大数据技术标准推进委员会
Big Data Technology and Standard Committee

数据库发展研究报告
(2023 年)

CCSA TC601 大数据技术标准推进委员会

2023年7月

版权声明

编写委员会

本报告的撰写得到了数据库领域多家企业与专家的支持和帮助，主要参与单位与人员如下。

❖ **主要编写单位**（排名不分先后）：

大数据技术标准推进委员会、中移动信息技术有限公司、华夏银行股份有限公司、北京科蓝软件系统股份有限公司、星环信息科技(上海）股份有限公司、天谋科技(北京)有限公司、云和恩墨（北京）信息技术有限公司、阿里云计算技术有限公司、华为云计算技术有限公司、深圳计算科学研究院、讯飞智元信息科技有限公司、中兴通讯股份有限公司、浪潮云信息技术股份公司、上海沄熹科技有限公司、浙江创邻科技有限公司、杭州沃趣科技股份有限公司、广州巨杉软件开发有限公司、天津南大通用数据技术股份有限公司、北京人大金仓信息技术股份有限公司、北京海致星图科技有限公司、上海爱可生信息技术股份有限公司、成都虚谷伟业科技有限公司、上海热璞网络科技有限公司、腾讯云计算(北京)有限责任公司、蚂蚁科技集团股份有限公司、蚂蚁区块链科技（上海）有限公司、北京庚顿数据科技有限公司、湖南亚信安慧科技有限公司、苏州库瀚信息科技有限公司、北京思斐软件技术有限公司、上海新炬网络信息技术股份有限公司、北京九章云极科技有限公司、深圳矩阵起源科技有限公司、武汉达梦数据库股份有限公司、四川蜀天梦图数据科技有限公司、武汉达梦数据技术有限公司、北京达梦数据库技术有限公司、北京万里开源软件有限公司、北京奥星贝斯科技有限公司、杭州拓数派科技发展有限公司、贵州易鲸捷信息技术有限公司。

❖ **编写组主要成员**（排名不分先后）：刘思源、齐丹阳、刘蔚、马嘉慧、马鹏玮、闫树、姜春宇、魏凯、袁畅、邢韦川、郑鸿健、杨明珉、郑展奋、赵春阳、徐珂、胡捷、王辉、陈曦、林海、田亮、郑贵德、郭帆、魏晗清、雷天洋、刘磊、吴丰泽、张星宇、乔嘉林、刘海、秦楚晴、黄向东、李轶楠、江宁、杨俊、张鹏志、王斌、谢炯、宋震、黎火荣、汪晟、陈吉强、刘颖男、冯程、朱松、樊文凯、张亚楠、隋景鹏、何睿、郭亮、陈伟红、杨锐、王义寅、王龙、黄佩、蒋昀岂、倪修峰、王慧敏、张晓阳、吕作晶、魏星、齐学成、韩银俊、王阳、刘刚、陈家伟、邓光超、金宁、周幸骏、苑晓龙、张晨、周研、马超、魏兴华、李春、张文件、吴炎、樊耀文、许建辉、武赟、杨上德、史新龙、冯文忠、白雪、王薇、贾欣泉、张俊峰、张秋举、胡一鸣、杨娟、沈游人、刘艺华、路新英、黄炎、苏鹏、明玉琢、苏德财、郭家文、江培锋、姜维莹、朱飞、陈亮、苏强、胡一鹤、崔安颀、林恒、郭智慧、吴晓晨、李阳、蒋志勇、徐岩、梁召远、王晋晖、贾孝芬、张桦、吕亚宁、顾鸿翔、杨国华、王磊、张远康、张亮、韩锋、潘娟、程永新、梁铭图、黄国标、郭萌萌、李慧静、黎超、程静、严恒、胡书能、王振宇、赖禧、张睿、陶天林、李庄庄、张永强、邓亮、徐欣、万亮、刘俊锋、齐益琛、李杨桅、徐爽、王栩、李阳、莫荻。

前　言

当前，数据正在成为重组全球要素资源、重塑全球经济结构、改变全球竞争格局的关键力量。数据库作为存储与处理数据的关键技术，在数字经济大浪潮下，全球数据库产业中新技术、新业态、新模式不断涌现。

2023 年，全球数据库产业、技术、应用呈现如下总体发展态势。

产业方面，全球产业发展热度持续保持高位，企业、产品数量再创新高。全球范围内，数据库市场规模约 833 亿美元，企业共 472 家，产品数量超 500 款。我国数据库市场规模 59.7 亿美元，占全球 7.2%，云数据库市场规模占比超过一半，数据库供应商数量达到 150 家，产品数量达到 238 款。

技术方面，数据库技术正围绕助力用户降本增效、护航数据要素安全流通、赋能新兴业务场景三个目标持续发展，呈现 12 个细分发展方向。分别为交易分析一体化、多模处理一体化、数据湖仓一体化、软硬协同一体化、AI 与数据库融合、云与数据库融合、密态数据库、区块链数据库、图联邦学习、向量数据库、图数据库、时空数据库。

应用方面，数据密集型行业应用聚焦深度优化，传统行业应用迎来创新变革。金融、电信等数据密集型行业在既有数据库应用基础上，正通过分布式改造等手段进行深度优化。制造业等传统行业正通过引入时序数据库、图数据库等创新技术，探索数据与实体经济深度融合的新模式。

本报告是中国通信标准化协会大数据技术标准推进委员会（CCSA TC601）继《数据库发展研究报告（2021 年）》、《数据

库发展研究报告（2022 年）》发布后的第三本数据库年度综合报告，内容涵盖数据库产业及市场、数据库产品及服务、数据库支撑体系、数据库技术发展趋势和典型行业数据库应用情况综述。由于水平所限，错误和不足之处在所难免，欢迎各位读者批评指正，本报告为内容简版，欲了解详细内容，请联系 liusiyuan@caict.ac.cn。

目 录

图 目 录

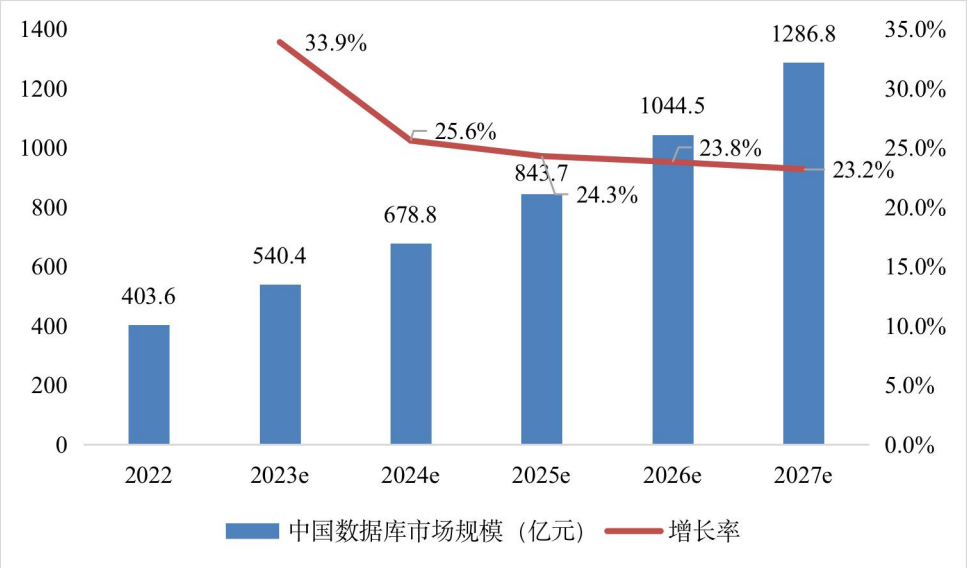
表 目 录

一、数据库产业发展情况综述

当前，全球产业生态加速变革，产品形态日益丰富；我国产业热度持续升温，创新能力不断增强。市场规模不断增高，产品提供商以中美两国为主；非关系型数据库产品是产业关注热点、产品数量占比过半；开源模式影响力再次增大，我国开源业态不断成熟。

(一)数据库产业及市场

根据中国通信标准化协会大数据技术标准推进委员会（以下简称：CCSA TC601）调研分析，我国数据库产业链包括数据库产品提供商、数据库生态工具提供商、数据库服务提供商、数据库安全供应商、数据库生态社区、数据库人才培养等多个环节，各领域参与者专攻术业，发挥竞争优势，积极拓展生态圈，为我国繁荣的数据库生态不断注入活力。

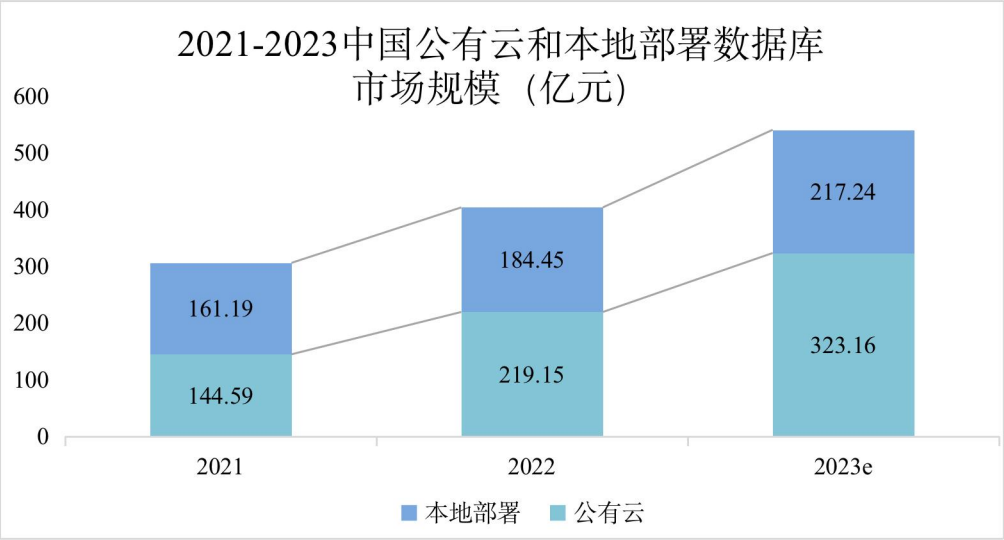


来源：CCSA TC601，2023 年 6 月

图 1 2022-2027 年中国数据库市场规模及增速

据 CCSA TC601 测算，2022 年全球数据库市场规模为 833 亿美元，中国数据库市场规模为 59.7 亿美元（约合 403.6 亿元人民币），

占全球 7.2%[1]。预计到 2027 年，中国数据库市场总规模将达到 1286.8 亿元，市场年复合增长率（CAGR）为 26.1%。



来源：CCSA TC601，2023 年 6 月

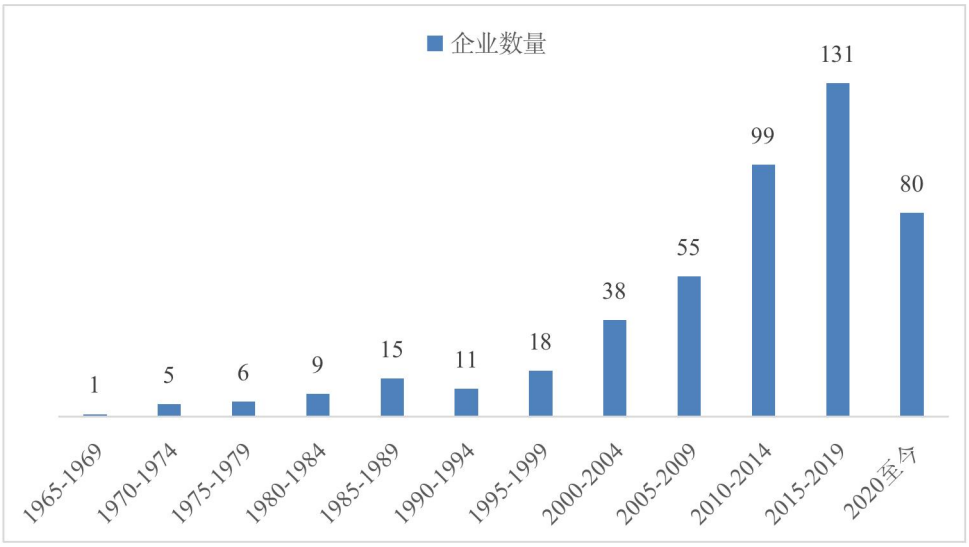
图 2 2021-2023 中国公有云和本地部署数据库市场规模

据 CCSA TC601 测算，按数据库部署方式划分市场规模，2022 年中国公有云数据库市场规模为 219.15 亿元，较 2021 年增速 51.6%，本地部署数据库市场规模为 184.45 亿元，较 2021 年增速 14.4%，公有云和本地部署模式市场规模分别占总市场 54.3%和 45.7%，2022 年公有云数据库市场规模首次过半，预计 2023 年公有云市场占比将进一步扩大达到 59.8%，规模达到 323.16 亿元，本地部署模式市场增速达到 17.8%，规模为 217.24 亿元。

