RV Educational Institutions ®
RV College of Engineering ®

Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi, Accredited By NAAC, Bengaluru And NBA, New Delhi

Go, change the world

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VALIDATING THE ANONYMIZED DATA USING AGGREGATE QUERIES IN PUBLIC CLOUD

INTERNSHIP REPORT

Submitted by,

HARDIK HIRAMAN PAWAR **1RV21CS046**

MOHAMMED RAZA **1RV21CS093**

Under the guidance of

Dr. Sowmyarani C N
Associate Professor
Dept. of CSE
RV College of Engineering

Veena Gadad
Assistant Professor
Dept. of CSE
RV College of Engineering

RESEARCH PROJECT: "Data Anonymization Techniques for Mitigation of Privacy Attacks in Privacy Preserving Data Publishing".

In partial fulfilment for the award of degree
Bachelor of Engineering in
Computer Science and Engineering 2022-2023

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Internship work titled **'VALIDATING THE ANONYMIZED DATA USING AGGREGATE QUERIES IN PUBLIC CLOUD'** is carried out by **HARDIK HIRAMAN PAWAR (1RV21CS046) and MOHAMMED RAZA (1RV21CS093)** who are bonafide students of RV College of Engineering, Bengaluru, in partial fulfilment for the award of degree of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belagavi during the year 2022-2023 for the Summer Internship-I (21CSI310). It is certified that all corrections/suggestions indicated for the Internship have been incorporated in the Internship project report. The Internship report has been approved as it satisfies the academic requirements in respect of internship work prescribed by the institution for the Research Project: Data Anonymization Techniques for Mitigation of Privacy Attacks in Privacy Preserving Data Publishing.

Signature of Guide 1
Dr. Sowmyarani C N

Signature of Head of the Department
Dr. Ramakanth Kumar P

Signature of Guide 2
Veena Gadad

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, HARDIK HIRAMAN PAWAR and MOHAMMED RAZA students of second semester B.E., department of CSE, RV College of Engineering, Bengaluru, hereby declare that the Internship Work titled **'VALIDATING THE ANONYMIZED DATA USING AGGREGATE QUERIES IN PUBLIC CLOUD'** has been carried out by us and submitted in partial fulfilment for the award of degree of **Bachelor of Engineering** in **Computer Science and Engineering during** the year 2022-23.

Further we declare that the content of the report has not been submitted previously by anybody for the award of any degree or diploma to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RVCE will be the property of RV College of Engineering, Bengaluru and we will be one of the authors of the same.

Place: Bengaluru

Date:

Name **Signature**

1. HARDIK HIRAMAN PAWAR (1RV21CS046)

2. MOHAMMED RAZA (1RV21CS093)

ACKNOWLEDGEMENT

We are indebted to our guide, Dr. Sowmyarani C N (Associate Professor) and **Veena Gadad (Assistant Professor), Dept of CSE** for his/her wholehearted support, suggestions and invaluable advice throughout our project work and also helped in the preparation of this report.

Our sincere thanks to **Dr. Ramakanth Kumar P.**, Professor and Head, Department of Computer Science and Engineering, RVCE for his support and encouragement.

We express sincere gratitude to our beloved Principal, **Dr. K. N. Subramanya** for his appreciation towards this project work.

We thank **VGST** for providing us an opportunity to work on this project and enhance our skills.

We thank all the **teaching staff and technical staff** of Computer Science and Engineering department, RVCE for their help.

Lastly, we take this opportunity to thank our **family** members and **friends** who provided all the backup support throughout the project work.

TABLE OF CONTENTS

Abstract

In this work, we present a methodology for storing and analyzing large datasets using Google Cloud Platform (GCP). We have chosen GCP as the online cloud database to store the original and masked datasets produced after applying our recently developed algorithm based on k & l-e diversity, called Incremental Diversity. This provides access to the dataset to the general public and any researcher who wishes to perform exploratory data analysis on the data or run a SQL query on the masked dataset via GCP's built-in BigQuery service. By choosing GCP as the online cloud database, we take advantage of GCP's ability to store large amounts of data in the cloud, which allows for easy access and retrieval of the data. We also discuss the use of BigQuery, a fully-managed, serverless data warehouse that enables super-fast SQL queries using the processing power of Google's infrastructure. We provide step-by-step instructions for creating a GCP bucket and storing datasets, as well as guidelines for sharing the dataset with others. The internship work report also contains the link to the GitHub repository containing the source code, dataset and query links for the GCP Bucket where the datasets are stored.

