Data Engineering Interview: Big Data Handling & Analysis

Instructions

Congratulations! You are now at the practical round of the rain Data Engineering interview process. The end is now closer than the beginning. The instructions for this round are as follows:

Together with this document, you have been provided with an anonymised dataset consisting of more than one hundred million network transactions in a compressed file that is named "***data-engineering-case-study.tar.gz***". Note that the uncompressed file takes up to 31.6 GB of memory. The actual size of the uncompressed flat file is 10.1GB and the compressed file is 2.6 GB.

Practical

Your instructions are to engineer the data so that it can be queried in an optimal way using SQL-like language. You may use any big data solution you want. Once you have completed this, use SQL to query the data and answer the following questions:

- How many transactions are there in the provided data set?
- What is the time period (in minutes) of the collected data?
- What is the destination IP address with the highest number of transactions?
- What is the destination IP and port combinations with the highest average session time and what is that average time in minutes?
- What is the busiest times of the day in terms of number of transactions? Aggregate the data in quarter-hourly (15 minutes interval starting at hh:00). Show trends – you can use any visualization tool for this part

All the information that you need to complete this assessment is in this document.

Theory

Data Transfer from FTPS to AWS:

- Considering that our data is currently hosted on an on-premises FTPS server, could you describe a detailed process for securely transferring this

data to AWS to facilitate analytics? What AWS services would you recommend for this process, and are there any best practices for ensuring data integrity and security during the transfer?

Choosing a Data Storage Solution on AWS:

- Given the need for high-performance querying capabilities on large datasets, which AWS database or data warehousing solution would you recommend? Include how these solutions support scalability and manage query performance effectively?

Data Structuring for Efficient Queries:

- In terms of structuring data to optimize queries related to transactions, session times, and IP addresses, what are the best practices? Should the data be normalized or denormalized, and what indexing strategies should be considered to enhance query efficiency?

By participating in the case study, you agree that any work submitted by you will be your own and that you will not receive help from another person. You also agree that you will treat this case study and associated dataset as confidential and that you will not make any copies and/or share this case study with anyone without the explicit consent of the Head of Operational Intelligence at rain South Africa. Good luck!