



What The? Another Database Model - Vector Databases Explained

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

 /in/wmcknight

www.mcknightcg.com
(214) 514-1444



McKnight Consulting Group Partial Technology Implementation Expertise

Big/Analytic/Vector/Mixed Data Management



Data Movement and APIs



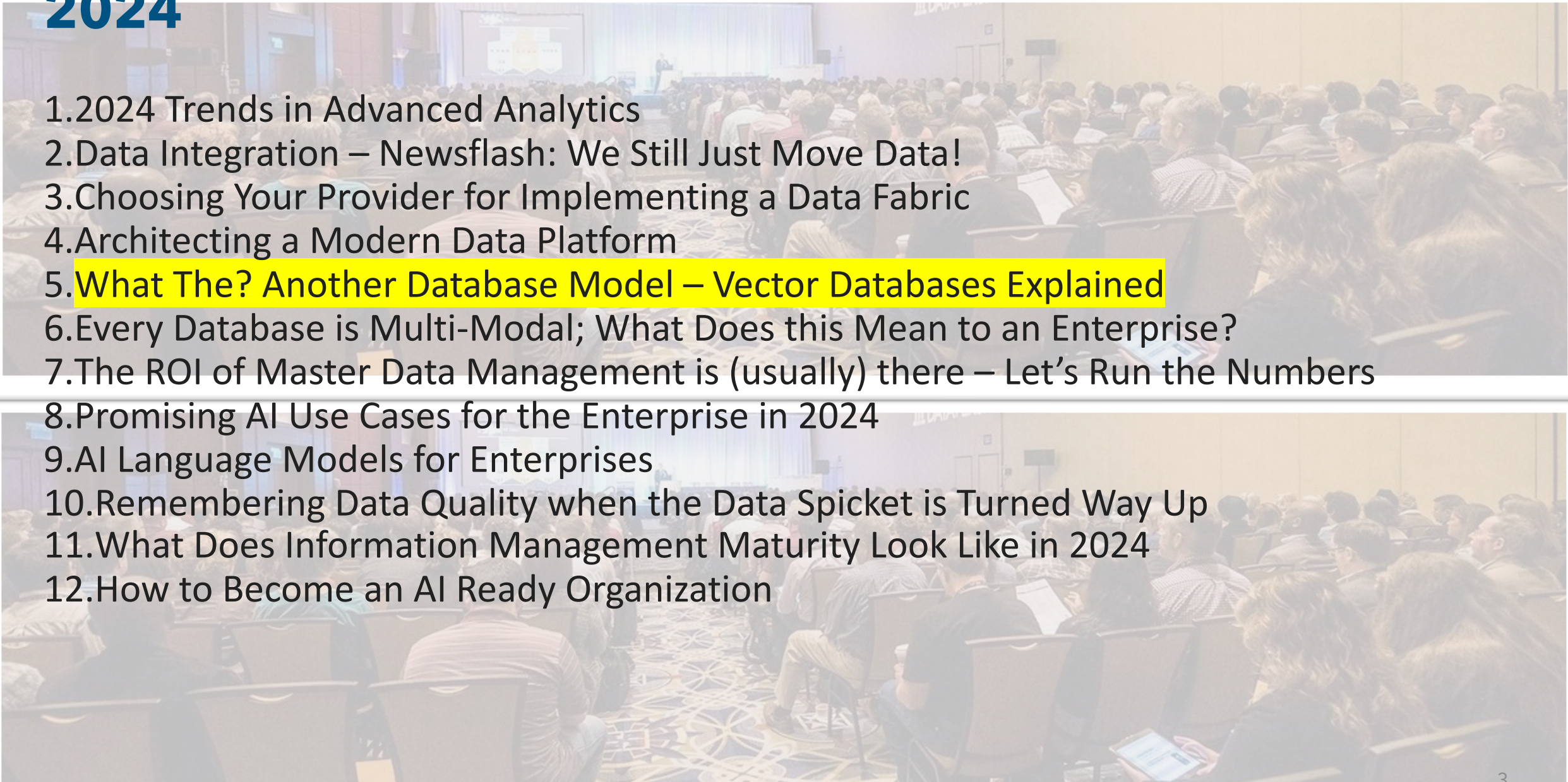
Data Management



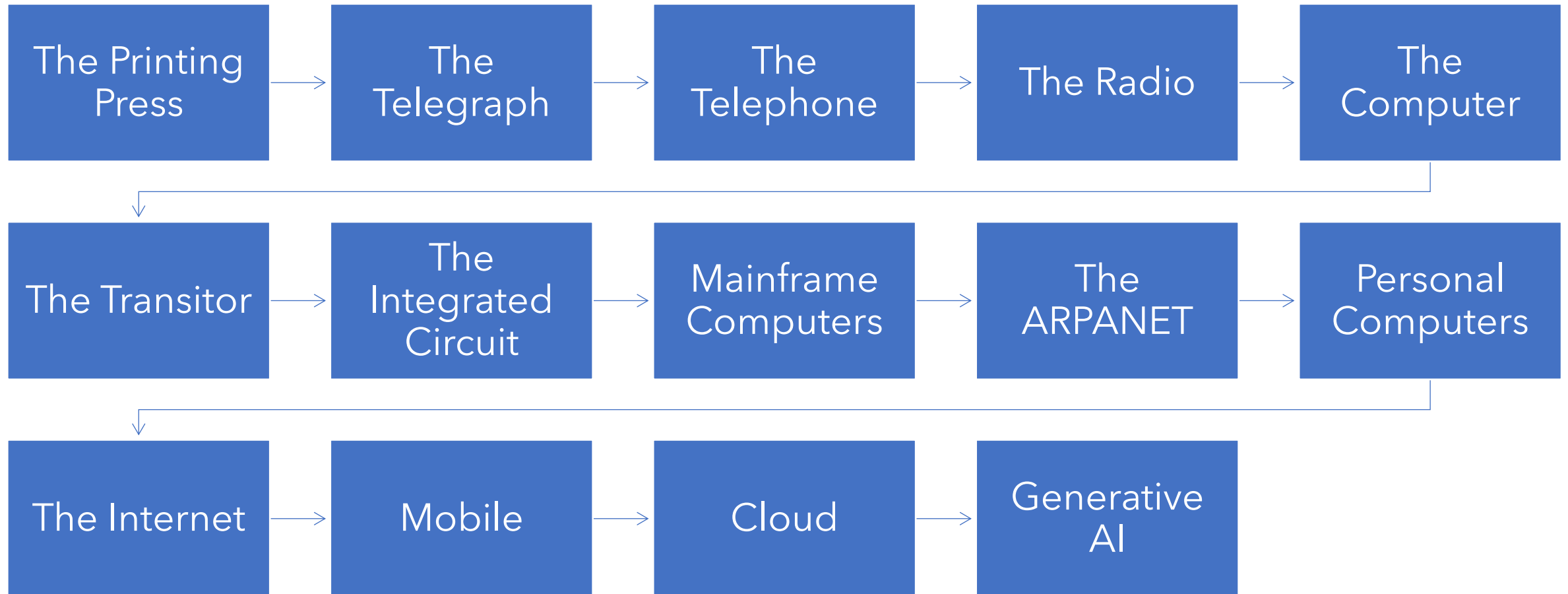
Operational/Transactional Data Management



Dataversity Advanced Analytics with William McKnight 2024

- 
1. 2024 Trends in Advanced Analytics
 2. Data Integration – Newsflash: We Still Just Move Data!
 3. Choosing Your Provider for Implementing a Data Fabric
 4. Architecting a Modern Data Platform
 5. What The? Another Database Model – Vector Databases Explained
