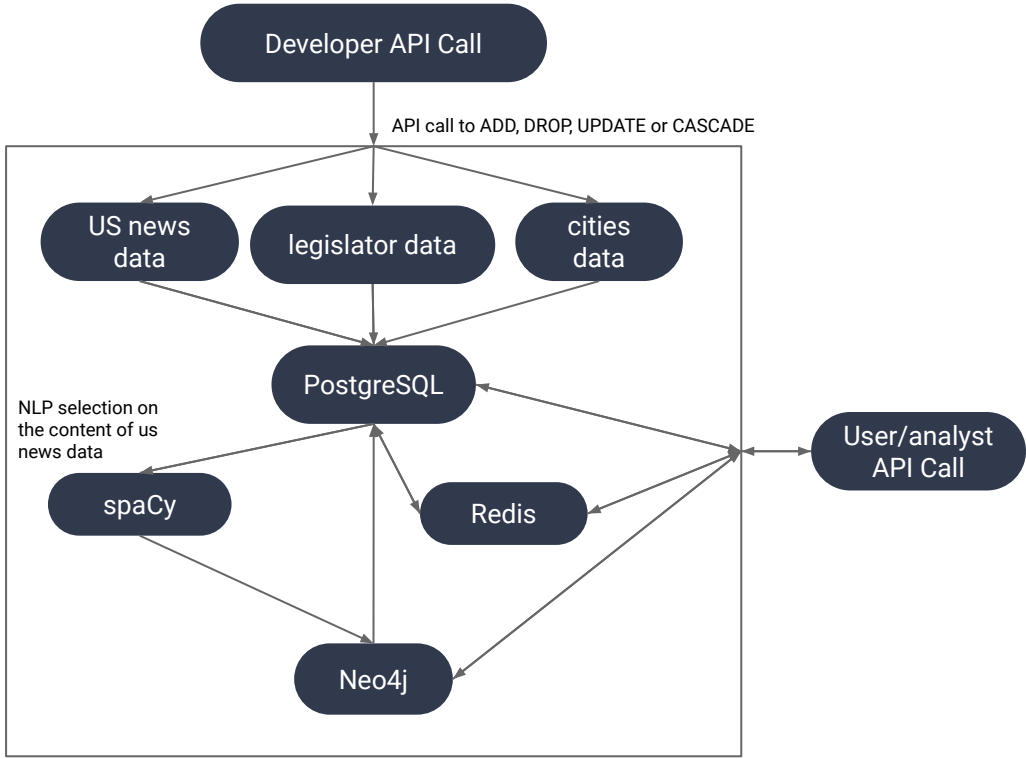
Integrated Cache Based NLP and Graph Analysis in US News

by Jiaxi Lei and Dennis Wu

What we've done in the project

- PostgreSQL
- Redis as cache to better optimize the query speed when DB size is huge and queries are similar
- spaCy as the non-destructive means to procure the NLP analysis of the news content
- Neo4j to analyze the network indicators concurrently

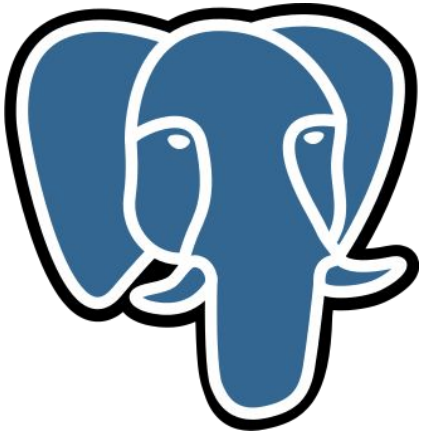


PostgreSQL

What’s our procedure?

Why Did We Choose PostgreSQL?

- Due to its stability, efficiency, and maneuverability, our group started off in PostgreSQL and implement it as the target database in data processing and results presentation
- However, in some industries like telecommunication and large-scale social media platforms, the size of the database can continuously grow, and numerous similar queries can be called on the data itself to perform analysis.



Redis as Cache

What’s on top of Postgres?

Caching PostgreSQL

We are using **Redis** as a TTL cache memory to provide faster querying service on the expected rather large Postgres database.

