Pipeline for NYC Subway Delays
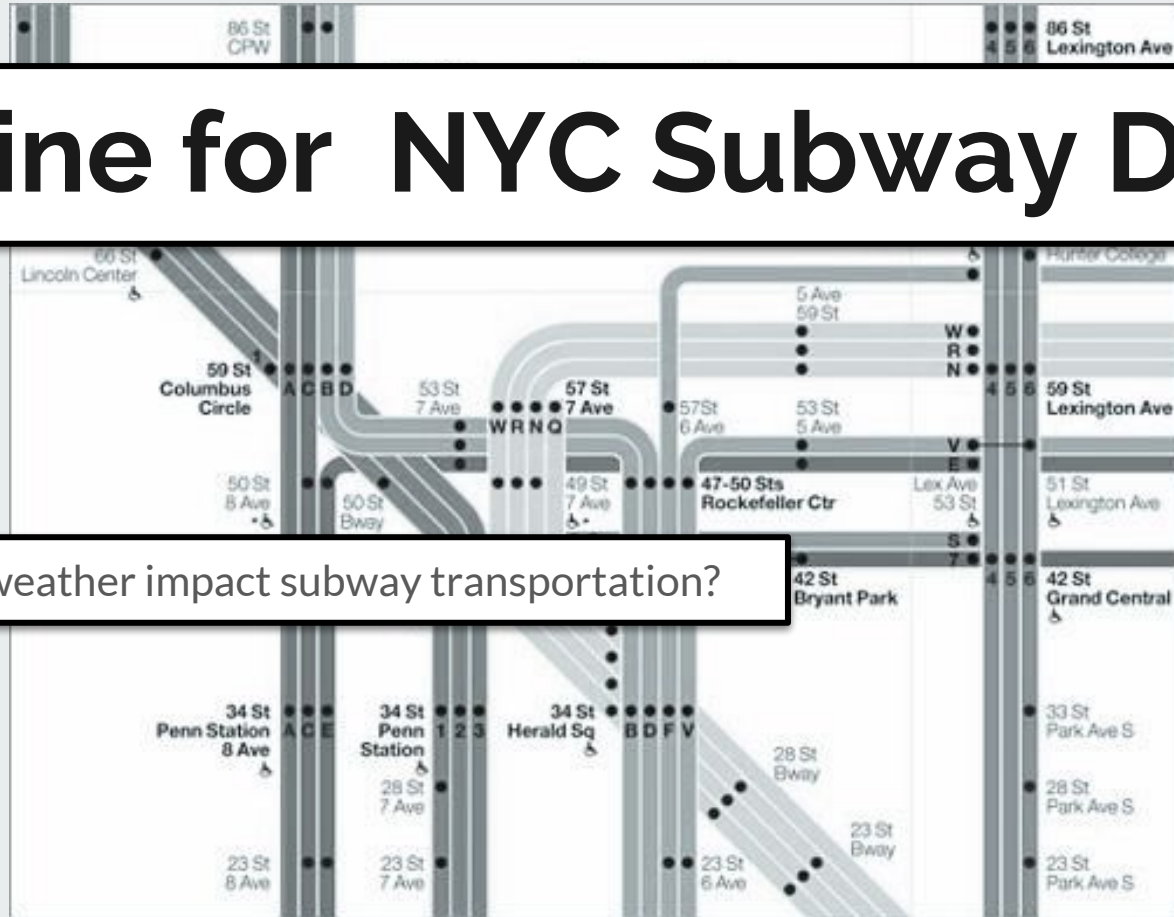
Does the weather impact subway transportation?

Motivation

Motivation

Data Sources

NYC Subway minute by minute status updates

```
<service><responsecode>0</responsecode><timestamp>5/13/2018 1:11:00 AM</timestamp>
<subway>
  <line>
    <name>123</name>
    <status>DELAYS</status>
    <text>
                &lt;span class="TitleDelay"&gt;Delays&lt;/span&gt;
                &lt;span class="DateStyle"&gt;
                &amp;nbsp;Posted:&amp;nbsp;05/13/2018&amp;nbsp;12:55AM
                &lt;/span&gt;&lt;br/&gt;&lt;br/&gt;
                Northbound [2] and [3] trains are running with delays because of
an investigation at &lt;STRONG&gt;116 St.&lt;/STRONG&gt;
                &lt;br/&gt;&lt;br/&gt;

                &lt;span class="TitlePlannedWork" &gt;Planned
Work&lt;/span&gt;
                &lt;br/&gt;
                &lt;a class="plannedWorkDetailLink" onclick=ShowHide(162679);&
gt;&lt;b&gt;TUNNEL RECONSTRUCTION | Until 12:01 AM, Sunday Jun 24, 2018&lt;/b&gt;
&lt;br clear=left&gt;&lt;b&gt;Weekend [2] [3] station closures and route
changes&lt;/a&gt;&lt;br/&gt;&lt;br/&gt;&lt;div id= 162679
class="plannedWorkDetail" &gt;&lt;br&gt;&lt;/b&gt;Normal weekend service is
scheduled to resume 12:01 AM Sunday, Jun 24, 2018.&lt;br&gt;&lt;br&gt;No service
at Park Place, Wall St, Clark St and Hoyt St; use nearby [4] [5]
stations&lt;br&gt;&lt;br&gt;No [2] [3] service between Manhattan and Brooklyn;
take the [4] or [5] instead.&lt;br&gt;&lt;br&gt;Free out-of system transfer is
available between South Ferry [1] [2] and Bowling Green [4]
[5].&lt;br&gt;&lt;br&gt;Weekend service map for &lt;a href=http://web.mta.info
/nyct/service/pdf/FF_Clark_St_web_map.pdf target=_blank&gt;&lt;font color=#0000FF&
gt;Lower Manhattan and Downtown Brooklyn&lt;/font&gt;&lt;
/a&gt;&lt;br&gt;&lt;b&gt;&lt;br&gt;&lt;/div&gt;&lt;/b&gt;&lt;br/&gt;
                &lt;br/&gt;&lt;br/&gt;
```

