Analysis of Congress Bills

By George Zhou, Kate Zhou

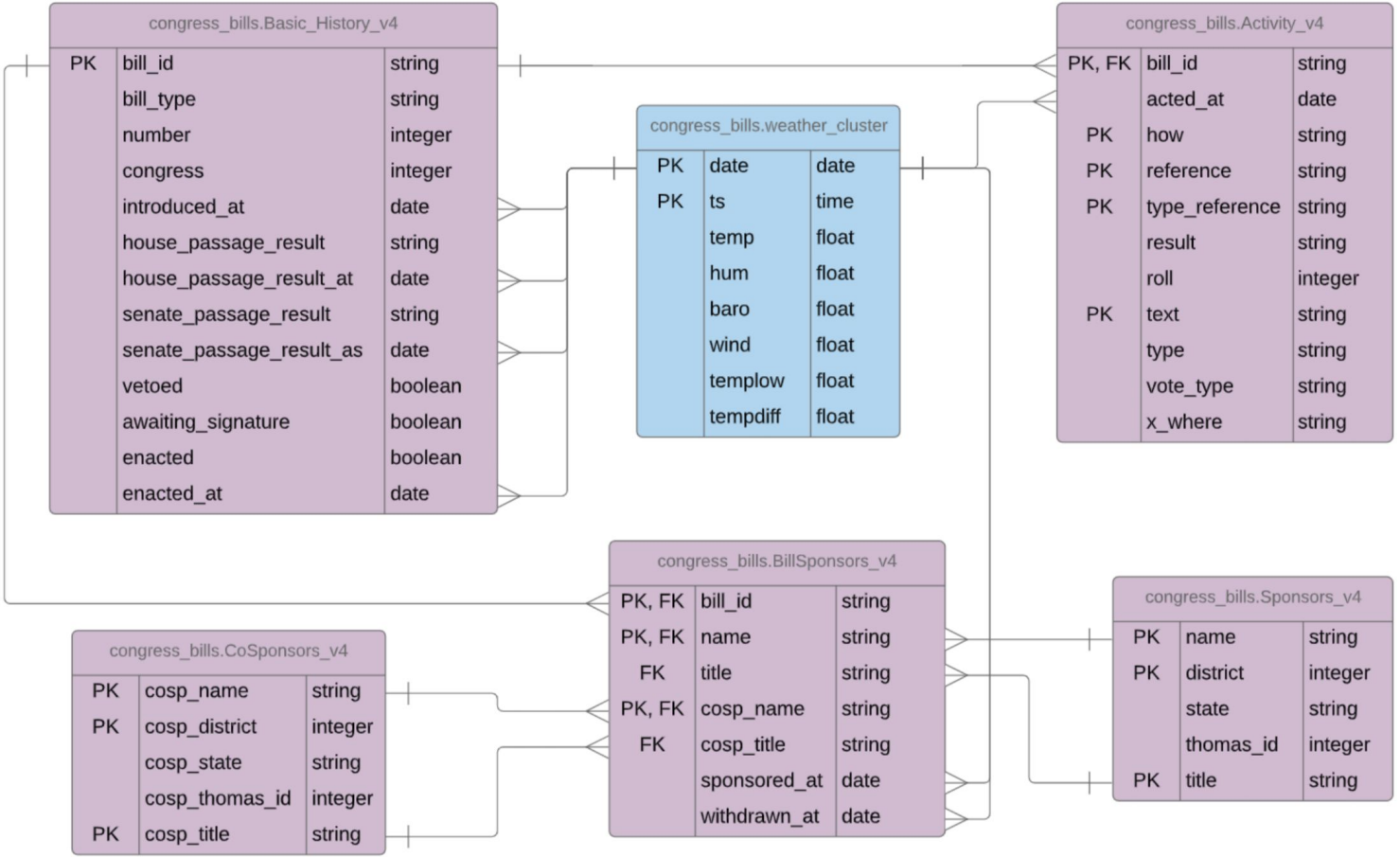
Summary

Dataset 1: U.S. Library of Congress - 107th to 112th Congress

Dataset 2: Weather Data from 2000-2012 in Washington D.C.

Question

What correlation, if any at all, is there between congressional activities and the weather in D.C?

