Indian Express Train Tracker – Project Documentation

Introduction

Indian Express Train Tracker is a containerized data engineering and visualization solution that monitors and analyzes real-time train movements in India.
 The project leverages Indian Railways API for live data ingestion, processes the data with PySpark, stores it in a PostgreSQL database, and visualizes insights via Apache Superset.
 All components are containerized and orchestrated with Docker Compose for seamless deployment and scalability.

Project Objectives

The primary goals of this project were to:

1. Automate the end-to-end train tracking and analytics process.
2. Standardize & clean raw API data for analytics readiness.
3. Persist processed datasets in a robust and query-friendly database.
4. Provide interactive and real-time dashboards for decision-making.
5. Ensure scalable deployment using Dockerized services.

What We Did – Step by Step

1. Data Ingestion

- Connected to the Indian Railways API to fetch real-time train running status and route information.
- Packaged API requests inside a Dockerized service for reliability and isolation.

---

2. Data Processing & Transformation (PySpark)

- Implemented a PySpark pipeline for:
    - Cleaning & normalizing raw API data.
    - Handling missing values and inconsistent time/date formats.
    - Standardizing schema for downstream processing.
- Generated analytics-ready datasets including:
    - Train route details.
    - Live status with delays and arrival/departure times.
    - Aggregated metrics for performance tracking.

---

3. Data Storage (PostgreSQL)

- Set up a PostgreSQL container as the data warehouse.
- Designed tables to store:
    - Raw ingested data.
    - Processed and aggregated datasets.
- Created indexes and optimized schema for faster queries.

---

4. Visualization & Analytics (Apache Superset)

- Deployed Apache Superset inside Docker for easy access and integration.
- Connected Superset to PostgreSQL to power dashboards.
- Built interactive visualizations such as:
    - Live Train Locations Map
    - Top Delayed Trains
    - Delay Trends by Route

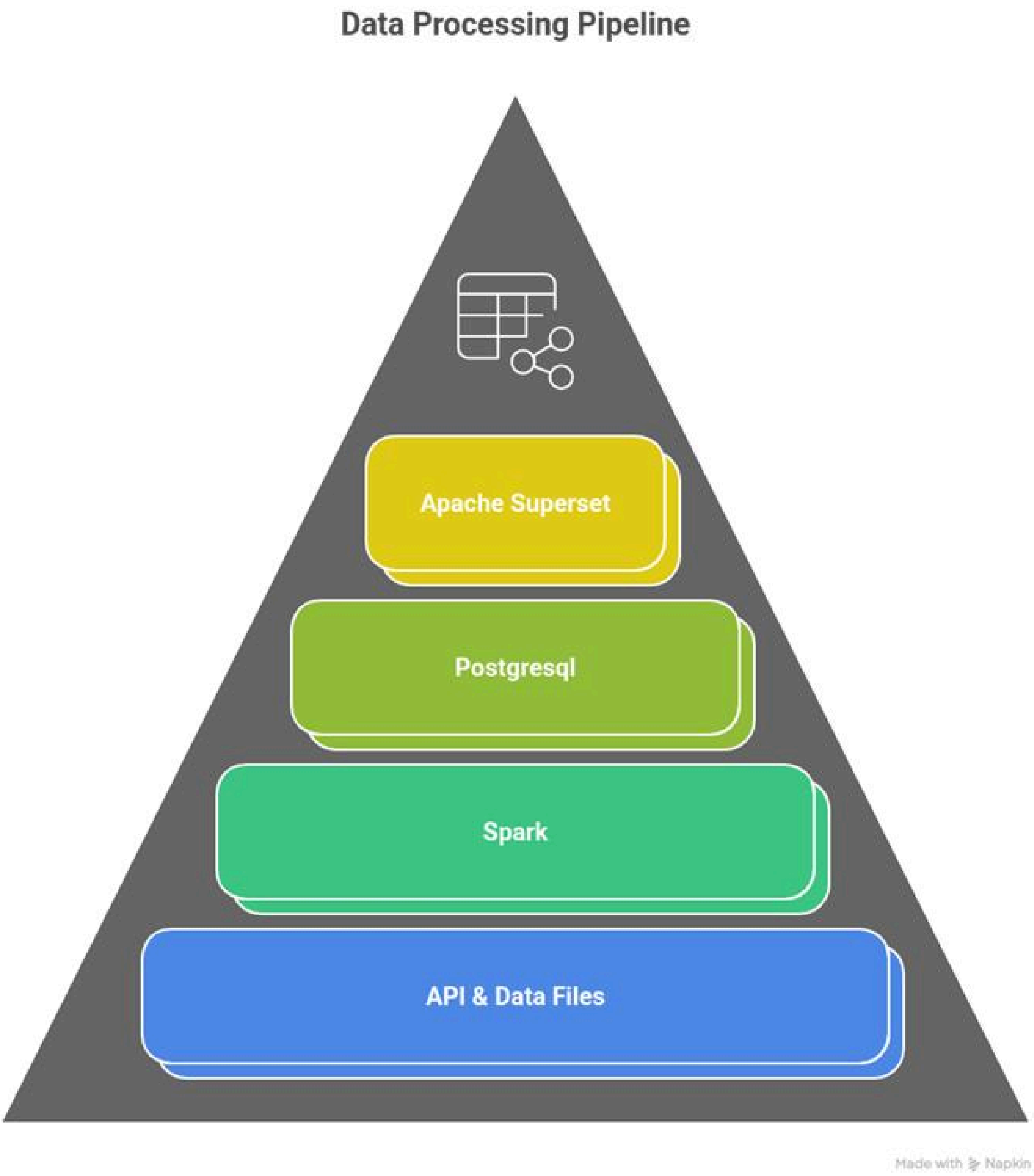
    - Daily Performance Metrics
- Implemented GeoJSON-based mapping for train routes.

