



NITTE
EDUCATION TRUST

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)

Nitte – 574 110, Karnataka, India

REPORT

ON

TWELVE WEEKS OF INTERNSHIP

Carried out at

NIVEUS-SOLUTIONS

Submitted to

NMAM INSTITUTE OF TECHNOLOGY, NITTE

(An Autonomous Institution under VTU, Belagavi)

In partial fulfillment of the requirements for the award of the

Degree of Bachelor of Engineering

in

Information Science & Engineering

by

Varun R Rao

USN 4NM20IS174

Under the guidance of

Dr. BOLA SUNIL KAMATH

Assistant Professor Gd-III



CERTIFICATE

*This is to certify that the “Internship report” submitted by **Mr./Ms.** Varun R Rao bearing USN 4NM20IS174 of ^{VIII} _semester B.E., a bonafide student of NMAM Institute of Technology, Nitte, has undergone twelve weeks of internship at niveus-solutions during **July 2023** fulfilling the partial requirements for the award of degree of Bachelor of Engineering in **Information Science & Engineering** at NMAM Institute of Technology, Nitte.*

Dr. BOLA SUNIL KAMATH

Name and Signature of Mentor

Signature of HOD

INDUSTRY CERTIFICATE



NIVEUS SOLUTIONS PVT. LTD.

 www.niveussolutions.com

 contact@niveussolutions.com

 (0820) 2520256

HR/Internship Letter/October / 2023-24

Date : 30 October 2023

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Varun R Rao has undergone internship with Niveus Solutions Pvt Ltd from **25th July 2023 to 23rd October 2023**.

During the period of internship, he has worked with the Customer Engineering team and has successfully met the objectives of the internship. We found Varun R Rao to be hardworking, sincere and displayed good conduct.

We wish him all success in future endeavors.

For M/s Niveus Solutions Pvt. Ltd

Yours sincerely,
For Niveus Solutions



Rashmi George - Chief Talent Officer

ACKNOWLEDGEMENT

I am thankful for the knowledge and skills I have acquired during my time at Niveus Solutions. The hands-on experience, the guidance of Mr. Mithesh Babu and Mr. Prasad Pai my mentors, and the collaborative work environment have provided me with a solid foundation for my future career.

I thank Dr. Abhishek Rao for referring my resume to the concerned authority of the company for getting my internship. I have learned not only about the industry and the role I was assigned but also about professionalism, teamwork, and dedication to excellence. I believe that the lessons I've learned and the relationships I've built during this internship will continue to benefit me as I move forward in my career. I look forward to staying in touch with the amazing people I've met at Niveus Solutions. It has been an invaluable experience, and I would like to extend my heartfelt appreciation for your guidance, support, and the chance to be a part of your organization.

TABLE OF CONTENTS

TITLE	PAGE NO
CERTIFICATE	i
INDUSTRY CERTIFICATE	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
ABSTRACT	1
INTRODUCTION	2
PROJECT DETAILS	3
CONCLUSION	16
REFERENCES	17

ABSTRACT

Niveus Solutions is a boot-strapped cloud engineering services organization. The company was founded by Suyog Shetty, Rashmi George, Roshan Bava, and Mohsin Khan in Karnataka, India. It is prominently well known for being a Google Cloud Platform Partner. The company specializes on cloud consulting, app modernization, infrastructure modernization, data modernization, platform migration, cloud-native application development, cloud security, and managed services.

The objectives achieved in this Internship were: -

- To Implement DLP in Google Cloud
- To develop a DocAI parser tool in Google Cloud Workbench
- To Integrate Big Query and Kafka in tiny bird tool

Learning outcomes of the internship include: -

- Understanding of Google Cloud Services and Architecture
- Computer Networking concepts
- Soft Skills and Team Work.

INTRODUCTION

Niveus Solutions Pvt. Ltd. is an independently financed cloud engineering services organization founded by esteemed professionals, namely Suyog Shetty, Rashmi George, Roshan Bava, and Mohsin Khan, in the state of Karnataka, India. The establishment of Niveus traces back to the year 2013 when the four founders, each possessing substantial experience garnered from esteemed corporations such as Infosys, Wipro, Cognizant, and Sapient, recognized a shared aspiration to establish a preeminent cloud engineering services enterprise. The founders were inspired by the exceptional talent reservoir present in the proximate educational centres of Udupi, Manipal, and Mangalore, which produce some of the nation's and the world's most brilliant minds. This realization fortified their belief that their collective vision could materialize into a world-class entity.

As a Google Cloud Platform Partner, Niveus leverages cloud technologies to help enterprises with cloud consulting, app modernization, infrastructure modernization, data modernization, platform migration, cloud-native application development, cloud security, and managed services. The organization empowers enterprises with the ability to harness the power of cloud services and build resilient infrastructures that scale.

PROJECT DETAILS

Objective 1: -

To Implement DLP in Google Cloud

In today's digital era, safeguarding personal data and ensuring compliance with data privacy regulations are paramount. Masking Aadhar numbers and other sensitive information is a crucial step in protecting personal data within organizations. Google Cloud Platform (GCP) provides a powerful solution for this through its Data Loss Prevention (DLP) API. This API offers a set of tools designed to automatically detect and protect sensitive information, such as Aadhar numbers, in your organization's data.

Methods to Mask Aadhar Numbers from an Image File:

1. Mask Aadhaar Number for Each Image using DLP Image Redact API:

Utilize the DLP Image Redact API to automatically detect and redact Aadhar numbers from image files.

This method ensures a seamless integration with GCP's DLP capabilities specifically designed for image data.

Go to the link: - <https://cloud.google.com/dlp/docs/reference/rest/v2/projects.image/redact/>

Convert the jpg image to base64 format and paste it in the data field and select info type as "INDIA_AADHAAR_INDIVIUDAL"

Click on execute and paste the base64 text in any tool and convert it back into jpg as shown in Fig 1 and output is show in in Fig 2.



