Satva –Document data extraction & analysis as a service: a SaaS platform
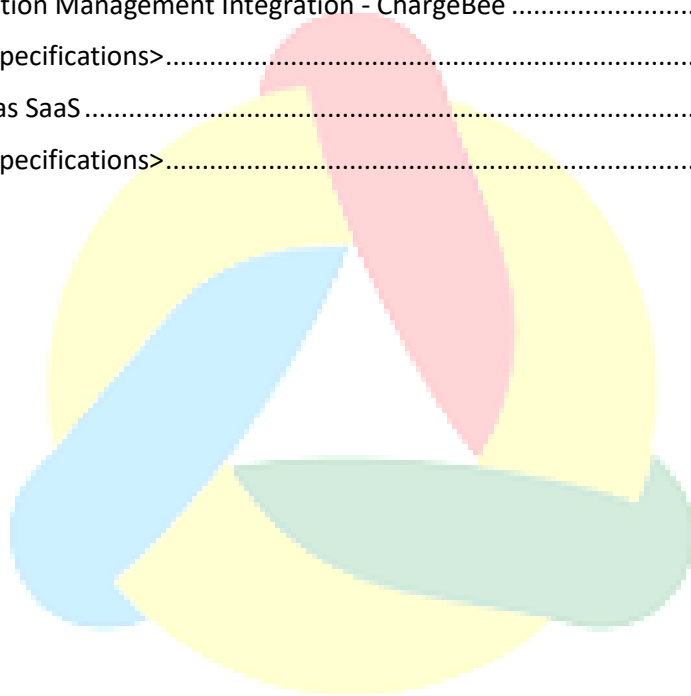
Product Specification document

Version 1.0

AARUNI

TECHNOLOGY • CHANGE • RESULTS

TABLE OF CONTENTS

1 OVERVIEW

Looking to extract data from financial documents?

Need a robust invoice scanner to power cognitive capture on invoices? Click below to check out Nanonets' automated invoice processing solution to supercharge your AP workflows!

Cognitive data capture has changed the way data is available to make business processes smooth and effective. Using a combination of techniques to recognize, extract and classify data has potentially paved the way for error-free maintenance of data that spurs expertise and efficient decision-making. It is wise to acknowledge that cognitive data capture solutions open a multitude of possibilities for business management and have become the need of the hour.

2 WHAT IS COGNITIVE CAPTURE?

Data capture is the process of extracting relevant information from both structured or unstructured documents and transforming it into machine-readable data. Artificial intelligence has transformed the way data capture is being done today. Cognitive capture is an advanced stage of machine learning and cognitive data is now pivotal for many successful business processes.

3 COGNITIVE DATA CAPTURE : DEFINITION

The term 'Cognitive" refers to the process used to recognize, process, store, or retrieve information to make informed decisions. In terms of artificial intelligence, it relates more to machine learning and the capability to make cognitive data instantaneous to improve business process efficiency.

Machine learning has been in existence for several decades and has constantly evolved by leaps and bounds. Cognitive capture is the height of innovation and uses what can be termed as 'intelligent processing'.

Cognitive data capture systems use machine learning by mimicking human capabilities. This technology works in the same way as humans do. It recognizes, extracts, classifies, and stores both structured and unstructured data, without the use of templates or any other rules. Deploying OCR techniques and natural language processing (NLP), it can look for words, phrases, or patterns and sorts data contextually and accurately.

Cognitive data keeps complex data simple and easy to use as document processing is not bound by types, fonts, or formats. Huge volumes of data whether structured data such as names, addresses,

numbers or unstructured data such as audio, video, images, emails can be handled smoothly to make data access logical and simple.

4 TYPES OF DATA CAPTURE

1. Manual Data Capture: One of the initial and popular methods used to collect data, manual data capturing is used in many businesses today. It refers to the method of collecting data from physical forms by manually entering them into software for access. This method is tedious, labor-intensive, time-consuming, and less efficient, which is why businesses are fast transitioning to using automated data capture.

2. Automated Data Capture: The current trend of using computerized technology to automatically convert any data from forms, emails, etc into readable digital data enables instant access anywhere, anytime. Businesses nowadays rely on using automated data capture to efficiently manage data requirements, reduce project and manpower costs, and enhance productivity. Many different forms of automated data capture exist, some of which are detailed in this section.

