



Earthquake Twitter Detection



BeCode-Brussels
27/01/2023

The Project

Provide DataForGood with an **internal dashboard** for **quickly detecting earthquakes** and their **locations** so they can act and help citizens impacted by these earthquakes.

Use Twitter as a primary source of information

- 350 million users.
- > 500 million tweets sent every day.



Challenges: massive, noisy, unstructured and dynamic data.

Engineering team



Data Engineer



Isidora
CUPARA



Ahmet
CICEK



Kevin
NOBLES



David
VALDEZ

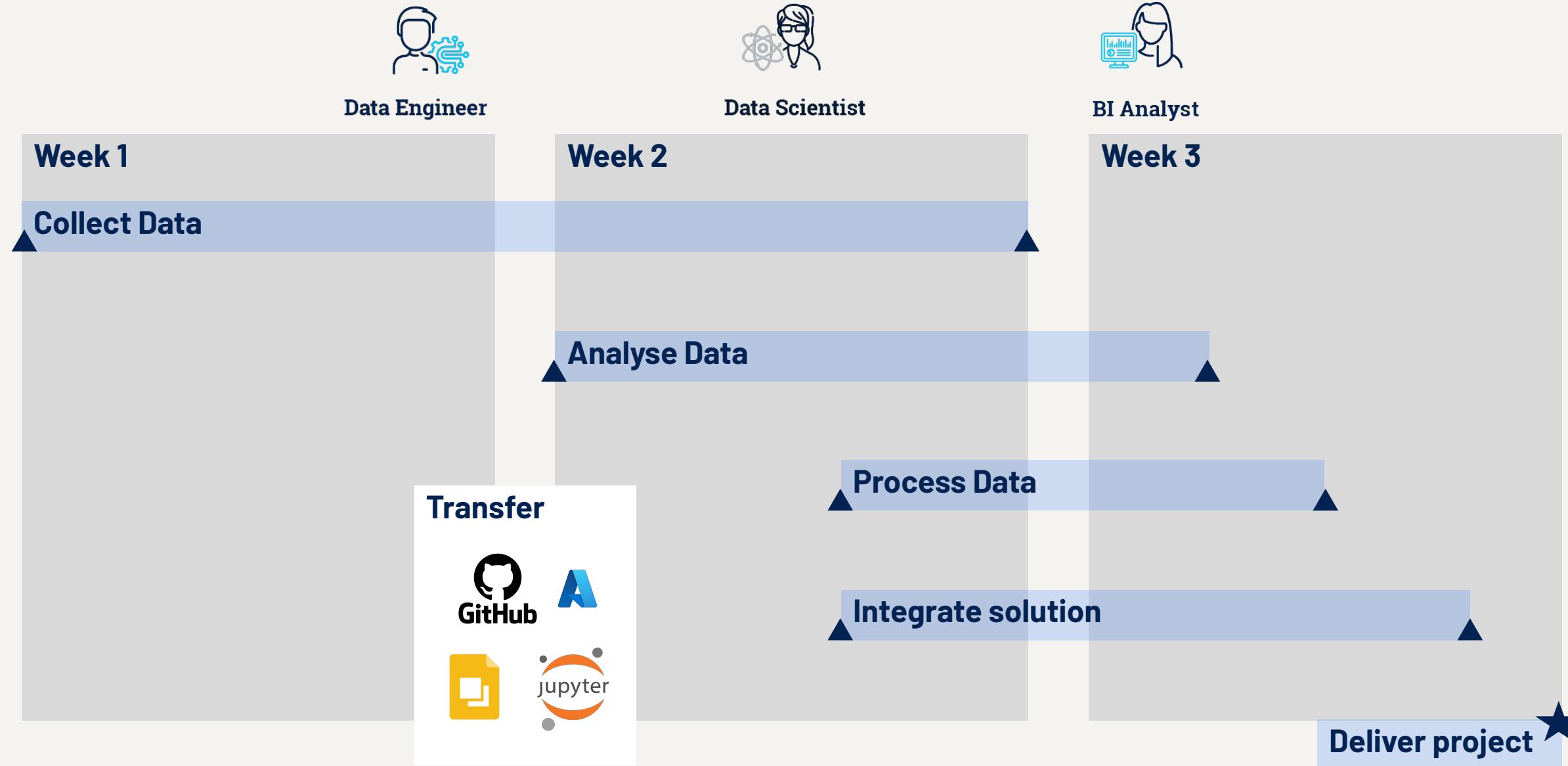


Louis
NOLLEVAUX



Antoine
VANNESTE

Management



Project Pipeline

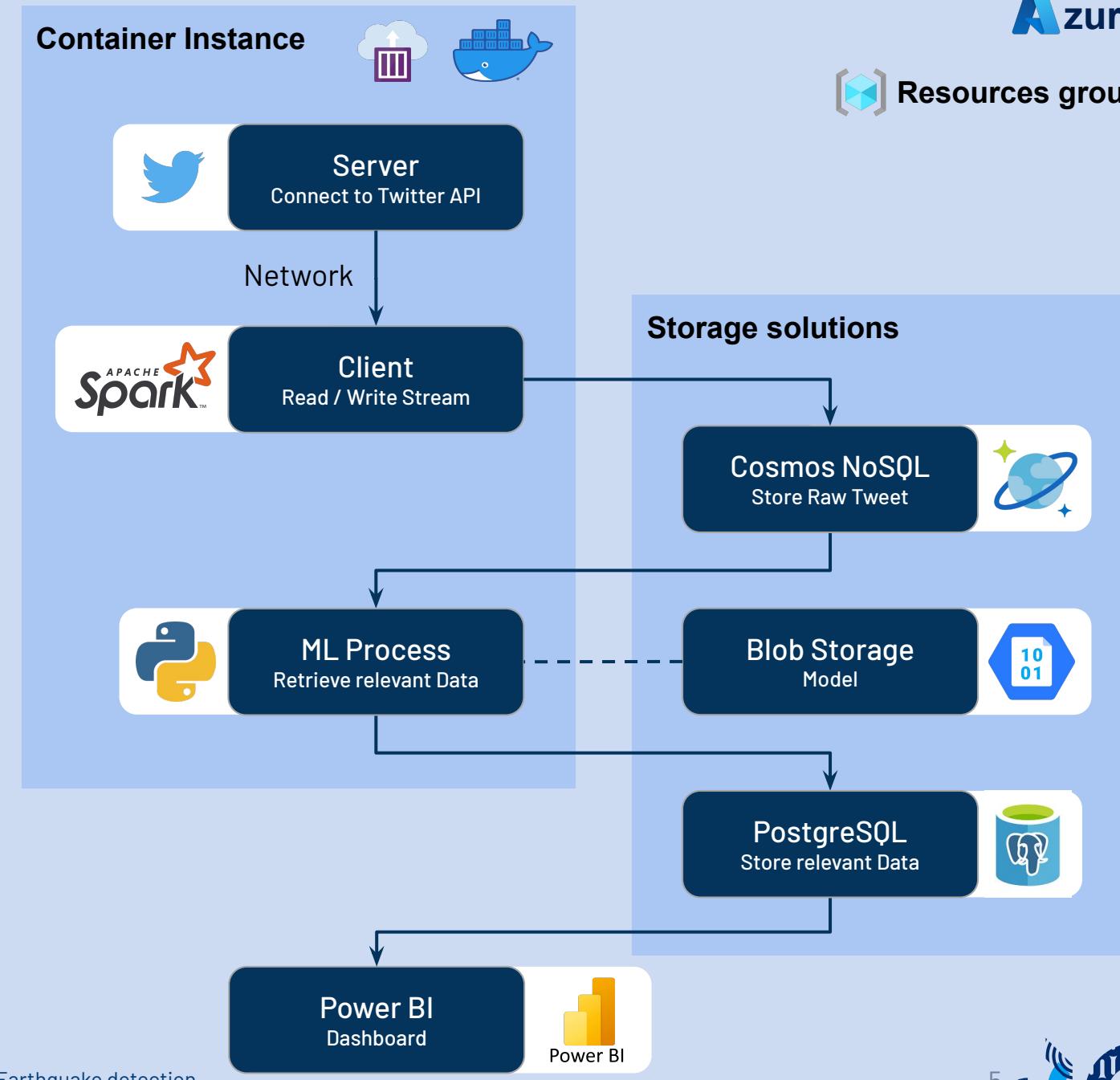
Azure



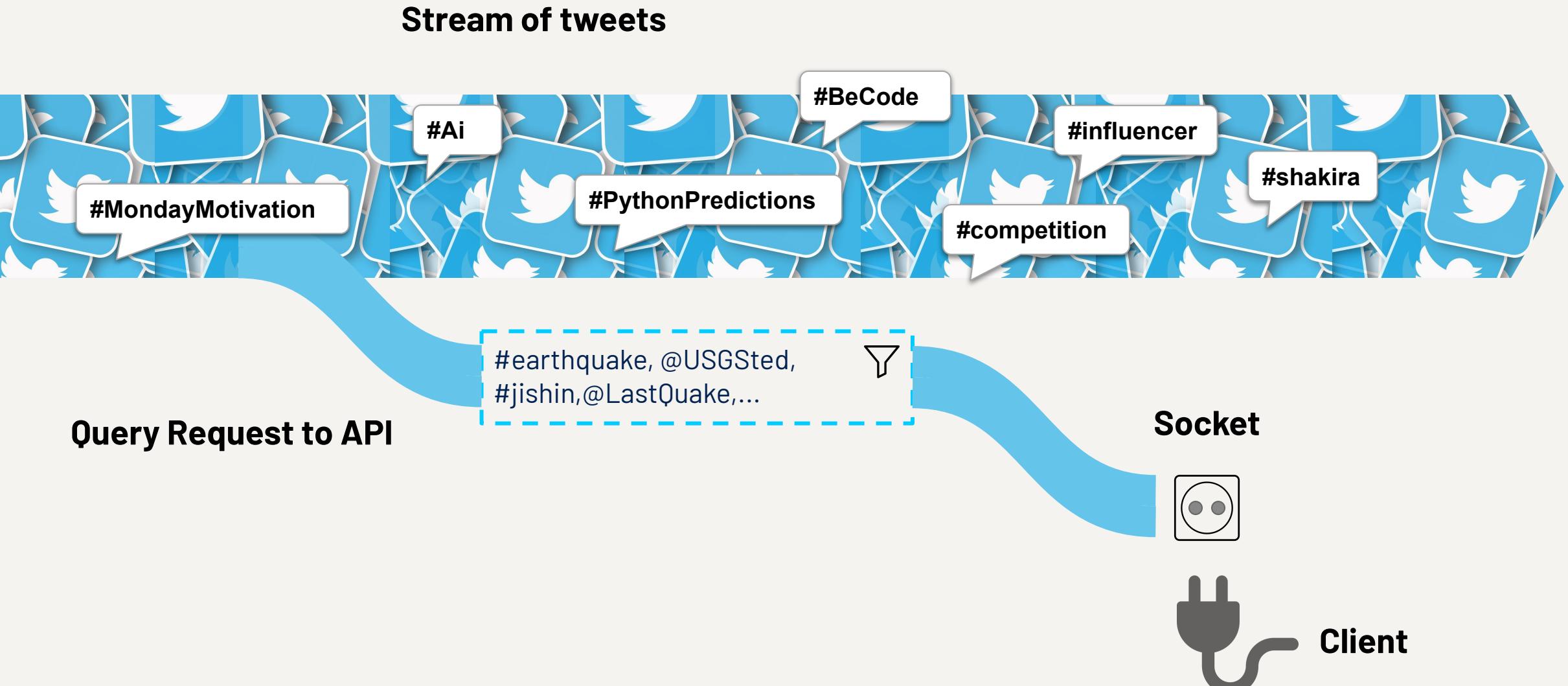
Easily deployed with containers

Easy scaling

Reproducible environment



Server/ Twitter API (A2)



Client/ Pyspark Data Stream



Every Earthquake ✅
@everyEarthquake

USGS reports a M2.4 earthquake, 39 km W of
Mentone, Texas on 1/17/23 @ 4:49:57 UTC
earthquake.usgs.gov/earthquakes/ev...
#earthquake

