

PostgresConf.CN 2024

中国 PostgreSQL 数据库生态大会 ● / ◆

基于PieDataCS的AI多模态数据平台

徐阳

OpenPie & 资深解决方案架构师

PCC POSTGRESCONF
CN 2024

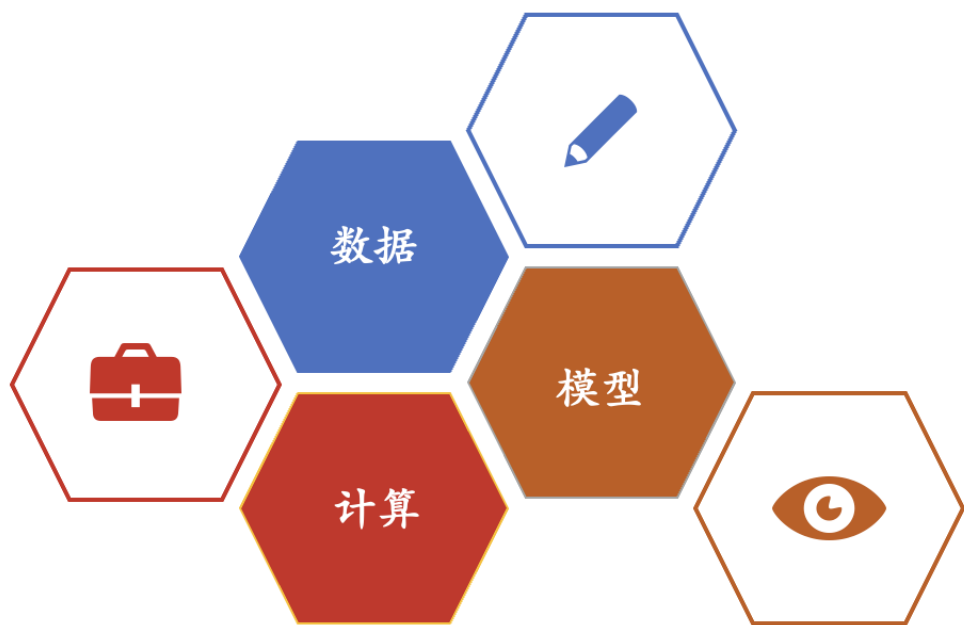
/ 2024.01.04 /



CHINA
POSTGRES
SQL
ASSOCIATION

PieDataCS的产品理念及定位

大模型数据计算系统，以云原生技术重构数据存储和计算，一份数据，多引擎数据计算，AI模型、数据和计算三者互为增强，全面升级大数据系统至大模型时代，赋能行业AI场景应用。



