MILVUS DATABASE

Author: Ly Van Vuong
Date: 08/01/2021

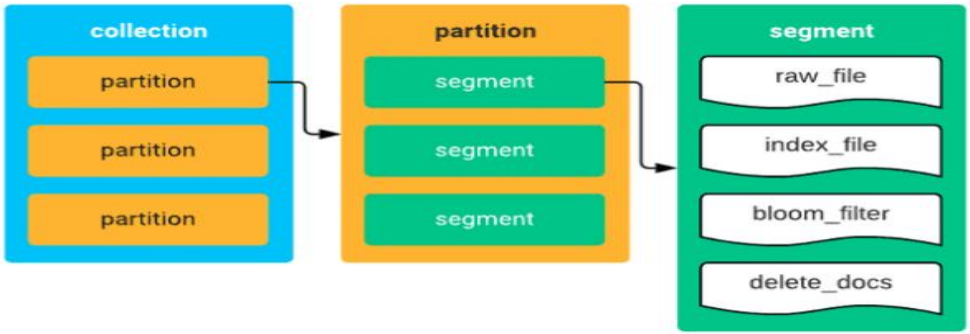
Contents

What is Milvus ?

- Milvus is an open-source vector similarity search engine designed that is highly flexible, reliable, and blazing fast.
- Milvus provides SDKs in Python, Java, Go, and C++, as well as RESTful APIs.
- Milvus comes in two distributions: CPU-only Milvus and GPU-enabled Milvus.

Concepts

- Partition and segment.
  - In a collection, you can divide the data into multiple partitions as needed.
  - When creating a collection, Milvus controls the size of a data segment according to the segment_row_limit (maximum number of entities a segment can hold).



Concepts

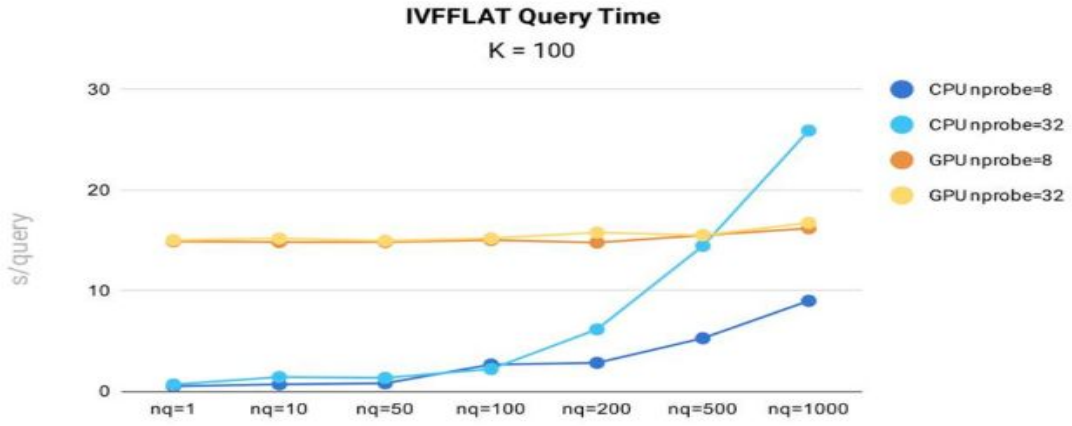
- Metadata
  - When querying data, Milvus must know the location and status information of each data file on the physical storage.
  - Metadata including the collection it belongs to, the number of entities it contains, the file size, the globally unique identifier, and the creation date, collection name, collection dimension, index type, partition label, and more.

Vector index

- Vector index is a time- and space-efficient data structure built on vectors through a certain mathematical model.
- Most of the vector index types of Milvus use ANNS (Approximate Nearest Neighbors Search)
- Some categories index milvus supports: FLAT, IVF_FLAT, IVF_SQ8, IVF_SQ8H, IVF_PQ, RNSG, HNSW, ANNOY.

