Google三大论文读后感

Google的三驾马车：Google fs，Mapreduce 和 Bigtable, 这三个产品Google并没有公布相关源码，但是发布了这三个产品的详细设计论文。Yahoo资助的Hadoop也有按照这三篇论文的开源Java实现:Hadoop对应Mapreduce, Hadoop Distributed File System (HDFS)对应Google fs, Hbase对应Bigtable。不过在性能上Hadoop比Google要差很多。

PART I

Bigtable是一个管理结构化数据的分布式存储系统，它被设计用来处理海量数据。Bigtable的设计目的是可靠地适应PB级别的数据和成千上万台机器。Bigtable已经实现了下面的几个目标：广泛的适用性、可扩展、高性能和高可用性。已经有超过60个Google的产品和项目在使用Bigtable，包括Google Analytics、Google Finance、Orkut、Personalized Search、Writely和Google Earth。这些产品使用Bigtable完成迥异的工作负载需求，这些需求从面向吞吐量的批处理作业到对终端用户而言延时敏感的数据服务。它们使用的Bigtable集群的配置也有很大的差异，从少数机器到成千上万台服务器，这些服务器里最多可存储几百TB的数据。

在很多方面，Bigtable和数据库很类似：它使用了很多数据库的实现策略。并行数据库和内存数据库已经具备可扩展性和高性能，但是Bigtable提供了一个和这些系统完全不同的接口。

Bigtable不支持完整的关系数据模型；与之相反，Bigtable为客户提供了简单的数据模型，利用这个模型，客户可以动态控制数据的布局和格式，也就是对BigTable而言，数据是没有格式的，用数据库领域的术语说，就是数据没有Schema，用户自己去定义Schema，用户也可以自己推测在底层存储中展示的数据的位置属性，比如具有相同前缀的数据存放的位置接近，在读取的时候可以将这些数据一次性读出来。数据用行和列的名字进行索引，名字可以是任意的字符串。虽然客户程序通常会在把各种结构化或半结构化的数据串行化到字符串里，Bigtable同样将数据视为未经解析的字符串。通过仔细选择数据的模式，客户可以控制数据的位置。最后，可以通过BigTable的模式参数动态地控制数据读或写。

Bigtable是一个稀疏的、分布式的、持久化存储的多维度排序Map，Map由行关键字、列关键字以及时间戳索引；Map中的每个value都是一个未经解析的字节数组。



表中的行关键字是任意字符串，在单一行关键字下的每一个读或者写操作都是原子的（不管在这一行里被读或者写的不同列的数目），这个设计决策能够使用户很容易地推测（reason about）对同一个行进行并发更新操作时的系统行为。

Bigtable通过行关键字的字典顺序来维护数据。表中一定范围内的行被动态分区。每个分区叫做一个”Tablet”，Tablet是数据分布和负载均衡的单位。这样做的结果是，读取一定范围内的少数行很高效，并且往往只需要跟少数机器通信。用户可以通过选择他们的行关键字来开发这种特性，这样可以为他们的数据访问获得好的本地性。

列关键字组成的集合叫做“列族“，列族构成了访问控制的基本单位。存放在同一列族下的所有数据通常都属于同一个类型（我们把同一个列族下的数据压缩在一起）。列族必须先创建，然后才能在列族中任何的列关键字下存放数据；列族创建后，其中的任何一个列关键字下都可以存放数据。我们的意图是，一张表中不同列族的数目要小（最多几百个），并且列族在操作中很少改变。与此相反，一张表可以有无限多个列。

列关键字的命名语法如下：列族：限定词。访问控制、磁盘和内存的计数都是在列族层面进行的。上述的控制权限能帮助我们管理不同类型的应用：一些应用可以添加新的基本数据、一些可以读取基本数据并创建派生的列族、一些则只允许浏览现存数据（甚至可能因为隐私的原因不能浏览所有现存列族）。

在Bigtable中，每一个数据项都可以包含同一数据的不同版本；这些版本通过时间戳来索引。Bigtable时间戳是64位整型数。时间戳可由Bigtable指定，这种情况下时间戳代表精确到毫秒的“实时”时间，或者该值由库户程序明确指定。需要避免冲突的程序必须自己生成一个唯一的时间戳。数据项中不同版本按照时间戳倒序排列，所以最新的版本可以被先读到。

Bigtable提供了建立和删除表以及列族的API函数。Bigtable还提供了修改集群、表和列族的元数据的API，比如修改访问权限。

Bigtable是建立在一些其他Google基础架构之上的。BigTable使用Google分布式文件系统(GFS)存储日志和数据文件。BigTable集群往往运行在一个共享的机器池中，池中的机器还会运行其它各种各样的分布式应用程序，BigTable的进程经常要和其它应用的进程共享机器。BigTable依赖集群管理系统在共享机器上调度作业、管理资源、处理机器的故障、以及监视机器的状态。