

FINAL PROJECT

GOOGLE CLOUD PLATFORM



PROJECT TOPICS:

- Create two buckets.
- Load the data one bucket to another bucket using cloud function.
- Create bigquery.
- Load the data from bucket into the bigquery using dataflow.

A Final Project Report Submitted in the fulfilment of the requirements for Final-Project Evaluation.

Submitted by

Anji L (2320471)

Ayyasamy M (2320474)

Bharath Kumar B (2320821)

Mushi Sai B (2320746)

Madhusri (2320460)

Venkata Swamy P(2320477)

COHORT ID: CSDAIA24GP003



ABSTRACTION:

This project focuses on automating the process of ingesting, processing, and loading data from Google Cloud Storage buckets into BigQuery using Dataflow and Cloud Functions. The workflow involves triggering Cloud Functions in response to new files being uploaded to a bucket, which then initiates a Dataflow job to transform and load the data into BigQuery. The project aims to demonstrate efficient data orchestration techniques in a cloud environment, showcasing the integration of various Google Cloud Platform services.

Key Components:

1. **Google Cloud Storage Buckets:** Used for storing raw data files.
2. **Cloud Functions:** Triggered by new file uploads, initiating the data processing pipeline.
3. **Dataflow:** Processes and transforms the data from Cloud Storage and loads it into BigQuery.
4. **BigQuery:** Serves as the data warehouse for the processed data.

Objectives:

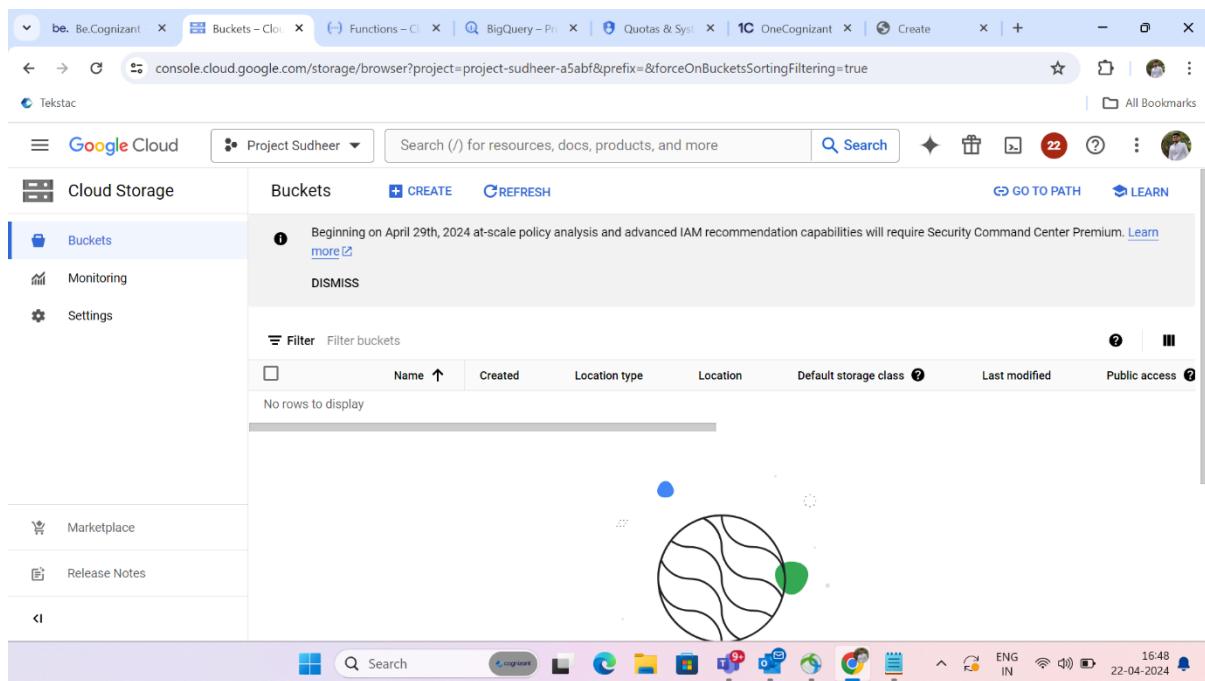
- Automatically trigger data processing tasks in response to new data uploads.
- Perform ETL (Extract, Transform, Load) operations on the data using Dataflow.
- Load transformed data into BigQuery for analysis and reporting.
- Demonstrate the scalability and efficiency of cloud-based data orchestration using Google Cloud Platform services.

GOOGLE CLOUD PLATFORM

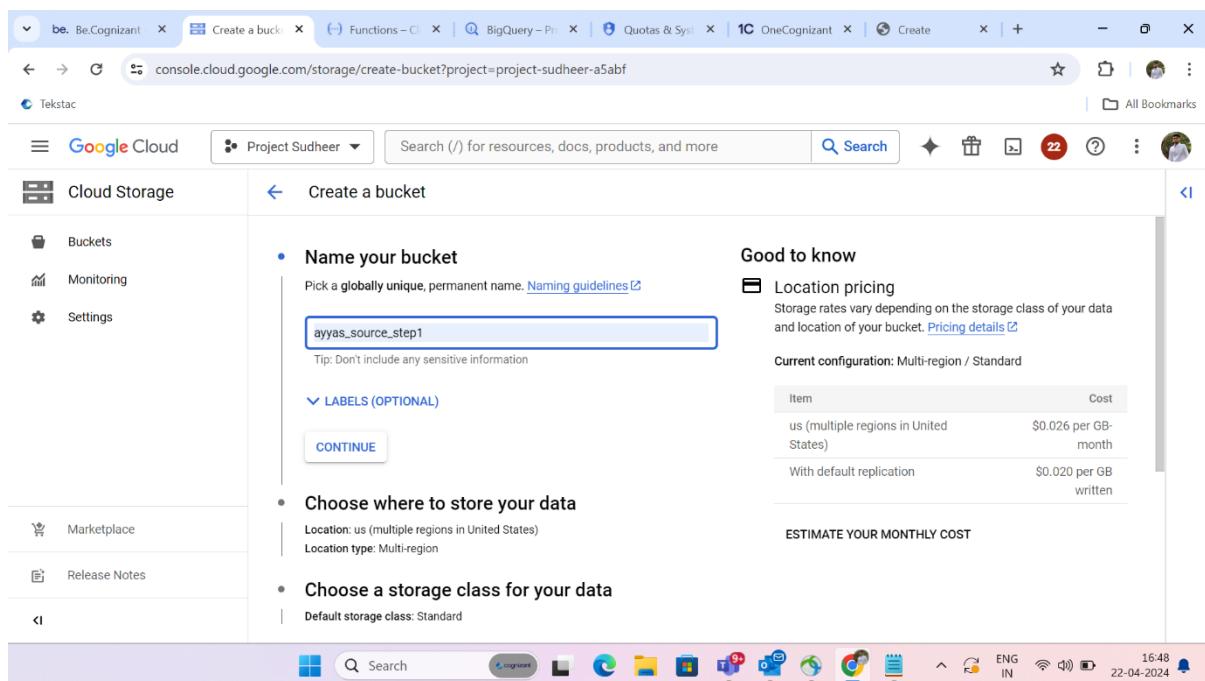
The screenshot shows the Google Cloud Platform (GCP) Welcome page. At the top, there is a navigation bar with several tabs: 'Be.Cognizant' (active), 'Welcome - Pr...', 'Functions - Cl...', 'BigQuery - Pr...', 'Quotas & Sys...', '1C OneCognizant', and 'Create'. Below the navigation bar is a search bar with the placeholder 'Search (/) for resources, docs, products, and more'. To the right of the search bar is a user profile section for 'Ayyas' (ayyasamy2002mv@gmail.com). The main content area features a 'Welcome' heading with the text 'You're working in Project Sudheer'. It displays the project number (901519289329) and ID (project-sudheer-a5abf). Below this are buttons for 'Create a VM', 'Run a query in BigQuery', 'Create a GKE cluster', and 'Create a storage bucket'. A sidebar on the left lists 'Quick access' items like 'Cloud overview' and 'Products & solutions'. A pinned product sidebar on the left lists 'APIs & Services', 'Billing', 'IAM & Admin', 'Marketplace', 'Compute Engine', 'Kubernetes Engine', and 'Cloud Storage'. A bottom navigation bar shows the URL 'https://console.cloud.google.com/storage?project=project-sudheer-a5abf'.

This screenshot is similar to the one above, showing the GCP Welcome page. However, the 'Cloud Storage' section of the pinned product sidebar is now expanded, revealing sub-options: 'Buckets', 'BigQuery', 'Create a GKE cluster', and 'Create a storage bucket'. The rest of the interface remains the same, including the navigation bar, search bar, user profile, and bottom navigation bar.

BUCKET CREATION :



SOURCE BUCKET:



Screenshot of the Google Cloud Storage 'Create a bucket' wizard.

The left sidebar shows 'Cloud Storage' with 'Buckets' selected. The main panel is titled 'Create a bucket'.

Location type:

- Multi-region (Highest availability across largest area)
- Dual-region (High availability and low latency across 2 regions)
- Region (Lowest latency within a single region)

A dropdown menu shows 'us-east1 (South Carolina)' selected.

ESTIMATE YOUR MONTHLY COST

Item	Cost
us-east1 (South Carolina)	\$0.020 per GB-month

CONTINUE

Choose a storage class for your data

Default storage class: Standard

Choose how to control access to objects

Public access prevention: On

System tray icons: Search, cog, file, browser, Microsoft Office, calendar, etc. Language: ENG IN. Date: 22-04-2024. Time: 16:48.

Screenshot of the Google Cloud Storage 'Create a bucket' wizard.

The left sidebar shows 'Cloud Storage' with 'Buckets' selected. The main panel is titled 'Create a bucket'.

Storage class:

Automatically or specify a default storage class based on how long you plan to store your data and your workload or use case. [Learn more](#)

Autoclass ?
Automatically transitions each object to Standard or Nearline class based on object-level activity, to optimize for cost and latency. Recommended if usage frequency may be unpredictable. Can be changed to a default class at any time. [Pricing details](#)

Set a default class
Applies to all objects in your bucket unless you manually modify the class per object or set object lifecycle rules. Best when your usage is highly predictable.

