End-to-End ETL Pipeline with Python, Airflow, Spark, Docker, S3, Snowflake & Google Looker Studio

🔹 "Welcome to this project on building an end-to-end ETL pipeline using **Python, Apache Airflow, Spark, Docker, S3, Snowflake, and Google Looker Studio**."

🔹 "In this project, I'll walk you through the complete process of extracting, transforming, and visualizing data from **YouTube API**, all automated with **Apache Airflow** and deployed in a **Dockerized environment**."

🛠 Step 1: Data Extraction from YouTube API

📌 "The process starts with extracting **trending video data** from YouTube using the **YouTube API**. This includes information like video titles, views, likes, comments, and categories."

📌 "Apache Airflow is used to schedule and orchestrate this process. A DAG (Directed Acyclic Graph) is set up to **automate the API calls at regular intervals**."

🔹 **Technologies Used:** Python, YouTube API, Apache Airflow

🔹 **Key Output:** Raw JSON data stored in an **S3 bucket** for further processing.

⚙ Step 2: Data Processing with Spark

📌 "Once the raw data is available in S3, the next step is **data transformation** using **Apache Spark**. Spark processes the data to handle missing values, filter out unnecessary information, and prepare structured datasets for analysis."

📌 "This ensures that we have **clean and well-structured data** ready for loading into **Snowflake**."

🔹 **Technologies Used:** Apache Spark, PySpark, AWS S3

🔹 **Key Output:** Processed data in a structured format, ready for Snowflake.

🎁 Step 3: Storing Data in Snowflake

📌 "The cleaned data is then **loaded into Snowflake**, a powerful cloud-based data warehouse, for efficient storage and querying."

📌 "Using **Snowflake's COPY command**, we efficiently move data from S3 to a structured Snowflake table, where it can be accessed for analysis."

🔹 **Technologies Used:** Snowflake, Snowflake Connector for Python, AWS S3

🔹 **Key Output:** Data stored in Snowflake tables, ready for reporting.

📊 **Step 4: Data Visualization with Google Looker Studio**

📌 "With the processed data now in Snowflake, we connect **Google Looker Studio** to Snowflake to create **interactive dashboards**."

📌 "This allows us to visualize insights such as **trending videos, engagement metrics, and audience interactions**, providing real-time analytics on YouTube content performance."

🔹 **Technologies Used:** Google Looker Studio, Snowflake

🔹 **Key Output: Real-time dashboards** with meaningful insights.

🛠 **Step 5: Automating & Deploying the ETL Pipeline with Docker & Airflow**

📌 "To ensure the entire ETL process runs **seamlessly**, we have **Dockerized the Apache Airflow setup**, making it portable and scalable."

📌 "Airflow DAGs are scheduled to **trigger Spark jobs**, manage data movement to Snowflake, and refresh Looker Studio dashboards automatically."

🔹 **Technologies Used:** Docker, Apache Airflow, Python

🔹 **Key Benefit: Fully automated and scalable ETL pipeline** with minimal manual intervention.

🚀 **Final Thoughts & Benefits**

✅ **End-to-end automation** of the YouTube data pipeline.

✅ **Scalable & cloud-based architecture** with Snowflake & S3.

✅ **Real-time insights** via Google Looker Studio.

✅ **Fully containerized deployment** using Docker & Apache Airflow.

🔹 "This setup enables businesses and content creators to analyze YouTube trends effectively, make data-driven decisions, and optimize content strategies."