“谷歌三宝”之BigTable

BigTable是Google在2006年发表的一篇论文。论文主要介绍了BigTable的组成，创建，运行过程，和实际应用方向。一下是我个人读过Big Table后的一些想法。

首先，Big Table是什么？我的理解是Big Table是一个谷歌为了方便存储和快速读取海量网页数据而研发的一个分布式存储系统。而后又将之应用到了更多的Google应用之上，例如googleMap上。Big Table构建在GFS，Chubby，SSTable等Google技术之上。BigTable最大的优势在于能够将数据存储量扩展到PB级和分布在上千台服务器上。我对Big Table的理解是一个利用关系型数据库的结构搭建的非关系型数据库。Big Table中沿用了很多关系型数据库中的术语。比如table，row，column等。但在我的理解中，这些术语更像是索引级别。后者按照Big Table中的官方说法，这些是键值（key-value）映射，Big Table的键有三级。在论文中的概念是Big Table是一个稀疏的，分布式的，持久化的，多维（个人理解是三维？）的排序映射。

其次，Big Table是怎样运行的？上面说了BigTable的键有三级，分别是行键（row key）、列键（column key）和时间戳（timestamp），行键和列键都是字节串，时间戳是64位整型；而值是一个字节串。可以用 (row:string, column:string,time:int64)→string 来表示一条键值对记录。这里注意时间戳（timestamp）这个键值是根据不同时间时数据的不同状态。在Big Table中行键可以是任意字节串，通常有10-100字节。行的读写都是原子性的。Bigtable按照行键的字典序存储数据。Bigtable的表会根据行键自动划分为片（tablet），片是负载均衡的单元。最初表都只有一个片，但随着表不断增大，片会自动分裂，片的大小控制在100-200MB。而在读取的时候，行作为一级索引，一个Table中有多个行。列作为二级索引，一个行中有多个列，为了方便管理，列被分为多个列族，也可以将列族看作一层新的索引。时间戳是三级索引，Big Table允许保存数据的多个版本。

最后，Big Table如何实现分布式？BigTable集群主要由三部分组成：一个供客户端使用的库，一个主服务器（master server），和许多片服务器（tablet server）。我自己的理解是一个供客户端使用的库是指客户端发出读写数据请求。而片服务器是GFS的代理，客户端的一些数据操作都通过片服务器代理间接访问GFS。主服务器是用来管理片服务器的负载均衡以及监控片服务器。举例来说，片服务器像是一大群代理商，而主服务器像是工会，调节货流平衡但本身不存储货物。而GFS就是供货商。而主服务器没有货物，即不存储数据，而片服务器只是将订单提交给供货商，也不存储数据。最后，数据在GFS中的存储方式是SSTable。

总的来说，Big Table更像是数据管理系统。其主要优势在于可管理的数据量级达到PB级，对数据的读取，写入速度很快。并且在压缩方面的性能也很好（没有理解）。