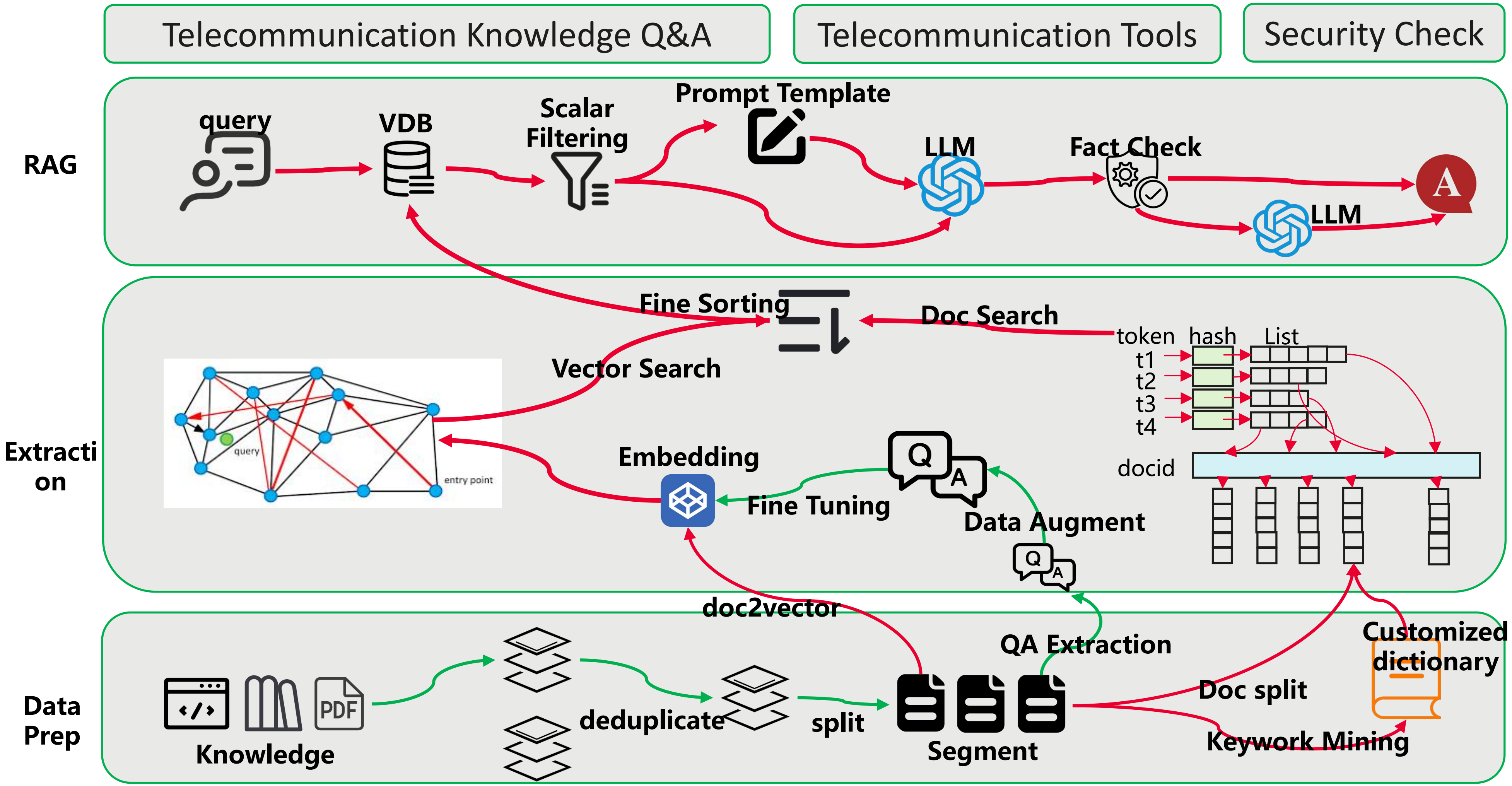
What Do We Really Need for Vector Database

Ji Sun

Huawei

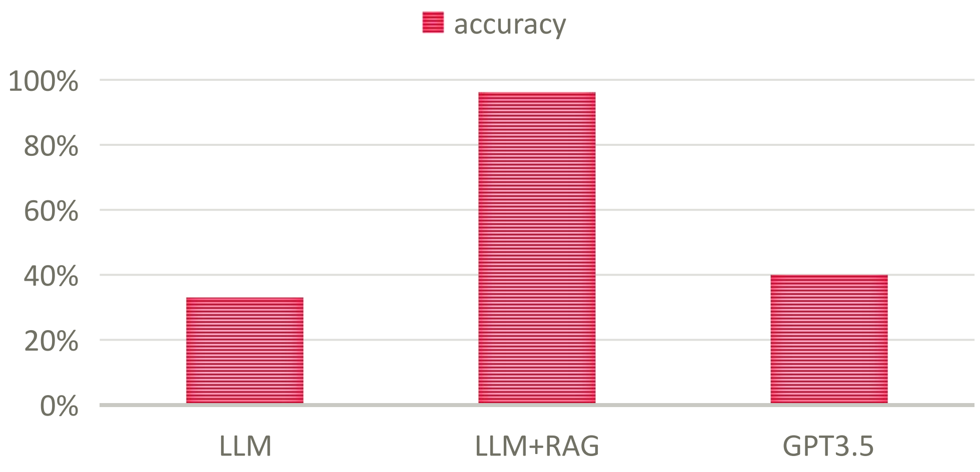
VectorDB is widely used in the LLM applications, and the apps are highly dependent on the performance, reliability, security and scalability of VectorDB

