Data Warehouse Project Report

Name: Saad Ullah

Abstract

This report describes the design, implementation, and functionality of a data warehouse project that integrates SQL data sources into AWS S3 (data lake) and Redshift (data warehouse) using Apache Airflow for ETL processes. The data is transformed into dimensional and fact tables, which are then visualized using Power BI for actionable insights.

Objective

The primary objective of this project is to design and implement a scalable, robust, and efficient data warehouse system for analytics. Key goals include:

1. Centralizing data from a SQL source to a data warehouse (Redshift) via a data lake (S3).
2. Automating ETL processes to load, transform, and store data incrementally.
3. Supporting business intelligence tasks by connecting the warehouse to Power BI dashboards.

Methodology

Architecture Overview

The project follows a multi-step architecture:

1. **Data Lake**: SQL source tables are written to AWS S3 in CSV format.
2. **Staging Zone**: Data from the S3 data lake is loaded into Redshift's staging zone.
3. **Raw Zone**: The staging zone data is incrementally loaded into the raw zone with a date column for tracking changes.
4. **Processing Zone**: Dimensional and fact tables are created using the raw zone data. SCD Type 1 is implemented for dimensions.

5. Dashboard: Data from Redshift is visualized in Power BI.

ETL Workflow

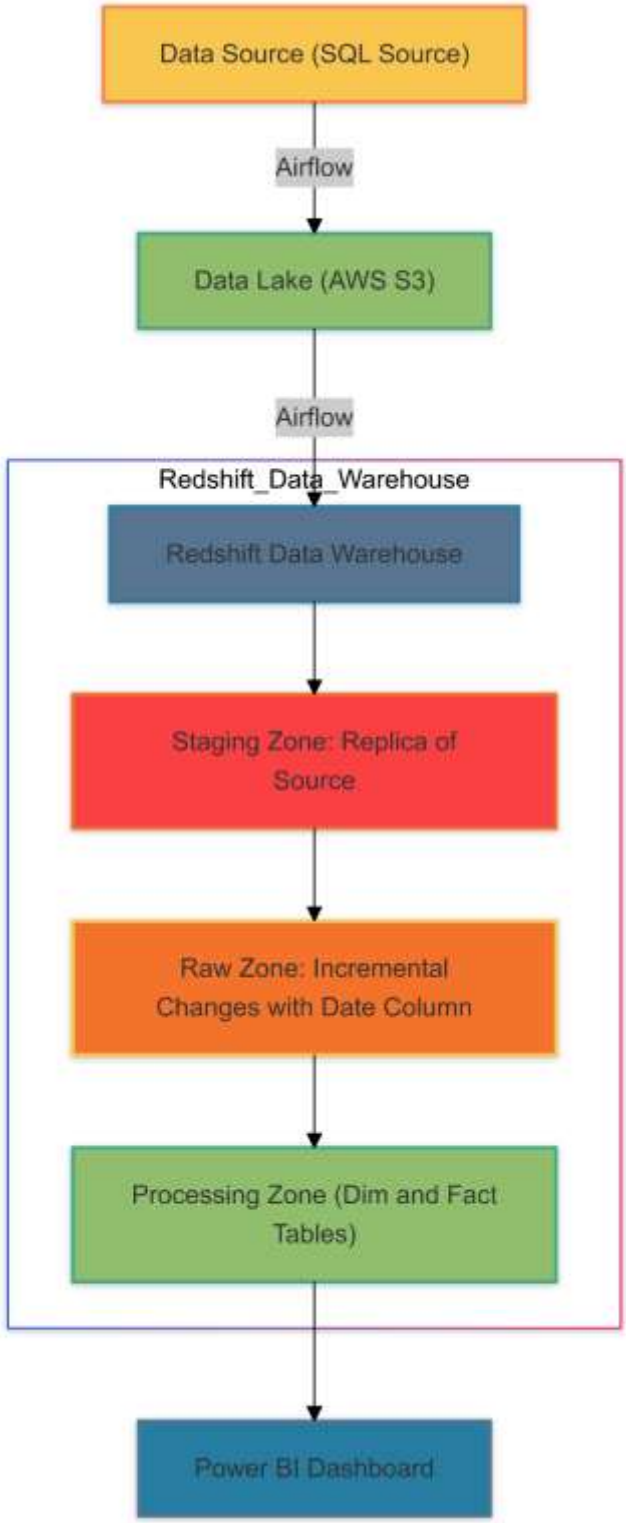
The ETL process is automated using Apache Airflow and involves the following steps:

1. Writing SQL Table to S3 (Data Lake):
    - Data from the SQL source is extracted and written as CSV files into an AWS S3 bucket using Airflow. This step ensures fast and reliable access to source data.
2. S3 to Redshift Staging Zone:
    - The CSV files in the S3 data lake are loaded into the staging zone of Redshift using Airflow.
3. Staging Zone to Raw Zone:
    - Incremental data from the staging zone is loaded into the raw zone. This involves inserting new records and updating existing ones while adding a date column to track changes.
4. Raw Zone to Processing Zone:
    - Data from the raw zone is transformed into dimension and fact tables. For dimension tables, Slowly Changing Dimension (SCD) Type 1 is applied to maintain the most current data.
5. Visualization in Power BI:
    - Redshift is connected to Power BI as a data source to create interactive dashboards.

Data Model

- Dimensions:
    - SCD Type 1 implementation.
    - Examples: Customer, Product, and Region dimensions.
- Fact Tables:
    - Fact tables include:

- Fact_Inventory_Transaction
- Fact_OrderDetails



Future Work

1. Optimization:
    - Implement partitioning and indexing in Redshift for faster query performance.
    - Evaluate column compression and distribution styles for large datasets.
2. Additional Data Sources:
    - Integrate data from NoSQL and API-based sources into the data warehouse.
3. Advanced ETL Processes:
    - Introduce delta lake features for real-time processing.
    - Explore using AWS Glue for serverless ETL workflows.
4. Enhanced Analytics:
    - Incorporate predictive analytics and machine learning models into the Power BI dashboards.