华为开发者联创日

技术无界 创想无限

# 向量数据库

赋能开发者高效构建私域大模型应用

# -- 目录 --

CONTENTS


# 李剑楠

上海爱可生信息技术股份有限公司
高级研发工程师

# 向量数据库原理解析

# 向量数据库

int, float, string, ...

01234
56789
$e$ $\pi$
ABCDE
2023.04.10

text

**Abstract**

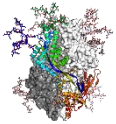
Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real-time data serving). Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products. In this paper we describe the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and we describe the design and implementation of Bigtable.

json


```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```


image
video
audio

domain specific

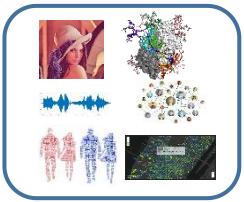


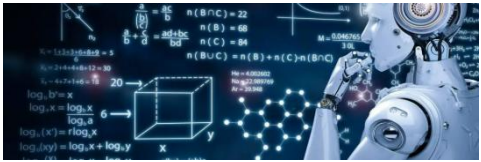
Structured data

Unstructured data

# 向量数据库

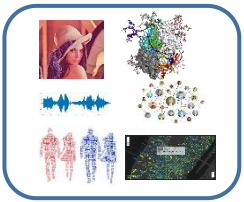
机器理解数据



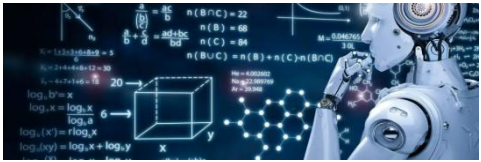


# 向量数据库

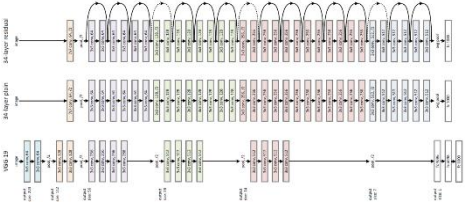
机器理解数据

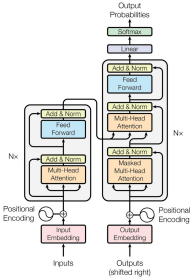


?