A typical pipeline of LLM application in the Telecommunication Field

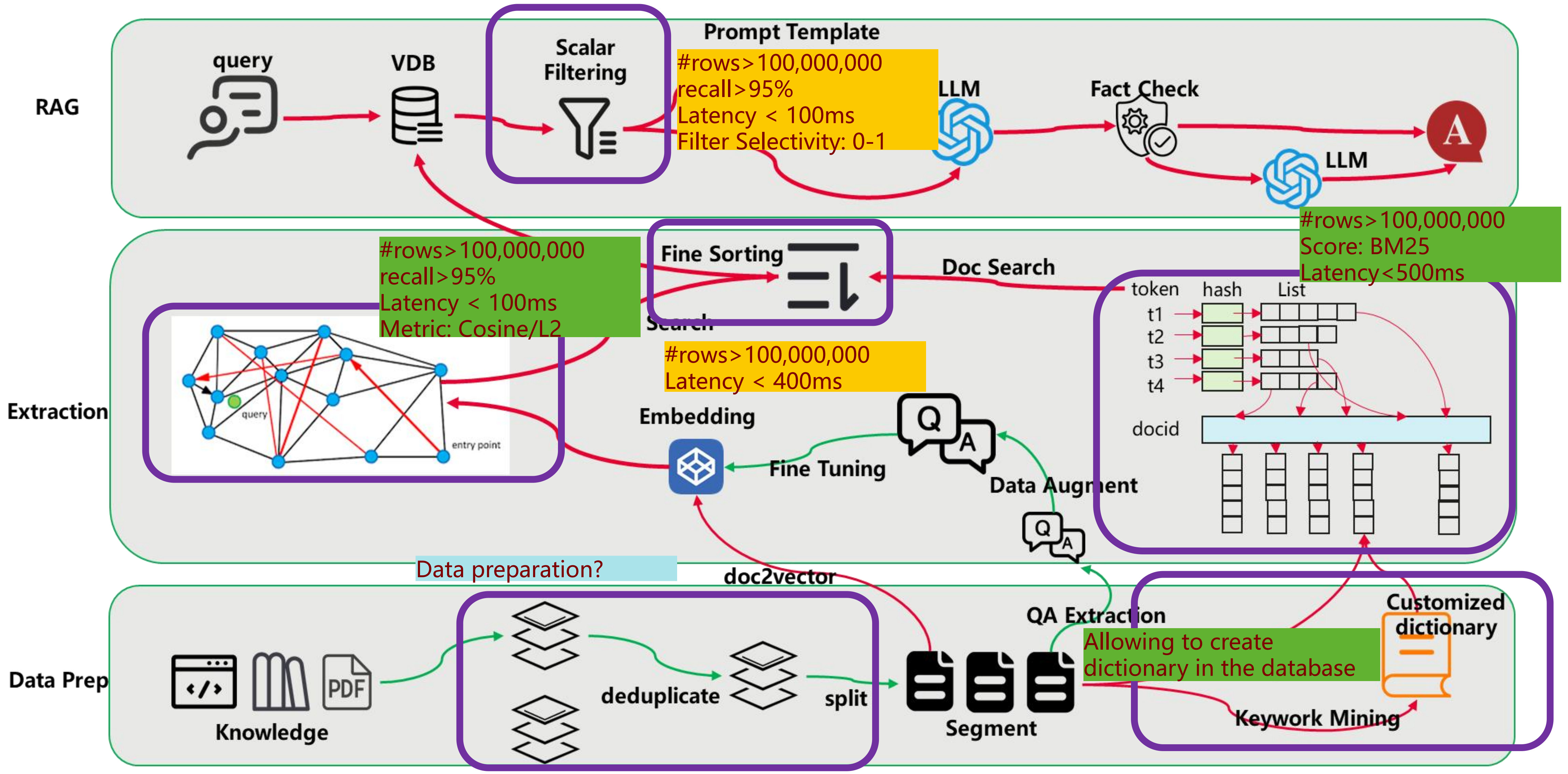


Details on this application

- **Pain Point:** LLM is not sensitive to the meaning of numeric data, and the conclusion is not trustworthy. For serious questions in the telecommunication field, LLM may produce nonsense words (even if they seem to be logical).
- **Data Source:** The source documents of telecommunication knowledges come from real product, including the product document, maintanence cases, external papers&journals
- **RAG:** users query the knowledge database by natural language, then the related knowledges are fed into LLM by prompt, and the LLM generate logical, smooth, correct and helpful answers.



