

数据库系统安装

◆ 项目提出

要使用数据库,一般有两种方式,方式 1 是在设备上安装好数据库管理系统,然后使用客户端进行本地或者远程的连接并获取数据库服务,这也是比较传统的做法,代表性的产品有 MySQL 数据库、达梦数据库、Oracle 数据库等。而方式 2 是采购远程的云数据库服务,这是随着网络通信能力的提升和云技术的发展在近些年才出现的一种产品服务形式。代表的产品有华为的 GaussDB 数据库。安装数据库是计算机专业技术人员必备的基础技能,从数据库管理系统的安装开始学习数据库技术也是非常有必要的。

◆ 项目分析

人们在开始接触数据库时,就会遇到各种各样的术语,如数据库、数据库技术、数据库系统、数据库管理系统等都是在行业中工作、交流经常要使用的术语。正确使用术语是专业人士的基本修养,也是后续开展工作和继续学习所必需的基础,所以应首先规范术语和概念。

其次要使用数据库,首先要确定使用什么产品,而数据库技术经过多年的发展和技术衍化,产品类型、品牌繁多,因此,需要从总体上对该项技术的起源、发展脉络以及今后的技术趋势有一个总体的了解,同时要对我国信创数据库技术发展有所了解,才能为具体的应用场景进行产品选型。

最后要确定获取数据库服务的方式。上文所述的第 1 种方式仍然是目前大部分企业更倾向于采用的。所以,在对数据库选型之后,数据库管理系统的安装和配置往往也是必不可少的步骤,也是专业人士必须掌握的基本技能。

任务 1.1 数据库管理系统的安装和配置

1.1.1 相关知识

1. 数据库的概念

数据库技术是信息系统的核心技术,是一种计算机辅助管理数据的方法,它研究如何组织和存储数据,如何高效地获取和处理数据。即数据库技术是研究、管理和应用数据库的一门软件科学。

(1) 数据。数据(Data)是人们用来反映客观世界而记录下来的可以鉴别的数字、字母、符号、图形、声音、图像、视频信号等的总称。我们这里所说的数据是经编码后可存入计算机中进行相关处理的符号集合。数据一般分为数值型数据和



数据库的概念

非数值数据两大类,数值型数据(如 32、78.91 等)主要用来进行科学计算(加、减、乘、除等运算),而非数值数据(如人的姓名、工作简历等)主要用来进行比较和查找、统计等操作。数据和信息密不可分,我们可以说信息是人们消化理解了的数据,其关系如图 1-1 所示。

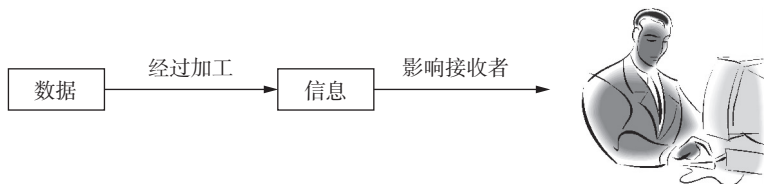


图 1-1 数据与信息的关系

(2) 数据库。J. Martin 给数据库(DataBase,DB)下了一个比较完整的定义:“数据库是存储在一起的相关数据的集合,这些数据是结构化的,无有害的或不必要的冗余,并为多种应用服务;数据的存储独立于使用它的程序;对数据库插入新数据,修改和检索原有数据均能按一种公用的和可控制的方式进行。”

通俗地说,数据库是长期存储在计算机存储器中、按照一定的数学模型组织起来的、具有较小的冗余度和较高的数据独立性,可由多个用户共享的数据集合。也就是说,数据库是按照数据结构来组织、存储和管理数据的仓库,并且其中的数据不是随意堆积在一起的内容,而是有组织有管理的数据聚集(图 1-2)。

