西南证券
SOUTHWEST SECURITIES

人工智能专题研究

向量数据库——AI时代的技术基座

西南证券研究发展中心
通信研究团队
2023年6月

核心要点

- **受大模型热潮催化，向量数据库方兴未艾。**NVIDIA CEO 黄仁勋在3月的NVIDIA GTC Keynote 中，首次提及向量数据库，并强调其在构建专有大型语言模型的组织中的重要性。大模型作为新一代的 AI 处理器，提供了数据处理能力；而向量数据库提供了存储能力，成为大模型时代的重要基座。向量数据库是一种专门用于存储和查询向量数据的数据库系统，与传统数据库相比，向量数据库使用向量化计算，能够高速地处理大规模的复杂数据；并可以处理高维数据，例如图像、音频和视频等，解决传统关系型数据库中的痛点；同时，向量数据库支持复杂的查询操作，也可以轻松地扩展到多个节点，以处理更大规模的数据。
- **百亿蓝海市场蓄势待发，向量数据库空间广阔。**据 Statista 数据，2021 年全球数据库市场规模为 800 亿美元，同比增长约20.3%。假设增速保持20%，预计到2025年，全球数据库市场规模将达到1658.9 亿美元。据中国信通院测算，2020年中国数据库市场规模约241亿元；预计到2025年，中国数据库市场规模将达688亿元，复合增长率为23.4%。随着AI应用场景加速落地，我们预计2025年向量数据库渗透率约为30%，则全球向量数据库市场规模约为99.5亿美元，中国向量数据库市场规模约为82.56亿元。
- **海外需求逐步爆发，新兴赛道群雄并起。**目前向量数据库的赛道仍处于发展初期，随着大模型日趋成熟，越来越多玩家瞄准向量数据库的机会并选择加入赛道，呈现百花齐放的竞争格局。向量数据库的头部企业包括Zilliz、Pinecone等，目前的主要的客户还是互联网厂商随着大模型应用的不断拓宽，预计向量数据库的公司将受到更多投资者青睐，迎来投资井喷期。 Zilliz目前已与Nvidia、IBM、Mircosoft等公司展开合作，在一级市场获得1.13亿美元投资；Pinecone先后上架Google云和AWS，逐步打开市场，在一级市场获得1.38亿美元投资。
- **风险提示**：AI技术更新迭代缓慢、专业领域落地效果不及预期、市场开拓不及预期等风险。

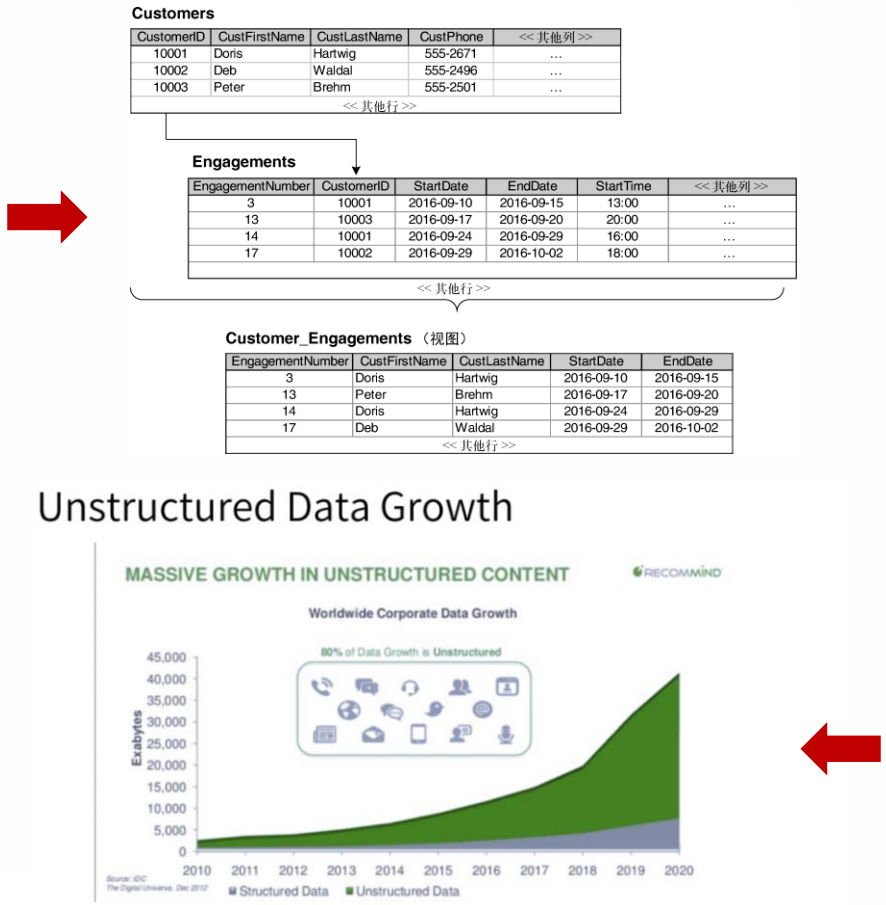
目 录

1.1 数据库分类

关系型数据库（SQL）vs. 非关系型数据库（NoSQL）

关系型数据库（SQL）

- **定义**：**依据“一对一、一对多、多对多”的关系模型创建数据库**，并将数据**以二维表格的形式储存**，各个表之间建立关系，通过这些关联的表格间分类、合并、连接或选取等运算来实现数据的管理。
- **发展情况**：1960s开始在航空领域发挥作用；因为其良好的一致性以及通用的关系型数据模型接口，使用范围广泛。
- **常见类型**：MySQL、Oracle、PostgreSQL等。
- **优点**：数据安全（磁盘）、数据一致性、二维表结构直观，易理解、使用SQL语句操作非常方便，可用于比较复杂的查询
- **缺点**：读写性能较差、不擅长处理较复杂的关系



图：关系型数据库和非关系型数据库规模对比情况

非关系型数据库（NoSQL）

- **起源**：2000年左右，互联网应用兴起，需要支持大规模的并发用户，并保持永远在线。一方面，**关系型数据库无法支持如此大规模数据和访问量**，升级CPU、内存和硬盘可以提高性能，但呈现明显的收益递减效应。另一方面，**数据库在机器间的迁移非常复杂**，需要较长的停机时间。**NoSQL因此应运而生，有效补充了SQL的适用范围，** NoSQL在Web应用领域提供了高可用性和可扩展性。
- **特点**：没有固定的表结构、**数据之间不存在表与表之间的关系、数据之间可以是独立的、NoSQL可用于分布式系统上。**
- **类型**：数据类型多样，针对不同的数据类型，出现了不同的 NoSQL，如**向量数据库**。

非关系型数据库是关系型数据库的有效补充

信息来源：《数据库分类及简介》，《数据库有哪些类型？如何根据应用场景选择？终于有人讲明白了》，IDC The Digital Universe，西南证券整理

1.1.1 数据库的分类——非关系型数据库

非关系型数据库按存储方式分为向量数据库、图形数据库、文档存储数据库、宽列数据库、键值存储数据库等，能够**实现非结构化或半结构化数据的处理和存储。**

	向量数据库	图形数据库	文档存储数据库
特点	将数据以向量形式存储，可实现向量数据的相似度搜索、聚类、降维等操作。	将数据以图的形式存储，以点、边为基础存储单元，每个节点代表一个实体，每条边代表两个实体之间的关系。	将数据以文档的形式存储，每个文档包含成对的字段和值。
优势	易处理高维度、高相似度、高并发的数据； 易与机器学习模型结合并提供智能化的服务。	易体现复杂的实体关系；支持高效的图遍历和分析。	非常灵活，可在文档中修改数据结构； 适用于处理半结构化或多变化的数据； 具有较高的性能，可快速传输、处理海量数据。
不足	技术成熟度较低，产品和相关应用较少	不适用于处理关系简单或无关系的数据；复杂性高，支持的数据规模有限。	缺乏严格的数据约束，需要小心谨慎地管理数据，避免数据出现质量问题。 通常不支持多文档操作，难以处理关联数据。

信息来源：《文档数据库介绍》，《揭秘文档数据库的优势与劣势》，《NoSQL 是什么？NoSQL 数据库详解》，西南证券整理

1.1.2 向量数据库的概述和原理

向量数据是什么？

- **“向量数据”**：向量数据是由多个数值组成的序列，可以表示一个数据量的大小和方向。通过Embedding技术，图像、声音、文本都可以被表达为一个高维的向量，比如一张图片可以转换为一个由像素值构成的向量。