10:58 AM · Jan 17, 2023 from Texas, USA · 162 Views

Every Earthquake ✅
@everyEarthquake

USGS reports a M2.4 earthquake, 39 km W of
Mentone, Texas on 1/17/23 @ 4:49:57 UTC
earthquake.usgs.gov/earthquakes/ev...
#earthquake

10:58 AM · Jan 17, 2023 from Texas, USA · 162 Views

Every Earthquake ✅
@everyEarthquake

USGS reports a M2.4 earthquake, 39 km W of
Mentone, Texas on 1/17/23 @ 4:49:57 UTC
earthquake.usgs.gov/earthquakes/ev...
#earthquake

10:58 AM · Jan 17, 2023 from Texas, USA · 162 Views

TWITTER STREAMING API



COSMOS DB NoSQL

Storage/ raw tweets

Why Azure Cosmos DB?

ELT Pipeline use case

Fully managed service

Easy to deploy

Schema-agnostic

Auto-indexing for every item

Scalable

```
1  [
2      "id": "1618160056402587648",
3      "tweet_content": {
4          "data": {
5              "attachments": {},
6              "author_id": "913948904730673157",
7              "conversation_id": "1618160056402587648",
8              "created_at": "2023-01-25T08:13:16.000Z",
9              "edit_history_tweet_ids": [
10                  "1618160056402587648"
11              ],
12              "entities": {},
13              "geo": {},
14              "id": "1618160056402587648",
15              "lang": "ja",
16              "possibly_sensitive": false,
17              "public_metrics": {
18                  "retweet_count": 0,
19                  "reply_count": 0,
20                  "like_count": 0,
21                  "quote_count": 0,
22                  "impression_count": 0
23              },
24              "text": "earthquake\n地震"
25          },
26          "includes": {
27              "users": [
28                  {
29                      "created_at": "2017-09-30T02:09:34.000Z",
30                      "description": "TOEFL60点くらいまでの単語を淡々とつぶやくbotです。リプライで単語を返します。",
31                      "id": "913948904730673157",
32                      "name": "TOEFL 60点単語bot",
33                      "profile_image_url": "https://pbs.twimg.com/profile_images/1091244118208958464/veBGVRUC_norm
```



Storage/ structured data

Why PostgreSQL Database?