3. OCR: Optical Character Recognition is a technology used by machines to recognize and read specific typefaces on printed documents including PDF files, scanned forms, etc., and digitize the data for use. OCR is commonly used in verticals such as insurance, finance, healthcare, and more, where there are high volumes of documentation involved and the information is very similar.

4. ICR: Intelligent Character recognition is an advanced version of OCR. This technology can read hand-printed characters written in different fonts and styles that people use and process them digitally for relevant use. ICR is used extensively by banks and financial institutions for automating customer information into their software.

5. Barcodes: Barcodes are encoded data made of black bars and white spaces in varying widths that are read by a scanner. This technology is efficiently used to classify inventory and collect data automatically, which speeds up the process to save time and effort. It is commonly used in most retail outlets and healthcare.

6. Magnetic Stripe Cards: Acting on the principle of automated data transfer, magnetic stripes contain encrypted data that are decoded by reader devices to receive information wirelessly. Credit/ debit cards, access cards are classic examples of magnetic swipe card technology.

7. Voice and Image Capture: If you are using Alexa or Siri, you are probably familiar with the voice capture technology. It uses speech recognition technology where signals are processed to interpret speech.

8. Biometric security systems use image capture for identifying relevant data. This kind of data capture uses technology to convert images or sounds into digital files that are cross-verified for authorization into secured systems. You can find biometrics being used at banks, airports, workplaces, and on mobiles for granting access.

5 BENEFITS OF COGNITIVE DATA CAPTURE

Cognitive data can optimize business workflows and ensure the smooth functioning of business processes. Here are some of the main reasons why using cognitive capture can be ideal for businesses.

1. Lowered costs

Cognitive data solutions can help in reducing business costs and yet maintain accurate data. By eliminating the need for manual labor and optimizing data capturing capabilities, it can lower operational costs efficiently.

2. Compatibility

Cognitive data capture intelligently processes any document. The technology is compatible across multiple document types and can automatically recognize and extract data from PDFs, forms, invoices, emails, etc, to streamline business processes and enhance productivity.

3. Speedy processes

The use of NLP combined with OCR techniques makes cognitive data capture faster and easier to use. With no set rules or templates, capturing data is faster and more efficient. Such technologies also facilitate instant data access to ensure the smooth flow of business processes.

4. Accuracy

Cognitive data capture is highly accurate and error-free. It can analyze an unlimited amount of data, be it structured or unstructured, and automatically extract specific information required for successful business transactions.

6 COGNITIVE DATA CAPTURE USE CASES

Cognitive data capture can be greatly beneficial to various industries. Below are some of the use case scenarios where cognitive data can be effective.

6.1 HEALTHCARE –

Maintaining detailed and diverse medical data in patient records. For example, every patient's data is collected from various sources such as the registration desk, doctor consultations, lab investigations, pharmacy, other referred departments, radiology, etc. This means a lot of information is being extracted and recorded for use. This process applies to numerous patient data across various departments. Cognitive capture technology offers a wholesome solution to streamline data collection for easy availability and use resulting in effective patient management.

6.2 INSURANCE –

Preventing frauds in claims processing. What this means is cognitive data capture can make the insurance claim process effortless and help insurance agents make conclusive decisions. AI-powered intelligent processing makes it possible to have hassle-free form-filling. Reviewing and verifying claims becomes easier as digital images and data extraction can be carried out on various documents yet making it relevant and accurate to the smallest detail. Billing errors are detected easily and cross-verification before a settlement can be automated to prevent fraud or false claims.

6.3 FINANCE/ ACCOUNTING –

 Help with multiple invoice processing in the Accounts Payable (AP) department. To elaborate, *AP departments in large organizations may receive invoices from multiple suppliers in the form of PDFs, emails, or as hard copies. Classifying and extracting data can become an error-prone and humongous task. Using cognitive data capture technology can help in automating invoice capture/processing and simplifying the entire process.* Extracting key information from any kind of document, cross-checking with appropriate Goods Received Note (GRN), delivery notes, classifying invoices for approvals all help in faster and error-free invoice processing.

