How to deal with slowly changing dimensions using Snowflake?

What is Slowly Changing Dimensions?

The terms "facts" and "dimensions" are used in data warehousing. A fact is a piece of numerical data, such as a sale or clicks. Facts are stored in fact tables, linked to a variety of dimension tables via foreign keys that act as companion tables to the facts. Dimension Attributes are the different columns in a dimension table that provide descriptive features of the facts.
A Slowly Changing Dimension (SCD) stores and maintains both current and historical data across time in a data warehouse. It is regarded as one of the essential ETL jobs for monitoring the history of dimension records, and it has been implemented. Customer, geography, and employee are examples of such dimensions.

SCD may be approached in a variety of ways. The most popular ones are:

Type 0: This is a passive method. When the dimensions change in this approach, no particular action is taken. Some dimension data can be kept the same as when it was initially entered, while others may be replaced.

Type 1: The new data overwrites the previous data in a Type 1 SCD. As a result, the existing data is lost because it is not saved elsewhere. This is the most common sort of dimension one will encounter. To make a Type 1 SCD, one does not need to provide further information.

Type 2: The complete history of values is preserved in a Type 2 SCD. The current record is closed when the value of a particular attribute changes. With the updated data values, a new record is generated, which then becomes the current record. Each record's adequate time and expiry time are used to determine the period during which the record was active.

Type 3: For some chosen dimensions, a Type 3 SCD maintains two copies of values. The previous and current values of the chosen attribute are saved in each record. When the value of any of the chosen attributes changes, the latest value is recorded as the current value, and the previous value is saved as the old value in a new column.

In this project, we use Snowflake Datawarehouse to implement different SCDs. Snowflake offers all sorts of services to build an efficient Data warehouse with ETL capability and support for various external data partners.

Data Pipeline:

It refers to a system for moving data from one system to another. The data may or may not be transformed, and it may be processed in real-time (or streaming) instead of batches. The data pipeline is extracting or capturing data using various tools, storing raw data, cleaning, validating data, transforming data into the query-worthy format, and visualizing KPIs, including Orchestrating the above process.

Dataset Description:

In this project, we use the faker library from Python to generate records of users and store the records in CSV format with the name, including the current system time.
The data includes the following parameters:

- Customer_id
- First_name
- Last_name
- Email
- Street
- State
- Country

Tech Stack:

➔ Languages: Python3, JavaScript, SQL
➔ Services: NiFi, Amazon S3, Snowflake, Amazon EC2, Docker

NiFi

Apache NiFi is a data logistics platform that automates data transfer across systems. It gives real-time control over data transportation from any source to any destination, making it simple to handle.

Docker

Docker is a containerization platform that is available as an open-source project. It allows developers to bundle programs into containers, which are standardized executable components that combine application source code with the OS libraries and dependencies needed to run that code in any environment.

Amazon EC2

In the Amazon Web Services Cloud, the Amazon Elastic Compute Cloud (Amazon EC2) offers scalable computing capability. The user will not have to buy hardware upfront if Amazon EC2 is used. Amazon EC2 allows developers to launch multiple virtual servers based on usage, set security and networking, and manage storage.

Amazon S3

Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance. Users may save and retrieve any quantity of data using Amazon S3 at any time and from any location.

Snowflake

Snowflake is a data storage, processing, and analytics platform that blends a unique SQL query engine with a cloud-native architecture. Snowflake delivers all the features of an enterprise analytic database to the user. Snowflake components include:

- Warehouse/Virtual Warehouse
- Database and Schema
- Table
- View
- Stored procedure
- Snowpipe
- Stream
- Task