Store processed/transformed data

Easy to use by data analysts

Quick integration with PowerBI

	id	text	geo_lat	geo_lng	date	time	label	magnitude	author_id	created_at	tweet_lang
	1617465854827831296	@terremoto_44 Exactly, If you're making it abo...	None	None	None	None	3	3.0	794618296956030976	2023-01-23T10:14:45.000Z	en
	1617466179022426113	RT @URDailyHistory: 23 Jan 1556: What is thou...	35.88941	109.13573	1556-01-23	None	0	3.0	704934261414109185	2023-01-23T10:16:02.000Z	en
	1617466449253269506	RT @anc_party: Many wonder why H.E @MusaliaMud...	None	None	None	None	3	3.0	1570368522517102595	2023-01-23T10:17:07.000Z	en
	1617466838576926722	@BIO99_BIO99 Can someone explain what this mea...	None	None	None	None	0	3.0	1224973611452313600	2023-01-23T10:18:40.000Z	en
	1617468180842123264	@DaisyKenyan_ They thought the deep state woul...	None	None	None	None	3	8.0	2747335167	2023-01-23T10:24:00.000Z	hi

	1617555952370348032	[Microearthquake Early Warning Wakayama Prefec...	None	None	None	00:40:39	0	7.0	406047243	2023-01-23T16:12:46.000Z	ja
	1617556776760524804	The National Seismological Center indicates th...	-30.83375	-71.25736	None	None	1	0.0	112431151	2023-01-23T16:16:02.000Z	es
	1617556882050396162	@Rubenempelotas @slifante That has caught my a...	None	None	None	None	3	1.0	1235540483671240704	2023-01-23T16:16:28.000Z	es

Storage/ ML Model

Why Azure Blob Storage?

Supports unstructured data

Relatively low cost

Accessible from anywhere

Ability to set different accessibility



Alternative Pipeline with Azure online services

Logic Apps Get tweets

Cognitive services Text analytics API

Embedded Power BI streaming dataset



Price Estimate / month

Azure (fully managed)

Logic Apps	\$180
Cognitive Services	\$650
Power BI Embedded	\$750
TOTAL (estimate)	\$1580

From [Azure pricing calculator](#)

Our solution

Container Registry	\$20
Cosmos NoSQL	\$100
PostgreSQL	\$150
Blob storage	\$20
Maintenance (2 days / month @ \$200-400 / day)	\$400 - 800
TOTAL (estimate)	\$690 - 1090



Final thoughts

Challenges

Design a good pipeline from day one

Work with stream (real time data)

Deploy and debug on the cloud

Organization

Improvements

Pipeline MVP with fake data ASAP

Naming convention guidelines

Deploy ML code into the cloud

Improve communication



Questions ?

Go to Demo

The screenshot shows the Microsoft Azure portal interface for the 'Bouman-earthquake' resource group. The left sidebar contains navigation links for Activities, Google Chrome, Microsoft Azure, Home, Resource groups, and the current 'Bouman-earthquake' group. The main content area displays the 'Overview' tab for the resource group, showing details like Subscription (move) to Microsoft Azure Sponsorship, Subscription ID, Tags, and Deployments (8 Succeeded). The 'Resources' tab lists 12 records, including Container registry, Synapse workspace, Azure Cosmos DB account, Storage account, Container registry, Azure Cosmos DB for PostgreSQL Cluster, Power BI Embedded, Container registry, Container registry, Container registry, Container instances, and Container instances. Each resource entry includes a checkbox, a name, its type, location, and a three-dot menu icon.

Name	Type	Location
antoinebouman	Container registry	West Europe
bouman-data-analyst	Synapse workspace	West Europe
bouman-earthquake-db	Azure Cosmos DB account	West Europe
boumanearthquakedl	Storage account	West Europe
davidcontainer	Container registry	West Europe
earthquake-tweets-structured	Azure Cosmos DB for PostgreSQL Cluster	West Europe
earthquakeembeddedbouman	Power BI Embedded	West Europe
isidora	Container registry	West Europe
isidoraearthquake	Container registry	West Europe
kevineqcontainer	Container registry	West Europe
maxim	Container registry	West Europe
xxx_isidora	Container instances	West Europe



