

Essential Google

——由 Google File System 所想到的

王仲远 (web 组)

引言：十年前，微软依靠 Windows 操作系统，成为 IT 业界的神话；而十年后的今天，Google 以其强大的互联网搜索能力征服了全世界，成为当今 IT 界最耀眼的公司之一。本文以 Google 的文件系统为切入点，介绍了 Google File System 的工作原理，论述了作者对 Google File System 与 Database 的一些思索和比较，试图探讨出一个全新的属于数据库人的研究方向。

一、李开复与 Google 中国

这是一个造星的时代，当一切成功都被神化以后，它外面所笼罩的美丽光环，使我们常常不能清醒地认识一个公司、一个人，他所起的作用，他所获得的真正成就。Google 就是这样一个公司，李开复就是这样一个人。

去年，李开复来我们学校青年大讲堂作报告，他信誓旦旦地说“微软将会是我度过余生的一个地方”。但一年之后，言犹在耳，可是物是人非：他由微软全球副总裁的身份变为了 Google 中国区总裁。顿时，李开复成为国人关注的焦点，Google 也成为了国人关注的焦点。



Google 中国李开复(左)和周韶宁(右,已离职)

李开复的加盟，使得 Google 成为大学校园里一大批学生向往的地方。Google 对我们来说，也似乎不再是那个遥不可及的在纳斯达克一上市就创造奇迹的地方，而是一个触手可及的承载着太多荣耀和梦想的地方。

Google 中国风暴席卷大学校园！

就如几年前的微软神话一样，Google 中国也是大家心中的神话：宽松的工作环境，随手可取的零食，带着宠物上班……进入了 Google 就意味着站在了 IT 时代潮流的浪尖。但是，Google 真的就是这么的完美无瑕吗？

二、Google 人才观

正当大家对 Google 充满无限遐想的时候，Google 举行了一个大型的实习生招聘活动。所有人都跃跃欲试，期待着 Google 再给大家一个惊喜。但是，当笔试卷子发下来时，所有人都惊讶了，笔试题目是如此的平凡无奇，全部是最基本的算法题和计算机知识。于是，许多人又叫嚣着“Google 神话破灭了！”。

但真的是神话破灭了吗，或者是 Google 并非是一个神化般的公司呢？

我们依然清晰地记着，去年 10 月，周韶宁踌躇满志的加盟 Google，与李开复成为 Google 大中华区联合总裁。但是，周韶宁向 Google 总部提出的一系列本地化策略并没有得到 Google 总部尤其是两位创始人谢尔盖·布林和拉里·佩吉的认同。Google 总部并不认可其一系列“激

进”的措施，这也成为周韶宁离职的重要原因之一。

从一系列事情中，我们可以看出一些端倪：Google 确实是在不断的创新之中，但是它所需要的人才或许并非是有着非常强的创新能力的人，而是那些有着非常扎实基础的计算机人才。

为什么呢？

这就得从 Google File System 说起……

三、Google File System

当我们使用 Google 进行关键字搜索，享受 Google 强大搜索所带来的便捷的时候；当我们赞叹 Google 地图搜索是如此之精确以至于能够清楚地看到我们所居住的房屋的时候；当我们已习惯使用 Google 个性化主页，让 Google 按照我们的想法随心所欲地提供我们想要的资讯的时候……是否有人静下心来思考这样一个问题：在这样强大的搜索背后，究竟是什么技术在支持呢？是什么系统在管理这样一个已超出我们所能想象的巨大的数据资源呢？

“世界上最出名的搜索引擎公司 Google 所使用的竟然不是数据库！”当听到这个消息的时候，对于我这个以为数据库已经是无所不能的即将进入数据库领域的人来说，确实是一个惊人的消息！数据不用数据库存储，而是用已经被我们所淘汰的文件系统来存储，这是一个让人费解的事情，这是一个让人动摇信念的事情。