(二)数据库产品及服务

1.从时间看，全球数据库发展经历两轮热周期

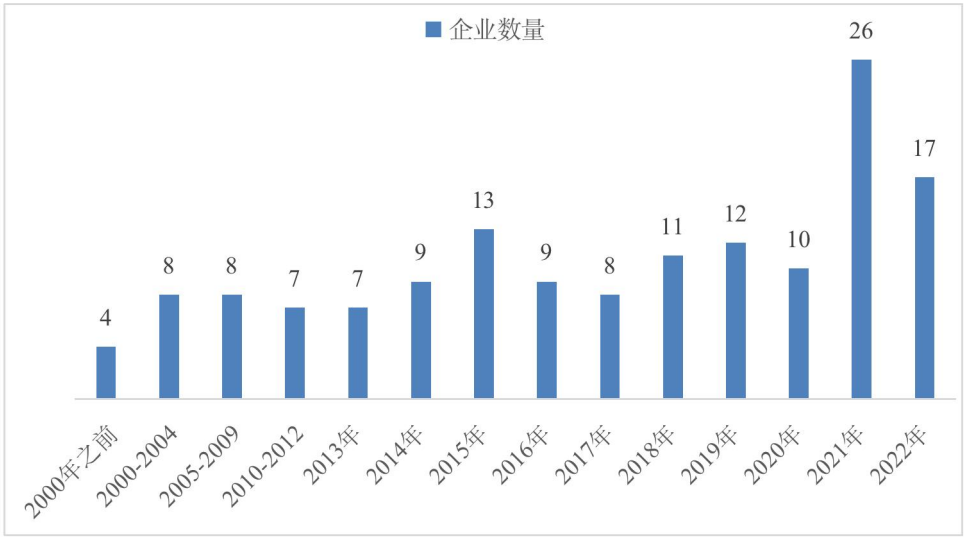
[1]《中华人民共和国 2022 年国民经济和社会发展统计公报》，国家统计局，2022 年全年人民币平均汇率为 1 美元兑 6.7261 元人民币。



来源：CCSA TC601，2023 年 6 月

图 4 全球数据库企业开展业务时间

全球数据库发展经历两次热潮，21 世纪后进入蓬勃发展期。从企业开展数据库业务时间看，全球数据库企业起步于 20 世纪 60 年代，随着 80 年代关系型数据库的理论突破和技术创新，全球数据库迎来第一波发展热潮。步入 21 世纪后，PC 互联网逐步向移动互联网发展，数据库的应用场景不断丰富，全球数据库在 2010-2019 年进入发展高峰期，新兴企业不断成立。这十年间，一共出现了 230 家企业，全球 48.7%的数据库企业均成立于这一时期。



来源：CCSA TC601，2023 年 6 月

图 5 中国数据库企业开展业务时间

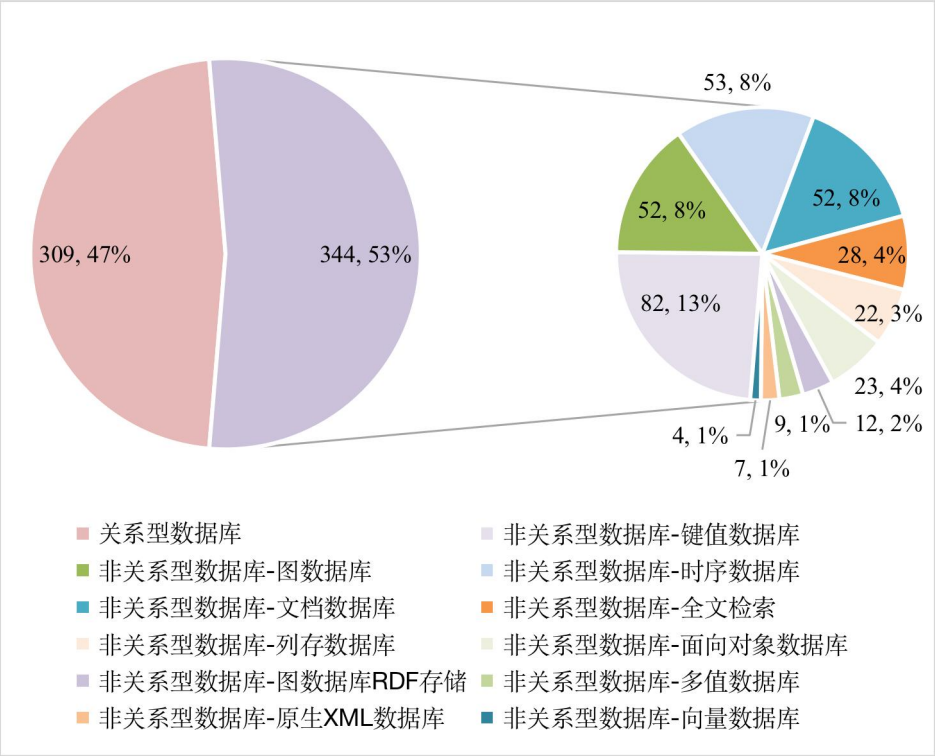
中国数据库产业始于 20 世纪末，并在 2013 年后迎来繁荣发展。截止 2023 年 6 月，我国数据库产品提供商共 150 家，2022 年新增企业数量仍然突破两位数，较 2021 年增速 12.8%。2014-2022 近十年时期迎来发展的高峰，其中 2015 年、2018-2022 年每年企业新增数量均为两位数，六年期间一共有 89 家企业成立，占总数比例 59.3%。

2.从地域看，美国和中国是全球数据库产业的主力军

美国和中国是全球数据库产业的主力军。据 CCSA TC601 统计，截止 2023 年 6 月，全球有共计 472 家数据库产品提供商，总部设在美国和中国的数据库厂商数量遥遥领先，分别为 157 和 150 家，占比 33.3%和 31.8%。全球数据库产品数量为 655 款。美国和中国的数据库产品数量以 242 和 238 款领先，占比分别为 36.9%和 36.3%。

北京为我国数据库产业贡献主要力量。中国 150 家数据库厂商总部大多集中在超一线城市。数量最多的前四名分别是北京、杭州、上海和深圳，数量为 80、15、12、8 个。天津、南京、广州、成都数据库企业数量均为 4 个，其中南京市和成都市由于高校资源丰富，成为很多数据库企业设立研发中心的青睐地点。

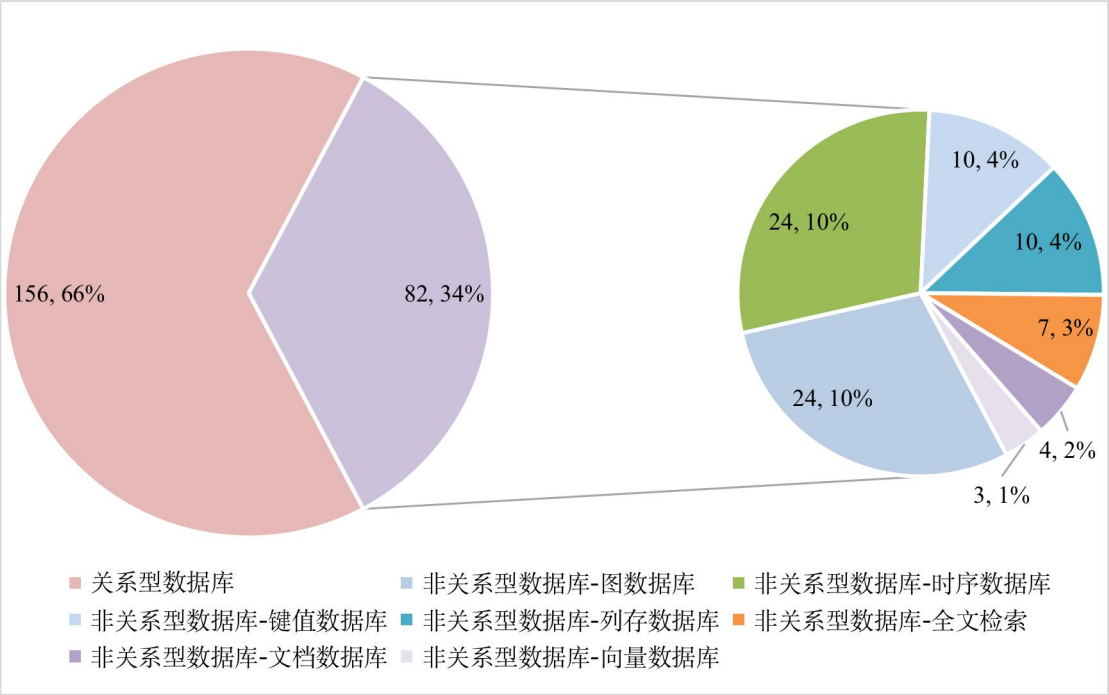
3.从类型看，非关系型数据库在全球范围占比略大



来源：CCSA TC601，2023 年 6 月

图 6 全球数据库产品类型分布

全球数据库产品数量整体分布呈现以非关系型及混合型数据库为主。据 CCSA TC601 统计分析，截止 2023 年 6 月，全球数据库产品共有 655 款。除了早期的两款网状数据库和层次数据库，在剩余的 653 个数据库产品中，关系型数据库为 309 个，非关系型数据库有 344 个，占比分别为 47.3%和 52.7%。非关系型数据库中，键值型数据库 82 个、时序数据库 53 个、图数据库 52 个，在非关系数据库中依次占比 23.8%、15.4%和 15.1%。

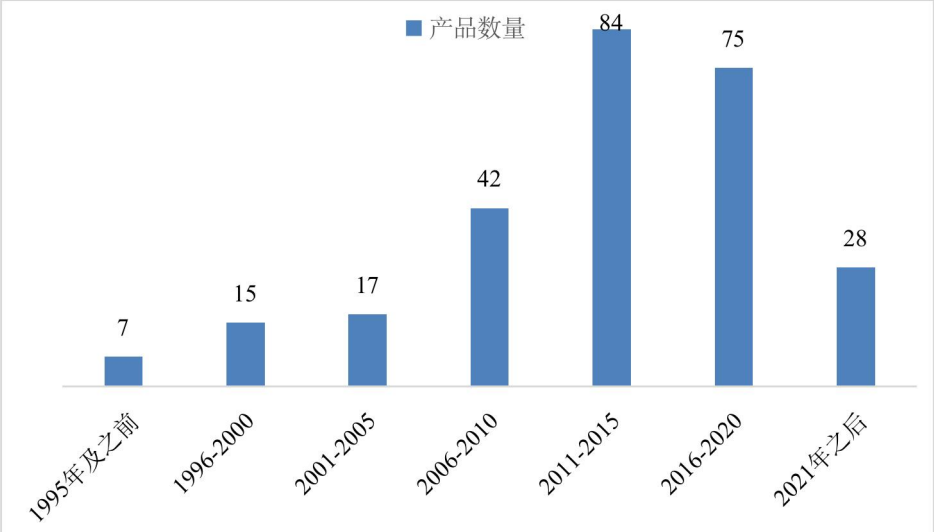


来源：CCSA TC601，2023 年 6 月

图 7 中国数据库产品类型分布

我国数据库产品数量呈现以关系型为主，非关系型数据库为辅的局面。关系型数据库 156 个，非关系型数据库有 82 个，占比分别为 65.5%和 34.5%。非关系型数据库中，图数据库 24 个、时序数据库 24 个、键值数据库 10 个、列存数据库 10 个，在非关系数据库中依次占比 29.3%、29.3%、12.2%和 12.2%。

