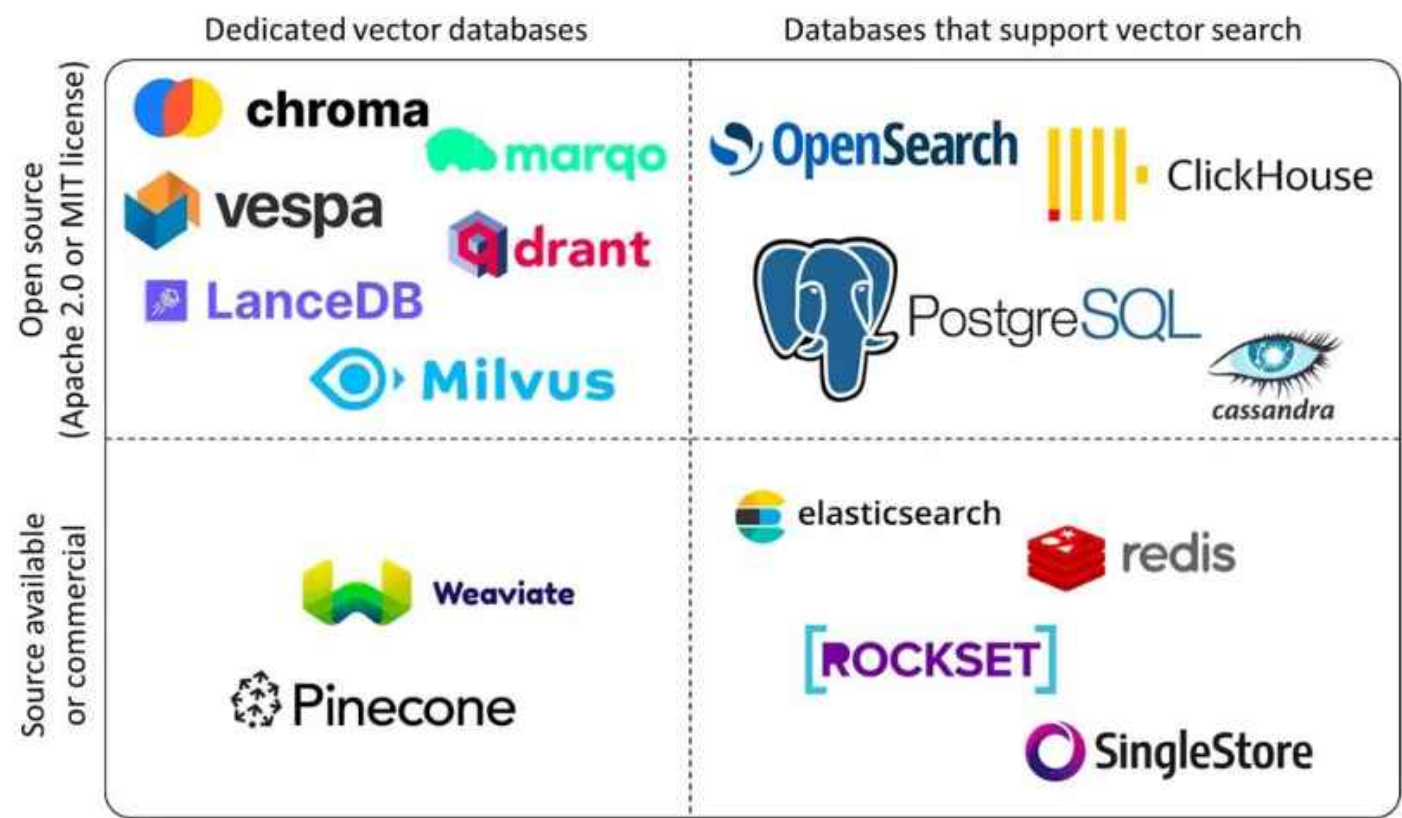# 向量数据库

简介：专门存储和查询向量数据的数据库系统

优势：

    a. 可处理高维数据。如图像、音视频

    b. 可处理大规模数据。亿级

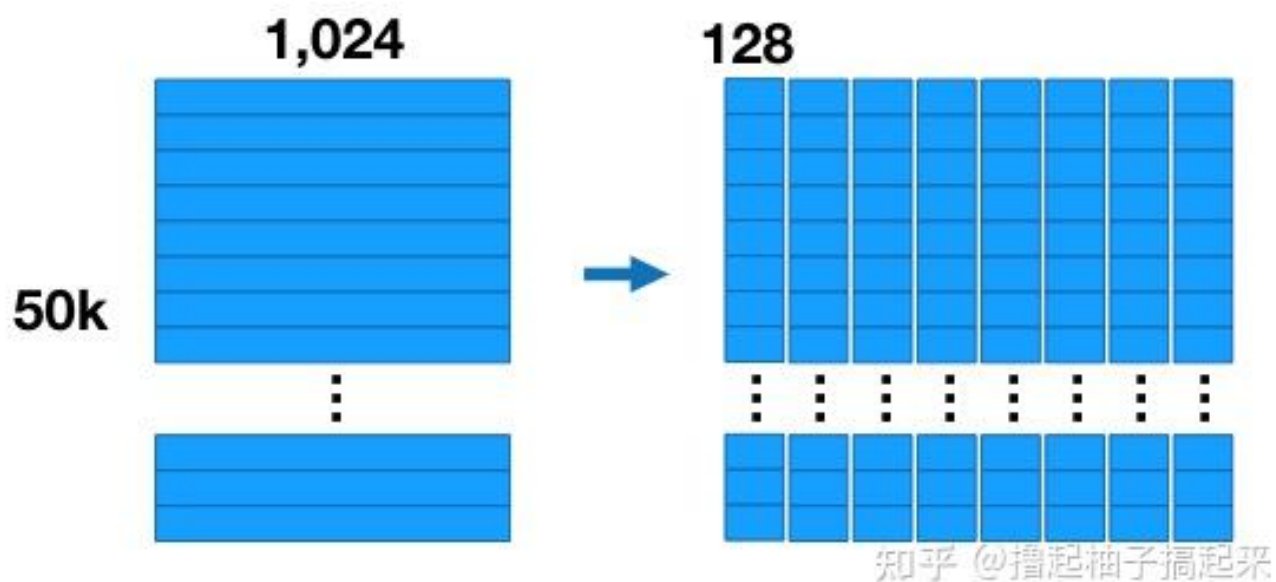   