Market Study

**Advanced Intelligent Document Processing**

*1. Market Study*

* **Parabola**

Parabola is a no-code, AI-powered data automation platform that enables users—especially non-technical teams—to build and automate data workflows using a visual, drag-and-drop interface. It connects to sources like PDFs, APIs, spreadsheets, and emails, performs data transformations, and outputs results to platforms like Google Sheets, Slack, or internal systems.

### **Core Capabilities-**

* **Visual Workflow Builder** Easily build workflows using a drag-and-drop interface—no coding required.
* **AI-Driven Automation** Supports plain-English descriptions and NLP-powered workflow creation.  
   You can describe what you want to automate, and Parabola will suggest or generate a suitable workflow.
* **Documentation & Audit Trails** Parabola documents your workflow automatically, producing clear logic chains and artifacts for audits and reviews.
* **Workflow Templates** Reuse pre-built workflow templates to quickly address common inefficiencies like:
  + Data cleaning
  + Reporting
  + System integration  
     These templates help you move faster with minimal setup.
* **Data Anomaly Detection** Automatically flags outliers or inconsistencies in your data to keep you alert and accurate.

### **Data Ingestion & Integration-**

* **From Emails**
  + Receive and process attachments (CSV, PDF, XLS)
  + Extract subject and body content
* **Cloud Storage Support**
  + Google Drive
  + Dropbox
  + OneDrive
  + Excel files
* **Database & FTP Integration**
  + Pull/send data from:
    - FTP
    - MySQL
    - PostgreSQL
    - Microsoft SQL Server
    - Amazon Redshift
    - Snowflake
    - MongoDB
* **E-commerce & Social Media Sources**
  + Shopify
  + Amazon Seller Central
  + TikTok Analytics

### **Data Output & Sharing-**

* **Export Options**
  + CSV
  + Excel
  + Google Sheets
  + Google Drive
* **Distribution**
  + Send generated files via email
  + Push results to Slack and other collaboration tools
* **Data Visualization**
  + Display your results as styled tables or charts for quick insights
* **Data Storage**
  + Store and compile structured data over time in reusable tables

### **AI Features for Smart Automation**

* **Extract with AI** Automatically convert free-form descriptions into structured data.  
   E.g., “Tina is a cat” becomes a column Pet Name = Tina.
* **Standardize with AI** Clean and unify messy values.  
   E.g., values like s,Small,sm → all converted to Small.

### **Data Transformation Tools**

* **Row Operations**
  + Sort, filter, limit, remove or merge duplicates
  + Fill in missing values
* **Find and Replace**
  + Modify text, numbers, dates, and casing formats
* **Column Operations**
  + Edit, merge, split columns
  + Split names into first/last name
  + Add timestamps and row numbers
* **Math & Logic Columns**
  + Calculate running totals, growth rates, min/max
  + Use if/else conditions to create conditional logic
* **Table Operations**
  + Combine multiple tables
  + Stack rows from different datasets
  + Perform group operations: sum, average, min/max
* **JSON Parsing**
  + Expand JSON objects into structured, tabular formats

1. **Roles**

| **Role** | **Description** |
| --- | --- |
| **Owner** | - Has full administrative control over the workspace.  - Can manage billing, team settings, and integrations.  - Can invite, remove, and assign roles to users. |
| **Admin** | - Has access to all workflows in the workspace.  - Can edit, run, and delete any workflow.  - Can invite other users and manage integrations. |
| **Editor** | - Can create, edit, and run workflows they own or are shared with them.  - Cannot manage billing or workspace settings. |
| **Viewer** | - Has read-only access to assigned workflows.  - Can view results and exported data but cannot make edits. |
| **Guest** (if enabled) | - Typically used for external collaborators.  - Can view or comment on specific workflows but has limited or no edit access. |

### **Workspace Permissions and Sharing**

* **Workflow-Level Sharing**:  
   Individual workflows can be shared selectively with specific teammates or roles, ensuring data separation within large teams.
* **Integration Access Control**:  
   Connection credentials to services like Google Sheets, Slack, or databases can be shared or restricted per user or role.
* **Change Tracking & Audit Logs**:  
   All workflow changes are logged, and users can see who made edits, when changes occurred, and what was modified—supporting compliance and transparency.

