Data Eng Take Home Project 2018

All of us at Weebly are really excited for you to start the next step in the hiring process!

First and foremost, this take home project is aimed to help us answer the question, "Will you succeed at Weebly?". We want to see you succeed, and we will help you along the way. We'll help you the same way as we would one of our current team members. Please call, text or email us any time of any day.

Project

Given the data available to you in the attached files, please complete two parts:

1) Build a Data Pipeline

Build a pipeline to store the data we provide in a data store of your choice, on your machine. This should be representative of a miniature data warehouse.

2) Anomaly Detection / Data Insight

Please identify potential anomalies in the data. Hopefully you have constructed your pipeline to make anomaly tracking a painless endeavor. Keep an eye out for issues like questionable data values, presence and lack thereof, and other anomalies that will be up to you to track down. Explain the logic, methodologies, tools and techniques used to identify the anomalies.

BONUS: Also please give insights / suggestions to this merchant for improving her/his business.

Please send your source code and video demo link to the recruiter.

Timeline

You have one week from the day we send you the project to complete it and send it to us. We are not going to time you by the hour. If we send you the project on a Tuesday, we would expect you to submit your work by end of day following Tuesday. We expect this project to be completed in a span of around 1-2 days maximum.

Video Presentation

Please prepare a 2-to-15 min video demo where you can show off your work to a few of the engineers, product managers, analysts and engineering leaders at Weebly. Please upload this video to YouTube or any other video hosting website of your choice and share the link with us.

Be sure to discuss your implementation (your code), the decisions/trade-offs, and your thought process on why you did what you did. And please share your wild ideas as well! That is, if there were no time and resource constraints, how would you do things differently.

It would be helpful to include:

- Your understanding of the requirements from this project.
- Architecture for current setup proposed in the project.
- Trade-offs made when choosing different technologies.
- Actual working product demonstrating the desired output.
- Proposed architecture for future setup when this problem need to be solved at TB/PB scale of data volume.
- Assumptions made during development of this project.
- Insights developed regarding Weebly’s business from the data and conversations.

You can use any screen capture software to create video. Some are listed here (https://blogs.systweak.com/2017/08/10-best-screen-recording-software-for-mac). OBS is a favorite of Weebly engineers (https://obsproject.com/download).

Interactions

Don’t be afraid to ask questions; we want to help you succeed and we are cheering for you! Reach out if you notice anything peculiar with the data or need additional clarification. **The primary goal for you is to provide the core functionality and performance with the full dataset provided.** If you are spending more than a couple of hours on one particular task, it might be time to give us a holler.

Tools

In regards to tools, libraries etc, please use what you think is the most appropriate tool for the job. Your goal is to impress us.

- Some popular tools for building pipelines are Spotify's Luigi, AirBNB's Airflow, Azkaban, Oozie, and so on. You're welcome to use any of the tools out there.
- We expect you to wisely choose a data warehouse store of your choice. There are many open-source choices in databases, please feel free to pick the one you like the best. In your video presentation, please do discuss any trade-offs your choice and the alternatives.

Please provide any source code and steps to reproduce your results. (ie command line prompts, documentation on steps that’s not covered in the source code, and config files.)

Help!

For any questions about this project, feel free to email, call or text either Gagan or Shuyu *any time of the day or night*.
Gagan Sandhu - Email: gagan@weebly.com Cell: 415-694-9443
Shuyu Wang - Email: shuyu@weebly.com Cell: 607-697-3557

Data

Here are some invoice data of an e-commerce store from a open data resource. The 3 tables attached include invoiceno, stockcode, quantity, invoicedate, unitprice, customerid and country.

customer_info.csv

invoice.csv

product_info.csv

Honor Code

Please feel free to use all online resources and books available out there but please do not collaborate with others. Feel free to collaborate with us as much as you like!