- Standard ?
Best for short-term storage and frequently accessed data
- Nearline
Best for backups and data accessed less than once a month
- Coldline
Best for disaster recovery and data accessed less than once a quarter
- Archive
Best for long-term digital preservation of data accessed less than once a year

CONTINUE

System tray icons: Search, cog, file, browser, Microsoft Office, calendar, etc. Language: ENG IN. Date: 22-04-2024. Time: 16:48.

Screenshot of a web browser showing the Google Cloud Platform (GCP) interface for creating a new bucket.

The URL in the address bar is `console.cloud.google.com/storage/create-bucket?project=project-sudheer-a5abf`.

The page title is "Create a bucket".

The left sidebar shows navigation links: "Cloud Storage" (selected), "Buckets", "Monitoring", and "Settings". Other links include "Marketplace" and "Release Notes".

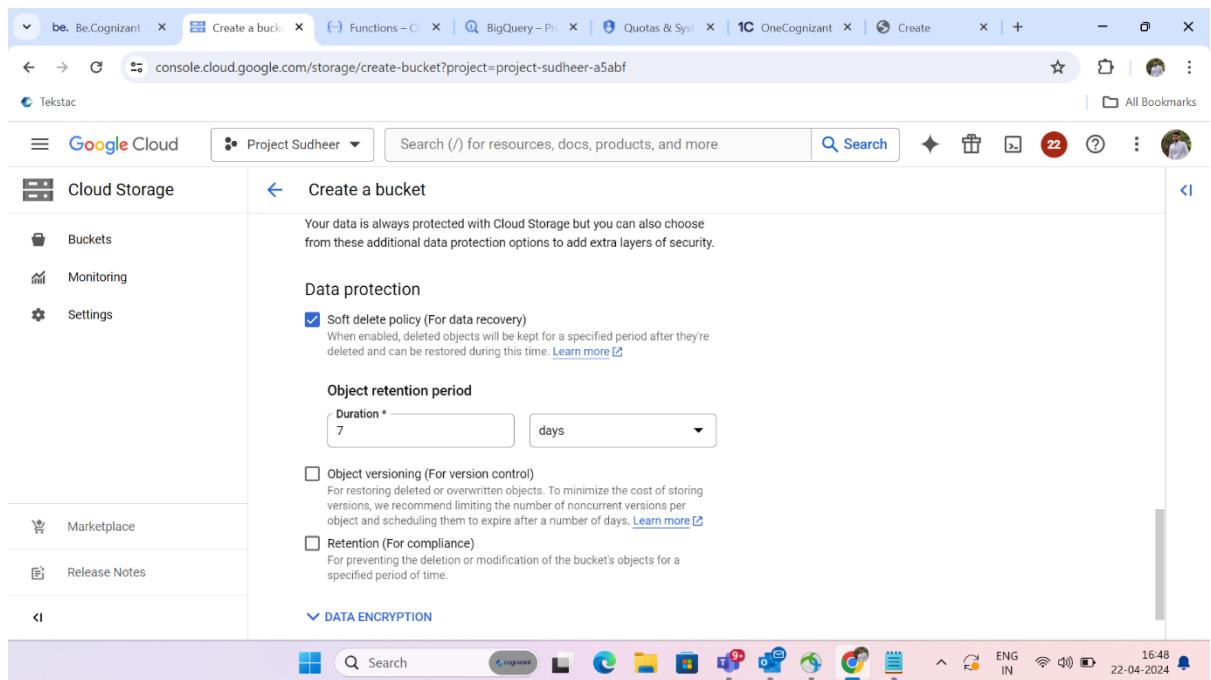
The main content area starts with a section titled "Choose how to control access to objects".

Prevent public access: A note states: "Restrict data from being publicly accessible via the internet. Will prevent this bucket from being used for web hosting." There is a checked checkbox labeled "Enforce public access prevention on this bucket".

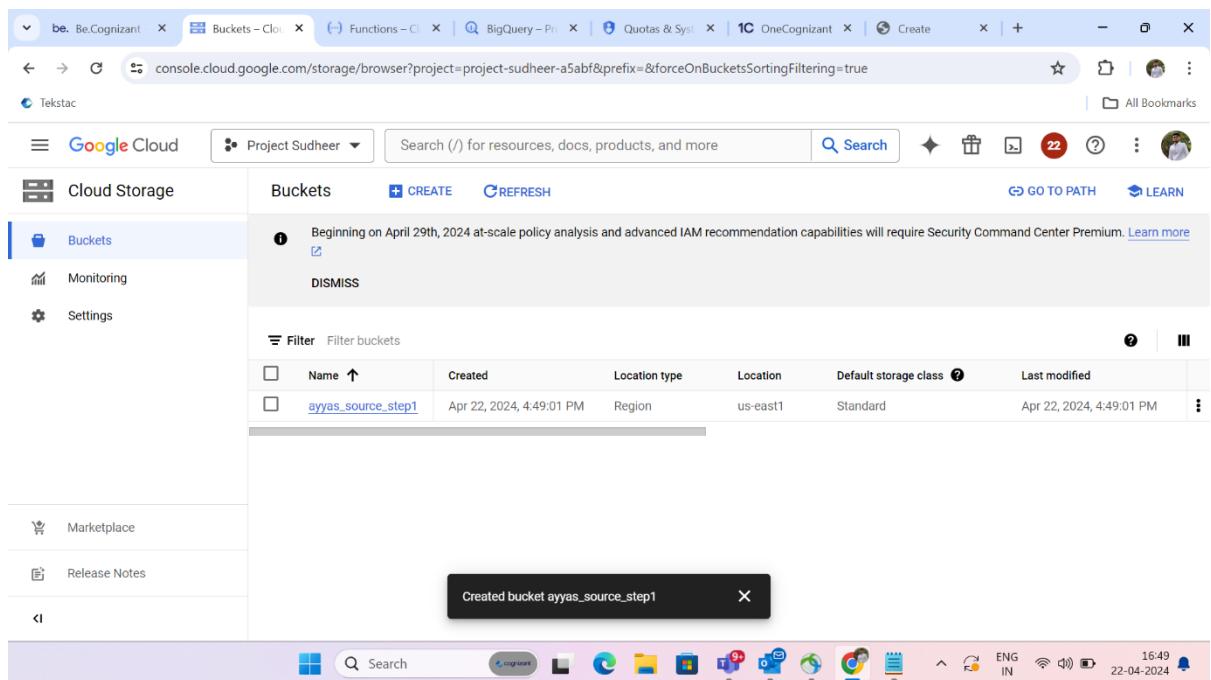
Access control: A radio button is selected for "Uniform", with a note: "Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days." There is also an unselected radio button for "Fine-grained".

A "CONTINUE" button is at the bottom of the form.

The system tray at the bottom right shows the date as 22-04-2024, time as 16:48, and language as ENG IN.



The screenshot shows the 'Create a bucket' page in the Google Cloud Storage interface. The left sidebar is titled 'Cloud Storage' and includes 'Buckets', 'Monitoring', and 'Settings'. The main content area is titled 'Create a bucket' and contains sections for 'Data protection' (with 'Soft delete policy (For data recovery)' checked) and 'Object retention period' (set to 7 days). Below these are options for 'Object versioning (For version control)' and 'Retention (For compliance)'. A 'DATA ENCRYPTION' section is also present. At the bottom, there are 'CREATE' and 'REFRESH' buttons.



The screenshot shows the 'Buckets' page in the Google Cloud Storage interface. The left sidebar is identical to the previous screen. The main content area lists a single bucket named 'ayyas_source_step1' with details: Created on Apr 22, 2024, 4:49:01 PM, Location type: Region, Location: us-east1, Default storage class: Standard. A success message 'Created bucket ayyas_source_step1' is displayed at the bottom. The status bar at the bottom right shows the date as 22-04-2024 and the time as 16:49.

DESTINATION BUCKET:

be. Be.Cognizant × Create a buck. × Functions – Cl. × BigQuery – Pro. × Quotas & Sys. × 1C OneCognizant × Create × - ×

console.cloud.google.com/storage/create-bucket?project=project-sudheer-a5abf

Tekstac

All Bookmarks

Google Cloud Project Sudheer Search (/) for resources, docs, products, and more Search

Cloud Storage Create a bucket

Buckets Monitoring Settings Marketplace Release Notes

Name your bucket Pick a globally unique, permanent name. [Naming guidelines](#)

ayyas_destination_step1

Tip: Don't include any sensitive information

LABELS (OPTIONAL)

CONTINUE

Good to know

Location pricing Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Multi-region / Standard