Data Scientist team



Data Scientist



Olivier
DE TIMMERMAN



Rafaella
PORTO



Piero
RUCCI



Olivier
LE DIBERDER



Tania
LEMOS



Alexis
VANDRIESSCHE



Julien
DE SMET



Esra
OGUZ

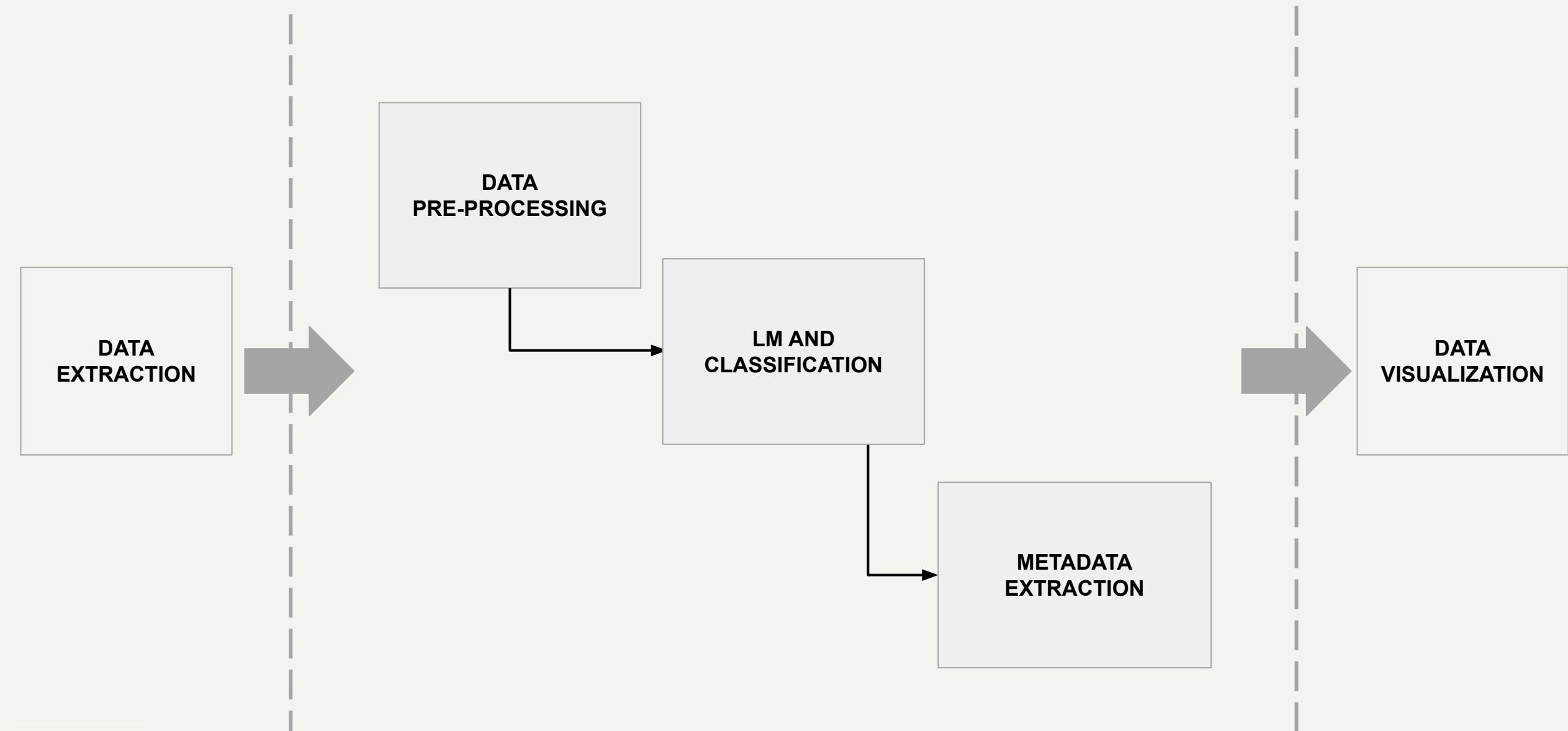


Jack
XIN



Carlton
NJUGUNA

Language Model



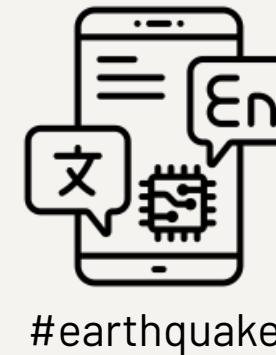
Language Detection & Translation



130 words:

Cloud Translation	
<input type="button" value="Edit"/>	<input type="button" value="Delete"/>
Text Translation: 5,200,000 characters	USD 94.00
Language Detection: 5,200,000 characters	
USD 188.00	

