W205 Final Project: Data Driven

Heather Feinstein
Cynthia Hu
Collin Reinking
Charlotte Swavola

Project Presentation
Project Repository

Goals

Empowering taxi drivers with street smarts from data

- How do I optimize my time driving for downtime, but maximize my fares per ride?
- Should I drive back to the area of high demand or wait for a pickup in my current location?
- Is there a strategy to maximize gratuities (tips)?
- Are fewer long-distance trip or more frequent short trips more profitable?
- How should I plan my week to maximize earnings?

Our Roadmap

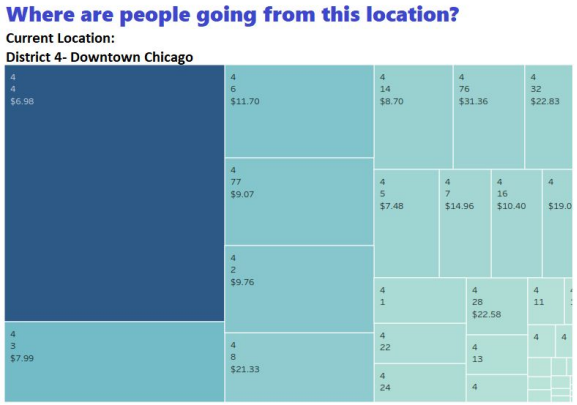
Phase I: Monthly Dashboard

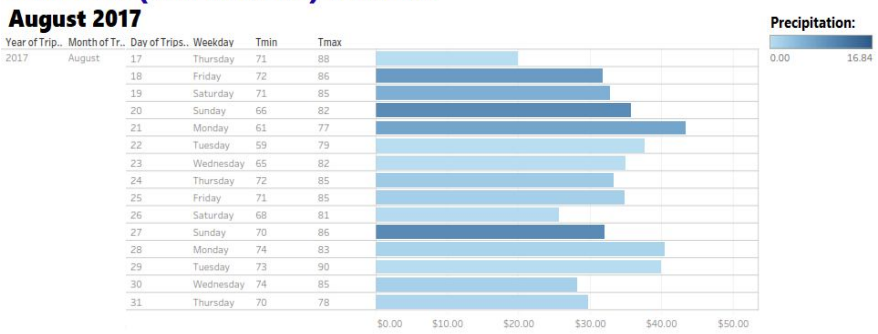
- Web-based dashboard
- Monthly updated analysis and weekly updated forecast
- Mainly descriptive analysis and simple ML forecast

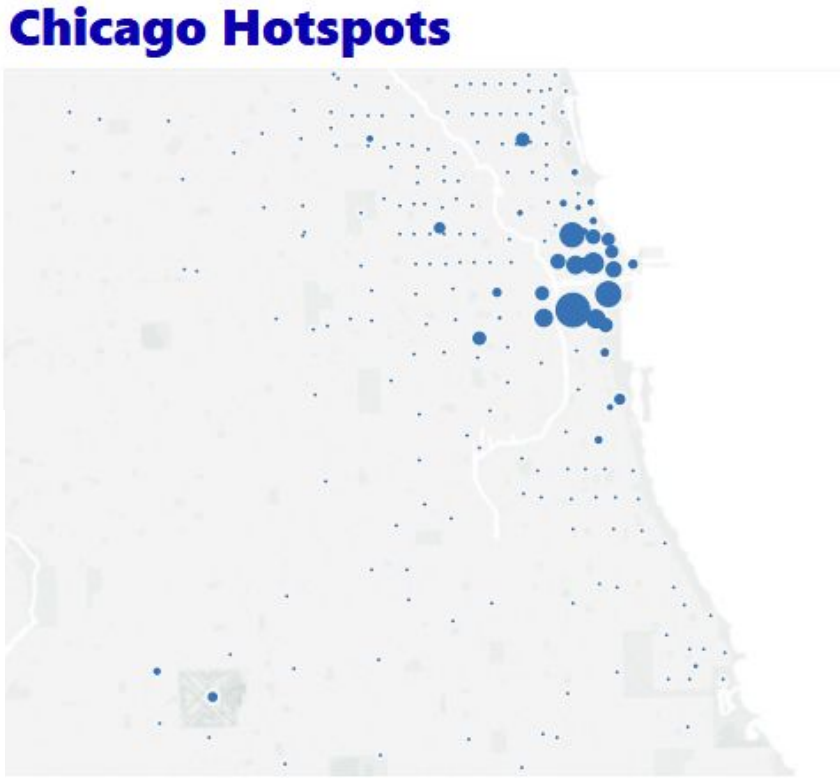
Phase II: Daily or Near-Real-time Application