Fig 1 Redacting Aadhar Number using Redact API



Fig 2 Redacted Output of Aadhaar Number using Redact API

2. Mask Aadhaar Number for Each Image using CLI (Cloud Shell):

Leverage the power of Google Cloud Shell to execute commands for masking Aadhaar numbers in image files.

This command-line approach provides flexibility and can be integrated into automated processes. Open cloud shell,

1) In Cloud Shell, run the following command to download the Cloud Data Loss Prevention Node.js Client repository:

git clone <https://github.com/googleapis/synthtool>

2) Once the project code is downloaded, change into the samples directory and install the required Node.js packages:

cd synthtool/tests/fixtures/nodejs-dlp/samples/ && npm install

3) To redact the aadhaar number from the image uid_image.jpg, run the following command:

node redactImage.js presales-286307 /home/varunrao_int/uid_image.jpg ""

INDIA_AADHAAR_INDIVIDUAL /home/varunrao_int/redacted-image.jpg as shown in Fig 3.



```
varunrao int@cloudshell:~/.../nodejs-dlp/samples (presales-286307)$ node redactImage.js presales-286307 /home/varunrao_int/uid_image.jpg "" INDIA_AADHAAR_INDIVIDUAL /home/varunrao_int/redacted-image.jpg
Saved image redaction results to path: /home/varunrao_int/redacted-image.jpg
varunrao int@cloudshell:~/.../nodejs-dlp/samples (presales-286307)$
```

Fig 3 Redacting Aadhaar Number using Cloud Shell Terminal

3. Mask Aadhaar Number for Each Image using Python Code:

Develop Python scripts using GCP's DLP API to programmatically mask Aadhar numbers in image files.

Python code allows for customization and integration with other processes within your organization. Results are shown in Fig 4.

Python Implementation: -

Run the python code either on terminal or vs code

Python Code:- redactionofaadhaarnumberpythonprogram.py

```
import google.cloud.dlp
def redact_image(
    project: str,
    filename: str,
    output_filename: str,
    info_types: list[str],
    min_likelihood: str = None,
) -> None:
    # Instantiate a client.
    dlp = google.cloud.dlp_v2.DlpServiceClient()
    # Prepare info_types by converting the list of strings into a list of
    # dictionaries (protos are also accepted).
    info_types = [{"name": info_type} for info_type in info_types]
    # Prepare image_redaction_configs, a list of dictionaries. Each dictionary
    # contains an info_type and optionally the color used for the replacement.
    # The color is omitted in this sample, so the default (black) will be used.
    image_redaction_configs = []
    if info_types is not None:
        for info_type in info_types:
            image_redaction_configs.append({"info_type": info_type})
    # Construct the configuration dictionary. Keys which are None may
    # optionally be omitted entirely.
    inspect_config = {
        "min_likelihood": min_likelihood,
        "info_types": info_types,
    }
    # Construct the byte_item, containing the file's byte data.
    with open(filename, mode="rb") as f:
        byte_item = {"type_": "IMAGE", "data": f.read()}
    # Convert the project id into a full resource id.
    parent = f"projects/{project}"
    # Call the API.
    response = dlp.redact_image(
        request={
            "parent": parent,
            "inspect_config": inspect_config,
            "image_redaction_configs": image_redaction_configs,
            "byte_item": byte_item,
        }
    )
```

```
# Write out the results.
with open(output_filename, mode="wb") as f:
f.write(response.redacted_image)
print(
"Wrote {byte_count} bytes to {filename}".format(
byte_count=len(response.redacted_image), filename=output_filename
)
)
redact_image("presales-286307",'uid_image.jpg','redacted_image.jpg',
["INDIA_AADHAAR_INDIVIDUAL"],"POSSIBLE")
```

```
user@IND040100320:~/Downloads/myfile$ /bin/python3 /home/niveus/Downloads/myfile/a43.py
Wrote 28040 bytes to redacted_image.jpg
```



Fig 4 Redacting Aadhar Number using Python Code

4. Mask Aadhaar Number for Multiple Images using Triggers and Jobs:

Implement triggers and jobs to automate the masking of Aadhar numbers across multiple image files.

This method is ideal for organizations dealing with a large volume of image data, providing efficiency and scalability.

Using Triggers and Jobs:

Go to the link: -

<https://console.cloud.google.com/security/sensitive-data-protection/landing/configuration/templates/inspect?project=presales-286307>
and create an inspection template as shown in Fig 5 and template is created as Fig 6.

Sensitive Data Protection					
OVERVIEW	DISCOVERY	INSPECTION	RISK ANALYSIS	CONFIGURATION	SUBSCRIPTIONS
TEMPLATES	INFOTYPES	DATA PROFILES			
INSPECT	DE-IDENTIFY	CREATE TEMPLATE			
Template ID	Display name ↑	Resource location	Creation time	Last updated	Actions
8598175705373805301	Auto-created profiler tem...	Global (any region)	Aug 2, 2023, 3:19:28 PM	Aug 2, 2023, 3:19:28 PM	⋮
aadhaarinspect	aadhaarinspect	Global (any region)	Aug 2, 2023, 1:36:50 PM	Aug 2, 2023, 7:56:56 PM	⋮
aadhaarinspect_test	aadhaarinspect_test	Global (any region)	Aug 3, 2023, 10:41:12 AM	Aug 3, 2023, 10:41:12 A...	⋮
aadharinspect	aadharinspect	Global (any region)	Aug 7, 2023, 8:33:11 AM	Aug 7, 2023, 9:09:44 AM	⋮
eti-accelerator-remove	eti-accelerator-remove	Mumbai (asia-sout...	Jun 17, 2022, 11:35:34 ...	Jun 17, 2022, 11:35:34 ...	⋮

