Implementation of Canadian Climate Action Tracker in GCP

CPSC_5207EL_01 : Intro to Cloud Technologies

Submitted by:

Ajmerry Hossain, Ahsan Md. Sajid Khan, Nafiza Anjum Khan, Bing Li, Xingbei Liu

Submitted to:

Jaspreet Bhatia

Bharti School of Engineering and Computation

Science, Engineering and Architecture

1. Abstract:

The Climate Action Tracker (CAT) report offers a comprehensive analysis of global efforts in tackling climate change. Through meticulous evaluation of national policies and emissions trajectories, it highlights the urgency of reducing greenhouse gas emissions to limit global temperature rise in accordance with the Paris Agreement. This report serves as a critical resource, emphasizing the need for immediate, ambitious actions from governments, industries, and individuals to mitigate climate change's impacts and build a sustainable future for all.

According to the globally agreed Paris Agreement, the aim of CAT is to hold warming well below 2°C, and pursue efforts to limit warming to 1.5°C [5].

CAT quantifies and evaluates climate change mitigation targets, policies and action. It also aggregates country action to the global level, determining likely temperature increases during the 21st century using the MAGICC climate model. CAT further develops sectoral analysis to illustrate required pathways for meeting the global temperature goals [5].

2. Overview:

The Canadian Climate Action Tracker, anchored on Google Cloud Platform (GCP) with emphasis on BigQuery, represents a pivotal initiative in real-time climate monitoring. Amidst a downward trend in emissions and persistent gaps between policies and targets, this project employs GCP's robust infrastructure and BigQuery's analytics prowess to deliver dynamic insights.

Key components include data integration via Google Cloud Storage, leveraging BigQuery for real-time analytics, and Google Data Studio for interactive visualizations. Security is ensured through Google Cloud IAM. Despite governmental acknowledgment of failures, the overall Climate Action Tracker (CAT) rating remains 'Highly Insufficient,' necessitating stronger targets.
The project addresses the urgency outlined by the Environment Commissioner, who stressed the need for coherent plans, highlighting policy inconsistencies such as the

Trans Mountain purchase. Canada's updated long-term strategy lacks a specific pathway and policies for achieving net-zero emissions, raising concerns about target credibility.

CAT's 'Highly Insufficient' rating underscores policy misalignment with the Paris Agreement, projecting 2°C warming under current policies. Successful implementation of planned measures could improve the rating to 'Insufficient,' though challenges persist. Emissions projections for 2030 fall short of 1.5°C compatibility, demanding accelerated policy implementation [5].

This project encapsulates a transformative initiative leveraging GCP and BigQuery, providing a comprehensive and real-time Canadian Climate Action Tracker. The integration aims to bridge policy gaps, enhance transparency, and contribute to urgent global climate action.

3. Project Objective :

The primary objective of the Canadian Climate Action Tracker is to offer a **centralized platform** for stakeholders, including policymakers, researchers, and the public, to **access real-time insights into Canada's climate actions**. Through GCP and BigQuery integration, the project aims to provide a **user-friendly interface, dynamic analytics, and visualizations, fostering data-driven decision-making**. The overarching goal is to contribute to national and global climate goals by enhancing transparency, accountability, and understanding of Canada's progress towards climate targets.

This overview sets the foundation for a transformative project that combines advanced technology with a commitment to addressing climate challenges, ultimately paving the way for informed and impactful climate action in Canada.

4. Centralized Platform Solution Architecture of CAT:

After a rigorous analysis of Climate Action Tracker's objectives, data sources, and data analytics requirements, we have outlined the following general solution architecture for the centralized CAT platform:

