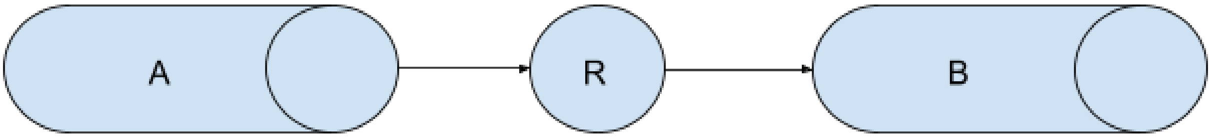
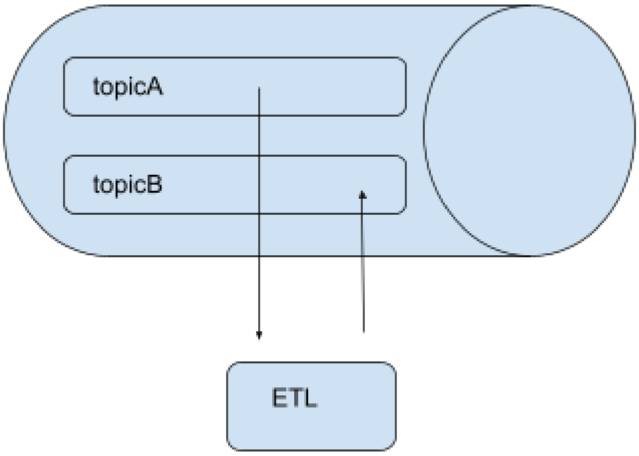
Data Engineering assignment

This assignment is about a real situation that occurred in our production environment.
For a certain period of time there has been an interruption in the replication mechanism R, responsible for propagating data from the Kafka cluster A to cluster B. This accident has caused a gap of data in cluster B.



Your assignment, in a simplified version, is to fill the gap by consuming data from a Kafka topicA, filter them and produce on topicB (topicA and topicB are in the same Kafka broker/cluster)



We have provided you with 30k records of user interaction data of 14th April 2021, from approximately 10:00 till 16:45 UTC.

The assignment should be structured in the following way:

1) load the data set in a Kafka broker/cluster in topicA, preferably using a container

2) consume from topicA and filter the records from 11:00 to 13:00 UTC

3) produce the data from point 2) in topicB

4) could you describe the potential issues in re-ingesting all the data from cluster A? Could you mention in which other ways you would fill the gap in cluster B?

Bonus Point: create a pipeline in the container such that both topics can be propagated to a data store and visualized (for example using the Logstash-Elasticsearch-Kibana stack).

Input data

Files:

- data/action_data.json.gz
- data/data_book.pdf

Delivery format

The preferred language is Python and the code base should be organized in modules and contain some unit tests.

For the sake of reproducibility you are encouraged to use and share:

- a virtual environment
- Dockerfile(s)

Together with the codebase we expect to receive a document where you communicate your approach and results.

The content of your project can be shared in:

- a compressed archive (without data set) or
- you can set up a private Git repository (pls do not upload the data set) to share with some of the XITE data guys

You will be evaluated on the following aspects:

- quality of your code,
- its reproducibility on another machine,
- the way you tackle the problem,
- how you communicate the results.

Have fun and good luck!