Top 25 Vector Database Interview Questions and Answers

5 min read · May 27, 2025

Sanjay Kumar PhD

1. **What is a vector database and how is it used in machine learning systems?**

A vector database is a specialized database designed to store and retrieve high-dimensional vectors — often generated as embeddings from machine learning models. It supports fast similarity search, enabling tasks like semantic search, recommendation, and anomaly detection. These databases use indexing techniques (e.g., HNSW, LSH) to optimize approximate nearest neighbor (ANN) search over millions or billions of vectors.

2. What are vector embeddings and how are they generated?

Embeddings are dense numerical representations of data (e.g., text, images, audio). They are generated using models like BERT (for text) or ResNet (for images). The idea is to capture semantic or contextual similarity, allowing similar items to be close in the vector space. Embeddings enable machines to "understand" complex data relationships through geometry.

3. What are common use cases for vector databases in production systems?

- **Semantic Search**: Retrieving documents/images based on meaning, not keywords.
- **Recommendation Engines**: Suggesting similar products, users, or content.
- **Anomaly Detection**: Finding outliers in network or behavioral data.
- **Multimodal Search**: Combining modalities (e.g., text + image) to find relevant matches.
- **Conversational AI**: Retrieving contextually relevant responses in RAG systems.

4. How does a vector database differ from a relational or NoSQL database?

Relational DBs organize data in rows and tables with strict schema, ideal for structured queries (e.g., SQL). Vector DBs focus on unstructured/semi-structured data, enabling similarity search over high-dimensional data using ANN algorithms, not joins or filters. NoSQL DBs offer schema flexibility but lack efficient vector indexing.

5. What is similarity search and how is it implemented in vector databases?

Similarity search retrieves items that are close to a given query vector. It is implemented using:

- **Distance Metrics** (e.g., cosine similarity, Euclidean)

- **Indexes** (e.g., HNSW, IVF, LSH)
- **Search Algorithms**: Exact search for high precision, ANN for fast results with acceptable recall.

6. What are some common similarity metrics used in vector databases?

- **Cosine Similarity**: Measures the angle between vectors (good for text).
- **Euclidean Distance**: Measures straight-line distance (good for geometric data).
- **Manhattan Distance**: Sum of absolute differences.
- **Jaccard Similarity**: For sparse vectors (like one-hot).

7. What is Approximate Nearest Neighbor (ANN) and why is it used?

ANN finds vectors that are "close enough" instead of the exact nearest. This speeds up searches over billions of vectors with minor accuracy trade-offs. Used for real-time systems where low-latency is critical.

8. Describe a real-world application where you used a vector database.

In a past role, I built a **semantic search engine** for customer support tickets. Using BERT for embedding generation, I stored embeddings in **FAISS**. When users typed a query, the system returned similar resolved tickets, reducing manual work and improving resolution time by 30%.

9. How does indexing work in vector databases?

Indexes are data structures used to organize vectors to enable fast retrieval. Common types include:

- **HNSW**: Graph-based, high accuracy, scalable.
- **IVF (Inverted File Index)**: Clusters vectors and searches within top clusters.
- **LSH**: Uses hash functions to group similar vectors.

Indexing reduces the number of vectors compared during queries, improving speed.

10. How does dimensionality affect performance and accuracy in vector DBs?

High-dimensional data offers better representational power but worsens performance due to the **curse of dimensionality**. Trade-offs:

- **Accuracy**: Increases with more dimensions (up to a point).
- **Speed**: Decreases due to harder indexing and comparison.
  Dimensionality reduction (e.g., PCA, UMAP) is often applied before storage.

11. What is vector normalization and why is it important?

Normalization scales vectors (e.g., to unit length). For cosine similarity, this removes the influence of vector magnitude, ensuring similarity depends only on direction. It standardizes vector space, especially important when embedding ranges differ.

12. What are the challenges in scaling a vector database?

- **Index Rebuilds** on inserts/deletes.
- **Latency**: With billions of vectors, naive search slows.
- **Memory Management**: Storing large embeddings (512–4096 dims).
- **Distribution**: Sharding while maintaining similarity guarantees.
- **Concurrency**: Handling multi-user access with consistent results.

