Migrating On-Prem PostgreSQL Transactional Data to the Cloud with Airflow, S3, and Athena

Company Overview

ShopEase Retail Ltd. is a growing retail and e-commerce company that operates across multiple regions in West Africa. Established in 2015, the company has rapidly expanded from a single local outlet into a multi-channel retailer with both physical stores and a robust online shopping platform.

The company serves thousands of customers daily, offering products ranging from household essentials to electronics. To manage operations, ShopEase relies heavily on data stored in its on-premise PostgreSQL database, which powers:

- Sales Transactions – recording customer purchases across online and physical stores.
- Customer Profiles – storing personal and behavioral data for loyalty programs and targeted promotions.
- Inventory Management – tracking stock levels across warehouses and retail outlets.
- Payments & Invoices – processing online and in-store payment records.

As customer demand has increased, so has the complexity of ShopEase's data landscape. Executives and consultants increasingly require fast, reliable insights into sales performance, customer behavior, and supply chain efficiency to support decision-making.

However, accessing these insights from the on-premise system has become challenging, prompting the move to a cloud-based analytics solution.

Current Business Challenge

On-Premise Limitations

Maintaining PostgreSQL servers on-prem is expensive and difficult to scale.

Accessibility

Consultants cannot directly connect to the on-prem database due to security and network restrictions.

Slow Reporting

Running heavy analytical queries on the transactional database slows down day-to-day operations.

Data Growth

As data volume increases, query performance continues to degrade.

AMDARI®

Project Rationale

Migrating data to the cloud solves these issues:

1 Scalability

AWS S3 scales cheaply with growing data.

2 Accessibility

Consultants and analysts can query data in Athena without needing DB access.

3 Efficiency

Athena queries Parquet files faster than Postgres transactional queries.

4 Cost Optimization

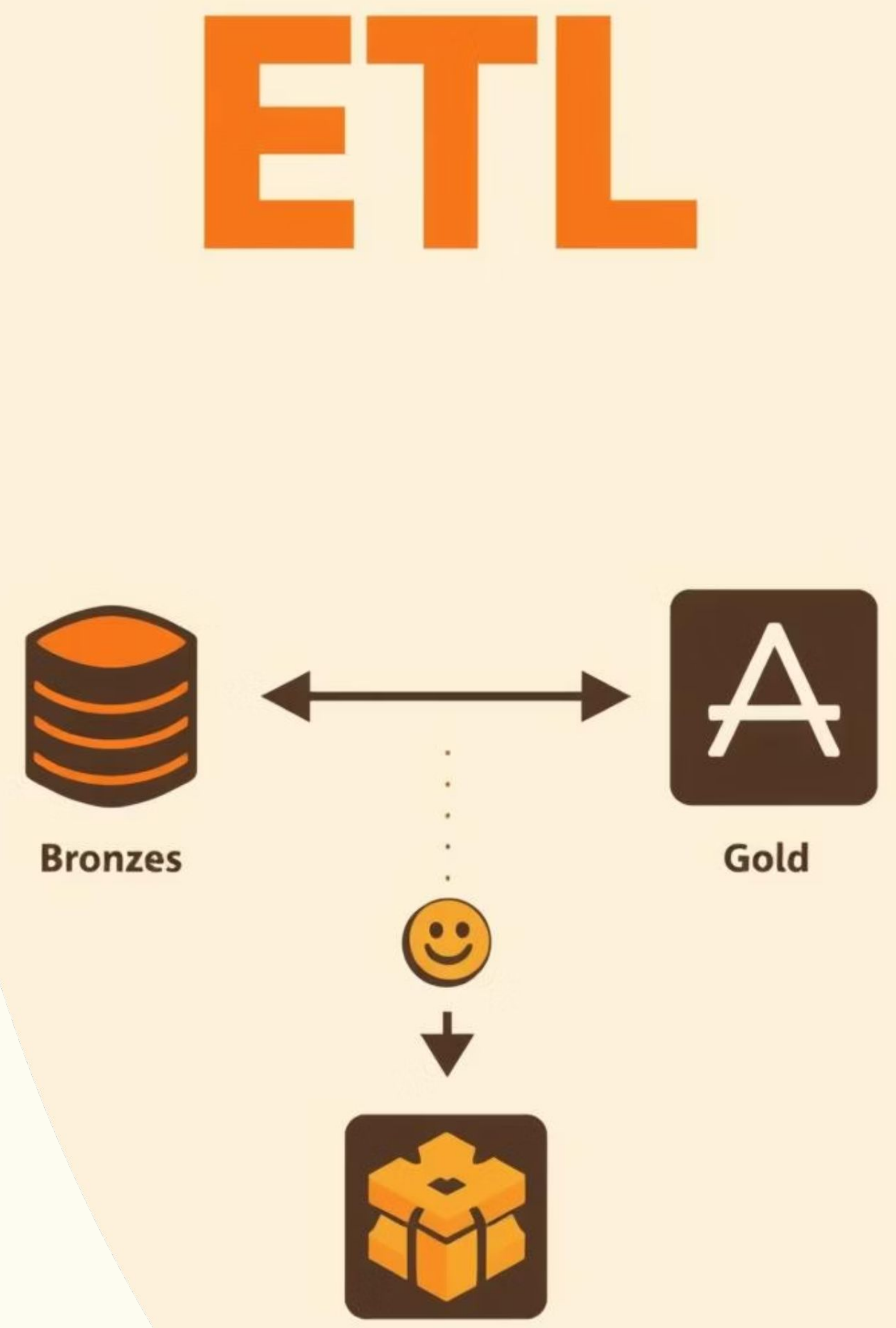
No need for constant on-prem upgrades; pay-as-you-go cloud costs.