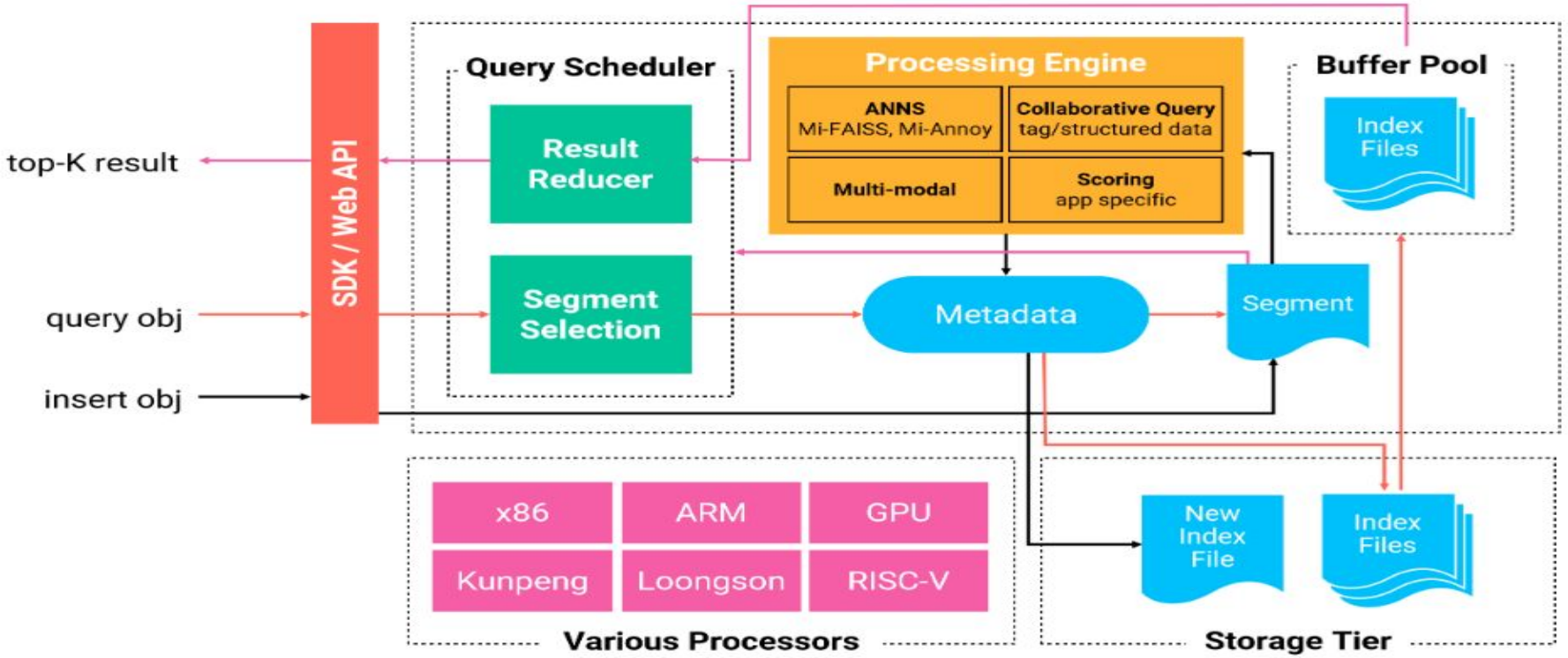
Vector index

- IVF_FLAT is the most basic inverted file index type and relies on a form of ANN search.
  - IVF_FLAT divides vector data into a number of cluster units (nlist), and then compares distances between the target input vector and the center of each cluster.
  - Milvus compares the distances between the target vector and the centers of all nlist clusters to get nprobe nearest clusters.
  - IVF_FLAT query time performance testing 1 billion 128-dimensional vectors:



Query time test results for IVF_FLAT index in Milvus.

Architecture



Distance Metrics

- Distance metrics are used to measure similarities among vectors.
- In Milvus have two distance metrics supported:
  - Inner product (IP)

$$p(A,B) = A \cdot B = \sum_{i=1}^{n} a_i \times b_i$$

  - Euclidean distance (L2)

$$d(\boldsymbol{a},\boldsymbol{b}) = d(\boldsymbol{b},\boldsymbol{a}) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$

Performance

- Timing when the client and the server are running on the same physical machine
  - Insert

Number of vector	Dimensional	Timing(s)
100 000	128	0.8
1 000 000	512	50

  - Search

Number of vector	Number of vector search(nq)	Build index	Timing(s)
1 000 000	10	No	0.14
1 000 000	10	Yes(nlist=100, nprobe = 20)	0.05
1 000 000	2000	No	0.29
1 000 000	2000	Yes(nlist=100, nprobe = 20)	0.18

Performance

- Note
  - If you want to use GPU for query, you need to set the value of gpu_search_threshold in milvus.yaml to be less than nq (number of vectors per query).
  - Milvus won't actually start build index task if the segment row count is smaller than `segment_row_limit`.

Advantages and Disadvantages

- Advantages
  - It supports adding, deleting, updating, and search of vector on a scale of trillion bytes.
  - Build index
  - Milvus support term, range, vector queries.
- Disadvantages
  - Each entity contains one ID field, one vector field, and multiple scala fields.
  - Only search with vector field.
  - Only get entity by ids.
  - Author, authen.

Discuss

Answer the question .

Reference

1. https://www.milvus.io/docs/v0.11.0/overview.md
2. https://github.com/milvus-io/milvus
3. https://medium0.com/unstructured-data-service/how-to-choose-an-index-in-milvus-4f3d15259212#e631