Choosing Health:

Making Big Data Work for Families

Andrew Lam, Ashton Chevallier, Eric Tsai
W205 Section 1
December 15, 2016

Motivation for the Project

In 2015, the United States spent...

3.2 trillion$

...on healthcare, accounting for

17.8%

as a percentage of GDP, and yet Americans will only live...

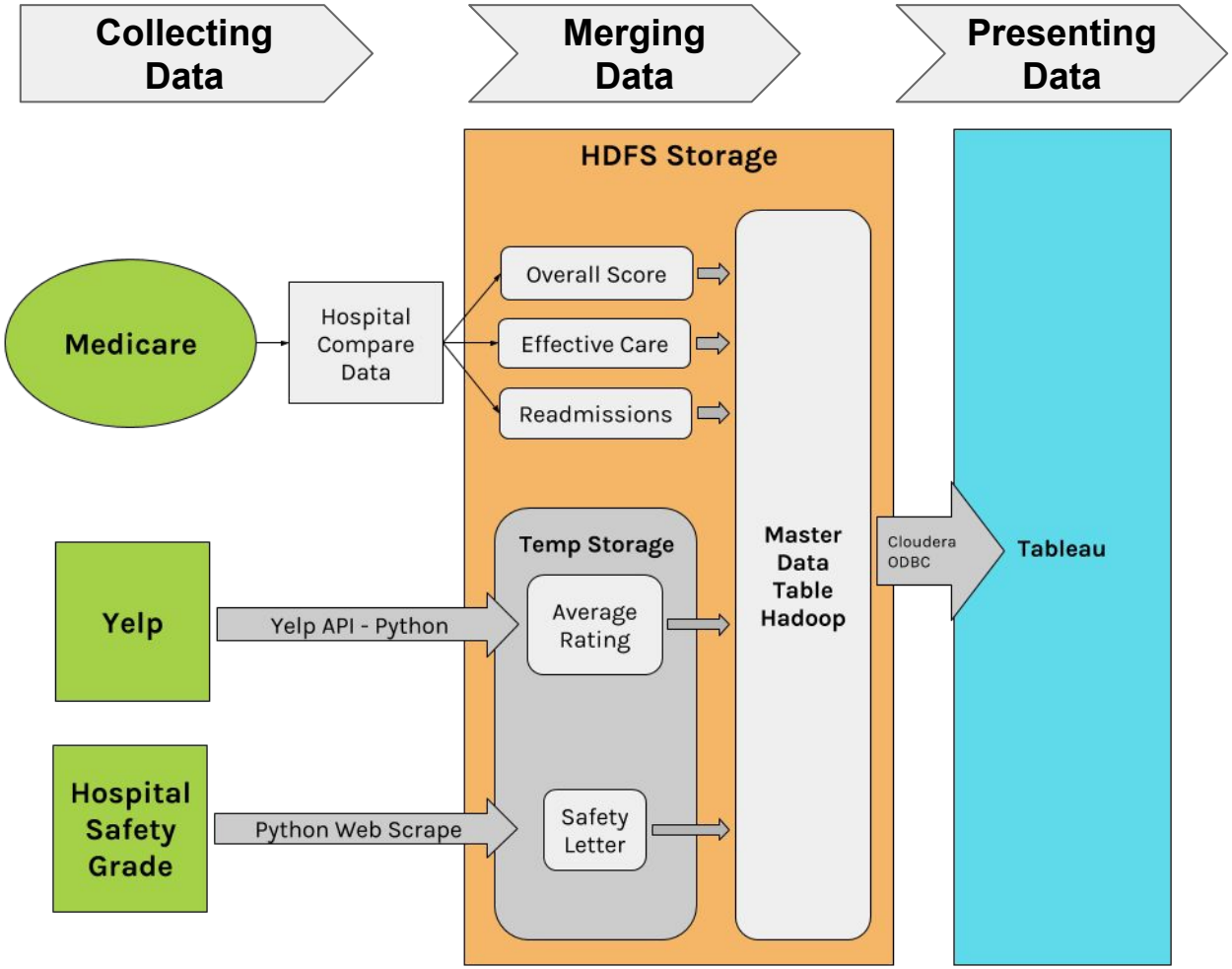
79years

...with the country ranking 31st in life expectancy

Goals & Scope

- **The goal of the project is to empower consumers with information about the quality of hospitals relative to the price paid**
- Although our intention was to include price information, we had difficulty obtaining enough data
- Therefore, this project primarily focuses on healthcare quality with an emphasis on the West Coast

Process Overview & High-Level Architecture



1. Collecting Data

- **Medicare**
  - Centers for Medicare & Medicaid Services (CMS)
- **Yelp**
  - Pulled via Yelp Fusion API
  - Used requests Python Package
- **Hospital Safety Grade**
  - Scraped data using Python

2. Merging Data

Challenge

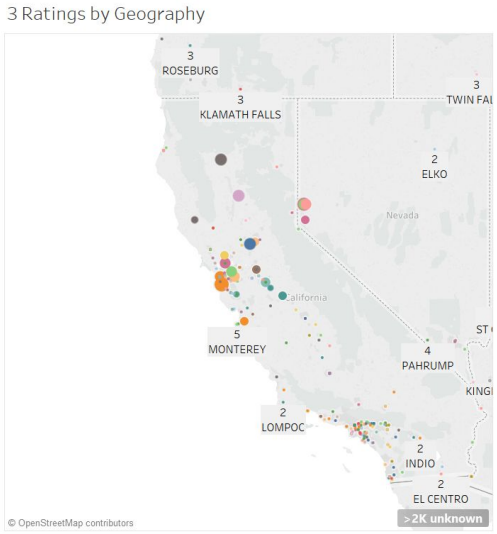
- Each data source had its own variation of the hospital name
- There was no primary key to use to match data between healthcare providers

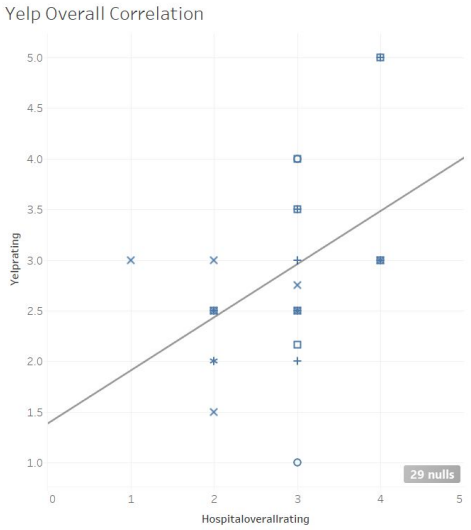
Solution

- Filter by zip code
- Within each zip code, perform Levenshtein matching on the name and address

3. Presenting Data

- EC2 acts as a Hive Server to enable Tableau (Cloudera Hadoop ODBC)
- Recoded ordinal variables to numbers for improved presentation





3 Ratings by Geography | Total Scores Detail | Average Hospital Score Map | Yelp Overall Correlation | MetaScore Detail | MetaScore by State | Trends | Details

Demo

Ideas for the Future

In-Depth Analysis

Acquire more granular metrics on price and performance from additional sources

More exploration of the data for greater insights

Improved Web-Serving

Create website that allows interactive updates to data

Create Tableau Webserver to provide direct web analytics

Faster Performance

Remove bottlenecks in performance in extraction and transformation

Register with Yelp to remove 25k request limit

Thanks!