## Offline Interview Assignment

There are N banks(merchants) and M customers. A customer can have an account in multiple banks. <https://drive.google.com/drive/folders/1qryhdlgNsmecWRy2haI8S3uC63wKk5X-?usp=sharing>

/transactions.csv - represents various transaction data

/CustomerImportance.csv - A higher weightage for a customer and transaction type means banks want to give higher benefits to those customers for those transaction types. Ignore the fraud field.

Create a mechanism X to invoke every second and create a chunk of next 10,000 transaction entries from GDrive and put them into a S3 folder.

Create a mechanism Y that starts at the same time as X and ingests the above S3 stream as soon as transaction chunk files become available, detects the below patterns asap and puts these detections to S3 , 50 at a time to a unique file. Each detection consists of YStartTime(IST), detectionTime(IST),patternId, ActionType, customerName, MerchantId.

Whichever fields for a given detection aren't applicable, leave them as “” empty string.

Use postgres to store any temp information.

### Patterns

PatId1 - A customer in the top 10 percentile for a given merchant for the total number of transactions with the bottom 10% percentile weight averaged over all transaction types, merchant wants to UPGRADE(actionType) them. Upgradation only begins once total transactions for the merchant exceed 50K.

PatId2 - A customer whose average transaction value for a given merchant < Rs 23 and made at least 80 transactions with that merchant, merchant wants to mark them as CHILD(actionType) asap.

PatId3 - Merchants where number of Female customers < number of Male customers overall and number of female customers > 100, are marked DEI-NEEDED(actionType)

## Submission

Share the below to [hiring@devdolphins.com](mailto:hiring@devdolphins.com) with heading “PySpark Data Engineer - Submission - <YourFullName>”

1. A github link
2. A downloadable s3 link where all output files are zipped into a single file
3. A loom.com video each for
   1. showing a live running demo of the above in action.
   2. explaining the code
   3. explaining the setup
   4. showing sample output data, intermediary files and postgres data, if any.
   5. Architecture of X and Y explanation

### Constraints

1. Use Databricks Pyspark
2. Get a Free AWS account / Azure account with a new email id if you existing one expired
3. Make assumptions and write them done, if any
4. Any tech you don't know, please learn and attempt it
5. Deadline - 2-3 days from when open this file

You will judged based on attention to detail, output accuracy, architecture, and maintainable code