 6. Every Database is Multi-Modal; What Does this Mean to an Enterprise?
 7. The ROI of Master Data Management is (usually) there – Let's Run the Numbers
 8. Promising AI Use Cases for the Enterprise in 2024
 9. AI Language Models for Enterprises
 10. Remembering Data Quality when the Data Spicket is Turned Way Up
 11. What Does Information Management Maturity Look Like in 2024
 12. How to Become an AI Ready Organization

The Big Technology Waves 1440-Present



AI Interest is at an all-time High and Growing

Hundreds of companies will be built around an API for something like ChatGPT



Startups will not be able to create the AI themselves, but they can use the APIs



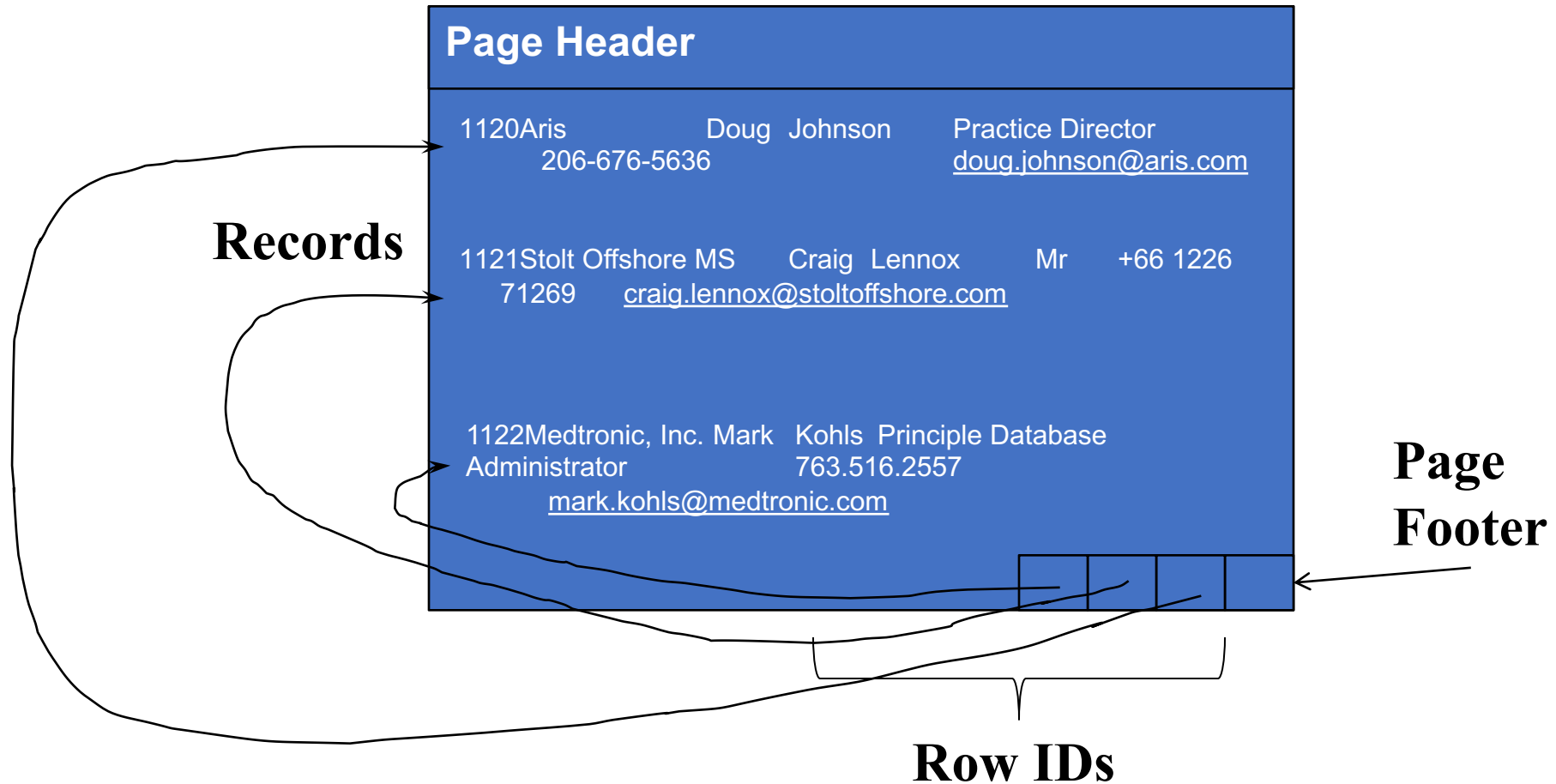
Nearly every industry and nearly every vertical is being transformed today