13. What is HNSW and why is it popular for vector search?

Hierarchical Navigable Small World (HNSW) is a graph-based ANN algorithm. It creates multiple layers of graphs, connecting similar nodes. Searches start from the top layer and move down, making it both accurate and efficient. Used in **FAISS**, **Milvus**, **Weaviate**.

14. Compare FAISS, Annoy, Milvus, and Pinecone.

Tool	Strengths	Use Case
FAISS	Fast, custom, local indexing	On-prem NLP/image tasks
Annoy	Low-memory, good for read-heavy	Recommender systems
Milvus	Distributed, hybrid search support	Production systems
Pinecone	Managed, scalable, serverless	Enterprise-level vector ops

15. How do you update or delete vectors in a database?

- **In-place update** (if supported) with automatic re-indexing.
- **Soft delete** (mark as inactive) for append-only systems.
- **Batch reindexing** for systems like FAISS.
  Best practice: use versioning to avoid query inconsistencies during updates.

16. What is dimensionality reduction and how is it used in vector DBs?

Techniques like PCA or UMAP reduce vector dimensions while preserving structure. This:

- Improves search speed
- Reduces storage
- Mitigates overfitting in similarity search
  Trade-off: Slight loss in precision.

17. What is vector quantization?

It compresses vectors by approximating them using a limited set of vectors (codebooks). Used in **Product Quantization (PQ)** in FAISS to reduce memory usage and improve performance for large-scale search.

18. How do you integrate vector DBs into machine learning pipelines?

- Store embeddings from model outputs into the DB.
- During inference, generate new embeddings and perform similarity search.
- Used for recommendations, classification, clustering, or retrieval-augmented generation (RAG).

19. What are ANN trade-offs in latency vs. accuracy?

- **Lower latency** → Fewer comparisons → Lower recall
- **Higher accuracy** → More comparisons → Higher latency
  Configurable via parameters like efSearch in HNSW or nprobe in IVF.

20. How is real-time querying supported in vector databases?

- **In-memory indexes** for sub-ms latency
- **Parallel search threads**
- **Asynchronous processing**
- **Batch updates** to reduce index rebuilds
  Pinecone and Milvus offer features like vector streaming and hybrid search.

21. How do you monitor and debug performance in a vector database?

- Track **latency per query**, **recall metrics**, and **throughput**
- Profile query logs to find slow queries
- Monitor CPU/memory usage
- Use benchmarks (e.g., ann-benchmarks dataset) for tuning

22. How do vector databases support multi-modal data?

By storing embeddings from different modalities (e.g., text, audio, images) in the same vector space or separate spaces. Fusion techniques like concatenation or projection can be applied before indexing.

23. How do you ensure privacy and security in vector databases?

- **Data encryption** (in transit + at rest)
- **Access control / RBAC**
- **Differential privacy** (embedding-level protection)
- **Audit trails** for access logging

24. What is a vector space model and how is it applied?

A vector space model represents items (documents, images) as points in a continuous space. It's used in:

- Information retrieval
- Semantic similarity
- Clustering/classification
  TF-IDF and word2vec models are traditional VSM applications.

25. How do you evaluate a vector search system?

- **Precision@k / Recall@k**
- **Mean Average Precision (MAP)**
- **Latency under load**
- **Throughput (queries/sec)**
- **Index build time and size**
  Use tools like ann-benchmarks, FAISS evaluation scripts, or custom test sets.