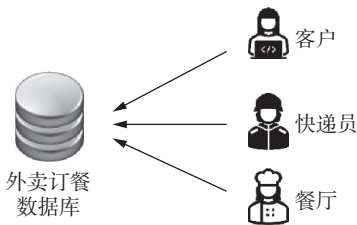


图 1-2 数据库服务

(3) 数据库管理系统。数据库管理系统(DataBase Management System,DBMS)是一个能够科学地组织和存储数据、高效地获取和维护数据的系统软件,是位于用户与操作系统之间的数据管理软件。它对数据库进行统一的管理和控制,以保证数据的安全性和完整性。数据库管理系统和操作系统一样,是计算机系统的基础软件,如图 1-3 所示。

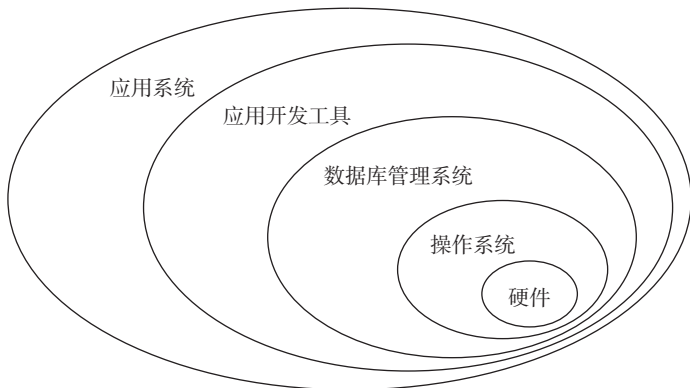


图 1-3 计算机系统层次结构图

数据库管理系统的主要功能如下。

① 数据定义功能:数据库管理系统提供数据定义语言(Data Definition Language,DDL),用户通过它可以方便地对数据库中的数据对象的组成与结构进行定义。

② 数据组织、存储和管理功能:数据库管理系统要分类组织、存储和管理数据,此功能涉及数据字典、用户数据、数据的存取路径等。数据库管理系统还要确定以何种文件结构和存取方式,在存储空间中组织这些数据,以及如何实现数据之间的联系。数据组织和存储的基本目标是提高存储空间利用率、方便进行数据存储,以及提供多种数据存储方式来提高存取效率。

③ 数据操纵功能:数据库管理系统还提供数据操纵语言(Data Manipulation Language,DML),用户可以使用它操纵数据,实现对数据的基本操作,如查询、插入、删除和修改等。

④ 数据库的事务管理和运行管理功能:数据库在建立、运行和维护时,由数据库管理系统统一管理和控制,以保证事务的正确运行,确保数据的安全性、完整性,保证多用户对数据的并发使用及发生故障后的系统恢复。

⑤ 数据库的建立和维护功能:包括数据库初始数据的输入和转换功能,数据库的转储、恢复功能,数据库的重组功能和性能监视、分析功能等。这些功能通常由一些程序或管理工具实现。

(4) 数据库系统。数据库系统(DataBase System,DBS)是指和数据库有关的整个计算机系统,包括计算机硬件、操作系统、数据库管理系统以及在其支持下建立起来的数据库、应用程序、用户和数据库维护人员等。有时也将人以外与数据库有关的硬件和软件系统称为数据库系统,广义数据库系统如图1-4所示。

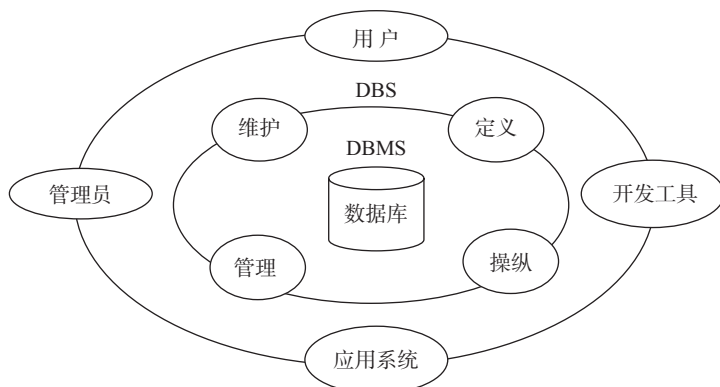


图 1-4 广义数据库系统

2. 数据库技术发展史

计算机的主要应用之一是数据处理,即对各种数据进行收集、存储、加工和管理等活动,其中数据管理是数据处理的中心问题,是对数据进行分类、组织、编码、存储、检索和维护的活动。数据管理技术伴随着计算机技术的不断发展,经历了3个发展阶段。

