2023年12月

弘则计算机

生成式AI驱动向量数据库加速发展

——对于AI产业趋势的思考

弘则研究科技组

电话：021-6194-6708

弘则弥道（上海）投资咨询有限公司

·科技 ·智造 ·消费 ·医药

核心观点提示

- **toB生成式AI应用均需外挂知识库以提升大模型精度，将驱动向量数据库的爆发。**向量数据库是非结构化数据的特殊形式，它的核心是将各种数据（如文档、音频和视频）转化为空间向量进行相似性搜索以提高不同数据类型的搜索效率和准确性，这使其在AI和深度学习领域中有着广泛的应用。生成式AI出现后，尤其是在toB场景中需要应用到向量数据库在大模型上外挂“企业知识库”，企业内部数据将被存储在向量数据库中，以提升大模型精度。

- **向量数据库潜在市场空间是传统结构化关系型数据库的数倍达到千亿美元。**据信通院统计数据，全球数据库市场规模在2020年为671亿美元，到2025年有望达到798亿美元，CAGR 3.5%，估算关系型数据库全球龙头Oracle收入规模小几百亿美元。仅考虑现有非结构化数据的向量化处理，估算需要的存储空间增量为之前的数倍。未来随着生成式AI应用增量数据的爆发，对于向量数据库的需求会更大。

- **产业处在发展早期，尚未形成寡头垄断，厂商具备错位竞争优势。**全球市场不同背景厂商以不同商业模式切入向量数据库赛道。美股上市公司中，MongoDB于今年12月在自身非结构化数据库业务基础上推出向量数据库产品Atlas Vector Search，Elastic于今年5月在自身搜索工具业务基础上推出向量搜索解决方案Elasticsearch Relevance Engine。最新业绩说明会中，两家公司均对向量数据库业务前景非常乐观。A股上市公司中，星环科技于今年5月推出向量数据库Hippo，目前已迭代到1.2版本，已有客户开始试用。

向量数据库作为B端Gen AI落地刚需，已经进入到商业化推广和落地阶段

向量数据库上市公司	
公司	产品更新和商业化更新
MongoDB	23年12月正式发布Atlas Vector Search功能，以向量数据库切入生成式AI产业； FY24Q2业绩会：“**向量数据库处于预览阶段，但已经看到大型客户的极大兴趣**，包括某咨询公司允许顾问在超过150万份专家纪要中进行语义检索。” FY24Q3业绩会：“**人工智能几乎存在于与各种规模的客户的每一次对话中。客户对矢量搜索公共预览版非常感兴趣。客户正在构建一系列人工智能用例，从语义搜索到检索增强生成。**例如，UKG为全球超过80,000多家客户提供服务，选择使用MongoDB Atlas Vector Search作为人工智能助手，帮助指导客户的员工、职能经理和人力资源主管。”
Elasticsearch	Elastic官方从2018年开始支持向量检索功能，23年5月推出ESRE（Elasticsearch Relevance Engine），目前作为8.8版本的一部分，所有功能会随白金级套餐和企业级套餐一起提供； FY24Q1业绩会：“我们看到围绕生成式AI的大量活动，许多客户选择ESRE作为使用我们的向量搜索和混合搜索功能构建生成式AI应用程序的平台。**目前有数百名付费客户使用 ESRE 进行向量搜索。**” FY24Q2业绩会：“Elastic Cloud同比增长31%，**这一增长得益于云消费的改善和在生成式AI领域的成功。**客户越来越多地将多种用例整合到Elastic平台上，取代了原有搜索解决方案。”
星环科技	23年5月正式发布行业大模型、向量数据库和大模型开发工具，向量数据库Hippo已迭代到1.2版本。11月，英特尔与星环科技联合发布AIGC向量数据库解决方案。**目前已有金融客户采购大模型开发工具，银行和券商等客户正在POC行业大模型和向量数据库**

美股财报指引+wind一致预期下估值水平变化（市值参考日期：23年11月13日）									
	市值（亿美元，亿元）	过去财年（FY23，2022）				当前财年（FY24，2023）		下一财年（FY25，2024）	
		收入	3年CAGR	毛利率	PS	收入预期	PS	收入预期	PS
MongoDB	269	12.8	45%	73%	21	26%	17	-	-
Elasticsearch	75	10.7	36%	72%	7	17%	6	-	-
星环科技	88	3.7	29%	57%	24	42%	17	42%	12

来源：互联网公开资料、弘则研究整理

Horizon Insights
弘 则 研 究

01 数据库发展复盘

70-80年代：数据库市场开始起步，Oracle、IBM、Microsoft三巨头并起，切分不同客户群体

70-80年代数据库的需求点

集中化的数据存储：
为了更有效地管理和利用这些数据，企业需要一个集中的地方来存储它们。这导致了关系数据库管理系统的出现，提供了结构化的方式来存储、查询和更新数据

数据的可靠性和完整性：
企业的数据是其最宝贵的资产之一，因此数据的可靠性和完整性至关重要。这需要数据库管理系统提供事务管理、备份和恢复等功能。

高效的数据访问：
随着数据量的增长，企业需要能够快速访问和查询数据。这要求数据库管理系统提供高效的查询优化和数据访问机制。

三巨头错位竞争

	Oracle	IBM	Microsoft
产品功能	第一个商业关系数据库管理系统。采用关系数据库模型，支持SQL查询，并提供强大的事务管理功能	DB2是为大型机设计的关系数据库管理系统，提供了高效的数据访问和强大的事务管理功能	SQL Server是一个关系数据库管理系统，支持SQL查询，并提供了一系列的数据管理和分析工具
商业策略	是为大型企业提供高性能、可扩展的数据库解决方案。它的客户群主要是大型企业和政府机构	利用其在大型机市场的领导地位，为其客户提供一站式的IT解决方案，包括硬件、软件和服务	通过与Windows操作系统的紧密集成，为中小企业提供数据库解决方案，SQL Server迅速获得市场份额
竞争壁垒	**技术领先**	**大型机市场的领导地位**	**Windows操作系统绑定**

来源：互联网公开资料、弘则研究整理

70-80年代：Oracle最初凭借技术和战略决策领先IBM推出产品抢占市场，微软起步较晚且主要客群集中在中小企业

Oracle与IBM的发展背景

1974年
IBM开始构建System R，历史上第一个使用SQL查询语言的数据库，但仅作为内部研究项目，当时IBM的战略重心仍在硬件业务

→

1977年
受到System R启发，Software Development Laboratories公司（Oracle前身）成立，最初想围绕IBM产品做协同工作

→

1979年
IBM不感兴趣合作后，Oracle开始自主开发产品，并推出Oracle V2，后续拿到CIA价值5万美元的合同

	Oracle	IBM
产品逻辑	SQL查询语言	SQL查询语言
战略地位	初创公司进攻市场的拳头产品	内部研究，验证关系数据库的理论和概念
商业模式	**开创了软件license的商业模式**	**硬件+嵌入软件，认为价值量应该通过硬件体现**

Oracle V2与IBM System R的对比

	Oracle V2	System R
查询场景	主要关注数据存储和基本的SQL查询功能，索引结构和查询优化策略在当时是比较先进的	引入了许多现代关系型数据库的核心概念，如R树索引、查询优化等
性能优化	主要关注于基本的查询功能，性能优化功能相对较少	引入了查询重写技术，可以自动将复杂的查询转化为更简单、更高效的形式
特性功能	提供了基本的SQL查询功能，没有太多的高级特性	引入了许多现代关系型数据库的核心概念和特性，如事务管理、并发控制等

- System R在理论和研究上引入了许多先进的概念，但**实际应用中需要付出维护复杂、资源消耗、兼容性等代价**
- Oracle V2推出时的商业化目的极其明确，使其**具备兼容性、易用性的优势，在实际应用中更加稳定高效**，早期进入市场后获得了先发优势，占领客户后迭代加速

来源：互联网公开资料、弘则研究整理

80-90年代：计算机技术和互联网的大规模应用驱动关系型数据库继续向高性能、高可靠性方向发展

80-90年代技术趋势		
技术发展趋势	**影响**	**数据库需求**
计算机和网络技术的普及	随着个人计算机和企业计算机的普及，数据量开始迅速增长。企业开始积累大量的业务数据，如销售数据、库存数据等	企业需要更大、更高性能的数据库系统来存储和管理这些数据。数据的备份和恢复、数据的安全性和完整性也成为了关键需求
分布式计算和网络技术的发展	随着局域网和广域网技术的发展，企业的数据开始分布在多个地点。企业的业务也开始跨越多个地理位置	企业需要分布式数据库技术来管理这些分布在不同地点的数据。数据的同步、数据的远程访问和数据的分布式查询成为了关键需求
数据分析和商业智能的兴起	企业开始重视数据分析和商业智能。数据不再仅仅是用来记录业务，而是用来支持决策和提供洞察	企业需要数据仓库技术来存储和管理用于分析的数据。数据挖掘、报表生成和数据可视化成为了关键需求
软件和应用的发展	随着软件和应用的发展，企业的业务流程变得更加复杂。企业开始使用ERP、CRM等复杂的业务应用	这些应用需要高性能、高可靠性的数据库系统作为后端。数据库的事务处理、并发控制和数据完整性成为了关键需求
开放系统和标准化的趋势	企业开始追求开放系统和标准化的解决方案。这使得企业可以选择最佳的技术和产品，而不是被锁定在某个厂商的技术生态中	SQL成为了标准的查询语言，被广泛应用于各种数据库系统。企业需要支持SQL的数据库系统，以确保与各种应用和工具的兼容性

- **个人计算机数据库的兴起：**在IBM PC之前，数据库主要运行在大型机和小型机上。但随着IBM PC的普及，开始出现为个人计算机设计的数据库系统，如dBASE、FoxPro和Paradox。这些数据库系统为个人和小型企业提供数据管理能力
- **客户端-服务器架构的普及：**在这种架构中，客户端（通常是个人计算机）负责用户界面和应用逻辑，而服务器（可能是另一台更强大的计算机）负责数据管理。这种分离使得数据可以集中管理，而应用可以在多个客户端上运行
- **数据库工具和应用的发展：**随着IBM PC的普及，开始出现了大量的数据库工具和应用，如查询工具、报表生成器和数据库开发工具。这些工具使得数据库开发和管理变得更加简单和高效
- **数据库的标准化：**IBM PC的普及促进了数据库技术的标准化。SQL成为了标准的查询语言，被广泛应用于各种数据库系统。这使得开发者可以使用统一的语言和工具开发和管理数据库，而不用担心不同数据库系统之间的兼容性问题
- **数据库市场的竞争加剧：**IBM PC的成功吸引了大量的硬件和软件制造商进入数据库市场。这导致了数据库市场的竞争加剧，促进了技术的发展和价格的下降

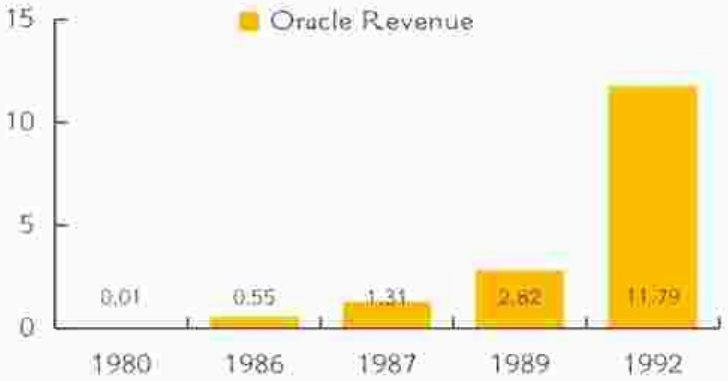
来源：互联网公开资料、弘则研究整理

80-90年代：有力竞争者增加，但最终输在商业策略退出市场

80-90年代新进入者及退出原因

厂商	优势	退出市场原因	结果
Sybase	• 客户端/服务器架构 • 金融行业优化	• 产品稳定性问题 • 与Microsoft合作关系破裂	• 被SAP于2013年收购
Informix	• 高性能和可靠性 • 对象关系特性	• 失败的收购 • 主要产品稳定性和性能问题	• 2001年被IBM收购，其技术被整合到DB2
Ingres	• 开放源代码 • 跨平台支持	• 市场营销和销售策略问题	• 作为开源项目重新出现，现在由Actian公司维护
Paradox	• 桌面数据库 • 集成开发环境	• 面临Microsoft Access等桌面数据库产品的竞争	• 被Borland收购，市场地位逐渐下降
dBASE	• 桌面数据库 • 集成开发环境	• 面临Microsoft Access等桌面数据库产品的竞争	• Ashton-Tate在1991年被Borland收购
FoxPro	• 数据访问速度 • 集成开发工具	• 面临Microsoft Access等桌面数据库产品的竞争	• 1992年被微软收购，整合到Microsoft的产品线中，但在2000年代初停止开发

厂商	产品变化	商业策略
Oracle	• 成为RDBMS市场的领导者，转型为"应用提供商"	• 将业务扩展到中小企业市场，在中端市场寻找增长机会
IBM	• 主要数据库为DB2 • 产品不仅限于高端市场	• 支持所有主要的操作平台 • 在NT-based RDBMS市场与Microsoft竞争
Microsoft	• SQL Server产品特点是易于管理、实施和成本效益	• 在中小企业市场与Oracle和IBM竞争，依赖操作系统



来源：互联网公开资料、弘则研究整理

进入到21世纪之后，分布式、非结构化、开源、转云成为重要趋势

产业发展趋势	
趋势	**原因**
分布式和非结构化	**数据多样性：**现代应用产生的数据类型和结构日益多样化，如社交媒体、日志、图片等
	技术进步：存储和处理非结构化数据的技术（如Hadoop、NoSQL数据库）得到了广泛的研究和应用
	业务需求：企业需要对非结构化数据进行深入分析，以获得更多的业务洞察和价值
开源	**成本压力：**企业寻求降低IT成本，而开源软件通常没有许可费用
	技术创新：开源社区鼓励技术创新和共享，加速了技术的发展和迭代
	透明性和信任：开源代码的透明性使企业能够更好地理解和信任所使用的技术
云数据库	**运维简化：**云服务提供了数据库的自动管理、备份和恢复，降低了运维复杂性
	全球化需求：随着业务的全球化，企业需要在多个地理位置提供服务，云数据库满足了这一需求
	资本投资减少：使用云服务，企业可以按需付费，避免了大量的前期硬件投资

- **市场更为分散：**关系型数据库时代，三巨头依靠技术或商业策略维持自身极高的护城河。分布式、非结构化、开源和云数据库趋势出现后，在新兴领域出现大量新进入者，比如MongoDB、Redis、Elastic、Pinecone、Milvus等
- **商业模式出现转型：**关系型数据库时代的商业模式多为license收费，随着客户采购更多服务器节点而增长，比如Oracle是按照服务器内存核数收取license费用。开源数据库厂商背后多有开源基金会等产业资金进行支持，因此license免费，主要收取后续技术支持等服务费用。上云趋势出现后，有些数据库厂商商业模式转变为云托管模式，购买数据库厂商的服务包括数据库产品和云上的存储/计算资源，数据库厂商再与云厂商进行成本结算

传统的结构化关系型数据库最重要的ACID特性使其在特定应用场景中非常重要

ACID事务特性

Atomicity

- 这意味着事务被视为一个单一的、不可分割的单位，它要么完全执行，要么完全不执行。如果事务的一部分失败，整个事务都会失败，并且数据库状态不会改变
- 例如，如果在银行转账过程中，从一个账户扣款成功，但向另一个账户存款失败，整个事务都会被回滚，确保资金的完整性

Consistency

- 事务确保数据库从一个一致的状态转移到另一个一致的状态。在事务开始之前和结束之后，所有的业务规则都必须保持为真
- 例如，银行转账应确保转账前后的总金额保持不变

Durability

- 一旦事务被确认，它的效果是永久的，即使在系统故障、崩溃或重启后也不会丢失
- 这通常通过将事务日志持久化到存储介质上来实现

Isolation

- 这确保并发事务的执行不会互相干扰。每个事务应该在一个隔离的环境中运行，好像没有其他事务并发执行一样
- 这可以通过多种隔离级别来实现，例如读未提交、读已提交、可重复读和串行化

ACID特性在许多应用场景中非常重要，尤其是在需要确保数据完整性和一致性的金融、医疗和零售等行业

来源：互联网公开资料、弘则研究整理

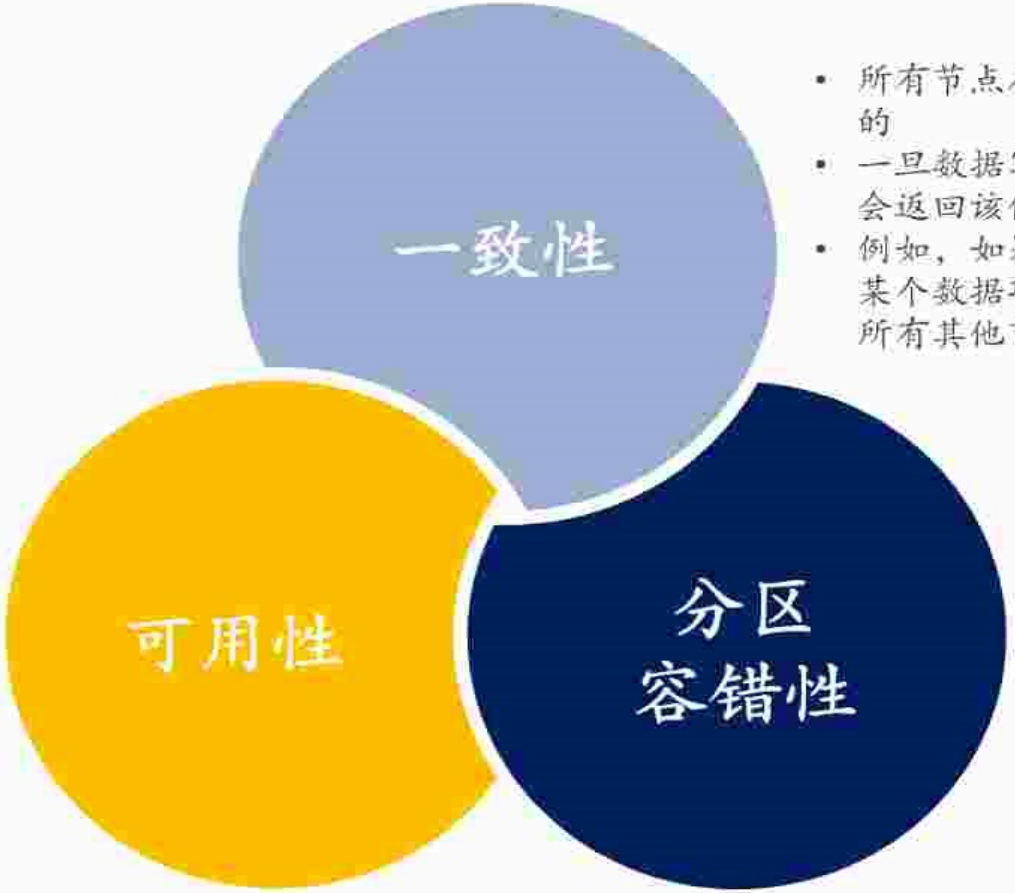
随着大数据、云计算等技术成熟，关系型数据库最重要的ACID特性开始制约其发展

新产业趋势	
趋势/需求	描述
数据量的爆炸性增长	互联网、社交媒体、物联网和移动设备导致数据生成的速度和规模迅速增长。传统的RDBMS在处理PB级别的数据时可能会遇到性能瓶颈
高并发和低延迟需求	互联网应用和服务需要能够支持数百万甚至数十亿的用户并发访问，同时要求低延迟的响应。传统的RDBMS可能难以满足这种高并发、低延迟的需求
弹性和可扩展性	云计算的兴起要求数据库能够轻松地在多个服务器和数据中心之间扩展。传统的RDBMS在水平扩展（横向扩展）上可能存在挑战
多样化的数据模型	不是所有的数据都适合关系模型。例如，社交网络数据、地理位置数据和时间序列数据可能更适合其他数据模型。NoSQL数据库提供了文档、键值、列族和图等多种数据模型，以满足这些特定的需求
数据结构的变化	在互联网和移动应用中，数据结构可能经常变化。传统的RDBMS需要固定的表结构，而NoSQL数据库通常更加灵活，允许数据结构的动态变化
分布式和全球化	为了提供全球化的服务和减少延迟，数据需要在全球多个地点存储和访问。传统的RDBMS可能不具备这种分布式和全球化的能力
成本考虑	开源和NoSQL数据库通常具有较低的总体拥有成本（TCO），尤其是在硬件、许可和维护方面。这使得它们在初创公司和互联网企业中尤为受欢迎