Eviction Policy: TTL

Using Time-To-Live eviction policy with a volatile-TTL on max memory

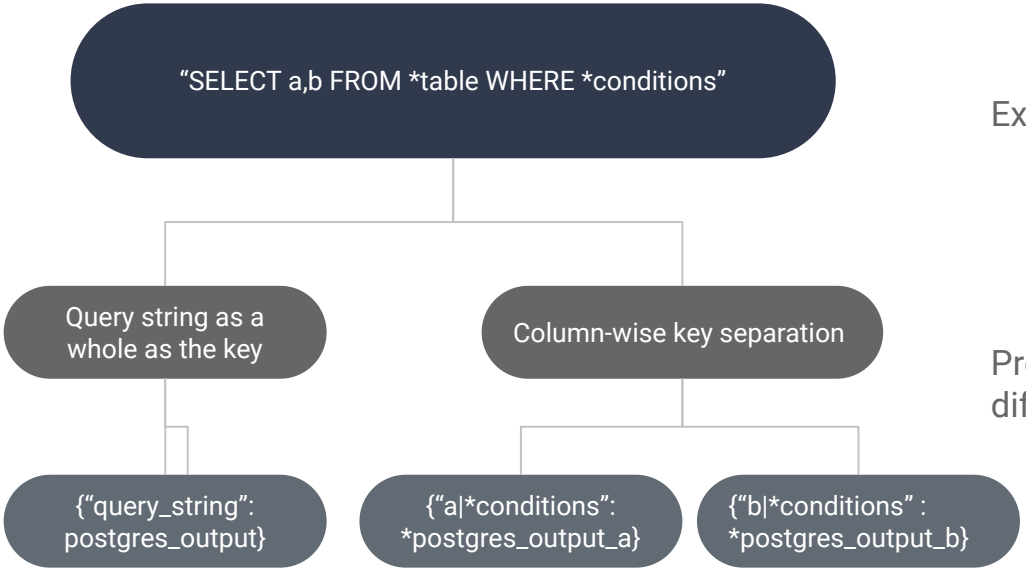
Configurations:

Default expiration time: 300s

Cache max-memory: 4mbs

Refresh live-time policy: every search

Caching PostgreSQL



Expects **same input** as querying postgres database

Provides two **different key:value storing methods** for different user preference

Caching PostgreSQL

We are using **Redis** as a TTL cache memory to provide faster querying service on the expected rather large Postgres database.

```
In [54]: %%time
cur.execute(query)
out = cur.fetchall()
out

CPU times: user 0 ns, sys: 1.63 ms, total: 1.63 ms
Wall time: 1.42 ms

Out[54]: [(4230220,
  '2020-09-08',
  ' How Trump’s Billion-Dollar Campaign Lost Its Cash Adva
  'Money was supposed to have been one of the great advant
Barack Obama in 2012 and George W. Bush in 2004. After get
his inauguration — earlier than any other modern president
advantage this year.\n\nIt seemed to have worked. His riva
resumptive Democratic nominee this spring, and Mr. Trump a
dvantage.\n\nFive months later, Mr. Trump’s financial supr
aised from the beginning of 2019 through July, more than $
gn are forecasting what was once unthinkable: a cash crunc
officials briefed on the matter.\n\nBrad Parscale, the for
ne an “unstoppable juggernaut.” But interviews with more t
review of thousands of items in federal campaign filings,
ligate habits as they burned through hundreds of millions
paign has imposed a series of belt-tightening measures tha
he advertising budget.')]

In [59]: %%time
out = redisLRU_get_pg(query, cur, r)
out

CPU times: user 2.17 ms, sys: 680 µs, total: 2.85 ms
Wall time: 3.88 ms
```

Extracting News

Who are mentioned in news articles? (and where?)

NLP on US News

Driven motivation:

In order to analyze legislator relationships from a graph database standpoint, we need to establish a procedure to reliably and faithfully setup the relations in legislator (person) nodes. We **extract keywords** in these articles, including name of **person and location**, to establish a **mentioned** relationship between legislators/locations and news articles. Here we chose to use the US News dataset which contains **10000 news articles** as a small-size demoing dataset.

We use **spaCy** Nature Language Processing API as our method to analyze news corpus.

In a real world scenario, we expect this step to be done in database management (at the process of adding new article data into system). However, in this project this NLP process is performed after populating Postgres as a performance enhancement.

NLP on US News

So what does Mr. Cuomo want to do? Pressure the federal government for help, despite the slim hopes of Democrats and Republicans reaching a deal that would solve all of New York's financial problems.

NLP keyword extraction:
extract unique location and person strings

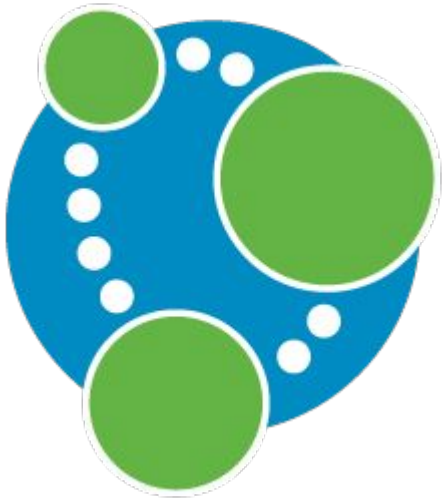
cuomo, new york

Neo4j Query

Oh, query.

Why Did We Choose Neo4j?

- With its high flexibility, we were able to update/query the data in more possible ways
- We could utilize its graph nature to analyze some indicators
- Updates in the target Postgres DB can be updated rather immediately.