(1) 人工管理阶段。从计算机出现到20世纪50年代中期,计算机主要用于科学计算。在这个阶段,数据是程序的组成部分,数据的输入、输出和使用都是由程序来控制的,数据在使用时随程序一起进入内存,用完后完全撤出计算机,如图1-5所示。人工管理数据阶段因为应用程序和数据之间的依赖性太强,程序员的负担很



数据库技术的发展

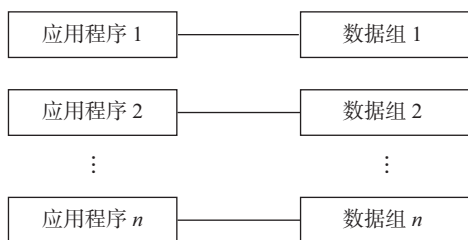


图 1-5 人工管理阶段数据和程序关系图

重,数据冗余量也很大。

(2) 文件系统阶段。到了 20 世纪 60 年代早中期,在这一阶段,按照一定的规则把成批数据组织在数据文件中,存放于外存储器上,由操作系统统一存取。在文件系统阶段,程序和数据之间的关系如图 1-6 所示。

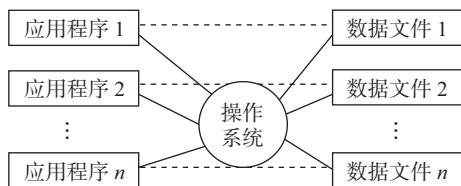


图 1-6 文件系统阶段程序和数据关系图

(3) 数据库系统阶段。自 20 世纪 60 年代后期,数据处理的规模急剧增长。同时,计算机系统中采用了大容量的磁盘(数百 MB 以上)系统,联机存储大量数据成为可能。为了解决数据的独立性问题,实现数据的统一管理,达到数据共享的目的,数据库技术得到了极大的发展。

在这个阶段,所有程序中的数据由 DBMS 统一管理,应用程序和数据实现了完全独立,数据得到高度共享,此阶段应用程序和数据之间的关系如图 1-7 所示。

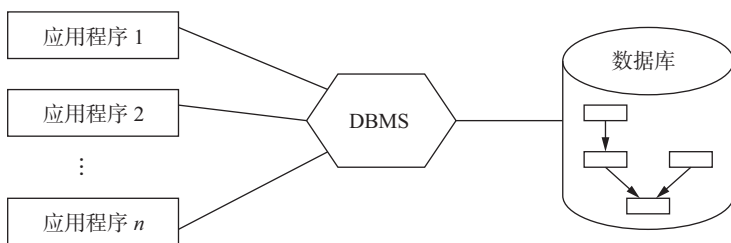


图 1-7 数据库系统阶段程序与数据之间的关系

数据库与之前的几种数据管理方式相比,具有以下优势。

- ① 整体数据结构化,从而减少了程序员的工作量。
- ② 数据共享程度高,系统弹性大,易于扩充。减少了数据冗余,节约了存储空间,避免了数据之间的不相容和不一致。
- ③ 数据独立性强。将数据从应用程序中独立出来,实际上就是把数据和应用程序解耦,原来的强耦合方式灵活性太低,开发量大,维护任务繁重。
- ④ 统一管理和控制。用户使用数据库系统,便于对数据进行统一管理和控制,包括数