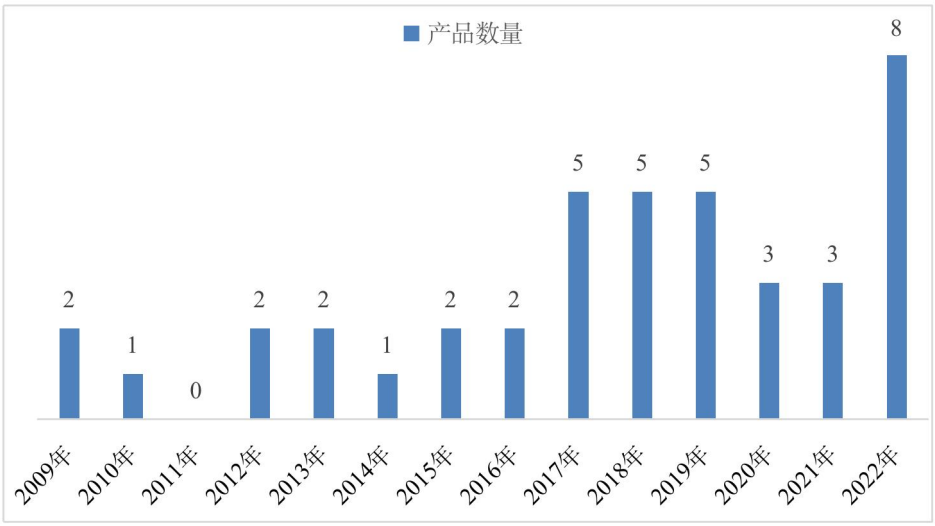
4.从模式看，开源模式在全球范围内发展势头迅猛



来源：CCSA TC601，2023 年 6 月

图 8 全球现存开源数据库的开源时间

全球开源数据库兴起于 20 世纪 90 年代。自 90 年代开源数据库不断推出，2001-2015 年，每隔 5 年，产品数量均呈 2-3 倍增长。开源数据库于 2006 年后迅速发展，目前共 268 款，占全部数据库比例 40.9%。其中在 2011-2020 年进入发展高峰期，大量开源数据库产品不断推出。这十年间，一共出现了 159 个产品，全球 59.3%的开源数据库均诞生于在这一时期。



来源：CCSA TC601，2023 年 6 月

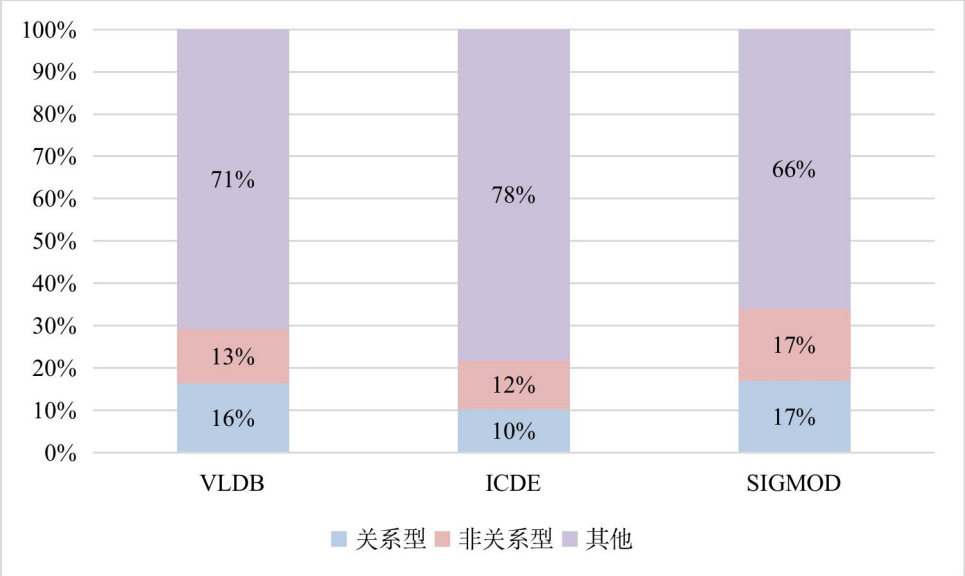
图 9 中国现存开源数据库的开源时间

我国开源数据库产品始于 2010 年前后，但开源数据库在总数中占比较小，开源数据库中七成为关系型数据库。我国数据库产品以商用为主，开源数据库产品共有 42 款，商用和开源占我国数据库产品总数分别为 82.4%和 17.6%。开源产品中，关系型数据库 29 个，非关系型数据库有 13 个，占比分别为 69.0%和 31.0%。我国开源数据库整体起步较晚，在 2017 年之后迎来发展高峰。2017 年至今，一共新增 29 款开源数据库产品，近 7 成产品采用 Apache 许可证 2.0 版。

近两年全球数据库开源生态发展态势良好，期间涌现出许多优秀的开源项目。从国外看，AWS 开源其搜索型数据库产品 OpenSearch，多模数据库 ArcadeDB 和向量数据库 Qdrant 陆续开源，Edgeless Systems 发布基于 MariaDB 的密态数据库 EdgelessDB，内存数据缓存系统 Dragonfly 以及端到端云原生数据库 SurrealDB 正式开源。**从国内看，**分析型数据库公司鼎石纵横和杭州石原子分别开源其产品 StarRocks 和 StoneDB，诺司时空开源其时序数据库产品 CnosDB，蚂蚁集团陆续开源单机版图数据库和图计算引擎 TuGraph。

(三)数据库支撑体系

1.创新方面，非关系型是热点，我国创新实力不断增强



来源：CCSA TC601，2023 年 6 月

图 10 2020-2022 年 VLDB、ICDE 和 SIGMOD 论文分布情况

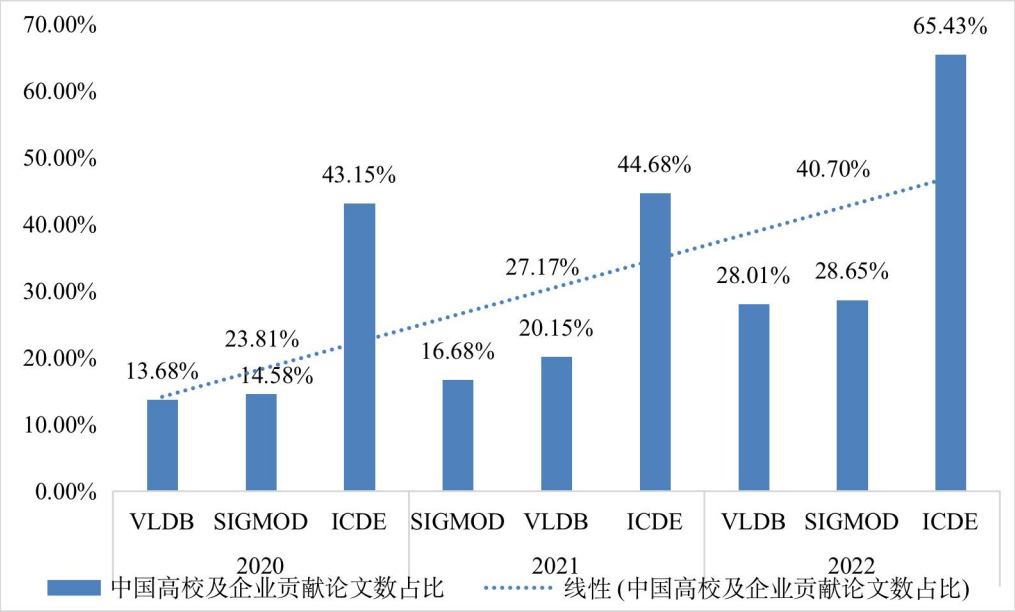
从 VLDB、SIGMOD 和 ICDE 三个数据库领域权威的学术会议研究方向看，**当前关系型数据库和非关系型数据库研究内容数量占比相当，非关系型数据库研究方向成为热点。**以 VLDB 为例，2020-2022 年，各领域论文总数分别为 110、81 和 483 篇，关系型和

非关系型数据库论文分别占三年论文总数量的 16%和 13%。SIGMOD 各领域论文总数分别为 87、87 和 350 篇，关系型和非关系型数据库论文总数均占 17%。ICDE 各领域论文总数分别为 75、85 和 574 篇，关系型和非关系型数据库论文总数占三年论文总数比例分别为 10%和 12%,非关系型数据库占比略微超过关系型数据库。

来源：CCSA TC601，2023 年 6 月

图 11 2022 年 VLDB、ICDE 和 SIGMOD 论文关键词云图

综合分析全球论文研究主题，2022 年三大顶会较为火热的研究方向有机器学习、异常检测、查询处理、数据科学、神经网络、联邦学习、差分隐私、云原生等等。此外，数据库领域如 HTAP、内存数据库、图数据库等方向也是每年不可或缺的研究主题。



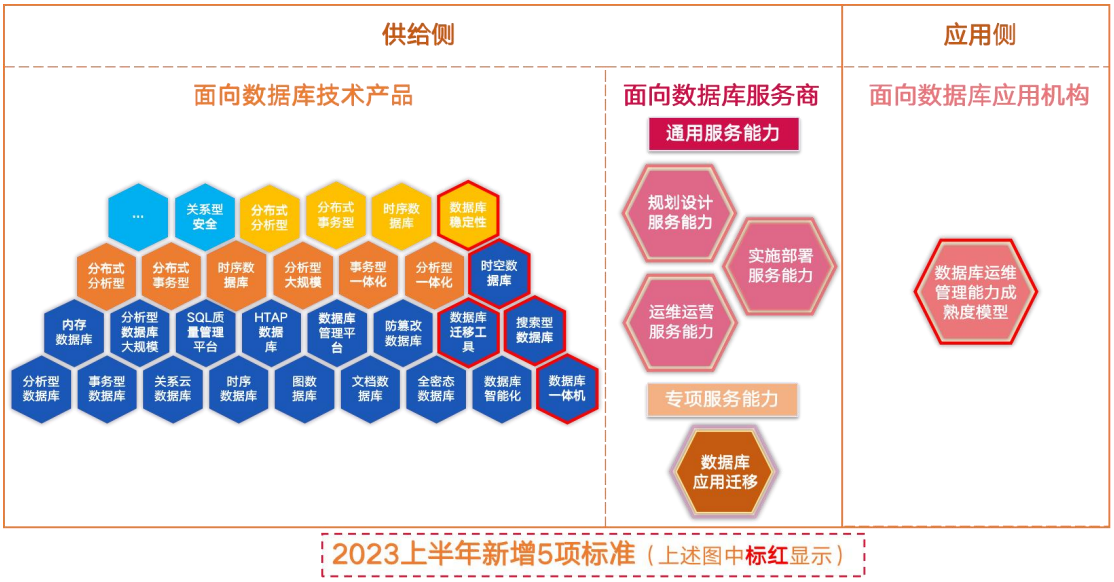
来源：CCSA TC601，2023 年 6 月

图 12 2020-2022 年中国高校及企业学术会议论文贡献情况

我国在全球三大数据库领域学术会议的影响力持续提升。高校及企业在 ICDE 论文贡献占比最高，三年依次为 43.15%、44.68%和 65.43%，三大会议每年贡献占比平均为 23.81%、27.17%和 40.70%，数量呈逐年上升趋势，且 2022 年增长幅度相较前两年十分明显。大部分由我国贡献的论文是以企业、高校合作或者高校间合作的方式发表到顶级会议上。2022 年入选三大顶会论文的企业有阿里巴巴、华为、腾讯、字节跳动、蚂蚁科技、美团、百度、快手科技等；科研机构有中国科学院、深圳计算科学研究院；入选 10 篇及以上论文的高校则有清华大学、香港科技大学、北京大学、香港中文大学、浙江大学、中国科学技术大学、华东师范大学、香港浸会大学、中国人民大学、哈尔滨工业大学、北京航空航天大学、复旦大学等，我国数据库入选高校数量不断扩大，学术国际影响力稳步提升。

2.标准方面，我国数据库产业标准引领作用初见成效

2021 年 10 月 10 日，国务院印发《国家标准化发展纲要》（以下简称《发展纲要》），《发展纲要》明确强调“开展数据库等方面标准攻关，提升标准设计水平，制定安全可靠、国际先进的通用技术标准”，首次在标准化顶层文件中将数据库领域标准化攻关的重要性提升到前所未有高度。纵观国内外数据库标准化进展，我国数据库标准化工作初见成效，从深度和广度均需推进大量工作，以不断适应产业日新月异的变化。