向量数据库是什么？

- 向量数据库是一种专门用于**存储和查询向量数据**的数据库系统。
- 向量数据库支持对向量数据进行各种操作，例如：

向量检索：根据给定的向量，找出数据库中与之最相似的向量，例如在图像向量数据库中，用户输入一张图片进行搜索时，先将这张图片转换为一个向量，通过向量之间的近似检索，找到与输入图片最相似的图片。
向量聚类：根据给定的相似度度量，将数据库中的向量分类，例如根据图片的内容或风格，将图片分成不同的主题。
向量降维：根据给定的目标维度，将数据库中的高维向量转换成低维向量，以便于可视化或压缩存储。
向量计算：根据给定的算法或模型，对数据库中的向量进行计算或分析，例如根据神经网络模型，对图片进行分类或标注。

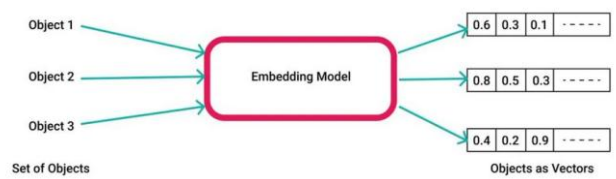
向量数据库有什么特点？

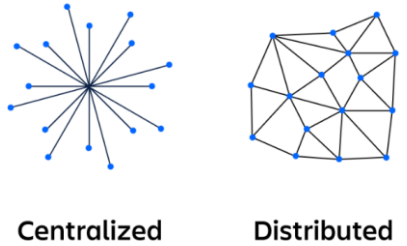
- **高维**：向量数据通常有很多元素，维度很高
- **稀疏**：向量数据中很多元素的值可能为零或接近零。
- **异构**：向量数据中的元素可能有不同的类型或含义。
- **动态**：向量数据可能随着时间或环境变化而变化。

信息来源：《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，《向量数据库(Vector Database)简介》，西南证券整理

1.1.2 向量数据库的概述和原理

向量数据库的部分核心技术





Embedding 技术：
针对问题：文本、图像、音频等非结构数据存储问题。
解决方法：利用Embedding技术将高维度的数据（例如文字、图片、音频）映射到低维度空间，即把图片、声音和文字转化为向量来表示，将这些向量存储起来就构成向量数据库。实现**Embedding过程的方法包括神经网络、LSH（局部敏感哈希算法）等。**

向量索引技术：
针对问题：向量数据维度很高，直接进行全量扫描或者基于树结构的索引会导致效率低下或者内存爆炸。
解决方法：采用近似搜索算法来加速向量的检索，通常利用向量之间的距离或者相似度来检索出与查询向量相近的K个向量，距离度量包括欧式距离、余弦、内积、海明距离，向量索引技术包括 k-d tree（k-dimensional tree），PQ（乘积量化），HNSW（可导航小世界网络）等。

分布式系统架构：
针对问题：向量数据规模庞大，单机无法满足存储、计算需求。
解决方法：使用分布式系统。分布式系统是计算机程序的集合，这些程序**利用多个节点的计算资源来实现共同的目标，**节点通常代表独立的物理硬件设备，但也可代表单独的软件进程或其他递归封装的系统。

硬件加速技术：
针对问题：向量数据计算密集，单纯依靠CPU的计算能力难以满足实时性和并发性的要求。
解决方法：利用专用硬件来加速向量运算，这些硬件包括GPU，FPGA，AI芯片等，用于提供更高的浮点运算能力和并行处理能力。

信息来源：《什么是分布式系统？分布式系统和微服务架构的概述》，《向量数据库介绍》，《语义索引（向量检索）的几类经典方法》，西南证券整理

1.1.2 向量数据库的概述和原理

Embedding的步骤

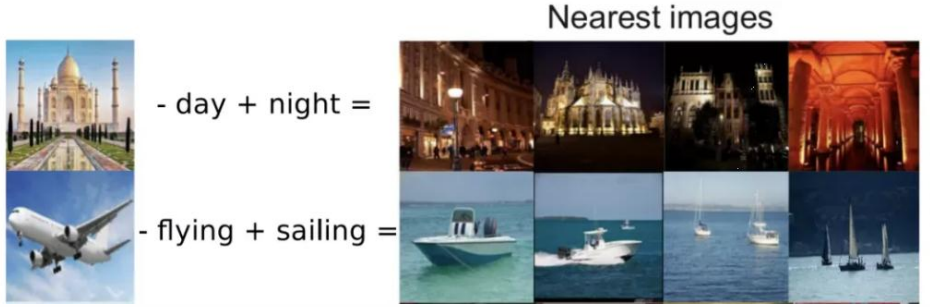
① **特征提取：将图片/音频转换成能够反映其内容或者属性的特征。**可以使用SIFT（Scale-invariant feature transform尺度不变特征转换），SURF（Speeded Up Robust Features加速稳健特征），HOG（Histogram of Oriented Gradients方向梯度直方图特征）等算法提取图片的边缘、角点、纹理等特征；可以使用MFCC（Mel频率倒谱系数）,LPC（线性预测分析）,PLP（感知线性预测）等方法提取音频的频谱、倒谱、能量等特征。

② **特征编码：将提取得到的特征进行编码，用一个固定长度的向量来表示。**方法包括BOW（词袋模型），VLAD (Aggregating local descriptors), Fisher Vector，GMM（高斯混合模型），HMM（隐马尔可夫模型），DNN（深度神经网路）等。

③ **特征压缩：将编码后的向量投影到低维度的子空间，进行向量压缩，使其能够用一个更低维度的向量来近似表示**，并保留尽可能多的信息。方法包括PCA（主成分分析算法），LDA（线性判别分析法），LSH（局部敏感哈希算法）。

Embedding的功能

- **语义搜索：**embedding向量是根据单词在上下文中的出现模式进行学习的，如果两个单词经常在上下文中一起出现，那么这两个单词**映射得到的向量在向量空间中就会有相似的位置。**用关键词进行语义搜索时，模型将关键词转化为embedding向量，然后在高维的向量空间里搜索这个embedding向量以及与其较为接近的向量，就可以得到与关键词相似的结果。
- **向量运算：**通过对embedding向量执行向量**加法和减法操作，可以推断出单词之间的语义关系。**例如，women的embedding向量可以通过下列运算得出：Embedding(woman)=Embedding(man)+[Embedding(queen)-Embedding(king)]
- **共享和迁移：**embedding向量可以**在多个自然语言处理任务中进行共享和迁移。**例如，在训练一个情感分析模型时，可以使用在句子分类任务中训练的嵌入向量，这些向量已经学习到了单词的语义和上下文信息，从而可以提高模型的准确性和泛化能力。



信息来源：《AI大模型领域的热门技术——Embedding入门介绍以及为什么Embedding在大语言模型中很重要》，《和Alex聊聊向量数据库与职业规划》，《深入解读向量数据库，Zilliz「三剑客」干货分享》，西南证券整理

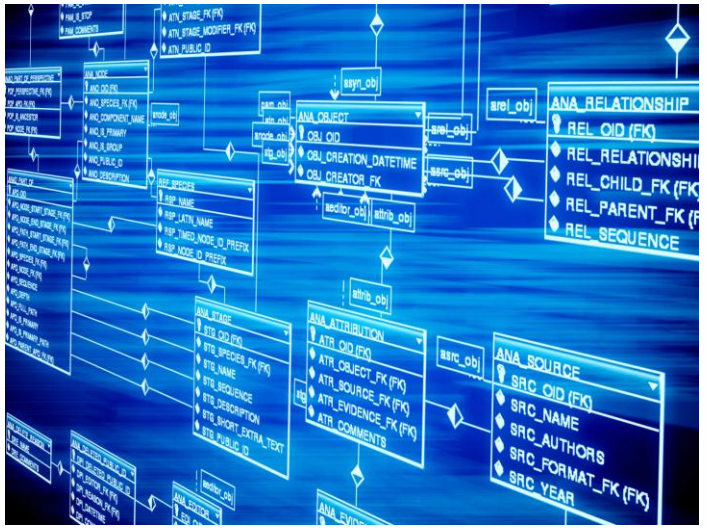
1.1.3 向量数据库的优势和不足

优点

- **处理大规模数据：**向量数据库的基本数据类型是向量，使用向量化计算能够比关系型数据库更快地处理大规模的复杂数据。
- **支持高维数据：**向量数据库可以处理关系型数据库中很难处理的高维数据，例如图像、音频和视频等。
- **支持复杂查询：**向量数据库支持复杂的查询操作，例如相似性搜索、聚类分析、降维等，并且速度快、准确度高，而关系型数据库中很难实现复杂操作。
- **易于扩展：**向量数据库可以利用分布式、云计算、边缘计算等技术轻松地扩展到多个节点，从而扩大数据处理规模，并提高向量数据的存储、管理和查询的稳定性。
- **高兼容性：**向量数据库支持多种类型和格式的向量数据，支持多种语言和平台的接口和工具。

不足

- **相对较新：**向量数据库是一种相对较新的技术，目前市场上的产品和应用还比较少。
- **学习成本高：**向量数据库需要掌握向量化计算的相关知识，学习成本较高。
- **适用场景局限：**向量数据库适用于处理大规模的复杂数据，而关系型数据库适用于处理简单数据。