来源：互联网公开资料、弘则研究整理

分布式：基于CAP理论，在一致性、分区容错性和可用性三者之间寻找平衡点

CAP理论



- 所有节点在同一时刻看到的数据是一致的
- 一旦数据写入成功，任何后续的读取都会返回该值或更新的值
- 例如，如果一个系统保证一致性，并且某个数据项在节点A上被修改，那么在所有其他节点上也应立即看到这个修改

- 每个请求（无论是读还是写）都会在有限的时间内返回一个结果，即使某些节点可能是不可用的
- 这意味着系统始终是在线的，但返回的数据可能不是最新的或不一致的

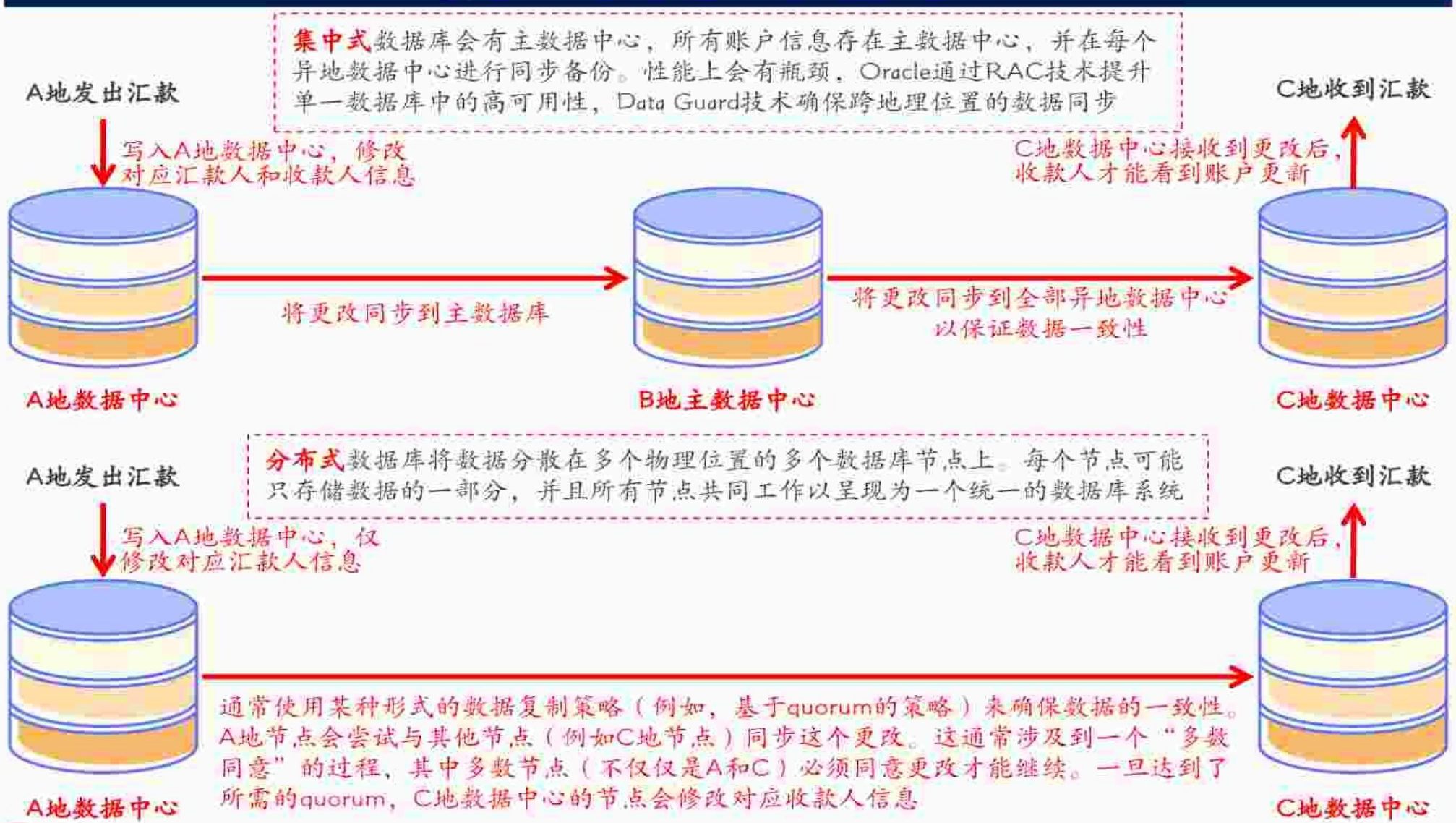
- 系统可以在网络分区（即节点之间的通信中断）的情况下继续运行
- 即使网络发生故障，导致数据存储的各个部分之间的通信中断，系统仍然可以正常响应用户的请求

CAP理论的核心观点是，分布式系统在面对网络分区时，必须在一致性和可用性之间做出选择。例如，一些系统可能会提供可调整的一致性级别，允许开发者根据需要选择更强的一致性或更高的可用性

来源：互联网公开资料、弘则研究整理

分布式：集中式数据库主数据中心来保证ACID特性，而分布式数据库牺牲掉主数据中心以加速数据流转

集中式和分布式的数据流（以银行汇款业务为例）



分布式：相较于集中式数据库，可以提供更强的扩展性、更低的网络延迟和更强的安全性

集中式数据库瓶颈（以前述场景为例）		
	集中式数据库的情况	分布式数据库的解决方案
网络延迟	A地发起汇款请求，需要首先发送到B地中心数据库，然后再将结果返回到A和C地，这个往返的过程会增加交易的响应时间	分布式数据库可以通过数据复制和地理分布来减少网络延迟的影响
单点故障	如果集中式数据库的服务器出现故障，整个系统可能会中断	通过数据复制和故障转移，分布式数据库可以实现高可用性，即使某些节点出现故障，系统仍然可以运行
扩展性	随着交易量的增加，集中式数据库可能会遇到性能瓶颈。垂直扩展（增加单个服务器的资源）有其限制，而水平扩展（增加更多的服务器）在集中式架构中可能更为复杂	分布式数据库支持水平扩展，可以通过增加更多的服务器来提高性能和存储容量
备份和恢复	集中式数据库需要定期备份，以防数据丢失。但随着数据量的增长，备份和恢复的过程可能会变得更加耗时	分布式数据库可以并行地在多个节点上进行备份和恢复，从而加速这一过程
安全风险	集中存储所有数据可能会增加安全风险。如果攻击者成功入侵了中心数据库，他们可能会获得大量的敏感信息	分布式数据库的分散存储可以降低单点攻击的风险，同时可以实现更细粒度的安全控制
成本	集中式数据库的垂直扩展和专有软件许可可能导致高昂的成本	分布式数据库通常支持开源和水平扩展，可以提供更低的总体拥有成本

来源：互联网公开资料、弘则研究整理

分布式：相比集中式数据库，现阶段分布式受制于数据一致性、迁移成本等，大规模商用仍存在落地难度

分布式数据库落地难点	
	描述
成熟度	许多分布式数据库是相对较新的技术，而传统的关系型数据库（如Oracle、SQL Server、DB2等）已经存在了几十年，它们已经在许多关键业务应用中得到了验证
复杂性	分布式数据库的设计和管理通常比单一节点的数据库更复杂。这需要数据库管理员和开发人员具有新的技能和知识
一致性和可用性	根据CAP理论，分布式系统必须在一致性和可用性之间做出权衡。某些业务场景，如金融交易，可能更倾向于选择保证强一致性的系统
迁移成本	对于已经在使用传统关系型数据库的企业来说，迁移到新的分布式数据库可能涉及高昂的迁移成本，包括数据迁移、应用程序更改和员工培训
工具和生态系统	传统的关系型数据库通常有一个成熟的工具和生态系统，包括备份、监控、性能调优等。而新的分布式数据库可能还在这方面迎头赶上
特定应用场景	虽然分布式数据库非常适合大数据、高并发和全球分布的应用，但并不是所有应用都需要这些特性。许多企业应用可能不会受益于分布式数据库的特点

- 根据Gartner的报告，尽管分布式数据库的采用率在增加，但传统的关系型数据库（如Oracle、Microsoft SQL Server和IBM DB2）仍然占据了数据库市场的大部分份额
- 互联网公司如Facebook和Google已经开发了自己的分布式数据库解决方案，例如Bigtable和Cassandra，以满足他们的特定需求
- 一些新兴的分布式数据库，如CockroachDB和TiDB，正在获得越来越多的关注和采用，这表明分布式数据库在某些场景中的优势
- 国内受到信创驱动，金融行业头部的银行、资管机构已经开始进行可研分析并逐步开始将集中式数据库替换成分布式数据库

来源：互联网公开资料、弘则研究整理

非结构化：互联网催生不同数据类型的数据爆发，传统关系型数据库面临困境

非结构化数据库特性	
	描述
数据模型	• 文档型：使用JSON或BSON格式存储数据，每个文档可以有不同的结构 • 键值型：使用键值对存储数据，适合快速读写 • 列族型：数据按列存储，适合大量数据的写入 • 图型：用于存储和查询图结构的数据
分布式架构	• 分片：数据被分成多个部分（或“分片”），每个分片存储在不同的服务器上 • 复制：数据的多个副本存储在不同的服务器上，以提高可用性和容错性 • 最终一致性：允许短暂的数据不一致
灵活的查询方式	• 提供丰富的查询API和语言，如MongoDB的查询语言
水平扩展	• 通过增加更多的服务器，可以轻松地扩展其存储和处理能力
内存存储和缓存	• 如Redis主要在内存中存储数据，提供超高的读写速度
CAP理论	• 在一致性可用性和分区容错性三者之间存在权衡
数据压缩和存储优化	• 采用特定的数据压缩技术和存储结构
事件驱动和实时处理	• 支持事件驱动的数据处理和实时查询

	传统RDBMS的特点和挑战	NoSQL数据库的特点和优势
可扩展性	为了保证ACID特性，采用单一、集中式的架构。虽然可以通过增加硬件资源来提高性能，但成本高且扩展性有限	采用分布式架构，更容易实现水平扩展，可以在多个服务器和数据中心之间分散数据和负载
灵活性	需要预定义的表结构。在互联网和移动应用中，数据结构可能经常变化，这使得RDBMS在适应这些变化上面临挑战	允许动态的、不固定的数据结构，更适合快速变化的环境
高并发和低延迟	为了保证ACID特性，可能需要在事务处理中加锁，这可能会影响并发性能和响应时间	放宽了ACID的一些要求，采用最终一致性模型，以实现更高的并发性和更低的延迟
成本	通常需要昂贵的许可费和硬件资源	企业和开发者开始寻找更经济、更灵活的方案

来源：互联网公开资料、弘则研究整理

云数据库：2012年左右移动互联网和物联网发展驱动企业向弹性平台转移

移动互联网和物联网驱使企业加速上云

- **CMO非常喜欢公共云。**毫无疑问，各地的营销部门都在内部IT之外部署Web和移动应用程序来与客户互动。通常，他们会求助于在PaaS（平台即服务）上构建这些应用程序的厂商，也许会在后端进行大数据分析，因为内部IT缺乏时间、意愿或技能来构建此类系统。这在IT参与规划和管理的情况下非常有效

- **物联网是云。**2012年4月，VMware分拆公司Pivotal作为下一代PaaS推出，得到了GE 1.05亿美元投资的支持，GE正忙于在广泛的工业产品中嵌入数百万个传感器。该平台的一个关键组件是GemFire事件处理软件，旨在处理来自所有这些传感器的遥测数据。11月，亚马逊添加了Kinesis服务，Salesforce宣布了Salesforce one集成平台，两者都可以用于类似的目的

- **云客户端很快就会占据统治地位。**云的最终目标与任何IT基础设施相同：交付应用程序。但在浏览器中运行的应用程序无法达到本机桌面或移动应用的水平。新的JavaScript框架正在缩小这一差距

混合云需求开始出现

- **多家厂商开始提供混合云方案。**长期以来，云的梦想就是让公共云成为内部基础设施的延伸。在实践中，“爆发”到云端往往是不切实际的。但如果至少能管理一部分本地和公有云资源，就可以减轻IT的负担。2013年的令人惊讶的事情之一是Microsoft在这个方向上采取了积极的行动，Windows Server 2012和System Center为Azure资源提供了更广泛的渠道。VMware虽然没有走得那么远，但计划采取类似的方法。当然，OpenStack的重点之一是为公共云和私有云建立一个框架

- **混合云管理成为关键。**全球云系统管理软件市场增长迅猛，2011年市场规模预计达到7.54亿美元，比2010年增长84.4%。排名前两位的供应商CA Technologies和VMware受益于市场对云系统管理软件的需求

云数据库：超融合技术的成熟极大简化混合云环境的部署和管理

分散式：服务器主要使用其内部的磁盘或外部的直连存储(DAS)

存储瓶颈：性能、容量等要求的提高

整合式: 直连存储→集中式共享存储模式

计算瓶颈：服务器增多、资源利用率、管理成本边际效应出现

计算虚拟化：服务器虚拟化提高资源利用率和运维效率；但服务器与存储之间使用的仍是三层架构

存储扩张瓶颈：存储设备数量递增，传统架构无法提供灵活的纵向及横向扩展能力

物理融合：将服务器、存储、存储网络交换机放置在同一个机架中以模块方式销售。

I/O瓶颈：存储未虚拟化，仍遭遇I/O瓶颈

超融合：存储虚拟化
摒弃了三层架构固有的问题，支持以更小的颗粒进行横向扩展，更高的可用性和可扩展性，很快成为新建、更新数据中心时的首选方案。

硬件定义 → **软件定义：计算、存储、网络虚拟化，打破瓶颈**

- **简化管理：**超融合基础设施（HCI）是一种将计算、存储和网络功能集成在一起的基础设施，可以简化IT管理，使部署和运行应用程序更加容易。这种模式对于混合云环境来说非常重要
- **灵活性和可扩展性：**HCI通过提供一种可以轻松扩展的基础设施，支持了混合云环境的灵活性和可扩展性。当需要更多的计算或存储资源时，可以简单地添加更多的HCI节点，而不需要进行复杂的硬件升级或配置更改
- **一致的操作体验：**HCI可以提供一种一致的操作体验，无论应用程序是运行在本地的HCI环境中，还是在云环境中

2012-2013年间，AWS Outposts，Microsoft Azure Stack和Google Anthos的混合云基础设施设计被推出，提供在公共和私有基础设施上一致的云服务、API和管理界面，**在某种程度上被视为超融合技术的一种形式。**

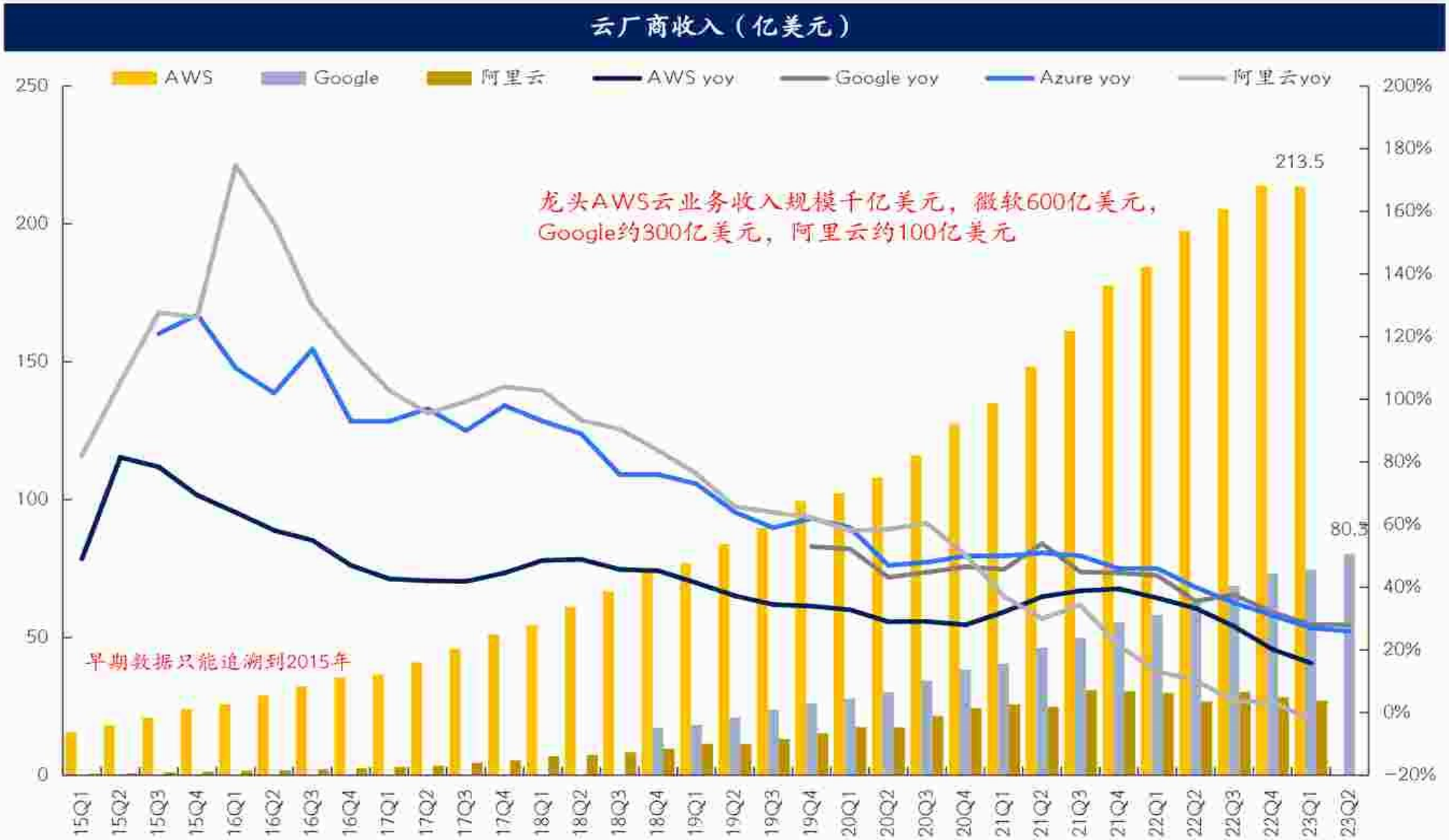
云数据库：早期微软Azure云的成长驱动力即本地SQL Server和Azure云上数据库的混合云解决方案

季报发布会引用			
	发言引用		发言引用
FY10Q4	进一步提高本地Windows Server，SQL Server和System Center产品与Azure平台之间的一致性	FY15Q2	商业云连续第六个季度实现三位数收入增长，高级Azure服务的收入大幅增长
FY11Q1	Windows Azure订阅量环比增长40%	FY15Q4	我们看到SQL的大量采用，所以这就是Azure DB，机器学习即服务。
FY11Q4	Windows Azure继续拥有强劲的客户势头，收入增长加速	FY16Q1	SQL的这一里程碑与我们在Azure中的快速增长及其在云魔力象限中的位置并列
FY13Q4	增加25%的企业客户，超过50%的财富500强企业使用Azure	FY16Q2	我们的服务器不是一个独特的部分，实际上是我们云的边缘，我们正在通过Azure Stack之类的东西来构建
FY14Q2	Azure客户净席位增长超过100%，70%的财富500强公司使用至少一项云服务	FY16Q3	我们在Azure中的高价值服务中添加了更多差异化服务，即人工智能、IoT和业务分析
FY14Q3	Azure收入增长150%以上，得益于新客户和使用率的增加	FY16Q4	我们显然支持我们所有的服务器。我们的每个服务器产品都有云注册权限，无论是SQL，还是Windows Server
FY14Q4	商业云收入增长147%，Azure大幅增长，今年存储翻了一番，计算量增加了两倍。随着核心服务的使用量增加，超过50%的Azure客户现在也在使用更高价值的服务	FY17Q1	这就是SQL Server 2016所代表的，因此我们拥有这些独特的功能，例如能够在SQL 2016中将数据库中的单个表一直延伸到云中，以获得无限的表容量，然后让您的应用程序和查询工作
FY15Q1	Azure实现了强劲增长；初创公司和ISV喜欢开放灵活的方法，并且正在Azure上快速构建，40%的收入来自初创公司和ISV		

