DATA 608 – DEVELOPING BIG DATA APPLICATIONS
WEATHER PREDICTION AND FORECASTING

GOAL

The project addresses the challenge of understanding short-term weather variations across ten major Canadian cities. By analysing three months of recent data, it identifies trends, anomalies, and differences between regions. This helps improve forecast accuracy and supports industries like agriculture and transportation in making better decisions based on real-time weather insights. It also enables a comparative analysis to detect regional weather patterns and seasonal changes. These insights can assist policymakers and businesses in planning for weather-related risks and optimizing resource management.

GUIDING QUESTIONS

Q1.Weather Pattern Analysis:

a) What are the **trends** in temperature, precipitation, and wind speed?
b) How do **weather variables** (e.g., humidity, pressure) correlate over time?

Q2.Forecast Accuracy & Reliability:

a) Can we evaluate **model performance** by comparing forecasts with actual outcomes?
b) What **factors** influence prediction reliability (e.g., location, season, data granularity)?

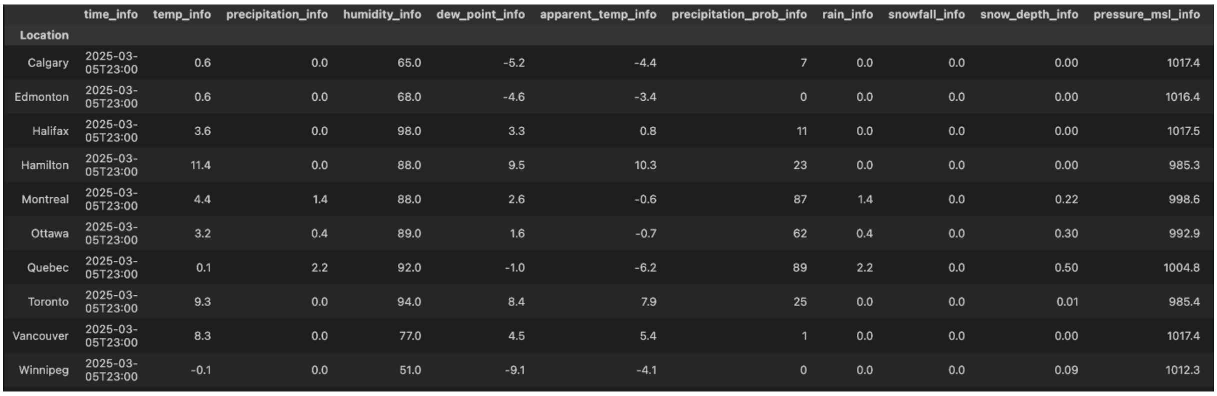
Q3.Extreme Weather Detection:

a) Can we predict **extreme weather events** (e.g., storms, heatwaves) using forecast data?
b) What **thresholds** can be established to identify anomalies?
c) How do **extreme events** vary across geographic regions?

Q4.Geographic & Temporal Insights:

a) How does **weather variability** differ across regions and time periods?
b) What regions are more prone to **climate volatility** based on historical trends?

ABOUT THE DATASET



	time_info	temp_info	precipitation_info	humidity_info	dew_point_info	apparent_temp_info	precipitation_prob_info	rain_info	snowfall_info	snow_depth_info	pressure_msl_info
Location											
Calgary	2025-03-05T23:00	0.6	0.0	65.0	-5.2	-4.4	7	0.0	0.0	0.00	1017.4
Edmonton	2025-03-05T23:00	0.6	0.0	68.0	-4.6	-3.4	0	0.0	0.0	0.00	1016.4
Halifax	2025-03-05T23:00	3.6	0.0	98.0	3.3	0.8	11	0.0	0.0	0.00	1017.5
Hamilton	2025-03-05T23:00	11.4	0.0	88.0	9.5	10.3	23	0.0	0.0	0.00	985.3
Montreal	2025-03-05T23:00	4.4	1.4	88.0	2.6	-0.6	87	1.4	0.0	0.22	998.6
Ottawa	2025-03-05T23:00	3.2	0.4	89.0	1.6	-0.7	62	0.4	0.0	0.30	992.9
Quebec	2025-03-05T23:00	0.1	2.2	92.0	-1.0	-6.2	89	2.2	0.0	0.50	1004.8
Toronto	2025-03-05T23:00	9.3	0.0	94.0	8.4	7.9	25	0.0	0.0	0.01	985.4
Vancouver	2025-03-05T23:00	8.3	0.0	77.0	4.5	5.4	1	0.0	0.0	0.00	1017.4
Winnipeg	2025-03-05T23:00	-0.1	0.0	51.0	-9.1	-4.1	0	0.0	0.0	0.09	1012.3

🌦 Free Open-Source Weather API | Open-Meteo.com

We'll be using time series weather data from the past three months for various cities across Canada using their longitude and latitude information, collected from the Open-Meteo API. This API sources its data from multiple national weather providers, including the **GEM (Global Environmental Multiscale) model** from the **Canadian Weather Service**. GEM provides data **updated every 6 hours**.

The weather data typically includes hourly observations of various variables such as **temperature at 2 meters**, **apparent temperature**, **relative humidity**, **dewpoint**, **precipitation probability**, **total precipitation** (including **rain**, **showers**, and **snow**), **snowfall**, **snow depth**, **sunshine duration**, **UV index**, and **daylight duration**. Additional details like **weather codes**, **sunrise**, and **sunset times** are also provided. For our analysis, we carefully **selected the most important parameters** to be included in the API request, ensuring we capture the most relevant and informative aspects of the weather data while keeping it efficient and focused. This combination of frequently updated data ensures a comprehensive and accurate representation of weather conditions over time.

