13 May, 2025

Unlocking Gen AI - Architectures, Technologies, and Industry Use Cases

- Anita Ayyagari, Senior Solutions Architect, CSA
- Krishnendu Chaklader, Principal Solutions Architect, CSA
- Nithin Hegde, Senior Solutions Architect, CSA

Where data & AI come to LIFE

Housekeeping Tips

- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Informatica

Feature Rich Success Portal

Bootstrap trial and POC Customers

Enriched Customer Onboarding experience

Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Agenda

- ❑ **Introduction to Gen AI**
- ❑ **Gen AI Architecture & Components**
- ❑ **Harnessing data: Fueling Gen AI success while navigating common data acquisition challenges**
- ❑ **Real-time Industry Use Cases**
- ❑ **IDMC Gen AI Integration Features**
  - ❑ ***CDI Connectors & Transformations***
  - ❑ ***CAI Connectors & Recipes***
- ❑ **Q&A**

What is **Gen AI all about?**

***ChatGPT**, launched by OpenAI, became wildly popular overnight and galvanized public attention.*

- ChatGPT
- Midjourney
- Gemini
- Microsoft Copilot

The Journey to Gen AI

2010
Translation of Natural Language

2014
Mastering the Meaning of Words

2017-2022
Large Language Foundation Models

2022
Conversational Large Language Foundation Models

2025
Advanced Generative AI Models

Practical Applications of Gen AI

Content creation and augmentation

Producing a "draft" output of text in a desired style and length

Knowledge base querying and discovery

Enabling users to generate answers to input, based on prompts

Tone

Text manipulation, to soften language or change text intent

Summarization

Summarized versions of conversations, articles, emails and webpages

Simplification

Breaking down complex information, creating outlines and extracting key content

Classification of content for specific use cases

Sorting by sentiment, topic, or other classifications

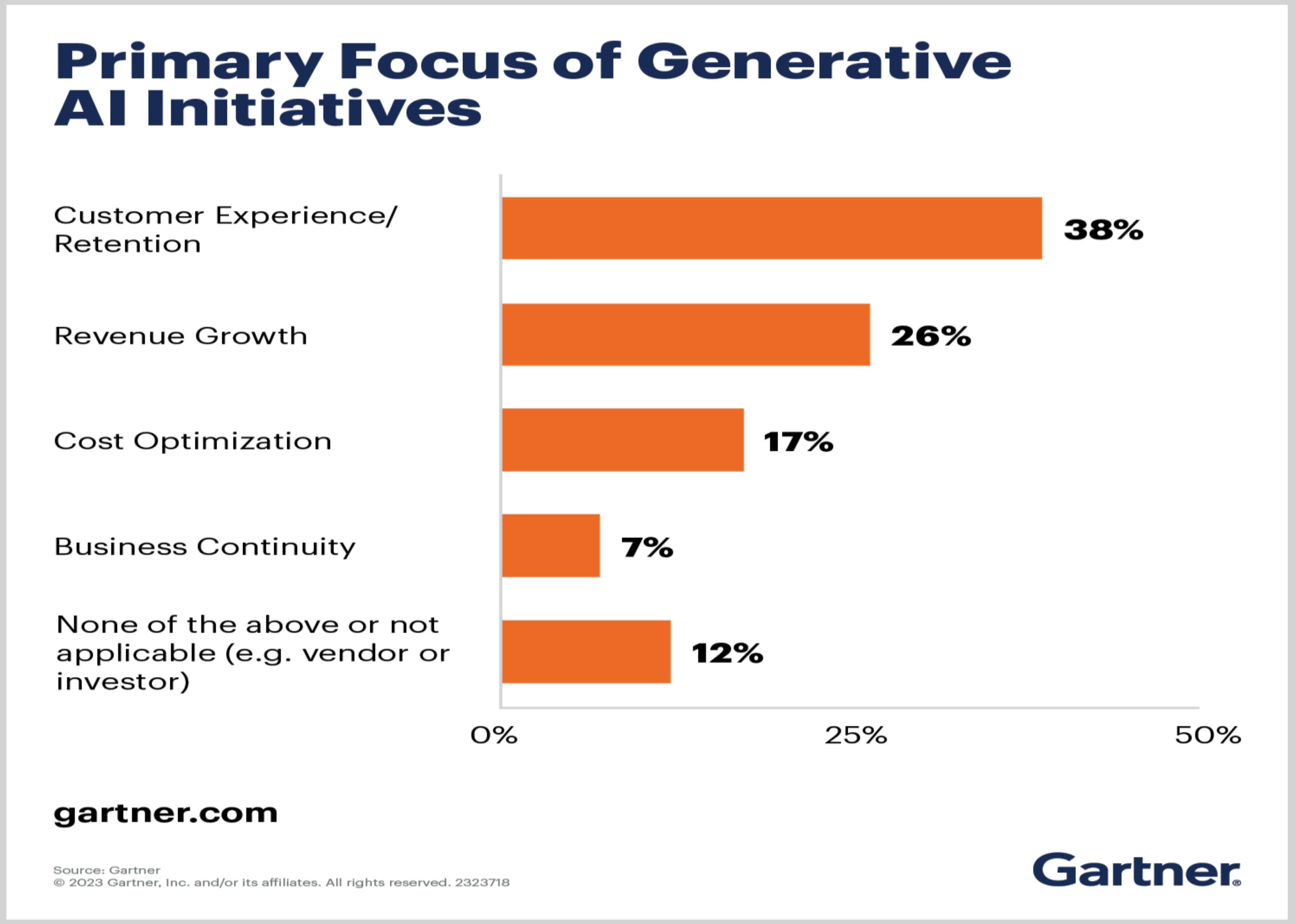
Chatbot performance improvement

Optimized "sentiment" extraction, whole-conversation sentiment classification and generation of journey flows from general descriptions

Programmatic coding

Code generation, translation, explanation and verification

Why Gen AI Matters?