--

Data Science Machine Learning Vector Database Genai Llm

Follow

Written by Sanjay Kumar PhD

1.2K followers · 445 following

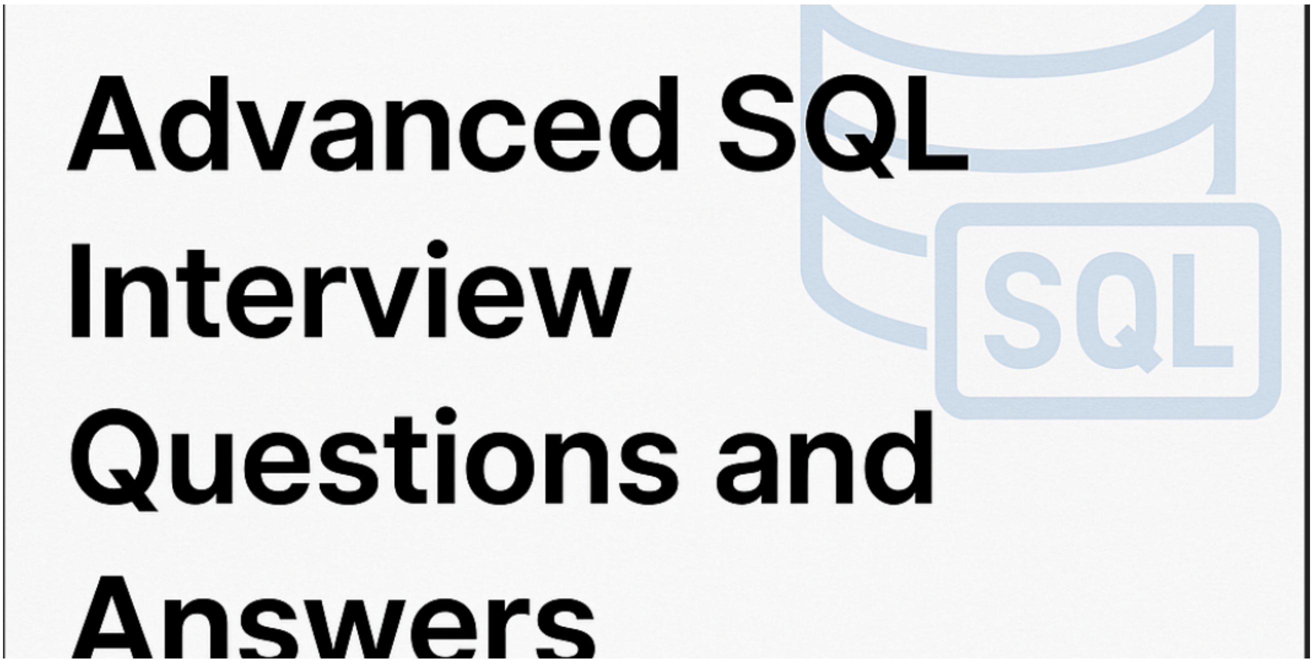
Data Science | Machine Learning | AI Product | GenAI | RAG | LLM | AI Agents | NLP | Analytics | Data Engineering | Deep Learning | Statistics

No responses yet

To respond to this story,
get the free Medium app.