信息来源：《一图文秒懂向量数据库、数据库发展过程》，《向量数据库(Vector Database)简介》，西南证券整理

1.1.4 向量数据库和传统数据库的区别

向量数据库与传统关系型数据库协同发展、相互补充。针对传统关系型数据库难以处理的大规模数据、低时延高并发检索、模糊匹配等领域，向量数据库通过数据的向量化来满足特定需求，尤其适用于人工智能领域。

	传统的关系型数据库	向量数据库
数据类型	数值、字符串、时间等传统数据类型	新的数据类型：向量数据 不存储原始数据
数据规模	小，1亿条数据对于关系型数据库来说规模很大	大，最少千亿数据是底线
数据组织方式	基于表格，按照行和列组织	基于向量，按照向量维度组织
查找方式	精确查找：点查/范围查 查询结果要么符合条件要么不符合条件	近似查找 查询结果是与输入条件最相似的，近似比较对计算能力要求非常高。
低时延，高并发	否	是
上层应用	较弱	对外提供统一的API，更适合大规模AI引用程序的部署和使用
下游应用场景	央企、国企。央企国企因工作内容要求容错率低，传统数据库能够提供更为准确的搜索结果。	互联网公司。向量数据库的结果正确率相对较低，成本低，互联网公司的场景容错率较高，能够包容这一缺陷。

信息来源：《向量数据库(Vector Database)简介》，《和Alex聊聊向量数据库与职业规划》，《向量数据库介绍》，西南证券整理

1.1.5 向量数据库存在的必要性

高效的检索

- **对于上千万或上亿规模数据的查找非常高效**，其他数据库难以提供大规模数据的快速查找。
- 能够检索特殊数据，如图片检索、人脸检索、人体检索、车辆检索等。

高效的分析

- **应用于平安城市：**把2个类似作案手法的案发现场周边的人做人像对比，找出同时出现在两个案发现场的人。
- 深圳平安城市项目到2018年底会部署20w摄像头，预计保留一年的人脸特征在千亿级别。

为上层应用提供坚实基础

- 数据库是一个通用的基础空间，基于数据库会有多种多样的的需求，可以**针对需求进行上层应用的设计**，技术研发也更有针对性。

信息来源：《Zilliz从开源数据库、AI开发到垂直应用的底层逻辑》，《向量数据库介绍》，西南证券整理

1.2 向量数据库的主要应用场景

➢ 向量数据库发展趋势

- **应用领域拓展**：目前主要应用于**图像搜索、音乐推荐、文本分类**等领域，未来可能运用于**语音识别、自然语言处理、智能推荐**等。
- **性能提升：**随着技术不断提升，向量数据库的性能将会进一步提升，会有**更快的查询速度、更高的并发处理能力**等。
- **安全性提高：**未来向量数据库会**加强数据加密，提高访问控制**，以提高数据库安全性。
- **云化趋势：**随着云计算技术的发展，向量数据库也将会趋向云化，**将向量数据库部署在云端，提供云服务。**

➢ 应用场景

- **检索：**传统的关键词搜索，搜索结果局限于输入的关键词，而向量数据库是基于文本生成embedding向量再进行检索，检索结果范围更广。
- **语义分析：**包括文字、图像、行为等多方面的语义分析。
- **以图搜图等**模糊数据匹配。
- **人工智能NLP场景**：包括搜索引擎和问答机器人。用向量数据库对问题进行相似性查找，可以得到相关结果。

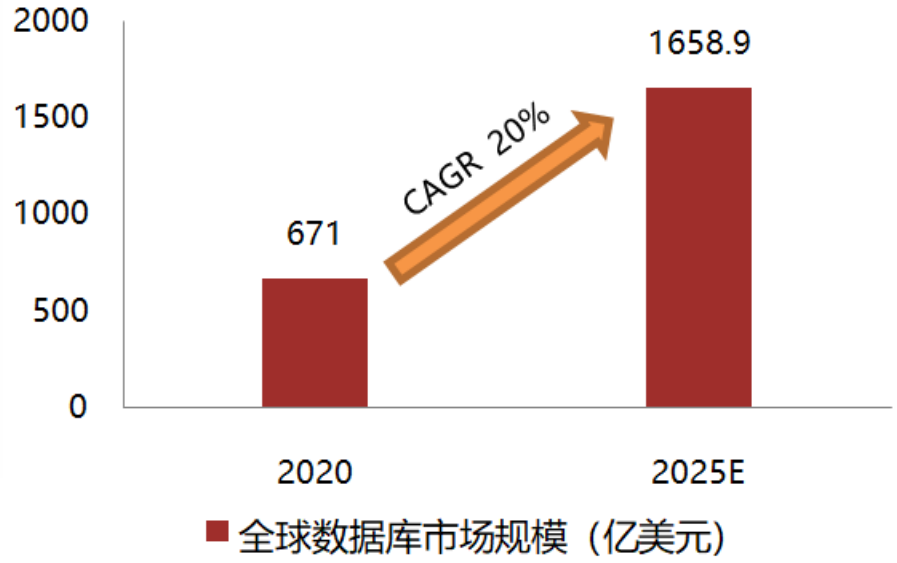
信息来源：《一图文秒懂向量数据库、数据库发展过程》，《Zilliz从开源数据库、AI开发到垂直应用的底层逻辑》，《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，《和Alex聊聊向量数据库与职业规划》，西南证券整理

目 录

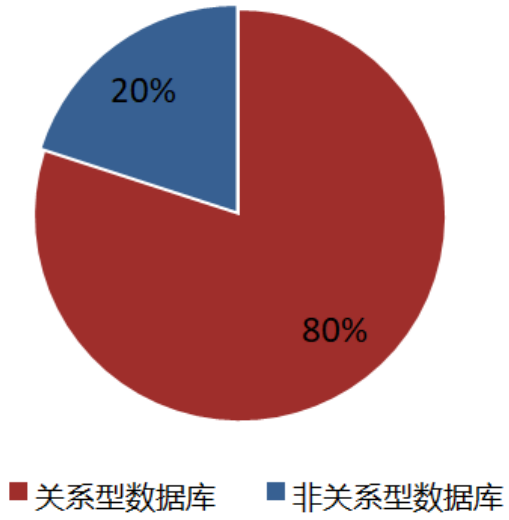
2.1 市场规模

全球数据库市场规模：根据 Statista 数据，2021 年全球数据库市场规模为800亿美元，同比增长约20.3%。假设增速保持20%，预计到2025年，全球数据库市场规模将达到1658.9 亿美元。
非关系型数据库市场份额：根据 Gartner 报告，关系型数据库在全球数据库总体市场中的占比约为80%，非关系型数据库在全球数据库总体市场中的占比大约为20%。
向量数据库市场规模：随着AI应用场景加速落地，我们预计2025年向量数据库渗透率约为30%，则全球向量数据库市场规模约为99.5亿美元。

2020-2025E全球数据库市场规模变化



2022年全球数据库市场份额

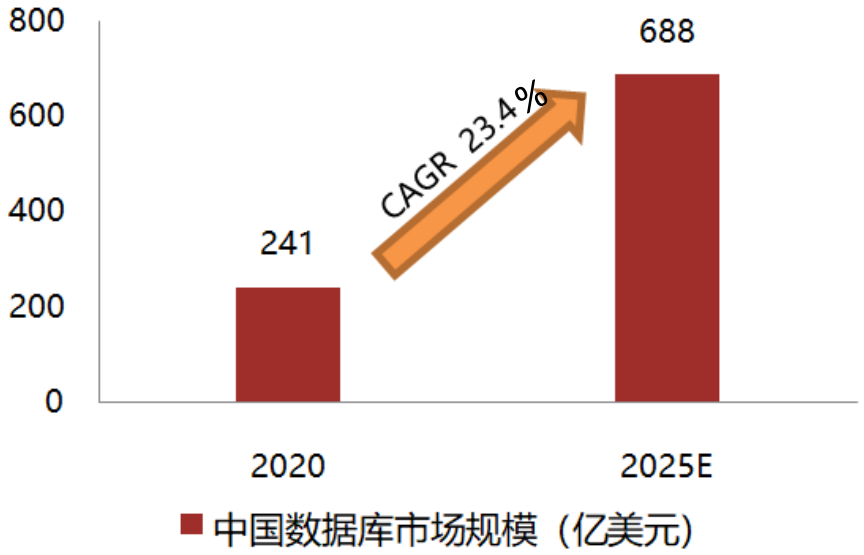


信息来源：Statista，Gartner，西南证券整理

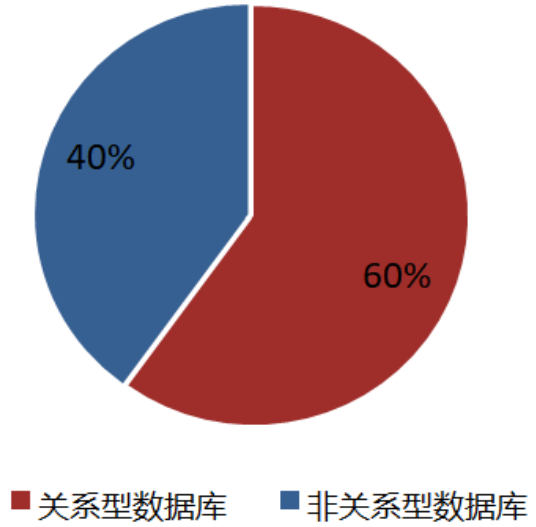
2.1 市场规模

中国数据库市场规模：据中国信通院测算，2020年中国数据库市场规模约241亿元。预计到2025年，中国数据库市场规模将达688亿元，年复合增长率为23.4%。
非关系型数据库市场份额：据IDC数据，关系型数据库在中国数据库总体市场中的占比超过60%，非关系型数据库在中国数据库总体市场中的占比约为40%。
向量数据库市场规模：随着AI应用场景加速落地，我们预计2025年向量数据库渗透率约为30%，则中国向量数据库市场规模约为82.56亿元。

