back market

HOME - BASED MISSION

'Your mission should you choose to accept it', is:

I. First: to pass the SQL basic knowledge test smoothly:

- *- Lets's consider the two following tables :*
    - ***Table USERS**: Contains general informations about users: USER_ID, USER_EMAIL, DATE_CREATION_PROFILE, DATE_LAST_VISIT, NUMBER_OF_VISITS,*
    - ***Table SALES**: Contains informations about every sales transactions on the web site: TRANSACTION_ID, PRODUCT_ID, USER_ID, MERCHANT_ID,DATE_SALES, PRICE*

- *Please prepare a sql request to compute the volume of sales (that is, sum of prices) per merchant per month (using mysql date functions only)*

- *Using a single sql request, extract the 10% of users making the greatest volume of transactions every month.*

- *Let's consider that a metric to measure the attractiveness of a product is the fact that one merchant only is selling it, and that he sells more than 10 products per day. Please, write a sql request to fetch the 'attractive' products of the last 3 months.*

II. Second: to write an ETL application in Python 3 to analyse some of the SAV related problems. This application will then be scheduled using Airflow, and run every day in order to provide two outputs: a '.csv' file and API able to send the data through POST request in asynchronous manner by 10 entries at the time.

Let me walk you through some steps:

(1) *(Theoretical Problem) First you need to fetch some data from the DB using Python, since at this time we can't provide you with any DB access, please just write a scheme of the necessary class (you can also mock it). The data necessary for further steps is provided in a 'csv' file, no worries !*

(2) *Prepare the application structure of your choice, which you think will be the best one for all similar problems. For the DB connection part: either write just scheme of a class to communicate with a DB or mock the DB response and use provided 'csv' file as the data source. In order to give you some hints, we are always keeping in mind that this kind of application should have some key-words guided design: Direct Acyclic Graph, UniTests, Logging, make-file, Airflow, dont-repeat-tourself, Asynch, Vectorize, Pandas, Numpy, Scipy, Hypothesis … I hope this will guide you through.*

(3) *The main challenge of the data analysis consist in extraction of mean responses times per country, per merchant and the number of exchanges between merchant and a client. But remember to only count the time difference (dt) and the exchanges number (exch_nr) between **the first** message of the client and **the first** response of the merchant (c -> m ONLY!)*

(4) *Once you have all the data analyses please save it as a 'csv' file and prepare an API Post requests working in an asynchronous manner.*

Feel free to spend as much or as little time on the exercise as you like ! Once you have it figured outLet us know!

Additional Questions:

How long did you spend on the coding test? What would you add to your solution if you had more time?

CSV file hints:

Every column containing '_user_id' comes from Django framework user identification system and corresponds to a unique id. The 'merchant_id' column however is a unique key (pk) notation for a merchant used at BackMarket. In this sense, 'merchant_user_id' is django's user id for a given merchant.