Open in app

More from Sanjay Kumar PhD



Sanjay Kumar PhD

Advanced SQL Interview Questions and Answers

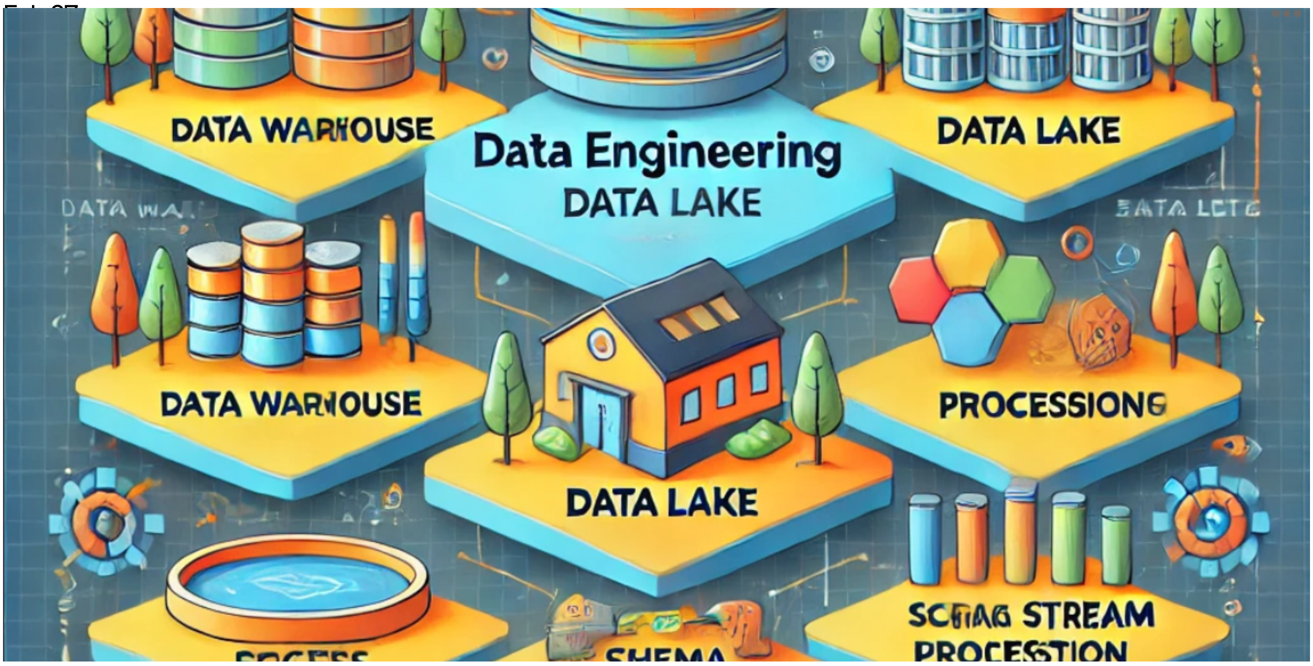
1. What is the difference between RANK(), DENSE_RANK(), and ROW_NUMBER()?