c. GPU 计算型近似搜索

有很多向量数据库，基本原理相似，主要介绍 FAISS

原理：

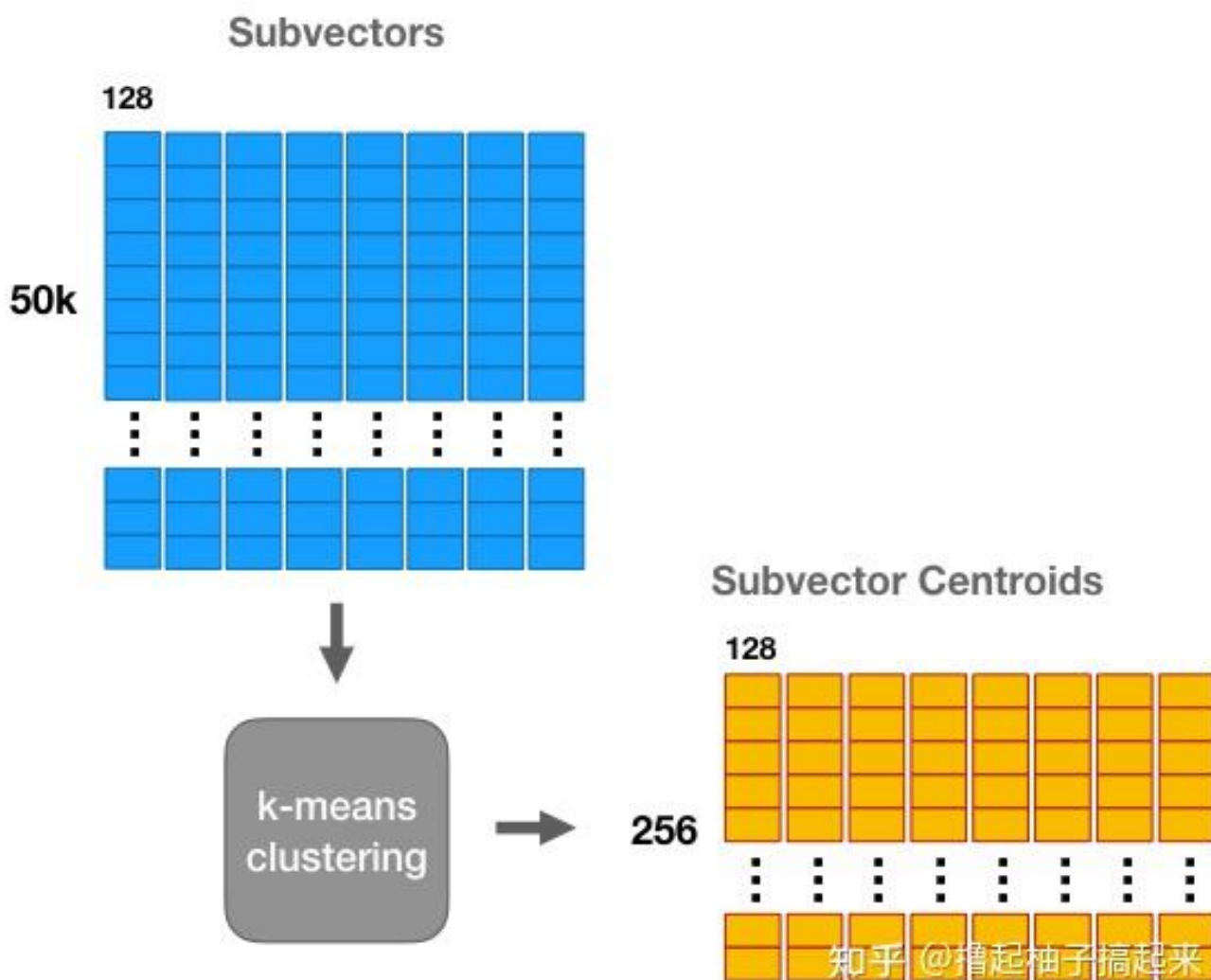## 1. 信息压缩 -> Product Quantization (乘积量化)