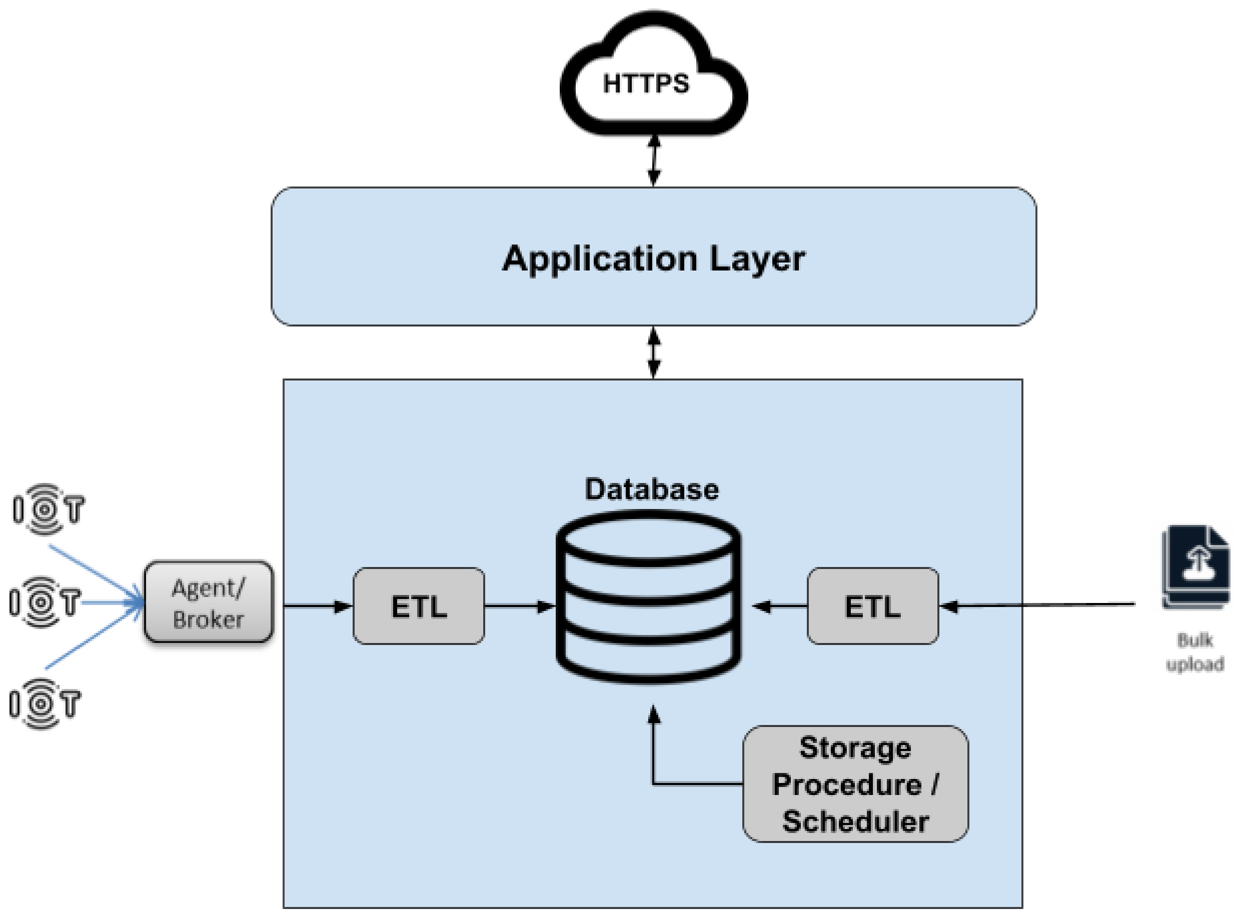


Figure: Generalized high level solution architecture for CAT platform

The overall solution architecture of the centralized CAT platform consists of the following modules:

Application Layer: This layer serves as the user interface accessible through web applications, facilitating the visualization of essential CAT-required data. It includes a dashboard that displays comprehensive figures and facts, enabling efficient tracking of CAT's ultimate goal.

Database: The database module stores both historical and current data relevant to the CAT platform.

Extract, Transform, and Load (ETL): This module comprises multiple processes responsible for gathering data from diverse sources. The project primarily focuses on two major data sources: Bulk-Upload and Real-Time data collection. Since, most of the data from data sources were clean and structured [1] [2], therefore we don’t have to perform transform operations to clean and restructure data.

5. Comparison of Different Cloud Platform

In this section, we have discussed the service offerings provided by three major cloud service providers: Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP).

Microsoft Azure offers a comprehensive suite of cloud services that include computing power, storage solutions, databases, networking, and an array of specialized tools for analytics, artificial intelligence (AI), and Internet of Things (IoT) applications.

Amazon Web Services (AWS) provides a vast selection of cloud services, encompassing computing, storage, databases, machine learning, and AI capabilities. AWS is well-known for its scalability, flexibility, and diverse range of tools catering to businesses of varying sizes and industries.

Google Cloud Platform (GCP) delivers a suite of cloud computing services that include computing, storage, databases, machine learning, and data analytics. GCP is recognized for its robust data processing, AI, and machine learning capabilities, and its extensive network infrastructure.

Each of these cloud service providers has its strengths and specialties, catering to different business needs and offering a variety of tools and services to support diverse applications and workloads in the cloud.

Compute services, database and storage services offer by the three service providers as following [7],

Compute Services

SERVICE	AWS	AZURE	GCP
VM (Compute Instance)	EC2 (Elastic Compute)	Azure Virtual Machine	Google Compute Engine
PaaS	AWS Elastic Beanstalk	App Service	Google App Engine
Container	AWS Elastic Container/Kubernetes Service	Azure Kubernetes Service (AKS)	Google Kubernetes Engine
Serverless Functions	AWS Lambda	Azure Function	Google Cloud Functions

Database & Storage Services

SERVICE	AWS	AZURE	GCP
RDBMS (Multiple Database Types – SQL, MySQL, etc..)	AWS RDS	Azure SQL/ Database for MySQL/PostgreSQL	Cloud SQL
NoSQL	DynamoDB, Simple DB	Azure Cosmos DB, Table Storage	BigTable, Cloud Datastore
Object Storage	S3 (Simple Storage Service)	Blob Storage	Google Cloud Storage
File Storage	Elastic File System	Azure File Storage	Google Filestore
Archive Storage	Amazon Glacier	Azure Archive Storage	Google Storage (Archive Storage)
Data Warehouse/Data Lake	Amazon Redshift	Azure Synapse Analytics	Google BigQuery

All three providers AWS, Azure, Google Cloud offer price discounts if we commit to using their services at least for one year. This pricing model is called Reserved Instances in AWS, Reserved Savings in Azure, and Committed use discounts in Google Cloud [6].

General purpose

Cloud provider	Instance type	Price	Discount
AWS	t4g.xlarge	$0.084	41%
Azure	B4ms	$0.1118	32%
Google Cloud Platform	e2-standard-4	$0.095092	37%
Oracle	VM.Standard3.Flex	$0.1038	1%

Compute optimized

Cloud provider	Instance type	Price	Discount
AWS	c6a.xlarge	$0.1010	38%
Azure	F4s v2	$0.1143	32%
Google Cloud Platform	c2-standard-4	$0.14072	41%
Oracle	VM.Optimized3.Flex	$0.1198	1%

General-purpose instances with a 1-year commitment receive quite similar discount rates in AWS and Azure. Still, AWS offers a cheaper alternative.

In both general purpose and compute optimized instances, Google Cloud Platform offers the biggest discounts [6].

6. Why GCP Platform?

There are several factors that we have considered while choosing GCP to implement centralized CAT. Though, the decision to choose GCP over Azure or AWS (or vice versa) depends on a thorough evaluation of factors, but for GCP we have found following features are more robust and flexible compared to other platforms,

- ➔ Strong Focus on Data Analytics and Machine Learning
- ➔ Networking and Infrastructure
- ➔ Innovative Technologies and Open Source Embrace
- ➔ Cost and Pricing Models
- ➔ Specialized Services
- ➔ Environmental Sustainability

7. CAT Platform Process Flow in GCP

To implement the centralized Canadian Climate Action Tracker (CAT) on the GCP platform, we have outlined the following process flow within the GCP platform. This flow aligns with the high-level solution architecture described in section 4.