Mar 26

Sanjay Kumar PhD

Microsoft Azure Interview Questions and Answers

Core Azure Concepts & Cloud Computing Basics

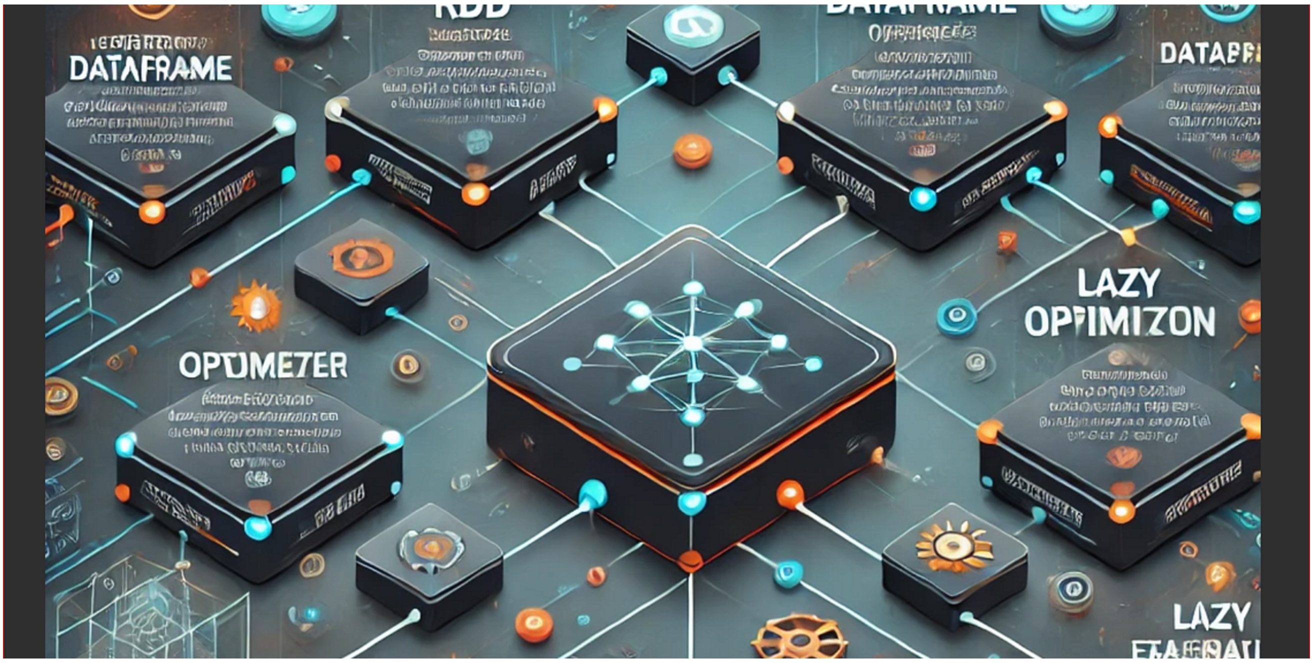


Sanjay Kumar PhD

Data Engineering Interview Questions and Answers

1. What is a Data Warehouse, and how is it different from a Data Lake?

Dec 26, 2024



Sanjay Kumar PhD

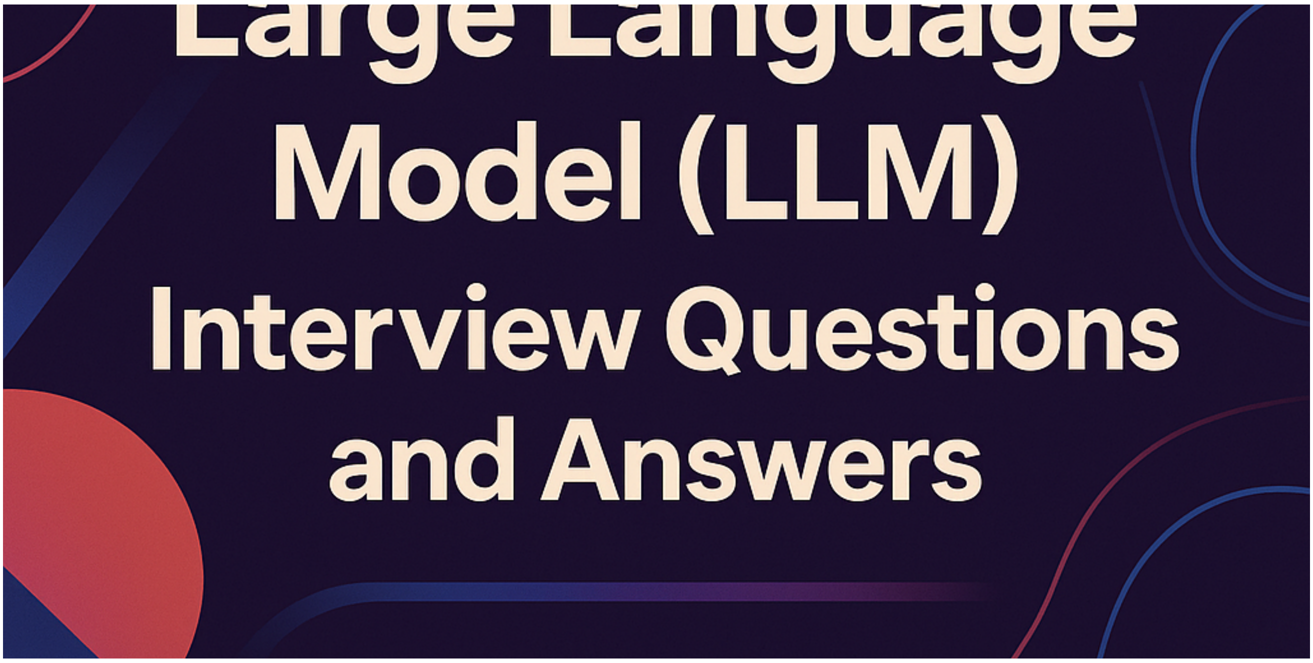
Apache Spark Interview Questions and Answers

What is Apache Spark?