---

5. Containerization & Orchestration

- Used Docker Compose to define and orchestrate:
    - PySpark container.
    - PostgreSQL container.
    - Apache Superset container.
    - API fetcher service.
- Ensured services start in the correct sequence for smooth pipeline execution.

Architecture



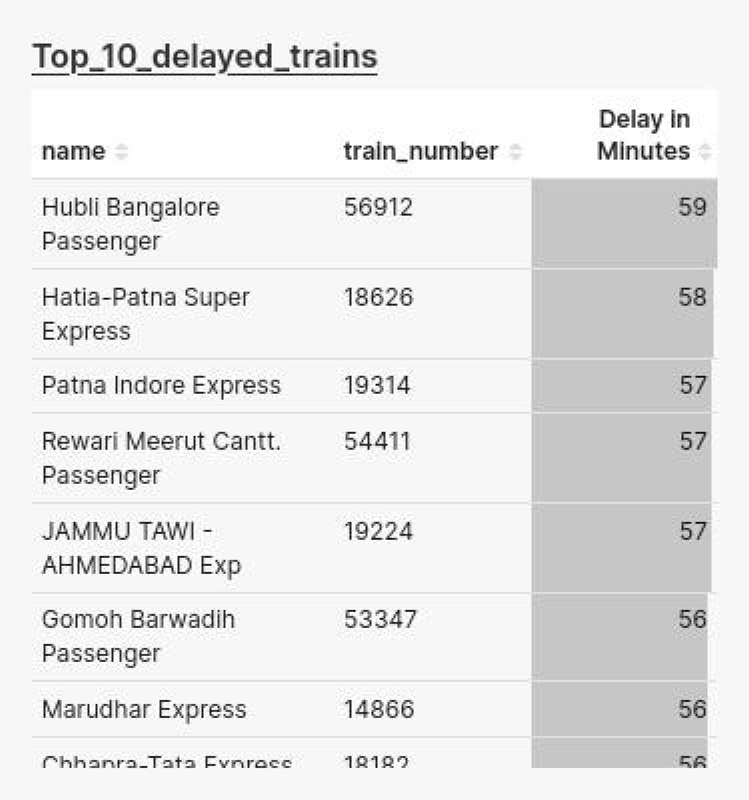
- **Indian Railways API** – Source of real-time train data.
- **PySpark** – Cleansing, transformations, and aggregation.
- **PostgreSQL** – Persistent storage for processed datasets.

- **Apache Superset** – Dashboards & analytics.
- **Docker Compose** – Service orchestration.

Project Structure

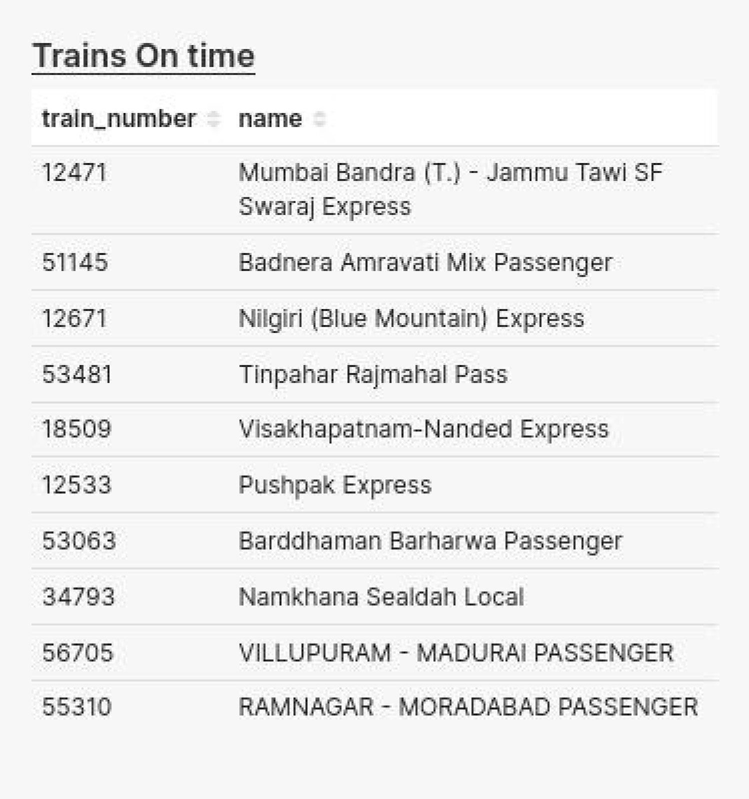
```
indian-train-live-tracker/
        ├── config/             # Superset configs
        ├── data/               # Local data files
        ├── db_backups/         # DB backup files
        ├── docker-compose.yml          # Orchestration file
        ├── indian-railway-tracker/ # API interaction & utilities
        ├── irctc-connect-main/         # Node.js service for API handling
        ├── notebooks/          # Analysis & development notebooks
        ├── postgres_data/              # Postgres persistent storage
        └── utils/              # Python helper scripts
```

