大数据相关技术综述

课程名称：大数据

学院专业：计算机学院计算机科学与技术专业

学生姓名：刘泽春

学　　号：2012060010017

2013年12月30日

Big Data是近来的一个技术热点，但从名字就能判断它并不是什么新词。毕竟，大是一个相对概念。历史上，数据库、数据仓库、数据集市等信息管理领域的技术，很大程度上也是为了解决大规模数据的问题。被誉为数据仓库之父的Bill Inmon早在20世纪90年代就经常将Big Data挂在嘴边了。

然而，Big Data作为一个专有名词成为热点，主要应归功于近年来互联网、云计算、移动和物联网的迅猛发展。无所不在的移动设备、RFID、无线传感器每分每秒都在产生数据，数以亿计用户的互联网服务时时刻刻在产生巨量的交互……要处理的数据量实在是太大、增长太快了，而业务需求和竞争压力对数据处理的实时性、有效性又提出了更高要求，传统的常规技术手段根本无法应付。

在这种情况下，技术人员纷纷研发和采用了一批新技术，主要包括分布式缓存、基于MPP的分布式数据库、分布式文件系统、各种NoSQL分布式存储方案等。

10年前，Eric Brewer提出著名的CAP定理，指出：一个分布式系统不可能满足一致性、可用性和分区容忍性这三个需求，最多只能同时满足两个。系统的关注点不同，采用的策略也不一样。只有真正理解了系统的需求，才有可能利用好CAP定理。

架构师一般有两个方向来利用CAP理论。

* Key-Value存储，如Amazon Dynamo等，可以根据CAP理论灵活选择不同倾向的数据库产品。
* 领域模型+分布式缓存+存储，可根据CAP理论结合自己的项目定制灵活的分布式方案，但难度较高。

对大型网站，可用性与分区容忍性优先级要高于数据一致性，一般会尽量朝着A、P的方向设计，然后通过其他手段保证对于一致性的商务需求。架构设计师不要将精力浪费在如何设计能满足三者的完美分布式系统，而应该懂得取舍。

不同的数据对一致性的要求是不同的。SNS网站可以容忍相对较长时间的不一致，而不影响交易和用户体验；而像支付宝这样的交易和账务数据则是非常敏感的，通常不能容忍超过秒级的不一致。

Cache篇

缓存在Web开发中运用越来越广泛，mem-cached是danga.com（运营LiveJournal的技术团队）开发的一套分布式内存对象缓存系统，用于在动态系统中减少数据库负载，提升性能。

memcached具有以下特点：

协议简单；基于libevent的事件处理；内置内存存储方式；memcached不互相通信的分布式。

memcached处理的原子是每一个（Key，Value）对（以下简称KV对），Key会通过一个hash算法转化成hash-Key，便于查找、对比以及做到尽可能的散列。同时，memcached用的是一个二级散列，通过一张大hash表来维护。

memcached由两个核心组件组成：服务端（ms）和客户端（mc），在一个memcached的查询中，ms先通过计算Key的hash值来确定KV对所处在的ms位置。当ms确定后，mc就会发送一个查询请求给对应的ms，让它来查找确切的数据。因为这之间没有交互以及多播协议，所以memcached交互带给网络的影响是最小化的。

MemcacheDB是一个分布式、Key-Value形式的持久存储系统。它不是一个缓存组件，而是一个基于对象存取的、可靠的、快速的持久存储引擎。协议与memcached一致（不完整），所以很多memcached客户端都可以跟它连接。MemcacheDB采用Berkeley DB作为持久存储组件，因此很多Berkeley DB的特性它都支持。

类似这样的产品也很多，如淘宝Tair就是Key-Value结构存储，在淘宝得到了广泛使用。后来Tair也做了一个持久化版本，思路基本与新浪MemcacheDB一致。

分布式数据库篇

支付宝公司在国内最早使用Greenplum数据库，将数据仓库从原来的Oracle RAC平台迁移到Greenplum集群。Greenplum强大的计算能力用来支持支付宝日益发展的业务需求。