6.4 BANKING –

Maintain customer records and help with individual loan processing. For example, banks handle thousands of customers who have different requirements like opening accounts, applying for loans, cash/cheque/ wireless transactions, and many more. Banks use a lot of data from different forms starting from a Know-your-customer (KYC) form, loan applications, challans, etc. that can be manually time-consuming and exhaustive. Cognitive capture can minimize time, effort, and labor to make these varied processes easier. By recording, classifying, and maintaining individual records of customers, it can provide instant access to banks to process applications, verify and sanction loans, update and transfer amounts between accounts ensuring smooth flow of operations.
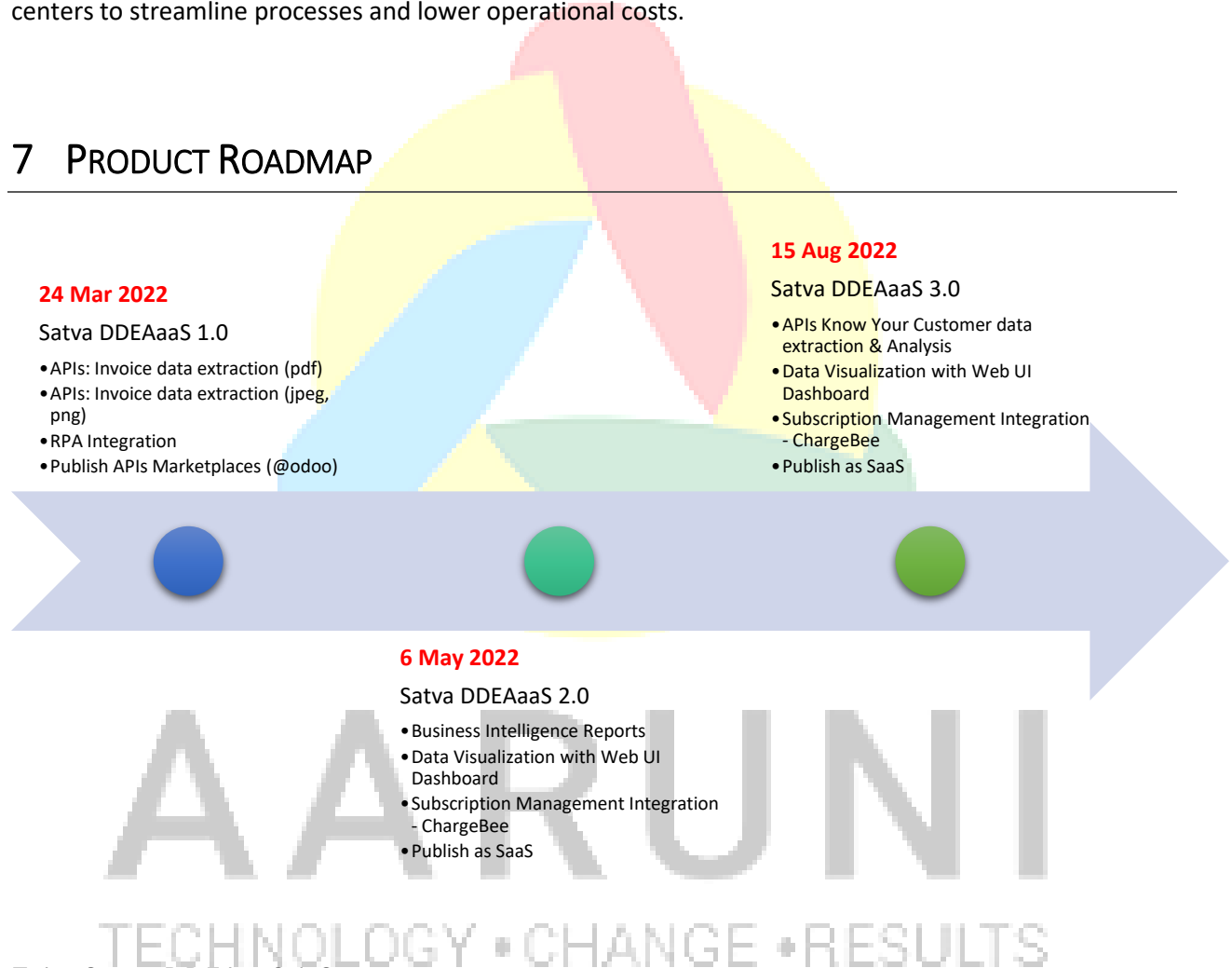
6.5 LOGISTICS AND TRANSPORTATION –

