

# 中国数据库行业研究报告

2021年



**中国数据库市场规模：**据艾瑞统计，2020年中国数据库市场总规模达247.1亿元，同比增长16.2%。2020-2022中国数据库市场预计将呈高增长态势，由多方面因素促成：1）政策利好，国家大力鼓励国产数据库厂商的发展；2）需求拉动，国产化和数字化转型带动需求的爆发增长；3）供给端传统、初创和跨界各类型厂商厚积薄发，产品和技术经历了多年工程实践的打磨走向成熟；4）国内企业对基础软件的付费意愿和IT支出占比在逐年提升，有利于市场的长期发展。



**中国数据库市场格局：**1）多类型数据库百花齐放，关系型占据绝对主流。从营收角度，2020年中国关系型数据库的市场份额达90%左右，NoSQL数据库更多地基于开源模式，产生二开和服务的费用。2）借助政策东风，国产厂商厚积薄发，市场版图快速扩张。受国产化影响，2020年国外数据库厂商的市场份额下降至52.6%，达梦金仓等传统国产厂商的市场份额上升至7.1%。3）公有云数据库增速放缓，未来仍有一定渗透空间。2020年中国公有云部署模式的数据库市场份额占比达32.7%，预计到2025年将达到47.2%，云厂商将成为中国数据库市场市占率最大的阵营。4）以NewSQL/NoSQL/SQL on Hadoop为典型路线的初创厂商不断涌现，成为中国数据库市场增长率最快的赛道，预计未来五年有10倍以上的成长空间。



**中国数据库市场挑战与趋势：**约2015年起，中国数据库市场进入了百花齐放、活跃创新的阶段，但仍面临多方挑战：1）分布式数据库在事务、性能等环节仍待优化；2）信创为国产厂商提供成长沃土，未来发展仍待市场磨炼；3）数据频繁迁移、多库长期并存为厂商提出新的诉求；4）CPU、内存等硬件变化为数据库设计提供更多的想象空间。综合供需两端视角，中国数据库未来发展将呈现多种趋势：1）多场景现状与融合需求长期并存；2）云数据库（包括公有、非公有各种形式）成为主流；3）湖仓一体服务联机交易和联机分析；4）开源成为产业互联网时代数据库厂商的破局之刃；5）人工智能延伸DBA的能力半径，优化数据库性能，是数据库下一步发展的目标。

产品与技术：数据库内涵与分类 1

供给与需求：数据库市场现状与选型 2

案例与启示：数据库典型厂商案例 3

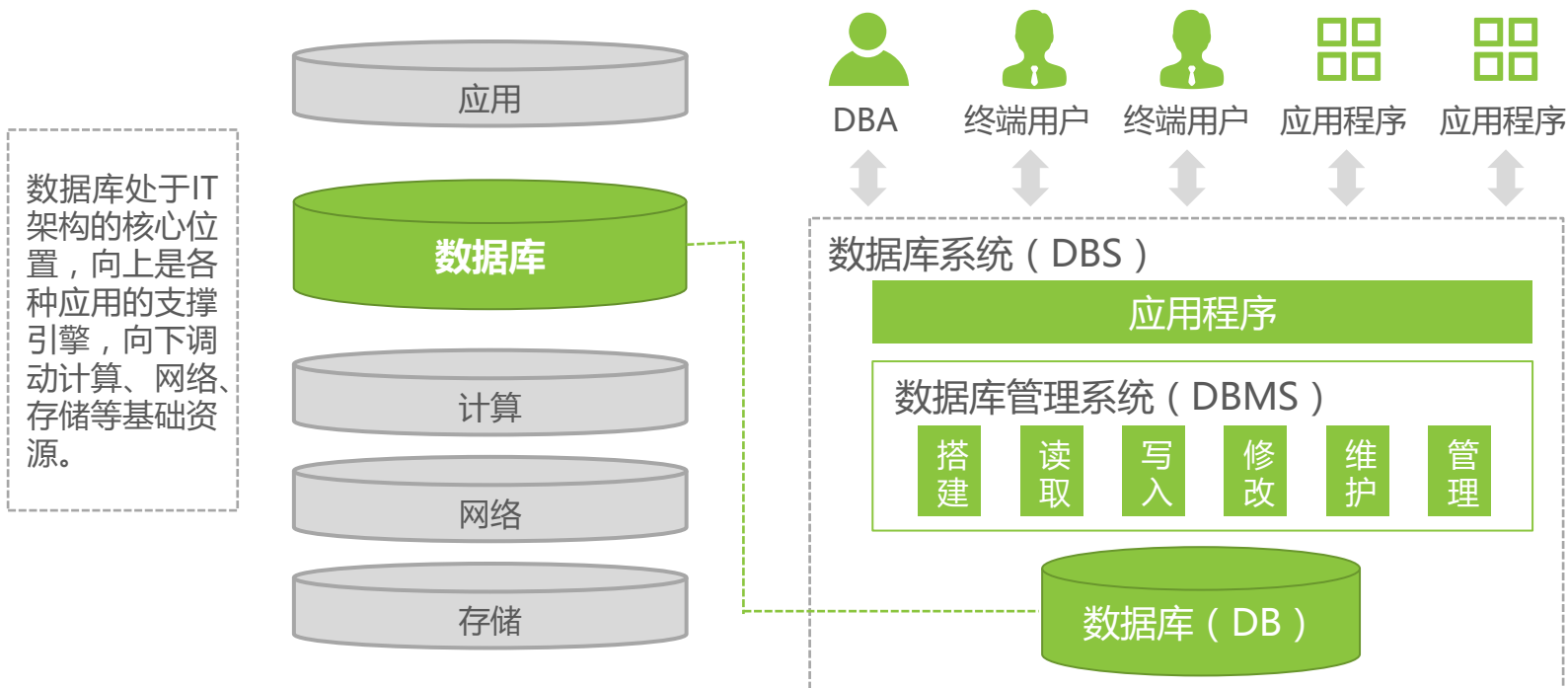
机遇与挑战：数据库未来发展趋势 4

# 数据库定义

## 数据库是由DBMS搭建管理的数据及数据间关系的集合体

数据是数据库中存储的基本对象，包括数字、图像、音频等形式，在进行逐级抽象后存储在数据库中。数据库是由特定软件，即数据库管理系统（DBMS）搭建、处理、维护的数据及数据间逻辑关系的集合体。它面向多种应用，可以被多个用户、多个应用程序所共享。DBMS是负责数据库搭建、使用和维护的大型系统软件，它对数据进行统一控制管理，以保证数据的完整性和安全性。数据库和数据库管理系统共同组成了数据库系统（DBS）。

### 数据库系统架构



# 按数据结构分类

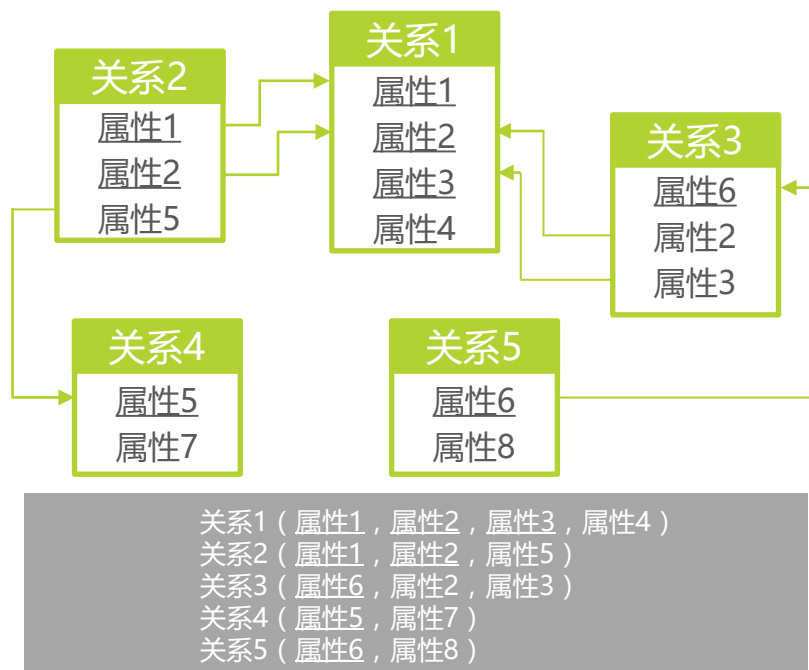
- 传统关系型数据库
- NoSQL数据库
- NewSQL数据库
- 多模数据库

# 传统关系型数据库

## 保证ACID特性，是当今主流的数据库类型

随着市场和技术的发展，关系模型因其特有的原子性、一致性、隔离性和持久性优势，取代了层次模型和网络模型，成为了当代主流的数据模型。关系型数据库建立在关系模型上，是多个关系（Relation）即二维表的集合。每个表有唯一的名字，表的每一行代表了一组值之间的联系，称为元组（Tuple），每一列是实体的描述，具有相同的数据类型，称为属性（Attribute）或者字段（Field）。为了唯一标识一张关系表，关系型数据库选定主码/键（Primary Key）来区分不同元组的候选码；同时为了维护数据库的完整性和数据的一致性，设置外码/键（属性）、参照关系（表）建立表之间的联系。

关系型数据库架构



注释：1、下划线表示主码；

2、此处关系型数据库指传统关系型数据库，支持横向扩展的关系型OLTP数据库（NewSQL）在后续讨论；

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

关系型数据库ACID特性

原子性  
Atomicity

为避免纠纷，数据库中的事务执行被视作原子不可再分，事务（例如转账）中的操作要么全部执行要么失败回滚（Rollback），一般通过日志机制实现。

一致性  
Consistency

为保证业务逻辑的一致性，数据库通过设置约束和触发器来保证其完整性约束不被破坏，即每个事务能够看到的数据总是保持一致。

隔离性  
Isolation

为防止事务之间的脏读、幻读、不可重复读，数据库通过加锁，保证多个事务并发访问时，事务之间是隔离的，互不干扰。

持久性  
Durability

为防止意外事故（例如断电）导致数据缺失，数据库保证事务对其所作的修改被永久保存，不会被回滚。

# 关系型数据库语言：SQL

## 作为RBDMS的标准语言，用于定义、查询、修改和管理

查询即向RBDMS寻求特定的信息，SQL（结构化查询语言），是RBDMS的标准语言，广泛应用于各主流关系型数据库，它包括DDL（数据定义语言）、DML（数据操纵语言）、DCL（数据控制语言）和TCL（事务控制语言）。SQL作为一种声明式语言，同时具有较好的可扩展性，不仅用于查询，还可以用来定义数据结构、插入、修改和删除数据、执行管理任务（安全、用户管理.....）等。

### SQL主要构成

类型	SQL			
	DDL	DML	DCL	TCL
任务	用于定义数据库的外模式、概念模式、内模式及其相互的映像，定义数据的完整性和约束	让用户或程序员使用，实现对数据库中数据的操作，包括过程式DML和声明式DML两类。	用来设置、更改数据库用户或者角色的权限	对关系型数据库管理系统里可能发生的事务进行处理
命令	CREATE ALTER DROP TRUNCATE COMMENT RENAME	SELECT INSERT UPDATE DELETE MERGE CALL EXPLAIN PLAN LOCK TABLE	GRANT DENY REVOKE	COMMIT SAVEPOINT ROLLBACK

### SQL语言特点

#### 语言功能一体化

SQL语言集数据定义、数据操纵、数据控制、事务控制功能于一体，在一次操作中可以使用任何语句，为数据库开发提供了良好的环境。

#### 高度非过程化的编程语言

SQL语言不涉及存取结构以及具体的执行过程，因而简化了编程的复杂性，提高了数据的独立性。

#### 面向集合的操作

SQL语言的操作对象和输出结果都是元组集合，可以实现对一组记录的增删查改。

#### 自含式+嵌入式语言

SQL作为自含式语言，可作为联机交互使用，每个SQL语句可以独立完成其操作；作为嵌入式语言，SQL语句可以嵌入到高级程序语言中使用。

#### 简洁易用

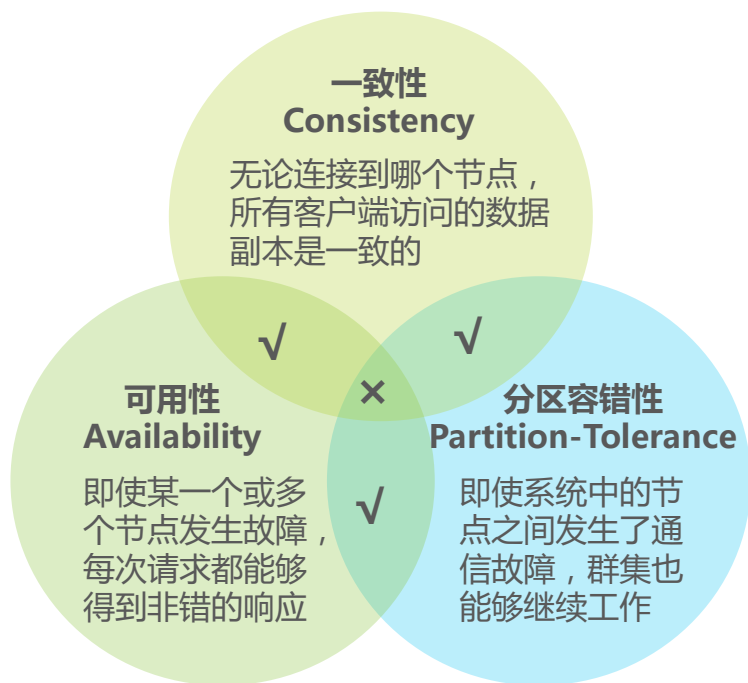
SQL用简单的语句（核心命令：CREATE, DROP, ALTER, SELECT, DELETE, INSERT, UPDATE, GRANT, REVOKE）和接近英语口语的语法完成了数据库创建管理维护的复杂功能。

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 针对数据库高可用、可扩展的需求，NoSQL兴起

NoSQL即Not Only SQL，NoSQL数据库指那些不使用关系模型、分布式、不保证遵循ACID原型的数据库。关系型数据库通过“强一致性”来避免数据库应用中出现的写入冲突（两个客户端同时修改一份数据）和读写冲突（某客户端在另一个客户端执行写入操作过程中读取数据）。“CAP定理”阐述了数据库系统的权衡问题，即当有可能发生“网络分区”时，必须在数据的“可用性”与“一致性”之间权衡。电商、社交网络等场景的容错度较高但需要实时可用，NoSQL数据库由于只要求达到“最终一致性”，可以轻松处理海量数据并实现高用户负载的扩展，在此类场景下应用较广。

### 数据库设计CAP定理



### NoSQL数据库BASE原则

基本可用  
Basically Available

NoSQL数据库在出现故障时，允许损失部分可用功能，或者降低响应速度，从而保证核心功能可用。

软状态  
Soft state

允许系统中的数据存在中间状态，即不同节点的数据副本同步的过程中存在不一致，并认为中间状态的存在不会影响到系统整体的可用性。

最终一致性  
Eventual consistency

在系统保证没有新的操作、无故障发生时，经过一段时间，数据库中的最终数据能够达到一致。其延迟的时间取决于网络延迟、系统负载和数据复制方案设计等因素。

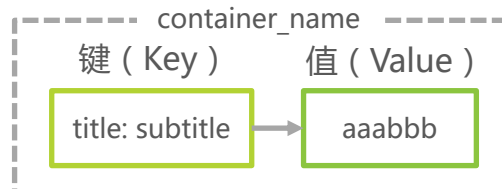


# NoSQL数据库存储结构

## 采用键值、宽列、文档、图等结构灵活存储

NoSQL数据库使用不同的数据存储模型来满足不同的场景需求，当今主流的NoSQL存储模型有键值对存储、宽列式存储、文档型存储和图形存储，以及扩展的RDF、时序、搜索引擎等。它们基于不同的场景需求，提出了相应的存储架构，从而满足传统关系型数据库所无法覆盖的场景。但是采用这些模型的 NoSQL 数据库并不提供规范化，本身在设计上是模式自由的（schema-free），因而保证了存储的敏捷性，可以依据业务变化而调整。

### 典型NoSQL数据库介绍

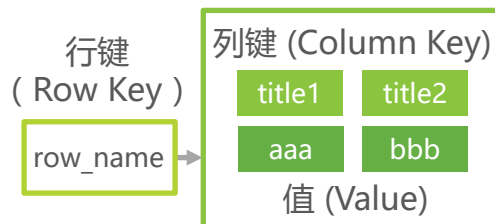


#### 应用场景：

Web应用程序和会话、PUB/SUB、内存中的数据缓存、购物车等

#### 键值数据库

键值 (Key-Value) 数据库以简单的键值对方式来存储数据，键是唯一的标识符。在这种存储结构下，数据库查询时只需要寻找单个键并返回其对应值即可，它通过牺牲数据的结构性，大大提高了读写的速度。

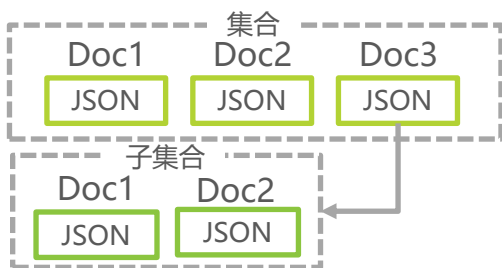


#### 应用场景：

时间序列、历史记录、地理信息等

#### 宽列数据库

宽列 (Wide-Column) 数据库将数据存储在记录中，以行键唯一标识该记录中的列，同时其一行中包含大量动态列，因此可以将其视为二维的键值数据库。它解决了部分列操作、数据压缩（不存储null）和数据过滤的问题。

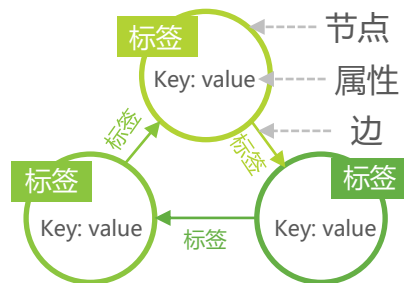


#### 应用场景：

内容管理、APP、目录、日志文件等

#### 文档数据库

文档 (Document) 数据库以文档格式（JSON、BSON、XML或YAML）储存和查询。文档没有一致的格式，因此文档数据库具有直观的数据模型、动态灵活的架构、横向可扩展的优势，但同时牺牲了一定的安全性和可靠性。



#### 应用场景：

社交网络、知识图谱、搜索引擎等

#### 图数据库

大数据时代带来两个明显的变化，即数据量的剧增和数据关联的复杂化。图 (Graph) 数据库使用图结构进行查询，并通过节点、边、标签和属性等方式来存储数据，可以较好的模拟现实世界中的实体及其复杂关系，具有敏捷、可扩展和高性能的特性，在大数据时代得到了广泛的应用。

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 关系型数据库 vs NoSQL数据库

## 在标准化一致性与扩展可用性方面各有优势，适用不同场景

20世纪70年代，为减少数据冗余、降低存储费用，关系型数据库诞生。21世纪初，随着移动互联网和新一代信息技术的发展，关系型数据库在大数据处理分析和读写性能方面的局限性逐渐凸显，NoSQL运动开展起来。NoSQL数据库解决了关系型数据库只能垂直扩展（即在硬件方面增强）的限制，通过分库分表的方式实现水平扩展，满足不断扩张的业务。两种数据库在数据完整性、横向扩展性、读写可用性、产品成熟性和架构灵活性等方面各有侧重，其适用的场景也有所不同。

### 传统关系型数据库 vs NoSQL数据库

#### （传统）关系型数据库优势

- 满足ACID原则，严格遵守数据完整性约束；
- 具有标准化语言SQL，查询操作方便；
- 二维表数据结构，减少数据的冗余，提高了存储空间的利用效率，维护也比较容易；
- 发展时间长，产品标准化、社区成熟、服务稳定

