

# REIMAGINE'23

Google Cloud Delivery Summit 2023 | May 31 – July 31

IDEATHON

*Casadona Warriors*

OBSCURER

#GrowandDisruptwithInfosysCobalt

# Contents

Team and Ideathon Intro	<a href="#">1</a>	API Endpoints with descriptions	<a href="#">13</a>
Team and Use Case	<a href="#">3</a>	API Testing on Postman	<a href="#">14</a>
Use Case Definition	<a href="#">4</a>	Deployment Process	<a href="#">15</a>
What is Obscurer?	<a href="#">6</a>	API Endpoint and Looker Dashboard	<a href="#">16</a>
Benefits of using Obscurer	<a href="#">7</a>	Error Handling and Incident Management	<a href="#">17</a>
Flow Diagram	<a href="#">8</a>	Product Demo	<a href="#">18</a>
How did we arrive at this solution?	<a href="#">9</a>	FAQ	<a href="#">19</a>
How does the Obscurer Data Pipeline work?	<a href="#">10</a>	Possible Usage in Other Industries	<a href="#">20</a>
Design and Architecture	<a href="#">11</a>	Appendix – Technology Stack	<a href="#">21</a>
FastAPI – Python Script Build	<a href="#">12</a>		

## Team and Use Case

Casadona Warriors	Team Name	Team Member Name	Mail Id	PU
		Sagnik Das	Sagnik.das03@infosys.com	DNA
		Somdutta Paul	Somdutta.paul@infosys.com	ENG
		Tania Rana	Tania.rana@infosys.com	IVS

### Use Case Name

Protected Healthcare Information (PHI) Data Pipeline  
for unstructured data management.

### Use Case description

Healthcare industry faces a consistent challenge due to magnanimous quantity of unstructured data like prescription, health reports and medical imaging data. However, it lacks a streamlined process through which we can convert both scanned printed documents and handwritten medical instructions. Obscurer would bring a change in the above process by incorporating AI by Google in the data pipeline and helps in conversion of unstructured data to a structured format and stores EHR (Electronic Health Record) information post PII deidentification. This data can be further used by BI tools for visualization and extraction of meaningful insights.

**OBSCURER**

Infosys  
cobalt

# Use Case Definition

## Open Use Case - Data Processing Pipeline on Google Cloud



### ➤ Context

Healthcare industry in the US is \$800+ million industry and it is expected to grow even more in the future. More than 80% of the healthcare data is unstructured due to minimal usage of IoT devices. Majority of EHR (Electronic Health Record) are not structured due to the fact that the scanned documents are never passed through a powerful OCR (Optical Character Recognition) engine.

### ➤ Business Relevance

Extracting text from data files and masking PHI information can help hospitals and healthcare processing units to unlock insights from their data and comply with data privacy regulations. Moreover, displaying metadata information and file processing status on a dashboard can help the organizations to monitor and optimize their data pipeline performance and efficiency.

### ➤ Business Solution

A data pipeline code written in Python/FastAPI (deployed on Google App Engine) can upload data to GCP Cloud Storage and convert them to text format using Google Cloud Document AI. Then it can mask PII information in the text using Google Cloud DLP API and gives the output text files as output. On the other hand, the metadata information and the file processing status can be stored and processed in GCP BigQuery and displayed on a Looker Dashboard for data analysis and reporting.

**OBSCURER**

Infosys  
cobalt



# Solution Architecture








Technical Solution



# What is Obscure?

Obscure is a data pipeline application that uses FastAPI and Google Cloud Platform to perform text extraction and PII deidentification on PDF documents or images in the healthcare domain. It also helps identify documents that contain medicine names or compositions.

## Supported File Types:

Name		File Extension(s)	MIME Type
Portable Document Format (PDF)		.pdf	application/pdf
Graphics Interchange Format (GIF)		.gif	image/gif
Text Format (TXT)		.txt	text/txt
Joint Photographic Experts Group (JPEG)		.jpg, .jpeg	image/jpeg
Portable Network Graphics (PNG)		.png	image/png
Tag Image File Format (TIFF)		.tiff, .tif	image/tiff
Bitmap (BMP)		.bmp	image/bmp

**OBSCURER**

Infosys  
cobalt

# Benefits of Using Obscurer

## ➤ Secure Storage



Obscurer uses a highly secure mode of cloud storage where there is minimal risk of data breach. Both the raw processed and deidentified files are stored in the same way inside fine-grained access controlled google cloud storage bucket.

## ➤ PHI (Protected Health Care Information) Redaction

Anonymity is critical while handling crucial health care information as these are highly confidential . But at times it might be required that data scientist and data analyst use these information to get data insights or create data models. For compliance with HIPAA, Obscurer would provide a seamless deidentified data output for usage by non-medical professionals.

## ➤ Smart Medicine Name Extraction



As there is a huge amount of unstructured data in the health care industry, there is a need to identify medicine names from a medical report, prescription , doctor slip etc. Obscurer helps in automatic medicine name extraction which can be used for both dashboard creation and for ai model building .

