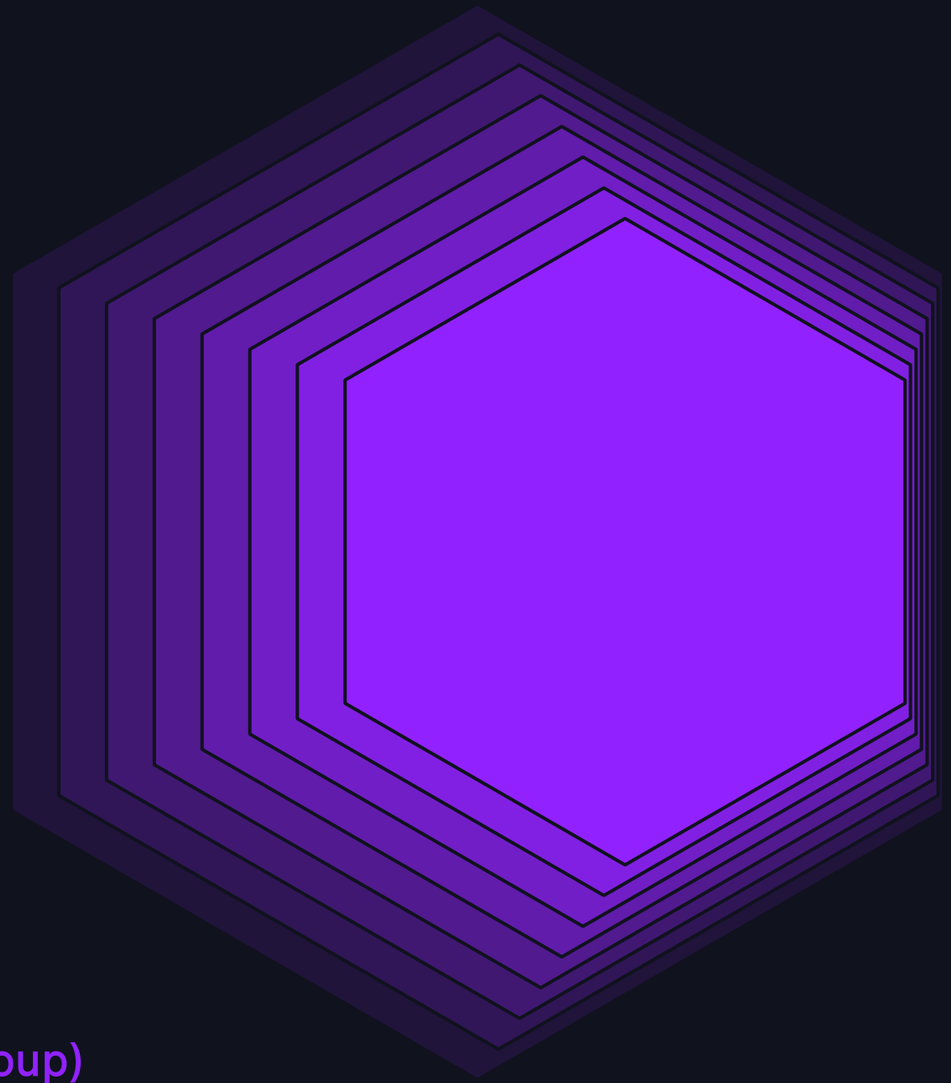
DATA+AI SUMMIT BY databricks

Augmenting Generative AI for complex Unstructured Data

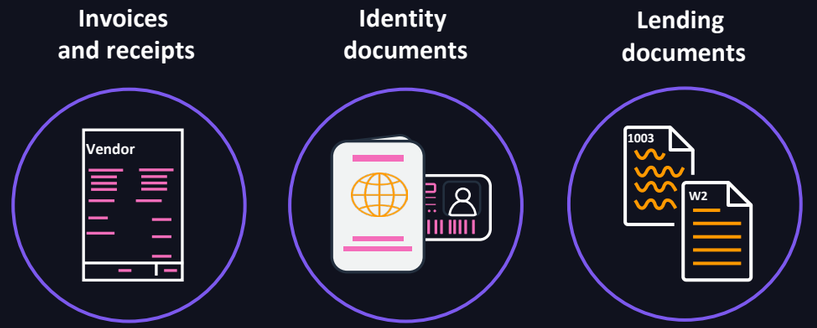


Venkat Viswanathan (AWS) and Mayur Rajdev (The Baldwin Group)
June 11th 2024

Complex Unstructured Data

Complex PDFs

Forms



Specialized documents

Handwriting

Tables

Signature

Industry Challenges when it comes to Complex Document Processing…

Manual Extraction	Time to Market	Classification
Scalability	Multi-Lingual Support	PII/PHI Security
Expensive	Decreased Employee Productivity	
Inconsistent Output	Lack of Automation	Continuous Training

Language is not a barrier to run your business any more

Stages of Complex Document Processing

Data Capture

Aggregating and organizing documents from different business workflow(s) within your organization

Classification

If more than one document type then classify each document and send to the appropriate document pipeline.

Extraction

Extract key business information

Enrichment

Getting insights and business value from your data

Review and Validation

Run business rules on your data and/or include human in the loop validation as needed.

Ready for Business Decisions

Send information to downstream apps or databases

Business decision activities: Comparing documents with baseline, identifying gaps, and creating insights

Key customer challenges

Data Capture

- Securely capture data
- Auto-correct for quality defects – distortion, dirt, rotated text, etc.

