MINI PROJECT REPORT

An internship report submitted in partial fulfilment of the requirements for the course "Minor Project-II (ECD 3991)" of degree

Bachelor of Technology

(Electronics and Communication Engineering)

(As a part of VI Semester Course)

by

AMAN RAJ 20BEC004

SARTHAK SHARMA 20BEC069

JATIN MANHOTRA 20BEC030

ANKIT KUMAR SINGH 20BEC010

SCHOOL OF ELECTRONICS AND COMMUNICATION ENGINEERING

FACULTY OF ENGINEERING

SHRI MATA VAISHNO DEVI UNIVERSITY

KATRA-182320 (J&K), INDIA 2023

DECLARATION

We, **AMAN RAJ , SARTHAK SHARMA, JATIN MANHOTRA, ANKIT KUMAR SINGH, 20BEC004, 20BEC069, 20BEC030, 20BEC010** hereby declare that the work, entitled "**Document AI**" which is being presented in the report by me in partial fulfilment of the requirements for the course "**Minor Project-II (ECD 3991)**" in Semester-VI, 2022-23 of degree **Bachelor of Technology in Electronics And Communication Engineering** from the School of Electronics Engineering, Faculty of Engineering, Shri Mata Vaishno Devi University, Katra is a bonafide record of our own work, carried out under the supervision of **Dr. Sachin Kumar Gupta** , Assistant Professor, School of Electronics & Communication Engineering, Shri Mata Vaishno Devi University, Katra.. The content presented in this study has not been submitted to any other university/institute for the award of any other degree.

Date: __ May ,2023

AMAN RAJ **20BEC004**

SARTHAK SHARMA **20BEC069**

JATIN MANHOTRA **20BEC030**

ANKIT KUMAR SINGH **20BEC010**

ACKNOWLEDGMENT

At the very outset, I am very much thankful to almighty God for giving me strength, courage and ability to accomplish the mini project as well as the report in a scheduled time in spite of various complications.

It gives me immense pleasure to thank a large number of individuals for their cordial cooperation and encouragement which has contributed directly or indirectly in preparing this report. First of all, I would like to express my gratitude to my supervisor **Sachin Kumar Gupta, Assistant Professor, School of Electronics & Communication Engineering, Shri Mata Vaishno Devi University, Katra** for his guidance and feedback which made everything clear to me to complete this project and report.

CERTIFICATE

This is to certify that the project entitled "**DOCUMENT AI**" being submitted by "**AMAN RAJ 20BEC004 , SARTHAK SHARMA 20BEC069, JATIN MANHOTRA 20BEC030, ANKIT KUMAR SINGH 20BEC010**" to School of Electronics & Communication of "Shri Mata Vaishno Devi University, Katra" for the award of the degree of "Bachelors of Technology" in "Electronics & Communication Engineering", is a bonafide research work carried out by them under my supervision and guidance. Their project report has reached the standard of fulfilling the requirements of regulations relating to degree. The report is an original piece of research work and embodies the findings made by the research scholar himself.

The results presented have not been submitted in part or in full to any other University/Institute for the award of any degree or diploma

Supervisor

Assistant Professor, School of Electronics & Communication Engineering

Shri Mata Vaishno Devi University, Katra -182320, Jammu and Kashmir (India)

Table of Contents

ABSTRACT

This report stresses on the work experience I have gathered as working in this project along with my team mates. We have explained all about Unstructured data and how to store it as most of the business insights are derived from unstructured data. We have use Google cloud platform to demonstrate our project with the screenshots provided for more clarification along with the links.

Chapter 1 Problem Statement Overview

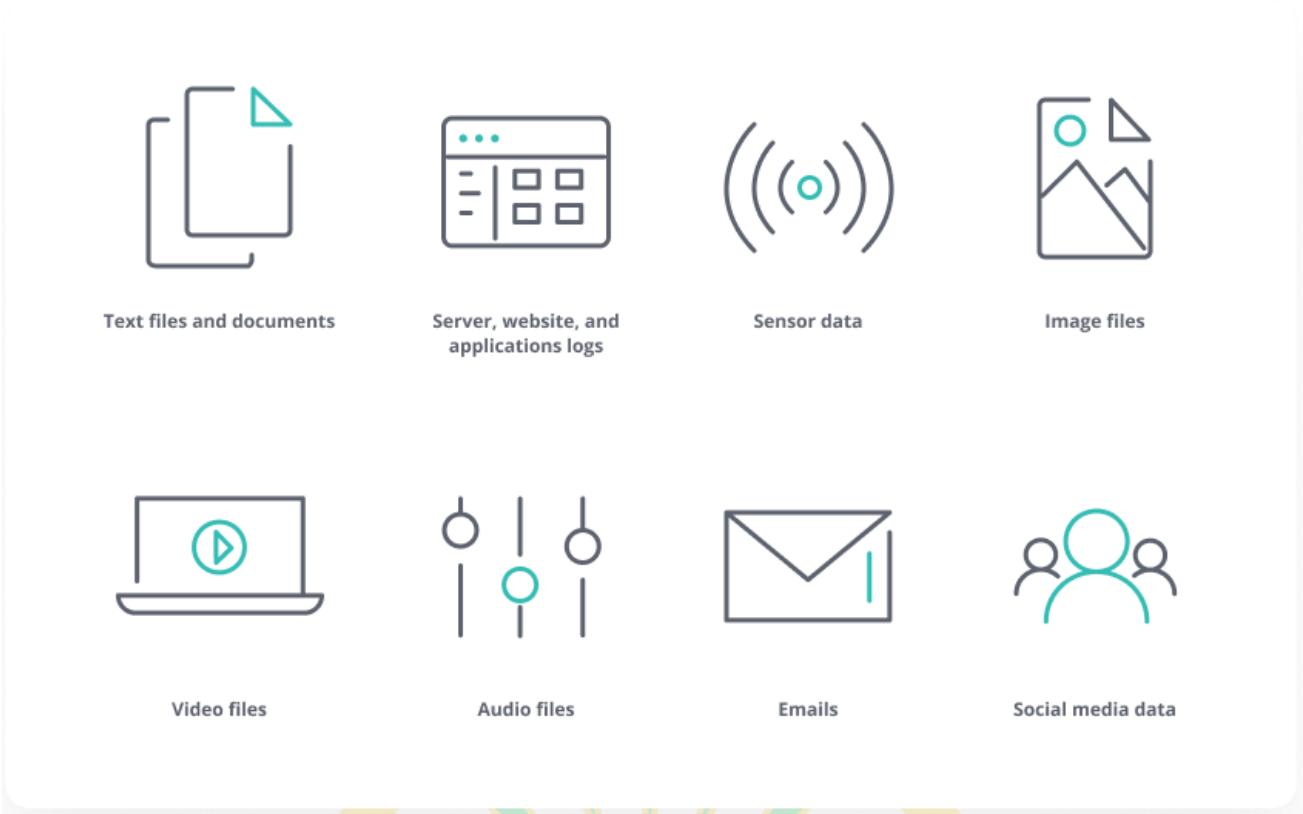
INTRODUCTION

Most businesses are now sitting on Document goldmines. These documents are contracts, PDFs, emails, customer feedback, patterns. These documents are increasing over time.

Following are some example of the business which has millions of contract documents. They need to read these documents at a different times of their lifecycle for analyzing it. This needs a lot of processing time and error prone.

- Mortgage Providers
- Insurance Companies
- School

These are unstructured data.



We are all very aware that computers are now an integral part of our lives. Today, technology has advanced so much that a computer can perform tasks just like humans and even have a high rate of success in doing so. This is all possible due to **Artificial Intelligence**.

Artificial Intelligence (AI) is the field of technology that enables the machines to automatically perform tasks that would otherwise require human intelligence. AI is a huge spectrum in the field of Computer Science and is developed and programmed through machine learning and deep learning.

AI is used in many fields for day-to-day application, making our lives easier than before. The world of Business is one such field where Artificial Intelligence is widely used. AI can help any business in three fields: automating business processes, gaining insight through data analysis, and engaging with customers and employees.

There is a **huge competition** between different companies in the market and every company wants to be at the top of their game. Successful multinational companies use features of AI like Automation, Big Data Analytics and Natural Language Processing to get insights of their business and make it more efficient and relevant to their customer base. Even smaller companies incorporate AI into their businesses to be successful.

Unstructured Data

Unstructured data is the data which **does not conforms to a data model** and has no easily identifiable structure such that it can not be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

Unstructured data is information that is not **arranged according to a pre-set data model or schema**, and therefore cannot be **stored in a traditional relational database or RDBMS**. Text and multimedia are two common types of unstructured content. Many business documents are unstructured, as are email messages, videos, photos, webpages, and audio files.

From **80 to 90 percent of data generated** and **collected** by **organizations, is unstructured**,, and its volumes are **growing rapidly** — many times faster than the rate of growth for structured databases.

Unstructured data stores contain a wealth of information that can be used to guide business decisions. However, unstructured data has historically been very difficult to analyze. With the **help of AI** and machine learning, new software tools are emerging that can **search through vast quantities** of it to **uncover beneficial and actionable business intelligence**.

Characteristics of Unstructured Data:

- Data neither conforms to a data model nor has any structure.
- Data can not be stored in the form of rows and columns as in Databases
- Data does not follows any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure
- Due to lack of identifiable structure, it can not used by computer programs easily

Unstructured Data vs. Structured Data

Let's take **structured data** first: It's usually **stored in a relational database** or RDBMS, and is sometimes referred to as relational data. It can be **easily mapped into designated fields** — for example, for zip codes, phone numbers, and credit cards, respectively. Data that conforms to **RDBMS structure is easy to search, both with human-defined queries and with software**.

Unstructured data, in contrast, **doesn't fit into these sorts of pre-defined data models**. It **can't be stored in an RDBMS.** And because it comes in so **many formats, it's a real challenge** for conventional software **to ingest, process, and analyze**. Simple content searches can be undertaken across textual unstructured data with the right tools .

Beyond that, the **lack of consistent internal structure** doesn't conform to **what typical data mining systems can work with**. As a result, **companies have largely been unable to tap into value-laden data like customer interactions, rich media, and social network conversations**. Robust tools for doing so are only now being developed and commercialized.

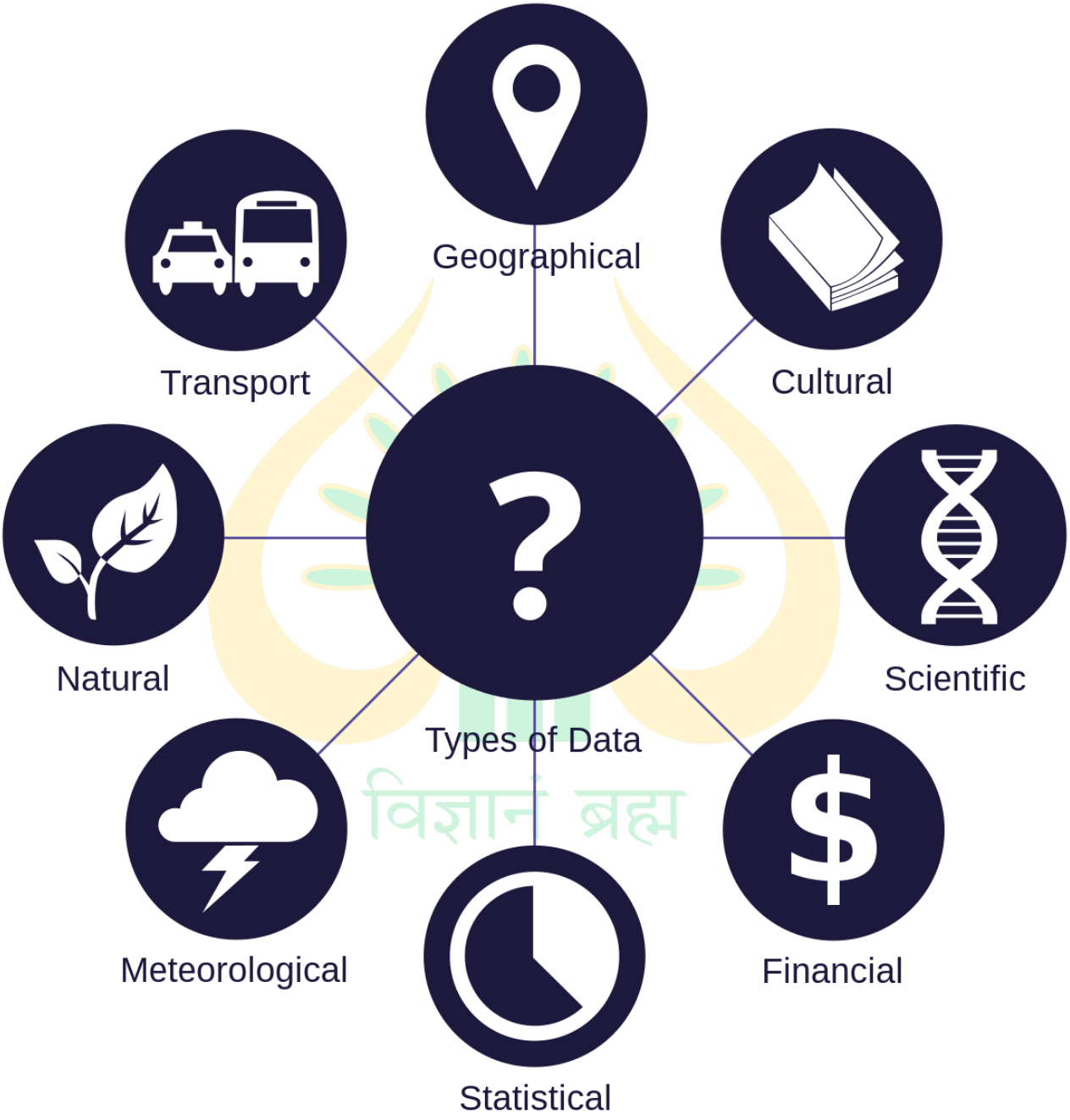
What Are Some Examples Of Unstructured Data?

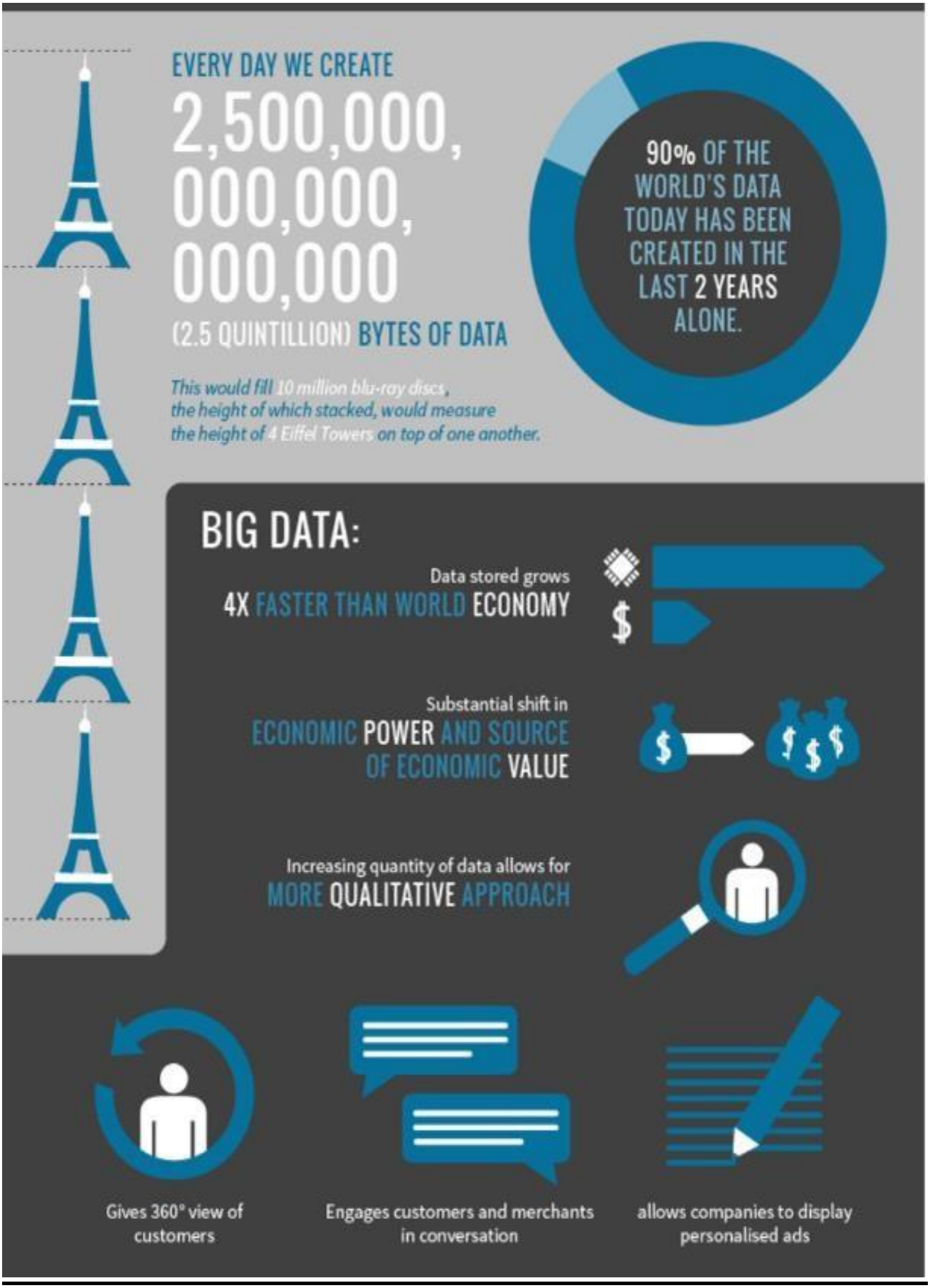
Unstructured data can be created by people or generated by machines.

Here are some examples of the human-generated variety:

- **Email:** Email message fields are unstructured and cannot be parsed by traditional analytics tools. That said, email metadata affords it some structure, and explains why email is sometimes considered semi-structured data.
- **Text files:** This category includes word processing documents, spreadsheets, presentations, email, and log files.
- **Social media and websites:** Data from social networks like Twitter, LinkedIn, and Facebook, and websites such as Instagram, photo-sharing sites, and YouTube.

- **Mobile and communications data:** Text messages, phone recordings, collaboration software, Chat, and Instant Messaging.
- **Media:** Digital photos, audio, and video files.




EVERY DAY WE CREATE
2,500,000,
000,000,
000,000
(2.5 QUINTILLION) BYTES OF DATA
This would fill 10 million blu-ray discs,
the height of which stacked, would measure
the height of 4 Eiffel Towers on top of one another.
90% OF THE WORLD'S DATA TODAY HAS BEEN CREATED IN THE LAST 2 YEARS ALONE.
BIG DATA:
Data stored grows
4X FASTER THAN WORLD ECONOMY
Substantial shift in
ECONOMIC POWER AND SOURCE OF ECONOMIC VALUE
Increasing quantity of data allows for
MORE QUALITATIVE APPROACH
Gives 360° view of customers
Engages customers and merchants in conversation
allows companies to display personalised ads

What is Unstructured Data Used For?

Simple content searches can be performed on textual unstructured data. Traditional analytics tools are optimized for highly structured relational data, so they're of little use for unstructured sources such as rich media, customer interactions, and social media data.

Big Data and unstructured data often go together: **IDC {**International Data Corporation**}** estimates **that 90% of these extremely large datasets are unstructured**. New tools have recently become available to analyze these and other unstructured sources. **Powered by AI and machine learning,** such platforms function at near real-time speed and educate themselves based on the patterns and insights they uncover. These **systems are being employed against large unstructured datasets to enable never-before-possible applications** like:

- Analyzing communications for regulatory compliance
- Tracking and analyzing customer social media conversations and interactions
- Gaining reliable insights into widespread customer behavior and preferences

Chapter 2
Document AI

Document AI overview

Document AI is a document understanding solution that takes unstructured data (documents, forms, etc.) and makes the data easier to understand, analyze, and consume by providing structure through content classification, entity extraction, advanced searching, and more.

Document AI or **Document Intelligence** is a technology that uses natural language processing (NLP) and machine learning (ML) to train computers to simulate a human review of documents

NLP enables the computer to 'understand' the contents of documents, including the contextual nuances of the language within them, before extracting the information and insights contained in the documents. The technology can then categorize and organize the documents themselves.

Document AI is used to process and parse forms, tables, receipts, invoices, tax forms, contracts, loan agreements, financial reports, etc.

Document AI uses machine learning and Google Cloud to help you create a scalable, cloud-based document understanding solution.

Using Document AI, you can:

- Convert images to text
- Classify documents
- Analyze and extract entities

Document AI, in beta, offers a scalable, serverless platform to automatically classify, extract, and enrich data from your scanned documents. It converts unstructured data into structured data.

Internally it uses the same deep machine learning technology that powers Google Search, Google Assistant, Natural Language Processing API to derive valuable insights from your unstructured documents.

Key Features

Document AI utilizes machine learning to extract information from documents in digital and print forms. Document AI is able to accurately identify text, characters, and images in different languages, thus enabling users to gain insights from the unstructured documents. Using the data from the documents allows Document AI users to make better and faster decisions regarding the documents. The technology makes the process of analyzing documents more efficient by automating and validating the data for the workflows.

Common Uses

- Freeing up employees for higher-value tasks.
- Using AI to check for anomalies in new invoices from old customers.
- Spotting fake currency and fraudulent checks.
- Fast-tracking the mortgage workflow process.
- Automating the monitoring of loan portfolios to manage credit risks.
- Enabling firms to automate the impact assessment of regulatory changes on their contracts.
- Analyzing previously inaccessible data siloed in documents to make informed business decisions.
- Streamlining the consumption of receipts on a worldwide scale.
- Increasing the reliability of business information by decreasing errors resulting from manual data entry.

Chapter 3 The Project

USE CASE

Consider organisation like **SHRI MATA VAISHNO DEVI UNIVERSITY** which has multiple field offices and the head office runs the HR / Payroll system. New joiner's details are filled in a form by field managers and then forms are being sent to head office. The operator enters the details manually to the system and then Employee's email, training, access card, laptop, and other formalities get sorted.
This end-2-end process takes days and till then employee sits idle. In my opinion new joiners (be it fresher or experienced) have always eager to demonstrate his/her talent or skill so they should not idle in their early days.

In this senerio, we will see how to automate document processing and end-to-end new joiner process:

- Field managers fill the form and upload the scanned form
- The application extracts the data from document and store into database
- Then alerts to other processes, new joiners employee_id
- Other processes like ID Cards, Laptop, Desk, etc will fetch the details from the database using employee_id

COMPONENTS

The following **Serverless** components are used in this architecture.
This means that you will **pay per use**, without any up-front costs. Also, no servers need to be configured or maintained.

1. Front-end app to upload the scanned document on **Cloud Run**.
2. The document is stored in **Google Cloud Storage**.
3. This triggers the **Cloud Function**.
4. The Cloud Function calls **Document AI** to fetch the entities.

5. The Cloud Function reads the response generates employee_id, email, and stores data to **Cloud Firestore.**
6. The new-joiner notification is sent to **Cloud Pub/Sub** topic.
7. This topic has multiple subscribers: Desk Service, Laptop Service, ID Card Service. These services will fetch the details from **Firestore**.
8. The Service deployed using **Cloud Run** which has end-point to GET the details of an employee from Cloud Firestore.
9. All components are logging data to **Stackdriver**.

Cloud Run: Google Cloud Run is a fully managed platform that takes a Docker container image and runs it as a stateless, autoscaling HTTP service.

Google Cloud Storage: Cloud Storage is a service for storing your *objects* in Google Cloud. An object is an immutable piece of data consisting of a file of any format. You store objects in containers called *buckets*. All buckets are associated with a *project*, and you can group your projects under an *organization*.

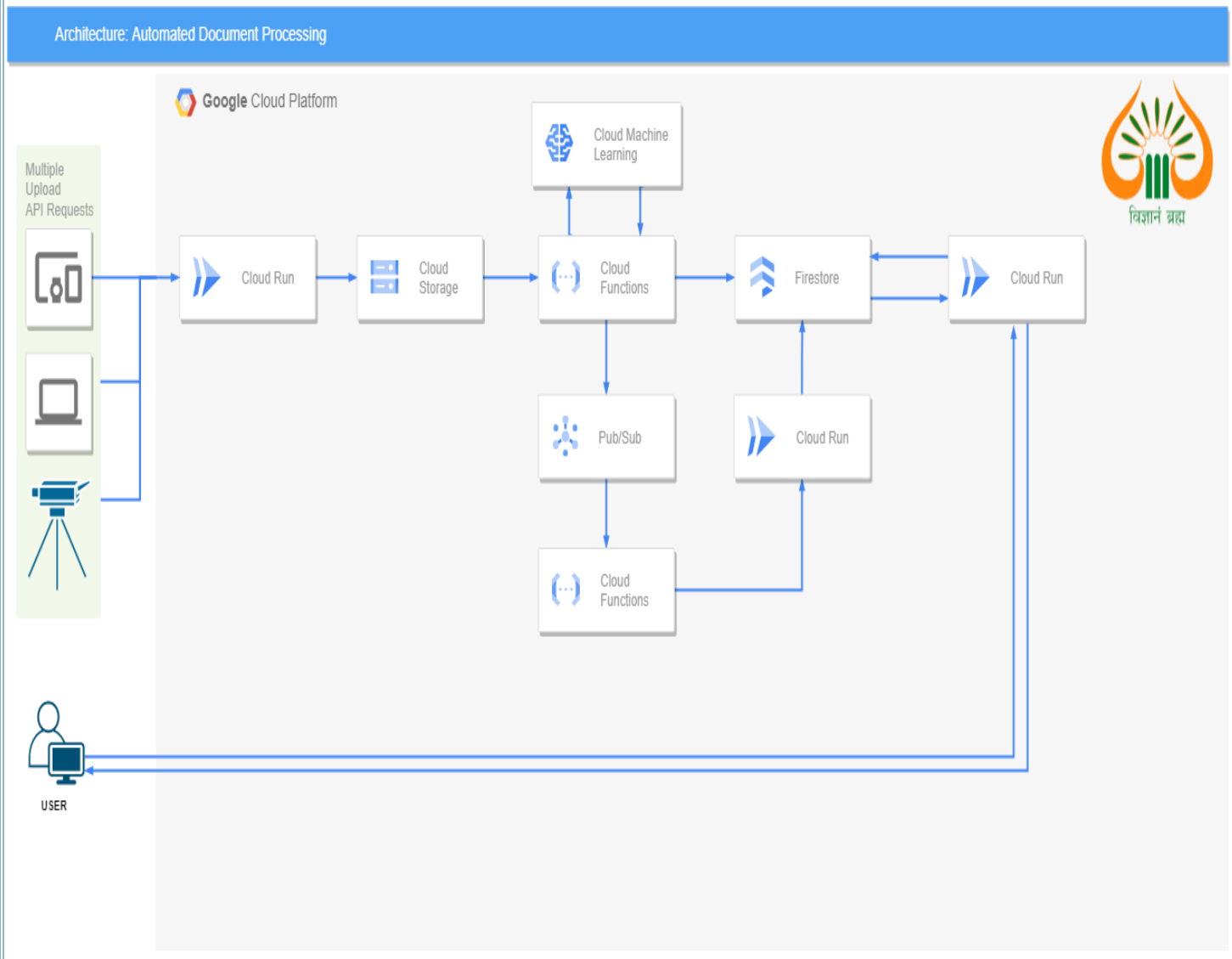
Cloud Function: Google Cloud Functions is a serverless execution environment for building and connecting cloud services. With Cloud Functions you write simple, single-purpose functions that are attached to events emitted from your cloud infrastructure and services.

Cloud Pub/Sub: Pub/Sub allows services to communicate asynchronously, with latencies on the order of 100 milliseconds. Pub/Sub is used for **streaming analytics and data integration pipelines to ingest and distribute data**.

Stackdriver.: Google Stackdriver was a **monitoring service that provided IT teams with performance data about applications and virtual machines** (VMs) running on the Google Cloud Platform (GCP)

Firestore: Cloud Firestore is **a NoSQL document database** that lets you easily store, sync, and query data for your mobile and web apps - at global scale.

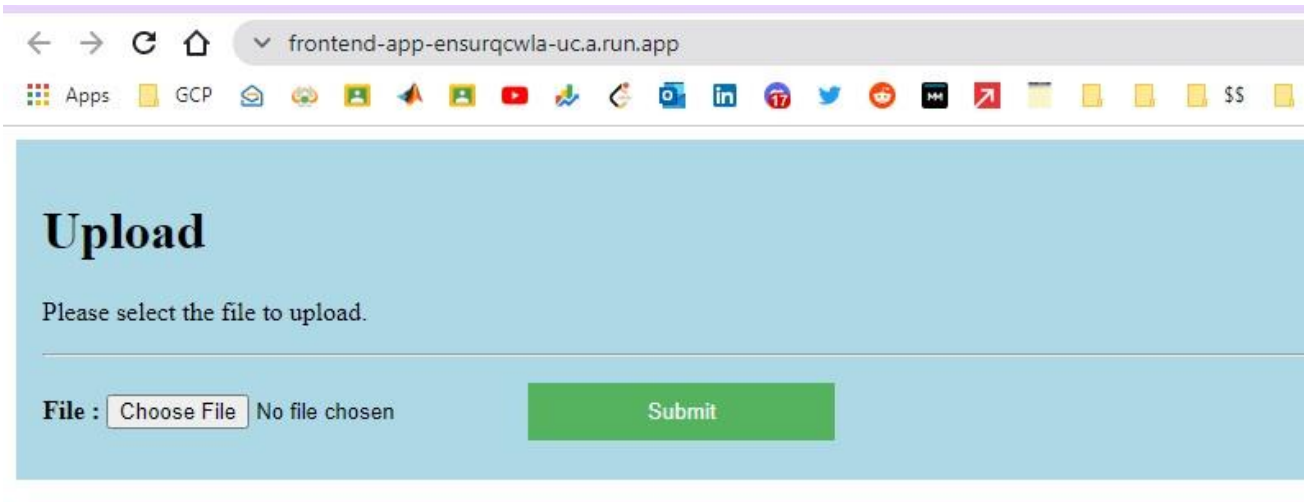
Architecture Diagram



Project Link: https://github.com/yugstar/document_ai_gcp/tree/main/document-ai-in-gcp-master-main/document-ai-in-gcp-master-master

Access the application

1. Upload the sample file on the sample frontend app. The example (sample.pdf) is available in the repository.



2. The entities are extracted using Document AI and add it into Firestore.

3. The Subscribers (e.g. ID_Cards Service, Desk Service, Laptop Service) will get notification of new joiner and it fetches the details using the API.

Conclusion

This way we have automated end-to-end Document processing in **Google Cloud Platform.**

<u>Aim is to reduce all these senerios that we face in our daily life</u>