**Assessment for Software Development Internship at DataGrokr**

Hi,

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

DataGrokr provides solutions to our global clients in the fields of Cloud enablement, IT Automation (DevOps), data management and Big Data Analytics. We work on several different technologies and we will have to adapt very quickly to our client demand. Learning agility and ability to deal with new technologies on the fly is a must-have requirement to succeed at DataGrokr. If you are wedded to a specific technology, DataGrokr is not the right place for you.

As such, our selection process is geared to identify candidates who’ll enjoy this type of work and thrive in this environment. As a part of our selection process, we ask interested candidates to complete a technical assignment. The details for the technical assignment are provided in the attached document. The document contains detailed instructions on what you need to do as well as well as the deliverables expected.

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 and Python 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 using online resources.

Your submissions are due to us by end of day February 28th. Based on your submissions, we will shortlist some candidates and schedule F2F/Skype/Hangout interviews in the following week. We hope to start our internship in late March / early April.

Good luck and we hope you enjoy the assignment and learn something new in the process irrespective of the outcome of your internship.

**Note to students in 4th year**: We are only going to consider candidates who are available for 6 months and who could potentially accept full-time offers at the end of the internship. We will be flexible in terms of start and end dates and your academic commitments (exams, vivas, etc.). if you are not planning on taking a full-time offer or have another full-time offer you’d like to consider, you should consider leaving the opportunity to another deserving candidate.

DataGrokr Team.

**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 install MySQL database 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 Python
* Section 3: You will demonstrate your knowledge of Python and do some data analysis using pandas.

**Section 1: Environment setup and data loading.**

1. Install MySQL (server and client) on your laptop:
2. Load the Northwind dataset into the cluster
   1. 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/>)
   2. We have sampled down the files and create a zipped file. You can download the files from [here](https://drive.google.com/file/d/11D6aORzzGdiPhWj8JSC8QdB41L_0A5k9/view?usp=sharing).
   3. The ER diagram is also present along with the files. Please review the data model and understand the relationship between the entities.
   4. Create DDL statements using the ERD diagram and create those files in MySQL. Load the files into those tables. The table names, column names and data types should match with what is provided. If your schema doesn’t match with ours, we will not be able to test your code.
3. Deliverables:
   1. DDL statements
   2. Loading scripts for populating data into the tables

**Section 2: Working with data and SQL.**

1. For this task you must write the SQL statements that will provide the necessary results.
   1. Create a report showing the title of courtesy and the first and last name of all employees whose title of courtesy begins with "M".
   2. Tax rates have gone up. If the cost of freight is greater than or equal to $500.00, it will now be taxed by 10%. Create a report that shows the order id and customer details, freight cost, freight cost with the new tax for all orders of $500 or more. (Orders and Customers).
   3. Make a list of products and the number of orders in which the product appears. Sort the results based on products present in the greatest number of orders.
   4. Group the number of orders based on the freight cost using the following breaks: < $51, $51 - $100, $101 - $150, > $150
   5. Identify the customers who have no orders placed.
2. 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.
3. Deliverables:
   1. Queries / SQL statements

**Section 3: Working with data using Python / pandas**

1. For this task you must write a Python program that connects to the database and pull the data into a pandas dataframe, does the necessary operations in pandas and writes the final output to a S3 bucket.
   1. Make a list of products and total up the number of actual items ordered. Sort the dataframe in decreasing order of items orders and save the file to S3 including the column headers.
   2. Group all the orders by the age of employees as of ‘2000-01-01’. The ages should be broken up as - under 30, 31 – 40, 41 – 50, above 50. Plot a histogram for the above results and save the image to S3.
   3. We have new tax rates based on the country of the order. Retrieve all order details and update the revised freight cost. Save the new order details into a file on S3.
      1. [USA, Canada] – 15% tax
      2. [France, Germany, Belgium] - 10% tax
      3. Remaining countries – 5% tax
2. Please check the output first in your local machine before uploading the files to S3 as you will not have read access on the S3 bucket only write access. The S3 files should be named with the following naming convention
   1. \<your firstname\_lastname>\query1.csv
   2. \<your firstname\_lastname>\query2.jpg
   3. \<your firstname\_lastname>\query3.csv
3. S3 details:
   1. S3 bucket name: dg-assessment
   2. Access Key: AKIAQAQEAS4WJ7F2BSAK
   3. Secret Access Key: 1rL7qFO9sT5fRaJyCCdgEt7SSj9pWrtKL0gptGnt
4. Deliverables:
   1. Python program
   2. Names of the files uploaded to S3

**Deliverables:**

1. Your final submission should be sent to *help-me-help-you@datagrokr.com*. Your submissions are due to us by end of day **29th February 2020.**
2. Please send your contact details (email, phone number) and attach your latest resume with the submission.

If you have any questions during the assignment, send your questions to [*help-me-help-you@datagrokr.com*](mailto:help-me-help-you@datagrokr.com)*.*

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