SI Partner Data Cloud Architect
Technical DE Demo exercise

As part of the interview process, we'd like for you to show off your technical hands-on building skills. Please use the Electric Vehicle Population Dataset (attached json file) and as a Data Engineer/Architect we would like you to:

Part -1 (Current Architecture)

- Analyze the data and describe all the fields you will be extracting to derive meaningful insights from the dataset.
- Read the JSON file which is in the cloud storage (Azure Blob/ADLS Gen-2 or S3) and parse and extract all the elements of the Json object.
- Use the lakehouse/data lake architecture and create objects accordingly.
- Store the master data in a cloud database (e.g. Snowflake DB, Azure SQL, or any DB)
- Identify the approvals, submit information from the Json file and parse the information.
- Get the column names for all the fields used in the dataset and load it in the destination (mentioned above).
- Provide an approach to dynamically parse the columns required for the insights. Tomorrow if a new column is added how will you process that information? Provide an approach to restrict the addition of new columns automatically.
- Create Python modules/Scala objects to segregate the various transformations that you are building. Your main method should use the functions defined in those Python modules/Scala Objects. {Looking for approach of how you modularize the code/implement}
- Perform simple data quality checks.
- Ingest the extracted information into Snowflake DB or Spark Delta tables. Files for these Delta tables must be on the cloud storage.
- Showcase 2-4 insights derived from the extracted data using queries or visualizations (See below for example insights).
- Check-in the source code to Github repo.
- Describe the steps involved in orchestrating the Pyspark/Scala code {Architecture Diagram}. High level steps for deploying the code in different environments (UAT/Prod).

Example of insights that could be derived: (use your own scenarios to derive the insights)

- Which one of the car make is more efficient?
- Is there any relationship between the choice of EV make and city?
- Which Plug-in Hybrid Electric Vehicle (PHEV) is preferred by buyers?
- Based on the data, which car make and model would you recommend?

Part-2 (Future state architecture)

- *Build the same solution using Snowflake Snowpark APIs. This need not be a full blown solution on Snowpark. A partial implementation of the solution is good enough, this is to analyze how much you were able to understand the concepts of Snowpark*

We'd like you to build a technical demo using Snowflake (Leverage Snowflake trail account) Overall, this effort typically takes about 3-5 business days of your time, so please plan accordingly.

Important Disclaimer

- Leverage free or trial accounts to demonstrate the demo.
- Snowflake is not responsible for any costs associated with building the demo and does not reimburse the costs.

What do we look for in this exercise?

- Your hands-on technical building abilities
- Your Snowflake knowledge (basic understanding is OK)
- Good knowledge on Python, PySpark, Spark, Pandas
- Your knowledge on data lake/Lake House architecture
- Breadth and depth of your knowledge in data engineering and data ops
- Your ability to demo solutions of Data Engineering pipeline using Spark and Snowpark APIs (Python/Scala/Java)
- Your understanding of the Snowpark frameworks and architectural design patterns you chose to use.
- Check-in the source code to Github repo.

Deliverables

1. An end-to-end demo using Spark for building a data lake on EV population dataset. And a sample walkthrough of the solution built on Snowpark.
2. A reference data lake /Lake House architecture slides outlining the solution built using Spark and Snowpark(before & after… before using Spark, after using Snowpark).

Format (60 minutes)

10 minutes - Use 1-2 slides to provide an overview of the use case and solution architecture.
35 minutes - Demo to walk us through an end-to-end pipeline solution talking about data ingesting options, data loading, transformations using Spark APIs. Small demo on Snowpark implementation of the written Spark code.
15 minutes - Q&A from a panel of interviewers.

Audience

We'd like for you to assume that the audience has both technical personas (like developer, software engineer), and business personas (Director/VP level). Tailor your presentation to cater to both these audiences.

References and Helpful Resources

Quick start and HOL for getting started with Snowpark Python

https://github.com/Snowflake-Labs/sfguide-getting-started-snowpark-python/tree/main/customer-churn-prediction

https://quickstarts.snowflake.com/guide/getting_started_with_snowpark_python/index.html?index=..%2F..index#0

A Dive into Slowly Changing Dimensions with Snowpark

https://quickstarts.snowflake.com/guide/snowflake_transformer/index.html?index=..%2F..index#0

Snowpark Python with DBT HOL

https://quickstarts.snowflake.com/guide/data_engineering_with_snowpark_python_and_dbt/index.html?index=..%2F..index#0

Spark to Snowpark API Mapping

https://medium.com/snowflake/migrating-from-pyspark-to-snowpark-python-series-part-1-a75058c1e579
https://medium.com/snowflake/migrating-from-pyspark-to-snowpark-python-series-part-2-9c87120097f3

Flatten JSON

https://docs.snowflake.com/en/sql-reference/functions/flatten

Json Source File:
ElectricVehiclePopulationData.json