## ➤ Unified Dashboard Experience

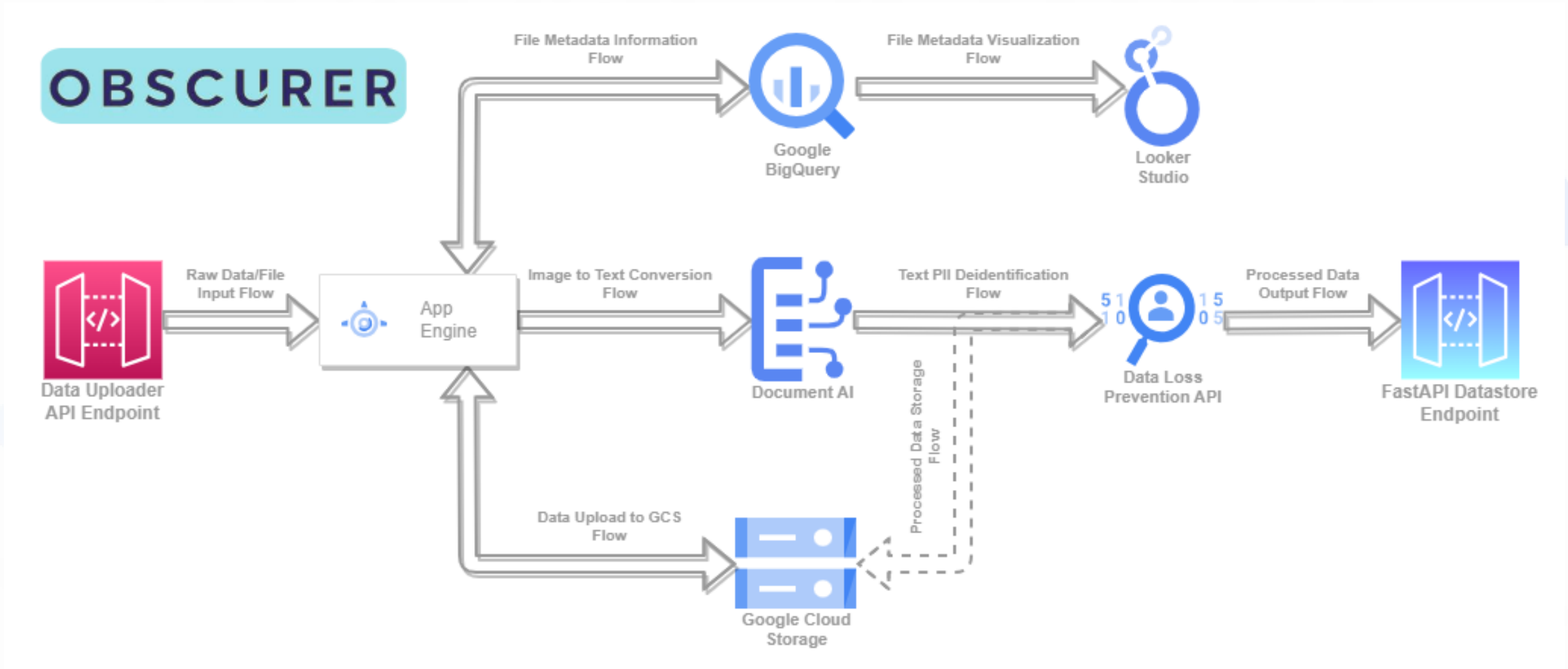


Although obscurer is primarily a data pipeline module however it also provides a viewer dashboard which gives an easy analytics view for the data being processed along with the summary .

**OBSCURER**

Infosys  
cobalt

## Flow Diagram





## How did we arrive this solution?

Requirement

- Conversion of unstructured data to structured data
- Masking of confidential information
- Centralized storage and fast retrieval of secured data

Planning

Design

- Google App Engine is suitable for deployment of a FastAPI application on cloud
- Google Cloud Document AI provides seamless conversion of graphical data and documents
- Cloud DLP helps in masking of confidential information
- FastAPI provides the user interaction endpoint through Swagger UI
- All the backend services are intertwined within the App Engine

Implementation

- FastAPI application will be deployed and run as PaaS

## How does the Obscurer Data Pipeline work?

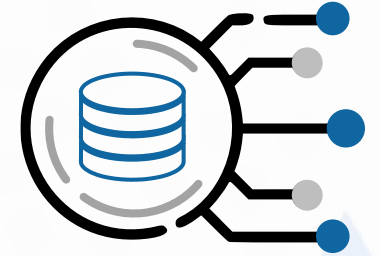
User uploads a document or an image to the API through an http client such as Postman or through the integrated Swagger UI

The data processing pipeline asynchronously starts automatically as soon as the 'Upload' is triggered.

User can check the status of the data processing through the Looker dashboard.

Once the processing is complete user can fetch the PII deidentified document in text format by choosing one of the two fetch functions.

User can also check if the stated document contains a medicine name by using Looker or an appropriate API Endpoint.