1

灵活扩展的数据引擎，支持关系型数据库SQL、Spark/Flink等流批一体处理、向量数据计算、全文检索以及时空时序等。

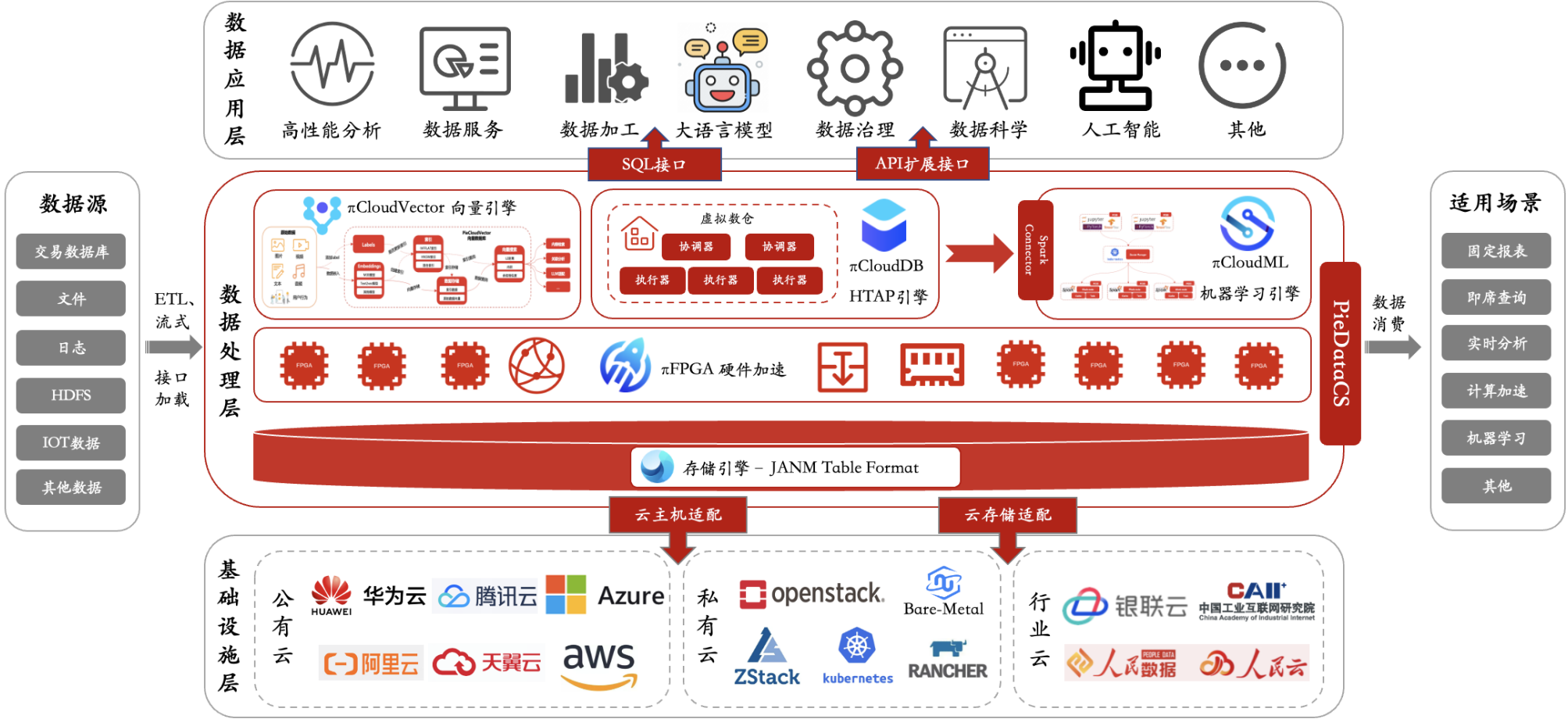
2

具备整体数据平台方案，支持多模数据处理（结构化、半结构化以及非结构化数据），实现数据共享和分析。

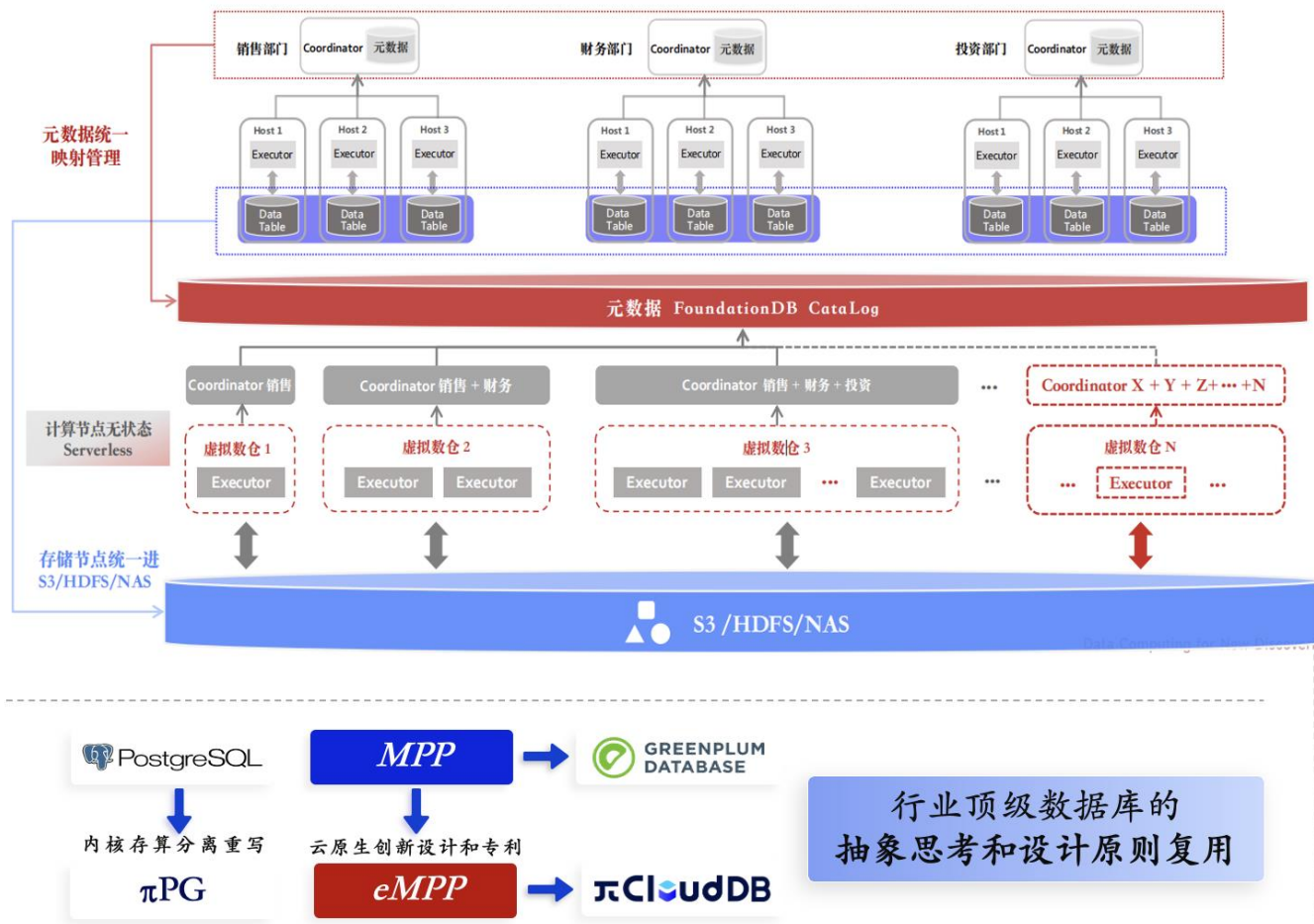
3

软件优化 + 新硬件（FPGA）加速，实现数据全链路性能飞跃，让数据存储、SQL查询、向量计算以及机器学习等能力全面升级。

PieDataCS整体架构



国内首创虚拟数仓技术



打破数据孤岛，解决数据多副本

物理数仓整合到云原生数据计算平台，根据数据授权动态创建虚拟数仓，打破数据孤岛，解决数据多副本问题，帮助企业降低数仓管理复杂度，以更低的成本实现存算资源在云上更灵活的配置。

提升数据计算资源利用效益

数据计算资源按需扩缩容，实现计算资源配置最优化，提升数仓的敏捷性和弹性，打开无限数据计算空间，支撑更大模型所需的数据和计算。更好地赋能业务发展并走向绿色。

坚如磐石 | 高安全 高在线 高可靠

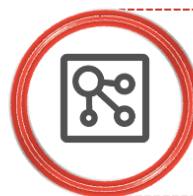
TDE技术保证了所有数据在落盘前完成加密，服务器无感知技术（Serverless）利用云上无限计算资源和弹性保证了虚拟数仓永远在线可用，S3存储和跨云灾备能力保证了永不丢数。

分布式SQL优化器

达奇

PieCloudDB Optimizer

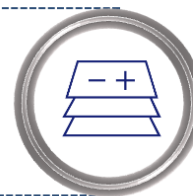
- 基于eMPP架构的云原生分布式优化器；
- 为海量数据集上的复杂OLAP查询提供最优的查询计划。



聚集下推

通过把聚集操作下推到连接操作之前去执行，极大的减少连接操作需要处理的数据量，使得查询性能显著提升。在很多场景下经过严格的测试，聚集下推会取得百倍或千倍的性能提升。

