Project Proposal: AI-Powered Data Processing & Analysis Platform

Expected Date of Completion

February 15, 2025

Problem Statement

Manual data processing and analysis are time-consuming, error-prone, and often fail to capture the full potential of available information. Organizations struggle with disparate data sources, inefficient workflows, and a lack of robust analytical tools, hindering their ability to extract meaningful insights and make data-driven decisions. This project aims to develop a platform that addresses these key challenges by streamlining data workflows, automating key analytical tasks (including complex calculations, financial statement generation, and content summarization), and providing secure access to critical information through a unified, user-friendly interface. The platform will integrate diverse data sources (documents, communications, APIs, etc.), automate agreement management, and offer advanced features like enhanced search and content chunking, ultimately empowering organizations to optimize operations and improve decision-making.

Problem Overview

Organizations struggle to effectively leverage their data due to its increasing volume, variety, and distribution across disparate sources. Manual processing is inefficient and error-prone, failing to capture the full potential of available information. Existing solutions often lack seamless integration, robust analytical capabilities, and adequate security measures. This project addresses these challenges by developing a unified platform that streamlines data workflows, automates key analytical tasks, and provides secure access to critical insights, enabling organizations to make data-driven decisions and maximize the value of their information assets.

Proposed Solution

This project proposes a unified data platform to address the challenges of managing and analyzing diverse data sources. The platform will integrate various data types (documents, communications, APIs), automate key analytical tasks (calculations, financial statements, summarization), and provide secure access to insights via a private RAG space. Key features include automated agreement management, a user-friendly interface with secure login, and a scalable, modular architecture. This solution will streamline workflows, improve efficiency, and empower users to leverage data for informed decision-making.

Key Features

- **Seamless Integration:** Connects to diverse data sources, including documents (PDFs with robust tabular data parsing), communications (WhatsApp chats with emoji encoding), multimedia content (YouTube), financial records, and data accessed via APIs.
- **Automated Analysis & Reporting:** Automates complex calculations, generates financial statements (P&L, Sales, Balance Sheet) with download options (CSV, XML), and provides intelligent content summarization.
- **Secure RAG Space:** A private, secure Retrieval-Augmented Generation (RAG) environment, incorporates optimized OCR, enhanced content chunking, and hybrid search for accurate and efficient information retrieval. Access is restricted to authorized users.
- **Automated Agreement Management:** Manages agreements (NDAs and other types) with the flexibility to add new agreement types as needed.
- **Enhanced User Experience:** Features a user-friendly interface with dynamic heading modification, a separate registration/login modal (supporting Gmail and LinkedIn), and seamless Google Drive integration.
- **Robust and Scalable Architecture:** Built with a modular, scalable codebase optimized for future modifications and expansion, ensuring long-term adaptability.
- **Advanced Data Processing:** Employs efficient PDF parsing algorithms for complex tabular data and includes a dedicated WhatsApp chat module for processing uploaded text files.

Tech Stack

Authentication & Access Control

- OAuth 2.0 (Gmail, LinkedIn) for secure login.
- JWT (JSON Web Token) for session management.
- Role-Based Access Control (RBAC) for user permissions.

Frontend Development

- React.js for a modern, dynamic, and responsive UI.

Backend Development

- Node.js with Express.js for API handling and backend logic.

Database & Storage

- MongoDB for flexible, scalable data storage.

File Processing & API Integration

- Google Drive API for file attachments and management.
- YouTube API for extracting video metadata and transcriptions.
- WhatsApp Data Processing Module for chat file analysis.
- PDF.js + PyMuPDF for efficient PDF parsing.

AI-Powered Features

- Gemini API for document summarization and NLP.
- OCR (Tesseract.js) for text extraction from images/PDFs.
- Private RAG System for intelligent document retrieval.

Data Processing & Interpretation

- D3.js & Chart.js for graph and report visualization.
- Pandas (Python-based API calls) for financial data calculations.

Search & Optimization

- Hybrid Search (FAISS + ElasticSearch) for fast, factual content retrieval.
- Custom Chunking Algorithm to improve data structuring.

Security & Compliance

- Secure Registration/Login Modal for access control.

Deployment & Scalability

- Vercel, Render

Timeline with Milestones

S.No	Task	Details/Description	Deadline
1	Requirement Analysis & Project Scope Finalization	Gather all client requirements, clarify ambiguous points, and finalize the project scope.	Feb 4th
2	User Authentication (Gmail & LinkedIn Login)	Implement login functionality using OAuth (Google and LinkedIn) for seamless login. Ensure security and proper integration.	Feb 5th
3	Google Drive Integration	Enable linking of Google Drive for file attachment and management. Use Google API for integration.	Feb 5th
4	Data Fetching from APIs	Implement API integration to fetch external data based on user needs (consider integration with RESTful APIs).	Feb 6th
5	Data Processing (WhatsApp, YouTube, PDFs)	Create modules to process WhatsApp chat files, YouTube data, and PDF files. Focus on data parsing and text extraction.	Feb 6th

6	Data Interpretation (Graphs & Calculations)	Develop features to generate graphs and conduct complex calculations (such as financial reports). Ensure data is visualized accurately.	Feb 7th and 8th
7	Pre-calculated Financial Reports (PnL, Balance Sheets)	Implement automated generation of financial reports based on user data input. Add download functionality for CSV and XML formats.	Feb 9th
8	Document Management (NDAs, Agreements)	Automate handling of NDAs and contracts, allowing flexibility to add new agreements. Implement document upload and processing.	Feb 10th
9	Gemini Integration (for Summarization)	Integrate Gemini for text summarization and content extraction from uploaded documents and chat data.	Feb 11th
10	Chunking & Hybrid Search Implementation	Implement a new chunking technique for data cohesion and hybrid search for improved factual accuracy.	Feb 11th
11	WhatsApp Chat Module (Text & Emoji Encoding)	Develop a module to process WhatsApp chat files including emojis. Ensure proper text encoding and processing.	Feb 12th
12	Efficient PDF Parsing Algorithm	Implement the most efficient algorithm to parse PDFs, particularly for handling complex tabular data seamlessly.	Feb 12th
13	User Interface (UI) Design & Updates	Redesign/update the UI to ensure smooth user navigation, including dynamic heading modifications and user-friendly interaction.	Feb 13th
14	RAG Space Integration (Private Access)	Implement the secure Retrieval-Augmented Generation (RAG) space, ensuring that it's restricted to authorized users only.	Feb 13th
15	Registration/Login Modal (Secure Access)	Implement a secure, private registration and login modal with user access control (especially for the RAG space).	Feb 13th
16	Output Export (CSV & XML)	Develop the export functionality for generating and downloading reports in CSV and XML formats.	Feb 14th

17	System Testing & Bug Fixing	Perform end-to-end testing to ensure all functionalities work as expected. Address any bugs or issues.	Feb 15th
18	Documentation (System Architecture & Workflows)	Provide comprehensive documentation covering system architecture, workflows, and platform functionalities for future reference and understanding.	Feb 15th
19	Review & Final Adjustments	Review the platform with the client, gather feedback, and make any necessary adjustments.	Feb 15th

Deliverables

- Integrated user authentication (Gmail/LinkedIn) and Google Drive connectivity.
- Modules for processing diverse data (documents, chats, YouTube, APIs) and generating financial statements.
- A secure RAG space with advanced search and content chunking.
- Automated agreement management.
- User-friendly interface with dynamic headings.
- Scalable, modular codebase and comprehensive documentation.
- Dedicated WhatsApp chat processing module.

Prepared by: Mratyunjay Chouhan
Date: February 4, 2025