### **Collaboration Features-**

* **Commenting & Documentation** Team members can leave comments and collaborate asynchronously within a workflow.
* **Auto-documentation** As workflows are built, Parabola generates clear, step-by-step artifacts for easy understanding across teams.
* **Versioning** (if available in your plan)  
   Helps teams revert to previous versions of workflows if needed.

### **Example Use Case:**

A data analyst (Editor) creates a reporting workflow; a team manager (Viewer) reviews the output weekly; and a workspace Admin manages connections to data sources like MySQL and Slack.

* **Nanonets:**

Automation of the entire flow of Intelligent Document Processing:

Ingestion, Processing, **Classification** (Receipts, Invoices, Tables, Passports, Drivers License, Bill of Lading and Custom Document), Extraction, Validation and Post processing.

1. **Effortless extraction**: Automatically import and extract data from different file types: .pdf, .jpg, .jpeg, .tiff
2. **Ingestion** can happen from **multiple sources** such as: Email, Google Drive, third-party applications, and even physical documents that require scanning.
3. For the **Classification** of Uploaded Document, Users can select the available Classes and also have the option to create a new class.

This class is called Custom Document and at least 10 files need to be uploaded for the model training.

1. **Smart processing**: With AI, Nanonets can tackle even the most complex documents, whether in different layouts, languages, or currencies.
2. Interprets handwriting
3. **Advanced OCR**
4. Object detection,image classification, multi label classification
5. Supports different languages: English, Spanish, Arabic etc
6. **Real-Time Analysis**:

Tracking the status of the uploaded document. Dashboards that have key metrics are also visible.

1. **Continuous improvement**: The AI learns from your data and improves over time. This means its performance improves with each interaction, helping you continually improve your document processing.
2. Live assistant on how to use nanonets
3. **Customizable Workflows**-User can add documents at any stage,before getting the final result they can give conditions(if,find and replace,format date), they can assign reviewers making their review mandatory or optional and define rules for flagged files
4. Admins could potentially get an Email for documents which have low confidence after extraction

* **Docsumo:**

Helps Automate the entire document workflow

1. Has the feature to allow user to either **enter the classification** allowed or let the system classify the document for further processing
2. The document **ingestion can happen from different sources**:

Google drive, Google Sheets, Microsoft OneDrive, Dropbox

1. **Smart filters** to filter on the uploaded documents
2. All the uploaded documents of the same classification are displayed in a structured table with columns and rows
3. **Integrated with Google Sheets**. The uploaded documents can be viewed in a consolidated manner in Google Sheets.

(Eg: For the classification Bank Statement, the invoices, Invoice number, Credited Amount, etc are displayed in a tabular form)

1. User friendly UI

* **Abbyy:**

1. Upload files by scanning QR through mobile
2. Create your own ocr by providing language,export document type(pdf,txt,xlsx,json,docx),bar codes
3. Ability to connect to same excel sheet while exporting multiple documents(invoices)
4. Abbyy has classification skill(to decide the type of document-receipt,passport,invoice..),document skill(to extract data from the uploaded document),process skill(workflow orchestration layer defines how documents flow through the system)
5. Targets:
   * Google sheets, Google translation API
   * Anthropic Claude- Connects ABBYY Vantage's IDP capabilities with Claude AI for enhanced data analysis like fraud detection
   * Outlook

* **Rossum:**

Ability to handle if uploaded document is rotated and if same document has more invoices

* User actions:
  + User can define confidence threshold
  + While rejecting a file, user sends an email to admin in which he can use predefined templates defining the reason for rejection
  + User can define fields to be captured for particular classification(invoice,passport,receipt etc)
  + User can decide automation with options-automate(no human review),never(human review for every file),confidence(human review if less than defined threshold)
* Results display:
  + On clicking the file user can see key value pairs
  + Confidence score for every result (key-value pair) is provided
  + Different statistics are provided-time per doc,ontime vs late documents,document turnaround time for overall files
* Source: Email
* **AgentOps**

Build, monitor, debug, and deploy AI agents and LLM applications. Provides observability into the agent lifecycle.