1. Introduction

Google Cloud Platform (GCP) is a cloud computing platform and infrastructure created by Google. It provides a variety of services including computing power, storage and application services, as well as machine learning and internet of things (IoT) capabilities. GCP allows users to build, test and deploy applications on Google's highly-scalable and reliable infrastructure. Additionally, GCP offers a range of tools for managing and analyzing data, as well as for security and compliance. Some of the most popular services on GCP include:

- Google Compute Engine (GCE), which provides virtual machines that run in Google's data centers
- Google Kubernetes Engine (GKE), which allows users to easily deploy and manage containerized applications
- Google Cloud Storage, which provides object storage for unstructured data
- Google BigQuery, which is a fully managed, cloud-native data warehouse for big data analytics
- Google Cloud AI Platform, which provides a variety of machine learning services, including TensorFlow and the AI Platform Training and Prediction service.

All of the services on GCP are designed to work seamlessly with each other, allowing developers to easily build and scale applications. GCP also offers a range of security and compliance features, such as encryption and access controls, to help customers keep their data safe.

It's also worth mentioning that Google Cloud Platform is in constant competition with other cloud providers like Amazon Web Services (AWS) and Microsoft Azure. GCP's strengths lie in Big Data and Machine Learning, while AWS is most commonly used for enterprise applications and Azure is used by many companies due to its strong hybrid capabilities.

2. Literature Survey

Google Cloud Platform is high-performance infrastructure for cloud computing, data analytics and machine learning. [1] discusses high-level architecture of GCP and its effect on application development and deployment. Also, it goes through exercises on utilising various services in GCP, focusing on computation and storage services.

[2] talks about deployment of an application using GCP. Cloud computing is an emanating technology being utilized in every field. Traditional Organisation adopts IT infrastructure, which is not expandable according to their needs. So, organizations are moving towards cloud for boosting their work and minimizing cost. Cloud computing is used in deployment of hospital management system.

[3] gives an overview of GCP. It reviews the important elements of GCP and introduces computing in google cloud. It also discusses the computing services of google cloud.

[4] describes the services offered by GCP for Internet of Things (IoT). IoT focuses on connecting the real-world build of equipment, machines, and gadgets to the virtual world of Internet for the sake of allying devices with each other producing details from the collected data. Gadgets have little power of computing and capacity of storing data. Google cloud computing has very large capability of storing data and computing. Thus, the coalition between IoT and cloud computing can be a good solution.

[5] showcases GCP Model, a deduced formal specification of GCP. GCP permits builders and designers to utilize its services by retrieving Application Programming Interfaces (APIs). But these APIs are defined only informally. The builder or designer would not recognize accurately the posture of the provider. The writings of API are in English prose and thus have many disadvantages. This GCP Model helps in avoiding chaos and depicts resources without complexity.

[6] describes a logistics information management system based on Google cloud computing platform. With the evolution of cloud computing, technical materials can be stored on the web server. Data can be transferred from computer to web server and vice versa. From this view, cloud computing as a magnificent platform for the logistics system, provides information for

clients, issues a comfortable payment method, and supervises and takes care of client data. And it can aid clients minimize their investments.

[7] tells how elevated performance computing in the GCP can be used in an urgent and emergency circumstances to operate huge traffic data. This proposal gives a solution to an urgent demand for handling calamities utilising large data processing and good performance computing.

Rest of the report discusses the methodology to create buckets on GCP and store datasets, and result.

3. Methodology

We have chosen GCP as the online cloud database to store the original and masked datasets produced after applying our recently developed algorithm based on k & l-e diversity, called Incremental Diversity. This provides access to the dataset to the general public and any researcher who wishes to perform exploratory data analysis on the data or run a SQL query on the masked dataset via GCP's built in BigQuery service, which can help to further advance research in this field.

By choosing GCP as the online cloud database, we are taking advantage of GCP's ability to store large amounts of data in the cloud, which allows for easy access and retrieval of the data. This is particularly useful for datasets that are large and complex, such as the original and masked datasets produced by the Incremental Diversity algorithm.

BigQuery is a fully-managed, serverless data warehouse that enables super-fast SQL queries using the processing power of Google's infrastructure. It allows to store and query massive datasets by creating a SQL-like syntax for querying data. BigQuery allows to share data with others, it is fully managed, which means that there is no need for infrastructure provisioning, setup, or configuration.

The following are the steps to create your very own bucket on GCP and store datasets.

1) Click on the **View All Products** on the left menu.

