Ken Kim, March 16, 2023

EV Population Data Pipeline Implementation & Demo

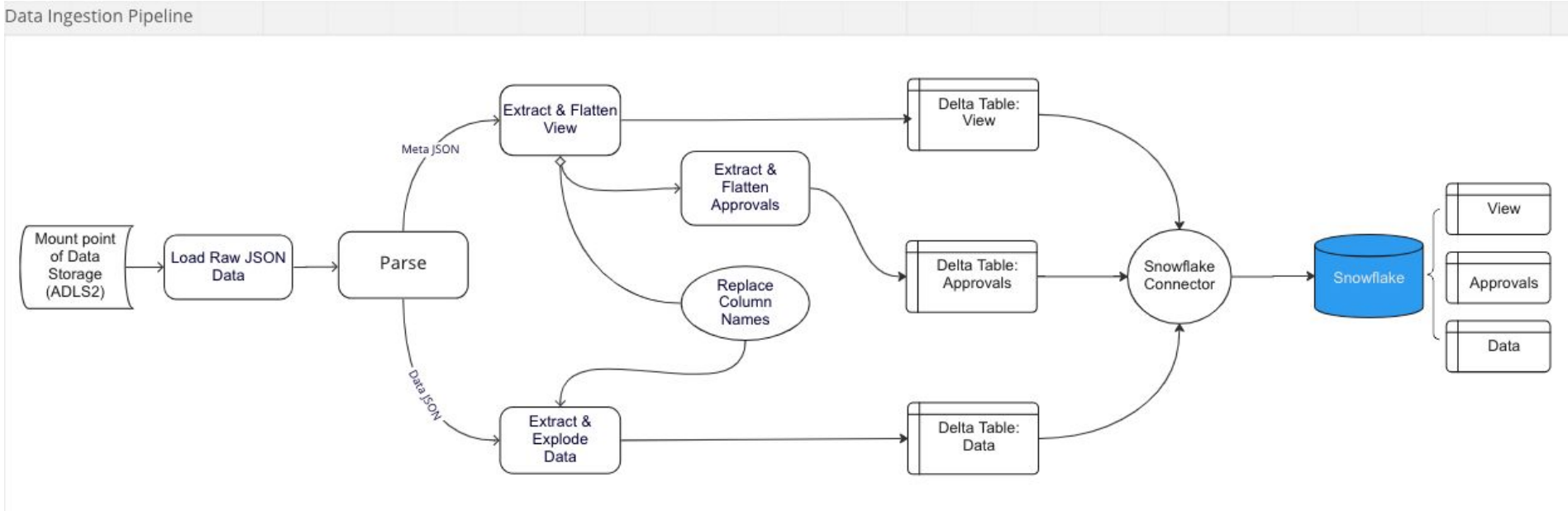
TOC

Overview

As a part of the interview process for the Snowflake SI Partner Data Cloud Architect role, the hiring team would assess and review the candidate's technical hands-on skills in Spark, data processing, and data pipeline design.

This assessment requires a demo implementation exercise, which requires the candidate to:
- build a data pipeline that parses the raw data from an Electric Vehicle Population Dataset (provided in a JSON file),
- extract the necessary information, and ingest it into Delta tables and then into DB tables.
- create insightful analysis based on the datasets.