Fig 5 Inspect Template

aadharinspect

projects/presales-286307/locations/global/inspectTemplates/aadharinspect

Basic info

Display name	aadharinspect
Description	inspect the aadhar
Built-in infoTypes	INDIA_AADHAAR_INDIVIDUAL
Minimum likelihood	Possible
Content types	Unspecified
Include quote	Off
Exclude infoTypes	Off
Create time	Aug 7, 2023, 8:33:11 AM
Update time	Aug 7, 2023, 9:09:44 AM

Fig 6 Completion of inspection template

Similarly create two de identifcation templates under deidentify as shown in Fig 7 and Fig 8

Deidentify_aadhar

projects/presales-286307/locations/global/deidentifyTemplates/Deidentify_aadhar

CONFIGURATION

TEST

Basic info

Create time	Aug 7, 2023, 8:48:43 AM
Update time	Aug 7, 2023, 8:48:43 AM
Transformation type	InfoType
Display name	Deidentify_aadhar
Description	masks aadhar number

Configuration

Transformation rule	Redact the following infoType(s): "INDIA_AADHAAR_INDIVIDUAL".
---------------------	---------------------------------------------------------------

Fig 7 Configuration of Deidentification Template

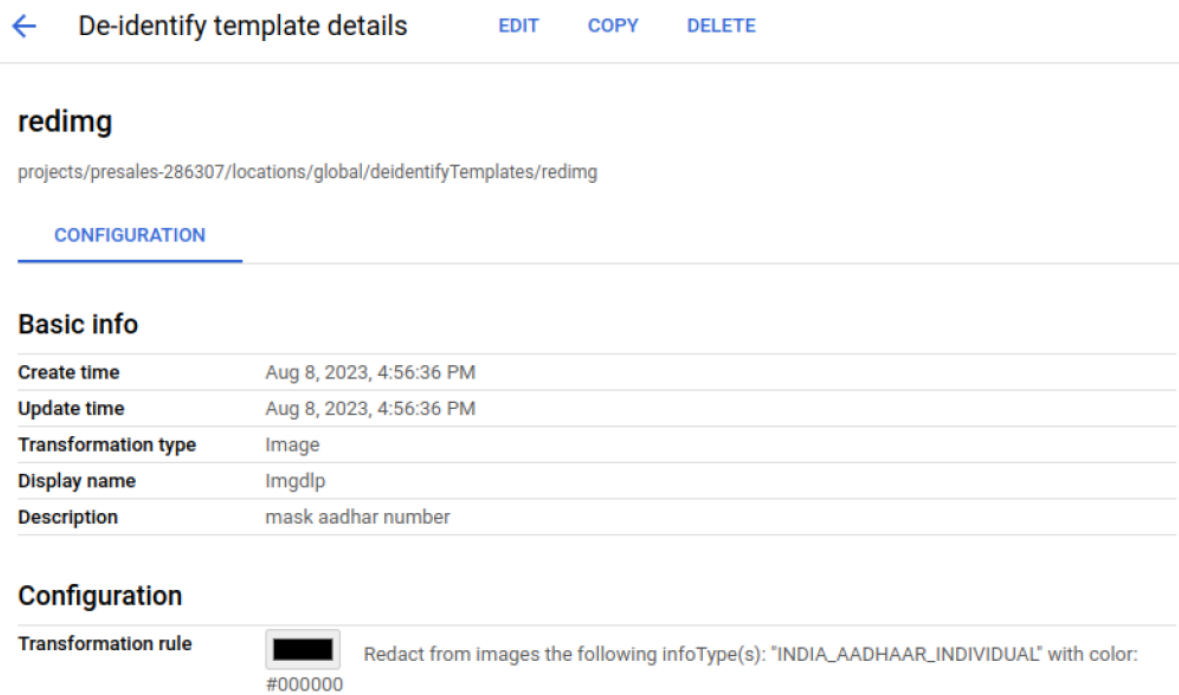


Fig 8 Completion of deidentification template

Create a bucket for input containing sample aadhaar image and another bucket for output as shown in Fig 9 and Fig 10