http://web.mta.info/status/serviceStatus.txt

Data Sources

DarkSky weather API

|	
latitude:	40.73061
longitude:	-73.935242
timezone:	"America/New_York"
currently:	
time:	1526188794
summary:	"Overcast"
icon:	"cloudy"
nearestStormDistance:	20
nearestStormBearing:	271
precipIntensity:	0
precipProbability:	0
temperature:	53.82
apparentTemperature:	53.82
dewPoint:	50.12
humidity:	0.87
pressure:	1020.17
windSpeed:	5.41
windGust:	5.6
windBearing:	15
cloudCover:	0.96
uvIndex:	0
visibility:	9.59
ozone:	327.45

https://api.darksky.net

Question

Given the recent uptick in delays on the NYC Subway, does weather play any significant role?

Overall Approach
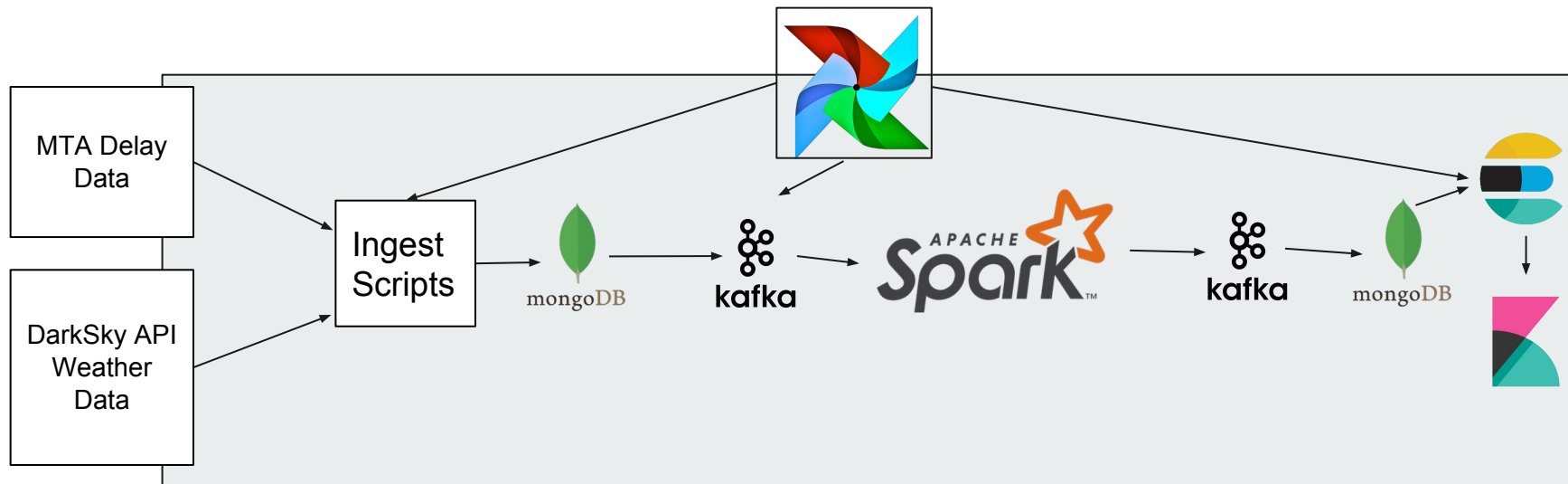
Ingest

1. Extract data once every 90 sec.
2. Parse and store raw in MongoDB
3. Send new records to Kafka topic

Processing

1. Deduplicate subway data and explode records (1 observation per line, per status update)
2. Round down Subway and weather timestamps and join into one dataframe
3. Extract encoded values (1 or 0) for certain features (e.g., signals, nypd)

Visualize

1. Send data to Kafka topic for pickup

2. Script sends from Kafka MongoDB

3. Index with Elasticsearch

4. Visualize with Kibana

DEMO
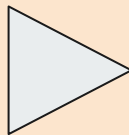
Conclusions

- Overall, it does not appear that weather is a major cause of NYC Subway delays

- Weather might exacerbate other primary sources of subway delays (signal problems, overcrowding)

- More data would be needed to get a better idea of weather's actual impact

Considerations

- More that might be done (e.g., compare with turnstile data)

- Airflow and Kibana are excellent tools

- Largest challenge was just making everything work together given constraints