Pipeline Architecture



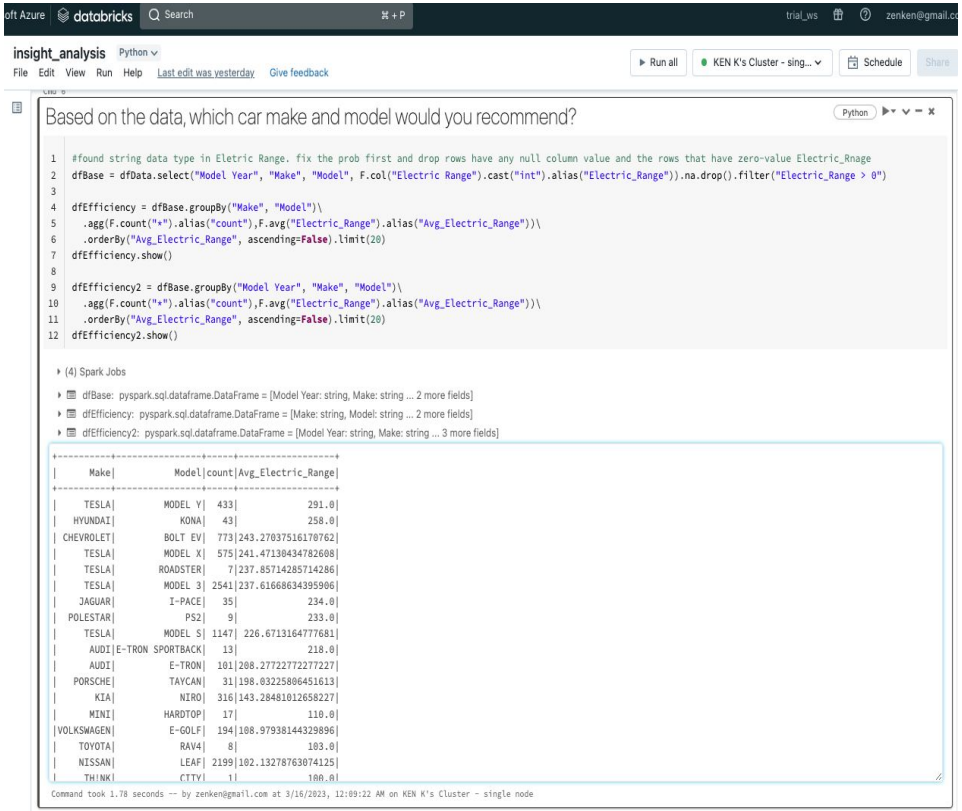
Demo

01: Databricks/Spark

Using Databricks Notebooks, created a Job pipeline which:

- Loads the raw data from ADLS2 storage
- Creates Spark dataframes to parse and extracts JSON objects, meta/view, meta/view/approvals, meta/view/columns and data[]
- Creates Delta Tables to flatten and store the parsed data
- Inserts the same data into Snowflake tables using Snowflake python connector

And created insights analysis example queries.



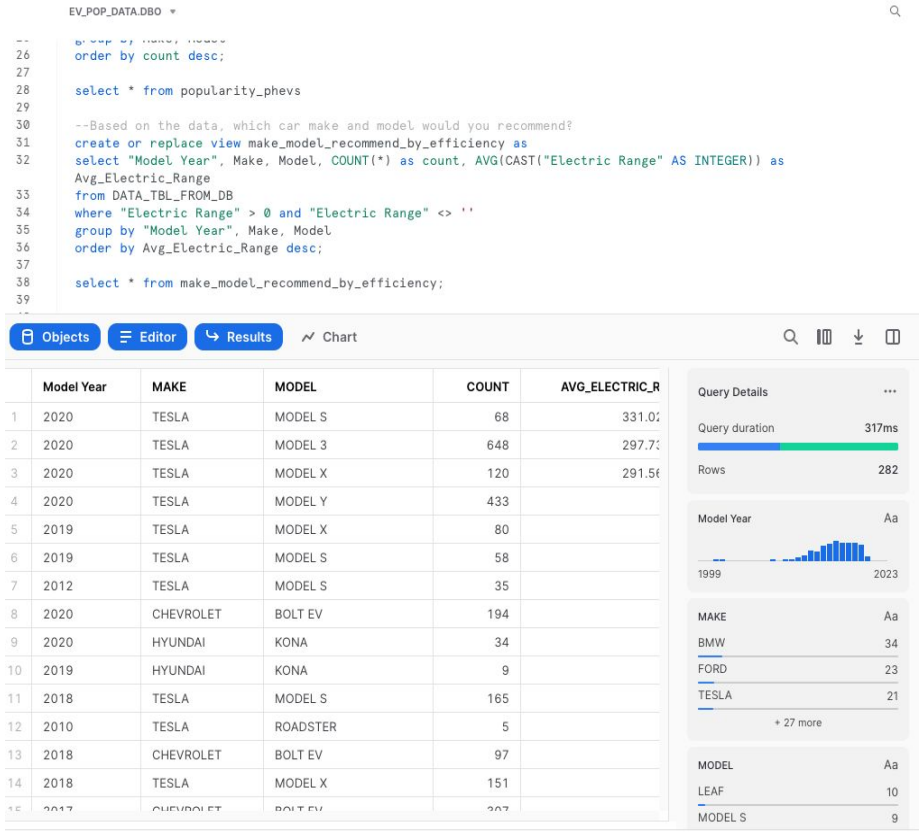
Demo

02: Snowflake/Snowpark

Having the same extracted data loaded into Snowflake tables, generated view for the same insight analysis queries.

Created a Snowpipe to process the same JSON data ingestion to compare to Databricks Job Workflow.

Used Snowpark API to perform the same Dataframe



Snowpark vs Databricks Spark

Snowpark and Databricks Spark are both powerful tools for data processing and analysis. However, there are several benefits to using Snowpark over Databricks Spark:

Performance: Snowpark is designed to optimize performance for Snowflake's data warehouse environment. This means that Snowpark jobs can run faster and more efficiently than Databricks Spark jobs when working with Snowflake.

Integration with Snowflake: Snowpark is tightly integrated with Snowflake, which makes it easy to write Spark code that can interact with Snowflake data warehouses. This means that you can use Snowpark to load data into Snowflake, extract data from Snowflake, and run complex analytical queries on Snowflake data.

Flexibility: Snowpark is a more flexible tool than Databricks Spark, as it can be used in a variety of environments, including on-premises data centers and public clouds. This means that you can use Snowpark to build data processing pipelines that meet your specific business needs, regardless of your organization's IT infrastructure.

Overall, Snowpark offers several benefits over Databricks Spark when working with Snowflake data warehouses. If your organization is already using Snowflake and needs a high-performance, flexible tool for data processing and analysis, Snowpark is definitely worth considering.

Thank you.