**Google三篇大数据论文读后感**

龚云基 19301087

在老师的推荐和要求下，我拜读了google有关大数据的三篇论文，了解了一些和大数据相关的知识。Google的三篇论文介绍了Google MapReduce, Google File System, Google Bigtable三个系统。

1.Google MapReduce

Google MapReduce是一个编程模型，也是一个处理和生成超大数据集的算法模型的相关实现，其架构的程序能够在大量的普通配置的计算机上实现并行化处理，并实现运行在规模可以灵活调整的由普通机器组成的集群上。

传统的方法处理大量数据集不仅花销成本高、出错率高，处理速度也不尽如人意，Google MapReduce的出现解决了这个问题。MapReduce采用先分布后合成的方式首先创建一个map函数处理被拆分了的数据集合，再创建一个reduce函数合并所有的value值。只需关心如何分割输入数据，计算机集群的调度与错误处理还有管理集群中计算机的通信往来，MapReduce就能充分利用分布式系统的丰富资源。这种编程模型利用分布式计算思想，把大量数据拆分给多个低性能电脑，处理后的结果汇总输出。我们只需要输入需要定义任务、去除错误的数据，静待结果，这不仅因为使用低性能的电脑省去了昂贵的成本，多台电脑的处理也增加了容错率，并行处理大大减少了处理时间。

由于MapReduce易于使用，大量不同类型的问题都可以通过其简单的解决，并且可以在数千台计算机组成的大型集群上灵活部署运行，使得它在数据集处理中占据十分重要的地位。MapReduce对于现在的我来说还接触不到，但作为软件专业的学生，以后在从事软件方面的工作中，MapReduce分布式系统的处理方法在我做数据处理时有借鉴作用。

2.Google File System

Google File System是一个面向大规模数据密集型应用的、可伸缩的分布式文件系统。GFS 虽然运行在廉价的普遍硬件设备上，但是它依然了提供灾难冗余的能力，为大量客户机提供了高性能的服务。GFS 作为存储平台已经被广泛的部署在 Google 内部，存储我们的服务产生和处理的数据，同时还用于那些需要大规模数据集的研究和开发工作。目前为止，最大的一个集群利用数千台机器的数千个硬盘，提供了数百TB的存储空间，同时为数百个客户机服务。

GFS的设计实现满足Google迅速增长的数据处理需求，其具有高度扩展性，单个GFS集群文件系统可扩展至几千个节点，数PB的容量；高可靠性，GFS运行在普通x86服务器上，但GFS通过多副本等机制保证可靠性；满足POSIX语义，应用程序无需任何修改即可使用GFS。正因为这些特点，在Google内部，GFS无论是作为研究和开发的存储平台，还是作为生产系统的数据处理平台，都得到了广泛的应用。

Google设计时放松了对GFS一致性模型的要求，减轻了文件系统对应用程序的苛刻要求，大大简化了 GFS 的设计。并引入了原子性的记录追加操作，从而保证多个客户端能够同时进行追加操作，不需要额外的同步操作来保证数据的一致性。这种不苛求一致性的设计与传统的文件系统大相径庭，也让我明白在软件设计开发时要敢于打破现有的思维枷锁。

3.Google Bigtable

Google 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。Bigtable具有适用性广泛、可扩展、高性能和高可用性。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；相反，Bigtable 为客户提供了简单的数据模型，使用这些简单模型客户可以完成适当的操作。

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 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删除Tablet 服务器。Bigtable 使用到的其它的Google的基础构件，在系统出现瓶颈或效率低下的情况时，能够快速的解决这些问题。

Bigtable服务器成员协议和接口遇到问题对我们软件专业的学生也有一些参考意义。Bigtable服务器成员协议在遇到网络问题时会大大降低系统的可用性，也会大大增加 Master 服务器恢复的时间，最终只能放弃协议。协议的制作也许对我们来说很遥远，但接口的实现却是当下的问题。在用户已经习惯于使用支持通用事务的关系型数据库的接口的情况下，由于Bigtable提供的编程接口并不常见，新的使用者有时不太确定使用 Bigtable 接口的最佳方法。这就警醒我们在设计接口时需要考虑易用性。

4.总结

云计算、物联网、社交网络等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长,大数据时代正式到来。大数据作为新一轮工业革命中最为活跃的技术创新要素，正在全面重构全球生产、流通、分配、消费等领域，对全球竞争、国家治理、经济发展、产业转型、社会生活等方面产生全面深刻影响，大数据将在数字经济发展中发挥更加重要的创新作用。在历史的浪潮下，我们作为软件专业的学生，似乎无可避免的需要接触甚至应用大数据，这也要求我们需要对其有基本的了解。

拜读Google有关大数据的这三篇论文，我了解了大规模的管理计算系统，也让我在以后从事软件方面的工作时，会借鉴它们的设计思路，不被现有框架束缚，采用分布式系统处理数据，设计接口时考虑易用性问题。计算机软件专业具有专业性和划时代性，给人类社会带来的变革体现在方方面面，大数据作为21世纪以来计算机火爆的全新领域，也为人类当下的发展和未来的进步提供了充足的动力。Google三篇论文让我对于大数据有了一个基本的了解和认识。但还处在很浅显的程度。专业性的学习需要大量的积累和实践，现如今的我还远远不够，在很多地方还需要积累，见识，提升和突破。大数据的发展给我们软件专业学生提供了很好的发展前景和机遇，同时也带来了挑战和难度。我们的专业学习还很长，需要努力的地方还很多，现在，还只是起步，未来，也同样需要我们去拼搏。