Building Your GenAI Application on InterSystems

Revolutionizing How You Use Health Data

Introduction

Generative AI applications represent a transformative leap in artificial intelligence, enabling systems to create new content, ideas, or solutions rather than merely analyzing or reacting to existing data. These applications are powered by models like Generative Adversarial Networks (GANs) and Transformer-based architectures such as GPT and BERT, which can generate human-like text, images, music, and even complex designs. From personalized marketing content and product recommendations to automated video production and code generation, the versatility of Generative AI is revolutionizing industries by delivering hyper-customized, scalable, and creative outputs at unprecedented speeds.

In healthcare, examples abound. **Generative AI** allows providers to automatically transcribe and summarize conversations with patients. It enables researchers to efficiently scan, summarize, and query large pools of information. And it aids in drug discovery by simulating biomolecules and predicting how they fit together. By combining this innovation with robust data platforms, healthcare workers and businesses can unlock new possibilities, ranging from multi-modal AI systems that integrate text, images, and speech, to enhanced decision-making processes based on generative insights.

InterSystems IRIS® for Health is a comprehensive platform for building and scaling Generative AI healthcare applications by seamlessly integrating robust data management, high-performance analytics, and real-time health informatics interoperability capabilities. Its multi-model database supports structured, unstructured, and multi-modal data, making it ideal for the diverse data requirements of Generative AI. It supports massive scalability through efficient caching, partitioning, and sharding.

InterSystems®
Creative data technology

InterSystems IRIS for Health offers seamless integration with the Python ecosystem, enabling developers to utilize popular frameworks such as LangChain and LlamaIndex for Generative AI. This compatibility allows for the ingestion and searching of unstructured medical and other documents using **tokenization** and **vector embeddings** from natural language processing (NLP), streamlining the development process and enhancing the functionality of AI-based applications. Additionally, the platform's vector search and vector datatype allow for scalable indexing and querying of high-dimensional vector embeddings of data tokens, essential for applications like semantic search, personalization, and anomaly detection.

InterSystems IRIS for Health incorporates native vector search capabilities that leverage optimized chipset instructions, specifically SIMD (Single Instruction, Multiple Data), to enhance performance in vector operations. These optimizations are integrated within the core database engine, enabling efficient similarity computations directly through SQL functions such as **vector_dot_product** and **vector_cosine**.

The latest release of InterSystems IRIS for Health builds on the groundbreaking Vector Search capabilities introduced in 2024.1.0, delivering a **3-4x speed improvement** in raw vector search performance. This leap in efficiency is achieved through optimized low-level vector similarity operations and a refined storage model for vector data.

Generative AI and the Role of Python

Python serves as the cornerstone of AI and machine learning (ML) development due to its simplicity, versatility, and rich ecosystem of libraries and frameworks. Its intuitive syntax allows developers to focus on solving complex problems rather than grappling with the intricacies of the programming language itself. Python's robust libraries, such as LangChain, LlamaIndex , TensorFlow, PyTorch, scikit-learn, and pandas, provide powerful tools for building and deploying AI/ML models, handling data preprocessing, and performing advanced analytics.

InterSystems has developed several Python integration libraries including: LangChain-Iris, Llama-Iris, and SQLAlchemy-iris to make it easy to use Python with the InterSystems IRIS for Health. The platform enhances the integration through **Embedded Python**. This feature allows Python code to execute natively within the database engine. This integration offers several advantages:

- **Seamless Language Interoperability** – Embedded Python enables developers to write methods in Python within InterSystems IRIS classes, allowing Python and ObjectScript code to coexist and interact seamlessly. This design facilitates the use of Python's extensive libraries alongside InterSystems IRIS's robust data management capabilities.
- **Direct Access to InterSystems IRIS Features** – Through the IRIS module, Python developers can directly interact with InterSystems IRIS features that include accessing and manipulating objects called globals, calling ObjectScript methods, and executing SQL commands. This direct access streamlines development by eliminating the need for intermediate layers or external connectors.
- **Unified Development Environment** – By embedding Python directly into the InterSystems IRIS kernel, developers can harness Python's extensive ecosystem of libraries to develop data-intensive, mission-critical applications efficiently.

InterSystems and IPA's subsidiary BioStrand collaborated to create an innovative integration of vector search with LENSai™, an AI-driven healthcare application, to identify novel therapeutic targets more quickly, streamlining the drug discovery process and reducing the time from discovery to clinical trials.