#### （传统）关系型数据库局限

- 架构刚性，前期需要进行完备的设计，后续修改成本高；
- 缺乏横向可扩展性，需要解决跨服务器JOIN等问题；
- 海量数据和高并发条件下读写效率较低；
- 为维护事务一致性，传统关系型数据库需要付出较大的开销

#### NoSQL数据库优势

- 采用动态架构，无需开发人员提前设定数据架构，可以随时更改，敏捷灵活；
- 可扩展性高，通过横向扩展提高可用性，无明显单点故障；
- 存储模式简单，大多NoSQL数据库可以实现极高的性能

#### NoSQL数据库局限

- 为提高可用性，在数据一致性方面有所牺牲；
- 没有标准化的查询语言，学习和使用成本较高，不适合复杂查询；
- 相关理论和技术成熟度较低；
- 缺乏第三方生态系统，公司需要自己开发BI和分析工具

#### 适用性：

对数据安全性和事务支持方面有高度要求的场景，例如电信、电力、金融机构等行业的核心系统

#### 适用性：

对读写性能要求高，且需要处理非结构化、海量的数据，且数据增长无法预期的场景，例如电商、社交网络、搜索引擎等

## 在底层解决了分布式问题，并满足ACID事务要求

NoSQL虽然在可扩展性和可用性方面表现优秀，但是无法满足事务一致性的要求。许多对数据完整性要求较高且数据量较大的企业，只能在应用关系型数据库的同时，购买功能更强大的硬件，或者定制化开发中间件来实现分布扩展。2011年，451 Group的Matthew Aslett 提出了“NewSQL”术语用以定义新出现的“可横向扩展的OLTP关系型数据库”。NewSQL数据库作为一种新兴并正在发展的解决方案，各产品在架构、特性和功能方面各有不同，当前并无通用的定义。但大多NewSQL数据库共有的两个特点：1）保持NoSQL数据库的可扩展性和高性能，2）采用以SQL为主要接口的关系数据模型，保持事务ACID特性。NewSQL并非颠覆式的创新，而是将业界和学术界已有的技术，例如面向内存(memory-oriented)的数据存储、分片、MVCC（多版本并发控制）、TO（时间戳）并发控制、二级索引、广域网数据复制机制、故障恢复机制等，集中到一个架构内。企业采用NewSQL数据库需要较高的硬件和学习成本，且需要承担产品不成熟带来的未知风险。

### 传统关系型数据库 vs NoSQL vs NewSQL

	传统关系型数据库	NoSQL	NewSQL
可扩展性	纵向扩展	水平扩展	水平扩展
关系模型	支持	不支持	支持
ACID事务	支持	不支持	支持
性能	海量数据读写性能差	高性能处理海量数据	高性能处理海量数据
SQL语言	支持	不原生支持	支持
模式自由	不支持	支持	不支持
OLTP	支持	效果较差	支持
OLAP	轻量查询	支持	支持
成熟度	高	较高	较低

注释：Matthew Aslett在2016年发表的论文中将NewSQL数据库分为三类：全新架构的NewSQL数据库、基于中间件-分片的NewSQL数据库和基于云环境和全新架构提供DBaaS的NewSQL数据库，此页只讨论一、三两类，第二种在后续分布式数据库处讨论。

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 通过改变原数据和存储结构来扩展多模 (Multi-Model)

多模型 ( Multi-Model ) 一词于2012年被 Luca Garulli 第一次提出，是一种可以在多个模型中存储和查询数据的数据库，为异构数据提供了较好的解决方案。数据库扩展原有模型的路径主要有四种：新存储方式+新数据模型、原存储方式+新数据模型、新接口+原存储模型、原存储模型。第一种典型代表是支持XML的数据库，它们使用原生XML方法来高效地存储和查询；第二种的典型代表是文档数据库，通过采用特殊的边集合来扩展图结构中的边信息，例如Arango DB和MongoDB；第三种在原关系型存储层上搭建了新的一层，采用相同的方式存储不同类型数据，但是增加了对新数据类型的增删查改支持；第四种即将所有的数据结构简化为Key-Value形式存储。多模数据库的发展时间较短，当今市场上数据库对多模的支持程度不同，在数据模型、索引、多模查询优化策略等方面的能力参差不齐。

### 多模数据库（典型）扩展路径

	数据库管理系统	原数据模型	扩展支持
新存储方式+新数据模型	PostgreSQL	关系	关系、键值对、JSON、XML、嵌套数据/ UDT /对象
	SQL Server	关系	关系、JSON、XML、图、嵌套数据/ UDT /对象
	IBM DB2	关系	关系、XML、图、RDF、嵌套数据/ UDT /对象
	Oracle DB	关系	关系、JSON、XML、RDF、嵌套数据/ UDT /对象
原存储方式+新数据模型	MySQL	关系	关系、键值对、嵌套数据/ UDT /对象
	ArangoDB	文档	键值对、JSON、图
	MongoDB	文档	键值对、JSON、嵌套数据/ UDT /对象
新接口+原存储模型	阿里Lindorm	原生多模	宽列、时序、文档、搜索引擎
	Couchbase	文档	键值对、JSON

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

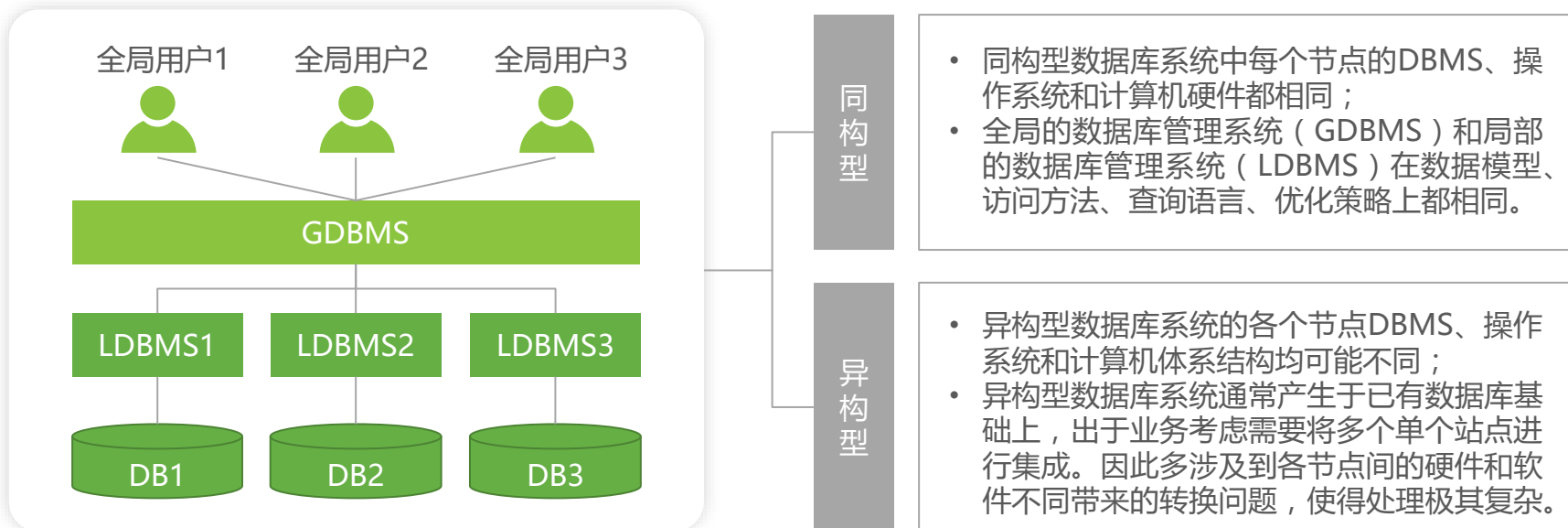
# 按架构分类

- 分布式数据库
- 单机数据库

## 单机性能有限条件下，解决数据量快速增长的最佳解决方案

分布式数据库系统的诞生远早于NoSQL和NewSQL数据库，它产生于20世纪70年代末，在80年代由于计算机功能和网络技术的增强而进一步成长。分布式数据库系统即利用计算机网络将物理上分散的多个数据库连接起来组成一个逻辑上统一的数据库，为业务应用提供完整的联机事务处理。随着数据量爆发式的增长以及应用负载的快速增加，单一服务器模式越来越难应对当今应用对数据和事务处理的需求，分布式成为热门的解决方案。分布式数据库的实现形式大致可以分为同构和异构两种。同构分布式数据库系统中，所有的站点都使用相同的数据架构、DBMS、操作系统和计算机体系结构；异构分布式数据库系统中，不同的站点使用不同的数据模型、DBMS、操作系统和硬件，通过应用程序接口、全局模式和联邦数据库系统结构实现不同数据库之间数据信息、硬件设备和人力资源的合并与共享。

### 分布式数据库实现形式



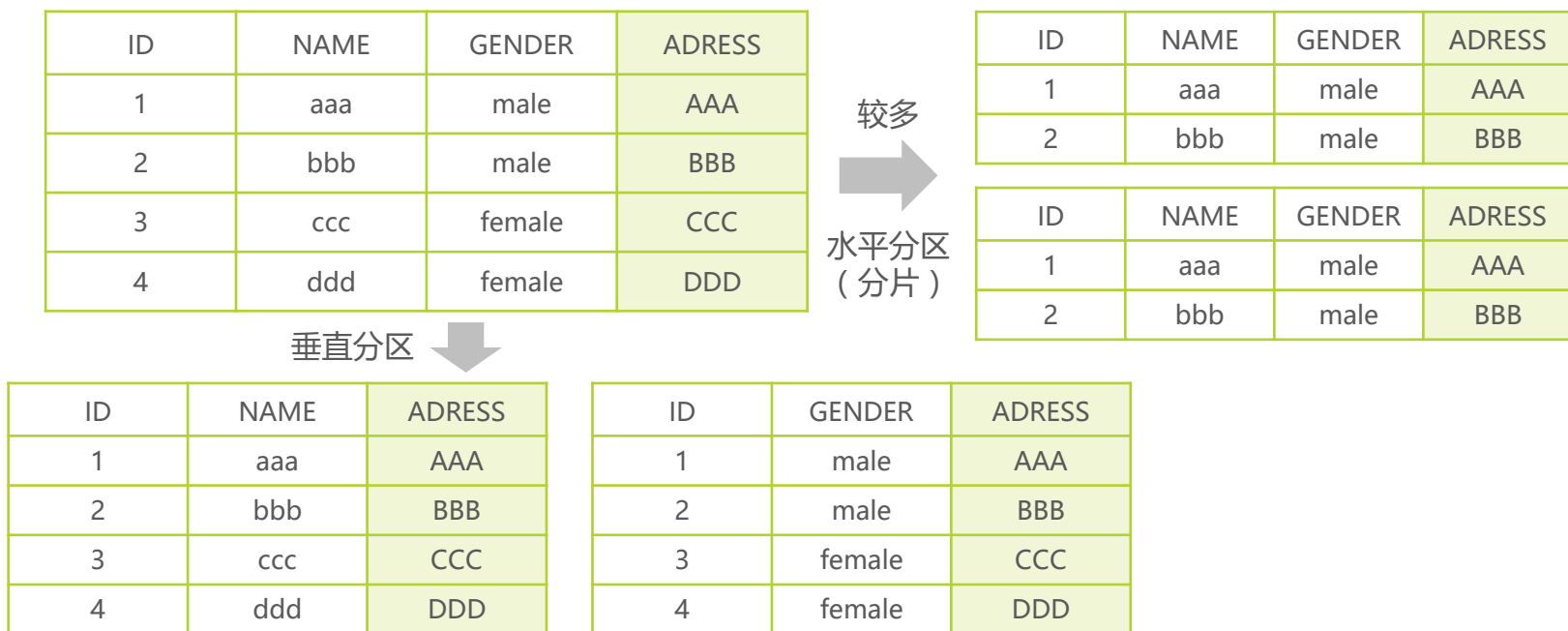
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 分布式核心技术（一）：复制/分区

## 通过“数据冗余”和“数据分割”实现分布式的扩展

分布式数据库的实现方式一般包括两种：复制 (Replication) 和分区分片 (Partitioning/Sharding)。一种是将数据复制到多个服务器上，从而每份数据都能在多个节点中找到；另一种是将不同的数据分片存放在多个服务器中，每一个数据子集都专门由一台服务器负责。复制提供了冗余的能力，包括主从复制（唯一节点负责写入，其他节点保持同步、负责读取）和对等复制（任何节点均可写入，相互协调、同步数据）。随着数据量的增加，出于负载均衡的目的，架构师对数据库进行分区，分区包括垂直分区（列）和水平分区（行）；分片 (Sharding) 是对数据库系统的水平分区，包括基于键值的分片、基于范围的分片、基于目录的分片等方式。

### 水平分区（分片）与垂直分片



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

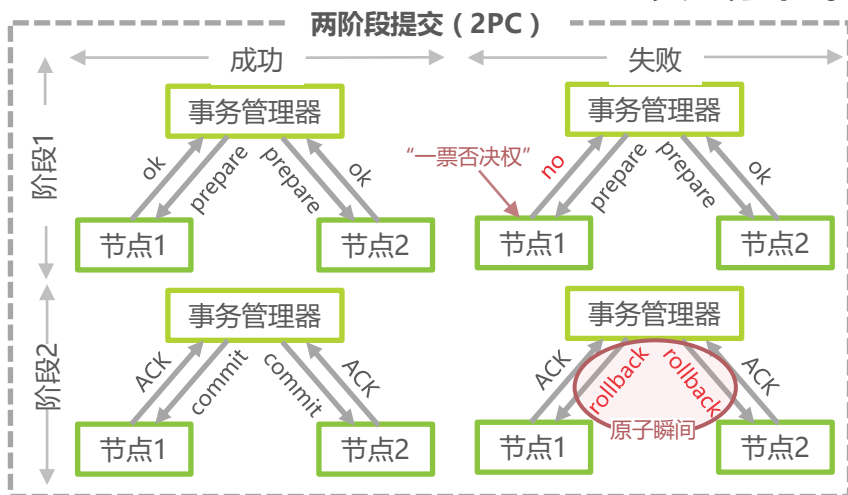


# 分布式核心技术（二）：分布式事务

## 通过机制设计保证分布式环境下的事务ACID特性

单体系统到分布式系统的变化增加了数据库实现ACID特性的难度，但许多环境下企业仍要求较强的一致性。经过多年的发展，各数据库厂商提出了多种分布式事务解决方案，例如两阶段提交（2PC）/三阶段提交（3PC）、TCC方案、可靠消息最终一致性（本地消息表方案-eBay、RocketMQ 事务消息方案-阿里/Apache）、最大努力通知方案等。

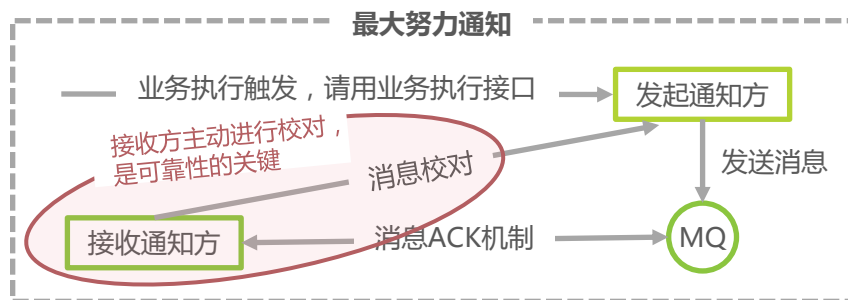
### 典型分布式事务解决方案



**两阶段提交（2PC）**是一种在多节点之间实现事务原子提交的算法，它把事务处理的过程分为prepare-commit两个阶段，增加事务处理器来保证所有节点要么全部提交，要么全部回滚。

#### 优势与局限：

2PC协议原理简单，保证了**强一致性**。但是由于机制完全依赖事务管理器管理且过于悲观导致了单点问题（事务管理器一旦崩溃将全局崩溃）、堵塞问题以及事务处理的延迟。针对以上问题，3PC协议在2PC的两阶段中插入了一个准备阶段并引入了超时机制，解决了2PC阻塞的问题，但仍可能出现数据不一致。



**最大努力通知**关注交易后的通知事务，发起方通过一定机制，最大努力将业务处理结果通知到接收方，若消息接收不到，则接收方主动调用接口查询业务处理结果。

#### 优势与局限：

最大努力通知方案下被动方的处理结果不影响主动方的处理结果，适用于跨企业系统间的操作。它是分布式事务中要求最低的一种，适用于一些仅要求**最终一致性**，且时间敏感度低的业务。

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

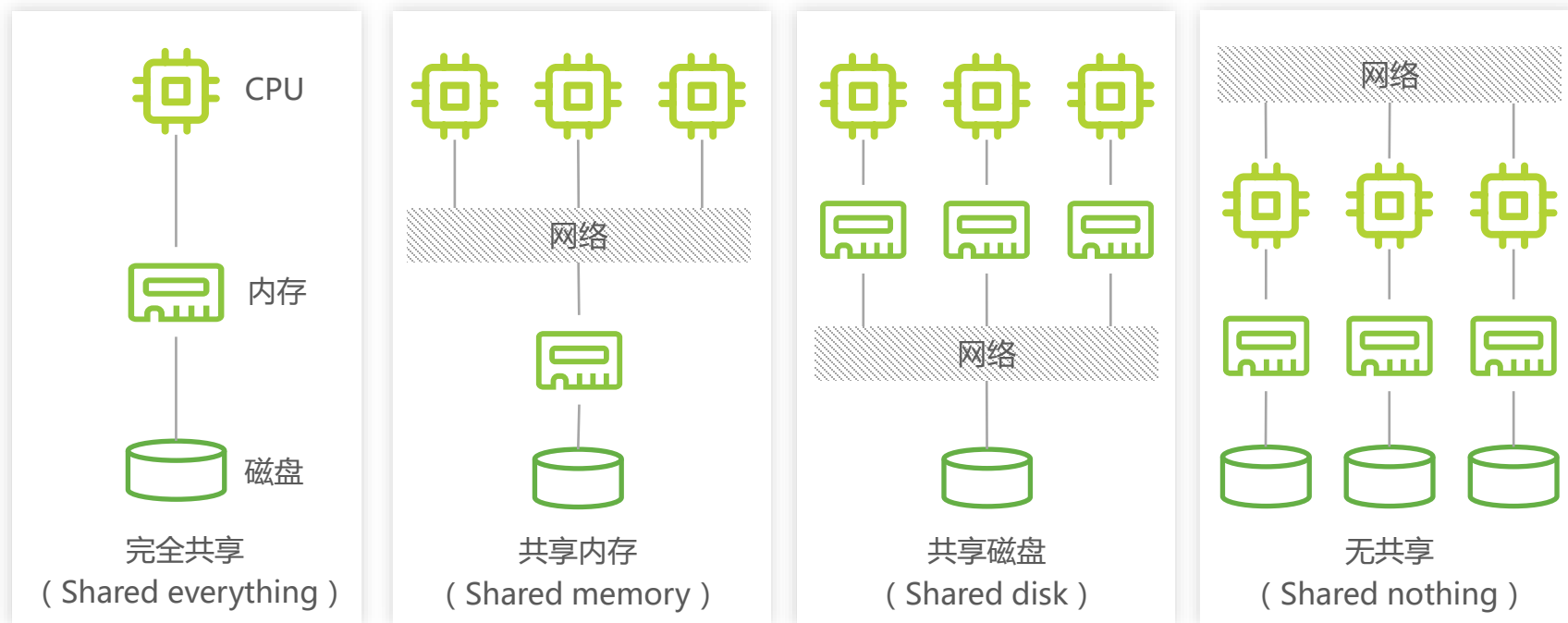


# 分布式架构创新（一）：无共享

## 完全共享→无共享：不同层次横向扩展的实现

数据库构架设计中主要有完全共享 (Shared Everything)、共享内存 (Shared Memory)、共享磁盘 (Shared Disk) 和无共享 (Shared Nothing) 四种。完全共享模式针对单个主机，拥有完全透明共享CPU、内存和磁盘，并行处理能力较差；共享磁盘和共享内存模式允许增加节点提高并行处理能力，但是随着数据量级的扩大，内存访问和网络带宽之间冲突增强，系统处理速度反而变慢。无共享模式下数据库的每个处理单元独立运行，并控制自己的内存和磁盘资源，相互之间通过协议通信。无共享架构并行处理和扩展能力较好，数据库通过增加低成本的计算设备作为系统的节点，获得了无限线性扩展的可能性，因而得到了广泛的应用。

