Title: End-to-End Data Pipeline for Multi-Tenant BI Platform

🧩 Objective:

Simulate a real-world BI project by building a basic, yet modular **ETL pipeline**, with a **multi-tenant-ready data model**, and a **dashboard** showing key metrics.

📅 Duration:

2 Days

- **Day 1:** Data Ingestion + Transformation
- **Day 2:** Data Warehouse + Dashboard

🔧 Task Overview:

Day 1 – Data Ingestion & Transformation

1. **Data Sources:**
   - CSV file: Sample marketing campaign data (leads, clicks, conversions)
   - Excel: Region-wise sales targets
   - API (mocked): Customer metadata from a REST endpoint
2. **Requirements:**
   - Ingest data using one of:
     - Apache NiFi

    - Apache Airflow (with Python or Bash operators)
    - Custom Python ETL (for air-gapped simulation)
  - For API mocking: Use Postman mock server or Python `http.server`
3. **Transform the data:**
  - Clean & merge datasets using:
    - `pandas` or `dask`
    - `dbt-core` (optional bonus for SQL-first transformation)
  - Standardize date formats, remove nulls, derive fields like CTR (Click-through rate), ROI
  - Create multi-tenant schema design using tenant_id/customer_id in all data

---

Day 2 – Warehouse + Dashboards

1. **Data Warehouse Options (choose one):**
  - **ClickHouse** (ideal, fast OLAP)
  - **PostgreSQL** or **DuckDB** (if lightweight)
  - **Apache Druid** (bonus: for real-time simulation)
  - Setup schema per tenant or include `tenant_id` as a partition key
2. **BI Visualization:**
  - Use one of:
    - **Metabase** (preferred for ease and multi-tenant config)
    - Apache Superset
    - Redash
    - Grafana (if working with metrics)
3. **Dashboards Must Include:**
  - Per-tenant campaign summary (CTR, Spend, Revenue)
  - Conversion funnel chart
  - Time-series chart (e.g., conversions/day)

4. **Optional Enhancements:**
    - Schedule email reports (Metabase or Superset)
    - Validate incoming data with `Great Expectations` or simple schema checks

---

📂 Deliverables:

- Codebase (GitHub or zip)
- Setup instructions (README)
- Sample data files & mock API script
- One dashboard link/screenshot per tenant
- Short Loom/Youtube walkthrough (optional bonus)

🎯 Skills Being Evaluated:

- Python ETL scripting or orchestration (Airflow/NiFi)
- API & file handling
- Modular transformation using pandas/dbt
- OLAP modeling (ClickHouse/PostgreSQL)
- Multi-tenant data handling
- Dashboarding, filtering, RBAC

✅ Success Criteria:

- End-to-end pipeline works with realistic sample data

- Modular code structure with README
- Dashboards reflect per-tenant separation
- Bonus: use of modern tools like dbt, NiFi, Metabase with RBAC