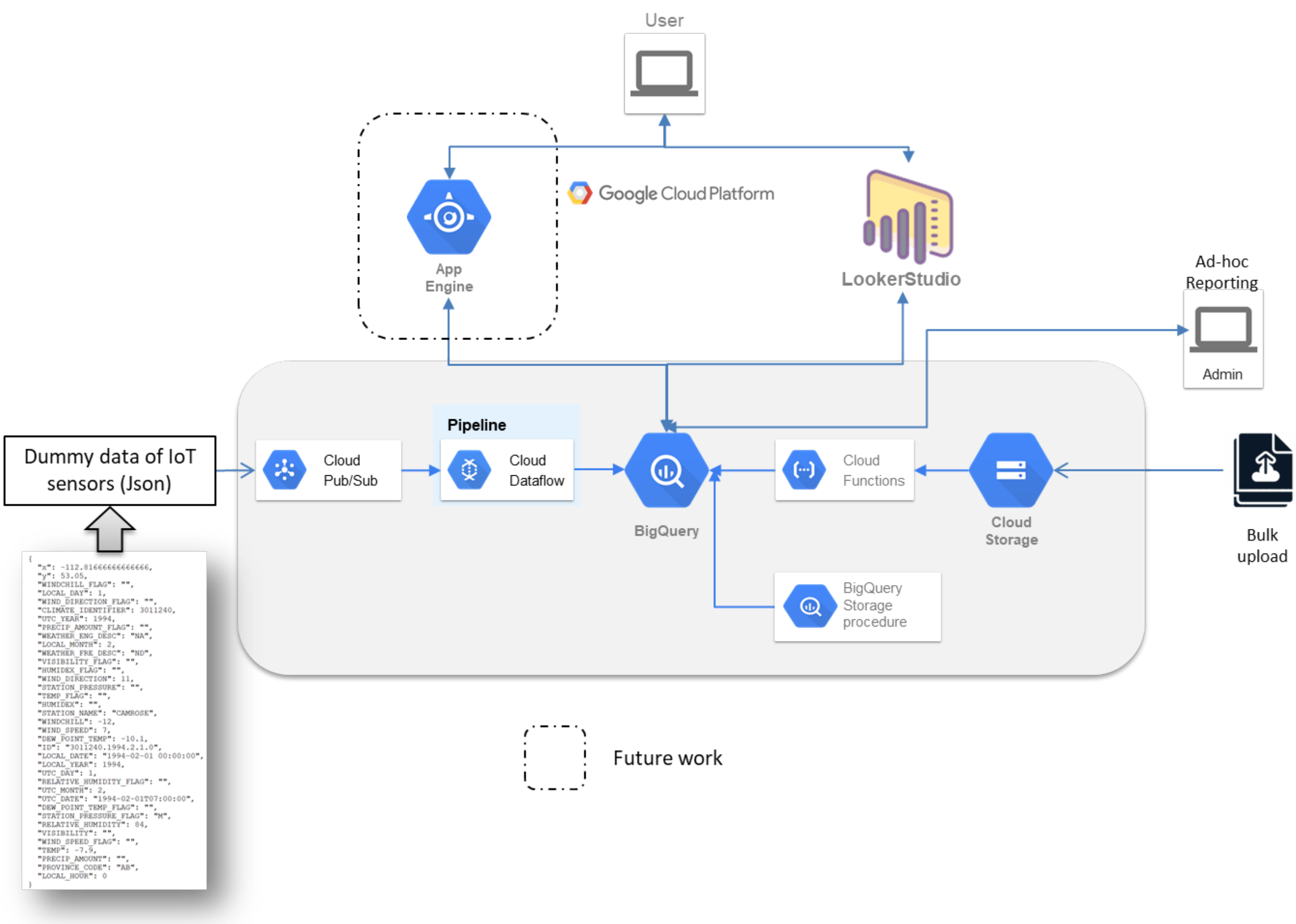


Figure: High level Process-Flow Diagram of Centralized CAT within the GCP Platform

The process flow of the centralized CAT in GCP has been further modularized into the following functional modules:

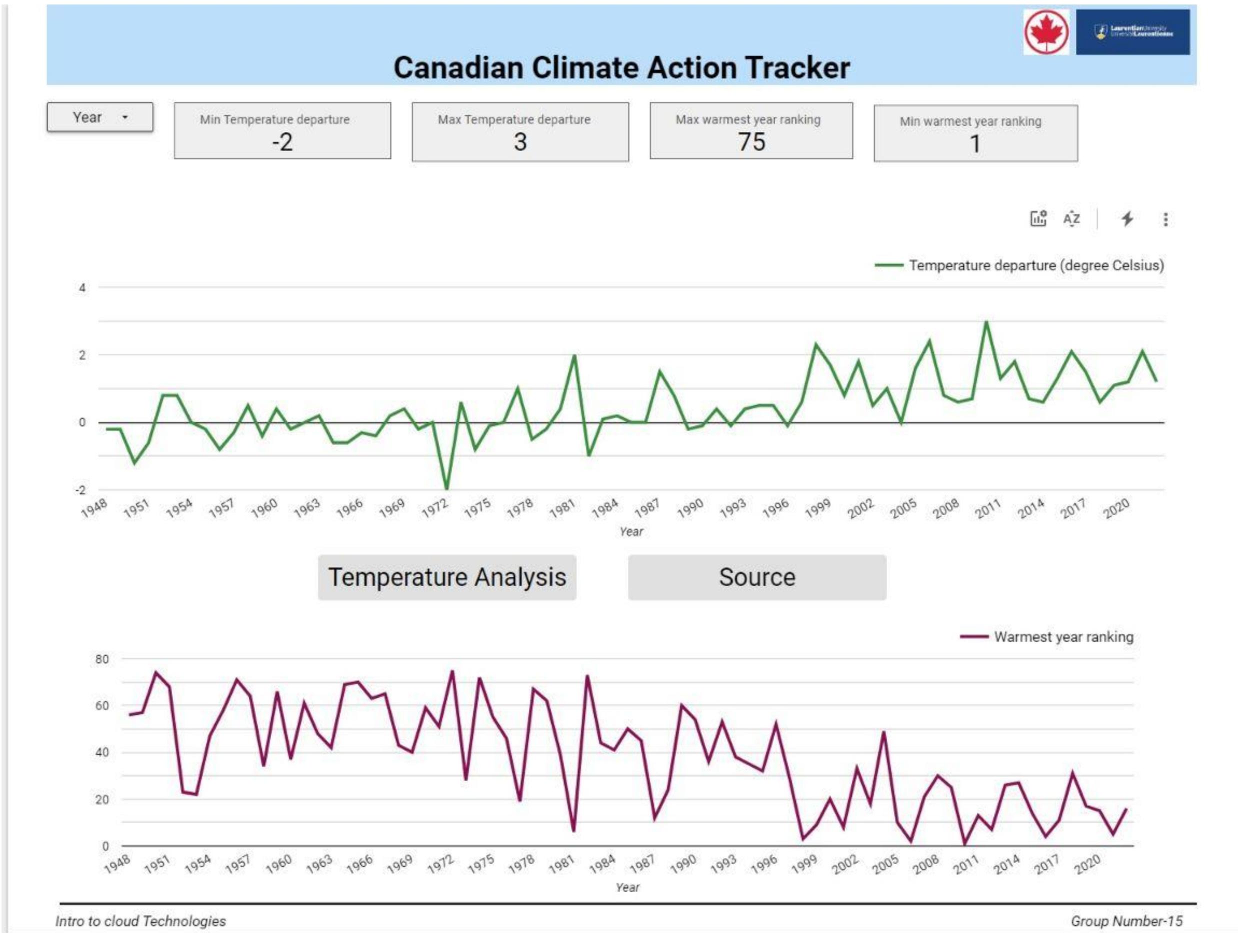
- Dashboard
- Data Collection & Processing
    - Bulk Upload
    - Real-time Data Collection
    - Scheduler/Storage Procedure
- Ad-hoc Reporting and Analysis

8. Implementation of CAT platform in GCP

In this section, we will describe the implementation steps for all the functional modules mentioned in section 7.

8.1 Dashboard:

This web-based dashboard for Climate Action Tracker facilitates the visualization of temperature change data through interactive charts, graphs, and customizable displays. It provides real-time insights and comparative analysis of temperature trends over different time frames and regions. The dashboard allows users to focus on specific temperature parameters and explore detailed data for deeper insights, aiding in understanding patterns and correlations crucial for climate analysis and decision-making.



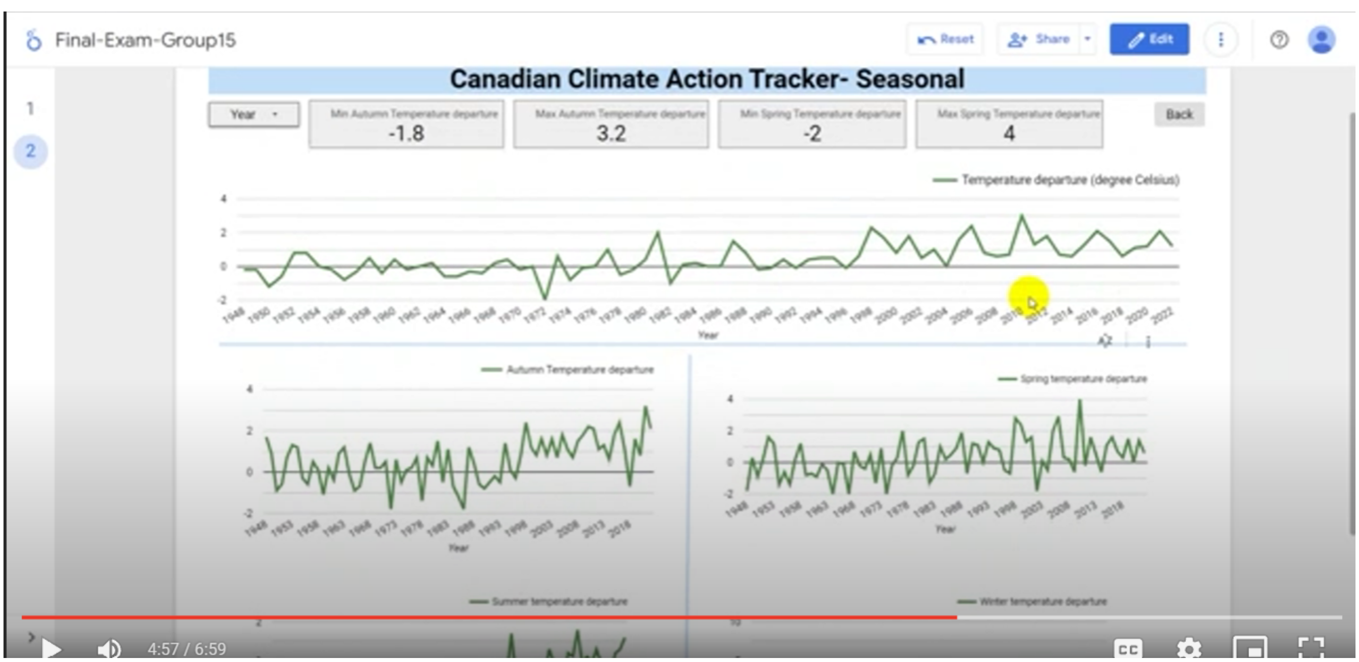


Figure : Dashboard view of CAT through ‘Lookerstudio’

To connect LookerStudio to BigQuery, we have performed following steps:

Set Up a Google Cloud Platform Project:

Initially we have created a GCP platform under a project name “g15project” which consists of BigQuery with a defined dataset named “CAT_Bulk”, Cloud Storage with a bucket name “cat_group15_bulk”. Created all the required access to access data according to the roles and policy.