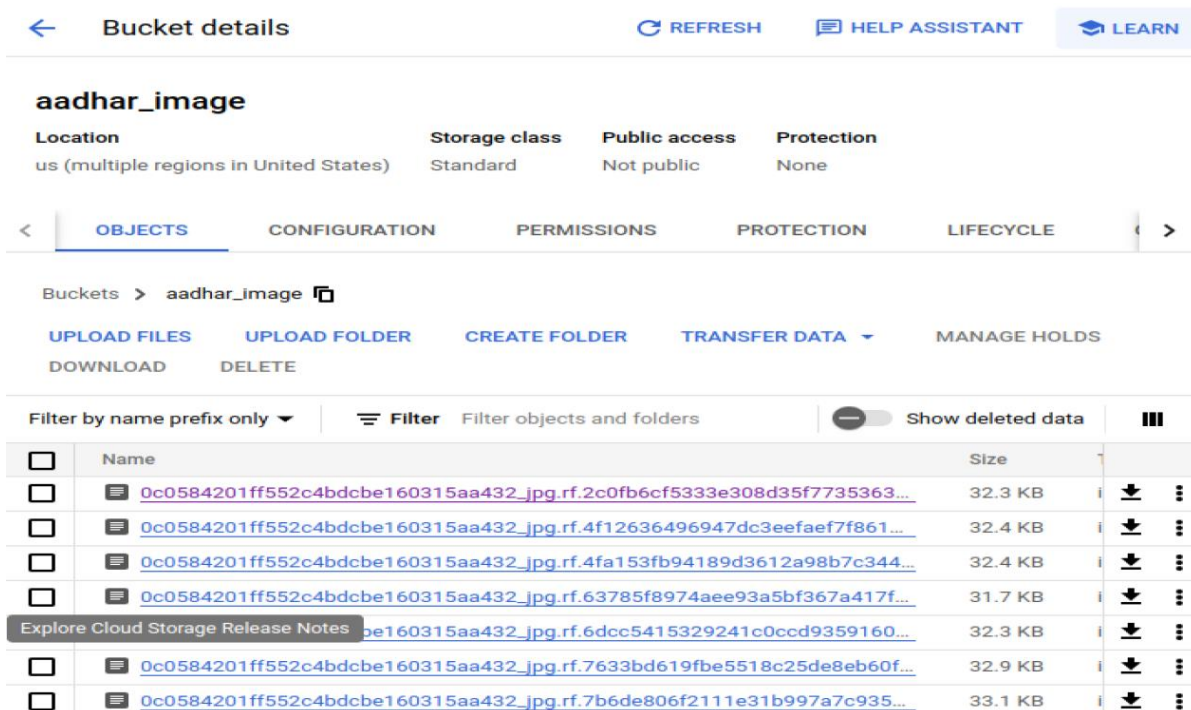


Fig 9 Input bucket to store Aadhar images

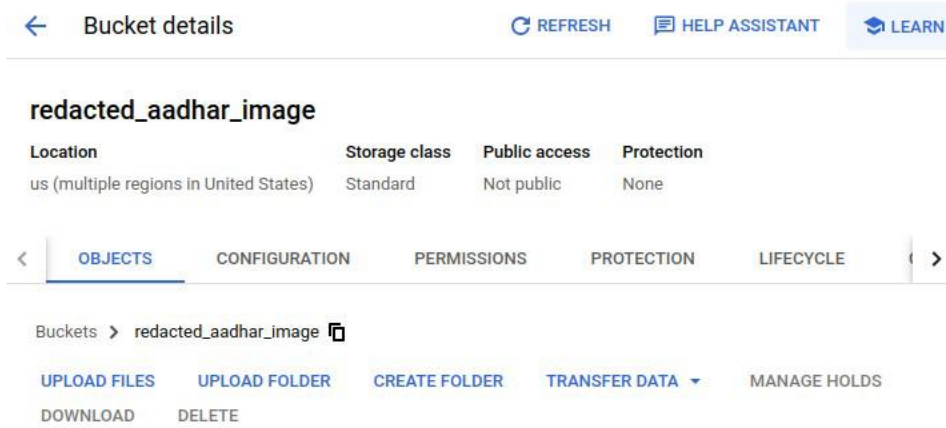


Fig 10 Output bucket for storing redacted Aadhar image

To create the job/trigger click on Create Job and Job Triggers as shown in Fig 11

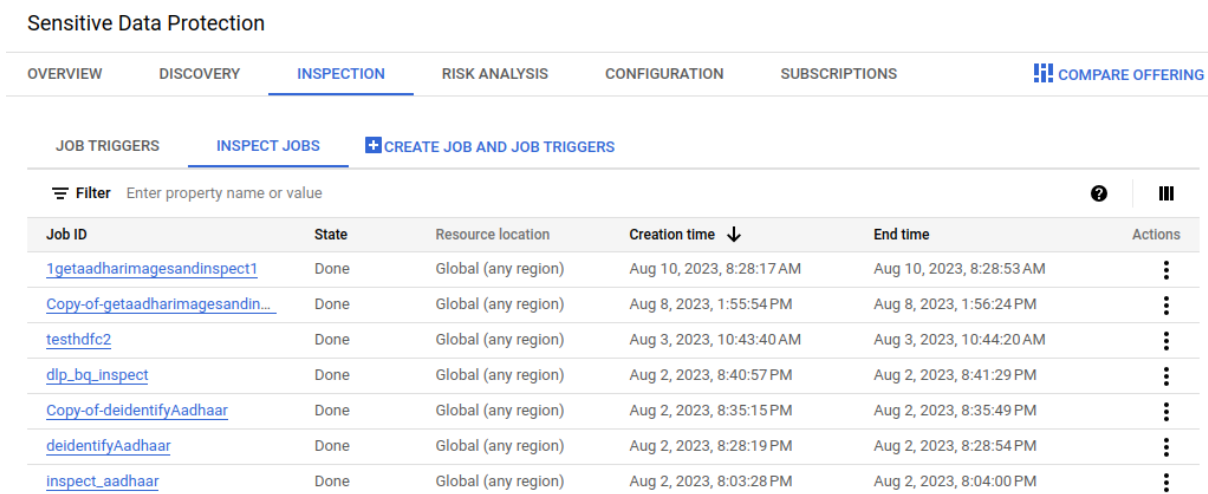


Fig 11 Job creation for masking Aadhar image

And configure a job manually with below configuration and click on “create” and “confirm” as shown in Fig 11 and Fig 12. Fig 13 shows the final output

```
{
  "jobId": "getaadharimagesandinspect",
  "inspectJob": {
    "actions": [
      {
        "deidentify": {
          "fileTypesToTransform": [
            "TEXT_FILE",
            "IMAGE",
            "CSV",
            "TSV"
          ],
          "transformationDetailsStorageConfig": {},
          "transformationConfig": {
            "deidentifyTemplate": "projects/presales-286307/locations/global/deidentifyTemplates/Deidentify_aadhar",
            "imageRedactTemplate": "projects/presales-286307/locations/global/deidentifyTemplates/redimg"
          },
          "cloudStorageOutput": "gs://redacted_aadhar_image"
        }
      },
      {
        "inspectConfig": {
          "infoTypes": [
            {
              "name": "INDIA_AADHAAR_INDIVIDUAL"
            }
          ],
          "minLikelihood": "POSSIBLE",
          "customInfoTypes": []
        },
        "inspectTemplateName": "projects/presales-286307/locations/global/inspectTemplates/aadhaarinspect",
        "storageConfig": {
          "cloudStorageOptions": {
            "filesLimitPercent": 100,
            "fileTypes": [
              "TEXT_FILE",
              "IMAGE",
              "WORD",
              "PDF",
              "AVRO",
              "CSV",
              "TSV",
              "EXCEL",
              "POWERPOINT"
            ],
            "fileSet": {
              "url": "gs://aadhar_image"
            }
          }
        }
      }
    ]
  }
}
```

←

Job details

COPY

CANCEL

DELETE

getaadharimagesandinspect

Container: gs://aadhar_image

projects/presales-286307/locations/global/dlp/jobs/i-getaadharimagesandinspect

✓ Done

OVERVIEW


CONFIGURATION

VIEW FINDINGS IN BIGQUERY



?


