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

Data Engineer

Data Scientist

BI Analyst

Week 1

Week 2

Week 3

Collect Data

Analyse Data

Process Data

Integrate solution

Deliver project

Transfer

GitHub

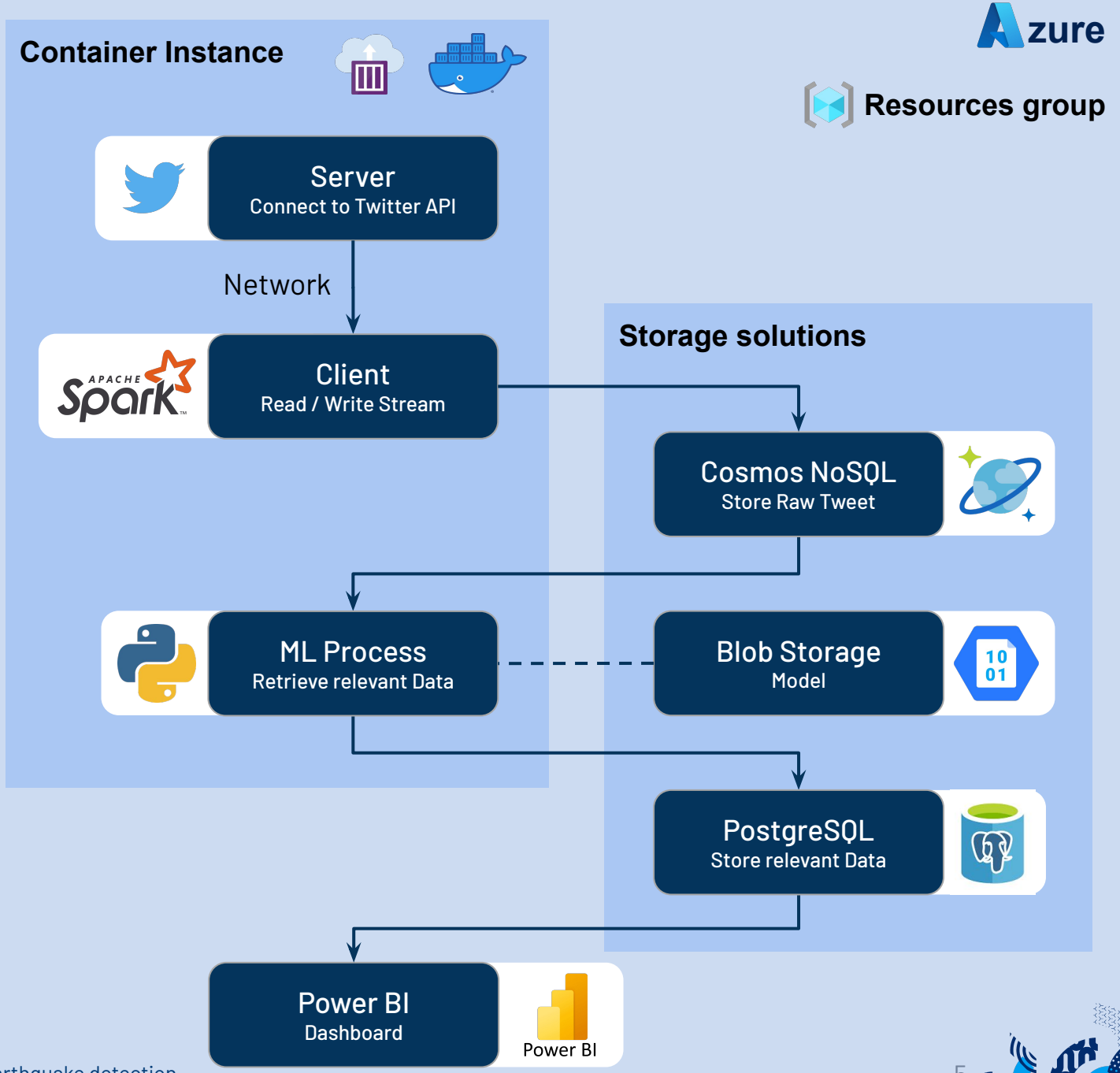
jupyter

Project Pipeline