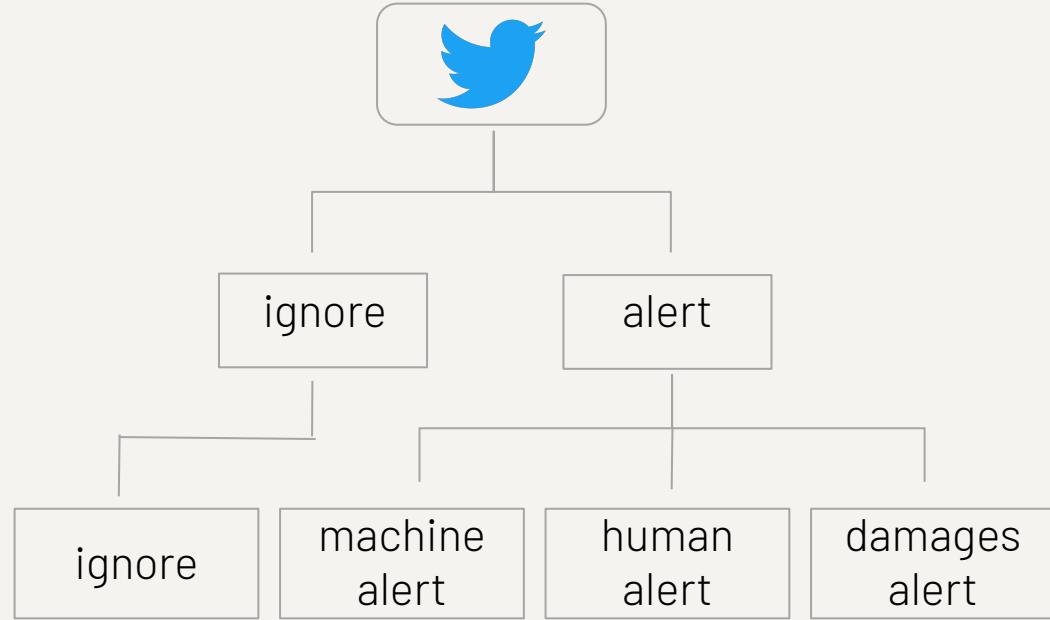
deep_translator →



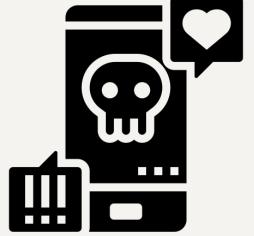
280 words:

Cloud Translation	
<input type="button" value="Edit"/>	<input type="button" value="Delete"/>
Text Translation: 11,200,000 characters	USD 214.00
Language Detection: 11,200,000 characters	
USD 428.00	

Twitter Classifier Model



Difficulty to manually classify tweets to train the model



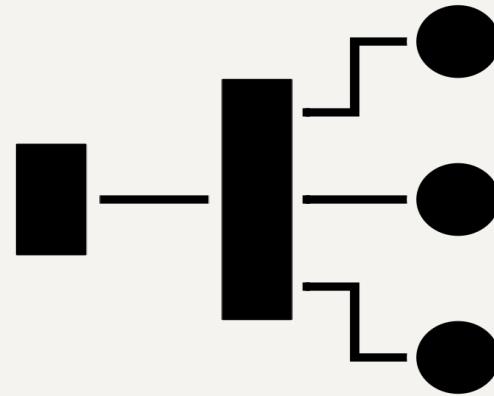
Time to search, understand and apply state-of-the-art models



CLASSIFIER - Model selection

Vectorization & Classification
models

BERTweet



Why:

- **Powerful** model
- Bidirectional (good understanding of the **context**)
- Best in class on **Tweet data**

Drawbacks:

- Specialized for **English**
- Large number of **parameters**
- **Black box** system
- **Heavy**



Notes

Measures

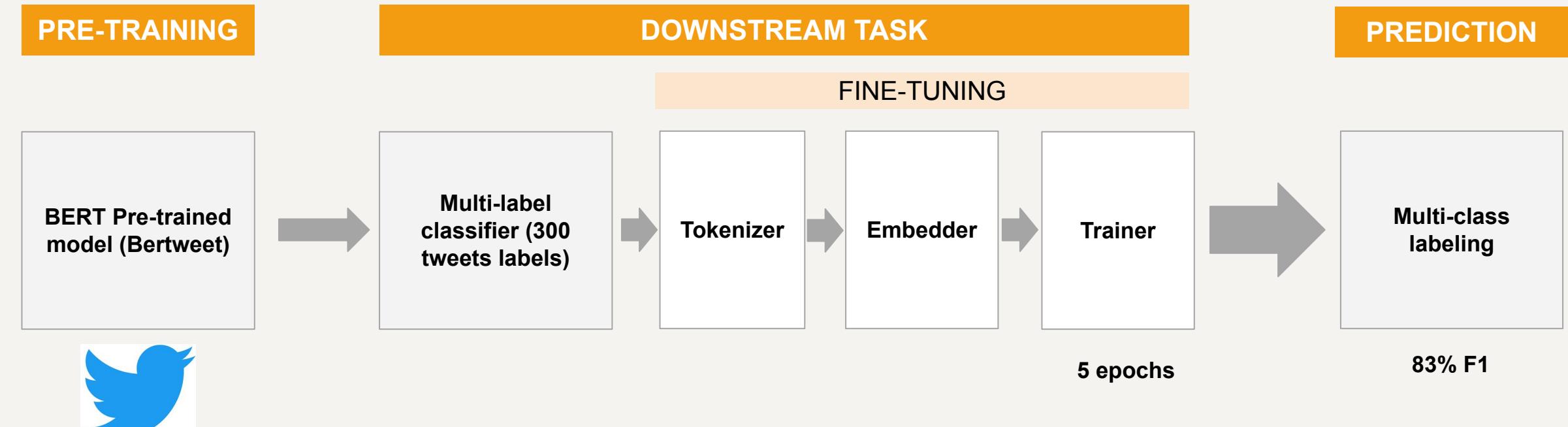
Ready to use products

Focus on business questions

Investigate the reason of the errors



Bertweet classifier



873M English Tweets (cased)