Gen AI Architecture & Components

Gen AI Terminology

Text Generation	Large Language Models (LLM)	Natural Language Processing (NLP)	Machine Learning	Prompt

Prompt Engineering	Retrieval Augmented Generation (RAG)	Vector Databases	Image and Audio Generation
Human in the Loop (HITL)	Deepfake Technology	Responsible AI	Model Tuning

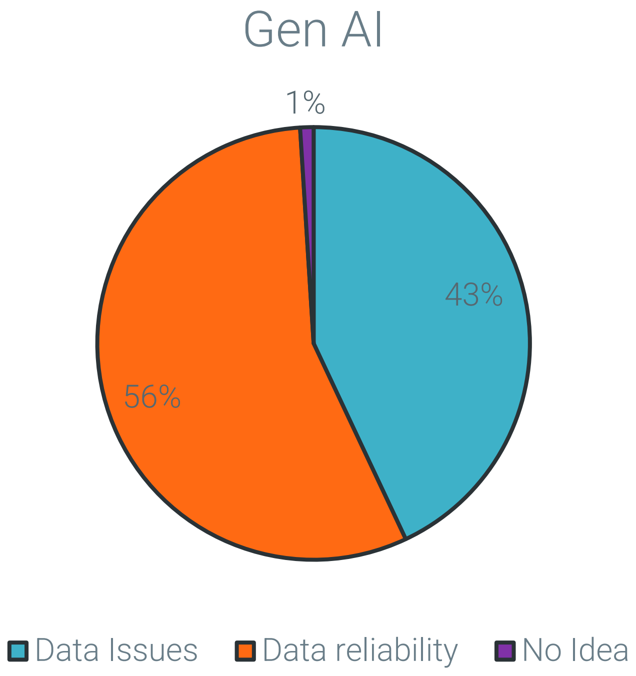
Layers of Gen AI Architecture

Application layer	• Application/UX
Data layer	• Data Platform • API Management Layer
Orchestration Layer	• Large Language Model Ops (LLMOps) • Prompt Engineering
Model Layer	• ML and LLM Foundation Models • Model Fine-Tuning and Model Hub
Infrastructure Layer	• Compute • Cloud Platform

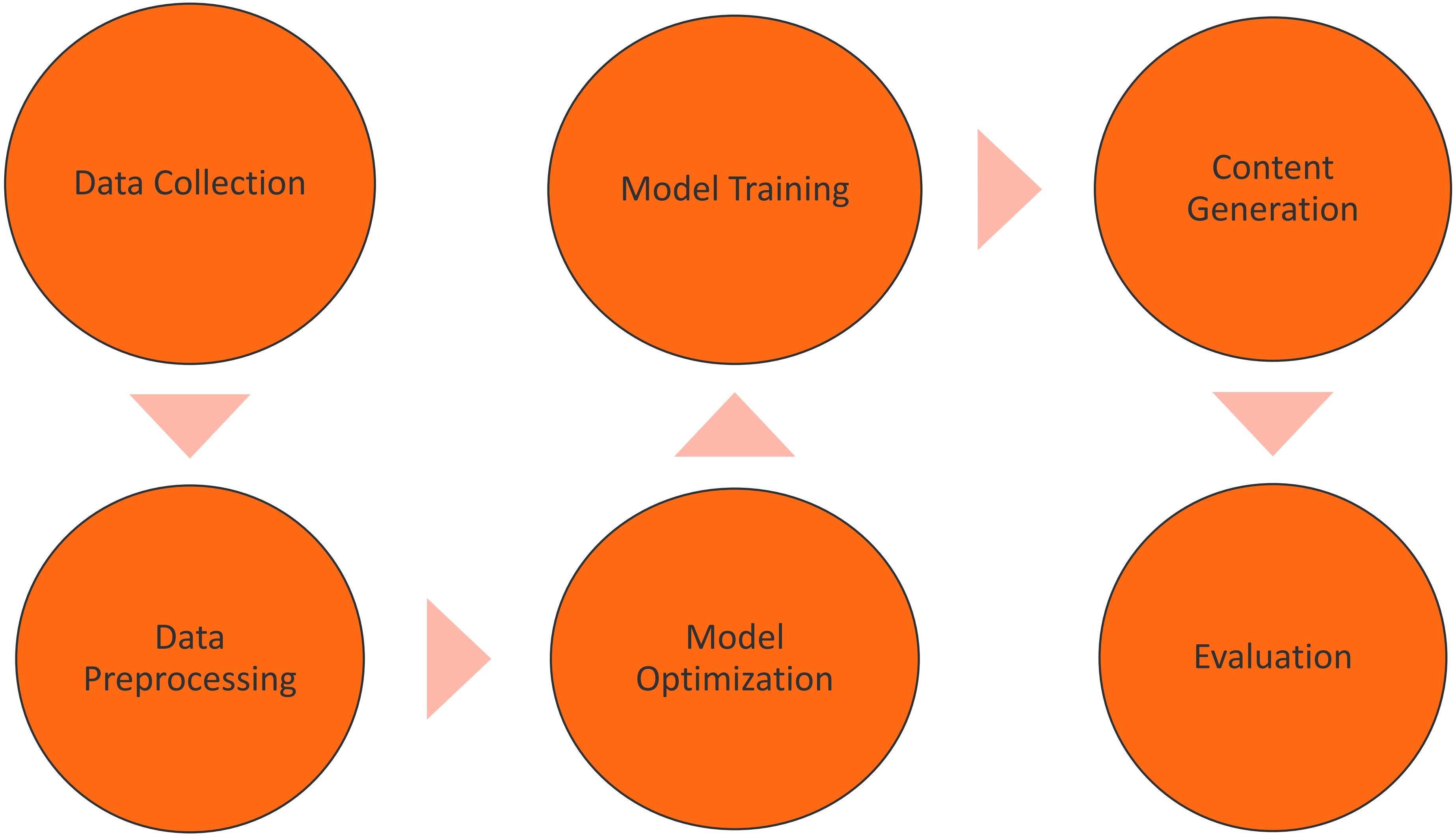
Navigating Common Data Acquisition Challenges

Key takeaways for CDOs:

- 43% of data leaders agree data issues are a roadblock to demonstrating Gen AI's value
- 56% describe data reliability as a key barrier to advancing Gen AI pilots



Harnessing Data: Fueling Gen AI Success



IDMC Accelerates AI Implementations

Getting the Right Data the Right Way at the Right Time

Cloud Data Governance and Catalog, Cloud Data Quality

- Discover data sources and AI models
- Understand data quality and provenance
- Remediate data quality

Master Data Management

- Ingest, Integrate Raw data sources to Master key data domains like customer, product, supplier, etc.