预计算



根据对象存储的特性，为每一个数据块预先计算好块内数据的聚集值。通过把聚集操作下推到对数据表的扫描之上，可以直接利用数据块内预计算的结果，从而避免了大量的计算。



文件裁剪

利用每个数据块内预计算的结果，提前过滤掉不符合条件的数据块，从而减少对数据的扫描和后续的计算。



多表连接的最优
顺序搜索



多阶段聚集



分区表的静态
和动态裁剪



相关子查询的
提升转换



CTE和递归
CTE的优化



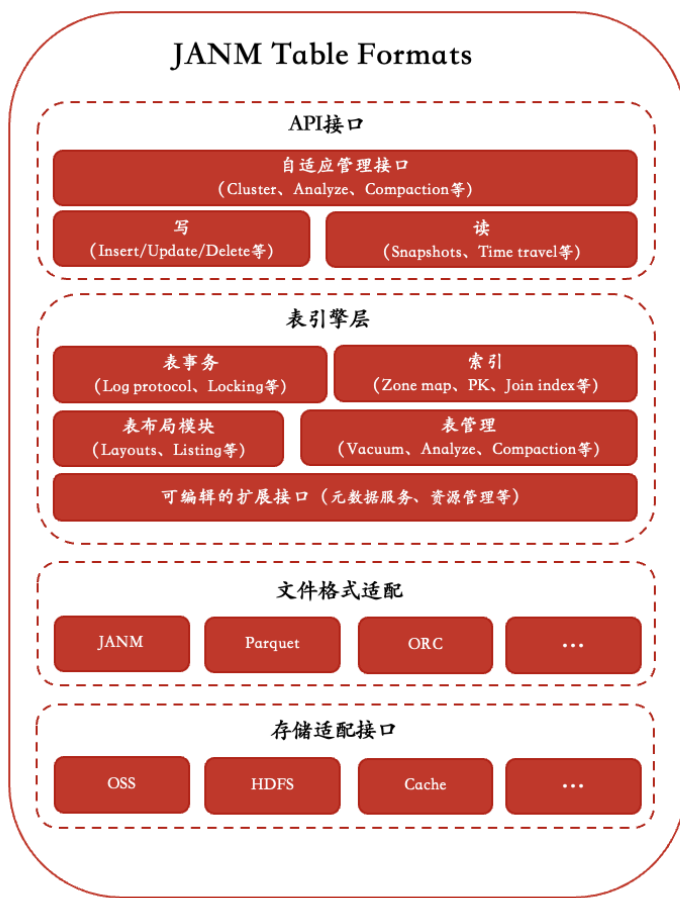
其他相关优化

JANM存储引擎设计

简墨Table Formats，数据存储引擎，构建统一的数据访问协议，将包含JANM表格式以及其他存储引擎或表格式的访问统一起来，打破数据孤岛，为企业提供一种可扩展且可靠的方式来管理海量的数据表格。

原始表 (table)

Column 1	Column 2	...	Column N



存储兼容性:

- 抽象API
- 对象存储、文件系统、缓存等任意引擎

文件格式:

- 自研Janm (简墨) 文件格式, 行列混存
- 数据压缩
- 主流文件格式兼容, 如Parquet/ORC等
- 统一访问接口
- 记录文件的更改

表格式定义:

- 统一表格式标准, 可扩展的接口
- 文件级别MVCC, 并发控制
- 多种索引类型, Data Skipping
- 支持对表细粒度的优化操作
- 事务性操作, 日志即数据

API接口:

- 完善的表操作应用服务
- 基于快照的操作
- 自适应的数据管理
- 可快速注册到任意服务中

PieCloudVector向量数据计算引擎



数据存储:

- 原始数据
- 向量数据
- 向量压缩

索引管理:

- 支持主流向量索引
- 索引缓存加速
- 支持BM25索引

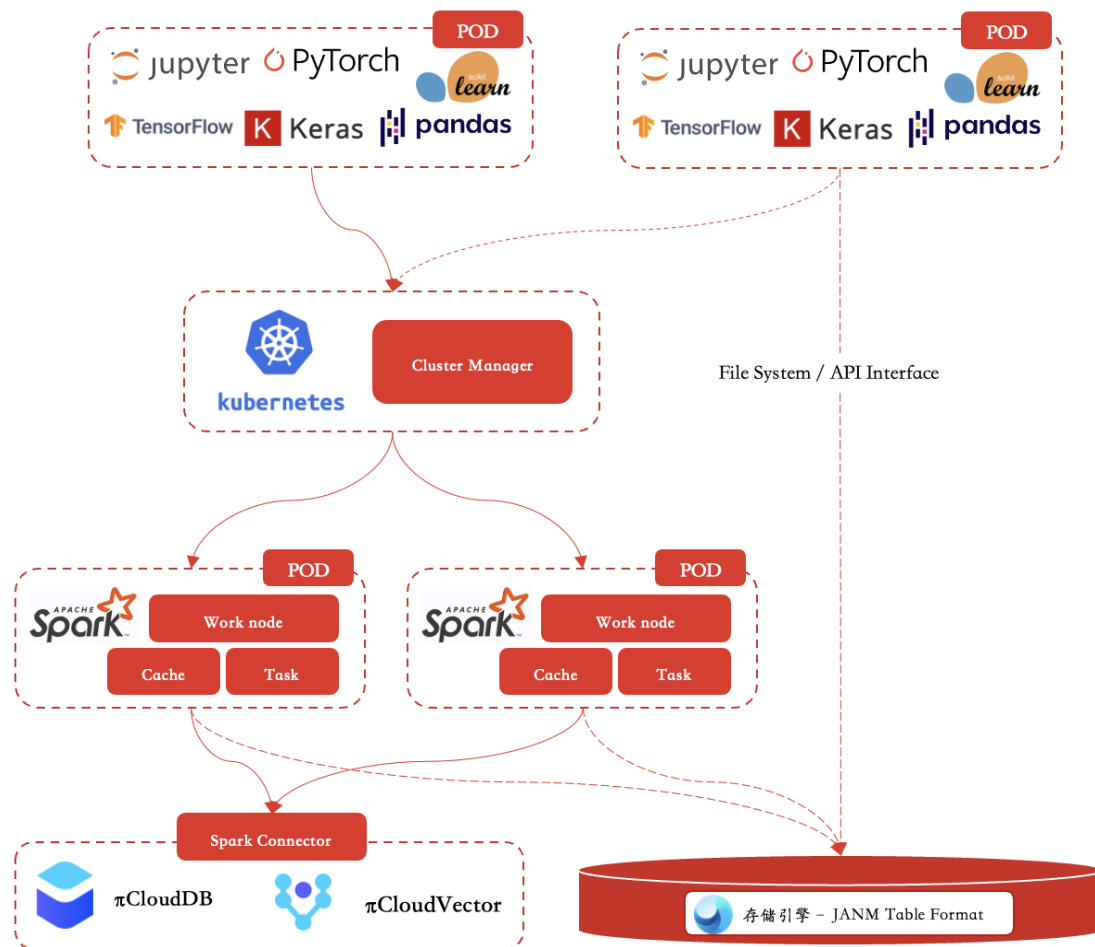
数据检索:

- 支持主流的ANN算法
- 支持乘积量化, 可牺牲部分精度加速搜索
- SIMD/GPU加速
- 支持全文检索和BM25算法评分
- 支持标量和向量数据融合查询
- 支持全文检索和向量检索混合查询

内置扩展:

- 文本chunk算法和Embedding算法
- Rerank算法
- API接口
- 兼容主流的LLM应用开发框架, 如LangChain、FinGPT等

PieCloudML大模型机器学习引擎



开发平台：

- 集成Jupyter工具
- 交互式建模
- 支持Python、R等语言

...