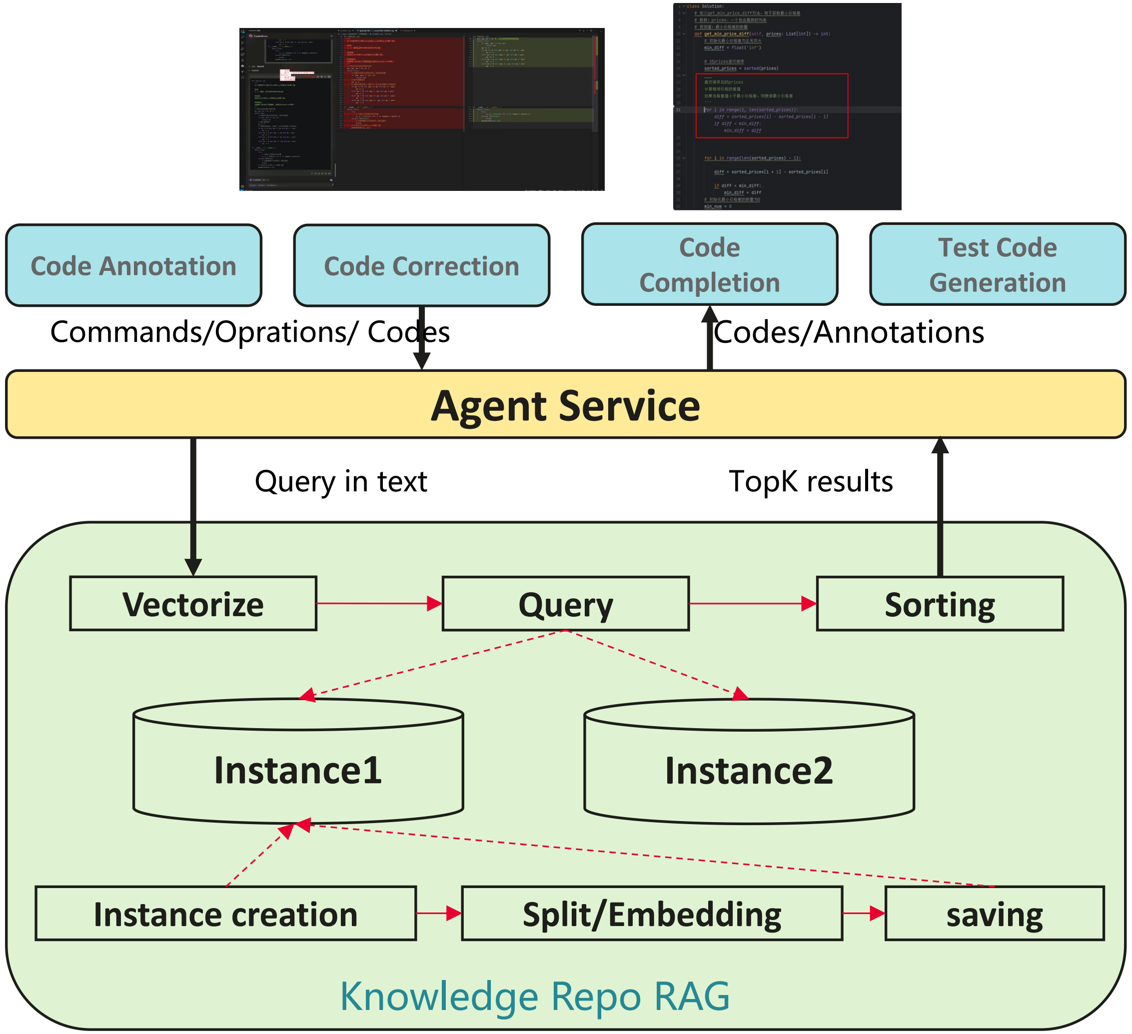
Vector Database Requirements



Applications with vector databases deployed on the cloud

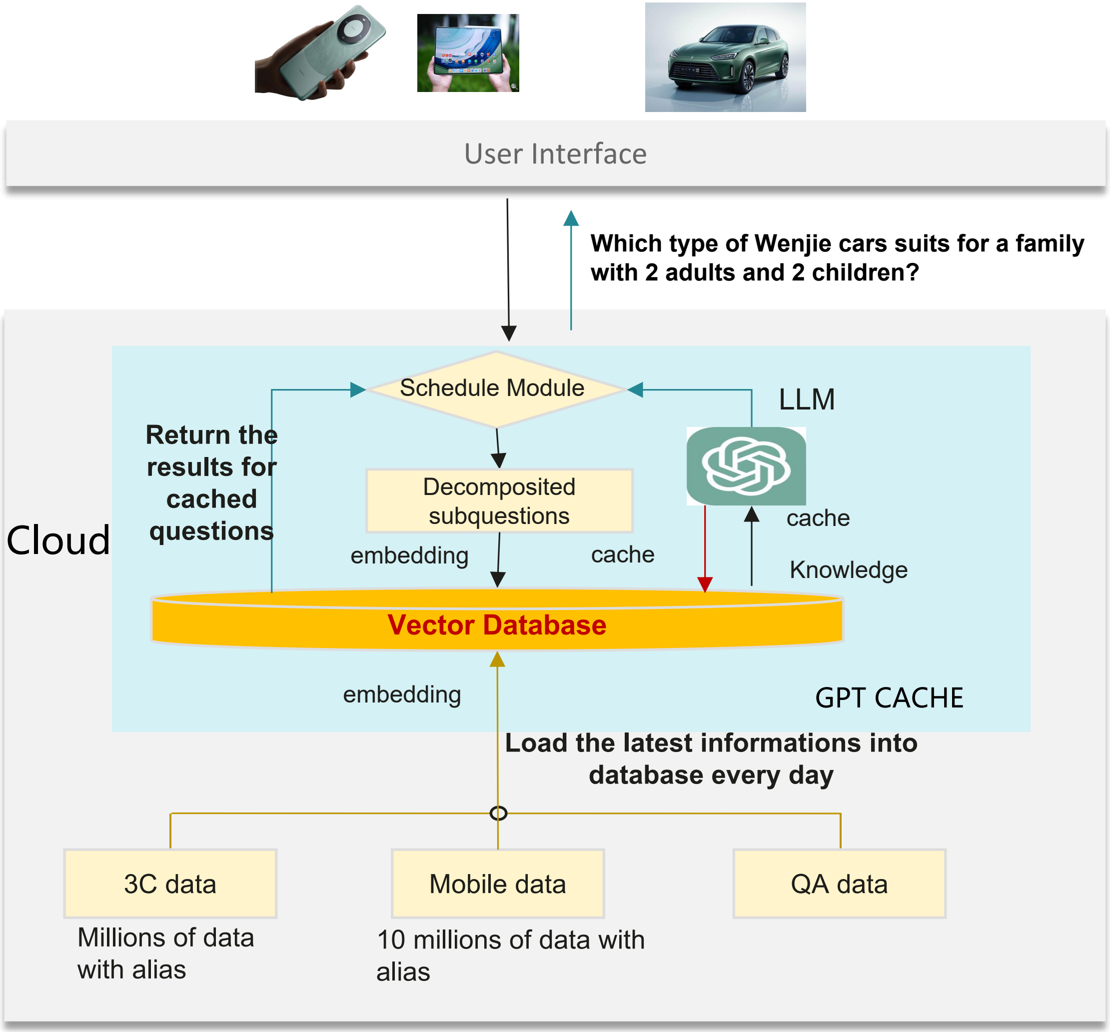
Developer Copilot

- Millions of data
- accurate search(Top5)
- multiple tasks, different requirements
- heterogeneous inputs/outputs



Voice Assistant

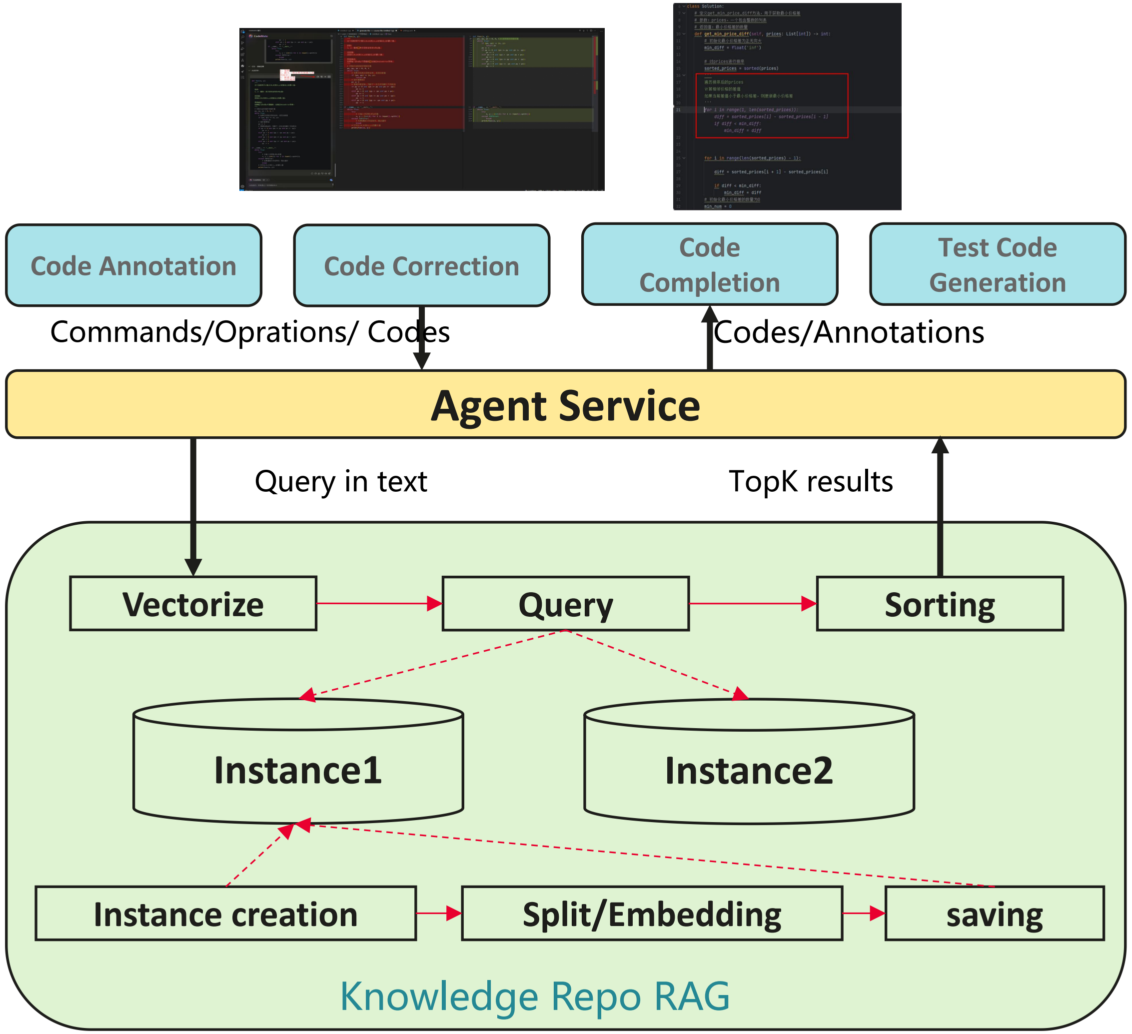
- Public service on cloud, privacy matters a lot
- High concurrency, high throughput is required
- Complicated questions, decomposition is required
- Relies on Time-to-live to manage the cache