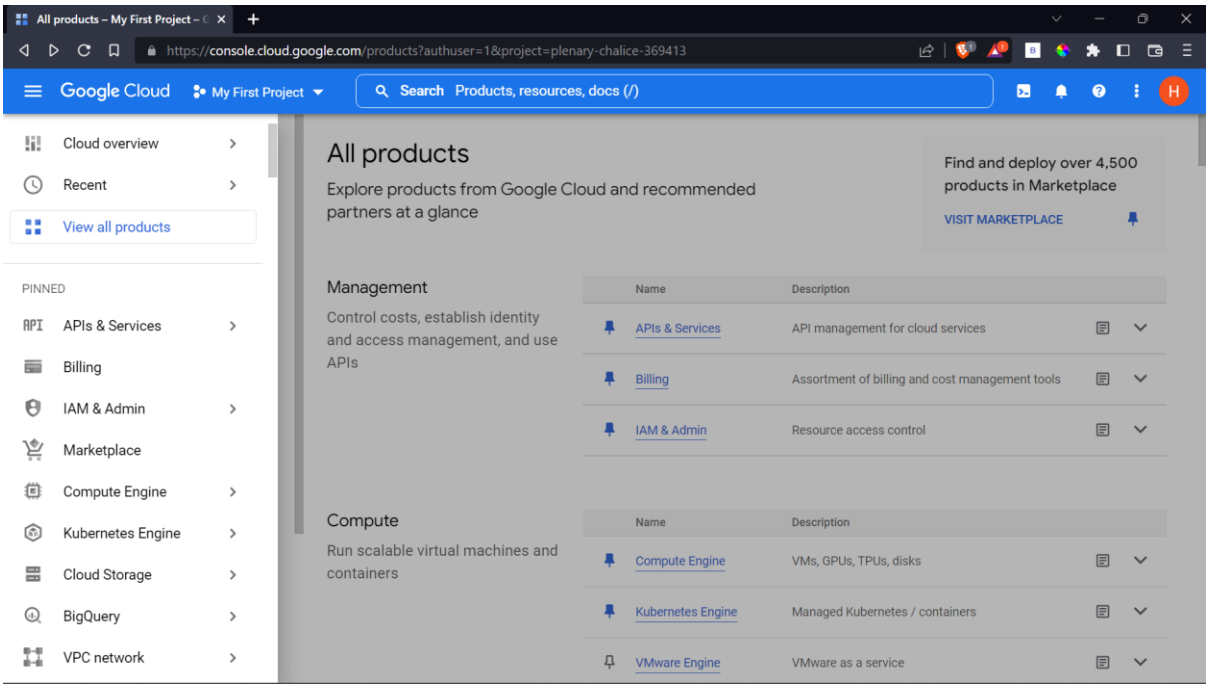

Fig 3.1: View All Products Screen

2) Click on **Cloud Storage**

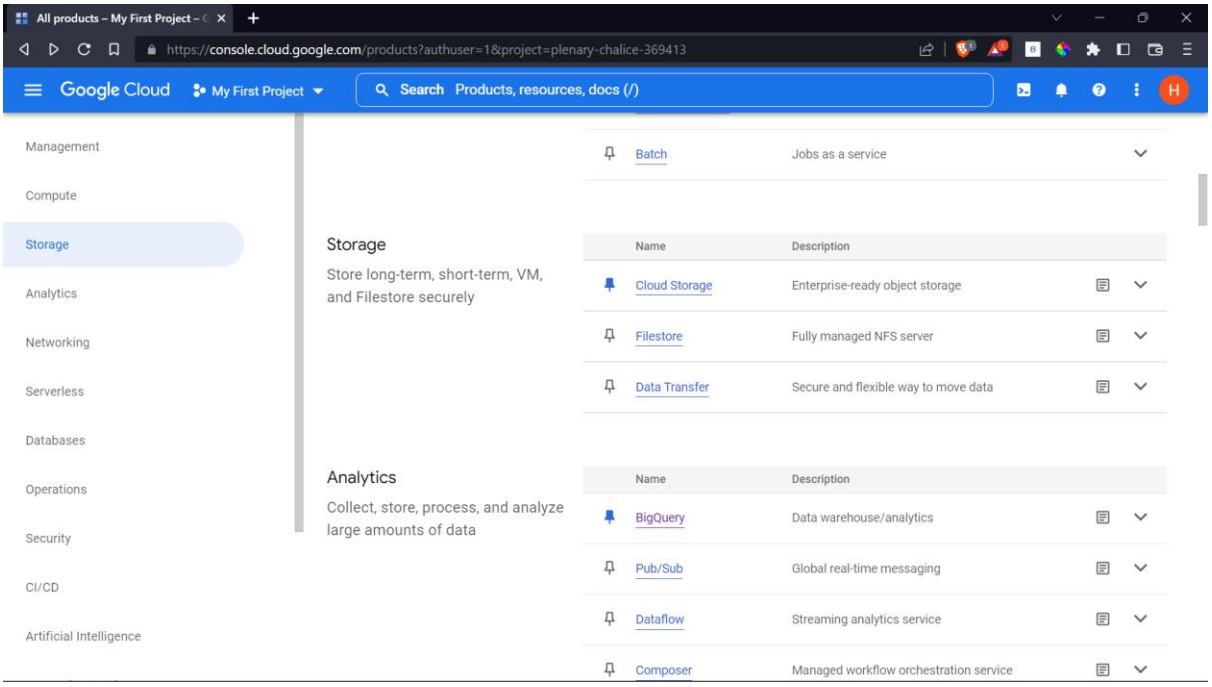

Fig 3.2: Cloud Storage Screen

3) Click on the **Create** Button

