### 论文阅读报告

高玮 18301094

在读过谷歌的Bigtable，GFS，MapReduce三篇论文后，我学习了一些关于大数据和分布式系统粗浅的、最为基础的知识，我希望凭借这些知识在这篇报告中分享一些我对Bigtable的了解。

Bigtable是一个分布式的结构化数据存储系统，它可以处理PB级的，分布在数千台服务器上的数据。Bigtable有如下特点：适用性广泛、可扩展性、高性能和高可用性。在很多方面上，Bigtable与数据库很相像，但是Bigtable不支持完整的关系数据模型，而是为客户提供了可以动态控制数据的分布与格式的简单数据模型。存储的数据都视作字符串，而Bigtable本身不对这些字符串进行解析。因此，Bigtable的数据模型不受传统数据库的关系模型约束。

Bigtable的数据模型本质是一个稀疏的、分布式的、持久化存储的多维度排序map。一条数据的索引包括行关键字、列关键字和时间戳。Bigtable中的行关键字可以动态分区为数个tablet，当操作只读取行中很少几列数据时效率很高。数个列关键字可以组成叫做“列族”的集合，存放在同一列族下的所有数据通常都属于同一个类型，可以有效地进行相近类型的数据的查询。表中的每一个数据项都可以包含同一份数据的不同版本，并通过时间戳进行索引。

Bigtable还支持一些其他特性：Bigtable支持单行上的事务处理，用户可以对存储在一个行关键字下的数据进行原子性的读-更新-写操作。Bigtable允许用户把数据项用作整数计数器。最后，Bigtable允许用户在服务器的地址空间内执行脚本程序。

Bigtable使用Google的分布式文件系统存储日志文件和数据文件，其文件格式是SSTable格式的，是一个持久化的、排序的、不可更改的Map结构。SSTable使用块索引来定位数据块：在打开SSTable的时候，索引被加载到内存，每次查找都可以通过一次磁盘搜索完成，也可以将整个SSTable放入内存中，就不必访问硬盘。

Bigtable还依赖一个高可用性的、序列化的分布式锁服务组件。Chubby。Bigtable借助Chubby完成以下任务：1. 存储Bigtable数据的自引导指令的位置；2. 查找tablet服务器，以及在tablet服务器失效时进行善后；3. 存储Bigtable的模式信息；4. 存储访问控制列表。

Bigtable的实现：1. Tablet的位置：使用一个三层的、类似B+树的结构存储Tablet的位置信息：第一层是一个存储在Chubby中的文件，它包含了root tablet的位置信息，root tablet包含了一个特殊的METADATA表里所有的tablet的位置信息。METADATA表的每个tablet包含了一个用户tablet的集合。2. Tablet的分配：在任何一个时刻，一个tablet只能分配给一个tablet服务器，当一个tablet还没有被分配、并且刚好有一个tablet服务器有足够的空闲空间装载该tablet时，master服务器会给这个tablet服务器发送一个装载请求，把tablet分配给这个服务器。3. Tablet的服务：tablet的持久化信息保存在GFS上，更新操作提交到REDO日志中。将这些更新操作中最近的那些存放在一个叫做memtable的缓存中，较早的更新存放在一系列SSTable中。当恢复一个tablet时，tablet服务器从METADATA中读取redo point，他们指向含有该tablet数据的已提交的日志记录，并以此重建memtable。4. 空间收缩：当写入操作超过一定数量，memtable的尺寸到达一个门限值时，这个memtable被冻结并转换成一个SSTable，写入GFS，然后一个新的memtable被创建。这是为了减少tablet服务器使用的内存，以及在服务器灾难恢复过程中减少必须从提交日志里读取的数据量。

Bigtable的优化：1. 局部性群组： 客户程序可以将多个列族合并成一个局部性群组，可以提高读取操作的效率；2. 压缩： 客户程序可指定每个SSTable的块使用用户指定的压缩格式来压缩，这样在只读取SSTable的一小部分数据的时候就不必解压整个文件了。3.缓存： tablet服务器使用二级缓存的策略提高读操作的性能。。扫描缓存是第一级缓存，主要缓存 Tablet 服务器通过 SSTable 接口获取的 Key-Value 对；Block 缓存是二级缓存，缓存的是从 GFS 读取的 SSTable 的 Block。

Bigtable的实际应用项目包括：Google Analytics、Google Earth、个性化查询等等。

通过Bigtable设计的Google自己的数据模型使得Google的系统极具灵活性。另外，Bigtable也因为其完全受控制的实现过程，使得当系统出现瓶颈或效率低下的情况时，能够快速地解决这些问题。

在大数据的时代中，各种系统对数据处理性能的需求将不断发展。谷歌这一系列的创新和有化，使得大数据处理变得更加简单可靠，依靠大数据处理的产业进入了一个新的发展阶段。