Applications with vector databases deployed on the cloud

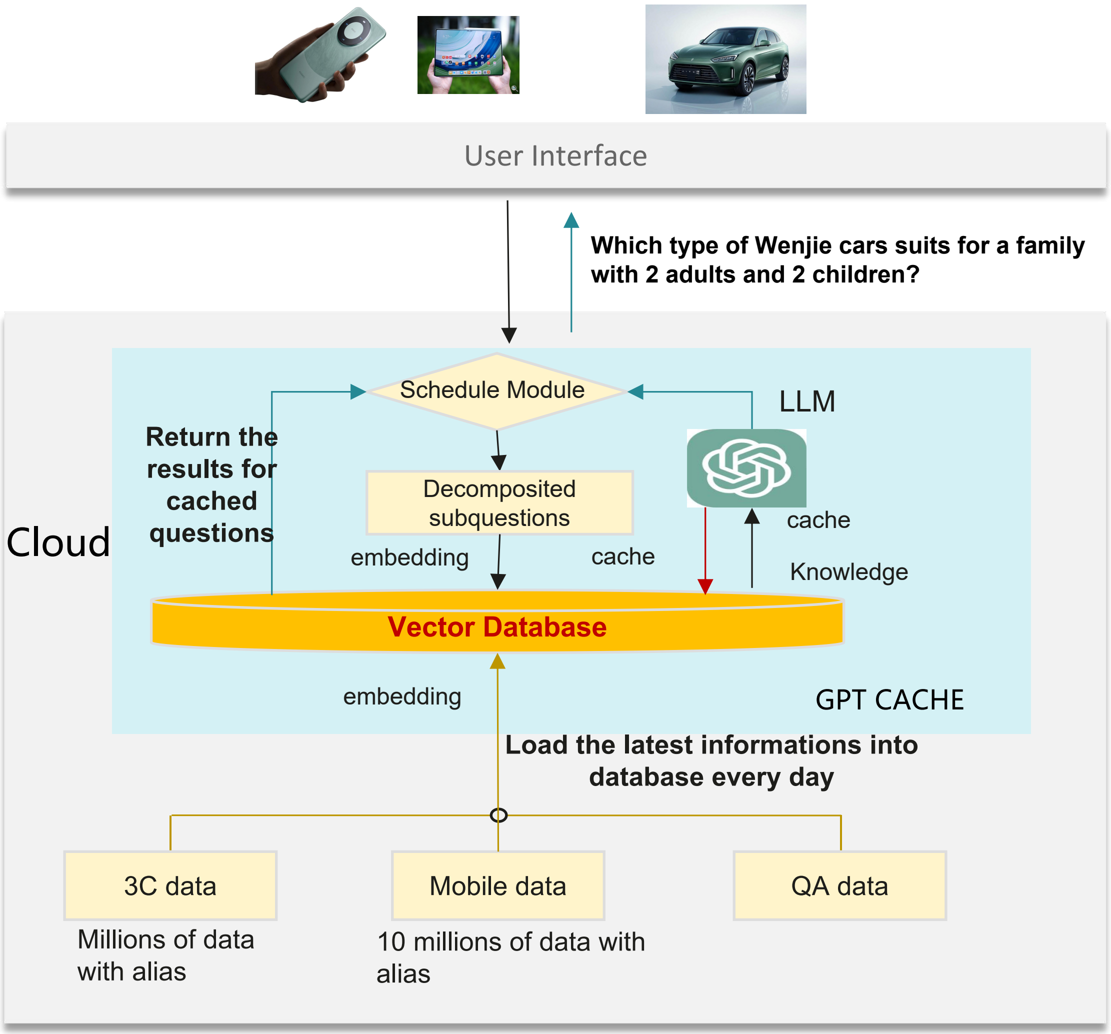
Developer Copilot

- Light-weighted index
- Fast I/O, Computing
- Disaggregation
- Resource Scheduling/Scaling



