Milvus

Build Up the Unstructured Data Service

Jun Gu

09.2020

ZILLIZ

Speaker bio

Jun Gu

Database engineer, SME

LF AI — Voting member in Technical Advisory Council (TAC)

ZILLIZ — Partner, Chief Evangelist

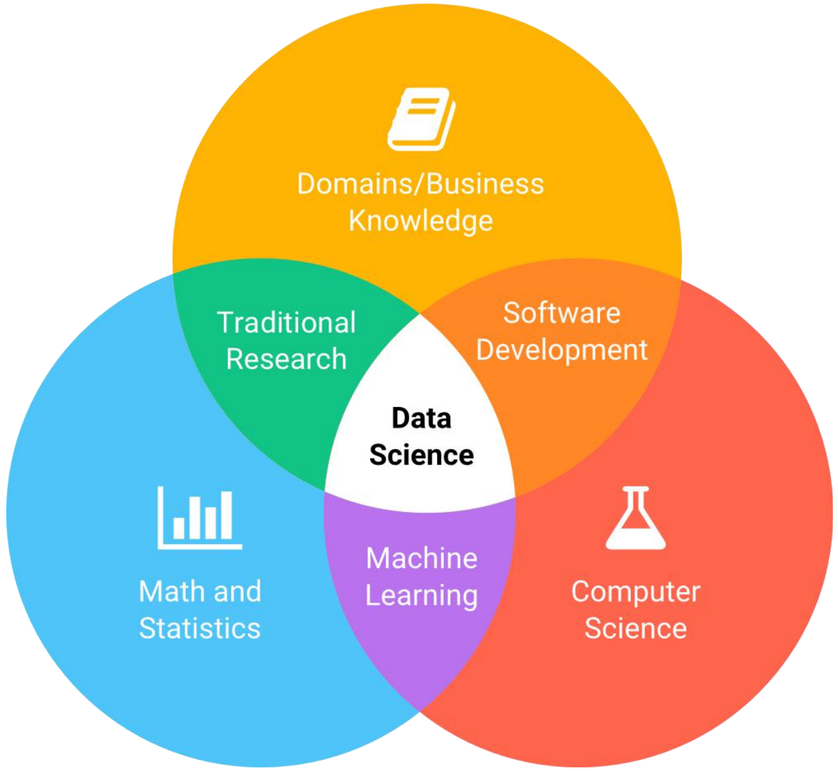
Career history

HUAWEI, Morgan Stanley, IBM, ICBC

Education

Peking University 1898

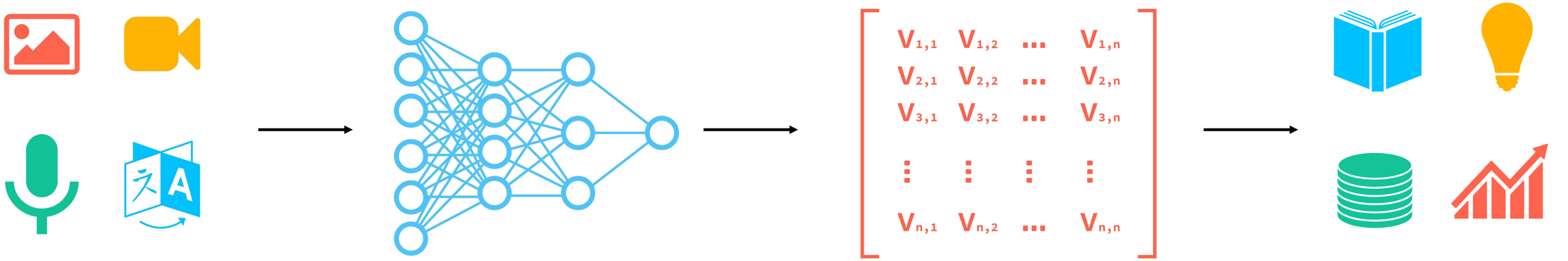
Zilliz: Who we are



- Open source software company based in Shanghai
- Mission: Reinvent data science
- Main contributor of Milvus project

Unlock the treasure of unstructured data

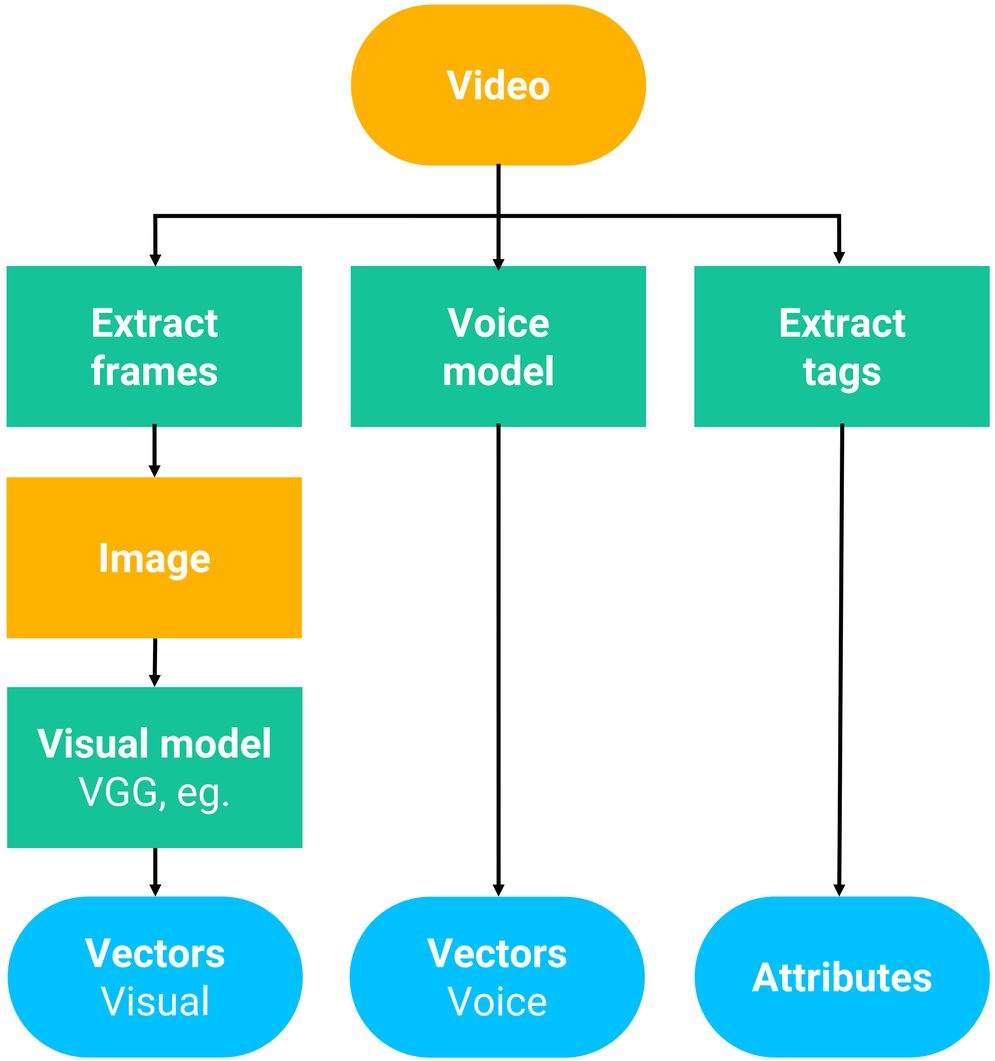
AI algorithms transform image, video, voice, natural language into vectors, and enables understanding and utilization of unstructured data at scale.