Azure云的数据库服务很难单独量化拆分带来的影响，但从微软季报发言中仍能看到数据存储服务带来的上云驱动力：

- Azure最初主要针对的是大型企业和政府机构，这些用户通常已经是微软的现有客户，使用微软的其他产品，如Office、Windows Server和SQL Server等
- 微软的Hyper-V虚拟化技术是一个由微软开发的虚拟化平台，可以作为独立产品使用，也可以作为Windows Server的一个功能。虚拟化技术和其他迁移工具使得微软整套云端和本地的服务具备高集成性和可迁移性
- 因此，Azure最初主要提供混合云解决方案，使得深度使用微软全产品线的大型企业用户可以在保留现有IT基础设施的同时，逐步迁移到云

云数据库：作为云厂商提供的云上打包服务之一，很难单独量化拆分数据库产品带来的影响，但其一定随着云业务同向成长



来源：互联网公开资料、弘则研究整理

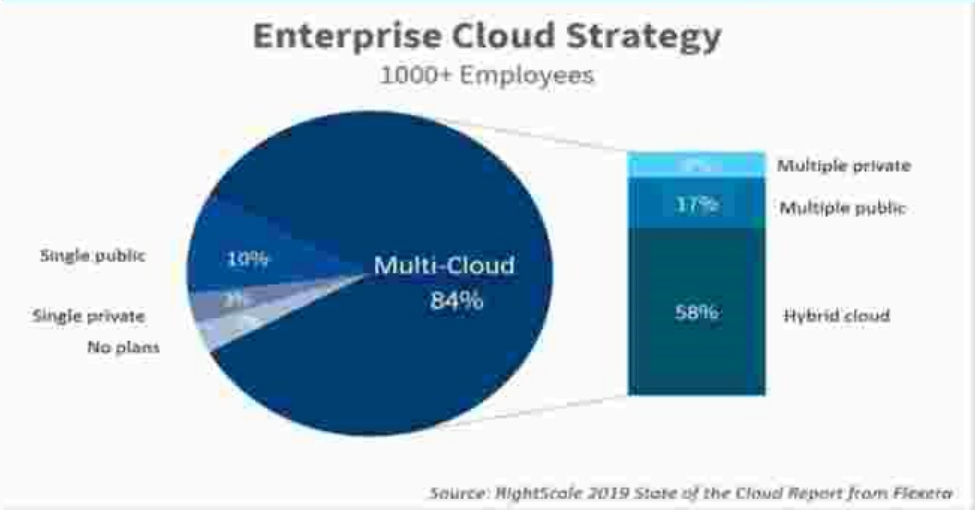
开源：对于厂商来说，开源更多是战略选择而非被迫转型

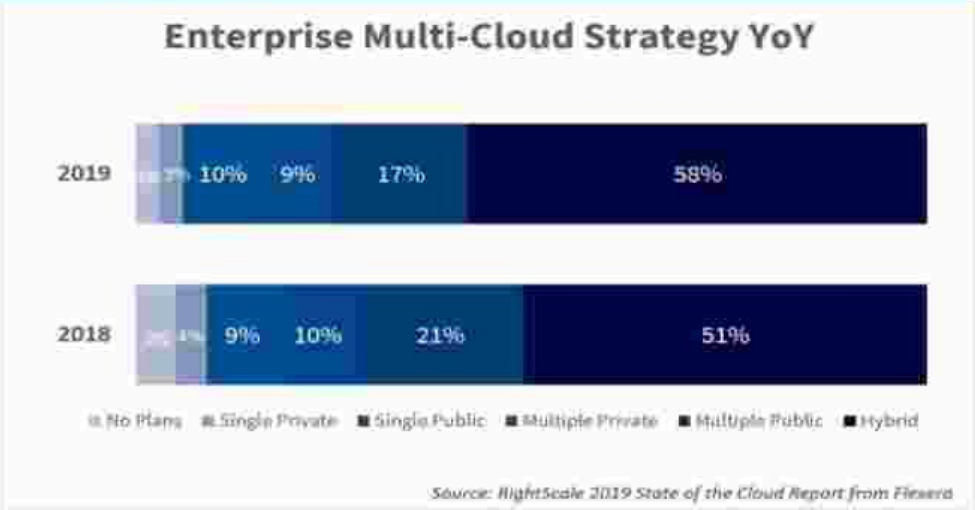
	描述
市场渗透	通过提供开源版本的产品，厂商可以迅速扩大其市场份额，吸引更多的用户。这为厂商创造了一个广泛的用户基础，从而为其后续的商业化策略提供了基础
社区驱动	开源模式鼓励社区的参与，这意味着厂商可以利用全球范围内的开发者为其产品带来创新。这种模式大大加速了产品的开发和改进且避免“重复造轮子”
品牌建设	成功的开源项目可以为厂商带来良好的声誉和品牌知名度。例如，Red Hat、Canonical和Docker等公司通过其开源项目建立了强大的品牌
生态建设	开源有助于建立行业标准，从而吸引更多的合作伙伴和开发者加入到厂商的生态系统中
竞争	通过开源某些关键技术或平台，厂商可以策略性减少竞争，将竞争对手转化为合作伙伴
人才	开源项目通常吸引了大量的开发者和贡献者。这为厂商提供了一个优质的人才库，从中挑选和招聘人才

来源：艾瑞咨询、互联网公开资料、弘则研究整理

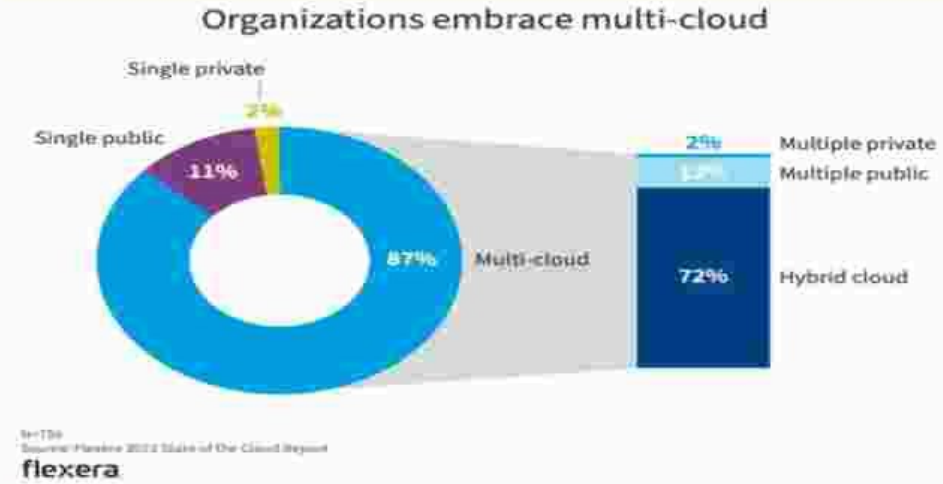
相比国内，海外企业多采用多云策略，因此云厂商战略重心集中在IaaS层，并不会向上进入细分垂类SaaS场景

海外企业多云策略（2019年Flexera调研报告）





海外企业多云策略（2023年Flexera调研报告）



- **避免供应商锁定：**企业不想被单一厂商锁定，可以确保他们有更大的灵活性和谈判能力
- **风险分散：**使用多云服务分散风险，确保当一个服务出现问题时，业务可以继续运行
- **满足特定需求：**不同的云服务提供商可能在某些特定领域或功能上有优势。企业可能会选择最适合其特定需求的云服务
- **合规性和数据主权：**在某些地区或行业，数据可能需要存储在特定的地理位置或满足特定的合规要求。多云策略可以帮助企业满足这些要求

因此，海外云厂商的战略重心集中在IaaS层各种技术的突破比如机器学习、AI等，以提供给企业更有吸引力的存储、计算等能力，SaaS场景引入垂直合作伙伴。国内云厂商则会打包提供全部IaaS+SaaS服务。

来源：互联网公开资料、弘则研究整理

商业模式从本地license+技术支持，转向开源、云托管商业模式

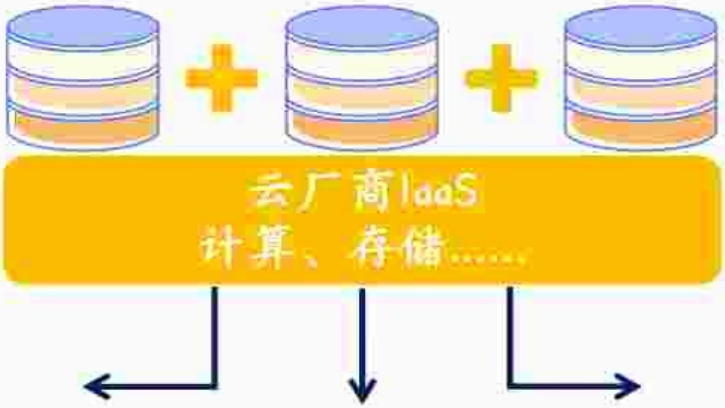
传统：本地license+技术支持模式（Oracle首创纯软件的商业模式，在此之前所有软件都是以嵌入硬件销售的方式体现价值）

- Oracle报价按照服务器硬件性能报价（几核CPU）
- 硬件扩容的时候需要向Oracle增购license
- Oracle也提供技术支持和咨询

开源：增值功能+技术支持模式

- 开源license免费，但存在功能限制，比如速度瓶颈、一些高级功能等
- 需要增值功能时才升级付费
- 额外还有服务收费，比如数据维护服务、更新、安全补丁、技术支持等

云托管模式（MongoDB）



- 用户可以在云上采购数据库厂商服务，或直接采购数据库厂商服务
- 价格包含云基础设施费用，数据库厂商再与云厂商结算

来源：互联网公开资料、弘则研究整理

Horizon Insights
弘 则 研 究

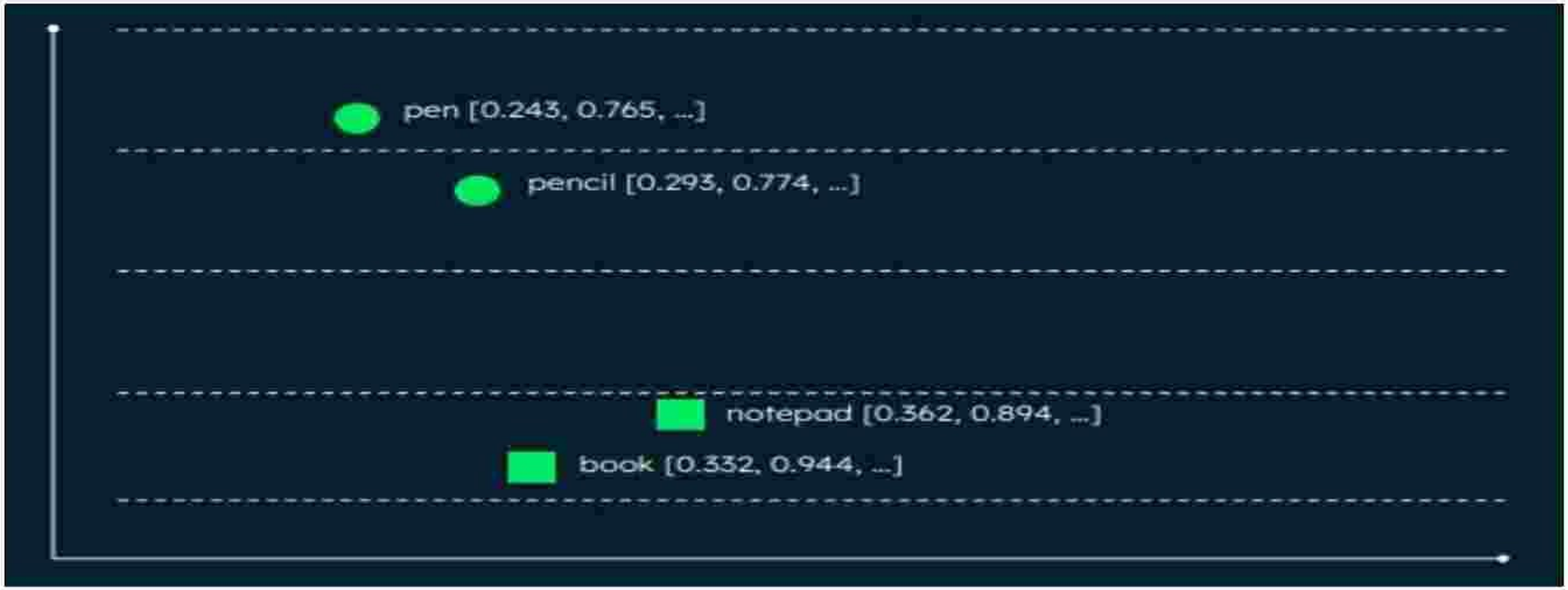
02 生成式AI催生向量数据库需求

相比其他类型数据库，向量数据库更擅长进行大数据量和多数据类型的快速检索

向量数据库优势			
	向量数据库	关系型数据库	非结构化数据库
数据存储	专为高维向量数据优化	结构化数据存储	灵活数据模型，支持非结构化和半结构化数据
查询效率	高效的相似性搜索和语义搜索	成熟的查询语言(SQL)，适合结构化查询	通常支持简单的查询，适合大数据分析
扩展性	通常支持水平扩展，适合大规模数据	有些支持水平扩展，但更多的是垂直扩展	高度的水平扩展能力，适合大数据环境
数据模型	通常为高维向量	严格的模式和完整性约束	灵活的数据模型，无需预定义模式
事务支持	一般不支持或支持有限	完全支持事务	通常不支持或支持有限
应用场景	推荐系统、图像/声音搜索、语义文本搜索	金融系统、企业资源计划、客户关系管理	社交媒体平台、大数据分析、物联网数据存储
系统成熟度	通常较新，但正在迅速发展	非常成熟，有多年的发展历史	通常较新，但在大数据和云计算领域有快速的发展

来源：互联网公开资料、弘则研究整理

所有数据格式均可以转换成高维向量，通过向量相似性比较进行快速检索



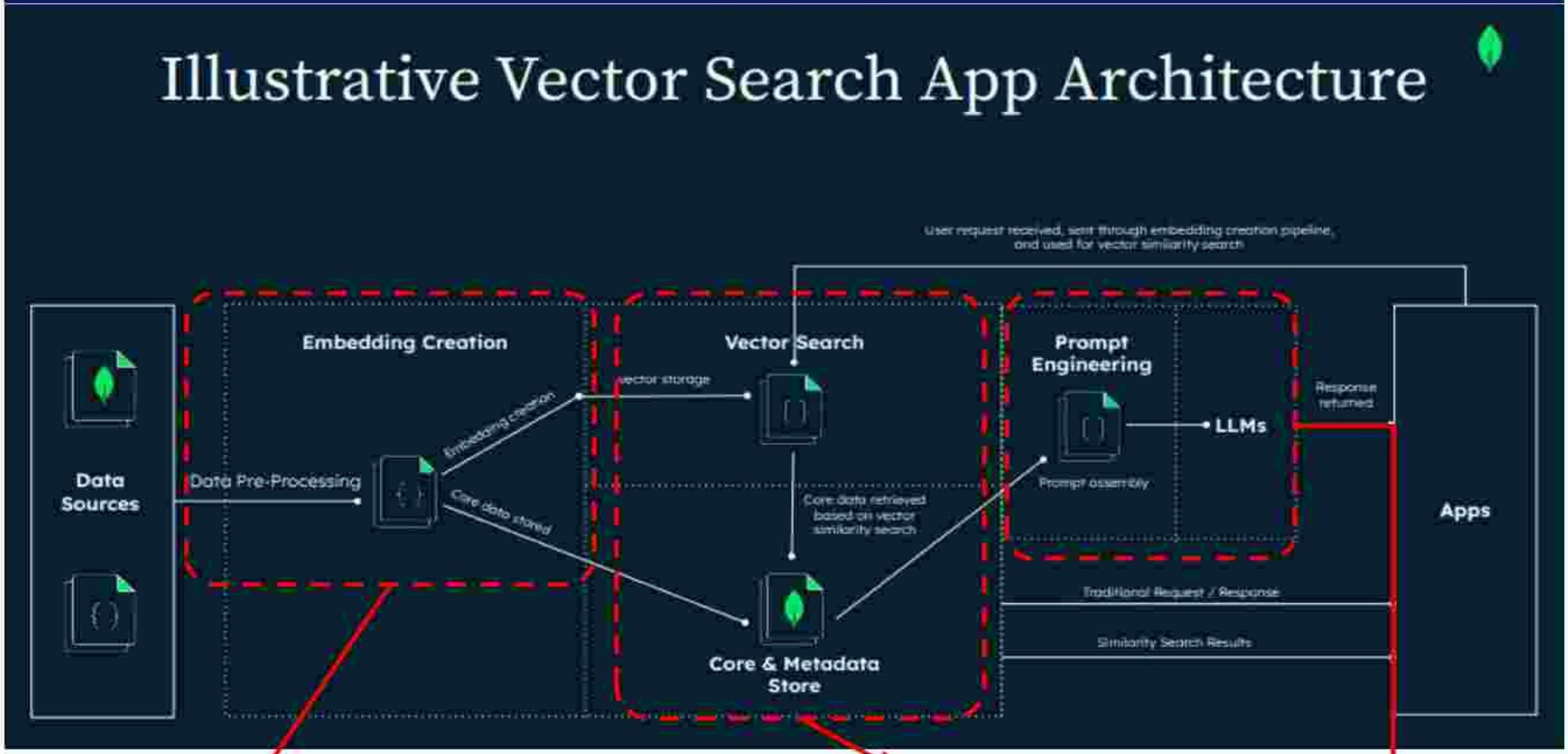
- 源数据可以是文本、代码、图片或视频等
- 向量数据是一种数学表达形式，它由一组有序的数值组成，这些数值可以表示空间中的一个点、一个方向或者一个速度等。在向量数据中，每个数值都有其特定的含义，例如在二维空间中，一个向量可以由两个数值表示，分别对应x轴和y轴的坐标；在三维空间中，一个向量可以由三个数值表示，分别对应x轴、y轴和z轴的坐标。比如人脸识别比对，图片要被转化成1000+维向量
- 通过计算两个向量之间的距离或夹角，我们可以得到这两个向量的相似性。这个特性在很多应用中都非常有用，例如在推荐系统中，我们可以通过计算用户的兴趣向量和商品的特征向量之间的相似性，来推荐用户可能感兴趣的商品。越相似的向量在空间中的位置会越相近
- 定性判断向量化数据量大小：视频 > 音频 ≧ 文档

语义搜索不仅是匹配关键字，而是试图理解真正意图，带来更准确、更有上下文的搜索结果

语义搜索vs传统分词搜索		
	语义搜索	传统分词搜索
核心技术	基于向量搜索，机器学习和人工智能	基于文本匹配和查询扩展
搜索目的	理解查询的深层意义和上下文	直接匹配关键词或扩展的词汇
处理上下文	能够根据搜索者的地理位置、搜索历史等信息调整结果	通常不考虑这些额外的上下文信息
搜索结果的相关性	根据查询的意图和上下文排名结果	主要基于关键词的频率和位置匹配
处理同义词和多义词	能够理解词语在不同上下文中的意义，并据此返回结果	通常使用同义词表或词汇扩展工具，可能不总是理解上下文中的真正意义
对查询的理解	能够区分如“chocolate milk”和“milk chocolate”这样的查询，即使关键词顺序或形式相同	可能只是简单地匹配关键词，而不理解它们的真正意思
学习和适应能力	通过机器学习不断改进，根据用户的反馈和行为适应	通常基于固定的算法和规则，没有持续学习和适应的能力
用户体验	提供更准确和有上下文的结果，从而提高用户满意度	依赖于用户精确输入，可能返回与用户实际意图不匹配的结果