据的安全性保护、数据的完整性检查、并发控制、数据恢复等。

3. 数据模型

数据库领域中常用的数据模型有4种,分别是层次模型、网状模型、关系模型和面向对象模型。其发展过程如图1-8所示。



数据模型

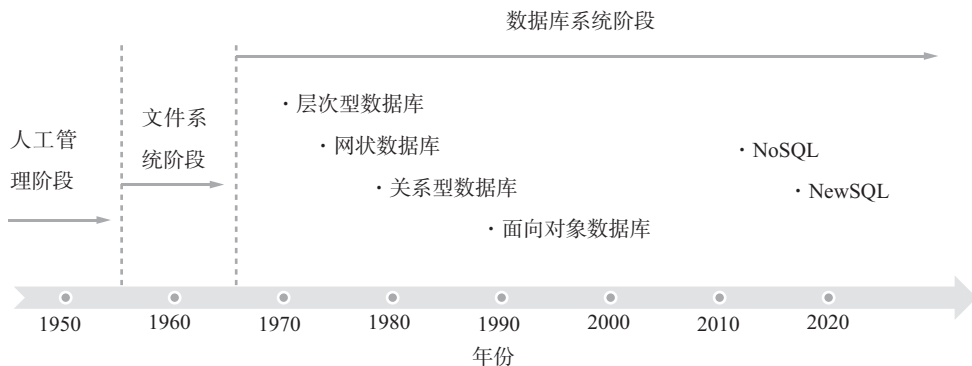


图 1-8 数据管理技术发展的时间线

层次模型和网状模型的数据库系统在20世纪70年代至80年代初非常流行,在当时的数据库产品中占据了主导地位,但现在已经完全被关系模型的数据库产品所取代。20世纪80年代末以来,面向对象的方法和技术在计算机程序设计语言、软件工程、信息系统设计等领域得到了普遍应用,也就促进了数据库中面向对象数据模型的研究和发展。

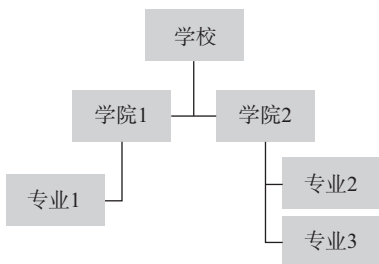


图 1-9 层次模型

(1) 层次模型。层次模型是数据库系统中最早出现的数据模型,它可以用树状(层次)结构表示实体类型及实体间联系的数据模型,曾经得到广泛的应用,其数据结构如图1-9所示。现实世界中许多实体之间的联系本来就呈现出一种很自然的层次结构,如家族关系、行政机构等。

(2) 网状模型。在数据库中,把满足以下两个条件的基本层次联系集合称为网状模型。

① 允许一个以上的结点无双亲。

② 一个结点可以有多个的双亲。

在网状模型数据库中也是以记录为数据的存储单位,而一个记录又包含若干数据项。网状数据库是导航式(Navigation)数据库,在查找语句中不但要说明查找的对象,而且要规定存取路径,其结构如图1-10所示。

利用网状数据库模型对于层次和非层次结构的事务都能比较自然的模拟,在关系数据库出现之前网状DBMS要比层次DBMS用得普遍。在数据库发展史上,网状数据库曾经占有重要地位。

(3) 关系模型。关系模型是用二维表的形式表示实

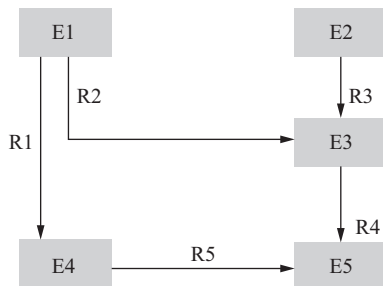


图 1-10 网状模型

体和实体间联系的数据模型。关系模型是当前最主流的数据模型,它的出现使层次模型和网状模型逐渐退出了数据库历史的舞台。

关系数据库理论出现于 20 世纪 60 年代末到 70 年代初。1970 年,IBM 的研究员 E. F. Codd 博士发表《大型共享数据银行的关系模型》一文并提出了关系模型的概念。后来 Codd 又陆续发表多篇文章,奠定了关系数据库的基础。

关系数据模型提供了关系操作的特点和功能要求,但不给 DBMS 的语言给出具体的语法要求。对关系数据库的操作是高度非过程化的,用户不需要指出特殊的存取路径,路径的选择由 DBMS 的优化机制来完成。Codd 在 20 世纪 70 年代初期的论文论述了范式理论和衡量关系系统的 12 条标准,用数学理论奠定了关系数据库的基础。Codd 博士也以其对关系数据库的卓越贡献获得了 1981 年 ACM 图灵奖。

