Bigtable 阅读整理

学号：18301085

姓名：姚聪

Bigtable 是一个分布式的结构化数据存储系统，它被设计用来处理海量数据：通常是分布在数千台普通服务器上的 PB 级的数据。在 Google 上使用最广泛的存储结构就是 Bigtable 存储结构。

Bigtable 的几个特性：适用性广泛、可扩展、高性能、高可用性。

Bigtable 是一个稀疏的、分布式的、持久化存储的多维排序 Map。

Map 的索引是行关键字、列关键字以及时间戳 Map 中的每个 value 都是一个未经解析的 byte 数组。

Bigtable 的行结构：

1. 表中的行关键字可以是任意的字符串（目前支持最大 64KB 的字符串，但是对大多数用户，10-100 个字 节就足够了）。
2. 对同一个行关键字的读或者写操作都是原子的（不管读或者写这一行里多少个不同列）。
3. 表中的每个行可以动态分区。每个分区叫做一个"Tablet"，Tablet 是数据分布和负载均衡调整的最小单位这样的话读取少量数据时效率很高。

Bigtable 的列结构:

1. 列关键字组成的集合叫做列族，它是访问控制的基本单位。
2. 列关键字的命名语法为：列族：限定词。列族的名字必须是可打印的字符串，而限定词的名字可以是任意字符串。
3. 访问控制、磁盘和内存的使用统计都是在列族层面进行的。

Bigtable 的时间戳:

1. 表中的每一项数据都可以包含同一份数据的不同版本，不同版本的数据通过时间戳来索引，Bigtable 时间戳的类型是 64 位整型。。
2. 最新的数据排在最前面。
3. 每个列族有两个设置参数，Bigtable 可以通过这两个参数可以对废弃版本的数据的自动数据收集。

Bigtable 的 API：

1. Bigtable 提供了建立和删除表和列族的 API 函数。
2. 调用 Apply 函数对 Webtable 进行原子操作时，他会为其增加一个描点，同时删除另一个描点。

Bigtable 的构件:

BigTable 内部存储数据的文件是 Google SSTabl 格式的 SSTable 是一个持久化的、排序的、不可更改的 Map 结构，而 Map 是一个 key-value 映射的数据结构，key 和 value 的值都是任意的 Byte 串。

从内部看，SSTable 是一系列的数据块（通常每个块的大小是 64KB，这个大小是可以配置的）。SSTable 使用块索引（通常存储在 SSTable 的最后）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：首先使用二分查找法在内存中的索引里找到数据块的位置，然后再从硬盘读取相 应的数据块。也可以选择把整个 SSTable 都放在内存中，这样就不必访问硬盘了。

对 Bigtable 的介绍:

Bigtable 包括了三个主要的组件：链接到客户程序中的库、一个 Master 服务器和多个 Tablet 服务器。针对系统工作负载的变化情况，BigTable 可以动态的向集群中添加（或者删除）Tablet 服务器。

每个 Tablet 服务器都管理一个 Tablet 的集合（通常每个服务器有大约数十个至上千个 Tablet）。每个 Tablet 服务器负责处理它所加载的 Tablet 的读写操作，以及在 Tablets 过大时，对其进行分割。

一个 BigTable 集群存储了很多表，每个表包含了一个 Tablet 的集合，而每个 Tablet 包含了某个范围内的行的所有相关数据。初始状态下，一个表只有一个 Tablet。随着表中数据的增长，它被自动分割成多个 Tablet，缺省情况下，每个 Tablet 的尺寸大约是 100MB 到 200MB。

Table 的位置:

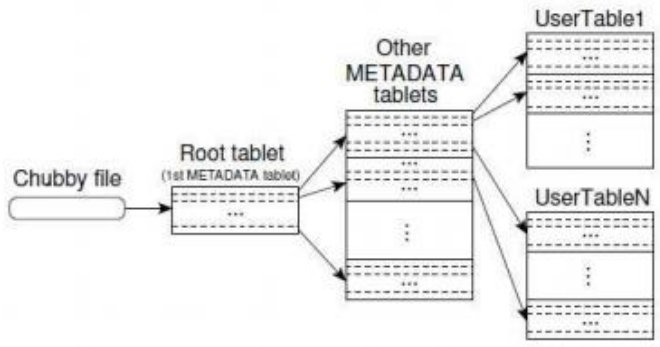

Figure 4: Tablet location hierarchy.

在 METADATA 表里面，每个 Tablet 的位置信息都存放在一个行关键字下面，而这个行关键字是由 Tablet 所在的表的标识符和 Tablet 的最后一行编码而成的。METADATA 的每一行都存储了大约 1KB 的内存数据。在一个大小适中的、容量限制为 128MB 的 METADATA Tablet 中，采用这种三层结构的存储模式，可以标识 2^34 个 Tablet 的地址（如果每个 Tablet 存储 128MB 数据，那么一共可以存储 2^61 字节数据）。

Tablet 分配:

当集群管理系统启动了一个 Master 服务器之后，Master 服务器首先要了解当前 Tablet 的分配状态，之后才能够修改分配状态。Master 服务器在启动的时候执行以下步骤:

1. Master 服务器从 Chubby 获取一个唯一的 Master 锁，用来阻止创建其它的 Master 服务器实例；
2. Master 服务器扫描 Chubby 的服务器文件锁存储目录，获取当前正在运行的服务器列表；
3. Master 服务器和所有的正在运行的 Tablet 表服务器通信，获取每个 Tablet 服务器上 Tablet 的分配信息；
4. Master 服务器扫描 METADATA 表获取所有的 Tablet 的集合。、

Google Analytics：

Google Analytics 使用的两个表：

Row Click 表（大约有 200TB 数据）的每一行存放了一个最终用户的会话。行的名字是一个包含 Web 站点名字以及用户会话创建时间的元组。这种模式保证了对同一个 Web 站点的访问会话是顺序的，会话按时间顺序存储。这个表可以压缩到原来尺寸的 14%。

Summary 表（大约有 20TB 的数据）包含了关于每个 Web 站点的、各种类型的预定义汇总信息。一个周 期性运行的 MapReduce 任务根据 Raw Click 表的数据生成 Summary 表的数据。每个 MapReduce 工作进程都 从 Raw Click 表中提取最新的会话数据。系统的整体吞吐量受限于 GFS 的吞吐量。这个表的能够压缩到原有 尺寸的 29%。