Logistics involves dealing with a vast amount of unstructured data such as billings, third-party vendor information, location, dates, and numerous other details. Ai-based cognitive capture can deal with these massive data to extract vital information needed to complete a particular transaction. *In transportation,*

it can provide access to shipping data, help in optimization of routes to lower shipping costs, generate corresponding information for validating invoices, and track delivery completion.

6.6 EDUCATION:

The education sector is huge and uses a huge amount of data relating to various institutions, universities, subjects, students in its processes. Cognitive capture can help automate document-driven processes and make them simpler to use. For example, it can help institutions manage student information, maintain reports, prepare or validate student transcripts and centralize the data across centers to streamline processes and lower operational costs.

7 PRODUCT ROADMAP

24 Mar 2022

Satva DDEAaaS 1.0

- APIs: Invoice data extraction (pdf)
- APIs: Invoice data extraction (jpeg, png)
- RPA Integration
- Publish APIs Marketplaces (@odoo)

15 Aug 2022

Satva DDEAaaS 3.0

- APIs Know Your Customer data extraction & Analysis
- Data Visualization with Web UI Dashboard
- Subscription Management Integration - ChargeBee
- Publish as SaaS

6 May 2022

Satva DDEAaaS 2.0

- Business Intelligence Reports
- Data Visualization with Web UI Dashboard
- Subscription Management Integration - ChargeBee
- Publish as SaaS

7.1 SATVA DDEAAAS 1.0

Satva API extracts, categorizes, and enriches all the details from unstructured consumer purchase receipts, invoices, and bills down to line items (SKU-level purchase data) at scale (in large volumes), without templates or humans-in-the-loop.

<Detailed UI/API specifications>
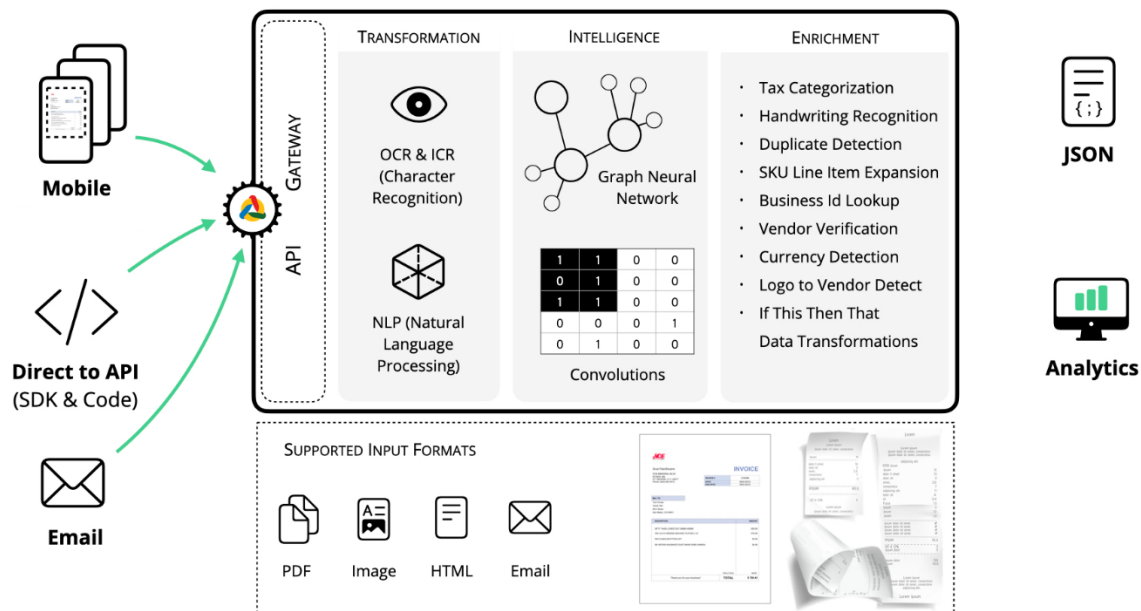
7.1.1 APIs: Invoice data extraction

