Stage 2 – System Design Documentation

A. Data Model Finalization

Entities:

- Stations
- AQI_Readings
- Weather_Readings
- Traffic_Readings
- Industrial_Zones
- Users
- Citizen_Reports
- Pollution_Events
- Event_Classifications

Key Relationships:

- One Station → Many AQI_Readings
- One Station → Many Weather_Readings
- One Station → Many Traffic_Readings
- One Station → Many Pollution_Events
- One Pollution_Event → One Event_Classification
- One User → Many Citizen_Reports

Indexes:

- Index on (station_id, timestamp) for AQI_Readings
- Index on (station_id, timestamp) for Weather_Readings
- Index on (station_id, timestamp) for Traffic_Readings
- Index on timestamp for Citizen_Reports
- Index on event_id for Event_Classifications

Timestamp Policy:

- All timestamps stored in UTC
- ISO 8601 format
- Fixed polling interval (e.g., 5 minutes)

- Rolling average window: 1 hour

Data Retention Policy:

- Raw readings retained for 1 year
- Aggregated summaries retained indefinitely
- Citizen media stored in secure cloud storage

B. Classification Engine Specification

Input Parameters:

- AQI
- PM2.5
- NO2
- Wind speed (m/s)
- Industrial distance (km)
- Traffic index (0–1)
- Nearby station spike count
- Citizen report count
- Farm influence index (0–1)

Score Categories:

- Industrial Score
- Vehicular Score
- Regional Transport Score
- Agricultural Burning Score

Wind Modifier Logic:

- If windSpeed < 1 m/s:
- IndustrialScore *= 0.6
- RegionalScore *= 0.8
- FarmScore *= 0.7
- VehicularScore += 10

Classification Decision Logic:

- If highest score < 30 → Uncertain
- If (highest – second highest) < 10 → Mixed Contribution
- Else assign label of highest score

Confidence Formula:

Confidence (%) = ((TopScore – SecondScore) / TopScore) × 100

C. API Integration Plan

External Source Handling:

- Authentication: API keys stored in environment variables
- Rate Limits: Implement throttling respecting provider limits
- Response Structure: Normalize into internal schema
- Error Handling: Retry with exponential backoff; log failures
- Fallback Strategy: Use cached last-known data if API fails

Data Ingestion Blueprint:

- Fetch AQI data at fixed interval
- Fetch weather and traffic data
- Normalize and validate responses
- Store in database with UTC timestamps
- Trigger spike detection
- Run classification engine
- Store classification results