### 分布式架构创新：不同层次的数据共享



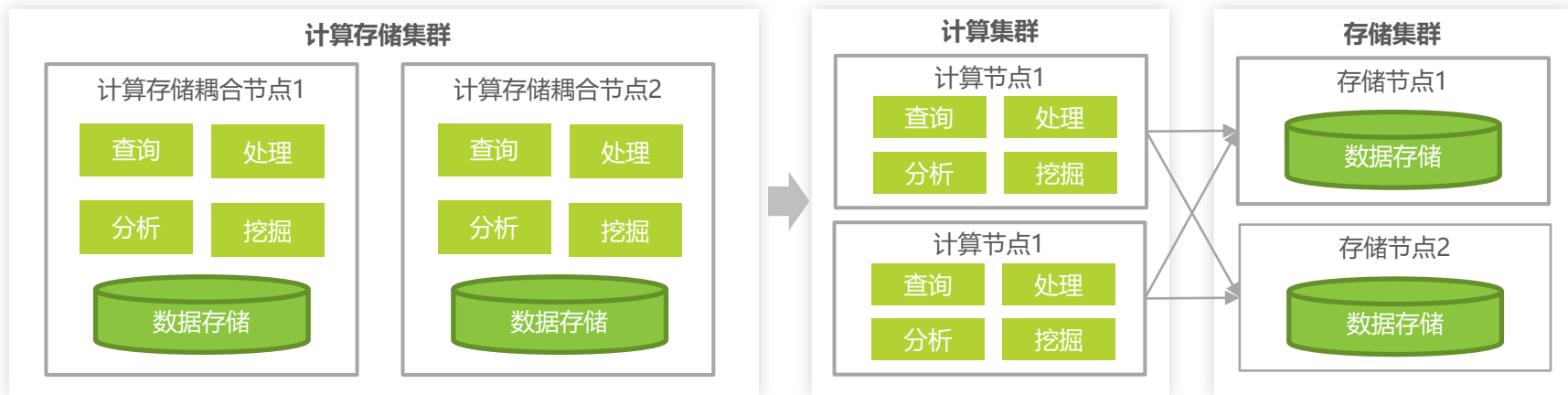
来源：艾瑞咨询研究院自主研究及绘制。

# 分布式架构创新（二）：计算存储分离

## 为分布式数据库资源弹性扩展的诉求提供了新思路

除了基于无共享模式进行分区分片，在云计算时代，一种新的创新架构被提出，即计算-存储分离架构（大多NewSQL数据库采用此种架构）。近十年互联网的发展，网络的性能得到了大幅度的提升，高效压缩算法和存储结构的优化也减少了IO数量，在数据本地化优化较好的数据计算集群中，大量网络带宽处于闲置状态，然而存储和计算耦合的架构不能很好的实现弹性。云计算提供了解决思路，它的核心思想包括分层和虚拟化：对IT架构分层后，每一层可以按各自的能力进行极限扩展；虚拟化后按租户隔离，可以提供高效率的弹性计算，降低了成本。计算-存储分离架构即“云”的模式和形态之一，将数据计算和存储进行分层，并通过高速网络连接。在这种架构下，数据库可以更加充分的利用不对称的存储资源和计算资源，让不同层都可以按照各自最优的模式进行横向扩展。

### 计算存储分离架构



比较：

计算和存储在一个集群里，性能表现较好

比较：

架构灵活、易扩展；优化利用率，有效的降低了成本

# 分布式数据库 vs 单机数据库

## 分别适用于海量数据场景和核心系统的高性能高可靠处理

单机数据库是企业的最初选择，它可以利用位于系统中心的服务器统一管理所有的共享资源，并处理来自用户的请求。单机数据库积累了大量的实践经验，在强一致性、稳定性、迁移成本和运维管理方面都更胜一筹，而且各资源独立，应用隔离性好，数据安全性高。分布式数据库在灵活性和扩展性方面具有优势，一方面分布式给予了每个部门根据其应用程序的特定需求选择软硬件的自由，不必因为共享IT架构而做出妥协；另一方面分布式IT架构天生自带可扩展属性，能够根据业务规模实现无限弹性扩展。

### 分布式数据库 vs 单机数据库



#### 分布式

#### 优点

- **弹性扩展**：通过横向扩展解决了单机性能上限和业务数据量增长不匹配的问题；
- **高度可用**：即使系统中的某些节点不可用（断电、系统崩溃等），也不影响其他节点正常工作，保证了面向用户的高可用；
- **成本控制**：企业可以选取较低配置的硬件。

#### 缺点

- **复杂性**：多节点横向分布提升了架构设计、运维、迁移的难度；
- **安全性**：远距离访问和网络通信传输带来了安全和保密方面的风险；
- **数据完整性**：多节点读写对事务性提出挑战。



#### 单机

#### 优点

- **简单性**：数据集中存储和处理，无需处理多个节点之间的协作，架构设计简单，易满足ACID事务需求；
- **可靠性**：集中式数据库发展时间长，产品在容灾设计、系统运维等方面都有较成熟的解决方案，稳健可靠易维护。

#### 缺点

- **性能扩展**：面对海量的数据存储需求，集中式数据库想要提升性能只能依赖硬件的提升（纵向扩展），在扩展空间方面具有局限性；
- **成本高昂**：高性能的硬件（服务器等）意味着高价，这为企业增添了成本负担。

# 按部署模式分类

- 云数据库
- 本地数据库

## 包括云厂商托管的开源数据库和云原生数据库

云数据库是在云计算的大背景下发展起来的一种新兴的共享基础架构的方法，它极大地增强了数据库的存储能力，消除了人员、硬件、软件的重复配置，让软、硬件升级变得更加容易。现阶段云数据库主要包括两种：一种是托管在云厂商上的“传统”数据库，例如阿里云、腾讯云上的MySQL、PostgreSQL、MongoDB、Redis等；一种是基于云环境的云原生数据库，例如AWS的Aurora、阿里云的Lindorm和PolarDB等。

### 云数据库类型

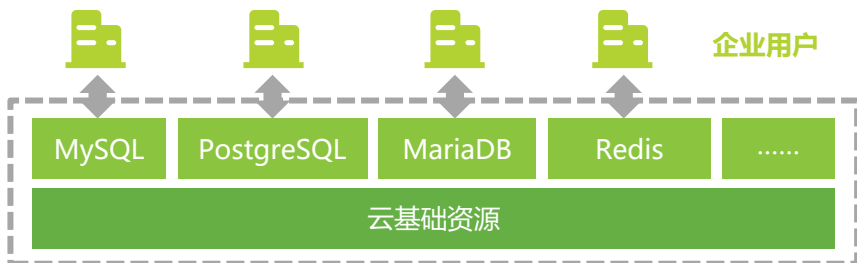
#### 云数据库

##### 云厂商托管的数据库

- 阿里云-RDS MySQL版
- 腾讯云-云数据库MySQL
- AWS-Amazon RDS
- .....

在该模式下，云服务厂商负责管理与维护基础设施，并提供优化、备份、恢复、监控等全套解决方案。企业用户无需购买服务器、交换机等软硬件，后续也无需投入大量的人力成本去运维，可以更专注于企业的应用开发。

云托管数据库与开源版和商业版数据库相比，是一种“开箱即用、弹性扩展、省钱省力、高度可用”的解决方案。



##### 云原生数据库

- 阿里云-PolarDB
- 腾讯云-TDSQL
- AWS-Aurora
- 华为云-GaussDB
- .....

随着云计算和数据库技术的进一步结合，2015年左右云原生数据库诞生，它是基于云环境设计的新型数据库，天生匹配云环境和分布式事务。其核心是存储与计算分离，同时还具备高性能、高可扩展、一致性、容错、易于管理和多云支持等特性。

各家云厂商在提供托管服务的同时都在加快自研速度，开发自己的云原生数据库。



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## “迁移”成为企业数据库上云最不容忽视的隐性成本

云模式下企业可以降低自己从头建设数据库及后期运维的成本，吸引了大批企业规划上云。但实际应用中，绝大部分企业都并非从零开始，都具有一定的数据库基础。对于想要应用云数据库的企业，“上云迁移”成为了其最大的门槛。“如何保证数据安全完整，如何建立失败回滚标准，如何对数据库重新进行设计，如何进行数据模型的转换，如何对新架构做调优……”，这些问题都需要企业谨慎考虑，具备一定的难度。虽然各公有云厂商针对上云迁移都提供了相应的工具，但由于迁移的复杂性，也催生了许多提供咨询、选型、规划、迁移、运维、优化等服务的中间厂商，近年来发展迅速。

### 数据库上云迁移步骤



根据企业自身具体需求判断选择云数据库还是自建数据库：

- 对于一些大型企业，出于安全性和个性化的考虑，通常采用自建本地数据库的方式
- 对于一些IT预算有限的中小企业，云数据库提供了可行的解决方案

- 收集需求
- 判断解决需求需要哪些能力
- 评估哪些数据库需要迁移（建议从非关键业务系统、非核心生产系统入手）
- 评估应用程序配合迁移数据库需要作出的改变
- 建立成功的评判标准和失败回滚原则

- 数据库备份（热备份or冷备份、部分备份or全部备份）
- 重新设计数据库（可选）
- 复制并将数据（包括备份后对原始数据的更改）重新存储在云中
- 移交后检查：数据验证、端到端测试（验证基本功能）、性能测试、安全评估

- 性能优化：负载测试、分布优化
- 可用性优化：容灾恢复计划、日志和系统检测、变更检测、系统测试
- .....

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 云端部署vs本地部署

## 云数据库在成本、易用性等方面具有优势，近年来快速增长

云数据库极大地利用了云计算“资源池化”的优势，在成本、可用性、易用性、扩展性和并行处理方面较传统数据库有绝对优势。云数据库即开即用，用户可以根据自身的业务情况弹性开支、灵活调整；无需从头采购基础软硬件，无需考虑专业人员（DBA）部署，节省了人力物力；同时云数据库大多支持热备架构，可以实现故障秒级自动切换，备份、恢复更加灵活。但同时，由于云环境的特性、产品的不成熟性和市场的混合部署需求，云数据库在数据质量、数据迁移、数据融合、性能优化和规范标准方面仍有改进的空间。

+

### 低成本

多租户模式，用户之间共享资源且只用按需付费，节省了成本

### 高可用

高水平的容错能力，一个节点崩溃，其他节点也可以继续工作

### 易用性

不需要关心底层服务器、系统等部署和运维，开箱即用

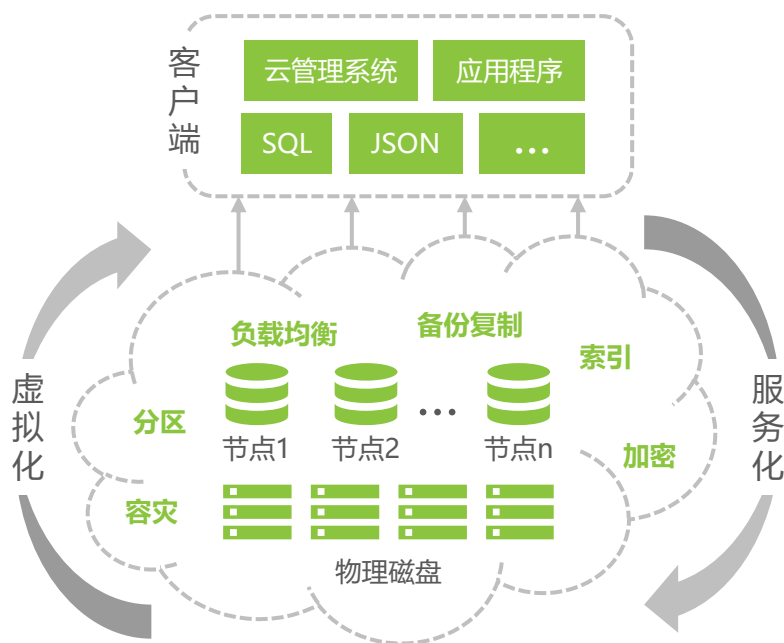
### 动态可扩展

具有无限可扩展性，可以满足不断增加的数据存储需求

### 大规模并行处理

并行处理能力强，面对海量数据，几乎可以做到实时的响应

## 云数据库的核心优势与改进空间



### 数据质量

云数据库在大数据环境下，容易产生脏数据，影响事务一致性

### 数据迁移

将大量、复杂的企业内部数据库数据迁移上云存在一定困难

### 数据融合

本地数据与云数据长期并存，需要有效的融合机制，统一管理

### 性能优化

云环境为动态负载均衡、资源分配管理提出了新的要求

### 规范标准

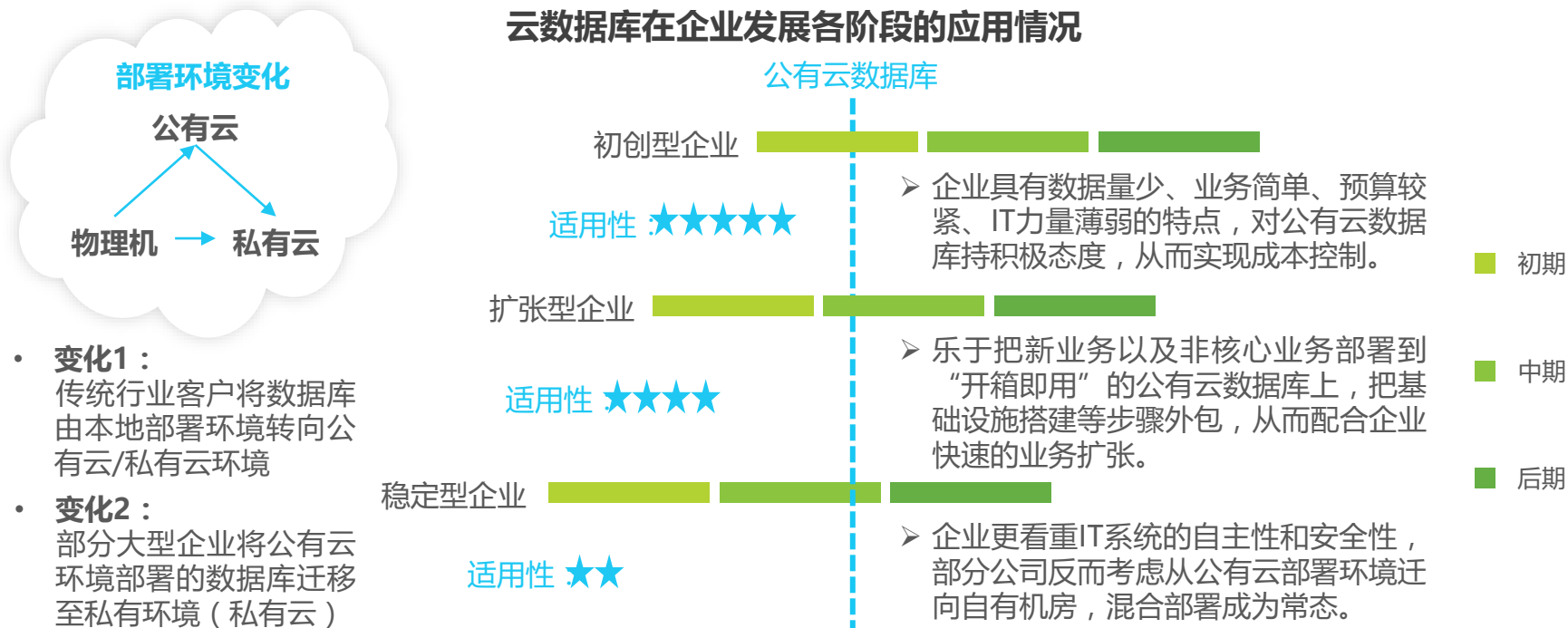
各大厂商独立发展云数据库，在查询语言、语言模型和安全等方面缺乏统一的规范标准



## 部分企业出现反向迁移情况，混合部署成为未来常态

一方面，尽管上云是大势所趋，但是由于数据库基础软件的特性和公司战略考虑，在一定时间内，云数据库很难完全替代本地数据库，混合部署成为企业的必然选择。现阶段绝大部分企业都具有一定的IT基础，业务数据都存储在本地自建的数据库里，经过了几十年的积累，具有复杂和海量的特点。短时间内让企业放弃原本投入了大量成本的本地数据库，把海量复杂的数据全面迁移上云，是不现实且不划算的。另一方面，企业私有云部署成为当下的热门选择，公有云数据库市场增速放缓。当企业业务发展到一定规模，对核心系统自主可控的要求也相应地提升，这一阶段的企业反而出现了反向迁移的现象，更多地考虑把部分业务数据从公有云迁移到私有云部署的环境里。

### 云数据库在企业发展各阶段的应用情况



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。



## — 按功能分类

- OLTP事务型数据库
- OLAP分析型数据库
- HTAP混合型数据库

# OLTP (事务型) vs OLAP (分析型)

## 分别满足企业联机交易和分析决策的需求，适用范围不同

面对事务处理和分析决策的需求，OLTP (Online Transactional Processing) 事务型数据库和 OLAP (Online analytical processing) 分析型数据库应运而生。OLTP系统主要使用关系模型，保证强一致性，面向一线业务人员，支持多并发、实时、快速地增删查改，例如银行交易、零售电商、车票预订等；OLAP系统可以高速多维分析来自数据仓库、数据集市或者数据湖的数据，可使用关系型或者非关系型的数据库，主要面向分析师和管理者，支持对历史数据的复杂分析操作，从而赋能企业商业智能决策。

### OLTP vs OLAP

	OLTP	OLAP
产品定位	支持实时交易数据的存储、更新、共享	通过数据分析现状，发现趋势，支持决策
操作	基于INSERT, UPDATE, DELETE命令	基于SELECT命令聚合数据用以分析
数据量	实时数据，通常较小	聚合历史数据，较大
并发访问量	高	低
响应时间	毫秒	秒，分钟或者小时（取决于处理的数据量）
存储结构	通常为行存储	通常为列存储
备份	需要定期备份以确保业务连续性	可以从OLTP数据库重新加载丢失的数据，以代替常规备份
可视化	日常业务交易列表视图	多维视图满足分析需求
典型适用场景	快速处理高并发、小批量的数据	使用复杂的查询处理大量数据
主要用户	银行柜员、收银员、仓库管理员等	数据分析师、业务分析师、高管等

