Technical Screening – Data Engineering

Thank you for your interest in IKEA!

With this exercise, we would like to probe your technical experience for a role as Data Engineer.

Your submission
You will need to answer questions and analyze data. This will involve writing some code, preferably in Python but choose what you feel most comfortable with. You are welcome to use a Big Data framework, if you like (*e.g.* Spark, Apache Beam).

Please submit your answers to questions, code and your output data. Please ensure that

(a) your logic and pipeline is clear without us executing any code or setting up a development environment. If you submit a notebook, please convert *e.g.* to PDF for submission.
(b) your code is well commented so we can follow your logic, and
(c) you submit your output data as a CSV file.

The following exercise will be structured around a hypothetical, but relevant case for IKEA

The case:
As a Data Engineer at IKEA you contribute to improving the everyday life of the many people by enabling co-workers to work with better data.

IKEA wants to create a fun and meaningful experience for families visiting our stores. Imagine that IKEA has upgraded our Småland facilities (the play area at the stores) and we want to let customers with small children know, but we do not want to bother customers who are not likely to have children.

You are working closely with a team of Data Analysts who want to predict which customers are likely to have children, as the available data is incomplete and only some customers have indicated if they have children.

You are asked to support the analysts with preparing the data, which is transaction line data. See attached sample.

IKEA leverages GCP, Azure, AWS, etc. When answering the questions, you can assume that the solution will need to run on one of these platforms. You can also assume that the actual data is available as both a daily batch and a near real-time stream.

There are 5 required questions and 4 optional questions
Where your answer includes code, please include comments to explain your decisions.

Question 1:
According to you, what are the main building blocks of a data pipeline?

Question 2:
Write a function to return an iterator which can be piped to transformation functions. The iterator should accept a parser as an argument and create a basic parsed output. Please do *not* use the pandas python package.

Question 3:
In the attached data, an age column contains age buckets. How would you transform this feature to make it more suitable for modelling? What different options would you consider? Pick one to implement and use the iterator from Question 2 to apply the transformation.

Question 4:
To enable modelling for the data analyst working on the prediction, please transform the dataset into a structure where each record in the dataset corresponds to a unique customer. What schema design do you choose and why?

Please remember to submit your transformed data as part of your submission to this exercise.

Question 5:
Suppose you are given the task to build the data pipeline for this analytical use case. You can assume that the customer data is coming from a relational database and the transactions are coming from a Pub/Sub stream.
What are the important design criteria for the data pipeline?

How would you design and deploy it?

[Optional] Question 6:
What else would you do to the dataset? Would you think of other transformations to apply to support analysis? You don't have to implement your suggestions, just share your thinking.

[Optional] Question 7:
Is there something we did not ask you, that you think is important to do or consider for this data?

[Optional] Question 8:
Please share any feedback you might have on this test.