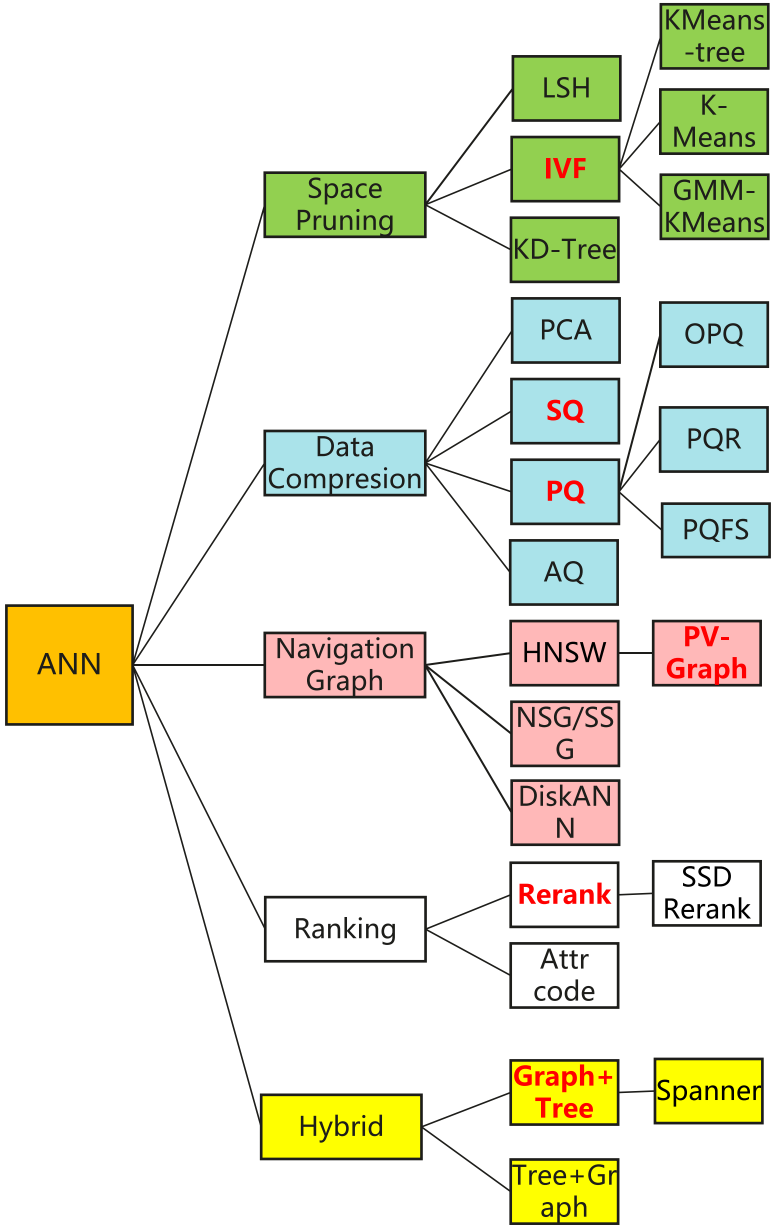
Voice Assistant

- Privacy protection
- Multi-tenant/Multi-read
- Query in batch
- Updatable



An algorithm view of vector similarity search

Algorithm optimization for approximate nearest neighbors

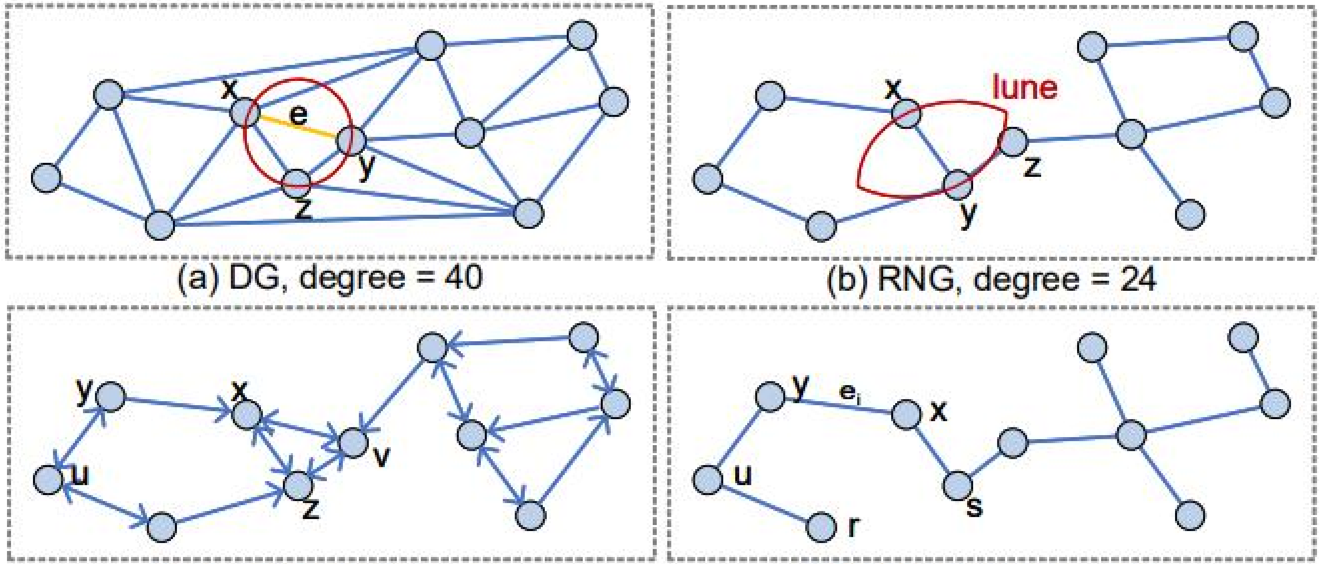


Brute-force KNN takes over 2 minutes for 100 million vectors.

1. The recall is not necessarily 100%, the second nearest neighbor maybe better because of the limitation of embedding methods.

2. Nearest neighbors have locality naturally, if close, continue to search; otherwise, jump out.

Graph-based Algorithms



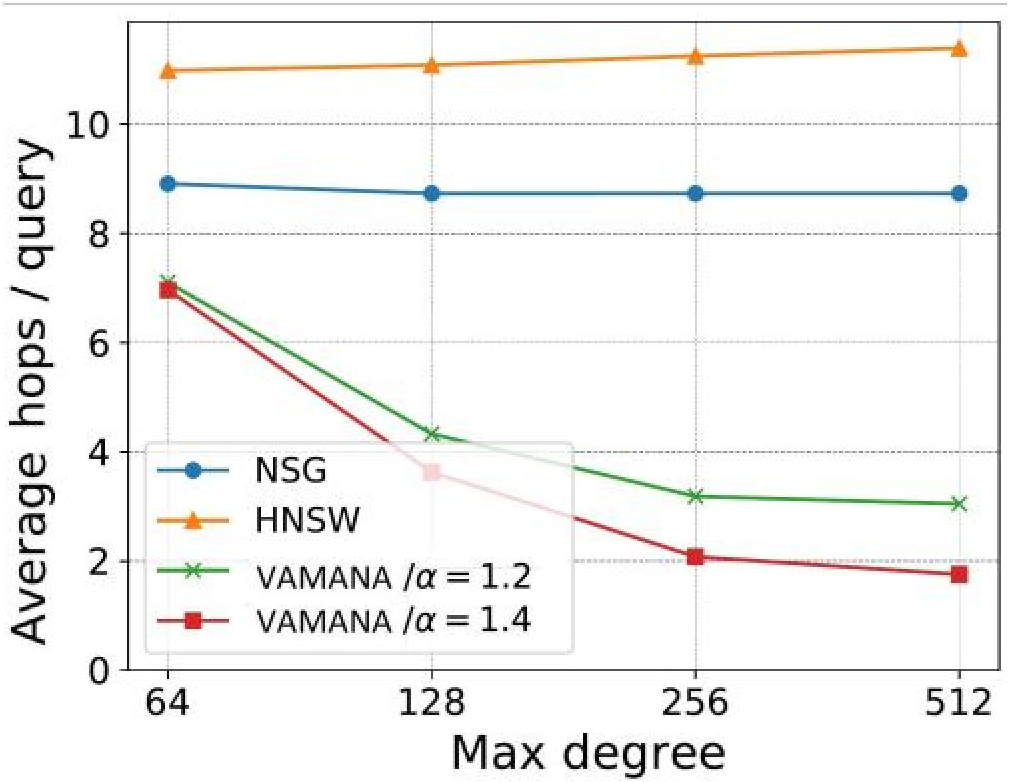
(a) DG, degree = 40
(b) RNG, degree = 24
(c) KNNG, degree = 22
(d) MST, degree = 20

	KGraph	NSW	HNSW	Vamana	NSG
Global Connectivity	Middle	Middle	Middle	Good	Good
Search Complexity (in memory, experimental)	$O(\|S\|^{0.54})$	$O(\log^2(\|S\|))$	$O(\log(\|S\|))$	$O(\|S\|^{0.75})$	$O(\log(\|S\|))$
Space Consumption	Middle	Middle	Middle	Good	Good
Construction Complexity (in memory, experimental)	$O(\|S\|^{1.14})$	$O(\|S\|\cdot\log^2(\|S\|))$	$O(\|S\|\cdot\log(\|S\|))$	$O(\|S\|^{1.16})$	$O(\|S\|^{(1+c)/c}\cdot\log(\|S\|)+\|S\|^{1.14})$

A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search.
Mengzhao Wang, Xiaoliang Xu, Qiang Yue, Yuxiang Wang. VLDB, 2021

DiskANN shows the best #hops when searching

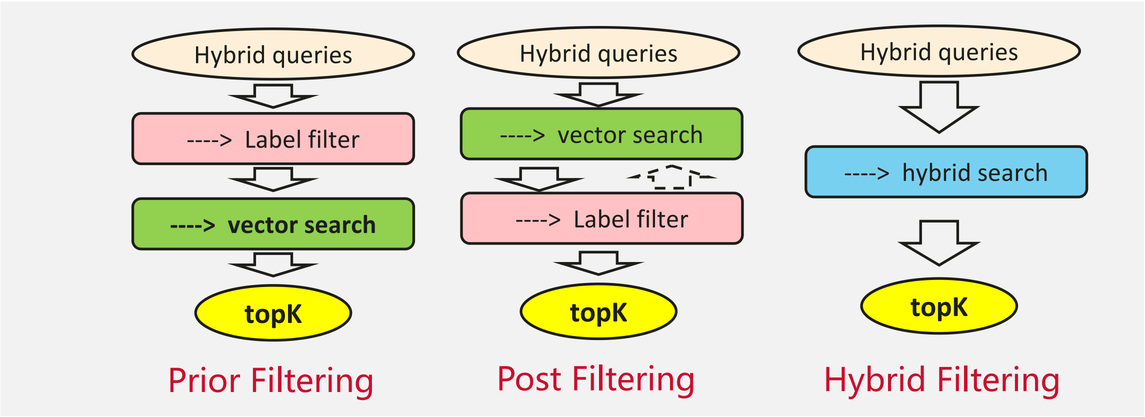
Average number of hops vs maximum graph degree
for achieving 98% 5-recall@5 on ANN_SIFT1M