1. API Gateway with TLS Certificate
   a. Access micro services through private IP address
   b. Protect from DDoS attacks and malwares, Keys, Tokens, IP filtering
   c. Authentication, authorization & fault tolerance
   d. Load balancing & routing (based on headers, paths, and params)
   e. Response Cache (with mention a URL and threshold time to cache the responses): Hold of all the static contents and can directly serve the client.
   f. Protocol adopter to benefit web socket or HTTP/2 (new version of http)
2. Transformation
   a. Convert an Image or a PDF to text (html, email, txt files) using OCR (optical character recognition). Text inside the image file can be either typed font or handwritten.
   b. Extracting fields from the text. Some invoices can contain over 50 different fields that need to be extracted. Also each product or service line on the invoice can be structured using 5-10 columns.
3. Intelligence
   a. Generate insights from extracted data using neural network (CNN/RCNN/GNN) and Convolutions.

   <Elaborate description>

4. Enrichment
   a. Data Enrichment and Fraud Detection.
      i. Tax categorization
      ii. Handwriting recognition

      iii.   Deduplication
      iv.   SKU Line item expansion
       v.   Business Id Lookup
      vi.   Logo to vendor detection
    vii.   Vendor Verification
   viii.   Currency detection
      ix.   Date format detection
       x.   Locale detection for geolocation/geofencing
      xi.   Data transformation (if then that)

7.2 STEPS

1. Detect target folder
2. Verify
   a. No of Folders
   b. No of files
   c. File extensions
   d. File sizes
   e. No of documents in each file
   f. Detect duplicate copies of the same document
3. **Document layout analysis (e.g. OCRopus, OCRFeeder)**
   a. Index documents by their structure or pictorial content.
   b. Identify and categorize the regions of interests
      i. Segmentation of text zones from non-textual ones
      ii. Arrangement in correct reading order
   c. **Geometric layout analysis**
      i. Detection and labeling of the different blocks as
         1. Text body,
         2. Illustrations,
         3. Math symbols
         4. Tables
   d. **Logical layout analysis**
      i. Detection and labelling of text zones
         1. Titles
         2. Captions
         3. Footnotes
4. OCR Analysis

7.3 METHODOLOGIES

Any of the following approach has issue. Noise and skew. Noise refers to image noise, such as salt and pepper noise or Gaussian noise.

1. **Bottom-up approach**
   a. Iteratively parse a document based on the raw pixel data

  i. Parse a document into connected regions of black and white

  ii. Group into

    1. Words,

    2. Text lines

    3. Finally into text blocks

 b. the advantage is to parse the global structure of a document directly

2. **Top-down** approach

 a. Iteratively cut up a document into columns and blocks based on white space and geometric information

 b. Require iterative segmentation and clustering and is slow process

7.3.1 RPA Integration

<Detailed UI/API specifications>

7.3.2 Publish APIs Marketplaces (@odoo)

<Detailed UI/API specifications>

7.4 Satva DDEAaaS 2.0

7.4.1 Business Intelligence Reports

<Detailed UI/API specifications>

7.4.2 Data Visualization with Web UI Dashboard

<Detailed UI/API specifications>

7.4.3 Subscription Management Integration – ChargeBee

<Detailed UI/API specifications>

7.4.4 Publish as SaaS

<Detailed UI/API specifications>

7.5 Satva DDEAaaS 3.0

7.5.1 APIs KYC (Know Your Customer) data extraction & Analysis

<Detailed UI/API specifications>

7.5.2 Data Visualization with Web UI Dashboard

<Detailed UI/API specifications>

7.5.3 Subscription Management Integration - ChargeBee

<Detailed UI/API specifications>

7.5.4 Publish as SaaS

<Detailed UI/API specifications>

8 Technology Stack

1. API : Python, Django/Flask
2. Mobile app: Android
3. Web App : Reac.js, Material UI
4. API Gateway