Análise de Dados com Base em Processamento Massivo em Paralelo

Twitter Data Lake

André Perez
ICMC/USP
andre.marcos.perez@usp.br
linkedin.com/in/andremarcosperez/

USP

ICMC USP SÃO CARLOS

CeMEAI
CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

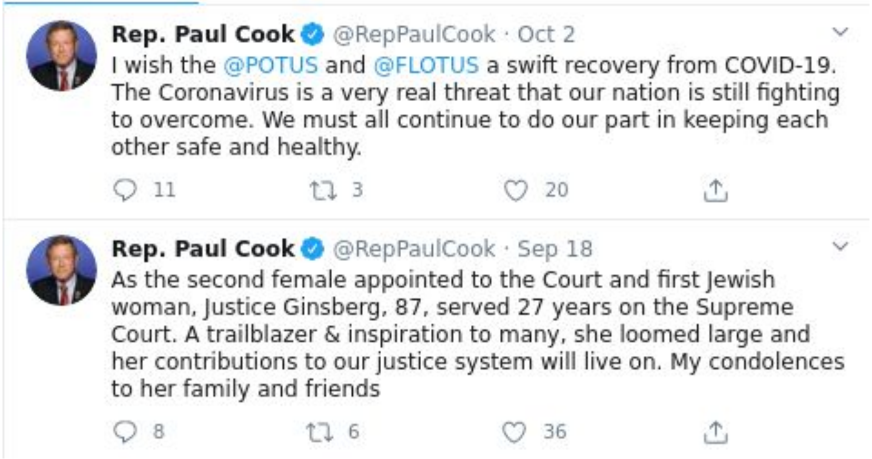
Agenda

- Contexto
- Objetivo
- Arquitetura
- Links

Agenda

- Contexto
- Objetivo
- Arquitetura
- Links

Política dos EUA no Twitter

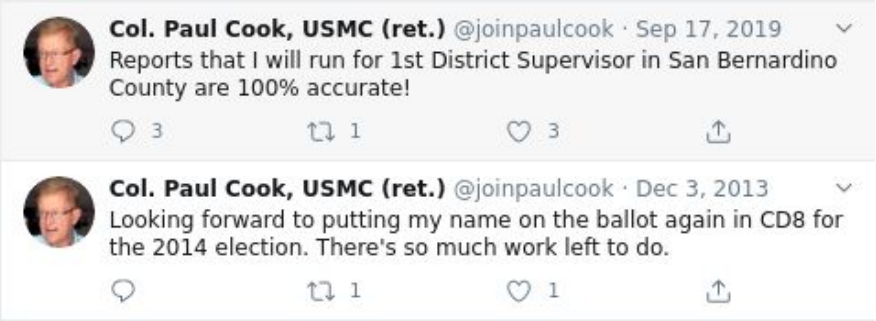


Rep. Paul Cook @RepPaulCook · Oct 2
I wish the @POTUS and @FLOTUS a swift recovery from COVID-19. The Coronavirus is a very real threat that our nation is still fighting to overcome. We must all continue to do our part in keeping each other safe and healthy.

11 3 20

Rep. Paul Cook @RepPaulCook · Sep 18
As the second female appointed to the Court and first Jewish woman, Justice Ginsberg, 87, served 27 years on the Supreme Court. A trailblazer & inspiration to many, she loomed large and her contributions to our justice system will live on. My condolences to her family and friends

