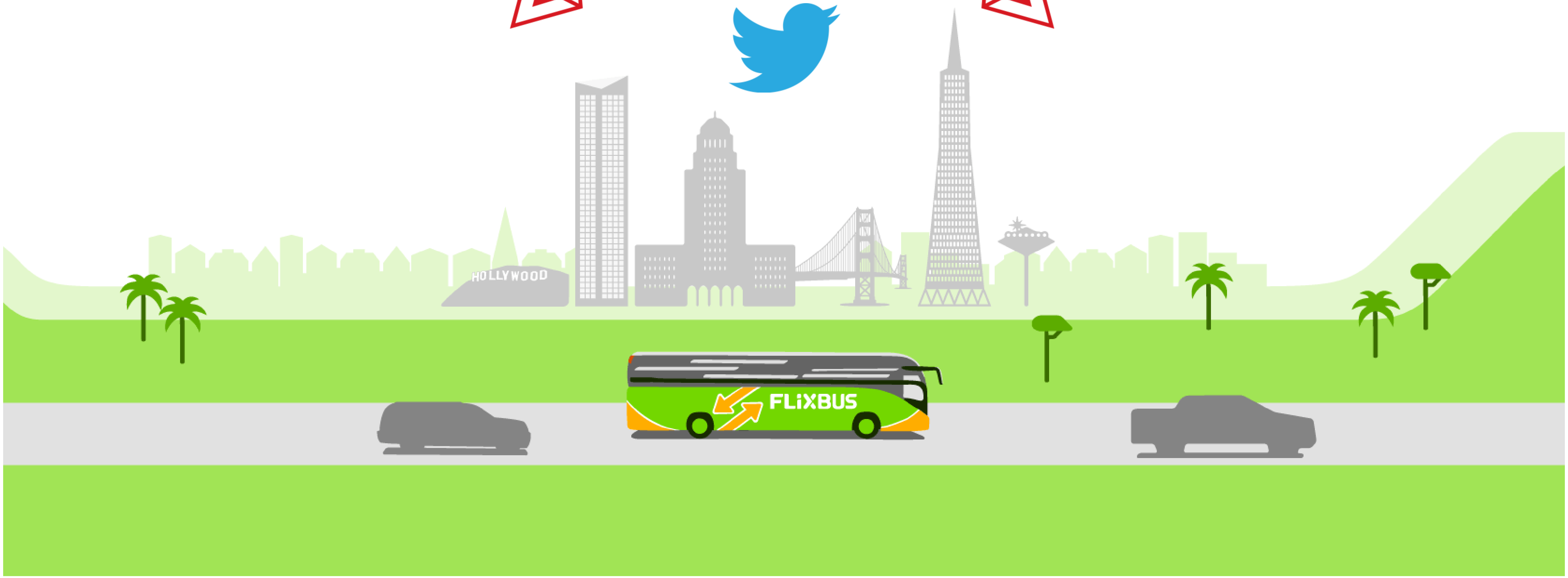
FLIX

THE CHALLENGE



FLIXBUS

01 The Challenge
Business requirements to this task

02 Methodology
Thinking and planning before any action

03 Development
All steps to achieve the goal

04 Conclusion
Sharing some ideas of improvement

1. Search for tweets mentioning #FlixBus (last week)
2. Store the raw data
3. From the raw data extract some attributes
4. Apply anonymization on User id
5. Prevent some pipeline issues
6. Feature of reprocessing (day, range)

Methodology

- Gathering requirements
- Tech Stack definition
- POC
- Development
- Documentation