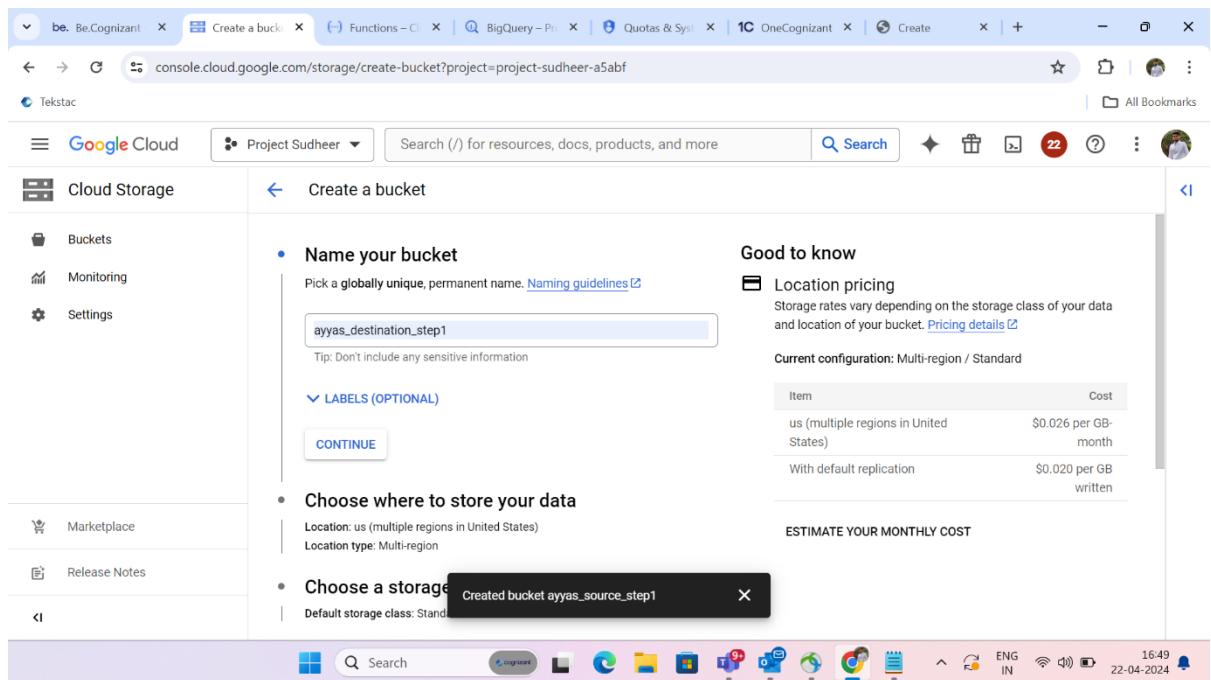
Item	Cost
us (multiple regions in United States)	\$0.026 per GB-month
With default replication	\$0.020 per GB written

ESTIMATE YOUR MONTHLY COST

Created bucket ayyas_source_step1

Default storage class: Standard

Search ENG IN 16:49 22-04-2024



Screenshot of the Google Cloud Storage 'Create a bucket' wizard.

Location type:

- Multi-region
- Dual-region
- Region

Current configuration: Region / Standard

Item	Cost
us-east1 (South Carolina)	\$0.020 per GB-month

ESTIMATE YOUR MONTHLY COST

Choose a storage class for your data

Default storage class: Standard

Created bucket ayyas_source_step1

Choose how to control access

Public access prevention: On

Windows taskbar at the bottom showing various icons and system status.

Screenshot of the Google Cloud Storage 'Create a bucket' wizard.

Name your bucket: ayyas_destination_step1

Choose where to store your data:

Location: us-east1 (South Carolina)
Location type: Region

Choose a storage class for your data:

A storage class sets costs for storage, retrieval, and operations, with minimal differences in uptime. Choose if you want objects to be managed automatically or specify a default storage class based on how long you plan to store your data and your workload or use case. [Learn more](#)

- Autoclass Automatically transitions each object to Standard or Nearline class based on object-level activity, to optimize for cost and latency. Recommended if usage frequency may be unpredictable. Can be changed to a default class at any time. [Pricing details](#)
- Set a default class Applies to all objects in your bucket unless you manually modify the class for individual objects.

Good to know:

Location pricing: Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Region / Standard

Item	Cost
us-east1 (South Carolina)	\$0.020 per GB-month

ESTIMATE YOUR MONTHLY COST

Windows taskbar at the bottom showing various icons and system status.

Screenshot of the Google Cloud Platform (GCP) console showing the 'Create a bucket' wizard.

The left sidebar shows the 'Cloud Storage' section with 'Buckets' selected. The main panel is titled 'Create a bucket'.

Set a default class:

- Standard: Best for short-term storage and frequently accessed data
- Nearline: Best for backups and data accessed less than once a month
- Coldline: Best for disaster recovery and data accessed less than once a quarter
- Archive: Best for long-term digital preservation of data accessed less than once a year

CONTINUE

Choose how to control access to objects:

Public access prevention: On
Access control: Uniform

At the bottom right of the window, there is a status bar showing: ENG IN 16:49 22-04-2024

Screenshot of the Google Cloud Platform (GCP) console showing the 'Create a bucket' wizard.

The left sidebar shows the 'Cloud Storage' section with 'Buckets' selected. The main panel is titled 'Create a bucket'.

Choose how to control access to objects:

Prevent public access:

Restrict data from being publicly accessible via the internet. Will prevent this bucket from being used for web hosting. [Learn more](#)

Enforce public access prevention on this bucket

Access control:

Uniform: Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

Fine-grained: Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)

CONTINUE

At the bottom right of the window, there is a status bar showing: ENG IN 16:49 22-04-2024

The screenshot shows the 'Create a bucket' page in the Google Cloud Storage interface. The left sidebar shows 'Cloud Storage' selected. The main area has a title 'Create a bucket' with a back arrow. Under 'Data protection', 'Soft delete policy (For data recovery)' is checked, with a note: 'When enabled, deleted objects will be kept for a specified period after they're deleted and can be restored during this time.' Below it is an 'Object retention period' section with a duration of 7 days. There are also unchecked options for 'Object versioning (For version control)' and 'Retention (For compliance)'. A 'DATA ENCRYPTION' section is collapsed. At the bottom are 'CREATE' and 'CANCEL' buttons.

CHECK BUCKET:

The screenshot shows the 'Buckets' page in the Google Cloud Storage interface. The left sidebar shows 'Buckets' selected. The main area has a title 'Buckets' with '+ CREATE' and 'REFRESH' buttons. A message at the top says: 'Beginning on April 29th, 2024 at-scale policy analysis and advanced IAM recommendation capabilities will require Security Command Center Premium.' with a 'Learn more' link. Below is a 'DISMISS' button. A 'Filter buckets' section allows filtering by Name, Created, Location type, Location, Default storage class, and Last modified. Two buckets are listed: 'ayyas_destination_step1' and 'ayyas_source_step1', both created on April 22, 2024. The bottom of the screen shows a taskbar with various icons and system status.

CLOUD FUNCTION:

The screenshot shows the Google Cloud search interface. The search bar at the top contains the query "cloud functions". Below the search bar, there are four tabs: ALL, DOCUMENTATION & TUTORIALS, RESOURCES, and MARKETPLACE & APIs. The ALL tab is selected. On the left, there is a sidebar with "Filter by" options like Product or Page, Documentation or tutorial, Marketplace and APIs, Organization, Folder, Project, and Resources. Under "Resource filters", it shows "Project, folder, or org: Project Sudheer" and "Resource type: Any". The main search results area displays several cards. The first card is titled "Cloud Functions" and describes it as a service for running code with zero server management. The second card is titled "AI/ML Image Processing on Cloud Functions" and describes it as a products & solutions. The third card is titled "Cloud Run" and describes it as a managed serverless platform. A tooltip for Cloud Run says "Deleting function successfully started". On the right side of the search results, there are tips for speeding up searches and instructions for finding resource types and API keys.

The screenshot shows the Google Cloud Functions list page. The top navigation bar includes the project name "Project Sudheer" and a search bar. Below the navigation, there are tabs for "Cloud Functions" (selected), "Functions", "+ CREATE FUNCTION", and "REFRESH". There is also a "RELEASE NOTES" and "LEARN" link. A "Filter" button is present. The main area shows a table with columns: Environment, Name, Last deployed, Region, Recommendation, Trigger, Runtime, Memory allocated, Executed function, and Actions. A message "No rows to display" is shown. In the center of the page is a decorative graphic of a globe with colored dots (red, blue, yellow) and small circular icons. At the bottom, a welcome message "Welcome to Cloud Functions!" is displayed.

Screenshot of the Google Cloud Functions "Create function" wizard, Step 1: Configuration.

Basics

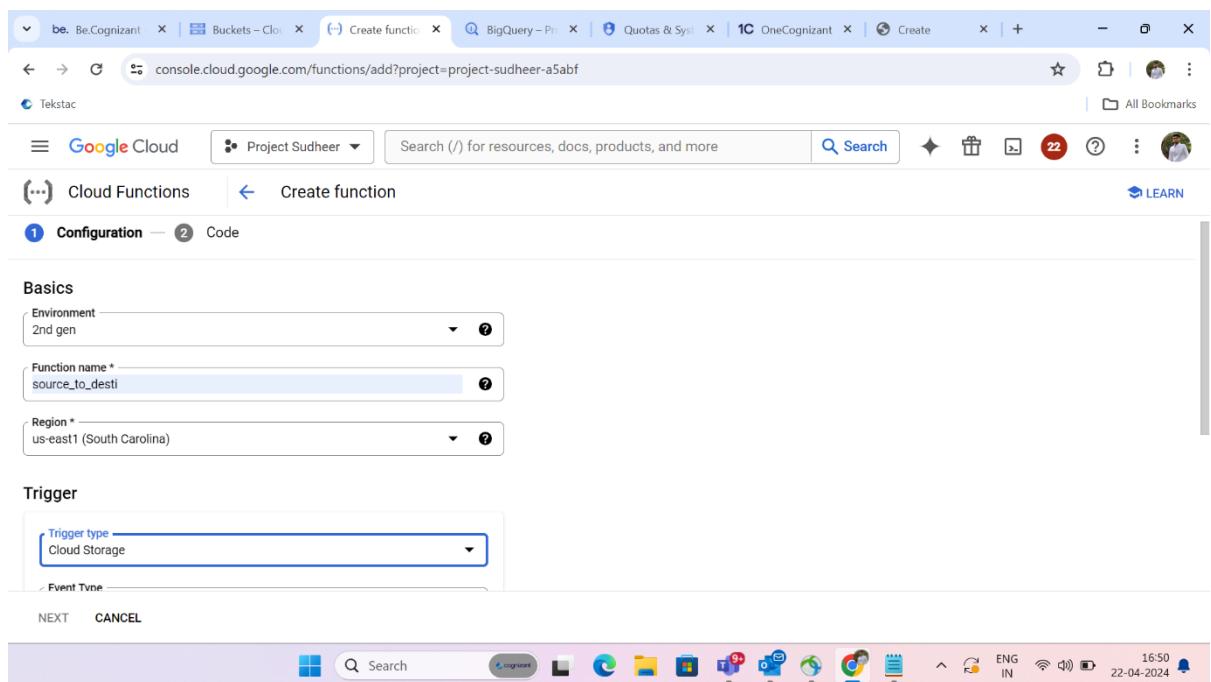
- Environment: 2nd gen
- Function name*: source_to_desti
- Region*: us-east1 (South Carolina)

Trigger

Trigger type: Cloud Storage

Event Type:

NEXT CANCEL



Screenshot of the Google Cloud Functions "Create function" wizard, Step 1: Configuration.

