TECHNICAL CHALLENGE

EMERGING TECH PRACTICE – DATA ENGINEERING RECRUITMENT

1.1 Summary

The purpose of this challenge is to assess a candidate's skills within the Data Engineering technology landscape. It will simulate, in a small-scale lab, a caricature of what a typical task a Data Engineer can face on a daily basis.
This document is organized as follows:

- **Exercise:** A business description of the challenge
- **Deliverables:** What CGI expects from the challenge
- **Resources:** Where to download the exercise's resources
- **Evaluation:** Timeframes and evaluation criteria

1.2 Exercise

A Retail client wants to analyse the amount of sales generated on its stores, to better understand its business and assess possible opportunities to refine and optimize its internal processes.
As such, the Retail client presented to CGI the following use cases:
– Considering only the first 100 days of sales, which were the 5 stores that produced more revenue;
– Considering only the first 100 days of sales, for each store what were the two best selling products

To tackle this use case, the Retail client supplied CGI with two data sources:

1. Sales_train_validation.csv: Contains the historical daily unit sales data per product and store [d_1 - d_1913], for the sake of these exercises consider just the first 100 days.

2. Sell_prices.csv: Contains information about the price of the products sold per store and date.

3. Calendar.csv: Contains information about the dates on which the products are sold.

The Retail client expects CGI to deliver a small report, describing the use case, accompanied with technical documentation to support its IT team, and the technical solution.

1.3 Deliverables

Technical solution

This section will test the candidate's technical skills, within a slice of the technology stack that CGI uses when delivering Data Engineering pieces to clients, as well as software development general awareness.

The solution must be developed using Apache Spark, preferably Python but the candidates can also choose to complete it using Spark SQL. Candidates can choose between using an IDE like Pycharm or notebooks like Jupyter, Zeppelin. If you are not familiar with the above, you can choose to complete the assessment using community edition of databricks which is free to signup (please check the link below):
https://community.cloud.databricks.com/login.html

We expect the candidate to follow development best practices, such as Test Driven Development (if possible use unittest python library), clean code, and code readability (commentaries, meaningful variables). Every asset the candidate develops (e.g., test suite) should be packaged and sent as part of the overall solution.
The source code must be sent to CGI in a zip file in case you are using an IDE.

If you are using jupyter/databricks, please make sure the notebook is legible and readable with clear description of the problem / solution. Additionally, you can also provide visual insights using charts. Finally send the notebook file over to CGI in any of these formats (*.ipynb, *.py, *.sql) .

Business insights

This section will test the candidate's general consultancy skills and ability to communicate value to the business.

A small report, with no more than two pages (one page being ideal), detailing the candidate's findings. We expect the report to have a brief description of the technical solution and a business-oriented description of the use case findings.

1.4 Resources

This exercise uses data sets from Kaggle that can be downloaded at the following link.
https://www.kaggle.com/c/m5-forecasting-accuracy/data
(You must signup/sign in before you can download the datasets)

Multiple resources to install and run PySpark exist on the internet, regardless of OS, but if the candidate is having trouble setting up a development environment, he/she can reach out to CGI for help. As mentioned above, if you are not familiar with setting up Interactive development environments, we recommend you to sign up for free community edition of databricks, which is fairly straight forward.

1.5 Evaluation

CGI will give a response, within five working days, after the candidate hands-over the expected outputs. The response will comprise of an overall assessment of each deliverable and feedback regarding if the candidate passed on to the next stage of recruitment.

Regarding evaluation, CGI will assess:

- Code functionality (does it generate the expected results)
- Code quality (software development best practices)
- Report quality (is it clear and delivers value)

At CGI, we value innovation, and encourage the candidate to explore the boundaries of the challenge's scope, which can be manifested in, but not exclusively, technical enhancements (e.g., providing the solution in a non-scripting approach, following software development best practices) or better business insights (e.g., build upon the requested use case and discover new and interesting insights).

Experience the Commitment and have fun!