**OBSCURER**

Infosys  
cobalt



## Design

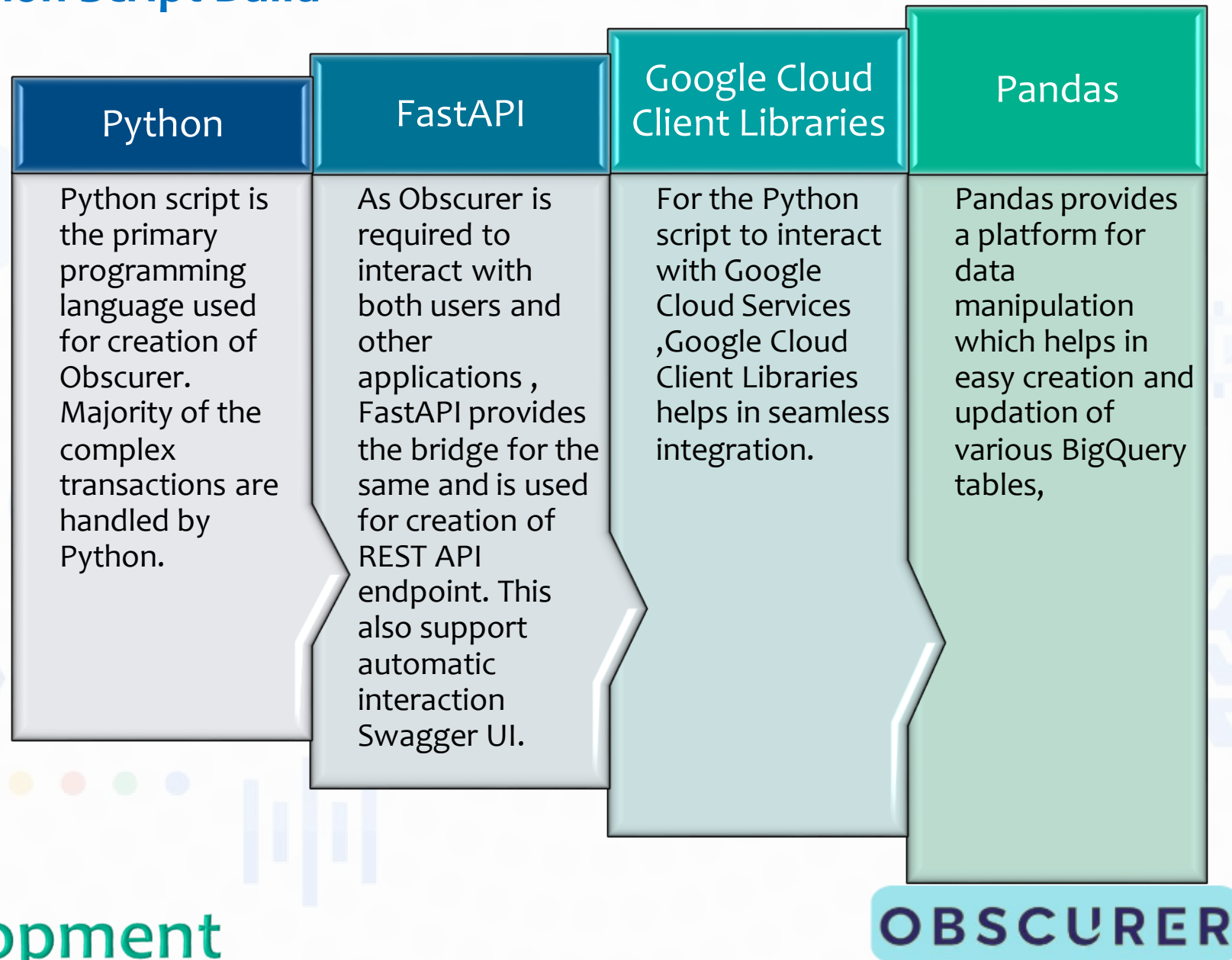
The Data pipeline code designed as a Python code that uses the FastAPI framework to create a webservice that can accept data files as input and written text files as output. This script also uses the google-document-ai and google-cloud-dlp libraries to interact with Google Cloud Document AI and the Google Cloud DLP API respectively. This script also uses the google-cloud-storage and google-cloud-bigquery libraries to interact with the GCS and GCP BigQuery services respectively.



## Architecture

The data pipeline code is deployed on Google Cloud App Engine which are serverless platforms that runs stateless containers or applications on a fully managed environment. Google Cloud App Engine automatically scales up or down the instances based on the demand and only charges for the resources used. The data pipeline code connects to various GCP services such as Cloud Storage, Cloud Document AI, Cloud DLP and BigQuery. The data pipeline code also connects to the Looker dashboard using the BigQuery connector that allows Looker to access and visualize the metadata information and file processing status in BigQuery. The data pipeline code can be triggered periodically or on demand by the data engineer or by an external application. The data analyst and the business user can access the looker dashboard using a web browser and view the metadata information and the file processing status for each files.

## FastAPI – Python Script Build



## API Endpoints with descriptions :



URL	Endpoint	Parameters	Endpoint Description	Type of Request
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/upload	Request body	Endpoint for uploading and start of data pipeline. The endpoint is used when transferring the media data itself.	POST
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/update_metatables		Endpoint is useful for manual metadata updation. If the resource contains any data fields, those fields are used to store metadata describing the uploaded file.	PATCH
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/update_bq_schema		Define an endpoint to run all the SQL files in parallel. It is use for repairing the existing views in the Big Query.	PATCH
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/fetch	name	Endpoint useful for fetching data as JSON. It is used for fetching the deidentified file as JSON format.	POST
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/download	name	Endpoint useful for downloading text. It is used to download documents and images .	POST
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/processed_files_list		Endpoint is useful for fetching list of files processed. In these endpoint we can view the list of files that are already processed.	POST
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/count_files_processed		Endpoint is useful for fetching count of files processed. In these endpoint we can view the count of files that are already processed.	POST
https://gcds-oh3219u9-2023.uc.r.appspot.com/	/fetch_medicine_names	filename	Endpoint is useful for fetching the medicine name from the files processed from BiqQuery.	POST



# API Testing on Postman

Response time and latency for each API endpoint has been duly monitored and measures have been taken to improve performance

Multiple API calls has been made rigorously by using Postman collections to check if Obscurer is able to handle multiple request at once

Different types of files has been uploaded to check if the application is able to handle the content types as pre-decided

For the files processed deidentified texts have been extracted and tested if they are correctly processed

File counts have been manually verified on one side calling the 'File Count API' and the other side writing a script on BigQuery