Findings	Bytes scanned	Errors
11	357.75 KiB	0

Transformations	Bytes transformed	Transformation errors
11	92.98 KiB	0




Bucket details


 REFRESH
  HELP


Buckets > redacted_aadhar_image 

[UPLOAD FILES](#)
[UPLOAD FOLDER](#)
[CREATE FOLDER](#)
[TRANSFER DATA](#)

[DOWNLOAD](#)
[DELETE](#)

Filter by name prefix only 

 Filter
 Filter objects and folders

 Show

Go to the output bucket location to verify results:-

←

Bucket details

REFRESH

HELP ASK

Buckets > redacted_aadhar_image

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

DOWNLOAD




DELETE

Filter by name prefix only

Filter Filter objects and folders

Filter

Sho

<input type="checkbox"/>	Name
<input type="checkbox"/>	 0c0584201ff552c4bdcbe160315aa432.jpg.rf.2c0fb6cf5333e308d35f7735363...
<input type="checkbox"/>	 0c0584201ff552c4bdcbe160315aa432.jpg.rf.4f12636496947dc3eeafef7f861...
<input type="checkbox"/>	 0c0584201ff552c4bdcbe160315aa432.jpg.rf.4fa153fb94189d3612a98b7c344...

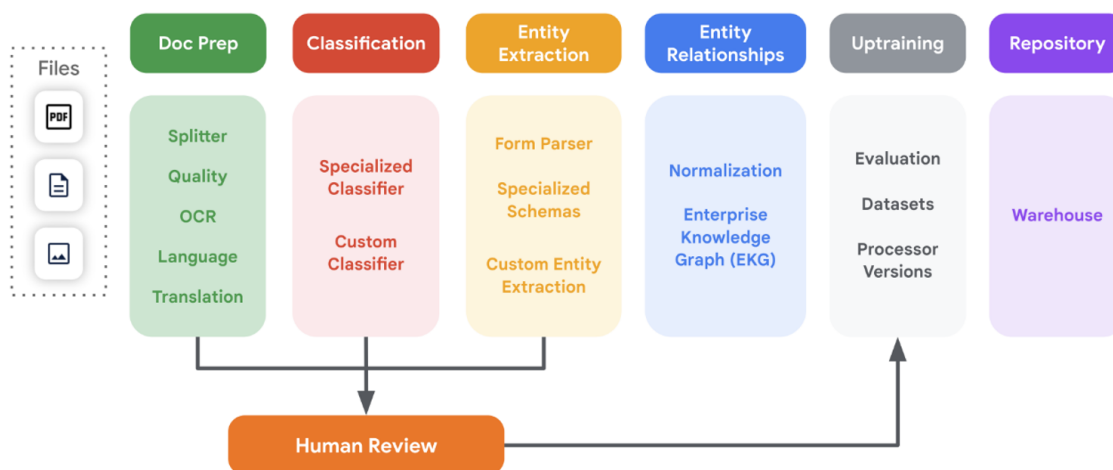


Objective 2: -

To develop a DocAI parser tool in Google Cloud Workbench

Developers seeking to create high-accuracy processors for document extraction, classification, and splitting can leverage generative AI solutions to streamline tasks. This innovative approach enables the efficient extraction and structuring of data within minutes. The use cases for such processors span various industries and scenarios, providing solutions to challenges like digitizing books, simplifying medical intake processes, and automating expense report generation. Refer Fig 14 for processing steps.

Fig 14 Document processing steps:-



Prepare training data as shown in Fig 15

The screenshot shows the Document AI interface. On the left, the 'Processors' tab is selected, displaying a list of training data. The data includes fields like 'supplier_name', 'supplier_address', 'invoice_date', 'receiver_name', 'receiver_address', 'total_amount', and 'invoice_type'. The 'total_amount' field is highlighted with a red box, showing the value '25,775.00'. Below the list, there is a 'MARK AS LABELED' button. On the right, a 'Commercial Invoice' document is displayed. The invoice is from 'Tabcodefg Limited' and includes details such as 'AIRWAY BILL NO.', 'INVOICE NO.', 'INVOICE DATE', 'DATE OF EXPORT', 'EXPORTER/SHIPPER', 'SHIP TO / CONSIGNEE', 'COMPANY NAME', 'ADDRESS', 'CONTACT NAME', 'PHONE/ FAX', 'EMAIL', and 'COUNTRY OF EXPORT'. The 'Total Value' is highlighted with a red box, showing the value '25,775.00'.