A Customer Experience

> "By combining InterSystems Vector Search with IPA's LENSai, we're empowering developers and researchers in the Life Sciences with unparalleled tools for extracting value and insights from complex datasets, driving forward the potential for AI in every application within the healthcare and life sciences sectors," said Dirk Van Hyfte MD, PhD, Co-Founder and Head of Innovation of BioStrand.

Utilizing the InterSystems native support for Python, developers build sophisticated AI applications. For instance, an IRIS-GenLab application integrated the Flask web framework, SQLAlchemy ORM, and InterSystems IRIS for Health to showcase functionalities such as machine learning, natural language processing (NLP), large language models (LLMs), and Generative AI APIs. This application employed a chatbot using Torch, named entity recognition with spaCy, sentiment analysis, text generation via Hugging Face's GPT-2 model, and integration with OpenAI's ChatGPT.

To learn more about Developing an AI Powered IRIS Application in Python, please see the on-demand video at **https://www.intersystems.com/uk/events/intersystems-uki-tech-talk-developing-an-ai-powered-iris-application-in-python/**

InterSystems Vector Search and Embeddings

Vector representations, often referred to as embeddings, are a cornerstone of modern AI, enabling machines to understand and process complex data such as text, images, and audio. Via tokenization, or analysis of unstructured data into simpler elements, these embeddings transform raw input data into dense, numerical representations that capture the underlying patterns, relationships, and semantics in a way that computers can efficiently analyze.

To implement relational SQL-based data capture, InterSystems created the **EMBEDDING** datatype, revolutionizing how vector embeddings are generated and stored. With EMBEDDING columns, embeddings for text or other content are created automatically, removing the need for manual coding to invoke models and manage embeddings. Simply define a source column and specify the embeddings model or service to use—InterSystems handles the rest.

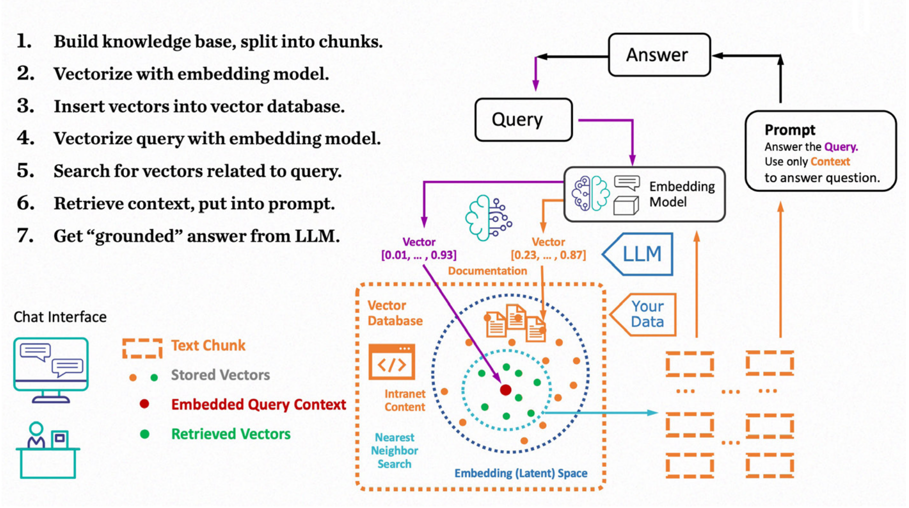
Our **Vector Search** capabilities empower application and solution providers to perform fast and efficient indexing and querying of high-dimensional data, such as embeddings used in AI applications. Designed for performance, similarity search leverages advanced indexing techniques to handle complex queries on large datasets with minimal latency. Its scalability ensures seamless handling of vast amounts of data, making it ideal for applications like semantic search, recommendation systems, and personalization engines. Integrated into the InterSystems IRIS for Health data platform, Vector Search combines high-performance analytics with enterprise-grade reliability, enabling developers to build AI-driven solutions that deliver real-time insights and enhanced user experiences, regardless of dataset size or complexity.

InterSystems vector datatype is designed to support AI and machine learning applications by enabling efficient storage and processing of high-dimensional data. This datatype allows developers to store vector embeddings. With native support for vector operations, such as similarity measurements using cosine distance or dot product, the vector datatype facilitates advanced analytics and semantic searches.

InterSystems distinguishes itself in the Generative AI landscape through the tight integration of vector capabilities within its core multi-model database engine, encompassing relational, object-based, JSON, and other data models. This unified approach enables seamless management of both structured and unstructured data, eliminating the need for separate vector and other databases, reducing data movement and copying, and simplifying application architectures and accelerating application performance.