Unstructured data **Deep learning models** **Vectors** **Knowledge, insight, $**

Milvus LF AI INCUBATION PROJECT

The flow-based AI applications



The sample pipelines for video processing

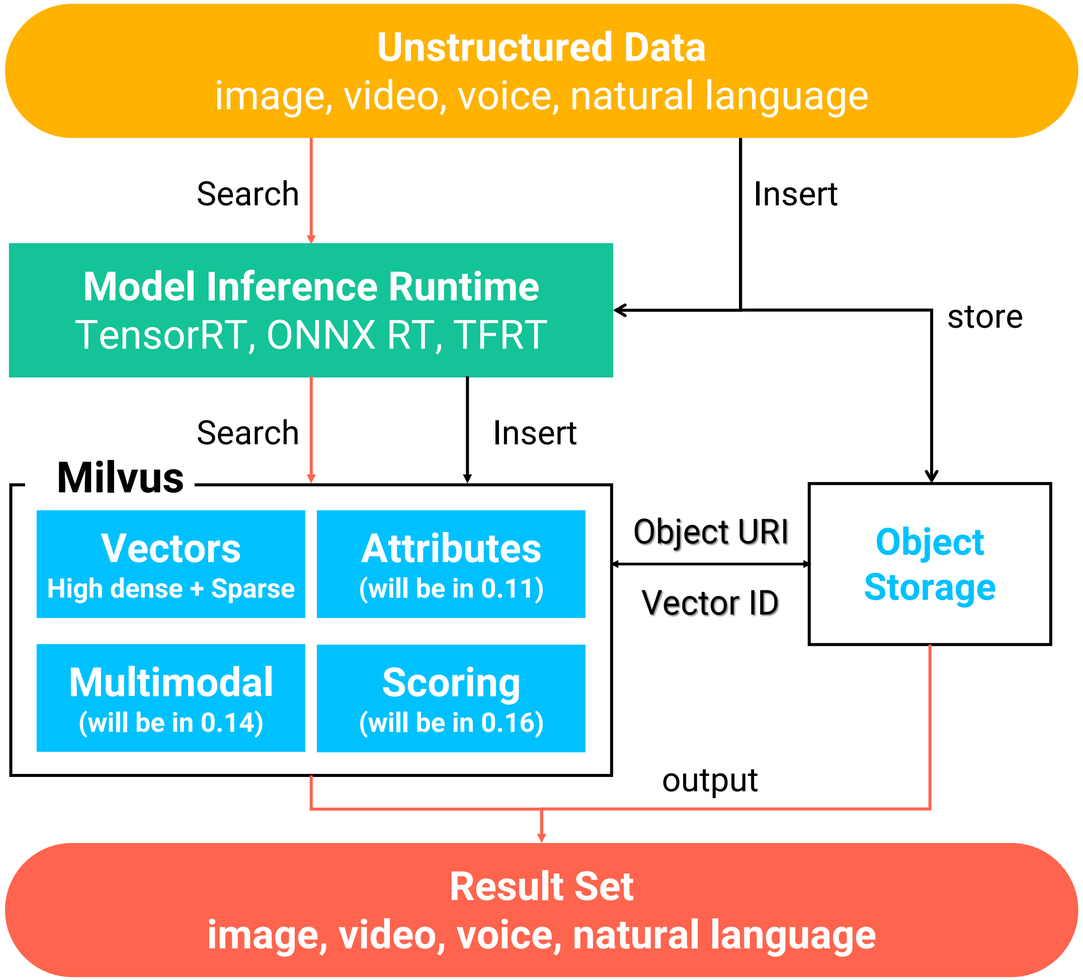
The most popular way

- Flexible
- Easy to compose, web-based UI
- Sample piplelines

The challenge

- Data fragmentation

The unstructured data service (UDS) for AI



Inference Layer

Data Service Layer

Why Milvus: Vectors are different

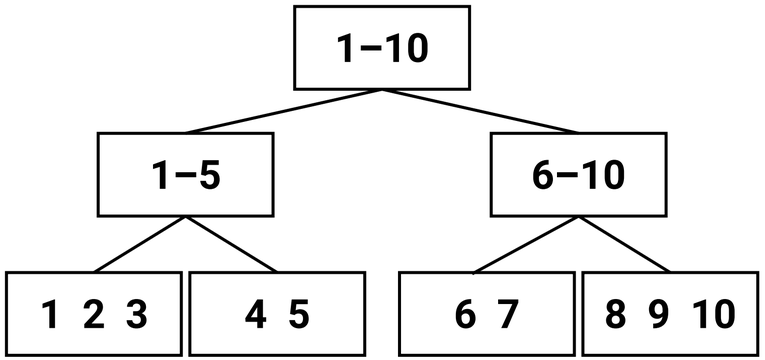
Numbers

Arithmetic operation

$+ - \times \div$

Number comparison

$a <=> b$



Operation

Organization

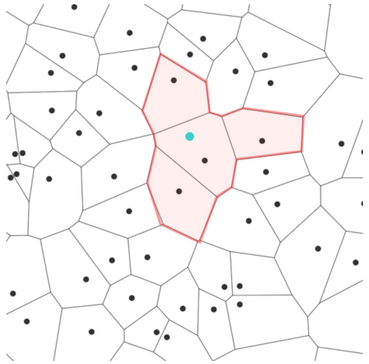
Vectors

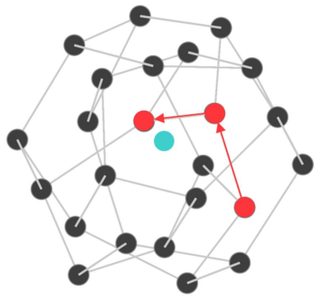
Similarity (eg. Euclidean distance)