信息压缩是信息论中提高效率必要的方式，向量中一般用降维：PCA 或者 PQ

例如，50k 条 prompt 转换为 embeddings 之后，shape=(50k, 1024) # *注：不同的 text_encoder 生成的 embedding 维度不同，还有 512，768 等*
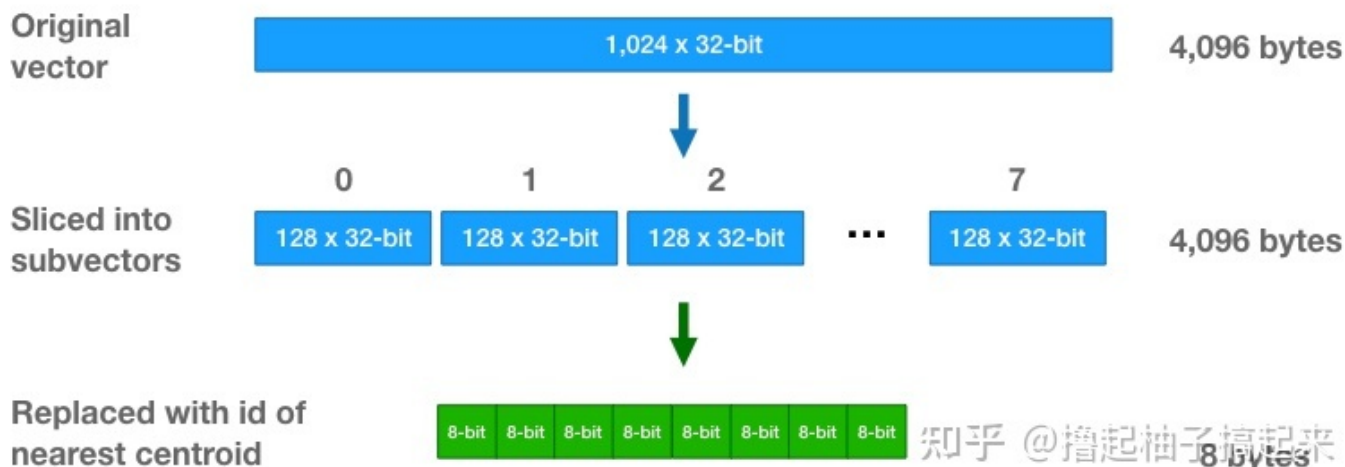
    a. 首先在 embedding_dim=1024 维度将 embeddings 切分为 m=8 组 sub_embedding

b. 在每一组 sub_embedding # shape=(50k, 128) 内，做 k-means 聚类 # k=256 便于 uint8 保存

c. 这样，每个单条 embedding 就被压缩编码为 uint8 * 8，也可以叫索引。

| Original vector | 1,024 x 32-bit | 4,096 bytes |

Sliced into subvectors — 0, 1, 2, ... 7

| 128 x 32-bit | 128 x 32-bit | 128 x 32-bit | ... | 128 x 32-bit | 4,096 bytes |