Cloud Data and API Integration

- Ingest, Integrate and Prep training, fine-tuning, and validation data
- Create vector embeddings for RAG
- Auto-classify and Mask Sensitive

Model Training and Fine-Tuning

- Python/Notebooks
- GPU Infrastructure
- Foundation LLMs

Model Serving and Monitoring

- Model Serving
- Model Accuracy and Performance Metrics

Cloud Data and API Integration

- Retrieval Augmentation Generation (RAG)

Cloud Data Governance and Catalog

- AI Model Governance
- Underlying data Lineage

Informatica

Gen AI with Cloud Data Integration

Real-Time Industry Use Cases of IDMC with Gen AI

- ❑ **Healthcare :**

Organizations across the healthcare sector are harnessing the power of AI to transform diagnostics, enhance operational efficiency, and medical research forward. IDMC improves the quality and relevance of AI-generated solutions automate routine tasks like handling customer queries by retrieving similar past interactions, billing, and processing insurance claims, reducing administrative burdens (Convert Patient Admission Data to FHIR R4) and allowing healthcare staff to focus more on patient care.

- ❑ **Finance and Insurance :**

Gen AI can significantly streamline and enhance loan processing within financial institutions. Businesses can streamline the loan approval process at scale to save time and drive accuracy. Aside from automated loan approval against predefined criteria, it also improves risk assessment outcomes, provides documentation for regulatory compliance and improves the efficiency of manual reviews by providing detailed background information. Gen AI can automatically extract relevant information from documents such as loan applications, income statements, and tax returns, reducing the need for manual data entry and speeding up the processing times.

Understanding the Generative Pre-trained Transformer

- **Generative Pre-trained Transformers (GPT)**
  - The Transformer architecture uses a self-supervised learning approach to train large-scale neural networks on vast amounts of text, weighing the importance of different words relative to each other, agnostic of position in a sequence.
  - This parallel processing approach enables the underlying model to capture long-range dependencies and contextual relationships in the training data set much more effectively than previous architectures like Recurrent Neural Networks or Generative Adversarial Networks.
  - The Transformer then attempts to predict the next word in a sequence, learning patterns and relationships in unlabeled data, generating coherent and contextually relevant output based on the prompt its given and its training dataset.
- **Chunking and Embedding**
  - The Transformer breaks down large text into smaller, more manageable tokens (words, sub-words etc.) for highly efficient processing and retrieval. This process is called **Chunking**.
  - **Embedding** refers to techniques that converts words, phrases or other data into numerical vector representations that capture semantic meaning in a high-dimensional space.
- **Vectors, Dimensions, Index and Vector Databases**
  - A **vector** is a mathematical representation of data points as arrays of numbers in a multi-dimensional space, where proximity between vectors indicates similarity.
  - A **dimension** refers to the number of values in a vector, with higher dimensions allowing for more nuanced representations of complex data relationships.
  - Vector embeddings are often stored in a data structure called an **index**, that enables fast and efficient similarity searches.
  - Specialized databases, optimized for storing, managing and querying large collections of vector embeddings enable similarity searches and semantic matching. These are called **Vector databases**.
  - Transformers calculate similarity between vectors, basing their retrieval on meaning rather than keywords.

When AI Gets It Wrong: Addressing AI Hallucinations

❑ What is AI Hallucination ?

Hallucinations are instances where an AI model generates outputs or responses that are factually incorrect, misleading, or completely fabricated

❑ How do AI hallucinations occur?

Quality and Incomplete of the training data

Model limitations and Complexity

Lack of proper grounding

❑ How to prevent AI hallucinations ?

Curate High-Quality Training Data

Introducing Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG)

- Retrieval-Augmented Generation, or RAG, is a technique in the field of Gen AI that combines retrieval-based approaches with generative models to improve the performance of language models.
- Contextual Relevance helps to contextualize LLM with Enterprise data, including unstructured or semi-structured data.
- Access to extensive data expanding its knowledge and capability to handle specialized queries.
- Drastically reduce hallucinations and improve prediction accuracy – better than basic prompt engineering.
- Easily adapts to new data – RAG can adapt in situations where data could evolve over time.
- Interpretability using RAG – the source of the LLM's answer can be pinpointed.
- Domain-specific Applications – particularly useful for industries such as Finance, Healthcare, or Legal, where up-to-date, precise information is crucial for decision-making processes.

