BikeShare Demand Forecasting

Tom Yedwab ⚙ Adam Spitzig ⚙ Chris Murray

Overview of the problem

- Bike sharing is a popular solution for commuters in cities
- Check out a bike from one station and return to another
- However, problems can occur if:
    - Bikes are not available at a station (station is empty)
    - Docks are not available for bike return (station is full)
- Bike demand highly correlated with weather

⇒ *Predict bike sharing demand based on weather forecast*

Acquisition and organization of data

Bike sharing data (CSV format):

http://www.bayareabikeshare.com/open-data

2 years → 2.6 GB (trip data only 77 MB)

Weather data (JSON format):

https://developer.forecast.io/docs/v2

2 years → 30 MB

Example bike sharing data

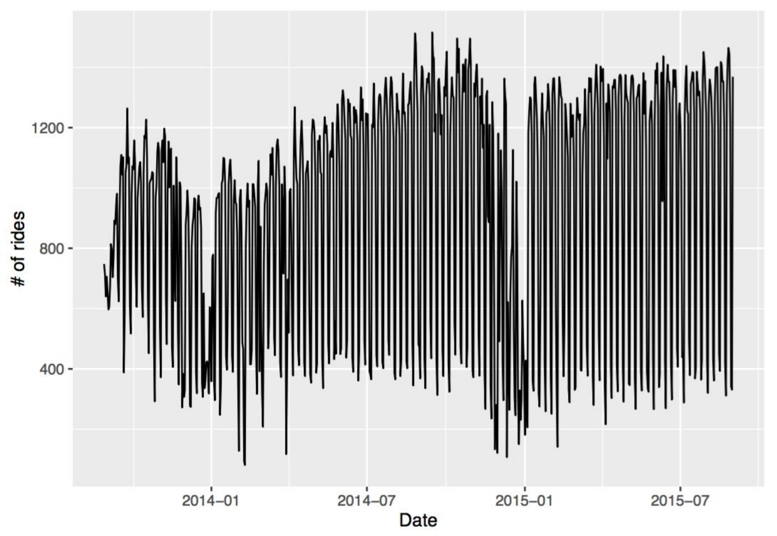
Trip ID	Duration	Start Date	Start Terminal	End Date	End Terminal	Bike #	Subscription Type	Zip Code
4576	63	8/29/2013 14:13:00	66	8/29/2013 14:14:00	66	520	Subscriber	94127
4607	70	8/29/2013 14:42:00	10	8/29/2013 14:43:00	10	661	Subscriber	95138
4130	71	8/29/2013 10:16:00	27	8/29/2013 10:17:00	27	48	Subscriber	97214
4251	77	8/29/2013 11:29:00	10	8/29/2013 11:30:00	10	26	Subscriber	95060
4299	83	8/29/2013 12:02:00	66	8/29/2013 12:04:00	67	319	Subscriber	94103

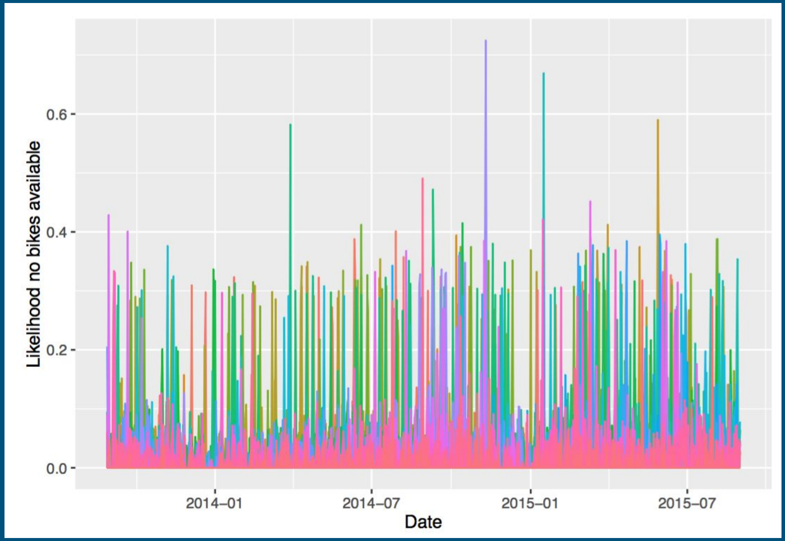
Station ID	Name	Lat	Long	Dock count	Landmark	Installation
66	South Van Ness at Market	37.774814	-122.418954	19	San Francisco	8/23/2013
67	Market at 10th	37.776619	-122.417385	27	San Francisco	8/23/2013
68	Yerba Buena Center of the Arts (3rd @ Howard)	37.784878	-122.401014	19	San Francisco	8/23/2013

