Capstone Project Report

Cost Estimation Application

Prepared for : City of Stockton

In association with : University of the Pacific – Data Science Department

Submitted by :

Project Group 7

Purva Mugdiya

Usha Pavani Thopalle

Sahana Reddy

Vamshi Krishna Perabathula

Chaitanya Vishnu Radhakrishna

Chen-Li Lee

Under the guidance of :

Prof. Arshad Khan

School of Engineering and Computer Science

University of the Pacific

Date : 1st May, 2025

DECLARATION

We hereby certify that the project titled *"Cost Estimation Application"* is the result of our own work, completed as part of the Capstone Project for the Master of Data Science program at the University of the Pacific. This work was carried out under the guidance of Prof. Arshad Khan between January and April 2025. All information presented in this report is based on our original research and analysis and is accurate to the best of our knowledge.

CERTIFICATE

This is to certify that the project work carried out by the above-mentioned students is, to the best of my knowledge, accurate and authentic. The project was undertaken under my supervision as part of the

Capstone Project for the Master of Data Science program at the University of the Pacific. The work presented is original, has not been submitted for any other degree or qualification at any institution, and meets the requirements for the award of the Master's degree in Data Science.

Name of the Mentor : Prof. Arshad Khan

Designation : Professor

School of Computer Science and Engineering

University of the Pacific, San Francisco, CA.

Date : 8th April, 2025

ACKNOWLEDGEMENT

We extend our sincere gratitude to Prof. Arshad Khan for his consistent support and guidance throughout our program. His hands-on sessions, especially on data visualization best practices, greatly strengthened our ability to communicate insights effectively—an essential skill for this project.

We also thank Dr. James Hetrick, Director of the Data Science Program, for the opportunity to undertake this Capstone Project. His leadership and dedication to the program have fostered a supportive and intellectually stimulating environment, encouraging us to translate classroom learning into practical, real-world solutions.

Finally, we are grateful to our peers and families for their continuous encouragement and support throughout this journey. Their belief in us played a key role in our progress and the outcomes we've achieved.

Table of Contents

1. Executive Summary.. **6**

2. Background.. **7**

3. Project Objective **7**

4. System Description *8*

4.1 Data Ingestion and Input Handling 8

4.2 Cost Estimation Engine........ 10

4.3 Report Generation and Visualization 11

5. Cost Components and Estimation Logic........ *13*

5.1 Overview of Cost Components 13

5.2 Calculation Logic 13

5.3 Cloud Storage Rate Handling........ 14

5.4 Manpower Effort Levels 14

5.5 Dynamic Cost Estimation Output........ 14

6. Technology Stack and Tools Used........ *16*

6.1 Frontend and User Interface........ 16

6.2 Backend Processing........ 16

6.3 Cloud Pricing Integration 16

6.4 Development Environment 17

6.5 Output & Reporting Formats 17

7. Data Input and User Interaction........ *17*

7.1 Input Methods 17

7.2 User Interface Elements........ 18

7.3 Feedback and Visual Summaries........ 19

7.4 Customization Options 20

7.5 Estimation Results and Interactivity........ 20

7.6 Session History and Tracking........ 20

8. Advanced Analytics and Intelligent Features........ *20*

8.1 Interactive Dashboards........ 21

8.2 Visualization Types 21

8.3 Report Generation 21

8.4 Use Cases for Reporting........ 21

8.5 Smart Features and Enhancements 22

9. Session History and Audit Trail........ *23*

9.1 Automatic Logging of Estimations 23
9.2 History Viewer Features 23
9.3 Reportable Audit Trail 24
9.4 History Management Options 24
10. Use Cases and Potential Impact *24*
10.1 Departmental Use Cases 24
10.2 Strategic Benefits 25
10.3 Long-Term Impact 25
11. Limitations *25*
11.1 Technical Limitations 25
11.2 Estimation Assumptions 26
11.3 Functional Constraints 26
11.4 Usability Considerations 26
12. Recommendations and Future Enhancements *27*
12.1 Functional Enhancements 27
12.2 Advanced Estimation Capabilities 27
12.3 Integration Opportunities 27
12.4 Analytics & Intelligence 27
12.5 User Access and Collaboration 28
13. Appendix *28*
13.1 Application Process 28
13.2 Sample cost reports 31
13.3 Glossary 33

1. Executive Summary

The Cost Estimation Application is a custom-built digital tool developed for the City of Stockton to streamline and optimize the cost evaluation process associated with large-scale data digitization projects. Designed as a web-based platform using Streamlit, the application simplifies the estimation of storage, processing, and manpower costs related to the conversion of physical documents into digital formats.

At its core, the application allows users to upload PDF documents or manually input page data. It then calculates key cost components including cloud storage, OCR processing, scanning, software licensing, and manpower effort—factoring in both static and live cloud pricing from Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. The tool supports regional pricing and provides fallback rates when live data is unavailable.

To enhance transparency and usability, the application includes features such as:

- File metadata extraction and summarization.
- Custom pricing options to suit varying contract needs.
- Session history tracking with CSV and PDF report generation.
- Visual analytics using interactive charts for cost and trend analysis.

Built with flexibility in mind, the tool empowers City of Stockton departments to make informed, data-driven budgeting decisions for their digitization initiatives. It ensures consistency in estimates, reduces manual calculation errors, and facilitates clearer communication with stakeholders.

2. Background

The City of Stockton, like many public sector organizations, manages a vast amount of physical documentation across departments—ranging from permits and contracts to historical records and public service files. Traditionally, these documents have been stored in filing cabinets, onsite archives, and offsite storage facilities. This approach, while long-standing, presents several operational challenges including space constraints, delayed retrieval times, and the risk of physical degradation or loss.

When digitization projects are proposed to modernize this documentation, budgeting is typically done manually. Departments estimate costs based on rough assumptions—often relying on outdated pricing tables, informal vendor quotes, or prior project experiences. These manual estimates are not only time-consuming but also prone to inaccuracies, leading to budgeting shortfalls or over-allocation of funds. Key cost factors such as OCR processing, scanning rates, cloud storage pricing (which varies by provider and region), software licensing, and manpower requirements are frequently overlooked or inconsistently accounted for.

Given the growing push for digital transformation and the increasing complexity of cloud pricing models, there was a clear need for a centralized, accurate, and user-friendly tool to assist city staff in generating realistic cost estimates. The Cost Estimation Application was developed in response to this need—offering a smarter, data-driven alternative to manual budgeting for document digitization efforts.

3. Project Objective

The primary objective of the Cost Estimation Application is to equip the City of Stockton with a robust, user-friendly platform that facilitates accurate and efficient budgeting for data digitization initiatives. As government agencies continue to transition from paper-based to digital systems, there is a growing need for precise cost estimation tools that can handle the complexities of such projects.

This application addresses that need by automating the estimation process and removing the guesswork traditionally involved in manual budgeting. It allows users to estimate costs based on either the number of pages or the file size of documents, with support for both manual entry and direct PDF uploads. When PDFs are uploaded, the tool automatically extracts file size, page count, and metadata, saving time and ensuring accuracy.

A key feature of the application is its ability to calculate costs across multiple dimensions. It considers:

- Cloud storage fees, adjusted for retention period and specific provider pricing.
- OCR (Optical Character Recognition) processing costs per page.
- Scanning costs for physical-to-digital conversion.
- Manpower costs adjusted based on selected effort level (Low, Medium, High).
- Software licensing fees, which vary by cloud provider.

The application supports three major cloud storage platforms—Amazon S3, Google Cloud Storage, and Microsoft Azure—and includes region-specific pricing for each, retrieved dynamically via live APIs where possible. When live data is unavailable, the system falls back to predefined default rates, ensuring that users can continue with their estimation process uninterrupted.

In addition, the tool provides options for custom pricing, allowing users to override default cost values and manpower multipliers. This makes the application flexible enough to accommodate special contracts, vendor-specific terms, or internal budgeting policies.

Beyond estimation, the application includes features for report generation, session history tracking, and interactive data visualization. Users can download detailed cost reports in CSV or PDF format, revisit past estimates, and analyze trends using charts and graphs. These capabilities not only improve internal transparency but also support better communication of cost structures across teams and stakeholders.

Overall, the Cost Estimation Application serves as a comprehensive digital solution to support the City of Stockton's modernization efforts—delivering consistency, transparency, and accuracy in the planning and execution of digitization projects.

4. System Description

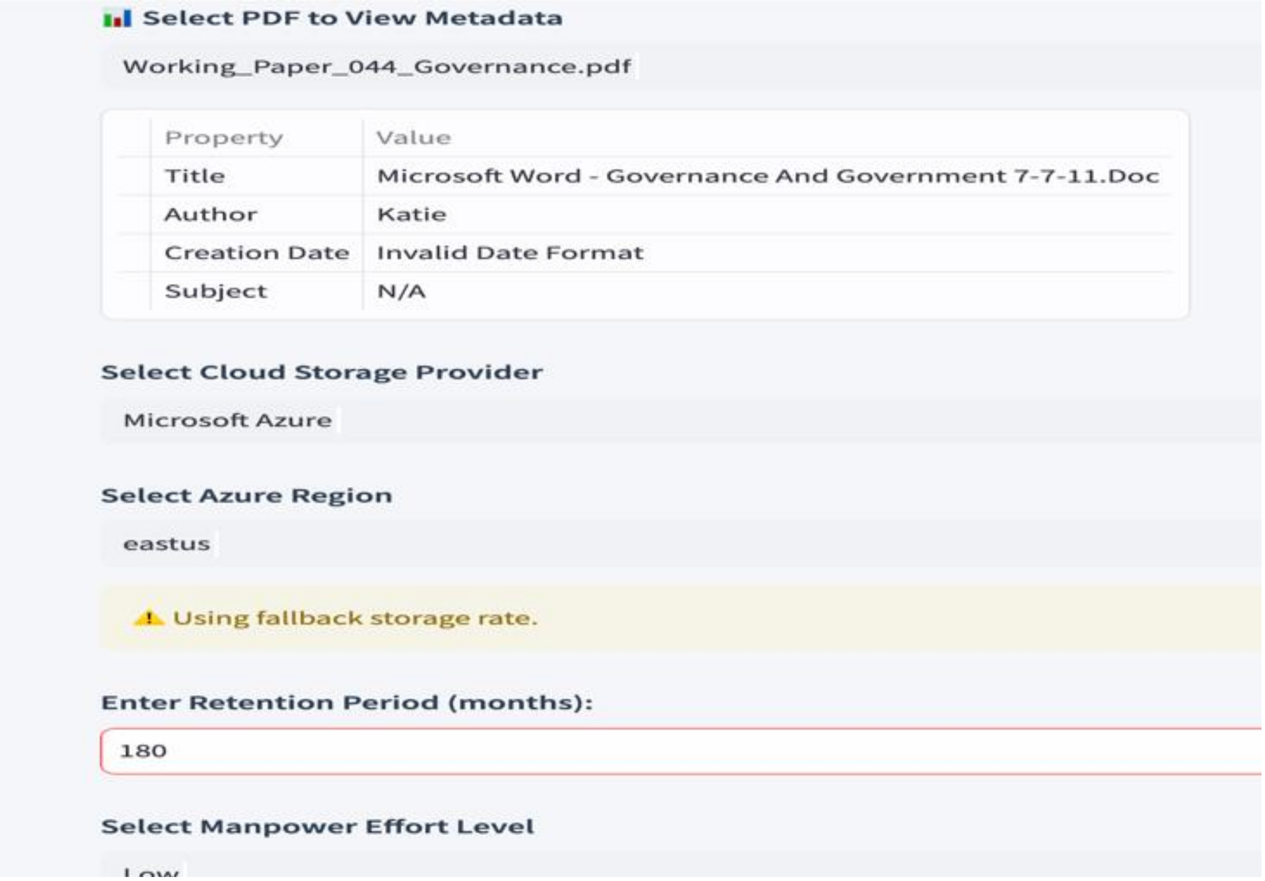
The Cost Estimation Application is a user-friendly, modular web tool that simplifies cost analysis for document digitization projects. Built using Streamlit, it guides users from document input to cost estimation, report generation, and visual comparison—while supporting live pricing, custom inputs, and smart automation.