2020-2025E中国数据库市场规模变化



2022年中国数据库市场份额



信息来源：中国信通院，IDC，西南证券整理

2.2 竞争格局

➢ 目前竞争格局

目前整个向量数据库的赛道仍处于培育阶段，受AI大模型热潮催化，向量数据库刚刚引起国内市场的关注，目前主要使用者是互联网巨头公司。

➢ 未来竞争格局

多元化竞争：最早涉及到向量数据库领域的是**Meta**的一个解决向量计算的计算库，后期逐渐出现例如**Zilliz**做开源向量数据库的公司。不同公司的向量数据库将实现不同的卖点，比如开源、托管、集群服务等。随着AI普及，向量计算的需求将大幅提升，越来越多玩家或将涌入赛道，呈现百家争鸣的状态。

向量数据库与传统数据库不会互相取代，而是会在不同的场景下发挥各自的优势。向量数据库的出现，也会促进传统数据库对向量数据类型的支持。

未来发展方向

向量运算	**未来传统关系型数据库也会逐渐往向量运算上提供相应的能力**，比如通过开发向量计算插件代码生成的方式去做向量计算。
形态并生	**未来的竞争格局会呈现两种形态并生的情况，即原生向量数据库和传统数据库均可支持向量计算插件并存的状态，且各有优劣。**原生向量数据库引擎主要面向大规模向量数据的存储和检索，对AI应用更具有针对性和亲和力。在传统数据库的基础上提供向量计算的能力也将是未来的一个切入点。

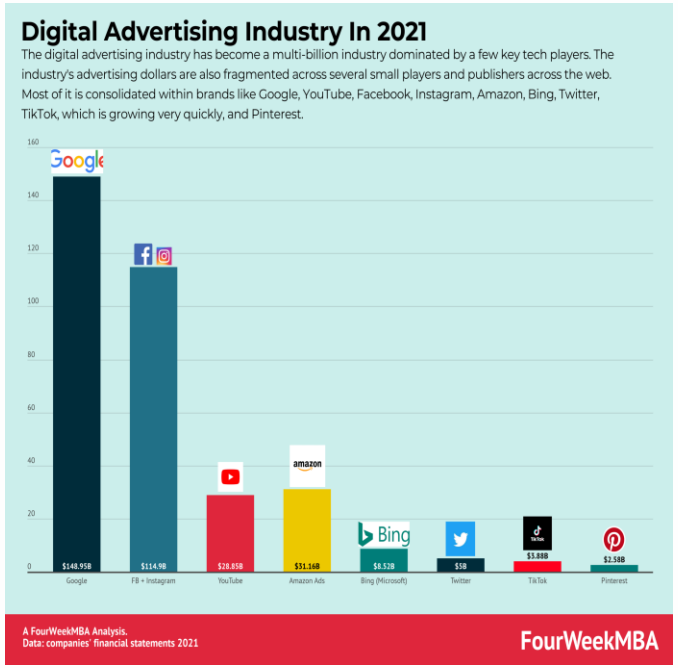
信息来源：《你真的需要一个（专门的）向量数据库么》，知乎，公司官网，西南证券整理

2.2 竞争格局

向量数据库与整个AI的生态联系十分紧密。随着大模型日趋完善，向量数据库受到了较多青睐，越多越多玩家选择加入这个赛道。**赛道目前处于群雄并起的阶段，**从融资、技术的角度上来讲，目前尚未有寡头角色出现。

向量数据库未来会出现双寡头模式

- **上一个技术时代云数据或者数据分析平台的最终格局基本上都是双寡头模式。**双寡头会占据市场的 60%以上的市场份额，后面市值在几十亿美金的云数据分析公司至少还有 20 家左右。市场格局基本上呈现出寡头和长尾分布的模式。
- Zilliz 创始人认为，**未来向量数据库领域可能也会呈现类似的格局，**即出现1到2家500-1000 亿美金的公司，它们可能会做通用的方案，解决通用类场景。同时在几十到百亿美金之间可能还会有 10 家左右，它们会专注在细分垂直领域，而从过去美国市场的情况来看，几乎都延续了这样的格局。
- **现在整个市场才刚开始，只能从前几个技术时代做简单的预测。**AI的产品形态和场景变化太快，目前主要还集中在文本大模型领域，创始人表示 Zilliz 很早就已经开始做多模态的大模型支撑，下一波的视频、图片以及生物医药等领域的大模型会很快到来，因此未来的向量数据库存储的记忆将不再限于文字，它会存储图片、视频甚至化学分子式等等。



信息来源：《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，《为什么各大VC最近都在投向量数据库》，西南证券整理

2.3 向量数据库中美市场对比

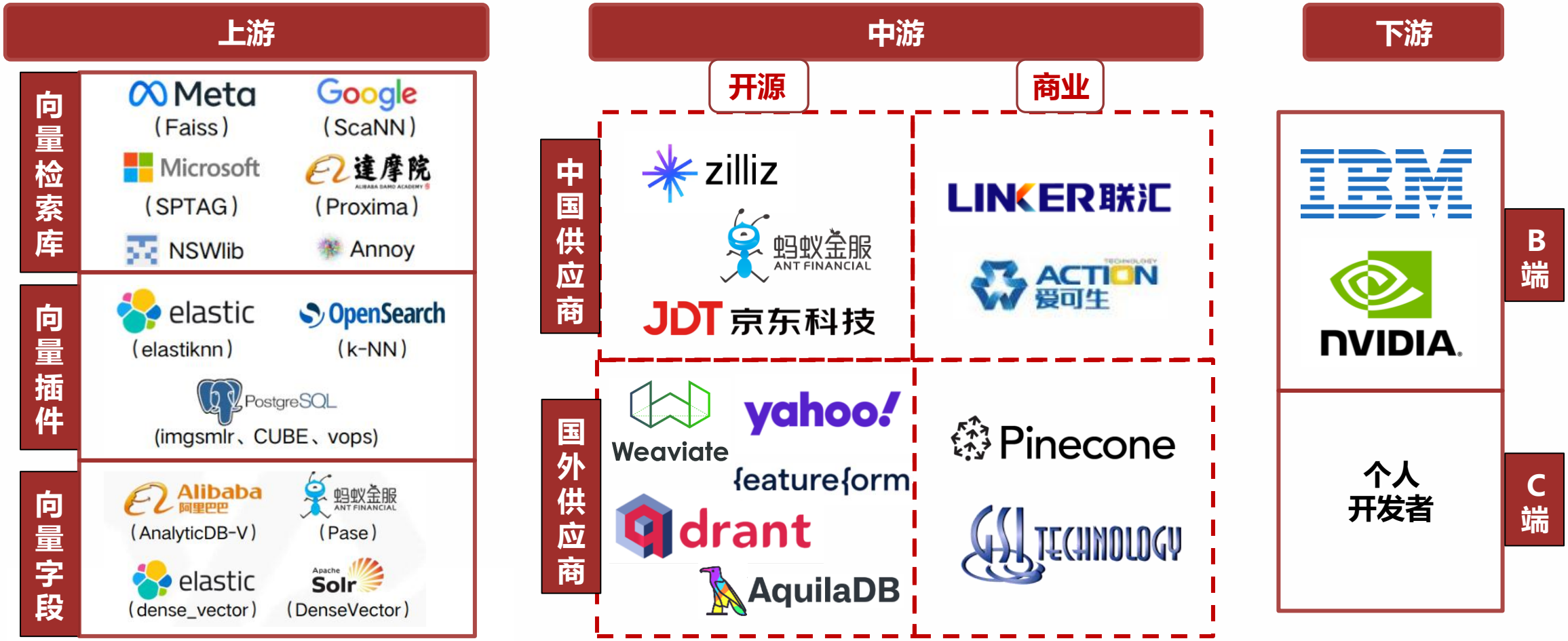
与美国市场相比，中国市场主要存在差距的原因：1）开源社区尚未成熟；2）用户端尚未成熟；3）需求滞后