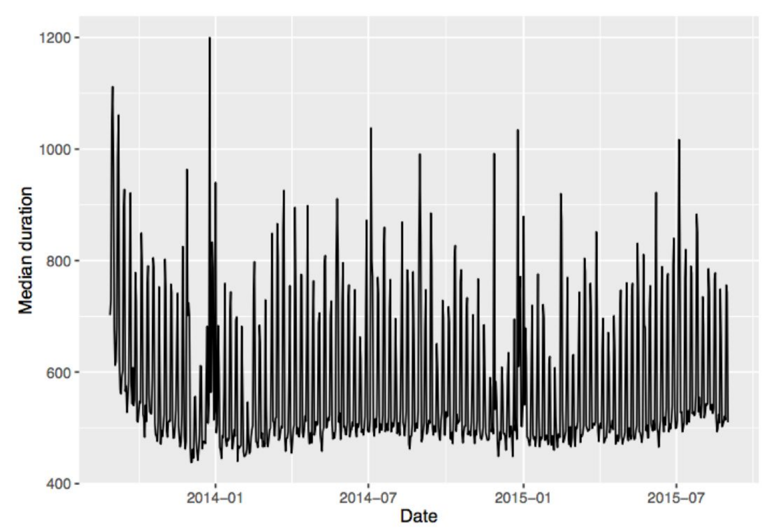
Example weather data

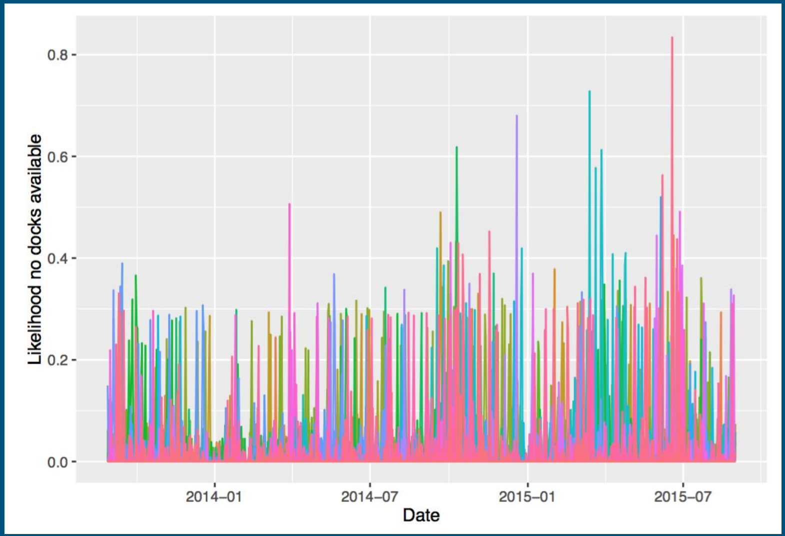
```
{
  "latitude": 37.7877,
  "longitude": -122.4016,
  "timezone": "America/Los_Angeles",
  "offset": -8,
...
 "daily": {
    "data": [
      {
        "time": 1421049600,
        "summary": "Clear throughout the day.",
        "icon": "clear-day",
        "sunriseTime": 1421076352,
        "sunsetTime": 1421111525,
        "moonPhase": 0.73,
        "precipType": "rain",
        "temperatureMin": 49.31,
        "temperatureMinTime": 1421082000,
        "temperatureMax": 58.94,
        "temperatureMaxTime": 1421103600,
        "apparentTemperatureMin": 48.66,
        "apparentTemperatureMinTime": 1421071200,
        "apparentTemperatureMax": 58.94,
        "apparentTemperatureMaxTime": 1421103600,
        "windSpeed": 1.12,
        "windBearing": 126,
        "pressure": 1022.47
      }
```