8 6 36



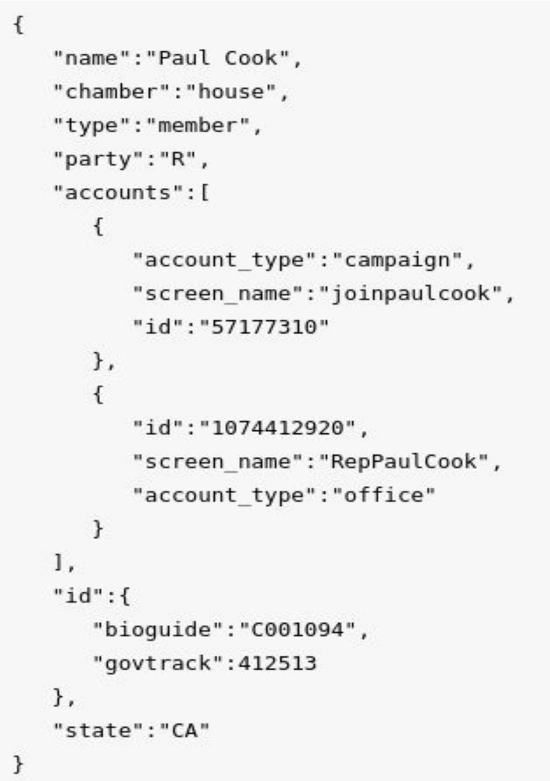
Col. Paul Cook, USMC (ret.) @joinpaulcook · Sep 17, 2019
Reports that I will run for 1st District Supervisor in San Bernardino County are 100% accurate!

3 1 3

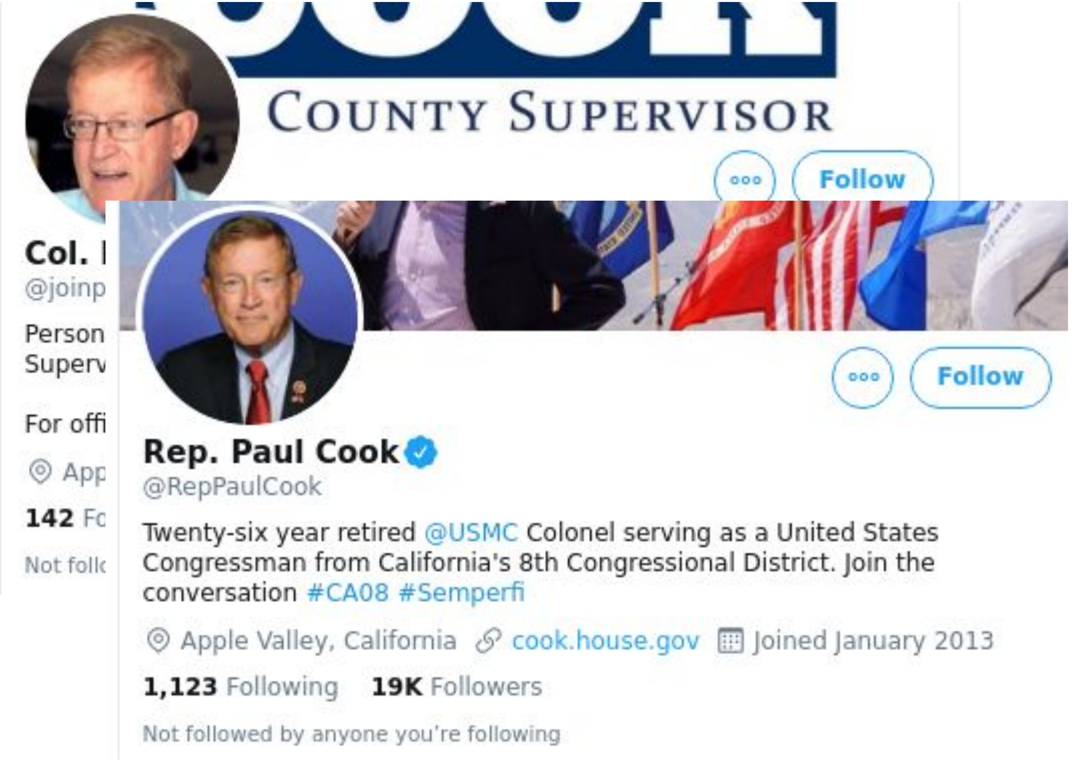
Col. Paul Cook, USMC (ret.) @joinpaulcook · Dec 3, 2013
Looking forward to putting my name on the ballot again in CD8 for the 2014 election. There's so much work left to do.