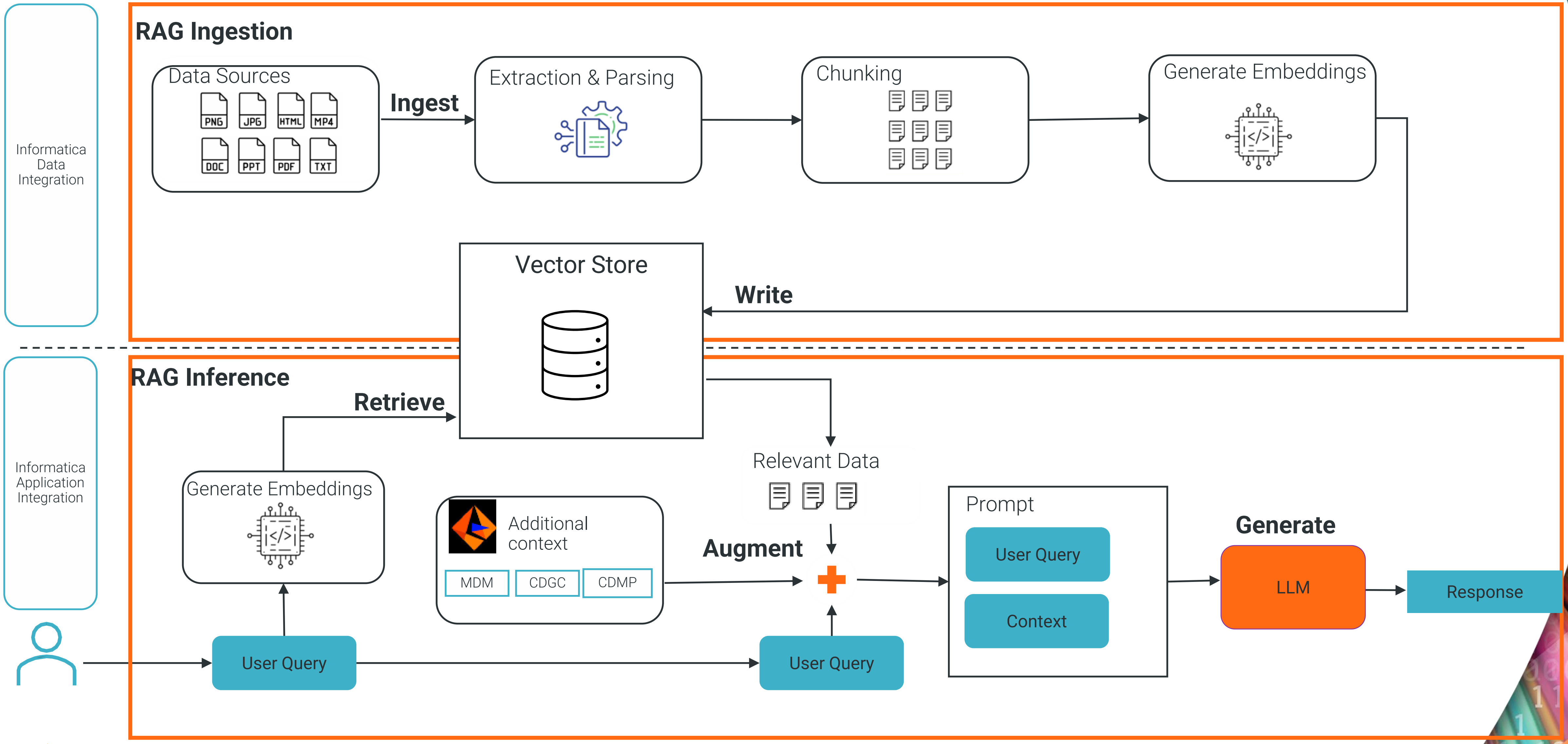
How does IDMC support integration with Gen AI?

- Provides OOTB support for specialized transformations such as **chunking** and vector **embedding**.
- Provides OOTB support for LLMs/embedding models.
- Native connector for **Pinecone Vector DB** (Note: Support for Azure Cosmos DB as Vector DB, ~May 2025).
- IDMC supports both RAG ingestion and RAG consumption pipelines.
- Organizations can build efficient RAG systems that access extensive, high-quality datasets, contributing to more accurate and contextually relevant AI-driven outputs.
- IDMC services can clean, enrich, and prepare data for AI models, ensuring that retrieved data is primed for effective generation and analysis.

Gen AI Requirements in Modern Data Architecture

Unstructured Data	Gen AI Transformations	Vector Database	Large language Models (LLM)	AI Orchestration
Connectors and metadata for documents, images and videos	Chunking and embedding unstructured data	Integration with Pinecone vector database	Integration with major LLM providers, including local LLMs	Framework to build Gen AI inference pipelines

Informatica RAG Ingestion and AI Inference Architecture



Gen AI Transformations

- Chunking transformation

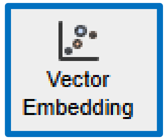
Chunking transformation splits large pieces of text into smaller segments, or chunks.

Sentence splitting

Word splitting

Fixed length character splitting

- Vector Embedding transformation

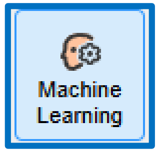


Vector Embedding transformation to generate vector embeddings for input text, capturing the semantic meaning of the text in a vector format.