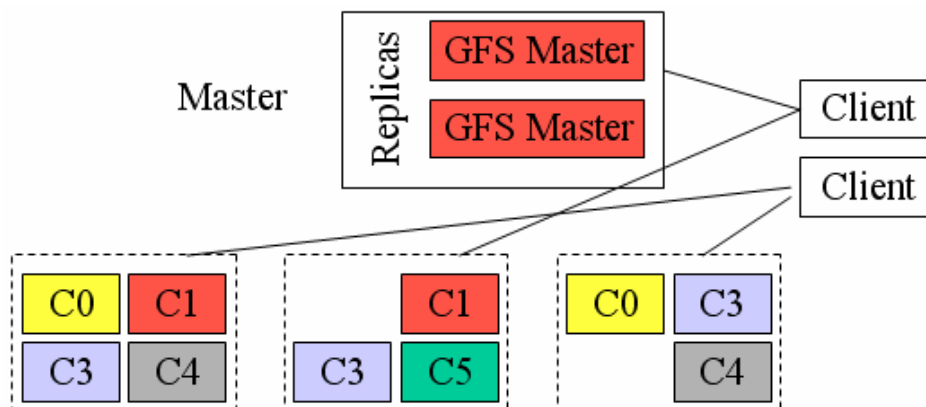
那 Google 为什么不用数据库呢？

Google 说“现有的数据库没法满足我们海量数据存储的需求，即使有，存储及查找代价也会让人无法忍受”。Google 每天所面对的，是成千上万台服务器，是上千 TB 的数据，是每秒数百万的读/写。而且，在这样的情况下，还要实现高效的查询。因此，Google 理直气壮地说：“数据库，No！”

于是 Google 利用极其便宜的 PC 机，来代替昂贵的高性能服务器，并且重新拾起被数据库人即将遗忘的文件系统，成为他们内部的数据管理系统，他们兴奋地说：“Google File System 能够达到我们全部的目标，能够实现高效的存储，具有非常强的容错性！”这些话，不禁让人遥想起 50 年前，当文件系统代替人工管理时，或许也是这般的兴奋。但仅仅十年，数据库就取代了让人激动不已的文件系统，成为数据管理的主要工具。

诚然，在当今网络盛起的时代，面对着 Internet 上数十亿的网页、上百 TB 的卫星图片，传统的关系数据库显得有些吃力甚至都无法管理这样的海量数据。因此，RDBMS 已经无法适应网络时代的需求，需要有新的突破。在 Google 内部，这种突破就是 Google 引以为豪的 Google File System！

那么，Google File System 与以往的文件系统有什么区别吗？



Google文件系统^[1]

Google的文件系统是一个大规模的分布式文件系统，它能够处理大规模的分布式数据。它包括控制服务器（Master）和块服务器（Chunkservers），两者之间的信息传输通过GFS的客户端（Client）实现。控制服务器负责管理元数据，它主要存储文件和块的名空间、文件到块之间的映射关系以及每一个块副本的存储位置；块服务器存储块数据，一个典型的块大小为 64M，它通过懒惰算法（Lazy space allocation）来管理存储在它上面的块。控制服务器通过文件系统客户端向块服务器发送数据请求，而块服务器则会将取得的数据直接返回给文件系统客户端^[2]。在块服务器中，一个块可能有多个备份，这样做的目的是为了保障数据的安全性，当然，也能够实现负载均衡。

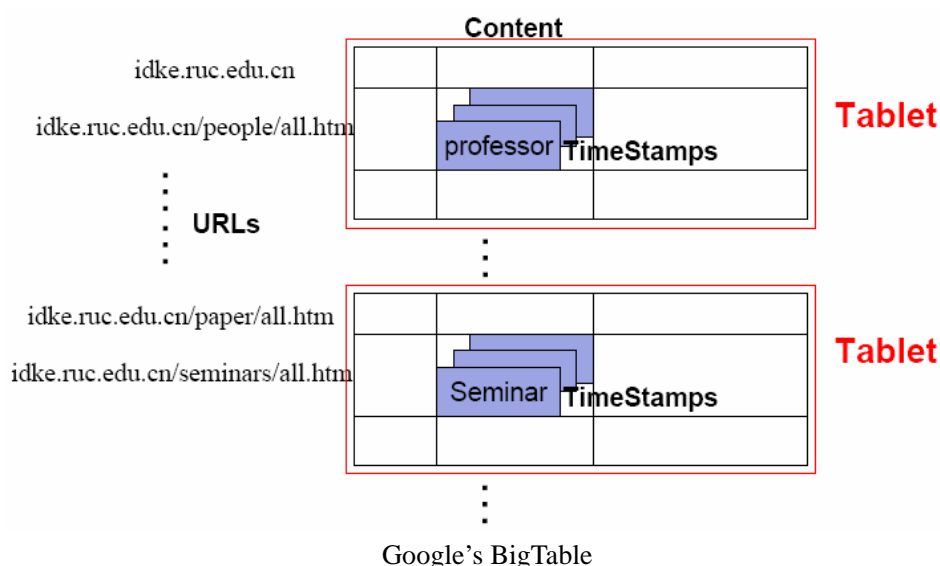
这就是 Google 文件系统的体系结构。然而，光有一个良好的体系结构是不够的，只有具体实现才是硬道理，因此我们需要好好看看建立在 Google File System 上的 BigTable 是如何工作的。