1 1

Congress Tweets Automator



```
{
    "name":"Paul Cook",
    "chamber":"house",
    "type":"member",
    "party":"R",
    "accounts":[
       {
          "account_type":"campaign",
          "screen_name":"joinpaulcook",
          "id":"57177310"
       },
       {
          "id":"1074412920",
          "screen_name":"RepPaulCook",
          "account_type":"office"
       }
    ],
    "id":{
       "bioguide":"C001094",
       "govtrack":412513
    },
    "state":"CA"
}
```



Col. I
@joinp
Person
Superv
For offi
App
142 Fo
Not follo

Rep. Paul Cook

@RepPaulCook

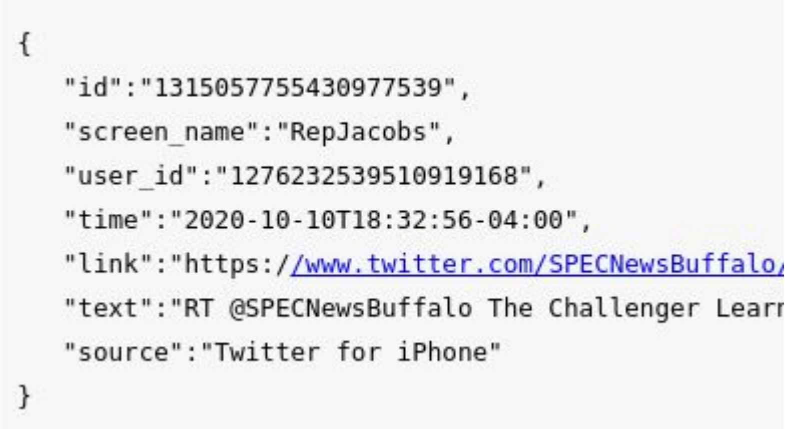
Twenty-six year retired @USMC Colonel serving as a United States Congressman from California's 8th Congressional District. Join the conversation #CA08 #Semperfi

Apple Valley, California cook.house.gov Joined January 2013

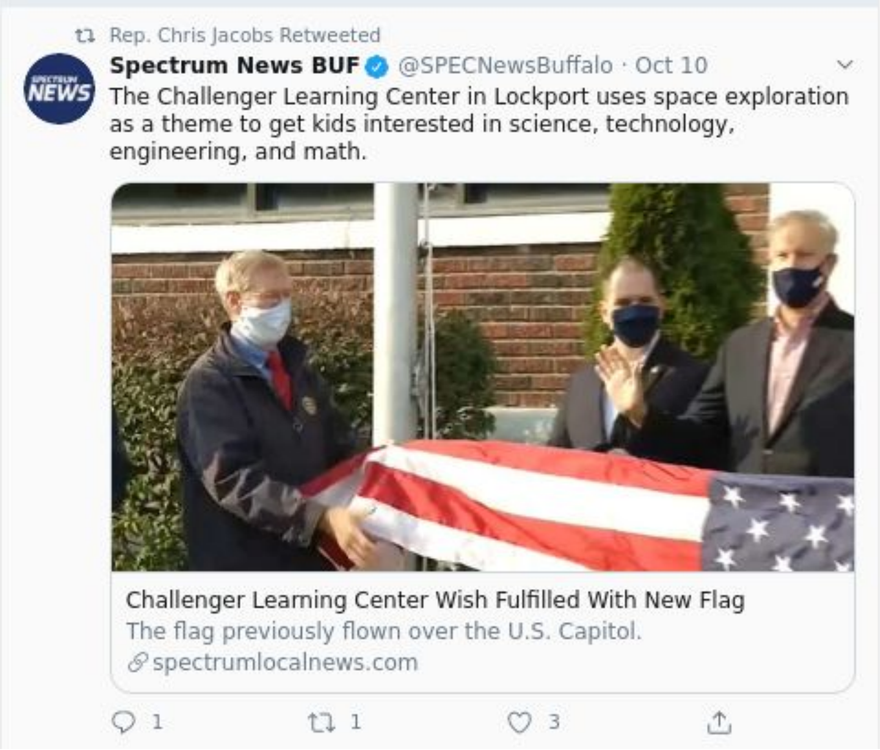
1,123 Following **19K** Followers

Not followed by anyone you're following

Congress Tweets



```
{
    "id":"1315057755430977539",
    "screen_name":"RepJacobs",
    "user_id":"1276232539510919168",
    "time":"2020-10-10T18:32:56-04:00",
    "link":"https://www.twitter.com/SPECNewsBuffalo/
    "text":"RT @SPECNewsBuffalo The Challenger Learn
    "source":"Twitter for iPhone"
}
```



Rep. Chris Jacobs Retweeted

Spectrum News BUF @SPECNewsBuffalo · Oct 10

The Challenger Learning Center in Lockport uses space exploration as a theme to get kids interested in science, technology, engineering, and math.

