Data Engineer Applicant Task

Task Description

As part of the application process, we would like to get more familiar with the way you (the candidate) work, think, and organize your code. The following task description gives some context to the supplied dataset. We would like you to write a Spark program that loads the data, analyzes data quality, provides a summary report, and reports your findings. We will evaluate the quality of the code that you write. You can analyze the data however you like, as long as you clearly explain the rationale behind your decisions.

Terminology

Granify is an eCommerce company, and as such our analysts work using the language of eCommerce. Here are some terms that are used in this task description:

- **Shopper** - an individual using an eCommerce website
- **Session** - the experience of a shopper on an eCommerce website within a single continuous period of time (if a shopper visits a site multiple times, sessions are split anywhere that there was at least a 30 minute break)
- **Conversion** - a session which resulted in a purchase (one conversion can have multiple transactions)
- **Marketing Strategy** - Any modification to an eCommerce website that is targeted to a shopper and is executed with the intent that it will increase the likelihood of a purchase

Fields

ssid - session identifier: used to link logs between files, it is a key composed of three values in the following format: user_id:site_id:session_start_time (session start time is taken from client side).

st - server timestamp: timestamp of when a web request was recorded on the server side

gr - determines assignment of a session to a control or experiment group

ad - indicates which marketing strategy a shopper was exposed to

Data Assumptions

- A shopper can have more than one session (each session separated by at least a 30 min break)
- Each session should have exactly one session log
- There is one marketing strategy per session
- Each session has a corresponding features log

Report format

After loading the data, we expect you to summarize and group it and prepare:

1) a populated table (tsv format) with the following header:

> Session start date at hourly granularity, site_id, gr, Ad, browser, number of sessions, number of conversions, number of transactions, sum of revenue
>
> Notes:
>
> Each row will contain aggregated data (key being first five columns)
>
> Session start date at hourly granularity: 1464742123 -> 2016-06-01 00:00 (UTC)

2) a list of means and standard deviations for each feature per every (site_id, ad) pair

Expected outcome

- Source code for Spark program to generate reports
- Report regarding data quality
- Reports with data summary

We expect the task to take 4-8 hours (assuming that you are familiar with the tools you are using and have some uninterrupted time to dedicate to the task). You may not be able to finish all data quality checks that you plan, so feel free to list additional procedures you would try using the dataset (along with an explanation of what you might hope to gain/discover by employing them).

Feel free to contact marcin@granify.com in case you have doubts or questions!