Classification

- Accurate identification and separation of documents where content across loan jackets may be highly variable

Extraction

- Capture data structures (tables, key-value pairs, entities, implied elements)
- Nested Data (page sections, duplicate fields)
- Data variations (SSN vs. Social Security #)

Enrichment

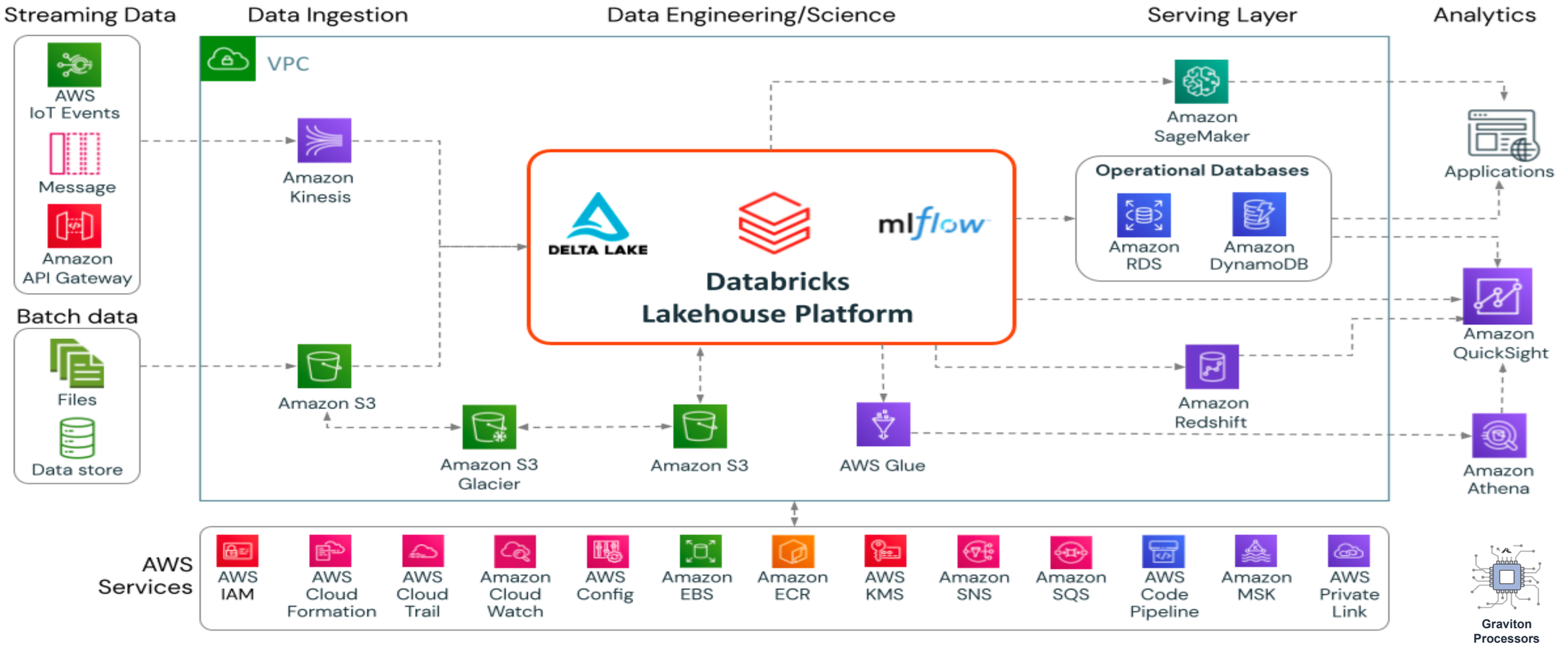
- Accurately identify PII (name, date of birth, addresses, etc.) out-of-the-box
- Flexibility to train custom models when necessary.