Word embedding

BERT embedding

- Machine Learning transformation

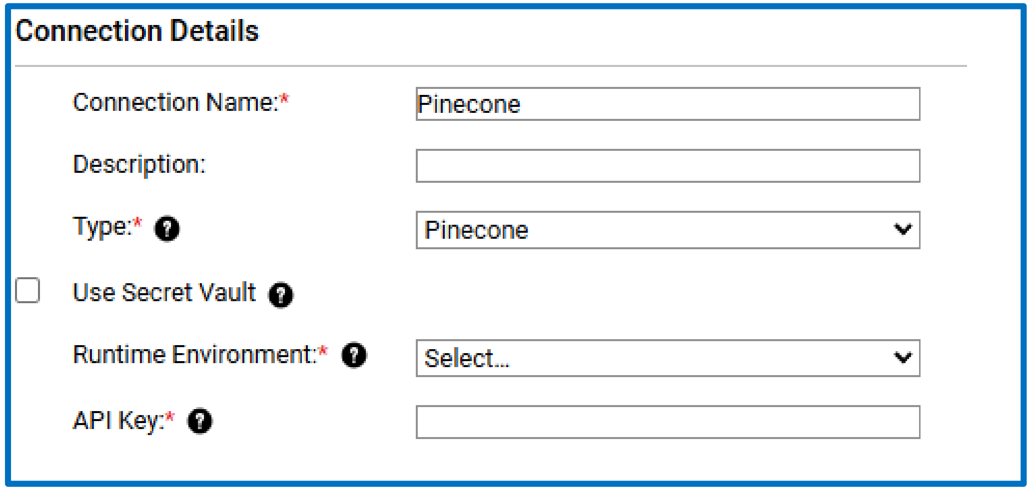


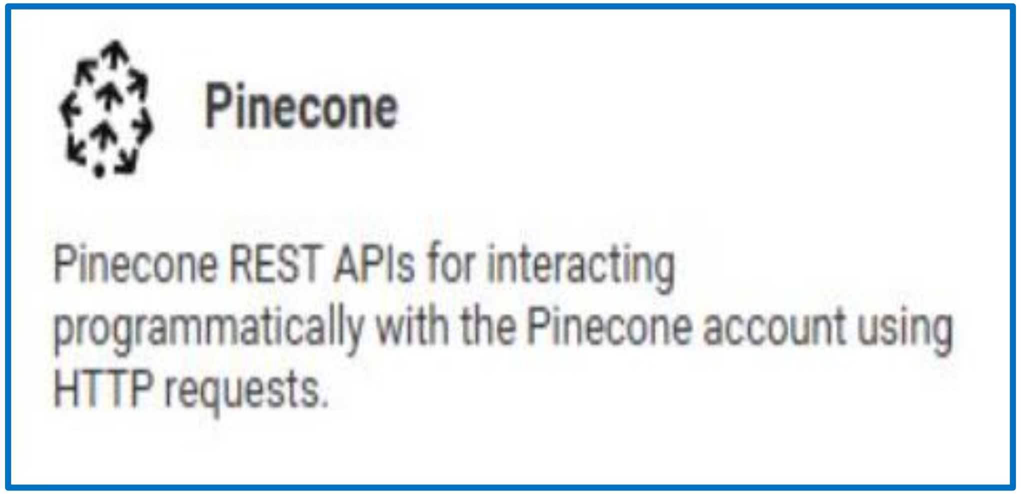
Machine Learning transformation runs a machine learning model and returns predictions.

Vector DB Connector (Pinecone)

- Pinecone Connector

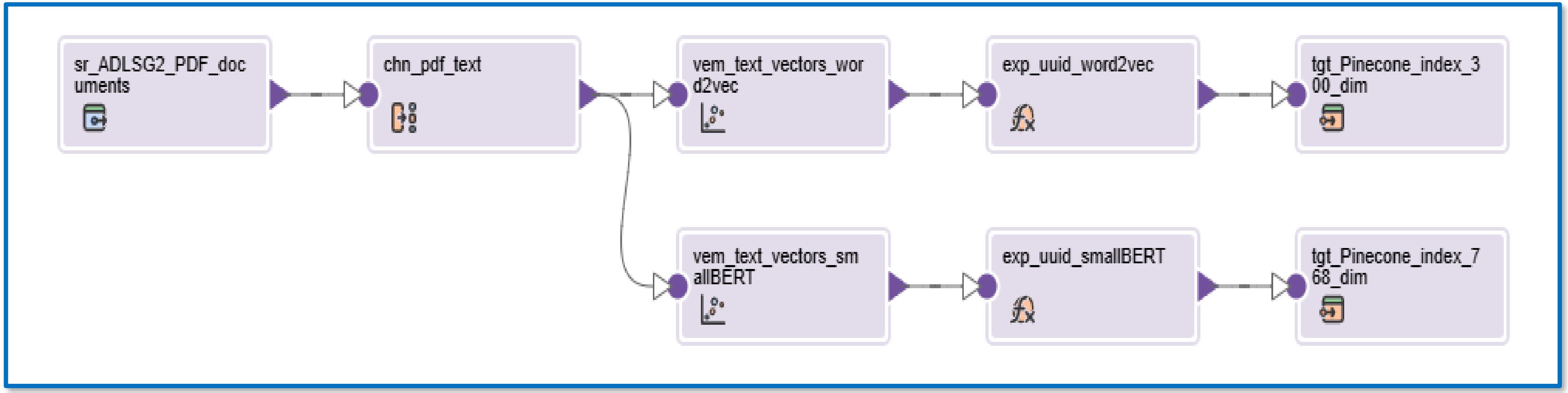
  Pinecone Connector streamlines the process of ingesting data into Pinecone, performing vector searches, and integrating search results into your application workflows.





Note: The API key of your Pinecone account to authenticate access to the Pinecone APIs.