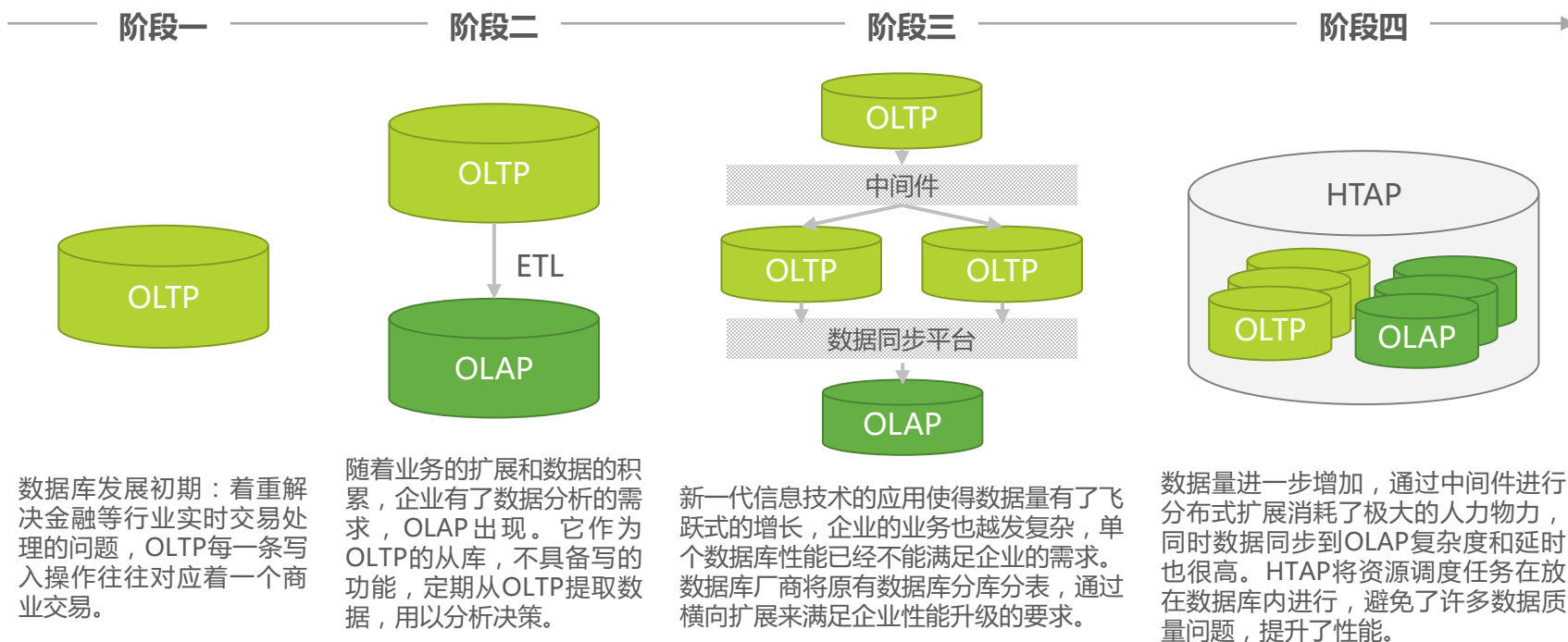
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# OLTP→OLAP→HTAP (混合型)

## 随企业业务扩展、市场需求变化而变化

随着数据价值的进一步挖掘，企业对数据库系统也提出了新的要求。现阶段企业为满足交易处理和分析的需求，往往采用 OLTP+OLAP 的组合方案。但二者之间往往存在时延，无法满足企业实时分析的需求；同时管理两个平台往往需要组建两支团队，运维成本高。HTAP (Hybrid Transactional/Analytical Processing) 混合型数据库基于新的计算存储框架，能够同时支撑OLTP和 OLAP 场景，避免传统架构中大量数据交互造成的资源浪费和冲突。此外，HTAP 基于分布式架构，支持弹性扩容，可按需扩展吞吐或存储，轻松应对高并发、海量数据场景。

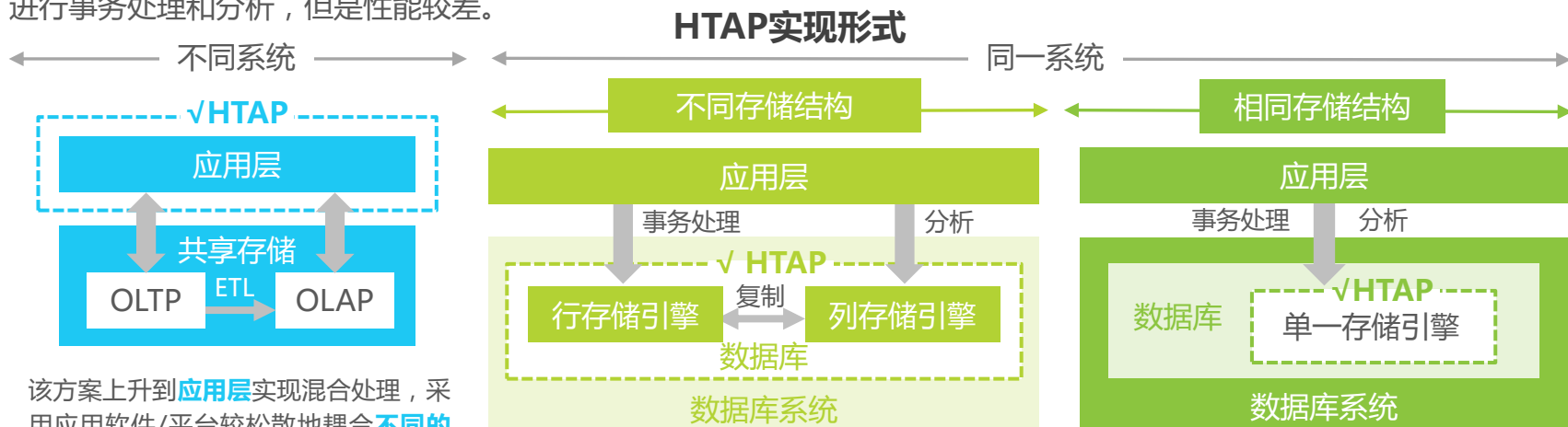
### 数据库发展历程——功能角度



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 提供打包事务处理和分析功能的一站式解决方案

当今市场上HTAP的实现形式主要包括三种：一种运用应用软件实现OLTP和OLAP系统的松耦合，底层共享存储来缩短数据同步的时间；一种让不同数据结构的存储引擎分别负责事务处理和分析，存储引擎在物理上隔离（分布式），并遵从一定的协议（例如Raft）在引擎之间进行实时复制，是当今HTAP数据库采用的主流解决方案；最后一种使用单一存储引擎进行事务处理和分析，但是性能较差。



该方案上升到**应用层**实现混合处理，采用应用软件/平台较松散地耦合**不同的系统**，对外整体呈现HTAP能力，例如 SAP HANA Platform+HANA+Spark SQL。此架构下，OLTP和OLAP共享同一存储，OLTP系统中的实时交易数据通过ETL转化到OLAP系统，从而实现了事务的快速（仍有一定延迟）分析。

e.g. SAP HANA Vora

该方案在**数据库层**，即**一个系统**内实现HTAP。该类型数据库系统提供不同的存储引擎，分别负责不同功能板块，例如行存储引擎负责实时事务处理，列存储引擎负责分析。此架构是当今HTAP数据库市场的主流采用方案，许多单一存储引擎的数据库厂商都在近年来提供了其他引擎的扩展。

e.g. 阿里 HybridDB、PingCAP TiDB、HyPer、Peloton、SAP HANA、Oracle Times Ten、MemSQL、IBM dashDB等

该方案希望在**最底层**就实现混合处理，整个架构下仅使用**一套系统和单一存储引擎**。用户可以在一个请求中既实现实时交易处理又实现分析，实现真正的HTAP。然而现实应用中，无论只使用行存储还是只使用列存储，都无法实现性能最优，仍待进一步的探索。

e.g. Hive、H<sup>2</sup>TAP（学术项目）、Impala + kudu等

# 按存储介质分类

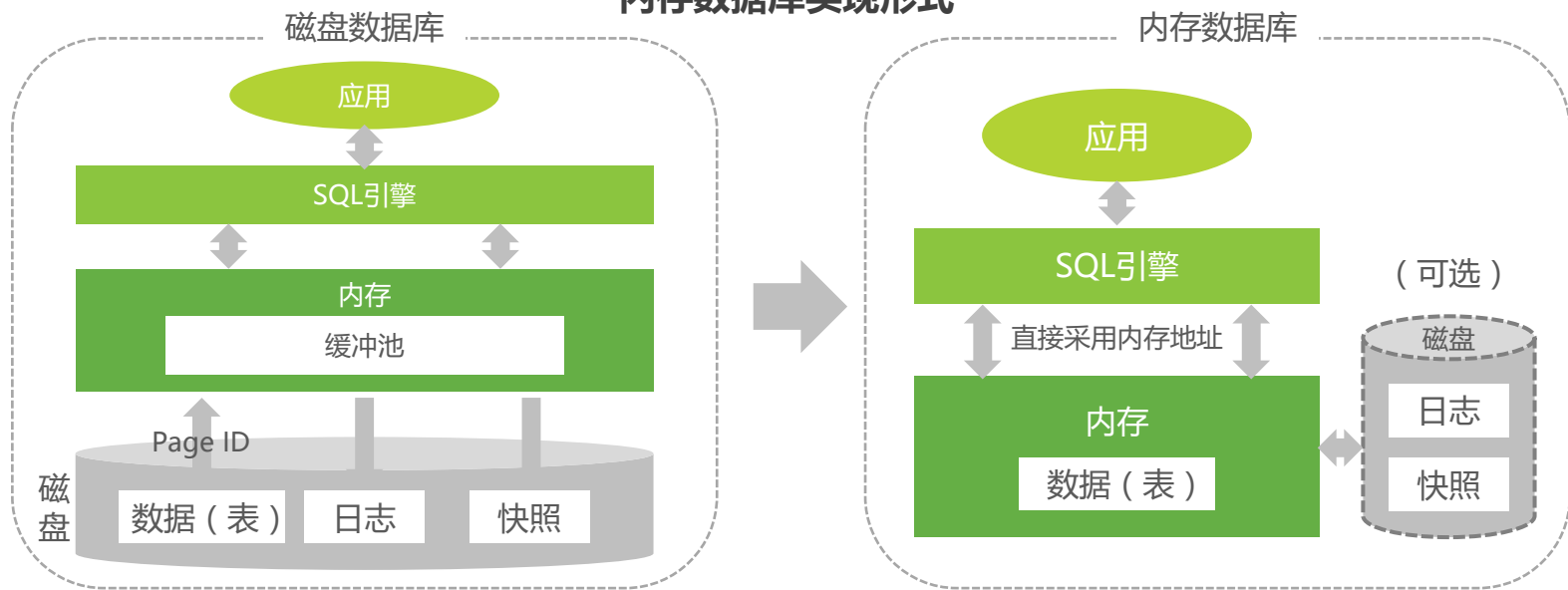
- 内存数据库
- 磁盘数据库

# 内存数据库

## 全部数据存储在内存中，具备更极致的读写性能

在数据库发展早期，由于硬件性能的局限，数据库系统通常采用基于磁盘的设计，数据在内存中进行相应处理并以磁盘块为单位存储在磁盘上。而内存数据库（IMDB）是一种将全部数据存储在内存中，无需进行磁盘I/O即可对数据进行增删查改，具备高读写性能的数据库。其设计理念最早可以追溯到IBM于1976年推出的IMS/VS Fast Path 数据库，它体现了数据分层的思想，将活跃数据放在物理内存中进行访问和管理。随着互联网的发展，用户对数据量、操作频率和响应速度有了越来越高的要求，而磁盘数据库面对多并发、高频率的访问时暴露出越来越多的问题；同时内存的容量不断增加，单价越来越低，计算机操作系统地址空间得到更大的支持，把全部数据放到内存中具备了可实现性。各商业、开源的内存数据库纷纷问世，内存数据库进入了高速发展的阶段。随着未来非易失内存NVM（实现内存存储的所有数据在电流关掉后也不会消失）的发展与成熟，内存数据库的应用范围将会得到进一步的跃升。

### 内存数据库实现形式

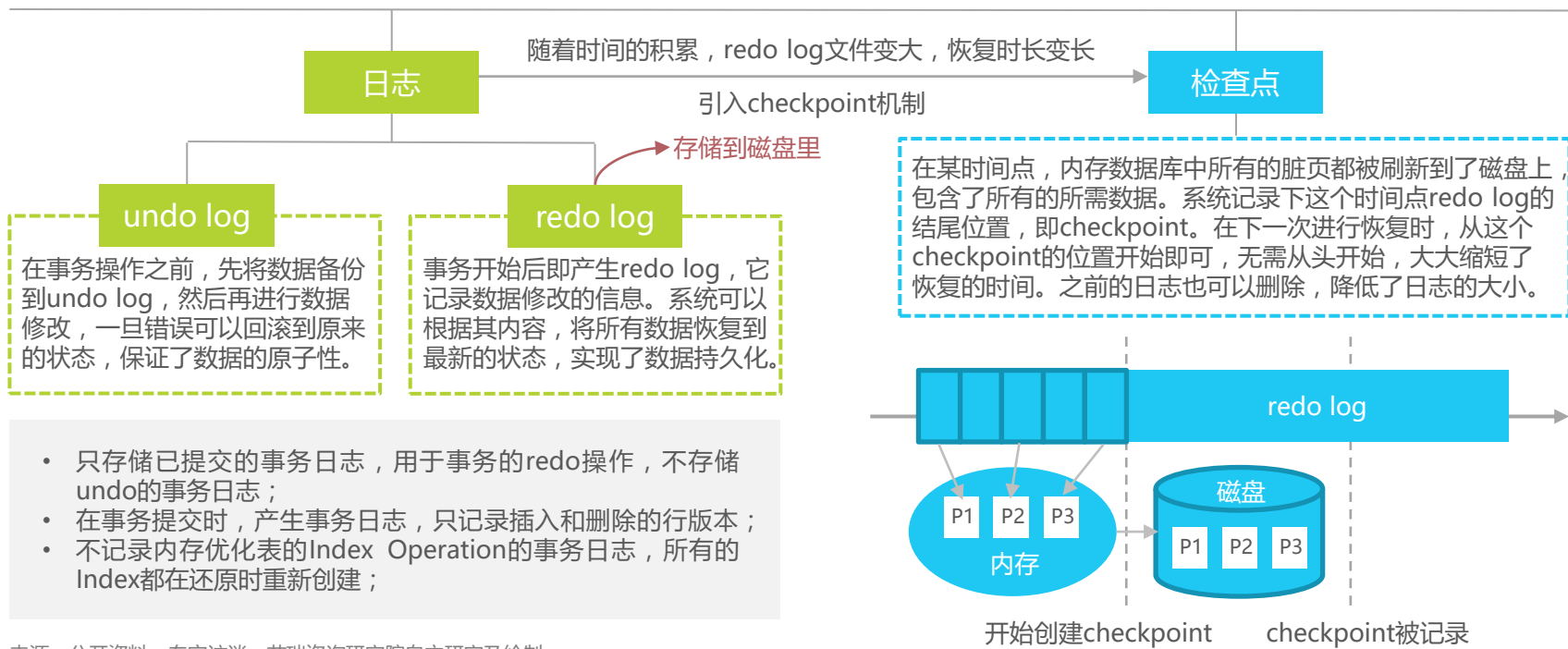


来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 通过事务日志和检查点机制，满足“高性能+持久性”双需求

由于现阶段NVM尚未达到应用水平，而存储在DRAM中的数据在重启后则会丢失，不能满足用户持久存储数据的要求。因此，内存数据库需要考虑数据的持久化问题。当前主要的方法包括日志机制 (Log) 和检查点机制 (Checkpoint)。日志即将每一次数据的更新操作（增删查改）记录在 Log Records文件中并写入磁盘；检查点即采用一定策略，周期性地将在内存中的数据同步到磁盘里。两种持久化方式都可以单独使用，但在实践中通常采用两者结合的方案。检查点可以配合相关日志进行数据库的恢复，二者的结合可以减少检查点对正常事务的影响，减轻系统恢复的开销并缩减日志文件的大小，实现恢复速度的大幅提升。

### 内存数据持久化机制



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 磁盘数据库 vs 内存数据库

## 在安全和性能方面各有优劣，往往搭配处理冷热数据

内存数据库具有“实时性能、IT架构/数据结构简单、灵活扩展”的优点，在对读写性能有极致要求的场景有着广泛地应用，例如电信计费、嵌入式控制系统、呼叫中心应用程序和电商秒杀平台等。但由于内存本身特性，以其为架构中心的产品在“数据持久性、容量限制、成本控制”方面较传统的磁盘数据库不具备相对优势。对数据遗失容忍度较低的企业还需要考虑相应的数据持久化方案。另外非易失内存（NVM）及其适配架构、产品还并不成熟。因此许多企业为满足多重约束，现阶段主要采取“磁盘数据库+内存数据库”配套使用的解决方案，分别处理冷热数据。

### 磁盘数据库 vs 内存数据库

#### 磁盘数据库

##### 优点

- 支持ACID事务特性，数据完整性好
- 数据库可用性高
- 发展时间较长，产品及配套工具成熟度高

##### 缺点

- 需要缓冲处理，占用大量系统资源
- 数据存取速度慢
- 数据存取时间不一致且难以预测

#### 内存数据库

##### 优点

- 避开了数据访问时磁盘的I/O瓶颈，存取速度快，系统性能高
- 直接采用内存地址查询，数据结构简单
- 并发控制表现较好

##### 缺点

- RAM介质掉电数据丢失，安全性较差
- 需要额外的日志和快照机制进行灾备
- NVM发展不成熟
- 较磁盘价格更高

#### 应用场景：

对数据读写性能要求不高的常规场景

#### 应用场景：

对读写性能有极致要求的电商秒杀、商城目录、视频直播、电信计费、新闻查询、嵌入式控制系统等场景



## 按商业模式分类

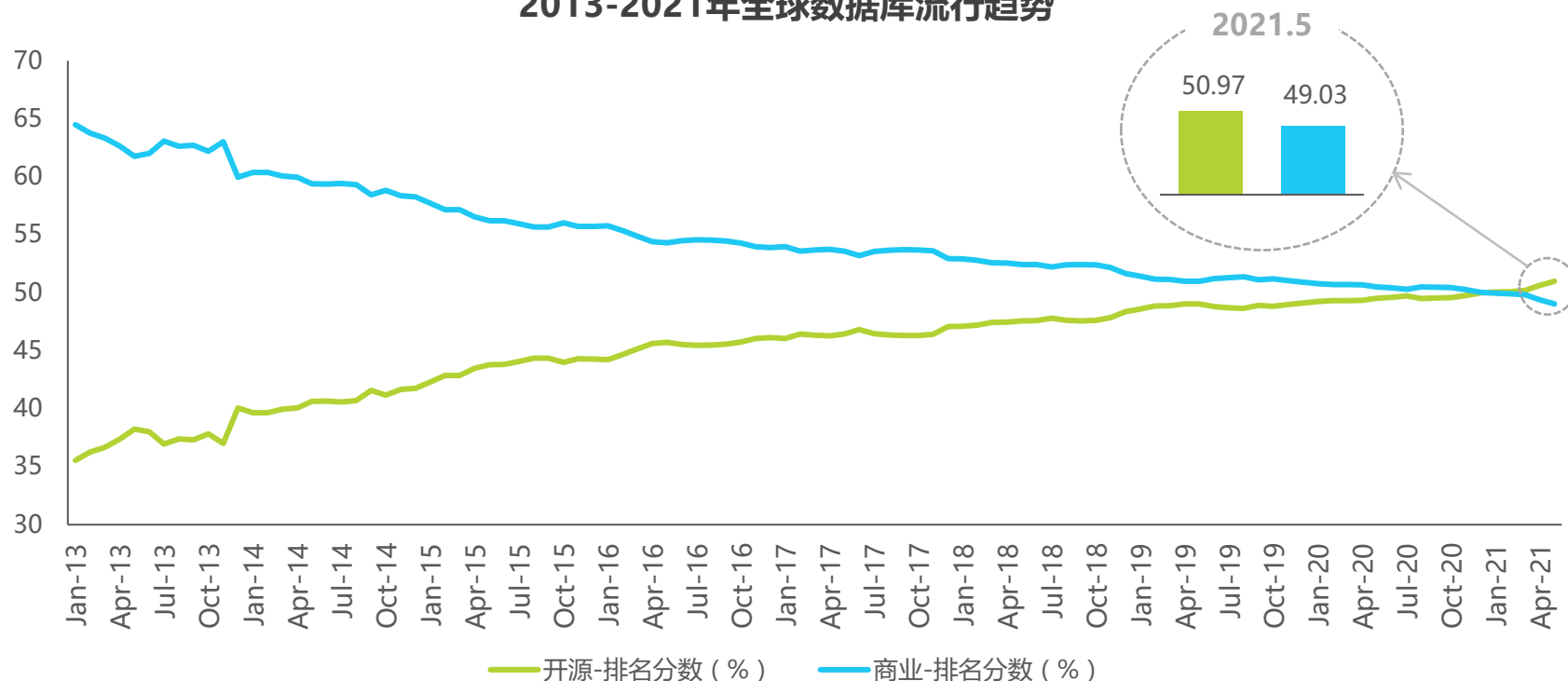
- 开源数据库
- 商业数据库

# 开源数据库

## 流程度逐年上升，规范性和配套设施逐渐完善

与闭源的商业数据库相反，开源数据库是免费的社区数据库，其源代码对外开放，开发人员可以在其原始设计基础上修改或使用。它以较低的成本、丰富的产品和活跃的社区支持为日益复杂的企业需求提供了相应的解决方案。从DB-Engines全球数据库管理系统排名看，开源DBMS流程度逐年上升，2021年1月首次超过商业数据库。