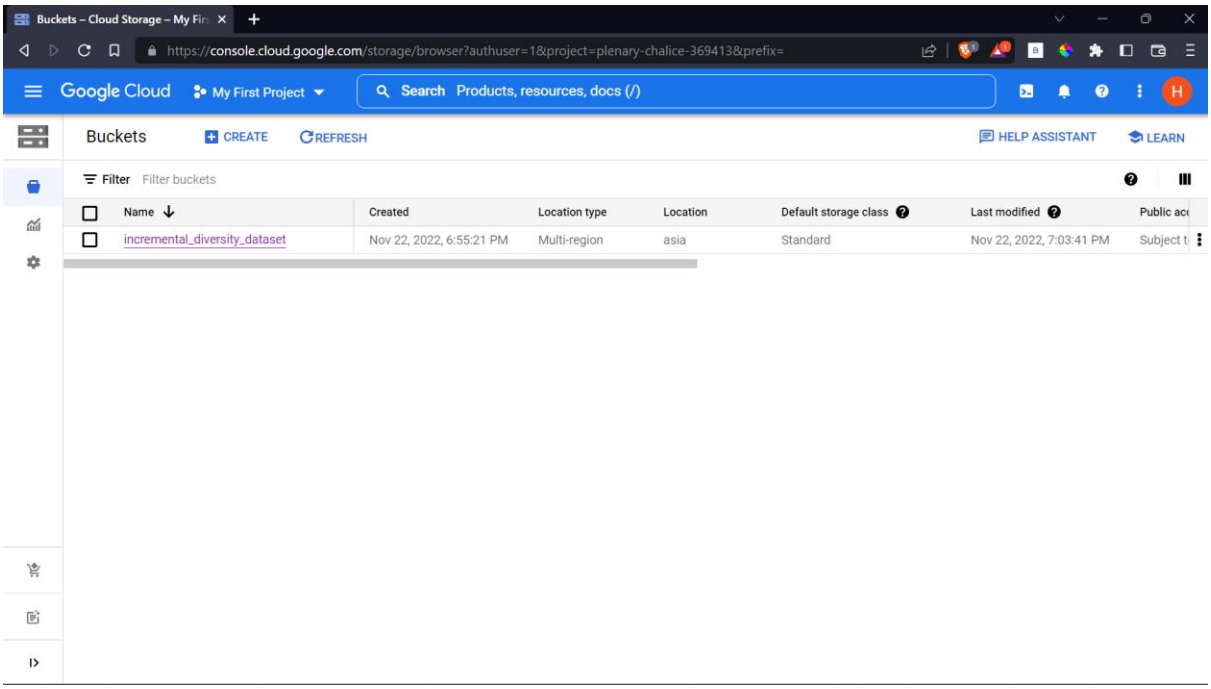

Fig 3.3: Create Bucket Screen

4) Fill the bucket details

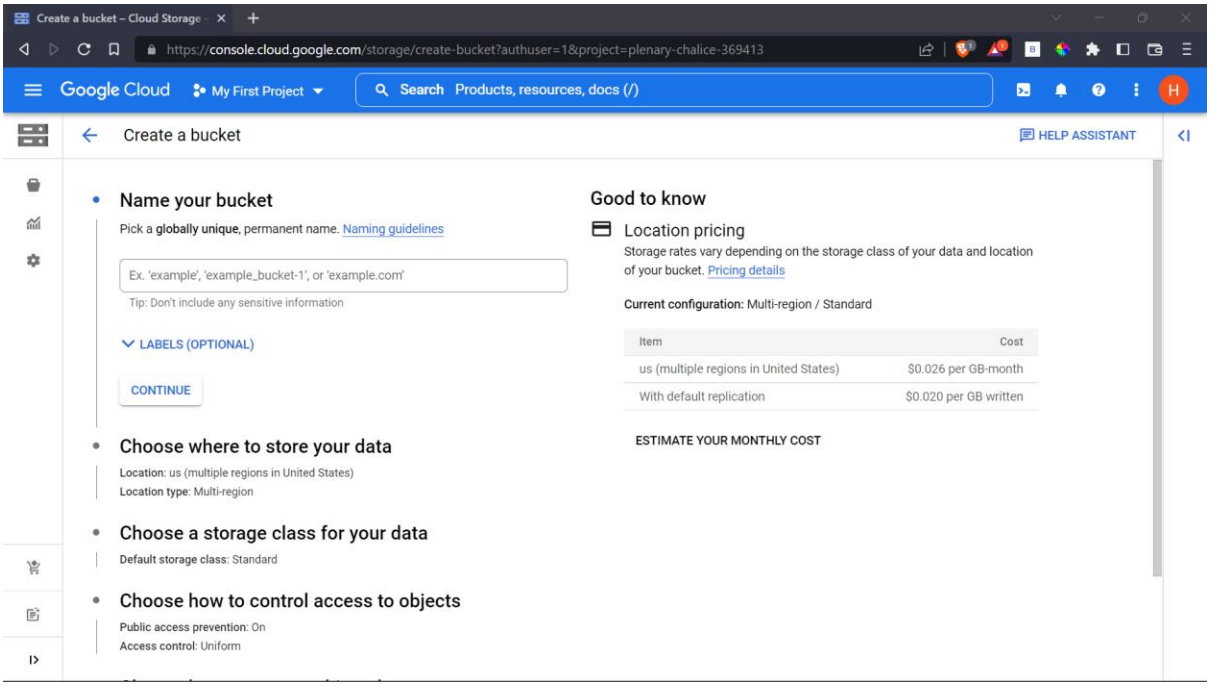

Fig 3.4.1: Bucket Details Screen

