[**1. Introduction 2**](#_wlekg7jd0xl0)

[**2. Objectives 2**](#_647z8xauw65f)

[2.1 Technical Objectives 2](#_kdekso9cquq)

[2.2 Business Objectives 2](#_ze5u5z3xi0di)

[**3. Methodology 2**](#_lwjesbv1ujky)

[3.1 Bronze Layer Methodology (Raw Ingestion) 2](#_9w9gw9211tc7)

[3.2 Silver Layer Methodology (Data Cleaning & Standardization) 2](#_1gpy5c5t5ur5)

[3.3 Gold Layer Methodology (Business Modeling & KPIs) 3](#_quj6pjlki53w)

[3.4 Deployment Methodology (AWS S3 Delivery) 3](#_lg0hk51z2yf3)

[**4. Architecture Overview 3**](#_r83zpa245l8e)

[4.1 Bronze Layer: 3](#_thwh8d4fxlq9)

[4.2 Silver Layer 3](#_um3m7f2vqcaz)

[4.3 Gold Layer: 3](#_a83a9j8dq37u)

[**5. Implementation 4**](#_rbf1bl9z4x2d)

[5.1 Environment Setup 4](#_gzb4zwm9sc8s)

[5.2 Bronze Layer Implementation 4](#_irya6abbpll7)

[5.3 Silver Layer Implementation 4](#_czqywaxsnvm0)

[5.4 Gold Layer Implementation 4](#_cqhm8f88ggvc)

[5.5 AWS S3 Deployment 4](#_ia2ok9k48o9n)

[**6. Results 4**](#_8ub9wstjnbsf)

[6.1 Technical Results 4](#_wqw7l36hpecn)

[6.2 Business Results 4](#_sps5rqkpj116)

[**7. Conclusion 5**](#_q7bxj3i5pq3t)

# 

# 1. Introduction

This document describes the design and implementation of a complete data pipeline built using the Medallion Architecture, a multi-layered data engineering approach that organizes data into Bronze, Silver, and Gold layers. The solution leverages Python, Snowflake, dbt, and AWS S3, ensuring a structured, scalable, and production-ready data platform.

# 2. Objectives

### 2.1 Technical Objectives

* Establish an end-to-end data pipeline based on the Medallion Architecture.
* Automate ingestion of datasets into Snowflake using Python.
* Programmatically set up Snowflake warehouse, database, schemas, and tables.
* Structure all raw data in the Bronze layer without transformation.
* Build reliable and reusable transformations in dbt for the Silver and Gold layers.
* Produce clean and analytics-ready datasets with validated quality.
* Generate business KPIs, aggregated views, and data marts in the Gold layer.
* Deploy final curated outputs to AWS S3 for wider consumption.

### 2.2 Business Objectives

* Improve data accuracy and trust across reporting systems.
* Ensure consistent, reproducible transformations and lineage.
* Enable rapid delivery of high-quality data for analytics teams.
* Reduce manual processing time and operational overhead.
* Build a scalable foundation for future datasets and business logic.

# 3. Methodology

### 3.1 Bronze Layer Methodology (Raw Ingestion)

Raw data is ingested exactly as received, without modifications. Python handles file ingestion and loading into Snowflake. Required Snowflake resources such as warehouse, schemas, and tables are created dynamically. Raw files are stored in Snowflake stages before being loaded into Bronze tables.

### 3.2 Silver Layer Methodology (Data Cleaning & Standardization)

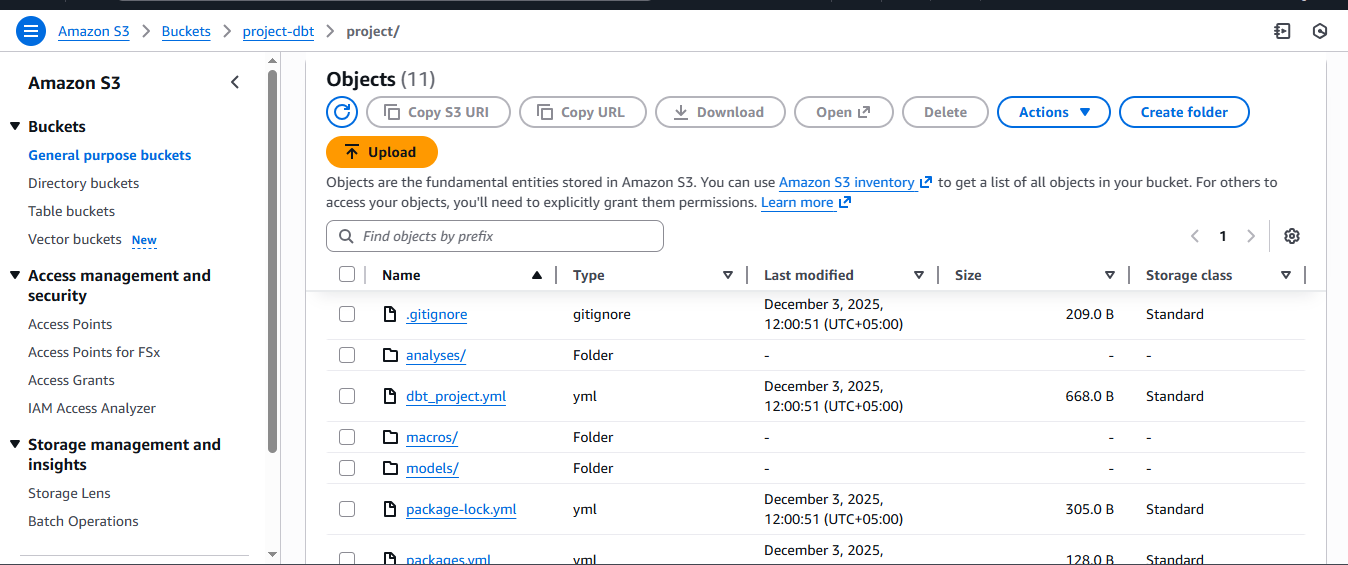
dbt is used to clean, standardize, and validate data. This layer applies conventions, fixes data types, removes duplicates, and ensures consistent formatting. dbt tests provide data quality checks.