Fig 15 Training data Commercial invoice

Train the model as shown in Fig 16

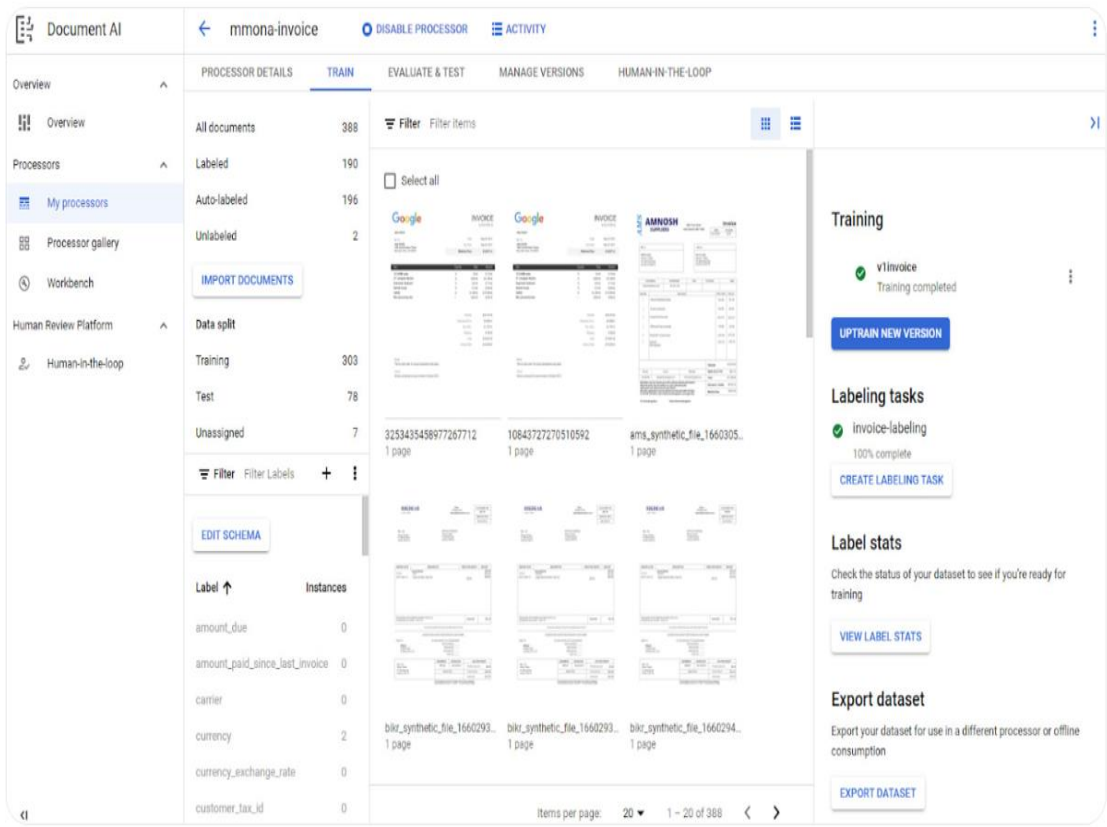


Fig 16 Model is being trained using the dataset

Evaluation of Performance metrics as shown in Fig 17

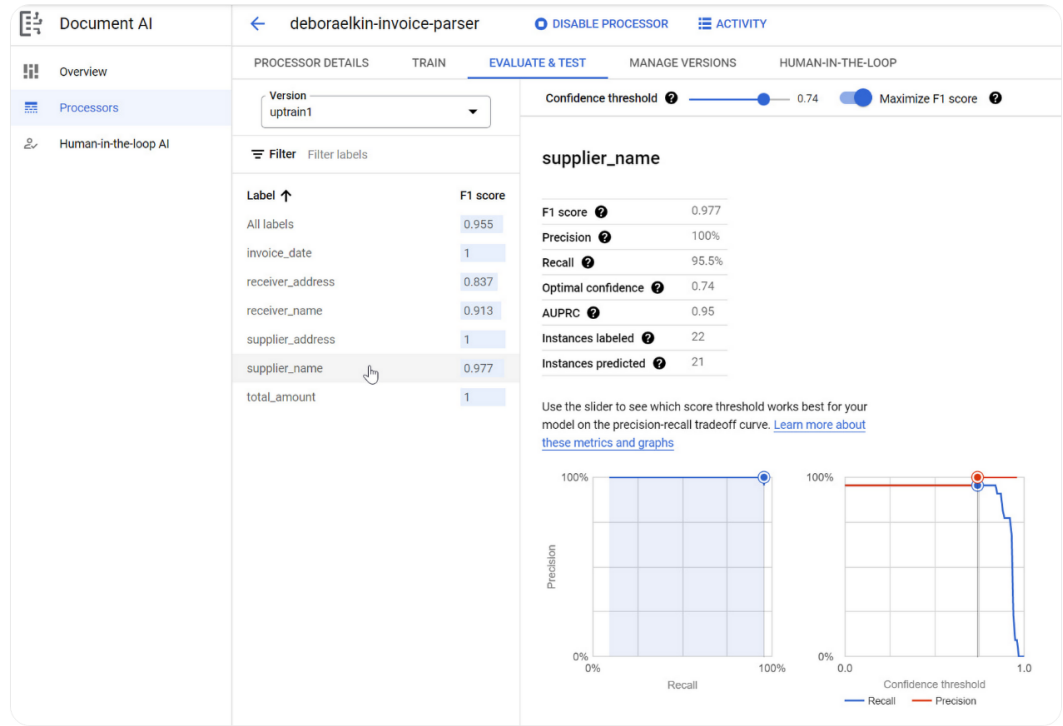


Fig 17 Evaluation of Performance metrics

- **To integrate Big Query and Kafka in tiny bird tool**

Tiny bird unifies the batch and streaming data. It is a real-time data analytics platform
Ingests data at scale, supports SQL and Git, and publishes API endpoints.

