Vector Compression Cheatsheet

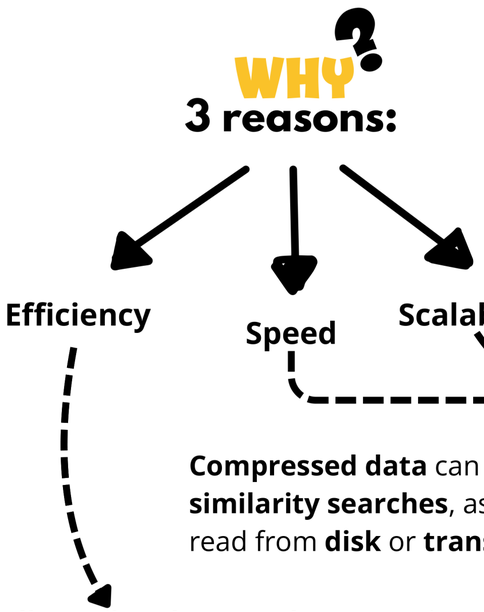
ashishpatel2604

What is Vector Compression?

Vector compression in vector databases refers to the process of reducing the size of high-dimensional vector data without significantly compromising its quality or the ability to perform similarity searches efficiently.

- **Objective:** Maintain data quality and perform efficient similarity searches.
- **Importance:** Essential for managing large-scale vector data.
- **Applications:** Crucial in machine learning and AI, where data volume is immense.

Why Vector Compression?

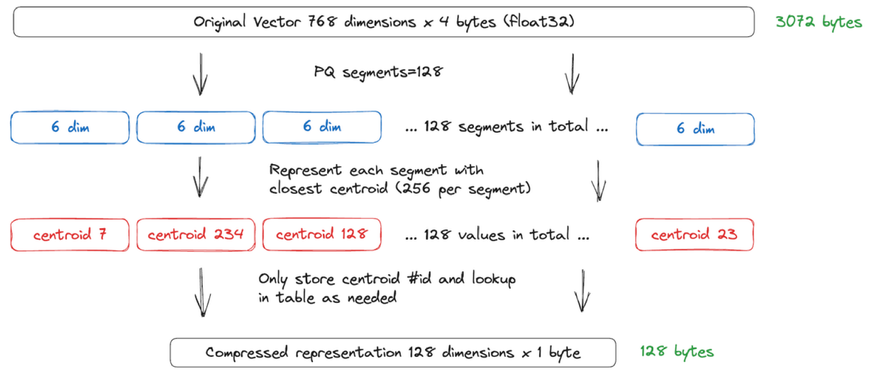


With **compressed data**, vector databases can **scale more effectively**, supporting **larger numbers of vectors** without a proportional increase in resource usage.

Compressed data can be processed **faster** during **similarity searches**, as **less data needs** to be read from **disk** or **transferred over networks.**

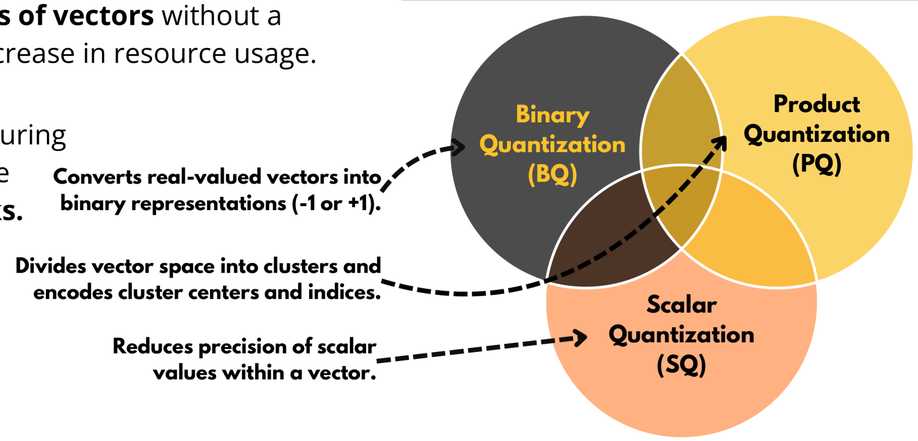
High-dimensional vector data can take up a lot of **storage space**. **Compression** reduces the **storage requirements**, making it feasible to **handle larger datasets** within the available resources.

How Vector Compression Works?



1. **Identify High-Dimensional Vectors**: Start with your high-dimensional data (e.g., text, images).
2. **Choose a Technique:** Select a dimensionality reduction method, such as Principal Component Analysis (PCA) or autoencoders.
3. **Apply Dimensionality Reduction:**
   a. **PCA:** Compute principal components and project data onto them.
   b. **Autoencoders:** Train a neural network to compress and reconstruct the data.
4. Transform Vectors: Use the trained model to convert high-dimensional vectors into lower-dimensional ones.
5. **Retain Essential Information:** Ensure the lower-dimensional vectors maintain key relationships and distances.
6. **Store Compressed Vectors:** Save the compressed vectors for efficient storage and faster searches.

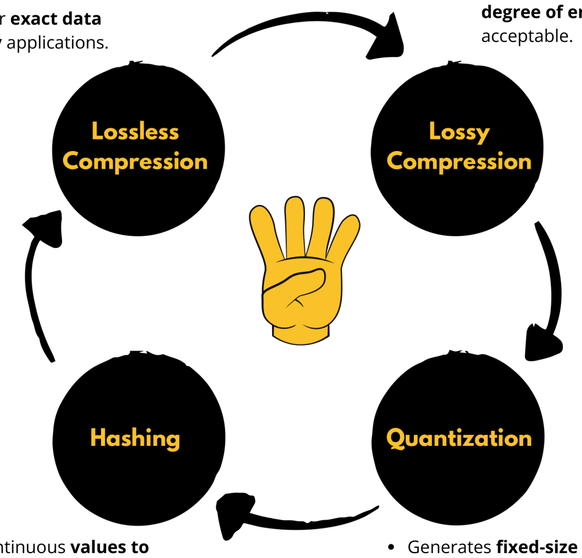
Famous Quantization Methods



Types of Vector Compression

- Ensures **perfect reconstruction** of original data.
- Useful for **exact data recovery** applications.

- Allows some **data loss for higher** compression ratios.
- Suitable when a **small degree of error** is acceptable.



- Maps continuous **values to discrete** ones.
- Reduces data precision for **efficient storage** and processing.

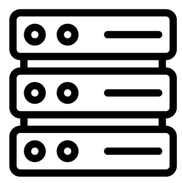
- Generates **fixed-size hash values** for vectors.
- Efficient for **large-scale datasets** but may reduce accuracy in similarity searches.

Application of Vector Compression

Machine Learning

Search Engine

Recommendation



Vector Database