关系模型有严格的数学基础,抽象级别比较高,而且简单清晰,便于理解和使用。关系数据模型是以集合论中的关系概念为基础发展起来的。关系模型中无论是实体还是实体间的联系均由单一的结构类型——关系来表示。在实际的关系数据库中的关系也被称为表。一个关系数据库就是由若干个表组成的。关系数据库的技术和理论是本门课程的主要学习内容。

采用关系模型建立的数据库即关系数据库,具有以下特点。

- 组织数据的结构单一:在关系模型中,无论是数据还是数据之间的联系都是以我们熟悉的二维表(关系)形式来表示的,这种表示方法不仅让人容易理解,而且便于计算机操作和实现。
- 采用集合运算:在关系模型中,运算的对象是关系,运算的结果还是关系,而关系可以看作行(元组或记录)的集合,所以对关系的运算可以转化为对集合的运算。
- 数据完全独立:因为关系数据库系统中的数据是由关系数据库管理系统(DBMS)进行管理的,对于程序员来说,不需要知道数据存放的具体位置和组织形式等方面的内容,只需要告诉系统要进行什么样的操作,由系统自动完成相关的任务,即程序和数据库高度独立。
- 数学理论支持:在关系模型中,每个关系都是集合,对关系的运算有集合论、数理逻辑作为基础,关系结构可以用关系规范化理论进行优化。总之,关系模型具有严格的数学定义,具有成熟的数学理论为依据,它是目前为止最简单有效、最受欢迎、最广泛应用的数据模型。

随着数据库应用领域的扩展以及数据对象的多样化,传统的关系型数据库模型开始暴露出许多弱点,如对复杂对象的标识能力差,语意表达能力较弱,对文本、时间、空间、声音、图像和视频等数据类型的处理能力差等。例如,多媒体数据在关系型数据库中基本上都以二进制数据流形式存放,但对于二进制数据流通用的数据库标识能力差,语意表达能力差,不利于检索、查询。

为了适应新的需求,也逐渐出现了许多新的数据模型。

(4) 面向对象模型。面向对象模型是一种新兴的数据模型,它采用面向对象的方法来设计数据库。面向对象的数据库存储对象是以对象为单位,每个对象包含对象的属性和方法,具有类和继承等特点。

数据库可以将类似的对象归并为类。在一个类中的每个对象称为实例。同一类的对象

具有共同的属性和方法,对这些属性和方法可以在类中统一进行说明。消息传送到对象后,可以在其所属的类中找到这些变量,称为类变量。在一个类中,可以有各种各样的统计值,如某个属性的最大值、最小值、平均值等。这些统计值不属于某个实例,而是属于类,因此也是类变量。

随着互联网的迅速发展,其他的一些数据模型也相继出现,如基于可扩展标记语言(Extensible Markup Language, XML)的 XML 数据模型,用资源描述框架(Resource Description Framework, RDF)来描述和注解互联网资源的 RDF 数据模型等。