- Daily or Near-Real-time update
- Advanced ML Fare forecast with improved accuracy
- Integrate both traffic and weather data

Product - Dashboard I



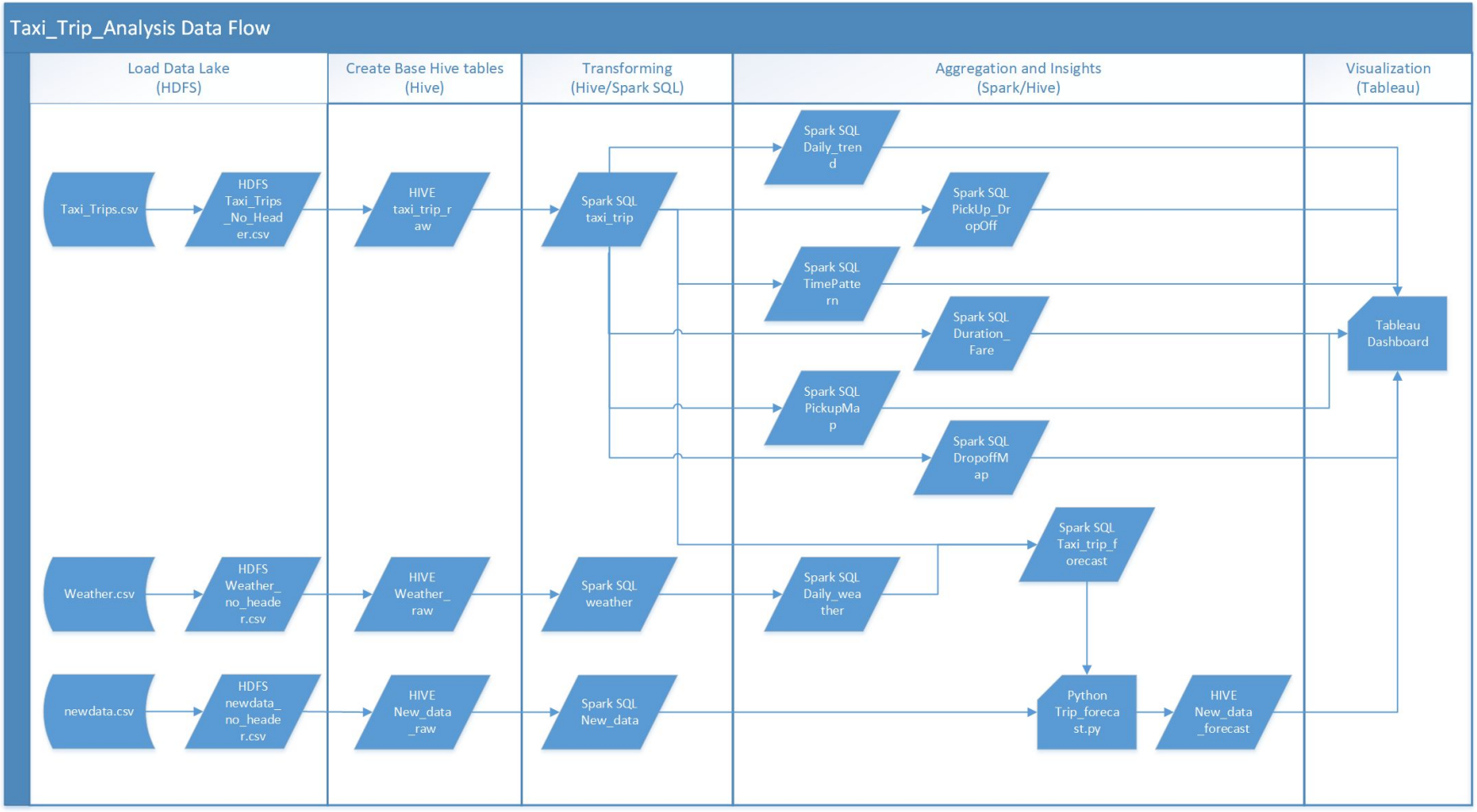




Limitations and Next Solutions

Limitation	Solution
Large data transfers take time	Use SODA API to filter initial download
SQL transformations take time	SODA API result will reduce the amount of processing at transformation step
SQL aggregations take time	Reducing the historical data and updating aggregation with data updates
Current implementation does not update	SODA API to download update(all data since last download), develop methods to update aggregations
Forecasting on mocked up data	Invest in service such as wunderground.com to programmatically access weather forecast data.

Architecture



Implementation

1. Set up EC2 Instance with 200GB EBS volume attached
2. Install Python3
3. Clone the project repository
4. Run setup scripts to Load data and save files into HDFS
5. Create base Hive tables
6. Transform datasets
7. Aggregate datasets for visualization and forecast
8. Run Python forecast
9. Connect Tableau

Appendix

Data sets used

1. Chicago Taxi Trip Data https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

   Time Period: 2013 - Current

   Frequency: Data are updated monthly

   Data Retrieve: Download as CSV, JSON, RDF, TSV or XML; Access through SODA API

2. Chicago Daily Weather Data
   https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:US170006/detail

Data Dictionary - Taxi Trip I

Column Name	Description	Type
Trip ID	A unique identifier for the trip.	Plain Text
Taxi ID	A unique identifier for the taxi.	Plain Text
Trip Start Timestamp	When the trip started, rounded to the nearest 15 minutes.	Date & Time
Trip End Timestamp	When the trip ended, rounded to the nearest 15 minutes.	Date & Time