The weather data from the Open-Meteo API also includes **WMO (World Meteorological Organization) weather interpretation codes (WW)**, which provide standardized descriptions of various weather conditions. These codes classify weather phenomena based on intensity and type, making it easier to interpret the data.

WHAT THE IDEAL END PRODUCT WOULD LOOK LIKE?

The ideal end product will be a web-based application or dashboard deployed on the Streamlit platform, designed to answer key guiding questions with real-time data visualizations and predictive modeling (such as forecasting the next day's temperature). We will compare forecasted results using various machine learning models. The application will also provide a detailed comparison of weather across different cities in Canada. The complete data engineering pipeline will be followed, starting from data generation and ingestion, setting up an EC2 instance, and storing the data in S3 storage. The data will be dynamically updated every 12 hours and automatically refreshed in both the S3 storage and the dashboard.

Data Generation: Weather data will be generated using the Open-Meteo API, which will provide real-time and historical weather information for various cities across Canada.

Data Ingestion: The data will be ingested into an EC2 instance, where it will be processed and stored for further analysis.

Data Storage: The ingested data will be stored in S3, providing a scalable and durable storage solution. The data will be updated dynamically every 12 hours to ensure the dashboard always reflects the most current information.

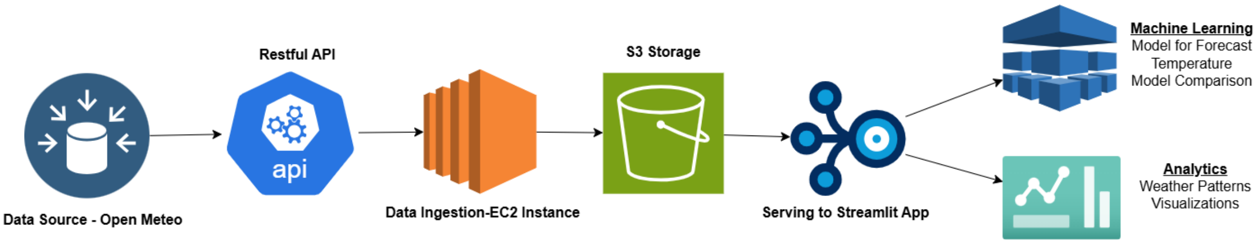
Data Transformation: The data will undergo transformation processes such as:

- **Cleaning**: Removing inconsistencies or irrelevant data to improve quality.
- **Serialization**: Converting data into formats (e.g., JSON, Parquet) for easier manipulation and storage.
- **Compression**: Reducing the data size to optimize storage and transfer speed.

Serving to Streamlit Web Application: Once the data is cleaned and transformed, it will be served to the Streamlit web-based app, where it can be displayed as real-time visualizations.

Analytics: The web-based app will display various analytics, including comparisons of weather patterns across different cities in Canada. This will include visualizations of temperature, humidity, precipitation, and more.

Machine Learning Models for Forecasting: Using machine learning models, the application will forecast weather conditions, such as predicting the next day's temperature. Multiple models will be compared to evaluate their accuracy and effectiveness.

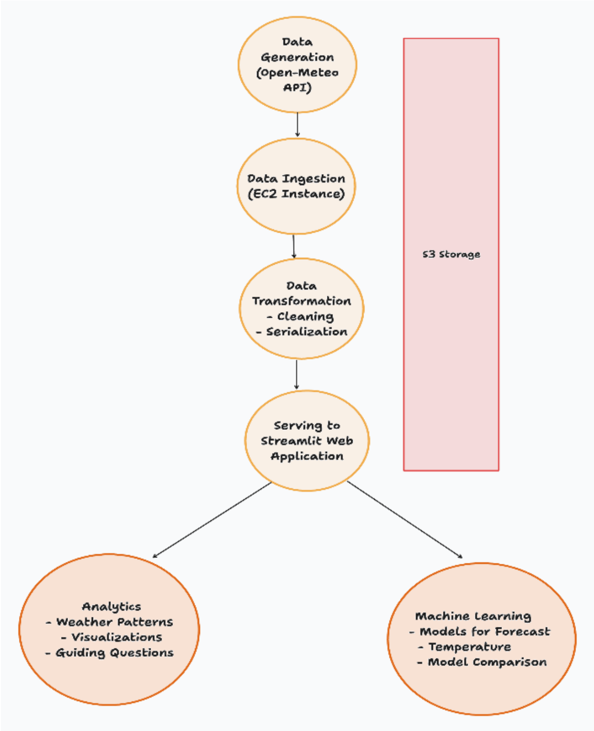


Data608_IODiagram

WORK DISTRIBUTION

Data Engineering Lifecycle	Part	Ayush	Hritvik	Satyam	Venkatesh
Data Generation	RestAPI Access	Equal	Equal	Equal	Equal
Ingestion, Transformation & Serving	EC2 Instances	Equal	Equal	Equal	Equal
Storage	S3 Buckets	Equal	Equal	Equal	Equal
Analytics	StreamLitt Dashboard	1 guiding question each	1 guiding question each	1 guiding question each	1 guiding question each

Archives


Data Generation (Open-Meteo API)
Data Ingestion (EC2 Instance)
Data Transformation
- Cleaning
- Serialization
Serving to Streamlit Web Application
Analytics
- Weather Patterns
- Visualizations
- Guiding Questions
Machine Learning
- Models for Forecast
- Temperature
- Model Comparison
S3 Storage