Grant access to the principal in Tinybird tool as shown in Fig 18

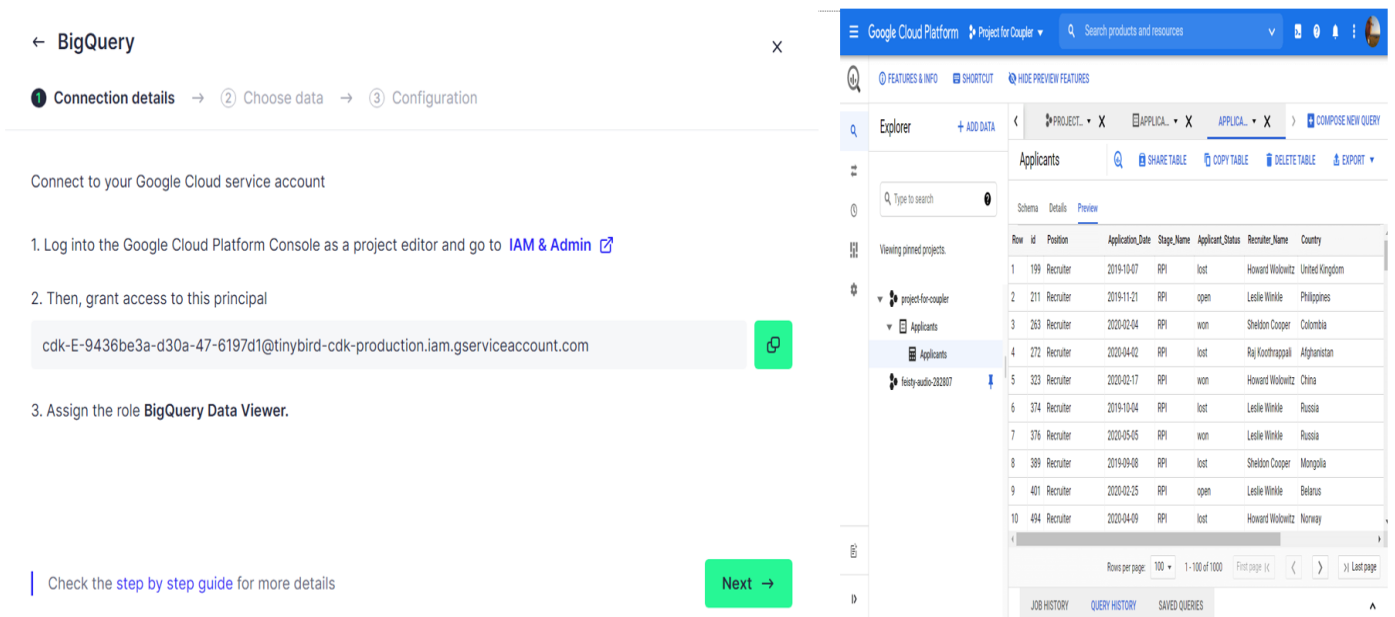


Fig 18 Granting access to principle from google cloud console in tinybird

Connect to kafka bootstrap server from confluent cloud as shown in Fig 19

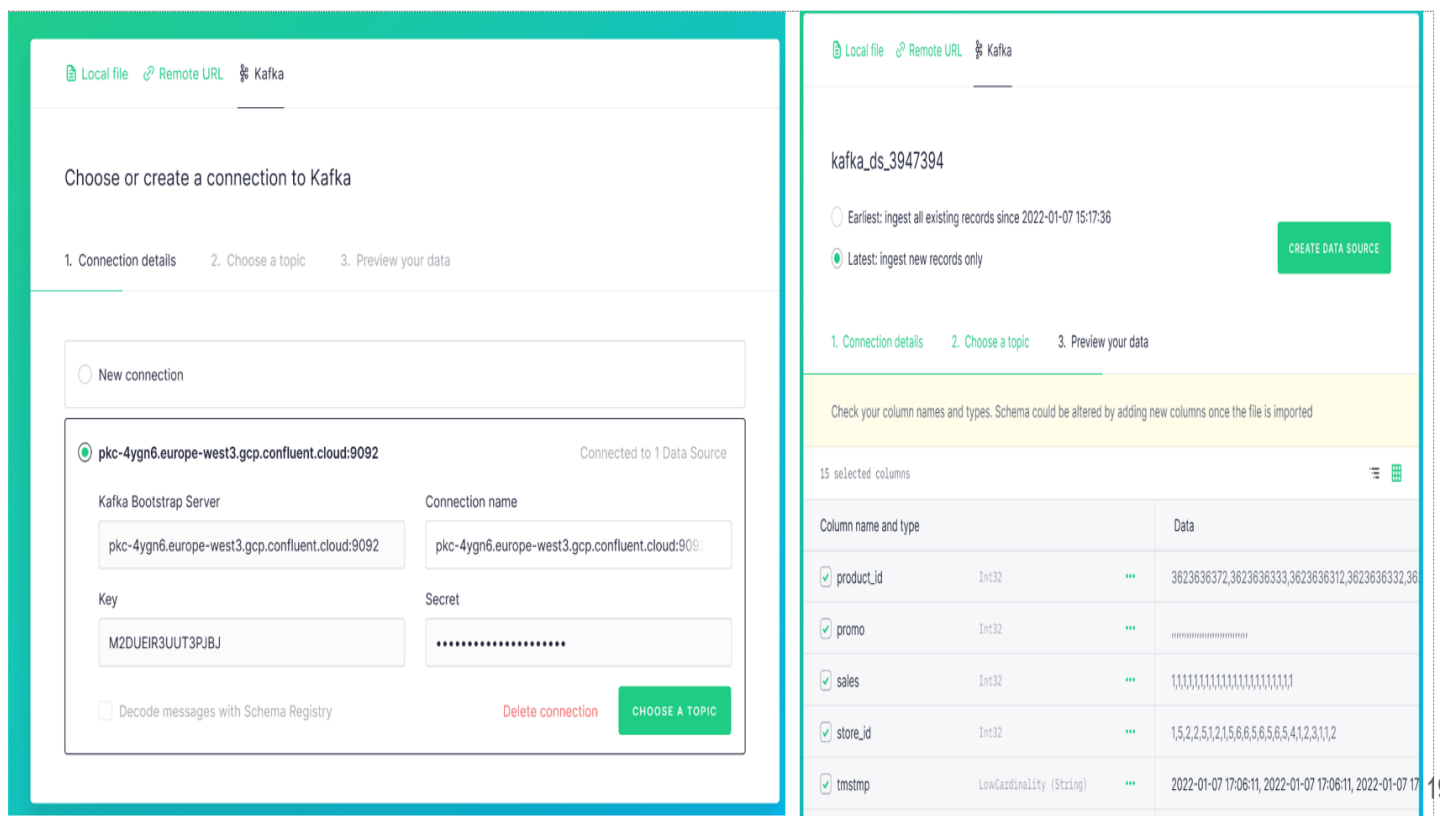


Fig 19 Connecting to kafka bootstrap server in tinybird

Table from Bigquery and stream from kafka successfully integrated as shown in Fig 20

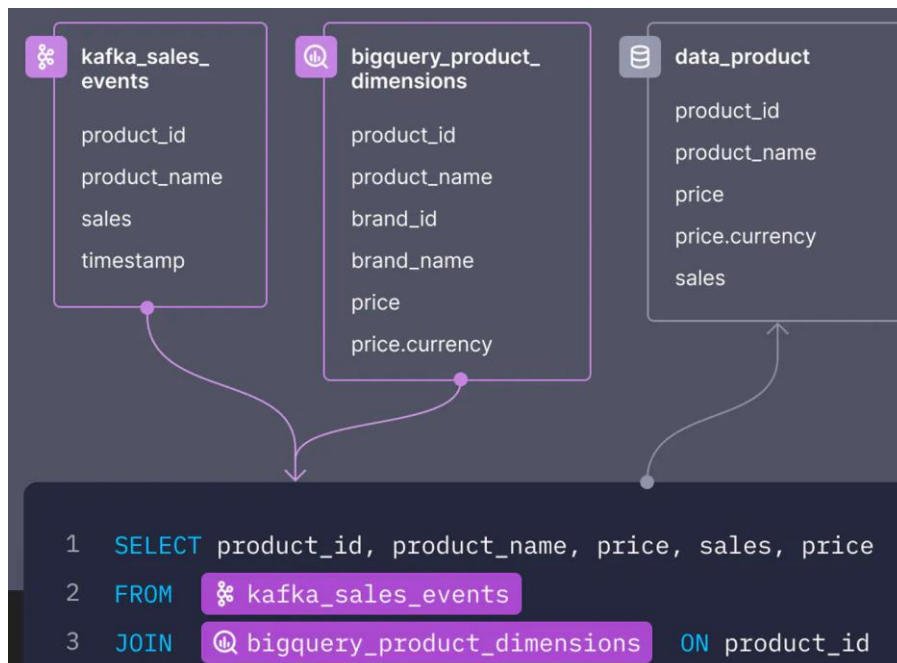


Fig 20 Integration of table from Bigquery and stream from Kafka
Create API endpoint of the above query as shown in Fig 21

Events Kafka Remote URL Local file

1 Copy the snippet, 2 run the script and 3 a new Data Source will be created.

Using our Events API

CURL PYTHON JS GO RUST PHP RUBY JAVA

```
fetch(
  'https://api.tinybird.co/v0/events?name=events_example',
  {
    method: 'POST',
    body: JSON.stringify({
      date: new Date().toISOString(),
      city: 'Pretoria'
    }),
    headers: { Authorization: 'Bearer p.eyJ1IjogIjUxNjVlMTFkLTNiMDYtNDExNy05MTg1LTNmZjY1kNGNjMGQzZiIsICJpZC'
  }
)
```

Read our docs Copy snippet

Fig 21 API endpoint configuration using fetch request

CONCLUSION

In conclusion, my internship journey has been marked by the successful achievement of key objectives. Implementing Data Loss Prevention (DLP) in Google Cloud has deepened my understanding of data security, emphasizing the importance of safeguarding sensitive information. The development of a DocAI parser tool in Google Cloud Workbench showcased my ability to harness advanced document processing tools, enabling the extraction of valuable insights from various types of documents.