1. **Visual Event Tracking** and Session Replays: LLM calls, tools invocations, multi-agent dialogues.
2. **Error Monitoring** and Auditing: Full data trail, logs errors and any failures.
3. Cost and Token Usage: **Analytics and dashboard** for optimising the LLM usage.
4. **Various Framework** integration: CrewAI, LangChain, OpenAI, Anthropic, etc. Also supports Python and TypeScript.

* **Bedrock Data Automation:**

1. **User-Friendly Interface**: Amazon Bedrock offers an easy-to-use interface that allows you to define output schemas and set precise business rules with minimal effort.
2. **AI Integration**: confidence scoring, and toxic content detection, ensuring ethical data processing and analysis.
3. Seamless Amazon **Bedrock Knowledge Base Integration** for querying upon uploaded documents.
4. **Scalability** with a Single Interface API: Its single inference API enables efficient handling of production-scale workloads, ensuring smooth performance as your business scales.

Functional and non-Functional Requirements

## **1. Functional Requirements**

### **1.1 Document Ingestion and Organization**

**1.1.1 Uploading Documents** (Priority: High)

* Users can upload documents through multiple channels such as direct upload (via UI), email ingestion, or integration with cloud storage (e.g., Google Drive).
* Supported file formats: .pdf, .csv, .png, .jpg, .tiff, .txt.

**1.1.2 Creation of DocVault** (Priority: High)

* Users must create a DocVault to upload documents.
* For email ingestion, the system prompts users to select the relevant DocVault destination.

**1.1.3 Tagging** (Priority: Low)

* Users can manually assign tags to documents or rely on auto-tagging based on content analysis.
* Users can search for documents using tags (e.g., project name, department) across all DocVaults they have created.
* This allows for efficient retrieval of related documents regardless of their storage location, improving cross-vault visibility and organization.

**1.1.4 Search and Filtering** (Priority: High)

* Searching and filtering the documents present in the DocVault after the extraction has been performed.
* Search and filter operations span across all DocVaults created by the user.

Eg: filter on a particular classification(Invoice) across all DocVaults created by the user.

* Users can search for a DocVault created by them.

**1.1.5 DocVault Sharing** (Priority: Medium)

* Users can share their DocVaults with other registered users by specifying email addresses and access levels (e.g., view, edit, or admin).
* The system will store and manage sharing permissions, allowing collaborators to securely view or modify the contents based on their assigned role.
* Shared users will see the DocVault in a separate "Shared With Me" section. The document access and actions will be governed by the permission granted.

### **1.2 Classification and Grouping**

**1.2.1 Custom Classification Creation** (Priority: High)

* Users can define custom classifications beyond predefined ones.
* If auto-classification classifies a document type with Low Confidence, the system will prompt the user to create a new classification.
* Classifications we are providing:
  + Passport, Driver’s License, Bill of Lading, Invoice, Receipt, Bank Statements.

**1.2.2 Adding extra fields to the Predefined Classifications** (Priority: High)

* Users can add additional fields to the already existing classification before extraction.
* The available fields in the predefined classification are displayed in the UI of the system.
* Eg: Adding Restaurant Rating to the Receipt Classification.

**1.2.3 Data Analysis and Aggregation** (Priority: Low)

* Users can upload multiple documents of the same type (e.g., invoices, receipts).
* The system automatically groups and consolidates extracted data into a unified summary for trend analysis and metric tracking.

**1.2.4 AI-based Document Segmentation** (Priority: Medium)

* The system uses layout and semantic analysis to extract and separate meaningful sections (e.g.,getting terms,conditions,signatures,other fields in a legal document) from complex documents.

