Data Engineer (Direct-to-Consumer) Challenge

This challenge consists of 2 separate problems that we would like you to consider. The first problem is more practical and requires some coding, whereas the second is a theoretical question.

To successfully complete this assignment please provide us with:

- The SQL queries you developed and the corresponding result set of those queries for the first challenge
- A draft description for the second challenge

Feel free to use any medium you feel comfortable with to share the code snippets and your answers to the additional questions (pull request, email, Docs, Notebooks, pptx, etc.).

You should not have to spend more than four hours on this assignment.
You can find the provided data in the data folder of this repository.

Challenge 1

Your were approached by the digital marketing team and one of our data analysts to help them creating a data table to make them able to answer their business questions and to unveil certain patterns. The team is keen on figuring out how our products and channels are performing to assess and compare trends. Their questions include ones like

- how the company performs on a daily basis, i.e. how many products we sold daily
- which day of the week performs the best (name of the day)
- how our products perform in terms of item model names
- how our item categories perform overall
- how our marketing channels perform in terms of sales
- how social media platforms perform in terms of conversions
- how our products perform in terms of gender split

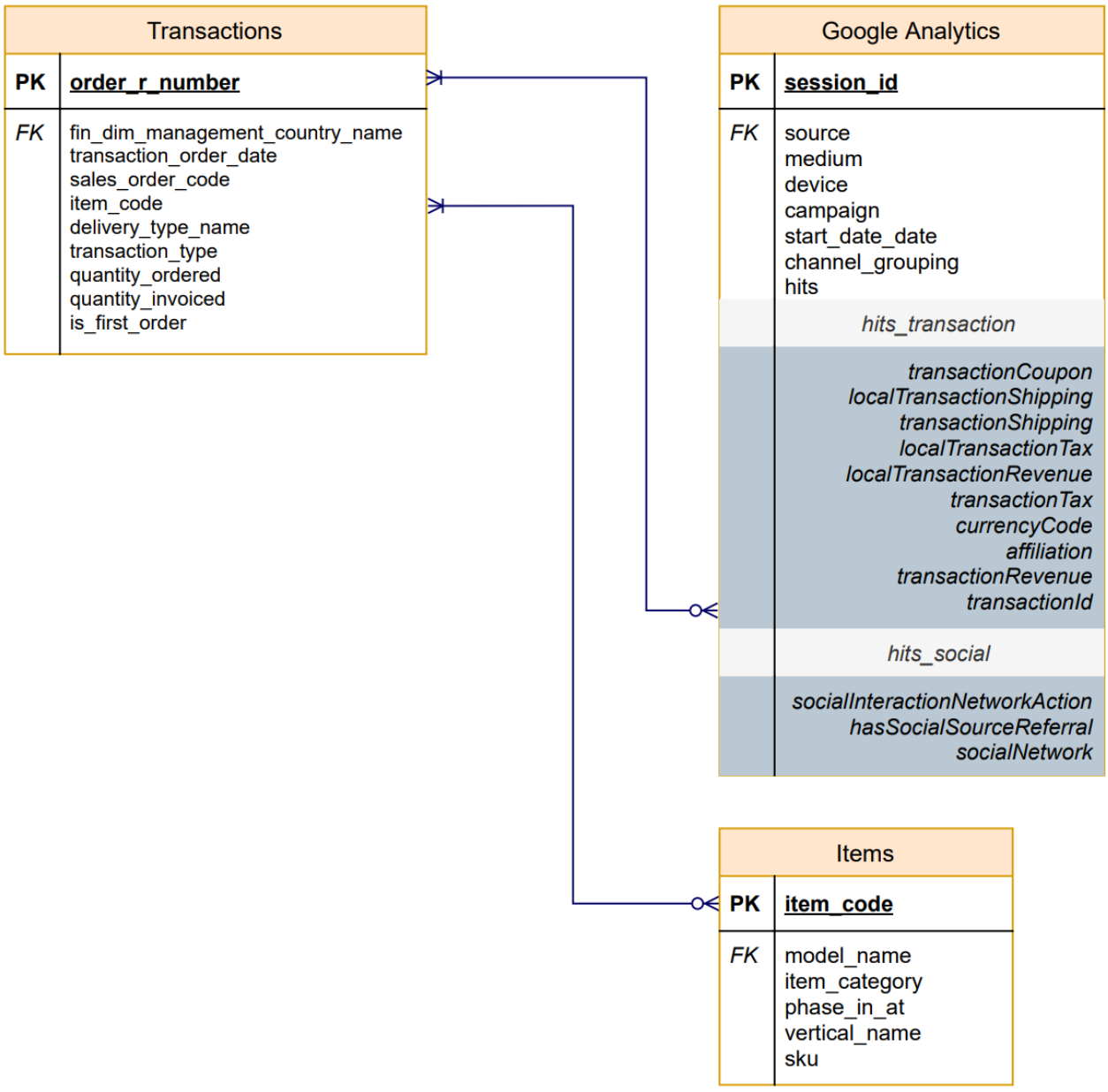
- how many items are included in an order
- how many orders were first ones for the respective customers.

It is the responsibility of the data analyst to analyse the data and find the exact answers to these questions. However, she needs an input data table to do so.

It is your task to come up with a unified, ready-to-use data table that contains all the required data to make the analyst be able to answer the listed questions. It is on you to determine the types of feature engineering you need to conduct and define the fields to make the data available for answering the questions.

You have a

- **transactions dataset** (a sample of it) with fields associated to a purchase happened on our web shop,
- **items dataset** with some of our SKUs currently offered and
- **some fields from Google Analytics** tracking data for the corresponding time interval.



The tables organised as follows. The transaction table (csv) has a unique ID per transaction called the order_r_number. Each transaction has 2 records by default; a "Sales Order" transaction_type marking an order happened on the website, and a "Sales Invoice" type referring to an actually delivered and invoiced order. As the business partner is only curious about fully completed orders (i.e. delivered and invoiced), you need to filter the data accordingly. In case a customer ordered more items in one transaction, each item appears in a new row under the same order_r_number ID. The item_code field refers to the ordered products and is a key for the items table.

The item table (csv) includes product-related data using the item_code as a primary ID like the name of the product (model_name) or whether the product is shoes or apparel (item_category). It also includes an SKU field containing item model name, gender and item colour in a string.

The google analytics table (json) is a nested table containing data on user behaviour on the website like the device on which the user visited the site (device), the marketing campaign the user clicked through the website (campaign) or the marketing channel the user arrived like email or paid display (channel_grouping). Inside the double nested fields of hits_transaction, the transactionId field refers to the transaction ID (order_r_number) of a purchase. The double nested fields of hits_social contains some information about the social media platforms users arrived to the site.

To answer the above stated questions, please, make sure you show that you are able to use different SQL functions like grouping, sorting, ordering, splitting strings, using regex, applying window functions and date functions where there is an opportunity for that.

Challenge 2

In this challenge, we are just curious how you approach things. You do not need to code anything or prepare charts, dashboards, metrics, features etc. We would like to ask you to answer briefly with a couple of sentences or bullet points.

- In regards to the previous challenge, please, also address the following question. If you would need to come up with some charts, a dashboard or an overall metric to measure, how would you aggregate the data you have created for the analyst?
- Looking at the provided Google Analytics tracking data in a deeper manner, you have realised that it is an event-style data recording past occurrences that are not prone to variation. What would be the best data table materialisation strategy in this case and why?

If you have any questions in terms of the data, or the challenges, please, feel free to contact adam.madacsi@on-running.com.