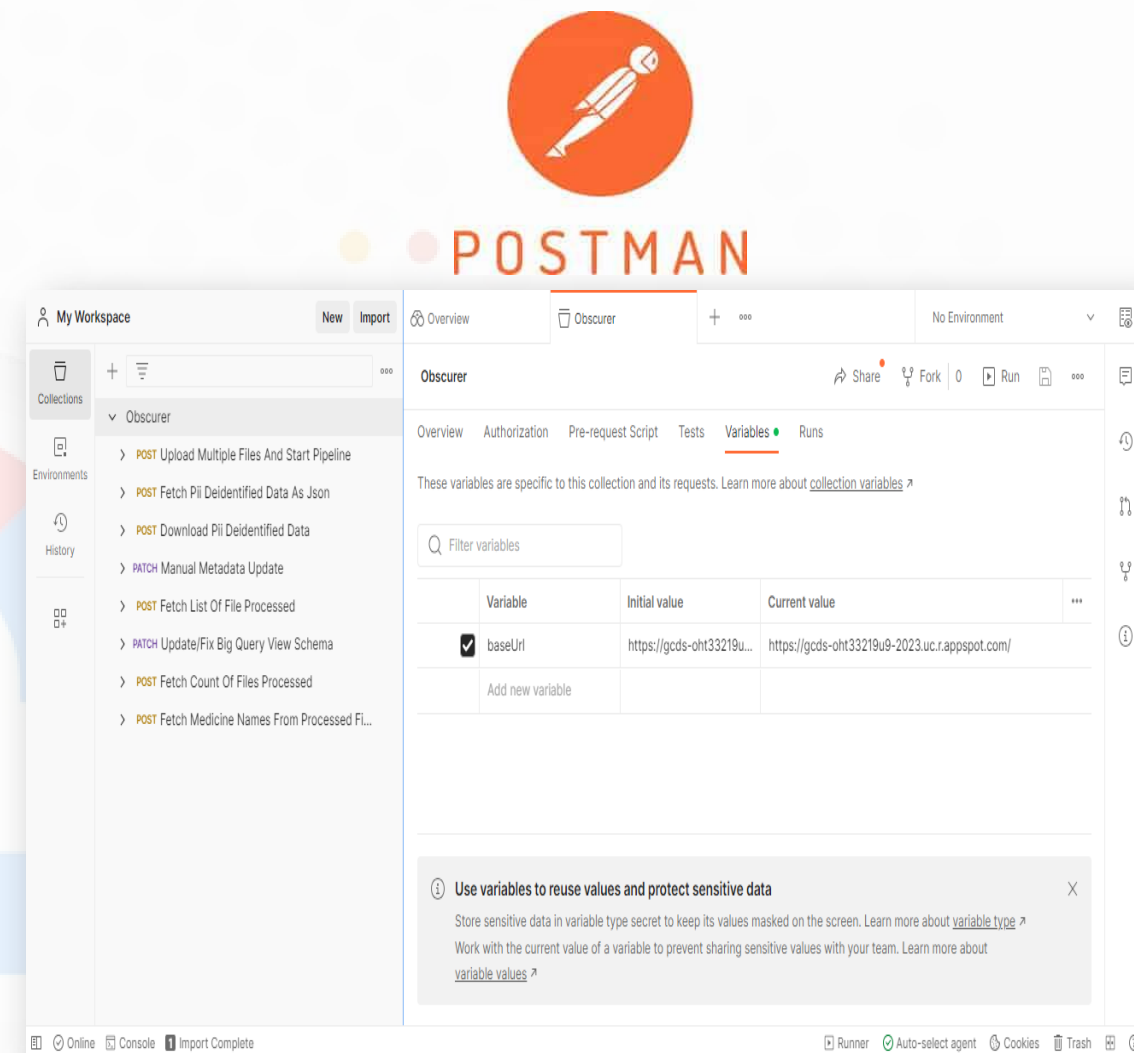
Pass ✓

Pass ✓

Pass ✓

Pass ✓

Pass ✓

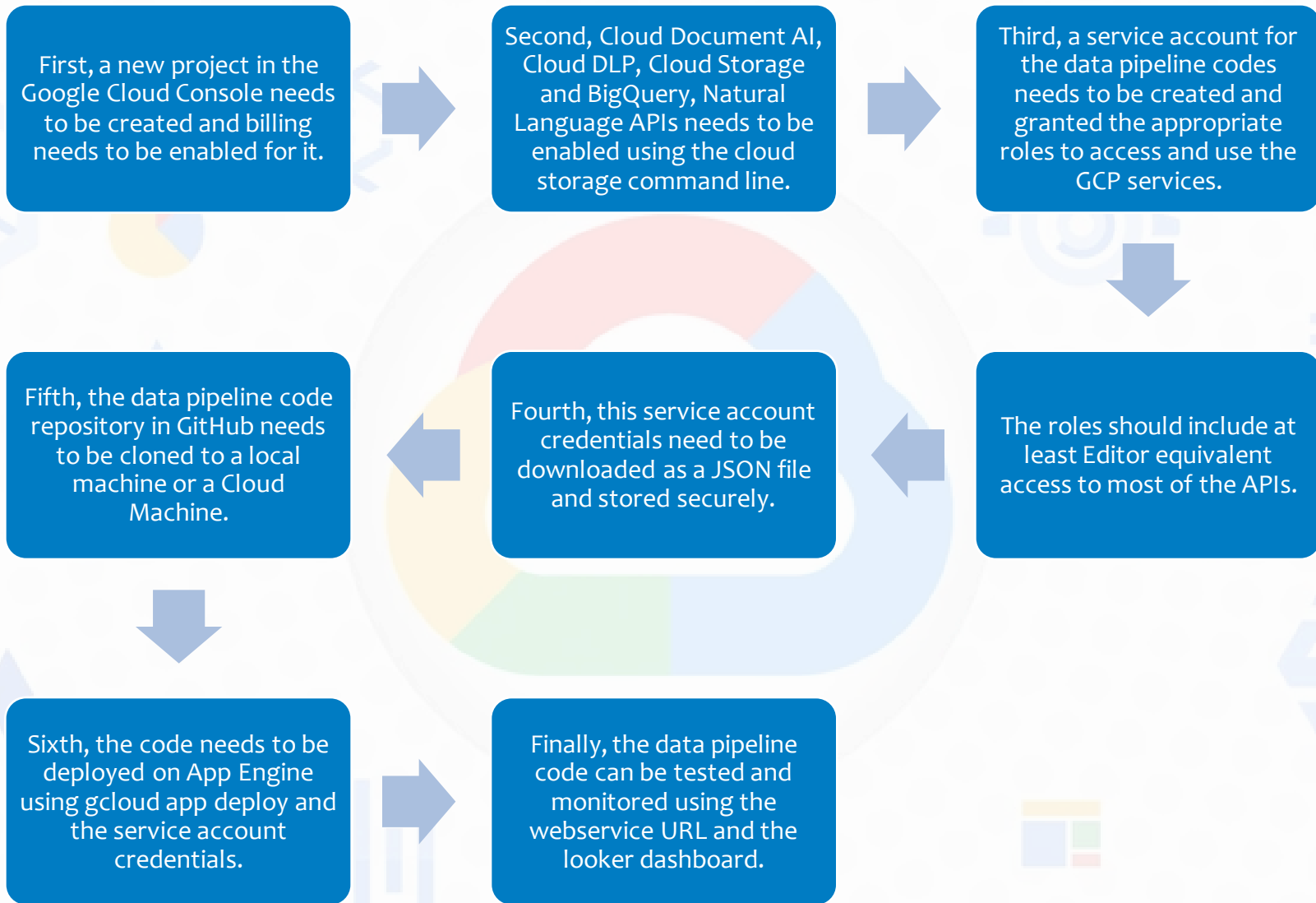


## ⑤ Testing

OBSCURER

Infosys  
cobalt

# Deployment Process



# API Endpoint and Looker Dashboard

## Obscure<sup>1.0.0</sup> <sup>OAS 3.1</sup>

/openapi.json

Obscure is a data pipeline application that uses FastAPI and Google Cloud Platform to perform text extraction and PII deidentification