## 🚆 *Indian Railway Delays Data Pipeline – Project Summary*

**Author:** Sanket Aba Adhav
**Role:** Data Engineer

### 🧠 Overview

This project demonstrates an **end-to-end AWS Data Lake architecture** that processes and analyzes Indian Railway train delay data.
The pipeline is built using **AWS Glue, S3, and Athena** and follows the **Medallion Architecture**: Raw → Staging → Curated.

### 🧱 Architecture Flow

1. **Raw Layer (Bronze)** — Raw CSV file stored in S3
2. **Staging Layer (Silver)** — Cleaned, transformed data (Parquet)
3. **Curated Layer (Gold)** — Aggregated analytical data
4. **Glue Crawlers & Catalog** — Schema discovery and registration
5. **Athena** — Query engine to analyze curated data

### 🧩 Tools & Services

| Component | Service Used |
|---|---|
| Storage | AWS S3 |
| ETL Processing | AWS Glue (PySpark) |
| Metadata | AWS Glue Catalog |
| Security | AWS IAM |
| Querying | AWS Athena |

### ⚙️ Data Flow

| Step | Description | Output |
|---|---|---|
| 1️⃣ Raw Ingestion | CSV upload to S3 | train_delays.csv |
| 2️⃣ Transformation | PySpark cleaning (Glue Job #1) | train_delays_cleaned.parquet |
| 3️⃣ Aggregation | Avg delay metrics (Glue Job #2) | avg_delay_per_train, avg_delay_per_route |

| Step | Description | Output |
|---|---|---|
| 4️⃣ Schema Registration | Glue Crawlers | Registered tables in Glue Catalog |
| 5️⃣ Query & Insights | Athena SQL Queries | Top delayed trains/routes |

### 📊 Example Insights

| Metric | Example Result |
|---|---|
| 🚆 Most Delayed Train | Rajdhani Express (24.5 min delay) |
| 📍 Most Delayed Route | NDLS → BCT |
| 🕒 Average Delay | 14.2 min |
| 🗓️ Worst Day | Monday |

### 🌟 Highlights

- ✅ Implemented AWS Medallion Architecture
- ✅ Automated schema registration using Glue Crawlers
- ✅ Used PySpark for transformation & aggregation
- ✅ Queried curated Parquet data with Athena
- ✅ Cloud-native, fully serverless pipeline

### 📈 Outcome

A scalable, modular, and cloud-native **AWS Data Lake pipeline** ready for analytics and BI dashboards.

**Sanket Aba Adhav**
*Data Engineer | AWS & PySpark Enthusiast*