NYDIG

Data Engineering Candidate Project

Instructions

This project consists of two parts, data modeling and coding. We want you to design a data model for blockchain data as it comes from our Bitcoin node, and then write a script that iterates through the provided data and transforms it into the model you designed and dumps it to S3 in the AWS free tier. You will then create glue tables on the data that would allow it to be queried from AWS resources and create some aggregate metrics to measure on-chain activity. You will be evaluated on the quality and usability of your data model, the utility of your data pipeline design, and the quality of your code.

Design:

1. You are provided 5 blocks worth of blockchain data as it comes from the Bitcoin node, in large json files
2. Develop a data model that allows all data in the provided json files to be contained in flat files (CSV)
    a. All data from the jsons must be contained within your data model, do not drop data
3. Design some aggregate metrics on the blockchain data to measure the network activity
4. Design a data pipeline that will do the transformations and calculate the aggregates in a production system using AWS technologies and create a diagram of the pipeline.
    a. Justify your design choices, ie: why this kind of database? How do you think about data quality? Specifically for event based time-series data? What things would you check for? How do you determine something is wrong? How, when, and who needs to be alerted to an issue? Etc.

Code:

1. Write Python code that will transform the provided json files to the data model / schemas you designed and transfer them to an S3 bucket in an appropriate Hive structure.
2. Create Glue tables on the normalized data and develop some aggregate metrics that tell us about the amount of traffic on-chain for those 5 blocks.
3. Check all code into a github repo that the NYDIG team can review, including table definitions and views that contain the aggregate metrics
    i. We will have a follow-up meeting for you to walk us through your solution in AWS.

Deliverables:

1. Diagram / documentation (write up) of your data model and data pipeline solution
2. Github repository with all code used
3. An informal walkthrough of your AWS solution

This project is ambiguous by design. We believe in deep collaboration at NYDIG and the intention of this exercise is to get a feel for your thought process. We encourage you to approach this project as if you were already working for NYDIG. Steve Collins will be available to answer questions and be a sounding board as you work through this project. We do not expect you to spend more than six to eight hours in aggregate. We would like you to aim to have this completed within seven business days of receipt but feel free to take more time if needed. Please send your deliverables to steve.collins@nydig.com and kyle.roessler@nydig.com when finished and we will schedule time for you to walk us through your solution. Good luck!