4. 数据库的发展特点

数据库已经成为计算机信息系统和智能应用系统的重要基础和核心技术之一,如图 1-11 所示。

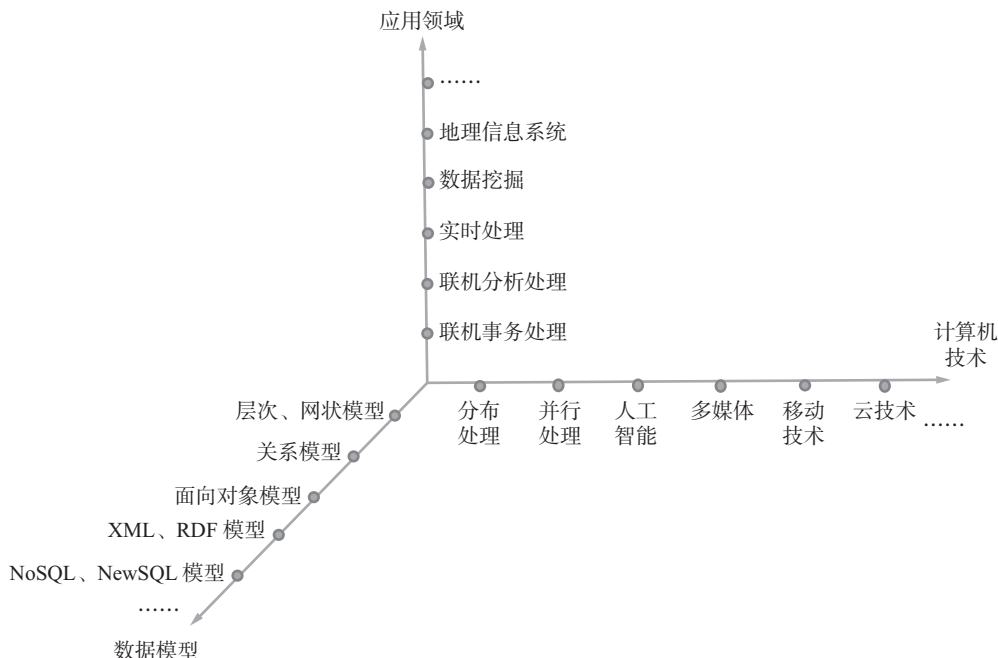


图 1-11 数据库系统发展的三个维度

数据库系统的发展有以下三个维度。

(1) 数据库的发展集中体现在数据模型的发展上。数据模型是数据库系统的核心和基础,所以数据库系统的发展和数据模型的发展密不可分,数据模型的划分维度是数据库系统划分的一个重要标准。

(2) 与其他计算机技术交叉、结合。新的计算机技术层出不穷并和其他计算机技术交叉、结合,是数据库系统发展的一个显著特征,如与分布式处理技术结合产生的分布式数据库,与云技术结合产生的云数据库等。

(3) 面向应用领域发展新数据库技术。通用数据库在特定领域无法满足应用需求,需要根据相关领域的特定需求来研发特定的数据库系统。

5. 数据管理技术的新挑战

虽然不断涌现了许多数据模型,但是这些数据模型都因为缺乏便携性和通用性等问题,

未能替代关系型数据模型成为通用的数据库产品的基本模型。

随着大数据时代来临,大数据的 4V 特性对传统关系型数据库提出了全面挑战。大数据具有 4V 特征,首先,数据规模大(Volume),大数据通常指 100TB(1TB=1024GB)规模以上的数据量,数据量大是大数据的基本属性。其次,数据种类繁多(Variety),随着传感器种类的增多及智能设备、社交网络等的流行,其需要处理的数据包括结构化数据、半结构化数据和非结构化数据。再次,数据处理速度快(Velocity),数据从生成到消耗,时间窗口非常小,可用于生成决策的时间非常短。最后,数据价值密度低(Value),数据呈指数增长的同时,隐藏在海量数据的有用信息却没有一个相应比例的增长。恰恰相反,挖掘大数据的价值类似沙里淘金,需要从海量数据中挖掘稀疏珍贵的信息。例如,商场的监控视频,连续数小时的监控过程中有可能有用的数据仅仅只有几秒,大数据特征与管理需求,如图 1-12 所示。于是,NoSQL 技术顺应大数据发展的需要,得到了蓬勃发展。



图 1-12 大数据特征与管理需求

(1) NoSQL 数据库。随着互联网应用的蓬勃发展,很多场景下,并不需要传统关系型数据库提供的强一致性以及关系型数据模型。相反,由于快速膨胀和变化的业务场景,对可扩展性(Scalability)以及可靠性(Reliable)更加需要,而这个又正是传统关系型数据库的弱点。自然地,新的适合这种业务特点的数据库 NoSQL 开始出现,其中最具有代表性的是 Amazon 的 Dynamo 以及 Google 的 BigTable,以及它们对应的开源版本,如 Cassandra 以及 HBase。由于业务模型的千变万化,以及抛弃了强一致和关系型,大大降低了技术难度,各种 NoSQL 版本像雨后春笋一样涌现,基本上成规模的互联网公司都会有自己的 NoSQL 实现。NoSQL 意为“不仅是 SQL 技术”,也就是 Not Only SQL。