2013-2021年全球数据库流行趋势



注释：1、DB-Engines排名是按照当前流行程度的排名，较实际使用情况具有一定的超前性，具体指标包括网站上系统提及的频次、Google trends、IT论坛上系统讨论的频率、提及系统的工作机会数量、专业网络中提及系统的配置文件数和系统在社交网络中的提及次数；

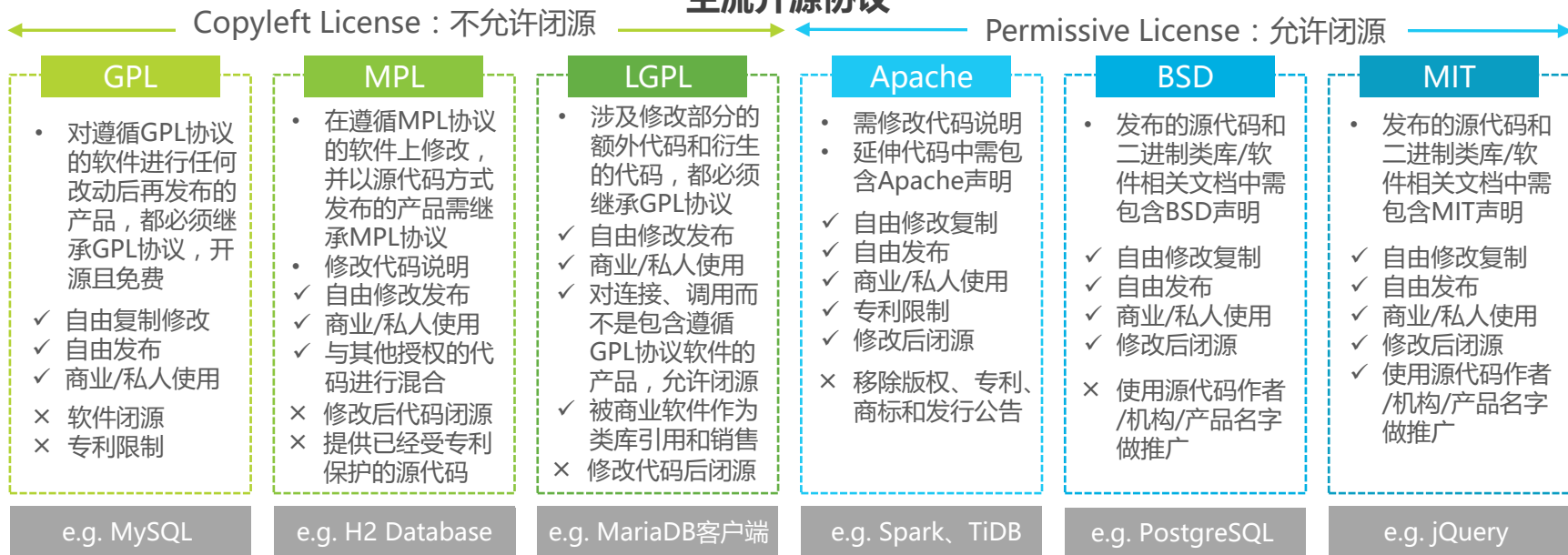
2、开源、商业-排名分数 (%) 为相对值。

来源：DB-Engines，艾瑞咨询研究院自主研究及绘制。

## 为开发者和开源软件保驾护航

当今开源数据库应用的开源许可协议主要包括两派：一派是以GPL、MPL、LGPL协议为代表的Copyleft License，严格执行开源精神，不允许修改代码后闭源，其中GPL更是做了进一步的要求，不允许修改后的新代码更改开源协议。另一派是以BSD、MIT、APACHE、木兰开源协议为代表的Permissive License，允许修改代码后闭源，因此较受商业公司青睐。近年来，由于云数据库托管服务的扩张，越来越多的企业客户流向了云服务商，使得开源社区活跃度下降，开源开发者的利润空间被进一步积压，对开源生态造成了较大侵袭。针对此现象，许多开源数据库（例如：MongoDB、CockroachDB、Redis Labs、Elastic、Confluent 和 TimescaleDB等）都采取了相应的措施，或是改用了商业化限制更严格的许可协议，或者自己提供收费的企业版，或是采取产品开源、服务收费的模式。然而，如何维持开源生态健康发展，在开源和商业化之间寻求平衡，还有待各方面因素的协商和共同努力。

### 主流开源协议



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 开源数据库 vs 商业数据库

## 在“低成本自主性”和“易用稳定”方面各有优势

开源数据库虽然避免了高昂的 License 费用和服务费用，但在易用性、配套能力、服务能力、版本更新方面存在一定的缺陷，同时产生了开发、部署、迁移等额外的成本。云计算时代来临后，云厂商提供开源数据库托管服务，将服务器、数据库维护升级、人力运维等底层工作包揽过来，为企业提供较高性价比的解决方案。但是，公有云托管的开源数据库较契合中小企业简单部署、运维、调优、低价等诉求，不能满足金融、政企等大型组织对安全可靠、数据一致性、高响应速度等方面的要求。因而现阶段许多企业在权衡成本和安全性等各方因素后，倾向选择“开源数据库+商业数据库”的组合。

### 开源数据库 vs 商业数据库

	前期：选型采购	中期：开发部署	后期：运维使用
开源	<b>选型成本</b> 开源产品/社区繁多造成选型障碍  <b>评测成本</b> 需要结合需求做大量的POC测试	<b>开发成本</b> 自研案例缺失  <b>部署成本</b> 问题不确定性  <b>迁移成本</b> 源码、接口不兼容问题	<b>扩展限制</b> 企业级性能差  <b>Bug优化</b> 部分Bug较多  <b>配套升级</b> 缺乏配套产品  <b>人力投入</b> 大量人力运维
商业	<b>产品费用</b> 商业license价格高昂  <b>咨询费用</b> 前期咨询产生的费用	<b>服务费用</b> 包括个性化定制、部署、培训、后续运维的费用	

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## — 其他数据产品

- 数据仓库/数据集市
- 数据湖
- 大数据平台
- 数字中台

## 逐步放大数据价值：存储→交易→分析→赋能

数据相关产品随着市场需求的变化而逐步发展，逐渐挖掘数据价值 (D2V)。最初企业意识到数据存储、事务处理的价值，数据库诞生。后又增加了对数据分析的需求，商业智能、数据仓库/集市应运而生，它是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策和信息的全局共享。之后，数据的体量进一步增长，大量的结构性数据和非结构数据产生，企业开始通过大数据平台分析来自数据库、数据湖和外部的数据，为企业赋能。但是企业各部门之间存在信息孤岛，数据转化为商业价值难，数字中台面世，通过数据业务化和业务数据化，形成企业内的数据闭环。关于数字中台更多内容，详见我们即将发布的《2021年中国数字中台行业研究报告》。

### 数据相关产品全局一览



来源：艾瑞研究院自主研究及绘制。

产品与技术：数据库内涵与分类

1

供给与需求：数据库市场现状与选型

2

案例与启示：数据库典型厂商案例

3

机遇与挑战：数据库未来发展趋势

4

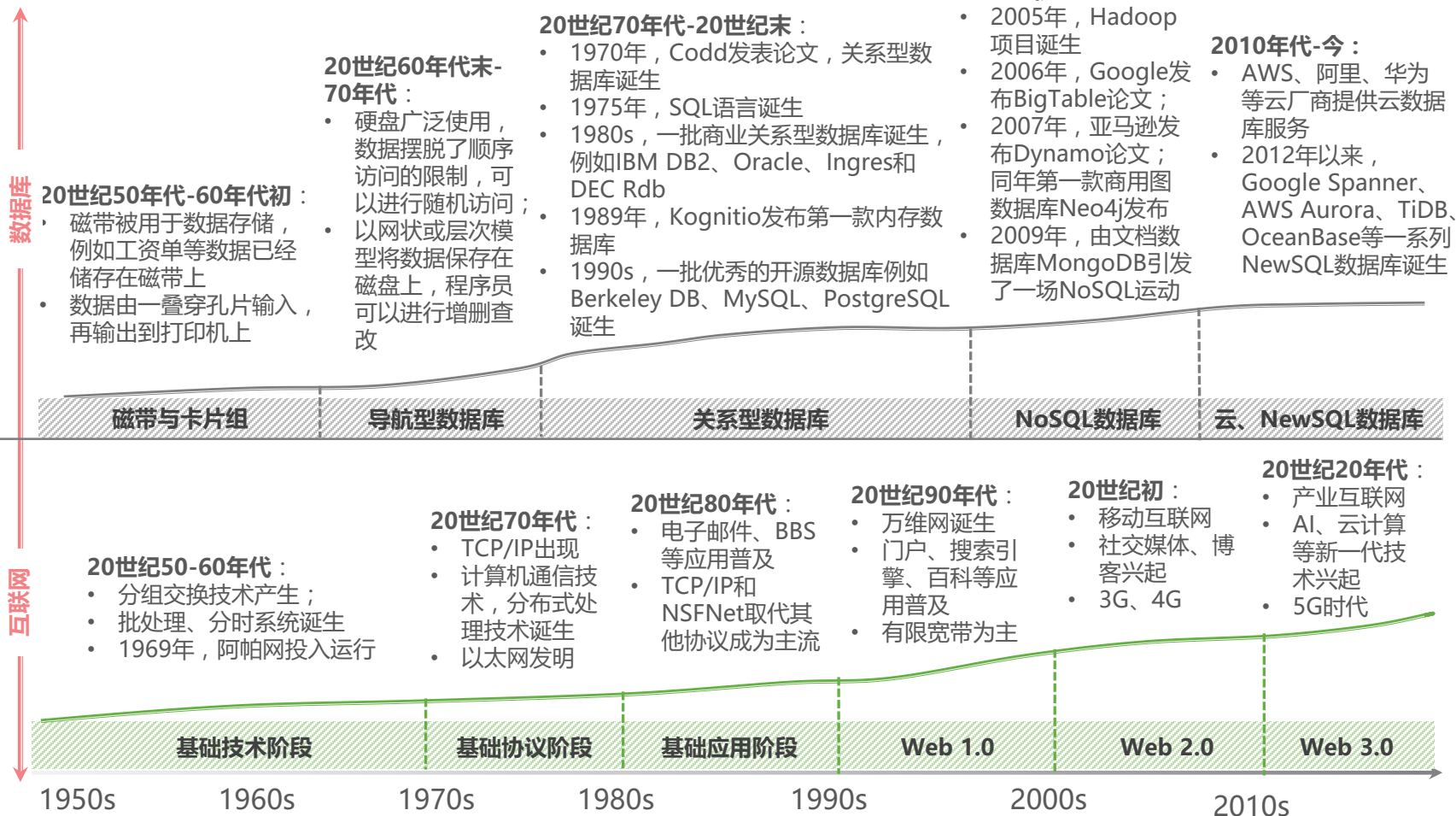
# 供给端



# 全球数据库发展历程

## 数据库与互联网发展相互促进，技术和产品趋于成熟和完善

### 全球数据库发展历程总览



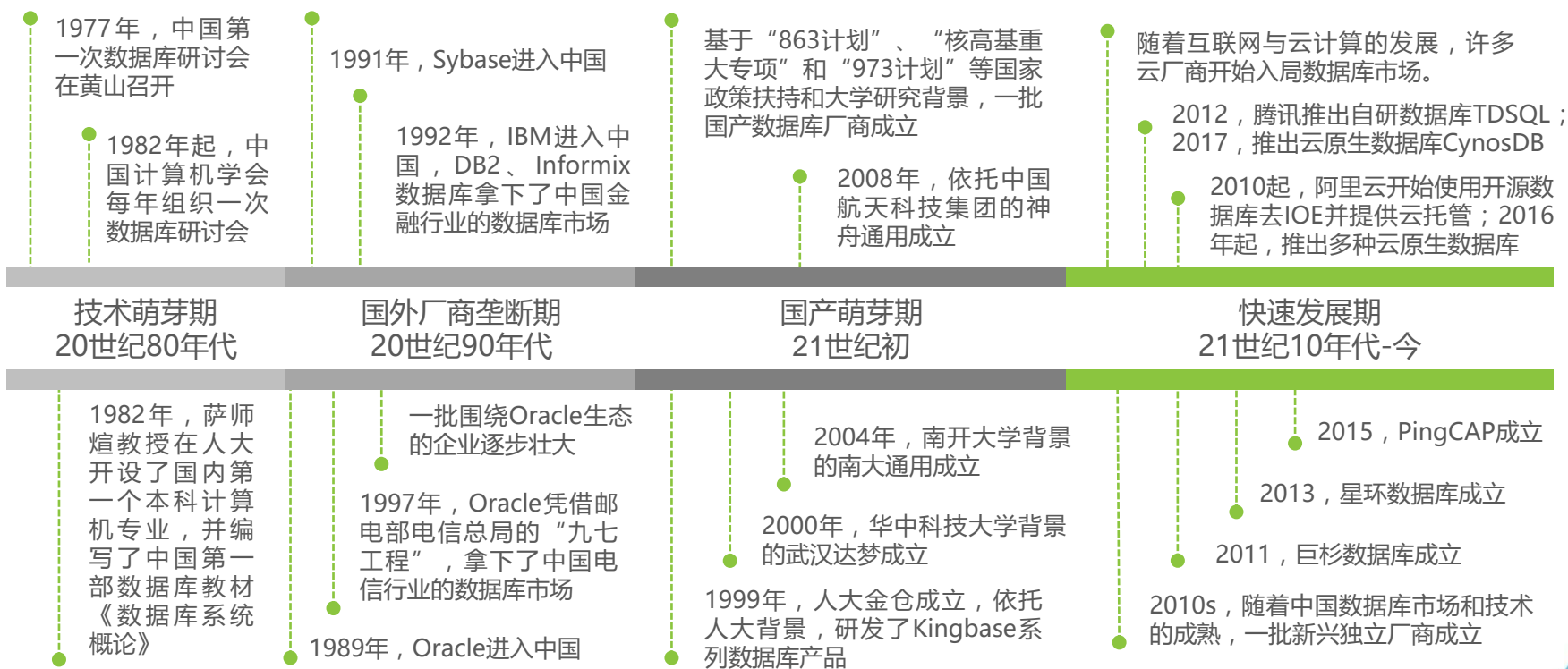
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 中国数据库发展历程

## 受益于市场需求和技术沉淀，进入百花齐放的快速发展期

从20世纪80年代起，我国数据库市场开始逐步发展起来。经历了初始的技术萌芽期和国外厂商垄断期，21世纪初，基于863计划、核高基计划等国家政策支持，一批拥有高校背景的国产厂商成立，打破了Oracle和IBM一统天下的格局。2010s，随着市场需求的增长、技术的沉淀，一批云厂商和新兴独立厂商开始提供数据库产品。近年来，借助国产化热潮，许多软件厂商、集成商、运营商等也开始入局，发展自己的数据库能力。

### 中国数据库发展历程总览



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 中国数据库产业图谱

## 中国数据库产业图谱

国内厂商

### 传统厂商



### 初创厂商



### 云厂商



### 跨界厂商



国外厂商

### 商业数据库



### 开源数据库



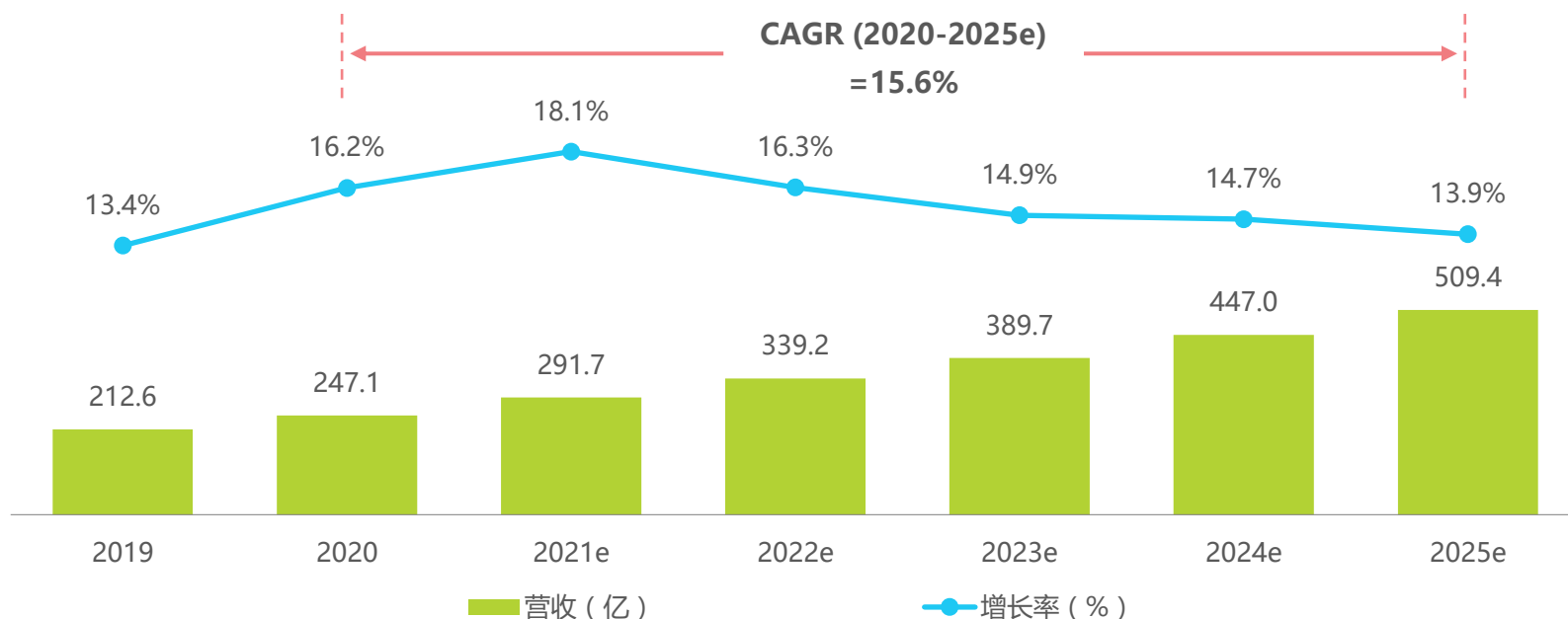
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。  
注：图谱仅展示部分典型厂商，图谱展示顺序不代表市场份额排名。