向量数据库厂商主要提供向量化工具、向量和源数据的键值对存储和查询

向量数据库功能



向量化工具：将源数据向量化处理

- 键值对存储：同时存储向量和源数据
- 向量比对：向量相似性比对，可以使用不同索引类型

提示工程

来源：互联网公开资料、弘则研究整理

向量数据库技术原理开源通用，用户侧对于技术差异的感知并不明显，更多比拼生态社区、服务等软性能力

向量数据库竞争点		
	技术差异点	用户感知
向量化工具	• 向量化维度：向量化工具能把用户问题向量化到不同的细度，但并不是维度越多就说明向量化的能力越强，很可能会出现维度少的向量化工具更准 • 召回率：指的是有多少个不同形式、不同提问方法的问题可以返回同一个答案，代表的是语义向量的准度。目前GPT的textada应该是召回率最高的，能做到60%以上，一般开源的word2vector模型只能做到30%-40%	• 向量化工具是可插拔组装的，比如用Milvus的数据库和GPT的textada向量化工具也是可以的 • 从用户感受上，细化向量化工具的维度在边际上的感受是在递减的，语义理解的准确性做到98%还是99%基本在使用上没什么区别
向量比对	• 是否支持一些查询的索引类型，比如欧式距离、邻近算法、平铺查询等 • 查询速度和并发能力	• 技术原理基本通用，现在门槛没那么高

- 向量数据库所提供的向量化工具、向量比对能力是不具备技术层面硬壁垒的，技术原理基本通用且开源，比如向量化工具基本可以在GitHub上下载源代码
- 国内厂商从团队启动，人员规模不过百，可以平移NoSQL团队过来，基本三个月就可以出产品
- **厂商核心差异在于开源生态社区：**海外开源的起家都在于开源社区绑定了大量的程序员。像MongoDB、MySQL这些厂商，初级开发者、刚毕业的大学生这类群体，基本都比较通用，在大学课程里就包含这些厂商的知识。大量开发者会支持后续的产品迭代。MongoDB在传统NoSQL领域很强，没有人不知道，没有人不会用

Hugging Face榜单侧重技术指标，但尚未发现某一向量化工具在所有测试任务中占主导地位；DB-Engine则侧重品牌认知的排名

Hugging Face向量化工具榜单

Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (56 datasets)
1	bge-large-en-v1.5	1.34	1024	512	64.23
2	bge-base-en-v1.5	0.44	768	512	63.55
3	gte-large	0.67	1024	512	63.13
4	gte-base	0.22	768	512	62.39
5	e5-large-v2	1.34	1024	512	62.25
6	bge-small-en-v1.5	0.13	384	512	62.17
7	instructor-xl	4.96	768	512	61.79
8	instructor-large	1.34	768	512	61.59
10	multilingual-e5-large	2.24	1024	514	61.5
9	e5-base-v2	0.44	768	512	61.5
11	e5-large	1.34	1024	512	61.42
12	gte-small	0.07	384	512	61.36
13	text-embedding-ada-002		1536	8191	60.99
14	e5-base	0.44	768	512	60.44
58	e5-small-v2	0.13	384	512	59.93

DB-Engine向量数据库榜单

☑ Include secondary database models　　20 systems in ranking, September 2023

Rank Sep 2023	Rank Aug 2023	Rank Sep 2022	DBMS	Database Model	Score Sep 2023	Score Aug 2023	Score Sep 2022
1.	1.	1.	PostgreSQL	Relational, Multi-model	620.75	+0.37	+0.29
2.	2.	2.	MongoDB	Document, Multi-model	439.42	+4.91	-50.21
3.	3.	3.	Redis	Key-value, Multi-model	163.68	+0.72	-17.79
4.	4.	4.	Elasticsearch	Search engine, Multi-model	138.98	-0.94	-12.46
5.	5.	5.	Cassandra	Wide column, Multi-model	110.06	+2.67	-9.06
6.	6.	6.	OpenSearch	Search engine, Multi-model	12.61	+0.17	+3.80
7.	7.	↑ 8.	Kdb	Multi-model	8.94	+0.52	+0.91
8.	8.	↓ 7.	Datastax Enterprise	Wide column, Multi-model	6.69	-0.34	-1.48
9.	9.	9.	SingleStore	Relational, Multi-model	5.67	-0.37	-1.59
10.	10.		Pinecone	Vector DBMS	3.01	+0.40	
11.	11.		Chroma	Vector DBMS	2.30	-0.15	
12.	↑ 13.	↓ 11.	Weaviate	Vector DBMS	1.61	+0.26	+1.45
13.	↓ 12.	↓ 10.	Milvus	Vector DBMS	1.38	-0.03	+0.92
14.	14.		Vald	Vector DBMS	0.86	-0.03	
15.	15.	↓ 12.	Qdrant	Vector DBMS	0.83	+0.19	+0.71
16.	↑ 17.		Vespa	Multi-model	0.46	-0.09	
17.	↓ 16.		Deep Lake	Vector DBMS	0.42	-0.18	
18.	18.		MyScale	Multi-model	0.23	+0.09	
19.	19.	↓ 13.	JaguarDB	Multi-model	0.02	+0.00	-0.05
20.			Transwarp Hippo	Vector DBMS	0.00		

基于康奈尔大学的论文《MTEB: Massive Text Embedding Benchmark》作为评判标准：文本嵌入通常在来自单个任务的一小组数据集上进行评估，而不涵盖它们在其他任务中的可能应用，这使得该领域的进展难以追踪。为了解决这个问题，我们引入了大规模文本嵌入基准（MTEB）。MTEB 涵盖 8 个嵌入任务，涵盖总共 58 个数据集和 112 种语言。通过对 MTEB 上 33 个模型的基准测试，我们建立了迄今为止最全面的文本嵌入基准。

DB-Engine的分数计算方法

- 网站上提及的次数：以Google和Bing搜索引擎查询中的结果数来衡量
- 对系统的兴趣度：Google Trends中的搜索频率
- 相关技术讨论频率：IT问答网站Stack Overflow和DBA Stack Exchange上的相关问题数量和感兴趣用户数
- 提及的工作机会数量：职位搜索引擎Indeed和Simply Hired上的职位数量
- 职业网站上提及的简介经历数量：使用LinkedIn来衡量
- 社交网络中的相关性：统计了Twitter推文的数量

生成式AI的出现驱动向量数据库发展，软件应用均需要借助向量数据库进行相似性搜索，进而生成更精准回答

应用在执行指令时必须调用“软件说明书”或企业内部数据，目前有两种技术路径

- 大模型只起到“翻译”的功能，将自然语言翻译成软件应用能听懂的机器语言
- 第一种方式是**“微调”（fine-tuning）**，相当于给大模型准备一本纸质教材，但每次做教材改版的时候都需要重新训练和学习
- 第二种方式是**“嵌入”（embedding）**，相当于把教材做成活页，在客户本地部署一个数据向量库，可以随时对数据进行调整

来源：互联网公开资料、弘则研究整理

大模型的应用场景中，无论C/B端，只要涉及到个性化、专业化场景，均需要应用到向量数据库

	C端	B端
训练	训练不需要对原数据集进行保存，形成的知识会以参数文件的形式进行存储，想要调用大模型可通过一段Python代码读取参数文件即可	企业内部应用多采用嵌入而非微调的方式以节省成本，内部知识数据会存储在向量数据库中，供通用/行业大模型进行调用以与企业用户交互
推理	不管C/B端推理场景，多轮对话场景必须要用到向量数据库以保存对话内容，在未来重新开启对话时才会有“记忆”。数据库用量会随着对话数据量同向增长	

随着生成式AI将大规模落地，作为刚需配套的向量数据库赛道将迎来加速，目前国内外市场均处在较早期阶段，市场竞争格局极分散

- **海外市场：**海外市场存在“云中立”产业逻辑，故海外云厂商不会降维进入该领域。目前海外较为知名的厂商有Milvus、Pinecone（为ChatGPT提供向量数据库）等；传统非结构化数据库厂商如MongoDB也在今年6月发布向量搜索等功能，目前处在市场宣传阶段
- **中国市场：**一级市场有较多厂商在转型做向量数据库，但国内并无“云中立”的产业逻辑，因此国内云大厂如阿里、腾讯、华为等均具备向量数据库产品以补全自身一揽子打包的云服务解决方案

来源：产业调研、互联网公开资料、弘则研究整理

C端训练场景中，不需要对原数据集进行保存，形成的知识会以参数文件的形式进行存储，因此不需要向量数据库

大模型训练

提供教材，喂数据　　机器学习、深度学习　　形成知识储存在神经网络

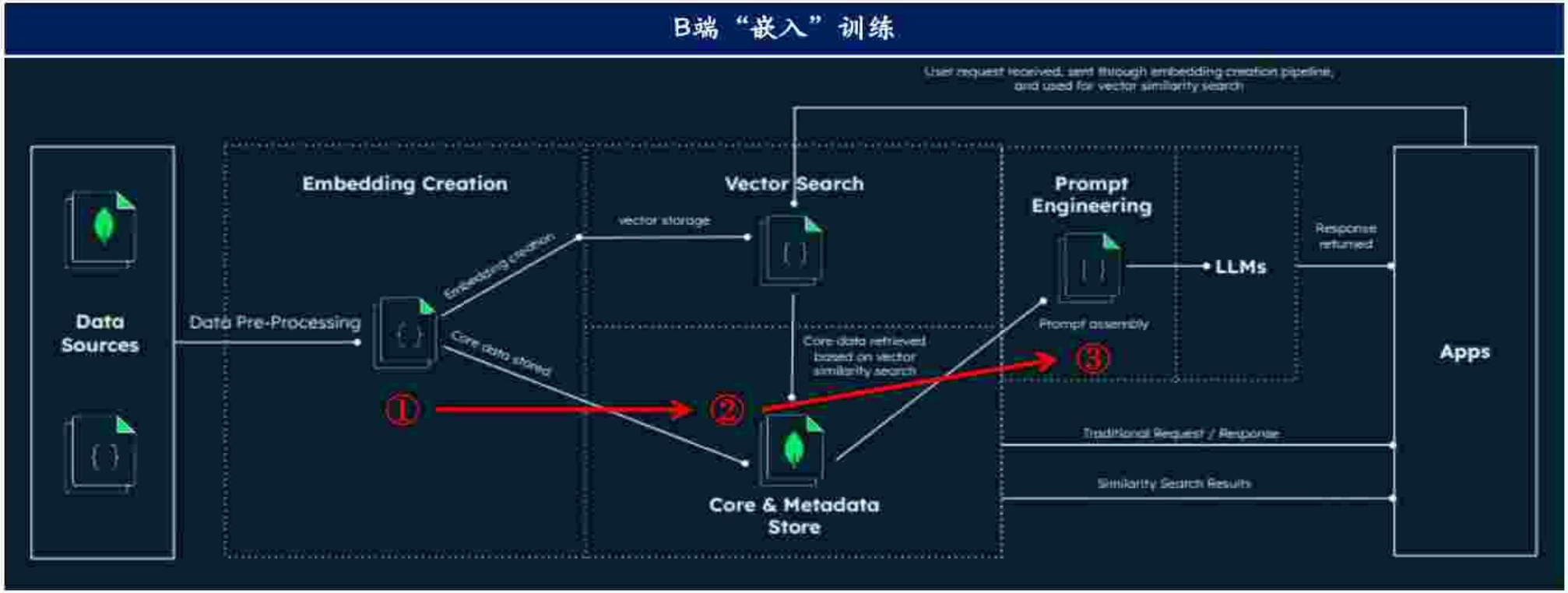
大模型本身并没有数据库，所有知识被拟合成高维公式，以参数文件的形式存储，调用大模型本身是一段Python代码读取参数文件

训练大模型时，不断把知识放进去增加模型的维度和参数，然后拟合成一个高维模型。当公式被拟合出来之后，原始信息被合成了公式的参数，以文件夹形式存储，不保存原始数据。应用大模型时用Python程序读取文件夹里的参数

垂类大模型做微调相当于重新训练一遍大模型，所以同样不需要向量数据库

强调泛化能力的场景，或者增加的样本数据量很大的时候会选择微调的方式，但不适合需要高频追加数据的场景。一般来说，增加的样本数据量（比如某些行业大模型增加垂直行业的知识）超过通用大模型原有训练数据集的2%以上，才会愿意用微调的方式，增加的数据太少对通用大模型造不成影响。微调也有两种方式，可以选择全参数调整也可以选择冻结一部分参数进行调整。这种方式都不涉及向量数据库

B端训练场景中，企业内部知识数据会存储在向量数据库中，用来提升大模型的回答精确度



“嵌入”而非“微调”的原因：

1）“微调”不适用企业内部场景：企业内部知识的样本量对于通用大模型的样本量来说量级太小了，不足以对大模型造成影响，fine-tuning之后可能还是查询不到想要的答案

2）限制大模型泛化能力来提升问答准确度：比如银行智能客服场景，开卡、提额度等功能，一定是答案唯一的。当客户提出问题，优先去向量数据库搜索唯一的答案，然后通过大模型总结回答给客户。

① 将各种源数据（文本、代码、图片或视频等）转成向量数据（市面上有很多向量化处理工具）

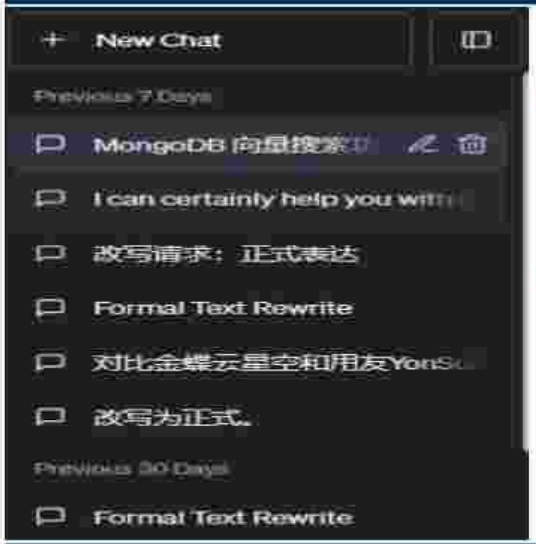
② “嵌入”向量数据并存储源数据，向量数据和源数据类似键值对，一一对应存储

③ 构建对应的提示工程

来源：产业调研、互联网公开资料、弘则研究整理

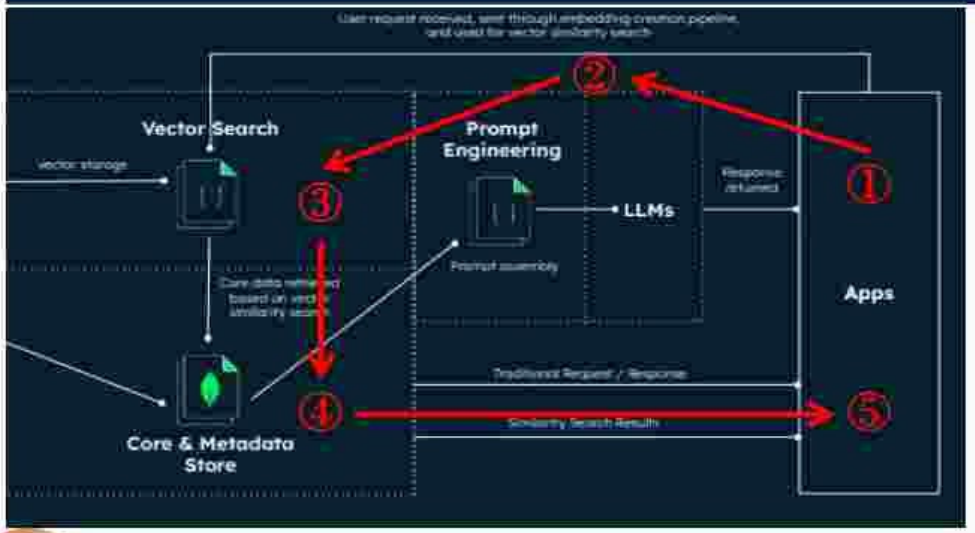
C/B端推理场景中，多轮对话场景必须要用到向量数据库，且数据库用量会随着对话数据量同向增长

C端推理场景



- 只要是问答类型的大模型，不论底层是谁的模型，微调还是嵌入的方式做的训练，基本都需要向量数据库来存储每一个用户的上下文回答，以便能让大模型越来越懂每一个用户。因为每一个用户长期的提问数据是不能内化到大语言模型里的，所以肯定需要一个地方去存储这些数据，这个就是通过向量数据库来解决的
- C端场景可能会更强调对于多轮对话的长期存储，数据量增长会推动向量数据库需求增长。相当于每个C端用户在通过多轮对话的方式去训练专属于自己的个人助理，可以了解健身数据、饮食数据等个人习惯和偏好，这样大模型可以给更精准的推荐。所以C端场景会更希望多轮对话的信息可以长期存储，这些数据量的增长会推动向量数据库需求的增长

B端推理场景



① 在应用端进行自然语言提问（LangChain技术框架会做规则判断和逻辑编排，判定是否需要调用向量数据库进行回答）
② 若不需要，则直接由大模型回答或进行互联网搜索回答；若需要用到内部知识，则向量化工具将提问向量化
③ 在向量数据库中进行向量相似性搜索
④ 找到对应的内部知识源数据作为论据支撑
⑤ 反馈到大模型进行生成式回答

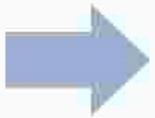
来源：产业调研、互联网公开资料、弘则研究整理

关系型数据库市场规模大几百亿美元；受到生成式AI驱动的向量数据库应用场景更多，潜在市场空间将超过关系型数据库

向量数据库市场规模

关系型数据库

据信通院统计数据，全球数据库市场规模在2020年为671亿美元，到2025年有望达到798亿美元，CAGR 3.5%，Oracle、MySQL、SQL Server等都是关系型数据库，估算关系型数据库全球龙头Oracle收入规模小几百亿美元



非关系型数据库

- 键值数据库
- 宽表数据库
- 文档数据库
- 图数据库
- 内存数据库
- 时序数据库
- 搜索数据库
- 向量数据库
- ……

- 关系型数据库遵循ACID规则。主要集中在强一致性场景，比如银行交易、零售电商、车票预订等
- 非关系型数据库放宽或取消了一些ACID的规则以达到更好的性能和更大的灵活性，扩展性和并发读写性能更高，更适合互联网应用的场景，比如Facebook、微博等。因此非关系型数据库的应用场景更广阔，数据量更大，理应具备更大的市场空间
- 实际商业化角度是，非关系型数据库的应用场景（日志、互联网内容管理、实时数据分析、移动应用等）相比于关系型数据库的应用场景（核心业务数据完整性、核心业务系统等），数据的商业化价值更低，因此企业更倾向于开源免费的非关系型数据库，导致非关系型数据库的商业化困难
- 生成式AI的出现带来了数据价值的深度挖掘，企业应用于AI应用的数据可能会带来更大的潜在价值，因此向量数据库的应用场景和商业化潜力将是非关系型数据库中最高的