**1.2.5 Prompt-Based Queries and Automation** (Priority: Medium)

* Users can type natural language prompts like: "Extract payment summary from all invoices in June".
* The system interprets the request, identifies relevant documents and fields across all DocVaults created by the user, and executes the appropriate workflow without manual configuration.

### **1.3 Data Extraction, Validation, and Transformation**

**1.3.1 Filter on DocVault** (Priority: High)

* Users can define value-based filters on multiple files in a DocVault

Eg: apply a filter to show invoices that have Total Transaction above

**1.3.2 AI-Powered Standardization** (Priority: Low)

* The system detects and suggests normalization for inconsistent values of extracted fields of a document(e.g., "sm", "S", and "Small" -> "Small").
* Users can review and approve which standardizations to apply.

**1.3.3 Data Transformation and Structuring** (Priority: Low)

* A full set of no-code transformation tools is provided:
  + **Row operations**: sort, filter, limit, merge, or deduplicate rows
  + **Find and Replace**: clean or reformat values
  + **Column operations**: merge/split columns, add timestamps, generate row numbers
  + **Math & Logic Columns**: perform aggregations, growth calculations, and logic-based transformations

**1.3.4 Fraud Detection** (Priority: Medium)

* For identity documents (passports, driver’s licenses), the system uses an AI-based fraud detection agent to verify authenticity.

**1.3.5 Human-In-The-Loop (HITL)** (Priority: High)

* Fields below a confidence threshold are flagged for manual review.
* Users can accept, reject, or correct values via an interactive interface.

**1.3.6 Document summarization** (Priority: Medium)

* Summarization of the document after extraction to get a high level idea of the information present in it.

**1.3.7 Document translation** (Priority: Low)

* Document can be uploaded in any of the horizontally aligned languages like German,French,Spanish,English etc
* User will get the results in uploaded document language and he has the option to translate results to other languages

### **1.4 Output and Integration**

**1.4.1 Exporting Result** (Priority: High)

* Users can export processed and structured data to:
  + Google Sheets
  + Google Drive
  + Amazon S3
  + Email
  + (Planned) Amazon DynamoDB and RDS

**1.4.2 Notification Preferences** (Priority: Medium)

* Users can choose to receive email notifications upon successful extraction or if a document fails to process.
* Notification settings can be configured at the DocVault level, enabling personalized workflow alerts.

**1.4.3 Visualizing Extracted Information** (Priority: Low)

* The system will provide users with the ability to view the extracted information in charts, graphs, and tables for better understanding.
* Users will be able to explore summaries, comparisons, and trends based on their uploaded documents.
* These visuals can be used for easier communication and decision-making.

### **1.5 Workflows**

**1.5.1 Agentic Workflows** (Priority: Low)

* Ask the agent to automate tasks requiring time and manual effort. Eg,. Create a DocVault with name <name> and ingest data from emails with email-id: <email address> which contains the word INVOICE. Also give me monthly and quarterly analytical reports on these invoices.

### **1.6 Admin Capabilities**

**1.6.1 Analytics and Monitoring** (Priority: Medium)

* Admins and users can access analytics on:
  + Number of documents processed
  + Accuracy of extractions
  + Workflow efficiency metrics

## **2. Non-Functional Requirements**

### **2.1 Performance**

**2.1.1 Scalability**

* The platform shall scale seamlessly based on user demand, ensuring consistent performance regardless of workload size.

### **2.2 Usability**

**2.2.1 User Interface**

* The system must feature an intuitive, user-friendly interface requiring **no formal training** for basic operations.

### **2.3 Security**

**2.3.1 Data Protection**

* All data must be **encrypted in transit and at rest**.
* The platform must implement input validation to **prevent injection attacks** and ensure data integrity.

### **2.4 Maintainability**

**2.4.1 Documentation**

* The system shall provide:
  + **API documentation** for developers
  + **Component-level documentation** for maintainability and onboarding