Trigger

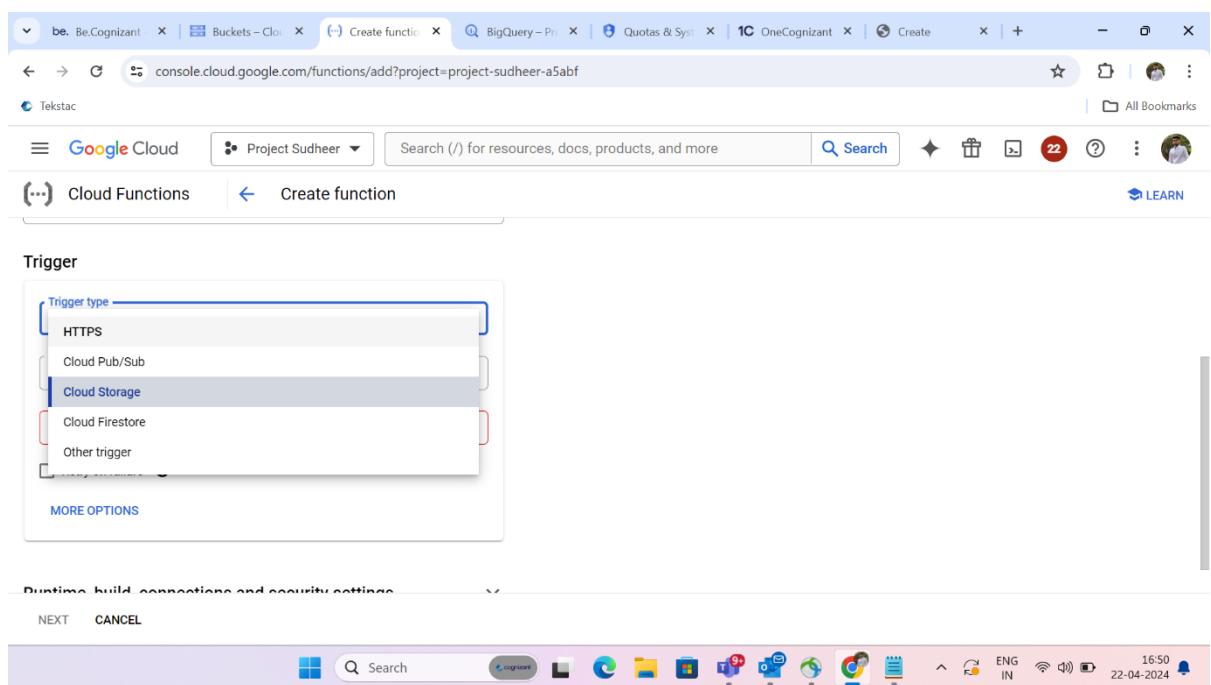
Trigger type: Cloud Storage

Event Type:

MORE OPTIONS

Runtimes, build, connections and security settings

NEXT CANCEL



Screenshot of the Google Cloud Functions "Create function" wizard, step 1: Trigger.

The "Trigger type" dropdown is set to "Cloud Storage". The "Event Type" dropdown shows a list of events, with "google.cloud.storage.object.v1.finalized" selected.

Below the trigger configuration, there is a "Runtime, build, connections and security settings" section which is currently collapsed.

At the bottom of the main window, there are "NEXT" and "CANCEL" buttons.

The taskbar at the bottom of the screen shows various application icons and the date/time: 22-04-2024 16:50.

Screenshot of the Google Cloud Functions "Create function" wizard, step 2: Select bucket.

The "Region" dropdown is set to "us-east1 (South Carolina)".

The "Trigger" section shows the same configuration as the previous screenshot: "Trigger type" is "Cloud Storage" and "Event Type" is "google.cloud.storage.object.v1.finalized".

The "Bucket" input field is highlighted in red, indicating it is required. To its right is a "BROWSE" button.

A modal window titled "Select bucket" lists two buckets: "ayyas_destination_step1" and "ayyas_source_step1".

At the bottom of the modal are "SELECT" and "CANCEL" buttons.

The taskbar at the bottom of the screen shows various application icons and the date/time: 22-04-2024 16:51.

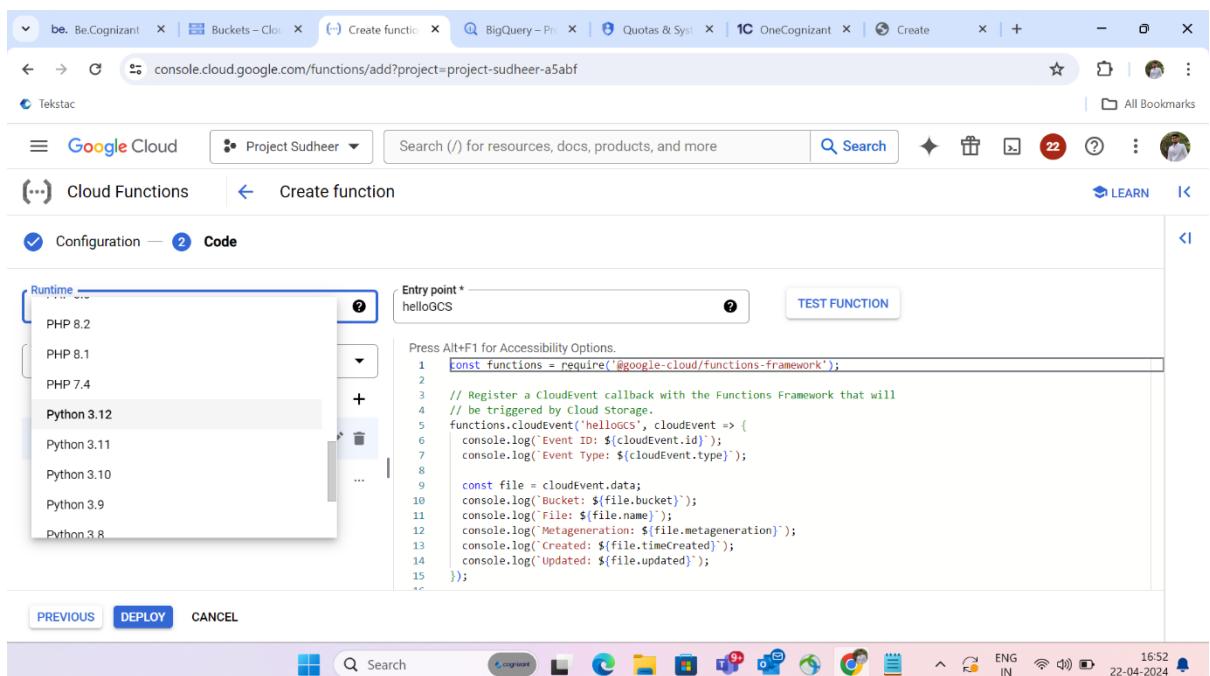
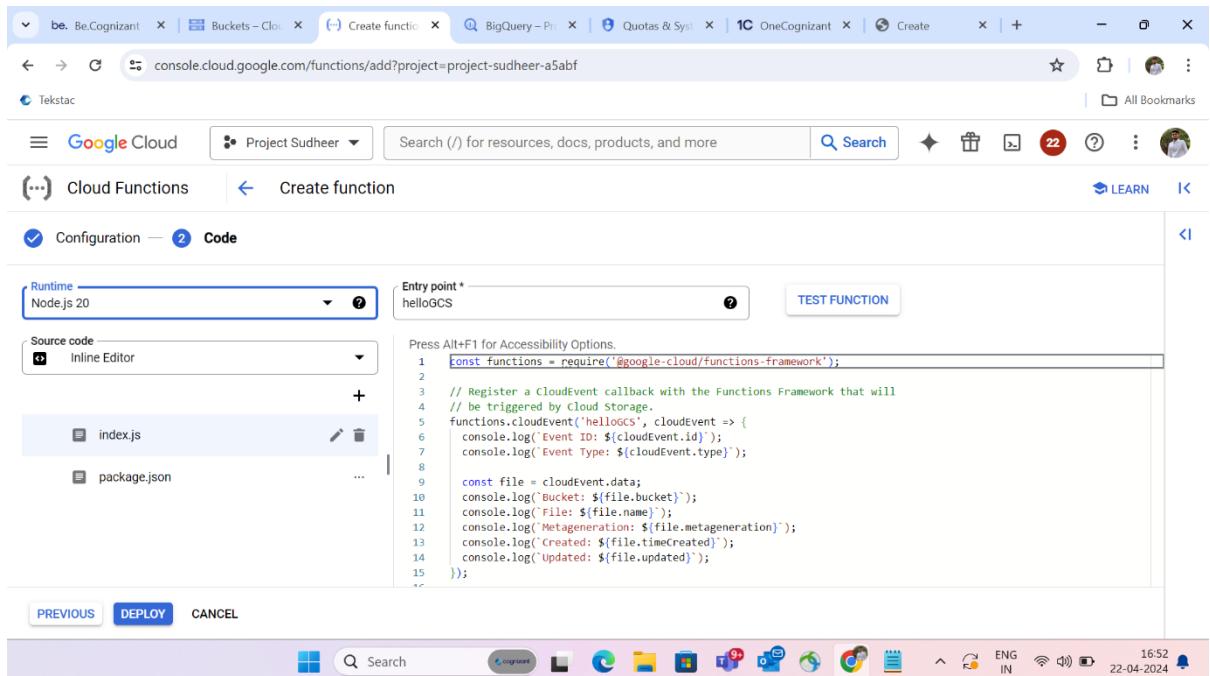
The screenshot shows the 'Create function' wizard in the Google Cloud Platform console. The current step is 'Runtime, build, connections and security settings'. The 'RUNTIME' tab is selected. The configuration includes:

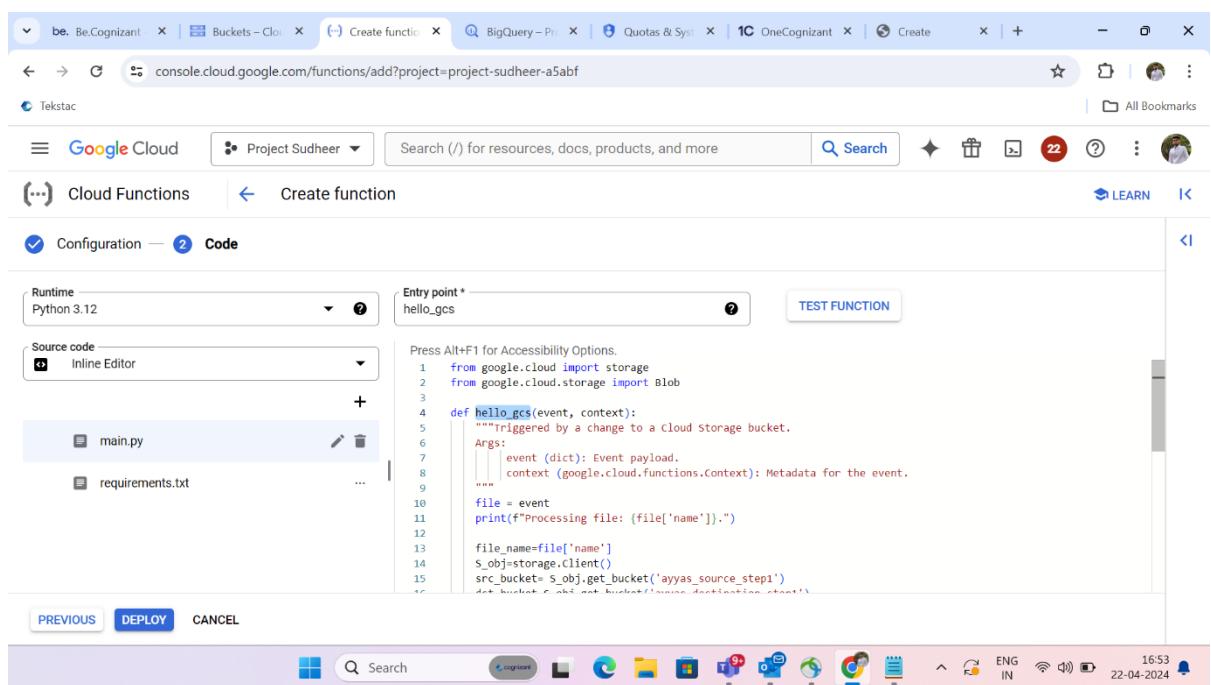
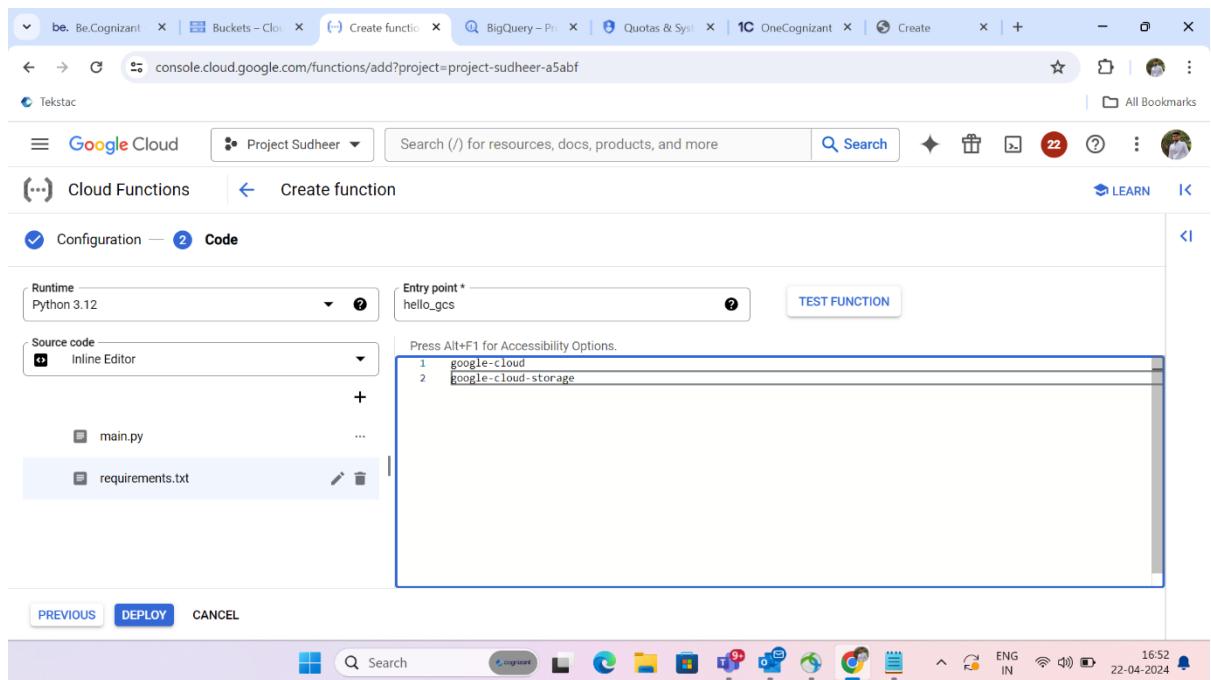
- Memory allocated: 256 MiB
- CPU: 0.167
- Timeout: 60 seconds
- Concurrency: Maximum concurrent requests per instance: 1

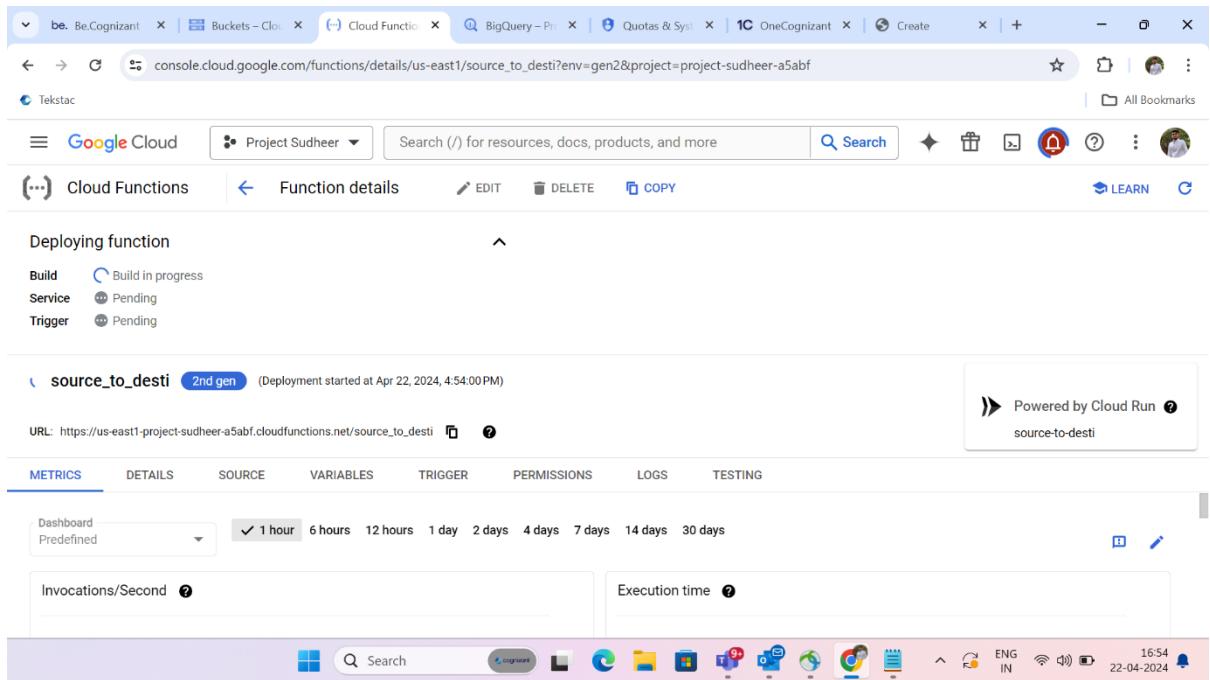
Below these settings is the 'Autoscaling' section, which is currently empty. At the bottom of the page are 'NEXT' and 'CANCEL' buttons.

This screenshot is identical to the one above, but it includes a tooltip for the 'Timeout' field. The tooltip provides information about the default timeout duration and the maximum allowed value for different trigger types.