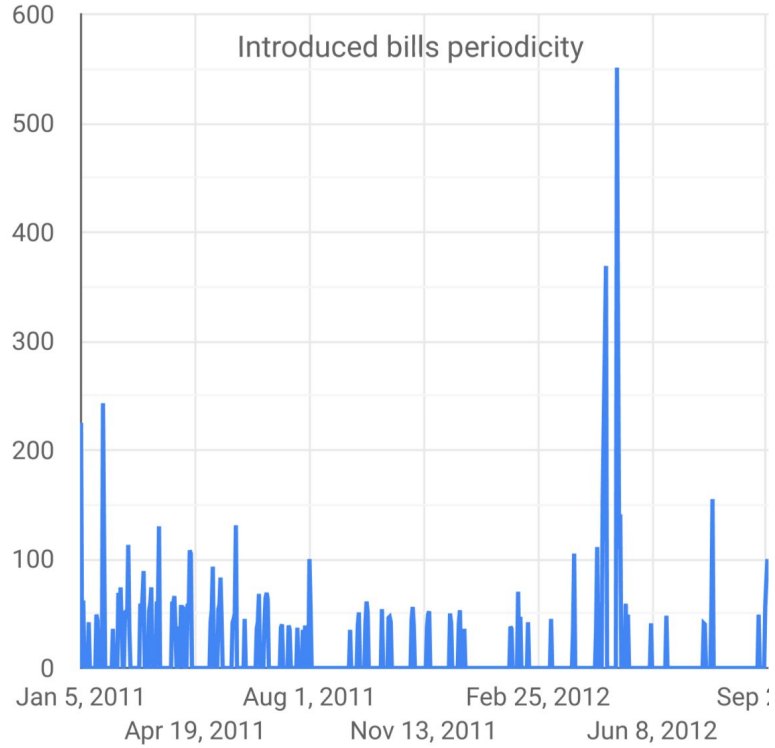

congress_bills.Basic_History_v4
PK bill_id string
bill_type string
number integer
congress integer
introduced_at date
house_passage_result string
house_passage_result_at date
senate_passage_result string
senate_passage_result_as date
vetoed boolean
awaiting_signature boolean
enacted boolean
enacted_at date
congress_bills.weather_cluster
PK date date
PK ts time
temp float
hum float
baro float
wind float
templow float
tempdiff float
congress_bills.Activity_v4
PK, FK bill_id string
acted_at date
PK how string
PK reference string
PK type_reference string
result string
roll integer
PK text string
type string
vote_type string
x_where string
congress_bills.BillSponsors_v4
PK, FK bill_id string
PK, FK name string
FK title string
PK, FK cosp_name string
FK cosp_title string
sponsored_at date
withdrawn_at date
congress_bills.CoSponsors_v4
PK cosp_name string
PK cosp_district integer
cosp_state string
cosp_thomas_id integer
PK cosp_title string
congress_bills.Sponsors_v4
PK name string
PK district integer
state string
thomas_id integer
PK title string