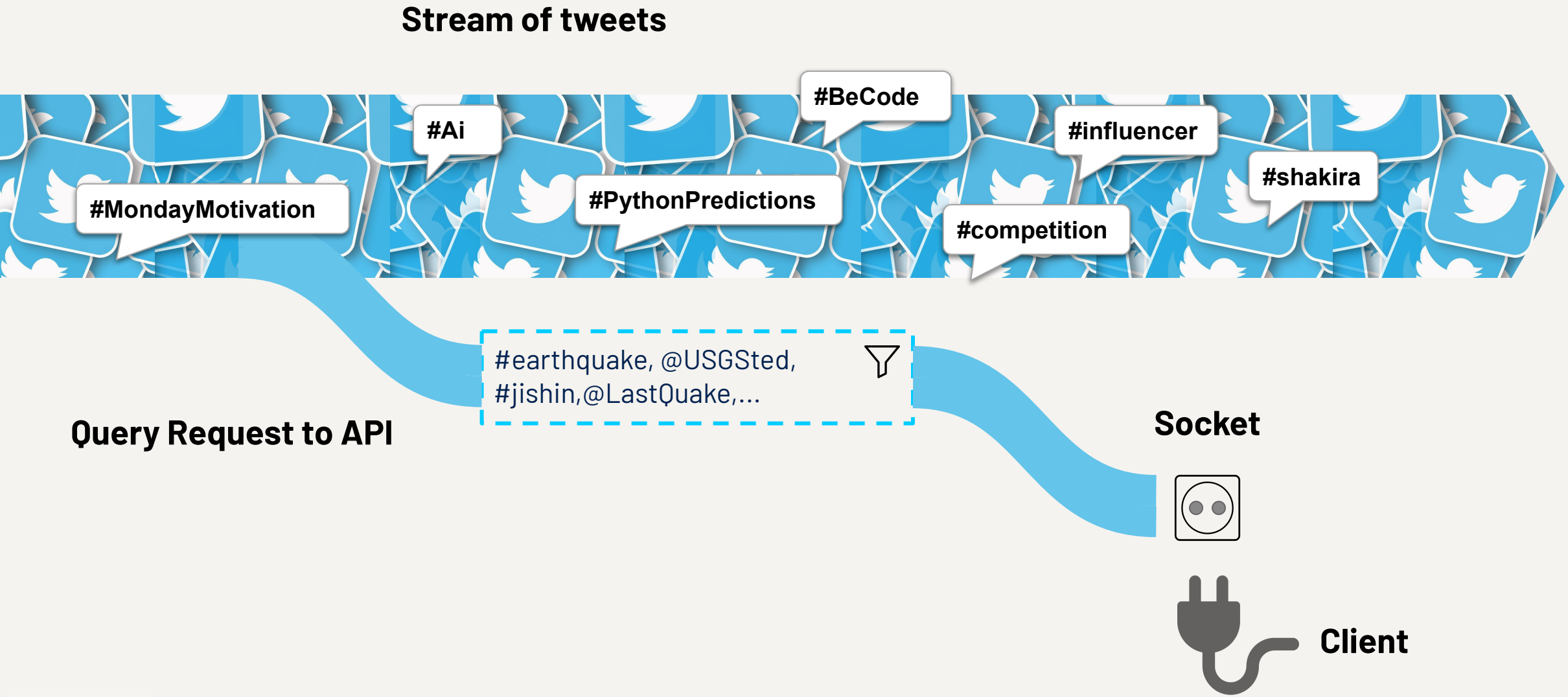
Easily deployed with containers

Easy scaling

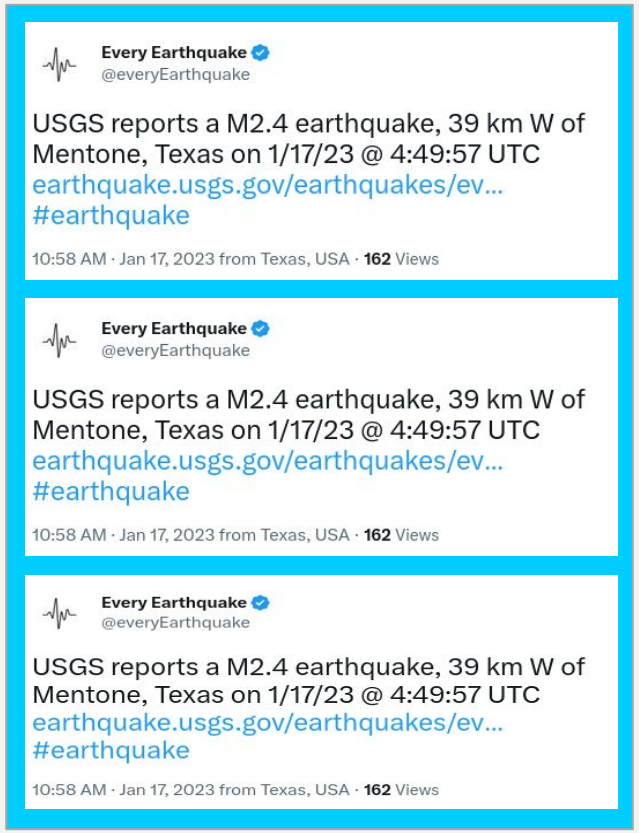
Reproducible environment



Server/ Twitter API (A2)



Client/ Pyspark Data Stream



TWITTER STREAMING API

PySpark

COSMOS DB NoSQL

Storage/ raw tweets

Why Azure Cosmos DB?