on PDF documents or images in the healthcare domain. It also helps identify documents that contain medicine names or compositions.

Contact Developer - Sagnik Das, Tania Rana, Somdutta Paul

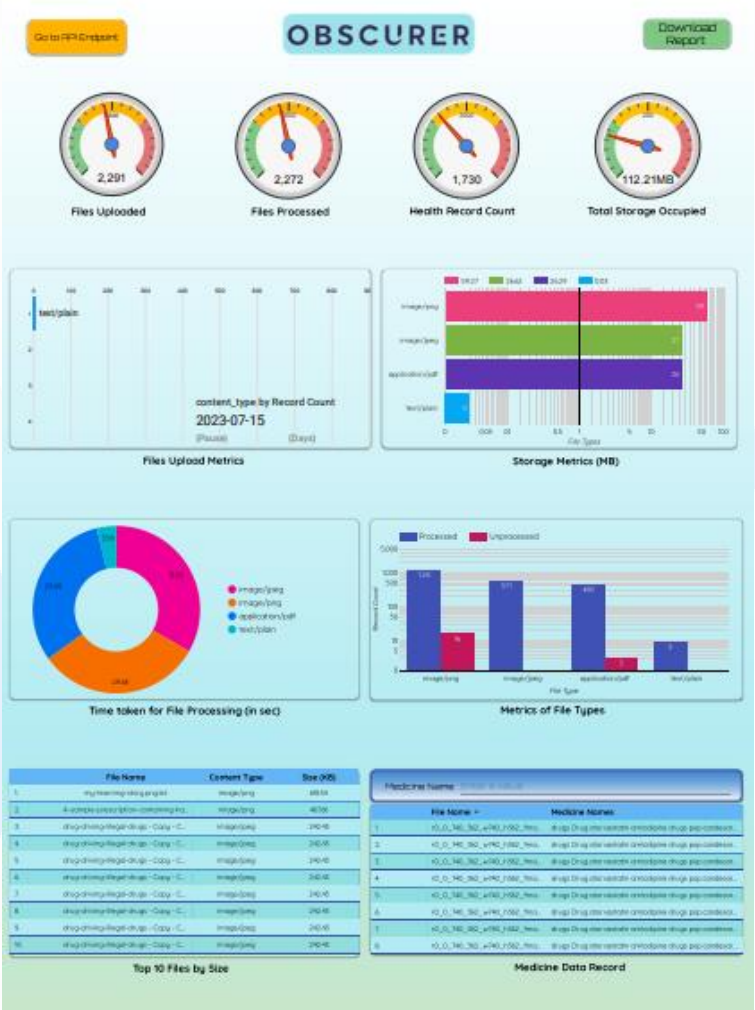
### Data Pipeline

POST	/upload	Upload Multiple Files And Start Pipeline
PATCH	/update_metatables	Manual Metadata Update
PATCH	/update_bq_schema	Update/Fix Big Query View Schema

