Eighty20 Data Engineering Challenge

# Birthdays

Attached is a list of a particular company’s clients’ birthdays (Q1\_birthday\_challenge.txt). There is something very interesting in the data. What is it? How do you think this happened?

Show how you read in and analysed this data. The tools you use should reflect what you would normally choose for ETL and validation of client data feeds.

# Database Design

[Submit basic diagram and discussion document for presentation at the interview]

We have just been engaged by a new financial services provider to help them understand their customers better. This client has multiple products that they offer their customers but we’re just interested in their credit card division. They can only give us a limited data feed (they’re still working on building their own mature data environment). Your job is to design a database for us to load their data into so that our analysts can interrogate it.

## Data Feed

There are 3 different feeds that we will be receiving (they will be dumped onto our SFTP server as zipped csv files) with details as follows:

* + Point of Sale (POS) Transactions
    - This file contains a record for every time a credit card is swiped at a retailer point of sale device
    - We get the timestamp, credit card number, transaction amount and unique POS identifier
    - The file lands at 7am every morning and contains all transactions from midnight the previous day to midnight this morning
  + POS devices
    - This file contains any new POS devices that the customers’ credit cards have been used at
    - We get the unique POS identifier, store name and merchant name (a merchant such as Checkers might have multiple stores)
    - The file lands at 8am on the 5th of every month and contains only newly identified devices since the previous file (existing known devices are not included)
  + Account Balances
    - This file contains the account balances of all active credit cards
    - We get the date, unique customer identifier, credit card number and balance (a positive balance means they owe us money)
    - The file lands at 8am on the 3rd of every month and contains the balance as at midnight on the 1st of that month

## Additional information

* + Customers may have more than one credit card
  + Credit cards can expire or be stopped if they are lost or stolen
  + Stopped or expired cards do not appear in the Account Balances Feed
  + When a card is replaced, any outstanding balance is transferred to the new credit card
  + Credit card repayments are not explicitly given in the data feed and must be inferred from the POS transactions and account balances
  + Our client currently has ~100 000 customers with ~1.5 credit cards each but they are looking to grow to ~1 million customers in the next year
  + They currently process ~ 50 000 transactions a day but estimate that this might peak at

~2 million transactions a day in the next year

* + There will almost certainly be some errors in the data, especially in the first few feeds

## Requirements

Our analysts want to be able to answer the following types of questions:

* + How are customers using their credit cards? E.g. average transaction value, average transaction frequency etc.
  + How is this usage changing over time?
  + What is the average credit card repayment rate?
  + How much do customers spend at different merchants?

## Deliverables

* + Design a database that will be used to house these data feeds as we receive them
  + The design should account for any requirements of the ETL process and any views or derived tables you provide on the data to make life easier for the analysts
  + Describe how your design is optimised to allow the analysts to quickly query the data for what they need. Include any questions you might want to ask the analysts in order to help you make future improvements
  + Describe how the ETL process would work. You do not need to go into the finer details, but clearly describe any issues that need to be dealt with (such as late arriving data) and how your design accommodates this
  + You have a sneaky suspicion that there’s something important missing from this data. Design for what you know but list out any concerns you may have