Sample Mapping with Gen AI Transformations



Gen AI with Cloud Application Integration

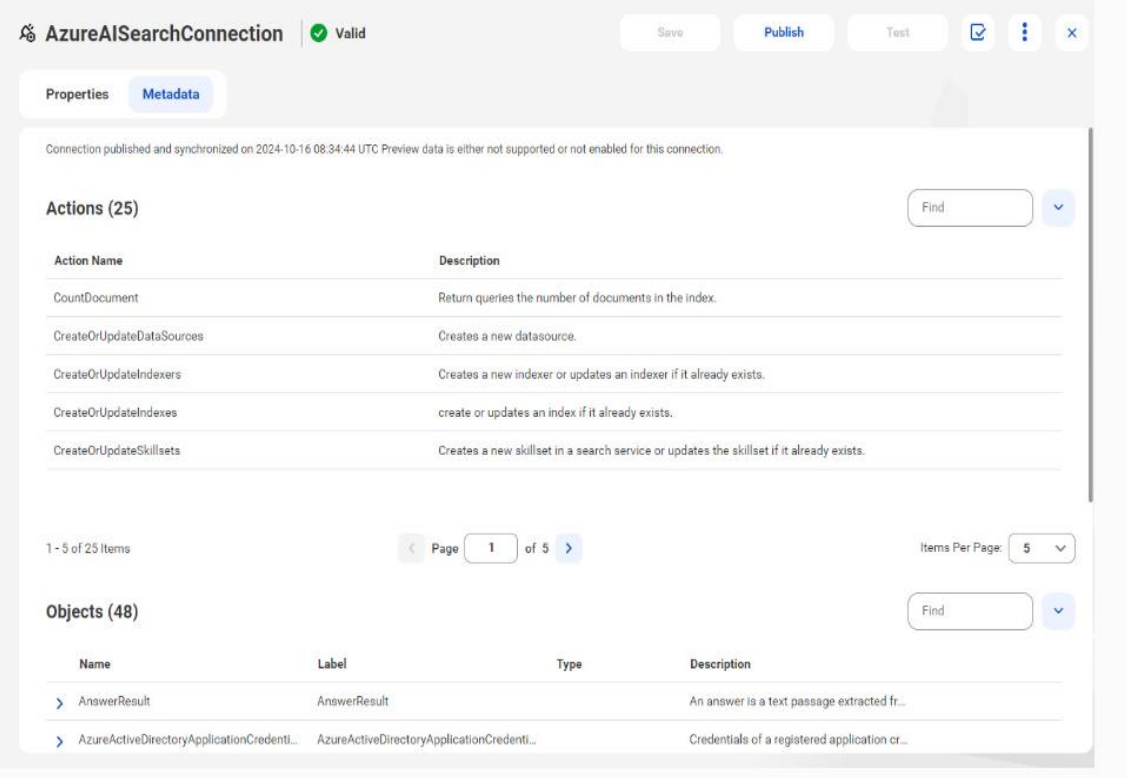
Gen AI connectors in CAI

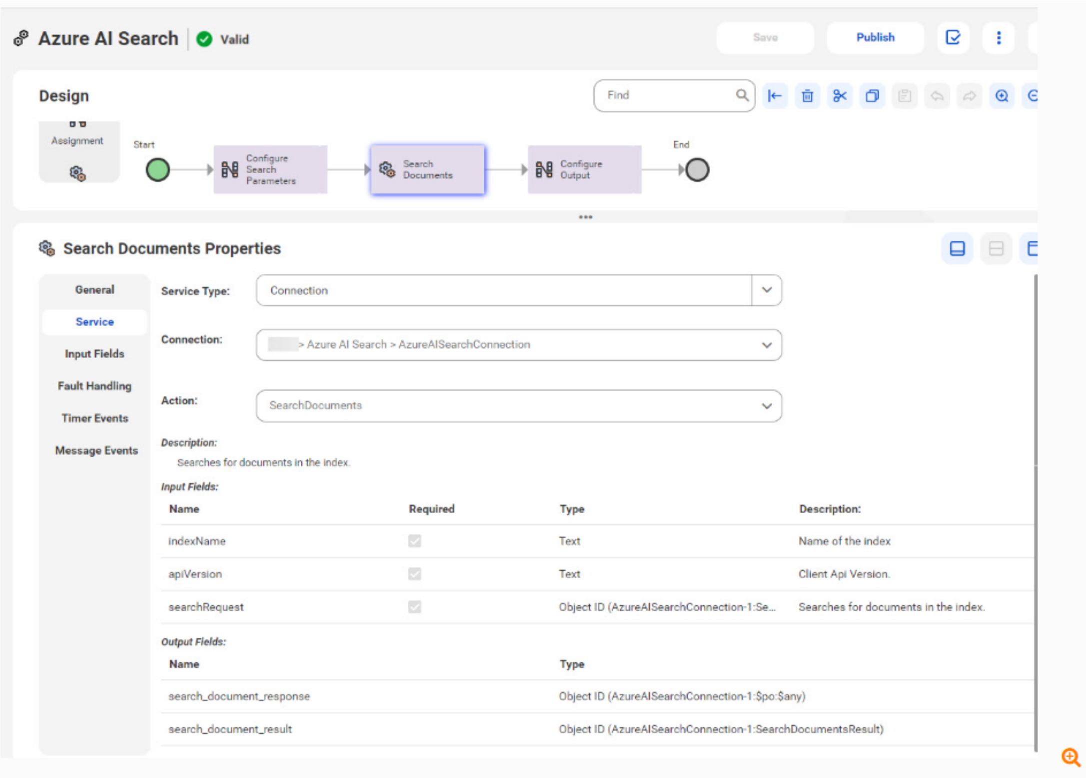
Realtime integration of data into Gen AI platforms

- ❑ The supported connectors within CAI are listed below; while the concepts are the same, each one has its own specific terminology:
  - ❑ Amazon – Bedrock
  - ❑ Microsoft– Azure AI Search, Azure Open AI
  - ❑ Open AI
  - ❑ Google– Vertex AI, Gemini
  - ❑ Local – Ollama
  - ❑ Pinecone

Azure AI Search/Azure Open AI connector

- ❑ Azure AI Search helps developers integrate modern search capabilities into their applications

