Assessment for DE Internship at DataGrokr

Thank you for your interest in the Data Engineering 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 follow 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:** Setting up PySpark in Colab and loading some data sets.
- **Section 2:** Analyze the given dataset and answer the question using Spark(pySpark).
- **Section 3:** Create a class to expose generated result-set from Section 2 to data consumers.

Section 1: Environment setup and data loading

1. Open this link to create a new Colab notebook (You need to sign in to your google account if not signed). Follow the below steps to setup spark in your notebook
    - Spark is written in the Scala programming language and requires the Java Virtual Machine (JVM) to run. Therefore, our first task is to download Java.
        - ```
          !apt-get install openjdk-8-jdk-headless -qq > /dev/null
          ```
    - Next, we will download and unzip Apache Spark with Hadoop 2.7 to install it.
        - ```
          !wget -q https://archive.apache.org/dist/spark/spark-3.1.2/spark-3.1.2-bin-hadoop2.7.tgz
          ```
        - ```
          !tar xf spark-3.1.2-bin-hadoop2.7.tgz
          ```
    - Setup Environment variables for Java and Spark
          ```

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.1.2-bin-hadoop2.7"
```

- Then we need to install and import the 'findspark' library that will locate Spark on the system and import it as a regular library.

```
!pip install -q findspark
import findspark
findspark.init()
```

- Now, import SparkSession from pyspark.sql and create a SparkSession, which is the entry point to Spark.

```
from pyspark.sql import SparkSession

spark = (SparkSession
  .builder
  .appName("<app_name>")
  .getOrCreate())
```

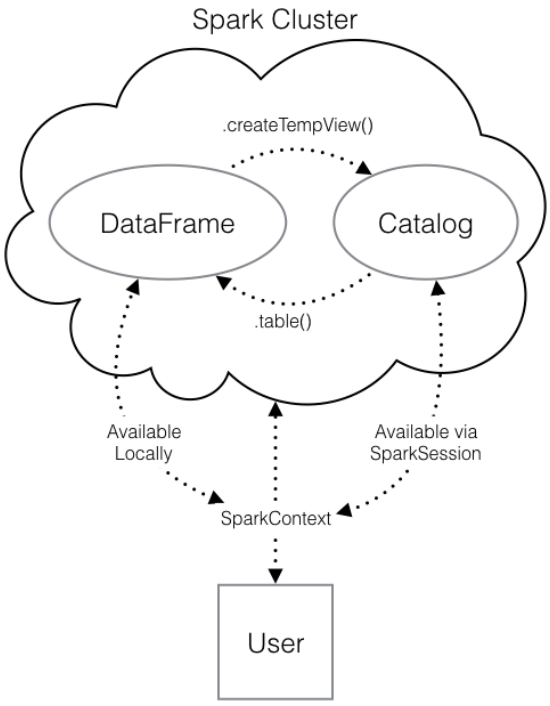
2. Download the files needed to complete this assignment from here and upload them in content directory of colab notebook.

3. Create dataframes for each of the datasets. Give proper column names and datatypes. (Refer the schema provided with the data for reference) Check out spark.read.format from here to create spark dataframe

4. Your SparkSession has an attribute called `catalog` which lists all the data inside the cluster. This attribute has a few methods for extracting different pieces of information.
   One of the most useful is the .listTables() method, which returns the names of all the tables in your cluster as a list.
   Register those dataframes as tables using .createOrReplaceTempView() , this method registers the DataFrame as a table in the catalog, but as this table is temporary, it can only be accessed from the specific SparkSession used to create the Spark DataFrame. To read about it refer to this link

5. You can use pyspark dataframe functions like .select() or you can use .sql() to provide the sql query inside it to solve the problem. Check out this link for more info.

6. Check out this pyspark documention if you are new to it.
7. Check out the diagram to see all the different ways your Spark data structures interact with each other.



Section 2: Using pySpark for data analysis

1. Find the top 3 venues which hosted the most number of eliminator matches?
2. Return most number of catches taken by a player in IPL history?
3. Write a query to return a report for highest wicket taker in matches which were affected by Duckworth-Lewis's method (D/L method).
4. Write a query to return a report for highest strike rate by a batsman in non powerplay overs(7-20 overs)
   Note: strike rate = (Total Runs scored/Total balls faced by player) *100, Make sure that balls faced by players should be legal delivery (not wide balls or no balls).
5. Write a query to return a report for highest extra runs in a venue (stadium, city).
6. Write a query to return a report for the cricketers with the most number of players of the match award in neutral venues.
7. Write a query to get a list of top 10 players with the highest batting average Note: Batting average is the total number of runs scored divided by the number of times they have been out (Make sure to include run outs (on non-striker end) as valid out while calculating average).

8. Write a query to find out who has officiated (as an umpire) the most number of matches in IPL.

9. Find venue details of the match where V Kohli scored his highest individual runs in IPL.

10. Creative Case study:
    Please analyze how winning/losing tosses can impact a match and it's result? (Marks for Visualization also)

Section 3: Expose Data

1. Create a database (use any relational DB preferably SQLite) and load data from the dataset (Section 1) into db.

2. Create a Class Database. Class will need to have

    1. Constructor to initialize dB connection and other variables
    2. Methods implemented to return result-set for each query in Section 2, result-set should be returned as a json/dict object.
    3. Exception handling.
    4. get_status method to ping database connectivity.

3. Feel free to create any additional classes or data structures you deem necessary.

4. Evaluation: Here is an input example

    ```
    from database import Database
    db = Database ()
    qry1_result = db.get_query1_result()
    ```

5. Please follow industry standards while writing the code and include basic schema and data validations.

6. Preferred Programming language – Python

Deliverables:

1. For Section 1, Section 2 and Section 3: A single colab notebook where you have developed the code for the Section 1, Section 2 and Section 3. Download the colab notebook and email it to us. We will run the colab notebook on our end and correct your submissions.

2. Your code will be evaluated not just on the final answers but on code quality and unit tests

    - Follow coding standards (PEP-8)
    - Appropriate error/exception handling
    - Modular function design

3. Your final submission should be sent to *dataengineering@datagrokr.com*. Your submissions are due to us by end of day 23nd Feb 2022 and subject should follow following pattern <Source e.g.: Collegename> Data Engineer.
4. Please include your up-to-date resume, named as Firstname_Lastname.pdf

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

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