Companies are using these techniques in software and statistical models to make predictions and drive businesses forward in a way that they're not able to with only humans



The Relational Database Data Page

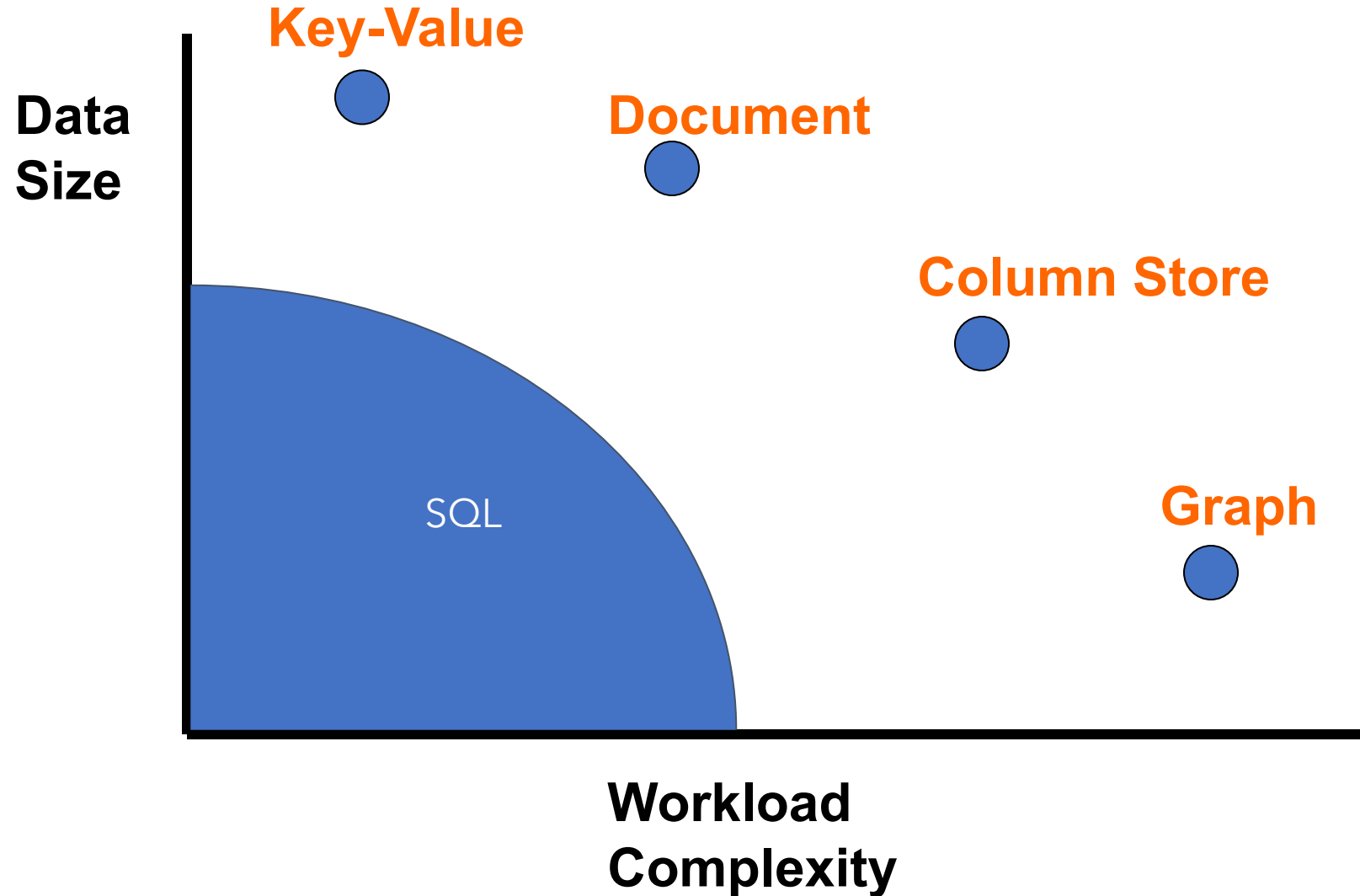


Why NoSQL for Big Data

- More data model flexibility
 - JSON as a data model
 - No “schema first” requirement; load first
- Faster time to insight from data acquisition
- Relaxed ACID
 - Eventual consistency
 - Willing to trade consistency for availability
 - ACID would crush things like storing all sensor reads
- Freedoms

