Propuesta y Arquitectura del Proyecto BlueMar

Introducción

Este documento describe la propuesta de arquitectura para optimizar y automatizar el manejo y análisis de datos en la empresa **BlueMar**, utilizando herramientas modernas de Big Data y Data Analytics. El objetivo es construir una arquitectura escalable que facilite la ingesta, limpieza, transformación, análisis y visualización de los datos, integrando procesos orquestados y tecnologías robustas.

Objetivos

1. Automatizar la ingesta y limpieza de datos provenientes de **Google Sheets** (otras fuentes de datos por definir).
2. Centralizar y almacenar datos en un **Data Warehouse en la nube** (Snowflake).
3. Implementar transformaciones escalables y modulares utilizando **dbt**.
4. Garantizar la calidad de los datos mediante verificaciones automáticas con **Soda**.
5. Desarrollar dashboards interactivos para análisis y visualización utilizando **Streamlit** directamente en Snowflake.
6. Escalar la arquitectura para integrar nuevas fuentes de datos en el futuro.

Herramientas y Tecnologías

Componente	Herramienta/Servicio	Propósito
Orquestación	Airflow (Cosmos, Astronomer CLI)	Automatización y orquestación de tareas en el pipeline.
Almacenamiento en la Nube	S3	Almacenamiento de datos en crudos.

Componente	Herramienta/Servicio	Propósito
Data Warehouse	Snowflake	Almacenamiento centralizado de datos transformados.
Transformación de Datos	dbt	Transformaciones modulares y escalables.
Calidad de Datos	Soda	Verificaciones automáticas de calidad y limpieza de datos.
Visualización	Streamlit	Dashboards dinámicos e interactivos.

Arquitectura Propuesta

1. Ingesta de Datos

- Los datos se obtienen de **Google Sheets** mediante Airflow y se almacenan como crudos en un bucket de **S3**.

2. Almacenamiento Inicial en Snowflake

- Los datos crudos se cargan desde **S3** hacia una tabla en la capa **Raw Layer** de Snowflake.

3. Calidad de Datos

- **Soda** verifica y valida la calidad de los datos en la capa Raw Layer:
  - Identifica valores faltantes, inconsistencias y otros problemas.
  - Documenta métricas de calidad para asegurar confiabilidad.

4. Transformación

- **dbt** realiza las transformaciones de los datos en Snowflake:
  - Limpia y estandariza formatos.
  - Integra datos de múltiples fuentes.
  - Genera modelos analíticos listos para consumo.

5. Análisis y Visualización

- Los datos transformados se utilizan para construir dashboards interactivos en **Streamlit**:
  - Los dashboards están conectados directamente a Snowflake.
  - Permiten a los usuarios explorar métricas clave y patrones en tiempo real.

Diagrama de Arquitectura

Mockup del Diagrama