Development

tweepy™

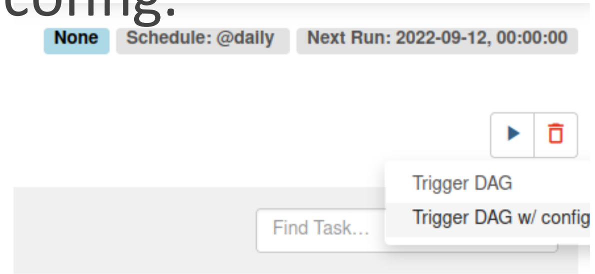
Apache Airflow

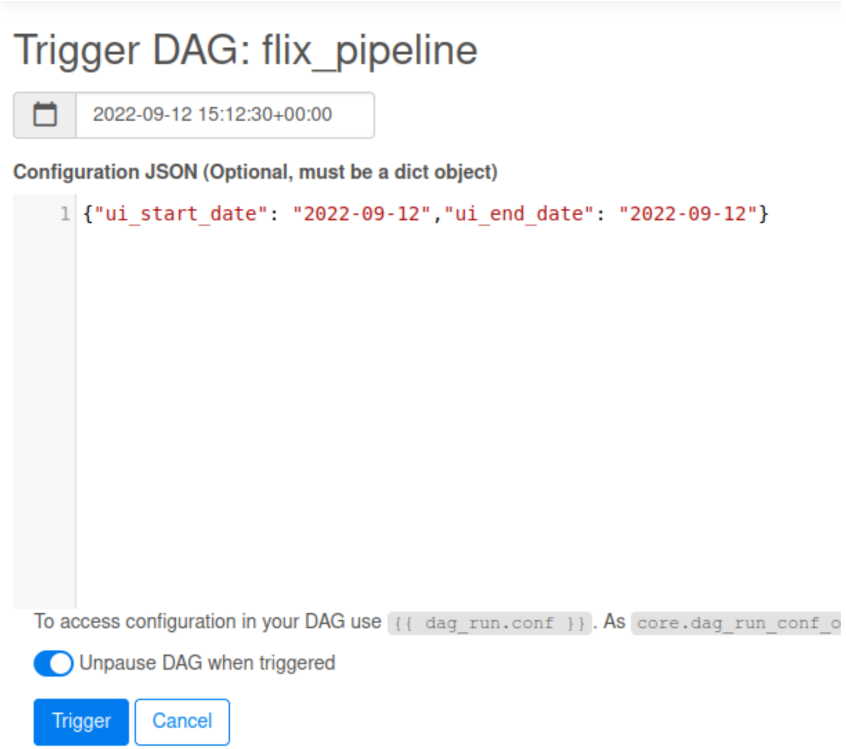
MINIO

Development

Reprocessing using single date or range

From UI click on arrow signal and choose the option with config.





Trigger DAG: flix_pipeline

2022-09-12 15:12:30+00:00

Configuration JSON (Optional, must be a dict object)

```
{"ui_start_date": "2022-09-12","ui_end_date": "2022-09-12"}
```

To access configuration in your DAG use `{{ dag_run.conf }}`. As `core.dag_run_conf_o`