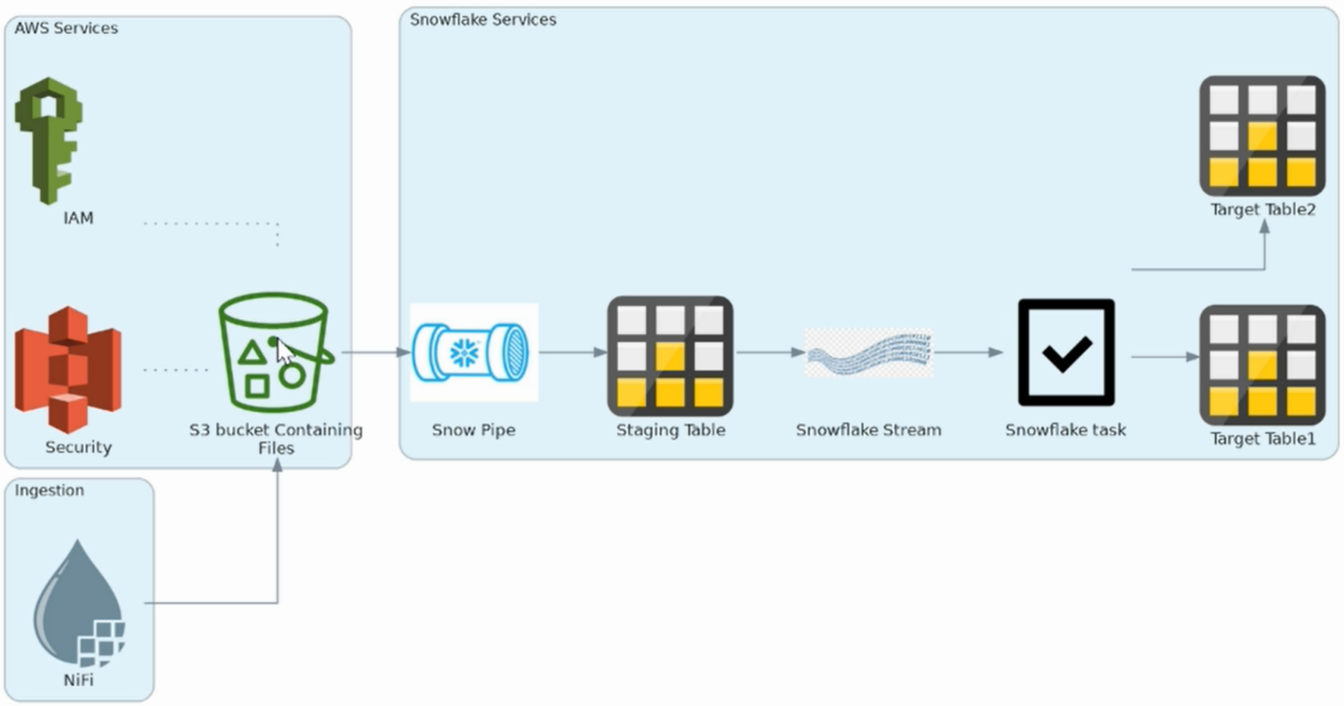
Approach:

- Test data created using faker library and saved in CSVs.
- Data is ingested using NiFi and pushed to Amazon S3.
- A Snowpipe automation tool loads new data from S3 to the staging table.
- Data manipulation language changes are recorded using Snowflake streams in the staging table to decide the operation to be performed.
- Based on the changes, Tasks and stored procedures are triggered to implement SCD Type-1 and Type-2.

Key Takeaways:

- Understanding the basics of SCD and its different types.
- Visualizing the complete Architecture of the system
- Understanding the project and how to use AWS EC2 Instance and security groups.
- Introduction to Docker.
- Docker Installation and execution.
- Usage of docker-composer and starting all tools.
- Creation of Access key.
- Creation of S3 bucket.
- Test Data preparation.
- Understanding basics of NiFi.
- Integrating NiFi with S3.
- Implementing NiFi flow setup.
- Introduction to different Snowflake components.
- Implementation of different Snowflake components.
- Implementation of SCD Type-1 and Type-2.

Architecture:



Implementation of Slowly Changing Dimensions using Snowflake services