美国市场	国内与美国市场相比出现的问题
➢ 开源环境更加成熟	➢ 开源社区尚未成熟，推行难度大
➢ 北美的用户端相对更成熟，用户与企业协同性更强，更容易商业化。	➢ **国内用户的付费意愿相对较低。** ➢ **国内用户（大B端和G端）定制化需求强，**产品标准化程度比较低，市场更看重渠道。
➢ 美国人工智能基础设施较早发展，叠加下游落地场景丰富；数据库作为中间层产品，需求较大	➢ **市场尚未成熟；**应用端发展先于中间层

数据来源：《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，西南证券整理

2.3 向量数据库中美市场对比

产业链上游包括向量检索库、向量插件、向量字段等数据供应商以实现检索功能；中游即向量数据库服务提供商；由于向量数据库又分为开源和商业，下游使用者可分为个人开发者及付费企业。

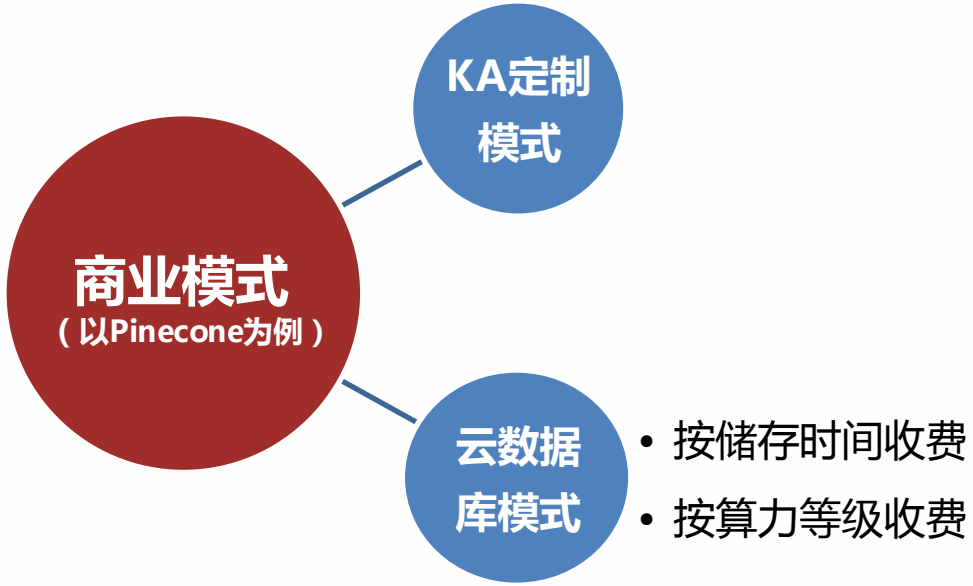


信息来源：《中国数据库行业分析报告》，公司官网，西南证券整理

2.4 向量数据库的商业模式

目前的商业模式：开源社区+定制化服务

目前市场仍处于摸索前期，商业模式分为KA定制和云数据库模式（按照存储和计算资源收费）。



未来商业模式的驱动因素

➢一、取决于未来应用端的开展

需求量不仅局限于市场对向量数据库技术或产品的需求，还依赖整个产业链的循环。如：大语言模型的崛起意味着市场对非结构化数据、对向量数据库的需求会增加。应用足够丰富，才会有可能产生对向量数据库的需求。现在仍处于偏早期阶段。

➢二、取决于数据源的数量

没有足够的数据源等价于没有足够多的算力，无法做出好的应用。即使AI大模型崛起，没有足够多的数据，向量数据库在大模型中也缺乏用武之地。

信息来源：the data exchange, gradientflow.com, 西南证券整理

目 录

3.1 Zilliz

Zilliz成立于2017年，是一家以美国为主要商业化市场的科技创业公司，并于中国上海、北京、深圳均设有机构。Zilliz拥有强大的专业团队，其主要相关产品为Milvus开源向量数据库。**Zilliz作为全球向量数据库领域的开拓者，以技术为核心，专注于在非结构化数据的分析中挖掘其价值，研发为人工智能服务的向量数据库系统，使更多的企业、组织、个人用更低的成本开发人工智能应用并从中获得便利。**目前Zilliz已与多家知名企业合作，其目标也是致力于成为全球领先的向量数据库领域企业，并希望为人工智能领域的发展助力。

公司发展历程

- 2017年：Zilliz成立；完成天使轮融资，金额高达数千万人民币
- 2018年：完成A轮融资，金额高达1000万美元
- 2019年：Zilliz推出开源向量数据库Milvus
- 2020年：完成B轮融资，金额高达4300万美元
- 2022年：完成B+轮融资，金额高达6000万美元

信息来源：公司官网，企查查，快出海，西南证券整理

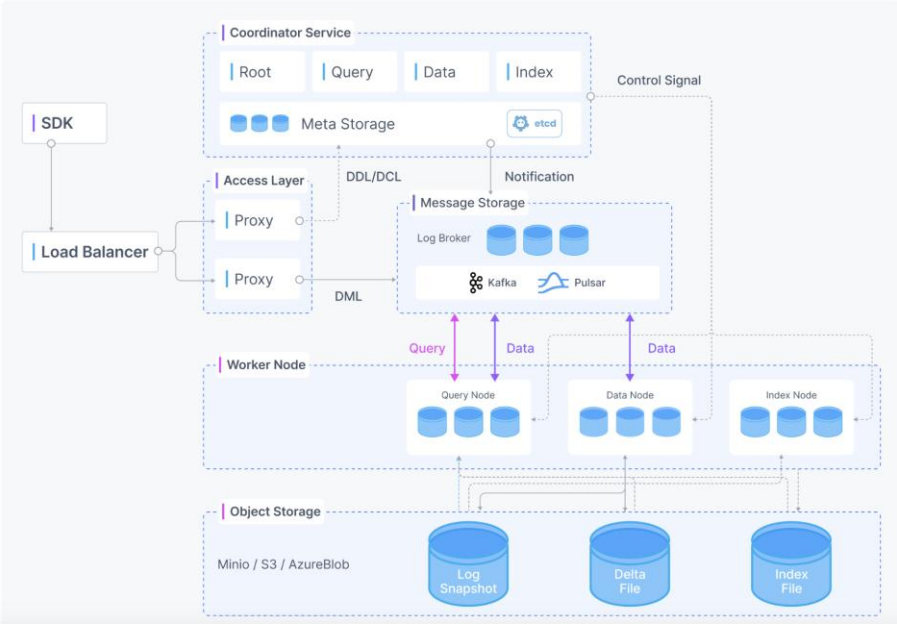
3.1.1 Zilliz核心产品——向量数据库Milvus

Milvus是一款开源的、分布式的、面向生产环境的向量数据库，是完全面向于人工智能的一款数据库。**Milvus可以存储、索引和管理由深度神经网络和其他机器学习模型生成的十亿以上的嵌入向量，核心功能是针对向量数据进行相似度搜索。**目前Milvus的发展健康平稳，其增长主要依赖于Github的自然流量，2.0版本也做到了高性能，高可用和高扩展性。

数据或向量的组织形式在Milvus里分成的collection（集合）和partition（分区）两个级别，每个collection和partition下面都可以进行向量的存储，同时也支持着对应的标量数据的存储。Collection可以被理解为传统数据库中的“表”，是Milvus中最基本的数据组织单位，负责管理和存储向量数据。Partition是collection中的子集，每个collection中会有多个partition，用于对数据进一步的分类存储。**Collection和Partition赋予了Milvus对大规模向量数据更灵活的管理能力，使用户在使用过程中更加的高效。**

Milvus的六大优势

高易用性	通过使用Milvus，**用户能够在一分钟内构建大规模相似性搜索服务，同时支持多语言。**
超高性能	Milvus硬件性能极高，**使检索速度实现了10倍的性能提升。**
高可用性	Milvus已被超过千家企业和用户实测和使用，**适用于各种不同的情况，并同时具备高弹性和高可靠性。**
高扩展性	Milvus的分布式和高吞吐量的特性使其适合于服务大规模向量数据。
云原生性	Milvus的云原生性使其将存储和计算分开，使用户灵活扩展。
功能丰富	**Milvus向量搜索支持多种数据类型。**