AI模型/编码手段

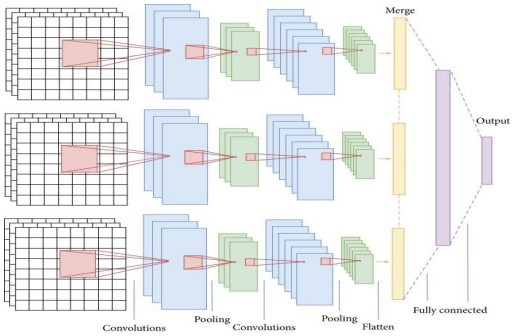




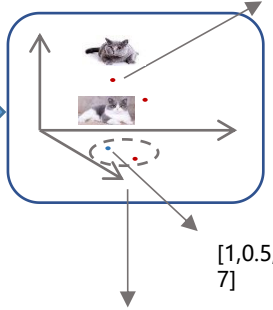
# 向量数据库

query

卷积神经网络




向量空间

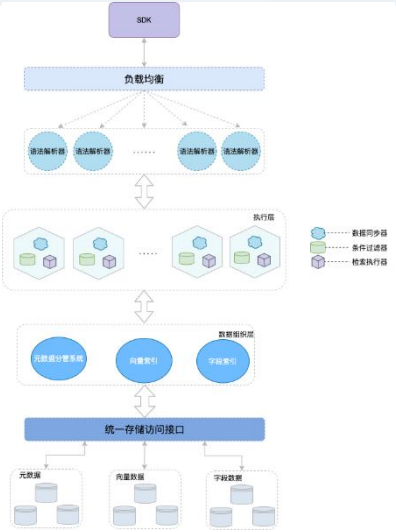


[1.4,2.1,3.1]

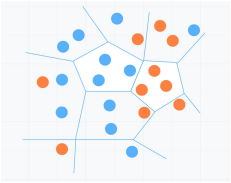
[1,0.5,0.7]

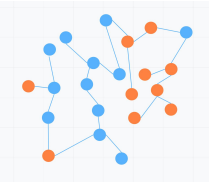
result

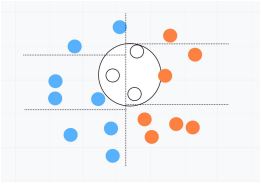
# 向量数据库原理解析

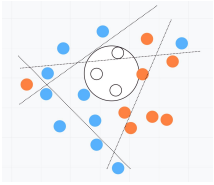



产品架构图





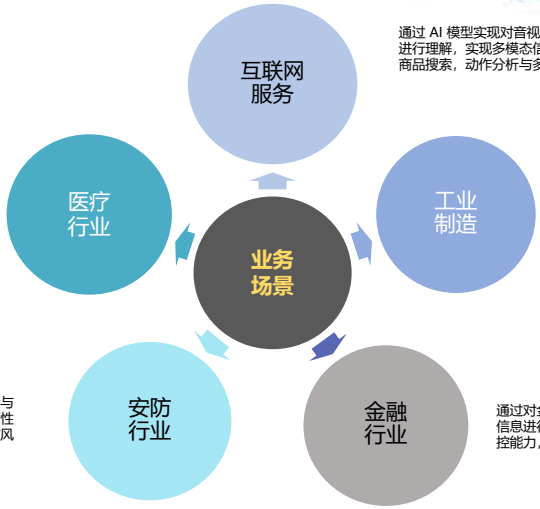




ANNS(Approximate Nearest Neighbor Search)
算法

# 向量数据库应用场景

# 向量数据库应用场景




通过 AI 模型实现对音视频，图像，文本等非结构化数据进行理解，实现多模态信息检索能力，例如以图搜图，商品搜索，动作分析与多模态数据推荐等泛互联网场景。

通过对生物分子结构进行分析，实现蛋白质性质预测，智能病理分析，智能问诊，缓解目前医疗行业资源紧张的问题。

通过对工业数据进行理解与分析，将工业生产中的流程，图像，视频等数据转化为高维向量进行存储与检索，赋能工业图像检测，质量监控与良率分析等多角度工业场景。

通过对智能物联网设备收集到的数据进行存储与分析，解决目前安防行业需要大量人工与重复性工作的问题，提高智能安防，数据智能归档与风险监测等业务场景的服务能力。

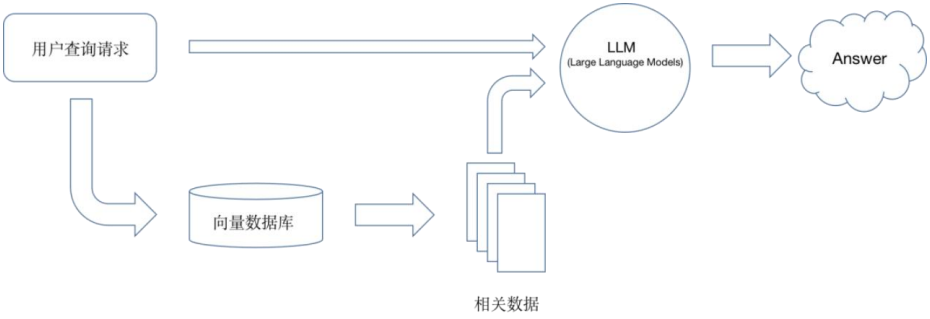
通过对金融行业中的用户，行为，图像，视频等信息进行高维向量提取，加速金融支付，提高风控能力，保障金融行业服务质量稳步提升。

# 大模型时代下向量数据库应用场景

## 舆情看管

# 大模型时代下向量数据库应用场景

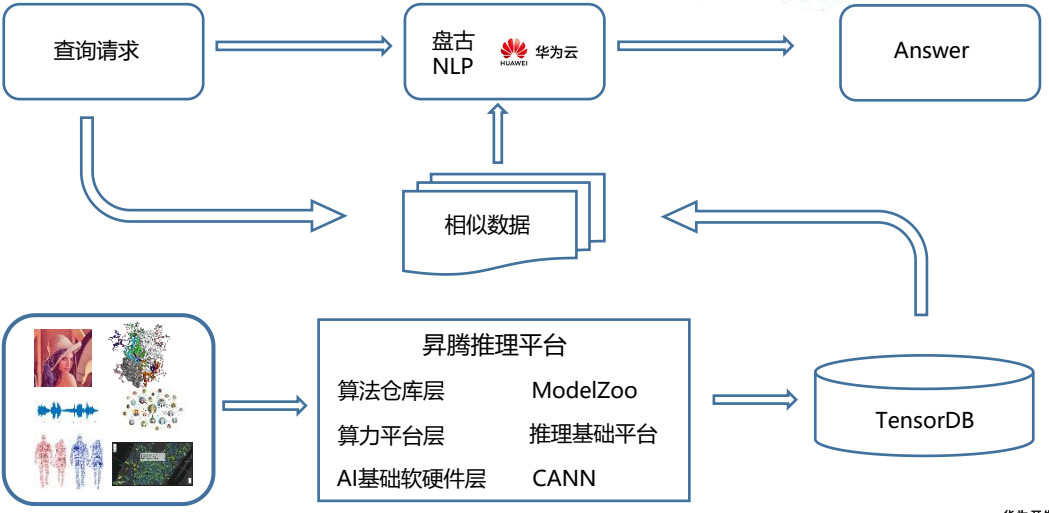