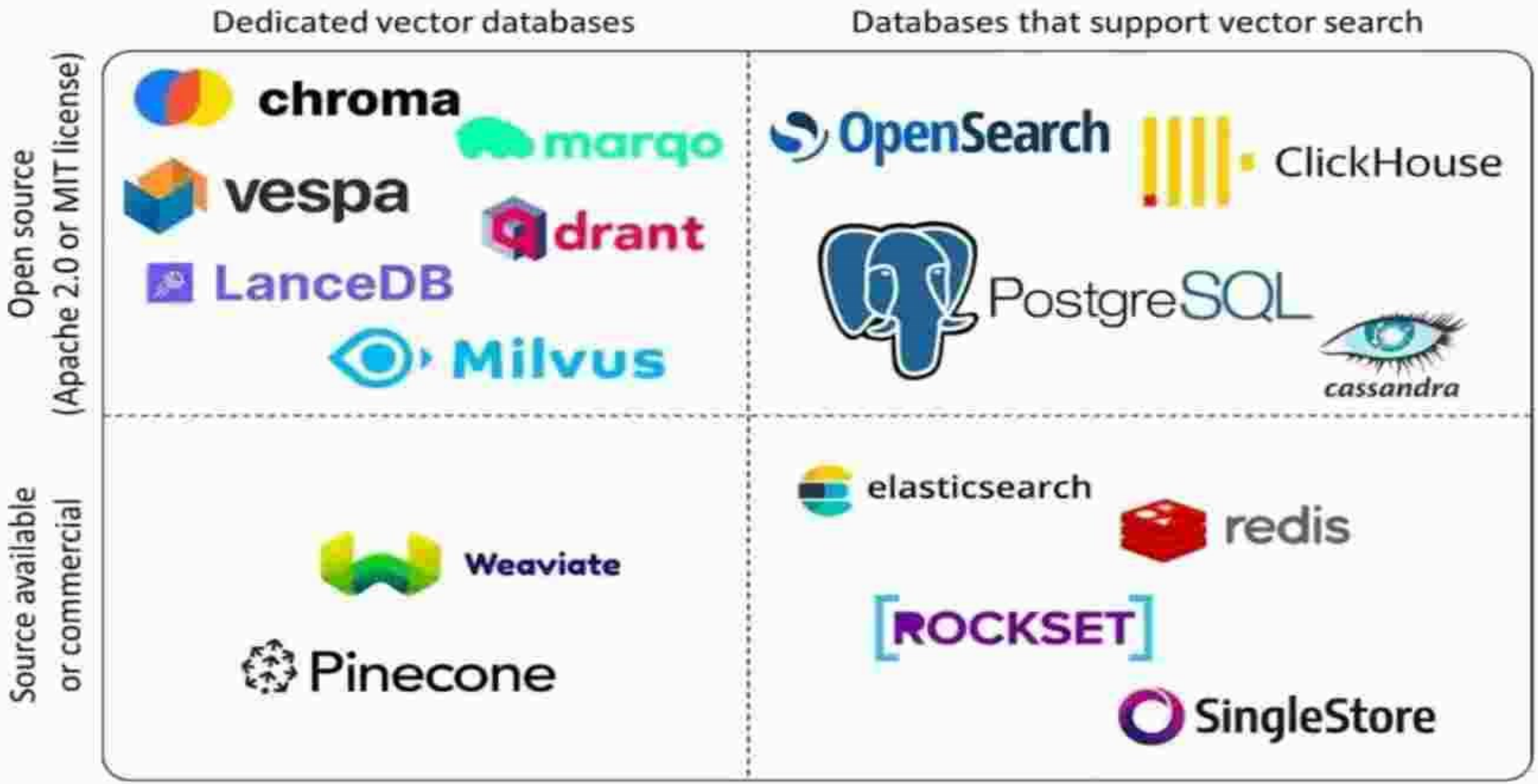
来源：产业调研、互联网公开资料、弘则研究整理

Horizon Insights
弘 则 研 究

03 各厂商具备错位竞争优势

海外市场不同背景厂商以不同商业模式切入向量数据库赛道

海外市场向量数据库厂商图谱



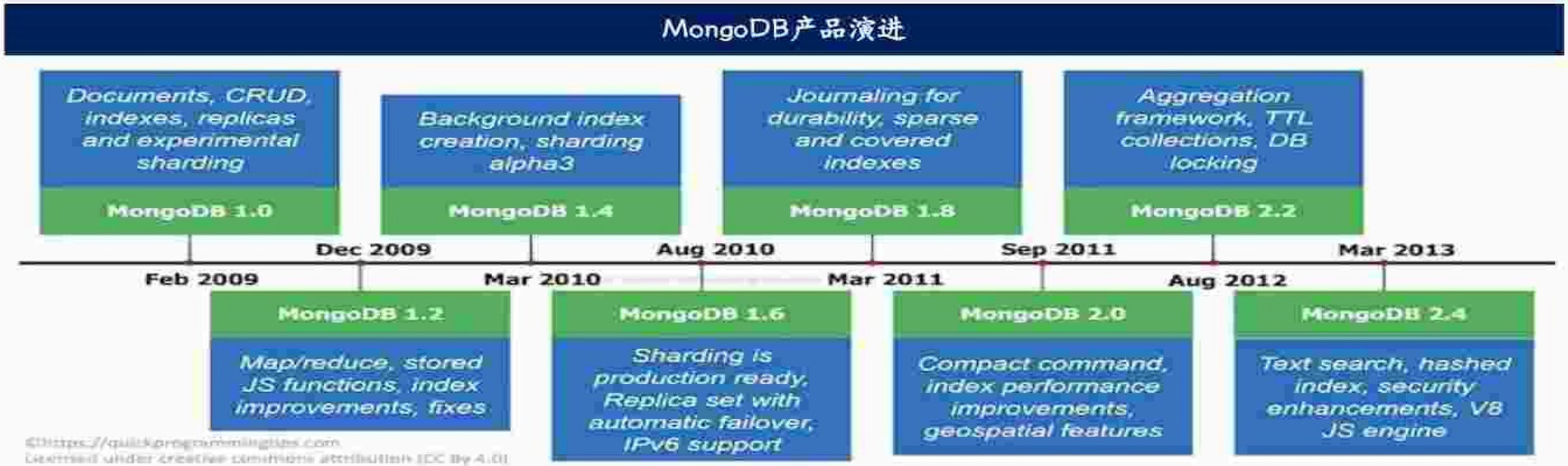
中国市场云厂商将向量数据库作为云服务矩阵中的一个SKU提供给客户以提升全面服务能力

来源：36氪、互联网公开资料、弘则研究整理

MongoDB顺应NoSQL和大数据需求诞生，培养生态社区逐步实现技术支持+云托管模式的商业化

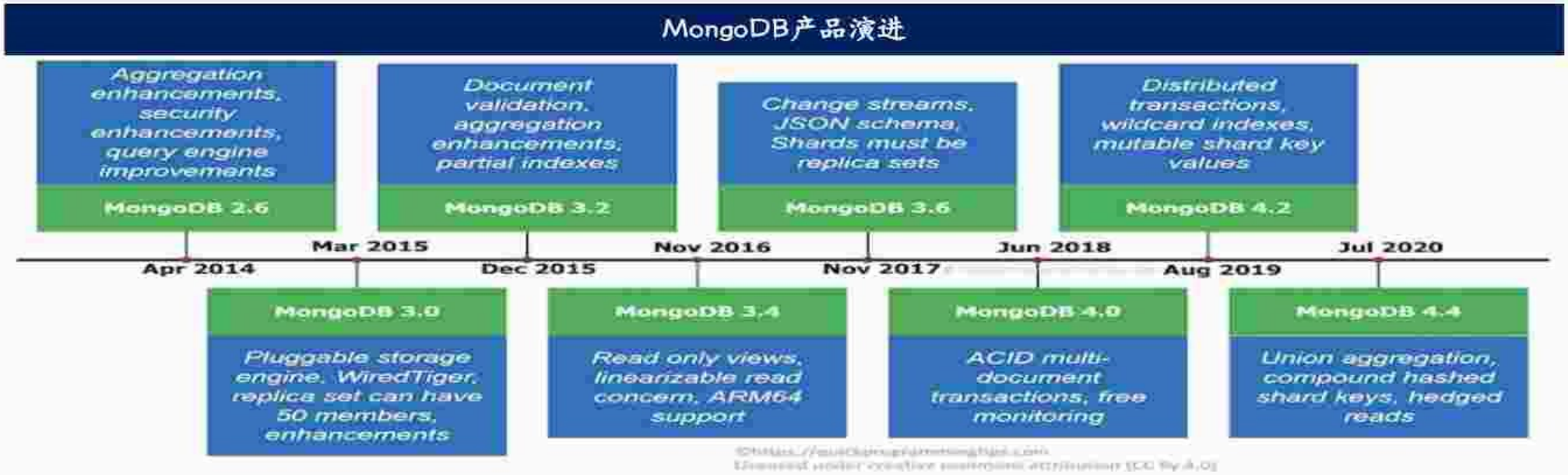
- MongoDB有别于当时其他的数据库产品，使用和安装都非常方便，在代码中通过API就可以操作数据，在当时引起不小的轰动
- 10gen一直通过开源社区和MongoDB大学扩大影响力，吸引程序员入驻社区，在社区按照不同地区成立不同的用户组，不同的用户组每年都会举办一次MongoDB大会。知名科技博客Business Insider上将MongoDB宣传成程序员必备技能之一，掌握好这门技术，不愁找不到工作。同时还和很多在线教育网站合作开展MongoDB的培训课程，从2012年起开始提供付费技术支持
- 2016年推出Atlas服务，开始和公有云厂商合作，提供云托管服务

互联网内容服务推动大型数据库应用，MongoDB早期版本提供大规模数据处理、分片集群（水平扩展能力）等功能



- 创始人团队首次创业时创立了著名的在线广告公司DoubleClick，几年之内广告流量达到了每秒40万条。当时成熟的数据库基本上都是基于单机架构的传统关系型数据库如Oracle、MS SQLServer等。即便Oracle支持一些集群部署，其扩展性也仅限于2-4台服务器的范围
- 在关系型数据库中，当数据量达到一定程度，单个节点服务器资源充分饱和无法保证及时的服务响应时间时，通常会采用分区分表的数据库优化方案。但是这些方案都是侵入式的，很多时候意味着应用程序需要做较大的改动，来配合数据库端的改动
- 2010年推出具有分片集群的1.6版本，在水平伸缩能力上要强于传统关系型数据库。MongoDB 的自动分片，可以在一个集群的几个分片服务器内自动进行数据的分布和均衡。在尽可能把数据均匀的分布到多个存储节点的同时，为应用开发者提供无缝的体验。开发者无须关心数据的具体位置，程序也不需要因为分片与否而进行修改

MongoDB补足功能以适应全业务场景，同时简单易用



- 3.2版本中增加了操作符: $lookup，意味着作为NoSQL数据库，MongoDB开始支持关系型数据库的核心功能：关联。从3.2开始，可以一次同时查询多个MongoDB的集合（表），不用像以前那样，如果有多表查询需要在代码中发起多个数据库查询，然后在内存中进行手工关联
- 2018年推出4.0版本具备多文档ACID强事务机制，之前MongoDB对事务的支持仅限于单文档内。无法保证原子性和出错回滚机制，很多交易性的业务会有意避开MongoDB。而随着4.0的发布，MongoDB可以用来支撑几乎所有的业务场景
- 2020年推出的4.4版本解压缩之后仅3个可执行文件（总大小约为 150MB，对于任意类型的 MongoDB 部署，都只需要这几个组件）：1）mongo：MongoDB Shell，使用基于 JavaScript 的命令与服务器发生交互；2）mongod：运行 MongoDB 的主文件，可以作为单个数据库实例、分片集群的成员或分片集群的配置服务器运行；3）mongos：一个路由器应用程序，用在具有水平伸缩能力的数据库服务器集群中

收购补全MongoDB能力并加深和开发者社群关系

MongoDB历次收购			
时间	标的	主要业务	收购意义
2014.12	WiredTiger	存储引擎	WiredTiger作为一个现代、高性能、高吞吐量的存储引擎，极大地提高了MongoDB在高写入量工作负载下的性能。WiredTiger还为MongoDB带来了压缩、记录级锁定、多版本并发控制（MVCC）、多文档事务以及对非常高插入工作负载的日志结构合并树（LSM trees）的支持
2018.10	mLab	DBaaS	mLab目前在其平台上拥有大约100万个托管数据库，包括免费和付费层。这次收购将加深MongoDB与以开发者为中心的初创公司社群的关系，并有助于MongoDB Atlas的快速扩张
2019.04	Realm	云计算移动数据库	这次收购加强了MongoDB与专注于移动和无服务器开发的开发者社群的关系。Realm拥有超过10万名活跃开发者，其解决方案已被下载超过20亿次。这次收购与MongoDB全球云数据库Atlas以及无服务器平台Stitch非常契合

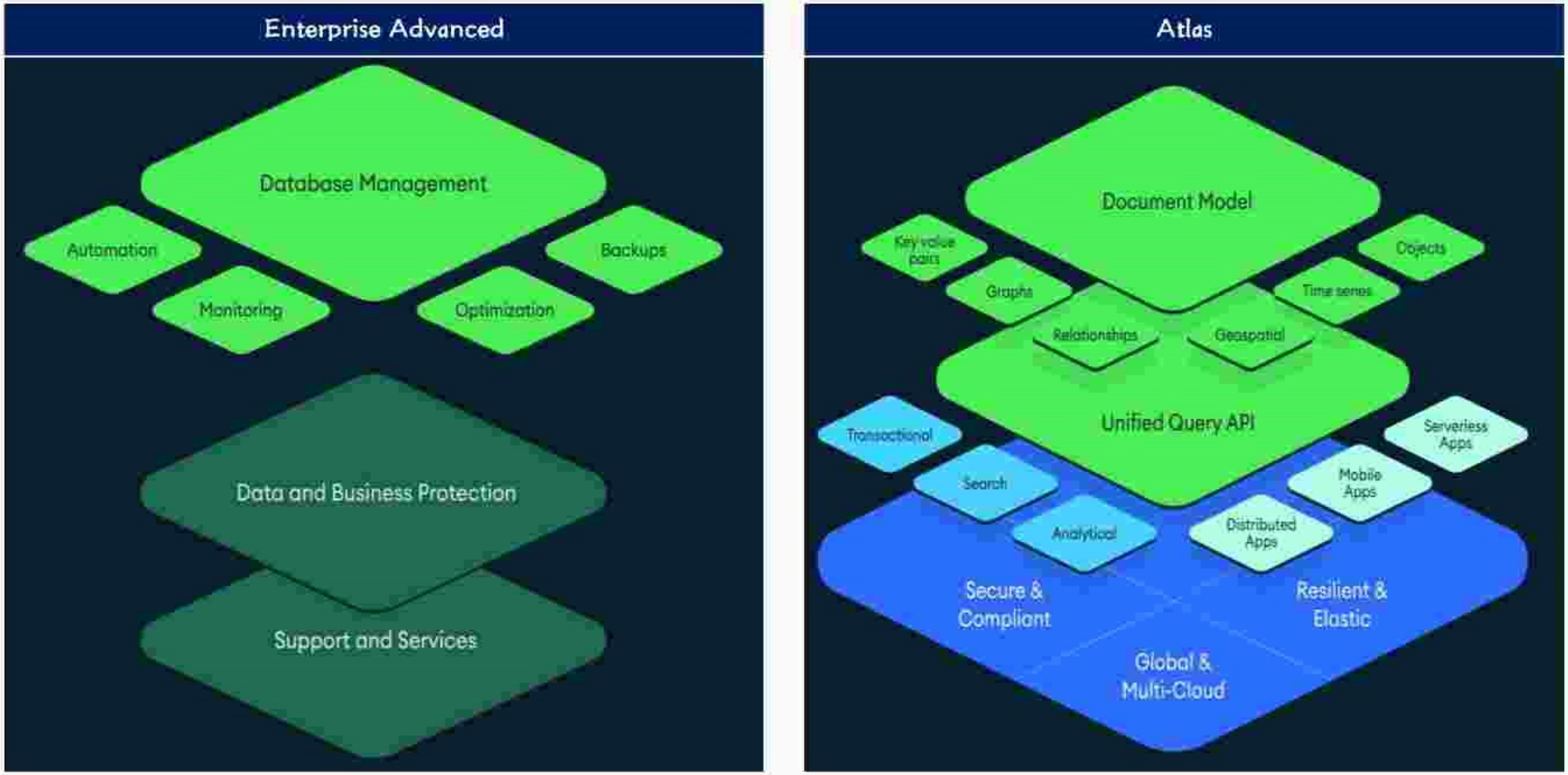
来源：互联网公开资料、弘则研究整理

MongoDB提供Enterprise Advanced、Atlas和Community版本

MongoDB产品版本		
版本	功能	详情
Enterprise Advanced	企业数据库服务器	存储、组织和处理数据，并方便对数据的访问和更改。包括高级安全功能、审计功能、企业标准的认证和授权、加密和内存存储引擎
	企业管理能力	Cloud Manager Premium和Ops Manager管理工具，允许运营团队运行、管理和配置MongoDB，包括能够对大约100个系统指标进行监控和报警，备份数据并将其恢复到任何时间点以进行灾难恢复，以及自动执行常见的操作任务，如升级、扩展和配置更改
	分析集成	提供集成，允许数据和业务分析师使用其现有的商业智能和分析工具分析其平台上运行的应用程序中的数据。对于与Tableau等商业智能产品的集成，分析师可以使用其MongoDB Connector for BI产品，其中包括其最新发布的ODBC驱动程序，以支持与Microsoft Excel的连接。该公司还为Spark和Hadoop提供开源连接器，这些产品经常用于数据分析
	技术支持	通过企业级服务水平协议为客户提供技术支持
Atlas	云托管数据库	提供了一个弹性的、可管理的产品，包括自动配置和愈合、全面的系统监控、可管理的备份和恢复、默认安全等功能。MongoDB Atlas让客户从管理数据库和相关底层基础设施的复杂性中解脱出来，从而可以专注于应用和终端用户体验，并创新服务自己的客户，把握新的商业机会
社区版	开源免费版	包含开发人员使用 MongoDB 所需的核心功能。使用MongoDB Atlas直接从社区版获得收入，并通过向上销售用户到其企业高级订阅包间接获得收入

MongoDB可以适应不同行业的不同用例

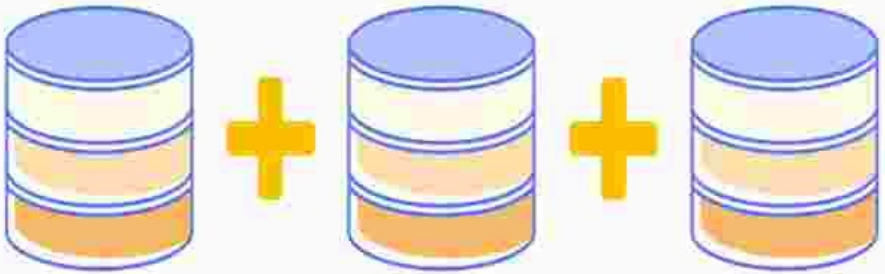
MongoDB在2016年推出Atlas向云数据库转型



- MongoDB Enterprise Advanced：可以在云端、内部部署或混合环境中运行。它提供了专有的商业数据库服务器和企业管理功能，使用户可以完全掌控自管理的MongoDB环境的管理和安全性
- MongoDB Atlas：是一个多云开发者数据平台，主要为用户提供云端的数据库服务，使得用户可以更加方便地使用和管理MongoDB

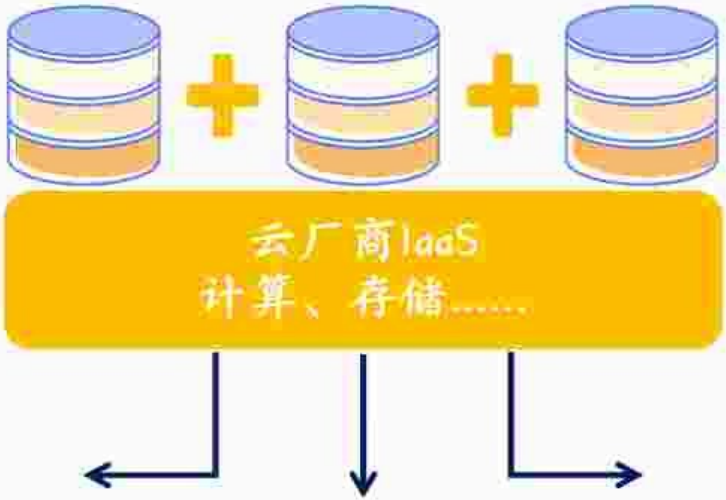
商业模式仍为订阅

Enterprise Advanced：license订阅+技术支持模式



- 报价按照服务器硬件性能报价（几核CPU）
- 客户可以选择其Cloud Manager Premium产品（适用于希望通过云端管理其平台的客户）或Ops Manager（适用于内部部署的客户）
- 订阅期间为客户提供技术支持

Atlas：云托管模式



- 用户可以在云上采购数据库厂商服务，或直接采购数据库厂商服务
- 价格包含云基础设施费用，数据库厂商再与云厂商结算

Dedicated Cluster

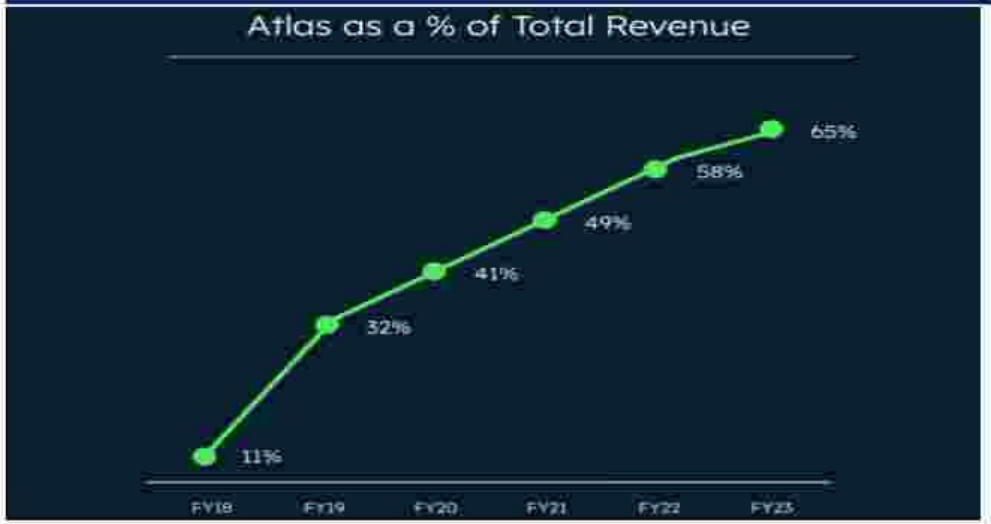
Pay-as-you-go! Clusters are billed hourly with monthly invoices.