机器学习框架：

- 深度学习库：TensorFlow/PyTorch/Keras等
- 通用机器学习库：Scikit-Learn
- 数据分析工具包：Pandas
- 集成Spark计算引擎

...

数据访问接口：

- 提供Spark Connector连接器
- JDBC/ODBC等标准SQL Driver
- JANM提供Posix/S3 API等接口

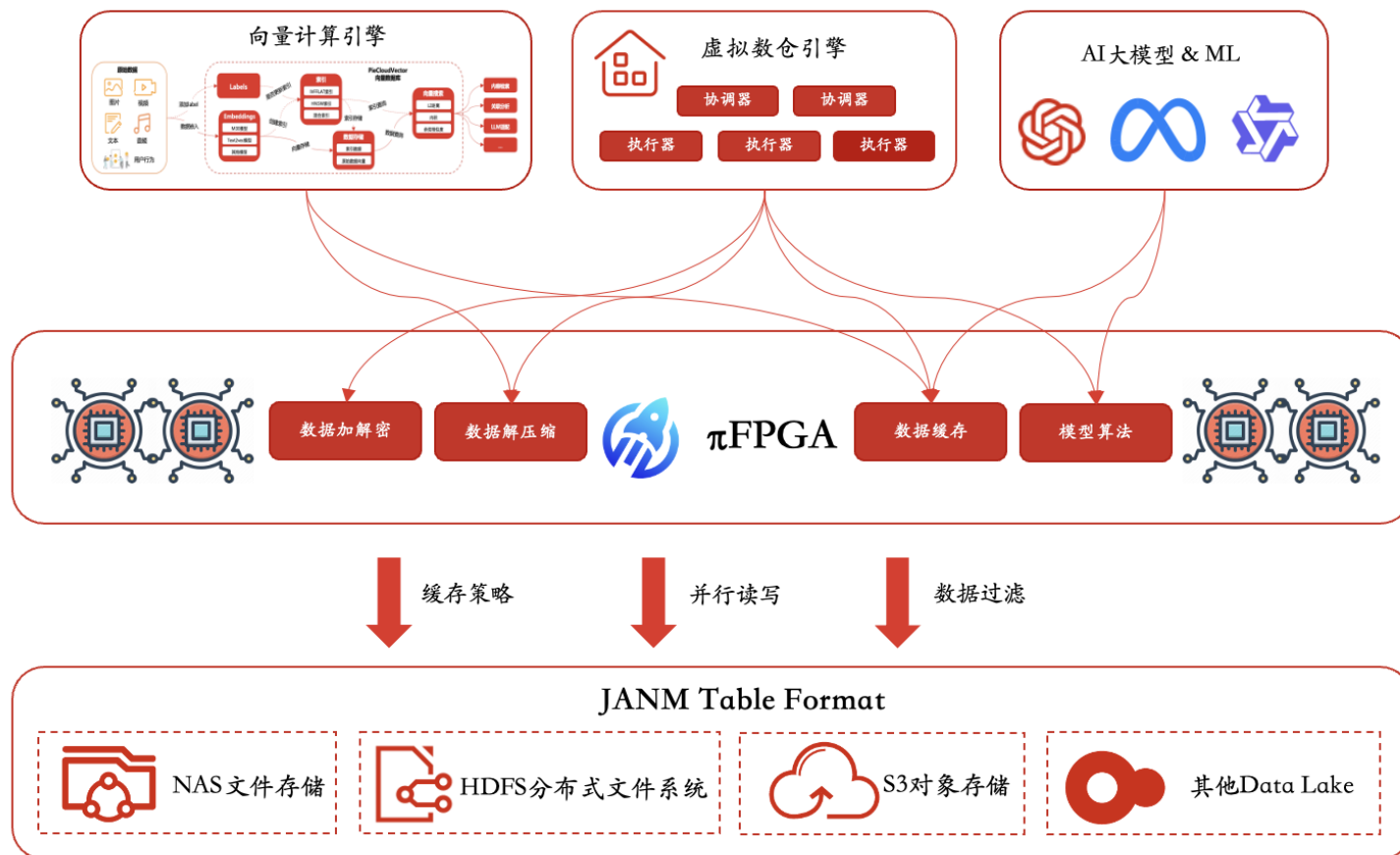
...

LLM结合：

- Ray分布式计算任务调度编排
- 模型微调（P-tuning V2/LoRA/Freeze/SFT等）

...

PieFPGA硬件加速引擎



存储加速:

- 加密/解密
- 压缩/解压缩
- 智能缓存

数据库加速:

- 数据排序
- 数据过滤
- 聚合函数

...

特定领域加速:

- 数据挖掘
- 图像信号处理
- 计算摄影学
- 机器学习
- AI大模型
- 数据预处理
- 数据后处理
- 地理空间

...