5 Future-Proofing

Provides a foundation for dashboards, BI, and advanced analytics.

Project Objectives

- Extract data from PostgreSQL (on-prem) and land it in S3 Bronze.
- Transform raw CSVs into Parquet format in S3 Silver.
- Load business-ready datasets into S3 Gold, registered with Glue.
- Query datasets using AWS Athena for daily reporting.
- Automate the entire process using Apache Airflow on Docker.



Tech Stack

- **PostgreSQL (On-Prem):** Transactional data source.
- **Airflow (Docker):** Orchestration and scheduling of ETL jobs.
- **Python + AWS Wrangler:** Data extraction, transformation, and Glue/Athena integration.
- **AWS S3:** Data lake with Medallion layers (Bronze, Silver, Gold).
- **AWS Redshift:**
- **GitHub:** Version control and collaboration.

Medallion Architecture

Bronze Layer

- Raw data extracted directly from PostgreSQL.
- Stored in S3 as CSV (minimal changes).

Silver Layer

- Cleaned and standardized data.
- Converted to Parquet for performance.
- Partitioned by date or business key.

Gold Layer

- Business-ready datasets.
- Aggregated tables such as daily sales, top customers, product revenue.
- Directly queryable in Athena for dashboards and reports.

Project Scope & Deliverables

1

Postgres Dump File

A backup of on-prem transactions (orders, customers, products) shared with consultants.

2

Automated ETL Pipeline

DAGs in Airflow:

- Extract (Postgres → S3 Bronze)
- Transform (Bronze CSV → Silver Parquet)
- Load (Silver → Gold + Glue/Athena registration)