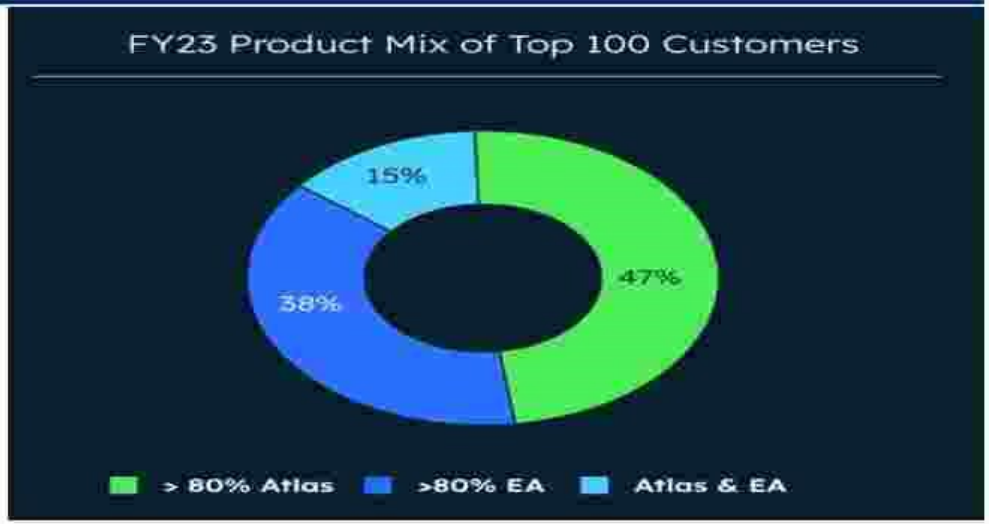
aws Amazon Web Services | Microsoft Azure | Google Cloud Platform

Cluster Tier	Storage	RAM	vCPUs	Base Price
M10	10 GB	2 GB	2 vCPUs	$0.08/hr
M20	20 GB	4 GB	2 vCPUs	$0.20/hr

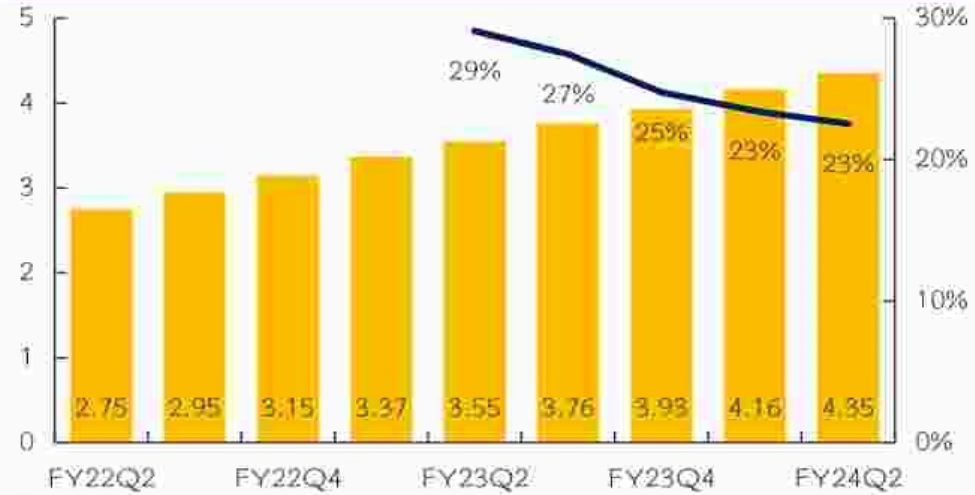
Atlas客户数占总客户数90%以上，收入占比快速提升至65%，跨多云数据库优能力显现

Atlas收入占比





Atlas客户数（万家）



- Atlas从2016年底推出，收入占比快速提升
- 头部百大客户中，47%的企业客户中，8成MongoDB的采购是Atlas，验证海外企业“多云策略”，Atlas的跨云应用管理能力显现
- 超90%客户均有采购Atlas

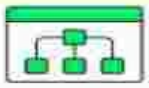
来源：互联网公开资料、弘则研究整理

客户将工作负载从传统关系型数据库迁移至MongoDB，带来客户数和单客户ARR共同增长

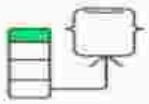
关系型数据库迁移难点

Update schema	Rewrite code	Migrate data
Determine how the existing relational schema should be best represented in MongoDB document model.	Update or rewrite application code to support new user requirements, modern tech stack and updated schema.	Perform one-time or continuous replication of data, allowing cutover from the legacy to the modernized app.

MongoDB迁移工具



Design an effective MongoDB schema, derived from an existing relational schema.



Migrate data from Oracle, SQL Server, MySQL, PostgreSQL, and Sybase ASE to MongoDB, while transforming to the target schema.

Generate code artifacts to reduce the time required to update application code.

- MongoDB自身提供专业的咨询和咨询服务以帮助客户迁移
- MongoDB大学提供相关迁移工具的课程
- 埃森哲、Infosys等生态伙伴同样可以帮助客户搭建新的数据库体系

来源：互联网公开资料、弘则研究整理

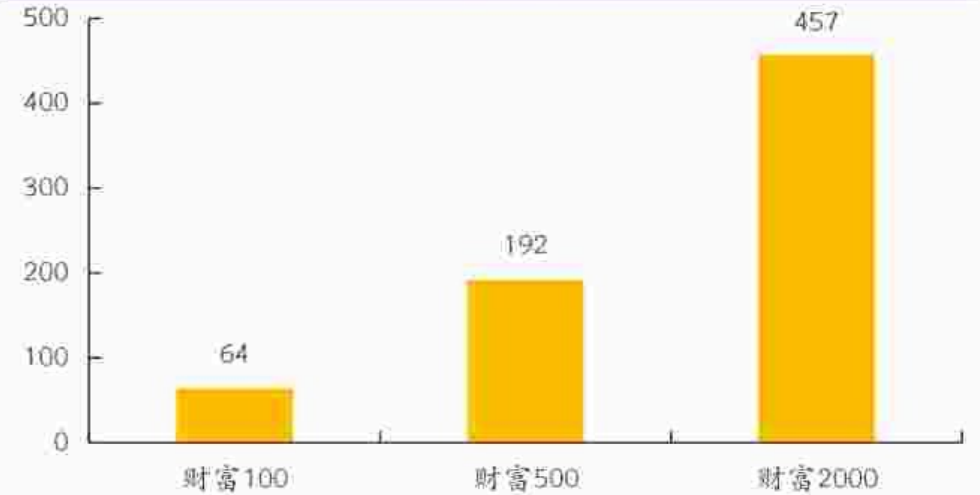
MongoDB持续推动客户数增长，成长空间巨大

MongoDB用户数（万家）

直销收入占比

销售人员数量（人）

头部客户覆盖数（家）



来源：互联网公开资料、弘则研究整理

MongoDB引导客户将更多负载迁移，以此推动单客户ARR提升

ARR>$1m客户数（家）

ARR>$100k客户数（家）

MongoDB客户三年ARR扩张情况

- MongoDB从文档数据库起家，补充其他工作负载类型，比如搜索、时序等，包括在今年6月推出的向量搜索。同时，2016年推出Atlas，为客户提供更全面简单的云托管服务。以此引导客户将更多的工作负载迁移至MongoDB Atlas之上，推动ARR增长，平均ARR三年扩张一倍
- 现阶段主要负载来源仍是传统关系型数据库
- 在财富100强和500强的数据库IT投入中，MongoDB目前仅占1.8%和1.7%
- Net ARR Expansion Rate高于120%

来源：互联网公开资料、弘则研究整理

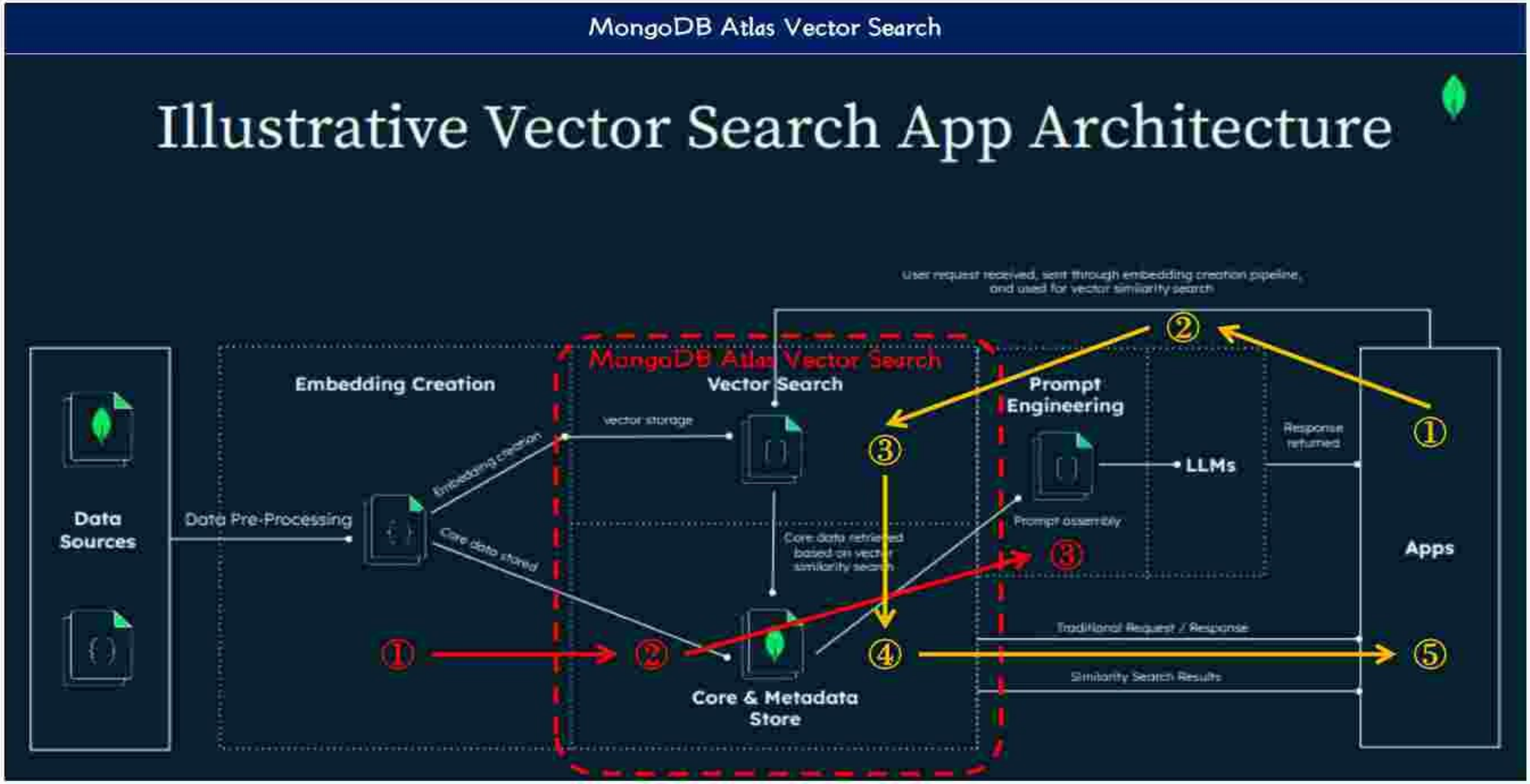
生成式AI的出现在需求侧和供给侧共同给MongoDB带来正向的增长驱动力

生成式AI对于需求侧的影响		
	C端	B端
训练	训练不需要对原数据集进行保存，形成的知识会以参数文件的形式进行存储，想要调用大模型可通过一段Python代码读取参数文件即可	企业内部应用多采用嵌入而非微调的方式以节省成本，内部知识数据会存储在向量数据库中，供通用/行业大模型进行调用以与企业用户交互
推理	不管C/B端推理场景，多轮对话场景必须要用到向量数据库以保存对话内容，在未来重新开启对话时才会有“记忆”。数据库用量会随着对话数据量同向增长	

生成式AI对于迁移工具的帮助		
迁移工具功能	具体能力	进展
SQL查询转换	• 从连接的关系数据库中导入存储过程和嵌入式SQL查询 • 迁移工具使用生成式AI将这些转换为MongoDB查询 • 根据在迁移工具中设计的架构创建MongoDB查询	开发中
AI重写代码	评估：搜索并理解代码库，以了解重构应用程序所涉及的努力和风险 代码转换：建议应用程序代码与应用程序架构建议一同使用，以最大限度地利用MongoDB 测试：验证转换后的应用程序在MongoDB上的表现是否符合预期	未来提供

来源：互联网公开资料、弘则研究整理

MongoDB在今年6月推出向量相似性搜索功能Atlas Vector Search



来源：互联网公开资料、弘则研究整理

MongoDB在今年6月推出向量相似性搜索功能Atlas Vector Search

部署和使用流程

红色为部署流程

①将各种源数据（文本、代码、图片或视频等）转成向量数据（市面上有很多向量化处理工具）

②“嵌入”向量数据并存储源数据，向量数据和源数据类似键值对，一一对应存储

③构建对应的提示工程

黄色为使用流程

①在应用端进行自然语言提问（LangChain技术框架会做规则判断和逻辑编排，判定是否需要调用向量数据库进行回答）

②若不需要，则直接由大模型回答或进行互联网搜索回答；若需要用到内部知识，则向量化工具将提问向量化

③在向量数据库中进行向量相似性搜索

④找到对应的内部知识源数据作为论据支撑

⑤反馈到大模型进行生成式回答

具体应用场景

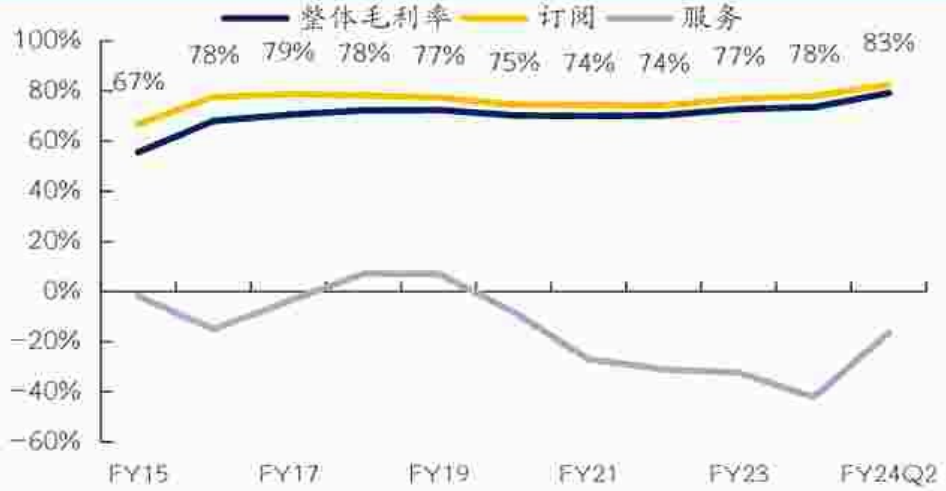
- 索引文本/图像/声音/视频、通过专有的增量数据增强基础LLMs并减少幻觉、问答系统、改进的推荐和相关性评分、动态个性化、对话式支持、同义词生成等

来源：互联网公开资料、弘则研究整理

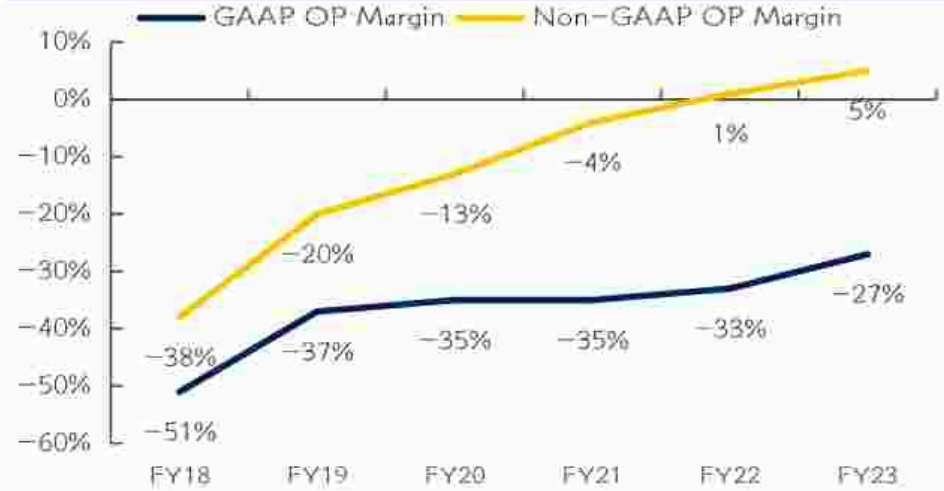
MongoDB全订阅收入，营收增速中枢在40%，经营性现金流比例改善明显，公司指引长期Non-GAAP OP Margin 20%+

MongoDB订阅业务营收（亿美元）

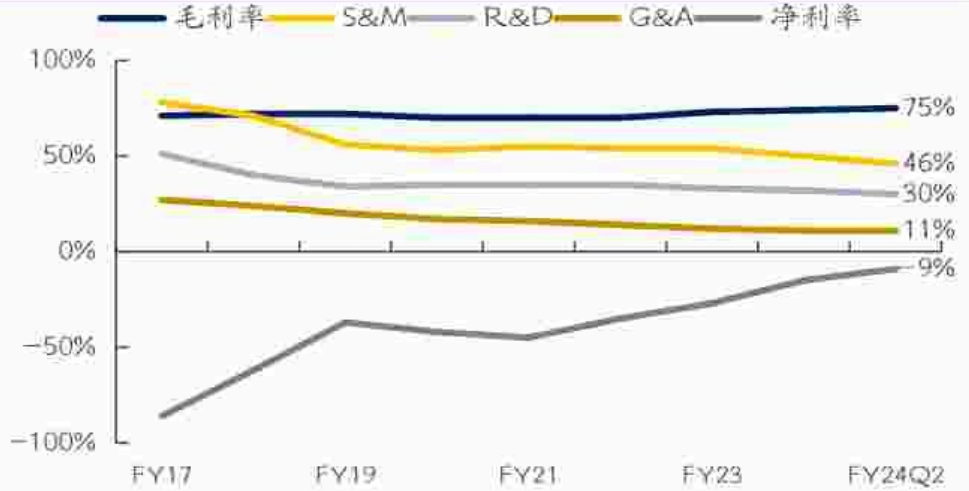
MongoDB毛利率



MongoDB经营性现金流比例



MongoDB费用率



来源：互联网公开资料、弘则研究整理

MongoDB连续两季度超预期，并上调全年收入指引

FY24Q1收入和客户数超预期，上调全年预期

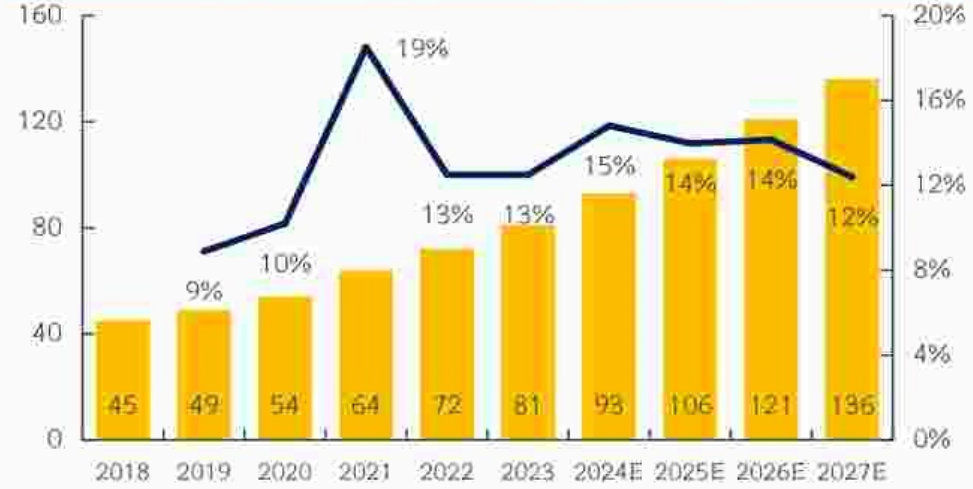
- 本季度公司实现营业收入3.68亿美元（+29%），超出华尔街预期的3.47亿美元元，订阅收入为3.547亿美元，同比增长29%，Atlas收入增长40%；客户数达到43,100个，超出华尔街预期的42,430个
- **业绩展望：**预计全年实现收入15.22-15.42亿美元

