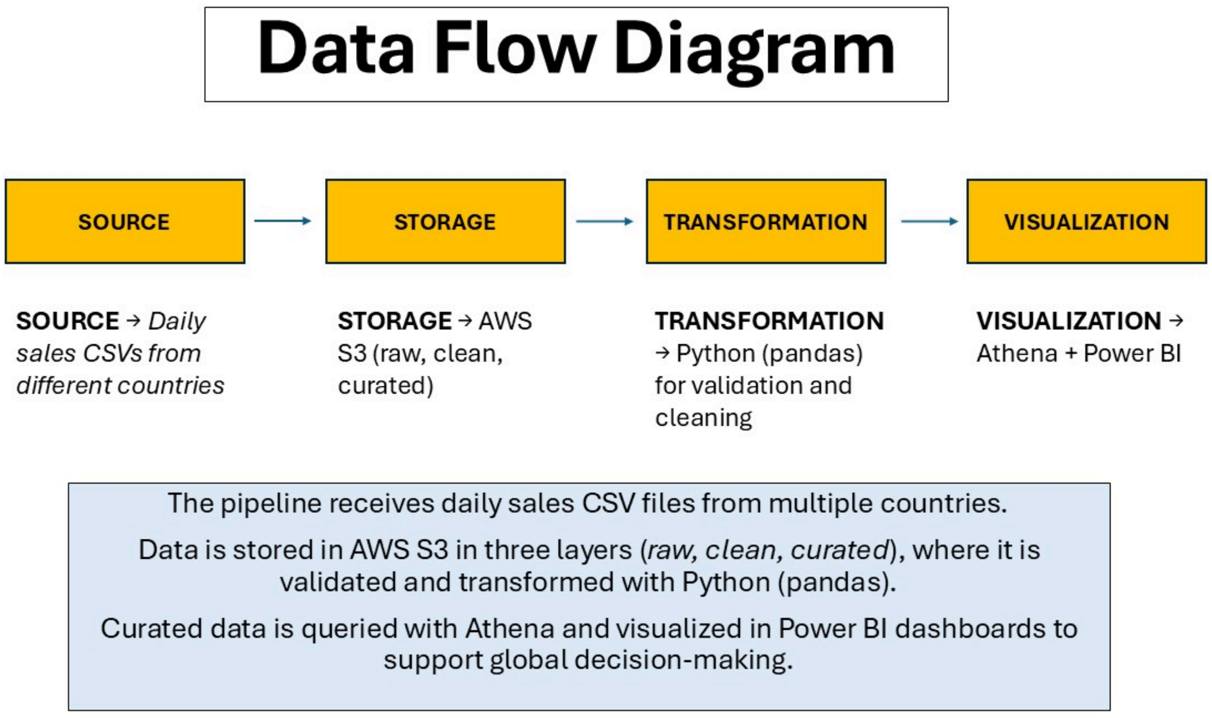
Week 1

Data Flow



Data Flow Diagram

SOURCE → *Daily sales CSVs from different countries*

STORAGE → AWS S3 (raw, clean, curated)

TRANSFORMATION → Python (pandas) for validation and cleaning

VISUALIZATION → Athena + Power BI

The pipeline receives daily sales CSV files from multiple countries.

Data is stored in AWS S3 in three layers (*raw, clean, curated*), where it is validated and transformed with Python (pandas).

Curated data is queried with Athena and visualized in Power BI dashboards to support global decision-making.

Glossary of Tools

Tool	Usage	Pipeline Stage	Complexity
S3 (AWS)	Cloud storage service used to keep files in buckets (raw, clean, curated).	Storage	Basic
pandas (Python)	Library to clean, transform, and analyze data in DataFrames.	Transformation	Intermediate
Athena (AWS)	Serverless SQL query engine to analyze data directly in S3.	Query / Storage	Intermediate

Power BI	Business intelligence tool to create interactive dashboards.	Visualization	Basic
Airflow	Workflow orchestrator to automate and schedule data pipelines.	Orchestration	Advanced
PostgreSQL	Open-source relational database to store structured data.	Storage	Intermediate

Roles in Data

Role	Main Tasks	Common Tools	Programming Level	Example Deliverable
Data Engineer	Build and maintain data pipelines, ensure quality and scalability.	SQL, Python, Airflow, Spark, AWS/GCP	High	Production-ready ETL pipeline
Data Analyst	Explore data, create reports and dashboards, generate insights.	SQL, Excel, Power BI, Tableau	Low–Medium	Sales dashboard in Power BI
Data Scientist	Apply statistics and ML to build predictive models.	Python (scikit-learn), R, SQL, Jupyter	High	Churn prediction model
Analytics Engineer	Design clean, versioned analytical models bridging DE and DA.	dbt, SQL, Snowflake, Looker	Medium–High	dbt models + standardized metrics

The role I am most interested in is **Data Engineer**, because I enjoy designing pipelines that ensure quality and scalability, similar to the sales pipeline with S3 and Athena. To reach this goal, I need to continue improving my **Python for ETL**, **advanced SQL**, and **cloud orchestration tools** such as Airflow or Step Functions. I also want to strengthen my knowledge in data modeling and governance to build reliable end-to-end pipelines.