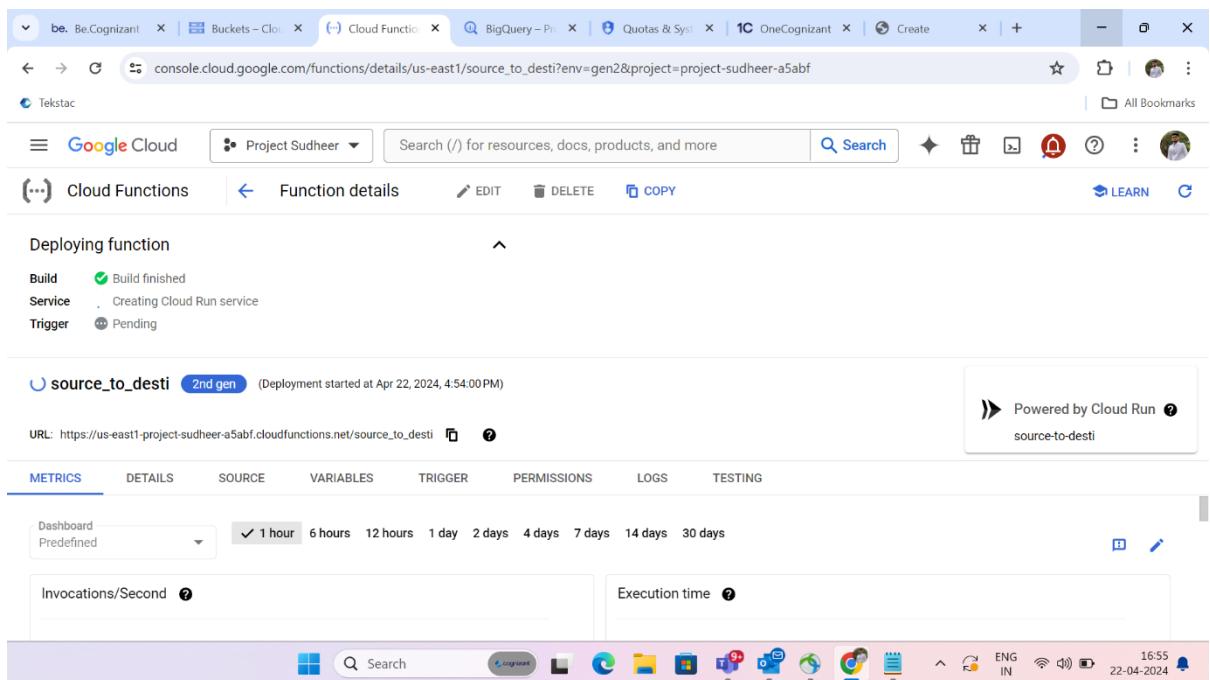
Timeout
If the Cloud Function has not completed by the timeout duration, then the Function will be terminated. The default timeout duration is 60 seconds. For Eventarc triggered functions, the maximum timeout that can be specified is 9 mins (540 seconds). For HTTPS triggered functions, the maximum timeout that can be specified is 60 mins (3600 seconds).







The screenshot shows the Google Cloud Functions "Function details" page for a function named "source_to_desti". The status bar indicates "Build in progress". The "Metrics" tab is selected, showing a chart for "Invocations/Second" and "Execution time". The URL for the function is https://us-east1-project-sudheer-a5abf.cloudfunctions.net/source_to_desti. A sidebar on the right says "Powered by Cloud Run source-to-desti". The taskbar at the bottom shows various application icons.



The screenshot shows the Google Cloud Functions "Function details" page for the same function "source_to_desti". The status bar now says "Build finished". The "Metrics" tab is selected, showing the same chart data as the previous screenshot. The URL remains the same. The sidebar still says "Powered by Cloud Run source-to-desti". The taskbar at the bottom shows various application icons.

The screenshot shows the Google Cloud Functions console for the 'source_to_desti' function in the 'Project Sudheer' project. The function is currently in the 'Creating Cloud Run service' stage. The deployment started at 4:54:00 PM on April 22, 2024. The URL for the function is https://us-east1-project-sudheer-a5abf.cloudfunctions.net/source_to_desti. The Metrics tab is selected, showing no data available for the selected time frame.

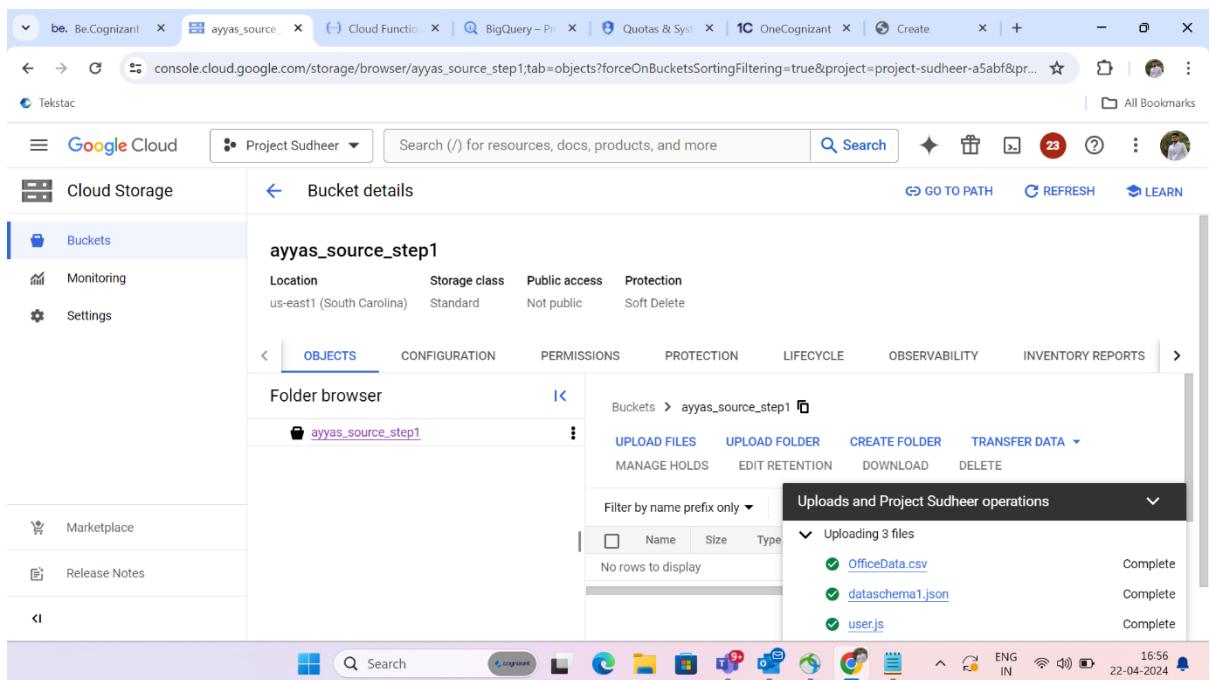
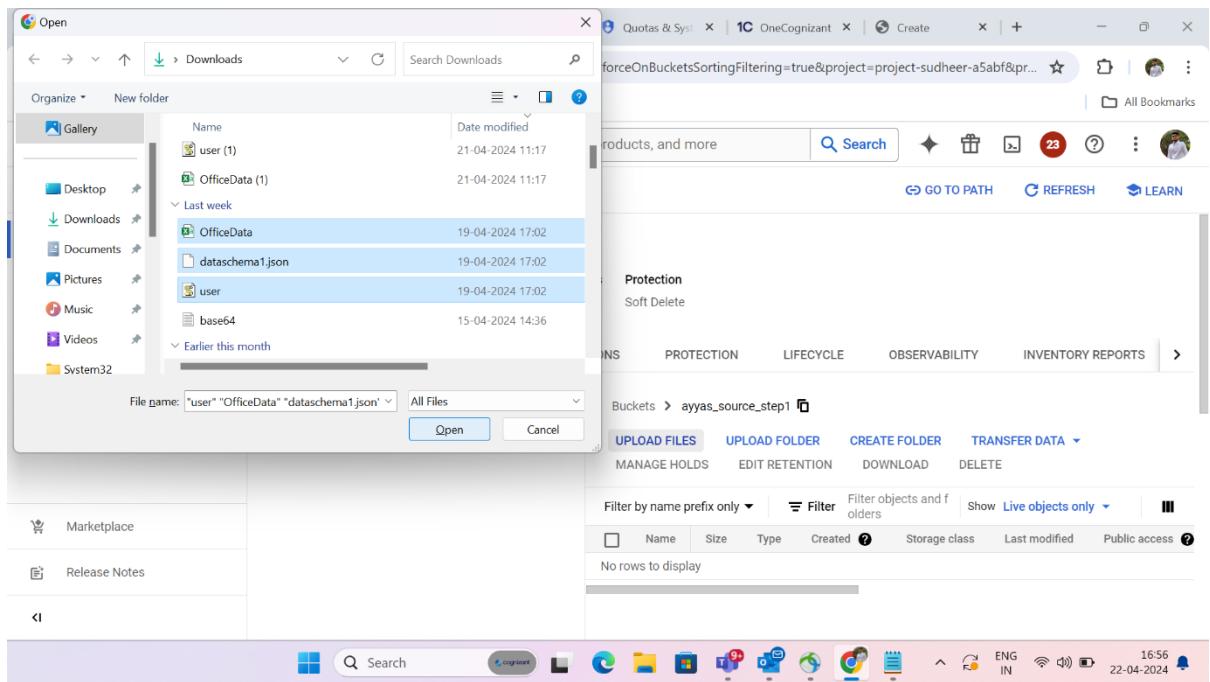
The screenshot shows the Google Cloud Functions console for the 'source_to_desti' function in the 'Project Sudheer' project. The function has been deployed successfully at 4:55:36 PM on April 22, 2024. The URL is now active. The Metrics tab shows no data available for the selected time frame.

Upload files to Buckets:

The screenshot shows the Google Cloud Storage Buckets page. On the left sidebar, under 'Cloud Storage', the 'Buckets' option is selected. The main area displays a table of buckets with the following columns: Name, Created, Location type, Location, Default storage class, and Last modified. Two buckets are listed:

Name	Created	Location type	Location	Default storage class	Last modified
ayyas_destination_step1	Apr 22, 2024, 4:49:51 PM	Region	us-east1	Standard	Apr 22, 2024, 4:49:51 PM
ayyas_source_step1	Apr 22, 2024, 4:49:01 PM	Region	us-east1	Standard	Apr 22, 2024, 4:49:01 PM

The screenshot shows the 'Bucket details' page for the 'ayyas_source_step1' bucket. The left sidebar shows the 'Buckets' option is selected. The main area displays the bucket's configuration, including its location (us-east1), storage class (Standard), public access (Not public), and protection (Soft Delete). Below this, there are tabs for 'OBJECTS', 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', 'OBSERVABILITY', and 'INVENTORY REPORTS'. The 'OBJECTS' tab is active, showing a 'Folder browser' with a single folder named 'ayyas_source_step1'. Below the browser are buttons for 'UPLOAD FILES', 'UPLOAD FOLDER', 'CREATE FOLDER', and 'TRANSFER DATA'. There is also a 'MANAGE HOLDS' button. At the bottom, there is a search bar and filter options.



