# AIG 130 – Project 1 – Group 5

## Group Members:

* Masoud Masoori
* Aliyyah Jackhan
* Mohammed Aadil Suhail Shaikh
* Jonathan Chacko Pattasseril

## Introduction:

* We were tasked with finding a Solution for a Shipping Company that would allow them to read PDFs and extract the necessary information from them.
* The PDF are a Very Specific Format and contain information about the Shipments that the Company is handling.
* The Company wants to automate the process of reading the PDFs and extracting the information from them.
* The Purposes of Extracting the Information varies from creating a Database, Reports, Dashboards, Financial Analysis, etc...
* The Company wants different Options/Solutions depending with Details like Cost, Scalability, Accessibility, Computational Power, Ease of Use, Implementation Challenges, etc...
* The Company also wants to know the Impact of all solutions in terms of Time, Efficiency, etc...
* After the POC (Proof of Concept) the Company wants to understand Future Implications and possible to Scale the Solution.

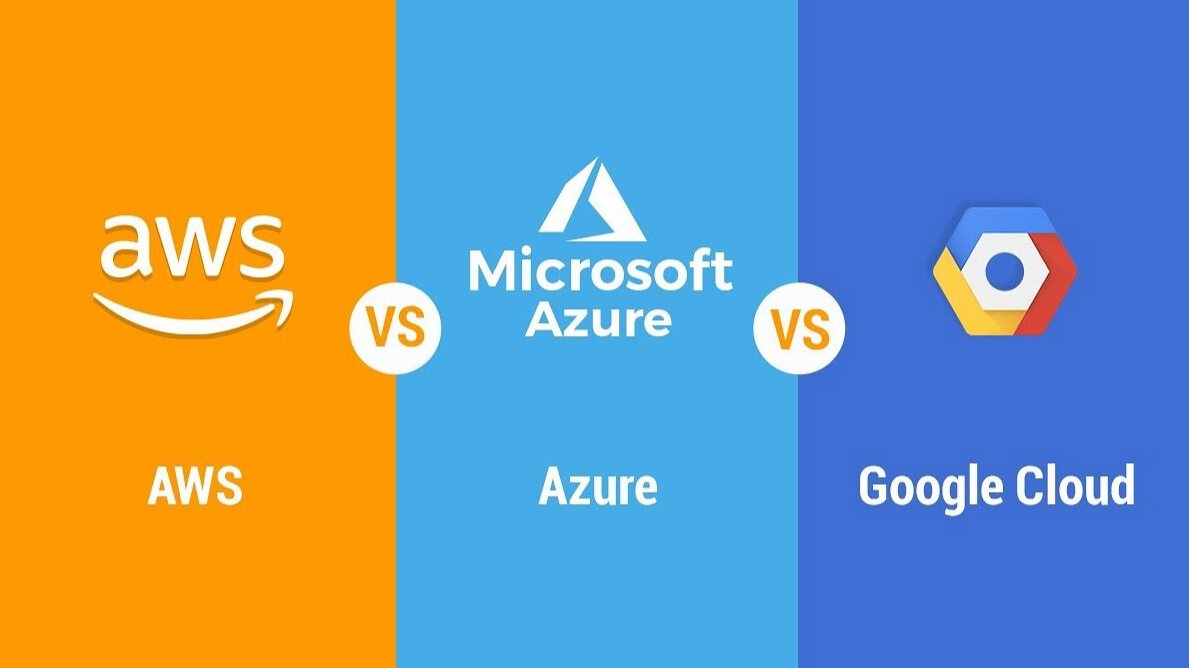
## Problem Statement:

* The Company receives a lot of PDFs with information about Shipments.
* If not done with high accuracy, the Company can lose a lot of Money and Time.
* Many of the PDFs are in a Scanned like Format & the Information cannot even be Copied and Pasted into a Spreadsheet.

## Solution we will Explore:

* We will use a Google Cloud Platform (GCP) Tool called Document AI which is part of their Vertex AI Suite.
* Document AI is a Machine Learning Tool that is able to read PDFs and extract the necessary information from them.
* Document AI in essence can Read PDFs, Classify the Data, Extract the Data, and even Translate the Data (if needed).
* We are going to Use a Custom Model to Train Document AI to Extract the Information from the PDFs.
* Since Document AI is a very powerful tool that can be used for many different purposes and can be easily integrated with other GCP Tools.
* We are going to Use Google Cloud Storage to Store the PDFs in Buckets for the Training, Testing and Validation of the Model.

## Comparison of Cloud Platforms:



While our primary focus is on ***GCP Document AI***, here’s how it compares to other leading solutions:

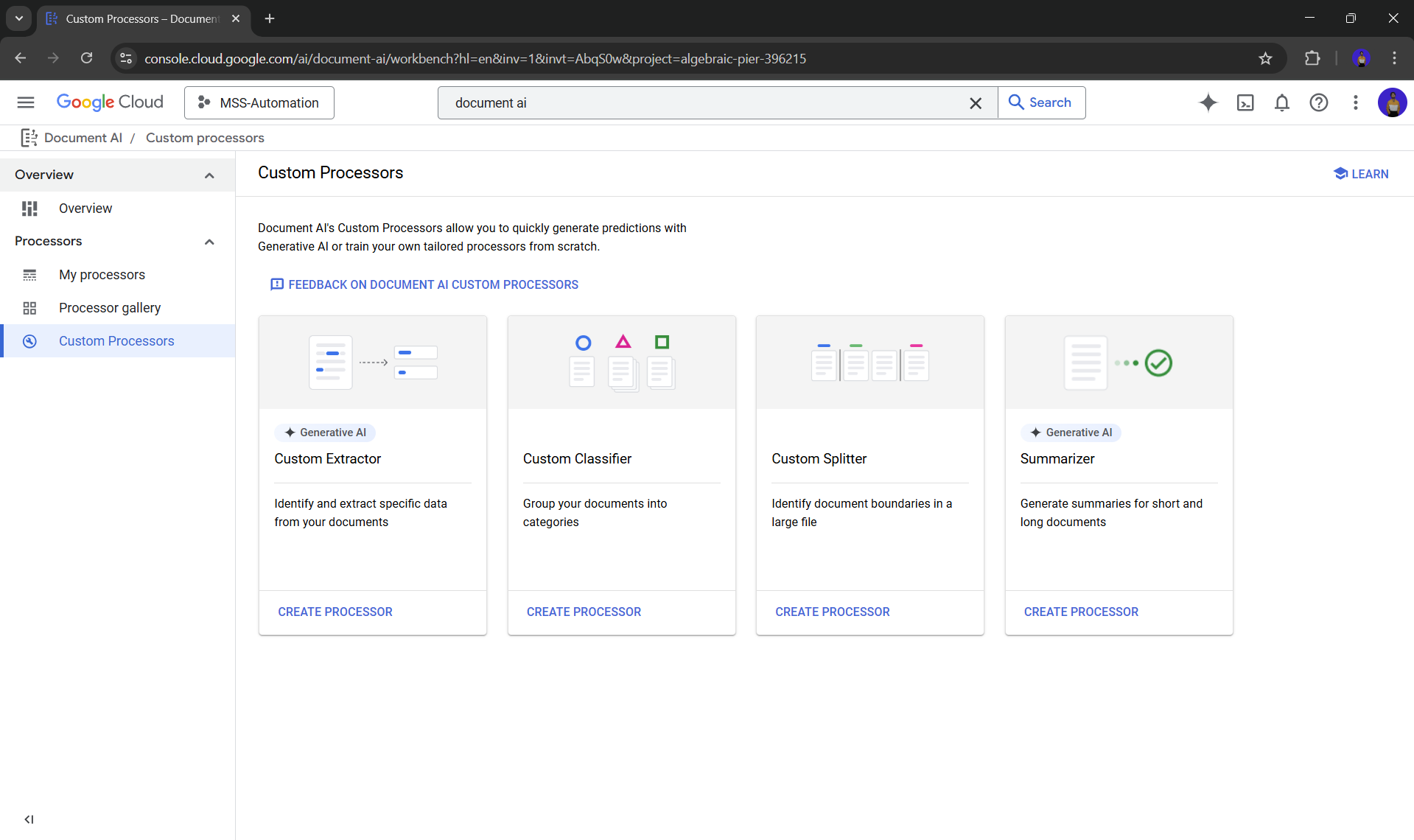
|  |  |  |  |
| --- | --- | --- | --- |
| ***Feature*** | [**AWS Textract**](https://aws.amazon.com/textract/) | [**Azure Form Recognizer**](https://azure.microsoft.com/en-us/services/form-recognizer/) | [**GCP Document AI**](https://cloud.google.com/document-ai) |
| *Logo* |  |  |  |
| *Machine Learning Capabilities* | Strong out-of-the-box extraction, but less customizable for niche document types. | Excellent layout detection with configurable templates. | Highly customizable via custom model training tailored to specific PDF formats. |
| *Integration* | |  | | --- | |  |  |  | | --- | | Deep integration within the AWS ecosystem (S3, Lambda, etc.). | | Seamless integration with Microsoft products and Azure services. | Native integration with Google Cloud Storage, Vertex AI, and other GCP services. |
| *Cost Structure* | Pay-per-use, which can be costlier with high-volume processing. | Competitive pricing with variable costs based on processing complexity. | Competitive, pay-as-you-go model, optimized when combined with other GCP services. |
| *Ease of Use* | Easy API but with limited customization options for domain-specific requirements. | Developer-friendly; some manual configuration required. | Intuitive training process using custom labeling, particularly suited for specialized formats. |
| *Scalability* | Highly scalable within AWS infrastructure. | Scalable with enterprise-grade performance in Azure. | Designed for high scalability, leveraging GCP’s robust global infrastructure. |
| *Customization* | Limited scope for custom model fine-tuning. | Customization possible but may require additional configuration. | Excellent customization for tailored document processing, leading to higher accuracy for specific formats. |

### Justification for Choosing GCP Document AI:

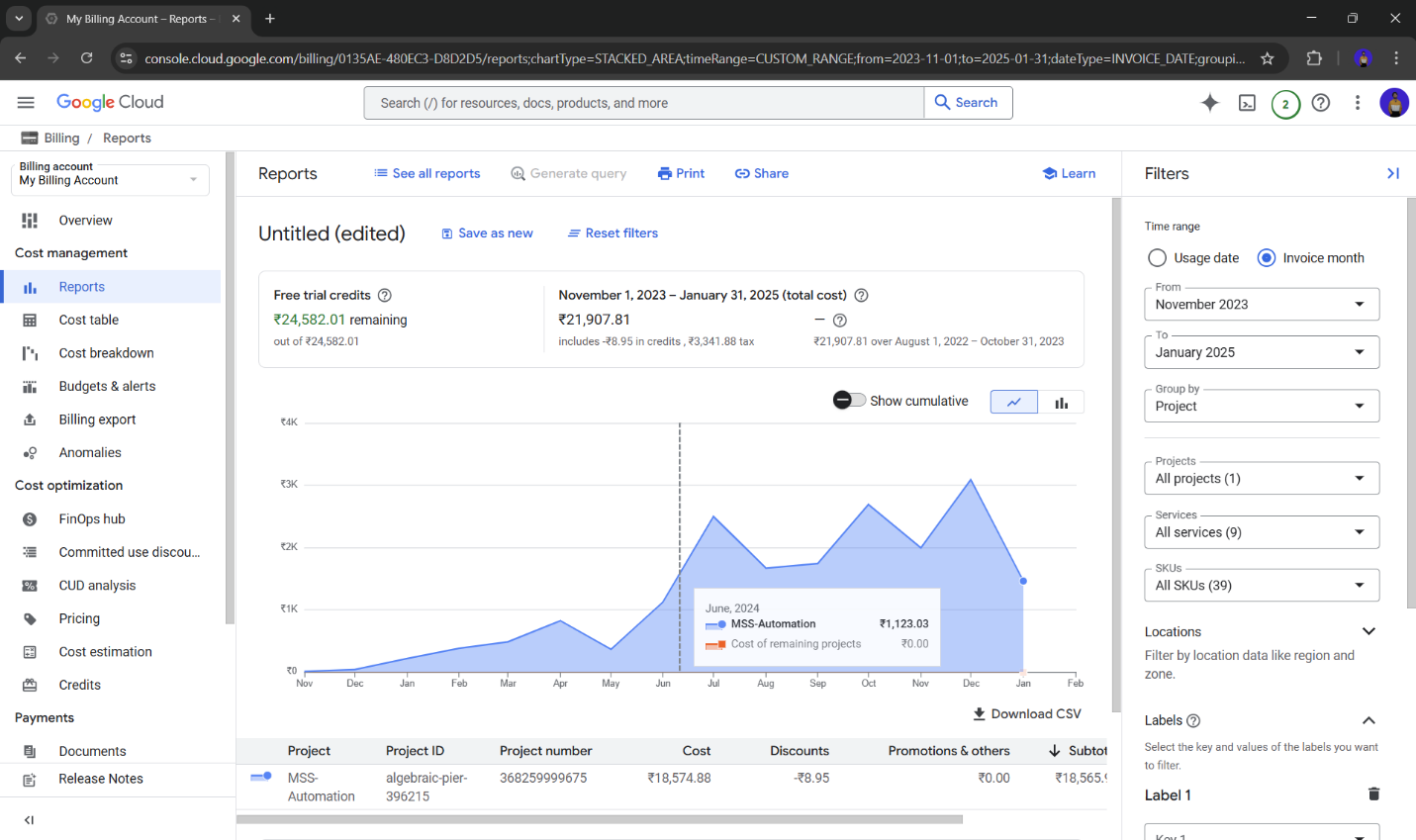
**Seamless Ecosystem Integration:** With tools like Google Cloud Storage and Vertex AI, the integration is smoother, facilitating a streamlined pipeline from data collection to deployment.



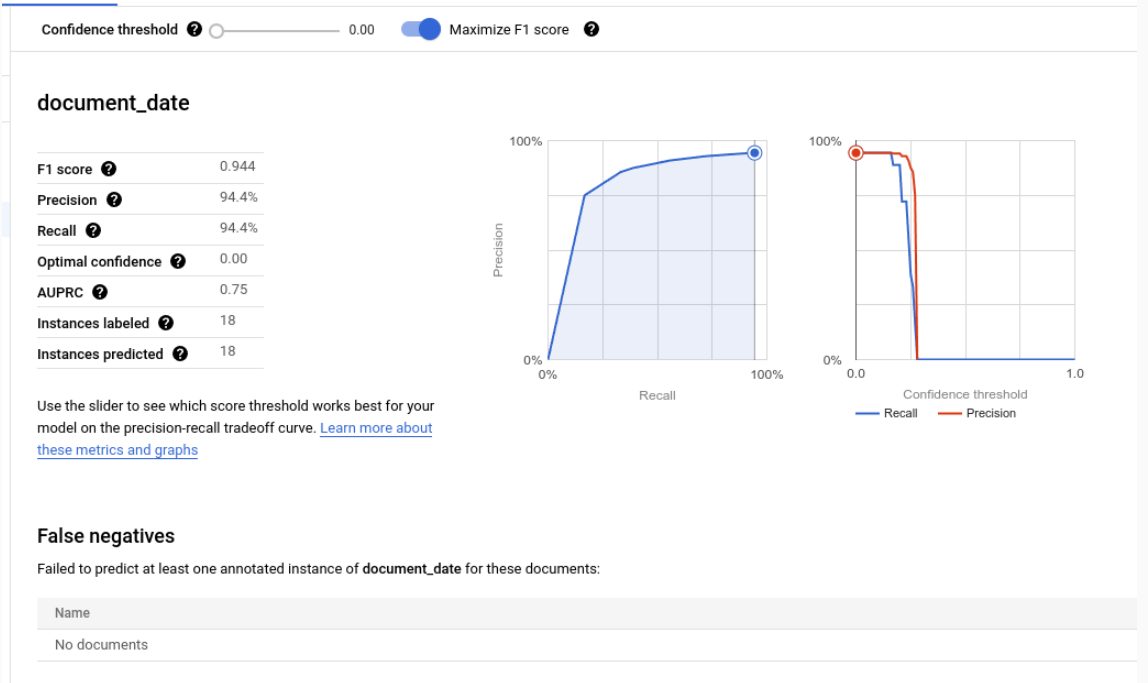
**Custom Model Training:** Our PDFs follow a very specific format. GCP’s ability to train custom models means we can precisely “teach” the system where and how to extract the relevant information.



**Cost Efficiency & Scalability:** The pay-as-you-go pricing combined with high scalability ensures that the solution remains cost-effective even as the document volume increases.



**High Accuracy:** Robust OCR and layout extraction capabilities, once fine-tuned, deliver superior accuracy and reduce errors.



### Economic & Business Impact Analysis

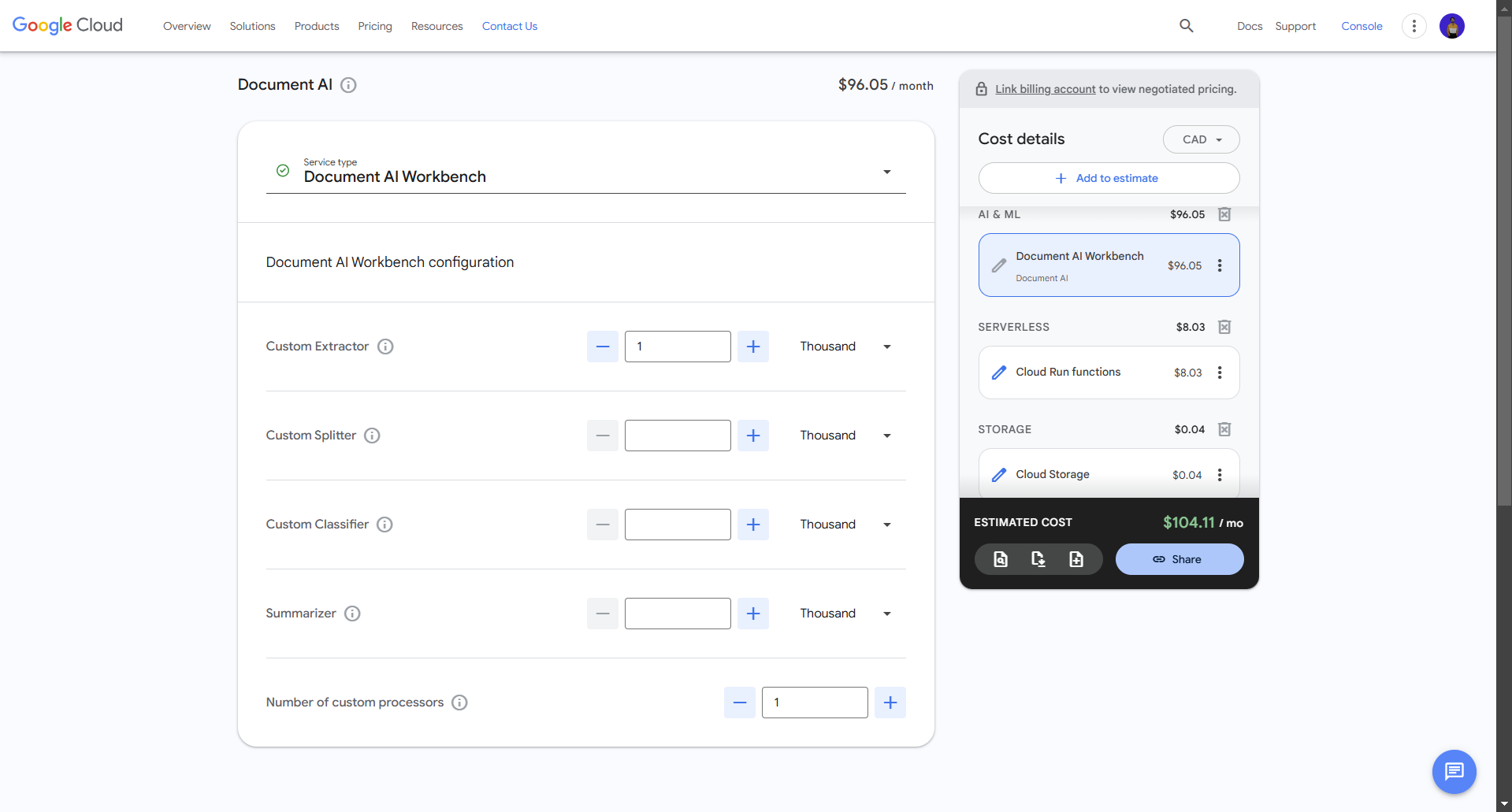
#### ROI & Time/Money Savings:

* **Time Savings:** Automating document processing can reduce manual handling time by up to 80%. For instance, if processing each PDF manually takes 10 minutes and the company handles 1,000 PDFs monthly, automation could save approximately 167 hours per month.
* **Cost Savings:** Decreasing labour costs and minimizing errors can lead to annual savings between **$50,000 and $200,000**.
* **ROI:** Depending on initial implementation expenses, the system could achieve breakeven within 6 to 12 months, with increased benefits as processing volumes grow.

#### GCP Running Costs:

* **Storage:** [Google Cloud Storage](https://cloud.google.com/storage/pricing) charges approximately **$0.020 per GB per month** for standard storage in regions like Iowa.
* **API Calls:** [Document AI](https://cloud.google.com/document-ai/pricing) pricing is based on the number of pages processed. For example, processing 1,000 pages may incur minimal costs, scaling linearly with volume.
* **Compute:** Additional compute services, such as [Vertex AI](https://cloud.google.com/vertex-ai) or [Cloud Functions](https://cloud.google.com/functions), operate on a pay-as-you-go basis.

NOTE: For detailed ROI projections and cost analyses, refer to [Google Cloud’s pricing calculator](https://cloud.google.com/products/calculator).



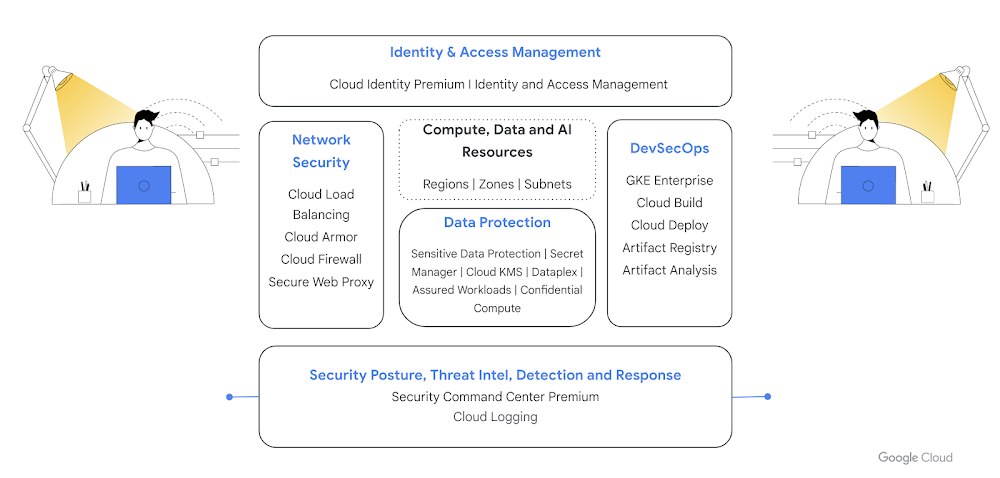
## Security & Compliance Considerations

### Data Security on GCP:

* **Encryption:** All data, both at rest and in transit, is automatically encrypted using industry-standard protocols.
* **IAM (Identity and Access Management):** GCP offers fine-grained [IAM controls](https://cloud.google.com/iam) to ensure that only authorized personnel and services can access sensitive shipment data.
* **Compliance Certifications:** GCP complies with key standards, including **GDPR, SOC 2, ISO 27001,** and **HIPAA**.

### Adherence to Industry Regulations:

* **GDPR:** GCP's data handling policies support compliance with GDPR requirements, such as data residency options, consent management, and breach notifications.
* **SOC 2:** Regular audits and certifications ensure that the system maintains robust security, availability, and confidentiality controls.



## Challenges & Mitigation Strategies

### Key Challenges:

* **Poor OCR Quality in Scanned PDFs:** Low-resolution scans and inconsistent image quality can lead to OCR inaccuracies.
* **Varying Document Formats:** Even minor format variations can disrupt the extraction process.
* **Preprocessing Needs:** Documents may require image enhancement, such as noise reduction and de-skewing, to optimize OCR performance.

### Mitigation Strategies:

* **Preprocessing Pipeline:** Implement image enhancement techniques before processing documents through the OCR engine. Tools like [OpenCV](https://opencv.org) can be utilized for these tasks.
* **Custom Model Fine-Tuning:** Employ detailed labelling to train the model, enhancing accuracy across various document formats.
* **Error Monitoring & Feedback Loop:** Establish continuous monitoring to identify errors, incorporating human validation for critical cases to provide feedback for ongoing improvements.

## Future Improvements & Scalability

### Expansion to More Document Types:

* **Modular Architecture:** Design the current pipeline to be modular, allowing integration of additional document types, such as invoices and contracts, with minimal adjustments.

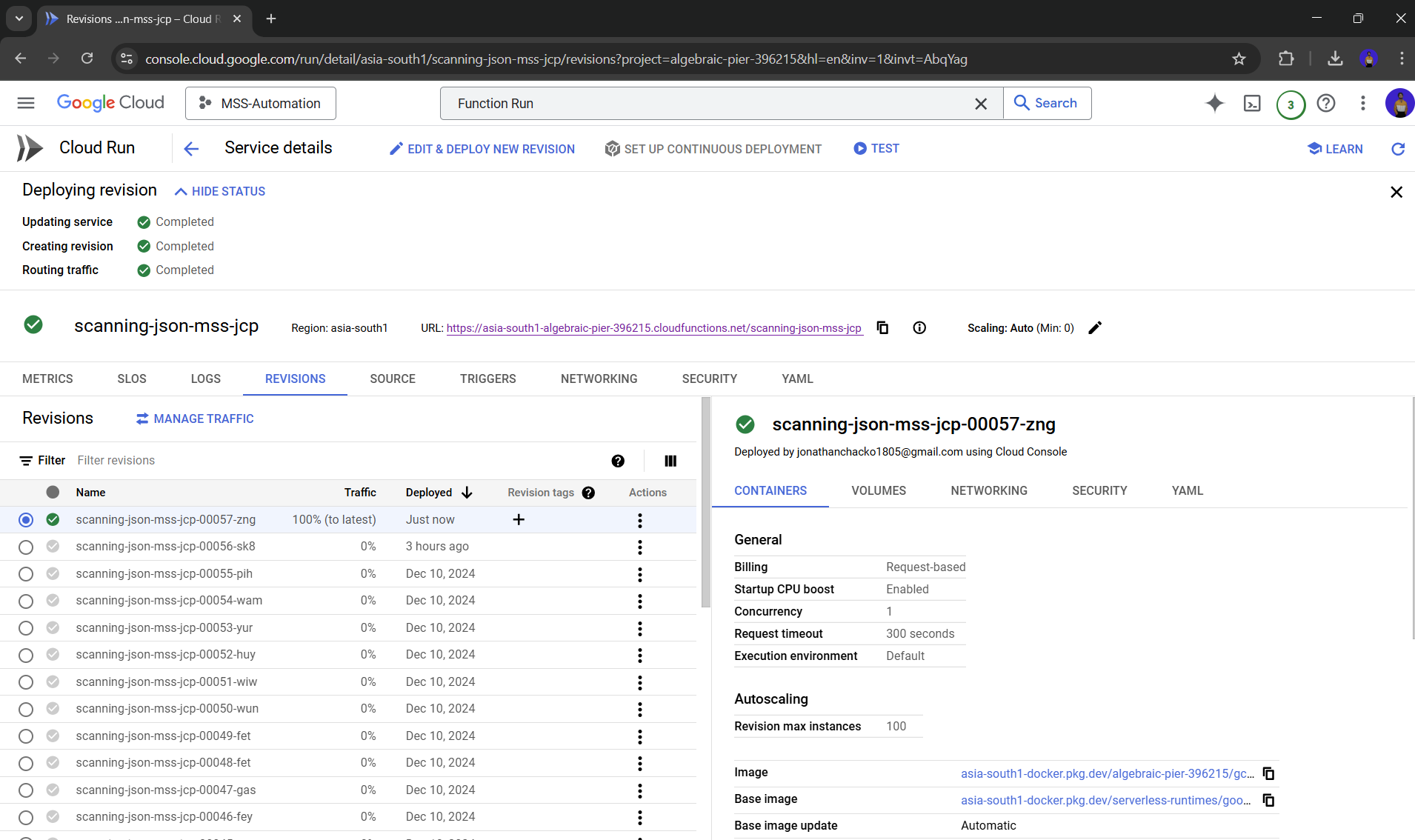
### Human Validation Integration:

* **Hybrid Approach:** Incorporate a human-in-the-loop system for critical documents to enhance accuracy and provide continuous feedback for model retraining.

### Real-Time Processing with Serverless Functions:

* **Serverless Triggers:** Utilize services like [Google Cloud Functions](https://cloud.google.com/functions) or [AWS Lambda](https://aws.amazon.com/lambda/) to automatically trigger document processing upon new file uploads, enabling near real-time data extraction.

#### Google Cloud Function Run Implementation:



## Cost Comparison Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Platform | Storage Cost (Approx.) | Document Processing Cost | Scalability | Customization Support |
| GCP Document AI | ~$0.026/GB/month | Competitive, pay-per-page pricing | High | Excellent (custom model training) |
| AWS Textract | ~$0.023/GB/month (S3) | Variable, can be higher with volume | High | Limited customization |
| Azure Form Recognizer | ~$0.02/GB/month | Variable pricing based on usage | High | Moderate customization |

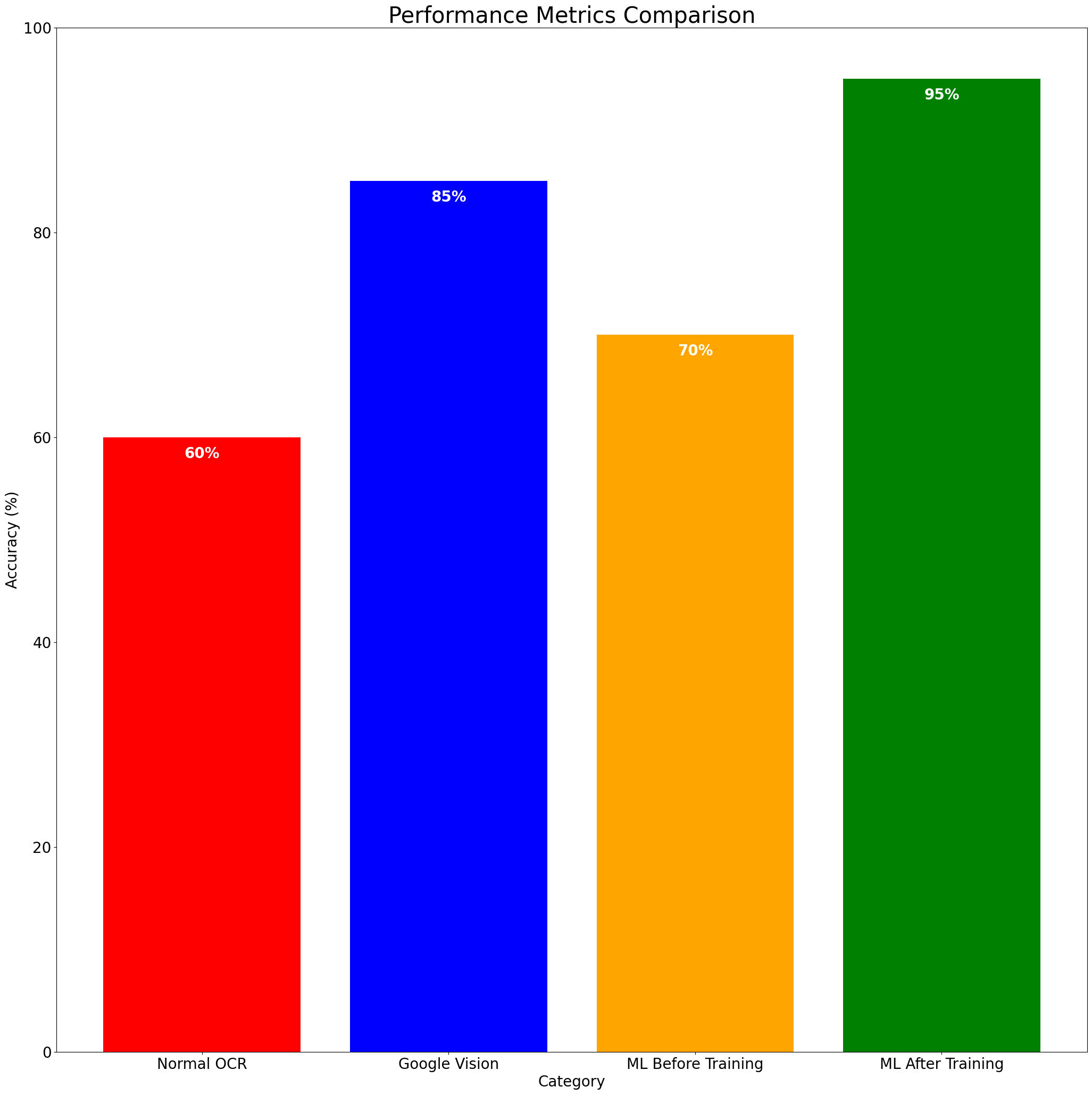
## Performance Metrics Chart:

* 1. With a Tool Developed without ML: (<https://smooth-ocean.tech/>)
* Normal OCR: ~60% accuracy
* Google Vision: ~ 85% accuracy

**NOTE=** This Tool Provides a User Interface for the Users to Draw Box’s & Label and Classify Data but it is Extremely Slow as it Requires Manual Labeling for Each Document by the User.

* 1. With the Machine Learning Model Integrated in the Workflow:
* Before Training: ~70% accuracy
* After Training: ~95% accuracy predicted

**NOTE=** The Training Uses a Custom Processor Model and Returns Labeled data in Correct Order.



As Real-Time Data Recorded by Jonathan for his Company.

This comprehensive approach not only meets the current requirements but also lays a robust foundation for future enhancements and scalability.

## Pipeline/Workflow:

### Data Collection and Preprocessing:

* + Collect the sample PDFs from the Company's Servers & Store them in Google Cloud Storage Buckets.
  + Store into different Buckets for Training, Testing and Validation of the Model.
  + Make sure that the PDFs are in the Correct Format and that all the Information is Readable & Valid.

### Model Selection and Training:

* + Use Document AI to Train the Model to Extract the Information from the PDFs.
  + Use the Custom Model to Train the Model to Extract the Information from the PDFs.
  + Use the Training, Testing and Validation Buckets to Train the Model.
  + Create Labels Corresponding to the Information that needs to be Extracted from the PDFs.
  + To Train the Model with the Documents use the Labels to Draw Boxes around the Information that needs to be Extracted.
  + Use the Labels to Train the Model to Extract the Information from the PDFs.

### Validation and Testing:

* + Use the Validation Bucket to Validate the Model and Check the Accuracy of the Model.
  + Use the Testing Bucket to Test the Model and Check the Accuracy of the Model.

### Deployment and Monitoring:

* + Deploy the Model to the Company's Servers and Integrate it with their Systems.
  + Monitor the Model for Accuracy, Performance, Errors, etc...
  + Make sure that the Model is working Correctly and that the Information is being Extracted Correctly.

### Evaluation:

* + Evaluate the Model for Accuracy, Performance, Errors, etc...
  + Evaluate the Model for Time, Efficiency, etc...
  + Evaluate the Model for Cost, Scalability, Accessibility, Computational Power, Ease of Use, Implementation Challenges, etc...
  + Evaluate the Model for Future Implications and possible to Scale the Solution.

## Conclusion:

This report analysed the implementation of Google Cloud Document AI for automating document processing, focusing on economic impact, security, challenges, and scalability.

### Key Findings:

* **Cost & ROI:** The solution significantly reduces manual effort, achieving up to 80% time savings and annual cost savings between $50K–$200K, with an estimated ROI breakeven within 6–12 months.
* **Security & Compliance:** Google Cloud ensures robust security through encryption, IAM controls, and compliance with industry standards like GDPR, SOC 2, and HIPAA.
* **Performance & Challenges:** Preprocessing techniques and model fine-tuning enhance OCR accuracy from ~70% to ~95%, mitigating issues like poor scan quality and document format variations.
* **Scalability & Future Growth:** The system is designed to support additional document types, integrate human validation for critical cases, and leverage serverless computing for real-time automation.

Final Thoughts:  
By leveraging cloud-based AI for document processing, businesses can streamline workflows, reduce costs, and enhance data accuracy while maintaining high security standards. With continuous improvements in AI models and cloud infrastructure, the solution is positioned for long-term scalability and efficiency gains.

## References:

* Google Cloud, "GCP Document AI," Google. <https://cloud.google.com/document-ai>
* Amazon Web Services, "AWS Textract," AWS. <https://aws.amazon.com/textract/>
* Microsoft Azure, "Azure Form Recognizer," Microsoft. <https://azure.microsoft.com/en-us/services/form-recognizer/>
* Google Cloud, "GCP Pricing Calculator," Google. <https://cloud.google.com/products/calculator>
* Google Cloud, "GCP Security & Compliance," Google. <https://cloud.google.com/security>
* Google Cloud, "Google Cloud Functions," Google. <https://cloud.google.com/functions>
* Google Cloud, "Google Vertex AI," Google. <https://cloud.google.com/vertex-ai>