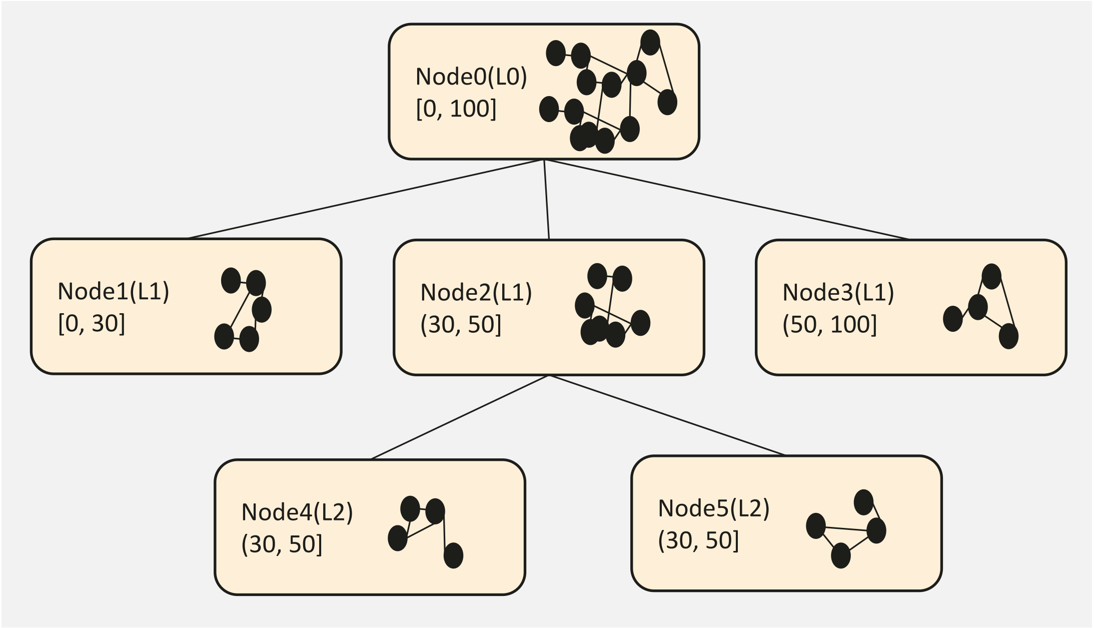


DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node.
Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, Harsha Vardhan Simhadri. NIPS, 2019

Multi-column scalar/vector hybrid index, support attribute filtering

Background: Vector is suitable for semantic search, scalar attribute is also important for access control, accuracy improve, and performance improve.



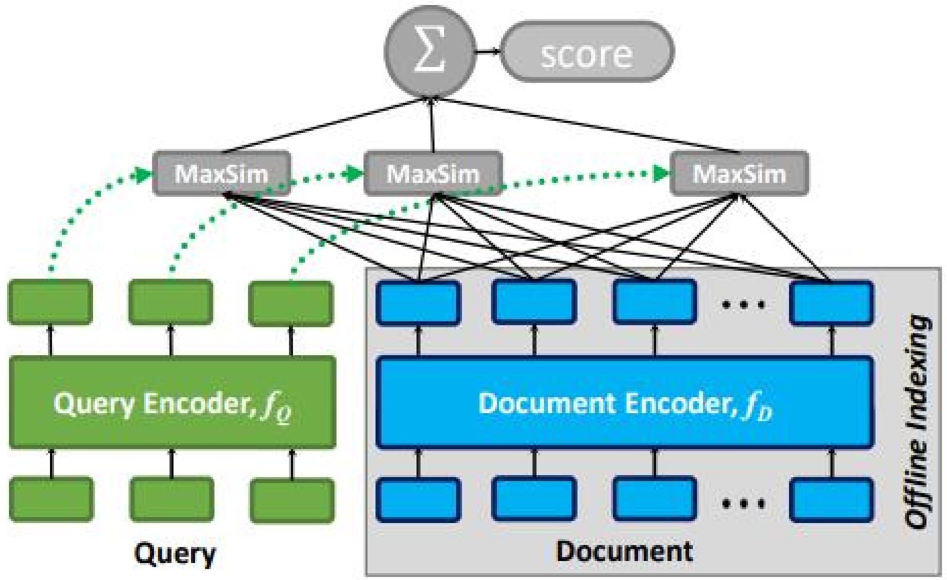


Scalar-vector hybrid query is supported by following 3 mainstream methods:

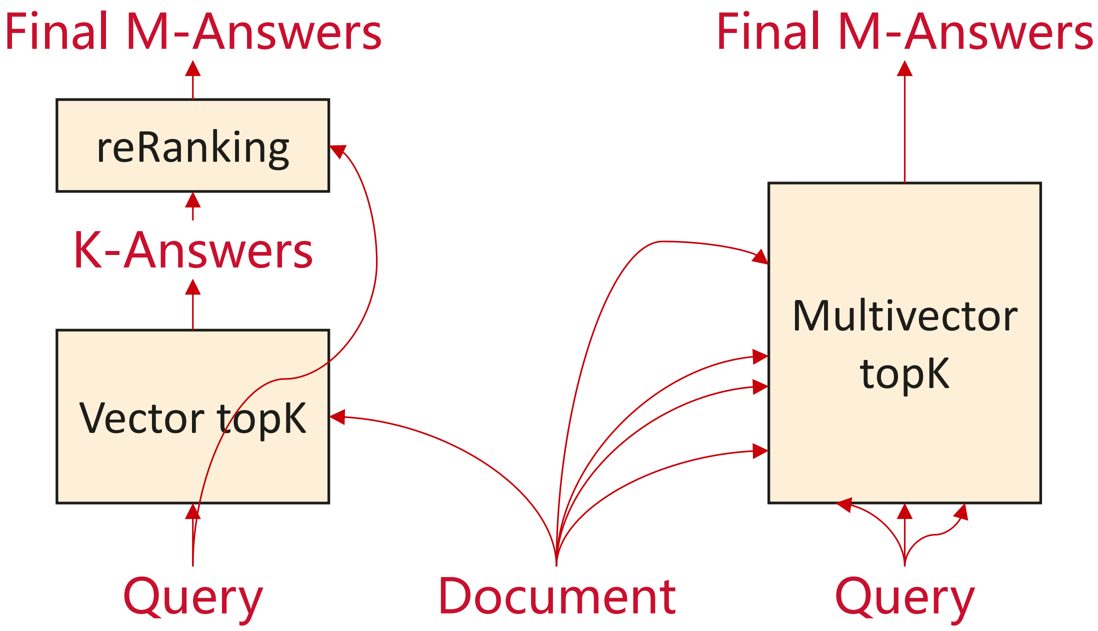
1. **Prior Filtering**. Sequence scan on selected vector is very slow. We can build index for each scalar label, but too many small index causes the range scan/ Full scan slow.
2. **Post Filtering.** Efficient when the selectivity of scalar condition is high. When the selectivity is low, the query will extend the candidate list constantly for enough result.
3. **Hybrid Filtering.** Consider scalar label when computing vector distances to find tuples whose vector is close to the query vector, and scalar matches the query condition. However, this method is not accurate enough.

