Build a Streaming Pipeline with DBT, Snowflake and Kinesis

Business Overview:

DBT (Data Build Tool) is a popular open-source command-line tool used in the data transformation and modeling process. It is widely used in data warehousing and business intelligence projects for building data pipelines, transforming data, and building data models.

DBT is designed to help data analysts and engineers create, maintain, and test modular SQL scripts that can be run sequentially to extract data from various sources, transform it into a structured format, and load it into a target system. It uses the concept of "models" to organize and define SQL queries that transform data from source to target. You can define dependencies between models and run tests to ensure the accuracy of the transformed data.

dbt Cloud refers to a cloud-based data transformation platform provided by dbt (data build tool), a popular open-source tool for transforming and modeling data in data warehouses. Dbt Cloud is designed to make it easier for data teams to collaborate, schedule, and monitor data transformations in a user-friendly and scalable environment.

Key features of dbt Cloud include:

- **Collaboration:** Dbt Cloud provides a platform for data analysts, engineers, and other team members to collaborate on data transformation projects. It often includes version control, code review, and documentation features to improve teamwork.
- **Scheduled Runs:** Users can schedule when data transformations should be executed. This helps to ensure that the data in the data warehouse is always up-to-date and reflects the latest changes from the source systems.
- **Monitoring and Alerting:** Dbt Cloud usually offers monitoring and alerting capabilities, allowing users to track the progress of their transformations, identify issues, and receive notifications if something goes wrong.
- **User Interface:** Dbt Cloud often comes with a user-friendly interface that makes it easier for non-technical users to create, manage, and monitor data transformations without writing complex SQL queries.
- **Deployment and Scaling:** It provides the ability to easily deploy dbt projects and scale data transformations as needed, without managing the infrastructure yourself.

- **Security:** Dbt Cloud typically offers security features to control access to data, projects, and resources.

Aim:
The main objective of this project series is to offer a complete comprehension of the Data Build Tool (DBT). This is the third part of the dbt series. In the second part of the series, we constructed an ETL pipeline utilizing various technologies such as dbt, Snowflake, and Airflow. By integrating these tools, we created a seamless flow for extracting, transforming, and loading data. Furthermore, to ensure efficient monitoring of each pipeline run, we incorporated Slack and email notifications using SNS (Simple Notification Service).

This project aims to construct a comprehensive Streaming Pipeline using the capabilities of DBT Cloud, Snowflake, and Amazon Kinesis to handle and process Stock Market Data. The project's objectives encompass various pivotal tasks, including setting up the DBT configuration within a cloud-based environment. It involves the creation of a dynamic data ingestion pipeline seamlessly integrated with a streaming data source, specifically Amazon Kinesis Firehose. The ingested data is then directed to a Raw Ingestion Layer hosted on Amazon S3, with data extraction from the yfinance library accomplished through Python code. The project also involves crafting a robust data transformation and consumption layer utilizing DBT Cloud's advanced features. The orchestration of the DBT pipeline within the DBT Cloud ecosystem ensures efficient execution, while the project's scope extends to streamlining Git merging and pipeline execution within the DBT Cloud environment. Additionally, the integration of alerting and notification functionalities provided by DBT Cloud enhances the project's overall functionality. In summary, this project strives to harmonize these technologies, resulting in a proficient Streaming Pipeline for real-time processing, transformation, and analysis of Stock Market Data.

Tech Stack
➔
Language: Python, SQL
➔
Tool: dbt Cloud
➔
Database: Snowflake
➔
Services: AWS EC2, Amazon Kinesis Firehose, AWS S3

Key Takeaways

- Understanding the DBT in detail
- Understanding the need for DBT
- Understanding the fundamentals of DBT
- Understanding the difference between dbt Cloud and dbt Core
- Understanding the dbt Cloud environment setup
- Understanding the integration of dbt Cloud with GitHub
- Connecting dbt Cloud with Snowflake database
- Understanding the Stock Market Data
- Understanding the yfinance library
- Creating AWS S3 bucket
- Creating Amazon Firehose Delivery Stream
- Testing dbt Cloud environment
- Understanding DBT tags
- Adding fact and dimension Airflow Tasks
- Creating external table and data view in Snowflake
- Performing Data Ingestion for Batch Analytics
- Building the Staging Models and environment variables in dbt Cloud environment
- Working with Macros and Multi environment deployment in dbt Cloud
- Building Models for Batch data analytics
- Understanding Data Strategy for Streaming data problem
- Creating Fact tables and Data Pipeline in dbt Cloud
- Understanding Alerting and Notification Features in dbt Cloud

Architecture Diagram: