数据库分类

数据库技术从理论研究到原型开发与技术攻关，再到实际产品研制与应用，形成了良性循环，成为了计算机领域的成功典范。数据库技术产生于20世纪60年代，至今已经有60年的历史。在这60年里，数据库研究日新月异，新技术、新系统层出不穷，应用领域也广泛深入，尤其是大数据时代的到来催生了技术和系统的不断演进。据统计，截止2022年6月，国产数据库产品已经高达250个。如今数据库系统是一个大家庭，当读者步入数据库领域时，面对众多的数据库技术和系统难免产生迷惑和混乱。本小节将从应用场景和数据模型两方面对数据库进行分类。

OLTP与OLAP

数据库的应用场景分为两类：在线事务处理（On-Line Transaction Processing, OLTP）和在线分析处理（On-Line Analynatic Processing, OLAP）。

- OLTP也称为事务型应用，是指对数据库的更新和查询操作，通常是对一个或一系列元组的查询和修改，如火车售票系统（12306）、银行交易系统、电商购物系统等。
- OLAP也称为分析型应用，是指对数据的查询和分析操作，通常是对海量的历史数据进行查询和分析，能够从历史数据中挖掘出有价值的信息支持企业的管理决策，如金融风险预测预警系统、证券股市违规分析系统，用户购物偏好分析系统等。

	OLTP应用	OLAP应用
数据	状态数据 可更新	历史数据 不可更新
操作类型	简单的增删改查操作 一次操作的数据量小	复杂的SQL查询 一次操作的数据量大
系统要求	支持事务ACID特性 具备高的事务处理性能	支持复杂查询功能 具备高的查询处理性能

图 9.1 OLTP应用与OLAP应用的区别

OLTP和OLAP两者之间的差异是很大的，首先数据是不同的，其次数据的操作方式是不一样的，最后对数据库管理系统的要求也是不一样的。图9.1中总结了两者的不同。

- 在数据方面，OLTP应用管理的是状态数据，即反映当前现实世界某一刻的状态，如银行账户的余额、学生本学期的选课记录和课程成绩、景区当日的售票情况等。OLAP应用管理的是历史数据，描述过去发生的事情，比如大学四年里学生的所有成绩、一个月内景区的售票情况、用户的历史购物记录等。
- 在数据操作方式方面，OLTP应用是对状态数据进行增删改查操作，基于事务来完成应用逻辑。状态数据会被不断更新，但是不会对它进行复杂查询。OLAP应用只对历史数据进行查询，历史数据不会被修改，但是会对其进行复杂的SQL查询，如聚集运算、多表连接等。通常，OLTP应用一次操作的数据量比较小，而OLAP应用一次操作的数据量很大。
- 在要求方面，OLTP应用要求数据库管理系统具备事务处理功能，能够保证事务的ACID属性，同时事务处理性能要高能够快速响应用户请求。OLAP应用要求数据库管理系统具备复杂查询的处理能力并且能够快速响应查询请求。在大数据时代下，OLTP和OLAP应用还要求数据库管理系统具备高可扩展性（满足数据量增长的需要）、容错性（保证分布式系统的可用性）和可伸缩性（能够进行按需分配资源）。

为了能够更好地支持OLAP应用，20世纪80年代数据仓库（Data Warehouse，DW）技术应运而生。数据仓库的定义是一个用以更好地支持企业（或组织）决策分析处理的、面向主题的、集成的、不可更新的、随时间不断变化的数据集合。数据仓库本质上和数据库一样，是长期存储在计算机内、有组织、可共享的数据集合。但是，数据仓库是专为数据的分析和处理而出现的技术。数据仓库中的数据来自于多个OLTP应用数据库系统，这些数据必须经过抽取（Extract）、清洗转换（Transform）之后加载（Load）到数据仓库。这一过程称为ETL过程，它的目的是将企业中分散、零乱、标准不统一的数据整合在一起。基于整合后的数据，企业可以构建面向主题的OLAP应用并用可视化图表展示分析结果。

数据仓库出现之后，数据仓库为OLAP应用服务，传统的数据库为OLTP应用服务，二者各司其职，泾渭分明。因此，数据库管理系统按应用场景可以分为面向分析型应用的数据库，如TeraData、SybaseIQ、Greenplum等，和面向事务型应用的数据库，如Oracle、MySQL、PostgreSQL等。另外，还有一类数据库管理系统，如TiDB、OceanBase等，它们目前以支持OLTP应用为主，未来计划支持混合事务/分析型应用（Hybrid Transactional/Analytical Processing，HTAP）。

SQL与NoSQL

数据库的数据模型分为两类：关系模型（SQL）和非关系模型（NoSQL）。关系模型是以一张二维表来描述结构化数据。非关系模型主要用于描述文本、图形图像、音频、视频、网页等半结构化/非结构化数据，也可以用于描述结构化数据。非关系模型包括键值（Key-Value）模型、列簇式（Column-Family）模型、文档（Document）模型、图（Graph）模型和时序（Time Series）模型。

按数据模型分类，数据库可以分为关系型数据库和非关系型数据库，数据仓库也可以分为关系型数据仓库和非关系型数据仓库，其中数据库服务于OLTP应用，数据仓库服务于OLAP应用。