Additionally, InterSystems IRIS for Health is optimized for low-latency query performance and high-speed data ingestion, facilitating real-time processing essential for applications like semantic search and **retrieval-augmented generation** (RAG). By leveraging built-in vector search functionalities and hardware acceleration, InterSystems ensures that AI-driven applications operate with maximum speed, scale, security, and reliability, effectively meeting the demands of modern enterprise environments.

Retrieval Augmented Generation



1. **Build knowledge base, split into chunks.**
2. **Vectorize with embedding model.**
3. **Insert vectors into vector database.**
4. **Vectorize query with embedding model.**
5. **Search for vectors related to query.**
6. **Retrieve context, put into prompt.**
7. **Get "grounded" answer from LLM.**

Vector Search Examples in Healthcare

These few examples built on the InterSystems platform barely scratch the surface of the range of applications solution and application developers can build with vector search.

Diagnostic Codes

The d[IA]gnosis application exemplifies the integration of InterSystems IRIS for Health vector data types and search functionalities to enhance ICD-10 diagnostic coding through Generative AI. By importing ICD-10 codes from CSV files and utilizing Embedded Python, the application vectorizes diagnostic descriptions using pre-trained language models, storing these vectors within the InterSystems IRIS database. The setup enables efficient **similarity searches** between free-text inputs and standardized codes, streamlining the coding process. The system featured a user-friendly front end for entering texts and orchestrating requests via REST APIs. This application demonstrated how InterSystems IRIS capabilities can be harnessed to develop Retrieval-Augmented Generation (RAG) applications that make healthcare diagnostics more accurate and efficient.

Fraud Detection

InterSystems IRIS for Health can serve as a comprehensive platform for **anti-fraud** initiatives by enabling the collection, enrichment, and unification of transactional and asset data to identify fraudulent patterns, for example, in insurance data. By integrating various data sources and formats, GenAI can automate significant portions of the needed intelligence work and facilitate advanced analytics. Leveraging InterSystems IRIS interoperability features, organizations can develop and implement anti-fraud algorithms using languages like R and Python, apply business rules, deduplicate data, and store multi-modal information for in-depth analysis. This unified approach streamlines fraud detection processes, enhances precision through AI algorithms, and fosters a multidisciplinary, self-service environment for effective fraud detection.

Customer Support

InterSystems built its own Yoky Support Assistant chatbot to leverage **vector-based semantic search** to enhance **customer support** for InterSystems TrakCare. By understanding the context of customer queries, it retrieves relevant knowledge from versrge datasets, providing accurate and timely responses to support advisors and improving efficiency and customer satisfaction.

