Eric Tsai, Andrew Lam, Ashton Chevallier
W205 Course Project
Progress Report
November 17, 2016

Project Idea Summary

Problem Statement

Health care data is hard to come by, understand and aggregate, which makes decisions for consumers even harder. Our application will conveniently store and aggregate relevant health data sources to make these important choices easier.

Project Goals

- Successfully import data sources into HDFS
  - Import Data.gov
  - Scrape Health Safety Scores
  - Pipe data from Yelp via API
- Correctly aggregate data together
  - Use ‘closeness’ models to match data between sets
  - Clean data into tidy format
- Display aggregate data in useful form

Data Acquisition and Organization Strategy

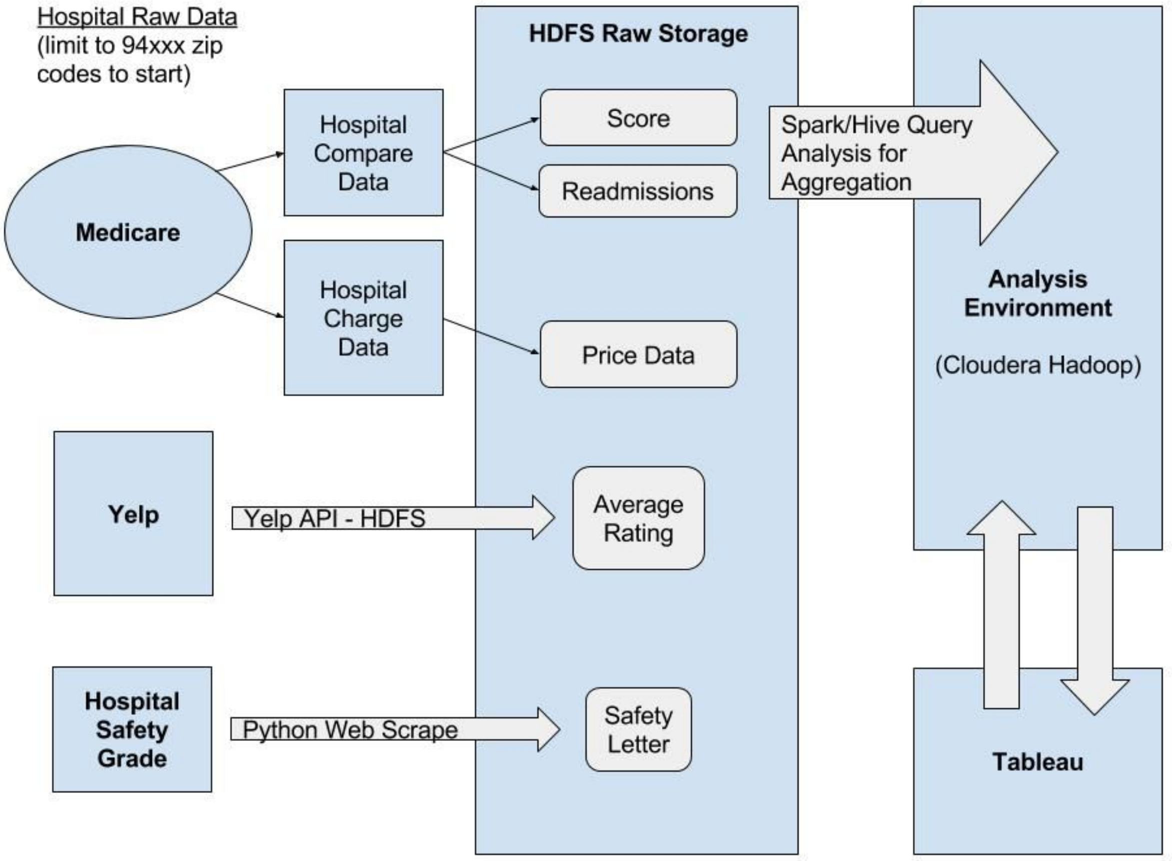
We will be sourcing and integrating data from multiple sources to bring in a diverse data set to work upon. We will be using Medicare’s vast data set as the base and enhance it with data from Yelp and Hospital Safety Grade.

	Data Set
General	Medicare Hospital Compare Data
Price	Medicare Hospital Charge Data
Quality	Medicare Physician Compare Data (Optional)
	Yelp
	Hospital Safety Grade

The Medicare data will be acquired by using wget (hospital compare data, charge data, and physician compare data). We will use Yelp’s public API to access Yelp’s user reviews and we will use web scraping to acquire the Hospital Safety Score data.

Once the data has been acquired, we will use fuzzy matching to align the data so that we can unify the data sets and score each provider and physician for quality of care.

Architecture Diagram



Initial Performance Evaluation

- Developed initial data acquisition and organization strategy
  - Downloaded data sources into HDFS, including Hospital Compare Data and Hospital Charge Data
  - Registered Yelp application for API access
- Created Github repository
- Created initial architecture diagram
- Moving forward, our plan is to begin our analysis with the 94xxx zip codes and iterate from there