### 3.3 Gold Layer Methodology (Business Modeling & KPIs)

This layer provides business-ready outputs. dbt builds aggregated views, KPIs, and domain-specific marts suitable for dashboards and analytics.

### 3.4 Deployment Methodology (AWS S3 Delivery)

Final curated outputs from the Gold layer are exported to AWS S3 for consumption by downstream systems, BI tools, or external stakeholders.



# 4. Architecture Overview

### 4.1 Bronze Layer:

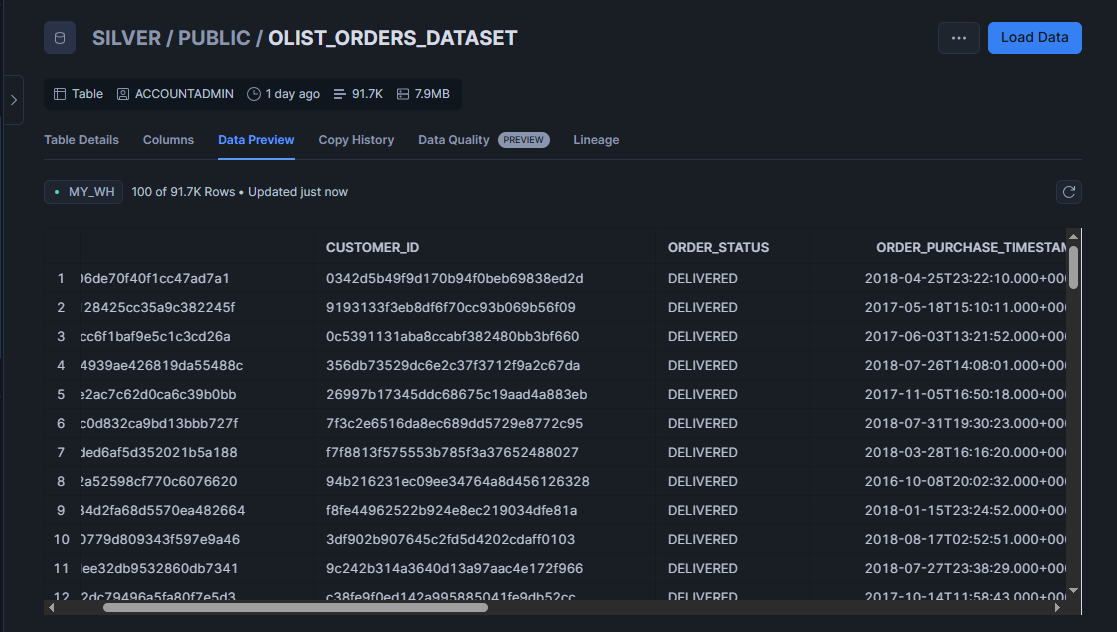
The Bronze Layer serves as the foundational storage zone of the Medallion Architecture, capturing data exactly as it arrives from source systems. This layer ensures full data fidelity, retaining all fields, formats, and records including inconsistencies or irregularities. By storing data in its original state, the Bronze Layer supports traceability, auditability, and reproducibility of downstream analytics.

In this project, the Bronze Layer stores raw files ingested through automated Python scripts that load source datasets directly into Snowflake without modification. These datasets may include CSV, JSON, Parquet, or other file formats. No cleaning, transformation, or filtering is performed at this stage. The Bronze Layer provides a trustworthy baseline for subsequent processing steps, ensuring all transformations are transparent and reversible.

### 4.2 Silver Layer

The **Silver Layer** represents the standardized and enhanced version of the Bronze data. This layer focuses on refining datasets into consistent, analysis-ready tables. Transformations applied in the Silver Layer include **data cleaning, type casting, deduplication, error handling, normalization, and alignment to business naming conventions**.