The screenshot shows the Google Cloud Storage interface. On the left, a sidebar lists 'Cloud Storage', 'Buckets' (selected), 'Monitoring', and 'Settings'. The main area displays 'Bucket details' for 'ayyas_destination_step1'. The bucket's location is 'us-east1 (South Carolina)', storage class is 'Standard', public access is 'Not public', and protection is 'Soft Delete'. Below this, a 'Folder browser' section shows three objects: 'OfficeData.csv', 'dataschema1.json', and 'user.js'. The 'user.js' file is highlighted. A toolbar at the top right includes 'GO TO PATH', 'REFRESH', and 'LEARN' buttons.

Create Dataset in Big query:

The screenshot shows the Google BigQuery interface. On the left, a sidebar lists 'BigQuery Studio' (selected), 'Data transfers', 'Scheduled queries', 'Analytics Hub', 'Dataform', and 'Partner Center'. The main area shows an 'Explorer' pane with a search bar and a list of resources under 'project-sudheer-a5abf'. A message at the bottom says "'Ayyas_finalData' deleted.". To the right, a 'Create dataset' dialog is open. It requires a 'Project ID' (set to 'project-sudheer-a5abf'), a 'Dataset ID' ('Project_result'), and a 'Region' (set to 'us-east1 (South Carolina)'). Other options include 'Default table expiration' (unchecked) and a 'Default maximum table age' of 'Days'. At the bottom are 'CREATE DATASET' and 'CANCEL' buttons.

A screenshot of the Google Cloud BigQuery Studio interface. On the left, the sidebar shows 'BigQuery Studio' selected under 'Analysis'. The main area displays a list of resources under 'project-sudheer-a5abf', including 'Project_result'. A context menu is open over 'Project_result', listing options: Open, Open in, Create table, Share, Copy ID, Refresh contents, and Delete. The 'Delete' option is highlighted with a black rectangle. The status bar at the bottom right shows the date as 22-04-2024.

A screenshot of the 'Create table' dialog in Google Cloud BigQuery Studio. The 'Source' section is set to 'Empty table'. The 'Destination' section shows 'Project * project-sudheer-a5abf' in the 'Project' field and 'Dataset * Project_result' in the 'Dataset' field. The 'Table' field is empty. The 'Table type' dropdown is set to 'Native table'. At the bottom, there are 'CREATE TABLE' and 'CANCEL' buttons. The status bar at the bottom right shows the date as 22-04-2024.

Create Tables:

The screenshot shows the 'Create table' dialog box in the Google Cloud BigQuery interface. The 'Source' section is set to 'Empty table'. The 'Destination' section includes 'Project' (project-sudheer-a5abf), 'Dataset' (Project_result), and 'Table' (result_data). The 'Table type' is set to 'Native table'. At the bottom, there are 'CREATE TABLE' and 'CANCEL' buttons.

The screenshot shows the 'result_data' table details page in the Google Cloud BigQuery interface. The table is located in the 'Project Sudheer' project under the 'Project_result' dataset. The table has a schema with one column named 'result'. The table was last modified on April 22, 2024, at 4:59:21 PM UTC+5:30. The table currently contains no rows.

The screenshot shows the Google Cloud BigQuery Studio interface. On the left, the sidebar has 'BigQuery Studio' selected under 'Analysis'. In the main area, there's a search bar and a tree view of resources. A dataset named 'result_data' is selected under 'Project_result'. The right panel shows the schema for 'result_data' with one column: 'id'. Below the schema, it says 'There is no data to display.'

Create Dataflow:

The screenshot shows the Google Cloud Dataflow Jobs page. The left sidebar has 'Jobs' selected. The main area lists several Dataflow jobs, including 'cloudstorage_to_bigquery' and 'storage_to_bigquery', along with their status, type, and creation time. The right side has sections for 'PRODUCTS & PAGES' (Dataflow, Jobs, Monitoring, Overview) and 'DOCUMENTATION & TUTORIALS' (Dataflow, Dataflow documentation, Run pipelines with Dataflow and Java, Run pipelines with Dataflow and Python).

The screenshot shows the Google Cloud Dataflow Jobs page. The URL in the address bar is `console.cloud.google.com/dataflow/jobs?referrer=search&project=project-sudheer-a5abf`. The page displays a table of jobs under the 'Jobs' tab. The columns include Name, Type, End time, Elapsed time, Start time, Status, SDK version, ID, Region, and Insights. There are filters for Running and Archived jobs, and a search bar at the top.

Name	Type	End time	Elapsed time	Start time	Status	SDK version	ID	Region	Insights
cloudstorage_to_bigquery	Batch	Apr 22, 2024, 4:20:59PM	2 min 3 sec	Apr 22, 2024, 4:18:56PM	Failed	2.54.0	2024-04-22_03_48_54-12805335723714367181	asia-south1	
cloudstorage_to_bigquery	Batch	Apr 22, 2024, 2:33:59PM	8 min 13 sec	Apr 22, 2024, 2:25:46PM	Succeeded	2.54.0	2024-04-22_01_55_44-10451570744809547121	us-central1	
storage_to_bigquery	Batch	Apr 22, 2024, 1:43:51PM	8 min 29 sec	Apr 22, 2024, 1:35:22PM	Failed	2.54.0	2024-04-22_01_05_19-5303186801584659231	asia-south1	
cloudstorage_to_bigquery	Batch	Apr 22, 2024, 12:00:26 AM	4 min 29 sec	Apr 21, 2024, 11:55:57PM	Failed	2.54.0	2024-04-21_11_25_54-7854230170929500880	asia-south1	
cloudstorage_to_bigquery	Batch	Apr 21, 2024, 6:43:27PM	4 min 38 sec	Apr 21, 2024, 6:38:49PM	Failed	2.54.0	2024-04-21_06_08_46-13010946342466047806	asia-south1	

The screenshot shows the 'Create job from template' dialog box. The URL in the address bar is `https://console.cloud.google.com/dataflow/createjob?project=project-sudheer-a5abf`. The dialog lists various template options under 'Text Files on Cloud Storage to BigQuery'. The 'Text Files on Cloud Storage to BigQuery' option is highlighted. At the bottom of the dialog, there is a 'RUN JOB' button and a note about equivalent REST or command line options.

Filter [Type to filter]

- Sourcedb to Spanner
- Spanner to BigQuery
- Text Files on Cloud Storage to BigQuery**
- Text Files on Cloud Storage to BigQuery with BigQuery Storage API & Python UDF support
- Text Files on Cloud Storage to Cloud Spanner
- Text Files on Cloud Storage to Firestore (Datastore mode)
- Utilities
- Bulk Compress Files on Cloud Storage

RUN JOB

Equivalent REST or command line

Screenshot of the Google Cloud Dataflow 'Create job from template' interface.

The 'Required Parameters' section includes:

- gs:// Cloud Storage Input File(s) ***: BROWSE button. Error message: "Error: value is required".
- gs:// Cloud Storage location of your BigQuery schema file, described as a JSON file with BigQuery Schema description. Example: { "name": "location", "type": "STRING" }, { "name": "name", "type": "STRING" }, { "name": "age", "type": "STRING" }, { "name": "color", "type": "STRING" }, { "name": "coffee", "type": "STRING" }] }**
- BigQuery output table ***: BROWSE button.
- gs:// Temporary directory for BigQuery loading process ***: BROWSE button. Description: "Temporary directory for BigQuery loading process (Example: gs://your-bucket/your-files/temp_dir)"
- gs:// Temporary location ***: BROWSE button.

The pipeline diagram shows a flow from "gs:// Cloud Storage Input File(s)" through "BigQueryCon...ToTableRow" to "Insert into Bigquery".

The right sidebar contains:

- How to use this Dataflow template**
- Cloud Storage text to BigQuery**
- Pipeline requirements**
 - A JSON file that describes your BigQuery schema must exist in Cloud Storage. Ensure that there is a top level JSON array titled "BigQuery Schema" and that its contents follow the pattern { "name": "COLUMN NAME", "type": "DATA TYPE" }. For example:

```
{
  "BigQuery Schema": [
    {
      ...
    }
  ]
}
```

Screenshot of the Google Cloud Dataflow 'Create job from template' interface, showing a file selection dialog.

The 'Required Parameters' section includes:

- gs:// Cloud Storage Input File(s) ***: BROWSE button. Error message: "Error: value is required".
- gs:// Cloud Storage location of your BigQuery schema file with BigQuery Schema description. Example: { "name": "location", "type": "STRING" }, { "name": "name", "type": "STRING" }, { "name": "age", "type": "STRING" }, { "name": "color", "type": "STRING" }, { "name": "coffee", "type": "STRING" }] }**
- BigQuery output table ***: BROWSE button.
- gs:// Temporary directory for BigQuery loading process**: BROWSE button. Description: "Temporary directory for BigQuery loading process (Example: gs://your-bucket/your-files/temp_dir)"
- gs:// Temporary location ***: BROWSE button.

The 'Select object' dialog shows a list of files in the folder "ayyas_destination_step1":

- OfficeData.csv
- dataschema1.json
- user.js

Buttons at the bottom of the dialog are **SELECT** and **CANCEL**.

The right sidebar contains:

- How to use this Dataflow template**
- Cloud Storage text to BigQuery**
- Pipeline requirements**
 - A JSON file that describes your BigQuery schema must exist in Cloud Storage. Ensure that there is a top level JSON array titled "BigQuery Schema" and that its contents follow the pattern { "name": "COLUMN NAME", "type": "DATA TYPE" }. For example:

```
{
  "BigQuery Schema": [
    {
      ...
    }
  ]
}
```

Screenshot of the Google Cloud Dataflow 'Create job from template' interface.