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

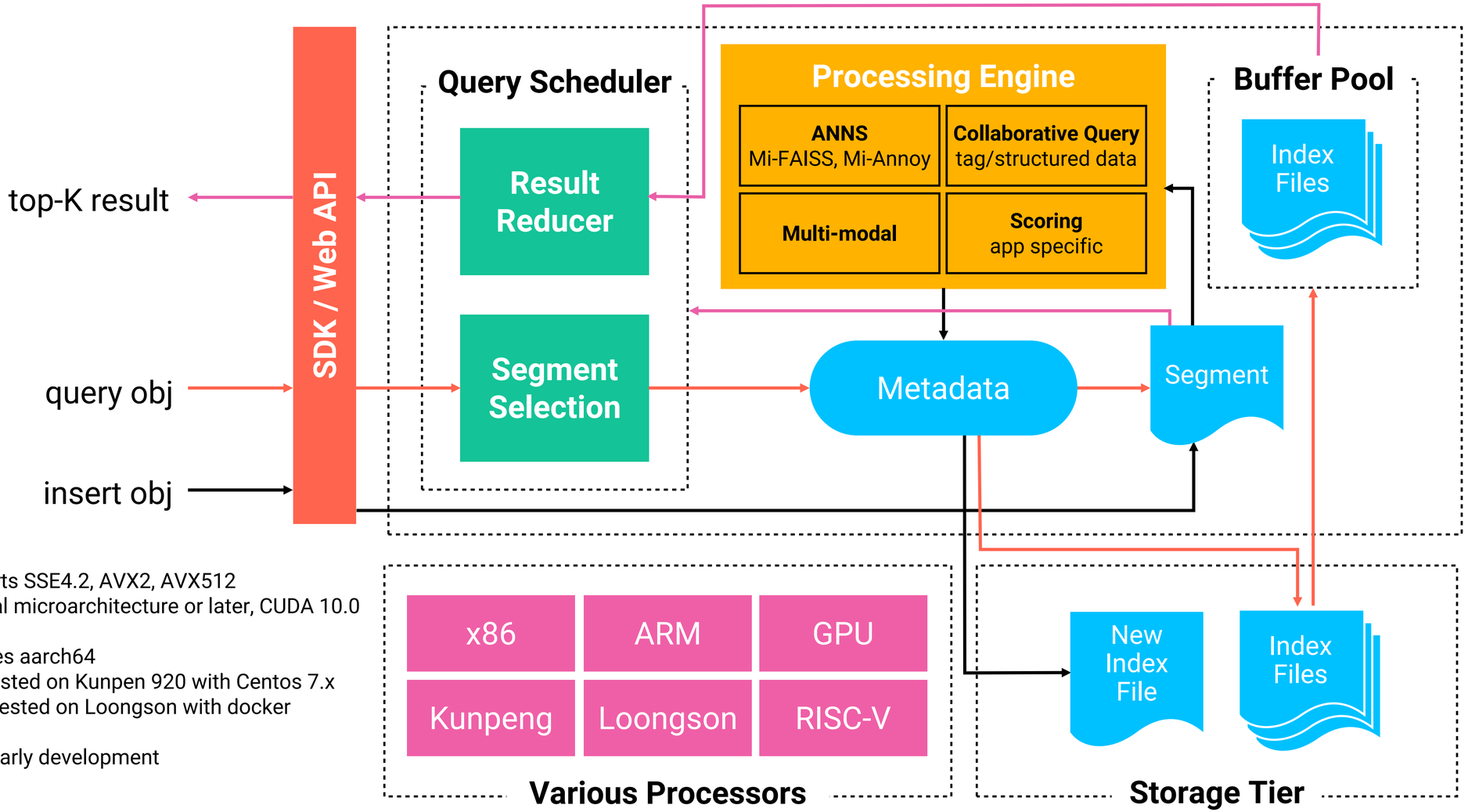
Similarity comparison

$$\mathrm{TopK}(A) = \arg\min_{B \in \gamma}(d(A,B))$$



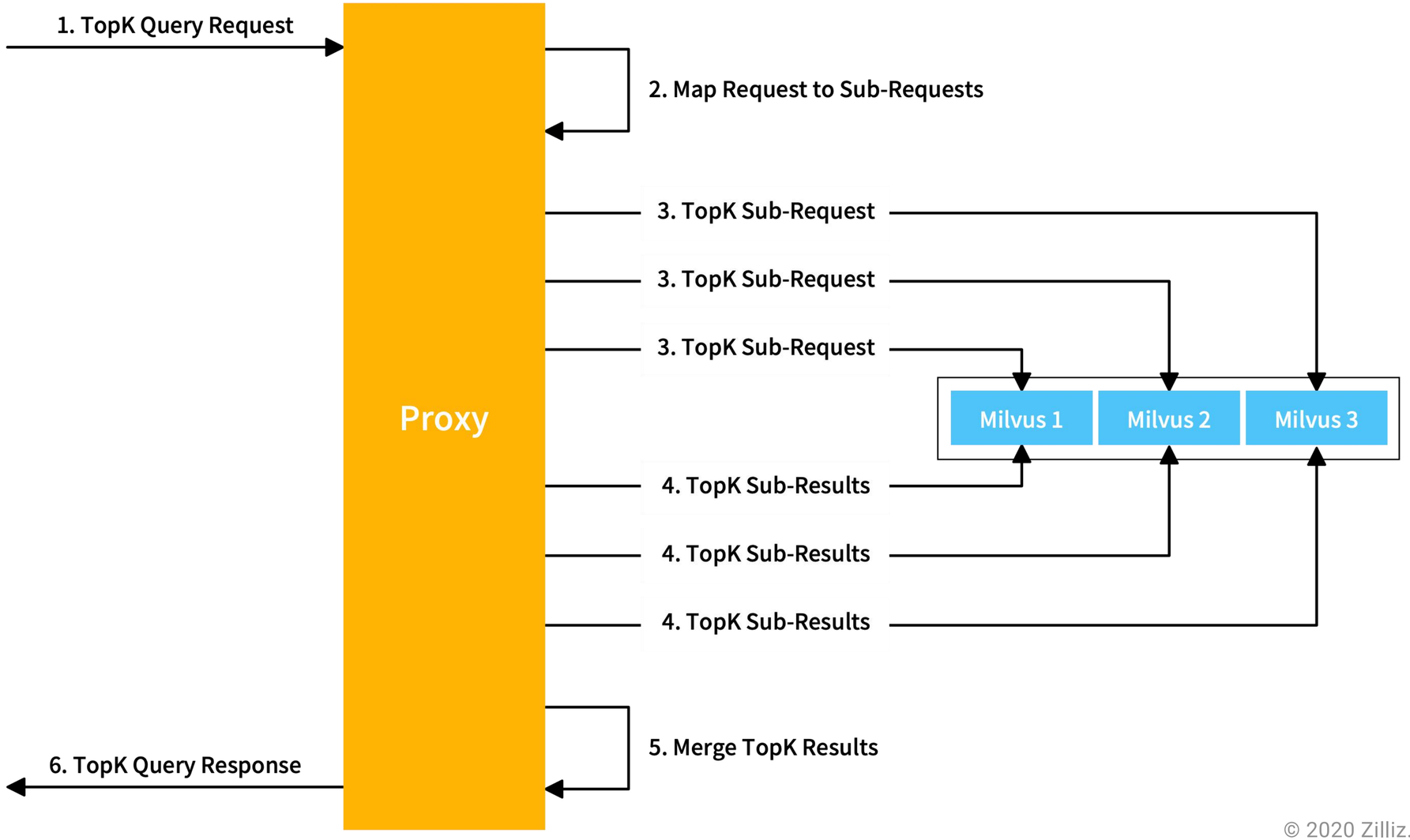


Milvus: The big picture