3

Data Lake Architecture

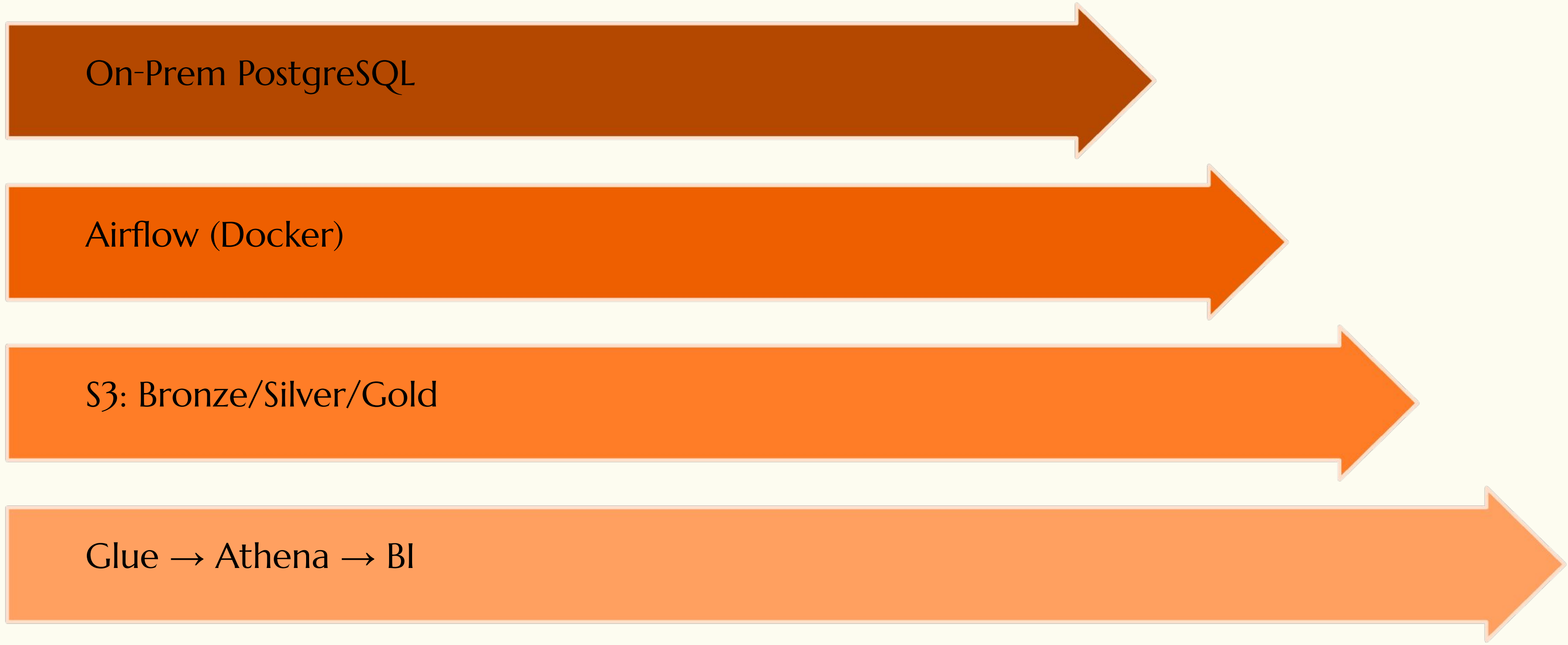
- **Bronze:** Raw PostgreSQL exports (CSV).
- **Silver:** Cleaned, Parquet, partitioned data.
- **Gold:** Business-ready tables (e.g., daily sales summary).

4

Athena Queries

- Daily sales report.
- Top 10 customers by spend.
- Revenue by product category.

Data Architecture (Simplified Flow)



