InsightStream: Cloud Data Warehouse & Analytics

Building a Modern ETL Pipeline with Airflow, Snowflake Internal Storage, dbt, and Power BI

Team Lead: Youssef Ahmed Mohamed Alkamashany

Abdelrahman Adel Abdelmola Abu Taleb

Abdullah Mohamed Ahmed

Omar Abdelgawad Mohamed

Project Overview

Project Proposal

Overview of the project, objectives, and scope for building a robust, modern ELT pipeline for data analytics using Snowflake's native capabilities and industry-standard transformation tools.

Business Objectives

- Establish a comprehensive ELT pipeline for ingesting diverse data sources directly into a cloud data warehouse.
- Enable advanced analytics and business intelligence through **Snowflake** and **Power BI** for data-driven decision making.
- Automate data workflows using **Airflow** and **dbt** to ensure data freshness, reliability, and scalability.

Technical Objectives

- **Data Ingestion:** Ingest API data directly into **Snowflake Internal Storage** (Internal Stages) for raw data storage, eliminating the need for external S3 buckets.
- **Data Transformation & Warehousing:** Utilize **dbt (data build tool)** to transform and model data within Snowflake, ensuring modularity and automated documentation.
- **Workflow Orchestration:** Implement **Airflow DAGs** to manage and automate the end-to-end ELT process.
- **Business Intelligence:** Connect Snowflake to **Power BI** for creating powerful, interactive dashboards and enterprise-grade visualizations.
- **Monitoring & Automation:** Set up monitoring for Airflow DAGs and dbt runs, ensuring data quality and pipeline health.
- **Documentation:** Provide comprehensive documentation for the architecture, dbt models, and Power BI reports.

Project Documentation Guidelines

Item	Student Deadline	Graduate Deadline
Project Planning & Management	2/20/2026	2/3/2026
Literature Review	4/20/2026	3/1/2026
Requirements Gathering	4/20/2026	3/1/2026
System Analysis & Design	5/1/2026	4/3/2026
Implementation (Source Code & Execution)	7/10/2026	5/17/2026
Final Presentation & Testing & Reports	7/17/2026	5/24/2026

LITERATURE REVIEW

Researching Best Practices for Modern ELT Pipelines

- Techniques for ingesting data from various APIs directly into **Snowflake Internal Stages**, optimizing for cost and security.
- Best practices for data modeling using **dbt**, focusing on Modular SQL, Macros, and automated testing frameworks.
- Effective use of **Apache Airflow** for orchestrating complex dbt jobs and multi-step data ingestion tasks.
- Designing compelling and interactive dashboards using **Power BI** for high-impact business intelligence.
- Strategies for ensuring **Data Quality and Governance** across the pipeline using dbt tests and Snowflake features.

Requirements Gathering Details

Functional

- Integration of diverse **API data sources**.
- Extraction of specific data points for business logic.
- Complex transformations using **dbt models**.
- Interactive analytical outputs in **Power BI**.

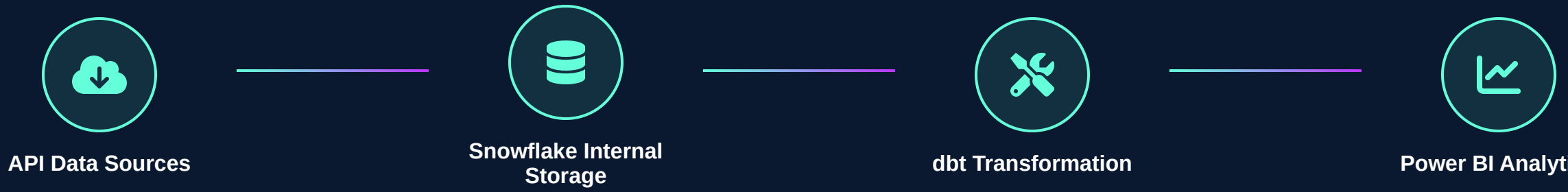
Non-Functional

- Low **data latency** and high throughput.
- Robust **data security** within Snowflake.
- Cost optimization for cloud resources.
- System scalability and disaster recovery.

Data

- Identification of all **API endpoints**.
- Support for multiple data formats (JSON, CSV).
- Handling varying **data volumes**.
- Defined frequency of updates for all sources.

System Analysis & Design



Orchestrated by Apache Airflow

System Architecture

A modern ELT approach where data is ingested directly into **Snowflake Internal Stages**. This eliminates external dependencies and reduces latency for raw data availability.

Data Model (dbt)

Utilizing **dbt** to manage the transformation layers: **Staging** (cleaning), **Intermediate** (joining), and **Marts** (business logic), all optimized for analytical performance.

Implementation: Tools & Technologies

Python

Core language for API interaction, data processing, and Airflow DAG development.

Snowflake

Cloud Data Warehouse utilizing **Internal Storage** for efficient raw data staging.

dbt

Data Build Tool for SQL-based transformations, version control, and automated documentation.

Apache Airflow

Workflow orchestration for scheduling and monitoring the end-to-end ELT pipeline.

Power BI

Enterprise BI tool for creating interactive dashboards and data visualizations.

Various APIs

External data sources integrated into the pipeline for real-time data ingestion.

Implementation Plan

01 **Cloud Environment Setup**

Configure Snowflake accounts, internal stages, and the Apache Airflow environment.

02 **Data Ingestion to Snowflake**

Develop Python scripts to fetch data from APIs and upload to Snowflake stages.

03 **dbt Project Setup**

Initialize dbt, configure profiles, and define sources.

04 **Transformation Layer**

Write dbt models to transform raw data into analytical tables.

05 **Airflow DAG Development**

Design DAGs to orchestrate ingestion scripts, dbt runs and tests.

06 **Power BI Dashboard Creation**

Develop interactive dashboards connected to Snowflake.

Final Presentation, Testing & Reports

Testing Procedures

Unit Testing

Test individual Python scripts and **dbt models** for logic accuracy.

Data Quality Testing

Use **dbt tests** (unique, not_null, relationships) to ensure data integrity.

Integration Testing

Verify end-to-end data flow from source APIs through Snowflake/dbt into Power BI.

Performance Testing

Evaluate **dbt run times** and Power BI query performance for scalability.

Presentation & Reports

Final Presentation

Showcase project architecture, **dbt lineage**, and key analytical insights.

Technical Report

Comprehensive documentation including **auto-generated dbt docs** and Power BI designs.

Live Demonstration

A live walkthrough showing data flowing from source to interactive dashboards.

Presentation & Reports Details

Final Presentation

- Showcase the project's **modern ELT architecture**.
- Visualize **dbt lineage graphs** for data transparency.
- Present key analytical insights derived from **Power BI**.
- Discuss challenges overcome during implementation.

Technical Report

- Comprehensive documentation of **design choices**.
- Inclusion of **auto-generated dbt documentation**.
- Detailed **Power BI dashboard designs** and logic.
- Cloud service configurations and security protocols.

Live Demonstration: End-to-End Data Flow

This project aims to deliver a robust, scalable, and modern ELT solution for data analytics, leveraging the power of **Snowflake**, **dbt**, **Airflow**, and **Power BI** to provide data-driven insights and support informed decision-making.

THANK YOU!

InsightStream Project Team