本地知识库




向量检索技术中核心的ANNS(Approximate Nearest Neighbor Search)算法在处理超大量规模的向量场景下，可以表现出很好的召回率和查询时延。在向量检索技术的加持下，我们可以为大模型提供更精准的提示词或上下文片段。例如将文章拆分成更多的片段、将同一段数据生成多个向量表示等。

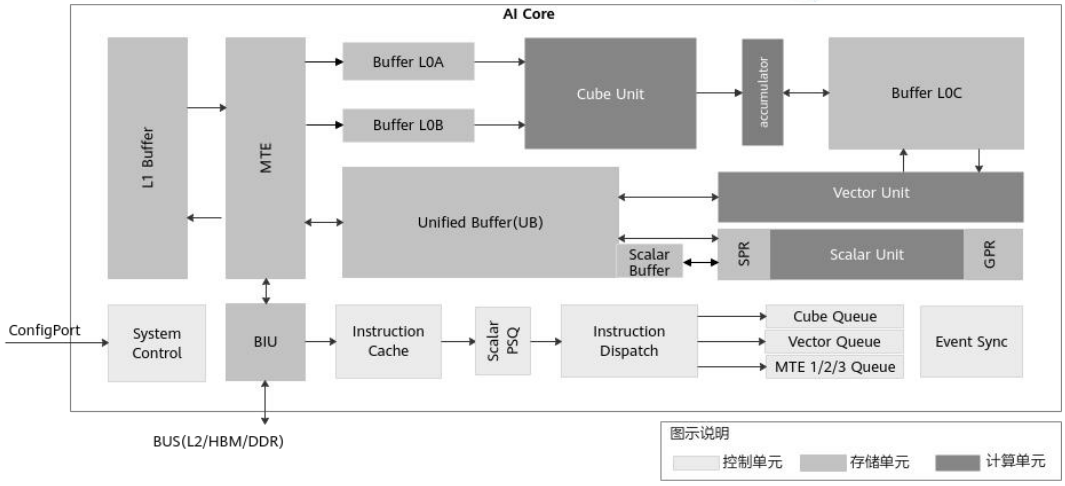
# 向量数据库、昇腾软硬件平台快速搭建大模型应用

# 基于昇腾、TensorDB快速构建大模型应用

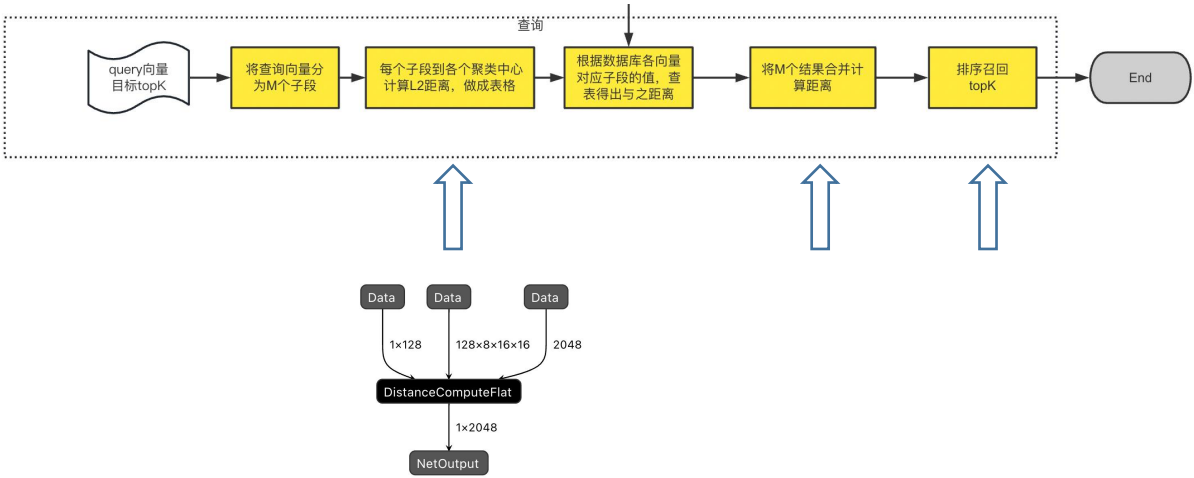


# 向量数据库与昇腾NPU硬件的结合

# 向量数据库与昇腾NPU硬件的结合



# 向量数据库与昇腾NPU硬件的结合



# 向量数据库与昇腾NPU硬件的结合

**使用MindSpore框架搭建的Kmeans训练模型**

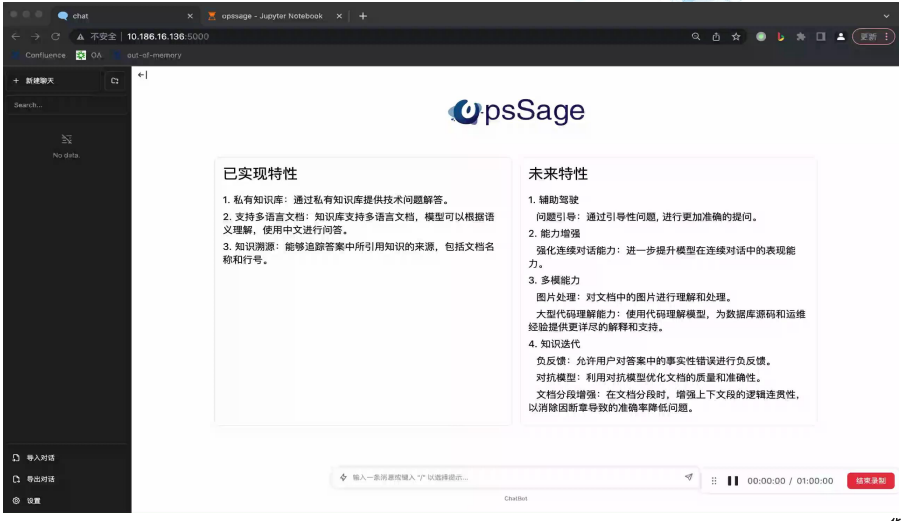
| 实现方式 | Pytorch-CPU | Pytorch-GPU | MindSpore-CPU | MindSpore-GPU |
|---|---|---|---|---|
| 训练速度（s/iter） | 0.08～0.8 | 0.13 | 0.14 | 0.06 |
| 查询召回率 | 98% | 98% | 99.3% | 99.3% |

**全应用流程中使用到的部分自定义算子测试结果**

| 性能指标 | 昇腾Tik-Matmul算子 | numpy实现Matmul | 昇腾DSL-Hamming算子 | numpy实现Hamming |
|---|---|---|---|---|
| 平均单次执行速度(ms) | **0.18** | 0.7 | **0.015** | 0.05 |

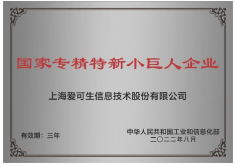
# 应用演示

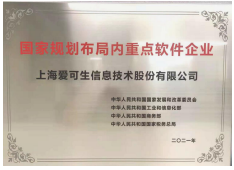
# 应用演示


chat
opssage - Jupyter Notebook
