浅谈NoSQL

张杰 / 宁夏工商职业技术学院

摘　要：NoSQL非关系型数据库已成为目前学术界和产业界研究的热点问题，它可解决传统关系型数据库不能解决的高并发读写、高可扩展性和高可用性问题。本文简单介绍了NoSQL的技术、基本数据类型以及与关系型数据库的比较。
关键词：非关系型数据库；关系型数据库；元数据；键值；数据模型

随着互联网web2.0网站的兴起，传统关系型数据库在大量数据写入处理方面暴露出难以克服的缺点，局限性显现为扩展困难、读写慢、成本高、有限的支撑容量。业界为了解决上述问题，推出了新类型的“NoSQL”数据库。

NoSQL是非关系型数据库，是一种与关系型数据库管理系统截然不同的数据库管理系统，它的数据存储格式可以是松散的、通常不支持Join操作并且易于横向扩展，由于非关系型数据库其本身的特点得到了非常迅速的发展。

1 NoSQL与关系型数据库的区别与联系

总的来说，在设计上，NoSQL非常关注对数据高并发地读写和对海量数据的存储等，与关系型数据库相比，它们在架构和数据模型方量面做了“减法”，而在扩展和并发等方面做了“加法”。大部分支持分布式集群，在某些特定场景下，能补充关系型数据库的缺点，但是缺乏统一的解决方案。传统的关系型数据库与NoSQL非关系型数据库的区别主要在于：

1.1 数据模式

区别于关系型数据库固定的二维表元组模式，NoSQL没有严格的数据模式，通常存储的是一对键值或数组，其结构不确定，在系统运行中可以动态更改。NoSQL的这种松耦合、可扩展的数据模式，有利于存储在日益广泛的Web应用中日趋重要的半结构化和非结构化数据。

1.2 可扩展性

从根本上说，NoSQL是伴随着分布式系统而产生的，在分布式系统下性能良好，支持横向扩展，能够很好的适应现代Web应用迅猛发展带来的海量数据。

1.3 事务完整性

关系型数据库通过ACID保证数据的完整性，ACID特性也是关系型数据库事务完整性最高级别的黄金标准。ACID，分别代表原子性、一致性、隔离性和持久性。NoSQL数据库优先考虑的是数据库的性能。这是NoSQL项目在企业中难以普及的原因。但另一方面，很多Web实时系统对事务完整性、读一致性要求并不高，某些时候对写一致性要求也不高。

其实NoSQL数据库仅是关系数据库在某些方面（性能，扩展）的一个弥补，单从功能上讲，NoSQL的几乎所有的功能，在关系数据库上都能够满足，所以选择NoSQL的原因并不在功能上。因此我们一般会把NoSQL和关系数据库结合使用，各取所长。

关系型数据库存储了数据之间存在的或潜在的关系，适合于企业数据的存储于查询；而NoSQL数据库更看重数据的存储，适合于现在的Web应用。

2 NoSQL的理论基础及应用分析

2.1 NoSQL相关理论

（1）CAP理论

CAP理论最早由Eric Brewer教授提出，Seth Gilbert和Nancy Lynch予以证明。CAP理论归纳了分布式系统的三个特性：

一致性：系统中的所有数据备份，在同一时刻都是同一值。

可用性：每个操作总能在确定的时间内返回，即系统随时都是可用的。

分区容忍性：在出现网络分区的情况下，例如断网，分离的系统也能正常运行。

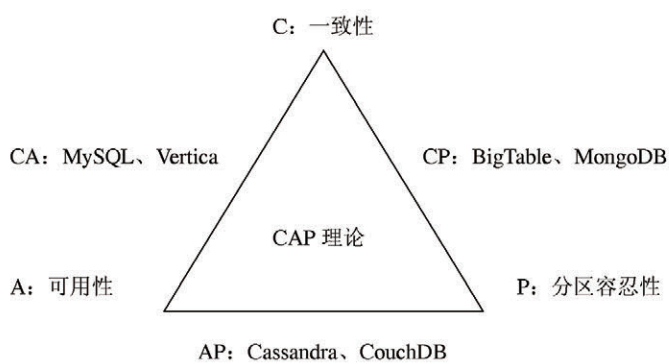

图1 CAP理论

中图分类号：TP311.138

CAP理论指出，分布式系统不可能同时实现所有这三个特性，系统必须做出权衡，至少牺牲一样以成全其他两样。图1对侧重点不同的数据库产品进行了说明。

（2）最终一致性

NoSQL中通常有两个层次的一致性：强一致性和最终一致性。强一致性指集群中的所有机器状态同步保持一致；最终一致性可以允许短暂的数据不一致，但数据最终会保持一致。根据CAP理论，强一致性无法实现与可用性、分区容忍性同步，而最终一致性是考虑用户体验的折中办法，也是与传统的RDBMS最大的不同。