来源：CCSA TC601，2023 年 6 月

图 13 CCSA TC601 数据库领域标准化工作体系

中国通信标准化协会大数据技术标准推进委员会紧跟国家战略，围绕数据库领域标准化工作，设立数据库与存储工作组（WG4）。自 2015 年起共推出 30 项标准，逐步构建以数据库产品、服务和应用为目标的标准体系。**产品能力方面**，从关系型和非关系型，构建了基础能力、性能和稳定性的技术标准；**服务能力方面**，围绕规划设计、实施部署和运维运营，推出国内首个面向数据库服务的团体标准《数据库服务能力成熟度模型》（标准编号：T/CCSA 418-2022），围绕数据库应用迁移和 SQL 质量管理平台，推出能力分级标准，其

中《数据库应用迁移服务能力分级要求》（标准编号：T/CCSA 335-2021）成功入选工信部 2022 年百项团体标准应用示范项目；**行业应用方面**，面向数据库应用方内部运维管理团队，推出《数据库运维管理能力成熟度模型》。CCSA TC601 见证了我国数据库标准化工作有序有力进行，成为国家在数据库领域重要的支撑单位。

二、数据库技术发展情况综述

数据要素时代，数据规模爆发式增长对数据库技术提出了新的挑战。数据库技术将在围绕三个目标持续发展，1）**助力用户降本增效**（交易分析一体化支撑多类业务，多模处理一体化实现一库多用，数据湖仓一体化降低存算成本，软硬协同一体化提升系统性能，AI 与数据库融合迸发无限潜力，云计算成为数据库重要驱动力）；2）**护航数据要素安全流通**（隐私计算保障密态数据安全流通，区块链技术赋能数据资产高度可信，图联邦学习技术打破图数据孤岛）；3）**赋能新兴业务场景**（AI 大模型催生向量数据库新应用，图分析技术洞察数据连接新价值，时空数据库释放时空数据新潜能）。

(一)助力用户降本增效

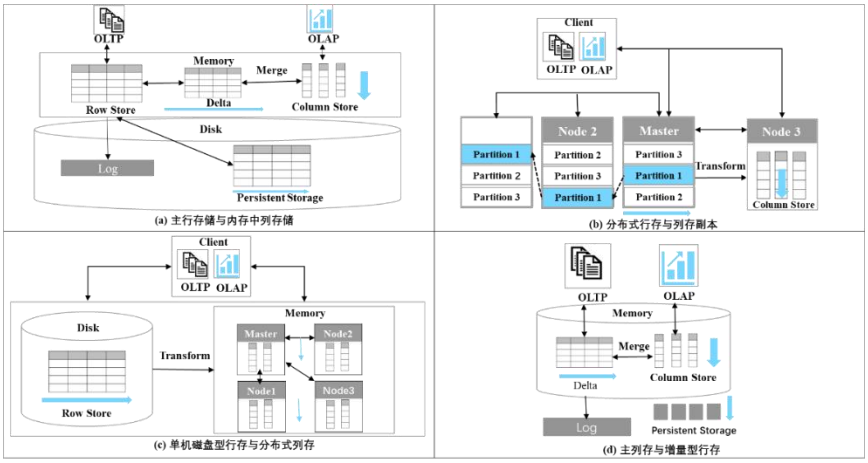
供给侧通过六类优化方式，助力数据库性能不断提升，以及运维、应用成本降低。

1.交易分析一体化支撑多类业务

HTAP（Hybrid Transaction / Analytical Processing，混合事务分析处理）的概念是指同时支持 OLTP 和 OLAP 场景。该技术可以实现一个平台上同时处理多个数据任务，支撑海量并发连接复杂混

合负载，提升系统弹性，降低开发运维复杂度和成本，提升数据使用粒度，提高组织数据处理的效率。

目前，业界主流的 HTAP 技术架构按存储类型划分，主要分为主行存储与内存型列存储、分布式行存与列存副本、单机磁盘型行存与分布式列存，以及主列存与增量型行存四种形态[2]。



来源：HTAP 数据库关键技术综述

图 14 四类 HTAP 数据库技术架构示意图

在技术实现方面，HTAP 在数据组织、数据同步、查询优化和资源调度等方面仍需持续突破。这些技术的解决方法在各种指标上互有优劣，例如效率、可扩展性和新鲜度，如下表所示。

表 1 HTAP 关键技术总览与优缺点比较

HTAP 技术类别	关键技术	代表性产品	主要优点	主要缺点
数据组织技术	基于主行存的内存列选择	MySQLHeatwave Oracle	事务性能高	分析性能低
	基于负载驱动的行列混合存储	/	存储代价低	系统复杂度高
数据同步技术	基于内存增量表与内存型列存的数据同步	Oracle, SQL Server, SAP HANA	性能高	扩展性低
	基于增量日志与持久化列存的数据同步	TiDB, F1 Lightning	扩展性高	合并代价高
查询优化技术	混合行/列存储扫描	TiDB, SQL Server	分析性能	搜索空间

[2] 张超, 李国良, 冯建华, 张金涛. HTAP 数据库关键技术综述. 软件学报, 2023, 34(2): 761–785.

			高	大
	异构 CPU/GPU 硬件加速	RateupDB, Caldera	分析性能高	事务性能低
	面向 HTAP 负载的索引技术	/	事务性能高	内存空间大
资源调度技术	基于负载驱动的资源调度	SAP HANA, Siper	性能高	新鲜度低
	基于新鲜度驱动的资源调度	/	新鲜度高	性能不高

来源：HTAP 数据库关键技术综述

在推广应用方面，HTAP 数据库仍面临多重挑战。一是 HTAP 将事务与分析处理相融合，需对数据库的结构进行大规模修改，这也增加了系统复杂性。二是 HTAP 数据库通常会应用在高度敏感的场景下，需有额外的安全措施保障数据机密性和完整性。三是 HTAP 数据库应用需集成包括分布式系统、高可用性、并发控制等技术，对于建设及运维团队的技术水平要求较高。

2.多模处理一体化实现一库多用

多模数据库技术是在 NoSQL 技术演进中发展起来的，由于需求不断变化、RDBMS 的扩展性不佳等诸多因素导致越来越多的开发者选择 NoSQL 数据库。但多个 NoSQL 数据库系统混用的方式为软件开发团队带来高额的学习成本和维护费用。多模数据库旨在提供多语言持久性的数据建模优势，通过使用单个数据库存储来降低操作的复杂性，更好地支持不同场景下的多种类型数据处理。多模数据库发展呈现两种形态，一是出现了多款原生的多模数据库系统，二是关系型数据库系统也陆续增加了对多模数据处理的支持。多模数据库不仅能够为多种数据模型提供该模型适用的查询接口，也可以通过一种语言实现对多种模型数据的同时查询。

表 2 多模数据库扩展策略

技术路径	数据库管理系统	存储类型

新存储方式	PostgreSQL	relational
	SQL server	relational
	IBM DB2	relational
	Oracle DB	relational
	Cassandra	column
	CrateDB	column
	DynamoDB	column
	Riak	key/value
	Cosmos DB	document
原存储模型扩展	MySQL	relational
	Vertica	column
	ArangoDB	document
	MongoDB	document
	OrientDB	graph
	Cache	object
原始存储策略加新型接口	Sinew	relational
	c-treeACE	key/value
	Oracle NoSQL Database	key/value
	Couchbase	document
	MarkLogic	document

来源：Multi-model Databases: A New Journey to Handle the Variety of Data

学术界对多模数据库的研究大致分为四阶段，2012 年之前的史前研究阶段、 2012 至 2017 年多模数据库开放探讨阶段、 2014 至 2019 年的系统研究阶段以及 2015 至今的细分研究阶段。1997 年，IBM Almaden Research Center 发表了一篇论文系统性地介绍了 Garlic system 的实现。1998 年美国的一篇专利系统性地提出管理多模型数据的统一数据库管理系统，该管理系统由物理存储层、语义数据模型层、逻辑数据模型层以及接口层[3]。2012 年开始，多模数据库系统开始受到学术届关注。2016 年 Serge Abiteboul 概括性地提出了数据管理领域未来几个重要的方向[4]，其中多模数据管理就是其中

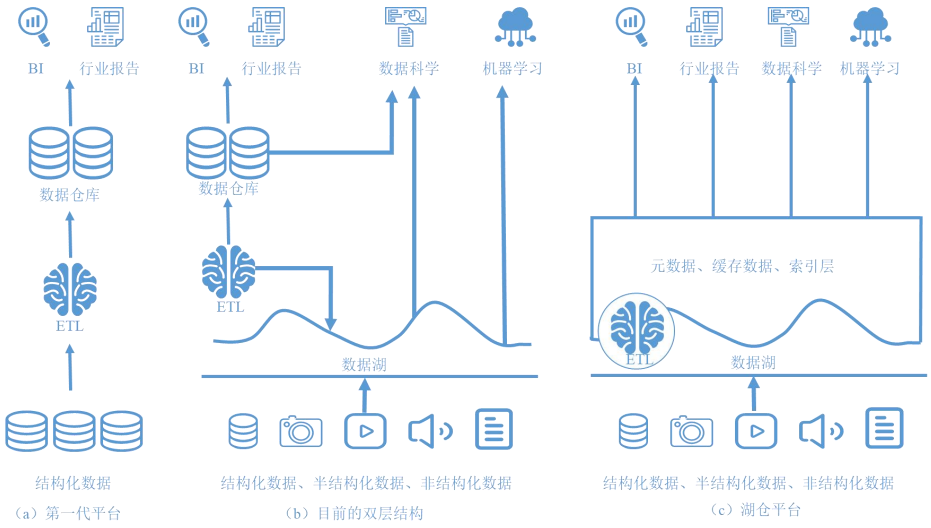
[3] 《Multi-model database management system engine for database having complex data models》
[4] 《Research Directions for Principles of Data Management》

之一。同年，陆嘉恒团队发表文章介绍了一款作者构想的多模数据管理系统的形态和能力[5]。2019 年，陆嘉恒团队提出多模数据库查询和优化以及多模数据库模式设计与优化的技术路径[6]。2015 年开始，学术界对多模数据库的研究进入细分领域研究阶段，主要分为多模数据库模式设计与优化、模式推导和多模演进等方向。

未来多模数据库应原生支持多种数据模型，有统一访问接口且兼容各行业数据规范，具备各模型自动化管理和转换能力的新型数据库系统，多模数据库将逐步形成新的规范和使用方式。

3.数据湖仓一体化降低存算成本

大数据平台技术架构持不断演进，以数据仓库（Data Warehouse）和数据湖（Data Lake）为两类经典代表，近年来这两项技术在不断演进过程中逐渐走向融合形成湖仓一体（Data Lakehouse）技术架构。



来源：Databricks

图 15 数据平台技术架构演进图

数据平台架构历经三个发展阶段。第一代是传统类型的数据仓库，通过 ETL 任务将结构化数据导入到通常为关系型数据库的数据

[5] 《RoadMap: UDBMS: Road to Unification for Multi-model Data Management》
[6] 《Multi-model Databases: A New Journey to Handle the Variety of Data》

仓库中进行商务分析及财务报表等工作。第一代数据仓库面临的问题是计算和存储高度耦合使得平台难以随着数据量的增长而不断增长，另一方面是无法支持非结构化数据。数据平台进入到第二代，也是当前最为流行的双层架构阶段，但这类架构存在难以保证数据湖与数据仓库中数据一致性等问题。为了解决以上问题，第三代数据平台架构湖仓一体架构应运而生。

表 3 数据湖支持数据仓库产品能力对比

时间	公司	产品	优势	缺陷
2011	Hortonworks	Apache Atlas	数据血缘追踪	/
2011	Hortonworks	Ranger	数据权限安全	数据湖中新引擎优先实现功能和场景，并非优先对接 Ranger，可能会产生安全漏洞
2018	Nexflix	Iceberg	提供 MVCC 等增强数仓能力	Iceberg 作为插件方式兼容并配合 HMS，数仓管理能力大打折扣
2018-2019	Uber& Databricks	Apache Hudi & DeltaLake	增量文件格式以支持 Update/Insert、事务等数据仓库功能	新功能打破了元数据湖多套引擎之间关于共用存储的简单约定，Hudi 发明两种表三中查询类型维持兼容性。