Challenger Learning Center Wish Fulfilled With New Flag

The flag previously flown over the U.S. Capitol.

spectrumlocalnews.com

Agenda

- Contexto
- Objetivo
- Arquitetura
- Links

Construir um pipeline de dados que ingere dados diariamente em um data lake para armazenamento e exploração.

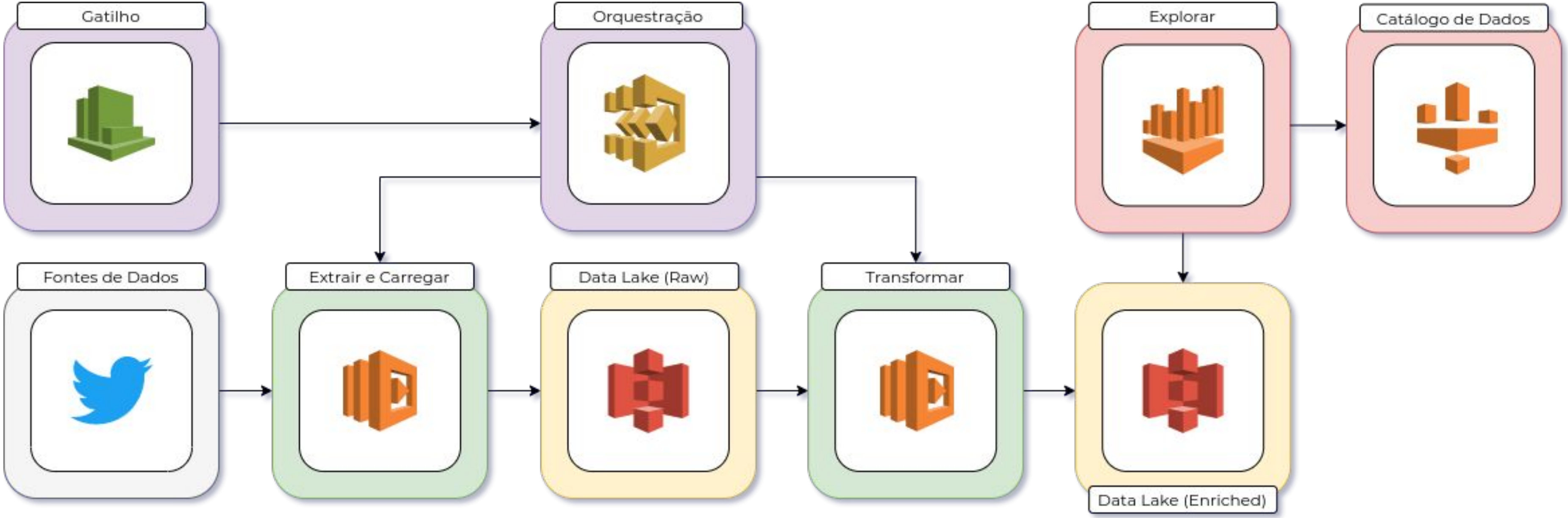
Agenda

- Contexto
- Objetivo
- Arquitetura
- Links

Arquitetura: Elementos

ELT	Data Lake	Motor de Consulta	Orquestração
AWS Lambda	AWS s3	AWS Athena	AWS Step Functions
		AWS Glue	AWS CloudWatch

Arquitetura: Pipeline



Agenda

- Contexto
- Objetivo
- Arquitetura
- Links

Links

Fonte de dados

- Congress Tweets Automator
- Congress Tweets

Amazon Web Services

- AWS Service Docs
- AWS Simple Calculator
- AWS Newest Calculator
- AWS to Azure (Microsoft) Map
- AWS to GCP (Google) Map

Links

Repositório de códigos

- Data Science MBA @ ICMC-USP

Artigos

- Demystify Hadoop Data Formats: Avro, ORC, and Parquet
- New in Hadoop: You should know the Various File Format in Hadoop