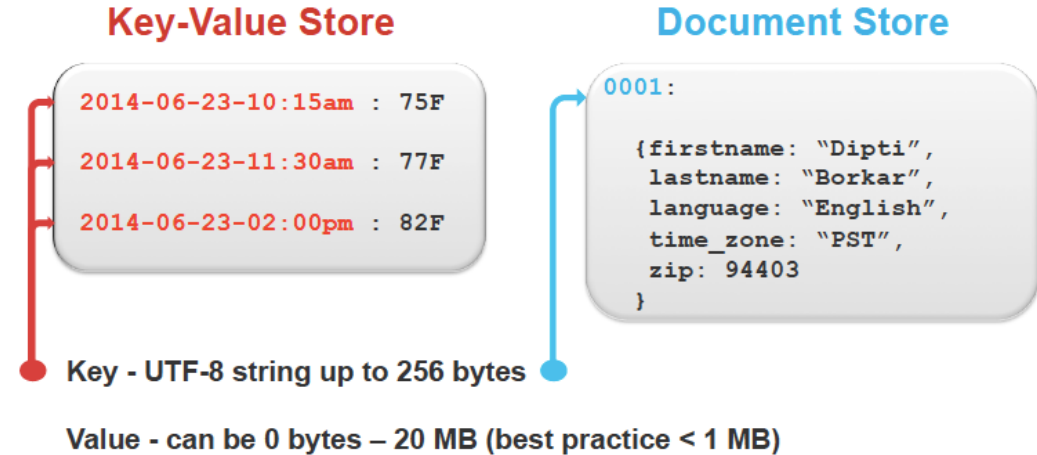


Operational Big Data Platform Selection



Many Data Types

- Web Crawlers
- Open Linked Data
- JSON
- XML
- Documents
- Binary
- Graph
- Log Files



Enter Vector Databases

- Conventional databases lack the structural capability to accommodate imprecise comparative inquiries such as "which items are comparable to this one?"
- The exploration of machine representations of datasets such as text, voice, image, and molecular structures is underway as ML and LLMs are applied to novel problems.
- Vector emerged to address new issues, much like the nosql generation of databases did.
- User inquiries evolved in tandem with the influx of machine representations of data.
- To address them, vector databases, a novel technology, were required.



The Struggle of Imprecise Search

- Ever searched for something with unclear detail?
- Struggling to search because you lack the exact keywords?
- Remembering actors but forgetting the movie title?
- Frustrated with outdated information in search results?
- Frustrated by generalized Large Language Models (LLMs)?
- If any of these sound familiar, then vector search can be your solution



Where are we in the new world?

- **Generative AI:** AI models that can create new data, like text, code, or images. LLMs (Large Language Models) are a type of generative AI that focus on text.
- **LLMs (Large Language Models):** These are AI models trained on massive amounts of text data. They can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. However, LLMs can sometimes struggle with factual accuracy and staying on topic.
- **Vector Databases:** These are a special kind of database designed to store and search for information represented as vectors. Vectors are mathematical structures that capture the relationships between data points. For generative AI, these vectors might represent the meaning or key aspects of text, code, or images.
- **RAG (Retrieval-Augmented Generation):** This is a technique that combines LLMs with vector databases. The process:
 1. The LLM receives a prompt or question.
 2. The RAG system uses the vector database to find similar information related to the prompt.
 3. The LLM uses the retrieved information to improve its understanding of the context and generate a more accurate and relevant response.

DBMS are adding Vector Support



- The demand for AI technology is experiencing explosive growth, driven by real-world applications solving critical problems.
- Vector data has a wide range of use cases, ultimately limited only by creativity and imagination.
- Isolated data sources (silos) can limit the accuracy and efficiency of analytics, including those involving vectors.
- Developers are overwhelmed by managing an abundance of individual tools and interfaces.
- Both developers and enterprises prioritize consolidation of their technology stacks for better manageability.
- The ideal scenario is a single interface offering broad capabilities across various data model problems, without sacrificing functionality.

Vector Search



Introducing Vector Search

- Search based on meaning, not just keywords
- Leverages machine learning models called encoders for powerful results
- Embeddings
 - Text, audio, images transformed into a numeric string called "vectors"
 - High-dimensional arrays with semantic meaning
 - Makes data available to AI
- Benefits of Vector Search
 - Semantic Understanding
 - Scalable
 - Flexible – anything can be vectorized
- **Example: [-0.0385810, -0.1348581, 0.0184810, -0.138542, -0.1984815, 0.12498134, 0.0124897, -0.021858, -0.0002384, -0.024911, 0.199248284,]**

What Are Vectors

- Numeric representations
- "The only thing we have to fear is fear itself" = $[-2.345, 7.812, -1.009, 4.567, 0.123, 9.998, \dots]$
- "To be or not to be, that is the question" = $[-3.141, 2.718, -8.000, 1.618, 4.200, 7.539, \dots]$
- "I think, therefore I am" = $[-5.827, 0.123, 9.000, -1.414, 3.142, 6.789, \dots]$



Vector Embeddings come from an Embedding Model

- Example Embedding Models are OpenAI, Cohere, Google Vertex, HuggingFace



Word2vec for text

- Word2Vec is a technique for representing words as numerical vectors.
- These vectors capture the semantic meaning and relationships between words.
- Word2Vec models are trained on massive datasets of text.
- The model analyzes the context in which words appear, learning their meaning based on surrounding words.
- Words with similar meanings have similar vector representations in the embedding space.
 - Imagine "king" and "queen" having vectors close together, while "king" and "car" would be farther apart.
- Word2Vec embeddings fuel various applications:
 - Recommendation systems suggesting similar products based on user searches.
 - Machine translation finding the closest meaning in another language.
 - Chatbots understanding the intent behind user queries.