Validation

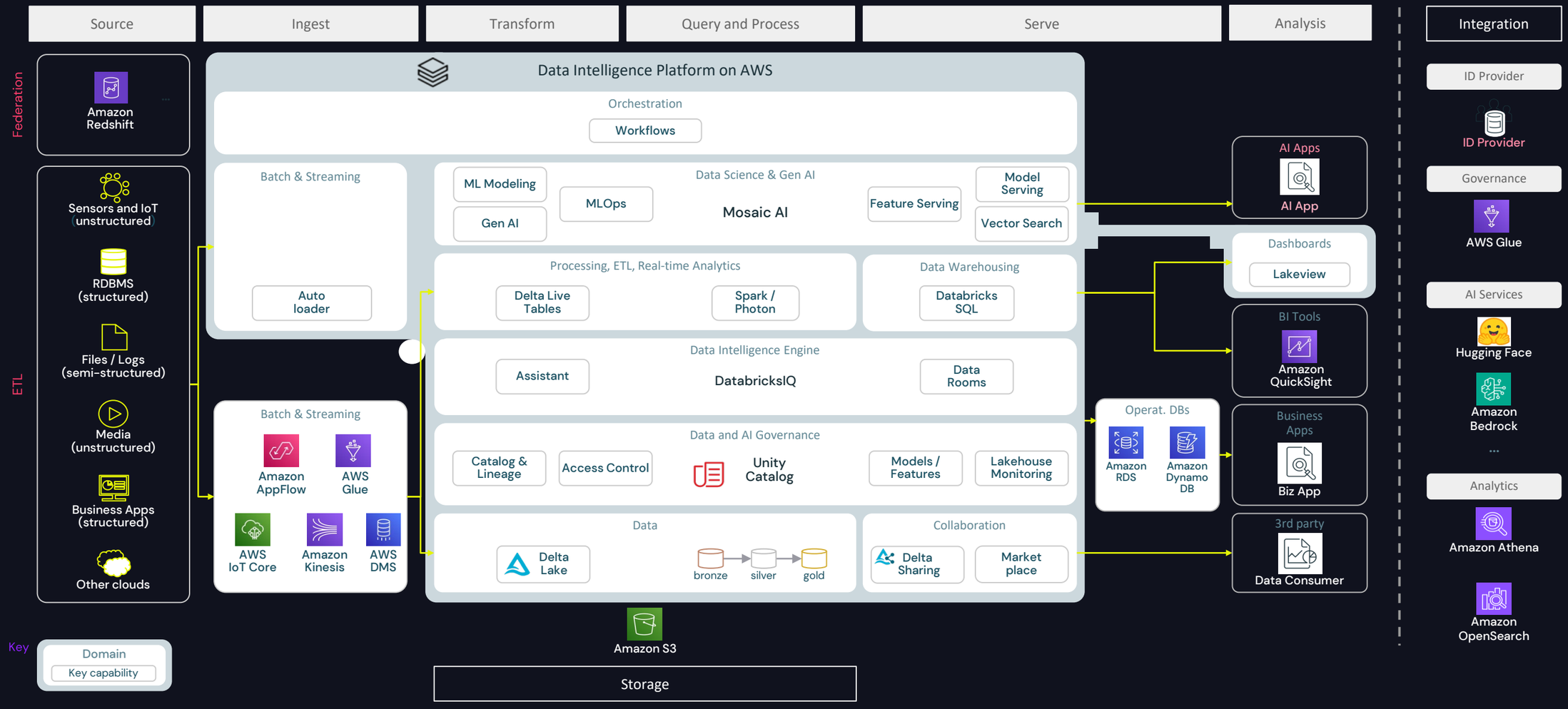
- Securely onboard your human review workforce to validate ML output
- Customize routing based on business rules, ML output metadata (e.g. confidence scores)

How AWS and Databricks together can help you speed up business processes, improve decision quality, and reduce overall costs with intelligent document processing (IDP)

Databricks on AWS: High-level Architecture



Databricks Data Intelligence Platform on AWS



Data Capture - Amazon S3

Aggregate and store documents securely

Scanned documents
PDF, JPEG, PNG, and TIFF

Captured by
cell phone cameras or digital scanners
Distortions, crumpled pages, rotated text, Ambient noise (dirt) and more

Files in common formats such as Word or .TXT

Automate document processing with Amazon Textract

Documents stored in Amazon S3

Amazon Textract

High volume OCR for digital or scanned documents in PDF, TIF, JPG, and PNG formats

Foundation model

Train/Tune an LLM using your documents

Amazon Textract Response Objects

Layout Aware

- Main Title as LAYOUT_TITLE
- Text in the top margin of the document as LAYOUT_HEADER
- Text in the bottom margin of the document as LAYOUT_FOOTER
- Titles below the main title that represent sections in the document. Returned as LAYOUT_SECTION_HEADER
- Page number of the documents. as LAYOUT_PAGE_NUMBER
- Any information grouped together in list form as LAYOUT_LIST
- Location of an image in a document as LAYOUT_FIGURE
- Location of a table in the document as LAYOUT_TABLE
- Location of form key-value pairs in a document as LAYOUT_KEY_VALUE
- Text that is present typically as a part of paragraphs in documents as LAYOUT_TEXT

Extract data

Signature Detection

Detects Signatures in a document (single page or multi page)

Captures page number where signature is present

Also outputs confidence scores and geometry info

Blocks:

PAGE, SIGNATURE

Example Output:

Signature Page 1 Accuracy 95%

Natural Language Query

Simple Natural Language Q & A Interface

Extract specific information of interest from documents

Designed to work with just about any type of document, no training needed

Extract key entities from documents using natural language questions

Classify Document – Amazon Comprehend

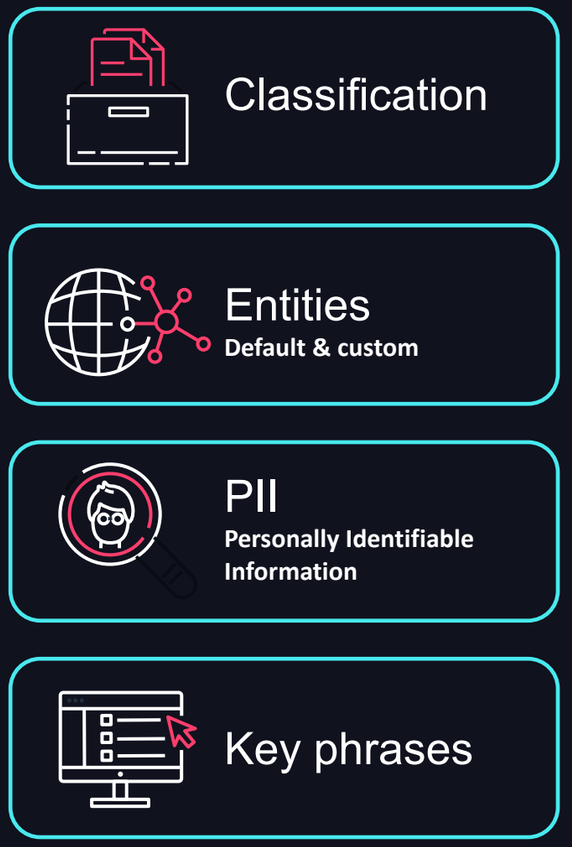
Identify document types

Train a custom classifier to organize and identify different types of documents (e.g. W-2 paystubs, 1003, 1040, identity documents and more) within a loan jacket to inform downstream routing and application of custom business processing rules.