# 中国数据库市场规模

## 政策和数字化释放需求，2020年市场规模达247.1亿

据艾瑞统计，2020年中国数据库市场总规模达247.1亿，较2019年增长16.2%，CAGR（2020-2025e）达15.6%。2020-2022中国数据库市场将呈高速增长，由多方面因素促成：1）政策利好，国家大力鼓励国产数据库厂商的发展；2）需求拉动，国产化和数字化建设带动需求的爆发增长；3）供给端多元厂商发力，传统、初创和跨界厂商厚积薄发，产品和技术经历了工程实践的打磨走向成熟；4）国内企业对基础软件的付费意愿和IT支出也在逐年提升，有利于市场的长期发展。

### 2019-2025e中国数据库市场规模及增速



注释：市场规模统计口径为国内外厂商在中国数据库销售的营收。其中包括DBMS基础软件的收入，必要配套工具的收入（数据迁移、数据备份等工具），项目定制化开发、实施、运维等服务的收入，数据库软硬一体机的收入；不包括单独售卖的硬件的收入，单独售卖的大数据平台的收入、单独售卖中间件及应用软件的收入。

注释：此处市场规模中包含云厂商托管开源数据库（MySQL、PostgreSQL、MongoDB、Redis等）所得的收入。

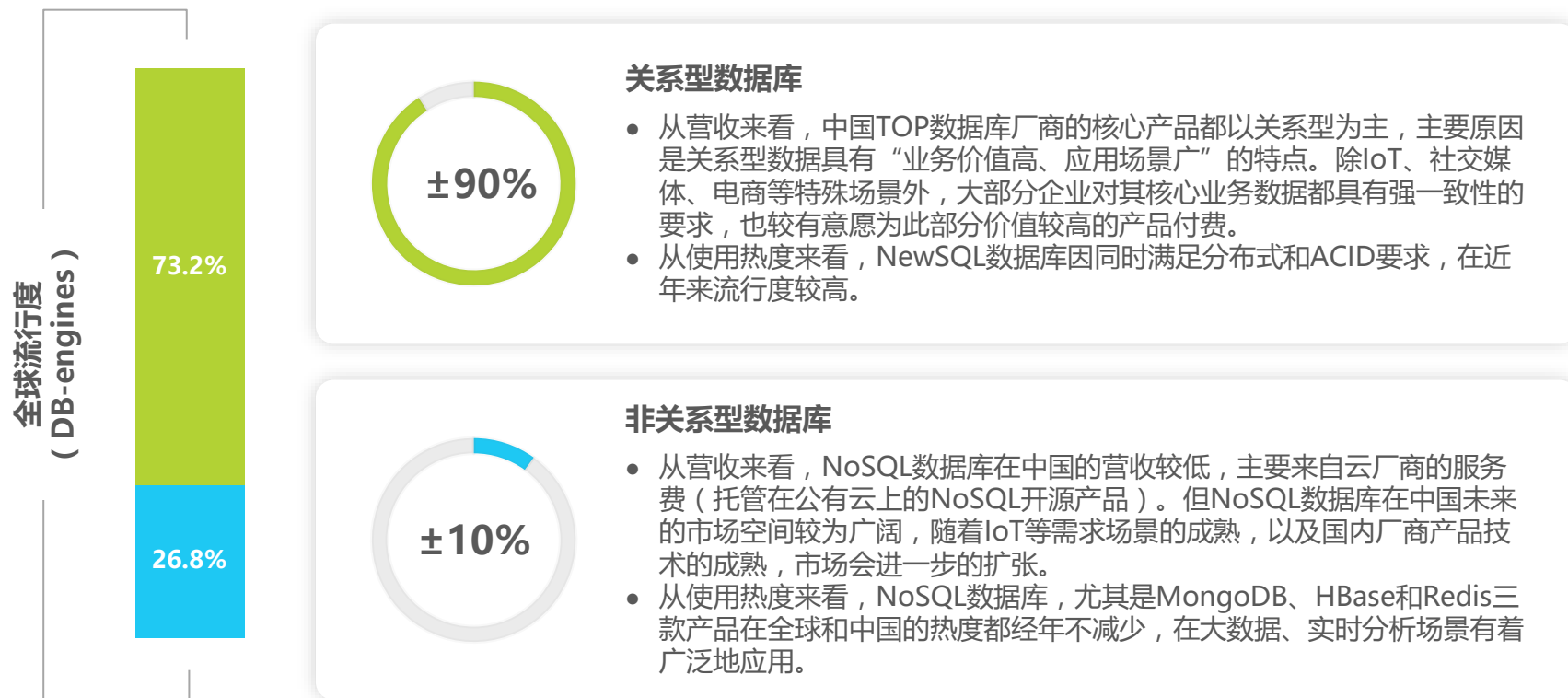
来源：根据公开资料、企业访谈，结合艾瑞统计模型核算。

# 中国数据库市场发展特点（一）

## 多类型数据库百花齐放，关系型占据绝对主流

随着互联网的发展，多种类型数据爆发式地增长，各种创新业务场景层出不穷，进而促进了供给端厂商技术和产品架构的创新。从2010s左右，多种类型和技术路线的数据库厂商纷纷成立，中国数据库市场进入了百花齐放的阶段。但从商业价值来看，中国数据库市场的营收仍主要来自关系型数据库，NoSQL数据库更多地是开源模式，产生二开和服务的费用。

### 2020年中国数据库市场份额：关系型vs非关系型



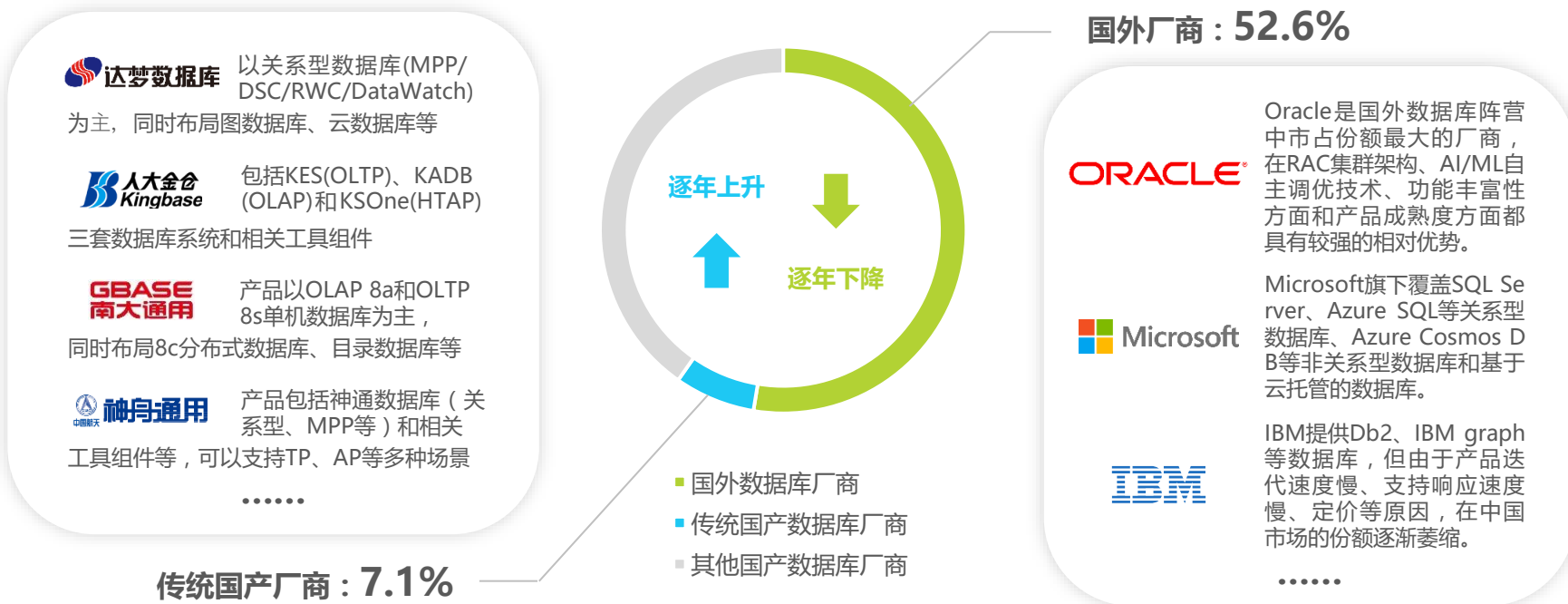
来源：db-engines.com、公开资料、企业访谈，结合艾瑞统计模型核算。

# 中国数据库市场发展特点（二）

## 借助政策东风，国产厂商厚积薄发，市场版图快速扩张

借助政策红利，国产厂商经过多年的技术研发和经验积累，市场份额在逐年提升。在国产阵营中，一批以“达梦、人大金仓、南大通用、神舟通用”为代表的，2000年左右成立的传统国产数据库厂商近年来开始发力，他们中有的从购买源码、借助开源走向自主研发，实力不断增强，在党政军市场有着较好的表现，同时也开始向能源电力、运营商、交通等其他行业快速拓展。此外，初创厂商、云厂商、ICT厂商等近年来也开始发力数据库市场，国产阵营日益强大。相比之下，国外数据库厂商如Oracle、Microsoft、IBM等，虽然在OLTP的核心场景还拥有较高的市占率，但整体市场份额在被逐渐侵蚀。

### 2020年中国数据库市场份额：国产vs国外



来源：根据公开资料、企业访谈，结合艾瑞统计模型核算。

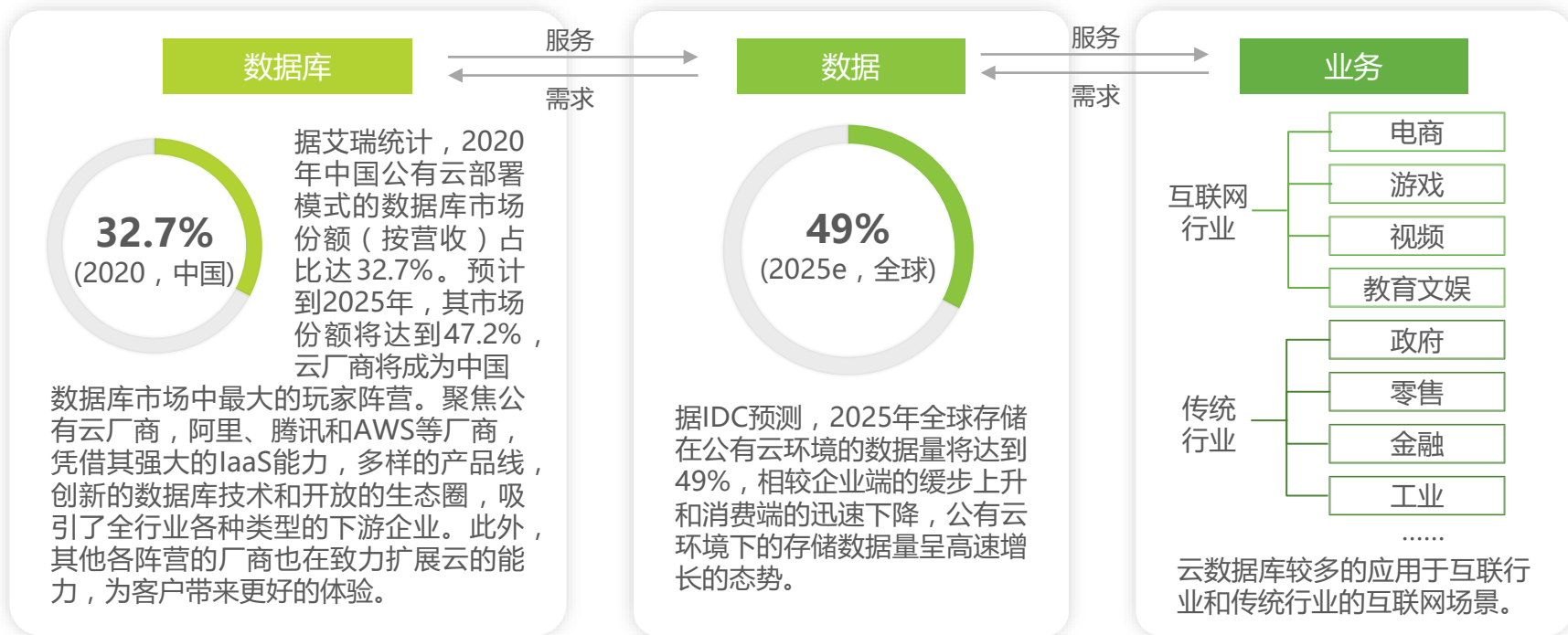


# 中国数据库市场发展特点（三）

## 公有云数据库增速放缓，未来仍有一定渗透空间

中国公有云部署模式的数据库在过去三年快速增长，于2020年达到了32.7%的市场份额，未来虽然增速会有所放缓，但仍有一定的渗透空间。从应用逻辑来看，数据库服务于数据，数据应用于业务最终产生价值；反之，业务端的创新和数据的变化也会反馈与数据库市场。从最终业务端来看，现阶段云数据库更多的还是应用于互联网行业，以及传统行业的互联网场景，未来随着产业端更多的业务创新，有望进一步拉动云数据库的需求。

### 2020年中国数据库市场份额：云上vs云下



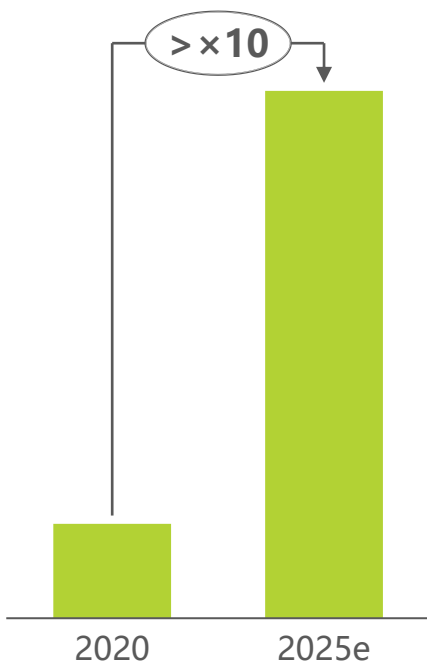
来源：IDC、公开资料、企业访谈，结合艾瑞统计模型核算。

# 中国数据库市场发展特点（四）

## 凭借HTAP、NoSQL等新技术，初创厂商不断涌现

一批2015年前后成立的初创型厂商借助NewSQL、SQL on Hadoop、NoSQL等新技术架构，以开源或垂直领域商业化的思路，逐步增强自己的市场影响力，在互联网、金融、物联网等行业有着较好的表现。从现阶段来看，其营收的市占份额较小，但增速较快，是中国数据库市场增速最快的一个赛道，预计到2025年可以实现高于十倍的扩张。随着市场的大浪淘沙，未来的初创数据库厂商赛道会趋于收敛，市场份额向一小批具有核心技术优势、抓住高价值应用场景的优秀厂商集中。

### 初创厂商典型发展路径



#### NewSQL

- 一部分新兴数据库厂商通过全新的分布式架构设计，在实现横向线性扩展的同时，维持了传统关系型数据的ACID特性。它解决了企业数据量爆发增长，单机不够用；多节点数据强一致；TP/AP统一管理，实时分析等诉求。

#### SQL on Hadoop

- 互联网的发展促生了NoSQL运动的兴起，许多应用程序都迁移到了NoSQL环境中，但随着时间推移，企业发现其更加需要传统关系型数据库所提供的能力。
- 针对此种需求，一部分大数据发家的初创厂商，基于Hadoop做SQL的优化，为企业提供SQL on Hadoop的结合解决方案。

#### NoSQL

- 随着IoT、车联网、社交媒体等特定业务场景的发展，催生了一批专注时序数据库、图数据库、内存数据库的初创厂商，也有着较好的发展。

来源：根据公开资料、企业访谈，结合艾瑞统计模型核算。



# 需求端

## 不同企业组织架构和选型要求不同，采购流程也有一定区别

### 企业数据库典型采购流程



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 综合考虑外围因素、产品技术相关因素和服务价格因素

参考企业的采购流程，一般数据库选型从前期到后期，会综合考虑外围因素（自身技术路线、资质、品牌声誉与行业案例、生态构建等），产品技术相关因素（一致性、兼容性、扩展性、性能、功能丰富性、安全性等），以及后期的价格服务因素（解决方案、性价比、服务响应速度、培训体系等）。

### 企业数据库选型核心指标

#### 前期：外围因素

##### ■ 历史IT架构与技术路线

除初创外，大多数企业都积累了一定的IT资源，不同IT路线之间的转换存壁垒，例如已经选择了Microsoft+.NET技术路线的企业很难再向Oracle+Java路线转型。

##### ■ 相关资质（e.g.国产、自主可控等）

对于政府、国企类型的客户，还会关注厂商的背景和资质，例如是否是国产背景，是否通过国家自主可控的测试等。

##### ■ 品牌声誉与行业案例

前期企业还会关注厂商在行业内的案例丰富性、其他用户的评价，综合声量和声誉进行筛选。

##### ■ 生态建设

企业还会考虑供应商的生态建设水平，例如MySQL在中国广受欢迎的重要原因之一即其借助开源模式，建立了广泛的生态。企业后期招聘人才、进行二开成本较低。

#### 后期：价格服务相关因素

##### ■ 性价比

在初步筛选后，企业会对各供应商进行比价，但价格往往不是B端用户（尤其是数据库软件付费企业）的重点关注指标。

##### ■ 解决方案/服务/培训

对于大型企业而言，厂商的后期服务也是其考核的重要指标。

#### 核心：产品技术相关因素

##### ■ 安全性

###### ■ 可靠性

系统可以无故障地持续运行的概率，一般通过MTBF、MTTR、MTTF等指标来衡量。

###### ■ 可用性

可用性指系统在给定时间内可以正常工作的概率，通常用SLA指标来衡量（俗称1个9、2个9、3个9等），供应商往往通过两地三中心、双活等方案来解决。

###### ■ 稳定性

在一个运行周期内、一定的压力条件下，持续操作时间内出错的概率，性能劣化情况等。

##### ■ 兼容性

包括与历史DBMS的兼容性，是否支持数据导出和迁移，开放接口的丰富性等。

##### ■ 性能

###### ■ 单节点/系统吞吐量

###### ■ 平均时延

###### ■ 执行时间

##### ■ 可扩展性

根据业务需求可以支持线性横向扩展的能力，读写分离支持等

##### ■ 事务特性

对于OLTP数据库，支持强一致性等事务特性是其核心能力

##### ■ 功能丰富性

数据库功能是否能够满足客户业务的多样需求

# 示例场景——金融

## 对“高可用/强一致/低时延”要求高，分布式改造是下一步重点

对于以银行为例的金融企业而言，其业务数据的价值较高，因而对数据库“高可用、强一致、低时延”的要求较为极致。在TP场景下，银行下一步选型的重点为分布式改造。初步来看，解决思路主要是“中间件+分库分表”or“原生分布式架构”。中间件路线方案成熟且性能表现较好，是现阶段大多数客户的选择；但原生分布式架构在扩展性方面存在天然的优势，在未来具有更广阔的发展空间。

### 典型银行业数据库选型



### 分布式数据库选型

#### A：中间件+分库分表

- 对数据取哈希打散，分发到每一个节点上，再用中间件进行全局的事务管理
- 优点：**1) 方案成熟；2) 整体能力接近单机数据库，适合银行低延时的要求
- 缺点：**1) 可拓展性差，数据分片后静态不可变；2) 对业务的侵入强；3) 中间件负担过重；4) 大量写入场景下成本高

#### B：原生分布式数据库

- 从底层架构上针对分布式进行优化
- 优点：**1) 实现真正的水平扩展；2) 对业务的影响小，可以在不停止业务的情况下进行；3) 查询效率高，满足轻量级实时分析
- 缺点：**1) 在并发量和延时方面存在一定缺陷；2) 架构较新，实践案例少