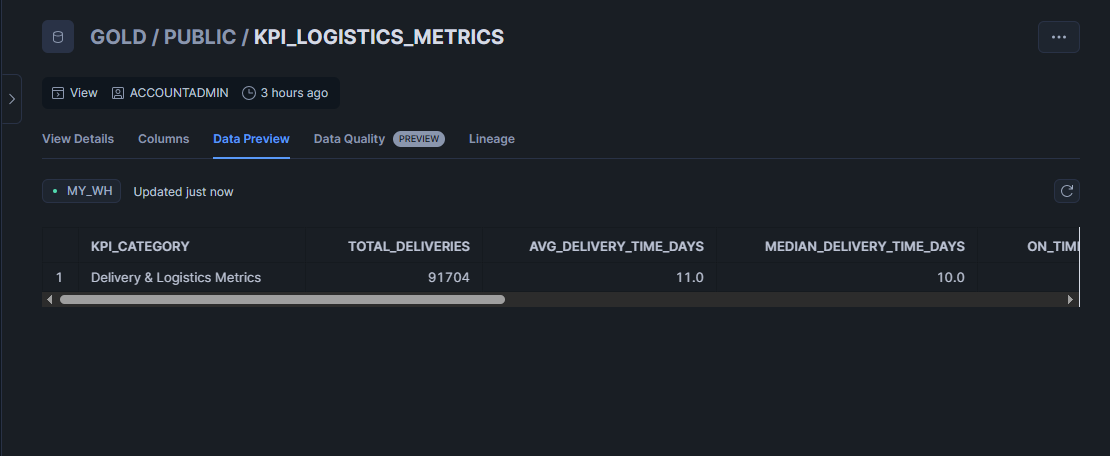
In this architecture, dbt (Data Build Tool) is used to manage all transformations applied to Silver datasets. dbt models ensure repeatability, version control, and clear lineage, enabling data engineers and analysts to understand how each field was produced. Data in the Silver Layer is accurate, validated, and formatted to support analytical requirements. This layer eliminates inconsistencies present in the raw data, making it suitable for more advanced business consumption in the Gold Layer.



### 4.3 Gold Layer:

The **Gold Layer** contains highly curated, business-ready datasets designed specifically for reporting, dashboarding, and decision-making. This layer includes **aggregated tables, business metrics, KPIs, analytical summaries, and domain-specific data marts**.

Building on the structured and validated data from the Silver Layer, the Gold Layer applies transformations aligned with business logic and strategic goals. Examples include customer acquisition metrics, revenue trends, sales conversions, performance insights, and time-based analytics. The Gold Layer is optimized for ease of use by analysts, BI tools, and executives. It represents the final, refined view of the organization’s data, ready for consumption across the enterprise.



# 5. Implementation

### 5.1 Environment Setup

Includes Snowflake configuration for warehouse, database, schemas, stages, and dbt project setup.

### 5.2 Bronze Layer Implementation

Python automates loading of raw files into Snowflake Bronze tables.

### 5.3 Silver Layer Implementation

dbt cleans and standardizes Bronze data into analytical structures, backed by tests ensuring reliability.

### 5.4 Gold Layer Implementation

dbt generates aggregated tables, business KPIs, and analytical models optimized for reporting.

### 5.5 AWS S3 Deployment

Gold layer outputs are exported to AWS S3 where they are organized for analytics and wider access.

# 6. Results

### 6.1 Technical Results

* The pipeline automates data loading and Snowflake setup, reducing manual steps and ensuring consistent ingestion.
* Silver Layer datasets are standardized, deduplicated, and quality-checked, providing reliable inputs for analysis.
* Gold Layer delivers curated KPIs and aggregated views that are ready for direct use in reporting and dashboards.
* The overall architecture supports growing data volumes, additional sources, and more complex transformations with ease.

### 6.2 Business Results

* Consistent cleaning and validation increase confidence in the data used across the business.
* Ready-to-use Gold datasets reduce time spent preparing data, enabling quicker insights.
* Automation replaces repetitive tasks, freeing teams for higher-value work.
* The architecture provides a strong base for expanding data sources, new KPIs, and advanced analytics in the future.

# 7. Conclusion

This project successfully delivers a scalable and modern data pipeline using the Medallion Architecture, integrating Python, Snowflake, dbt, and AWS S3 to ensure a streamlined, reliable, and high-quality data flow. The layered Bronze, Silver, and Gold structure strengthens data governance and enables consistent transformations from raw ingestion to business-ready outputs.

Automated ingestion and resource creation in Snowflake reduce manual effort, while dbt provides transparent, repeatable, and well-governed transformations. The final deployment to AWS S3 ensures broad accessibility and supports downstream analytics needs.

Overall, the solution enhances data accuracy, improves operational efficiency, and creates a strong foundation for future analytical and machine learning initiatives. It equips the organization with trustworthy, timely data to drive informed and effective decision-making.