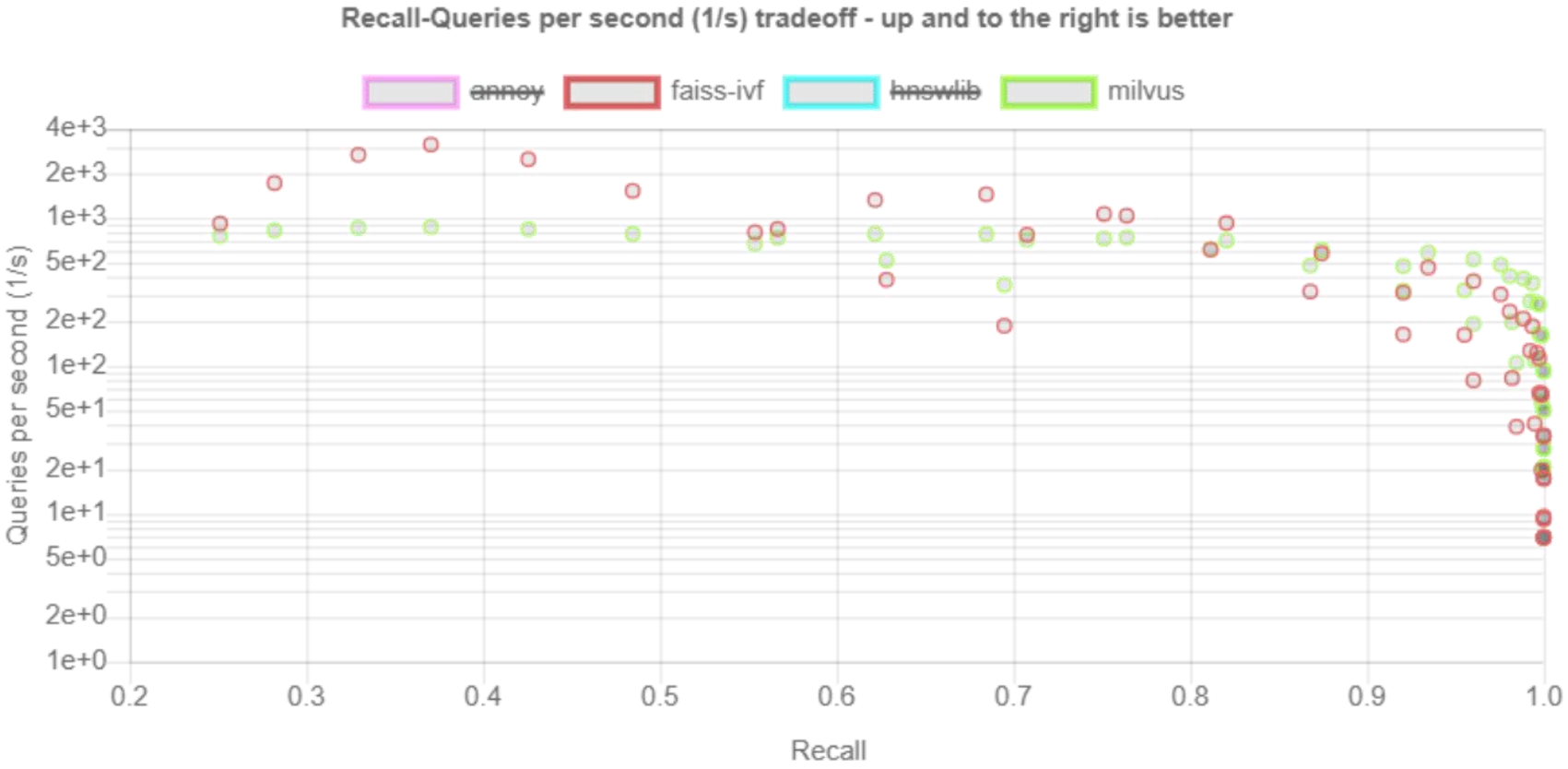
X86: supports SSE4.2, AVX2, AVX512
GPU: Pascal microarchitecture or later, CUDA 10.0 or later
Arm: requires aarch64
Kunpeng: tested on Kunpen 920 with Centos 7.x
Loongson: tested on Loongson with docker container
RSIC-V: in early development