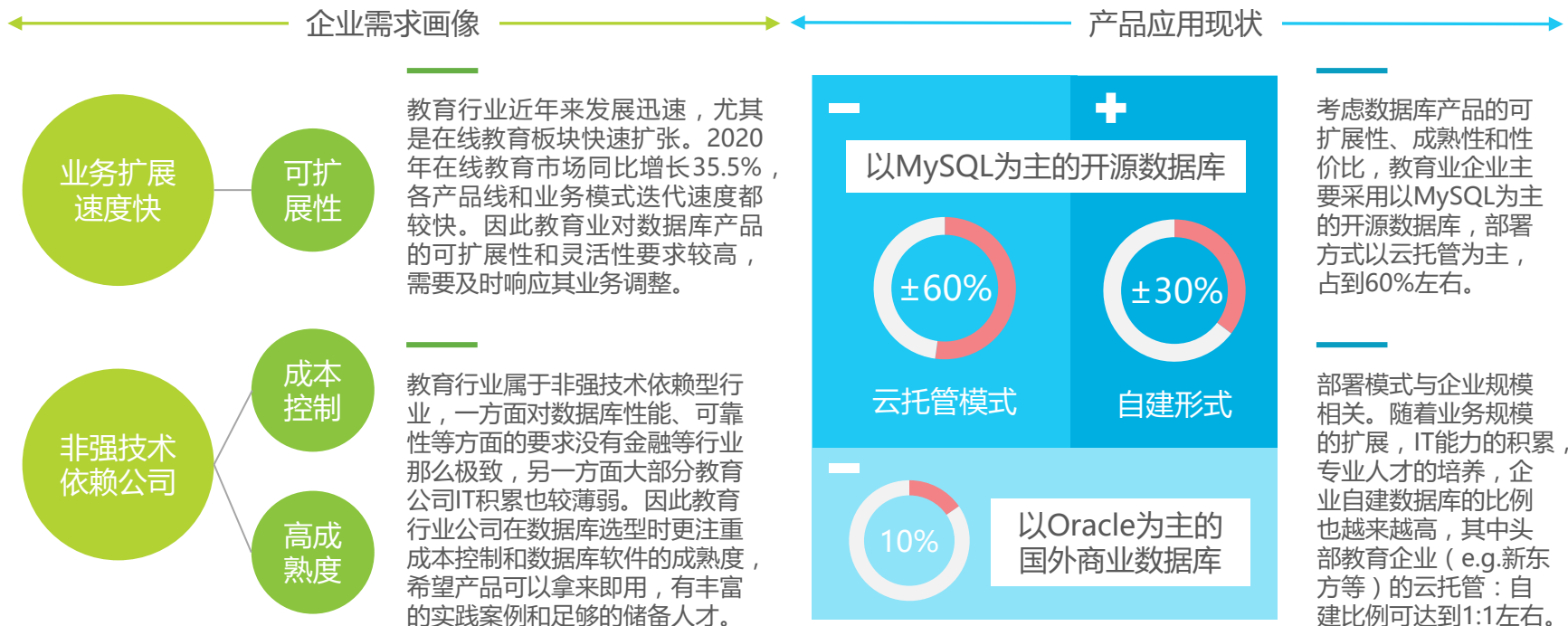
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 示例场景——互联网教育

## 业务扩张速度快，看重可扩展性、成本控制和成熟度

受利互联网和疫情，教育行业近年来呈“营收高增长，业务快迭代”发展特点。因此，企业在进行选型时较看重数据库的可扩展性，希望产品可以及时响应公司版图的扩张和业务的变更。同时，教育行业属于非强技术导向型，企业对数据库强一致、高性能和高可靠的要求并不极致，更多会考虑产品的成本控制和成熟度。企业在选型时表现较保守，虽然看好一些新产品（e.g. HTAP数据库、云原生数据库），但更希望数据库产品工程实践丰富，可以拿来即用，且专业人才招聘容易。因此MySQL数据库成为许多（互联网）教育企业的最佳选择，同时考虑成本和便捷性，云托管形式在业界也比较流行。

### 典型互联网教育业数据库选型



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

产品与技术：数据库内涵与分类

1

供给与需求：数据库市场现状与选型

2

案例与启示：数据库典型厂商案例

3

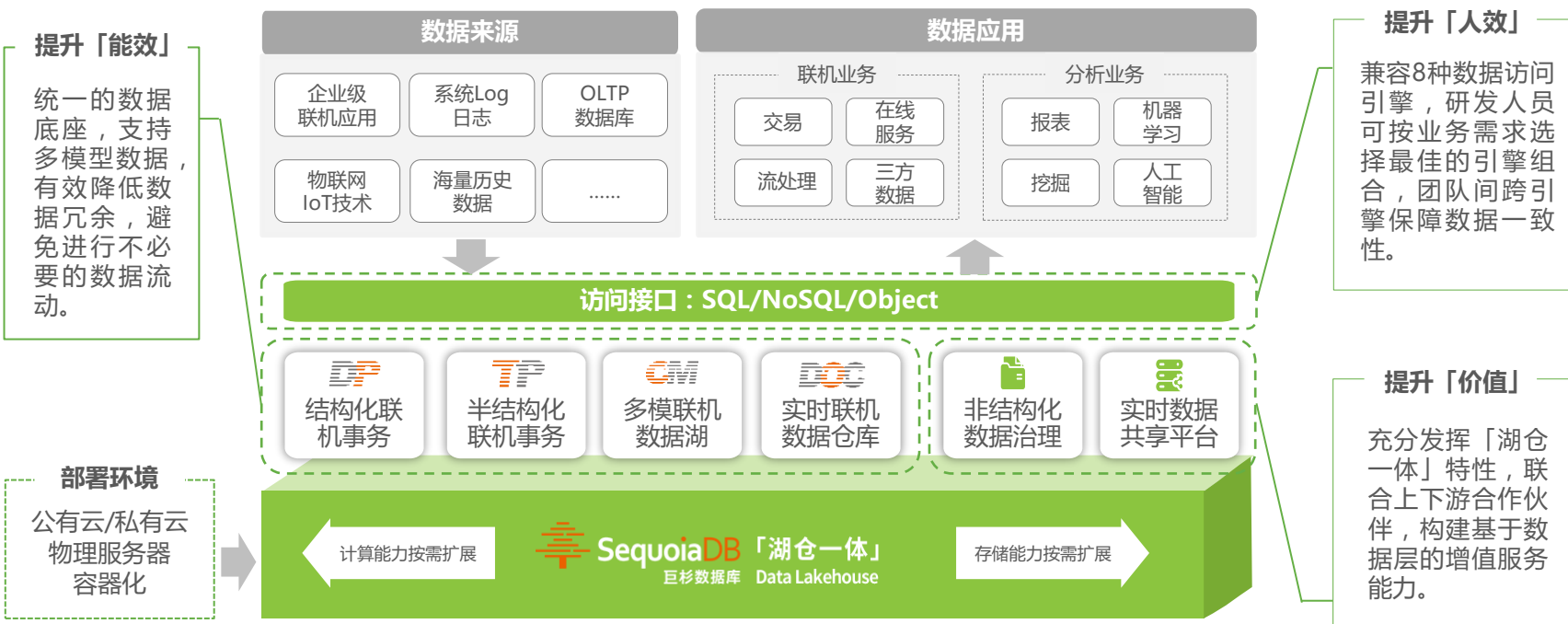
机遇与挑战：数据库未来发展趋势

4

## 从数据湖到湖仓一体，聚焦金融级数字化转型数据基础设施

巨杉数据库产品自2011年起投入研发，是国内少数实现全内核自研的分布式数据库独立厂商，10年来聚焦于金融行业。当前巨杉数据库已经在民生银行、广发银行、恒丰银行、渤海银行等股份制银行；广东省农信、吉林省农信、四川省农信等省级农信行；上海银行、长沙银行、广州银行等城商农商行；以及PICC人保、中国结算等超过100家头部金融银行业客户规模化生产上线。与传统为每个系统提供独立数据库的模式不同，基于「湖仓一体」架构，巨杉数据库打通了不同业务类型、不同数据类型之间的技术壁垒，实现交易分析一体化、流批一体化、多模数据一体化，最终降低数据流动带来的开发成本及计算存储开销，提升企业的运作的“人效”和“能效”。

### 巨杉数据库「湖仓一体」产品架构体系



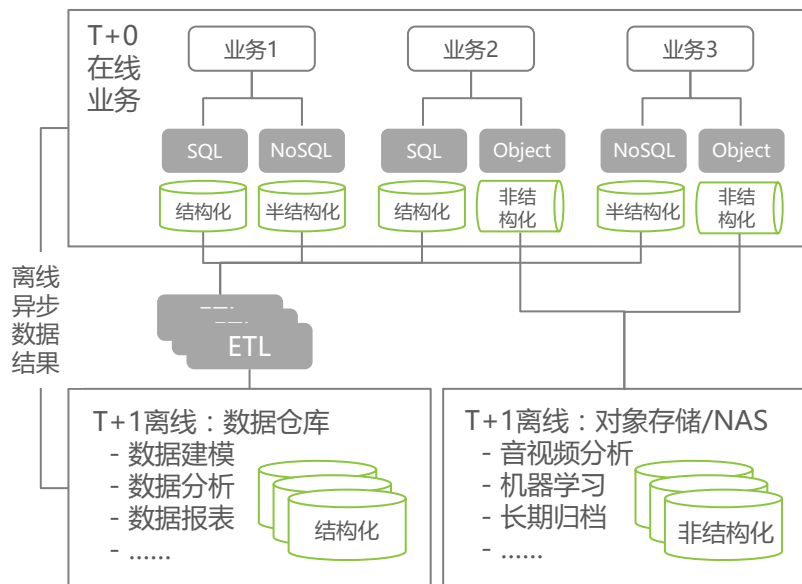
来源：艾瑞咨询研究院自主研究及绘制。



## 金融行业万亿级数据量「湖仓一体」融合数据平台典型架构

巨杉数据库在金融银行业生产环境中，运行时间最长的巨杉数据库集群已经超过7年，最大单客户集群规模达300台物理服务器，所管理的单集群最大数据量达到1万2000亿条。在巨杉数据库典型的应用架构中，企业通常基于其产品「湖仓一体」的架构特性，构建数据基础设施平台，以整合以往分散管理的结构化、半结构化、非结构化数据。巨杉数据库充分兼容包括：MySQL、MariaDB、PostgreSQL、MongoDB、Apache Spark、S3、NAS、SDB API在内的8种接口，特有的跨引擎事务一致性能力，可以有效简化多团队开发流程中对不同引擎及结构的ACID管理，提升业务开发、数据处理、运维管理多方面企业的综合数据管理效率。

### 传统数据平台 → 巨杉湖仓一体架构



**传统数据平台：**各业务及数据类型独立管理，数据不实时，运维复杂，难以实现统一容灾

**巨杉数据库「湖仓一体」架构：**多模型统一管理，数据T+0实时可用，易于运维管理，跨数据中心整体容灾

来源：艾瑞咨询研究院自主研究及绘制。



# 国内领先的企业级大数据和人工智能解决方案提供商

国双是中国领先的企业级大数据和人工智能解决方案提供商，致力于成为企业和政府组织数字化、智能化转型最值得信赖的合作伙伴，国双英文名称（Gridsum）意为网格求和，这一概念由创始人祁国晟于2003年提出，与当今主流的分布式计算不谋而合。国双自2005年成立以来，在工业互联网、智慧城市、智慧能源、智慧司法、智能营销、财税等领域为客户提供安全可靠的数字化、智能化解决方案和数据仓库等大型基础软件产品。

## 国双数据库产品与技术

### ★ 国双数据仓库

**特点** 数据查询、加载、批量处理速度快、可拓展性强、数据安全性强

- 数据接入便捷，适配多源异构数据库
- 数据处理门槛低，支持海量多维数据存储与处理
- 灵活便捷，主题（指标）数据一站式开发，多样化的数据开发、展示、分析工具
- 成熟健全的数据资产管理体系，保证数据治理质量高可用
- 数据管控安全保障平台可靠性

### 应用场景

- 企业精细化运营，建设企业大屏
- 构建管理驾驶舱，洞悉企业发展全貌
- 整合业务链协同，优化业务执行流程
- 打造供应商、客户画像，落地数据价值

### 技术支撑

#### 云原生数据库

- 国产化生态
- 性能极致
- 安全可靠
- 功能强大
- 智能易用

#### 图数据库

- 支持分布式存储与计算
- 无限大的数据量
- 支持常用的图运算与图嵌入运算

#### 时序数据库

- 高性能
- 通过插件适配其它数据库协议
- 数据聚合查询
- 高效检索

### 其他产品

#### 企业级大数据平台

#### 智能可视分析平台

#### 营销云系列

#### 智能运维管理平台



#### 企业级研发效能平台

#### 知识智能平台

#### 工业互联网平台 COMPaaS

## 打破数据壁垒，提升数据价值，赋能业务发展

国双结合多年企业管理及运营分析经验，基于大数据、人工智能、可视化、微服务等技术，形成了新一代数据仓库平台产品及解决方案，可通过对数据自动化加工、建模、智能分析，实现企业数据全在线，推进企业业务数据化向数据资产化、资产服务化演进，助力企业数字化、智能化转型升级，赋能企业高质量发展。国双数据仓库在2019年实现了国内大型央企的国产替代，实现了该集团公司的生产经营数据汇聚、共享和应用的云化服务，构建集团数据生态的核心系统。

### 国双产品服务领域与应用案例



#### 能源数据仓库

##### 问题

- 超40个生产经营管理相关的系统相互间数据不通，横向的数据共享、分析无法实现。
- 系统历史数据总量超2400TB，性能出现了瓶颈，无法快速扩容。
- 数据分析以报表和指标为主，缺乏数据可视化分析挖掘及预警预测类分析。

##### 解决效果

- **数据整合**：将数据全部整合到数据中台，打破了系统间的壁垒。
- **数据目录统一管理**：建立集团统一的数据目录，便于集中管理。
- **建立示范应用**：建立审计、物资、财务等示范应用并在集团内部推广。
- **分布式计算和建模分析**：利用分布式计算、建模、聚合等方法，对原有业务查询逻辑进行优化，提高查询效率。

#### 智能解决方案 大数据与人工

##### 问题

- 资产管理难度大
- 降本增效压力大
- 安全环保责任大
- 数字化转型发展任务重

##### 解决效果

- **大数据基础平台**：优化数据接入、数据治理与数据分析流程。
- **油气知识共享管理平台**：国双的油气行业专家建立了15万余专业分词，5800余条专业词条，为面向具体业务场景的知识共享奠定了有力的基础。
- **智能生产**：智能油气藏、智能井场、智能管道、智能站库、智能生产管控，提高生产效率，实现生产过程集中管控与智能化运行，推动流程标准化与轻量化

产品与技术：数据库内涵与分类

1

供给与需求：数据库市场现状与选型

2

案例与启示：数据库典型厂商案例

3

机遇与挑战：数据库未来发展趋势

4

## 分布式实践仍存在许多问题

横向扩展 (scale out) 的设计思想并非创新，但在历史上一直未被广泛应用。一方面是由于分布式环境使事务ACID特性难以实现，另一方面是由于单机性能升级即可满足企业的数据需求。但随着摩尔定律在某种程度上的失效和互联网的发展，硬件性能升级无法匹配海量数据的增长，分布式在近年来广受关注，许多厂商都推出了相应的分布式数据库产品。然而，在分布式的前提下，还有许多待探索的问题，例如分布式事务的解决、架构的创新、数据分片的智能化、企业级能力的提升等。

### 分布式现存问题

#### 分布式事务

##### ➤ 如何解决分布式事务问题？

分布式数据库将需要处理的事务进行拆分，再部署到不同的服务器上进行处理。对于单机较容易实现的ACID，分布式环境中出现了更多的难题。现阶段各家提出相应的解决方案，但2PC/3PC、TCC机制、事件队列/本地消息表机制、最大努力通知机制等解决方案都不完美，需要进一步的探索。

#### 架构创新

##### ➤ 如何针对分布式进行架构的创新？

各企业在进行分布式改造时，往往会面临“中间件+分库分表”或“NewSQL”的技术路线选择。传统的分库分表解决方案已经发展的较为成熟，在“高并发、强一致、低延时”的场景下表现也较好，但对业务的侵入性强，中间件负担过重，可扩展性较差。NewSQL路线从底层架构上就做了分布式的改造，可扩展性强，但在多并发和低延时上还存在一定的改造空间。

#### 数据分片

##### ➤ 如何科学高效地进行分片？

分布式通过分库分表进行数据的拆分使得各个表的数据量保持在阈值以下，从而应对高并发和海量数据。但是数据量和模式的增加了DBA和开发工程师工作的难度。如何选择合适的分片字段？如何选择合适的哈希函数？许多从业者都感受到了“人”能力的边界，进而寻求算法的创新来提升分片的效率和质量。

#### 企业级能力

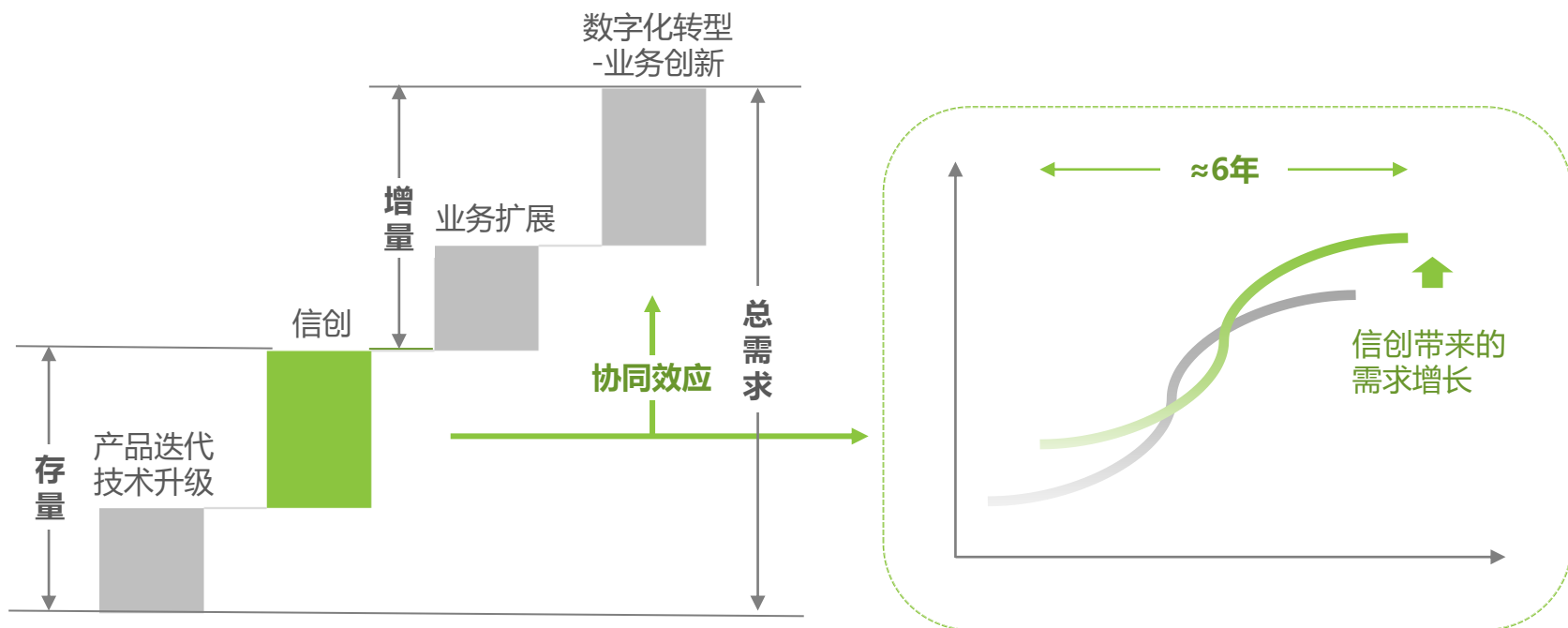
##### ➤ 如何加强分布式数据库本身的企业级能力？

传统的数据库为客户提供了很多企业级能力，例如存储过程、复杂查询。然而这些企业级能力在分布式下具有众多挑战（例如存储过程跨网络），现阶段还需要应用工具层的叠加才来满足企业客户需求，未来分布式赛道的各厂商还需进一步升级产品的企业级能力。