FY24Q2收入超预期，上调全年预期

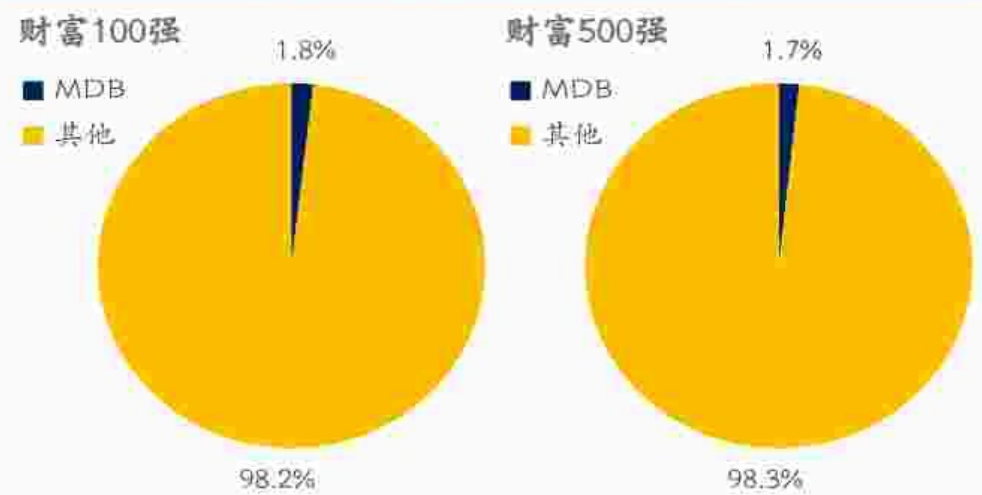
- 本季度公司实现营业收入4.24亿美元（+40%），大幅超出此前3.88-3.92亿美元（+28%）的指引，主要由于非Atlas业务（EA和许可授权）的强劲表现，以及略好于预期的Atlas收入表现。实现毛利率78%（+5pcts），主要由于毛利极高的EA和许可授权收入（包括阿里续签）大幅超预期。实现Non-GAAP营业利润7910万美元，对应op margin 19%（+23pcts），亦大幅超出此前3600-3900万美元的指引。客户数超过45,000个，环比增加1900个客户，同比增加8,000个。其中，直销客户6800个，同比增加1,400个
- **业绩展望：**公司预计Q3将实现收入4-4.04亿美元（+21%），实现Non-GAAP营业利润4100-4400万美元；预计全年实现收入15.96-16.08亿美元（+26%），较此前15.22-15.42亿美元显著提升。主要反映Q3起始ARR的提升，并继续预计Atlas的增长将受到困难宏观环境的影响，预计用量增长将与去年Q2放缓后的平均水平相符，但在Q3有轻微的季节性收益
- **AI用例：**向量数据库处于预览阶段，但已经看到大型客户的极大兴趣，包括某咨询公司允许顾问在超过150万份专家纪要中进行语义检索

数据管理软件市场空间千亿美元，MongoDB单客户ARR和客户数共同提升推动增长

IDC测算数据管理软件市场规模（十亿美元）



MongoDB占客户数据库投入占比



单客户ARR

行业β：1）生成式AI带来应用爆发和向量数据库需求；2）客户持续转移本地IT架构向多云端混合部署迁移；3）客户内部应用数量随着业务和地域扩张而增长带来数据库使用量增长

公司战略：1）向量数据库等新产品推出；2）AI加持的数据库迁移工具和代码重写工具；3）持续扩充销售和客户成功团队

公司战略：1）持续扩充销售团队并支持开发者生态以进行触达并转化免费客户

客户数

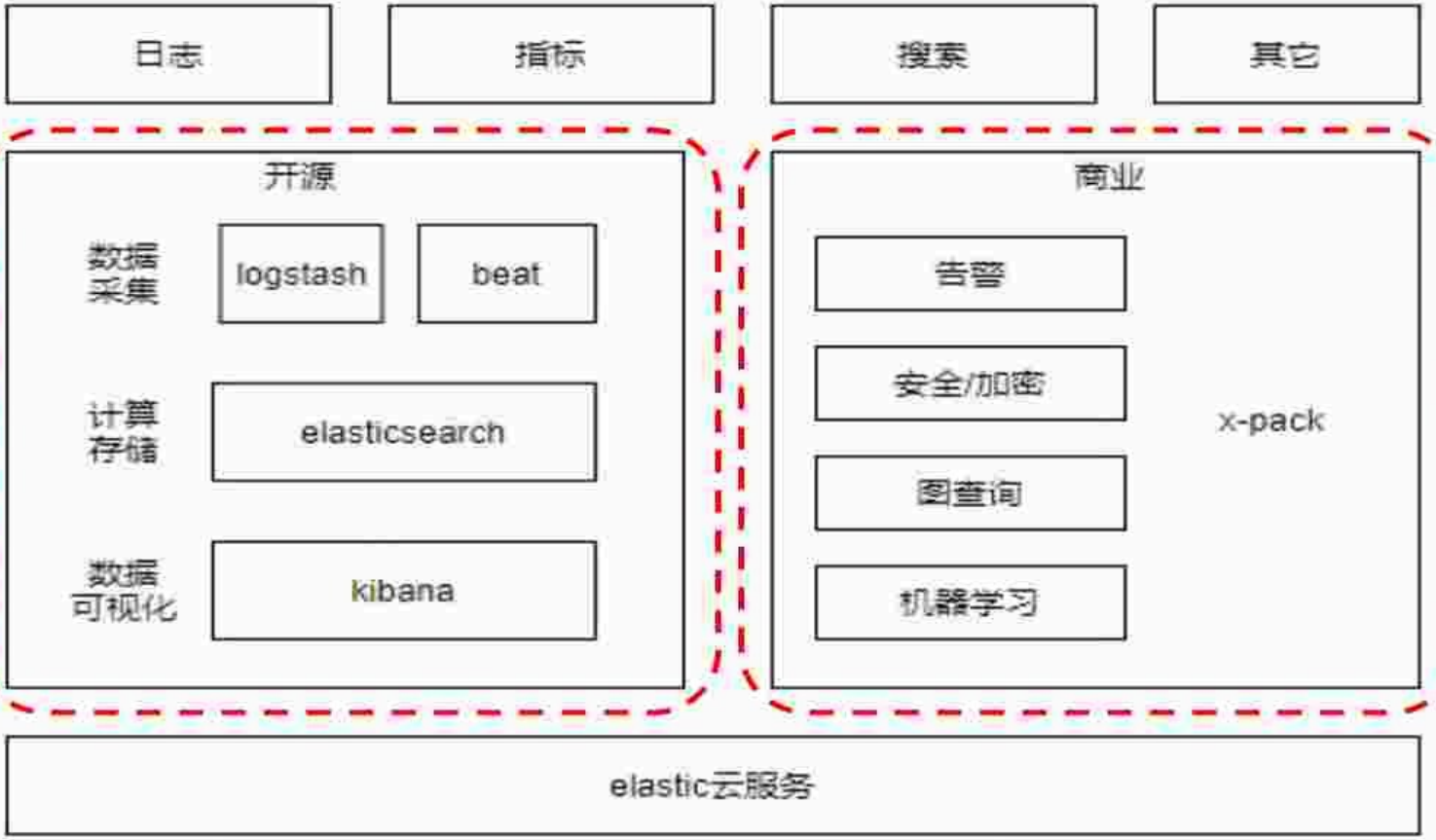
来源：互联网公开资料、弘则研究整理

Elasticsearch是专为搜索和分析场景优化的文档型搜索引擎

来源：互联网公开资料、弘则研究整理

目前形成搜索分析、日志记录、安全性和分析用例、数据可视化的整套数据解决方案

Elastic产品矩阵



ELK Stack由几个创始人的产品合并；Beat项目于2015年启动，目的是简化从各种数据源收集和传输数据的过程

商业组件x-pack包括后期收购公司的产品能力，形成完整解决方案

Elastic在2015年收购了Found，这是一家提供Elasticsearch托管服务的公司，这次收购后来促成了Elastic Cloud的发展

ELK Stack的结合为用户提供实时大数据分析解决方案，后期通过收购进一步补齐能力

Elastic功能延展

收购标的	功能
Logstash（2012）	开源可插拔数据采集工具，后成为ELK Stack的一部分
Kibana（2012）	开源UI，主要用于数据可视化，后成为ELK Stack的一部分
Found（2015）	基于AWS提供主机托管服务，后基于此推出Elastic Cloud SaaS产品系列
Prelert（2016）	机器学习算法分析数据，自动检测系统中异常行为和潜在的安全威胁
Swiftype（2017）	基于云的搜索平台，可以轻松地在网站和应用程序中集成。后期发展成为Elastic Cloud中的Elastic App搜索服务和Elastic网站搜索服务

- 最初的商业化阶段是优先引导用户的日志系统导入到Elastic上，1）日志是产生数据较多的地方，2）日志系统相比主要业务系统对业务影响不大。ELK构成了最基础的实时大数据分析
- 最初合并之时，三大产品的工程师团队各自为战，导致版本发布、兼容十分混乱："如果想使用Shield，您需要使用Elasticsearch 1.4.2……但前提是您不能使用Watcher。如果使用Watcher的话，则您需要使用Elasticsearch 1.5.2。而如果您使用Elasticsearch 1.5.2的话，其仅能与Kibana 4.0.x、Logstash 1.4.x、Shield 1.2.x和Watcher 1.0.x兼容。"该问题于2015年被解决
- 后续收购Found、Prelert和Swiftype等公司，逐渐形成现有的更完整的数据解决方案
- 比如github、stackoverflow等网站的搜索都是基于elasticsearch

目前形成搜索分析、日志记录、安全性和分析用例、数据可视化的整套数据解决方案

Elastic解决方案特性		
	特性	功能
搜索	搜索应用	为网站或App带来精细的API集和直观的仪表板。客户可以直接在Elasticsearch之上构建，或使用Elastic应用程序搜索框架快速构建和定制搜索应用程序
	工作区搜索	客户可以部署内部工作区搜索，无缝连接到其他生产力工具、CRM、云存储平台、协作工具、操作管理平台和内容管理系统，可以从更多的来源摄取任何类型的内容
可观测性	日志	大规模索引、搜索和分析结构化和非结构化日志，可视化从日志中提取的信息，以了解系统行为和趋势，优化性能
	指标	指标摄取、搜索、可视化和分析来自IT系统的数字和时间序列数据，包括应用程序、数据存储、主机、容器、云基础设施等
	APM	APM提供了对代码级别应用程序性能的洞察。开发人员可以对应用程序进行检测，并看到事务在服务之间从前端到后端的生命周期
	合成监控	客户和用户利用合成监控来跟踪和监控支持业务运营的主机、网站、服务和应用程序端点的可用性。通过主动监控，客户可以在终端用户报告之前检测到故障组件
安全	安全信息和事件管理	Elastic SIEM自动化威胁检测和修复，通过预建的Elastic Agent和Beats集成，SIEM可以从云、网络、端点、应用程序和其他系统摄取数据
	端点安全	将预防、检测和响应结合成一个单一的、自主的代理，甚至可以在隔离的环境中运行。端点安全包括对勒索软件、恶意软件、网络钓鱼、漏洞利用、无文件攻击和其他威胁的保护
	XDR	当SIEM和端点安全一起部署时，它们提供了强大的安全姿态和对潜在威胁的广泛可见性。XDR提供了一个统一的安全堆栈，保护端点、云和更广泛的环境
	云安全	云安全通过丰富的云姿态可见性和对云工作负载的运行时保护，保护云部署，具有预防、检测和响应能力，所有这些都集成在一个解决方案中

来源：互联网公开资料、弘则研究整理

Elastic Cloud云托管定价

Elastic Cloud定价

标准级

低至
95 美元/月[1]

免费试用

开始上手的理想之选

- Elastic Stack 核心功能，包括安全性
- Discover、字段统计信息、Kibana Lens、Elastic Maps 和 Canvas
- 告警和堆栈内操作

安全性

- 告警，包含检测引擎和预构建规则
- 集中化采集和代理管理
- 恶意软件防御和主机数据收集
- 实例管理
- 云安全态势管理 (CSPM) 和云漏洞管理 (CNVM)

黄金级

低至
109 美元/月[1]

免费试用

标准级所有功能，另加：

- Reporting
- 第三方告警操作
- Watcher
- 多堆栈监测

安全性

- 经优化的工作流，包括第三方事件响应工作流
- 检测告警外部通知和操作
- 高级托管管理配置

白金级

低至
125 美元/月[1]

免费试用

黄金级所有功能，另加：

- Elastic Stack 高级安全功能
- Machine Learning — 异常检测、监督式学习、第三方模型管理
- 跨集群复制

安全性

- Machine Learning 异常检测和预构建 SIEM 作业
- 行为勒索软件防护

企业级

低至
175 美元/月[1]

免费试用

白金级中的所有功能，另加：

- 可搜索快照
- 支持可搜索冷层和冻结层
- Elastic Maps Server

安全性

- 用于长期保留可搜索操作的归档文件的可搜索快照
- 主机响应操作
- 可供深入了解工作负载的云工作负载保护
- 用于在生成式 AI 方面提供指导的 Elastic AI Assistant

Welcom
wonderi

最低价测算基于云生产配置，120 GB 存储空间/2 个区域。按实例类型使用量定价

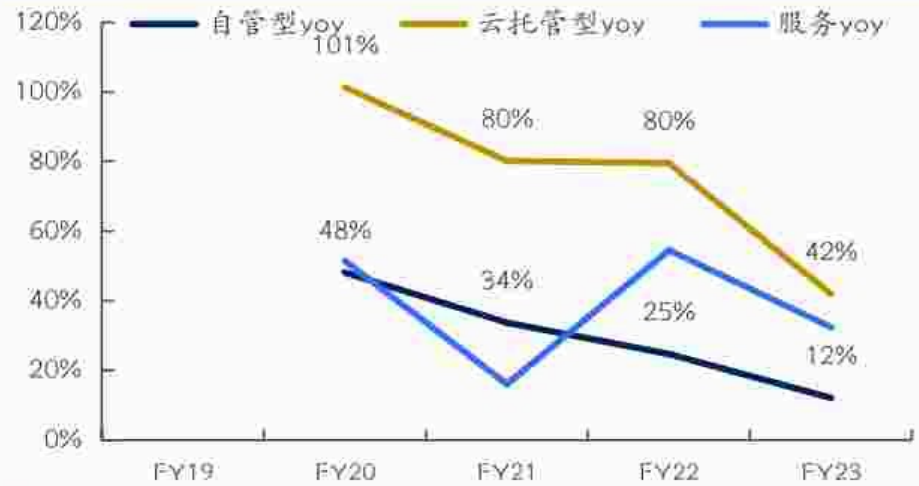
来源：互联网公开资料、弘则研究整理

订阅收入超90%，分为自管型本地部署订阅和全托管云订阅，全托管云订阅（Elastic Cloud）占比超40%

Elastic营业收入（亿美元）

Elastic分业务营收（亿美元）

Elastic分业务收入增速



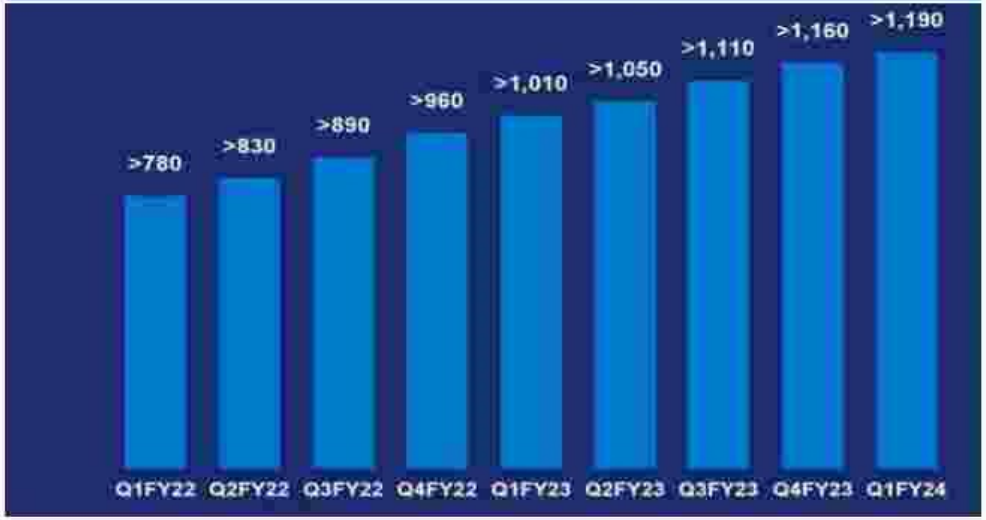
Elastic Cloud收入占比

来源：互联网公开资料、弘则研究整理

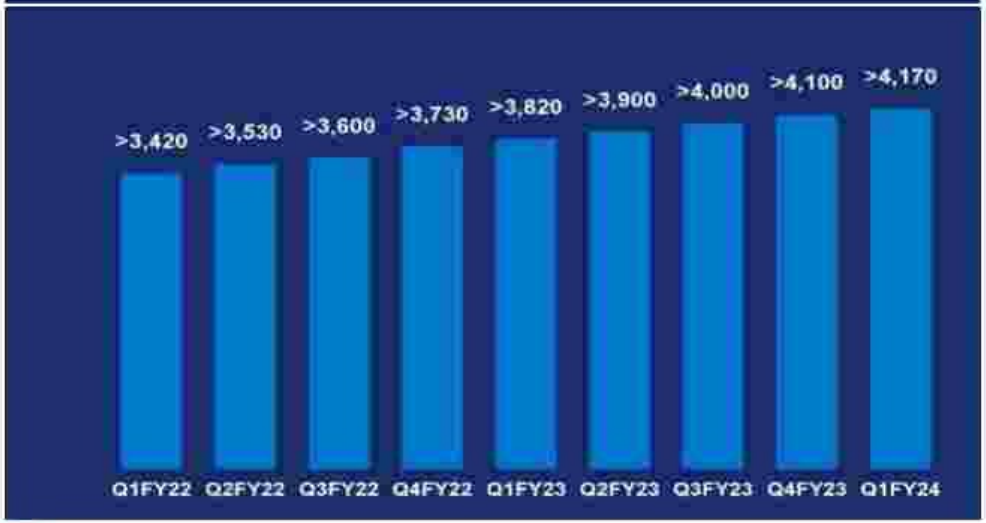
客户数和ACV价值量共同驱动增长

Elastic客户数（家）

ACV>$100K客户数（家）



ACV>$10K客户数（家）



Net Expansion Rate

~113%

Q1 FY24

Trailing twelve month measure

Includes only consumption, not commitments, for customers on Cloud consumption contracts