Make sure to select **Fine-grained** option under **Access control**

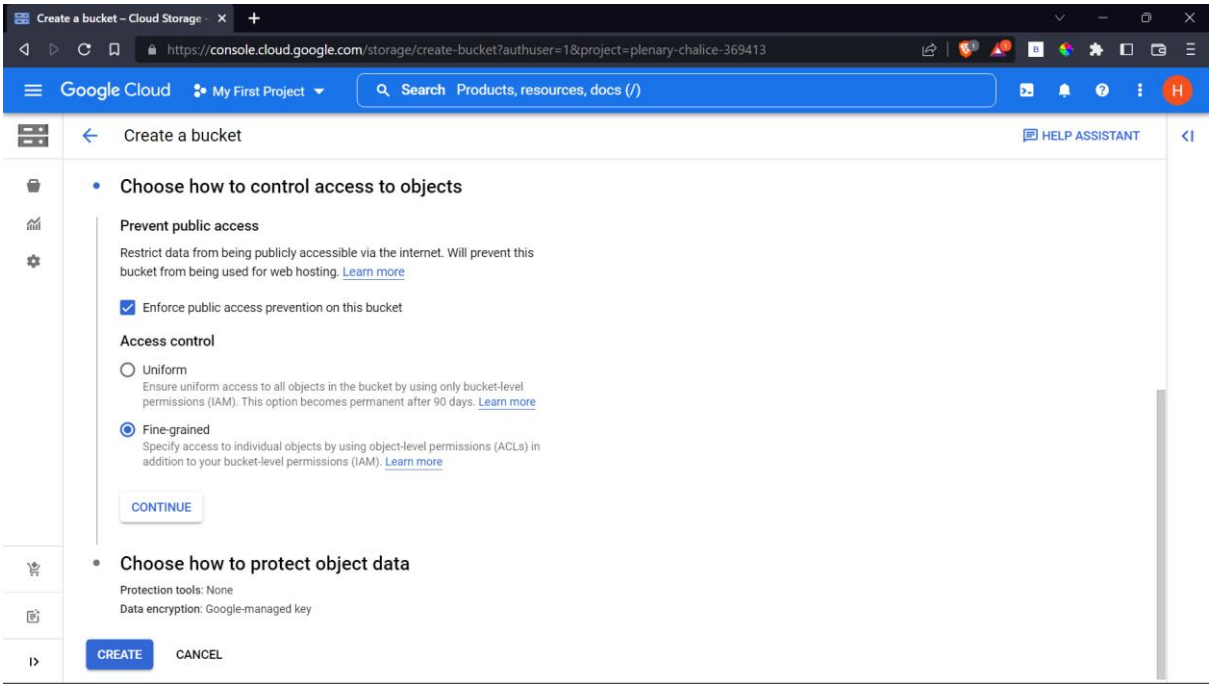

Fig 3.4.2: Control Access to Objects Screen

5) Click on the bucket name (incremental_diversity_dataset in this case).