Therefore, Neo4j enables us to generate a perspicacious view into the data itself. With its low response time and ready-to-go methods, we are able to create some great methods for the users.

APIs Created to Analyze the Network

- **Shortest Path Between the Legislators**

Inputs:

first

default = “Kyrsten Sinema”

second

default = “new york”

first: “Kyrsten Sinema”, second: “new york” shortestpath = 4

This gives us insight into the data that those Sinema and New York aren’t likely to be mentioned in the same document, so probably they are likely not related to each other.

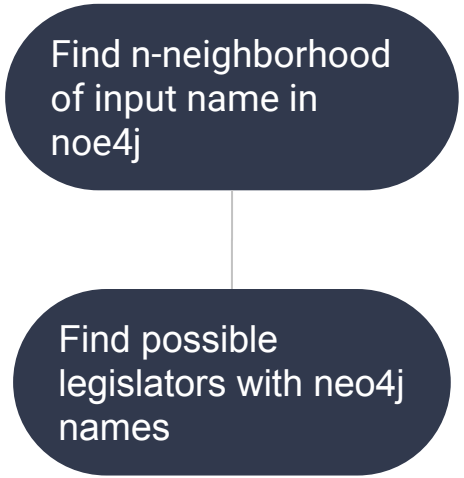
first: “Donald Trump”, second: “new york” shortestpath = 2

This means that the two nodes are connected by a news node.

Possible Legislators in N-Neighborhood

Out of all the people in one person's n-neighborhood, how many of them are potential current legislators?

This is made possible with our efficient caching query method and our graphical database.



Inputs:

name_string

default="Donald Trump"

n:

default=2

Possible Legislators in N-Neighborhood

name: “Kyrsten Sinema”, n: 2 total distinct count 7

{('Doug', 'Jones'), ('Joni', 'Ernst'), ('Kelly', 'Armstrong'), ('Kelly', 'Loeffler'), ('Mike', 'Kelly'), ('Robin', 'Kelly'), ('Trent', 'Kelly')}

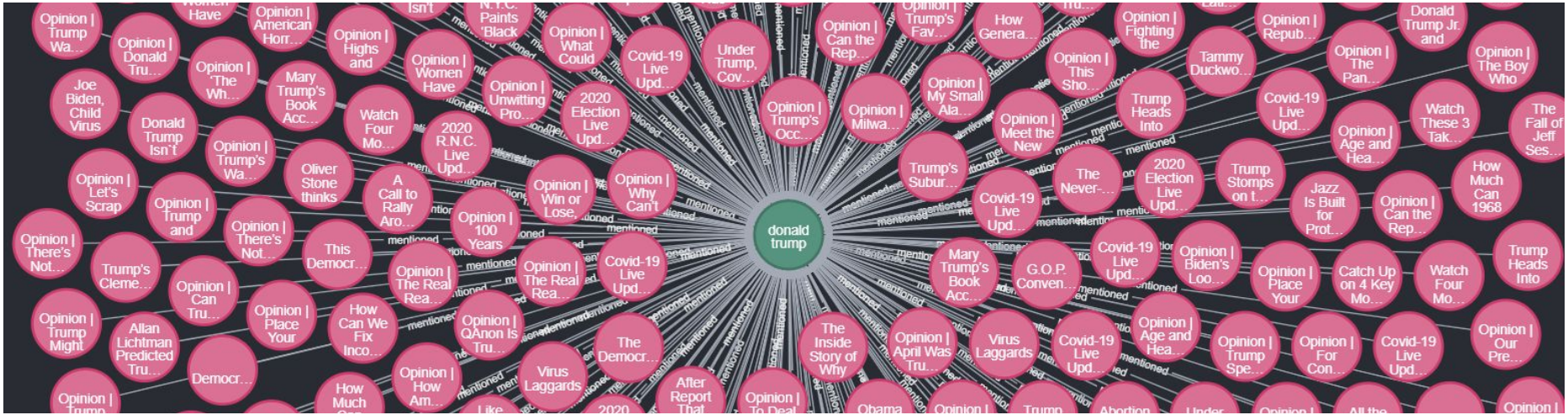
name: “Trump”, n: 2 total distinct count 357

name: “Trump”, n: 4 total distinct count 451

name: “Trump”, n: 6 total distinct count 460

Why so many Kellys?

Local Clustering Coefficient



Local clustering coefficient (LCC) gives different insight, even-to-say completely opposite insight, due to our Knowledge Graph setup. In our graph, there only exists one type of relation: **mentioned**, which must link a **news article node** to a **person/location node**.

Therefore, LCC can be a good indicator of the size of a clustering n-neighborhood.

Local Clustering Coefficient

Demo

Donald Trump: 0.00164

Kyrsten Sinema: 0.09942

Bill Stepien: 0.00921

Thanks!

Question time