Data analyst exercises - 3 - 4 hours

Scripting

Background

Imagine you've been asked to automate part of the work related to monthly settlements managed by the sales team. The team has a list of advertising campaigns and their budgets expressed in EUR. They want to enrich the list by adding a column that contains the budget converted to the local currency. The output should be delivered as a *csv* file.

Exercise

Solve the problem by writing a simple **Python** program. **Current exchange rates should be downloaded from the web API** (https://www.exchangerate-api.com). It doesn't require a key (https://www.exchangerate-api.com/docs/free) but if you prefer to use a different API feel free to do so. Generate the dataset using a provided input file.

List of deliverables:

- Source code
- A README containing instructions so that we know how to build/run your code
- Generated csv file containing requested dataset

PySpark

Exercise

Your goal is to answer the following questions using **PySpark**. You can find the data necessary for this exercise in the file called ***inventory.parquet***. You should read it as a Spark DataFrame and perform all operations without changing the data structure (e.g. converting to Pandas DataFrame).

Data description

- user_id - the unique identifier of each user
- is_logged_in - represents if a user is logged in during an event
- device_type - represents a device type used during an event
- event - either view or click on the ad
- is_mobile_app - represents if a user was exposed to the ad in the mobile app
- site - the site where the user was exposed to the ad

- order_id - the additional value passed during an event
- date - the date of the event in the format “yyyy-MM-dd”

Tasks:

1. Find what is the percentage of logged in users every day.
2. Which site has the most logged in users?
3. Calculate the share of logged in users who are using Mobile App.
4. Create a new column called **identity_type** which will take the following value:
    - If **device_type** is “Mobile Phone” and **is_mobile_app** is set to True then “Mobile Phone App”
    - If **device_type** is “Mobile Phone” and is_mobile_app is set to False then “Mobile Phone Web”
    - If **device_type** is “Desktop” then “Desktop”
    - Otherwise “Unknown”
5. Create a new column in the dataset called **max_order_id** which will show the maximum **order_id** for each **identity_type**. The DataFrame must persist the original number of records.
6. You have been notified by the Marketing team that they would like to know what was the number of clicks (**event** column equals to “click”) each day for a given campaign. They sent you the list of users taking part in this campaign (***selected_users.parquet***). Your goal is to filter the dataset to include only selected users and calculate the total number of clicks per day.
7. What was the number of clicks per day for users who weren’t in this campaign?

List of deliverables:

- Source code