4.1 Data Ingestion and Input Handling

Users can initiate estimations through:

1. Upload PDFs
    - Accepts multiple files.
    - Automatically extracts file size, page count, title, author, subject, and creation date using PyMuPDF.

- Prevents duplicate uploads.
- Figure 1.



📊 Select PDF to View Metadata

Working_Paper_044_Governance.pdf

Property	Value
Title	Microsoft Word - Governance And Government 7-7-11.Doc
Author	Katie
Creation Date	Invalid Date Format
Subject	N/A

Select Cloud Storage Provider

Microsoft Azure

Select Azure Region

eastus

⚠ Using fallback storage rate.

Enter Retention Period (months):

180

Select Manpower Effort Level

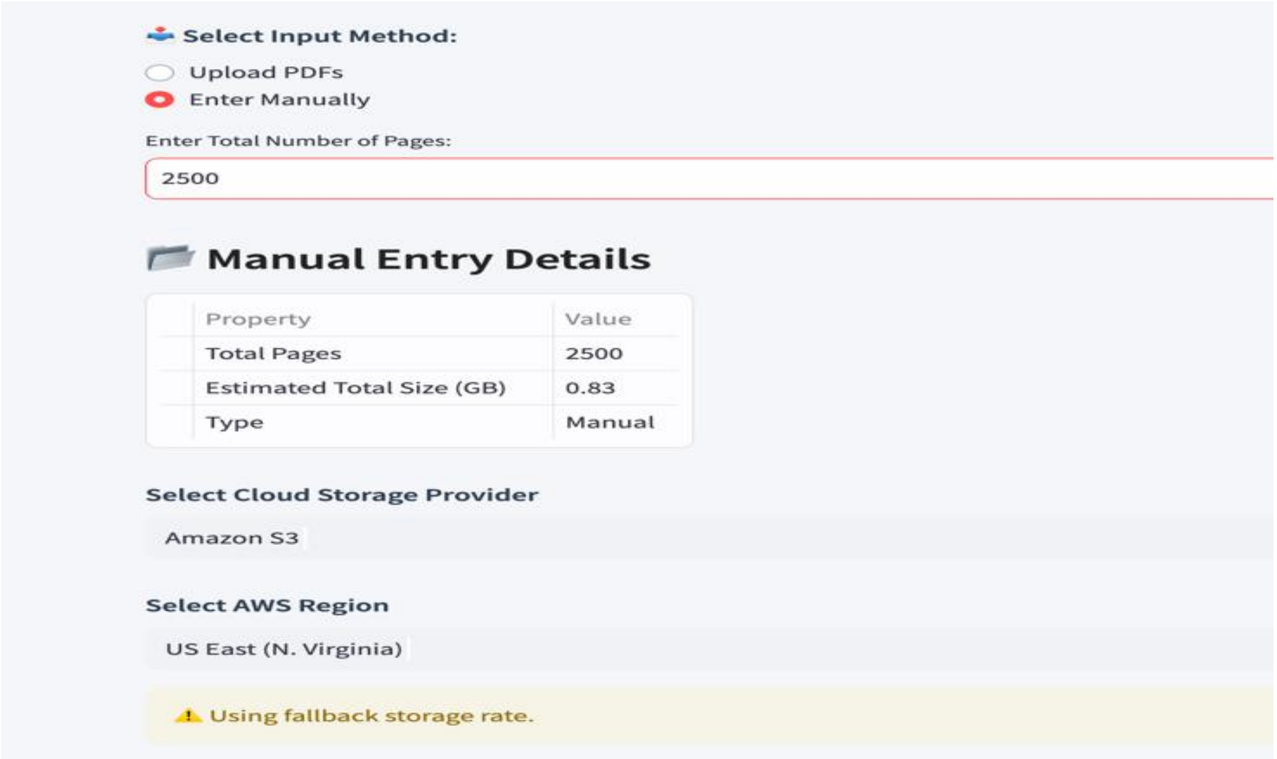
Low

Figure 1

2. Manual Entry
   - Users input page count directly.
   - System estimates file size based on an average page size of 350 KB.

Figure 2

    - Figure 2.



Select Input Method:

○ Upload PDFs
● Enter Manually

Enter Total Number of Pages:

2500

Manual Entry Details

Property	Value
Total Pages	2500
Estimated Total Size (GB)	0.83
Type	Manual

Select Cloud Storage Provider

Amazon S3

Select AWS Region

US East (N. Virginia)

⚠ Using fallback storage rate.

Once input is provided, the system displays:

- A file summary table (name, size, page count).
- Aggregated totals (pages, size in GB).
- Metadata for any selected file.

These dynamic previews support input validation and transparency before cost estimation.

4.2 Cost Estimation Engine

The system performs detailed cost estimation using both static defaults and real-time API pricing. Users are guided through the following steps:

- Select Cloud Provider (Amazon S3, Google Cloud Storage, Microsoft Azure).
- Choose Storage Region:
    - AWS: e.g., US East (N. Virginia).
    - Azure: e.g., East US.
    - GCP: e.g., US.
- Provide Retention Period (in months).
- Set Manpower Effort Level (Low, Medium, High).
- (Optional) Enable and configure Custom Pricing.
- Figure 1.

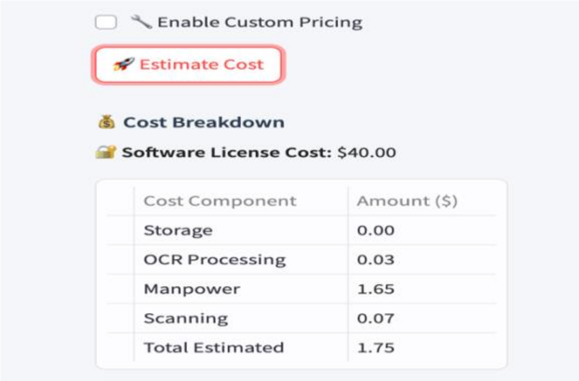
On submission, the system calculates:

Cost Component	Calculation Formula
Scanning	Number of pages x Scanning rate per page
OCR Processing	Number of pages x OCR rate per page
Storage	File size (GB) x Storage rate x Retention period (mo)
Manpower	Number of Pages × Effort multiplier
Software Licensing	Flat cost depending on provider

The total is displayed in a structured breakdown, and users can export this data immediately.

The system retrieves live cloud storage rates (AWS, Azure, GCP) when available. If real-time data cannot be fetched, fallback default rates are used. Users also have the option to input custom pricing for each cost factor, enabling tailored estimations for specific vendors or internal cost models.

Cost Component	Amount ($)
Scanning	24.00
OCR	12.00
Cloud Storage	18.50
Manpower	45.00
Software Licensing	40.00
Total Estimated	139.50



☐ 🔧 Enable Custom Pricing

🚀 Estimate Cost

💰 **Cost Breakdown**

🔐 **Software License Cost:** $40.00

Cost Component	Amount ($)
Storage	0.00
OCR Processing	0.03
Manpower	1.65
Scanning	0.07
Total Estimated	1.75

Figure 3

4.3 Report Generation and Visualization

The Cost Estimation Application provides a rich set of tools to transform estimation data into clear, actionable visual outputs and downloadable reports. This functionality supports budgeting, planning, grant writing, and cross-departmental communication.

- View dynamic charts showing cost distribution by category, Figure 4.
- Compare costs across storage providers and document types, Figure 5.
- Analyze manpower costs based on document complexity or volume.

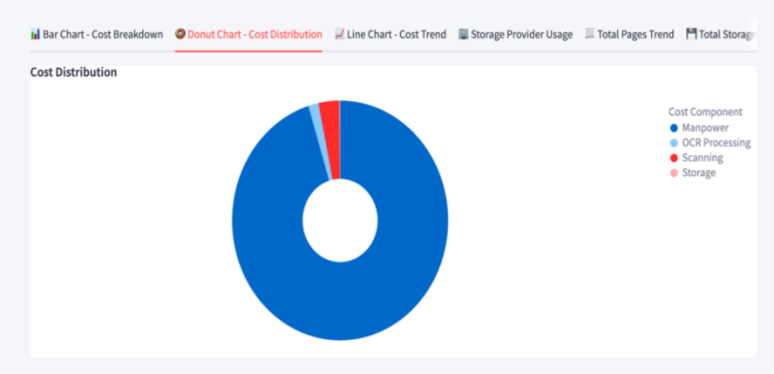


Figure 4

Export Formats

The system supports two export options for reporting:

Format	Use Case
CSV	Best for spreadsheet analysis, auditing, or importing into external tools
PDF	Ideal for presentations, formal reports, or grant documentation

Available report types include:

- Cost Breakdown Report (PDF or CSV) – Summarizes the current estimate by cost category, refer Figure 6.
- Session History Report (PDF or CSV) – Includes all past estimates with metadata, refer Figure 7.
- Filtered Report (PDF or CSV) – Contains only records matching user-defined filters.

Each report includes proper column headers and timestamps, ensuring traceability and usability.

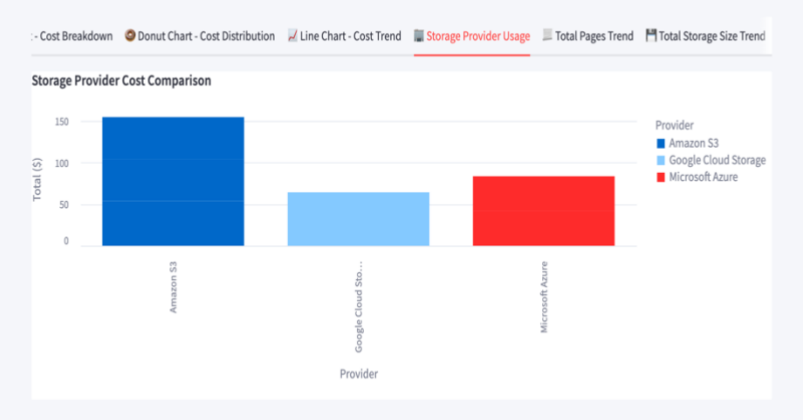


Figure 5

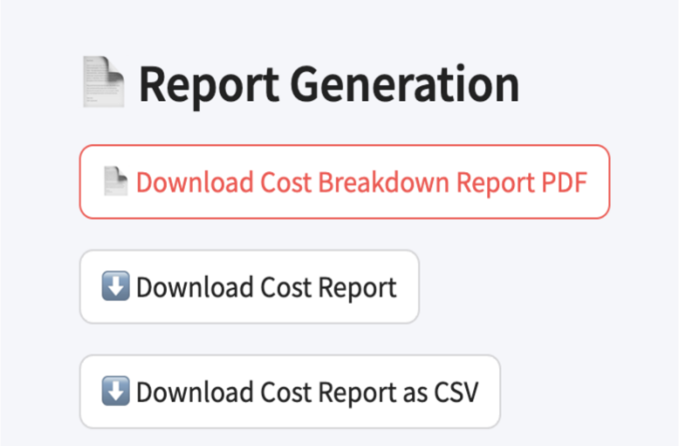

Report Generation
Download Cost Breakdown Report PDF
Download Cost Report
Download Cost Report as CSV

Figure 6

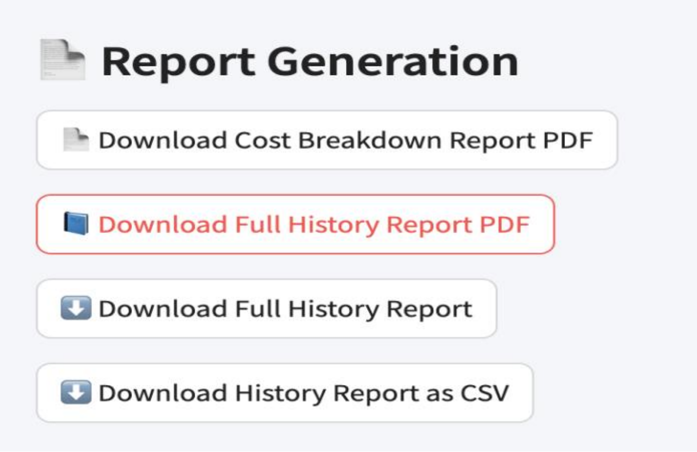

Report Generation
Download Cost Breakdown Report PDF
Download Full History Report PDF
Download Full History Report
Download History Report as CSV

Figure 7

Use Cases for Visualization and Reporting

User Type	Application
Finance Teams	Justify cost allocations and optimize department budgets
Procurement Officers	Compare provider rates and support vendor selection
Project Managers	Plan digitization timelines and assess resource needs
Grant Writers	Include structured, reliable cost estimates in funding proposals
IT or Data Teams	Analyze usage trends to inform infrastructure and storage planning

Strategic Value

These visualization and reporting capabilities not only enhance user understanding of cost dynamics but also:

- Increase Transparency: Clear charts and downloadable reports help communicate estimates to other teams and stakeholders.
- Enable Cost Optimization: Users can visually identify which cost component has the greatest impact and adjust project strategies accordingly.
- Support Data-Driven Decisions: By providing insight into past and predicted costs, departments can better plan for future digitization efforts.