Required Parameters:

- Cloud Storage Input File(s) ***: `gs://ayyas_destination_step1/OfficeData.csv`
- Cloud Storage location of your BigQuery schema file, described as a JSON ***: `gs://ayyas_destination_step1/dataschema1.json`
- BigQuery output table ***: (Empty)
- Temporary directory for BigQuery loading process ***: `gs://temp_dir`
- Temporary location ***: `gs://temp`

Select object dialog open, showing files in `ayyas_destination_step1` bucket:

- `OfficeData.csv`
- `dataschema1.json` (Selected)
- `user.js`

Pipeline requirements:

A JSON file that describes your BigQuery schema must exist in Cloud Storage. Ensure that there is a top level JSON array titled "BigQuery Schema" and that its contents follow the pattern `{"name": "COLUMN NAME", "type": "DATA TYPE"}`. For example:

```
{
  "BigQuery Schema": [
    {
      ...
    }
  ]
}
```

Screenshot of the Google Cloud Dataflow 'Create job from template' interface, showing a pipeline diagram.

Required Parameters:

- Cloud Storage Input File(s) ***: `gs://ayyas_destination_step1/OfficeData.csv`
- Cloud Storage location of your BigQuery schema file, described as a JSON ***: `gs://ayyas_destination_step1/dataschema1.json`
- BigQuery output table ***: (Empty)
- Temporary directory for BigQuery loading process ***: `gs://temp_dir`

Pipeline Diagram:

```

graph TD
    A[BigQueryCon...ToTableRow] --> B[Insert into Bigquery]

```

Pipeline requirements:

A JSON file that describes your BigQuery schema must exist in Cloud Storage. Ensure that there is a top level JSON array titled "BigQuery Schema" and that its contents follow the pattern `{"name": "COLUMN NAME", "type": "DATA TYPE"}`. For example:

```
{
  "BigQuery Schema": [
    {
      ...
    }
  ]
}
```

Screenshot of the Google Cloud Platform Dataflow interface showing the "Select table" dialog.

The dialog lists datasets in the project "project-sudheer-a5abf".

Name	Description	Project
result_data	Dataset: Project_result	project-sudheer-a5abf

Buttons at the bottom: SELECT, CANCEL.

Right sidebar: "Cloud Storage text to BigQuery" tutorial and "Pipeline requirements" section.

Screenshot of the Google Cloud Platform Dataflow interface showing the "Create job from template" configuration screen.

Form fields under "Optional Parameters":

- JavaScript UDF path in Cloud Storage: gs://... (BROWSE button)
- JavaScript UDF name: transform_udf1
- Max workers: 1
- Number of workers: 1
- Worker region: us-central1

Right sidebar: "Cloud Storage text to BigQuery" tutorial and "Pipeline requirements" section.

Screenshot of the Google Cloud Dataflow 'Create job from template' interface.

The left sidebar shows navigation links: Buckets - Cloud Storage, Cloud Functions, BigQuery - Projects, Create job from template, Quotas & System, Create, and All Bookmarks.

The main form is titled 'Create job from template' and includes the following fields:

- Number of workers: 1
- Worker region: us-east1
- Worker zone: (dropdown)
- Use default machine type
- Service account email: Compute Engine default service account
- Additional experiments: (dropdown)

The right panel contains a 'How to use this Dataflow template' section and a 'Cloud Storage text to BigQuery' section. The 'Cloud Storage text to BigQuery' section describes the template's purpose and provides pipeline requirements and an example JSON schema.

```
BigQuery Schema : [ { }
```

Screenshot of the Google Cloud Dataflow 'Create job from template' interface, showing different configuration options.

The left sidebar shows navigation links: Buckets - Cloud Storage, Cloud Functions, BigQuery - Projects, Create job from template, Quotas & System, Create, and All Bookmarks.

The main form is titled 'Create job from template' and includes the following fields:

- Network: Unspecified
- Network: Dataflow workers can be configured to use public or internal IP addresses. [Learn more](#)
- Network: Network to which workers will be assigned. If empty or unspecified, the service will use the network 'default'.
- Subnetwork: Subnetwork to which workers will be assigned, if desired. Value can be either a complete URL or an abbreviated path. If the subnetwork is located in a Shared VPC network, you must use the complete URL. [Learn more](#)
- Enable Streaming Engine: Dataflow Streaming Engine moves pipeline execution out of the worker VMs and into the Dataflow service backend. This setting only applies to jobs launched from a streaming template. For batch templates, you can set the 'shuffle_mode=service' experiment flag.

A large blue 'RUN JOB' button is prominently displayed.

The right panel contains a 'How to use this Dataflow template' section and a 'Cloud Storage text to BigQuery' section. The 'Cloud Storage text to BigQuery' section describes the template's purpose and provides pipeline requirements and an example JSON schema.

```
BigQuery Schema : [ { }
```

Screenshot of Google Cloud Dataflow Job Overview (Job Graph View)

The screenshot shows the Google Cloud Dataflow interface for a job named "cloud_storage...".

Job Graph:

```
graph TD; A[Read from source] --> B[JavascriptTextTransformer.TransformTextViaJavascript];
```

Job Info:

- Job region: us-east1
- Worker location: us-east1
- Current workers: -
- Latest worker status: April 22, 2024 at 5:04:25 PM
- Start time: April 22, 2024 at 5:04:25 PM
- Elapsed time: 24 sec
- Encryption type: Google-managed
- Dataflow Prime: Disabled
- Runner v2: Enabled
- Dataflow Shuffle: Enabled

Resource metrics:

Screenshot of Google Cloud Dataflow Job Overview (Table View)

The screenshot shows the Google Cloud Dataflow interface for the same job, but with a different view.

Job Graph:

```
graph TD; A[Read from source] --> B[JavascriptTextTransformer.TransformTextViaJavascript];
```

Job Info:

- Job region: us-east1
- Worker location: us-east1
- Current workers: -
- Latest worker status: April 22, 2024 at 5:04:25 PM
- Start time: April 22, 2024 at 5:04:25 PM
- Elapsed time: 39 sec
- Encryption type: Google-managed
- Dataflow Prime: Disabled
- Runner v2: Enabled
- Dataflow Shuffle: Enabled

Resource metrics:

Screenshot of Google Cloud Dataflow Job Overview

Job Name: cloud_storage_to_bigquery

Job ID: 2024-04-22_04_34_24-10158601627659603172

Job type: Batch

Job status: Running

SDK version: Apache Beam SDK for Java 2.54.0

Job region: us-east1

Worker location: us-east1

Current workers: 1

Latest worker status: Worker pool started.

Start time: April 22, 2024 at 5:04:25 PM GMT+5

Elapsed time: 3 min 16 sec

Job steps view: Table view

Job steps table:

Step name	Status	Wall time	Stages	Input
Read from source	Running	0 seconds	F144	-
JavascriptTextTransformer.TransformTextViaJavascript	Starting...	0 seconds	F144	Read
BigQueryConverters.JsonTableRow	Starting...	0 seconds	F144	Jav
Insert into Bigquery	Running	0 seconds	F125	Big

Logs: SHOW

Screenshot of Google Cloud Dataflow Job Overview

Job Name: cloud_storage_to_bigquery

Job ID: 2024-04-22_04_34_24-10158601627659603172

Job type: Batch

Job status: Running

SDK version: Apache Beam SDK for Java 2.54.0

Job region: us-east1

Worker location: us-east1

Current workers: 1

Latest worker status: Worker pool started.

Start time: April 22, 2024 at 5:04:25 PM GMT+5

Elapsed time: 3 min 50 sec

Job steps view: Graph view

Job steps graph:

```
graph TD; A[Read from source  
Running  
1 of 2 stages succeeded] --> B[JavascriptT...Javascript  
Running  
0 of 1 stage succeeded]
```

Logs: SHOW

Screenshot of the Google Cloud Dataflow Job Overview page for job ID 2024-04-22_04_34_24-10158601627659603172.

Job info

Job name	cloud_storage_to_bigquery
Job ID	2024-04-22_04_34_24-10158601627659603172
Job type	Batch
Job status	Running
SDK version	Apache Beam SDK for Java 2.54.0
A newer version of the SDK family exists and updating is recommended. Learn more	
Job region	us-east1
Worker location	us-east1
Current workers	1
Latest worker status	Worker pool started.
Start time	April 22, 2024 at 5:04:25 PM GMT+5
Elapsed time	3 min 56 sec

JOB GRAPH

Job steps view: Graph view

```
graph TD; A[BigQueryCon...ToTableRow] --> B[Insert into Bigquery]
```

Execution Details

Logs

Screenshot of the Google Cloud Dataflow Job Overview page for job ID 2024-04-22_04_34_24-10158601627659603172.

Job info

Job name	cloud_storage_to_bigquery
Job ID	2024-04-22_04_34_24-10158601627659603172
Job type	Batch
Job status	Running
SDK version	Apache Beam SDK for Java 2.54.0
A newer version of the SDK family exists and updating is recommended. Learn more	
Job region	us-east1
Worker location	us-east1
Current workers	-
Latest worker status	Stopping worker pool.
Start time	April 22, 2024 at 5:04:25 PM GMT+5
Elapsed time	4 min 2 sec

JOB GRAPH

Job steps view: Graph view

```
graph TD; A[BigQueryCon...ToTableRow] --> B[Insert into Bigquery]
```

Execution Details

Logs

Screenshot of Google Cloud BigQuery Studio showing the schema of the 'result_data' table.

The schema consists of the following fields:

Field name	Type	Mode	Key	Collation	Default Value	Policy
employee_name	STRING	NULLABLE	-	-	-	-
department	STRING	NULLABLE	-	-	-	-
state	STRING	NULLABLE	-	-	-	-
salary	STRING	NULLABLE	-	-	-	-
age	STRING	NULLABLE	-	-	-	-
bonus	STRING	NULLABLE	-	-	-	-

Screenshot of Google Cloud BigQuery Studio showing the preview of the 'result_data' table.

The table has 10 rows of sample data:

Row	employee_name	department	state	salary
1	James	Sales	NY	90000
2	Michael	Sales	NY	86000
3	Robert	Sales	CA	81000
4	Maria	Finance	CA	90000
5	Raman	Finance	CA	99000
6	Scott	Finance	NY	83000
7	Jen	Finance	NY	79000
8	Jeff	Marketing	CA	80000
9	Kumar	Marketing	NY	91000