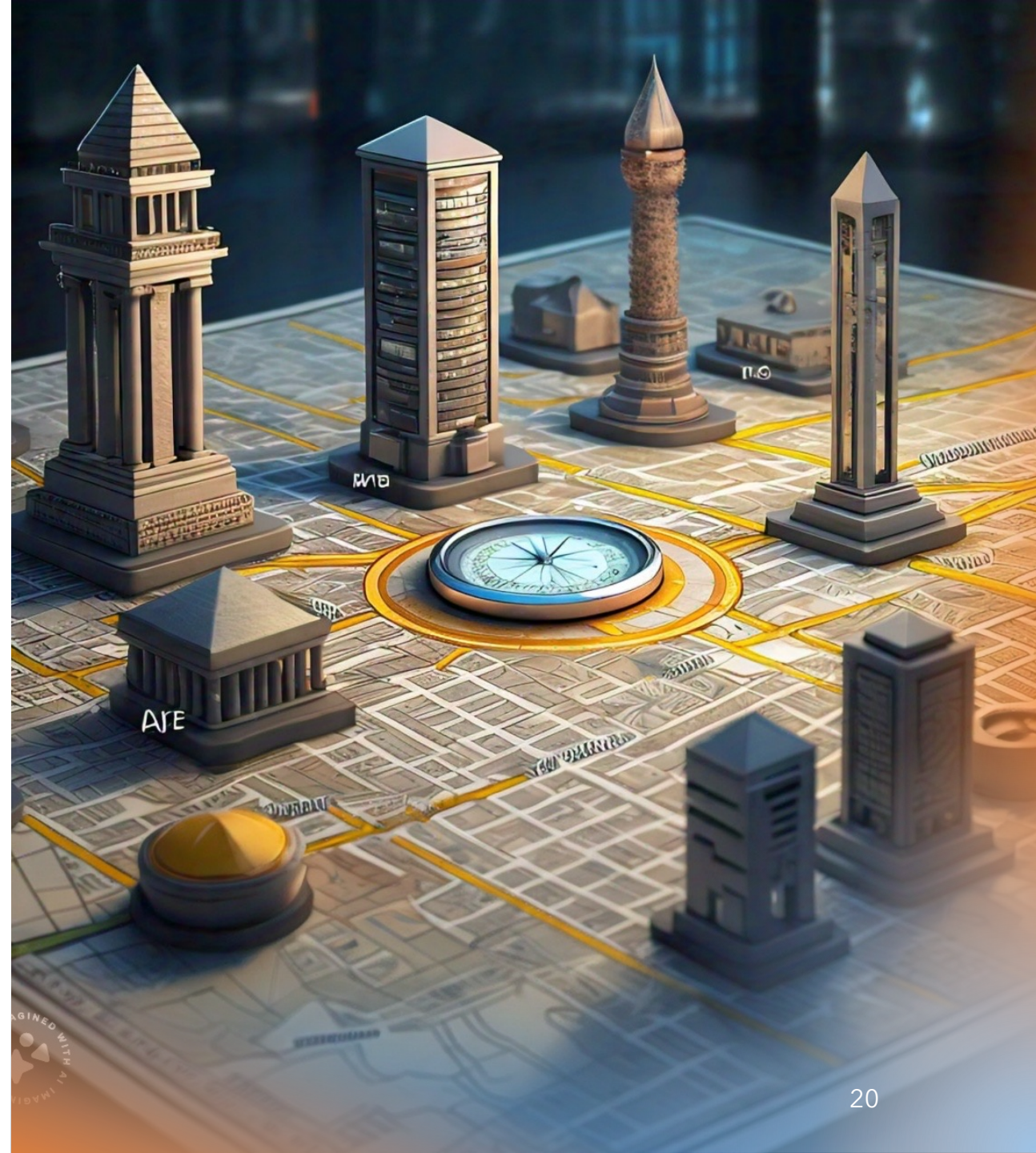
The Power of Similarity Search

- In high-dimensional space, vectors that are close to each other show information that is similar
- Adds to standard keyword search to get more detailed results
- Vector search helps LLM learn more than just what they learned in training
- Get useful results even if your queries aren't very specific

