

**1️⃣ External APIs / Data Sources**

Sources of raw, high-velocity or batch data.

* **Google Maps API**: Travel times, congestion
* **OpenWeatherMap API**: Weather data
* **GTFS Feeds**: Public transit schedules
* **Event Feeds**: Eventbrite, Ticketmaster
* **Historical Datasets**: Sidewalk Labs, Open Gov

🔍 *These provide both real-time and contextual signals to power downstream analytics and ML.*

**2️⃣ Ingestion Layer**

Handles **streaming and batch ingestion** of data into the platform.

* **Kafka / Pub/Sub**: Real-time traffic/event/weather updates
* **Airflow DAGs**: Hourly/daily batch ingestion from APIs or dumps
* **Pydantic / JSON Schema**: Validates and enforces input formats
* **Raw Zone (GCS)**: Stores all unprocessed data in partitioned buckets

📌 *Ensures timeliness, freshness, and traceability of data.*

**3️⃣ Processing Layer**

Applies transformation, cleansing, and enrichment.

* **Apache Beam / Spark**: Real-time and batch transformations
* **dbt**: Manages SQL-based modeling, dependency graphs, testing
* **Validation**: Great Expectations, Pytest, schema checks

💡 *This layer converts raw data into curated and analytical-ready forms.*

**4️⃣ Storage Layer**

Where all processed data is structured and stored.

* **BigQuery**: Central analytical warehouse
* **GCS Lakehouse**: Raw → Cleaned → Curated layers (Parquet)
* **Feature Store**: ML-specific enriched data (lagged features, geohash, time windows)

📊 *Supports analytics, ML, and BI layers.*

**5️⃣ ML Layer**

Predictive modeling and intelligent services.

* **Model Training**: XGBoost, Prophet, LSTM (trained on congestion history)
* **Vertex AI / MLflow**: For model lifecycle tracking, experiments, and serving
* **Model Serving**: Real-time API predictions via FastAPI or Vertex endpoints
* **Drift Detection**: Scheduled jobs to monitor accuracy and freshness

🤖 *The intelligence engine — predicts future congestion and commute times.*

**6️⃣ Insight Layer (Dashboards & Apps)**

User-facing applications for both policy and commuter needs.

* **Power BI**: Executive dashboards for city planners and policymakers
* **Streamlit**: Interactive commute planner with real-time insights for users
* **Leaflet.js**: Heatmaps for congestion hotspots and route maps

📈 *Translates data into actionable insights for end-users.*

**7️⃣ CI/CD Layer**

Automated workflows to manage changes, testing, and deployment.

* **GitHub Actions**: Run unit tests, dbt tests, DAG linting, ML retraining
* **Terraform**: Provision infrastructure as code
* **Deployment Pipelines**: Canary releases, staging, and production environments

🚀 *Enables fast, safe, and reliable development at scale.*

**8️⃣ Observability & Monitoring**

Ensures system health, performance, and data reliability.

* **Prometheus + Grafana**: Real-time metrics (latency, throughput)
* **Stackdriver**: Log-based monitoring and alerting
* **OpenLineage**: Track data pipeline lineage and dependencies

🛠️ *Critical for reliability, cost control, and SLA adherence.*

**9️⃣ Security & Access Control**

Governance and protection of data and services.

* **IAM Roles**: Fine-grained access control
* **Policy Enforcement**: Data anonymization, GDPR/CCPA compliance
* **Audit Logs**: Track user actions, data access events

🔐 *Ensures legal, ethical, and secure data use.*