四、BigTable

BigTable，顾名思义，就是一张“大表”，是一张稀疏多维图。Google 于 2004 年初开始研发，到现在它已经运行了两年多，基本上能够满足 Google 的需求：处理海量数据，实现高速存储与查找。

BigTable由行和列组成，每个单元（BigTable Cell）是一个三元组，由行、列、时间戳组成。在一个典型的单元格中，行可以是URLs，列可以是属性/规则，时间戳则是用来标识版本的，它可以在多个备份中，将最新的信息提供给用户^[1]。

BigTable 就如它的名字一样，是一张“大表”，以至于为了便于管理，需要将 Bigtable 按照行拆分成 Tablets。如果说 BigTable 是一块布，Tablets 就好像是从这块布上扯下的布条。即使这样，Tablets 也是不小的。每个 Tablets 大约有 100~200M，而每台机器大约会存储 100 个左右的 Tablets。由于 Google 采用的是廉价的 PC 机，而不是使用高端的服务器，因此采用 Tablets，就将 BitTable 化整为零，分布式地存放在各台 PC 机上。



同时，由于采用了 Tablets，也非常容易地实现了负载均衡和快速恢复。如果某台机器的某个 Tablets 经常被访问，则它可以将原来存储在它上面的其它 Tablets 转移到别的机器上，然后专门负责这个 Tablets，而这个 Tablets 也可以完全载入内存，提高访问速度。如果某台机器坏了，不难想象，这台机器上的 Tablets 只需要由其它 100 台机器，每台机器恢复一个

Tablets，系统就重建起来了，因而机器损坏的影响也会降到最低。这点其实是很重要的，因为 Google 采用的是最普通的机器。“如果你买了一台机器，也许用三年也不会有什么太大的问题；但如果你拥有上千台机器，你要做好每天 down 掉一台的准备”。Google 拥有许许多多的普通 PC，因此，每天都有机器不断损坏，也会又机器不断补充进来，在这种情况下，具有非常好的容错性是很重要的一点。

为了实现对数据的管理和恢复，日志是必不可少的。不难想象，如果每个 Tablets 就有一个日志，那对于这些日志本身的管理就将是一个巨大的工程，所以 Google 选择了同一台机器上的所有 Tablets 共享一个日志的方式。但这种方式虽然减少了日志的个数，却带来另一个问题：一个日志块将很快被写满，于是系统将非常频繁地开始一个新的日志块。看来，鱼和熊掌确实是不可兼得啊！

同时，由于 Tablets 采用的是不可修改的（immutable）的 SSTables 存储方式，因此系统将产生大量的冗余数据，面对这些冗余数据，Google 主要采取两种压缩方式进行数据压缩：BMDiff^[3]和 Zippy。这两种压缩方式，与 Gzip 和 LZW 等压缩方式相比，在压缩率上并没有什么优势，但它们都有一个很大的特点，这就是压缩速率和解压速率都非常快，而这正是 Google 所需要的。能够快速压缩以节约空间，又能快速地解压获得数据，这对 Google 来说，远比其他特性要重要得多。

至此，一个建立在 Google File System 上的 BigTable 系统就已呈现在我们面前，从这个系统中，我们可以看到 Google 的核心理念：**低价的机器，高速的处理，大量的冗余，极强的容错。**

采用了 BigTable 后，Google 完全实现了这些理念。

Google File System 虽然很好，但数据库原有的与文件系统相比的优势难道就荡然无存了吗？

五、Google File System 和 Database

从 Google File System 可以看出，虽然 Google File System 有一些自己的特色，有一个不同的应用背景——网络环境，但它仍具有普通文件系统最重要的特点：冗余与单一应用。

当我们数据库人指责文件系统缺陷时，总是忍不住指责其冗余性所带来的坏处：占用存储空间、造成数据不一致性。但这些问题却恰恰在 Google 中都不存在。Google 采用的是便宜的 PC 机，因此，他可以买很多很多机器来解决存储容量问题。至于不一致性，对于 Google 来说并不是一个大问题，他能够通过时间戳给用户提供的最新的信息，而且，由于 Google 应用的特殊性，他并不存在修改问题，他的数据一旦写入，就是不可修改的。此外，由于 Google 使用的是廉价 PC，面临着机器随时损坏的可能，使用冗余能够实现系统迅速的恢复。