Milvus: Distributed deployment



Milvus LF AI INCUBATION PROJECT

Milvus: The ANN benchmark

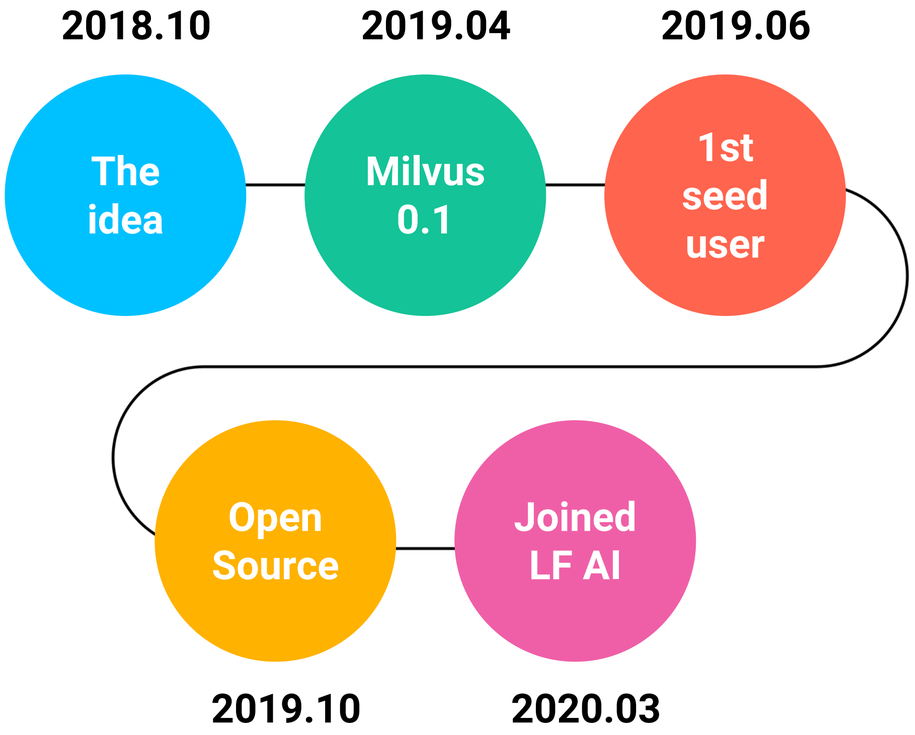


Milvus: 0.8.0
OS: Ubuntu 18.04
ECS: AWS c5.4xlarge (16c, 32GB), Intel XeonPlatinum 8275CL
Data set: sift-128-euclidean (1 million vectors)
More info: https://milvus.io/docs/benchmarks_aws

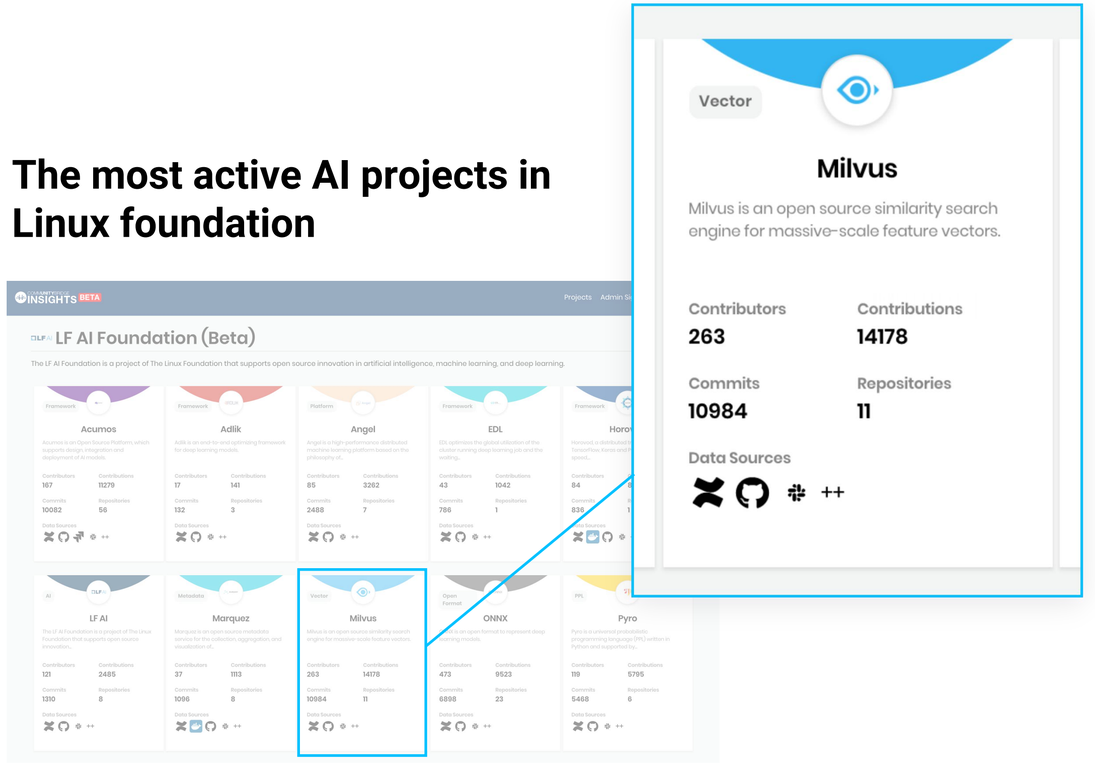
Milvus LF AI INCUBATION PROJECT

Special thanks to **ANN-Benchmarks** (developed by **Martin Aumueller**, **Erik Bernhardsson** and **Alec Faitfull**)

Milvus: The journey



The most active AI projects in Linux foundation



Milvus LF AI INCUBATION PROJECT

Progress

Unstoppable momentum since its debut.