On-Prem PostgreSQL

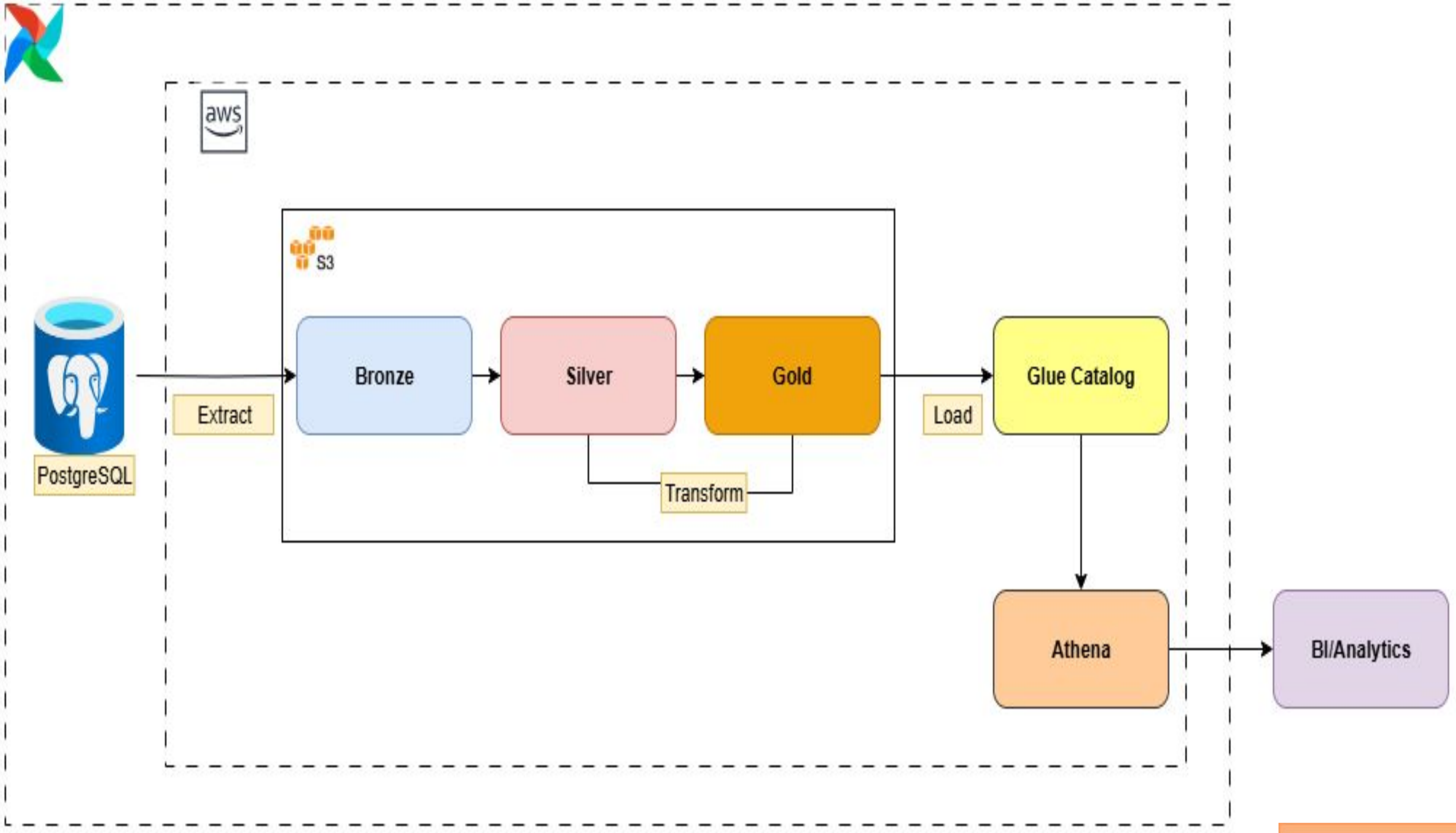
Airflow (Docker)

S3: Bronze/Silver/Gold

Glue → Athena → BI

Important Considerations

- **Partitioning Strategy** Use `order_date` or `created_at` to partition data in S3 for faster Athena queries.
- **Schema Management** AWS Wrangler automatically registers schemas in Glue for Athena.
- **Cost Awareness** Athena charges per query scanned — using Parquet and partitions reduces costs significantly.
- **Data Quality** Simple checks (row counts, duplicates, NULL values) included in the transform step.
- **Future Enhancements** Incremental loads and Change Data Capture (CDC) for near real-time pipelines.


aws
S3
PostgreSQL
Extract
Bronze
Silver
Gold
Transform
Load
Glue Catalog
Athena
BI/Analytics

Success Criteria

- Consultants can restore the Postgres dump and run the pipeline end-to-end.
- ETL jobs complete successfully and land data in Bronze, Silver, and Gold zones.
- Tables are queryable in Athena within minutes of pipeline completion.
- Business users/consultants can run reports without touching the on-prem DB.

AMDARI®