5. Cost Components and Estimation Logic

The Cost Estimation Application is built to provide a transparent and modular view of all cost drivers involved in digitization projects. Each component is calculated independently, using user inputs, predefined formulas, and, when available, real-time data from cloud service providers. This breakdown enables departments to identify the biggest cost contributors and plan accordingly.

5.1 Overview of Cost Components

Component	**Description**
Scanning Cost	Physical-to-digital conversion cost per page scanned.
OCR Processing Cost	Optical Character Recognition cost for converting scanned images into text.
Storage Cost	Cost for storing files in the cloud based on file size, retention period and provider.
Manpower Cost	Manual review, correction, or validation time multiplied by hourly rate.
Software Licensing Cost	Fixed cost per project, depending on the selected cloud provider.

5.2 Calculation Logic

Each cost is calculated based on the following logic:

1. Storage: Real-time API pricing is fetched via:
   - AWS Pricing API (S3 Standard Class).
   - GCP Cloud Pricing JSON.
   - Azure Retail Pricing API.
2. Fallback Rates are used when live data is unavailable:
   - AWS: $0.023/GB/month.
   - GCP: $0.020/GB/month.
   - Azure: $0.020/GB/month.

If Custom Pricing is enabled, users can override these values, along with scanning, OCR, and manpower multipliers.

5.3 Cloud Storage Rate Handling

The system supports real-time pricing from:

- Amazon Web Services (AWS) S3.
- Google Cloud Storage (GCP).
- Microsoft Azure Blob Storage.

If pricing APIs fail to return results, the tool falls back to standard rates:

Provider	Region Example	Source
Amazon S3	US East (N. Virginia)	AWS On-Demand JSON API
Google Cloud Storage	US / EU / Asia	Cloud Pricing Calculator JSON
Microsoft Azure	East US / West Europe	Azure Retail REST API

Pricing data is cached for 1 hour to reduce API load while maintaining accuracy.

5.4 Manpower Effort Levels

Manpower costs reflect the effort required for manual quality checks and data corrections, especially with low-quality scans or handwritten forms.

Effort Level	Multiplier (Per Page)	Use Case
Low	0.03	Clean, typed scans
Medium	0.05	Moderate-quality documents
High	0.08	Handwritten, degraded, or image-heavy files

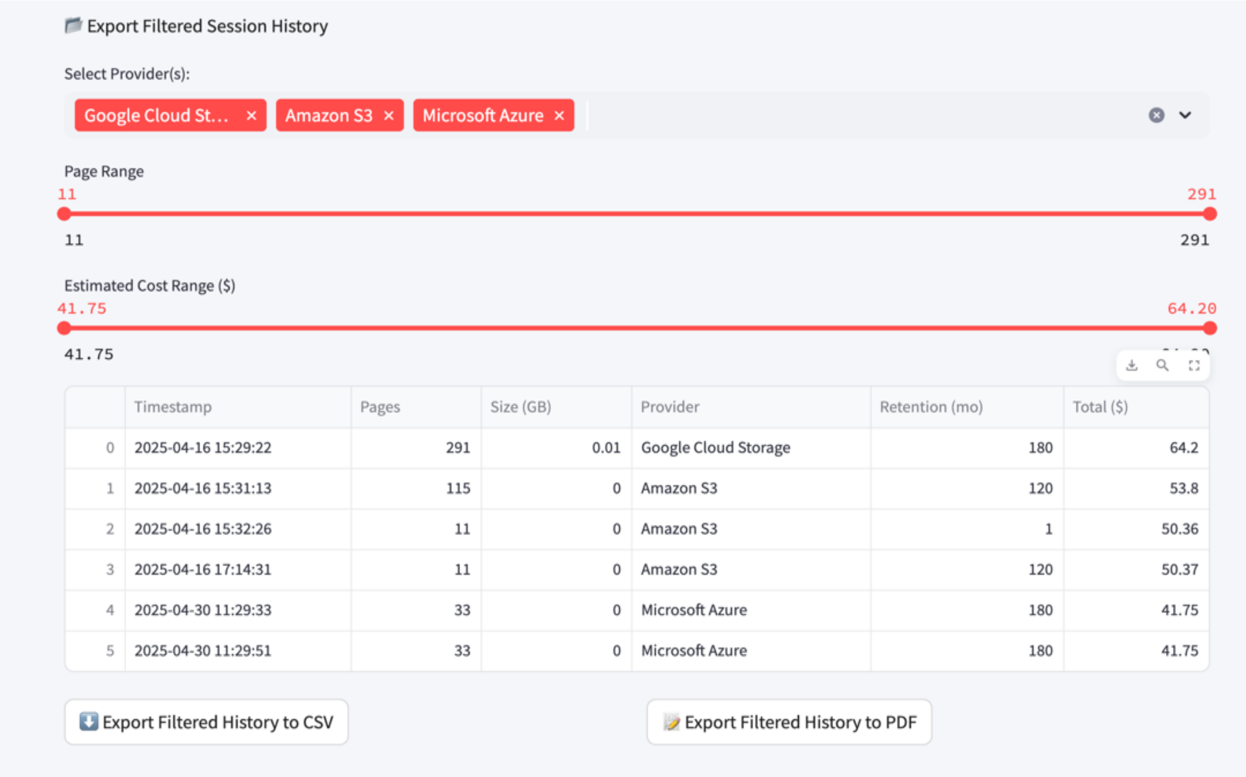
Sliders allow users to fine-tune these values in Custom Pricing mode.

5.5 Dynamic Cost Estimation Output

After computation, users receive:

- Cost Table: Component-wise breakdown.
- License Fee Display: Shown separately for clarity, Figure 3.

- Session Logging: Automatically saved to CSV files, Figure 8.


Export Filtered Session History

Select Provider(s):

Page Range

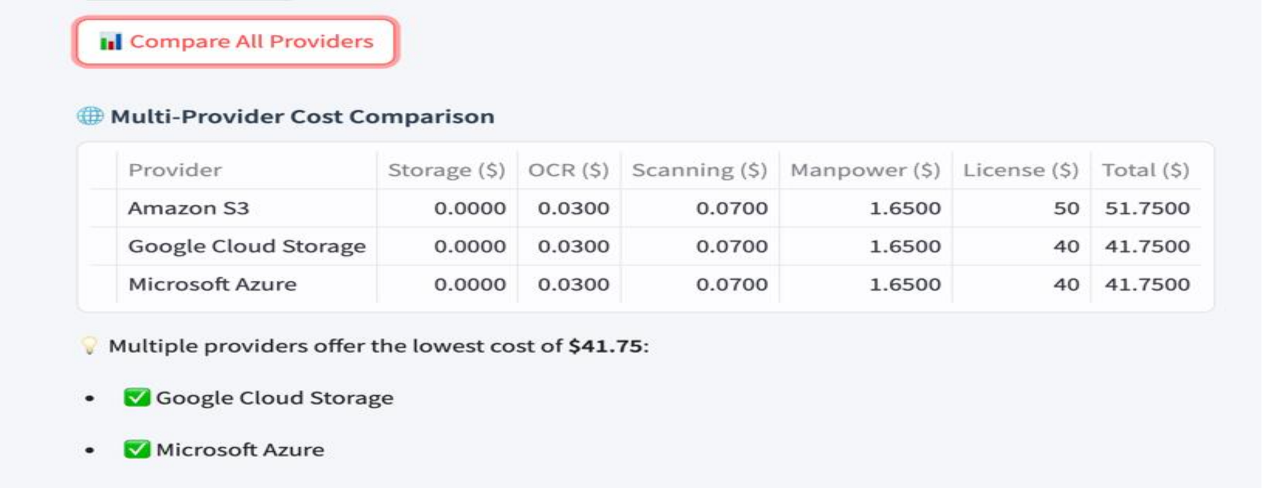
Estimated Cost Range ($)

	Timestamp	Pages	Size (GB)	Provider	Retention (mo)	Total ($)
0	2025-04-16 15:29:22	291	0.01	Google Cloud Storage	180	64.2
1	2025-04-16 15:31:13	115	0	Amazon S3	120	53.8
2	2025-04-16 15:32:26	11	0	Amazon S3	1	50.36
3	2025-04-16 17:14:31	11	0	Amazon S3	120	50.37
4	2025-04-30 11:29:33	33	0	Microsoft Azure	180	41.75
5	2025-04-30 11:29:51	33	0	Microsoft Azure	180	41.75

Export Filtered History to CSV
Export Filtered History to PDF

Figure 8

- Optional Multi-Provider Comparison: One-click feature to calculate equivalent estimates across AWS, GCP, and Azure, with a recommendation engine to highlight the cheapest option, Figure 9.



Multi-Provider Cost Comparison

Provider	Storage ($)	OCR ($)	Scanning ($)	Manpower ($)	License ($)	Total ($)
Amazon S3	0.0000	0.0300	0.0700	1.6500	50	51.7500
Google Cloud Storage	0.0000	0.0300	0.0700	1.6500	40	41.7500
Microsoft Azure	0.0000	0.0300	0.0700	1.6500	40	41.7500

Multiple providers offer the lowest cost of **$41.75**:

- ✓ Google Cloud Storage
- ✓ Microsoft Azure

Figure 9

This modular structure ensures adaptability to a variety of cost scenarios and user roles.

6. Technology Stack and Tools Used

The development of the Cost Estimation Application leverages a modern and efficient technology stack that supports scalability, rapid deployment, and user interactivity. The chosen tools were selected for their open-source accessibility, robust performance, and compatibility with cloud-based workflows. Together, these technologies enable real-time processing, dynamic visualization, and streamlined reporting.

6.1 Frontend and User Interface

Technology	Purpose
Streamlit	Core framework for building the web application UI. It allows fast prototyping with a minimal learning curve and provides built-in components for file uploads, forms, and visual rendering.
Altair	Used for creating interactive and responsive data visualizations such as bar charts, line graphs, and donut charts. Enables clear cost trend analysis and report visuals.

Streamlit's simplicity ensures that non-technical users can easily interact with the platform, while Altair enables intuitive visual storytelling.

6.2 Backend Processing

Technology	Purpose
Python	Core programming language used across the entire application. It powers data handling, calculations, API calls, and file processing.
PyMuPDF (Fitz)	For extracting metadata (title, author, page count, creation date) and structure from uploaded PDF files.
Pandas	Provides efficient data structures for cleaning, transforming, and analyzing cost-related inputs and history logs.
FPDF	Generates downloadable PDF reports, including cost breakdowns and session summaries.
Requests	Facilitates API calls to cloud providers to fetch live pricing data.
datetime	Handles timestamping for file names, history tracking, and retention-based calculations.
OS	Manages local file operations for storing uploads, downloads, history logs, and reports.

These libraries are tightly integrated, enabling smooth transitions between file input, processing, and output generation.

6.3 Cloud Pricing Integration

To ensure cost estimations are grounded in real-world rates, the application connects to live cloud pricing APIs when available:

Provider	Integration Source
Amazon Web Services	AWS pricing API (S3 OnDemanding pricing JSON)
Google Cloud	Cloud Pricing Calculator API (JSON endpoint)
Microsoft azure	Azure Retail Pricing API (Filtered REST query)

These integrations enable the system to retrieve updated cost-per-GB rates for different storage regions. If live rates are unavailable due to timeout or API changes, fallback rates are used to ensure continuity.