### Stream Data

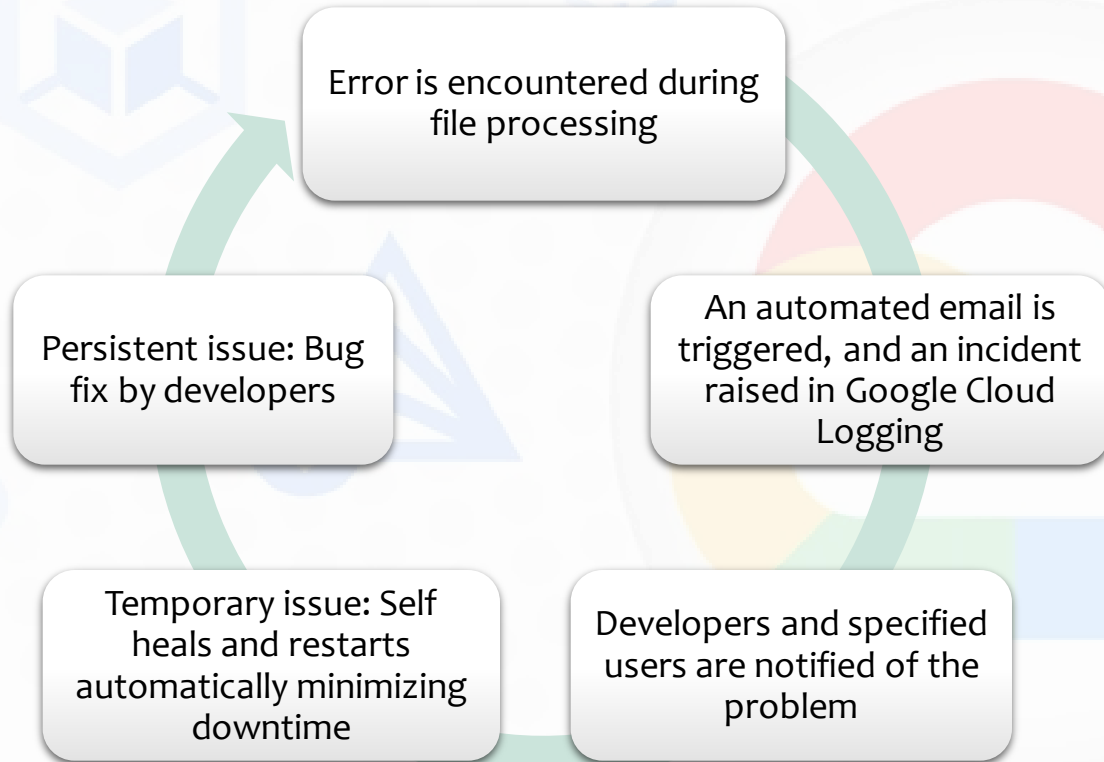
POST	/fetch	Fetch Pii Deidentified Data As Json
POST	/download	Download Pii Deidentified Data
POST	/processed_files_list	Fetch List Of File Processed
POST	/count_files_processed	Fetch Count Of Files Processed
POST	/fetch_medicine_names	Fetch Medicine Names From Processed Files

[Obscure – Swagger](#)



[Obscure - Looker](#)

## Error Handling and Incident Management



Google  
Cloud Logging

**Cloud Logging**, which is a real-time log-management system with storage, search, analysis, and monitoring support. Cloud Logging automatically collects logs from Google Cloud resources. We can also collect logs from the application. If we want to view the error rates in Google Cloud service, then we can view the Cloud Logging dashboard, which is preconfigured.

## Product Demo

- Obscurer API Endpoints
- Obscurer Looker Dashboard
- Glance of Obscurer Source Code



## FAQ

**Why do we need Obscurer in health care industry?**

Obscurer consolidates multiple GCP AI services which helps in confidential data handling with features like OCR, Redaction, Entity Identification.

**How do we ensure encryption for PHI information ?**

By default, GCP always encrypts all customer data at rest as well as in motion. This encryption is automatic, and it requires no action on the user's part. Persistent disks, for instance, are already encrypted using AES-256, and the keys themselves are encrypted with master keys.

**Why are we converting all the images and documents to text format?**

Healthcare industry handles a lot of both scanned images and handwritten notes which are hard to decrypt unless an OCR is performed. Obscurer thus, handles all the unstructured data in text format for easy usage in other application.

**Is Obscurer an End User product?**

No, Obscurer is a part and parcel of a larger data processing pipeline. Obscurer can be used wherever there is a large amount of unprocessed data, and a single brewed application is required for data preprocessing.

**How do we detect medicine names although we're not using a large medicine database internally?**

We're using Google Pre-trained AI models for entity detection from natural language text.

