Assessment for Software Engineer Internship at DataGrokr

Thank you for your interest in the Software Development Internship at DataGrokr.

We anticipate the selected candidates to be working in Data Engineering and Cloud related projects. As such for this given assignment, we'd like to test candidates' skills in those areas. Candidates who are already proficient in SQL, Python and Spark will have an edge in this assignment but even if you didn't know anything about any of these technologies you should be able to do this assignment by following along the instructions and studying the links provided.

Please note that this ability to learn new technologies and following instructions will be a key skill required in your day to day job at DataGrokr.

What you need to do:

The objective of the assignment is to test your proficiency in querying and data analysis. The assignment has 3 parts.

- Section 1: you will provision an environment in Databricks (a provider of managed Spark as a Service on the Cloud) and load some data sets.
- Section 2: you will analyse the relationships between the data sets using an ERD diagram. You will then be asked to answer some business questions about the data. You will answer these questions by writing SQL statements. You will execute these queries using Spark on the Databricks cluster provisioned in Step 1;
- Section 3: It is same as Section 2, except you will be using Spark API instead of Spark SQL.

Section 1: Environment setup and data loading.

1. Go to Databricks community cloud and create a free account to get access to a single node Spark cluster:
    a. Go to https://databricks.com/try-databricks link and select the "**COMMUNITY EDITION**" (click on "get started" button).
    b. Provide all details and sign up.
    c. Verify your email id and select a password to get more information
    d. You can watch this video on YouTube to get started with Databricks community cloud.
2. Load the Northwind dataset into the cluster
    a. Northwind database is a set of data sets that is shipped with Microsoft Access and is used in learning SQL. There are several resources available online to learn more about Northwind. (https://theaccessbuddy.wordpress.com/2011/07/03/northwind-database-explained/)
    b. We have sampled down the files and create a zipped file. You download the files from here.
    c. Once you have downloaded the files, create a new Jupyter Notebook and import the data into the cluster.
    d. Create dataframes for each of the data sets. Give proper column names and datatypes. (refer the ER diagram provided with the data for reference)

e. Register those dataframes as tables.
f. Please use PySpark and not Scala.
g. If you are new to Spark, refer to the Databricks and Spark documentation to learn about Notebooks, dataframes, loading data, etc. Below are some links that maybe useful for you to learn spark:
    i. https://docs.databricks.com/getting-started/spark/index.html
    ii. https://spark.apache.org/docs/latest/api/python/pyspark.sql.html
    iii. https://www.analyticsvidhya.com/blog/2016/10/spark-dataframe-and-operations

Section 2: Working with data and Spark SQL.

a. For this task you must run Spark SQL queries on the tables you registered in Section1. Please create a separate dataframe for each of the below queries. The queries progressively increase in complexity.
    i. Write a query to get Product list (id, name, unit price) where the unit price products cost less than $20. Filter out discontinued products.
    ii. Make a listing of all categories of products in the order of decreasing number of products in that category. Filter out discontinued products.
    iii. Make a list of customers who have not made any orders in the months of July - September.
b. Please refer to the ERD diagram provided in the zipped to understand the relationship between the tables. Pay attention when to use outer joins vs. inner joins.
c. Please write the queries in the same Notebook.

Section 3: Working with data and Spark API.

a. For this task you must use Spark dataframe API instead of Spark SQL and create separate dataframe for each of the below queries. The queries progressively increase in complexity.
    i. Write a query to get Product list (id, name, unit price) where products cost between $15 and $25.
    ii. Write a query to get Product list (name, unit price) of ten most expensive products.
    iii. Write a query to get Product list (name, unit price) whose prices are above the average price.
    iv. Write a query to get count of current and discontinued products.
    v. Give the names of employees who sell the products of more than 7 suppliers.

Deliverables:

a. Jupyter book where you have developed the code for the above 3 sections. Download the files and email it to us. We will run the Jupyter book on our end and correct your submissions.
b. Your final submission should be sent to *help-me-help-you@datagrokr.com*. Your submissions are due to us by end of day **25th August 2019.**
c. **Bonus**: Please use markdown cells to make your code descriptive and self-explanatory. Your Jupyter notebook should speak for itself and represent all the effort you put into the assignment.

If you have any questions during the assignment, send your questions to *help-me-help-you@datagrokr.com*.

Good luck and we hope you learn something new in this process!