6.4 Development Environment

Tools	Purpose
Jupyter	Used during early-stage prototyping and data validation
VS Code	Full-featured IDE for application development and debugging
Git/GitHub	For version control, collaboration, and deployment readiness

These tools allow for agile iteration and collaboration among development team members.

6.5 Output & Reporting Formats

The system supports exporting results in two main formats:

1. CSV: Useful for spreadsheet analysis, grant attachments, or archival tracking.
2. PDF: Formal reports with tabular summaries suitable for presentations and internal reviews.

Together, this technology stack ensures the Cost Estimation Application remains reliable, flexible, and easy to maintain, while offering users a seamless experience from input to insights.

7. Data Input and User Interaction

The Cost Estimation Application is designed with a user-first approach, ensuring that city staff and project planners can easily interact with the system regardless of technical background. It offers flexible input methods, real-time feedback, and customizable options throughout the estimation process.

7.1 Input Methods

Users have two primary ways to begin:

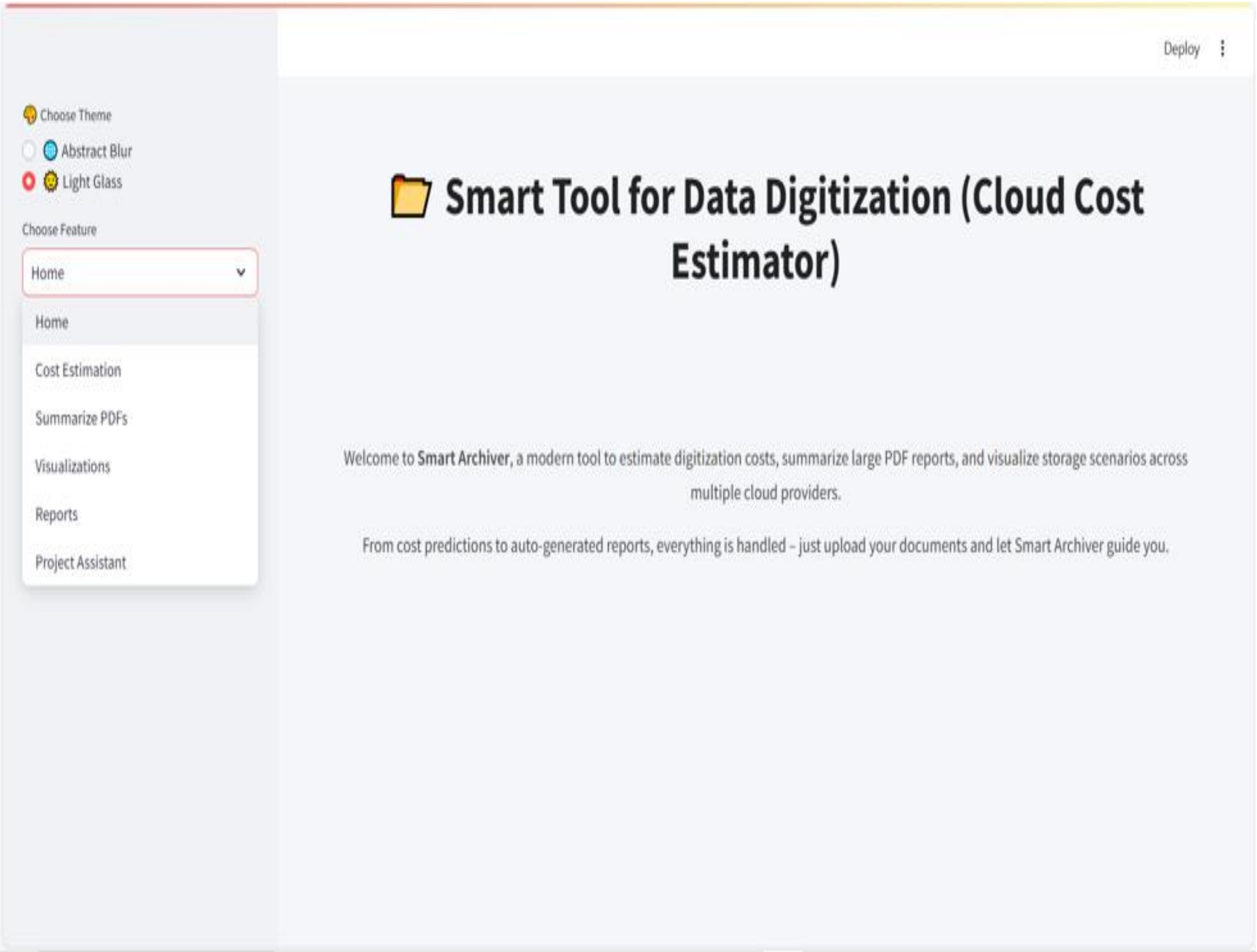
1. Upload Documents
   Users can drag-and-drop or browse to upload scanned files in the following formats: PDF, DOCX, JPG, PNG. Upon upload, the system:
   - Automatically extracts file size, number of pages, and metadata (e.g., title, author, subject, creation date).
   - A Displays individual file details and cumulative totals.
   - Warns if a file is uploaded more than once (deduplication logic).
   - Figure 1.
2. Manual Entry
   Ideal for early-stage planning, users can input:
   - File format.
   - Estimated number of pages.
   - Estimated file size (auto calculated based on page count).
   - Retention period in months.
   - Figure 2.

Both input methods feed into the same estimation engine, ensuring consistency in calculations.

7.2 User Interface Elements

The application uses an intuitive, responsive interface built with Streamlit. Key UI features include:

- Sidebar navigation with feature selector and theme switcher (Light Glass / Abstract Blur), Figure 10.



Choose Theme: Light Glass

Choose Feature: Home

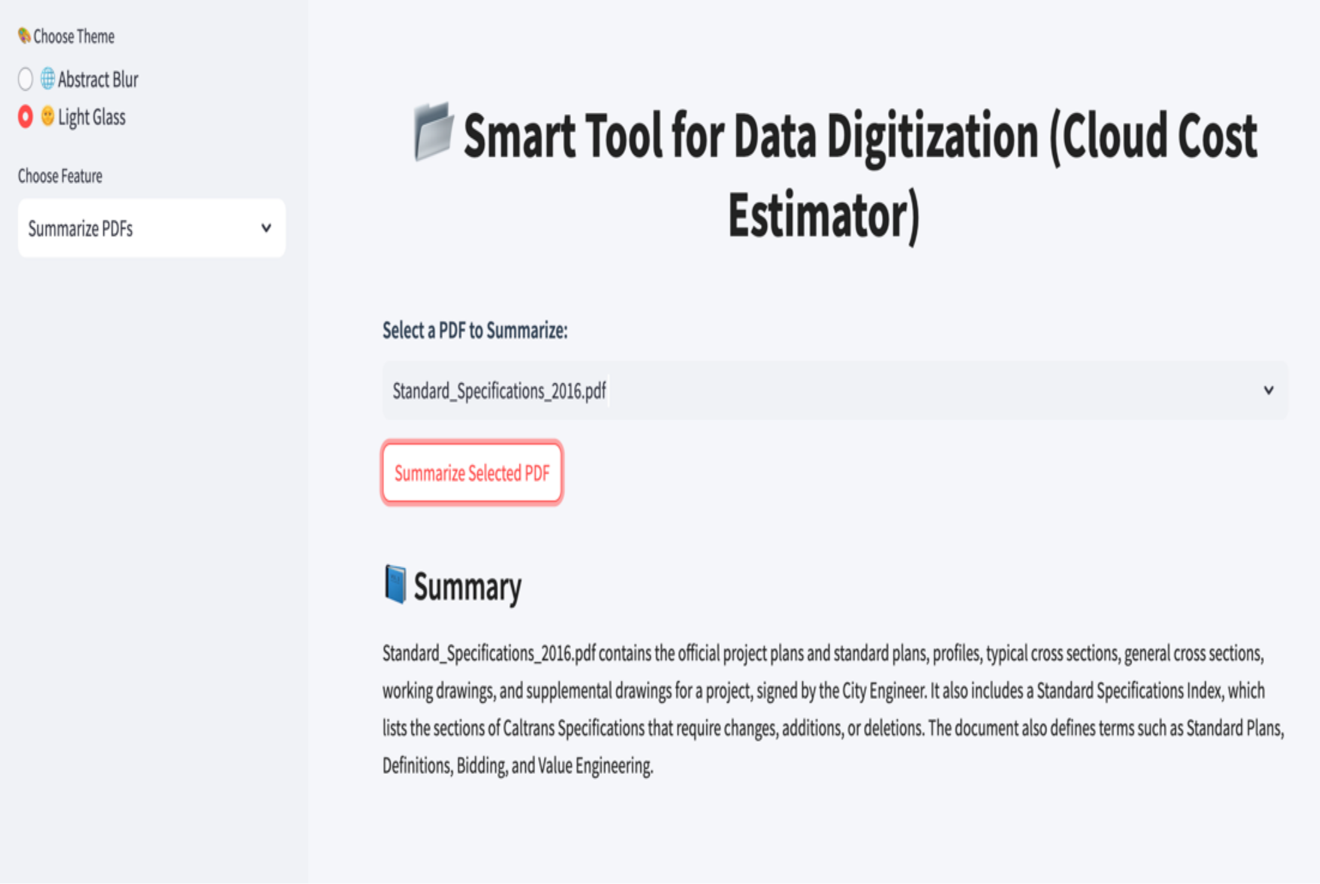
Smart Tool for Data Digitization (Cloud Cost Estimator)

Welcome to **Smart Archiver**, a modern tool to estimate digitization costs, summarize large PDF reports, and visualize storage scenarios across multiple cloud providers.

From cost predictions to auto-generated reports, everything is handled – just upload your documents and let Smart Archiver guide you.

Figure 10

- File uploader with summary preview, Figure 11.



Smart Tool for Data Digitization (Cloud Cost Estimator)

Select a PDF to Summarize: Standard_Specifications_2016.pdf

Summary

Standard_Specifications_2016.pdf contains the official project plans and standard plans, profiles, typical cross sections, general cross sections, working drawings, and supplemental drawings for a project, signed by the City Engineer. It also includes a Standard Specifications Index, which lists the sections of Caltrans Specifications that require changes, additions, or deletions. The document also defines terms such as Standard Plans, Definitions, Bidding, and Value Engineering.

Figure 11

- Manual input fields with validation.
- Region selector for AWS, GCP, Azure.
- Real-time metadata viewer for uploaded files.
- Toggle for enabling advanced custom pricing.

7.3 Feedback and Visual Summaries

Once documents are uploaded or values are entered, users receive:

Display	Details Shown
Uploaded File Summary	File name, size, and number of pages for each uploaded document
Cumulative Overview	Total page count, total file size (KB/GB), and file type
Metadata Table	Title, Author, Subject, and Creation Date for selected files
Input Validation Alerts	Warnings for duplicate files or missing entries

These summaries help users verify the completeness and accuracy of their input before estimation.

7.4 Customization Options

Advanced users can enable Custom Pricing, which reveals fields to override default values:

Customizable Field	Description
Storage Cost (per GB/month)	Override real-time or fallback storage pricing
OCR & Scanning Cost (per page)	Set vendor-specific or internal rate estimates
Software License Cost	Adjust fixed license fee per project/provider
Manpower Multipliers (Low/Med/High)	Fine-tune effort level impact using sliders

These options support scenario planning, RFP analysis, or internal budgeting needs.

7.5 Estimation Results and Interactivity

Upon clicking "Estimate Cost", users are presented with:

- A detailed cost breakdown (by component).
- Visual charts highlighting cost distribution.
- Multi-provider cost comparison (optional).
- Smart recommendation identifying the lowest-cost provider.
- Export options (PDF and CSV).
- Saved session data for future reference.

Results are dynamic and reflect all custom inputs and selections made during the session.

7.6 Session History and Tracking

The application logs each estimation and provides robust history management tools:

Feature	Functionality
View Session History	Shows all past estimations with provider, pages, size, and cost
Filter History	Filter by provider, page range, or cost range using sliders
Export History	Export filtered results as PDF or CSV
Auto-Persisted Logs	Saves all estimates to session_history.csv and master_history.csv

This ensures traceability for audits, project reviews, and multi-department use, Figure 7.