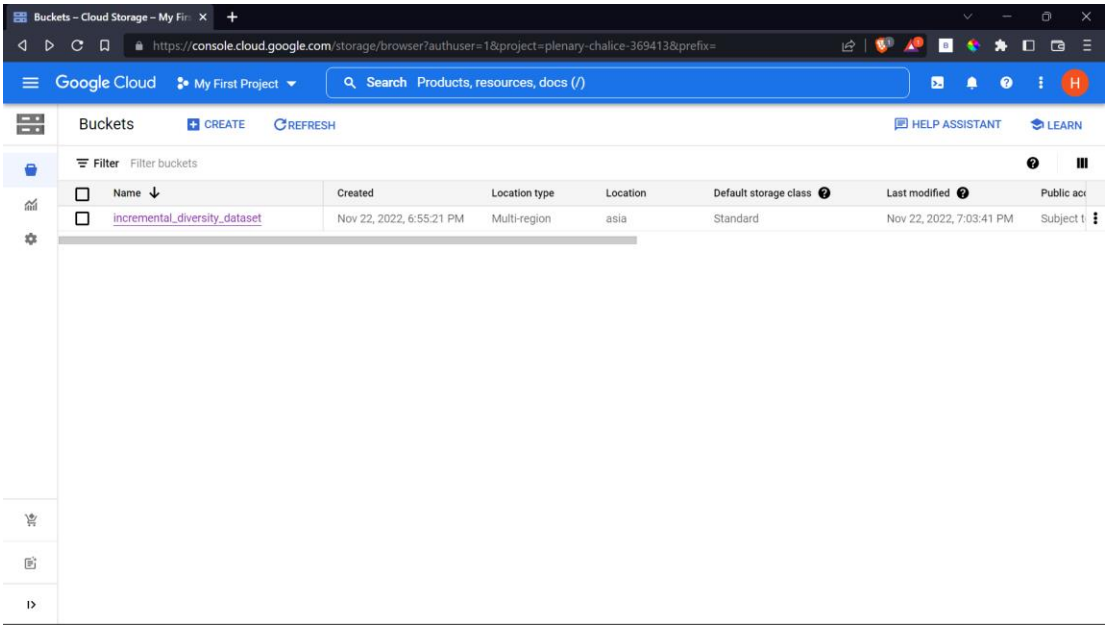

Fig 3.5: Available Buckets Screen

6) Click on **Upload Files**

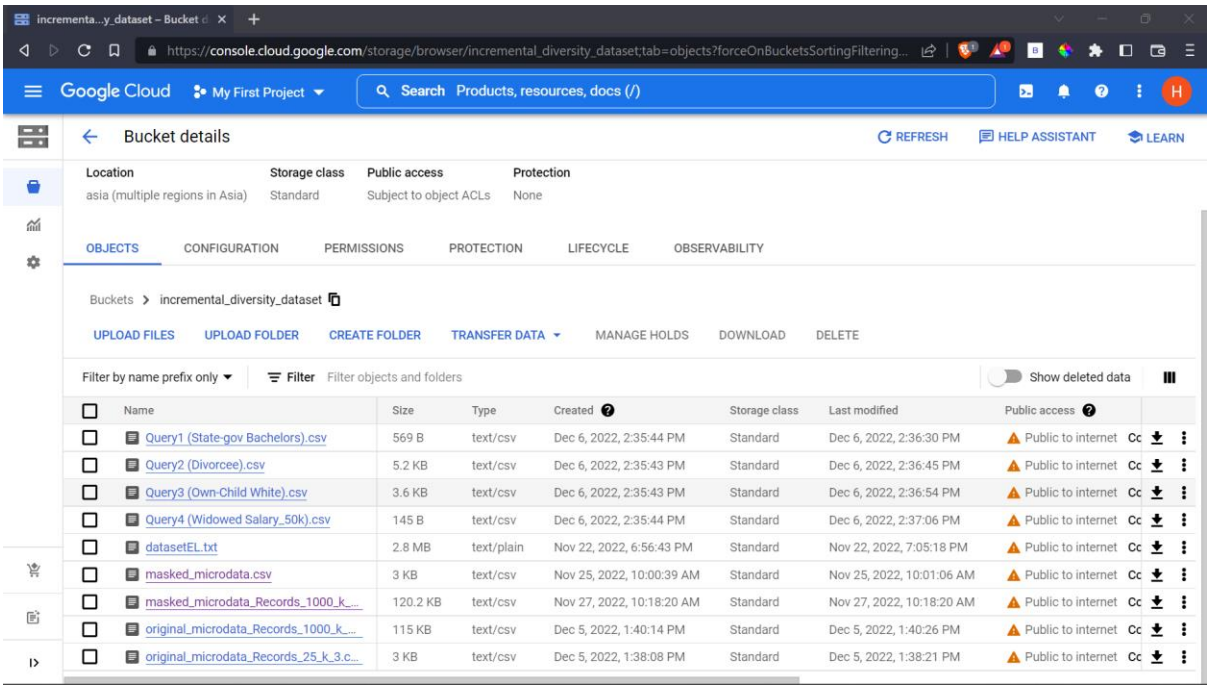

Fig 3.6: Uploading Files Screen

7) Click on the **three dot menu** next to the file.

