# MCA Insights Engine

## Problem Statement

The Ministry of Corporate Affairs (MCA) publishes company master data as state-wise CSV files on data.gov.in.
Each state file contains company-level information including:

- **CIN (Corporate Identification Number)**
- **Company name and class** (Private/Public, Limited by shares, etc.)
- **Date of incorporation**
- **Authorized and paid-up capital**
- **Company status** (Active, Strike Off, Amalgamated, etc.)
- **Principal business activity** (NIC code)
- **Registered office address and ROC code**

Tracking changes in this data is important for functions such as compliance, credit assessment, and risk monitoring.
However, there are practical challenges:

- The datasets are **large** and **updated frequently**, making manual monitoring infeasible.
- The raw MCA data often **lacks contextual information** (for example, company websites, director lists, or enriched sector labels), which limits downstream analysis.
- Evaluators and users need a reproducible, auditable way to see **what changed, when, and why**.

**Task:** Build a Python application that:

1. Consolidates and normalizes state-wise MCA data into a canonical master dataset,
2. Detects and logs daily company-level changes (new incorporations, deregistrations, and field updates), and
3. Enriches the changed records with public web data and an AI-powered insights layer so users can query the dataset conversationally and receive automated summaries.

## Scope & Guidelines

### Dataset

The project uses the **Company Master Data** published by the **Ministry of Corporate Affairs (MCA)** on data.gov.in.
This dataset provides a comprehensive record of all registered companies in India, organized by their respective **Registrar of Companies (RoC)** jurisdictions.

---

**Structure of the Dataset**

Each state-wise file contains detailed company-level information, including:

- **CIN (Corporate Identification Number)** a unique company identifier
- **Company Name and Class** (Private/Public, Limited by shares, etc.)
- **Date of Incorporation**
- **Authorized and Paid-up Capital**
- **Company Status** (Active, Strike Off, Amalgamated, etc.)
- **Principal Business Activity** (as per NIC code)
- **Registered Office Address** and **RoC code**

---

**State Coverage**

For this assignment, data from **five RoC jurisdictions** — *Maharashtra, Gujarat, Delhi, Tamil Nadu,* and *Karnataka* —
has been selected to represent diverse corporate ecosystems across India.

---

**Data Update Simulation**

To simulate a **daily update mechanism**, multiple snapshots of the same state-level dataset are compared over time to detect and log:

- **Newly incorporated companies**

- **Companies struck off or deregistered**
- **Field-level changes**, such as modifications in authorized capital, classification, or company status

This enables the system to automatically track company lifecycle events and maintain an up-to-date master record for each RoC.

## Core Tasks

---

### A. Data Integration

- Merge and clean data from five selected states.
- Standardize column structures, handle null values, and remove duplicates.
- Create a **canonical master dataset** combining data across all states to ensure consistency and comparability.

---

### B. Change Detection

Design and implement a system to **detect and log company-level changes daily**.

The system should:

- Identify key change categories:
  - **New incorporations**
  - **Status changes**
  - **Authorized or Paid-up Capital modifications**
- Generate structured change logs in CSV/JSON format with the following fields:

| CIN | Change_Type | Field_Changed | Old_Value | New_Value | Date |
|---|---|---|---|---|---|

- Track updates across **three consecutive daily snapshots** to ensure temporal accuracy.

- Maintain and automatically update a **master database** to reflect the most recent company information for each CIN.

---

**C. Web-Based CIN Enrichment**

For a representative sample of **50–100 companies** showing recent changes, perform data enrichment using publicly available online sources.

**Suggested sources:**

- ZaubaCorp
- API Setu (MCA Master Data)
- Indian Kanoon
- GST Portal
- MCA21 Corporate Data Management Portal

The enrichment process extracts and merges supplementary information such as:

- Sector and Industry classification
- Director names and company type
- Registered office address

**Final enriched dataset format:**

| CIN | COMPANY_ NAME | STATE | STATUS | SOURCE | FIELD | SOURCE_URL |
|---|---|---|---|---|---|---|

---

**D. Query Layer**

Develop an interface for **interactive data access and visualization**.

**Functional requirements:**

- Build a **Streamlit** or **Flask-based dashboard** with:
    - Search functionality (CIN or Company Name)
    - Filters by Year, State, and Company Status
    - Visualized change history over time
    - Display of enriched company-level information

- Include an **optional REST API endpoint** (/search_company) for integration with external applications or testing via Postman.

---

**E. AI-Powered Features (Enhanced Insight Layer)**

To provide an **intelligent and interactive user experience**, integrate an AI-powered layer that leverages structured MCA data for discovery and analysis.

**This includes:**

- Automated **AI Summary Generation** for daily change reports.
- A **conversational chatbot interface** that allows users to query and explore company changes, trends, and insights in natural language.

This layer transforms the system from a static tracker into an **AI-driven insight engine**.

### *1. AI Summary Generator*

Automatically generate **concise daily reports** after each data update.
These reports highlight key company-level changes and serve as quick analytical summaries.

**Reports summarize the following:**

- Total number of **new incorporations**, **removals**, and **field updates**
- Notable **status transitions** or variations in company data

**Example Output:**

Daily Summary

New incorporations: 124

Deregistered: 5

Updated records: 42

This summary can be generated automatically and stored as a .txt or .json file, or displayed directly in the dashboard interface.

---

**2. Chat with MCA Data**

Implement a **conversational query layer (Chatbot)** that enables users to interact with the MCA dataset using **natural language questions**.

**Example Queries:**

- "Show new incorporations in Maharashtra."
- "List all companies in the manufacturing sector with authorized capital above Rs.10 lakh."
- "How many companies were struck off last month?"

The chatbot may use **rule-based natural language interpretation** or **Retrieval-Augmented Generation (RAG)** using an **LLM of choice**.
It translates user questions into structured database queries and returns relevant records or summaries directly through the chat interface.

---

**Deliverables**

1. **Source code repository (GitHub)** which contains all scripts and documentation.
2. **Processed and cleaned MCA dataset** (CSV or Database).
3. **Three daily change logs** demonstrating incremental data evolution.
4. **Enriched dataset** merged with publicly available web information.
5. **AI summary reports** generated automatically (daily_summary.txt / .json).
6. **Streamlit or Flask dashboard** featuring both search and chatbot interaction capabilities.
7. **Comprehensive README** detailing setup, architecture, workflow, and enrichment logic.

**NOTE:**

We sincerely appreciate your time and effort in reviewing this submission.
Please note that we **do not expect you to invest extensive time** in building or executing a fully functional solution.
A **working proxy or representative implementation**, demonstrating the intended logic and integration with the **appropriate data sources that is mentioned**, will be considered sufficient for evaluation purposes.