8. Advanced Analytics and Intelligent Features

A core strength of the Cost Estimation Application lies in its ability to transform raw cost data into actionable insights through visual dashboards and intelligent automation. This section outlines the application's visual reporting capabilities, smart features, and export tools that enhance user experience and decision-making.

8.1 Interactive Dashboards

Upon completing an estimate, users are presented with a set of interactive visualizations powered by Altair. These dashboards are organized into clearly labeled tabs, enabling smooth navigation across different views. Visuals are updated dynamically based on session data and allow for real-time exploration of cost components and historical trends.

8.2 Visualization Types

Figure 1: Key Visualization Options

Chart Type	Purpose
Bar Chat – Cost Breakdown	Visualizes individual cost components for a single estimate
Donut Chart – Cost Distribution	Highlights the proportional impact of each cost category
Line Chart – Cost Trend	Shows how estimate totals have changed across past sessions
Storage Provider Usage	Compares total estimated costs by cloud provider.
Total Pages Trend	Displays document volume trends across sessions.
Storage Size Trend	Highlights changes in estimated storage requirements.
Cost Comparison Across Providers	Allows side-by-side evaluation of cloud options.
Multi-Provider Total Cost Comparison	Shows comprehensive cost data for all providers using the most recent estimation parameters.

Charts include hover-to-view tooltips, sorting, and dynamic scaling to support detailed analysis.

8.3 Report Generation

The application enables users to export both current and historical data in two primary formats:

- CSV: For integration with spreadsheets, audits, or procurement systems.
- PDF: For polished, formal reports suitable for meetings, budget justifications, or grant proposals.

Report types include:

- Cost Breakdown Report (latest estimate), Figure 6.
- Full History Report (all sessions), Figure 7.
- Filtered History Report (based on user-selected criteria).

8.4 Use Cases for Reporting

Audience	Purpose
Finance Teams	Justify budget requests and evaluate cost drivers
Procurement Officers	Compare cloud providers for contract negotiations
Project Managers	Plan digitization timelines and manpower needs
Grant Writers	Provide cost breakdowns for funding applications

By combining cost calculations with easy-to-understand visuals and exportable documentation, the application turns raw estimation into a strategic planning asset.

8.5 Smart Features and Enhancements

To expand beyond basic estimation, the Cost Estimation Application includes advanced smart features that improve usability, decision-making, and user support. These enhancements reflect the team’s focus on automation, interactivity, and accessibility for non-technical users.

1. PDF Summarization with AI

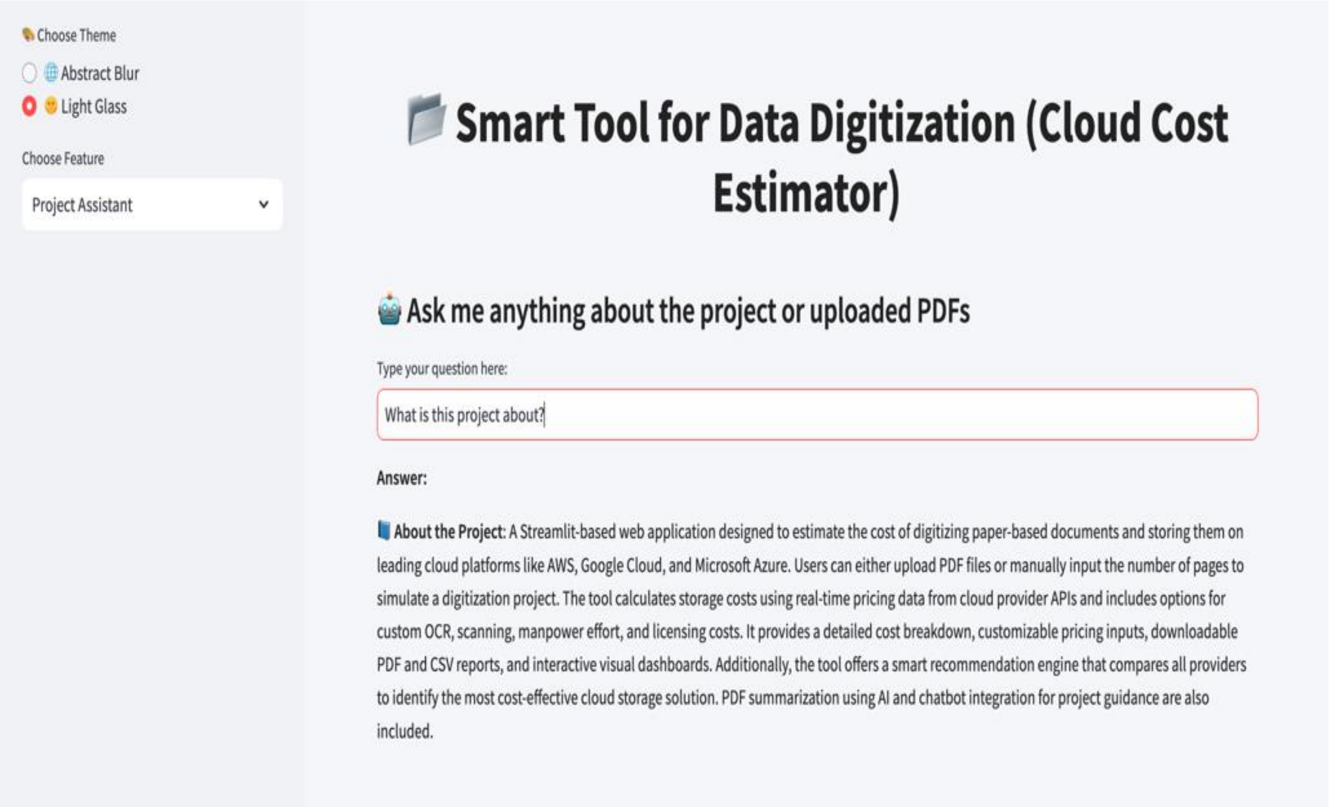
The application supports document summarization powered by a large language model (Mistral-7B via Hugging Face). Users can upload a PDF and receive a concise, natural-language summary of its content. This is especially valuable when working with large files or unfamiliar reports. The tool uses LangChain for document chunking and retrieval, ensuring summaries are context-aware and accurate, Figure 11.

2. Smart Cloud Provider Recommendation

After users perform a multi-provider cost comparison, the tool automatically identifies and recommends the most cost-effective cloud storage provider. This recommendation is based on total estimated cost (including storage, OCR, manpower, scanning, and license fees). It helps users make informed, budget-conscious decisions without manual analysis, Figure 9.

3. Integrated Chatbot Support

The application includes a built-in project assistant chatbot that helps users understand the tool’s features, technical architecture, and use cases. This assistant uses a semantic similarity model (MiniLM) to answer natural-language questions by referencing both project documentation and uploaded PDFs. It enhances accessibility for first-time users and supports self-guided onboarding, Figure 12.



Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature

Project Assistant

Smart Tool for Data Digitization (Cloud Cost Estimator)

Ask me anything about the project or uploaded PDFs

Type your question here:

What is this project about?

Answer:

About the Project: A Streamlit-based web application designed to estimate the cost of digitizing paper-based documents and storing them on leading cloud platforms like AWS, Google Cloud, and Microsoft Azure. Users can either upload PDF files or manually input the number of pages to simulate a digitization project. The tool calculates storage costs using real-time pricing data from cloud provider APIs and includes options for custom OCR, scanning, manpower effort, and licensing costs. It provides a detailed cost breakdown, customizable pricing inputs, downloadable PDF and CSV reports, and interactive visual dashboards. Additionally, the tool offers a smart recommendation engine that compares all providers to identify the most cost-effective cloud storage solution. PDF summarization using AI and chatbot integration for project guidance are also included.

Figure 12

4. Interactive Filtering and Export Tools

Users can apply dynamic filters to session history records before exporting reports. Filters include:

- Cloud provider.
- Page count range.
- Estimated cost range.

Filtered datasets can be downloaded as either CSV or PDF for audit, reporting, or budgeting purposes. This supports more focused reporting and scenario analysis across different project profiles.

9. Session History and Audit Trail

The Cost Estimation Application incorporates a comprehensive session tracking system that ensures all estimations are automatically saved, viewable, and exportable. This supports transparency, reproducibility, and long-term project tracking for budgeting and compliance purposes.

9.1 Automatic Logging of Estimations

Each time a user completes a cost estimation; the following details are saved:

- Total number of pages.
- File size in GB.
- Selected cloud provider.
- Retention period (months).
- Final estimated cost ($).
- Timestamp of the session.

These entries are saved in structured .csv files:

- session_history.csv (current user session).
- master_history.csv (all-time, persistent log).

Additionally, the system stores detailed cost breakdowns (per component) in master_cost_breakdown.csv for reference and reporting.

9.2 History Viewer Features

Users can access their estimation history through the "Reports" section. Features include:

- A searchable, sortable table of all previous estimates.
- Filter options to narrow records by:
  - Cloud storage provider.
  - Page count range (via slider).
  - Estimated total cost range (via slider).
- Dynamic table updates reflecting filter selections.

This enables users to revisit prior estimates, compare costs across scenarios, and replicate previous sessions for consistency.

9.3 Reportable Audit Trail

Filtered or full session histories can be exported using built-in export tools:

- CSV Export: Ideal for internal reviews, spreadsheets, or financial planning systems.
- PDF Export: Professional report format with structured columns and timestamped filenames.

The system also supports download of cost breakdown reports and multi-provider comparisons in both formats, offering flexibility for reporting across use cases.

9.4 History Management Options

Users have control over their historical data with the following features:

- Download Entire History: Export complete or filtered records.
- Download Filtered Subsets: Export only those estimates matching selected criteria.
- Clear Session History: Reset or clear saved estimates from the current session.
- Visual Trend Charts: View session history trends across page counts, providers, costs, and storage size.

These capabilities support internal audits, repeat project budgeting, and long-term tracking of digitization trends within departments.

10. Use Cases and Potential Impact

The Cost Estimation Application was designed with real-world workflows in mind, specifically tailored to address the operational challenges faced by municipal departments like those within the City of Stockton. Its modular, flexible design makes it applicable across various digitization projects—whether small-scale file conversions or large archival digitization efforts.

10.1 Departmental Use Cases

1. City Clerk's Office
   Use the tool to estimate costs for scanning and archiving legislative records, historical city council minutes, or voter registration files.
2. Planning & Development
   Budget for digitizing permits, zoning documents, blueprints, and land use files—many of which are high-volume and complex.
3. Public Works
   Digitize engineering drawings, infrastructure plans, and maintenance logs, which are often in image-heavy or oversized formats requiring specialized handling.
4. Human Resources

Estimate the cost of converting personnel files, timecards, and training records into a searchable digital archive.

5. Finance & Procurement
   Analyze storage and processing costs for invoices, contracts, and tax records to justify system upgrades or grant requests.

10.2 Strategic Benefits

Benefit	Description
Improvement Budget Accuracy	Automates calculations using real-time or customized pricing, reducing estimation errors.
Faster Project Planning	Instantly generates cost estimates with downloadable reports and visualizations.
Cross-Department Consistency	Standardizes the way costs are calculated and presented across different teams.
Audit-Readiness	Built-in session tracking ensures all estimates are logged and exportable.
Supports Grant Applications	Provides breakdowns and documentation often required in funding proposals.
Cost Optimization Insights	Visual comparisons highlight opportunities to reduce costs by adjusting provider, retention, or OCR settings.

10.3 Long-Term Impact

By introducing this tool into daily workflows, the City of Stockton can:

- Scale digitization efforts confidently without relying on external consultants for every budget review.
- Reduce manual overhead, freeing up staff time for more strategic work.
- Improve transparency and collaboration across city departments.
- Build a historical database of digital transformation costs to inform future planning and reporting.

11. Limitations

While the Cost Estimation Application offers a robust framework for document digitization cost analysis, like any system, it comes with certain limitations. Understanding these helps manage expectations and identify areas for future improvement.

11.1 Technical Limitations

Limitation	Details
API Reliability	Cloud pricing APIs (AWS, GCP, Azure) may become temporarily unavailable, leading the system to fall back on static rates.
File Type Support	The tool primarily optimizes for PDF documents. While it accepts images and Word files, extraction and processing may be less consistent.
File Size Constraints	Very large files (e.g., >200 MB) may experience slower processing or timeouts, depending on system capacity.