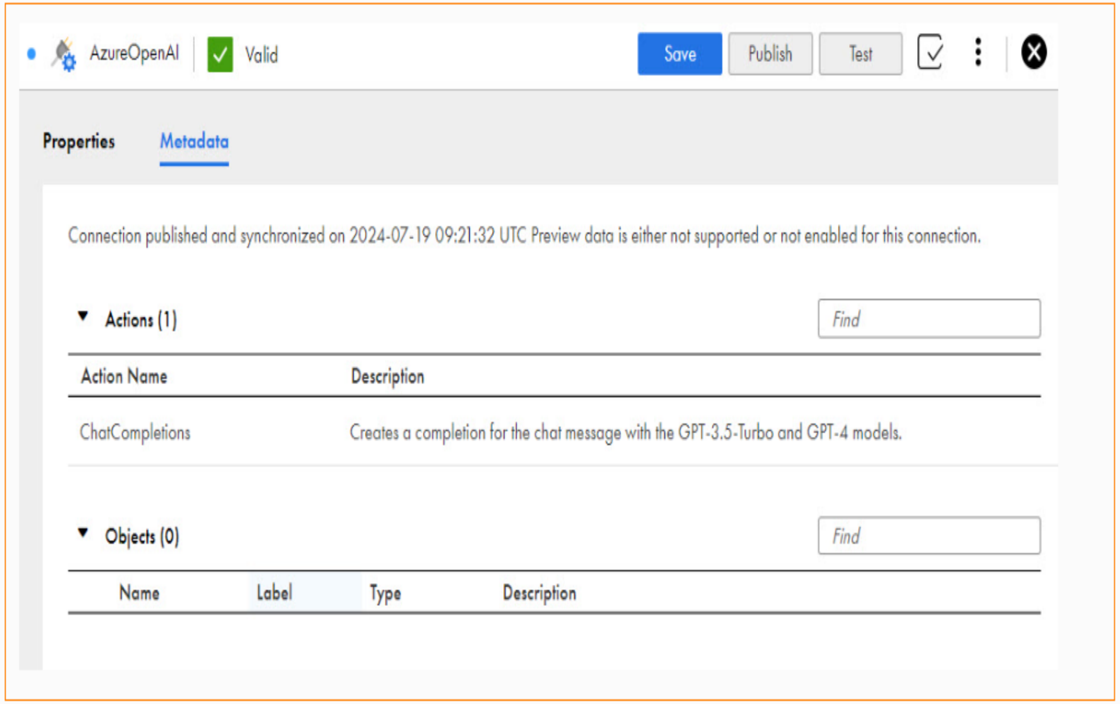


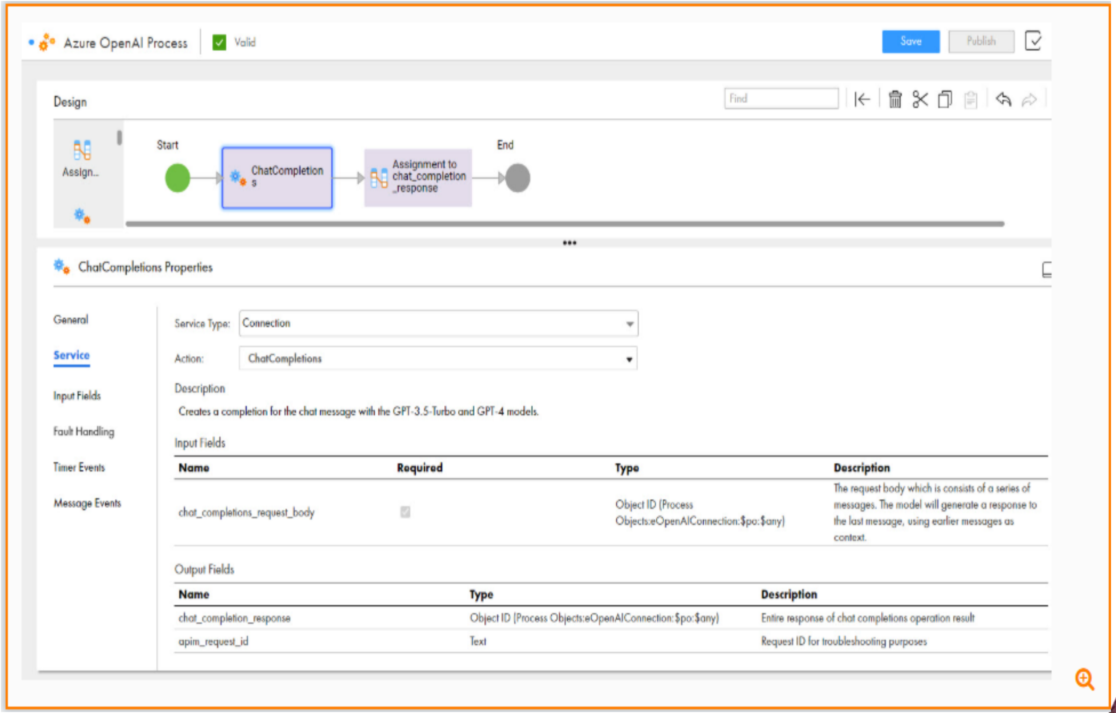


Informatica

Azure AI/Azure Open AI Connector

- Azure Open AI helps in seamless interaction between your Open AI-defined APIs and various Azure services