Elastic官方从2018年开始支持向量检索功能，23年5月推出ESRE（Elasticsearch Relevance Engine）

Elasticsearch向量检索功能的发展历程

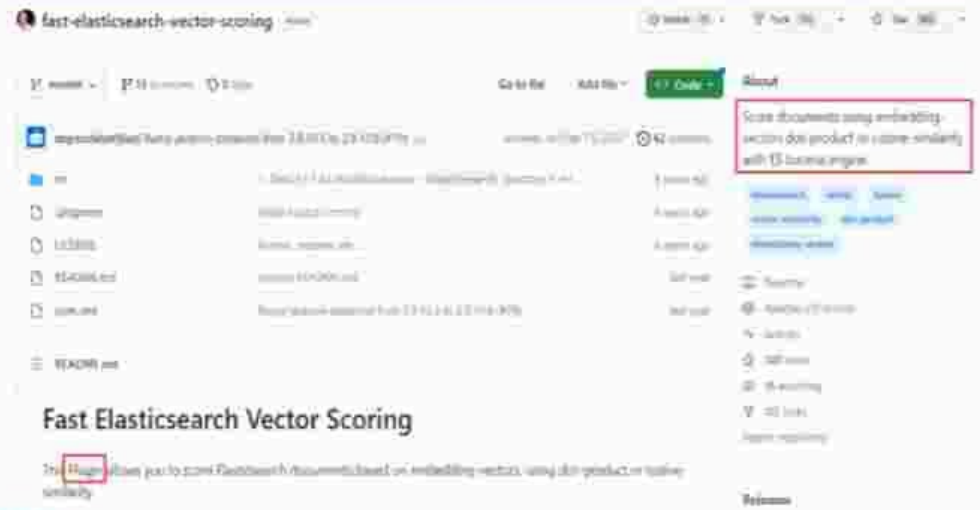
2016年
5.x版本中，爱好者们开始尝试通过插件和数学运算来实现向量检索功能，一些早期插件如elasticsearch-vector-scoring、fast-elasticsearch-vector-scoring

2018年
Elasticsearch 7.0版本正式增加对向量字段的支持，例如通过dense_vector类型，这标志着Elasticsearch正式进入向量检索领域，不再只依赖于插件

2018年
Elasticsearch 7.3版本后，官方引入了更复杂的相似度计算方法，比如余弦相似度、欧几里得距离等

2023年
Elasticsearch 8.8版本的向量检索支持维度从1024提升至2048，并推出ESRE

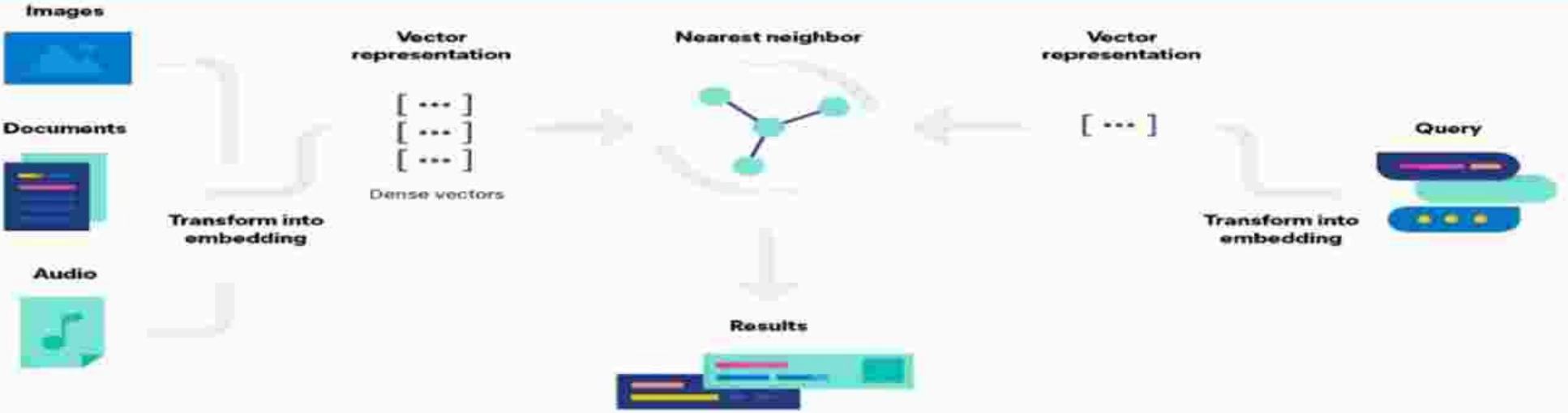
5.x版本中社区开发者开发的向量检索插件



- 最初的向量检索插件主要用于基本的相似度查询，比如文本相似度计算
- 7.3版本之后，引入了更复杂的相似度计算方法，提供更强大和灵活的相似度计算选项，主要场景在于：1）个性化推荐：通过余弦相似度分析用户的行为和兴趣，提供更个性化的推荐内容；2）图像识别和搜索：使用欧几里得距离快速检索与给定图像相似的图像；3）声音分析：在声音文件之间寻找相似模式，用于语音识别和分析

ESRE旨在使用Elasticsearch作为底层存储和搜索技术，帮助开发人员构建AI搜索应用程序；FY24Q1已有数百家客户使用

Elasticsearch一体化向量搜索引擎



矢量数据库

使用您自己的转换器模型

数据集成和采集库

RRF 混合排名

Elastic Learned Sparse Encoder

检索增强生成

- 作为8.8版本的一部分，ESRE的所有功能会随白金级套餐和企业级套餐一起提供
- 业绩说明会引言：“FY24Q1，我们看到围绕生成式AI的大量活动，许多客户选择ESRE作为使用我们的向量搜索和混合搜索功能构建生成式AI应用程序的平台。例如，一家总部位于美国的财富100强全球媒体和技术公司已将ESRE与他们自己本地托管的大型语言模型集成，使他们的票务系统现在能够针对客户的问题提供上下文答案。**目前有数百名付费客户使用 ESRE 进行向量搜索。**”

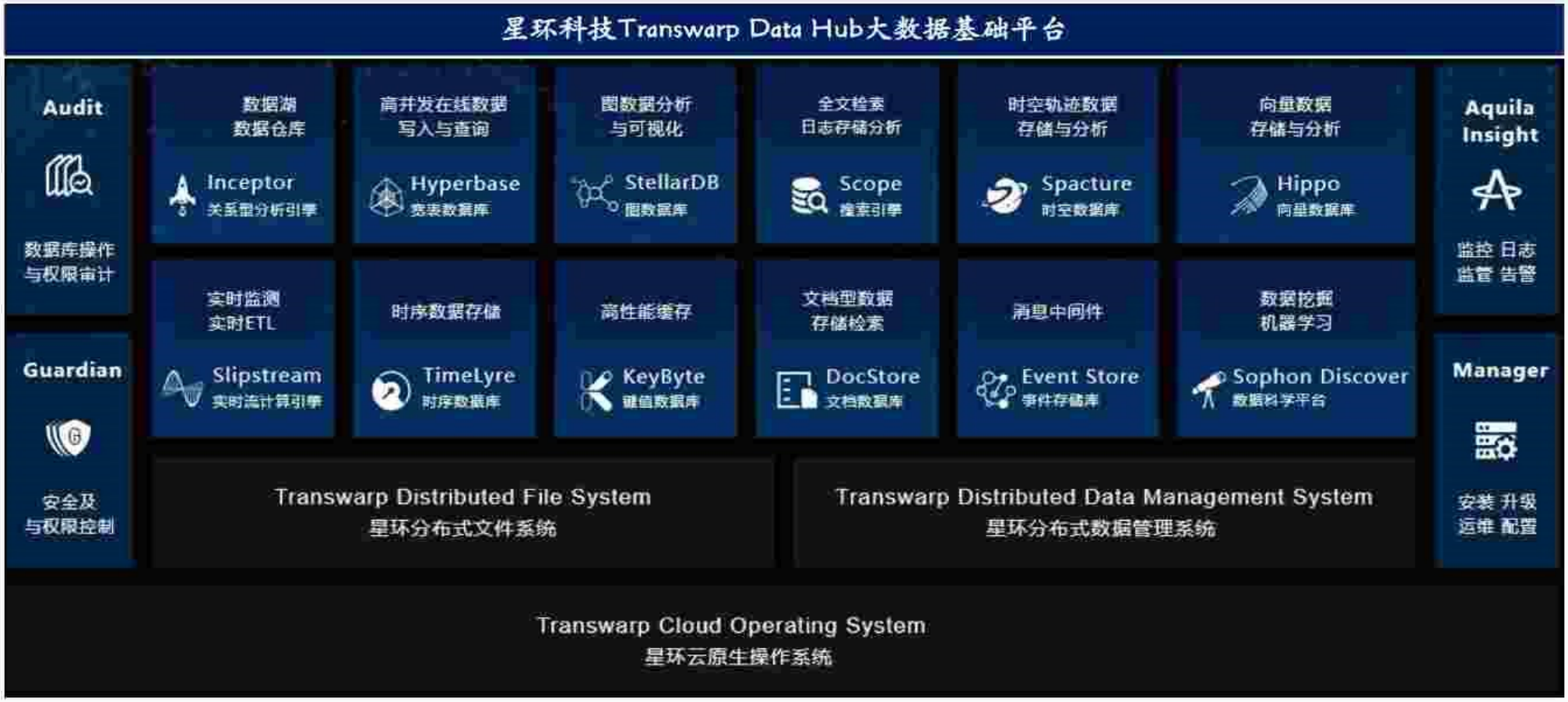
来源：互联网公开资料、弘则研究整理

星环科技于23年5月发布向量数据库产品

核心组件	特点
TDDMS	TDDMS支持弹性扩缩容、自动故障恢复、权限控制、多租户与冷热数据分层存储等功能，多副本机制实现数据服务高可用并保证副本之间的数据一致性
Vector Engine	支持海量向量数据的检索，具备高准确性与高性能的相似检索能力
Embedding Hub	内置的向量转化工具，提供标准化接口连通各类大模型并实现数据的向量嵌入
Model Cube	统一了模型生命周期中的模型上架、模型评估和模型部署，可纳管多模态、多类型的模型，可提高模型的可维护性和可操作性

来源：互联网公开资料、弘则研究整理

向量数据库Hippo将作为TDH中的SKU之一进行商业化



- TDH是公司自主研发的一站式大数据基础平台，包括多个大数据存储与分析产品，能够存储PB级别的海量数据，可以处理包括关系表、文本、时空地理、图数据、文档、时序、图像等在内的多种数据格式，提供高性能的查询搜索、实时分析、统计分析、预测性分析等数据分析功能
- 目前TDH已经在政府、金融、能源、制造业等十多个行业内落地，支撑如金融风控与营销、智慧制造、城市大脑、智慧交通等多种核心行业应用

来源：互联网公开资料、弘则研究整理

实际业务中TDH可作为基础软件或包装成解决方案进行销售

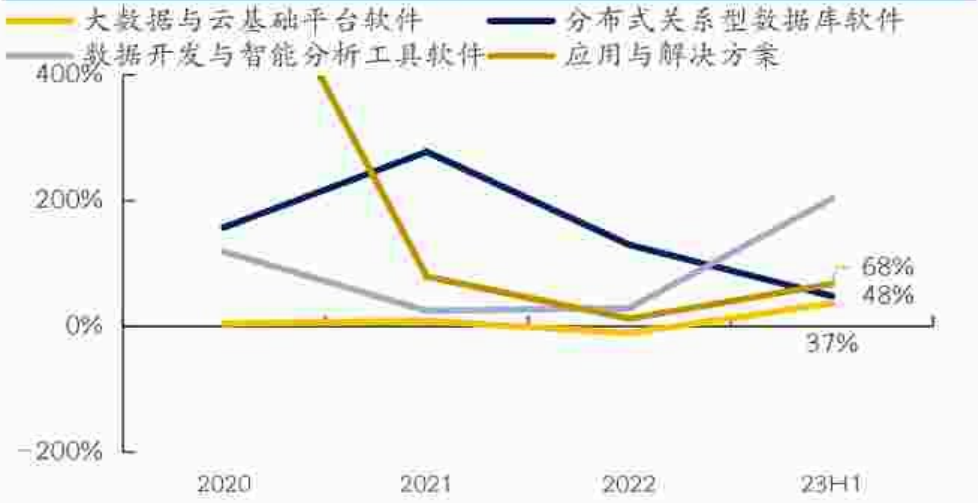
星环科技分业务营业收入（亿元）											
业务类别		细分类别	2019		2020		2021		2022		23H1	
			金额	占比	金额	占比	金额	占比	金额	占比	金额	占比
大数据基础软件业务	基础软件	大数据与云基础平台软件	1.29	74%	1.35	52%	1.46	44%	1.31	35%	0.33	24%
		分布式关系型数据库软件	0.01	1%	0.04	1%	0.14	4%	0.31	8%	0.10	7%
		数据开发与智能分析工具软件	0.15	8%	0.32	12%	0.40	12%	0.52	14%	0.25	18%
		合计	1.45	83%	1.70	65%	2.00	60%	2.14	57%	0.68	50%
	技术服务		0.23	13%	0.53	20%	0.64	19%	0.92	25%	0.31	23%
	合计		1.68	96%	2.23	86%	2.64	80%	3.06	82%	1.00	72%
应用与解决方案	数据应用解决方案		0.04	2%	0.30	11%	0.54	16%	0.60	16%	0.31	23%
	业务应用解决方案				0.01	0%	0.01	0%	0.01	0%	0.04	3%
	合计		0.04	2%	0.30	12%	0.54	16%	0.61	16%	0.36	26%
其他业务			0.03	2%	0.07	3%	0.12	4%	0.06	2%	0.03	2%
总计			1.74	100%	2.60	100%	3.31	100%	3.73	100%	1.38	100%

来源：互联网公开资料、弘则研究整理

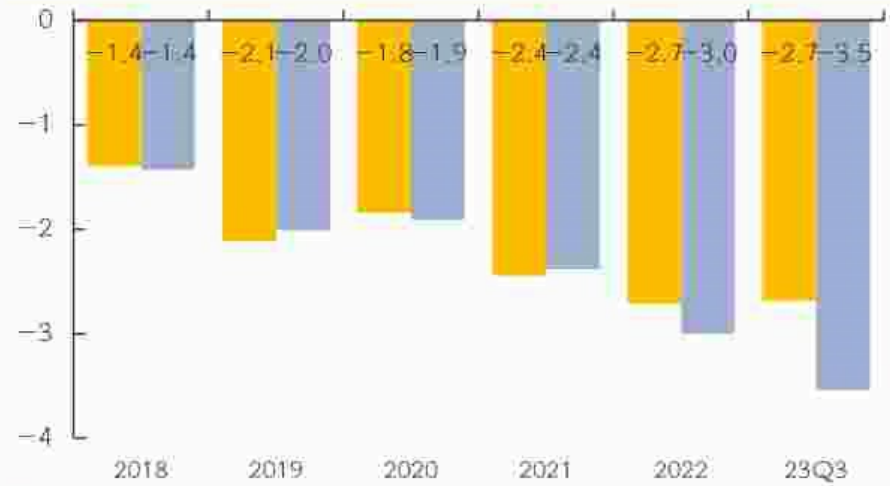
过去4年收入复合增速保持在30%以上，深耕行业和产品解决方案推动公司成长

星环科技营业收入和现金流收入（亿元）

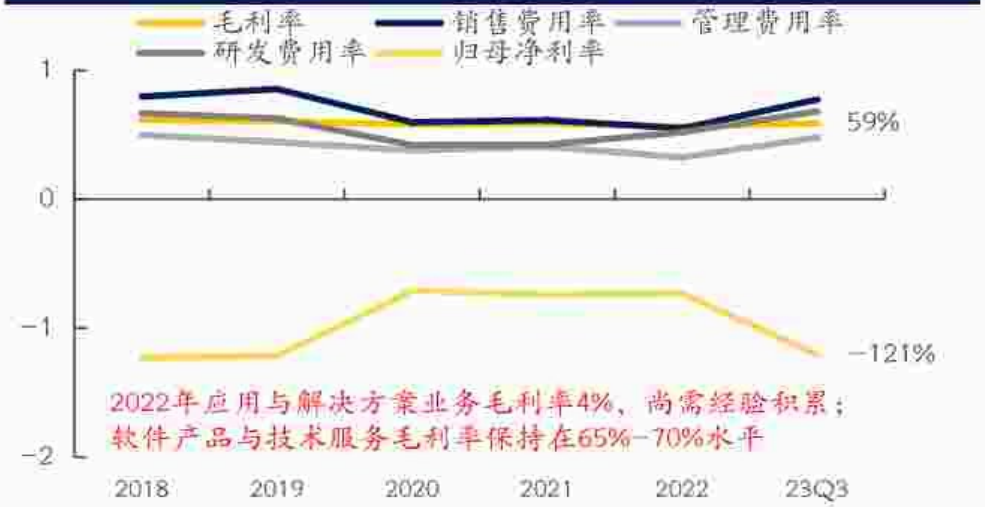
星环科技分业务增速



星环科技净利润和现金流净额（亿元）



星环科技利润率和费用率

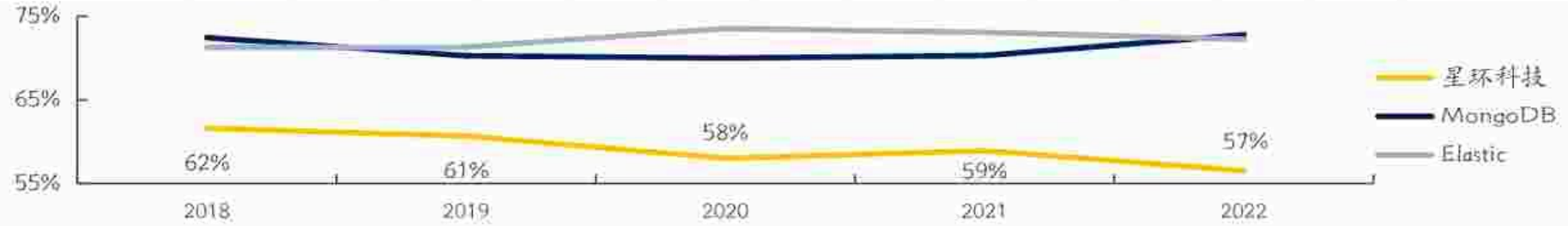


来源：互联网公开资料、弘则研究整理

高研发投入补全自研软件矩阵

可比公司研发人效及毛利率

可比公司人效（万元、万美元）																	
				销售					研发					总数			
年份	营收	净利润	薪酬	人员	薪酬	人均创收	人均创利	人均薪酬	人员	薪酬	人均创收	人均创利	人均薪酬	总人员	总人均创收	总人均创利	总人均薪酬
								星环科技									
2019	17,425	-21,135	27,247	382	10,334	46	-55	27	211	9,211	83	-100	44	720	24	-29	38
2020	25,999	-18,434	32,282	501	11,250	52	-37	22	215	9,393	121	-86	44	788	33	-23	41
2021	33,086	-24,468	44,935	622	15,839	53	-39	25	271	12,206	122	-90	45	1,024	32	-24	44
2022	37,262	-27,135	52,673	606	16,528	61	-45	27	354	16,858	105	-77	48	1,088	34	-25	48
								MongoDB									
2019	42,172	-17,552		789		53	-22		476		89	-37		1,813	23	-10	
2020	59,038	-26,694		1,171		50	-23		638		93	-42		2,539	23	-11	
2021	87,378	-30,687		1,713		51	-18		863		101	-36		3,544	25	-9	
2022	128,404	-34,540		2,249		57	-15		1,030		125	-34		4,619	28	-7	
								Elasticsearch									
2019	42,762	-16,717												1,936	22	-9	
2020	60,849	-12,943												2,179	28	-6	
2021	86,237	-20,385												2,978	29	-7	
2022	106,899	-23,616												2,886	37	-8	



来源：互联网公开资料、弘则研究整理

Horizon Insights
弘 则 研 究

THANKS

欢迎指正

免责声明

，不会仅因接收人/接受机构收到本报告而将其视为客户。
本报告根据国际和行业通行的准则，以合法渠道获得这些信息，尽可能保证可靠、准确和完整，但并不保证报告所述信息的准确性和完整性，也不保证本报告所包含的信息或建议在本报告发出后不会发生任何变更。本报告中所提供的信息仅供参考。报告中的内容不对投资者做出的最终操作建议做任何的担保，也没有任何形式的分享投资收益或者分担投资损失的书面或口头承诺。不作为客户在投资、法律、会计或税务等方面的最终操作建议，也不作为道义的、责任的和法律的依据或者凭证。无论是否已经明示或者暗示。在任何情况下，本公司不对客户/接受人/接受机构因使用报告中内容所引致的一切损失负责任，客户/接受人/接受机构需自行承担全部风险。

弘则弥道（上海）投资咨询有限公司
公司地址：上海市浦东新区世纪大道210号21世纪中心大厦1206室