Architecture Overview

1. Ingestion Layer

- Data sources: Apollo TSV, California CSV (cleaned), Excel Part2
- Processing Engine: PySpark (local mode) + DuckDB clean-up
- Output: Parquet dataset (~10M rows)

2. Processing Layer

- Deduplication (email → phone → LinkedIn → name+company)
- Normalization of columns to Master Schema
- Output stored in Parquet

3. Storage Layer

- PostgreSQL holds:
    - unified_people (final master dataset)
    - api_users
    - api_logs
    - credits

4. API Layer (FastAPI)

- Endpoints:
    - /search
    - /admin/add_credits
    - /logs
- Features:
    - API key auth
    - Rate limiting
    - Credit deduction
    - Query filters
    - Pagination

5. Authentication

- API keys stored in postgres(api_users)
- Middleware validates:
    - API key
    - credits remaining

    - rate limit

6. Deployment Layer

- Dockerfile (FastAPI)
- docker-compose.yml (API + PostgreSQL)