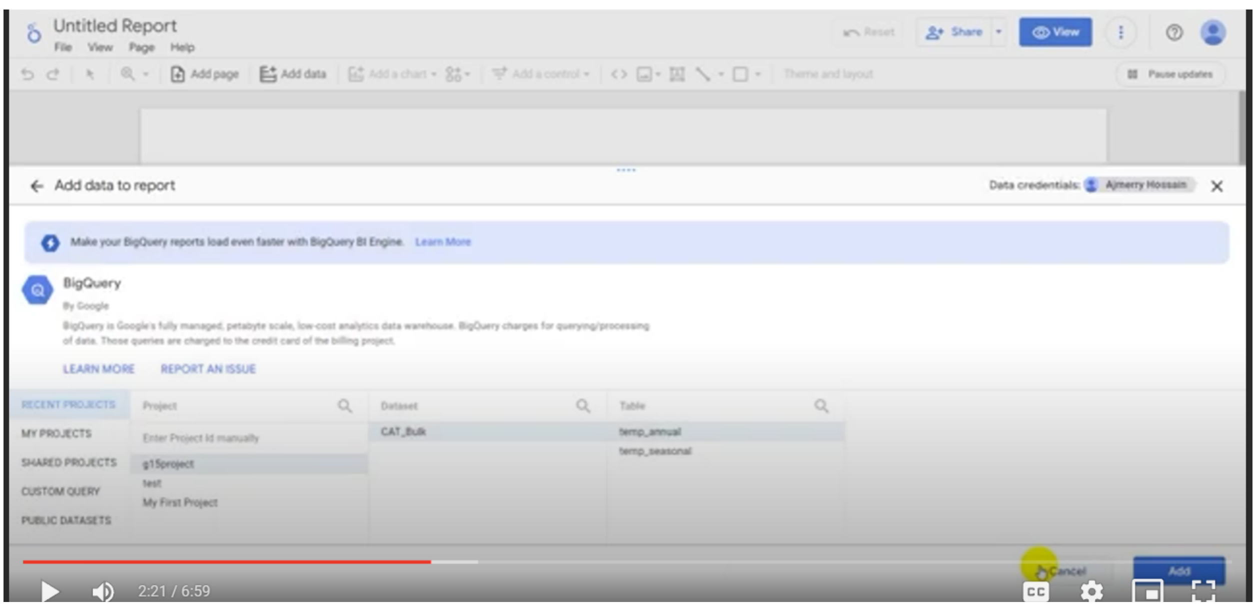
Connect Looker Studio with BigQuery:

We connect the Google Looker studio using the same credential of the google cloud project.

Report Creation:

Here we are following these steps:

1. Click on the Blank report for creating a new report.
2. We need to connect with the data here as we are showing the data from the Bigquery, so we selected BigQuery.
3. Then we need to choose the dataset and table under this BigQuery. Here we select CAT_Bulk and select the Temperature change annually Table from this dataset.



4. After selecting the dataset and table we need to select data to add in the desired report format.
5. To add another table we need to repeat the same procedure1-4.
6. After adding the data in the dashboard platform we can add charts, scoreboard other visualization forms to show the report.

7. Here we used line charts to show year wise temperature departure annually and seasonally. For visualizing data from multiple tables we used the Data blend option.
8. We also used filter and scoreboard visualization to filter the data and show some highlighting data from our database.
9. We used a button here to go to another page of the dashboard and another button to go to a link of the data source.
10. So from this dashboard we were able to show the trend of the year wise temperature data. Therefore, we can perform analyses, and generate visualizations based on the connected BigQuery data.

Video demonstration:
https://drive.google.com/file/d/1C0aBq-xnSnsC5XSGh9NGOzU1qCgoODHG/view?usp=sharing

8.2 Data Collection & Processing:

In this section we have described how we collect bulk data, real-time data and perform a scheduler inside the BigQuery module to process and summarize data for meaningful visualization.

Bulk Upload

For bulk-upload we have created an event-triggered process where a file uploaded to Cloud Storage triggers a Cloud Function that processes the data and inserts it into a BigQuery table “bulkdata”. To implement this we have performed following steps,

Set Up a Google Cloud Platform Project:

Initially we have created a GCP platform under a project name “g15project” which consists of BigQuery with a defined dataset named “CAT_Bulk”, required table to insert data, Cloud Storage with a bucket name “cat_group15_bulk”. Created all the required access to access data according to the roles and policy.

Create a Cloud Function:

We have written a Cloud Function using the python programming language that will be triggered by the Cloud Storage event whenever the file will be uploaded or finalized. This function contains following python code to process the data and insert it into the BigQuery tables,

```
import pandas as pd

from pandas.io import gbq

from google.cloud import bigquery


'''

Python Dependencies to be installed


gcsfs

fsspec

pandas

pandas-gbq


'''


def hello_gcs(event, context):

    """Triggered by a change to a Cloud Storage bucket.

    Args:

    event (dict): Event payload.

    context (google.cloud.functions.Context): Metadata for the event.

    """


    lst = []

    file_name = event['name']

    table_name = file_name.split('.')[0]
```

```
# Event,File metadata details writing into Big Query
dct={
'Event_ID':context.event_id,
'Event_type':context.event_type,
'Bucket_name':event['bucket'],
'File_name':event['name'],
'Created':event['timeCreated'],
'Updated':event['updated']
}
lst.append(dct)
df_metadata = pd.DataFrame.from_records(lst)
df_metadata.to_gbq('CAT_Bulk.bulkdata_meta',
        project_id='g15project',
        if_exists='append',
        location='us')

# Actual file data , writing to Big Query
df_data = pd.read_csv('gs://' + event['bucket'] + '/' + file_name)

df_data.to_gbq('CAT_Bulk.bulkdata',
        project_id='g15project',
        if_exists='append',
        location='us')
```

The above set of code will process and insert data into two tables, one is metadata of the uploaded file and another will load raw data consisting of temperature as well as other relevant information for CAT.