ELTL Pipeline use case

Fully managed service

Easy to deploy

Schema-agnostic

Auto-indexing for every item

Scalable


```
{
    "id": "1618160056402587648",
    "tweet_content": {
        "data": {
            "attachments": {},
            "author_id": "913948904730673157",
            "conversation_id": "1618160056402587648",
            "created_at": "2023-01-25T08:13:16.000Z",
            "edit_history_tweet_ids": [
                "1618160056402587648"
            ],
            "entities": {},
            "geo": {},
            "id": "1618160056402587648",
            "lang": "ja",
            "possibly_sensitive": false,
            "public_metrics": {
                "retweet_count": 0,
                "reply_count": 0,
                "like_count": 0,
                "quote_count": 0,
                "impression_count": 0
            },
            "text": "earthquake\n地震"
        },
        "includes": {
            "users": [
                {
                    "created_at": "2017-09-30T02:09:34.000Z",
                    "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。",
                    "id": "913948904730673157",
                    "name": "TOEFL 60点単語bot",
                    "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 thoug...	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

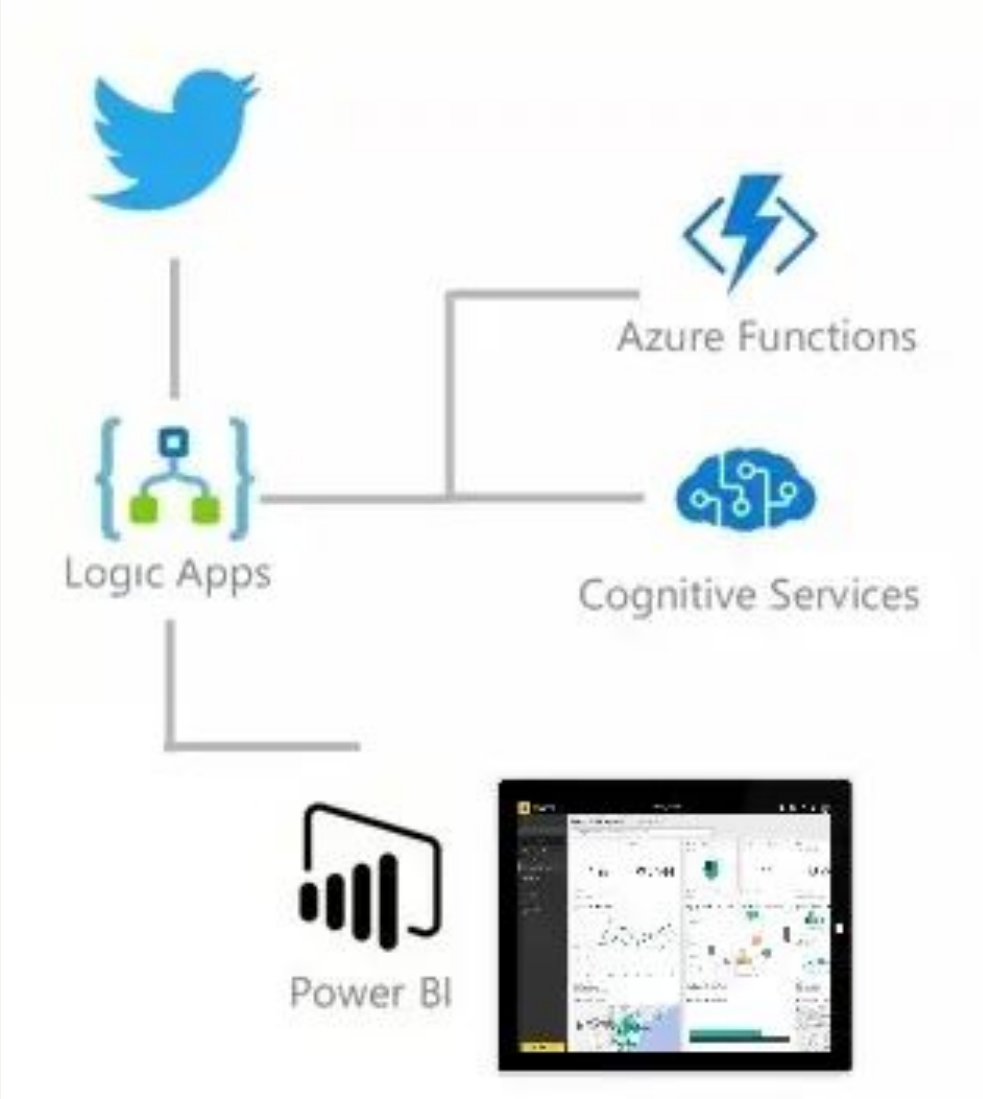
Ability to set different accessibility

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]

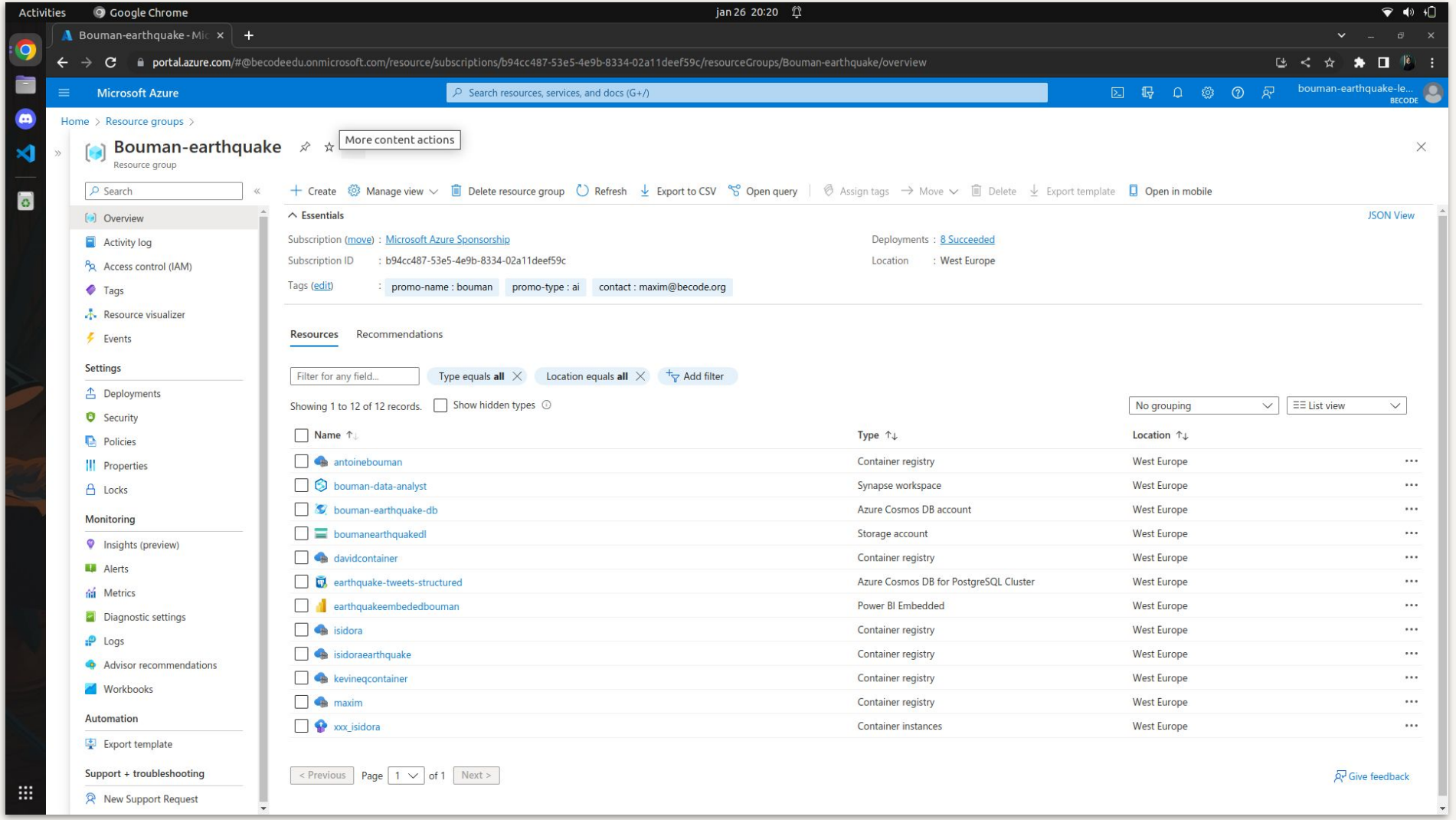
Our solution	
Container Registry	$20
Cosmos NoSQL	$100
PostgreSQL	$150
Blob storage	$20
Maintenance (2 days / month @ $200-400 / day)	$400 - 800
TOTAL (estimate)	**$690 - 1090**

Final thoughts

Challenges	Improvements
Design a good pipeline from day one	Pipeline MVP with fake data ASAP
Work with stream (real time data)	Naming convention guidelines
Deploy and debug on the cloud	Deploy ML code into the cloud
Organization	Improve communication

Questions ?

Go to Demo



Data Scientist team

