Democracy [http://datafordemocracy.org], [https://github.com/Data4Democracy] as a "stakeholder" for our work, meaning that we strive to create work products that adhere to their mission, technology conventions, and coding style; and can be integrated into their GitHub repository. Specifically, we hope to provide capability to their democratizing_weather_data project, while still meeting the requirements for the BIGDATA 230B Summer Project. Using kafka, we will stream weather data from multiple sources (multiple web api's) and provide a uniform data frame or csv output, amenable to multiple data science or machine learning workflows.

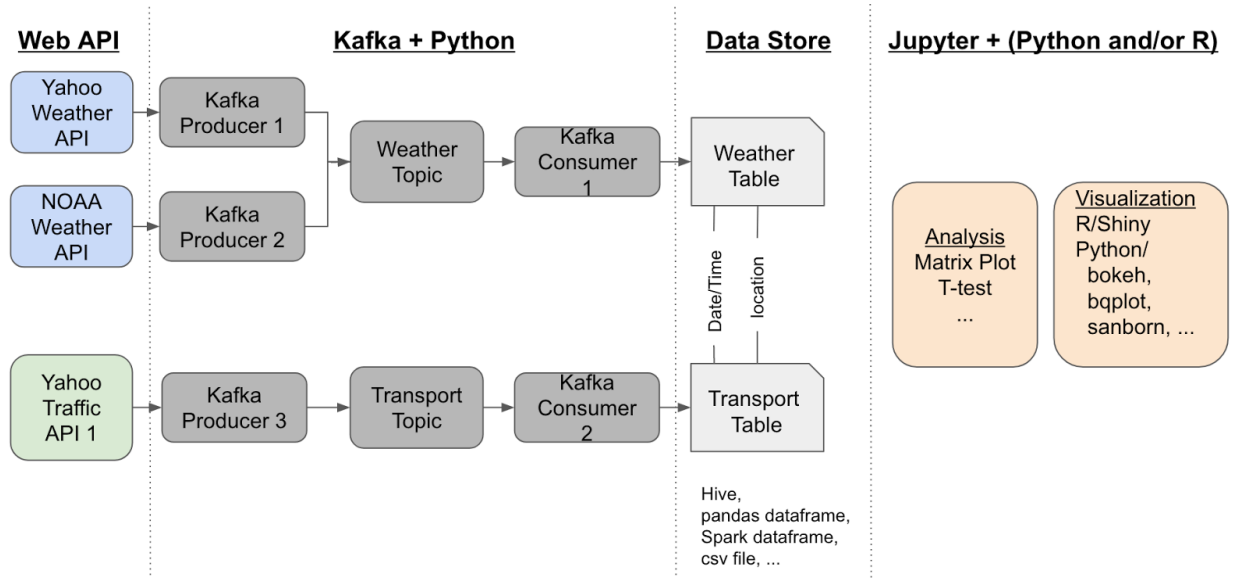
Data Sources

We will use weather and transportation API's such that both have a date/time and a location. We'll be joining on these keys later. Examples include:

1. **NOAA QCLCD Weather API** already in use by Data for Democracy
2. **Yahoo Weather API** simple, free/low-cost
3. **Yahoo Traffic API** simple, free/low-cost

Architecture & Tools

We'll be combining RESTful Web API's, Kafka, elements of Hadoop (Hive, HDFS, Spark) and Jupyter Notebooks. Our main implementation language will be python, with the potential for some of the anlysis in R. Visualization will be done in Jupyter using python and possibly R packages (shiny, bokeh, sanborn, ..) in keeping with Data for Democracy's preference for open source solutions and Jupyter as the main analytic & visualization enviornment.



Questions for Analysis

Which weather characteristics can predict traffic patterns? Examples of "traffic patterns" include:

- Average speed/commute time
- Busses arriving at their stops on time
- RideShare/Taxi volume