Initial bike share data analysis

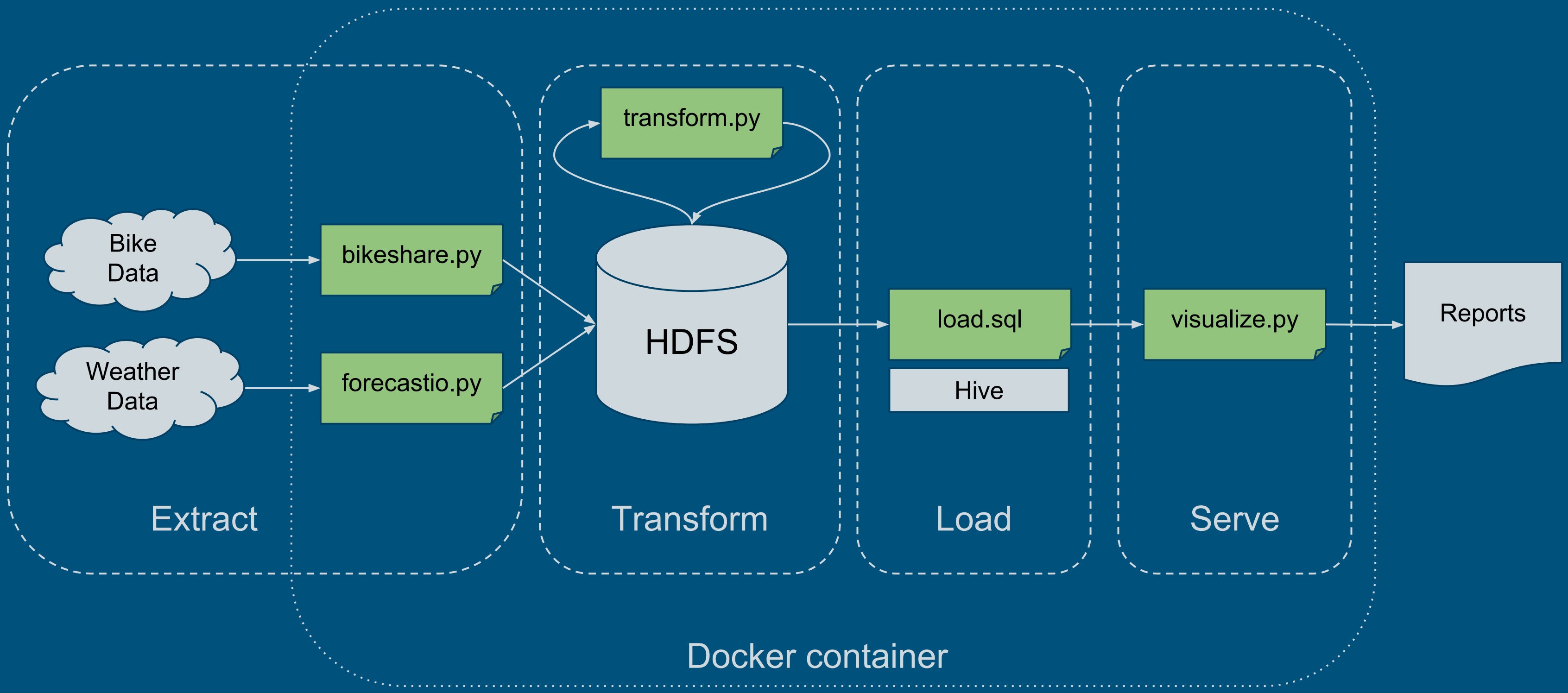








Overall architecture of the solution



Results

- Settled on data sources
    - Pivoted away from problematic public transportation data
- Data extraction code finished
- Initial data analysis complete
- Working on data transformation
    - Having trouble with Hive, switching to Spark

Roadmap for improving the solution

How to scale the solution

- Station status data is huge (2+ GB)
    - Pre-filtering can dramatically reduce size
- Aggregate trip data and station data over time to reduce size

How to evolve the project

- Use streaming weather data to provide real-time demand forecasting
- Include more cities

Additional data sources

- Stadium event data
    - Do people ride more bikes when there is a major public event?
- Road construction / Traffic data