# ANZ Code challenge

Create a workflow to produce / consume messages and store them. Schedule a daily job to generate reports to publish data usage and time spent

Disclaimer : Prior to this challenge, I have no knowledge about Kafka, so it took a considerable time for learning, configuring and experimenting Kafka.

Github repo : <https://github.com/Sangeeth-fb/ANZ>

Programming language : Python

Other stacks : Apache airflow, Kafka, confluent, Mysql

## Highlevel architecture

A close up of text on a white background

Description automatically generated

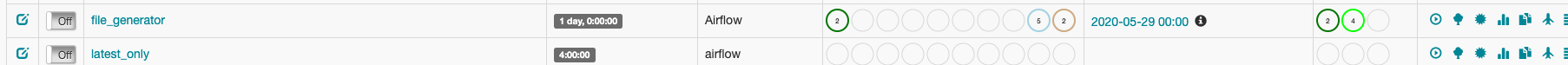
Components

|  |  |  |
| --- | --- | --- |
| Component | File | Details |
| Kafka Producer | **Log\_producer.py** | **Reads from mobile\_log.json , serializes the messages and sends it to kafka broker as “MobileLog” topic**  **Used python to generate the json serializer and sends the message to Kafka broker using Kafka-python libraries**  **The files can be either run standalone or scheduled in the airflow for demo purpose**  **Sample command to run or test – airflow run log\_producer generate\_stream –local 2020-05-29** |
| Kafka Consumer | **Log\_consumer.py** | **Deployed by running the python library which listens to the Kafka server. When a new set of messages are available in the server, the messages are consumed and send to Mysql if the messages are not corrupted**  **File can be run standalone** |
| Report generator | **Report\_generator.py** | **This library is scheduled to run at 12:00 am daily using Apache airflow.**  **The python operator will trigger the mysql commands to calculate the download / upload and time spent per url per subscriber**  **The scheduler is fault tolerant. If a pipeline fails it will retry as per settings and if it exhausts, it will send alert to defined mail.**  **The report generated will be in CSV which is timestamped for the day it generates report.**  **It can be also be triggered from apache web, cli ot standalone using python library .**  **Sample given below** |
| Mysql | **ANZDB – Database**  **MobileLog – Table**  **Schema :**  CREATE TABLE MobileLog  (mobileNum VARCHAR(20), url VARCHAR(100), SessionStartTime VARCHAR(30) , SessionEndTime VARCHAR(30), bytesIn BIGINT, bytesOut BIGINT, timestamp BIGINT, ds timestamp); | **MobileLog is partitions on datestamp. Reports will be generated for the day file is executed for. It can be back run for any day the report is needed by passing day as param and it will regenerate the reports for the day. If only one of the report is needed, the following command can be run**  **Airflow backfill report\_generator consumed\_data -s start\_Date -e end\_date**  **Please note the backfill is handy if you need to generate report for multiple days at once.** |

Apache scheduler

A screenshot of a cell phone

Description automatically generated



Deployment strategy

Workflow can be run as per commands specified in the component section. Due to time constraints, I have not included a Docker image, but I can add if that helps.

Assumptions and deviations

I am storing the data in the Mysql instead of a Datalake because of following reasons

* I could connect the Kafka in local machine to Google cloud using the connecter but as the google is billing around $674, I removed the process from workflow, Happy to explain the steps end to end
* I tried connecting the Kafka to MYSQL directly using JDBCSinkconnetor. It connects but brings down the whole cluster, mostly be due to community edition. This could have helped in avoiding the consumer library to store the messages in MySQL to make it better data transfer system. I very well understand the concept