- 关系型数据库：是传统集中式关系数据库和新型分布式关系数据库的总称，它们支持SQL语法和事务处理，保证事务的ACID特性。传统集中式关系数据库，如Oracle、DB2、MySQL，长期以来一直是企业OLTP业务系统的核心和基础，但是它的扩展性差、成本高，难以支持大数据时代海量数据的存储和高并发事务处理请求。新型分布式关系数据库（NewSQL），如VoltDB、OceanBase、Spanner、TiDB等，采用分布式架构、横向扩展（scale-out）以及数据分区和数据备份（一般三份）的方式实现了大数据提出的可扩展性和可用性等要求，同时充分利用多核CPU、大内存、固态硬盘（SSD）、非易失存储设备（NVM）、远程直接数据存取（RDMA）等技术来提高海量数据的事务处理性能和事务处理吞吐量。新型分布式关系数据库已经逐渐成为支持面向海量数据OLTP应用的主要力量；
- 非关系型数据库：是以互联网大数据应用为背景发展起来的非关系型、分布式、不保证满足ACID特性的数据管理系统。非关系数据库的种类多种多样，包括键值数据库，如Redis、RocksDB、DynamoDB等；列簇式数据库，如BigTable、Hbase、Hypertable等；文档数据库，如MongoDB、Couchbase等；图数据库，如Neo4j、AlibabaGDB、TGDB等、时序数据库，如TDengine、DolphinDB、IoTDB等。这些非关系数据库能够满足大数据时代海量数据的存储和处理需求，具有高可扩展性、高性能和高可用性。它们大都不支持SQL语法，也没有一种通用的操作语言，大都不支持事务处理，不保证数据的强一致性，只支持简单的增删改查操作。简单的增删改查操作足以满足互联网OLTP应用对半结构化数据和非结构化数据的处理请求；
- 关系型数据仓库：是以关系模型和SQL查询引擎为基础对海量结构化数据进行复杂查询和分析的数据管理系统。面向OLAP分析应用的关系数据仓库系统通常采用Shared Nothing的大规模并行处理体系（Massively Parallel Processing, MPP ）架构，支持较高的扩展性和处理性能，采用列存储的方式提高数据压缩率和数据读取的I/O效率，同时也会利用大内存、多核CPU等新硬件来提高对大数据分析的性能。常见的关系数据仓库有MonentDB、SAP HANA、Clickhouse、AnalyticDB、TDSQL-A等；
- 非关系型数据仓库：是以非关系模型和并行计算框架（MapReduce、Spark、Flink）为基础对海量半结构化/非结构化进行分析和挖掘的数据管理系统。在大数据时代，OLAP应用从对海量结构化历史数据的多维分析发展为对海量半结构化/非结构化数据的复杂分析和深度挖掘，以及对异构数据的分析挖掘。

OLAP应用从传统面向主题的分析应用扩展为面向数据的分析应用，从数据本身出发，构建模型，挖掘数据之间的内在联系，然后进行预测、推理等。这些支持OLAP分析应用的非关系数据仓库系统具有高度的扩展性、容错性和高效的数据处理性能。常见的非关系数据仓库有Hive、Impala、Presto、Druid、Kylin、Solr、ElasticSearch、Milvus等。这些非关系数据仓库有些不支持SQL查询接口，有些为了更好的用户体验支持SQL语法。

图9.2总结了数据管理系统在应用场景和数据模型两个维度上的分类。

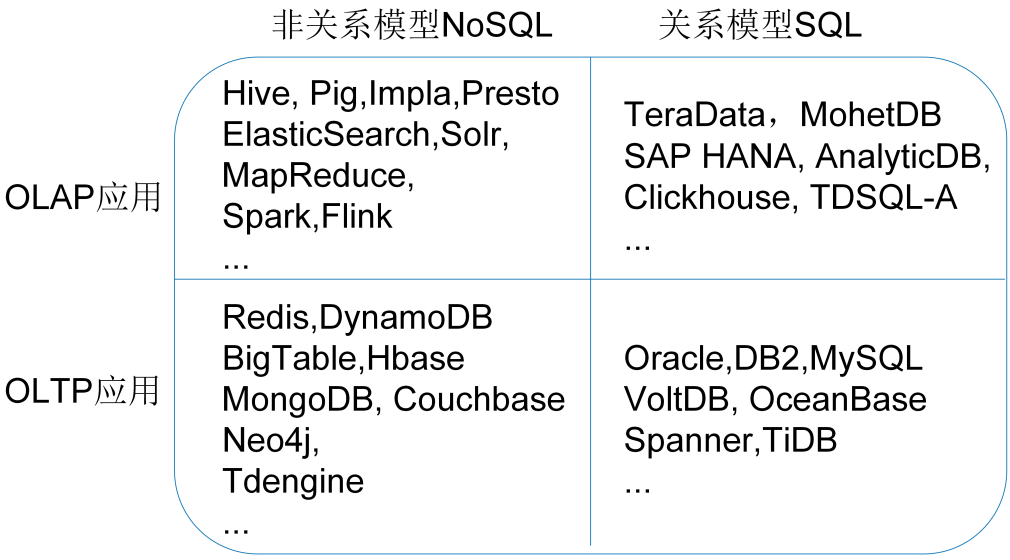

图 9.2 数据管理系统分类