Patient Engagement

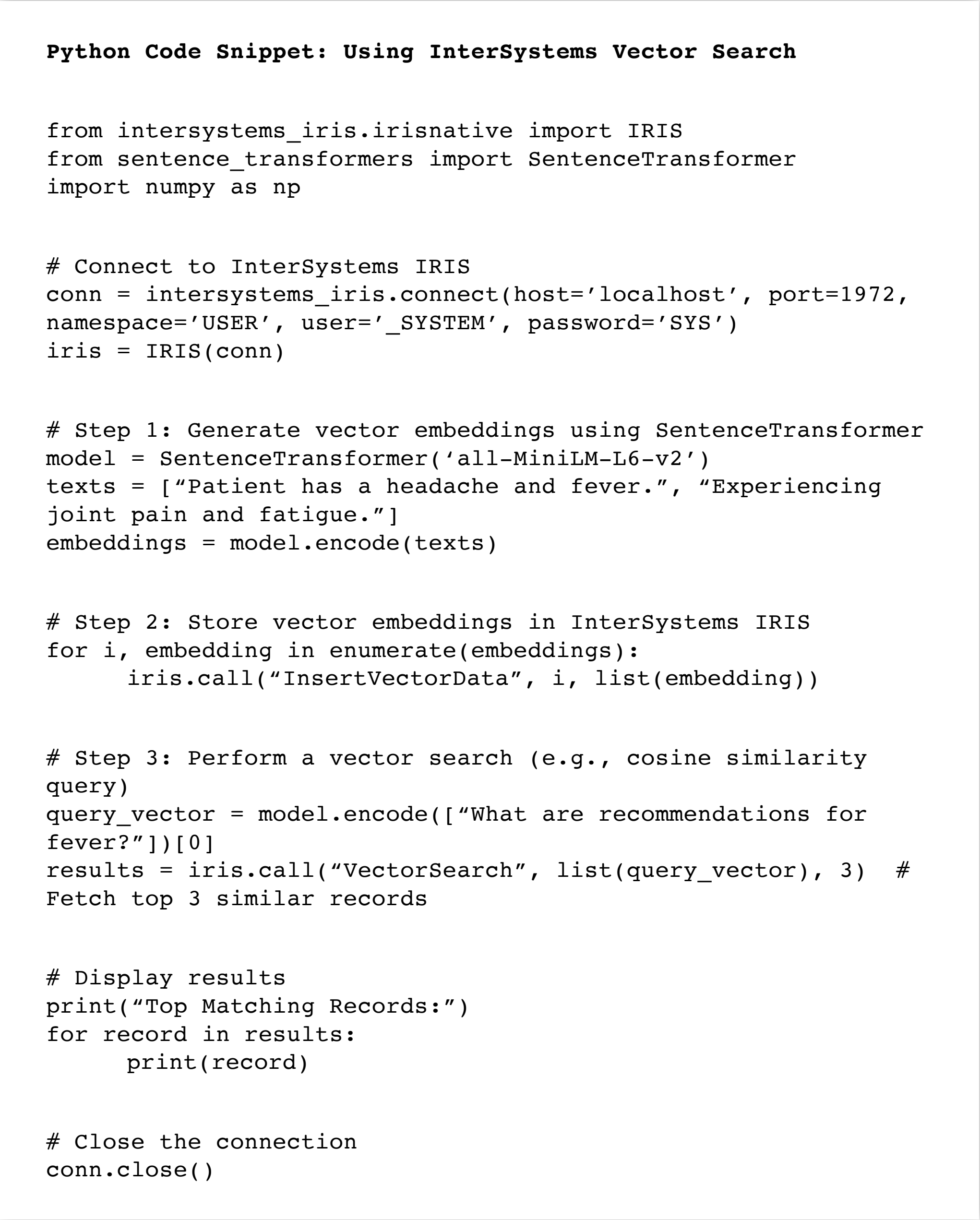
Imagine a healthcare application that enhances patient engagement by providing **personalized wellness recommendations**. Data from patient health histories and wearable devices is ingested into InterSystems IRIS for Health and preprocessed in Python. The data is encoded into vector embeddings to capture meaningful patterns, such as lifestyle habits and symptom trends.

When a patient enters a query about their health, the system uses vector search to find semantically similar records from the database, such as cases with similar symptoms or demographic profiles. The retrieved data is then fed into a pre-trained GPT-based AI model to generate personalized recommendations, such as exercise plans or diet modifications. This workflow ensures that recommendations are both contextually relevant and generated in real time.

Using Vector Search and RAG from Python

InterSystems IRIS for Health has added SQL syntax and Python classes to make development of generative AI applications intuitive and quick, as illustrated by the code snippet below. Most tools and frameworks for generative AI are Python-based. You can use this rich ecosystem directly from embedded Python.

For more on InterSystems IRIS Vector Search and the Python Ecosystem, please watch this video: **https://www.intersystems.com/resources/iris-vecto-search-python-ecosystem/**



Python Code Snippet: Using InterSystems Vector Search

```
from intersystems_iris.irisnative import IRIS
from sentence_transformers import SentenceTransformer
import numpy as np


# Connect to InterSystems IRIS
conn = intersystems_iris.connect(host=’localhost’, port=1972,
namespace=’USER’, user=’_SYSTEM’, password=’SYS’)
iris = IRIS(conn)


# Step 1: Generate vector embeddings using SentenceTransformer
model = SentenceTransformer(‘all-MiniLM-L6-v2’)
texts = [“Patient has a headache and fever.”, “Experiencing
joint pain and fatigue.”]
embeddings = model.encode(texts)


# Step 2: Store vector embeddings in InterSystems IRIS
for i, embedding in enumerate(embeddings):
      iris.call(“InsertVectorData”, i, list(embedding))


# Step 3: Perform a vector search (e.g., cosine similarity
query)
query_vector = model.encode([“What are recommendations for
fever?”])[0]
results = iris.call(“VectorSearch”, list(query_vector), 3)  #
Fetch top 3 similar records


# Display results
print(“Top Matching Records:”)
for record in results:
      print(record)


# Close the connection
conn.close()
```

High-Performance Data Architecture with InterSystems

InterSystems sets itself apart with its **high-performance** and **scalable** architecture, enabling developers to build and maintain large-scale, high-throughput applications with ease. Its proven ability to handle massive amounts of **transactional** (OLTP) and **analytical** (OLAP) workloads ensures reliability even in the most demanding environments. Whether supporting thousands of healthcare systems globally, InterSystems delivers exceptional performance and scalability, allowing developers to create applications that can grow seamlessly alongside business needs. Its database engine is optimized for speed and efficiency, making it the backbone of mission-critical solutions in industries where downtime or delays are unacceptable.