Browser Dependency	As a Streamlit-based web app, performance may vary slightly across browsers or under poor internet conditions.

11.2 Estimation Assumptions

Assumption	**Impact**
Standard Document Format	The tool assumes users will upload standard scanned documents in supported file types.
Metadata Accuracy	It assumes metadata such as page count and file size can be reliably extracted from uploads.
OCR Format Support	The OCR component is expected to handle multiple file types; however, quality may vary.
Manual Verification Required	For low-quality scans or handwritten content, manual review may be necessary and is accounted for under manpower cost.
NLP-Based Text Correction	The tool uses pre-trained NLP models to correct OCR errors, but perfect accuracy is not guaranteed—manual review may still be needed.
Flat Licensing and Processing Rates	Default rates are based on general market averages and may not reflect actual vendor pricing unless customized by the user.
Cloud Provider Choice	Users can select from AWS, GCP, or Azure. The system supports both live pricing and fallback rates for these providers.
Exportable Reporting	Reports are exportable in PDF and CSV formats. The assumption is that these formats meet user needs for documentation and planning.

11.3 Functional Constraints

Constraint	**Details**
Limited Document Classification	The system does not automatically classify document type or complexity (e.g., forms vs. text-heavy reports).
No OCR Output Display	OCR is included in cost estimation only; actual OCR text output is not shown or stored.
No System Integrations Yet	The application currently operates independently and does not integrate with internal city systems or procurement platforms.

11.4 Usability Considerations

- No Role-Based Access Control: All users have equal access to all features; there's no user-level restriction or login system in place.
- Limited Mobile Support: The UI is designed primarily for desktop use.
- English-Language Only: At this stage, the interface is available only in English.

Summary: These limitations do not hinder the application's core purpose—accurate cost estimation—but they do highlight opportunities for future refinement, especially around scalability, integration, and advanced processing.

12. Recommendations and Future Enhancements

To extend the utility, scalability, and impact of the Cost Estimation Application, several improvements are recommended. These enhancements focus on usability, performance, integration, and long-term alignment with citywide digital transformation goals.

12.1 Functional Enhancements

Recommendation	Rationale
Enable OCR Output Preview	Allow users to view a snippet of the processed OCR text to validate extraction quality and reduce reliance on manual verification.
Add Document Type Classification	Automatically detect whether a file is a form, report, invoice, or image-heavy scan to refine effort estimation and pricing logic.
Support Multi-language Interface	Make the application accessible to a broader set of city staff and partners by localizing UI in Spanish, Tagalog, and other common languages.
Mobile Compatibility	Improve layout and responsiveness for tablet and mobile access, especially useful for field operations or quick reference.

12.2 Advanced Estimation Capabilities

Feature	Impact
Fine Grained OCR Effort Scaling	Let users specify document clarity levels (e.g., clear, noisy, handwritten) to auto-adjust manpower cost estimates.
Batch Upload and Estimate Mode	Enable users to upload folders or multiple documents at once and generate consolidated cost estimates.
Provider Contract Rate	Allow departments to upload negotiated cloud pricing sheets to override standard or live pricing with contract-specific terms.

12.3 Integration Opportunities

System	Potential Benefit
City Procurement Systems	Push finalized cost estimates into internal budget workflows or RFQ templates.
Document Management Platforms	Automatically archive processed files and reports into centralized repositories like SharePoint or Laserfiche.
Grant Management Tools	Export reports in grant-ready formats to streamline proposal submissions.

12.4 Analytics & Intelligence

- Forecasting Module: Integrate predictive analytics to project future digitization costs based on document backlog, seasonal trends, or provider rate changes.
- Cost Benchmarking: Compare cost efficiency across departments or project types, helping the city prioritize high-value digitization efforts.

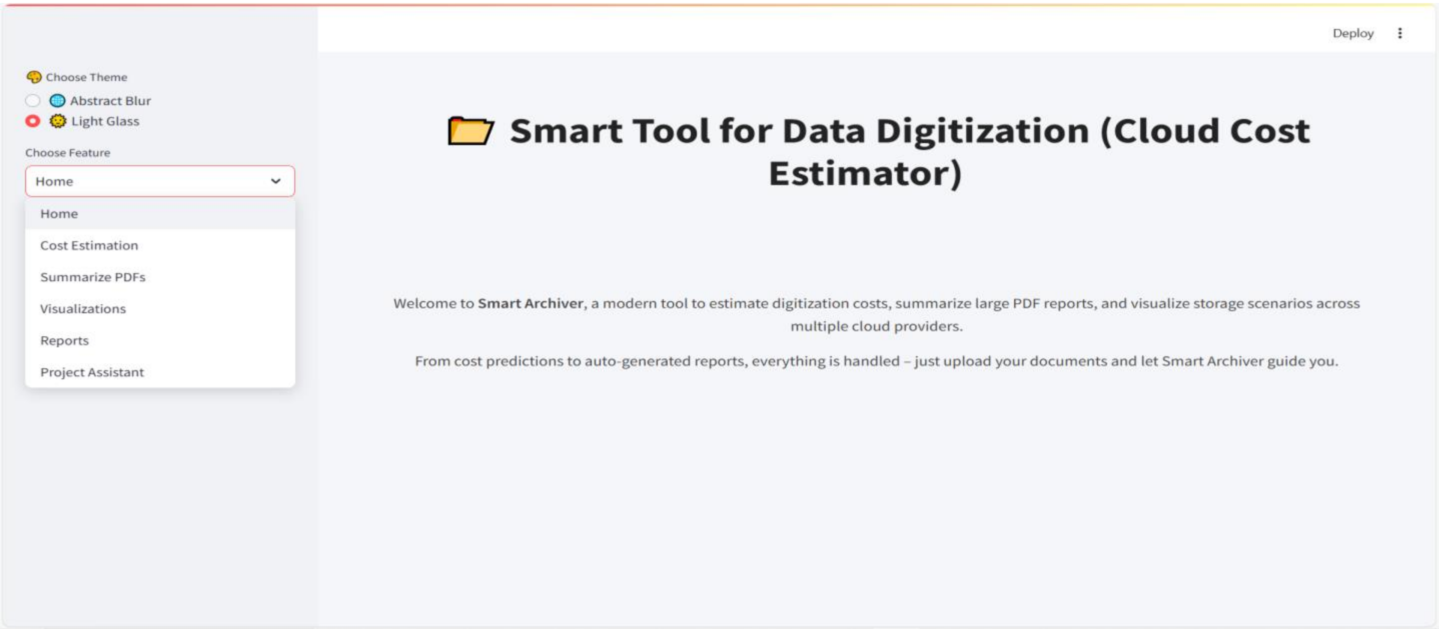
12.5 User Access and Collaboration

- Role-Based Access: Introduce user accounts with different permission levels (e.g., Editor, Viewer, Admin) for multi-department use.
- Collaborative Workspace: Enable users to save, share, or tag estimates by project or department for better coordination.

Vision: These enhancements would not only future-proof the application but also evolve it into a comprehensive planning platform—one that supports budgeting, execution, and accountability for digital transformation across the City of Stockton.

13. Appendix

13.1 Application Process



Choose Theme

○ Abstract Blur

● Light Glass

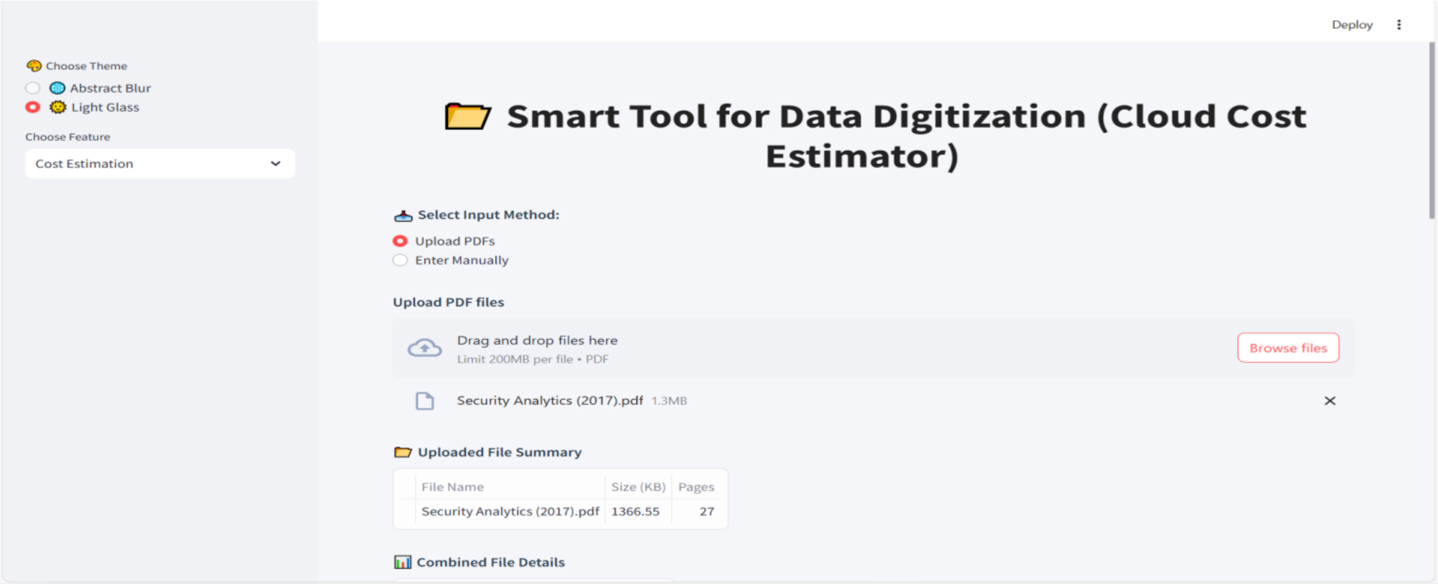
Choose Feature: Home

🗂 Smart Tool for Data Digitization (Cloud Cost Estimator)

Welcome to **Smart Archiver**, a modern tool to estimate digitization costs, summarize large PDF reports, and visualize storage scenarios across multiple cloud providers.

From cost predictions to auto-generated reports, everything is handled – just upload your documents and let Smart Archiver guide you.

Figure 13 - Key Features



Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature: Cost Estimation

📂 Smart Tool for Data Digitization (Cloud Cost Estimator)

Select Input Method:

● Upload PDFs

○ Enter Manually

Upload PDF files

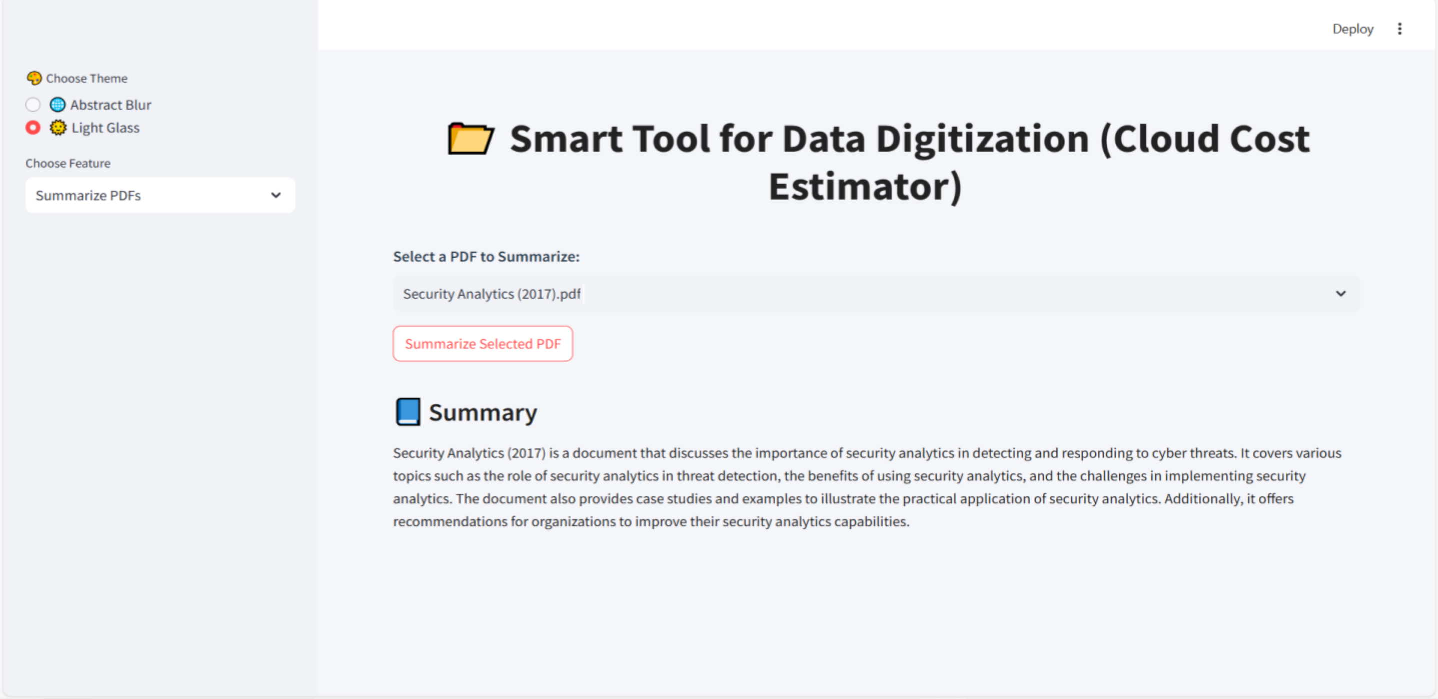
Security Analytics (2017).pdf 1.3MB

📂 Uploaded File Summary

File Name	Size (KB)	Pages
Security Analytics (2017).pdf	1366.55	27

📊 Combined File Details

Figure 14 - Cost Estimation



Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature: Summarize PDFs

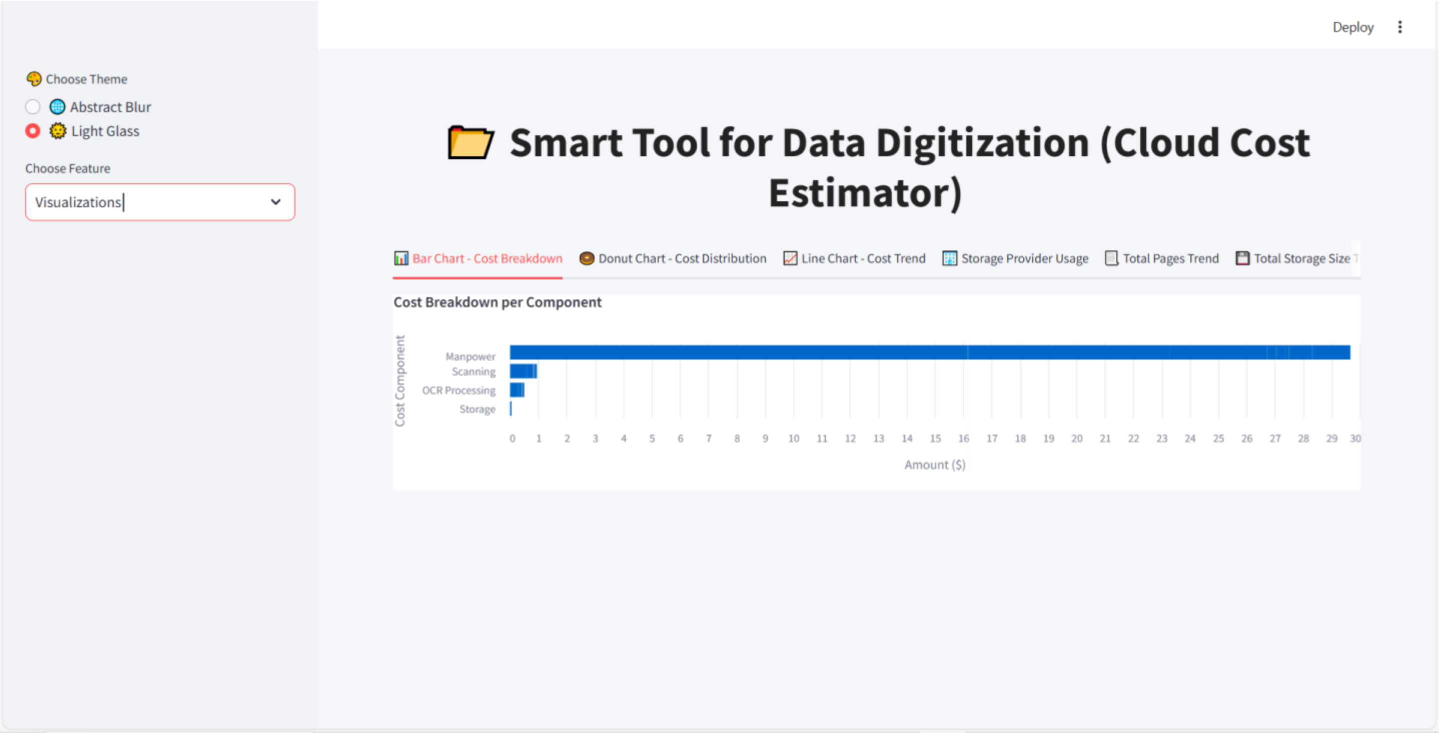
📂 Smart Tool for Data Digitization (Cloud Cost Estimator)

Select a PDF to Summarize: Security Analytics (2017).pdf

📘 Summary

Security Analytics (2017) is a document that discusses the importance of security analytics in detecting and responding to cyber threats. It covers various topics such as the role of security analytics in threat detection, the benefits of using security analytics, and the challenges in implementing security analytics. The document also provides case studies and examples to illustrate the practical application of security analytics. Additionally, it offers recommendations for organizations to improve their security analytics capabilities.

Figure 15 - Summarize PDFs



Deploy

Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature

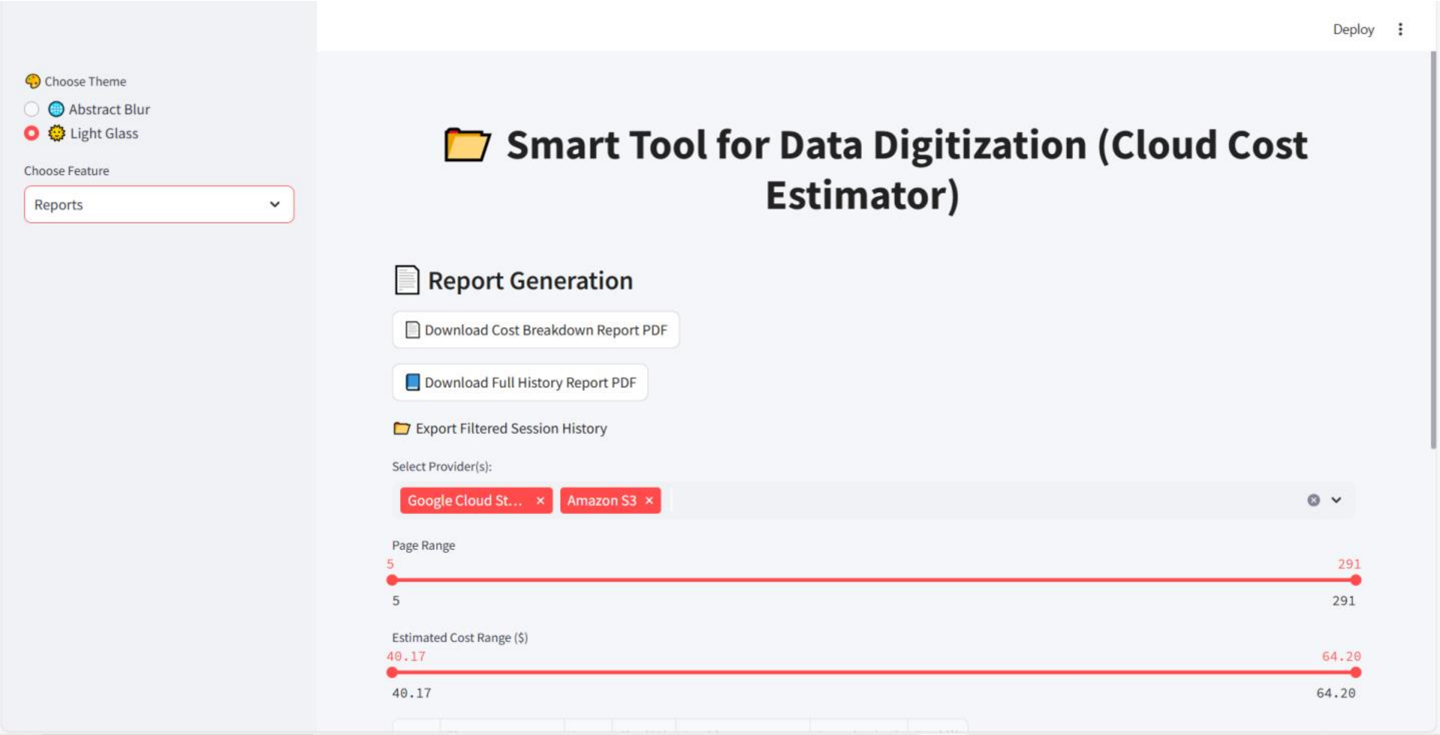
Visualizations

Smart Tool for Data Digitization (Cloud Cost Estimator)

Bar Chart - Cost Breakdown | Donut Chart - Cost Distribution | Line Chart - Cost Trend | Storage Provider Usage | Total Pages Trend | Total Storage Size

Cost Breakdown per Component

Figure 16 - Visualization



Deploy

Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature

Reports

Smart Tool for Data Digitization (Cloud Cost Estimator)

Report Generation

Download Cost Breakdown Report PDF

Download Full History Report PDF

Export Filtered Session History

Select Provider(s):

Google Cloud St... × | Amazon S3 ×

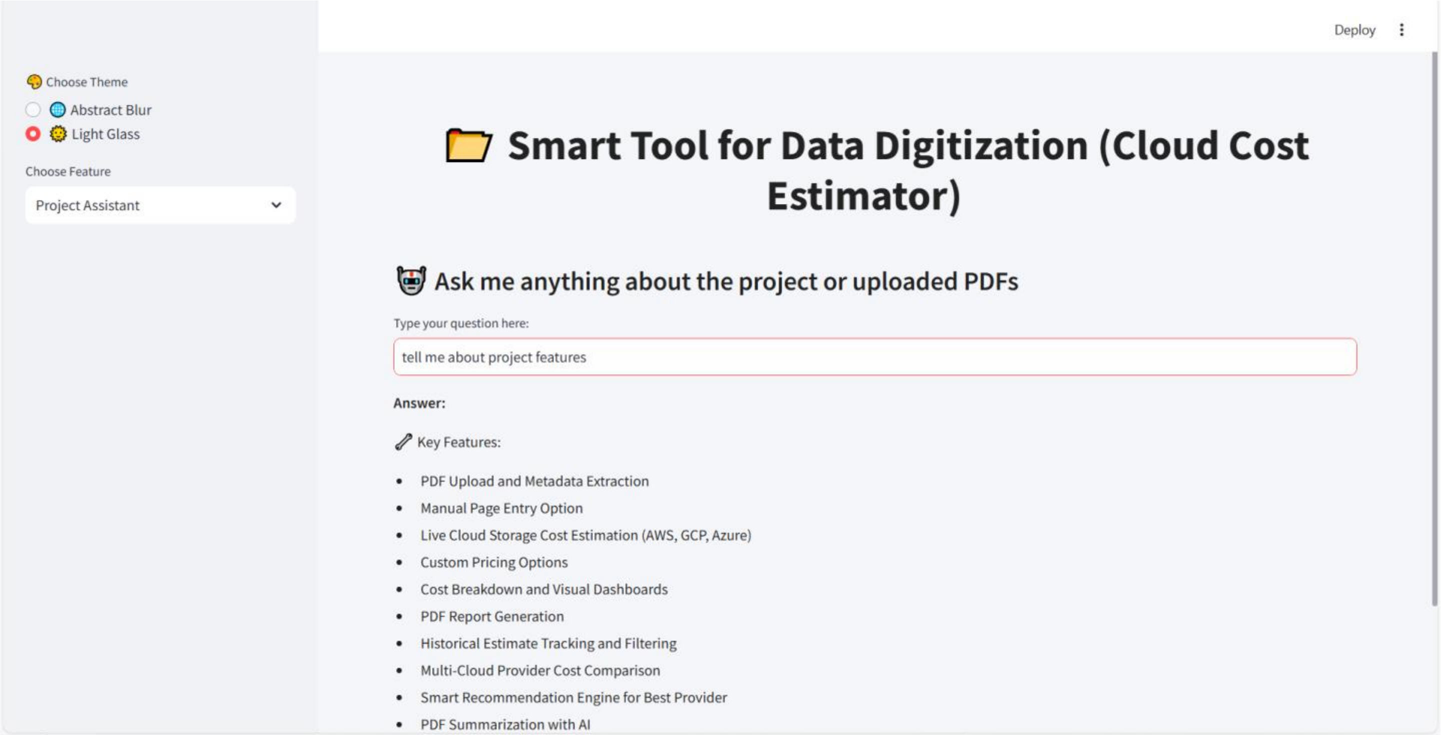
Page Range

5 — 291

Estimated Cost Range ($)