Key Dashboards



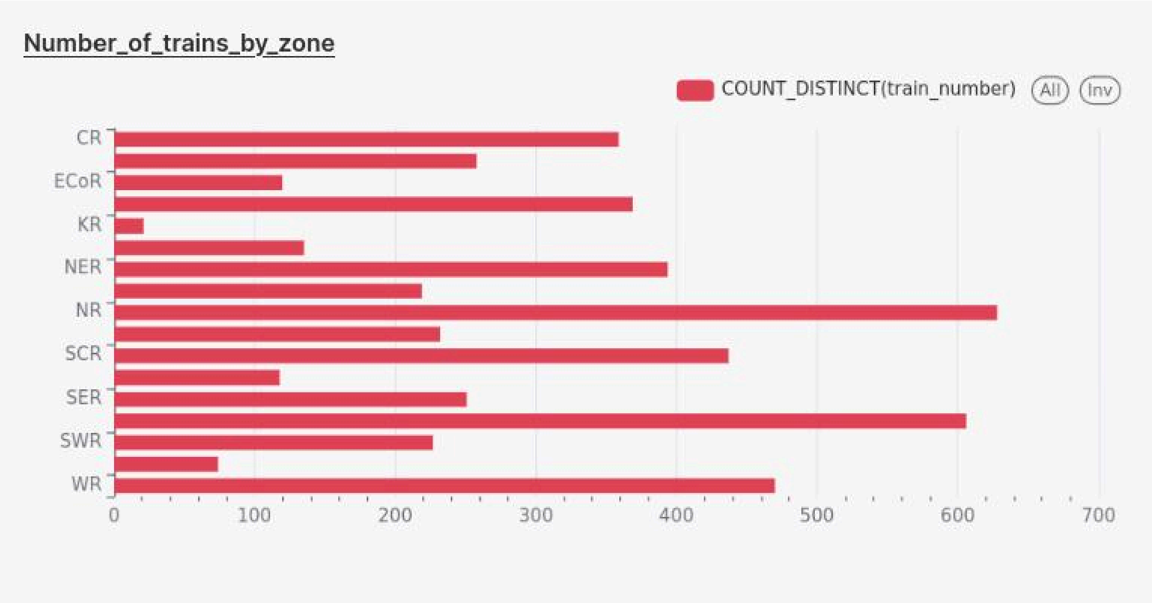
Top_10_delayed_trains

name	train_number	Delay in Minutes
Hubli Bangalore Passenger	56912	59
Hatia-Patna Super Express	18626	58
Patna Indore Express	19314	57
Rewari Meerut Cantt. Passenger	54411	57
JAMMU TAWI - AHMEDABAD Exp	19224	57
Gomoh Barwadih Passenger	53347	56
Marudhar Express	14866	56
Chhapra-Tata Express	18182	56



Trains On time

train_number	name
12471	Mumbai Bandra (T.) - Jammu Tawi SF Swaraj Express
51145	Badnera Amravati Mix Passenger
12671	Nilgiri (Blue Mountain) Express
53481	Tinpahar Rajmahal Pass
18509	Visakhapatnam-Nanded Express
12533	Pushpak Express
53063	Barddhaman Barharwa Passenger
34793	Namkhana Sealdah Local
56705	VILLUPURAM - MADURAI PASSENGER
55310	RAMNAGAR - MORADABAD PASSENGER



Future Enhancements

- Integrate **Apache Airflow** for scheduled ETL pipeline execution.
- Enhance **geo-visualization** for detailed route mapping.
- Implement **predictive analytics** to forecast train delays.
- Add **streaming data ingestion** for near real-time updates.

Outcome

By the end of this project, we achieved:

- **Consistent & clean datasets** for analytics.
- **Interactive dashboards** for real-time operational insights.
- A **scalable, containerized architecture** that can be deployed in any environment.