Interoperability is a hallmark of InterSystems, offering seamless integration with a broad range of data formats and external systems, enabling the smooth exchange of data across disparate systems, and ensuring that developers can unify information silos. With built-in support for healthcare standards, such as the widely used HL7® and FHIR (Fast Healthcare Interoperability Resources), as well as many others, InterSystems platforms are exceptionally well-suited for developers working in the healthcare sector. InterSystems also supports **real-time data streaming** and **analytics**, allowing developers to build solutions that react instantaneously to changes and deliver actionable insights as they happen. This combination of interoperability and real-time capabilities empowers developers to create sophisticated, integrated solutions that bridge the gap between diverse technologies and systems.

Security and **compliance** are integral to InterSystems offerings, providing developers with robust tools to meet the stringent regulatory requirements imposed on health data. With features such as data encryption, user access controls, and audit trails, developers can build applications that prioritize data protection and regulatory compliance. Additionally, InterSystems fosters a thriving developer ecosystem by offering a suite of tools for debugging, deployment, and collaboration, along with comprehensive documentation and support and a thriving developer community. This ensures that developers have the resources they need to efficiently build, test, and deploy innovative solutions, making InterSystems a trusted partner for developers in any IT industry.

Cost-Effective GenAI at Scale with InterSystems

InterSystems revolutionizes the development of Generative AI applications by providing a unified platform that integrates robust data management, high-performance analytics, and advanced vector search capabilities. Its support for multi-modal data and native Python integration enables developers to seamlessly combine AI frameworks with real-time data processing, creating intelligent applications across the healthcare sector. Whether it's embedding data for semantic search, training AI models, or building scalable, multi-modal solutions, InterSystems offers an ecosystem designed for efficiency and reliability.

If you're a health information application or solution developer looking to explore scalable and AI-enabled workflows, InterSystems provides an ideal foundation. Its vector datatype and search capabilities allow you to build personalized recommendations, fraud detection systems, or AI-driven customer support tools with ease. Start by diving into our comprehensive technical documentation or engaging with the developer community for tutorials and expert advice. Unlock the potential of Generative AI and create solutions that not only meet today's challenges but are prepared for tomorrow's opportunities.

InterSystems IRIS for Health: The InterSystems Difference

InterSystems Knows Healthcare

As a longstanding leader in healthcare data technology and standards-based interoperability, InterSystems has the experience to solve real-world healthcare challenges.

Industry-Leading Support

We're focused on making our customers successful and ready for any challenge, which is demonstrated by some of the highest customer satisfaction ratings in the category.

Comprehensive Healthcare Interoperability

Seamless Integration Connects You into the Greater Health and Care Ecosystem

Unique Architectural Approach

Our integrated, interoperable, multi-model, multi-lingual engine provides the highest performance and resiliency with the lowest TCO.

We Bring the Processing to the Data

Our approach means less moving of data. That means less potential for data errors, faster processing, greater security, and lower costs.

Highly Flexible

InterSystems IRIS includes the tools to solve unfamiliar problems and adapt as business needs change. Every aspect from data transformations to workflow can be tailored, and low-code tools let you put some customization in the hands of business users.

Powering the World's Most Important Applications

Our software powers mission-critical applications in almost every industry — from healthcare and financial services to supply chain and space exploration.

About InterSystems

InterSystems, a creative data technology provider, delivers a unified foundation for next-generation applications for healthcare, finance, manufacturing, and supply chain customers in more than 80 countries. Our cloud-first data platforms solve interoperability, speed, and scalability problems for large organizations around the globe to unlock the power of data and allow people to perceive data in imaginative ways. Established in 1978, InterSystems is committed to excellence through its award-winning, 24×7 support for customers and partners in more than 80 countries. Privately held and headquartered in Boston, Massachusetts, InterSystems has 39 offices in 28 countries worldwide.

For more information, please visit **InterSystems.com**.

InterSystems®
Creative data technology