Training a custom classifier is as easy as providing 50 samples of each document type.

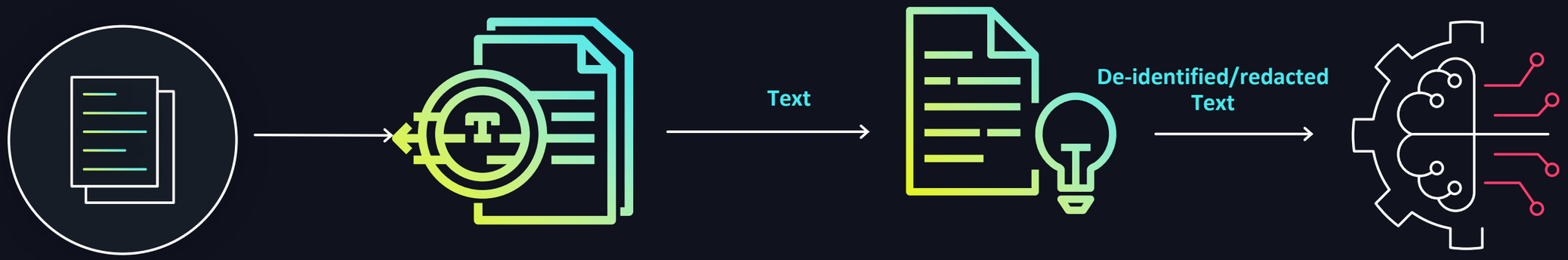
Real-time or asynchronous processing

Perform real time document classification when needed or reduce costs using asynchronous processes.



Sensitive data handling for generative AI use-cases

Redacting inputs can improve sensitive data handling with Generative AI models.



Redaction of sensitive data

DOCUMENT TEXT

Hi, my name is John Doe. For verification, the last 4 digits of my social are 6789 and my DOB is 01/01.

→

Hi, my name is [NAME]. For verification, the last 4 digits of my social are [SSN] and my DOB is [DATE_TIME].

→

Hi, my name is ********. For verification, the last 4 digits of my social are **** and my dob is *****.

Generative AI - Zero-shot and few-shot prompts

Get the desired output from an LLM with prompt engineering

AnyCompany Inc.
2022
10-K filing

Amazon Textract

What are the company's financial risk factors?

Foundation model

The company's financial risk factor includes high debt to income ratio, slowing sales…..

Response

Answer the question given the text and include sources.

Question: What is AnyCompany's D2I ratio?
Text: Any company's 10K report
Answer: D2I ratio is 0.9%
Source: Page 22
===
Question: What are the company's financial risk factors?
Text: Any company's 10K report
Answer:

AnyCompany Inc.
2022
10-K filing

Amazon Textract

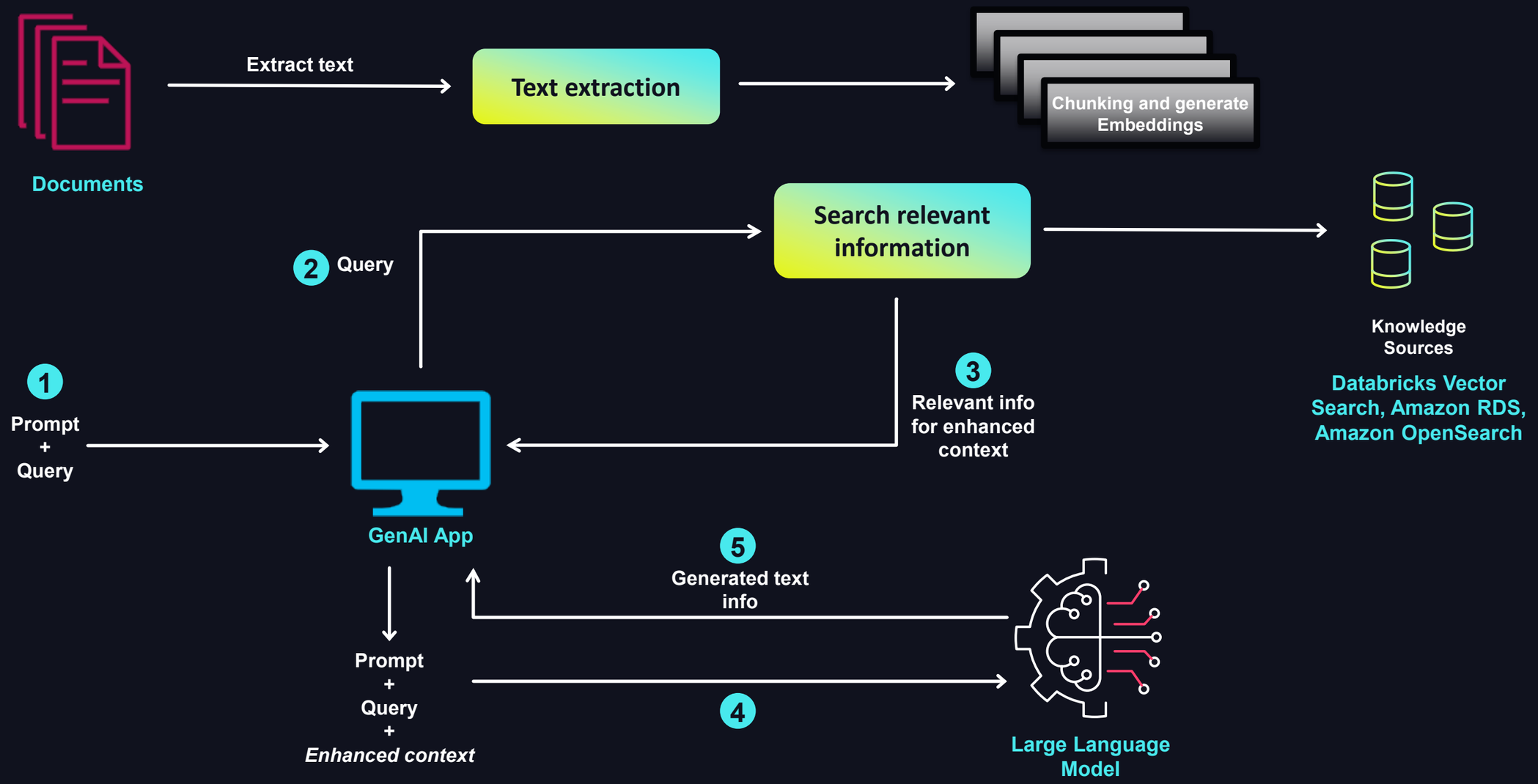
Foundation model

Answer The company's financial risk factor includes high debt to income ratio, slowing sales…..
Source: Page 15

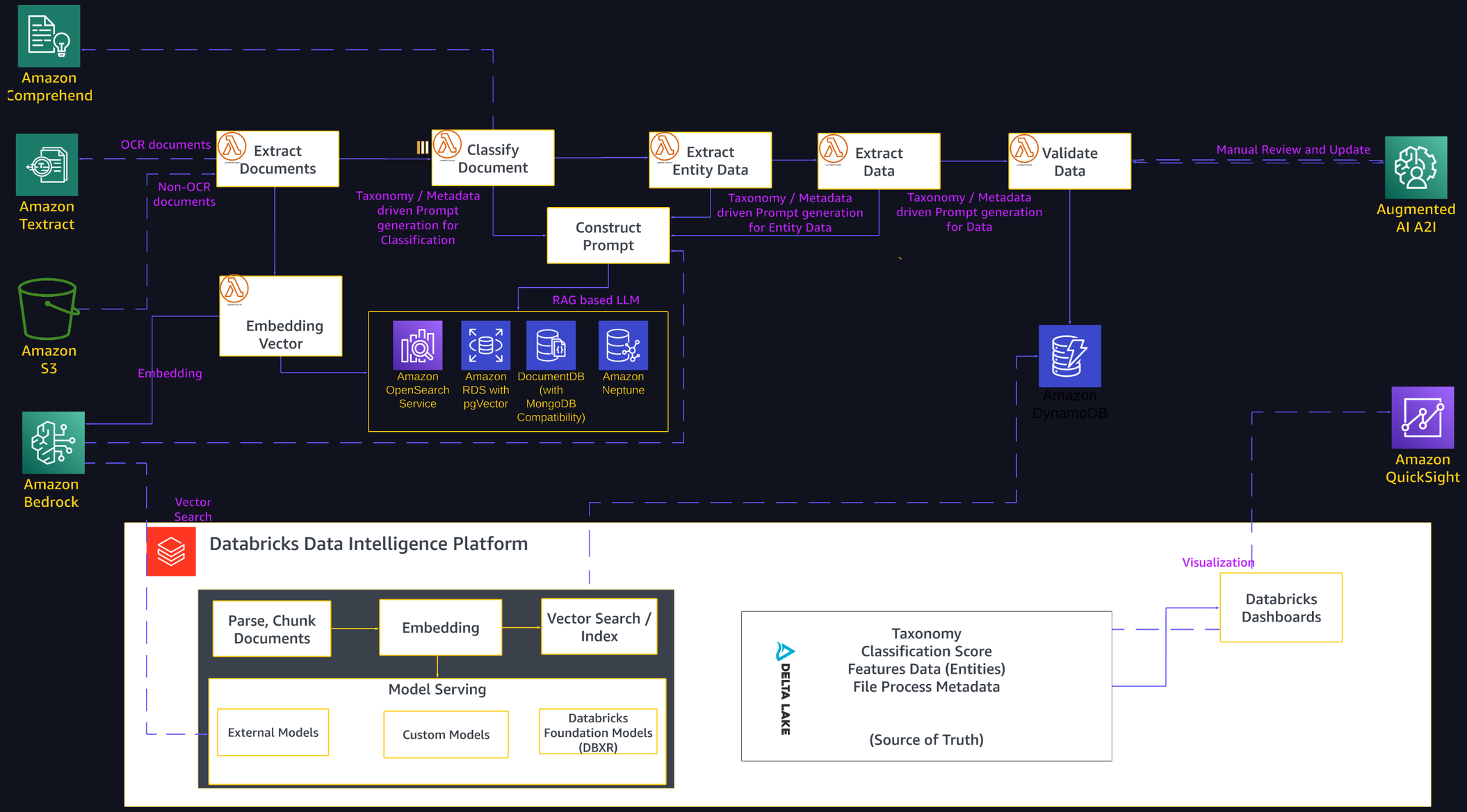
Response

Generative AI - Using Retrieval Augmented Generation (RAG)