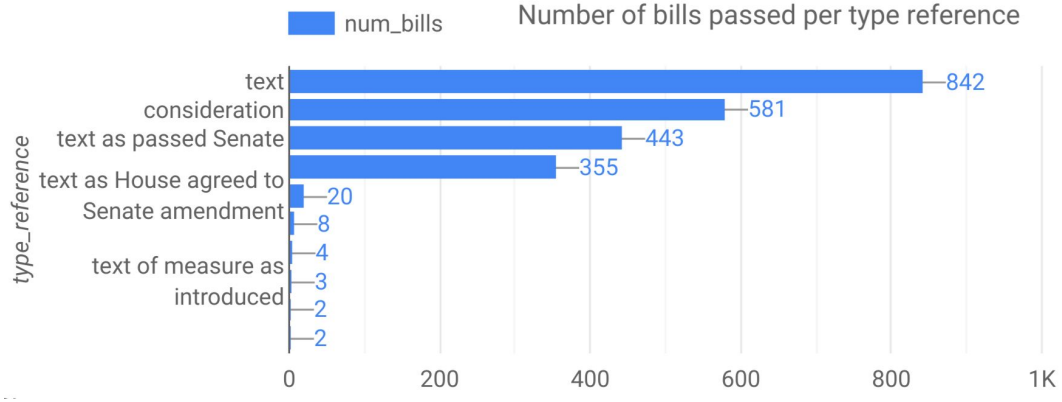
Apache Beam

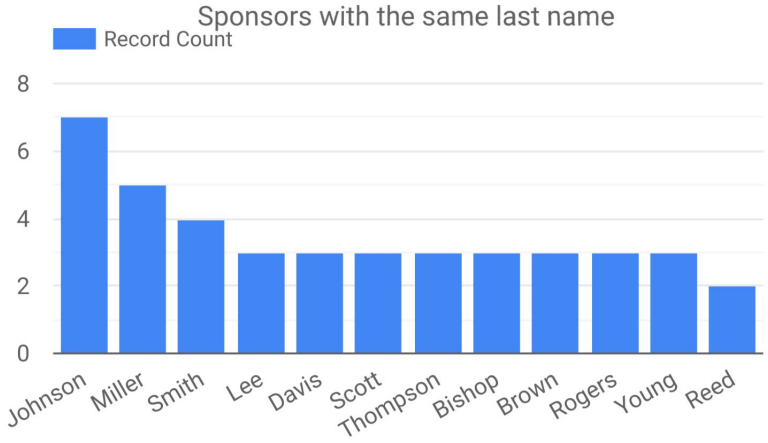
```
# Split names of sponsors into first and last names
class SponsorFirstLast(beam.DoFn):
    def process(self, element):
        values = element.get('name')
        name = values.split(',', 1)  # ["last", "first, [suffix] (title)"]
        last = name[0].strip()
        first = name[1].strip().split(' (')[0] # remove the (title)
        title = name[1].strip().split('(')[1].split(')')[0]
        # repeat for cosponsor
        values = element.get('cospons_name')
        cosp_name = values.split(',', 1)
        cosp_last = cosp_name[0].strip()
        cosp_first = cosp_name[1].strip().split(' (')[0]
        cosp_title = cosp_name[1].strip().split('(')[1].split(')')[0]
        bill_id = element.get('bill_id')
        sponsored_at = element.get('sponsored_at')
        withdrawn_at = element.get('withdrawn_at')
        return [(bill_id, first, last, title, cosp_first, cosp_last, cosp_title
                 sponsored_at, withdrawn_at)]
```

```
# Round values and create tempdiff
class TempDiff(beam.DoFn):
    def process(self, element):
        new_features = []
        features = ['date','temp','templow','tempdiff','hum','baro','wind']
        for feature in features:
            if feature == 'tempdiff':
                if new_features[1] is not None and new_features[2] is not None:
                    new_features.append(round(new_features[1]-new_features[2], 5))
                else:
                    new_features.append(None)
            elif feature == 'date':
                new_features.append(element.get(feature))
            else:
                val = element.get(feature)
                if val is not None:
                    val = round(val, 5)
                new_features.append(val)
        return [tuple(new_features)]
```