Sep 11, 2024

See all from Sanjay Kumar PhD

Recommended from Medium



Sanjay Kumar PhD

Top 25 Large Language Model (LLM) Interview Questions and Answers

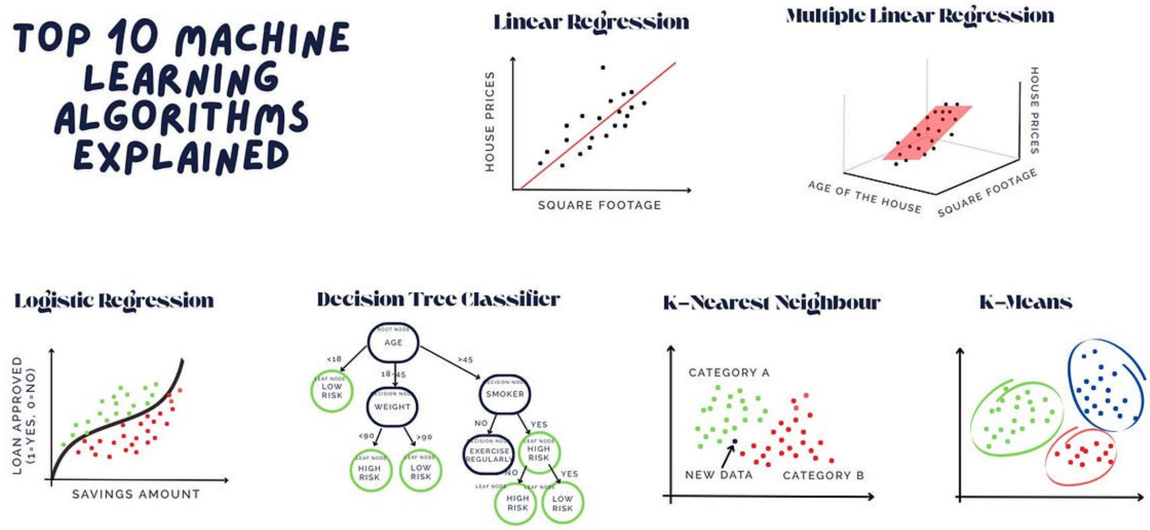
1. What is tokenization and why is it critical for LLMs?

Jul 17

In AI-ML Interview Playbook by Sajid Khan

A Quick Introduction to PyTorch (with a Hands-On Mini Project)

Learn the basics of PyTorch and build your first image classifier in under 100 lines of code



In Learning Data by Rita Angelou

10 ML Algorithms Every Data Scientist Should Know — Part 1

I understand well that machine learning might sound intimidating. But once you break down the common algorithms, you'll see they're not.

