KDB.AI - Enabling AI-driven Data Immediacy for GenAI Applications

Unchecked growth in data diversity

Data pundits have been measuring and predicting the explosive growth of structured data and streaming data across enterprises. However, what has been happening, almost unsuspectingly, is that unstructured data has been growing at a rate of 55% - 65% across enterprises. Today, 80% of enterprise data is unstructured, hiding a vast amount of valuable information. This leads us to the problem statement - there is an unchecked growth in data diversity and the current set of data platforms is largely unsuited to handle them efficiently.

Real-time enterprise decisions require data immediacy

Data immediacy is the requirement to get real-time access to all the key data points to make mission-critical business decisions that truly impact business performance, customer retention or other such key metrics. This is not just about latency pertaining to data access. While that is important, there are other aspects such as context, correlation, data store location, data quality, ability to handle diverse datasets, etc., that will impact the decision-making process.

This problem became multi-fold complex as enterprises started seeing the massive amounts of valuable information hidden in unstructured data such as PDFs, images, videos, and audio files. Now, enterprises realize they need multi-modal search capabilities to scour all this information, correlate the right data points within the right context, and provide key insights for critical business decisions. And, all this has to be done in real-time. Areas such as Risk & Compliance or Fraud Detection within the financial services sector have prime use cases for data immediacy.

Key considerations for new data architectures

As companies start realizing this problem and they start redefining their data architectures to prepare for more modern GenAI applications, there are five key considerations to be taken into account:

- **Accuracy** - What types of measures and features are being put in place to ensure that your system is delivering the most accurate information possible? With today's GenAI applications still being prone to hallucinations, this is a fundamental requirement.
- **Latency** - What does real-time mean to you? In some use cases, there is an acceptable level of latency when it comes to decision-making but in others like fraud detection, even a few milliseconds of latency can have severe fiscal consequences.
- **Initial setup cost/times** - When you have a use case (e.g. in Insurance) to scan and upload 600K PDFs to power up a GenAI system, the initial setup time / costs should be taken into account to ensure timely availability of the system to be in place.
- **Runtime** - Once your system is up and running, it is equally important to ensure that it can handle the scale of incoming users and data. This is also a critical factor when designing your data architecture. Its performance will directly affect other factors, such as latency and accuracy.
- **Data Freshness** - The value of your data is only as good as your ability to harness it in real-time. The longer you wait to process your data, the lower its value. So, in order to make the best real-time decisions, it is critical that you retain the freshness of your data and use it as soon as possible.

Vector databases to the rescue

Vector databases help with data immediacy and enable today's GenAI applications with contextual search at scale in several key ways:

- **Efficient storage and retrieval of high-dimensional vector data.** Vector databases are optimized to store and quickly retrieve vector embeddings that capture the semantic meaning of data like text, images, and audio. This allows GenAI systems to find relevant information based on similarity rather than just keyword matches.

- **Enabling semantic search.** Vector databases allow searching based on meaning and context, not just exact keyword matches. This is crucial for understanding natural language queries and retrieving relevant information for GenAI applications like chatbots and question-answering systems.
- **Serving as external memory for GenAI models.** Large Language Models (LLMs) used in GenAI are often stateless and lack long-term memory. Vector databases can store relevant information that can be quickly retrieved to provide context and reduce the risk of hallucinations or inaccurate responses.
- **Scalability to handle large datasets.** Vector databases are designed to scale to billions of data objects while still providing fast similarity search performance. This allows GenAI applications to work with massive amounts of structured and unstructured data.
- **Integration with the GenAI technology stack.** Vector databases can be tightly integrated with LLMs and other components of the GenAI stack. This allows developers to easily retrieve relevant information from the vector database and feed it into the GenAI model to enhance its outputs.

KDB.AI

KDB.AI is a powerful knowledge-based vector database and search engine that allows you to build scalable, reliable AI applications, using real-time data, by providing advanced search, recommendation, and personalization. Multi-Modal RAG (Retrieval Augmented Generation) is an advanced AI technique that combines the capabilities of LLMs with the ability to retrieve and utilize information from various data sources (audio, text, video, etc.) This approach is particularly beneficial for enterprises as it allows AI systems to provide more accurate, detailed, and contextually relevant responses from public and private datasets. KDB.AI fully supports multi-modal embeddings with seamless hybrid search capabilities and enables building RAG applications.

KDB.AI sets itself apart from the competition in the following ways -

- **Dynamic hybrid search:** Combine similarity, exact, and literal search in a single query where query results remain relevant with content changes.
- **Multimodal RAG:** Handle the data diversity challenges of GenAI by modeling unstructured data such as text, video, audio, & images.
- **Behavioral analytics:** Spot trends, patterns, and anomalies in your time-oriented data, such as data from IoT sensors, market-related insights, and more.

When should you use KDB.AI?

- Processing, searching, and analyzing unstructured data such as videos, images or audio
- Anomaly detection or forecasting future events
- Real-time decision-making with use cases like fraud detection
- Integrating with LLMs and RAG applications for more accurate search results
- Need to build applications that integrate both structured and unstructured data

Next-generation GenAI and RAG applications are how you deliver data immediacy in today's world and they require searching across structured and unstructured data sources. Vector databases such as KDB.AI help with processing such diverse data sources and in detecting anomaly detection or enabling predictive analytics.

Dinesh Chandrasekhar has been a data management veteran, thought leader, and data practitioner for 30+ years. As the chief analyst and founder of Stratola, he speaks and writes on the latest topics in data-in-motion, real-time analytics, IoT, Observability, GenAI, and more. Follow his work at www.stratola.com.