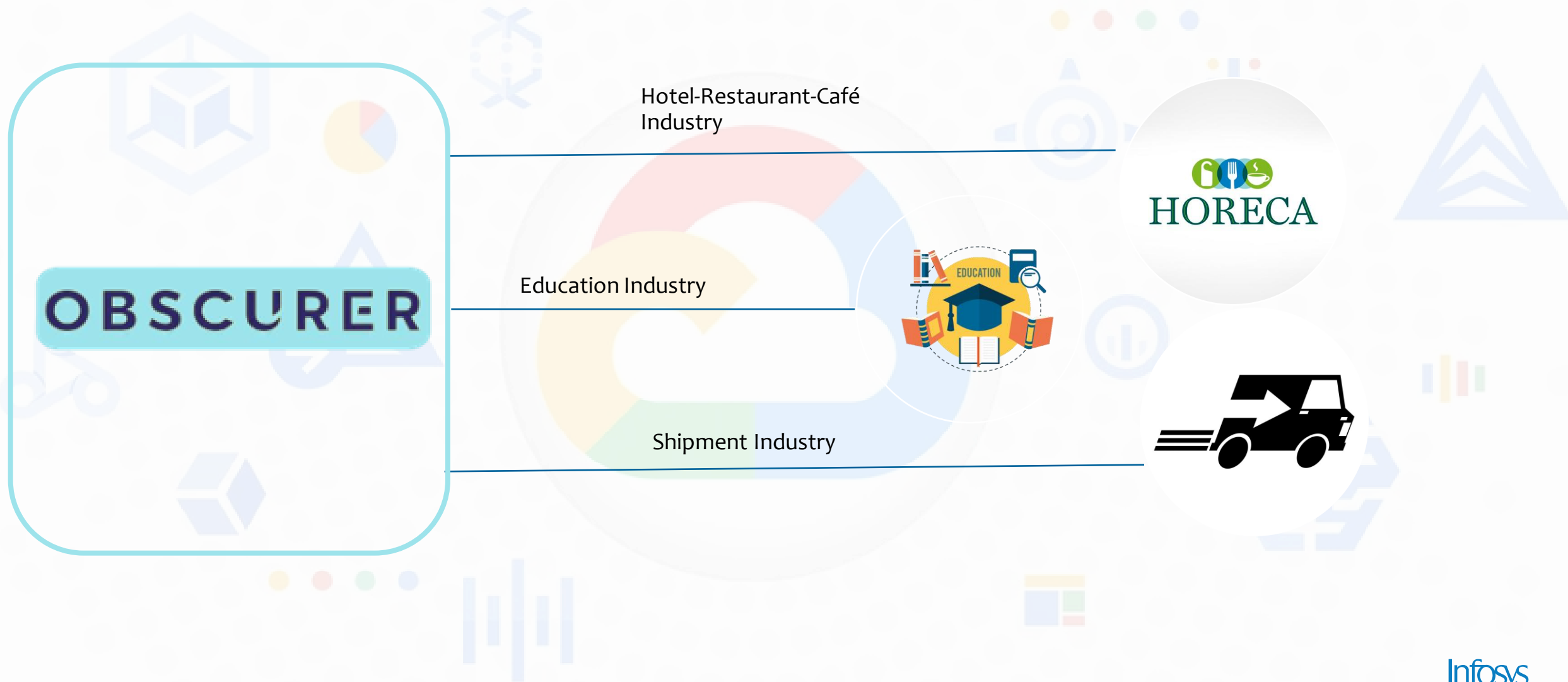
**Why do we use Fast API/Python in Obscurer?**

Fast API enables asynchronous API calls with very fast response time. Python also enables easy integration with all GCP services including but not limited to Google Cloud Document AI, Google DLP, Google BigQuery.