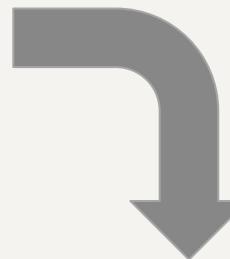
Data extraction from tweet text

spaCy

Every Earthquake ✅
@everyEarthquake

USGS reports a M2.4 earthquake, 39 km W of
Mentone, Texas on 1/17/23 @ 4:49:57 UTC
earthquake.usgs.gov/earthquakes/ev...
#earthquake

10:58 AM · Jan 17, 2023 from Texas, USA · 162 Views



USGS ORG reports a M2.4 CARDINAL earthquake, 39 km QUANTITY W of Mentone GPE , Texas GPE on 1/17/23 DATE @
4:49:57 UTC TIME <https://t.co/HA6XJ8XzF6> #earthquake

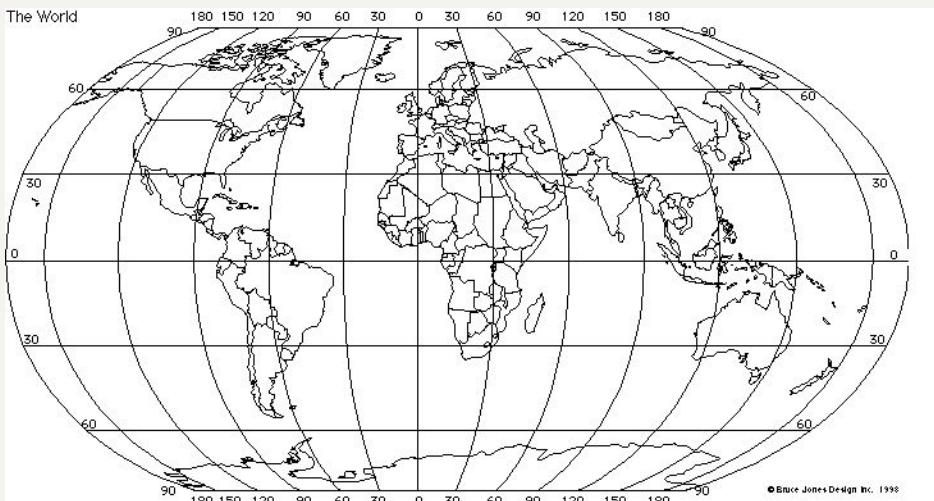


Data extraction from tweet text

spaCy

USGS ORG reports a M2.4 CARDINAL earthquake, 39 km QUANTITY W of Mentone GPE , Texas GPE on 1/17/23 DATE @
4:49:57 UTC TIME <https://t.co/HA6XJ8XzF6> #earthquake

Geopy



Country	City	State	Latitude	Longitude
United States	Mentone	Texas	31,7067958	-103,598792

Data extraction from links embedded in tweets



Every Earthquake (@everyEarthquake)

Tweeting every earthquake occurrence reported by USGS.

Built and maintained by David Barkman aka @cybler.

Embedded link

<https://t.co/HA6XJ8XzF6> #earthquake

Magnitude

Date & time

Latitude & Longitude

Depth



The screenshot shows the USGS TexNet website for the Earthquake Hazards Program. The main header features the USGS logo and the TexNet logo with a seismogram graphic. Below the header, a red oval highlights the title "M 2.4 - 39 km W of Mentone, Texas" and the timestamp "2023-01-17 04:49:57 (UTC) | 31.638°N 104.010°W | 5.4 km depth". To the left, a sidebar menu includes "Latest Earthquakes" (which is also highlighted with a red oval), "Overview", "Interactive Map", "Regional Information", "Felt Report - Tell Us!", "Technical", "Origin", and "Waveforms". The "Interactive Map" section shows two maps of the area around Mentone, Texas, with a yellow star indicating the event's location. The "Origin" section provides details like "Review Status: REVIEWED", "Magnitude: 2.4 ml", "Depth: 5.4 km", and "Time: 2023-01-17 04:49:57 UTC". The "View Nearby Seismicity" section allows users to set search parameters for time range, search radius, and magnitude range. Other sections include "Regional Information", "Felt Report - Tell Us!", and "Citizen Scientist Contributions".

27/01/2023 BeCode-Brussels, Python Predictions use case: Earthquake detection.

Potential Improvements: Extracting data from embedded links

Many “bots” exist and tweet information about earthquakes.

Data structures might differ.



EMSC (@LastQuake)

Independent Scientific Organization and provider of real-time earthquake info.

Significant earthquakes only on @LastQuake.



AllQuakes - EMSC (@EMSC)

EMSC provides rapid earthquake info.

This account displays all recorded worldwide earthquakes.



Earthquake Monitor (@EQAlerts)

Earthquakes alerts in near-real time worldwide (from M 3.2)



Potential Improvements: Computer vision



Based on a study presented at Stanford

Useful for damage severity

400K tweets - 7K pictures

Lack of pictures

Best performance: images + text

Reference:

Firoj Alam, Ferda Ofli, and Muhammad Imran, CrisisMMD: **Multimodal Twitter Datasets from Natural Disasters**, In Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM), 2018, Stanford, California, USA.