Deploy the Cloud Function:

We have deployed the above Cloud Function to Google Cloud Platform, associating it with the Cloud Storage bucket and specifying the trigger event when a new file is uploaded and finalized in the bucket.

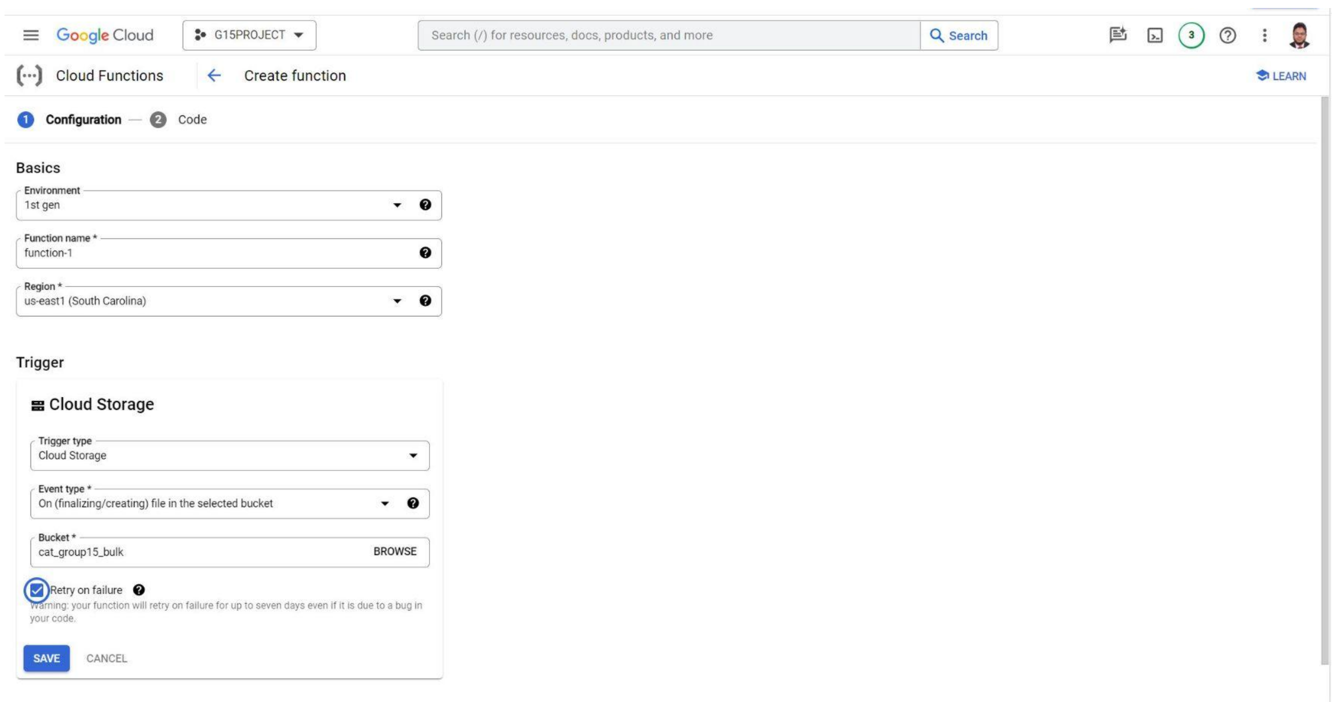
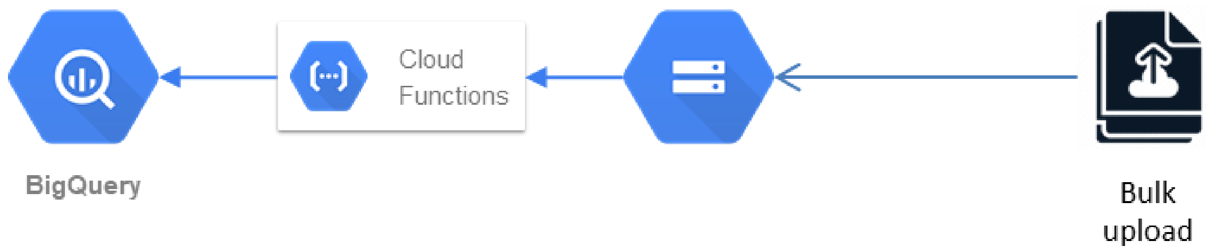


Figure: Configuration of Cloud Function.

The overall GCP process flow of bulk upload is as follows,

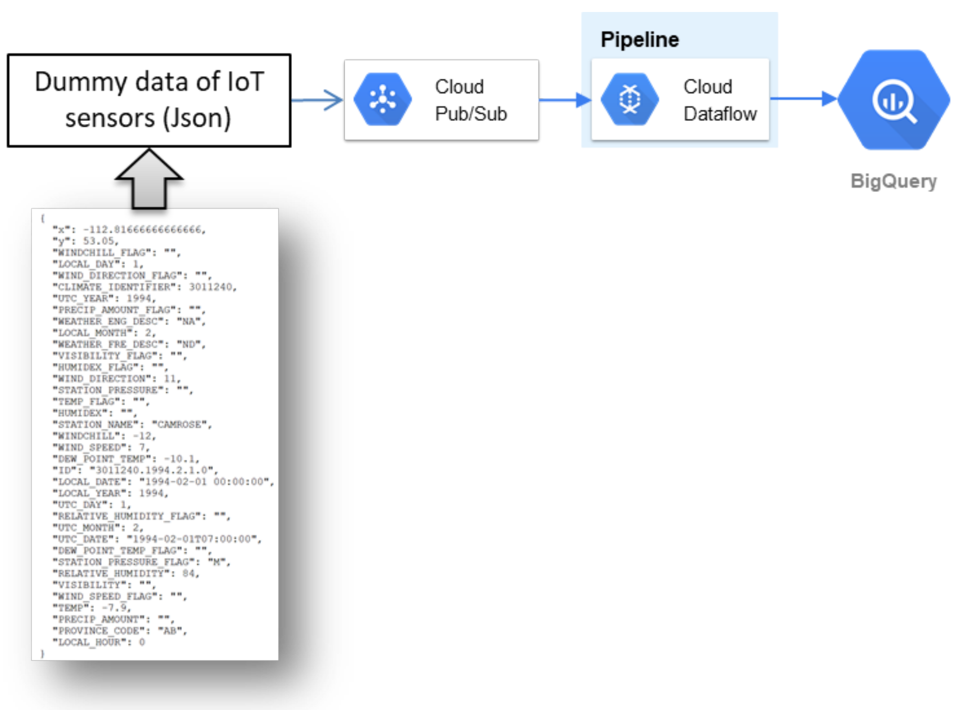


Video demonstration :

https://drive.google.com/file/d/19jixM6p5ET7riJzbf6o5d2FIncSTghum/view?usp=drive_link