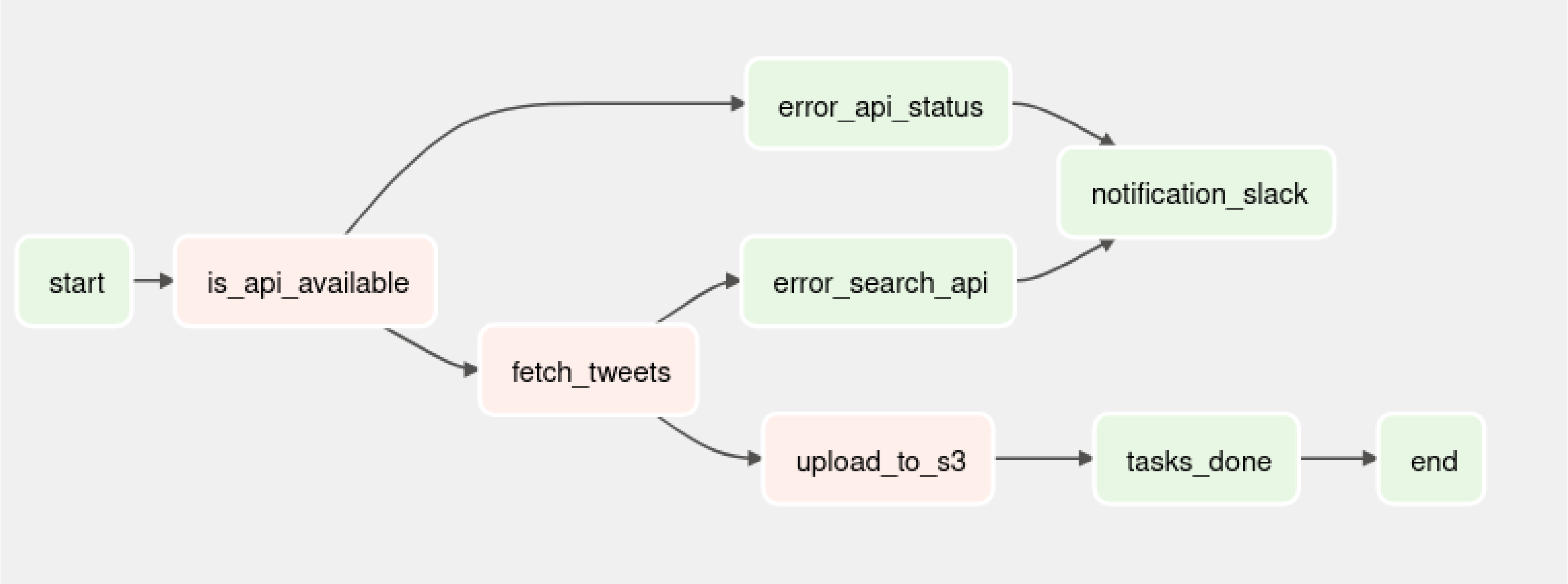
Unpause DAG when triggered

Fill the text box with both dates to ask data from a range or set only the ui_start_date attribute that it will be considered as a single day

FLIX

Development

Pipeline



1 Check if API is available

2 Fetch tweets from API

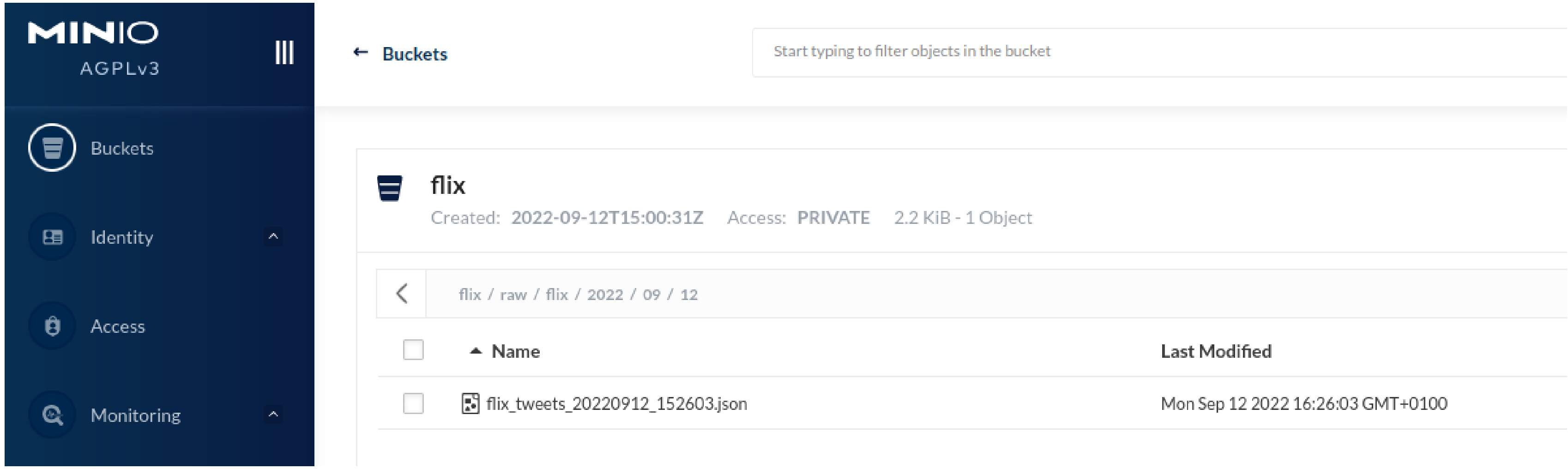
3 Upload data to S3

In case of error to check the API status or not found any tweet the task is going to fail and a notification should be send to Slack.

NOTE: Communication with SLACK is not implemented, it was added only as example to handle issues.

Data stored with attributes

After retrieve the data and extract what is required (also anonymization) the data has been write out to MinIO (S3) and is available on URL below http://localhost:9001



flix

Created: 2022-09-12T15:00:31Z Access: PRIVATE 2.2 KiB - 1 Object

flix / raw / flix / 2022 / 09 / 12

□	▲ Name	Last Modified
□	flix_tweets_20220912_152603.json	Mon Sep 12 2022 16:26:03 GMT+0100

Conclusion

All requirements were achieved but also at the end some possibilities of improvement were detected.

- Store data in a different format as Parquet
- Implement notification to Slack or other tool.
- Improve context of reprocessing to be dynamic
- Start to develop it to Tweeter V2 api version

THANK YOU!