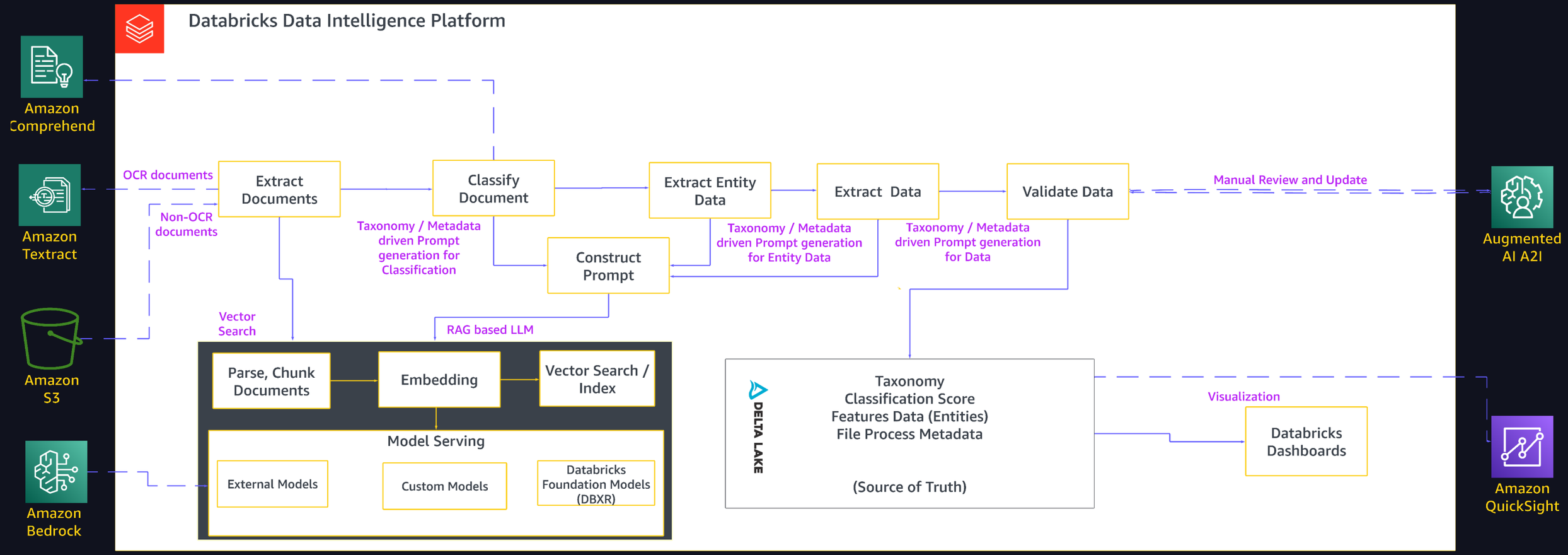
Augment prompts with relevant data in context



Reference Architecture – Using Databricks Approach (Batch, at-scale)



Reference Architecture – Using Databricks Approach (Batch, at-scale)