来源：CCSA TC601，2023 年 6 月

湖仓一体是一种开放式的数据管理架构，集数据湖的灵活性、可扩展性优势以及数据仓库的数据结构和数据管理功能于一体。主要优势包括以下几个方面，一是降低数据冗余，二是减少存储成本，三是减少报表分析师与数据科学家不必要的重复劳动，四是提升数据分析时效性，五是提升对其它数据技术的兼容性。当前，湖仓一体的技术路径主要分为以数据仓库中支持数据湖特性和以数据湖中支持数仓特性两种技术路径。在数据仓库中支持数据湖的功能主要是通过在数仓中建外部表来实现，目的是使数据仓库更加灵活，主要是以数仓为核心，支持访问数据库。这类技术路线的代表产品包括 Snowflake，阿里云 MaxCompute 和亚马逊 Redshift。数据湖中支

持数仓的功能主要是通过功能性开发实现，如多版本并发控制、自适应 Schema、提供文件级事务等来实现数仓功能，这类产品以 Databricks 的 DeltaLake、Uber 的 Apache Hudi 等为代表。

表 4 数据仓库支持数据湖产品能力对比

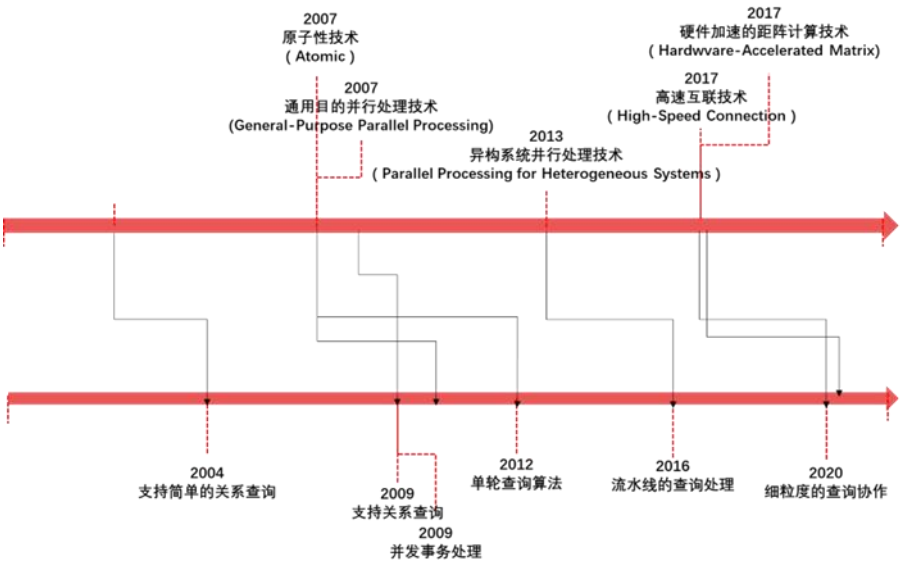
时间	公司	产品	优势	缺陷
2017	Redshift	Redshift Spectrum	支持数仓用户访问 S3 数据湖的数据	需要用户在数仓中通过创建外部表来将数据湖的开放存储路径纳入数仓的概念体系，无法完全自动化创建外部表、添加分区等。生产使用中较为复杂。
2018	阿里云	MaxCompute	外表能力，支持访问包括 OSS/OTS/RDS 数据库在内的多种外部存储	

来源：CCSA TC601，2023 年 6 月

当前，湖仓一体作为一种新兴技术架构，在企业落地方面还处于早期探索阶段，在部署方面仍面临多重挑战。一方面是由于团队缺乏前期数据治理经验，另一方面湖仓一体的高度复杂性使得湖仓之间存在如何协同的问题。怎样打通两套系统存储、保证元数据一致性、确保湖仓之间不同引擎数据交叉引用、如何保障数据安全等问题仍是湖仓一体未来发展过程中亟待解决的问题。

4.软硬协同一体化提升系统性能

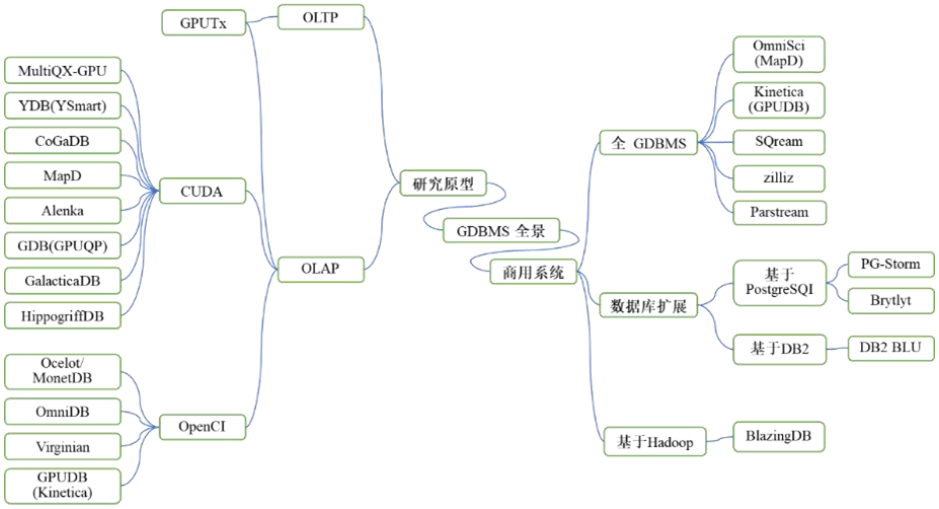
计算机软件和硬件的发展相辅相成、并行不悖，硬件技术的创新或产品成本变化，不仅会给传统的计算机体系结构和系统带来影响，也给系统软件，特别是数据库系统带来了新的机遇。一方面，伴随着硬件技术发展数据库技术不断进步，另一方面在数据库架构方面，硬件技术的发展也不断推进着数据库在分布式、云原生等方面的快速发展。此外，硬件技术的发展也促进了数据库与其它新兴技术的融合，提升了数据库安全性和智能性。



来源：中兴通讯股份有限公司，2023 年 6 月

图 16 FPGA 与 GPU 技术发展历程示意图

数据库技术方面，新型硬件使得数据库在数据计算、数据存储以及数据通信方面持续提升。**数据计算层面**，借助多核、GPU、FPGA、专用芯片等，可以实现并行优化、事务并发控制、查询加速、数据压缩加速、工作负载迁移等；**数据存储层面**，随着新型内存及 NVM 的出现和发展，内存和外存的界限变得模糊，存储及索引设计得到新的性能提升；**数据通信层面**，RDMA、CXL 协议带来网络传输高性能表现和 CPU 卸载能力，或将对数据库系统的网络通信架构设计带来颠覆性变化。**数据库架构方面**，新型硬件对于不同架构类型的数据库产生不同影响。一是使得集中式关系型数据库网络架构更加便捷、建设成本更加低廉。二是使得分布式数据库、云原生数据库等具有更强实用性。此外硬件技术的发展也使得分布式数据库节点之间的处理延时得到不断改进。



来源：GPU 数据库核心技术综述

图 17 GDBMS 系统全景图

目前，以 GPU 计算为核心的数据库技术（GDBMS）受到广泛关注，其具有吞吐量大、响应时间短、成本低廉、易于扩展等特点，可为人工智能、时空数据分析、数据可视化、商务智能等领域带来更大价值，有望改变数据分析领域的格局。GDBMS 按照商业模式分为研究原型（R-GDBMS: for research)和商用系统(C-GDBMS: for commercial)两大类，其中商用 GDBMS 可以进一步分为三类。一是支持 GPU 计算的传统数据库、二是非内存型 GDBMS 使用 GPU 完成全部或者大部分数据库关系运算、三是内存行 GDBMS 内存型 GDBMS[7]。

5.AI 与数据库融合迸发无限潜力

人工智能技术发展驶入快车道，为数据库与 AI 深度融合带来新机遇。2023 年 AIGC 技术的跨越式突破发展，不仅使大语言模型进入公众视野，更扩展了数据库与 AI 融合的发展空间。**一方面，生成式 AI 在数据库结构设计、架构设计、数据分析挖掘等方面可以不同**

[7] 裴威,李战怀,潘巍.GPU 数据库核心技术综述.软件学报,2021,32(3):859−885.

程度简化人员操作，提高开发、运维、分析的效率。例如 2022 年 12 月，数据库自动化和优化平台 OtterTune 宣布推出 OtterTune V1.5，2023 年 Databricks 将大型语言模型（LLMs）引入 SQL 和 MLflow2.3，国内 Bytebase 于 5 月推出基于对话式交互的 SQL 客户端 SQL Chat，阿里巴巴开源了支持自然语言与 SQL 互相转换的数据库开发工具 Chat2DB。**另一方面，多模态数据存储和计算的需求随着大语言模型出现而剧增，向量数据库在构建基于大语言模型的行业智能应用中扮演着重要角色。**2023 年除了 Qdrant、Pinecone、Weaviate、Milvus 等特化的向量数据库备受关注外，许多数据库厂商也开始在原有产品上拓展向量检索的能力，2023 年以来，AWS RDS PostgreSQL 和阿里云 PostgreSQL 14、15 版本新增支持 pgvector 插件，实时数据库 Rockset 增加向量嵌入功能支持，微软宣布 Cosmos DB 支持向量搜索功能。

随着以 ChatGPT 为代表的 AIGC 技术产品发展火热，数据库从业者不断思考 AIGC 技术与数据库相互赋能的途径。AIGC 技术对数据库的影响主要体现在数据开发与分析、数据库性能优化、数据库结构设计、数据库架构设计等方面。一些大型语言模型已可以初步创建复杂数据库的查询过程，使得用户更容易使用自然语言来与数据库进行交互检索。

来源：CCSA TC601，2023 年 6 月

图 18 AIGC 为数据库运维提供建议的示例

数据库开发与分析方面，数据库开发者和数据分析师可以通过大语言模型将自然语言转换为对应的 SQL 语句，从而对数据库进行开发与操作。数据库性能优化方面来看，AIGC 技术可以对数据对象或查询语句进行优化，提供一些通用性建议，同时可以根据具体语句给出进一步优化建议。

来源：CCSA TC601，2023 年 6 月

图 19 AIGC 为数据库结构设计提供建议的示例

数据库结构设计方面，AIGC 技术可以帮助 DBA 前置完成结构设计，DBA 提出简单的场景描述，大语言模型能够返回数据库结构定义，较大程度简化数据库结构的设计工作。从数据库架构设计方面来看，用户可以根据自身需求用自然语言进行场景描述，AIGC 技术能够提出推荐的数据库选型建议。当用户向大语言模型提供一定性能要求后，AIGC 技术还可以反馈推荐的规格和潜在架构优化点，进而有效减少数据库架构师的工作量，提升其工作效率。

来源：CCSA TC601，2023 年 6 月

图 20 AIGC 对数字进行判断的示例

AIGC 技术十分消耗算力，未来硬件发展使得数据库算力不断提升的同时，也会进一步激发数据库潜能。此外，最近同样火热的向量数据库迅速发展，有效支持多模态数据的存储、索引和查询。随着近几年大语言模型（LLM）的发展也扩展了向量数据库的应用场景，AI4DB 技术将会更快地在向量数据库中落地。