Moreover, integrating BigQuery and Kafka in the Tiny Bird tool reflects my proficiency in connecting cloud-based data warehousing with distributed streaming platforms. This achievement highlights my capacity to enhance data processing capabilities and facilitate real-time analytics. Overall, these accomplishments collectively underscore my evolving skills in cloud technologies and data management, providing a solid foundation for tackling the dynamic challenges of the information technology landscape.

Furthermore, this internship has not only equipped me with technical skills but has also honed my ability to navigate complex and interdisciplinary projects. Collaborating on these objectives has fostered effective communication, teamwork, and problem-solving skills. The hands-on experience gained in implementing DLP, developing a DocAI parser tool, and integrating BigQuery and Kafka in the Tiny Bird tool has not only expanded my technical toolkit but has also deepened my understanding of the broader implications and applications of these technologies in real-world scenarios. As I reflect on the internship experience, I am confident that these achievements will serve as a solid foundation for future endeavours in the ever-evolving landscape of cloud computing and data-driven technologies.

REFERENCES

- [1] <https://cloud.google.com/dlp/docs>
- [2] <https://cloud.google.com/document-ai/docs>
- [3] <https://www.tinybird.co/docs>
- [4] <https://www.confluent.io/product/confluent-platform/>
- [5] <https://cloud.google.com/free/docs/free-cloud-features#bigquery/>
- [6] <https://cloud.google.com/iam/docs/overview/>
- [7] <https://cloud.google.com/shell/docs/use-cloud-shell-terminal/>
- [8] <https://cloud.google.com/shell/docs/configuring-cloud-shell/>
- [9] <https://cloud.google.com/shell/docs/deploy-app-engine-app>