（3）BASE思想

BASE思想基于CAP理论，实际上是CAP理论中AP的衍伸，是对一致性进行概化处理，这是因为在越来越多的应用和实际案例中，公认为可用性和分区容忍性比一致性更需要严格设计。这些应用设计普遍倾向于降低一致性，突出可用性和数据冗余机制，也就是强调有序地将数据分散于不同节点中。

（4）分布式系统

分布式系统是建立在网络之上的软件系统，具有高度的内聚性和透明性。内聚性是指每一个数据库分布节点高度自治，有本地的数据库管理系统。透明性是指每一个数据库分布节点对用户的应用来说都是透明的，看不出是本地还是远程。在分布式数据库系统中，用户感觉不到数据是分布的，即用户不须知道关系是否分割、有无复本、数据存于哪个站点以及事务在哪个站点上执行等。

2.2 NoSQL基本数据类型

数据模型是定义数据如何输入和输出的一种模型。其主要作用是为信息系统提供数据的定义和格式。数据模型是数据库系统的核心和基础，现有的数据库系统都是基于某种数据模型而建立起来的。

（1）基于键-值的数据模型

Key-Value存储是最简单的NoSQL存储，是非结构话的数据存储模式。它将以一种算法把“键”映射到相应的“值”（数据），而不关心数据的内容。开发者需要自行组织和定义“值”的数据格式，并进行解析。这种存储系统不支持任何非“键”的查询，Dynamo是最典型的键-值存储系统。

表1 键值模型

实例	Dynamo、Redis、Voldemort
主要应用场景	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
数据模型	Key与Value间建立的键值映射，通常用哈希表实现
优点	查找迅速
缺点	数据无结构化，通常只被当作字符串或者二进制数据

（2）面向文档的数据模型

面向文档的存储系统的代表有CouchDB和MongoDB。文档存储的数据一般用json或类似json的格式，存储内容是文档型的。MongoDB的文档数据以bson格式存储，CouchDB的文档数据以json格式存储，文档可以存储列表、键-值对以及层次结构复杂的文档。文档型存储的灵活性和复杂性是一把双刃剑：一方面，开发者可以任意组织文档结构；另一方面，应用层的查询需求会变得比较复杂。

表2 文档模型

实例	CouchDB、MongoDB
应用场景	Web应用
数据模型	与键值模型类似，value指向结构化数据
优点	数据要求不严格，不需要预先定义结构
缺点	查询性能不高，缺乏统一查询语法

（3）面向列的数据模型

这种数据模型的特点是列式存储，即每一行数据的各项被存储到不同的列中，这些列的集合称为列簇，它可以对大量行少数列进行读取和更新。代表系统有BigTable和HBase等。BigTable是Google为了有效管理海量大规模结构化数据而设计的分布式存储系统，它可以用来处理PB级的海量数据并能分布在数千台普通服务器上。HBase和Cassandra的数据模型都借鉴自Google的BigTable。

表3 列式模型

实例	Bigtable、Cassandra、HBase
应用场景	分布式文件系统
数据模型	以列存储，将同一列数据存在一起
优点	查找迅速、可扩展性强，更容易进行分布式扩展
缺点	功能相对有限

（4）图结构数据模型

图结构存储是NoSQL的另一种存储形式。基于图理论的图数据库使用节点、属性和边的概念，其中节点类似于面向对象编程中的对象概念，代表人、商业、账户等任意项的实体；属性存储与节点有关的信息，例如使用Wikipedia作为节点，那么它的属性可以是网页、引用材料或者以W开头的单词，具体选择取决于实际应用；边被用来连接节点与节点或者节点与属性，表示两者之间的关系，最重要的信息存储在边上。Neo4J和HyperGraphDB是当前最流行的图结构数据库。

表4 图形模型

实例	Neo4J
应用场景	社交网络、推荐系统、关系图谱
数据模型	图结构
优点	利用图结构相关算法提高性能
缺点	功能相对有限，不好做分布式集群解决方案

2.3 NoSQL的优劣分析

NoSQL在应用中有着灵活、低成本高性能、以扩展等优势。

（1）灵活的数据模型

在传统的RDBMS的领域，分析数据，构建数据模型时存储数据前的必须工作。然而在实际企业中需求是随着时间的推移不断变化的，虽然在传统的关系型数据库中支持一定程度的重构造，但如果应用变化太大，重