Questions



Data Analyst Team



BI Analyst



Zoé
DELCORPS



Iryna
SAKHANDA

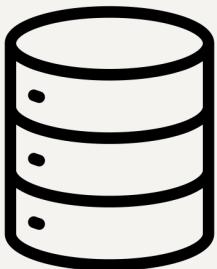


Martin
VELGE



Koumeyl
BELKHIDAR

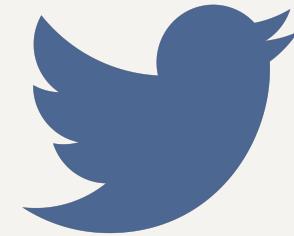
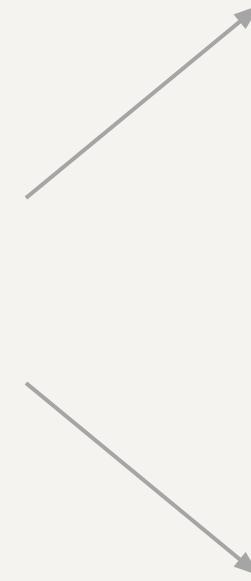
Data Streams



Database



Power BI



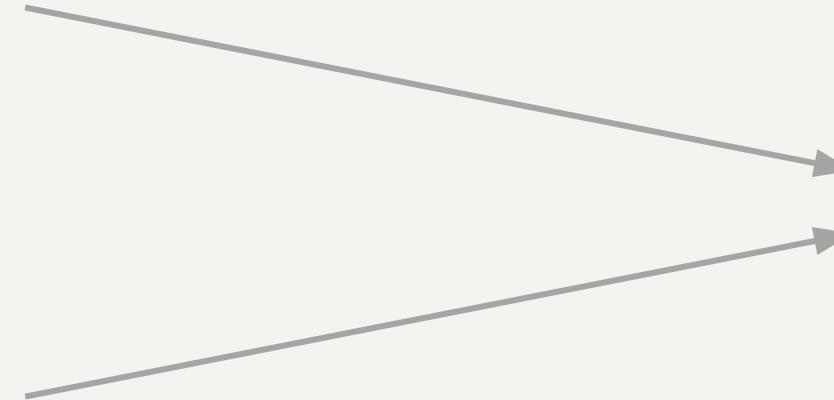
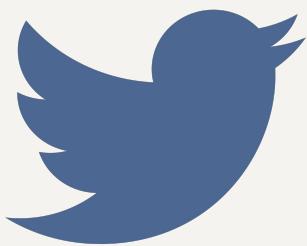
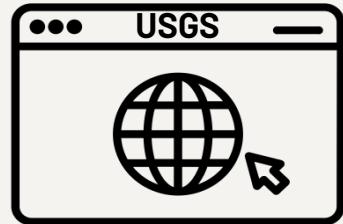
User & Tweets
Analysis



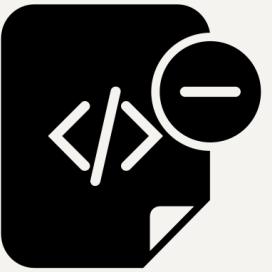
Earthquakes
Visualization



Power BI Connectivity



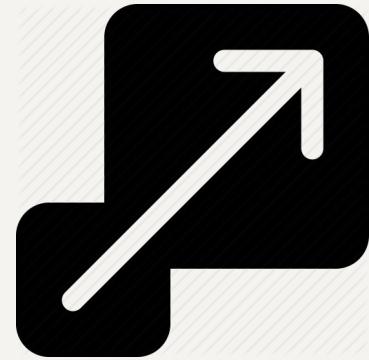
Power BI Abilities



Less coding



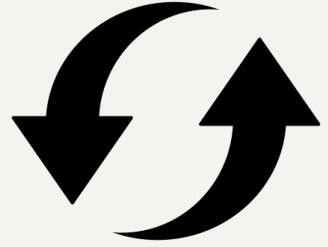
Less time and cost



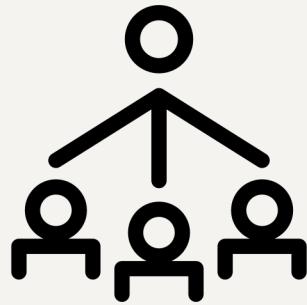
Not scalable



Potential Improvements



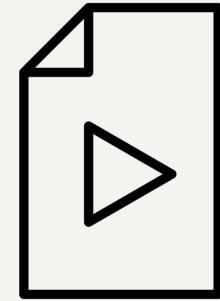
Automatic refresh



More sources:

Official sources (NGO's, Gov, ...)

Other channels (SoMe, Forums, ...)



Animated visuals

THANK YOU!



QUESTIONS?

P Y T H O N
P R E D I C T I O N S
A TOBANIA COMPANY



27/01/2023 BeCode-Brussels, Python Predictions use case: Earthquake detection.