信息来源：《云原生向量数据库Milvus（一）-简述、系统架构及应用场景》，《和Alex聊聊向量数据库与职业规划》，公司官网，西南证券整理

3.1.1 Zilliz核心产品——向量数据库Milvus

Milvus系统架构

四个层次	主要功能	
接入层（Access Layer）	接入层是由一组无状态代理组成的，**并作为系统的前层和用户的终端**，对客户端的请求进行验证并减少返还结果。	
协调服务层（Coordinator service）	协调者服务将任务分配给工作节点，并作为系统的大脑发挥作用，**负责集群拓扑结构管理、负载平衡、时间戳生成、数据声明和数据管理。**	**根协调器**（root coordinator）：跟协调器负责处理数据定义语言和数据控制语言请求，如创建或删除集合和分区。
		数据协调器（data coordinator）：负责管理数据节点的拓扑结构并维护元数据。
		查询协调器（query coordinator）：负责管理拓扑结构和查询节点的负载平衡，以及从增长段到密封段的交接。
		索引协调器（index coordinator）：负责管理索引节点的拓扑结构，建立索引，并维护索引元数据。
执行节点层（Worker nodes）	**执行节点层遵循协调者服务层的指示，执行来自代理的数据操作语言命令，是系统的四肢。**由于存储和计算的分离，执行节点是无状态的，可以促进系统的扩展和灾难恢复。	**查询节点**（Query node）：检索增量的日志数据，并在向量和标量数据之间运行混合搜索。
		数据节点（Data node）：负责处理突变请求，并存储数据日志打包形成的日志快照。
		索引节点（Index node）：主要负责建立索引。
存储服务层（Storage）	**存储是系统的骨骼，负责保证数据的持久性。**	**元存储**（Meta storage）：主要负责存储元数据的快照。
		对象存储（Object storage）：主要负责存储日志的快照等文件。
		日志代理（Log broker）：主要负责保证流式数据的持久性。

信息来源：《云原生向量数据库Milvus（一）-简述、系统架构及应用场景》，公司官网，西南证券整理

3.1.2 Zilliz的商业模式

Milvus本身是一个完全开源的架构，开源也是商业化非常好的契机。运营开源社区的前提是要创造一个开源社区，创造一个开源社区的重点在于要挖掘出社区的具体价值，即开发的产品如何帮助到用户。**Zilliz的大逻辑是：随着AI逐渐成为未来的大趋势，建立一个为AI服务的数据库是非常有价值的，也能够解决很多AI公司的痛点。**

Zilliz的基本原则是所有基础功能都会开源，并同时投入商业化产品的开发，运用开源手段持续扩大社区并收集用户反馈，不断打磨产品，并为有需求的客户提供更进一步的付费服务。所有用户可以免费体验Zilliz的基础软件，用户若对软件有更高性能的需求，也可以购买其商业化版本。Zilliz目前已经拥有超过300家企业用户选择购买其付费版本，计算单位为每小时0.25美元，存储费用为每GB每月0.02美元。

Zilliz目前商业化还在起步阶段，确定未来方向仍为时尚早。**从长远的角度来看，如果公司想获得爆发式增长，可能会以销售为主发力点并同时进行其他的辅助销售和市场营销策略。对于Zilliz的产品来讲，开源已经是最好的落地手段。**Zilliz的设想是：**当公司拥有一定的网络效应后，盈利模式将以云托管为主，即用户直接在公有云平台上购买公司已经部署好的服务，达到进一步降低成本的目的。**

Zilliz的主要合作企业

ebay 小米 Shopee

IKEA Baidu 百度 LINE

Tencent 腾讯 Walmart Save money. Live better. COMPASS

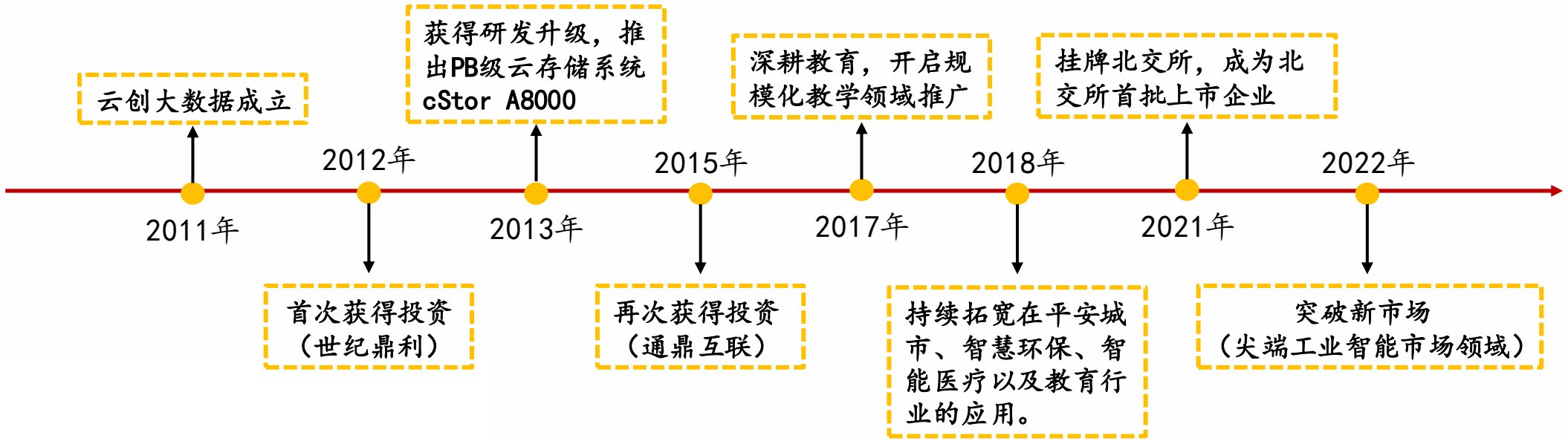
信息来源：公司官网，《Zilliz从开源数据库、AI开发到垂直应用的底层逻辑》，《和Zilliz聊聊开源商业化、向量数据库和软件全球化》，《和Alex聊聊向量数据库与职业规划》，《“零收入”依然淡定，一家基础软件公司的冒险与底气 | 甲子光年》，西南证券整理

3.2 云创大数据

南京云创大数据科技股份有限公司成立于2011年3月，作为一家高新技术企业，其主要业务为提供大数据存储产品、大数据处理产品及解决方案。公司具有完整的大数据价值链业务体系，以大数据的存储和处理为出发点，并结合云计算、人工智能等技术，帮助客户更便捷快速的处理和分析海量数据。云创大数据的产品已成功落地于各个领域，包括：公共安全、环境监控、学科教育等，致力于为客户提供更加全面的解决方案。随着大数据产业的不断发展，云创大数据不断创新升级，持续寻找行业的新机遇。

云创大数据
www.cStor.cn

公司发展历程



信息来源：公司订阅号，《北交所企业研究：云创数据（835305）——大数据存储与处理服务商》，西南证券整理

3.2 云创大数据

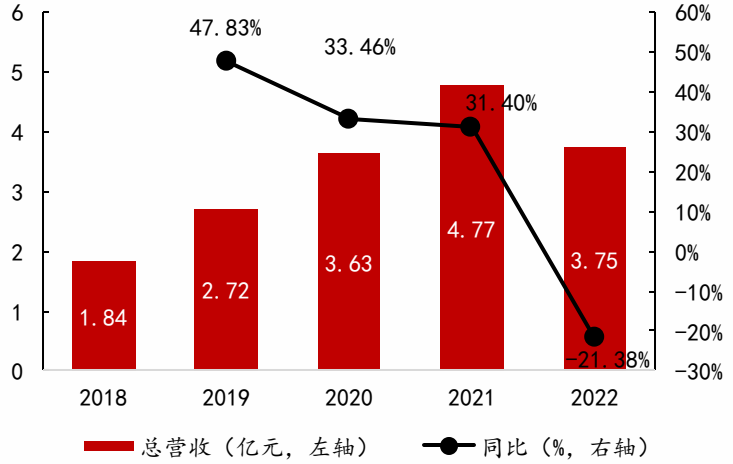
公司经营状况简介

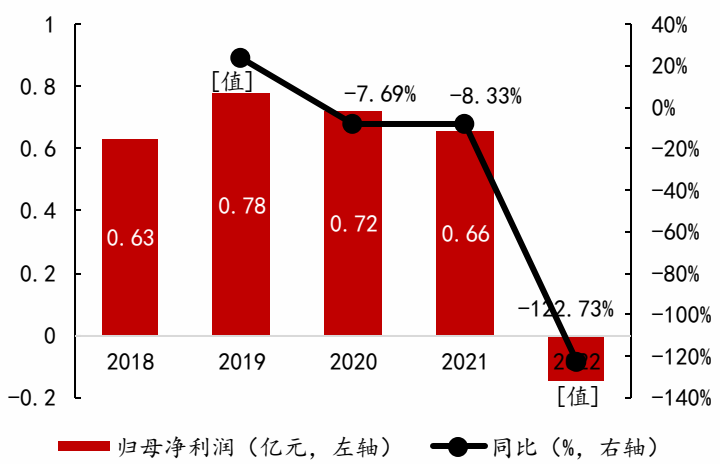
公司主要业务收入来自于华东地区，主营业务为大数据智能处理和大数据存储产品。自2018至2021年总营收呈稳定增长趋势，同比增长速度逐年变缓。2022年总营收相比2021年下降显著，跌至3.75亿元，归母净利润由盈转亏。公司毛利率也呈逐年下降趋势，2022年跌至-4.53%。**公司盈利能力下降主要受疫情及国际形势影响，应收帐款的逐年增加也对公司现金流造成重压。同时，公司极其重视产品研发，公司的研发费用2021年同比增加60.73%，2022年同比增加31.03%。**