6.云计算成为数据库重要驱动力

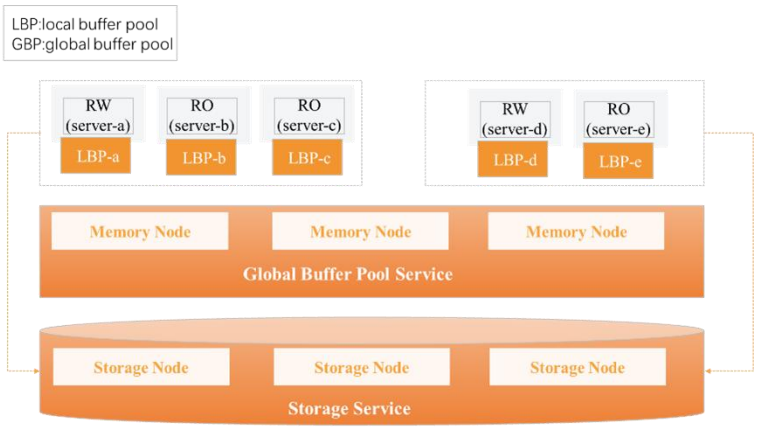
云被视为数字化转型的高度战略性平台，云计算成为数据库发展的重要驱动力。数据库产品及生态工具上云成为趋势，从全球范围看，目前，Elasticsearch、MongoDB、Databricks、Snowflake 等数据库厂商，已与微软、谷歌、亚马逊、阿里云、腾讯云、Clever Cloud、Aiven 等公有云厂商开展合作。**从国内范围看，**近两年部分数据库产品及生态公司如新数科技 ShinData DMP、沃趣科技 QFusion、飞轮科技 SelectDB、玖章算术 Ninedata、涛思数据 TDengine、悦数科技 NebulaGraph 等，已与阿里云、华为云联合推出 DBaaS 版本，持续完善公有云数据库产品及运维体系，为用户搭建高效、便捷、安全的数据库云生态应用场景服务。

DBaaS 提供弹性灵活的数据库管理解决方案，助力企业降本增效。DBaaS 模式最早由亚马逊提出，随后 Oracle、MongoDB、微软、谷歌、阿里巴巴、SAP、Redis Labs、IBM、腾讯、EnterpriseDB、Rackspace 等供应商纷纷推出相关服务。随着建立和管理多云环境正在成为国外用户趋势，互有竞争关系的甲骨文和微软甚至联合推出 Oracle Database Service for Azure，旨在为其共同客户的应用迁移上云降低复杂性，更是为 OCI（Oracle Cloud Infrastructure）在 DBaaS 方面与 AWS 的竞争提供支撑。根据 Forrester 调查数据显示[8]，33%的全球基础设施业务决策者已经在生产环境中部署 DBaaS 版本。企业支持的使用场景类型已大大增加，不仅限于简单的测试、开发和备份，更扩大到错综复杂的客户体验、物联网、移动和大数据等应用领域。**未来，DBaaS 将与其它技术更加深入融合。**随着 DBaaS 技术的普及和成熟，DBaaS 供应商逐步提供一些创新功能。例如通过人工智能技术实现数据库部署、运维、管理全流程的自动化，减少人为干预的同时加快部署，帮助企业迅速构建和支持庞大且更复杂的业务应用程序和操作型系统。

以无服务器架构（Serverless）为核心计算范式的云原生技术飞速发展，云原生数据库取得不断进步。越来越多的云原生数据库通过存储计算分离架构，实现资源池化和极致弹性，具备高扩展性、高可用性、跨地域规模、低成本等优势，可为用户提供真正具备秒级智能弹性扩容能力、随需而动的云原生数据库服务。云原生数据库 Serverless 关键技术以底层池化资源为基础，利用 RDMA 高性能网络高效管理、使用物理资源实现资源池化及弹性扩展、高可用、

[8] 《The Forrester Wave™: Database-As-A-Service, Q2 2019》

高性能、低成本的 Serverless 能力。Serverless 服务大部分以 API 形式提供，无需运维同时用户也无需关注后端使用情况。服务还能实现是实时弹性扩缩容，用户可以像使用自来水一样按使用量进行付费。最初的云数据库主要是模仿线下数据库使用方式，为用户提供数据库托管服务。但云上主机的型号选择并不灵活，很难根据用户业务及资源需求进行协调。云原生数据库计算和存储分离的架构很好地解决了这一问题，这也是数据库 Serverless 化基础。目前存在一些 Serverless 数据库在架构上分为三层，即接入层、计算层和存储层。



来源：《Serverless 数据库技术研究报告》

图 21 一种计算、内存、存储三层解耦架构示意图

云原生数据库可以广泛应用在可变工作负载或不可预测的工作负载场景中，使得用户无需按峰值容量或平均容量预置，从而避免为不常使用的资源付费以及由于容量不足导致的性能问题。在电商、电信运营商、金融等行业中能够帮助企业应对业务洪峰，助力系统平稳运行。未来，云原生数据库将在提升易用性、标准化计算资源、扩容无感知、快速调度资源、提升数据共享能力和数据库智能自治方面不断发展，从而更好地帮助用户降本增效。

公有云厂商发布数据管理服务助力数据价值不断放大，数据库企业收购初创公司布局 IDE 生态。2022 年 12 月，亚马逊在 re:Invent 全球大会上推出数据管理服务 Amazon DataZone，旨在让客户更快、更轻松地对存储的数据进行编排、发现、共享和治理。阿里云推出 DMS 产品提供一站式全链路数据管理与服务，进一步释放云原生技术红利。2020 年至今，MongoDB、Databricks 和 ClickHouse 分别先后收购数据库生态工具厂商 Compass、Redash 和 Arctype。国内 PingCAP 创始人也投资了数据库开发工具企业 Bytebase。各厂商着力打造自己的数据库 IDE，不断提升用户的数据库使用体验。

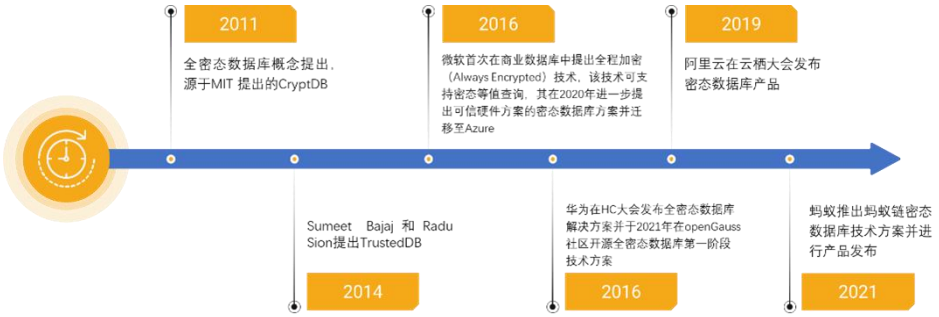
(二)技术融合护航数据要素安全流通

数字经济时代，数据要素被列为和土地、资本、技术和劳动力并列的第五大生产要素。在交易流通过程中数据要素的安全如何保障成为当前技术决策者重点关注的问题。数据安全与数据流通的关系好比矛与盾，更多地流通意味着更多的数据通道暴露，也为数据安全带来更大挑战。隐私计算、区块链及图技术等与数据库技术的结合为数据流通提供了更加安全可靠的解决方案。

1.隐私计算保障密态数据安全流通

隐私计算技术与数据库相结合产生的全密态数据库能够解决数据全生命周期的隐私保护问题，使得系统无论在何种环境下，数据在传输、运算以及存储的各个环节始终都处于密文状态。全密态数据库是指能够提供对应用透明的加解密能力，在数据库系统中数据的全生命周期以密文形式进行处理，同时密钥掌握在授权用户手中的数据库管理系统。当数据拥有者在客户端完成数据加密并发送给服务端后，在攻击者（包括黑客、超级用户等任何角色）借助系统

脆弱点窃取用户数据的状态下仍然无法获得有效的价值信息，从而起到保护数据隐私的作用。



来源：CCSA TC601，2023 年 6 月

图 22 全密态数据库发展历程图

目前，全密态数据库发展尚处于早期阶段。2022 年，CCSA TC601 WG4（数据库与存储工作组）组织编制国内首个全密态数据库技术标准，使业内各厂商对于关系型数据库密态存储与计算的技术架构、基本功能达成初步共识。国内目前以华为云 GaussDB、阿里云 PolarDB 以及蚂蚁科技集团的蚂蚁链数镜产品较为成熟，其中华为云全密态数据库已在华为公司流程 IT ERP 项目中落地使用。近年来，全密态数据库研究已从传统关系型数据库加解密研究，拓展至非关系型数据库如空间数据库的加解密研究。未来全密态数据库的性能提升、搜索型数据库、图数据库等加密技术将会成为专家学者们探索的下一个蓝海。

2.区块链技术赋能数据资产高度可信

近年来，随着数据资产可信流动的需求不断增强，业界对于数据全向追踪管理、防止数据篡改与作弊、实现多方认同的需求越来越迫切。区块链技术具有数据防篡改、数据可追溯、信息全透明、多方地位平等以及数据可共享的技术特征，是数据资产可信流动的必要技术保障，是数字世界不可或缺的根基。区块链技术能够很好

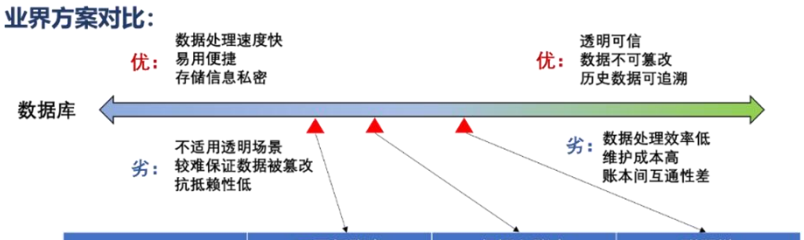
地弥补当前数据库缺乏防篡改能力、无法验证篡改行为、不具备抗抵赖性等问题，二者相结合形成的多方可信防篡改数据库技术方案能够更好地保障云上数据可信运维。

表 5 防篡改数据库典型产品

企业名称	产品名称	特性
华为	云数据库 GaussDB（for openGauss）	保障数据在云上的增、删、改全生命周期可追溯、可校验，为数据完整性提供更强有力的保护，极大降低用户使用门槛和业务切换的难度，全方位实时保障企业数据安全。
阿里云	Lindorm 防篡改数据库	通过构建可信数据结构为用户提供防篡改、可追溯、不可抵赖等完整性保证能力，并可结合三方签名服务使数据具备司法效力。结合 Lindorm 自身宽表、时序、计算、搜索、时空等多模引擎能力，在金融政企、供应链、物联网、车联网等领域，提供一体化、全方位、高可信的解决方案。
微软	SQL Server	通过实现加密保护和提供安全卫士（Security Sentinel）支持防止数据篡改，同时提供身份验证、授权、审计、角色管理等安全控制功能
Oracle	Oracle Database	提供了安全可信赖体系结构，支持内置的数据加密、身份验证、审计和访问控制等多种安全特性，可帮助用户实现防篡改和数据保护。
IBM	IBM DB2	提供了高级的数据安全和加密功能，包括数据压缩、必须的域限制等，同时还支持访问控制和审计，以提高数据的安全性。

来源：CCSA TC601，2023 年 6 月

目前数据库与区块链相结合的技术主要分为两类技术路径，一是单中心账本方案，采用区块链技术增强数据库防篡改特性，可以通过加密验证，不可变且透明，易用性较高。二是多方共识防篡改方案（即联盟链）：有准入机制的多方参与联盟链，联盟链成员使用多方共识共同维护链上数据，使用数据库增加数据存储、处理能力。业界主要技术方案包括仅插入账本、可更新账本以及联盟链方案。



	可更新账本	仅插入账本	联盟链
插入操作	支持	支持	支持
删除操作	任意删除	归档历史数据	-
修改操作	支持	不支持	不支持
中心化部署	支持	支持	不支持
多方协同	支持	不支持	支持
链码	不支持	不支持	支持

来源：华为云计算技术有限公司

图 23 业界防篡改数据库方案对比

未来，区块链技术与数据库技术结合将产生更多火花。区块链技术和数据库技术与可信硬件、高性能共识、KMS、零知识证明等技术不断融合，硬件可信账本、多方可信数据库、三方可信账本以及端侧可信账本等新兴技术将会为信息技术发展带来更多机会。