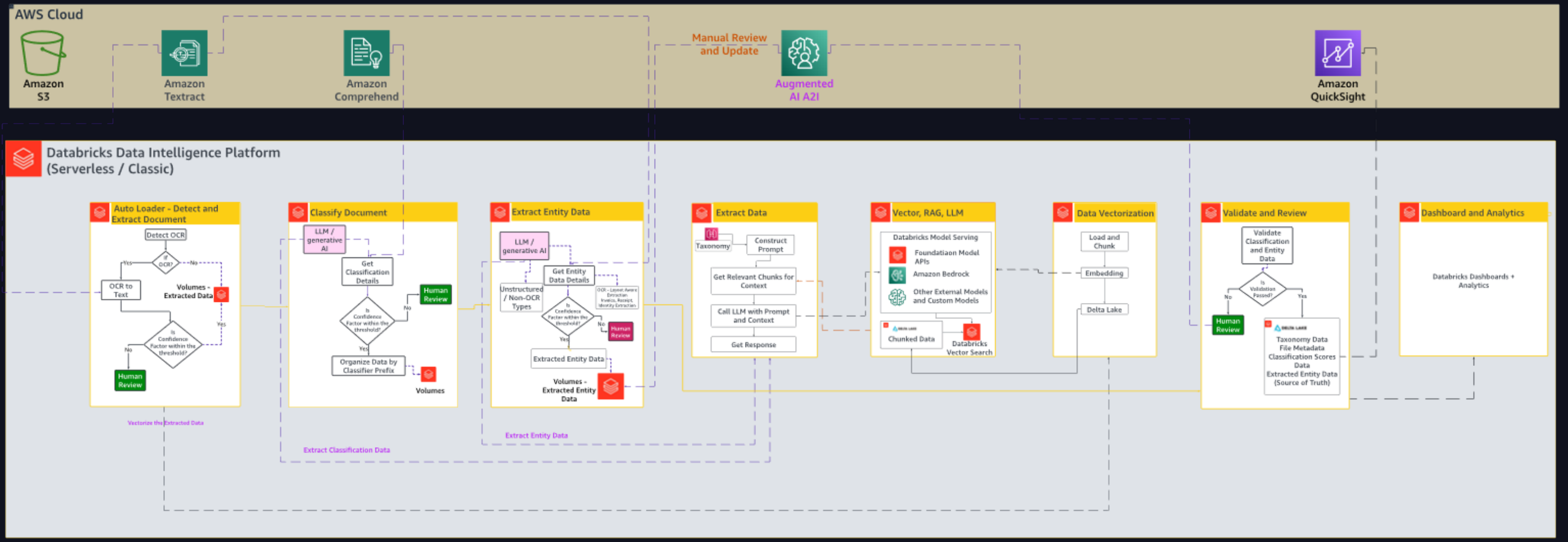
IDP Use-cases with Generative AI

A few very common use cases

- Document Q&A with Chatbots
- Document summarization
- Enhanced data extraction
- Document classification
- Automated content creation
- Medical record analysis
- Translation and localization
- Learning and development

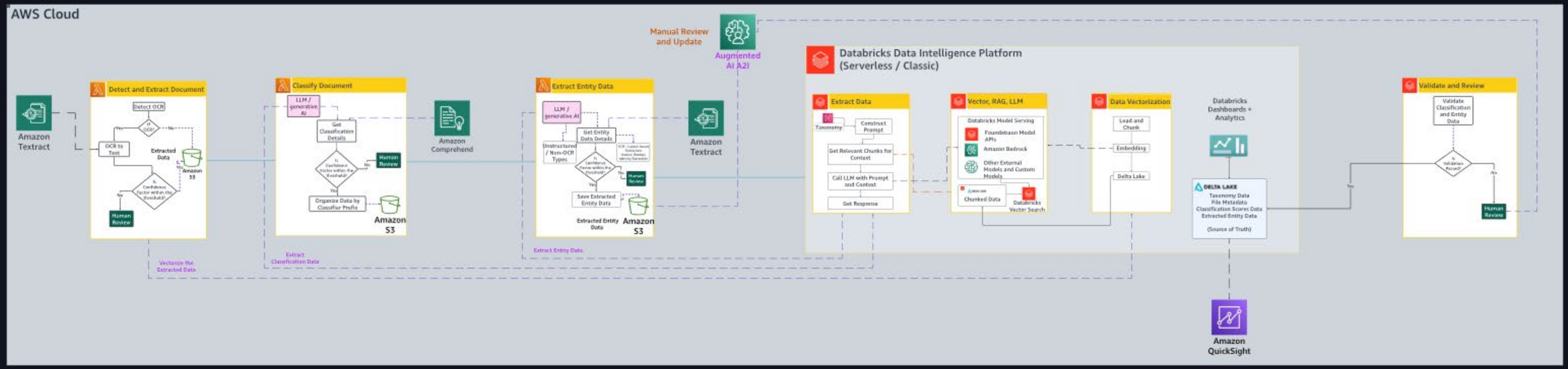
and more…

Reference Architecture – Using Databricks Approach (Batch, at-scale)



Delta Lake, Mosaic AI Vector Search, Deployment AI Gateway, DBRX and AWS AI Services and generative AI/LLM

Reference Architecture – Using AWS Lambda Approach (Real-time)



Amazon S3, AWS Lambda, SQS, SNS, Amazon Textract, Amazon Comprehend, Databricks' Delta Lake, Databricks Vector Store, Deployment AI Gateway, DBRX and 3P AI/LLM

Voice of Customer
Mayur Rajdev
Head of Architecture, AI & Automation
The Baldwin Group

End-to-End Enterprise Data Extraction

Input Channels


Clients


Colleague


Email


Copilot


Epic


Stratus

E2E Extraction

- Micro-Service managed by a rule-based engine
- LLM models is used to classify and extract information.
- Retrieval Augmented Generation models to feed as retrieve data.
- Innovative Taxonomy approach, describing documents and fields, minimizing the requirement of training to support documents