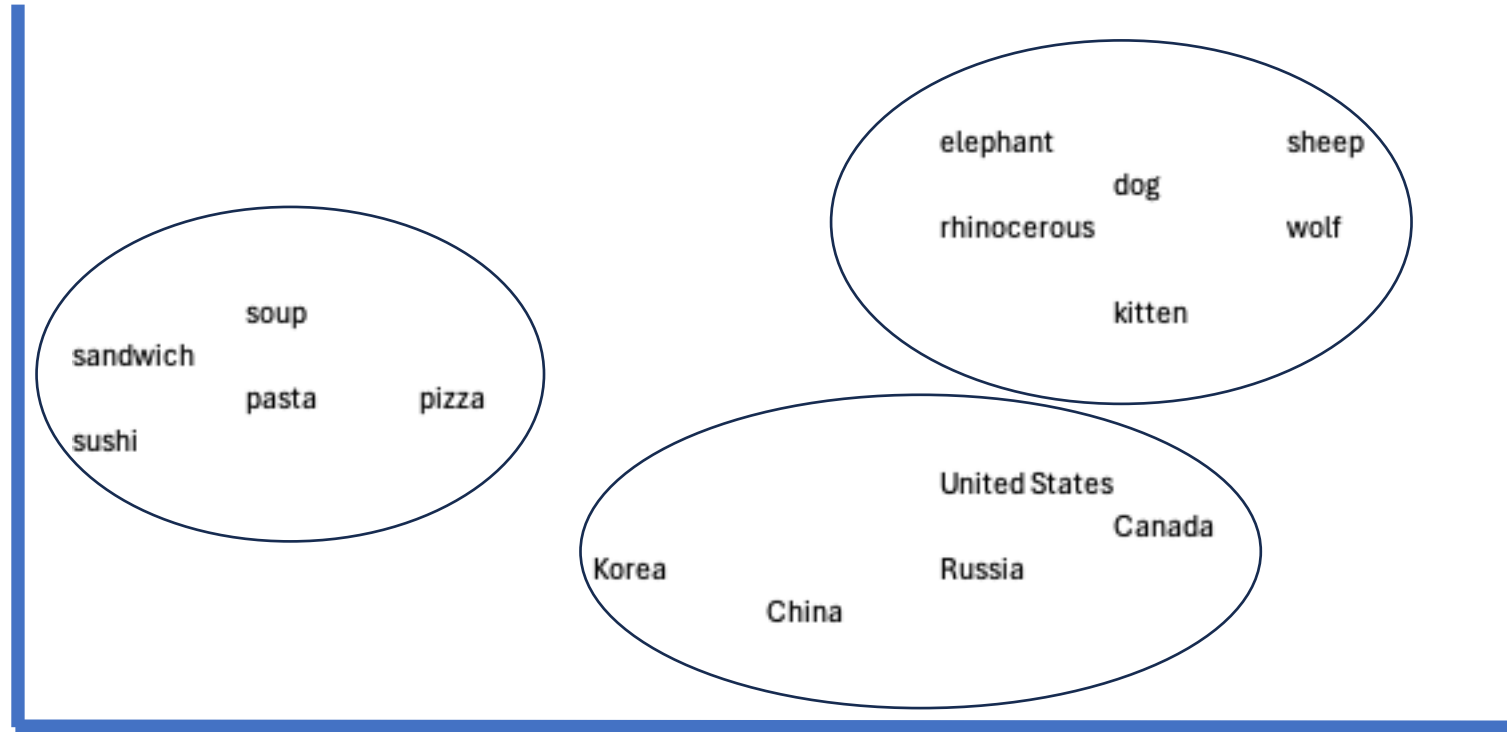


Vector Databases do other, normal database functions

- Database will partition or shard vector indexes for improved performance
- Database will transparently scale vector processing across the computers in a RAC cluster with full data consistency



Data Clustering in Space





Similarity Functions

- Eucliden
 - Measures the straight-line distance between two points in a multidimensional space.
 - Commonly used for numerical data with real number values. Larger distance indicates less similarity.
- Cosine
 - Measures the directional similarity between two vectors. Useful for data where the magnitude (length) of the vectors may not be important.
 - Values range from -1 (completely opposite) to 1 (identical).
- Dot Product
 - Calculates the product of corresponding components of two vectors and then sums them up.
 - Closely related to cosine similarity, but it considers the magnitudes of the vectors as well.
 - Larger dot product indicates greater similarity.

Vector Search



- By using the distance between high dimensional vectors, vector search enables us to search through data and find relevant results.
- When calculating the distance between your query and stored vectors at query time, it uses machine learning models to embed the data.

Simple Vector_Distance Function

SELECT...

FROM house_for_sale

WHERE price <= (SELECT budget FROM
customer ...)

AND city in (SELECT search_city FROM
customer ...)

ORDER BY vector_distance(house_vectors,
:input_vector);



Vector Databases vs Graph Databases

- Graph databases are better suited to processing data with complex relationships, whereas vector databases are better suited to handling high-dimensional data, such as images and video
- Graph databases are made for queries involving relationships, while vector databases excel at similarity searches
- Graph databases may not be as scalable due to the complexity of the data model

K-Nearest Neighbor

- Closest thing to exact search for vectors; it finds the perfect nearest neighbors.
- Typical approach to classification problems is k-nearest neighbors (KNN) Classification.
- KNN predicts the label (class) or value (regression) for a new data point by looking at its k nearest neighbors in the training data.
- Contrary to the notion of "exact search," KNN doesn't necessarily find perfect matches in the training data.
- It focuses on identifying the k data points most similar (closest) to the new point based on a chosen distance metric (e.g., Euclidean distance).