Real-Time Data Collection:

The overall process flow of implemented real-time data collection is as follows,



In this process once Pub/Sub will receive/generate a message it will be processed by Dataflow and will be inserted into the BigQuery table. To implement this we have performed the following steps,

Set Up a Google Cloud Platform Project:

Initially we have created a GCP platform under a project name “g15project” which consists of BigQuery with a defined dataset named “CAT_Bulk”, required table to insert data, Cloud Storage with a bucket name “cat_group15_bulk”. Created all the required access to access data according to the roles and policy.

Set up Pub/Sub Topic:

We have created a pub/sub topic for real-time data streaming.

Create a Dataflow Pipeline:

We implemented a Dataflow pipeline using a predefined job template named 'Pub/Sub topic to BigQuery'. We configured the template by filling in all the required fields, such as associating it with the Pub/Sub topic name, specifying the destination table in BigQuery where the processed data would be inserted, and defining a temporary location in Cloud Storage. After configuring all the parameters, we executed the job, and the entire process was set up to ingest data into BigQuery through Pub/Sub.

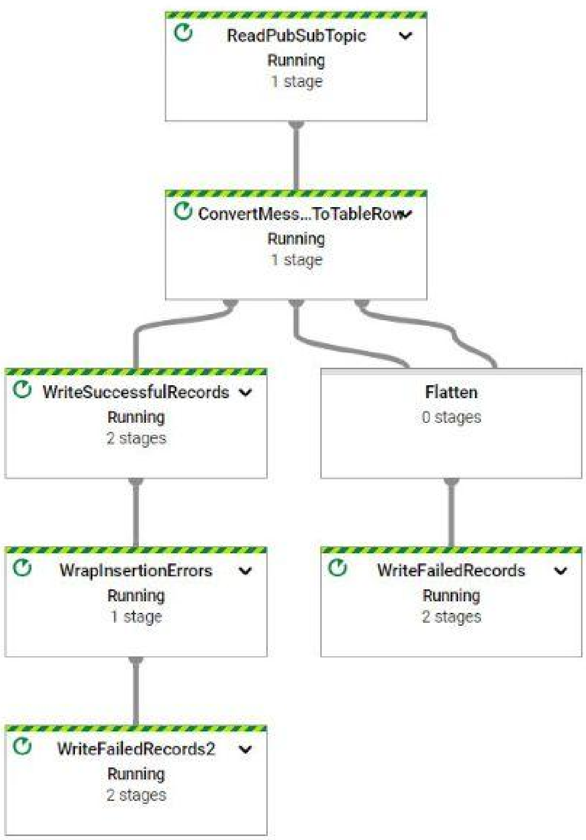


Figure : Running Dataflow job for real-time data collection trough Pub/Sub

Video Demonstration : https://drive.google.com/file/d/1Z0BmZb8eeJAHeMhF3IBs9Ma-Pk7iHTLG/view?usp=drive_link

Scheduler / Storage Procedure Job:

This task is implemented within BigQuery. It is necessary to process hourly and daily data to generate summary data as per the policy requirements of CAT. This can be achieved by executing a single SQL command through a scheduled job or by using a stored procedure. For this project, we have implemented a scheduler by directly editing SQL commands within the BigQuery studio and setting a schedule to run that SQL command.

