[***Cloud Project 1 - Script***](https://drive.google.com/drive/folders/1GyhWPI6XcrPund22PEgERXqdMVCwB_TI?usp=sharing)

|  |  |
| --- | --- |
| **Role** | **Name** |
| ***Presenter 1*** | **Aadil** |
| ***Presenter 2*** | **Aliyyah** |
| ***Presenter 3*** | **Jonathan** |
| ***Presenter 4*** | **Masood** |

**Segment 1: Introduction & Problem Statement (Presenter 1 – 3 Minutes)**

**[0:00 – 0:20] – Opening Shot**

* **Presenter 1 Dialogue:**  
  **“Welcome, everyone. Today we’re happy to present our demo on creating a PDF Reading Tool for a shipping company. We are going to kick things off by explaining the challenge and our overall approach.”**
* ***# Showing the Team Members***

**[0:20 – 1:30] – Explaining the Challenge**

* **Presenter 2 Dialogue:**  
  **“The shipping company receives a large volume of PDFs containing shipment details, many of which are scanned documents. Manual extraction not only consumes valuable time but also risks costly errors. Our task was to develop an automated solution that reliably extracts shipment data for database creation, reporting, dashboards, and financial analysis.”**

**[1:30 – 3:00] – Overview of the Proposed Solution**

* **Presenter 3 Dialogue:**  
  **“To tackle these challenges, we chose Google Cloud Platform’s Document AI, part of the Vertex AI suite. This tool not only reads and extracts data from PDFs but also supports custom model training to handle our specific document format. By integrating with Google Cloud Storage for data management and using robust OCR and labeling techniques, we streamline the extraction process, reduce errors, and save time.”**
* **Presenter 4 Dialogue:**  
  **“With that foundation, let’s move to our next segment where we compare cloud platforms and justify our solution choice.”**

**Segment 2: Cloud Platform Comparison & Justification (Presenter 2 – 3 Minutes)**

**[3:00 – 3:20] – Introduction**

* **Presenter 4 Dialogue:**  
  **“In this segment, we’ll explore how leading cloud platforms stack up against each other for document processing.”**

**[3:20 – 4:40] – Detailed Comparison**

* **Presenter 4 Dialogue:**  
  **“AWS Textract offers strong out-of-the-box extraction but is less customizable for niche document formats. Azure Form Recognizer provides excellent layout detection with configurable templates. However, Google Cloud’s Document AI stands out because it allows custom model training tailored to our specific PDF formats, making it the best fit for our needs.”**
* **Presenter 1 Dialogue:**  
  **“Additionally, GCP Document AI integrates seamlessly with other Google services such as Cloud Storage and Vertex AI, offering not only cost efficiency through a pay-as-you-go model but also the scalability to handle increasing document volumes.”**

**[4:40 – 6:00] – Justification & Transition**

* **Presenter 1 Dialogue:**  
  **“This comparative analysis clearly shows that our choice of GCP Document AI is driven by its superior customization, integration capabilities, and competitive pricing model. Next, we’ll dive into our pipeline and demonstrate the live process of data extraction.”**

**Segment 3: Pipeline/Workflow and Live Process Demonstration (Presenter 3 – 5 Minutes)**

**[6:00 – 6:20] – Introduction**

* **Presenter 1 Dialogue:**  
  **“When talking about our data pipeline—from data collection to model deployment—and show you live screenshots of the process.”**

**[6:20 – 7:40] – Pipeline Overview**

* **Presenter 1 Dialogue:**

**“Our workflow begins with collecting the PDFs into Google Cloud Storage. We first Split the Documents into Training, Testing & Validation with minimum 5 Documents per Section. After, we will select our model from GCP’s Documentation Custom Processors. For this scenario, we have selected ‘Custom Extractor’ as our model.”**

* **Presenter 2 Dialogue:**

**“We need to Create Labels in the Format of the Data we want to Collect, As you can see in this Document there is Heading, Body & Summary. So we create 3 Sections of Nested Labels and Under Each we Write the Category/Feature Names which we want to Collect, Classify & Store for Further Processing.”**

* **Presenter 3 Dialogue:**

**”””**

**Next we Upload the PDF’s from the Bucket to Document AI Proccessors.**

**Now we Need to Label the Data for ALL THE PDF we do so by selecting the Feild Name & Creating a Box around it to Show where the Data Generally is Located & what to Expect from it.**

**”””**

**[7:40 – 9:00] – Live Process Demonstration**

* **Presenter 4 Dialogue:**  
  **“For validation and testing, we use the validation PDFs to confirm the accuracy of model we have trained.”**
* **Presenter 2 Dialogue:**  
  **“After the model is deployed, we can do API integrations into custom made software for users to extract and utilize the data. In this scenario, we can expect the finance team to use the extracted data to create reports and validate existing reports which allows them to understand the company’s profits and losses.”**
* **Presenter 3 Dialogue:**  
  **“As the reports are coming from external sources, the company will need to monitor for cost using GCP’s custom processor, any variations of the PDFs, verifying the accuracy of the model to ensure it is working as intended.”**
* **Presenter 3 Dialogue:**  
  **“““**

**Since we have not built the Model Fully yet this Demo is a Pre-Built Demo on GCP’s Documentation Page on how we can Expect the Output.**

**As you observed, the model accurately locates and extracts the data. This not only speeds up processing but also minimizes human errors.”**

**This live demo is a key part of our solution, proving that the process is both efficient and scalable. Now, let’s discuss the economic impact, security considerations, and future improvements.**

**”””**

* ***# Continue by Showing the Team’s Call ONLY.***

**Segment 4: Economic Impact, Security, Challenges & Future Scalability (Presenter 4 – 4 Minutes)**

**[9:00 – 9:20] – Introduction**

* **Presenter 2 Dialogue:**  
  **“Now we’ll be covering how our solution impacts the business economically, ensures data security, and what our future scalability plans look like.”**

**[9:20 – 10:40] – Economic & Business Impact Analysis**

* **Presenter 2 Dialogue:**  
  **“Our automated solution dramatically reduces processing time. For instance, reducing a 10-minute manual task per PDF across 1,000 PDFs translates to roughly 167 hours saved monthly. With reduced labor costs and fewer errors.”**
* **Presenter 1 Dialogue:**  
  **“Cost-wise, GCP’s pay-as-you-go model ensures that expenses scale with usage. Storage costs are competitive and API call fees remain minimal.”**

**[10:40 – 11:40] – Security & Compliance Considerations**

* **Presenter 1 Dialogue:**  
  **“Security is Very Important & GCP provides robust encryption for all data and adheres to major industry standards including GDPR, SOC 2, and ISO 27001, which is critical for handling sensitive shipment data.”**

**[11:40 – 13:00] – Challenges, Mitigation, and Future Improvements**

* **Presenter 3 Dialogue:**  
  **“Certainly, every solution has its own set of challenges. To tackle problems like poor OCR quality in scanned PDFs, we’ve implemented preprocessing techniques such as noise reduction and de-skewing. Moreover, our custom model is regularly refined using human feedback, allowing it to adapt effectively to the diverse formats of documents.”**
* **Presenter 2 Dialogue:**  
  **“Looking ahead, our modular architecture will allow us to easily integrate other document types and incorporate real-time processing using cloud functions, ensuring the solution grows with the company’s needs.”**

**[13:00 – 14:00] – Final Summary**

* **Presenter 2 Dialogue:**  
  **“In summary, our solution using GCP Document AI offers high accuracy, seamless integration, significant cost and time savings, and robust security. With clear future improvements planned, we’re confident this tool will continue to provide value as document volumes grow.”**

**[14:00 – 15:00] – Wrap-Up & Call-to-Action**

* **Presenter 2 Dialogue:**  
  **“Thank you for joining our presentation.”**