Data Scientist

Olivier
DE TIMMERMAN

Rafaella
PORTO

Piero
RUCCI

Olivier
LE DIBERDER

Tania
LEMOS

Alexis
VANDRIESSCHE

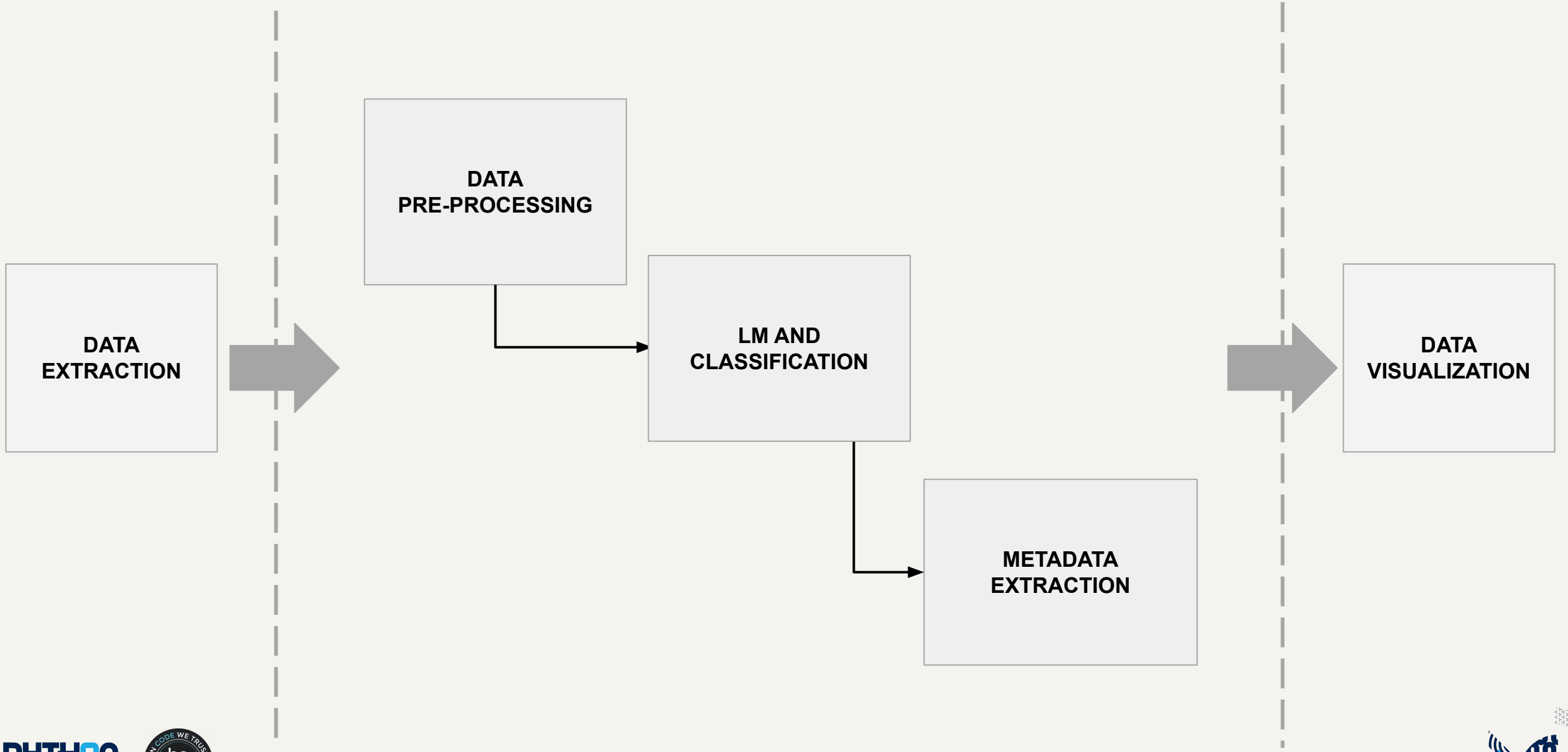
Julien
DE SMET

Esra
OGUZ

Jack
XIN

Carlton
NJUGUNA

Language Model



Language Detection & Translation

deep_translator

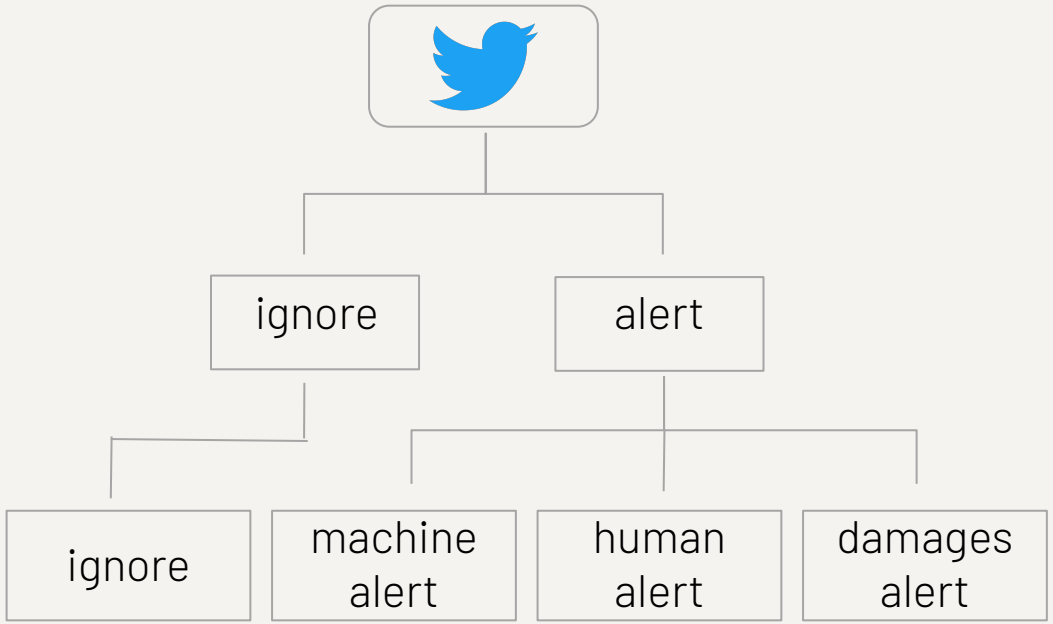
130 words:

Cloud Translation	
Text Translation: 5,200,000 characters	USD 94.00
Language Detection: 5,200,000 characters	USD 94.00
USD 188.00	

280 words:

Cloud Translation	
Text Translation: 11,200,000 characters	USD 214.00
Language Detection: 11,200,000 characters	USD 214.00
USD 428.00	

Twitter Classifier Model

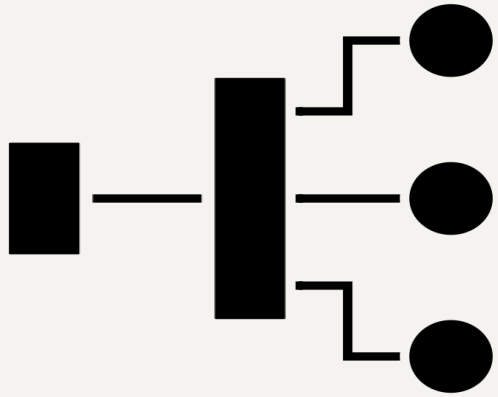