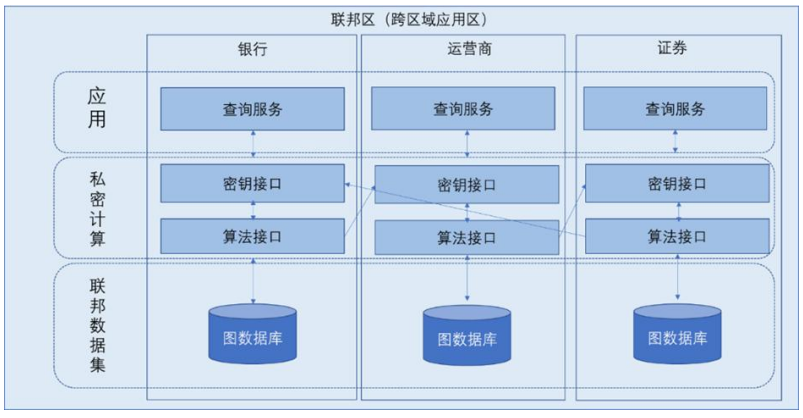
3.图联邦学习技术打破图数据孤岛

图联邦技术是为了解决数据孤岛、隐私保护和数据安全问题提出的概念，在保护用户隐私和公司数据的前提下，更好地发挥数据价值。图数据库（Graph Database）是一种使用图结构进行语义查询的数据库，通常使用属性图模型（包含节点、边和属性）来表示和存储数据。图数据库技术突破了传统关系型数据库对于数据之间关系的束缚，图联邦技术打破了“数据孤岛”的限制，图联邦数据库作为两者的交叉领域，存在着巨大的发展潜力。

来源：浙江创邻科技有限公司

图 24 一种图联邦数据库方案架构示例

梅特卡夫定律（Metcalfe’s law）表明[9]，数据的连通性越完整，获得的价值越高。图数据库因其能很好地处理复杂的数据关系，同时具有高效的复杂关联关系查询性能，因此天然善于处理复杂的网络关系从而帮助数据释放价值。由于技术限制、法律合规等多种因素的制约，传统的图数据库只能缓解企业内部部门之间的“数据孤岛”，对于企业之间的“数据孤岛”现象难以提供有效解决方案。图联邦数据库能够更好地管理、查询、集成和计算跨越不同数据源的图数据，对于促进图数据更好地流通有巨大价值。



来源：浙江创邻科技有限公司

图 25 一种图联邦数据库应用架构示例

[9] 一个网络的价值等于该网络内节点数的平方。即一个网络的价值和这个网络节点数的平方成正比。

图联邦数据库可以应用于社交网络分析、推荐系统、金融风险管理、生命科学等场景。目前，图联邦数据库在应用方面仍面临数据安全、法律合规以及底层数据库性能方面的挑战。如何在保证数据安全、保证合法合规的前提下进行数据共享，是图联邦数据库需要解决的重点问题之一。

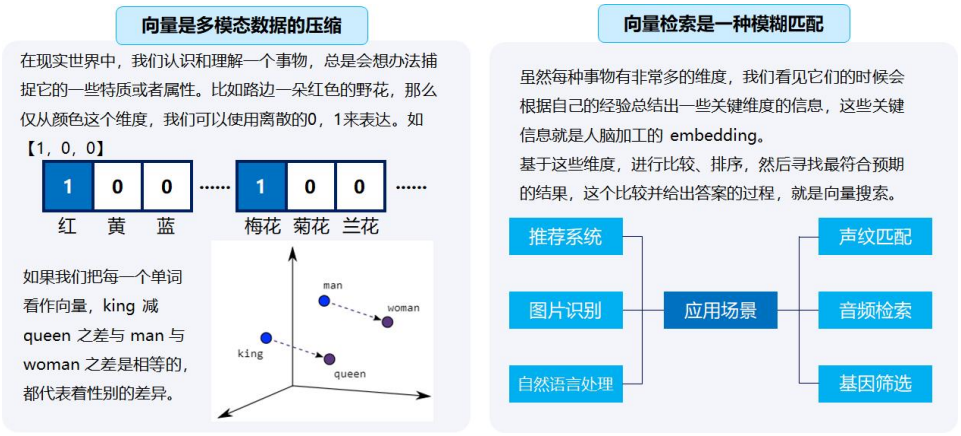
(三)技术革新赋能新兴业务场景

近年来，随着人工智能、云计算等技术不断发展，以及组织数字化转型持续深入，新兴业务场景驱动数据库技术不断革新。2023 年上半年，生成式人工智能（AIGC）引发业界对大语言模型的关注，向量数据库被认为是数据库未来十年最重要的新兴技术之一。智慧城市、智能电网以及车联网等新兴场景下产生的图数据和时空数据也对数据库的数据处理能力提出新需求。

1.AI 大模型催生向量数据库新应用

文本、图像、音视频等海量的非结构化数据占数据总量不断上升，预计 2025 年，将达到八成以上，这些数据需要通过机器学习算法从中提取出以向量为表示形式的“特征”。向量数据库便是为了解决对这些向量进行存储与计算的问题而兴起。向量数据库可以通过将向量的特征进行分组和索引，以实现高效的相似性搜索。同时，向量化技术可以帮助向量数据库将高维向量映射到低维空间，从而减少存储和计算成本。基于索引技术，向量数据库通过自身的各类向量操作，如向量相加、相似度计算和聚类分析等，使得用户能够对向量进行高效搜索。向量数据库的优势在于用统一的形式呈现所有类型的数据，降低了底层数据处理系统的复杂性。近几年大语言模型（LLM）的发展扩展了向量数据库的应用场景，在 LLM 中，

向量数据库可用于存储 LLM 训练产生的向量嵌入（Embeddings）。通过存储数十亿个表示 LLM 的大量训练的向量嵌入，向量数据库执行至关重要的相似性搜索，以找到用户提示和特定向量嵌入之间的最佳匹配。



来源：CCSA TC601，2023 年 6 月

图 26 向量数据库关键技术及应用场景示意

随着向量数据库关注度持续上升，众多传统数据库企业陆续投入资源研究该领域。目前全球已有 70%的向量数据库选择开源发展模式，超过一半的向量数据库具有云化部署能力。向量数据库公司在一级市场上获得众多投资者青睐。国内爱可生向量数据库 TensorDB 完成与昇腾 AI 基础软硬件平台的全面融合，基于昇腾 AI 完成深度优化，达到索引速度 10 倍提升的效果。

表 6 向量数据库企业投融资情况

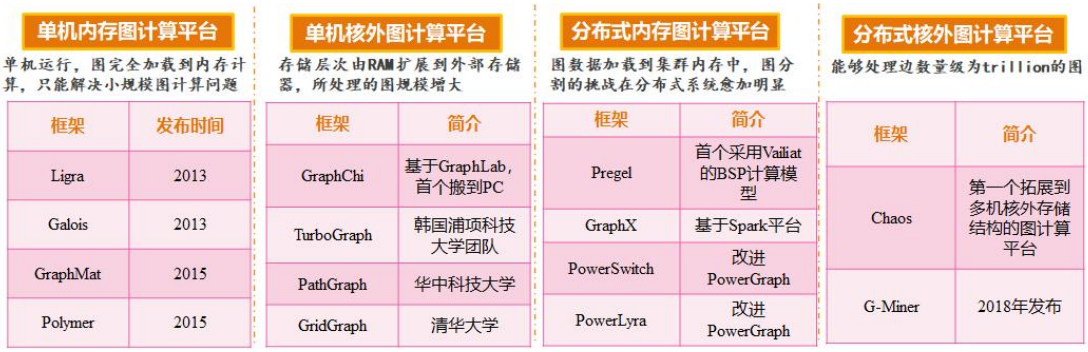
产品名称	所属组织	产品发布时间	投融资日期
Milvus	Zilliz	2019 年开源	2022 年 8 月完成 6000 万美元融资
Vearch	京东	2019 年 10 月	/
TensorDB	爱可生	2020 年	2021 年完成 B 轮融资近亿人民币
Om-iBASE	联汇科技	2020 年	2022 年 1 月完成 B++轮融资
Pinecone	Pinecone	2021 年 4 月	2023 年 4 月 B 轮融资 1 亿美元
Weaviate	Weaviate	2020 年 5 月	2023 年 4 月 B 轮融资 5 千万美元
Qdrant	Qdrant	2023 年 2 月	2023 年 4 月 750 万美金种子融资
Chroma	Chroma	2023 年 2 月	2023 年 4 月 1800 万美金融资

来源：CCSA TC601，2023 年 6 月

未来，向量数据库面临在可运维性、性能成本、离在线一体化、智能化、易用性以及标量数据处理方面的六大挑战。企业也在通过复用基础设施、与 GPU 等硬件相结合、与 Hugging face、OpenAI 等大模型生态对接和标量执行引擎研发等方面不断向更加完善的向量数据库演进。

2.图分析技术洞察数据连接新价值

随着数据自身丰富度不断增加，数据之间的关联性以及如何有效分析和处理数据之间的复杂关系成为从业人员研究的重点。当前图分析技术研究热点主要聚焦在图计算以及图神经网络两个方面。



单机内存图计算平台

单机运行，图完全加载到内存计算，只能解决小规模图计算问题

框架	发布时间
Ligra	2013
Galois	2013
GraphMat	2015
Polymer	2015

单机核外图计算平台

存储层次由RAM扩展到外部存储器，所处理的图规模增大

框架	简介
GraphChi	基于GraphLab，首个搬到PC
TurboGraph	韩国浦项科技大学团队
PathGraph	华中科技大学
GridGraph	清华大学

分布式内存图计算平台

图数据加载到集群内存中，图分割的挑战在分布式系统愈加明显

框架	简介
Pregel	首个采用Vailiat的BSP计算模型
GraphX	基于Spark平台
PowerSwitch	改进PowerGraph
PowerLyra	改进PowerGraph

分布式核外图计算平台

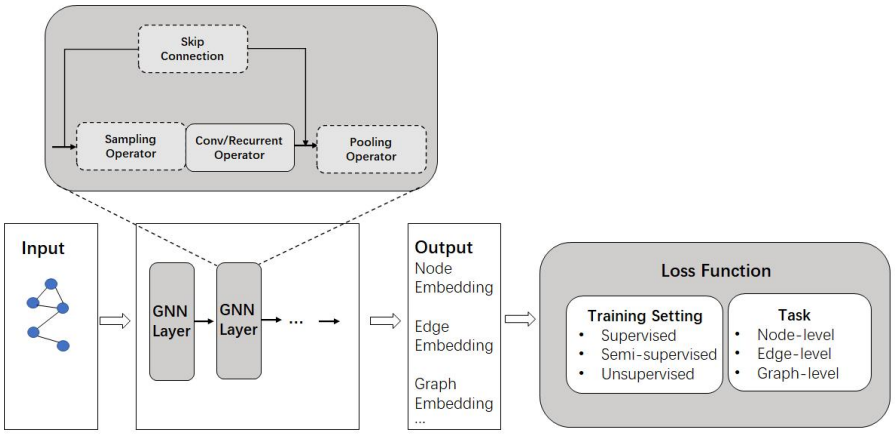
能够处理边数量级为trillion的图

框架	简介
Chaos	第一个拓展到多机核外存储结构的图计算平台
G-Miner	2018年发布

来源：CCSA TC601，2023 年 6 月

图 27 图计算平台分类方式及典型产品

常用的图计算模型有两种：BSP（Bulk Synchronous Parallel）模型和 Pregel 模型。BSP 模型是一种同步计算模型，将计算任务划分成多个超级步，每个超级步包含计算、通信和同步三个阶段。Pregel 模型是一种异步计算模型，将计算任务划分成多个迭代步骤，每个迭代步骤包含计算和消息传递两个阶段。



来源：Graph neural networks: A review of methods and applications

图 28 GNN 模型的一般设计流程

图神经网络（GNN）也是当前图机器学习最火的分类之一。传统神经网络主要是基于欧几里得空间的向量数据，其输入数据是经过预处理后的向量，通过层层传递计算，最终输出一个预测结果。而在图数据中，节点之间的关系通常是非线性的，所以需要一种能够处理图数据的神经网络模型，即图神经网络，它主要应用于节点分类、图分类、链接预测等任务中。尽管 GNN 在诸多领域取得巨大成就，但 GNN 模型在在鲁棒性、可解释性、图预训练以及复杂图结构方面仍面临多重挑战[10]。

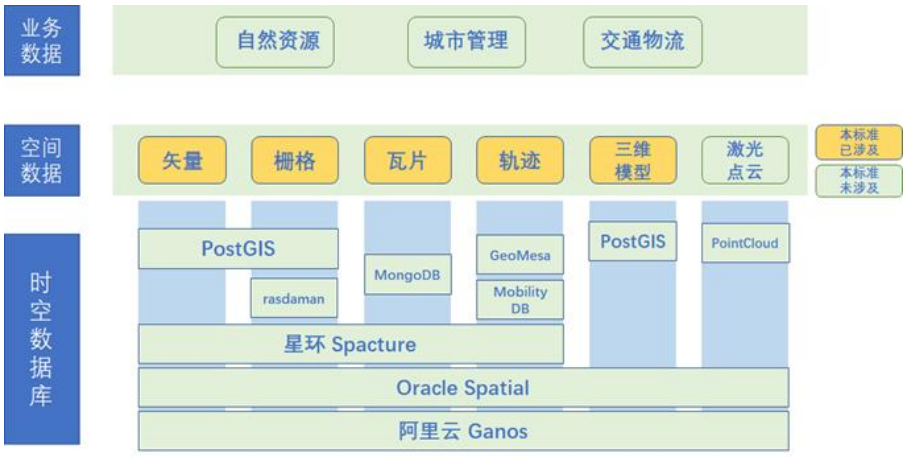
目前，一些图数据库已提供原生的图分析能力，无需将数据导出到外部计算平台，在图数据库内部即可完成图分析任务。相比于依赖外部计算平台的架构，原生的图分析可以免去同外部计算平台进行数据导入和导出的巨大开销，同时将计算的实时性由天或小时级别提高至分钟级甚至秒级。随着算力不断提升以及大模型技术持

[10] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun,Graph neural networks: A review of methods and applications,AI Open,Volume 1,2020,Pages 57-81,ISSN 2666-6510.