5.9K
Commits

3.9K
GitHub stars

104
Contributors

16
Release

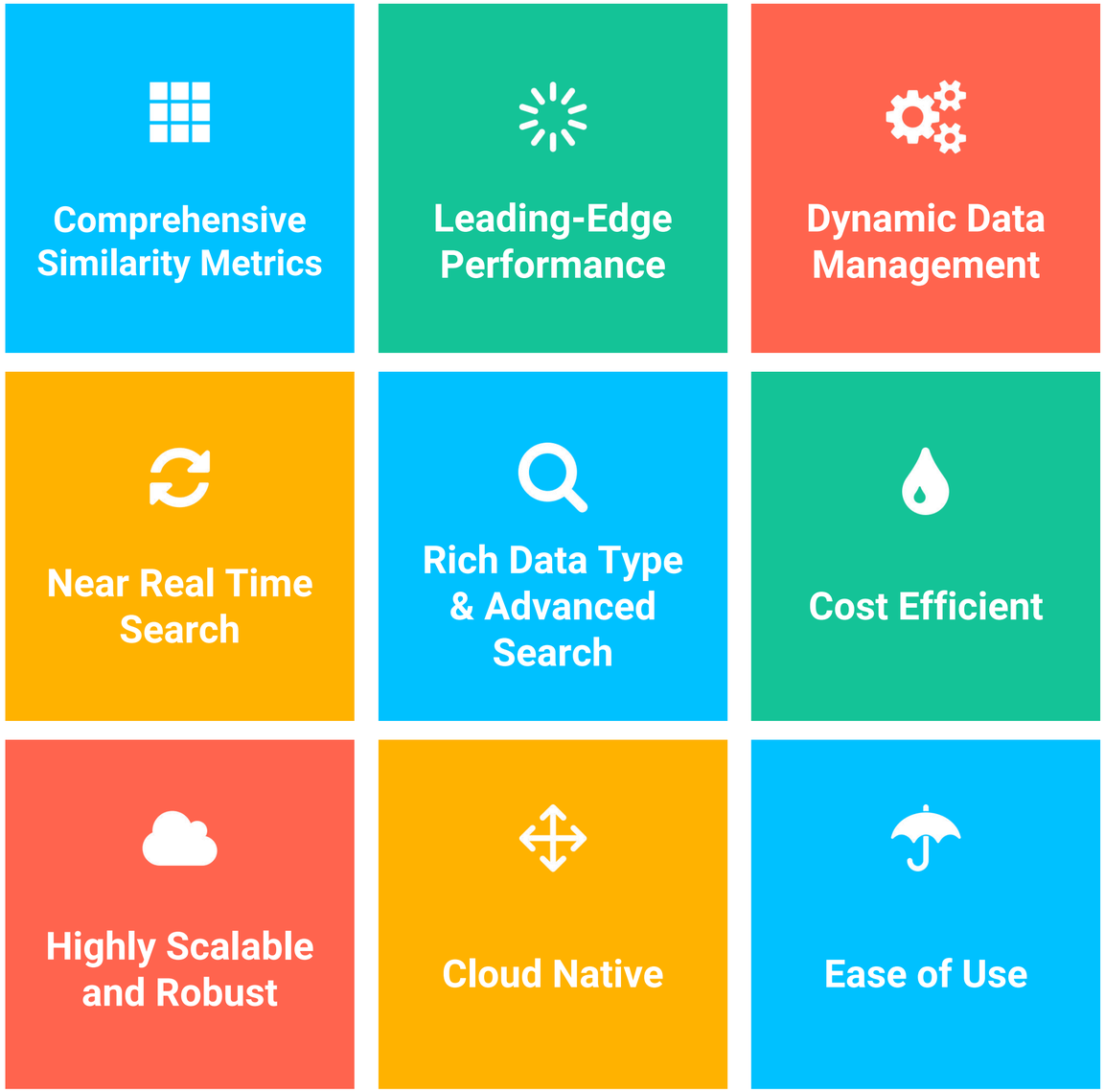
200+
Users

19
Patents filed

Milvus LF AI INCUBATION PROJECT

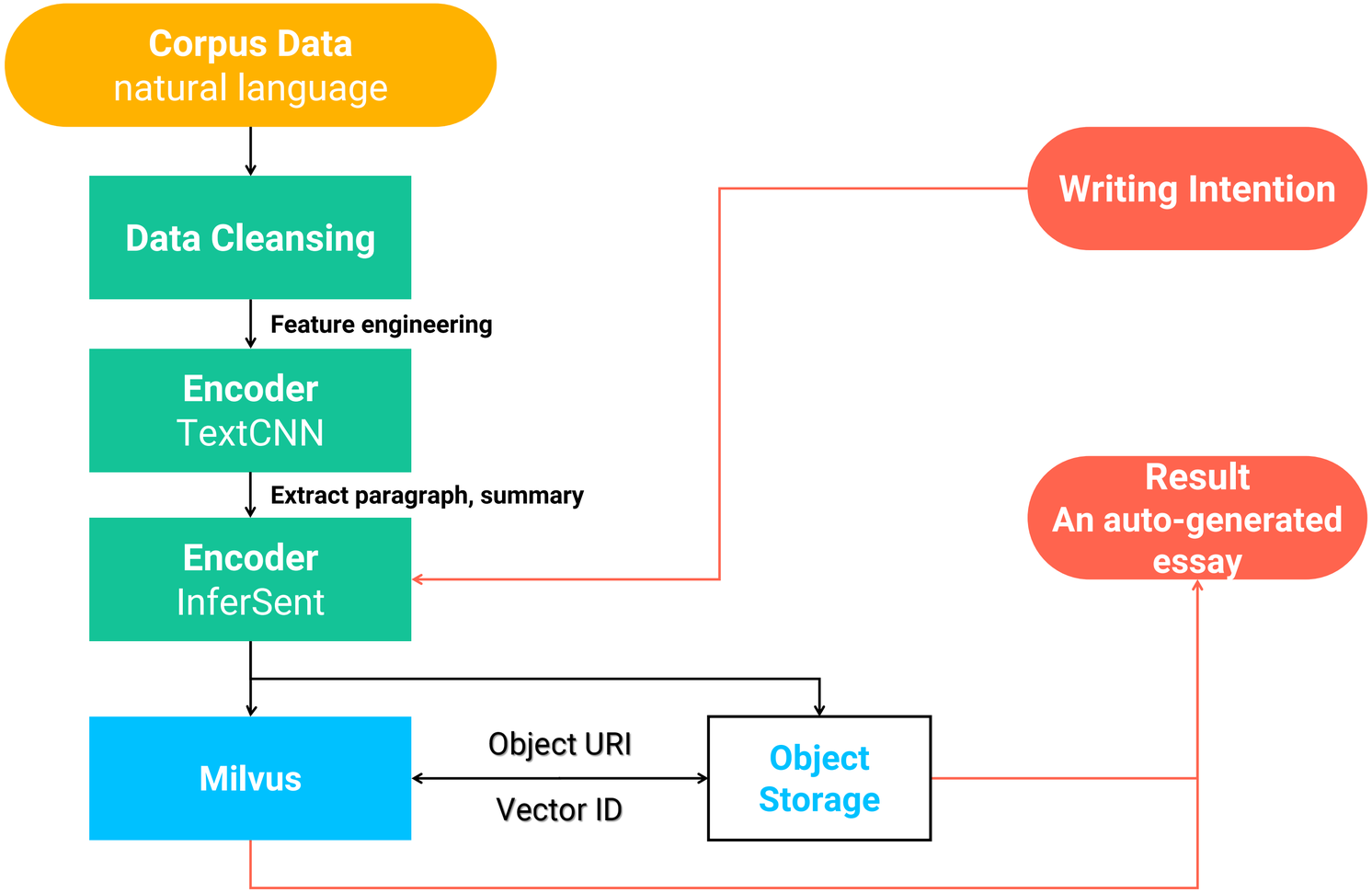
Milvus
Features & benefits

The world’s most advanced, our target ☺



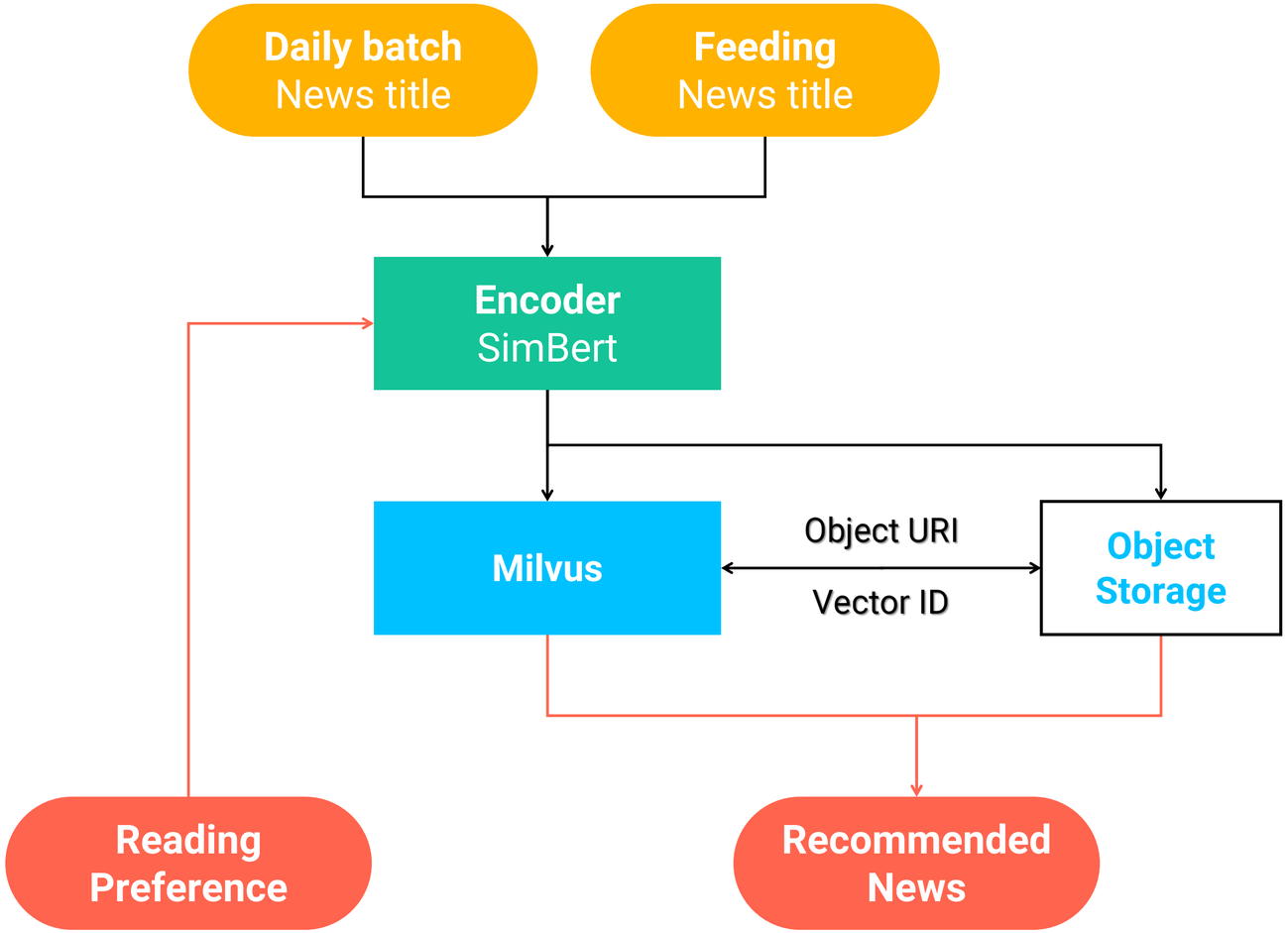
Milvus LF AI INCUBATION PROJECT

Use case: Inteligent writing assistant