至于文件系统常见原子性问题、完整性约束、同步性问题等，则由于 Google 目前的主要应用——关键字搜索，使得这些问题对 Google 来说已经不成问题了。

但，难道 Google File System 真的是无懈可击的吗？

不。

我觉得 Google 现在之所以使用文件系统使用得心应手与 Google 现在所从事的应用有关——即 Google 虽然提供了众多服务，但其核心仍然是对大量数据的搜索。这也正是问题的关键所在。

我们知道，数据结构化是文件系统与数据库系统最本质的区别^[4]。也就是，数据库中的数据都是结构化的，它能够针对不同的应用；而文件系统则常常只是针对某一特定应用，但应用发生改变时，需要建立另一个系统。

因此，我们不难猜想到，Google 的成功是有其特殊背景的：在 Web 广泛应用后，出现了大量 html 等不规整的数据，而面对这些数据，又有查询、处理的需求，因此面对这一特定应用，传统数据库已经不再适用，需要有新的系统来适应这种环境。而 Google 选择了专门为这种应用开发了一个系统，这就是 Google File System！

可以预见，在将来应用需求越来越多之后，例如面临的是语义网络或其它更多的扩展，Google File System 很可能就会力不从心。而红极一时的 Google File System，就会像 50 年前的文件系统一样，被其它能够应对更多应用的系统所取代。至于取代它的是不是 Web 上的 Database，还有待观察和实践。

六、Google's Database: Google Base

当然，Google 也会意识到他所存在的危机，也会寻求他自身的突破。

2005 年 10 月，号称是 Google 的数据库的 Google Base 在网民的关注下上线了。从首日就有 200 万相关网页的诞生，可以看出 Google 的影响力。面对 Google Base，Google 是这样说的：

Rankings of Top 20 Google Domains Week Ending 5/13/06		
Rank	Name	Market Share
1	Google	79.98%
2	Google Image Search	9.54%
3	Google Mail	5.51%
4	Google News	1.49%
5	Google Maps	0.82%
6	Froogle	0.46%
7	Google Video Search	0.45%
8	Google Groups	0.43%
9	Google Scholar	0.27%
10	Google Book Search	0.25%
11	Google Earth	0.22%
12	Google Desktop Search	0.18%
13	Google Directory	0.10%
14	Google Answers	0.09%
15	Google AdWords	0.07%
16	Google - Local	0.05%
17	Google Finance	0.03%
18	Google Calendar	0.01%
19	Google Talk	0.01%
20	Google Labs	0.01%

“Google Base 是 Google 的数据库，用户可以往里面添加自己的数据，然后 Google 会让这些数据出现在搜索范围之内”^[5]。

也就是说，Google 在试图利用自己的影响力来收集元数据。这其中涉及到一个满有趣的问题，也就是“先有鸡还是先有蛋”的问题。即到底是先有规整的数据，再有处理这些数据的工具呢；还是先有工具，再把原本不规整的数据转变为规整的数据呢？很显然，Google Base 采取的是前者。因此，虽然有人在说 Google Base 的出现，是对 Ebay、Amazon 等电子商务网站的挑战，是对网络社区的挑战，但我却觉得 Google 的野心远不止如此。我相信，Google Base 是在试图建立一个局部的语义网络。这是一个尝试，是一个对下一代网络的尝试，是一个对建立在 Semantic Web 上的搜索的尝试！它代表了 Google 对未来的摸索与试探，代表了 Google 希望在网络上保持霸主地位的野心与决心！

在 Google 今年年中公布的旗下服务访问量中，Google Base 并没有进入前 20 名，看来 Google Base 任重而道远啊

但这种试探的实际现状如何，我们无法得知，或许只有天知、地知、Google 内部的人知；这种试

探的结局会如何，我们也无法得知，恐怕只有天知、地知、举头的三尺神明知。

我们只能拭目以待……

七、结语

但无论如何，Google File System 对于我们数据库人来说，是一个机遇，也是一个挑战。因为我们身处在一个瞬息万变的时代，我们面临着即将到来的又一次科技上的巨大变革，我们是要做未来潮流的引导者，还是要做错过机遇的叹息人？

答案，就在我们每一个人的手中。

参考文献：

- [1] Jeff Dean: BigTable A System for Distributed Structured Storage
- [2] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: The Google File System 2003
- [3] Bentley, McIlroy: Data Compression Using Long Common Strings. DCC'99
- [4] 萨师宣,王珊: 数据库系统概论(第三版). 北京: 高等教育出版社, 2005: 9
- [5] <http://googlebase.blogspot.com/>