不同类型的 NoSQL 数据库产品虽然各有特点,但是都具备统一的特性,即非关系型的、分布的、不保证满足 ACID 特性。

在技术上,NoSQL 数据库具备以下三个特点。

① 对数据进行分区:能够将数据分布在集群的多个结点上,利用大量结点并行处理的方式来获得高性能,同时能够支持横向扩展方式,便于集群的扩展。

② 降低 ACID 一致性约束:允许暂时不一致,接受最终一致性约束遵循的是 BASE 原则。

③ 对各数据分区提供备份:一般遵循三备份原则(在当前结点、同一个机架不同结点、不同机架不同结点上保存三份数据,用于避免结点故障和机架故障所导致的数据不安全问题,备份数量越多,数据冗余量越大,综合考虑安全性和冗余性,三份数据是最合理的设定)来应对结点故障,从而提高系统的可用性。

四类常见的 NoSQL 数据库技术是按照存储模型划分的,包括键值数据库、列存储数据

库、文档数据库和图数据库。

主要 NoSQL 数据库简介见表 1-1。

表 1-1 主要 NoSQL 数据库简介

分 类	典型产品	典型应用场景	数据模型	优 点	缺 点
键值数据库	Tokyo Cabinet/ Tyrant, Redis, Voldemort, Oracle BDB	内容缓存,主要用于处理大量数据的高访问负载,也用于一些日志系统等	Key 指向 Value 的键值对,通常用 hash table 来实现	查找速度快	数据无结构化,通常只被当作字符串或者二进制数据
列存储数据库	Cassandra, HBase, Riak	分布式的文件系统	以列簇式存储,将同一列数据存在一起	查找速度快,可扩展性强,更容易进行分布式扩展	功能相对局限
文档数据库	CouchDB, MongoDB	Web 应用(与 Key-Value 类似, Value 是结构化的,不同的是数据库能够了解 Value 的内容)	Key-Value 对应的键值对, Value 为结构化数据	数据结构要求不严格,表结构可变,不需要像关系型数据库一样需要预先定义表结构	查询性能不高,而且缺乏统一的查询语法
图(Graph)数据库	Neo4J, InfoGrid, Infinite Graph	社交网络、推荐系统等。专注于构建关系图谱	图结构	利用图结构相关算法,比如最短路路径寻址、N 度关系查找等	很多时候需要对整个图做计算才能得出需要的信息,而且这种结构不太容易做分布式的集群方案

(2) NewSQL 数据库。NoSQL 也有很明显的问题,由于缺乏强一致性及事务支持,很多业务场景被 NoSQL 拒之门外。同时,缺乏统一的高级数据模型、访问接口,又让业务代码承担了很多的负担。人们开始寻找一种既具备 NoSQL 的可扩展性,又能够支持关系模型的关系型数据库产品。这种新型数据库主要面向联机事务处理场景,同时使用 SQL 作为主要语言,所以大家称此类产品为 NewSQL 数据库。NewSQL 只是一类产品的描述,并不是具有官方定义的词语。

NewSQL 数据库产品一般具有以下特点。

① 采用新架构:如采用多结点并发控制、分布式处理,利用复制实现容错、流式控制等技术架构。这类产品有 Google Spanner、H-store、VoltDB 等。

② 采用透明分片中间件技术:这类产品的数据分片过程对用户来说是透明的,用户的应用程序不需要做出变化。这类产品有 Oracle、MySQL、Proxy、MariaDB MaxScale 等。

③ 数据库即服务:云服务商提供的数据库产品,一般都有这类具备 NewSQL 特性的数据库产品。如 Amazon Aurora,阿里云的 Oceanbase,腾讯云的 CynosDB,华为的 GaussDB