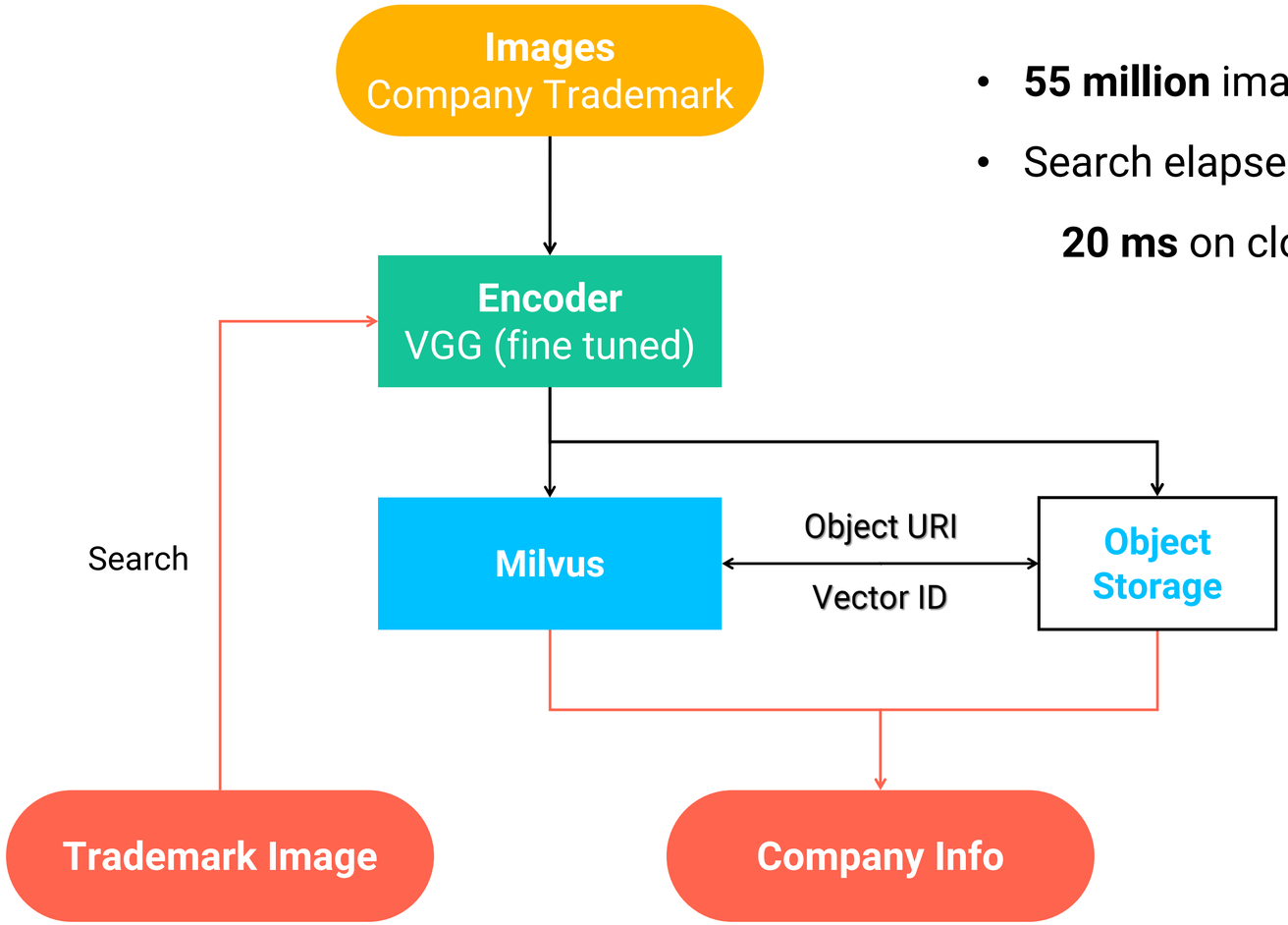


Milvus LF AI INCUBATION PROJECT

Use case: News recommendation on mobile



Use case: Image search for company trademark



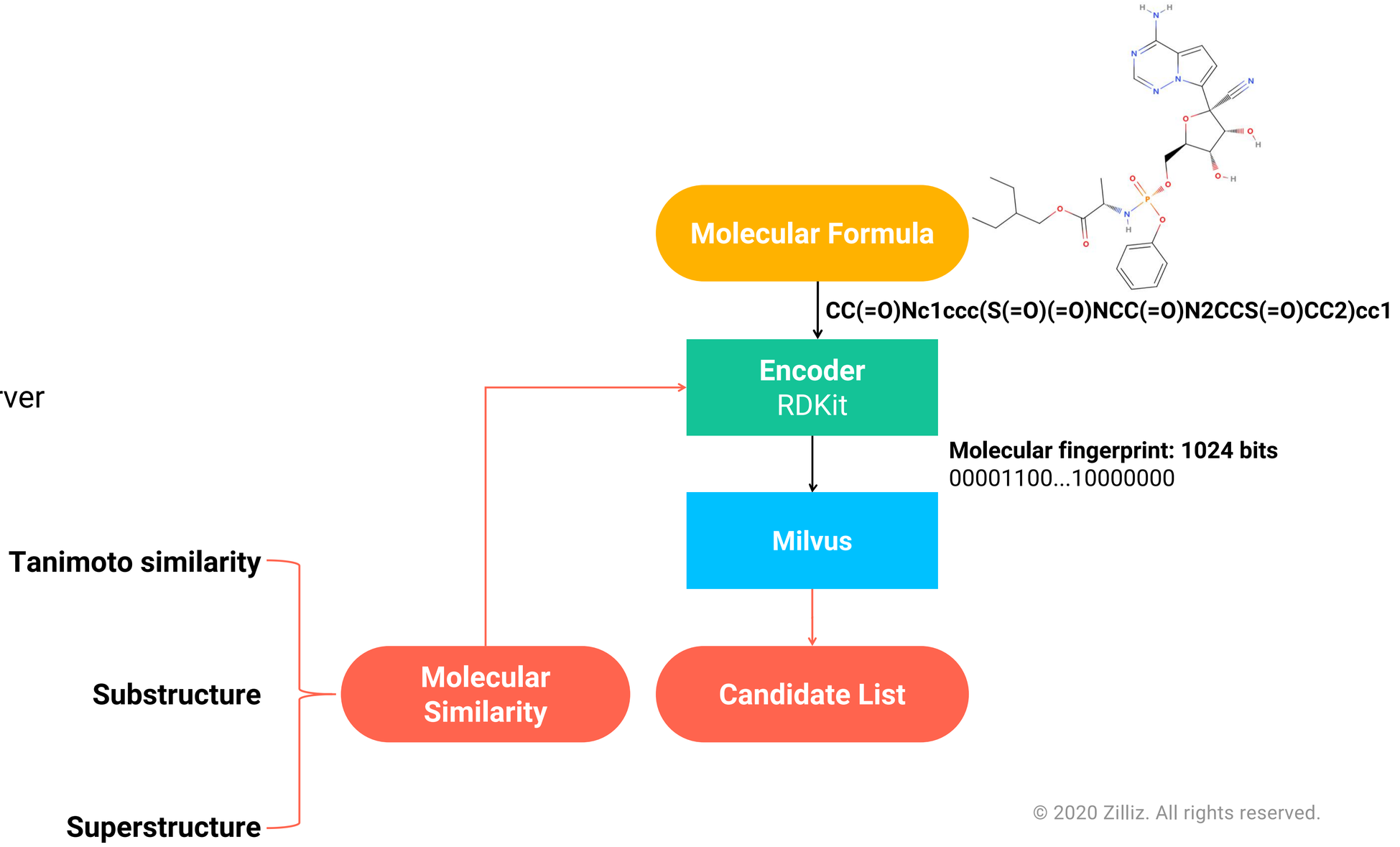
- **55 million** images
- Search elapsed time:

  20 ms on cloud GPU server

Use case: Pharmaceutical molecule analysis

- **800 million** molecules
- Search elapsed time:

  500 ms on single server



Useful Links

https://milvus.io

https://github.com/milvus-io/milvus

https://milvusio.slack.com

https://twitter.com/milvusio

https://medium.com/unstructured-data-service

https://zhuanlan.zhihu.com/ai-search

Performnance benchmark:
https://milvus.io/docs/benchmarks_aws

Live demo:
https://milvus.io/scenarios

- Content-based image retrieval system (以图搜图)
- Q&A chatbot powered by NLP (智能客服机器人)
- Molecular analysis (化合物分析)

Thanks!