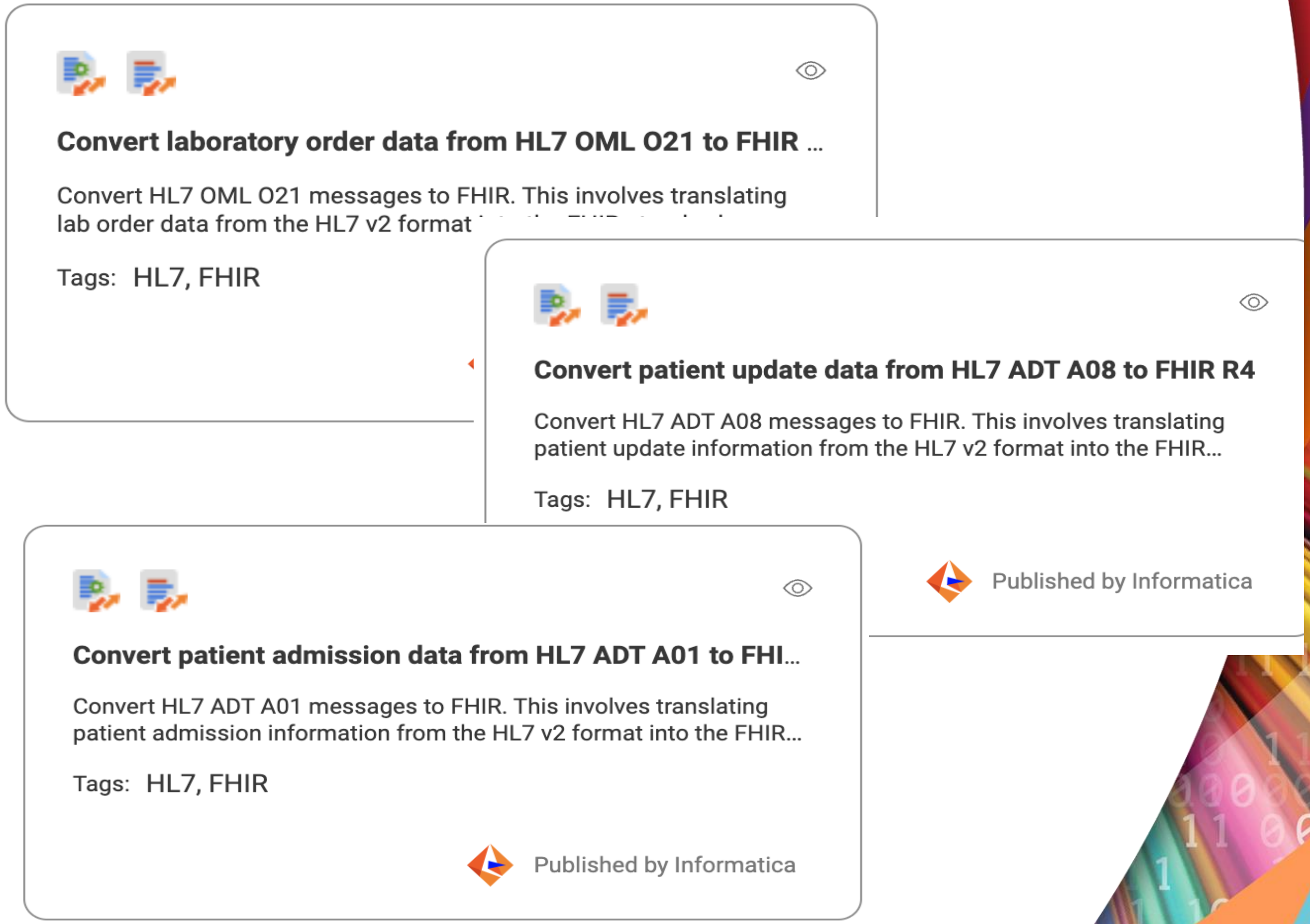




Recipes

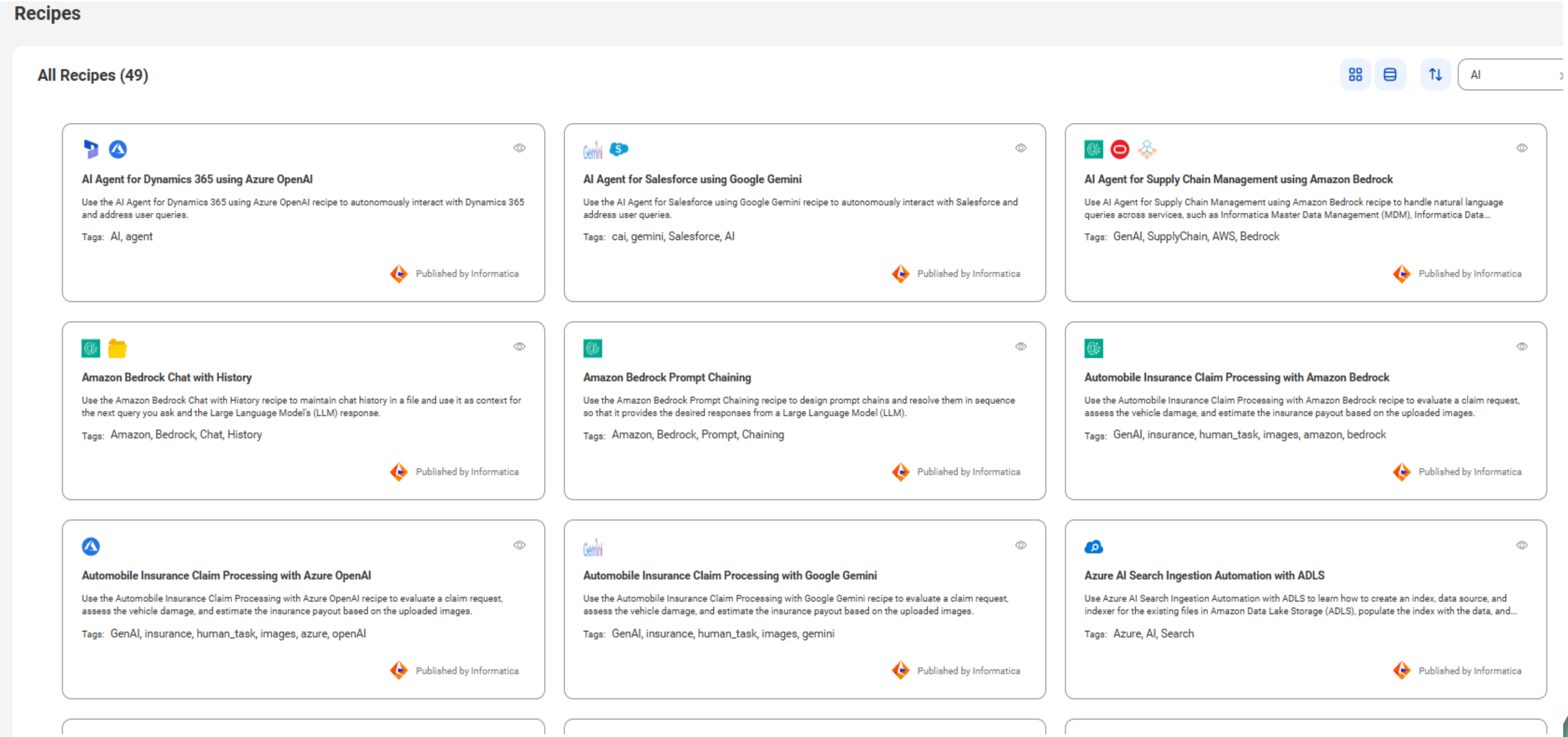
One-Click Automated Integration Process Setup

- ❑ Predefined integration solutions designed to connect, automate, and standardize business processes that require integrations across multiple applications
- ❑ Capability Overview
  - ❑ Prebuilt integration packages including Gen AI app creation using AI agents, RAG, prompt engineering frameworks
  - ❑ In-app user experience for seamless discovery & consumption of recipes
  - ❑ Complete documentation with specific how-to libraries and tutorials
- ❑ Benefits
  - ❑ Easy
  - ❑ Efficient
  - ❑ Enterprise Grade



Recipes

At a glance!



Recipe: Azure AI Search Ingestion Automation with ADLS

- Learn how to create an index, data source, and indexer for existing files in Azure Data Lake Storage (ADLS), populate the index with data, and then return the documents.

Recipe - Azure AI Search Ingestion Automation with ADLS Use

Recipe Content (7) ☆ Primary As

Asset Name	Description	Type
AzureAISearchConnection	Azure AI Search connection	App Connection
Create Index	The subprocess that creates an index or updates an existing index	Process
Create Indexer	The subprocess that creates an indexer or updates an existing indexer	Process
Azure AI Search Documents	The subprocess that searches for documents in the index	Process
Create Data Source	The subprocess that creates a data source or updates an existing data source	Process
Create Skillset	The subprocess that creates a skillset or updates an existing skillset	Process
☆ Azure AI Search Ingestion Automation with AD...	The process that creates the index, data source, and indexer for the files present in ...	Process

Q&A

Thank you