PROJECT BLUEPRINT — REAL-TIME WEATHER TRENDS & PREDICTION

Team:
Max Matteucci
Ethan Lebon
Ian Hedges
Caleb Brunton

Course:
MGMT 467 — Units 1–3 Integrated Term Project

---

1. BUSINESS PROBLEM

Many organizations depend on timely and accurate weather information for planning, logistics, and operations. While historical weather data supports long-term trend analysis, it does not reflect current conditions. Conversely, live weather feeds lack historical context.

This project addresses that gap by building a real-time weather analytics pipeline that combines historical data with continuously streaming live data. The system enables real-time monitoring, anomaly detection, and short-term forecasting using a unified cloud-based architecture.

---

2. DATA SOURCES

Historical (Batch):
- Kaggle historical hourly weather dataset
- Multi-year data covering temperature, humidity, wind, pressure, and timestamps
- Used for baseline analysis and machine learning training

Live (Streaming):
- Open-Meteo public weather API
- Provides current weather conditions for multiple cities
- Used for real-time monitoring and near-term prediction

All data sources are public and contain no personal or sensitive information.

---

3. ARCHITECTURE OVERVIEW

The system follows a serverless, event-driven streaming architecture on Google Cloud Platform.

High-level flow:
Public Weather API
→ Cloud Function (2nd Gen)
→ Pub/Sub
→ Dataflow (Streaming Template)
→ BigQuery
→ BigQuery ML
→ Looker Studio

Key characteristics:
- Fully serverless
- Horizontally scalable
- Near real-time data availability
- Decoupled ingestion and processing

(Dataflow Job Graph screenshot included as architectural evidence.)

---

4. PIPELINE COMPONENTS

Ingestion:
- A Cloud Function periodically calls the Open-Meteo API
- API responses are normalized into structured JSON
- Each payload is published as a message to a Pub/Sub topic

Streaming Processing:
- A managed Dataflow streaming template subscribes to Pub/Sub
- Messages are validated and transformed
- Clean records are written to a BigQuery streaming table

Storage:
- BigQuery stores live streaming data and historical batch data
- Tables are append-only and time-partitioned
- Supports both analytics and ML workloads

---

5. MACHINE LEARNING PLAN

BigQuery ML is used to train predictive models directly inside BigQuery.

Model objectives:
- Predict short-term future temperature
- Use historical weather patterns and recent live observations

Approach:
- Feature engineering from historical data (seasonality, time-based features)
- Integration of live weather attributes
- Linear regression model for interpretability
- Evaluation using built-in BigQuery ML metrics
- Model explainability via ML.EXPLAIN_PREDICT

---

6. KEY PERFORMANCE INDICATORS (KPIs)

The executive dashboard focuses on the following KPIs:
- Current temperature by city (live data)
- Short-term temperature trend (time series)
- Forecasted temperature (ML output)
- Anomaly indicator relative to historical norms

These KPIs directly align with the business objective of real-time situational awareness.

---

7. RISKS & MITIGATION

Identified risks:
- API rate limits or temporary outages
- Streaming job interruptions
- Data latency or missing values

Mitigation strategies:
- Serverless retries and monitoring
- Dataflow job restart capability
- Timestamp-based freshness validation
- Clear separat