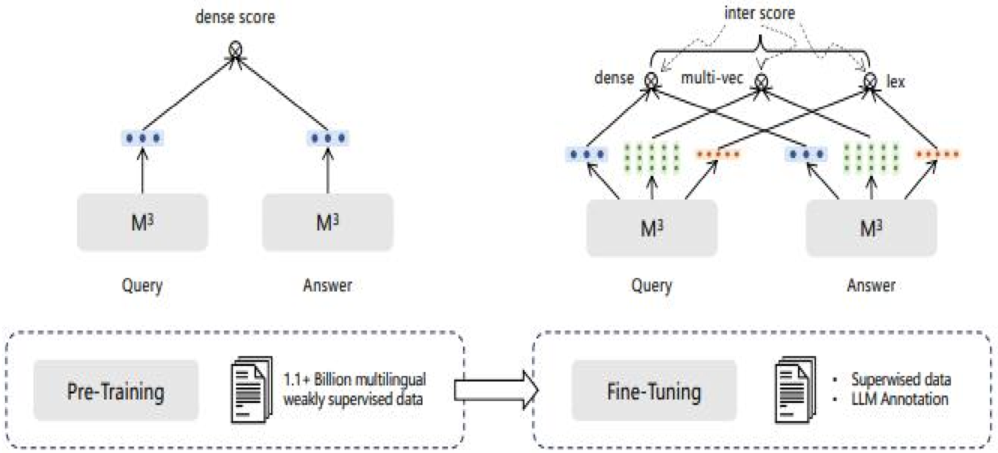
1. Full scan can be executed in the root node0, which has the same performance with the full graph;
2. Query with equal condition can search on the leaf node with higher performance;
3. Query with range condition can search on a small part of nodes for better performance

Light-weighted multi-vector similarity query for ranking

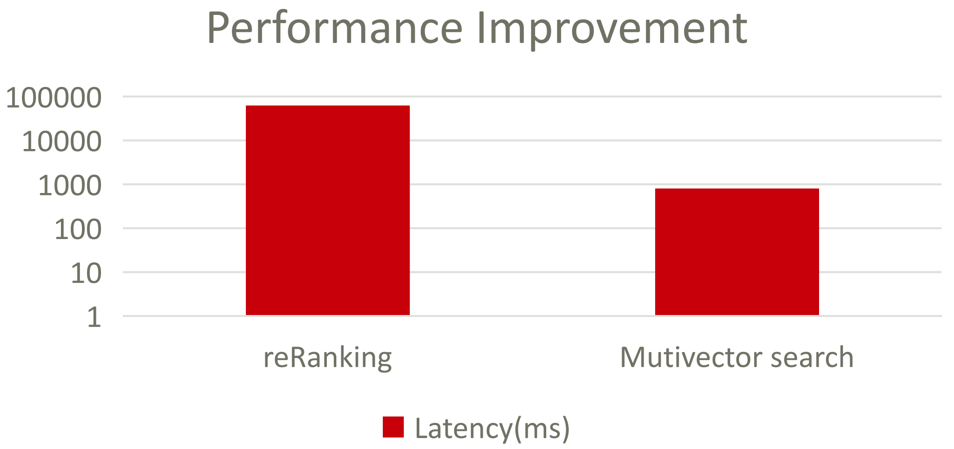


ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT





BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation

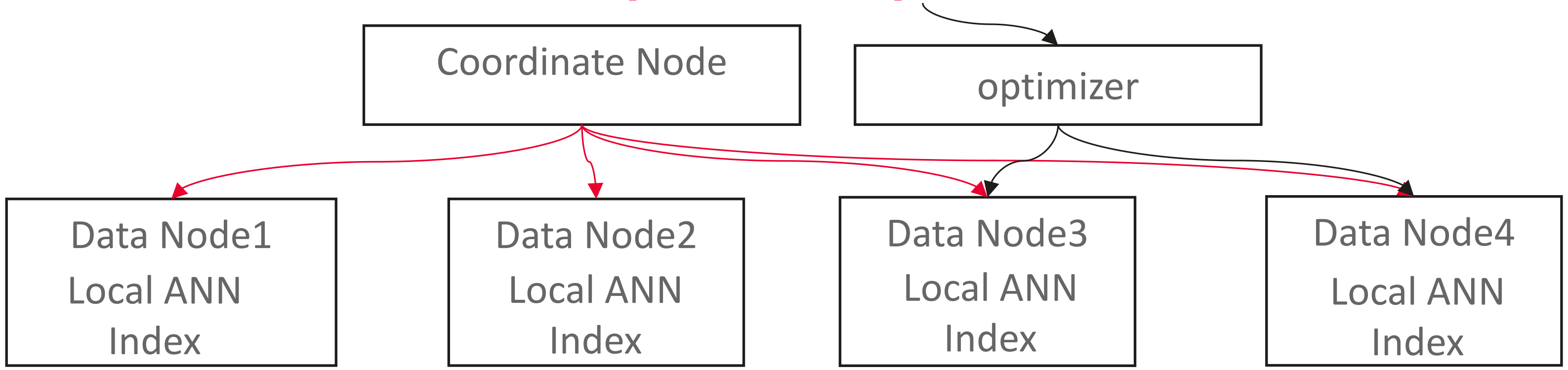


A system view of vector database

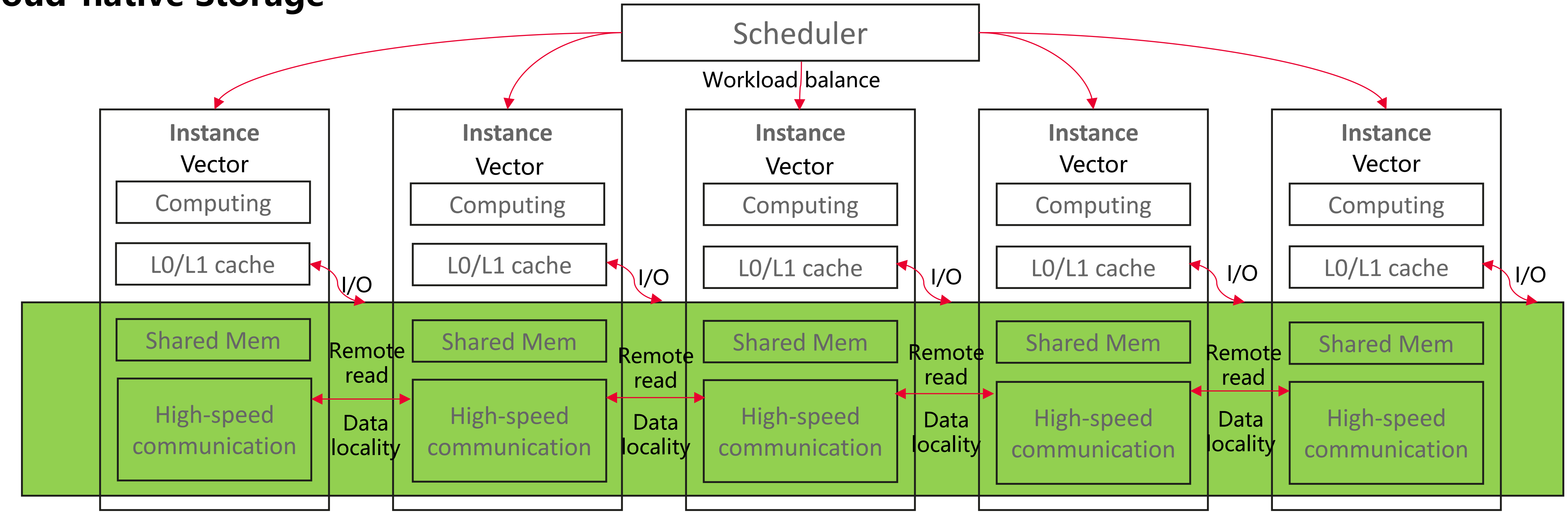
Vector Index on distributed database（10^9 < N ）