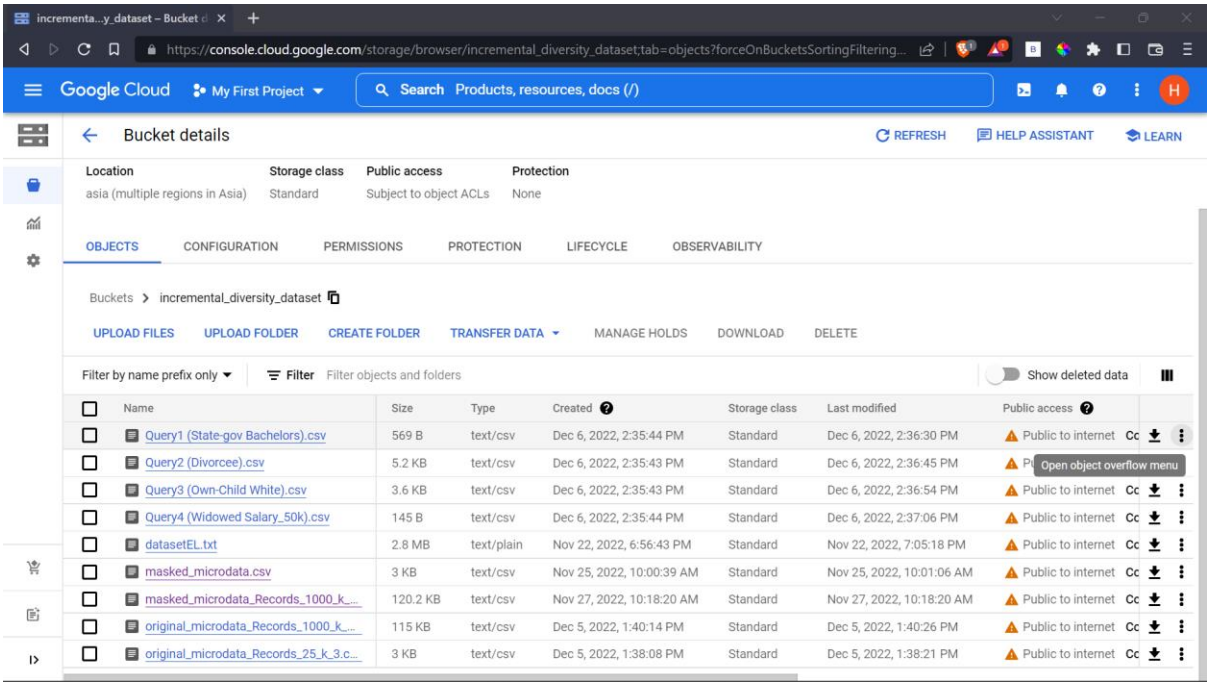

Fig 3.7: Editing Access Screen

Then click on **Edit access**

8) Click on **Add Entry** and fill the following detail:

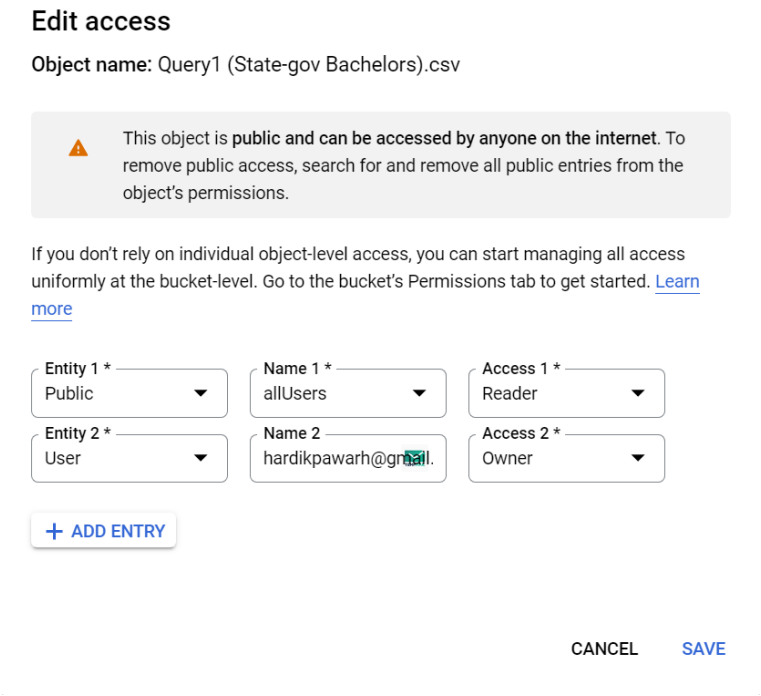

Fig 3.8: Add Entry Screen in Edit Access Window

Ignore the Entity 2. Add Entry of **type Public** and **Name** as **allUsers** and set **access** as **Reader**.

9) Copy the link available for the file under the **Public access** column by clicking on **Copy URL** option.