(DWS)和 GaussDB(for MySQL)。

6. 主流数据库产品

和编程语言有排行榜一样,数据库产品也有流行度排行榜,其排名每月变更一次,有全部数据库的排名,也有不同分类的排名,如关系型数据库、键值数据库、时序数据库、图数据库等专项排名,如图 1-13 所示。可以看出,在 2022 年 8 月前 20 名中出现了 8 个非关系型数据库,但关系型数据库仍然是主流产品。关系型数据库也在不断扩展自己的功能和特性。

Rank			DBMS	Database Model	Score		
Sep 2022	Aug 2022	Sep 2021			Sep 2022	Aug 2022	Sep 2021
1.	1.	1.	Oracle	Relational, Multi-model	1238.25	-22.54	-33.29
2.	2.	2.	MySQL	Relational, Multi-model	1212.47	+9.61	-0.06
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	926.30	-18.66	-44.55
4.	4.	4.	PostgreSQL	Relational, Multi-model	620.46	+2.46	+42.95
5.	5.	5.	MongoDB	Document, Multi-model	489.64	+11.97	-6.87
6.	6.	6.	Redis	Key-value, Multi-model	181.47	+5.08	+9.53
7.	↑8.	↑8.	Elasticsearch	Search engine, Multi-model	151.44	-3.64	-8.80
8.	↓7.	↓7.	IBM Db2	Relational, Multi-model	151.39	-5.83	-15.16
9.	9.	↑11.	Microsoft Access	Relational	140.03	-6.47	+23.09
10.	↓9.	↓9.	SQLite	Relational	138.82	-0.05	+10.17
11.	11.	↓10.	Cassandra	Wide column	119.11	+0.97	+0.12
12.	12.	12.	MariaDB	Relational, Multi-model	110.16	-3.74	+9.46
13.	13.	↑21.	Snowflake	Relational	103.50	+0.38	+51.43
14.	14.	↓13.	Splunk	Search engine	94.05	-3.39	+2.45
15.	↑15.	↑16.	Amazon DynamoDB	Multi-model	87.42	+0.16	+10.49
16.	↓16.	↓15.	Microsoft Azure SQL Database	Relational, Multi-model	84.42	-1.75	+6.16
17.	17.	↓14.	Hive	Relational	78.43	-0.22	-7.14
18.	18.	↓17.	Teradata	Relational, Multi-model	66.58	-2.49	-3.09
19.	19.	↓18.	Neo4j	Graph	59.48	+0.12	+1.85
20.	↑22.	↑22.	Databricks	Multi-model	55.62	+1.00	

图 1-13 数据库流行度排名

(1) 甲骨文公司的 Oracle 数据库管理系统。Oracle 是一个最早商品化的关系型数据库管理系统,也是应用广泛、功能强大的数据库管理系统。Oracle 作为一个通用的数据库管理系统,不仅具有完整的数据管理功能,还是一个分布式数据库系统,支持各种分布式功能,特别是支持 Internet 应用。对于应用开发环境,Oracle 提供了一套界面友好、功能齐全的数据库开发工具。Oracle 使用 PL/SQL 语言执行各种操作,具有可开放性、可移植性、可伸缩性等功能。Oracle Database 19c 及其后版本支持本地部署和云端部署,具有市场领先的性能、可扩展性、可靠性和安全性。Oracle 的官方网站的网址是:www.oracle.com。

(2) MySQL 数据库管理系统。MySQL 是一种关系数据库管理系统,在关系数据库中会将数据保存在不同的表中,而不是将所有数据放在一个大仓库内,这样就增加了访问速度并提高了灵活性。MySQL 所使用的 SQL 语言是用于访问数据库的最常用标准化语言,MySQL 软件由于体积小、速度快、总体拥有成本低,尤其是开放源码这一特点,从而成为最流行的中小型数据库系统。MySQL 的官方网站的网址是:www.mysql.com。

(3) 微软公司的 SQL Server 数据库管理系统。Microsoft SQL Server 是一种典型的关系型数据库管理系统,可以在许多操作系统上运行,它使用 Transact-SQL 语言完成数据操作。Microsoft SQL Server 是开放式的系统,其他系统可以与它进行完好的交互操作。在数据分析领域与时俱进,集成扩展了当今在高级数据分析领域最为流行的程序语言——R 语言。