**OBSCURER**

infosys  
cobalt

## Possible Usage in Other Industries



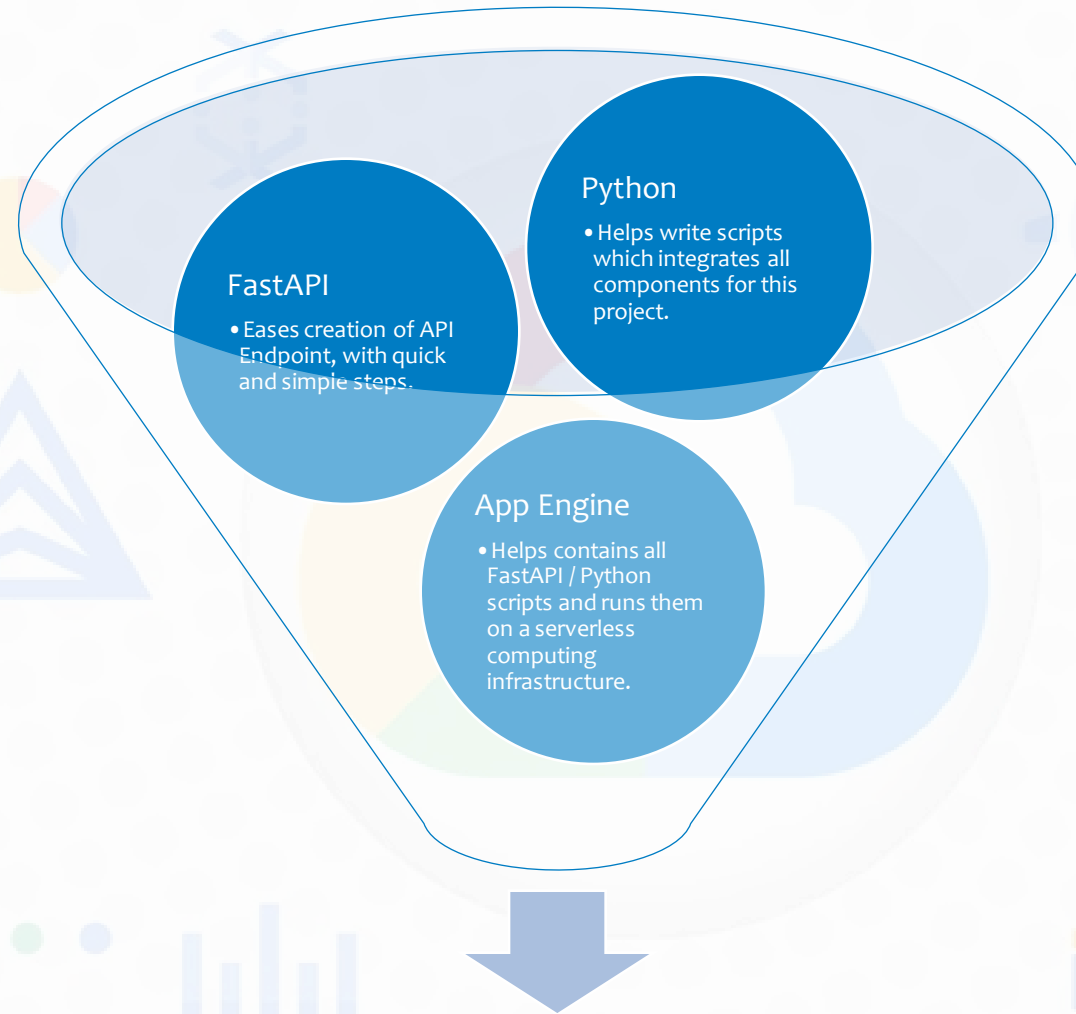
# Appendix >>

## Technology Stack

Technical Solution



## Backend Services



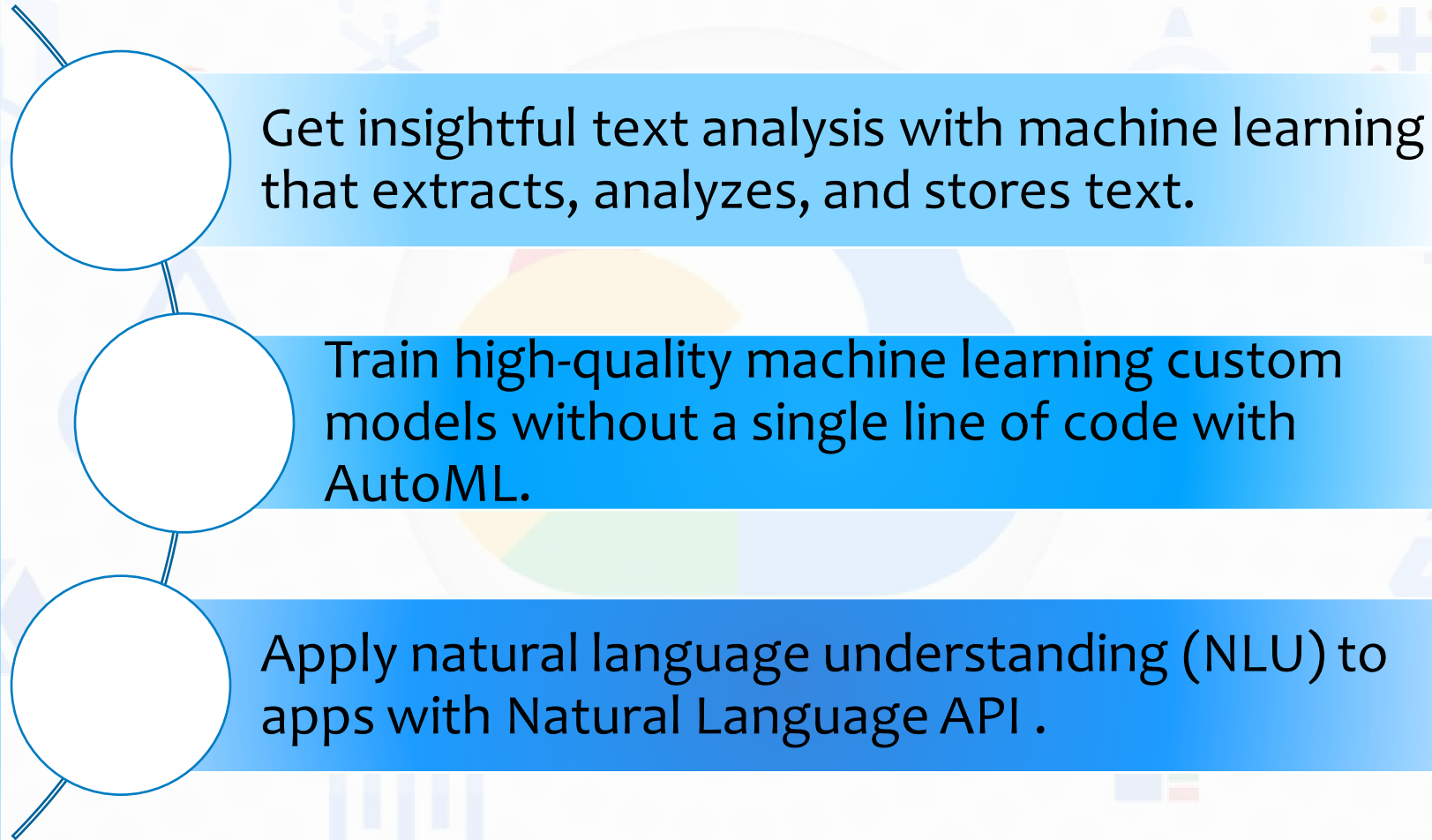
Data Pipeline Endpoint



OBSCURER

Infosys  
cobalt

# Google Cloud Language API



**OBSCURER**

Infosys  
cobalt



# Google Cloud Document AI



Helps convert documents and images to human readable text.

Easy integration with other Google cloud services.

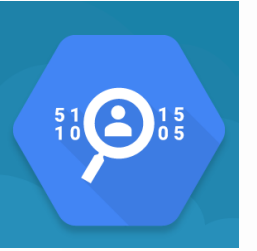
Manage the entire unstructured document lifecycle in one unified solution.

Use your document data to gain new insights about your products and meet customer expectations.

**OBSCURER**

Infosys  
cobalt

# Google Cloud DLP



**OBSCURER**

# Big Query



**OBSCURER**

# Looker Studio



Looker Studio



**OBSCURER**

Infosys  
cobalt



# Thank you

All images used are subject to respective owners' copyright

© 2023 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.