Replaced with id of nearest centroid

| 8-bit | 8-bit | 8-bit | 8-bit | 8-bit | 8-bit | 8-bit | 8-bit | 8 bytes |

2. 向量检索 -> Asymmetric Distance Computation (非对称距离计算)

向量检索中使用几乎所有的 ANN 方法都是：聚类全向量空间、查询向量快速锁定子空间、精确遍历。

    a. 查询向量切分 m=8 组

    b. 计算与每组 k=256 个聚类中心的距离，取最近

    c. 在簇中精确计算子向量距离，取最小

3. Prompt embeddings FAISS Search

    a. 用户定义：角色 A，角色 B，行为 -> 原始 prompt

角色 1 [杨幂 ▼] LoRA [couple_0614_001-000001 ▼]

角色 2 [杨幂 ▼] LoRA [couple_0614_001-000001 ▼]

行为 [＿＿＿] 图片张数 [＿＿＿]

[搜索]

    b. 预定义图库：图片，描述，标签 -> faiss.index

1	26	230912834	https://cdn.create.vista.com	Multicultural group of young people using digital devices while man holding coffee to go isolated on wh	12982378	IgorVetushko	['230912886', '619684330',	['Group', 'Chairs', 'Compute
2	27	7260594	https://cdn.create.vista.com	Composition of different smiling people	1008939	olly18	['32380407', '116932692', '1	['Fun', 'Group', 'White', 'Bea
3	28	224908070	https://cdn.create.vista.com	Successful businesspeople speaking with bullhorn isolated on white	13193658	AndrewLozovyi	['224891124', '224908174',	['Beautiful', 'Business', 'Fem
4	29	55311869	https://cdn.create.vista.com	Crowd of people in the city out of focus	3378121	Madrabothair	['55312407', '148994747', '1	['Focus', 'People', 'Unrecogn
5	30	305981468	https://cdn.create.vista.com	Panoramic shot of three friends using laptop and doing paperwork in office	12982378	IgorVetushko	['172089712', '475388304',	['Beautiful', 'Business', 'Girl',
6	31	31294803	https://cdn.create.vista.com	Large group of diverse people. Isolated on white	1010613	AndreyPopov	['607893274', '413680228',	['Fun', 'Group', 'Large', 'Whi
7	32	184131020	https://cdn.create.vista.com	Shocked girl gesturing with hands and looking at camera isolated on purple	12985848	EdZbarzhyvetsky	['574046972', '490145998',	['Beautiful', 'Person', 'Sign',
8	33	32380407	https://cdn.create.vista.com	Different people with different expressions	1008939	olly18	['156458886', '19066015', '4	['Color', 'Fun', 'Group', 'Whi
9	34	185179220	https://cdn.create.vista.com	Happy multicultural retro styled friends looking at camera isolated on grey	13194036	VitalikRadko	['185179214', '185179418',	['Beautiful', 'Happy', 'Female
10	35	48786423	https://cdn.create.vista.com	Large group of people in the shape of an arrow pointing up symbolizing direction , progress or growth.	1162342	digitalstorm	['91849974', '60178937', '70	['Bar', 'Group', 'Large', 'Back
11	36	216473576	https://cdn.create.vista.com	Beautiful young family spending time together in nature	12982378	IgorVetushko	['207412540', '216473714',	['Sport', 'Ball', 'Happy', 'Girl'
12	37	45169135	https://cdn.create.vista.com	Summer, holidays, vacation, happy people concept - group of teenagers looking down and showing fing	1017986	Syda_Productions	['32677917', '69826495', '70	['Game', 'Fun', 'Group', 'Vict
13	38	177722522	https://cdn.create.vista.com	Young african amercian businessman with laptop and businesswoman with tablet in hands looking at ea	13194036	VitalikRadko	['577648966', '447300016',	['Blue', 'Beautiful', 'Business

i. description embedding

ii. tags embedding

iii. image detection + tags embedding

c. 原始 prompt 查询，筛选对应 image 做 control 引导图