Output Channels


DataBricks

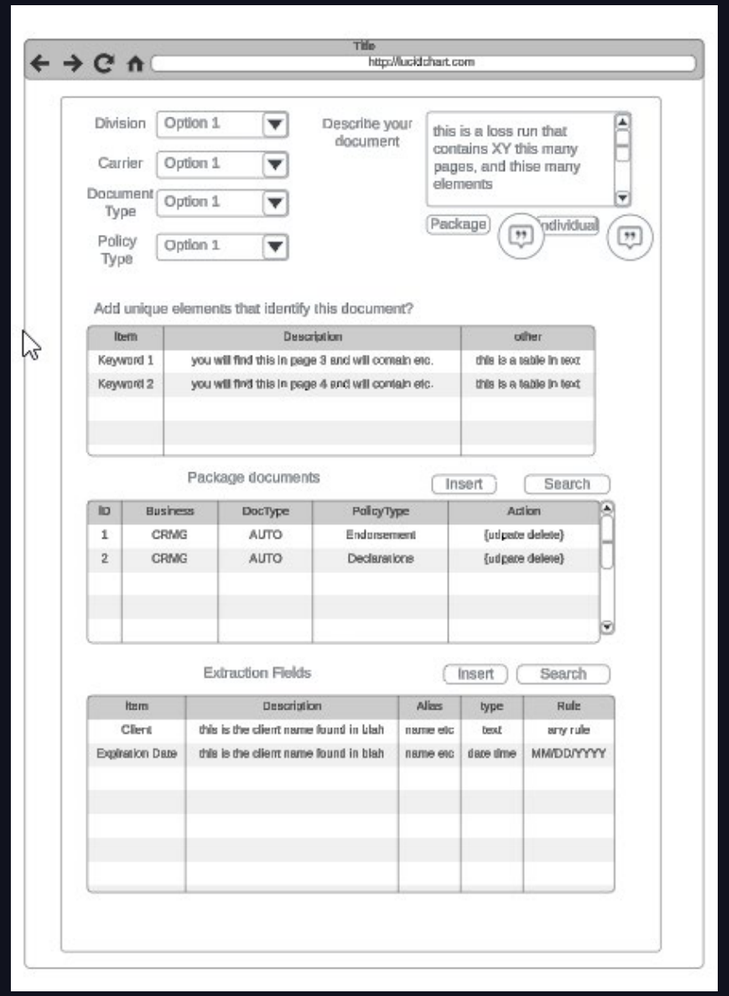

Epic Stratus


Copilot

Taxonomy – generative AI augmented approach

Overview:

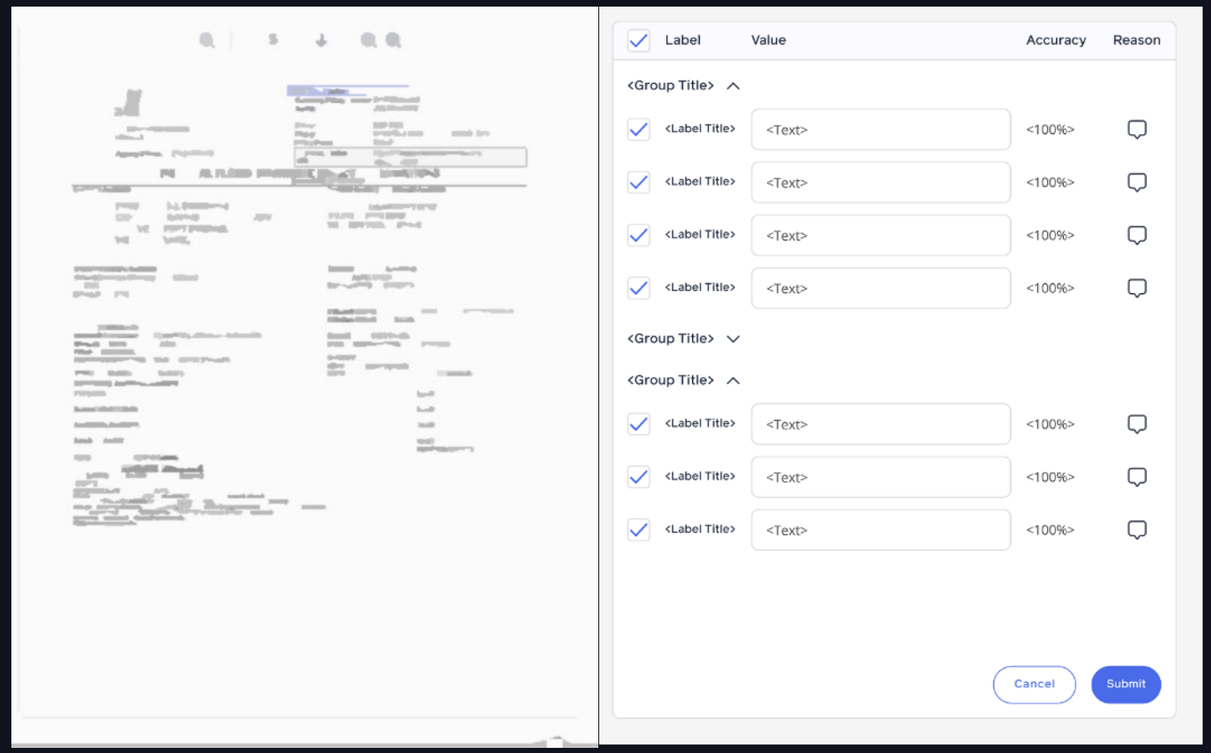
- The document Taxonomy is a representation of a document or group of documents that identifies and describes the document to be supported, it defines its characteristics, unique features to distinguish between other documents and entity fields to extract
- The taxonomy object is created using the Taxonomy Manager Ui, where the user can create and define the representation of a document
- This created JSON object will be used to create dynamic prompts using standard templates to feed the LLMs to classify and retrieve the relevant information from the document.



Classification and Extraction – generative AI augmented approach

Overview:

- The LLM will use the dynamic prompts generated by the taxonomy document, to classify the document into an individual file, or package as well as the type of document.
- Once classified, the system will use the defined entities from the taxonomy to individually search for each field in the document corpus and retrieve it in the proper format.
- Depending on the channel rules a Classification Station or Validation Station will be presented to a support team or to the business to confirm the accuracy of the classification or extraction and make the information available to the output channels and data lake.



Thank you!

Venkatavaradhan Viswanathan,
AWS, WW Databricks Architect
visvenky@amazon.com