续发展，图计算技术与图神经网络技术将会让图数据的价值进一步得到释放。

3.时空数据库释放时空数据新潜能

时空数据指在统一的时空参考下地球或者其它星体上的所有与位置有关的地理要素或者现象的数据集合。现实世界中超过 80%的数据与地理位置（空间）相关，而所有数据均含有时间属性[11]。实际业务场景中很多数据需要通过时空数据引擎进行处理。因此，能够实现海量时空数据管理、查询、统计与分析的时空数据库应运而生。



来源：CCSA TC601，2023 年 6 月

图 29 国内外典型时空数据库产品

时空数据库主要针对矢量、栅格、瓦片、轨迹、三维模型和激光点云空间数据进行处理。为了提高数据库对时空数据的管理能力，各家数据库厂商也面向不同需求开出了不同的引擎。国外如 Refractions 基于 PostgreSQL 开发的 PostGIS 可以对矢量、栅格及三维模型数据进行处理，Oracle Spatial 可以对全量空间数据类型进行处理。国内以阿里云 Ganos 引擎和星环 Spacture 为代表，在传统数

[11] Franklin, Carl and Paula Hane, “An introduction to GIS: linking maps to databases,” Database. 15 (2) April, 1992, 17-22.

据库基础上增加了对于时空数据的动态感知能力，更好地对于时空数据进行处理分析以支撑决策。

时空数据库能够通过一库统管的方式对于不同格式的数据进行处理，打破传统时空数据处理平台限制。2023 年上半年，中国信通院联合二十余家单位共同制定时空数据库技术标准，标准包括几何对象管理、影像与格网对象管理、移动对象管理、表面网格对象管理及地理网格对象管理五大能力域。未来时空数据库通过多模融合处理、与 AIGC 深入融合等方式，更好地释放时空数据价值。

三、数据库行业应用情况综述

数据库是应用系统运行的关键基础软件，近些年随着各行业数字化转型不断加速，我国数据库正朝着由边缘系统至核心系统、由重点行业向全行业应用铺开，下文以金融、电信及制造业为例，分别阐述我国数据库应用创新实践情况。

(一)金融行业核心系统改造升级进度加快

数据库作为金融系统的核心基础设施，历经数十年发展，为金融行业经营战略转型升级提供了有力的技术支撑。在战略指导下,国内金融机构积极探索分布式数据库在金融业务中的应用，并已经开始尝试在核心交易系统中进行分布式改造，取得显著成果。近一年，我国数据库在金融行业核心交易系统不断取得积极成果。

从技术架构看，金融行业使用的数据库仍以集中式为主，分布式数据库在中大型金融机构形成了有力补充。《金融业数据库供应链安全发展报告(2022)》调研数据显示，集中式数据库在金融业总体占比仍高达 89%，其中银行 80%，证券和保险业占比均超过 90%，集中式数据库在金融科技数字化进程中扮演重要角色。金融行业分

布式数据库总体占比达到 7%，银行业超过了 17%，证券业和保险业相对较低。此外，金融业逐步开始探索应用云数据库，且主要以私有云为主，《金融业数据库供应链安全发展报告(2022)》调研数据显示，云数据库在金融业占比大致在 3.97%。

金融行业在近几年的数据库迁移改造中，集中式数据库仍发挥着重要作用，新技术分布式和云原生成为新选择，共同推进了数据库在金融行业的广泛应用实践。除此之外，以图数据库为代表的新型数据库近年来也在金融行业应用实践中崭露头角。随着互联网金融兴起，图数据库成为金融机构在风险管理、反欺诈、推荐系统和市场分析等系统的关键选择。

(二)电信行业三类系统适配迁移加速推进

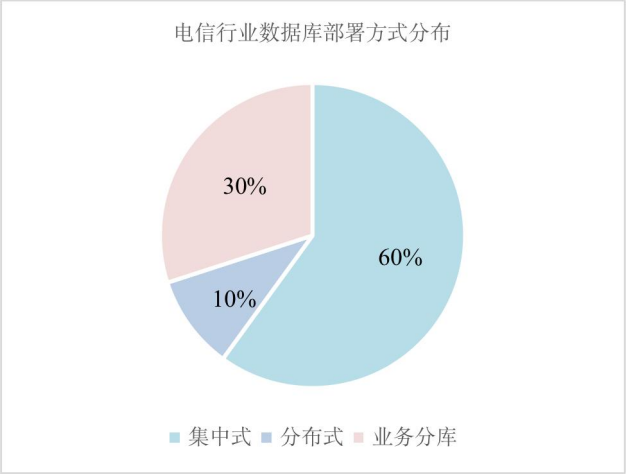
电信行业作为数字中国建设的基础性、战略性、先导性产业，是新型数字基础设施建设者和服务提供者；同时电信行业业务复杂，部分核心应用对性能和高可用性的要求极高，通过 IT 监管环境、数据业务复杂性、核心业务数据类型、成本敏感性四个维度对电信行业支撑系统体系三大域进行分析，分析结果如下表所示。

表 7 电信行业支撑体系三大域分析

	B 域	O 域	M 域
IT 监管环境	强	强	一般
数据业务复杂性	复杂	一般	弱
核心业务数据特点	强事务+分析	强事务	强事务
成本敏感性	一般	一般	一般

来源：CCSA TC601，2023 年 6 月

出于对数据安全等因素考虑，当前我国电信行业几乎全部采用私有云或自有机房部署的方式。在数据库部署类型方面，集中式数据库仍占据主导地位。B 域作为支撑体系的主要业务承载域，其业务数据复杂且有强事务要求，目前主要采用在事务一致性、维护等方面表现突出的集中式数据库。单库集中式部署占 60%左右。当面临对 B 域大量流水类冷数据，以及应对业务相对简单的 O 域和 M 域，通常采用分析型数据库。此外，分布式数据库近年逐渐成熟，其平滑扩展的特性适用于业务弹性较大的业务系统，但总体占比较少，在整个支撑体系中不超过 10%。早期分布式数据库不成熟、事务控制复杂、网络开销大等问题导致原生分布式数据库并未得到广泛应用，电信行业绝大多数业务数据具备业务隔离特性，通过业务层发起的纵向分库可以较好解决单库负载问题，因此基于业务分库模式的分布式部署方式在部分负载较大的核心业务系统得到应用。



来源：CCSA TC601，2023 年 6 月

图 30 电信行业数据库部署方式分布

自 2018 年起，我国数据库产品在电信行业非核心业务系统的应用逐步增加。公开资料显示，过去一年，三大电信运营商在各自支

撑系统中不断上线我国数据库产品。未来电信行业数据库发展趋势主要有三大方面，一是我国数据库应用创新进程将迈入深水区，对大型核心生产系统的适配改造将加速推进，二是数据库产品交付后供应商的维保、服务等将成为运营商选择数据库产品的重要考量点；三是运营商纷纷升级或发布自有数据库产品的同时，各类专业数据库将在特定场景得到应用，未来电信行业的数据库选型将进一步呈现多元化、专业化态势。

(三)制造业数据库创新应用具备广阔空间

制造行业是立国之本与强国之基，顺应时代发展走向，推动制造业的数字化发展，增强数字技术与自身业务的融合，加速数字产业与制造业的相互融合，不仅可以为企业发展持续注入新动力，更能持续推动我国制造业高质量发展。

受我国工业发展基础影响，我国工业数据环境面临着数据量不断激增、数据类型复杂、数据治理难等问题。导致一是数据量爆炸性增长对数据库性能提出新的要求，二是数据库需要融合行业特征提供特定的计算模型以适配业务需求，三是工业领域我国数据库产业生态尚未完全建立。目前工业领域各系统存在着相互隔离、信息孤岛等问题，数据库方面也没有形成统一的技术标准、服务标准、管理标准和安全标准，尚未实现在系统兼容、数据共享、信息安全以及互联互通等方面的模式创新。

表 8 制造行业典型系统及数据库类型分布情况

业务系统类型	业务系统	主流国外厂商	国内厂商	数据库类型
研发设计	EDA、PDM、PLM 等	Cadence、Synopsys、Siemens	概伦电子、华大九天、思尔芯等	Oracle
生产制造	MES、YMS、EAP、RTD、PMS、QMS 等	西门子、霍尼韦尔、GE、IBM 等	华为、上扬软件、赛美特、哥瑞利等	Oracle、DB2、

				SQLServer
经营管理	ERP、SCM、SRM、CRM、WMS 等	SAP、Oracle、Salesforce	用友、金蝶、鼎捷、浪潮等	Oracle、HANA

来源：CCSA TC601，2023 年 6 月

当前制造业在数据库应用创新方面呈现如下特点，一是工业领域数据量激增，数据库需求大，市场广阔。二是工业领域部分场景已开始试点运行我国数据库产品。国内已有 30 多款时序数据库产品在新能源发电、储能智慧运维等场景中落地应用。

随着新一代信息技术与制造业的深度融合，工业大数据在制造业的发展具备了一定的技术基础，未来将呈现三大趋势，一是向工业云和边缘计算发展。二是更加重视数据治理和质量控制。三是分布式数据库助力工业云化发展。随着工业互联网深入发展，企业需要将数据集中存储在云端，以便数据的共享和协同工作。分布式数据库能够支撑多节点的数据存储和管理，提高数据的可靠性和可扩展性。

四、总结与展望

在全球数字经济浪潮下，数据库作为承载数据存算的关键数据技术，正经历又一轮发展热潮。当前，我国数据库行业市场前景广阔，产业欣欣向荣，正在经历由“数量型”向“质量型”关键转变期。

产业层面看，全球数据库市场稳步增长，理论技术推陈出新，我国数据库企业增速仍在高位，开源数据库产品不断丰富，在云原生数据库、图数据库、全密态数据库等新兴赛道实现引领发展。数据库未来将支撑着更多关键业务系统运行和海量数据价值挖掘，这对于数据库的架构设计、生态建设等方面提出了更严苛的要求。**技术层面看**，数据库与 GPU、RDMA、NVMe 等新兴硬件不断融合发展，打破传统数据库边界，持续为在线业务创造价值。数据库也在持续与以 AI 大模型为代表的人工智能技术、云原生技术、区块链技术和隐私计算技术等新兴 IT 技术有机融合，以满足日益变革的新兴业务需求。**从应用侧看**，我国数据库应用创新实践迈入新阶段，其应用范围已从对能力需求较低的办公、邮件等外围系统，逐步向金融、电信等关键行业中，对性能需求极高、稳定性要求极强的账务、调度等核心系统深入。

“水滴石穿，日月生辉”。我国数据库产业发展过程中，需凝聚产业链各方力量，久久为功，持续推动数据库技术创新、生态培育和应用落地，从而支撑数字中国建设，构筑国家竞争新优势。

大数据技术标准推进委员会

地址： 北京市海淀区花园北路 52 号

邮编： 100191

邮箱： TC601@CCSA.org.cn

网址： www.tc601.com