Environmental Air Quality Monitoring & Analytics Platform

A civic-tech data platform that turns raw air-quality readings (PM2.5, O_3, NO_2) from public APIs and community sensors into clear, trustworthy insights for **public-health teams, schools, and residents**. Think: timely visibility during wildfire smoke events, neighborhood-level trends over weeks, and simple advisories like "limit outdoor activity near School X today." No PII, fast to implement, real community impact.

What it delivers:

- A **daily/near-real-time AQI map** with 7- and 30-day rolling trends by station and area.
- **Hotspot & exceedance detection** (where/when air is consistently unhealthy).
- An optional **"schools view"** to help principals/coaches make outdoor activity calls.
- A small **alerting** path (Slack/email) for threshold breaches during events.

Data we'll use:
Public AQ APIs (federal/state or community sensor networks), station metadata (lat/long, elevation), and optional layers like school locations or census tracts for an equity lens. We can start with two sources, then add more.

How it works (tech + flow):

- **Snowflake** is our warehouse: we land raw files in an external stage, bulk-load with `COPY INTO`, and keep a clean **dimensional model** (stations, time, pollutants; facts for readings and daily AQI). We handle incremental updates with `MERGE` so reruns don't duplicate data.
- **dbt** manages transformations, tests (unique/not-null/range checks), and lineage docs so our tables are reliable and explainable.
- **Airflow** orchestrates everything: hourly/daily DAGs per source plus a model-refresh DAG, with retries, catchup, and backfills driven by Airflow's logical date (easy to reprocess missing days).
- **Spark** runs a weekly batch job over history to flag anomalies (e.g., unusual spikes not explained by season/weekday patterns).
- **Streaming (stretch)** brings near-live sensor feeds (Kafka/Snowpipe Streaming) so dashboards update during fast-moving events.
- **NoSQL/Vector (optional)** keeps unstructured sensor logs/images for future analysis.

Team of 4 – clean split:

- **Pipelines Lead:** Airflow DAGs, schedules, backfills, notifications.
- **DW/Modeling Lead:** Snowflake schemas, staging/COPY, `MERGE` upserts, performance.
- **Analytics/Viz Lead:** dbt marts, window-function KPIs, dashboard, "schools view."
- **PM & QA Lead:** data-quality tests, runbooks, docs, and the demo storyline.

Milestones (keep us honest):

- **Week 1:** finalize sources, draft schema, outline DAGs.
- **Weeks 2–3:** Snowflake + external stage; first hourly ingest; dbt core models; basic dashboard.
- **Weeks 4–5:** add second source; daily AQI mart; anomaly scan (Spark); tighten tests.
- **Stretch:** streaming path + alerts; equity overlays (schools/tracts).
- **Final:** polish docs, run live demo.

Demo script (what we'll show live):
Kick off an Airflow run → watch raw files land and `COPY INTO` → show `MERGE` upserts and dbt tests passing → run a query with rolling AQI (7/30-day) → open the dashboard (hotspots + schools view) → show an alert example from a recent spike.