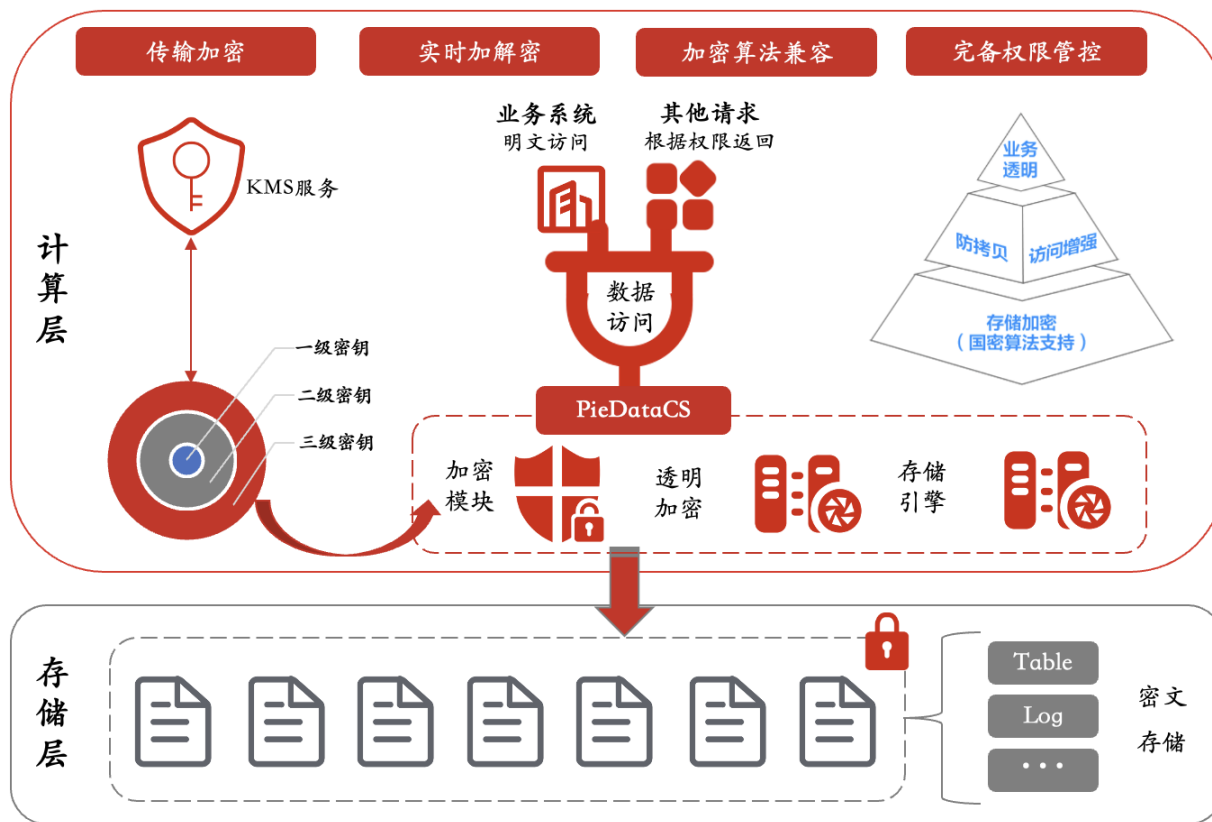
数据安全设计

国家法律与全国性标准

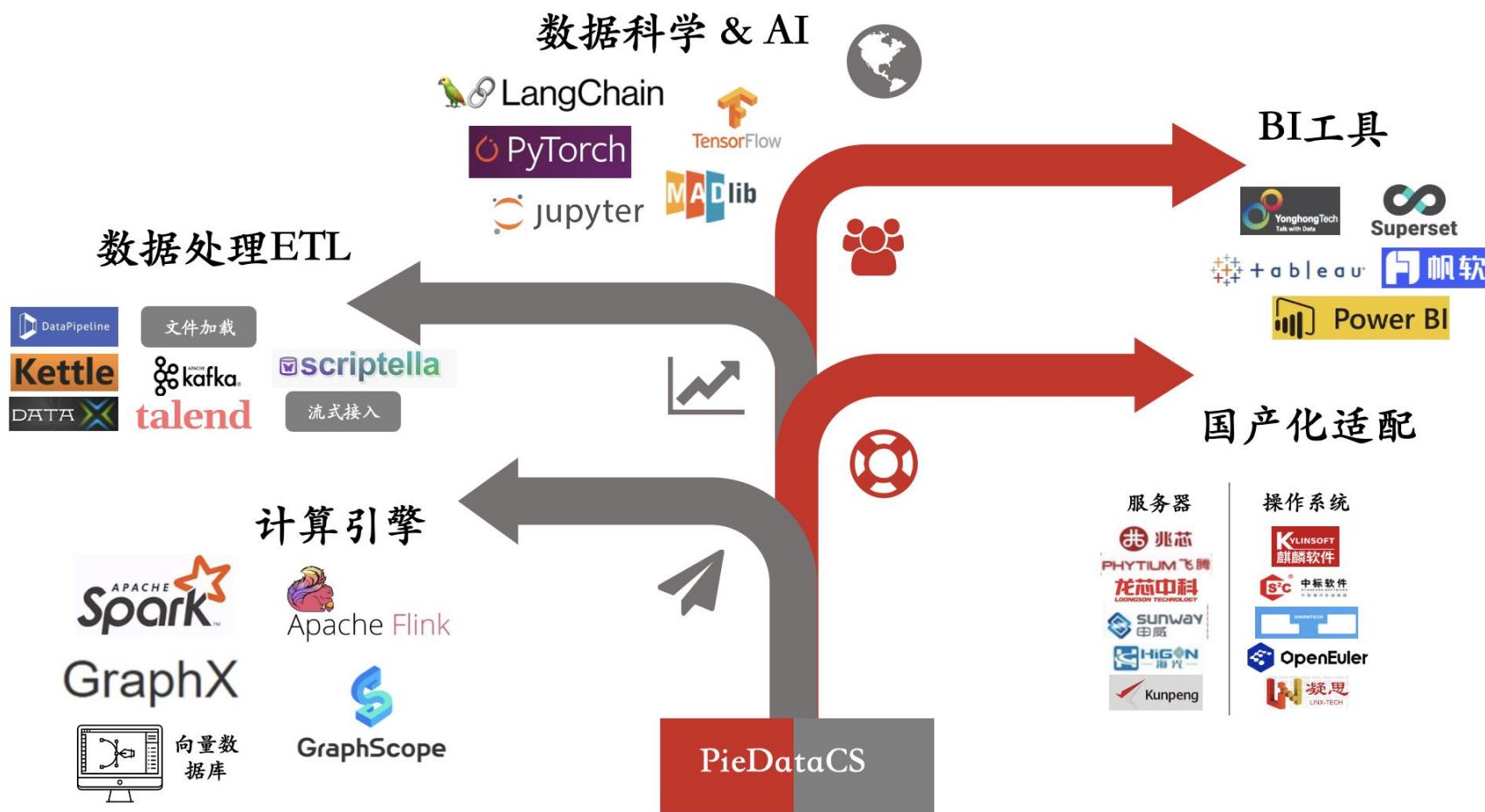


- 2019年：中国网络安全等级保护2.0（等保2.0）起步年。同年《新密码法》和《个人信息保护法（专家建议稿）》等发布。
- 《网络安全法》第四十二条：网络运营者不得泄露、篡改、毁损其收集的个人信息；未经被收集者同意，不得向他人提供个人信息。但是，**经过处理无法识别特定个人且不能复原的除外**。
- 《中国银行业“十二五”信息科技发展规则监管指导意见》中提出，完善敏感信息存储与传输等高风险环节的控制措施，对数据、文档的访问应建立严格的审批机制，对用于测试的生产数据要进行脱敏处理，严格防止敏感信息泄露。

满足等保合规要求



完善的产品生态



常见应用场景

PieDataCS应用场景



某企业知识库项目实践

场景应用层

知识库问答

智能客服

智能办公助手

智能研发助手

.....

