Spotify Analytics Data Pipeline 🎶

Project Overview

The Spotify Analytics Data Pipeline project is designed to handle data from Spotify's API, ensuring high-quality extraction, transformation, and loading (ETL) into a data warehouse. This pipeline facilitates accurate and comprehensive analytics, providing a scalable solution for data-driven decision-making.

Technologies and Tools Used

- **Python**: Scripting and custom ETL processes
- **SQL**: Data querying and transformation
- **Apache Airflow**: Workflow orchestration
- **Spotify API**: Data source
- **CSV and JSON**: Data storage and interchange formats
- **Git**: Version control
- **Docker**: Containerization for consistent environments
- **PostgreSQL**: Data warehouse
- **Data Modeling**: Structuring data for analytics

Project Phases and Key Milestones

Phase 1: Initialization and Planning

- Defined project scope, objectives, and success criteria
- Configured development environment with Python, Airflow, PostgreSQL, and Docker
- Set up Spotify API connectivity

Phase 2: Data Extraction

- Developed Python scripts for data extraction from the Spotify API
- Stored extracted data in CSV and JSON formats

Phase 3: Data Transformation

- Cleaned and preprocessed raw data using Python and SQL
- Developed a schema for data organization in PostgreSQL
- Implemented transformation rules for structured data

Phase 4: Data Loading

- Designed PostgreSQL tables for transformed data
- Loaded data into PostgreSQL using Python scripts
- Ensured data integrity with quality checks

Phase 5: Workflow Orchestration

- Created Airflow DAGs to automate and schedule ETL processes
- Defined task dependencies and schedules
- Set up monitoring and alerting mechanisms

Project Outcome

- **High-Quality Data**: Ensured data quality and integrity
- **Scalable Solution**: Designed to handle increasing data volumes
- **Automated Workflow**: Enhanced efficiency with Apache Airflow
- **Comprehensive Analytics**: Enabled accurate and comprehensive analytics

Conclusion

The Spotify Analytics Data Pipeline project successfully met its objectives, showcasing the effective use of modern data engineering tools and techniques. This robust solution supports data-driven decision-making based on Spotify's data and lays the foundation for future enhancements and scalability.