Assignment Data Engineer IKEA

Your submission

You will need to answer questions and analyze data. This will involve writing some code, our data engineering practice is mostly python but choose what you feel most comfortable with. Please ensure that

(a) your logic and pipeline(s) are clear without us executing any code or setting up a development environment.
(b) your code is well commented so we can follow your logic
(c) you can choose to submit your project as a private git repository (preferred) or send the compressed files over email.

The case:

As a data engineer in IKEA, one of the responsibilities is to provide Data Scientists with data. We have many machine learning models in different domains and the Data Engineers are the backbone of making this possible.

For a new product, we want to understand what the conversation is online about IKEA and what we can learn from this. We ask you to create a dataset that collects this conversation about IKEA using the Twitter API. The easiest is to filter on Tweets containing keyword “IKEA” but feel free to be creative! You can choose if you want the pipeline in batch or streaming. For this project, we would at a minimum require id, text and created_at from the Tweet. However, any additional data that you think is useful would be great to add to the data model. You can choose any form of storing the data but please keep the use case in mind.

We want the Data Scientists to be able to run different projects on this data so make sure to keep it flexible. We keep the assignment relatively flexible so you can use your preferred tooling and we hope it makes the assignment more fun. However, we do expect you to be able to explain your choices. You are free to choose between local or cloud implementations.

Minimal requirements:

- Working solution.
- Simple design diagram of the solution. Bonus points if you created one for batch and one for streaming.

Preferred requirements:

- A working (simplified) ci/cd to automate deployment.
- Applications can be run in a container.
- Data stored in a database that can be queried (the database can be run in a container too).
- Show that you understand testing.

Some additional ideas to extend the project:

- Build an advanced data model and be able to reason about the structure

- Store additional metadata provided by the Twitter API (such as topics)
- Sketch a more advanced set up with batch/streaming combination and explain the database choices that come with it.
- Extract features such as emoticons and store these separately.