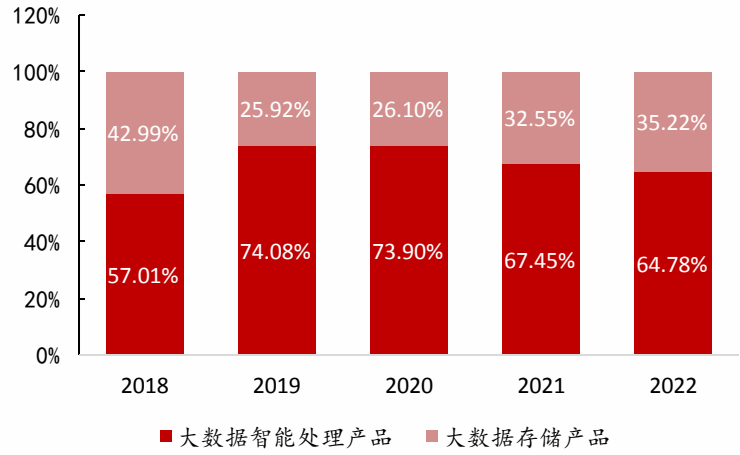
公司区域经营情况

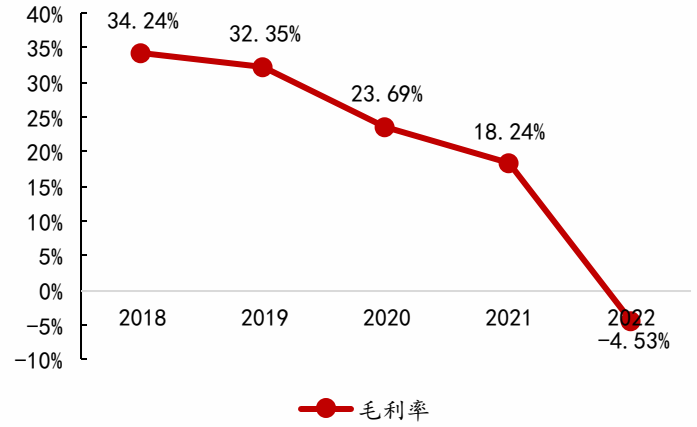
单位：亿元	2018	2019	2020	2021	2022
华东业务收入	0.83	1.37	1.91	3.18	2.71
西北业务收入	0.49	0.89	0.85	0.46	0.53
华南业务收入	0.29	0.22	0.39	0.42	0.25
其他	0.25	0.26	0.48	0.71	0.25

公司经营情况









数据来源：公司定期报告，wind，西南证券整理

3.2.1 云创大数据主要产品——cVector向量计算一体机

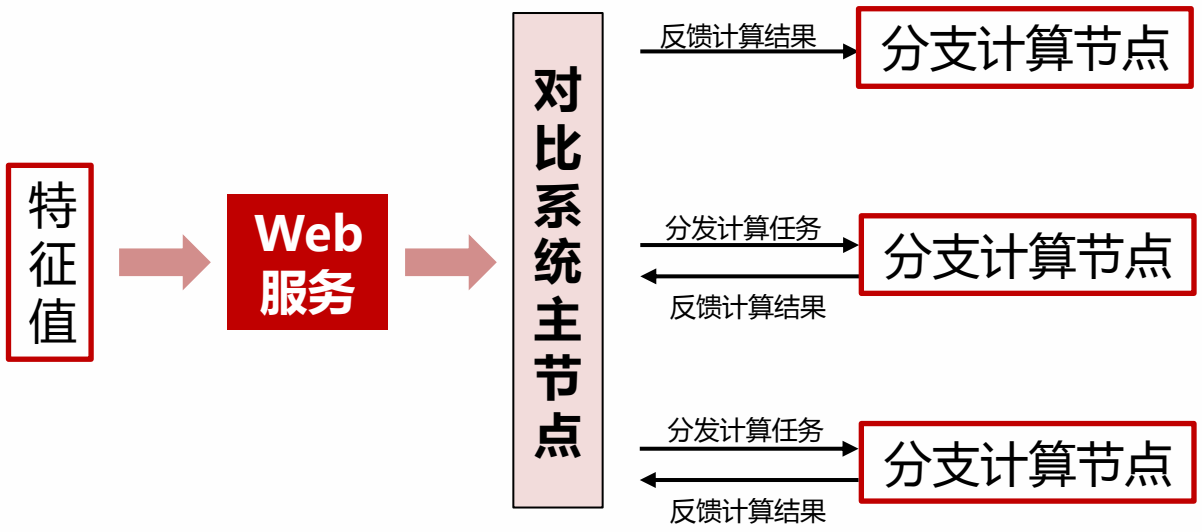
cVector向量计算一体机由云创大数据自主研发，是一款高性能的人脸特征向量高速对比计算一体机，支持亿量级别的大规模人脸1:N对比场景。与传统的基于GPU集群的人脸特征对比技术， cVector向量计算一体机同时融合了并行计算架构和高密度混合服务硬件支撑平台，可以满足用户在大规模人脸比对方面的需求，**具有高集成度、高性能、高兼容性、低成本以及性能弹性分配调度等优势。**

cVector向量计算一体机的主要应用场景为公安行业的人脸识别，以底层基础设施的形态，快速准确的为公安上层业务提供数据对比的结果。**同时也可用于公共场所敏感人员的人脸识别，满足场景高并发的需求，拥有极高的识别率。**

cVector向量计算一体机产品特性

五大产品特性
1秒7亿次人脸比对，毫秒级响应千万级数据
处理性能弹性配置，满足各种量级的业务需求
跨密级网间特征数据比对，实现安全交互
支持多算法综合比对，为客户使用提供便利
支持计算和存储分离，灵活调度，进一步降低成本

cVector向量计算一体机产品架构



信息来源：公司官网，西南证券整理

3.2.2 云创大数据的商业模式

云创大数据的商业模式主要包括四个方面：盈利模式、采购模式、研发模式和销售模式。

盈利模式	**云创大数据公司属于大数据行业，其终端用户主要为政府、学校以及企事业单位。**在明确了客户需求后，为其提供适当的产品或解决方案，根据项目的规模、复杂度和所需资源等因素，公司会制定合理的报价，并充分考虑相关成本，以确保获得合适的利润。
采购模式	**公司采购的内容主要为了满足日常产品开发和测试、项目实施所需的材料和第三方技术服务。**公司采购有十分严谨的流程，当采购需求提出后，经过多方审核，方可实施采购，保障采购成本处于合理状态。
研发模式	**云创大数据的研发工作主要包含两个方面：软件开发和硬件研发，**由多个部门共同合作完成。
销售模式	**公司目前共有三种销售模式：直接销售、经销商销售、系统集成商销售。** 直接销售：借助网络、行业专业展会等渠道推销产品并拓宽市场。 经销商销售：主要销售学科教育领域的大数据处理产品，在核查经销商公司的主体资格、企业诚信度等方面后与之达成合作。 系统集成商销售：系统集成商获得订单后与公司签订采购合同，并由其将公司产品集成至最终客户系统。

主要合作伙伴

清华大学 Tsinghua University

国家超级计算深圳中心 National Supercomputing Center in Shenzhen 深圳云计算中心 Shenzhen Cloud Computing Center

HUAWEI

中国移动 China Mobile

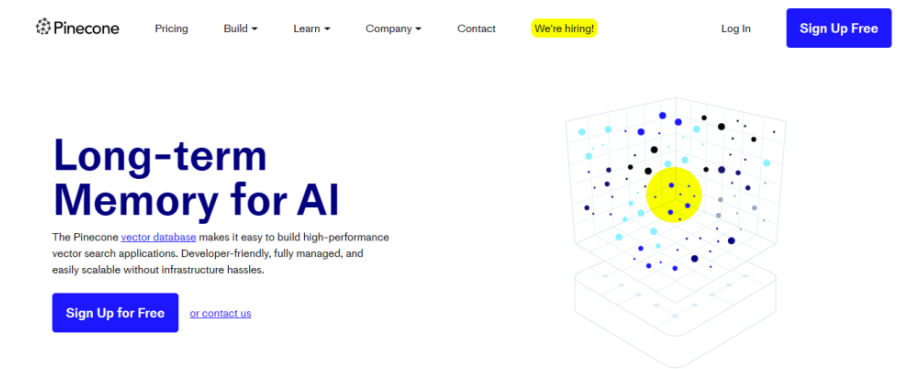
信息来源：《北交所企业研究：云创数据（835305）——大数据存储与处理服务商》，公司官网，西南证券整理