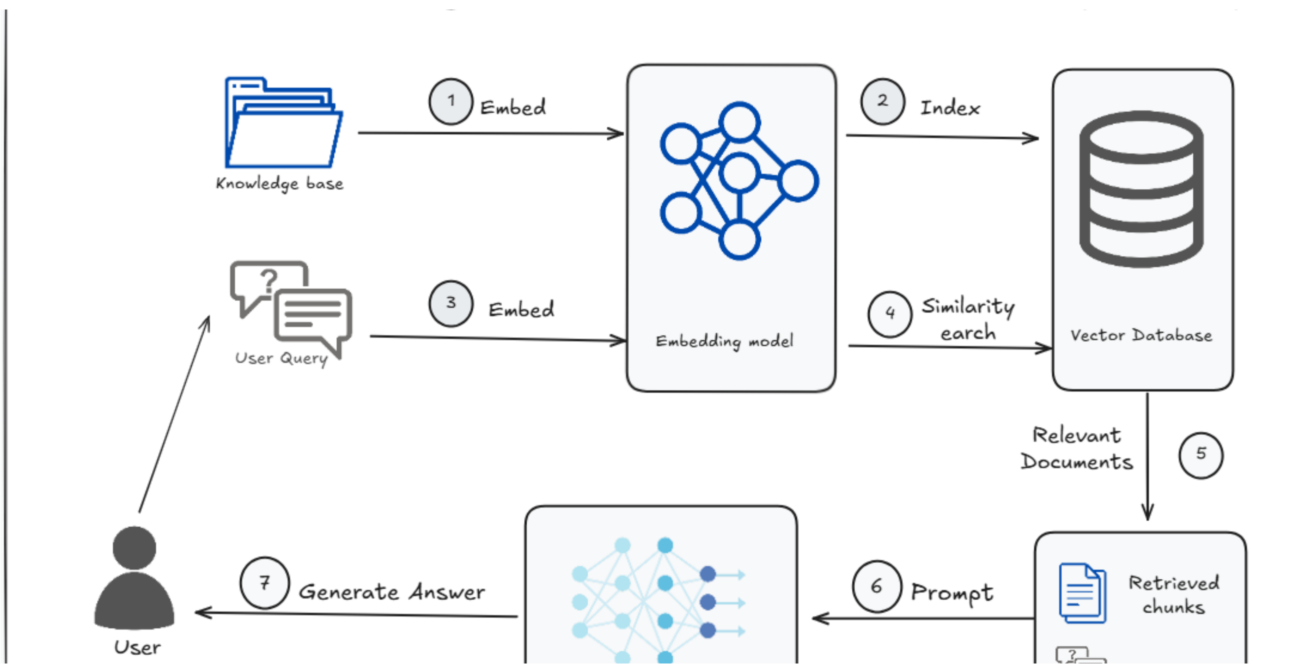
Jun 10

Anirban Mukherjee ✍️

🚀 50 Machine Learning Projects That Will Get You Hired in 2025 🤖

Not a member yet? Read for free here.

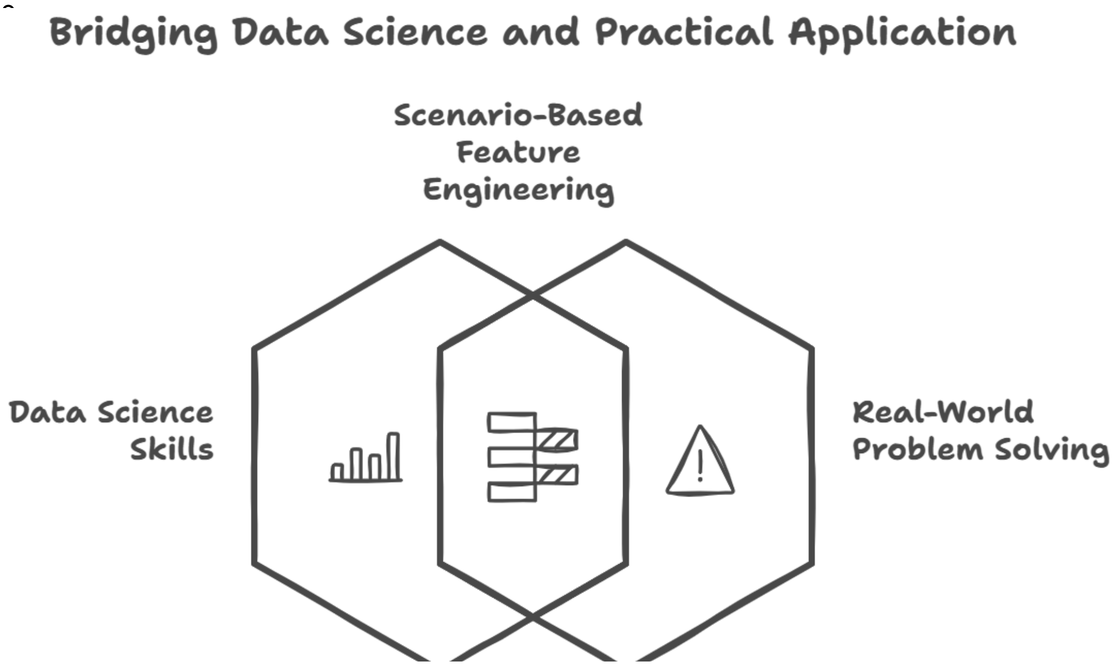
✨ May 23



In AI Advances by Anjolaoluwa Ajayi

21 Chunking Strategies for RAG

And how to choose the right one for your next LLM application



Vikash Singh

Scenario-Based Interview Questions for Senior Data Science Roles: Feature Engineering in Complex...

"Your model is only as good as the features you feed into it."

Mar 3

See more recommendations