Difficulty to manually classify tweets to train the model

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**

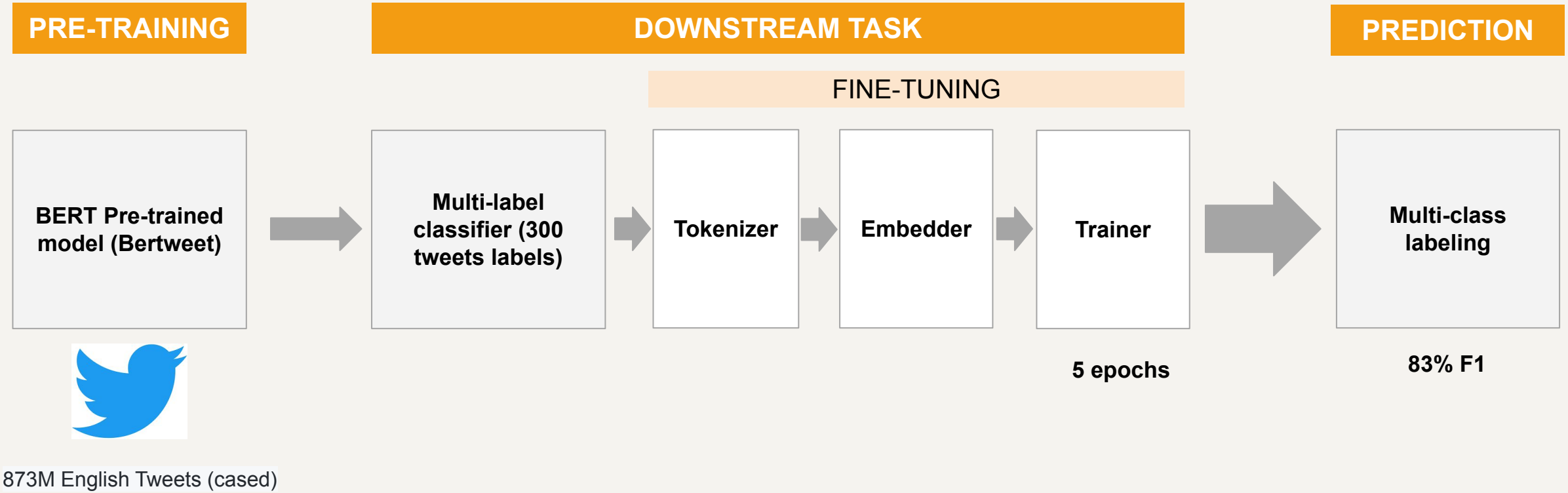
Drawbacks:

- 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



Data extraction from tweet text

spaCy

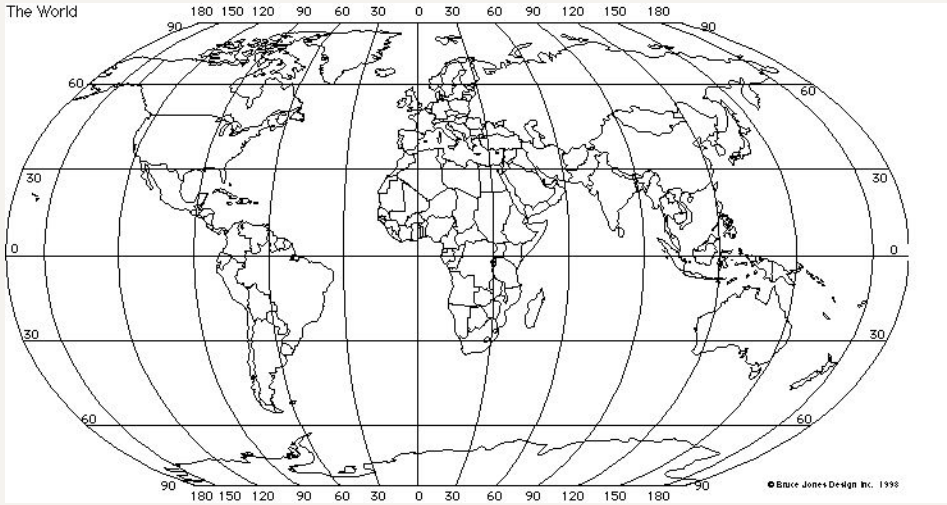
USGS ORG reports a M2.4 CARDINAL earthquake, 39 km QUANTITY W of Mentone GPE , Texas GPE on 1/17/23 DATE @ 4:49:57 UTC TIME https://t.co/HA6XJ8XzF6 #earthquake

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.

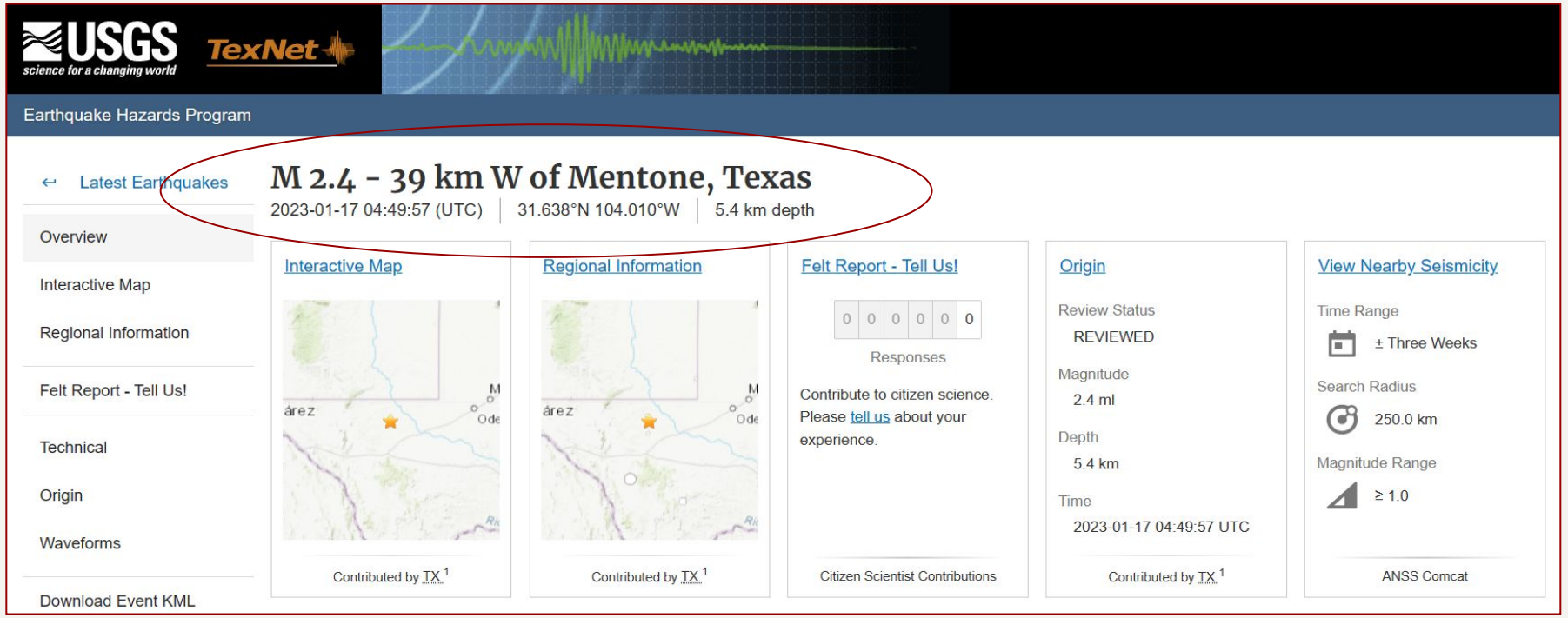
Embedded link https://t.co/HA6XJ8XzF6 #earthquake

Magnitude

Date & time

Latitude & Longitude