40.17 — 64.20

Figure 17 - Reports



Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature

Project Assistant

📂 Smart Tool for Data Digitization (Cloud Cost Estimator)

Ask me anything about the project or uploaded PDFs

Type your question here:

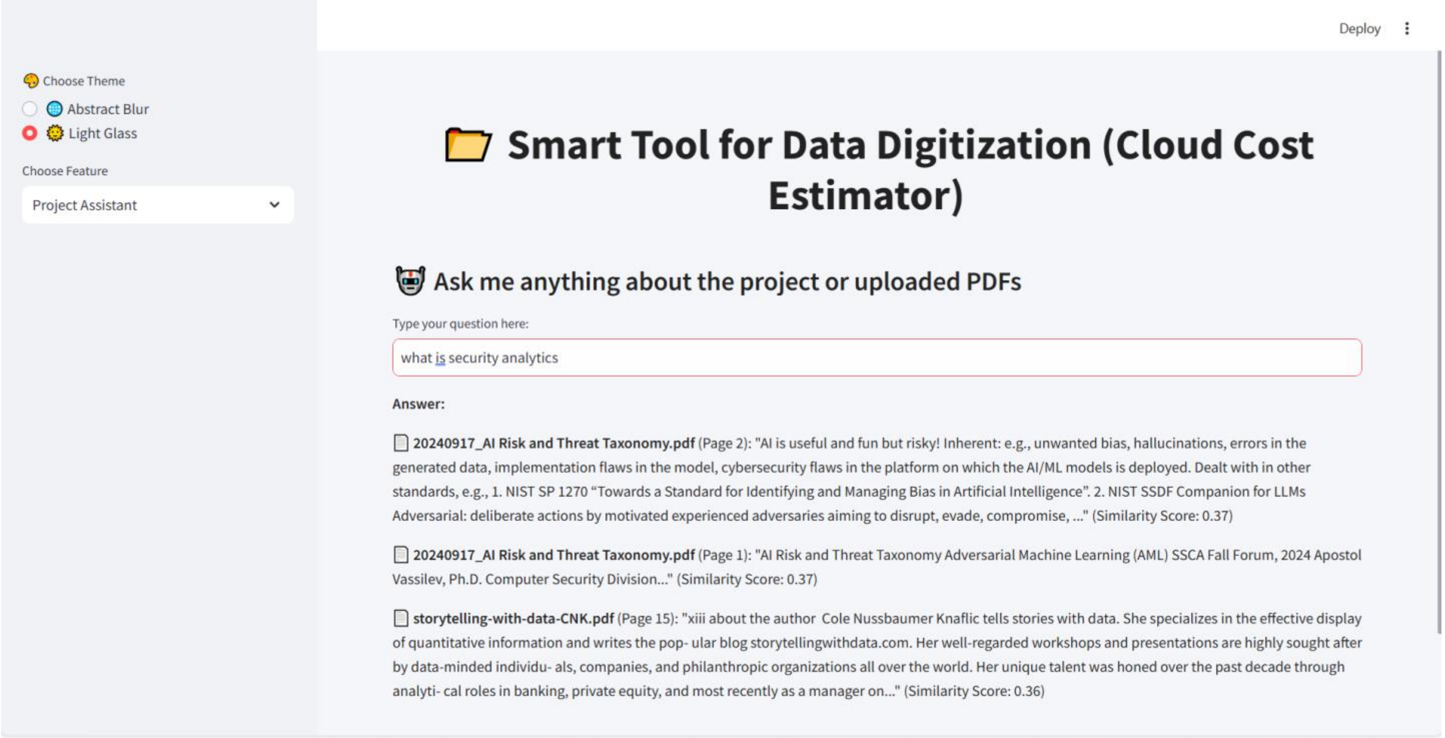
tell me about project features

Answer:

🖊 Key Features:

- PDF Upload and Metadata Extraction
- Manual Page Entry Option
- Live Cloud Storage Cost Estimation (AWS, GCP, Azure)
- Custom Pricing Options
- Cost Breakdown and Visual Dashboards
- PDF Report Generation
- Historical Estimate Tracking and Filtering
- Multi-Cloud Provider Cost Comparison
- Smart Recommendation Engine for Best Provider
- PDF Summarization with AI

Figure 18 - Project Assistant



Choose Theme

○ Abstract Blur

● Light Glass

Choose Feature

Project Assistant

📂 Smart Tool for Data Digitization (Cloud Cost Estimator)

Ask me anything about the project or uploaded PDFs

Type your question here:

what is security analytics

Answer:

📄 **20240917_AI Risk and Threat Taxonomy.pdf** (Page 2): "AI is useful and fun but risky! Inherent: e.g., unwanted bias, hallucinations, errors in the generated data, implementation flaws in the model, cybersecurity flaws in the platform on which the AI/ML models is deployed. Dealt with in other standards, e.g., 1. NIST SP 1270 “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence”. 2. NIST SSDF Companion for LLMs Adversarial: deliberate actions by motivated experienced adversaries aiming to disrupt, evade, compromise, ..." (Similarity Score: 0.37)

📄 **20240917_AI Risk and Threat Taxonomy.pdf** (Page 1): "AI Risk and Threat Taxonomy Adversarial Machine Learning (AML) SSCA Fall Forum, 2024 Apostol Vassilev, Ph.D. Computer Security Division..." (Similarity Score: 0.37)

📄 **storytelling-with-data-CNK.pdf** (Page 15): "xiii about the author Cole Nussbaumer Knaflic tells stories with data. She specializes in the effective display of quantitative information and writes the pop- ular blog storytellingwithdata.com. Her well-regarded workshops and presentations are highly sought after by data-minded individu- als, companies, and philanthropic organizations all over the world. Her unique talent was honed over the past decade through analyti- cal roles in banking, private equity, and most recently as a manager on..." (Similarity Score: 0.36)

Figure 19 - Semantic Search

13.2 Sample cost reports

PDF & CSV Formats

1. History Report

Timestamp	Pages	Size (GB)	Provider	Retention (mo)
2025-04-16 15:29:22	291	0.01	Google Cloud Storage	180
2025-04-16 15:31:13	115	0.0	Amazon S3	120
2025-04-16 15:32:26	11	0.0	Amazon S3	1
2025-04-21 14:59:55	6	0.0	Amazon S3	10
2025-04-22 13:21:48	5	0.0	Google Cloud Storage	10
2025-04-22 17:34:07	27	0.0	Amazon S3	10
2025-04-22 17:38:47	27	0.0	Amazon S3	10

2. Cost Report

Cost Component	Amount ($)	Provider	Timestamp
Storage	0.05	Google Cloud Storage	2025-04-16 15:29:22
OCR Processing	0.29	Google Cloud Storage	2025-04-16 15:29:22
Manpower	23.28	Google Cloud Storage	2025-04-16 15:29:22
Scanning	0.58	Google Cloud Storage	2025-04-16 15:29:22
Total Estimated	24.2	Google Cloud Storage	2025-04-16 15:29:22
Storage	0.0	Amazon S3	2025-04-16 15:31:13
OCR Processing	0.12	Amazon S3	2025-04-16 15:31:13
Manpower	3.45	Amazon S3	2025-04-16 15:31:13
Scanning	0.23	Amazon S3	2025-04-16 15:31:13
Total Estimated	3.8	Amazon S3	2025-04-16 15:31:13
Storage	0.0	Amazon S3	2025-04-16 15:32:26
OCR Processing	0.01	Amazon S3	2025-04-16 15:32:26
Manpower	0.33	Amazon S3	2025-04-16 15:32:26
Scanning	0.02	Amazon S3	2025-04-16 15:32:26

Total Estimated	0.36	Amazon S3	2025-04-16 15:32:26
Storage	0.0	Amazon S3	2025-04-21 14:59:55
OCR Processing	0.01	Amazon S3	2025-04-21 14:59:55
Manpower	0.3	Amazon S3	2025-04-21 14:59:55
Scanning	0.01	Amazon S3	2025-04-21 14:59:55
Total Estimated	0.32	Amazon S3	2025-04-21 14:59:55
Storage	0.0	Google Cloud Storage	2025-04-22 13:21:48
OCR Processing	0.01	Google Cloud Storage	2025-04-22 13:21:48
Manpower	0.15	Google Cloud Storage	2025-04-22 13:21:48
Scanning	0.01	Google Cloud Storage	2025-04-22 13:21:48
Total Estimated	0.17	Google Cloud Storage	2025-04-22 13:21:48
Storage	0.0	Amazon S3	2025-04-22 17:34:07
OCR Processing	0.03	Amazon S3	2025-04-22 17:34:07
Manpower	0.81	Amazon S3	2025-04-22 17:34:07
Scanning	0.05	Amazon S3	2025-04-22 17:34:07
Total Estimated	0.89	Amazon S3	2025-04-22 17:34:07
Storage	0.0	Amazon S3	2025-04-22 17:38:47
OCR Processing	0.03	Amazon S3	2025-04-22 17:38:47
Manpower	1.35	Amazon S3	2025-04-22 17:38:47
Scanning	0.05	Amazon S3	2025-04-22 17:38:47
Total Estimated	1.43	Amazon S3	2025-04-22 7:38:47

13.3 Glossary

Term	Definition
OCR (Optical Character Recognition)	Technology that converts scanned images of text into machine-readable, searchable data.
Cloud Storage	Online storage service that saves files remotely, accessed over the internet (e.g., AWS S3, GCP, Azure).
Streamlit	A Python-based open-source framework used to create interactive web applications with minimal code.
PyMuPDF (Fitz)	A Python library used for extracting metadata, text, and structure from PDF documents.

Altair	A declarative statistical visualization library in Python used for interactive charting in the application.
FPDF	A library used to generate PDF reports programmatically.
Session History	A log of previous estimations that includes pages, cost, provider, and timestamps for each session.
Retention Period	The duration (in months) for which digitized files are stored in the cloud.
Effort Level (Low/Medium/High)	User-selected scale to estimate manpower required for reviewing and correcting documents.
Custom Pricing	Feature that allows users to override default cost parameters such as OCR rate, storage cost, etc.
CSV (Comma-Separated Values)	A lightweight data format used to export reports that can be opened in spreadsheets like Excel.
PDF Report	A downloadable, presentation-ready format for cost breakdowns and historical logs.
Multi-Provider Comparison	A feature that estimates and compares total digitization costs across AWS, GCP, and Azure.
Smart Recommendation Engine	A system that identifies the most cost-effective cloud provider based on the current estimate.
Metadata	Descriptive information extracted from uploaded files, including title, author, and creation date.
Semantic Search	AI-based technique that finds relevant information by meaning, not just keyword matching.
LangChain	A framework used to integrate language models with document processing pipelines.
Mistral-7B	A large language model used in this application to summarize PDF documents in natural language.
Fallback Rate	Default pricing used when live cloud pricing APIs are unavailable.
Visualization Dashboard	A set of interactive charts that display cost trends, provider comparisons, and usage metrics.
Estimate Breakdown	A detailed summary of cost components such as OCR, scanning, storage, manpower, and licensing.
Session Filters	Tools that allow users to refine history logs by provider, page count, or cost range before exporting.
Chatbot Assistant	Built-in tool that answers user questions about the application using project knowledge and semantic models.