Spark Interview Exercise

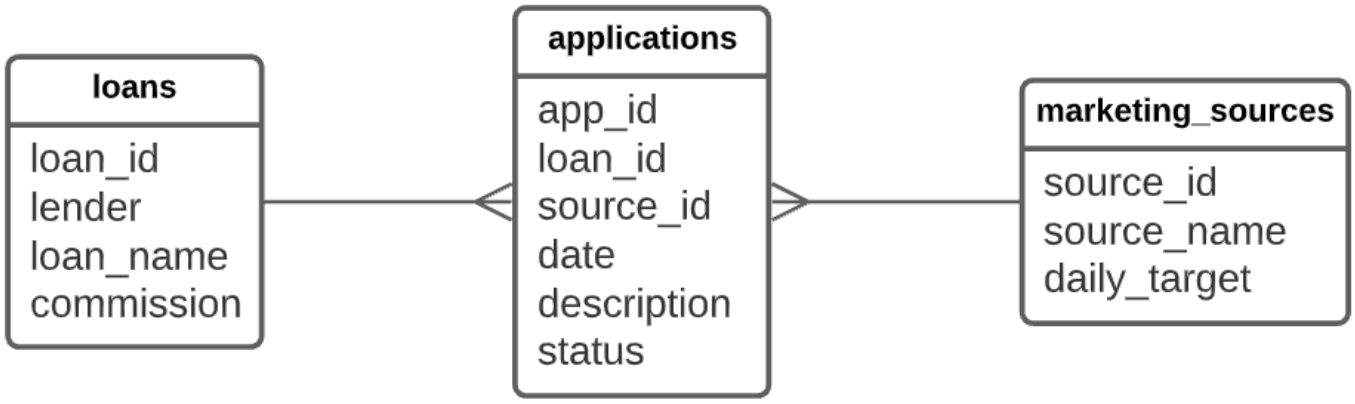
Working with Spark requires that you master several Skills including:

- Understand and transform data
- Test-driven development. Testing on real data in prod is not possible with Big Data
- Good balance on how implementation can affect Spark Jobs

The Dataset

Our dataset consists of three tables that describe a mini-version of a loan broking shop like Ocean Finance. Data is available in parquet and can be downloaded here: https://drive.google.com/file/d/1zTNAxVhtJcjBb8Vr__O16j6nvnF8uNiA/view?usp=sharing

The diagram shows the three aforementioned tables



Tables

loans

Every row contains a distinct loan type

- loan_id: This is used to identify the loan type and lender
- lender: The lender offering this loan. e.g. Darclays
- loan_name: A given name for the loan. e.g Matwest 19% Personal Loan
- commission: The commission the company receives for every successful application in GBP.

applications

Every row contains a distinct loan application

- app_id: The application form unique identifier
- loan_id: The loan type this application is referring to
- source_id: Refers to the marketing source this applicant came from
- date: Date of the application was submitted
- description: Application description
- status: submitted, approved, declined

marketing_sources

Contains a series of all Marketing sources we are advertising

- source_id: Unique marketing source identifier
- source_name: Facebook
- daily_target: number of applications we are expecting from this source

Exercises

Write a Spark job in Python that uses the provided dataset and answers/displays the following questions. Try to provide clear instructions in a README file on how to run this.

Exercise 1

How many applications have been submitted from the beginning of time?

Exercise 2

What is the average profit of all applications?

```
An application produces profit only if it is approved.
An application is approved when its status = `approved`
```

Exercise 3

Which marketing sources are the first and second most popular **for each loan type**?

Example output:

Loan	Most popular	Second most popular
Matwest 39%	Google Ads	Facebook Ads
...	...	...

Exercise 4

Provide a list that shows for each day what is the percentage of profit generated by each marketing source and to what percentage did they reach their daily target

```
daily_profit = for each loan (commission * applications on the day)
Daily Target % = daily_profit/daily_target
```

Date	Source	Profit	Daily Target %
2022-01-01	Facebook	£10000	50
2022-01-01	Google	£30000	70