**SkillBoost Analytics Platform Documentation**

**Overview**

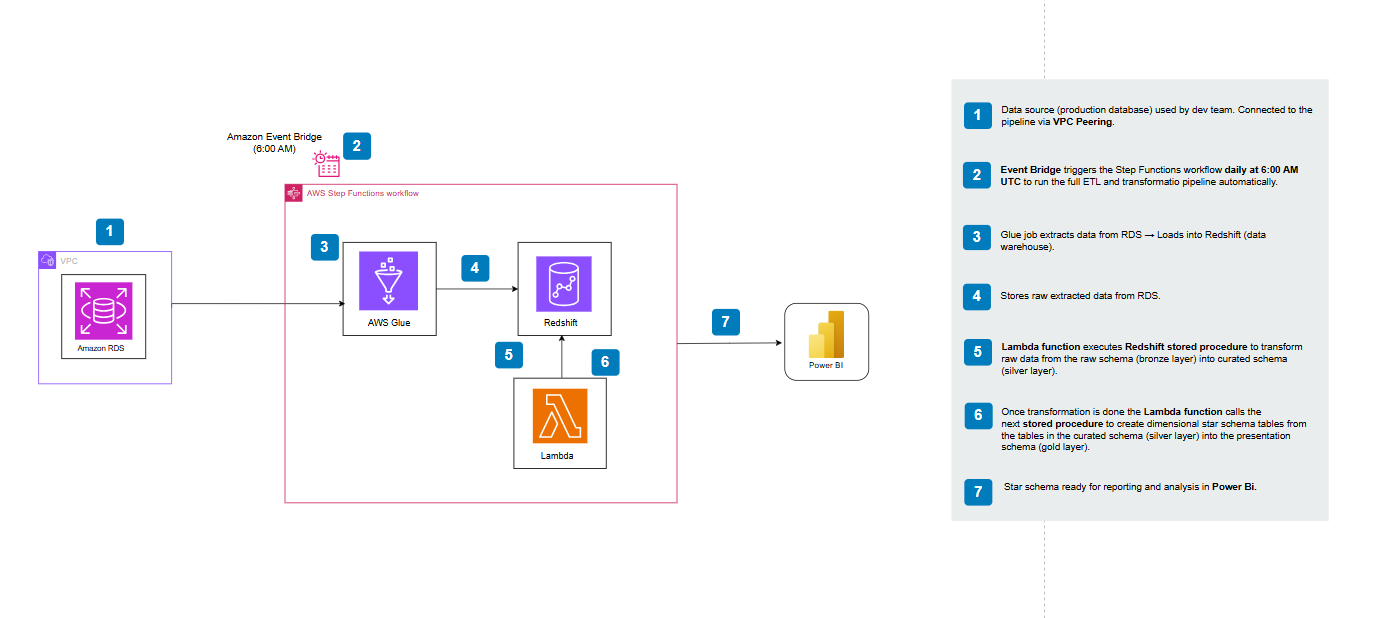
The **SkillBoost Analytics Platform** is a batch-oriented, modern data pipeline system designed to support data-driven decision-making. It ingests operational data from a PostgreSQL-based RDS source system into an Amazon Redshift data warehouse using an orchestrated ETL flow powered by AWS Glue and Step Functions. The architecture adheres to Medallion Architecture principles, providing structured layers of transformation: raw (bronze), curated (silver), and presentation (gold).

**Architecture Components**

**Core Services Used:**

* **Amazon RDS (PostgreSQL)** – Source operational database (external; Dev team AWS account)
* **AWS Glue** – Performs ETL from RDS to Redshift
* **Amazon Redshift** – Target data warehouse with three schemas:
  + raw (Bronze layer)
  + curated (Silver layer)
  + presentation (Gold layer)
* **AWS Step Functions** – Orchestrates the ETL workflow
* **Amazon EventBridge** – Triggers Step Function daily at 6:00 AM
* **AWS Lambda** – Triggers stored procedure execution in Redshift

**Architecture Design**



**Note:** The Step Function is scheduled via **Amazon EventBridge** to run **daily at 6:00 AM GMT**.

**Project Structure**

| **File** | **Description** |
| --- | --- |
| main.tf | Entry point that includes all modules |
| provider.tf | AWS provider and region setup |
| variables.tf | Input variables for Terraform |
| outputs.tf | Output endpoints (e.g., Redshift, Step Function) |
| iam.tf | IAM roles and permissions for Glue and Step Functions |
| glue.tf | Definition of AWS Glue job |
| connection.tf | JDBC connection config for RDS access |
| redshift.tf | Redshift schema and optional cluster setup |
| stepfunctions.tf | ETL orchestration logic with Step Functions |
| scripts/ | Reference ETL and SQL transformation scripts |

**Prerequisites**

* AWS CLI configured with credentials
* Terraform CLI installed
* Access to RDS instance (security group + VPC connectivity configured)
* Redshift cluster (optional if external cluster is used)

**Usage Instructions**

**1. Clone the Repository**

git clone https://github.com/charles-amali/SkillBoost-DataPipeline.git

cd SkillBoost-DataPipeline

**2. Initialize Terraform**

terraform init

**3. Apply Terraform**

terraform apply

This deploys:

* IAM roles
* Glue job
* Redshift schemas
* Step Function state machine
* Redshift cluster

**ETL Data Flow**

1. **Extract**: AWS Glue pulls data from RDS PostgreSQL
2. **Load (Bronze)**: Loads into Redshift raw schema
3. **Transform (Silver)**: Stored procedures clean and enrich data into curated schema
4. **Transform (Gold)**: Star schema into presentation schema for analytics
5. **Visualize**: Connect to Redshift via Power BI.

**Automation with EventBridge**

* An **Amazon EventBridge Rule** runs the Step Function **daily at 6:00 AM GMT**.
* This enables automated ingestion, transformation, and reporting.

**Future Enhancements**

* Add CI/CD pipeline for infrastructure and ETL scripts
* Automate Power BI dataset refresh