服务编排层

RAG与多路召回

Agent流程画布

模型运营

取数与数据分析

图谱构建

模型算法层

多模态大模型

大模型微调

机器学习模型(OCR等)

Rerank模型

Embedding模型

计算引擎层

SQL计算引擎

向量计算引擎

全文检索

图计算引擎

机器学习引擎

基础设施层

GPU算力资源

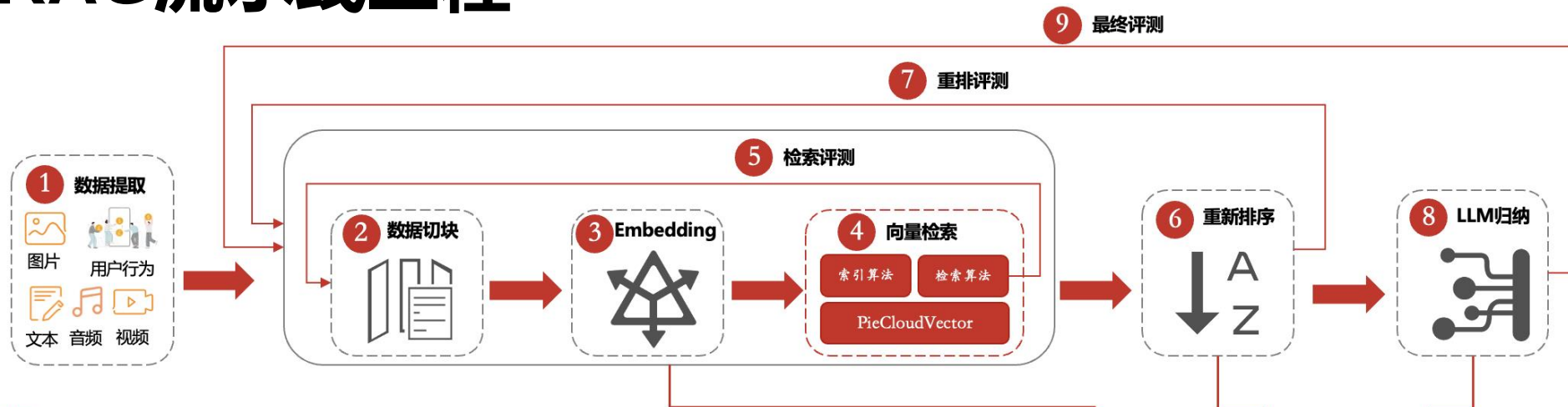
原始数据

湖仓一体架构

多模态数据

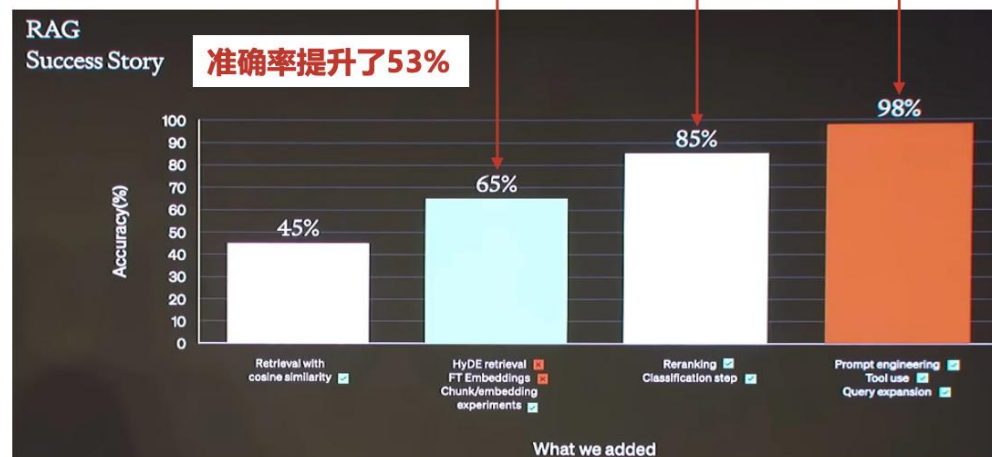
向量数据

传统RAG流水线工程



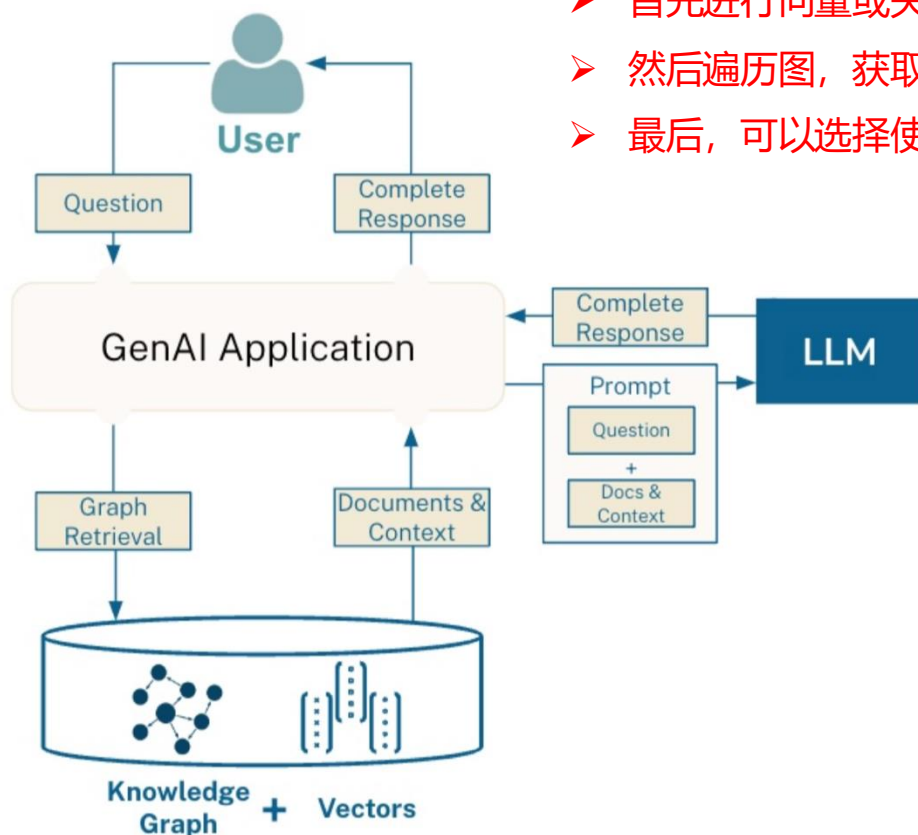
说明:

- 根据Embedding模型匹配最佳的切块大小, 切块大小会影响响应时间、可信度和相关度, 块越大, 响应时间越长, 可信度和相关度会提升;
- 不要微调Embedding模型, 成本非常高, 收效也不会超过10%, 推荐评测RAG Embedding模型, 当前免费可使用智源BAAI的bge-large-zh模型;
- RAG 检索评测指标主要是命中率、召回率和平均倒数排名MRR, 可参考Llamaindex的评测方法 (<https://blog.llamaindex.ai/boosting-rag-picking-the-best-embedding-reranker-models-42d079022e83>);
- 重排评测指标主要是HitRate (结果命中率) 和MRR (平均倒数排名), 推荐的评测模型有bge-reranker-base、bge-reranker-large、CohereRerank;
- RAG生成内容主要通过优化提示词 (prompt)、query expansion、tool use三种方法提升准确率, 指标包括总体答案相关性、答案溯源性、可信度和检索到的上下文相关性等, 评测RAG推荐采用RAGAS框架 (<https://docs.ragas.io/>)、TruLens和 LangSmith。



GraphRAG升级优化

- 首先进行向量或关键词搜索，以找到一组初始节点。
- 然后遍历图，获取这些节点相关的信息。这可以通过图数据库中的查询来实现，比如使用图遍历算法。
- 最后，可以选择使用基于图的排名算法(如 PageRank)对文档进行重新排名。

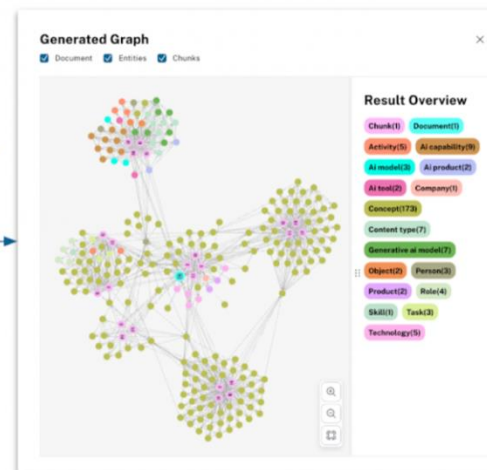


Automatically Build a KG for GenAI

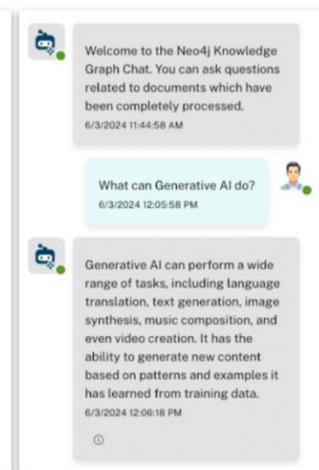
3 Simple Steps

- 1 **Connect to Neo4j**
- 2 **Upload Files** (pdf, YouTube, cloud storage, wikipedia)
- 3 **Generate Graph**