Hierarchical, Navigable Small World

- HNSW utilizes a graph-like structure with layers, enabling efficient traversal to find approximate nearest neighbors.
- Hierarchical Navigable Small World (HNSW) is an indexing strategy for Approximate Nearest Neighbor (ANN) search.
- HNSW enables fast retrieval of "mostly" nearest neighbors for K-Nearest Neighbors (KNN), improving efficiency for large datasets.
- HNSW utilizes a graph-like structure with layers, where similar data points are connected within layers.
- During a search, the algorithm traverses the layers of the graph, efficiently navigating towards potential nearest neighbors.
- HNSW offers a trade-off between finding the absolute closest neighbors (accuracy) and achieving high search speed.
- It is ideal for real-time applications with large datasets where KNN needs a performance boost.



GenAI and Vector Use Cases

Questions for Vectors

- Given a user's search query for a product (e.g., "running shoes"), which product descriptions in the online store have the **most similar** vector representations, indicating potential matches for the user's needs?
- Music Recommendation Systems: Based on a user's listening history and the vector embeddings of various songs, which songs are **most likely** to be enjoyed by the user due to their similar musical characteristics captured in the vector space?
- Customer Segmentation: Considering customer purchase history and demographics represented as vectors, which customers have **similar buying habits** and preferences, allowing businesses to create targeted marketing campaigns?
- Anomaly Detection in Sensor Data: In a network of sensors monitoring various parameters (e.g., temperature, pressure), **which sensor readings deviate significantly from the norm** in the vector space, potentially indicating an anomaly or equipment malfunction?
- Fraud Detection in Financial Transactions: By analyzing transaction features like amount, location, and time as vectors, which transactions have a **high probability of being fraudulent** due to their dissimilarity to typical spending patterns?
- Social Network Analysis: Considering user profiles and interactions represented as vectors, which **users within a social network have similar interests** and connections, revealing potential communities or clusters?
- Text Summarization: Given a lengthy document represented as a sequence of word vectors, which subset of sentences (also represented as vectors) best **captures the overall meaning** and key points of the document?
- Image Captioning: Based on the visual features of an image extracted and converted into a vector, can a model **generate a textual description** (caption) that accurately reflects the content of the image?
- Video Retrieval: Given a user's query describing a desired video scene (e.g., "car chase"), can a system efficiently **retrieve videos** from a large database by comparing the user's query vector with the vector representations of video content?
- Medical Diagnosis: By analyzing patient medical records and symptoms represented as vectors, can a system **identify similar cases** and relevant medical knowledge to assist doctors in diagnosis and treatment planning?

Applications of Vector Search in Industry - Common

- Find, analyze and summarize financial reports and documents
- Recommendation Systems
- Process and generate personalized bills and marketing content
- Virtual chatbots
- Enable staff to search laws, regulations and internal repositories
- Triage, categorize and summarize customer emails and text messages
- Process real-time data from manufacturing and inspection machines
- Suggest personalized content to users based on browsing history, behaviors and comparisons to similar users
- Customer service and support
- Marketing personalization
- Analyze and summarize product reviews
- Answering questions in the call center
- Automate the creation of marketing content
- Analyze anonymized call transcripts to improve customer service
- Automate supply chain negotiations

Applications of Vector Search by Industry - Specialized

- Fraud detection
- Streamline paperwork (i.e., healthcare)
- Identify the root cause of loss (i.e., insurance)
- Generate accurate legal content
- Generate protein sequences and DNA sequences (life sciences)
- Continuously monitor, summarize and analyze telemetry from equipment to reduce downtime and improve utilization
- Determine best ingredient combinations (food)
- Scan a broad range of media sources and highlight relevant mentions
- De-age actors (movies)
- Create new virtual worlds and characters (gaming)
- Review satellite imagery to detect unsustainable trends

Applications of Vector Search in Company

- Copilots
- Customer service
- IT support
- Employee onboarding
- Chip design
- Drug discovery
- Fraud investigation
- Screening resumes
- Answering common employee questions
- Curriculum design
- Demand planning for supplies
- RFP Automation
- Document and research summarization
- Patent drafting
- Chatbot for support
- Ticket classification
- Logistics optimization
- Supplier Analysis

Gen AI and LLM Success



- Organizations will successfully implement GenAI and large language models without facing knockout challenges related to data quality, governance, ethical compliance, and cost management
- These challenges can be mitigated through robust data collection processes, stringent governance frameworks, and continuous monitoring of ethical guidelines
- Additionally, organizations can leverage advancements in cloud computing and AI infrastructure to optimize cost management while ensuring seamless implementation of GenAI and large language models
- Generative AI apps, like chatbots, will become commonplace in daily tasks, with large language models being a powerful and user-friendly form of AI
- Major data and compute platforms are integrating these APIs into the enterprise, fostering creativity and innovation in the future



Analyzing Vector Databases

The Critical Dimensions of Vector Database Performance

- Throughput, latency, F1 recall/relevance, and TCO
- We tested throughput, which involved generating vectors and labels, inserting them into databases, and executing queries to measure performance.
 - The queries were of various types, such as nearest neighbor, range, KNN classification, KNN regression, and vector clustering.
 - We also performed latency testing, which measured the response time for each query.
 - Finally, F1 recall/relevance testing measured the database's performance in returning relevant results for a given query.