Greenplum数据引擎软件专为新一代数据仓库所需的大规模数据和复杂查询功能所设计，基于MPP（海量并行处理）和Shared-Nothing（完全无共享）架构，基于开源软件和x86商用硬件设计（性价比更高）。

分布式文件系统篇

谈到分布式文件系统，不得不提的是Google的GFS。基于大量安装有Linux操作系统的普通PC构成的集群系统，整个集群系统由一台Master（通常有几台备份）和若干台TrunkServer构成。GFS中文件备份成固定大小的Trunk分别存储在不同的TrunkServer上，每个Trunk有多份（通常为3份）拷贝，也存储在不同的TrunkServer上。Master负责维护GFS中的 Metadata，即文件名及其Trunk信息。客户端先从Master上得到文件的Metadata，根据要读取的数据在文件中的位置与相应的TrunkServer通信，获取文件数据。

在Google的论文发表后，就诞生了Hadoop。截至今日，Hadoop被很多中国最大互联网公司所追捧，百度的搜索日志分析，腾讯、淘宝和支付宝的数据仓库都可以看到Hadoop的身影。

Hadoop具备低廉的硬件成本、开源的软件体系、较强的灵活性、允许用户自己修改代码等特点，同时能支持海量数据存储和计算任务。

Hive是一个基于Hadoop的数据仓库平台，将转化为相应的MapReduce程序基于Hadoop执行。通过Hive，开发人员可以方便地进行ETL开发。

如图3所示，引用一张Facebook工程师做的Hive和Hadoop的关系图。

NoSQL篇

随着数据量增长，越来越多的人关注NoSQL，特别是2010年下半年，Facebook选择HBase来做实时消息存储系统，替换原来开发的Cassandra系统。这使得很多人开始关注HBase。Facebook选择HBase是基于短期小批量临时数据和长期增长的很少被访问到的数据这两个需求来考虑的。

HBase是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PC Server上搭建大规模结构化存储集群。HBase是BigTable的开源实现，使用HDFS作为其文件存储系统。Google运行MapReduce来处理BigTable中的海量数据，HBase同样利用MapReduce来处理HBase中的海量数据；BigTable利用Chubby作为协同服务，HBase则利用Zookeeper作为对应。

总结篇

近来NoSQL数据库的使用越来越普及，几乎所有的大型互联网公司都在这个领域进行着实践和探索。在享受了这类数据库与生俱来的扩展性、容错性、高读写吞吐外（尽管各主流NoSQL仍在不断完善中），越来越多的实际需求把人们带到了NoSQL并不擅长的其他领域，比如搜索、准实时统计分析、简单事务等。实践中一般会在NoSQL的外围组合一些其他技术形成一个整体解决方案。

准实时的统计分析

* 传输时统计分析，Stream Processing技术：FlumeBase、S4。
* 查询时统计分析，结果集较小时，可以直接在返回前做统计分析处理。

比如买家消费记录查询的HBase实现，Schema设计，rowkey=uid，column=搜索词和查询值，version=交易id。

搜索相关

* 充分利用NoSQL（比如HBase）内部数据的有序性、Row Key、Column Family、Version Timestamp。

我们用“HBase+二次索引”来实现实时营销的解决方案。

* 构建一个外围系统完成索引建立。

Google MegaStore

也有人开始尝试基于HBase的MegaStore实现

简单事务

* 事务处理服务 + NoSQL存储。
* 淘宝开发的千亿级海量数据库Oceanbase，通过update server角色执行将写操作限制在一台机器上，实现事务。
* Google MegaStore通过为用户提供机制，根据应用特点划分entity group，将事务涉及的数据分布到一台机器上，实现事务。

总之，大数据时代，信息和数据是最重要的资源，一旦信息自由、数据开放，就意味着信息和每一个公民之间是等距的，而且中间没有层级的过滤。数据的开放和流动，就代表着知识的开放和流动、代表着权力的开放和流动，这种开放和流动是多中心的、水平的。这个时代社会的主体结构将从“分层”转向“结网”，在网状传播力量的不断冲击下，个人的主体价值将得到前所未有的张扬。

大数据时代正在推动下一波社会化的浪潮：一个更开放的社会，一个权力更分散的社会、一个网状社会。