Shared-nothing Storage

Select * from table order by vec <-> '[1,123,0,0,...,3,1]' limit 10;

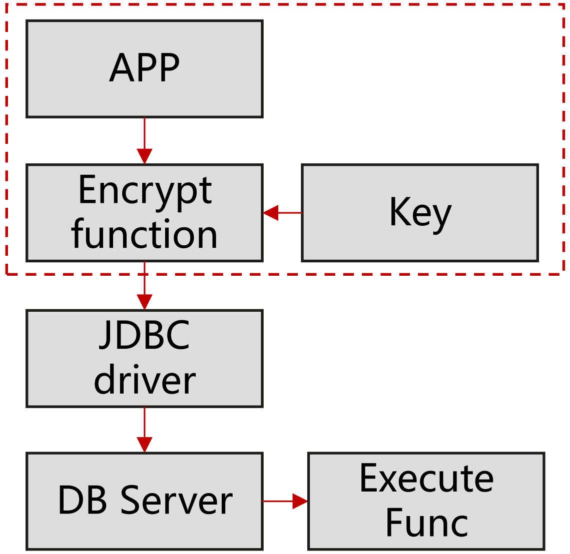


Cloud-native Storage



Data encryption for privacy protection

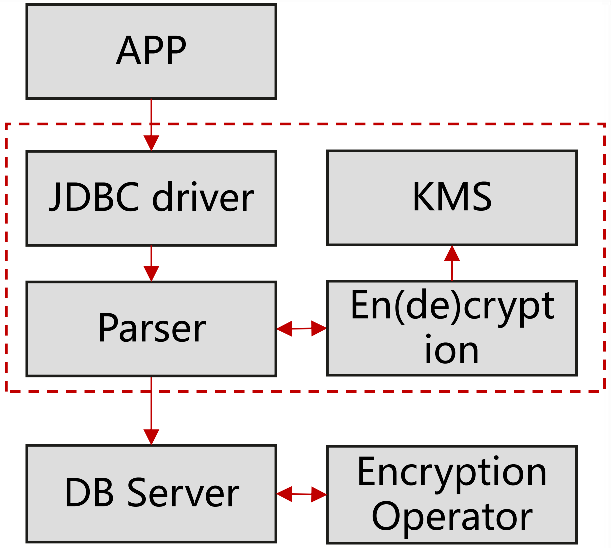
Functional Encryption



- column-level，user-side key，server-side encryption
- APP knows encryption, not support encrypted condition
- Data is decrypted in session, database cannot decrypted automatically, avoid data being stolen by other users.

VS

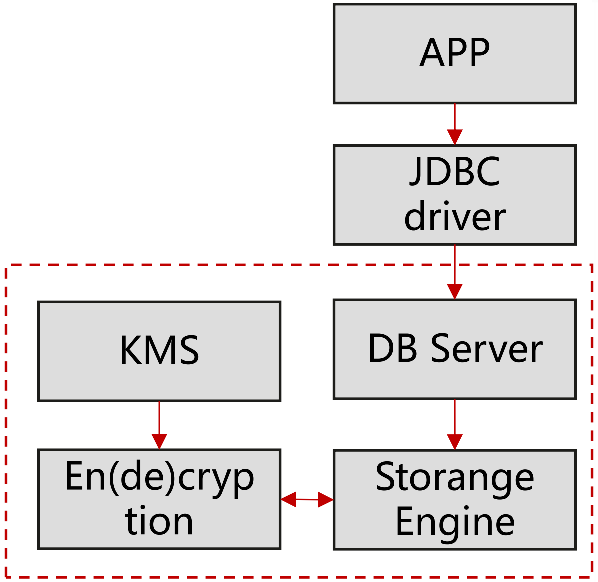
Fully Encryption



- Column-level, user-side key, driver embedded
- APP is not aware of encryption, support encrypted equal conditions.
- DB cannot decrypt data, avoid data leaking during OM activities.

VS

Transparent Data Encryption

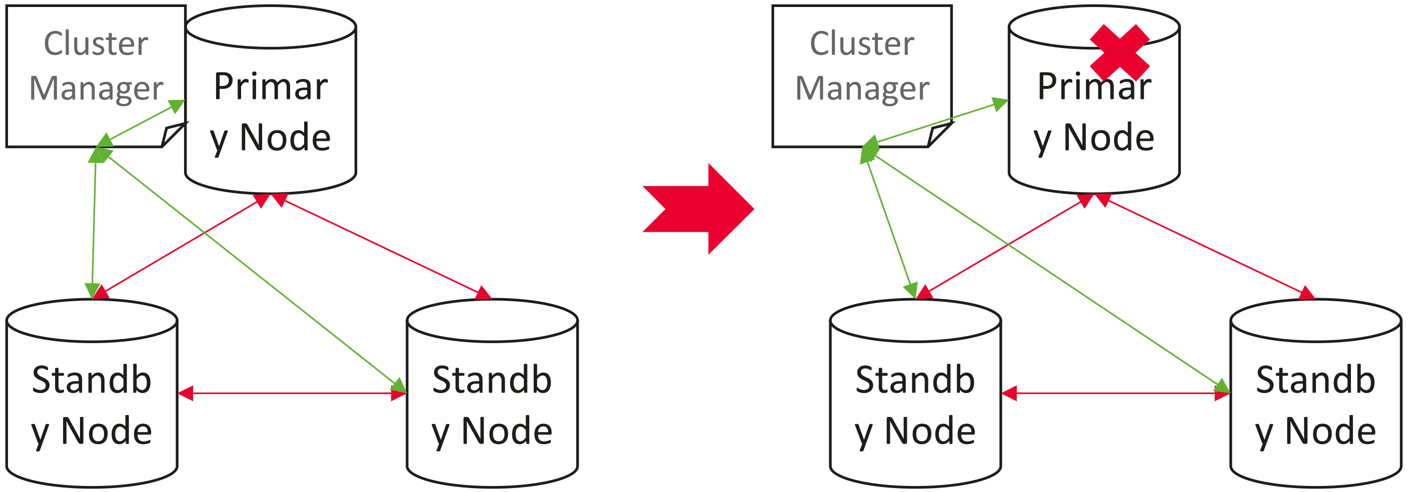


- Table-level, server-side key, server-side encryption
- APP is not aware of encryption, plaintext in the memory.
- Prevent data from being stolen on the disk.

High availability by data replication

XLOG Types	XLOG Description
XLOG_DISKANN_OPERATIONS	Index Operations
…	

Failure Recovery



1. Customizes xlog collections for vector index, supports synchronize between primary node and standby nodes;
2. Supports fast RTO < 10s for service recovery;
3. Supports data recovery, users can recover the data to any version when misoperation or disk damage happen (PITR).

Thanks!