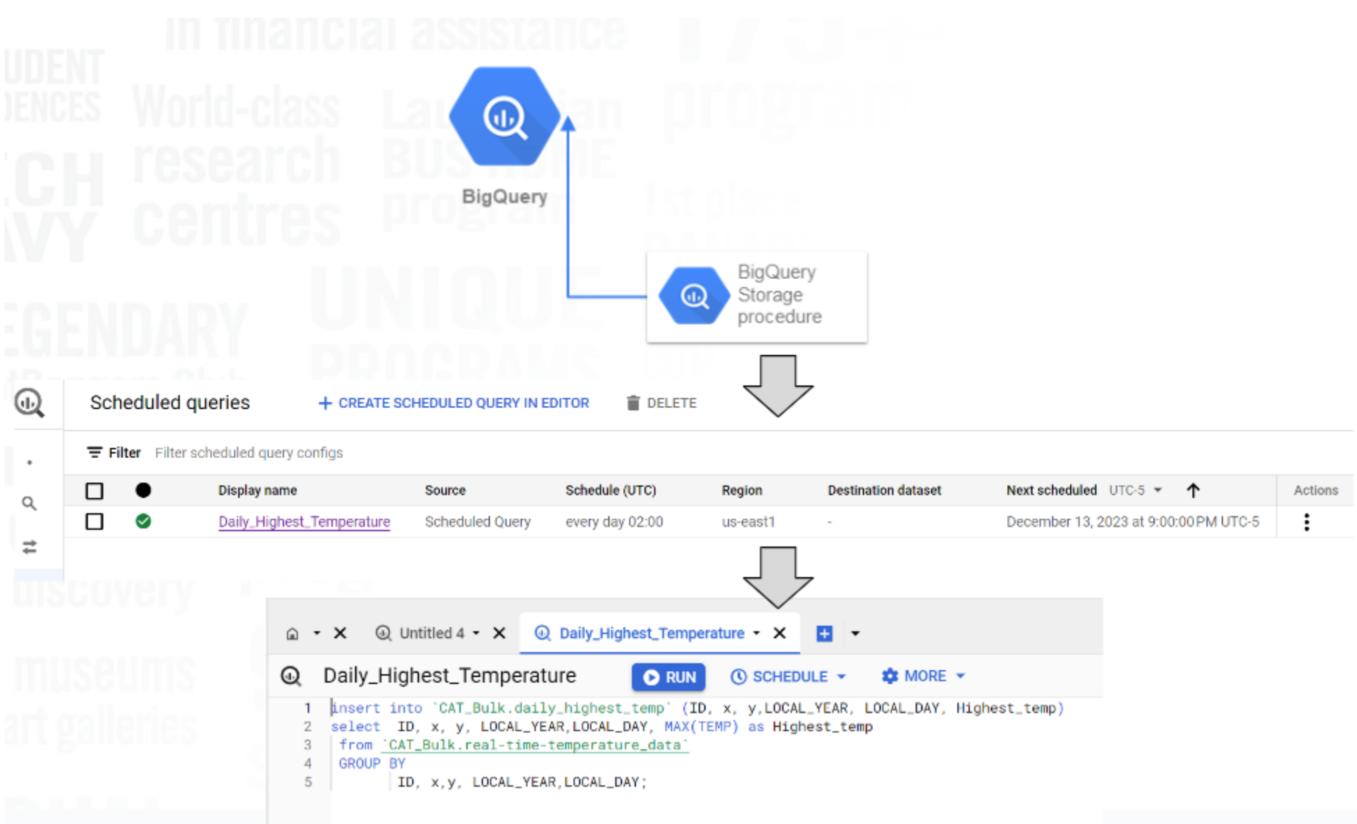


Figure: Schedule job set-up for daily highest temperature calculation

Video Demonstration : https://drive.google.com/file/d/17fPzZFYY7IJ4t74mWIdMmlkOsLpHcVVu/view?usp=drive_link

8.3 Ad-Hoc Reporting and Analysis

BigQuery is a serverless service offered by GCP. It offers a robust SQL-based querying environment and seamlessly integrates with various tools and platforms specialized in visualization, trend analysis, and predictive modeling. In cases where certain reports are not covered by the dashboard and customization is required, BigQuery supports various functionalities to address ad-hoc requirements. Those features are as follows,

- Integration with Data Visualization Tools
- BigQuery ML for Predictive Modeling
- SQL for Trend Analysis
- Advanced Analytics Extensions (Standard SQL)
- Real-time Data Analysis for Trend Monitoring
- Geospatial Visualization

9. Acknowledgements

In the completion of this project, we would first like to thank every team member for their hard work and mutual understanding. Although everyone has their own personal work and is very busy, they can still find time to discuss together from topic selection to the final project. In addition, although there are often different understandings among the group members from the beginning of the topic selection, every time a dispute arises, everyone chooses to discuss it together and make the final best choice. Everyone is very friendly, and very outstanding, and willing to selflessly share their learning experience and knowledge. Thank you to everyone for their efforts. At the same time, everyone’s communication skills and collaboration abilities have been exercised.I believe this valuable cooperation experience is a treasure for everyone.

Secondly, we would like to thank the distinguished professor, Jaspreet Bhatia, for her valuable guidance. Her help in our learning process is the key to the success of our project. Her clear explanation of the various functions of GCP in class not only promoted our understanding, but also provided great help for our project development. We especially thank the professor for each assignment. These assignments give us a deeper understanding of what we have learned in class. This approach not only exercises our technical application ability, but also directly promotes the progress of our project.

10. Recommendation and Future Plan

In charting the course for the future of the Canadian Climate Action Tracker within the Google Cloud Platform (GCP), a set of strategic recommendations is proposed to elevate its functionality and impact. First and foremost, the implementation of Google App Engine is advised to ensure the scalability and adaptability necessary to meet the evolving demands of climate data processing. Concurrently, the integration of IoT sensor devices will enable real-time data collection, fostering a more dynamic and accurate understanding of environmental shifts. Leveraging the power of BigQuery ML for predictive modeling stands as a pivotal step forward, enabling the Tracker to anticipate climate trends and inform proactive decision-making. Moreover, the inclusion of geospatial visualization tools, such as Google Maps Platform, enhances user engagement and comprehension by providing an interactive exploration of regional climate data. By synergizing these recommendations, the future of the Canadian Climate Action Tracker in GCP is envisioned as a comprehensive, agile, and data-driven platform at the forefront of climate monitoring and policy formulation. This strategic roadmap underscores a commitment to technological innovation, ensuring that the Tracker remains a vital tool in the global effort to address climate change.

11. References

[1]. Climate Data Canada. (2023, January 26). *Location — Climate Data Canada*.

https://climatedata.ca/explore/location/?loc=JBVRD&location-select-temperature=tx_max&location-select-precipitation=r1mm&location-select-other=frost_days

[2]. Canada, E. a. C. C. (n.d.). *Climate data extraction tool*.

https://climate-change.canada.ca/climate-data/#/downscaled-data

[3]. *Index of climate*. (n.d.). https://dd.weather.gc.ca/climate/

[4]. Canada, E. a. C. C. (n.d.-b). *Climate data extraction tool*.

https://climate-change.canada.ca/climate-data/#/hourly-climate-data

[5]. *The Climate Action Tracker*. (n.d.). https://climateactiontracker.org/about/

[6]. Gil, L. (2023, October 27). *Cloud Pricing Comparison: AWS vs. Azure vs. Google Cloud Platform in 2023*. CAST AI – Kubernetes Automation Platform. https://cast.ai/blog/cloud-pricing-comparison-aws-vs-azure-vs-google-cloud-platform/

[7]. Wickramasinghe, S. (n.d.). *AWS vs Azure vs GCP: Comparing The Big 3 Cloud Platforms*. BMC Blogs. https://www.bmc.com/blogs/aws-vs-azure-vs-google-cloud-platforms/