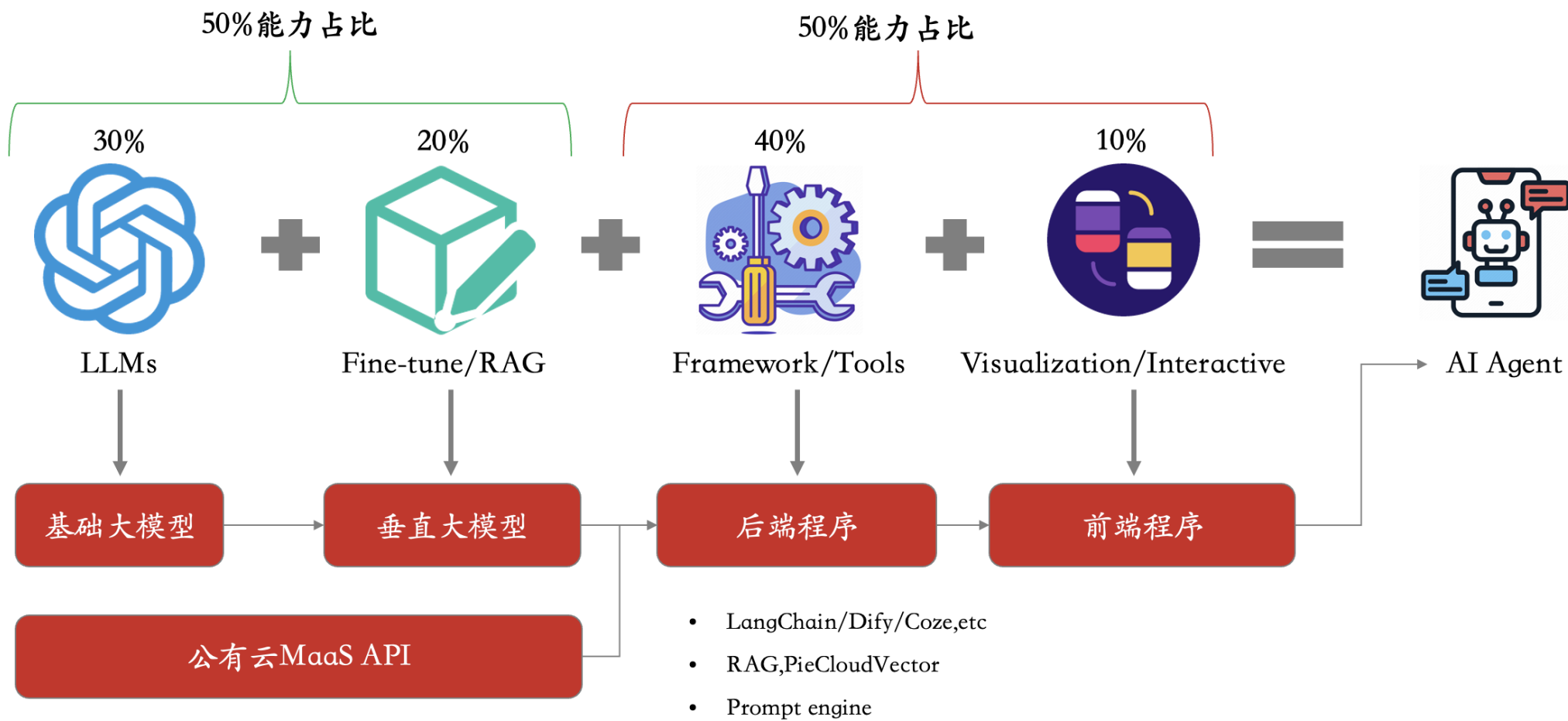
View & Explore Your Graph



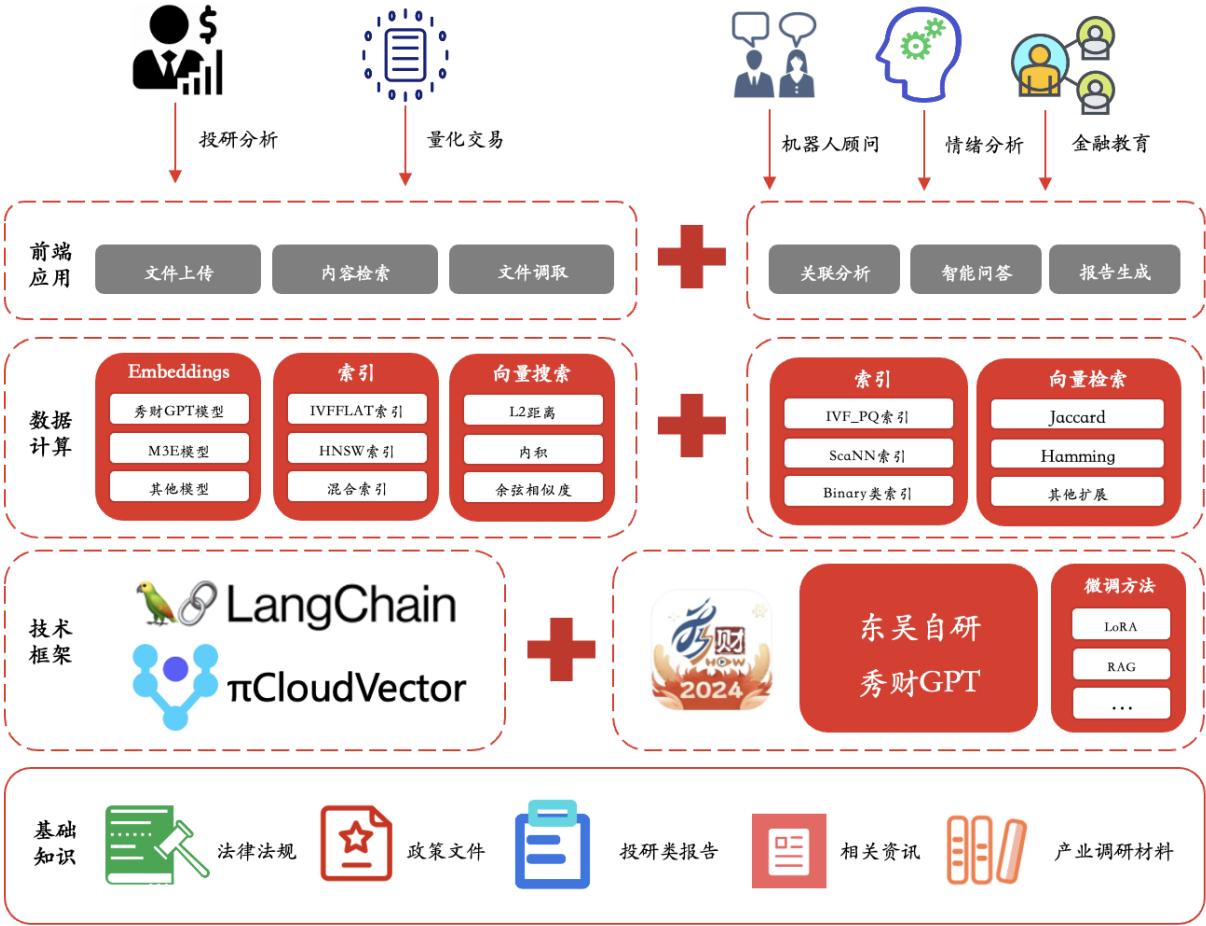
Power GraphRAG



AI Agent的能力占比分布



某证券客户基于AI大模型的应用实践



投研分析：

针对日常工作中的法律法规、政策文件以及研报报告进行快速检索，形成对应的分析报告，为客户提供投资相关的数据支撑。

量化交易：

基于 GPT 的大量金融数据训练，可以发现事件的情绪对资产的影响模式，可将这种模式用到量化交易策略中，由 GPT 实时产生交易信号以自动执行交易。这种数据驱动的量化交易策略可以更快地对市场变化进行响应。

机器人顾问：

根据个人需求和偏好实时提供金融建议，提高了获取建议的便利程度，降低了获得服务的成本。GPT 可以学习大量历史案例和研究报告，在此基础上，它可以总结出投资策略与建议。用户可以直接使用这些建议，或根据实际情况进行适当调整，这可以极大提高工作效率。

金融市场情绪分析：

对投资相关的言论和情绪表达等数据进行深度分析，获取市场情绪的指标，帮助投资者更好的把握市场走势，制定合理的交易策略，避免情绪犯错。

金融教育：

形成基础金融知识和产品知识库，针对用户和内部员工传授相关的投资策略和产品功能，用户可以与GPT交互，询问投资工具、风险和回报等具体问题，让用户对投资原理有更深入的了解。同时也可以让员工更快速的掌握更复杂的金融产品和工作技能。

业务运营报告：

结合内容运营的需求，把行情类的数据进行智能化和个性化创作，定时为客户推送实时行情、财经类的资讯以及精准的市场数据和分析，提升内容创作效率。



关注 OpenPie 公众号
获取更多干货视频



加入技术交流群
与社区小伙伴碰撞思想

PCC POSTGRESCONF
CN 2024