3.3.1 Pinecone公司简要介绍

Pinecone总部位于纽约，专为OpenAI的GPT-4等**大型语言模型(LLMs)**提供长期记忆服务。Vector Search专注于通过AI生成的内容表示进行存储和搜索。Pinecone为工程团队提供了**搜索基础设施**，以便在他们的应用程序中实施人工智能搜索，无需构建自身或修改旧的基础设施。**Pinecone是OpenAI、Cohere等LLM生成商的合作方**，如今已有1500个客户，下一步可能将与Shopify、Gong和Zapier等公司合作。

历史轨迹	2019年，Pinecone创始人Edo Liberty创立**Hypercube.ai**，提供基于深度学习的多媒体搜索解决方案； 2021年初，Hypercube.ai正式转化为**Pinecone**，专注于向量数据库领域研发。 2022年12月起，Pinecone先后上架**Google Cloud**和**AWS**，打开市场。
技术团队	工程师出自**Google、Databricks、Splunk**等知名科技企业 公司创始人兼首席执行官Edo Liberty获耶鲁大学计算机科学博士学位,曾任Yahoo的高级研究总监和纽约地区Yahoo研究实验室负责人，后加入AWS带领团队构建尖端的机器学习算法、系统和服务

Pinecone



信息来源：公司官网，西南证券整理

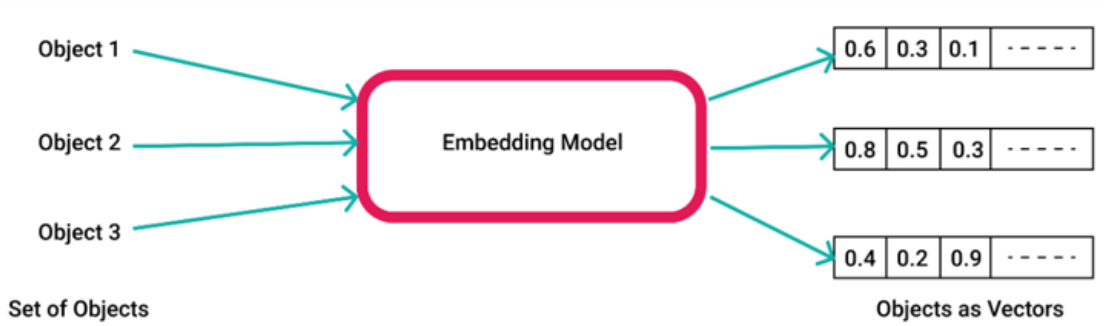
3.3.2 Pinecone——云原生向量数据库

Pinecone可**处理各类向量**，进行高性能、低延迟和可扩展的向量的**相似性搜索**并提供易用API。工程师可以使用AI模型快速构建可扩展的应用程序，并将其快速投入生产。相较于传统数据库，Pinecone具有以下**优势**：

- 可对数据库进行模糊搜索
- 大多数信息可转为向量，不局限于文字
- 可以用人类语言搜索，不局限于专有数据库操作语句

Pinecone特点

快速	**面对大量条目可保持低时延特性**
实时	可**实时更新数据变化**并索引
稳定	可保证**高可用性、安全性、一致性**
灵活	**支持多种语言和向量**模型，并且可以进行元数据过滤和排序等功能
可扩展	可**依需求**自动调整资源



Pinecone将某一个具象化信息如文字、图片通过机器学习模型或深度学习模型（即Embedding Model）转换成Vector。在获得每个Object的Vector之后，Pinecone可以使用此类Vector通过输入的文字、图片来完成对数据库的检索。

信息来源：公司官网，西南证券整理

3.4.1 Weaviate简要介绍

➢ **公司介绍**

Weaviate成立于2019年，目前在荷兰、美国、加拿大等地设有分支机构，拥有超过30名团队成员。Weaviate源自 **SeMI Technology**，从最初的**ING Labs**中分离而来。创始人、首席执行官Bob Van Luijt、首席技术官Etienne Dilocker具有资深工程师经历，同首席运营官Micha Verhagen一起专注于实现**民主化搜索**。
SeMI Technologies围绕其Weaviate开源解决方案提供**服务托管、服务许可协议和支持**等服务。

Weaviate产品特性

产品特点	主要内容
快速查询	Weaviate通常在不到100毫秒的时间内对数百万个对象执行最近邻(NN)搜索。
使用模块摄取任何媒体类型	使用最先进的AI模型推理（如Transformer）在搜索和查询时访问文本、图像等数据，管理数据矢量化的过程或提供导入使用者载体
组合矢量和标量搜索	Weaviate允许进行高效的矢量和标量组合搜索。Weaviate存储对象和向量，并确保两者检索有效，不需要第三方对象存储。
实时和持续性	在导入或更新过程中也可以搜索数据。每次写入都会写入预写日志(WAL)，以便立即持久写入
成本效益	大数据集不需要完全保存在Weaviate内存中。并且，其余可用内存可用于提高查询速度
对象	以图形方式在对象之间建立任意连接来模拟数据点之间的真实连接。使用GraphQL遍历
其他特性	水平扩展性、高可用性等

信息来源：公司官网，西南证券整理

3.4.2 Weaviate——开源向量数据库

➢ **应用场景**

- 使用Weaviate，可在**语义**上搜索数据
- 使用最先进的ML模型进行**文本和图像相似性搜索**（在一个Weaviate实例中结合存储和查询多种媒体类型）
- 将语义（向量）和标量搜索与需要的向量数据库相结合
- 将自身机器学习模型扩展到生产规模（HNSW算法和水平扩展支持近实时数据库操作）
- 快速且实时地对大型**数据集进行分类**

➢ **索引流程**

- 用户可以创建任意数量的索引，索引包含一定数量的分片，索引中的分片是独立的存储单元，每个分片均可以进行对象、倒置和向量存储，其中对象和倒置存储使用LSM树方法进行实现。
- 向量索引独立于这些对象存储，不受LSM分割的影响。

➢ **常用模块举例**

Weaviate**不自带任何模块**，各类功能将由**可选模块**进行执行。Weaviate还支持加载例如其他媒体类型的向量化、物体识别、拼写检查等外部模块

- **text2vec-contextionary** 将数据对象和对象的上下文进行向量化表示并保存到数据库中
- **text2vec-transformers** 利用sentence-transformers中丰富的嵌入模型为每个导入Weaviate的文本对象创建一个段落/文档嵌入，替代开发人员实现推理代码
- **img2vec-neural** 类似于 text2vec，使用大型预训练计算机视觉模型自动将图像矢量化，以无缝启用对任何图像的语义搜索

西南证券
SOUTHWEST SECURITIES

分析师：叶泽佑
执业证号：S1250522090003
电话：13524424436
邮箱：yezy@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
	持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
	中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
	回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
	卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
	跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
	弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券 SOUTHWEST SECURITIES

西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴东路166号中国保险大厦20楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区深南大道6023号创建大厦4楼

邮编：518040

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	汪艺	销售经理	13127920536	wyyf@swsc.com.cn
	崔露文	销售经理	15642960315	clw@swsc.com.cn	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn
	谭世泽	销售经理	13122900886	tsz@swsc.com.cn	陈阳阳	销售经理	17863111858	cyyyf@swsc.com.cn
	薛世宇	销售经理	18502146429	xsy@swsc.com.cn	李煜	销售经理	18801732511	yfliyu@swsc.com.cn
	刘中一	销售经理	19821158911	lzhongy@swsc.com.cn	卞黎旸	销售经理	13262983309	bly@swsc.com.cn
	岑宇婷	销售经理	18616243268	cyryf@swsc.com.cn	龙思宇	销售经理	18062608256	lsyu@swsc.com.cn
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	徐铭婉	销售经理	15204539291	xumw@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	胡青璇	销售经理	18800123955	hqx@swsc.com.cn
	杨薇	高级销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	销售经理	18500981866	wangyuf@swsc.com
	王一菲	销售经理	18040060359	wyf@swsc.com.cn	路漫天	销售经理	18610741553	lmtyf@swsc.com.cn
	姚航	销售经理	15652026677	yhang@swsc.com.cn	巢语欢	销售经理	13667084989	cyh@swsc.com.cn
广深	郑龑	广深销售负责人	18825189744	zhengyan@swsc.com.cn	张文锋	销售经理	13642639789	zwf@swsc.com.cn
	杨新意	销售经理	17628609919	yxy@swsc.com.cn	陈紫琳	销售经理	13266723634	chzlyf@swsc.com.cn
	龚之涵	销售经理	15808001926	gongzh@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	丁凡	销售经理	15559989681	dingfyf@swsc.com.cn				