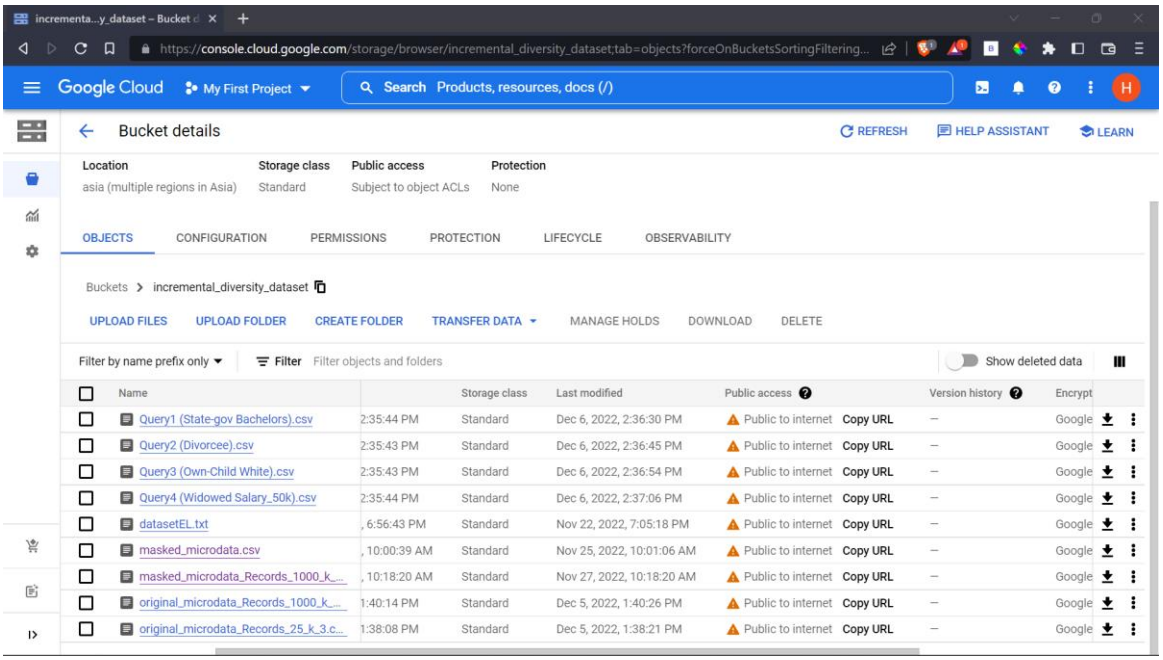

Fig 3.9: Copy Link Screen

10) Done, now you can share the link to the dataset!

4. Result

Various permutations and combinations of queries were run on the masked dataset, which is discussed in the Aggregate Queries paper.

The link to the GitHub repository via which the aforementioned GCP Bucket can be accessed with is:

https://github.com/Hardvan/Website-Incremental-Diversity

To run the website on your local machine:

1) Clone the repository
2) npm install the required modules
3) Change directory to website/my-app
4) Run “npm start” command in the terminal
5) Your website should be up and running!

OR

Directly view the website with the link provided:
https://hardvan.github.io/Website-Incremental-Diversity/

The website has been developed primarily by using the JavaScript library called React, involving the use of React’s Functional and State Components. React is a popular JavaScript library for building user interfaces. React allows developers to create reusable UI components and manage the state of their applications. The user can upload their own custom dataset but it comes with additional guidelines and restrictions that it should contain “quasi-identifier” attributes (which are attributes that can be used to identify individuals in a dataset but are not unique, such as age, gender, and zip code) along with the presence of “Disease” sensitive attribute as the hierarchical tree for the above mentioned sensitive attribute has already been constructed in the code, and dealing with any other primary sensitive attribute requires the construction of a brand new hierarchical tree, which the user can find cumbersome to provide.

After uploading a custom dataset, run the anonymise.py file located in the backend folder to get the masked dataset, which has been anonymized using k & l-e diversity on your very own custom dataset.

5. References:

[1] Lewis Tseng, Haochen Pan and Yingjian Wu. Tutorial: Google Cloud for Beginners: Architecture, Storage, and Computation, 2020.

[2] Ambika Gupta, Pragati Goswami, Nishi Chaudhary and Rashi Bansal. Deploying an Application using Google Cloud Platform, 2020.

[3] Dan Sullivan. Overwiew of Google Cloud Platform, 2019.

[4] Paula Pierleoni, Roberto Concetti, Alberto Belli and Lorenzo Palma. Amazon, Google and Microsoft Solutions for IoT: Architectures and a Performance Comparison, 2019.

[5] Stephanie Challita, Faiez Zalila, Christophe Gourdin and Philippe Merle. A Precise Model for Google Cloud Platform, 2018.

[6] Jiahai Liu, Gao Yang, Haihua Wu and Linwei Zheng. Logistics information management system based on Google cloud computing platform, 2012.

[7] Brandson Posey, Adam Deer, Wyatt Gorman, Vanessa July, Neeraj Kanhere, Dan Speck, Boyd Wilson and Amy Apon, 2019.