Visualizations

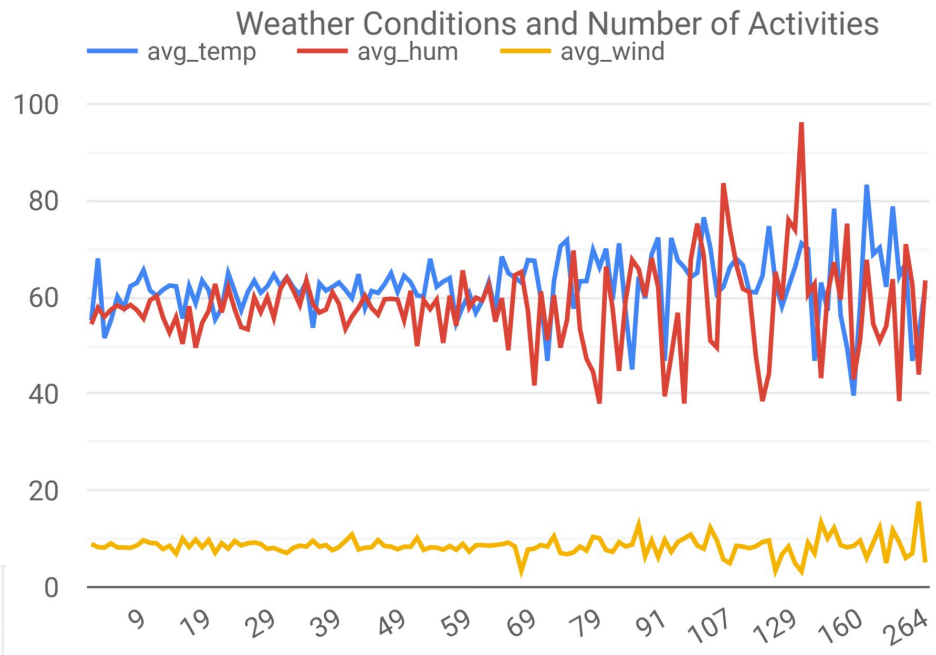






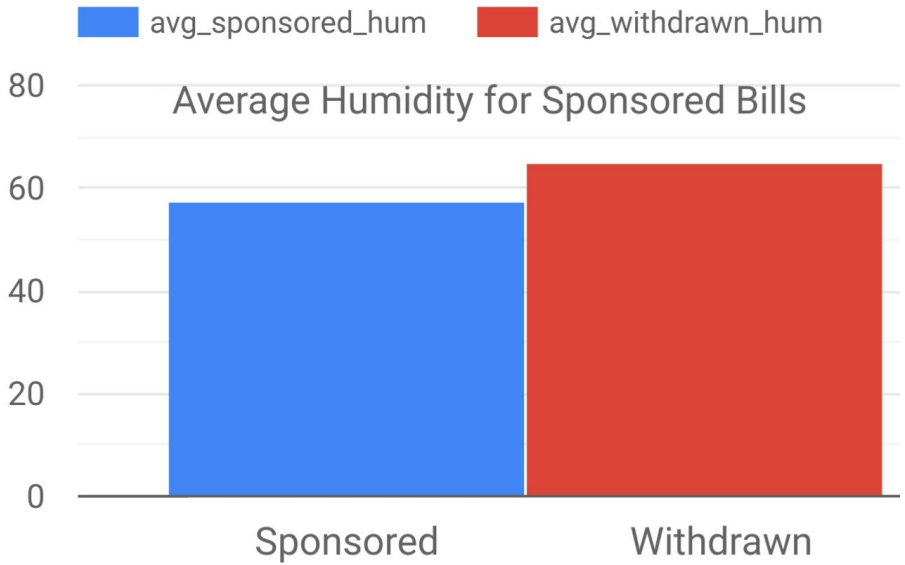
Cross-Dataset Query 1

```
-- average weather conditions for days with the same congress activity level
create view congress_bills.weather_activities as
(select num_acts,
        round(avg(temp), 3) as avg_temp,
        round(avg(hum), 3) as avg_hum,
        round(avg(wind), 3) as avg_wind
 from
    (select b.num_acts as num_acts, a.temp as temp,
          a.hum as hum, a.wind as wind
     from congress_bills.Weather_cluster a
     join
       (select acted_at, count(*) as num_acts
        from congress_bills.Activity_v4 group by acted_at) b
     on a.date=b.acted_at)
 group by num_acts
 order by num_acts desc)
```



Cross-Dataset Query 2

```
-- on average it is 8% more humid on dates the sponsors withdrew the bill than the dates they sponsored it
create view congress_bills.hum_sponsored_withdrawn as
(select c.avg_hum as avg_sponsored_hum,
        d.avg_hum as avg_withdrawn_hum
 from
   (select avg(b.hum) as avg_hum
    from congress_bills.BillSponsors_v4 a
   join
     congress_bills.Weather_cluster b
   on a.sponsored_at = b.date) c
 cross join
   (select avg(b.hum) as avg_hum
    from congress_bills.BillSponsors_v4 a
    join
      congress_bills.Weather_cluster b
    on a.withdrawn_at = b.date) d)
```



Airflow


create_weather_table
group_by_weather_table
tempdiff_weather_table

Future

Analyze congress data with more data:

- Major world events
- National disasters
- International politics
    - Comparing side-by-side with another country

Thank you!

Any Questions?