不安全 | 10.186.16.136:5000
更新
Confluence
QA
aut-of-memory
新建聊天
Search...
No data
导入对话
导出对话
设置
OpsSage
已实现特性
1. 私有知识库：通过私有知识库提供技术问题解答。
2. 支持多语言文档：知识库支持多语言文档，模型可以根据语义理解，使用中文进行问答。
3. 知识溯源：能够追踪答案中所引用知识的来源，包括文档名称和行号。
未来特性
1. 辅助驾驶
问题引导：通过引导性问题，进行更加准确的提问。
2. 能力增强
强化连续对话能力：进一步提升模型在连续对话中的表现能力。
3. 多模能力
图片处理：对文档中的图片进行理解和处理。
大型代码理解能力：使用代码理解模型，为数据库源码和运维经验提供更详尽的解释和支持。
4. 知识迭代
负反馈：允许用户对答案中的事实性错误进行负反馈。
对抗模型：利用对抗模型优化文档的质量和准确性。
文档分段增强：在文档分段时，增强上下文段的逻辑连贯性，以消除因断章导致的准确率降低问题。
输入一条消息或键入 "/" 以选择提示...
00:00:00 / 01:00:00
结束录制
ChatBot

# 公司介绍

# 公司介绍









- **国家规划布局内的重点软件企业、国家级专精特新“小巨人”企业。**
- **国家科技部国家重点研发计划-“网络协同制造和智能工厂” 重点专项。**
- 已拥有80多项软件著作权和产品登记证书和**5项已授权核心技术发明专利**（另45+项数据技术专利申报中）。
- 核心产品为多库融合、数据智能应用等产品。
- 数据库标杆客户：工商银行、交通银行、农业银行、兴业银行、中国人寿、太平洋保险、中国人保、国家电网、上汽大众、中国移动、中国电信、华为等50+世界五百强国计民生核心业务的后台数据技术。
- 承担了工信部5所、工业互联网创新中心（上海）、航天科技集团、国家电网等多个重点数据项目。
- 2020年，公司获中国信息协会2020-2021年度信息技术应用创新优秀解决方案奖。
- 2021年上海市重点支持的信创-基础软件企业。

# 社区介绍

https://opensource.actionsky.com/

# 谢 谢