Trip Seconds	Time of the trip in seconds.	Number
Trip Miles	Distance of the trip in miles.	Number
Pickup Census Tract	The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips.	Plain Text
Dropoff Census Tract	The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips.	Plain Text
Pickup Community Area	The Community Area where the trip began.	Number
Dropoff Community Area	The Community Area where the trip ended.	Number
Fare	The fare for the trip.	Money
Tips	The tip for the trip. Cash tips generally will not be recorded.	Money

Data Dictionary - Taxi Trip II

Column Name	Description	Type
Tolls	The tolls for the trip.	Money
Extras	Extra charges for the trip.	Money
Trip Total	Total cost of the trip, the total of the previous columns.	Money
Payment Type	Type of payment for the trip.	Plain Text
Company	The taxi company.	Plain Text
Pickup Centroid Latitude	The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Number
Pickup Centroid Longitude	The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Number
Pickup Centroid Location	The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Point
Dropoff Centroid Latitude	The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Number
Dropoff Centroid Longitude	The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Number
Dropoff Centroid Location	The location of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Point

Data Dictionary - Weather

Data Type	Description	Start	End	Coverage[2]
TAVG	Average Temperature.	10/30/1958	8/10/2017	100%
TMAX	Maximum temperature	1893-01-01	8/10/2017	100%
TMIN	Minimum temperature	1893-01-01	8/10/2017	100%
TOBS	Temperature at the time of observation	1/1/1901	8/10/2017	100%
PRCP	Precipitation	1870-10-15	8/10/2017	100%
SNOW	Snowfall	1893-01-01	8/10/2017	100%
SNWD	Snow depth	1893-01-01	8/10/2017	100%

Q & A