The Data Scientist has developed an unsupervised model to help analyse traffic flow and conversions into our online shop.

The model is expected to analyse web-traffic data (sourced from the data warehouse) and output some model results daily.
The model is delivered to you as a python module.
Model output from the main function is expected to be a dataframe.

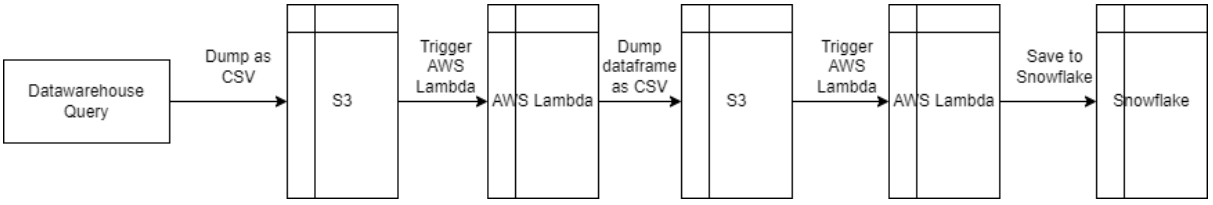
You have been tasked to deploy the solution, and allow BI to develop Tableau dashboards based on all the data science model's output for downstream business users.
Discuss how you would implement this system, and provide a simple systems diagram.

Answer:

This is clear that the data source is a data-warehouse with an expected T-1 latency.

Scenario 1:

We are not clear of the data volume. Suppose the data volume is small enough:



So we will dump the query results from the csv into S3 to trigger a aws lambda function to run the DS python module. Then the returned dataframe will be saved as CSV and then trigger another function to save to snowflake.

Scenario 2:

We are not clear of the data volume. Suppose the data volume is too large to process:

We adopt strategy number 1 which is an ETL strategy. We can create a Spark cluster on Amazon EMR. Using Spark RDD Pipe, we can read the datawarehouse data through a Spark Query and then create a Spark RDD. For each set of the RDD, we write a Spark RDD pipe to trigger the python module's core DS logic and return the serialized message back to spark. Spark can then write it directly to snowflake via snowflake spark connector.

Another strategy to explore could be ELT strategy. Since the final output will end up in some form of a data warehouse like Snowflake, we can do ENTIRE ELT in snowflake environment via Snowpark API.

https://docs.snowflake.com/en/developer-guide/snowpark/index