Depth



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

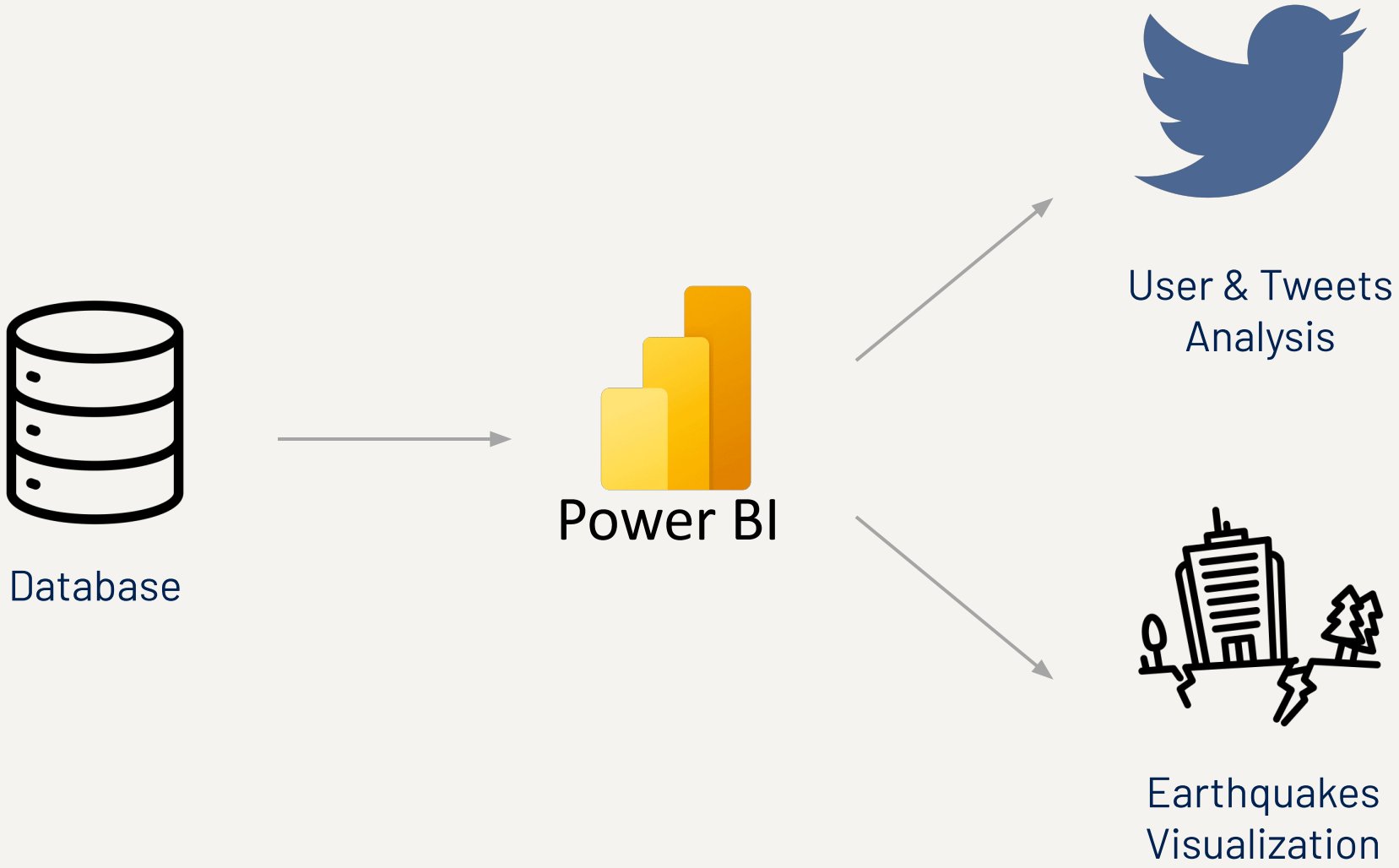
Zoé
DELCORPS

Iryna
SAKHANDA

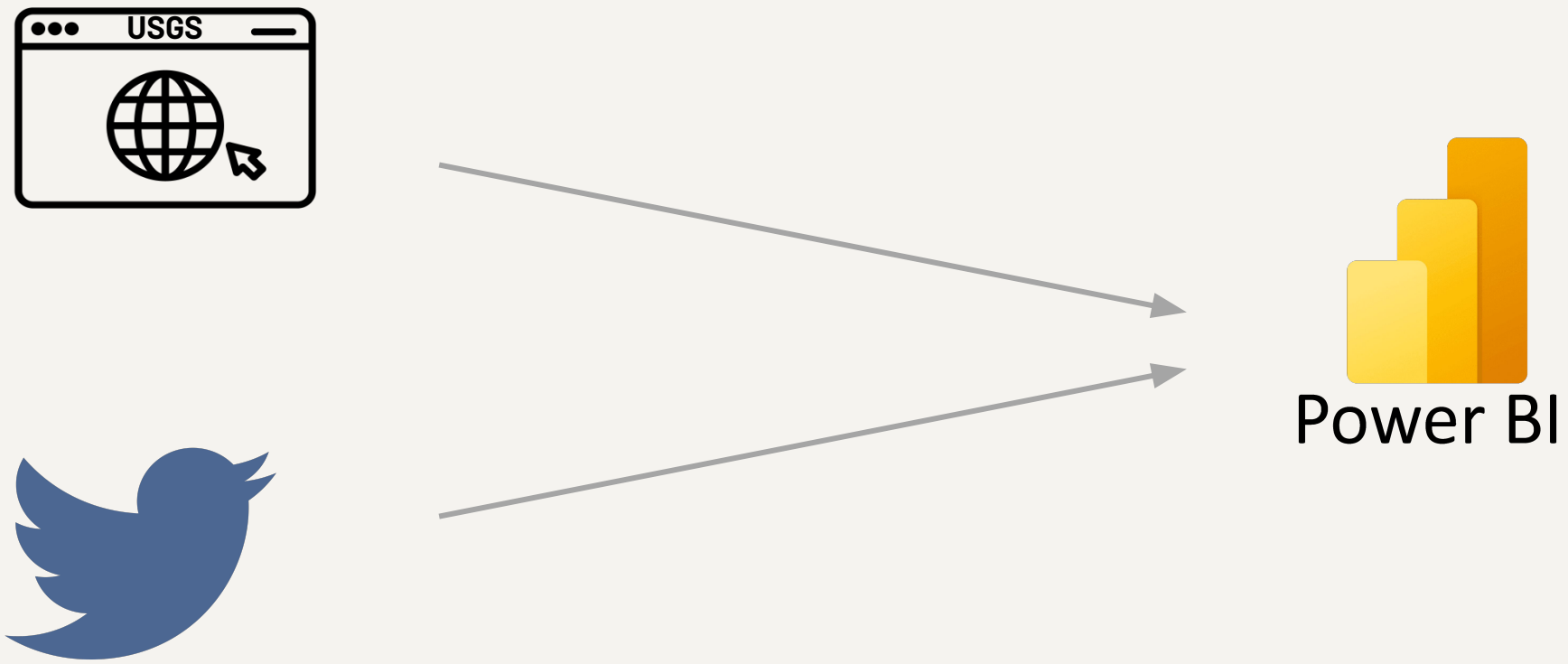
Martin
VELGE

Koumeyl
BELKHIDAR

Data Streams



Power BI Connectivity



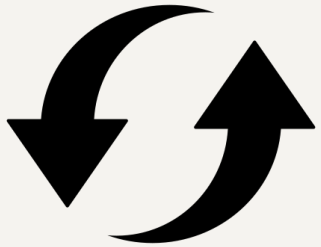
Power BI Abilities

Less coding

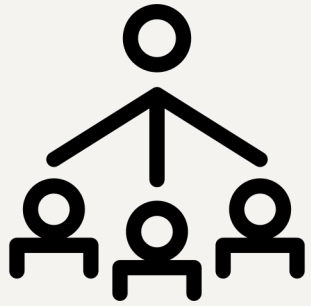
Less time and cost

Not scalable

Potential Improvements



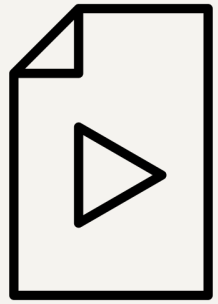
Automatic refresh



More sources:

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

Other channels (SoMe, Forums, ...)



Animated visuals

THANK YOU!

PYTHON PREDICTIONS
A TOBANIA COMPANY

QUESTIONS?