Testing Datasets

Dataset	# of Dimensions	# of Vectors
GloVe 25	25	1,183,514
GloVe 50	50	1,183,514
Last.fm	65	292,385
DEEP1B	96	9,990,000
GloVe 100	100	1,183,514
GloVe 200	200	1,183,514
E5 Multilingual	384	100,000
E5 Small	384	100,000
Vertex AI Gecko	768	100,000
E5 Base	768	100,000
E5 Large	1,024	100,000
OpenAI Ada-002	1,536	100,000

Source: GigaOm 2023

Liveness and Relevancy

- 2 workloads: liveness and relevancy.
- Liveness is defined as creating an empty table and storage-attached index (SAI) on Astra DB and an empty pod on Pinecone, loading and ingesting one of the previously defined data sets, and immediately following up with a battery of search queries.
- Each system does indexing on its own terms.
- Thus, the indexing state immediately after ingesting is what we are testing.
- The relevancy test simply repeats the search query battery after the data is fully loaded and indexed.

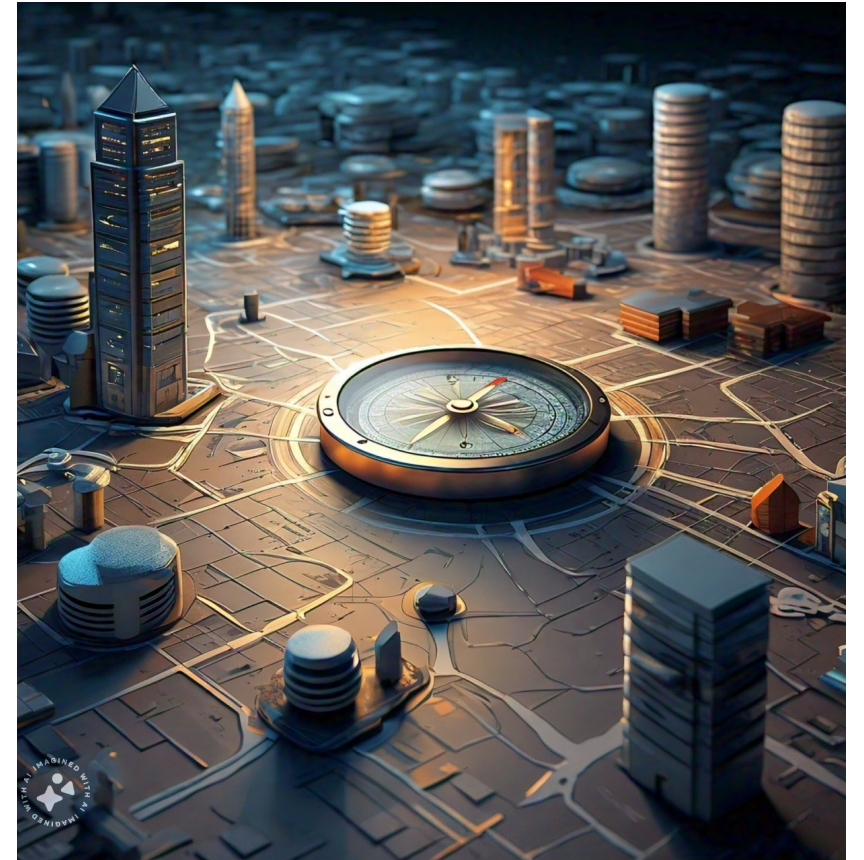
Indexing Performance

- Time to index!
- It took some time before the data was fully indexed and made available for searching.
- Want predictability across datasets



Search Performance and Recall (Accuracy)

- We tested search performance during active ingest (liveness phase) and after the data was fully indexed (relevancy phase)
- Of course, the fastest query execution times are irrelevant if the data returned is unreliable and inaccurate
- $F1 \text{ is } 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$



TCO

- With any total cost of ownership projection, you have to make numerous assumptions.
- For our calculations, we used the following constants for our RAG-based use case.

TCO time period: three years

Dataset Dimensionality: 1,024 dimensions

Cardinality (new vectors per month): 10,000,000

Average queries per second: 50 QPS

Write Size: 5,120 bytes per vector ingested

Astra DB Storage-Attached Index Size: 5,120 bytes per vector ingested Read Size: 51,200 bytes per query

TCO Scenarios

- **Scenario 1: New Dataset Every Month**

- In this scenario, we consume 10 million new vectors every month. At the beginning of each month, we add the new vectors.

- **Scenario 2: Continuous New Dataset Every Week**

- In this scenario, we still consume 10 million new vectors each month, but the new datasets are added weekly to bring the model up-to-date more often.

- **Scenario 3: Near-Real-Time Data Ingest**

- In this scenario, we still ingest 10 million new vectors each month, but the data is added as it is received, so it can be leveraged in near-real time.

Summary

- The concept behind vector databases and their ability to handle unstructured data is relatively simple.
 - Assume your company possesses an extensive collection of textual documents. You want to develop a chatbot capable of responding to inquiries that pertain to the specified documents; however, you do not want the chatbot to have to read every document to do so.
- Storing documents in a vector database is optimal for this and numerous other types of applications, where data is stored as high-dimensional vectors - mathematical representations of the features or attributes of an object.
- Vector databases are designed to efficiently store and search high-dimensional data. They use graph embeddings, which are low-level representations of items, created using machine learning algorithms.
- Embeddings are ideal for fuzzy match problems and work with a variety of algorithms. Items that are near each other in this embedding space are considered similar to each other in the real world.
- Vector databases need to be scalable, performant, and versatile, and they are increasingly popular for a variety of applications involving similarity and generative AI.
- Many databases are also adding vector search functionality to their data cloud platform.





What The? Another Database Model - Vector Databases Explained

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

 /in/wmcknight

www.mcknightcg.com
(214) 514-1444