构也无济于事。NoSQL数据库打破了这种数据模型的限制，允许在一个数据单元中存入其想要的任何结构，数据单元间的联系是扁平的，一般也不受模型的限制。但在注重模式自由的同时，也需要注意管理数据的完整性。

（2）建立在低成本上的高性能

简单的数据模型，令NoSQL本身的扩展性极强，节点易于扩展；分布式的结构，使其能够适应低成本、不稳定的机器，实现低成本、高性能，因而NoSQL数据库通常通过使用廉价服务器集群来管理暴增的数据与事务。廉价服务器集群的方案，相对高性能机器的RDBMS的集群有更多的数据节点，因而能够提供了更廉价、更可靠、更多备份的服务。

（3）易扩展

RDBMS通常采用纵向扩展的方式解决数据库负载增加带来的性能不足，即购买性能更好的服务器代替旧服务器。这针对负载缓慢增加的情况是较有效的，然而如果负载一直增加，不可能每次都更换服务器，横向扩展的思想应运而生。所谓横向扩展就是把负载均衡的分到不同的主机上。虽然RDBMS也提供了横向扩展的功能，但对程序来说是半透明的，在横向扩展的时候，会大量的修改程序，甚至会停机，而NoSQL数据库在设计之初就考虑到了横向扩展，它对程序来说是透明的，可以随时添加节点、删除节点。

NoSQL也有很多不成熟的地方，开发上也有不少劣势，主要表现：

1）数据模型和查询语言没有经过数学验证

SQL这种基于关系代数和关系演算的查询结构有着坚实的数学保证，即使一个结构化的查询本身很复杂，但是它仍然能够获取满足条件的所有数据。NoSQL系统没有使用SQL，其使用的一些模型还未有完善的数学基础，这是NoSQL系统较为混乱的主要原因。

2）不支持ACID特性

这既是NoSQL的优势，同时也是其缺点。ACID特性是系统在中断的情况下也能够保证在线事务能够准确执行，而NoSQL无法实现这一功能。

3）功能简单

NoSQL系统提供的功能普遍比较简单，因而增加了应用层的负担，如果在应用层实现ACID特性，编写代码的工作极为痛苦。

4）没有统一的查询模型

NoSQL系统一般提供不同的查询模型，这在一定程度上增加了开发者的负担。

3 NoSQL在教学资源管理系统中的应用

“十二五”期间，教育部积极推进信息化教学资源的建设工作，教学资源库、精品资源共享课、视频公开课是高等学校信息化教学资源建设的重要内容。信息化教学资源的建设，旨在保护资源知识产权的原则下，最大程度的搜集和整理各种有价值的教学资源，使其在最大范围内实现整合和共享，使得教学资源呈现开放性和扩展性，从而实现优质资源高效利用、高度共享。

目前，高校网络教学资源主要包括基本教学资源、网络课程、虚拟实训、考试系统等，网络教学资源的主要目的是充分利用网络环境，既满足“助学”功能，又能发挥“助教”的作用。在系统的开发与设计过程中，将计算机技术与现代教育理念紧密结合，将各专业、各课程的知识点、技能点以媒体通过文档、演示文稿、图片、视频、动画等形式呈现，满足不同专业、不同层次学生的学习需要，同时为教师备课、教学交流提供良好的支持，有利于提高教学质量。信息化教学资源建设工作已经取得了显著成效，但随着数据量的急剧增长，系统对数据管理技术的要求不断提高，系统的一些问题也逐渐暴露出来，如系统扩展的局限性、结合教学需要的数据库存储和管理问题，以及系统的可靠性问题。

针对现有教学资源管理系统存在的问题，通过NoSQL数据库均可提出解决方案。

随着web应用的普及和数据的爆炸式增长，NoSQL已成为目前学术界和产业界研究的热点。相比而言，NoSQL较关系型数据库在处理海量数据等方面具有一定的优势，目前已在国内的部分网站建设中得到应用，如视觉中国网站的建设，相信在未来的发展中将会更好的适应web环境的要求。

参考文献：

[1]吕明育，李小勇.NoSQL和关系数据库的比较分析[J].微型电脑应用，2011，27(10).

[2]陆嘉恒.大数据挑战与NoSQL数据库技术[M].

[3]陈莉莹.双锴.NoSQL数据库综述[J].中国科技论文在线.

[4]潘凡.从MySQL到MongoDB——视觉中国的NoSQL之路[J].程序员，2010(6)：79-81.

作者简介：张杰（1966.07-），女，满族，辽宁大连人，副教授，硕士，研究方向：计算机基础应用。

作者单位：宁夏工商职业技术学院 信息技术系，银川 750021