## 信创为国产厂商提供成长沃土，未来发展仍待市场磨炼

信创即信息技术应用创新，是在复杂国际政治背景下，国家政策引导的新一轮信息产业创新。“信创”无论是对产业端还是需求端企业都是一个重要的契机。国产数据库厂商借助政策东风，有利于其拓展市场，将产品放到实际场景中打磨，不断更新迭代，实现自己技术实力真正的弯道超车；传统行业企业、政府等也可以借此契机，实现数字化转型和业务的创新发展。但信创并非一日之功，从产业发展规律来看，新一轮的技术变革往往需要长达6年左右的实践和积累，需要上下游厂商和企业共同的长期努力。

### 信创带来存量市场的高速增长



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

## 数据迁移、多库管理提出新的挑战

面对业务形态多样、商业模式多变、需求变化频繁的当代市场，数据库和应用系统存在的形式也愈发的丰富。一个企业往往拥有多个系统，从本地到云端，从关系型到非关系型，从OLTP到OLAP，从国外品牌到国产品牌，数据库之间的跨库查询、数据导出迁移、结构变更等操作已成为常态。数据迁移频繁、多库并存的现状，使得企业后期的使用成本（运维成本、人力成本、多技术栈学习成本、迁移成本、二开成本等）大幅提高，也为数据库厂商提出了“统一管理”的新挑战。

### 数据迁移及多库管理的痛点



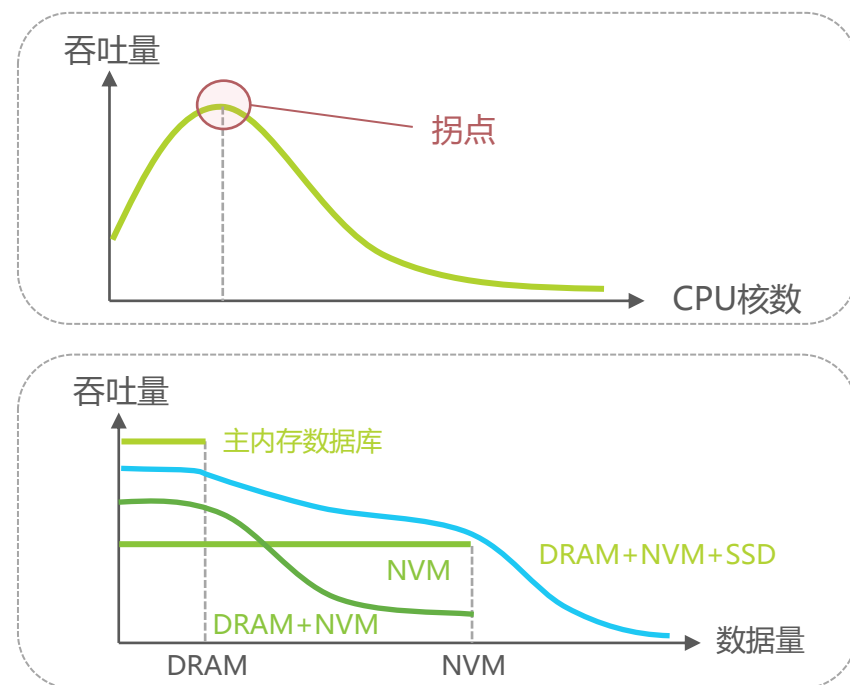
来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 挑战四

## CPU、内存等硬件变化为数据库设计提供更多的想象空间

数据库系统遵循“木桶理论”，硬件和软件作为系统的核心组件，互相制约，互相促进。而新型硬件的发展为数据库软件的发展带来了新的挑战。例如多核CPU技术走向成熟，但实际应用中并发控制出现冲突，使得核数增加带来的性能增益出现限制，如何进行多核CPU调度优化为厂商提出了新的难题。同时，大容量内存和高速硬盘走向普及，NVM非易失内存也逐渐成熟，内存的潜力释放，如何搭配新存储介质设计新的数据库架构也有待探索。因此，一些数据库领先企业如Oracle、阿里等都开始探索数据库软硬一体机的设计与实践。

### 软硬件协同设计



由于竞争带来的效率问题，CPU核数的简单叠加并不能实现数据库处理能力的无限扩张，增长曲线存在拐点。在拐点之前，吞吐量随着核数的增加而增加；在拐点之后，吞吐量则随着核数的增加而下降。当今学术界虽然提出了一些解决方案，例如MOCC（主操作控制中心），但距实际应用还存在一定的距离。

非易失内存具有掉电不易失、高速读写负载等优点，使得“把全部数据放到内存里”的设计思路变得切实可行，也为数据库系统提供了更大的设计空间。现有设计思路主要包括三个方向：直接使用NVM存储，使用NVM存储并添加DRAM作为缓冲区，以及DRAM处理热数据+NVM处理暖数据+SSD处理冷数据方案。

软硬件协同设计

来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。



# 趋势一：多场景融合

## 结合细分场景的多样发展是必然选择，用户简单化需求驱动的一体化融合也不容忽视

从产品视角来看，不同场景具有不同的特性，对数据库读写性能、吞吐量、一致性等方面的要求各有不同。为支持不同场景下的不同要求，数据库多样化是必然的选择。例如，物联网场景下写入的数据量特别大，对实时性的要求特别高，但数据天然是时间有序的且具有静态特征，因此时序数据库会较传统的事务型数据库更有优势。

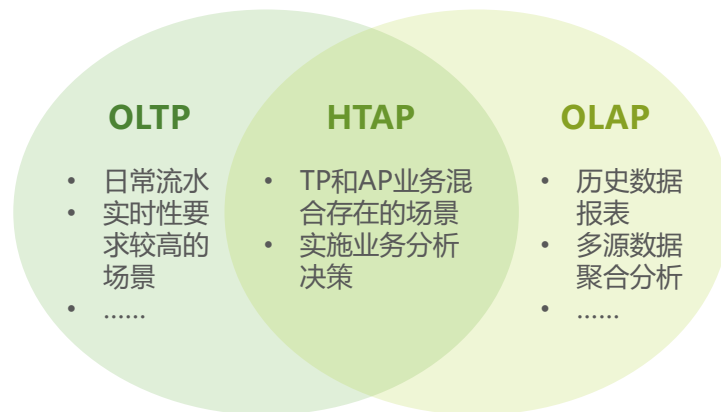
从用户视角来看，除部分头部互联网公司外，其他大中小型企业的IT人员比例都不高，对口数据库团队的人数会更少。对于他们而言，各种日新月异的技术栈、多种类型的数据库产生了极大的学习成本和维持成本，因此希望能够有一体化的产品/平台直接使用或者统一管理。对于该种类型的企业而言，会更倾向采购能满足其70%全部需求的一体化产品，而非能100%满足其部分需求的多个产品。当今市场上的HTAP/NewSQL数据库、多模数据库、统一管理平台等即满足了企业简单化一体化的需求，因此在多场景大背景下的“融合”也是不容忽视的趋势。

### “多场景”和“融合”的趋势同时存在

#### 不同场景适用不同的数据库类型



#### HTAP数据库对TP和AP功能进行了融合



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。



## 趋势二：DBaaS

### 解决弹性伸缩问题，为供应商和企业提供更多的想象空间

随着企业业务规模的扩张，数字化转型的推进，其对数据库系统也提出了新的要求：传统建设模式项目周期长，不能匹配业务上新的速度；企业部署多套数据库系统，系统间割裂，缺少统一的管理平台；资源采购和体系规划按现有规模建设，难以随业务的变化而弹性伸缩等。DBaaS (Database as a Service) 即将IT基础资源以服务化的方式提供给数据库，以及多租户和动态调整来解决成本和响应问题。部分对数据自主性和安全性要求较高的大型企业，可以选择以私有云或者专有云的方式进行数据库的云化改造。

#### DBaaS成为大势所趋

##### ■ 弹性伸缩

企业无需考虑基础资源层的问题，可以根据自身的业务需求变化进行弹性的扩展。

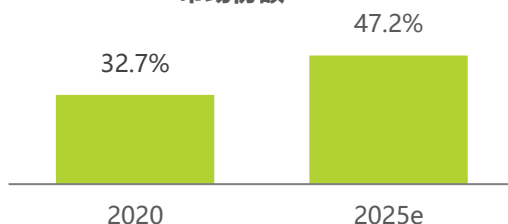
##### ■ 节约成本

无论是架设在公有云/专业云还是私有云上的数据库，相较传统的自建机房都可以一定程度上降低成本。

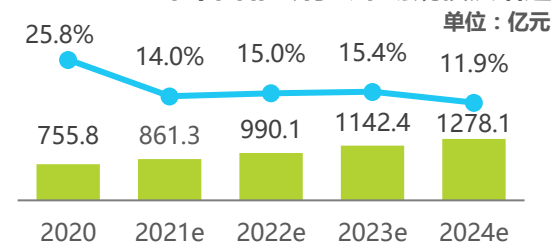
##### ■ 业务/技术创新

云的方式屏蔽了底层的部署，软件开发者可以根据业务部门创新需求，随时进行软件的开发创新

2020&2025年中国公有云数据库  
市场份额



2020-2024年中国非公有云市场规模及增速



据艾瑞统计预测，公有云数据库市场将进入平缓增长期。从市场份额来看，2025年预计将由2020年的32.7%扩张到47.2%，虽仍是数据库领域增速较高的一个赛道，但较上一个5年增速有所下滑。

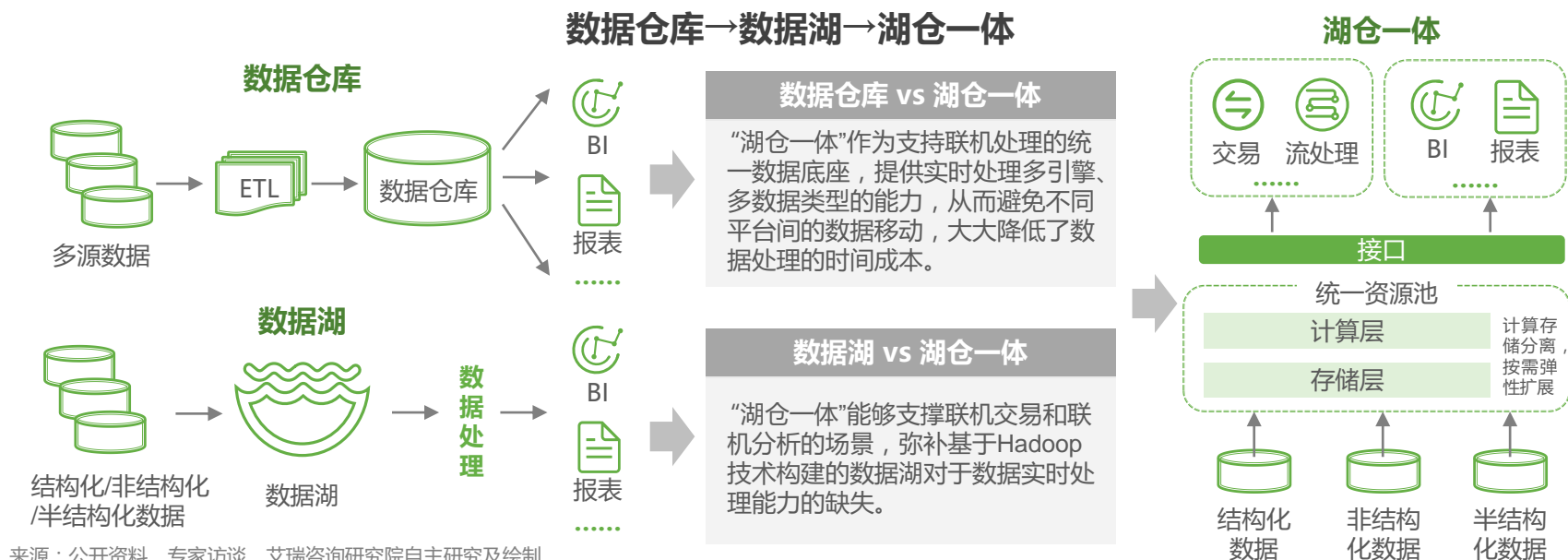
地方政府、科研院校、以及金融、运营商、工业、能源电力等传统行业企业近年来都开始了容器化上云改造的尝试，从测试场景、互联网创新业务入手，运用云弹性的能力，赋能数据库系统性能、扩展性各方面的升级。

# 趋势三：湖仓一体

## 架构创新，同时实现海量大数据的联机交易和联机分析

1980s以来，数据仓库技术不断发展，尤其MPP架构使得DBMS能够处理大量数据，满足企业通过数据分析来支持商业决策的需求。但随着互联网的发展，许多企业需要同时处理非结构化数据，半结构化数据以及海量结构化数据。数据湖随之诞生，它可以直接存储各种格式的原始数据，根据用户需求进行计算，具有灵活弹性的优点。但是，数据湖虽然适用于存储多元化数据，却缺少一些企业级功能，在实际执行时也存在许多挑战：数据缺少加工，难以实现实时分析，数据查询性能差；不支持ACID事务等。

面对企业海量大数据场景下的联机交易、非结构化数据治理的需求，以及数据仓库/数据湖架构的局限，以Snowflake、Databricks、阿里云、巨杉数据库为代表的新一代“湖仓一体”数据库厂商快速崛起。湖仓一体架构下打通了数仓和数据湖，并融合了两种架构的优势，底层多套存储系统并存且互相数据共享，形成了资源池，上层各引擎可以通过一体的封装接口访问，实现了联机交易和联机分析的同时支持。



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 趋势四：开源

## 开源模式成为产业互联网时代数据库厂商的破局之刃

从产业发展角度来，开源模式一方面提高了数据库产品开发的“效率”，它将源代码开放，避免了研发人员对基础程序的重复开发；另一方面它也有助于产品的技术“创新”，开源社区最大程度上汇聚了全球的资源力量，为开发者提供了交流切磋的空间，从而加速创新灵感的诞生。对于厂商而言，尽管从表面上看，企业部署开源并不能获得直接的利益，但在这个过程中，它们却可以布局产品的生态建设（包括人才培养、市场教育、实践反馈、企业文化、产品影响力、配套周边产品等），从而为自己获得有利的战略地位。在当今移动互联网向产业互联网发展的转折点，开源模式未尝不是各云厂商、传统厂商、新兴厂商扩展市场的破局之刃。

### 开源模式的驱动因素



来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

# 趋势五：智能化

## 人工智能延伸DBA的能力半径，优化数据库性能

随着移动互联网到产业互联网的发展，数据每日呈指数级的增长，且呈现多模态特性。面对复杂海量的数据，越来越多种类的数据库出现，需要调试的范围越来越广。但大多优化任务仍落在DBA身上，需要其进行手动调优，致使人工能力逐渐跟不上数据库的发展。而人工智能可以弥补人能力的不足，解决许多存在多年的数据库问题，例如资源的调度、索引的设计和优化、查询的优化、负载均衡设计、缓存失效等。AI 通过优化算法，对任务进行有效地预测、分析和自动化，从而减少了人工成本并大大提高数据库的性能。尤其是未来随着云上数据库更大范围内的普及，智能资源调度将成为各供应商需要面对的下一个课题。

### 人工智能适用性及应用实例

#### 数据复杂性

数字时代数据具有“多模态、海量、指数式增长”的特点，这使得DBA难以了解数据全貌，把握数据规律。人需要借助机器学习延展自身的能力半径。

#### 数据库复杂性

为满足企业多种需求，数据库采用不同的索引、数据结构、设计架构、存储介质和优化方案，使得数据库变得越来越复杂，可调试的内容也越来越多。



机器学习等AI技术从数据中自动分析获得规律，并利用规律对未知数据进行预测。它可以广泛应用于数据库的各方面，增强算法，以科学的模型操作海量数据，提高处理效率。

#### ✓ 学习索引

把索引看成一个映射函数/模型，通过机器学习预测被查询值的位置，从而减少IO，提高查询速度，节省内存。

#### ✓ 查询优化

避免DBA大量的手动调优，通过强化学习、神经网络、线性模型等方法实现连接顺序优化、查询性能预测、索引选择等。

#### ✓ 存储选择

通过深度学习分析，在关系、图、键值对等结构中自动选择最优的存储结构，提高数据库空间的应用率。

#### ✓ 负载预测

分析预测数据库工作负载，并基于算法自动为数据库系统的所有组件构建最佳数据结构和算法。

#### ✓ 缓存优化

针对异步数据合并造成缓存失效问题，运用机器学习算法预测可能访问的数据并提前加载到缓冲池，提高缓存命中率。

# 报告说明

## 致谢

本报告撰写过程中，艾瑞拜访了诸多优秀数据库厂商和下游客户，并与企业负责人进行了深入的交流沟通。

他们为报告的撰写提供了大量的有益帮助、指导和启发，在此对所有受访人及所处企业表示最真诚的感谢和祝福！



# 艾瑞新经济产业研究解决方案



## 行业咨询

- 市 场 进 入 为企业提供市场进入机会扫描，可行性分析及路径规划
- 竞 争 策 略 为企业提供竞争策略制定，帮助企业构建长期竞争壁垒



## 投资研究

- IPO行业顾问 为企业提供上市招股书编撰及相关工作流程中的行业顾问服务
- 募 投 为企业提供融资、上市中的募投报告撰写及咨询服务
- 商业尽职调查 为投资机构提供拟投标的所在行业的基本面研究、标的项目的机会收益风险等方面的深度调查
- 投后战略咨询 为投资机构提供投后项目的跟踪评估，包括盈利能力、风险情况、行业竞对表现、未来战略等方向。协助投资机构为投后项目公司的长期经营增长提供咨询服务

# 关于艾瑞




艾瑞咨询是中国新经济与产业数字化洞察研究咨询服务领域的领导品牌，为客户提供专业的行业分析、数据洞察、市场研究、战略咨询及数字化解决方案，助力客户提升认知水平、盈利能力和综合竞争力。

自2002年成立至今，累计发布超过3000份行业研究报告，在互联网、新经济领域的研究覆盖能力处于行业领先水平。

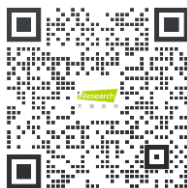
如今，艾瑞咨询一直致力于通过科技与数据手段，并结合外部数据、客户反馈数据、内部运营数据等全域数据的收集与分析，提升客户的商业决策效率。并通过系统的数字产业、产业数据化研究及全面的供应商选择，帮助客户制定数字化战略以及落地数字化解决方案，提升客户运营效率。

未来，艾瑞咨询将持续深耕商业决策服务领域，致力于成为解决商业决策问题的顶级服务机构。

## 联系我们 Contact Us

 400 - 026 - 2099

 [ask@iresearch.com.cn](mailto:ask@iresearch.com.cn)



企 业 微 信



微 信 公 众 号



# 法律声明

## 版权声明

本报告为艾瑞咨询制作，其版权归属艾瑞咨询，没有经过艾瑞咨询的书面许可，任何组织和个人不得以任何形式复制、传播或输出中华人民共和国境外。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。

## 免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。



# 为商业决策赋能

EMPOWER BUSINESS DECISIONS



艾 瑞 咨 询