Hadoop in China 2011 参会感言

许利杰

2011 年 12 月 4 日星期日

【摘要】

1、 身体最重要

2、 MapReduce 已经沦为白菜技术

3、 对比工业界，学术界表示压力很大

4、未来

【身体最重要】

2011 年 12 月 2 日和 2011 年 12 月 3 日在北京会议中心举行了特别盛大的“Hadoop in China 2011”大会，据说参会人有 1000 左右。会议讨论主题涉及分布式文件系统、大数据分布式处理、大数据分析、NoSQL、虚拟化和学术研究方面的内容。参会人可以分为学术大牛、学术与技术大牛、技术大牛、学术大牛粉丝、技术大牛粉丝。我是学术大牛粉丝也是技术大牛粉丝。

会议时间选的很无语，早上 7:00 起床下楼发现已经开始飘雪，果断回去换了厚点的裤子后，感觉能抗一点了。半个小时后才打上 Taxi，再半个小时候到了一个鸟不拉屎的地方（五环），旁边就是高速公路。幸好订了会议餐，不然中午没地方吃饭，会议只提供一瓶冰冷矿泉水，第一天还没给我发。一天下来又困又累，报告信息量大，还非常紧凑。最悲催的是下午听完已经 5:30，没的打，走了 20 分钟坐上拥挤的地铁，又碰上周五下班高峰，回去啃了点面包就睡了。加上最近牙还肿痛，感觉“身体是革命的本钱”是真理中的真理。

与《华尔街》里表现的“华尔街拼的不是智力而是体力”一样，IT 从业者（需要大量的时间学习各种技术、编写和调试各种程序，更多时候拼得也是体力。因此身体健康强壮是根本。

【keynote】

第一天的 Keynote 包括 Lucene 和 Hadoop 的创始人 Doug Cutting，主要介绍了 Apache Hadoop 的项目历史、发展和未来。Doug Cutting 高高瘦瘦，中午吃饭时候还见到他用筷子的样子，貌似很熟练。系统和数据库牛校威斯康星麦迪逊分校（WISC）的 Miron Livny 教授对比了 Condor 和 Hadoop 项目，以及 Condor 的未来发展。08 年我才知道 Condor 这个项目，有点类似寻找外星人的网格计算，Condor 也设置工作流调度。这个项目 97 年就已经是 v5.62 版了，值得注意的是 Google 的 MapReduce 论文的 related work 就提到了 Condor。可见在大数据之前的集群计算里，这个来自高校的研究项目有多大影响力。第三个 keynote 来自 Google 的学术与技术大牛 Grzegorz Malewicz，也是号称处理 Google 20%计算的图计算引擎 Pregel 的主要作者，这哥们在很多地方讲过，来中国好几次了吧，Pregel 没有仔细研究过，所以听的晕晕乎乎。最后一个 EMC 的哥们是来推销产品的吧。

第二天的 Keynote 第一个是 eBay 的 Hadoop Leader，介绍了 eBay 的 Hadoop 使用情况、集群和做的改进，自己维护了一个 Hadoop 的代码分支，支持多租户和多 Namenode，并介绍了新的商品搜索引擎 Cassini（构建在 Hadoop/HBase 上）。第二个是 Yahoo Research 巴塞罗那研究院的 Flavio Junqueira，介绍了分布式协作系统 ZooKeeper 的基本概念、用途、历史和发展，还鼓励志愿者加入开发。讲稿中引用了大量的论文，有很大的研究价值。然后中移动的业务支撑经理代表运营商做了关于 Hadoop 在处理用户数据方面的应用，以前的关系型数

据库的 Group By 性能比较低，而且磁盘故障率较高。他们看中的是 HDFS 的强大容错功能。他们也在 Hadoop 应用的各个方面（包括 Namenode 问题、性能问题、调度方面）做了很多优化，有自己的分支。最后的学术大牛俄亥俄州立大学的计算机系主任张晓东教授直接站在哲学高度，尝试建立大数据处理的理论模型，定义了抽象的操作来表示数据的各种处理方式，取得了一定的成果。报告高度很高、学术味很浓，工作意义也比较大，在今年的 ICDE、ICDCS、SoCC 等牛会上都有斩获，也让我更加理解 PhD 中 Ph 的意义。

【分会场情况】

我听的是第一天的《云计算研究》和第二天的《大数据技术与应用》分会场。

第一天的《云计算研究》：

第一个出场的是复旦的学术新星陈海波，他从多核的角度深入剖析 MapReduce 在多核上的优化执行方法，效果比原有的斯坦福的 Phoenix 系统好，正在试图将 Hadoop 和自己的工作结合，搞出一个 CHadoop，核心是将目前的一个 MapReduce task 再转化为更适合多核运行的 MapReduce tasks。我感觉工作很有价值，因为目前的 MapReduce task 确实没有在多核方面优化很多，我也向他要了联系方式，邀请他有时间可以到我们所做一个学术报告。

第二个是拥有 8 年操作系统内核经验的华为技术大牛杨晓伟，从各个方面介绍了华为为了构建桌面云而使用的虚拟化技术。他们的系统虽然构建在开源虚拟机 Xen 上，但在很多方面进行了改进，甚至与硬件厂商合作开发驱动。整个讲稿图非常多，改进后的性能对比图也很多，颇有做学术的味道。可以考虑邀请到实验室做做报告，详细交流一下。

第三个是计算所的副研詹剑锋，锋哥讲话很有趣，之前听过他在研究生院讲的几节课，很有收获。这次他讲大数据的测试床，目的是构建一个可以供科学研究的真实的大数据实验环境。这个要是做好了，很有价值，毕竟一般的科研人员很难深入企业去搞研究。 第四个是风趣幽默的清华教授武永卫，从头到尾，一刻也不停地吐字，声音抑扬顿挫。他带领团队做了一个社区化的 FTP，进而变成一个分布式文件系统，需求很明确，功能也很明确。值得注意的是他的讲话很有逻辑性、而且通俗易懂，看似只是技术演讲，但不经意也透露出深厚的学术背景，很厉害。他的项目目前已经支持智能手机，软件在清华早已普及。

最后一个是来自曙光的王勇，介绍了他们自己开发的一整套大数据硬件、软件解决方案，讲话较慢。由于当时已经很困了，也没抓住重点。分会场的主席是 MSRA 的学术大牛文继荣研究员，搞 IR/ML/DM 方面的应该都清楚，蔡登大牛的早期论文就是他指导的。

第二天听的是《大数据技术与应用》分会场。

在吃饭时碰到了 11 楼的文沛同学，根据师兄第一天的经验，让其帮忙占了座，事实证明十分明智。我到会场几分钟后，就开始有人站着或者坐在地板上听了。

本场的主角是互联网公司，依次介绍了各种后台自己开发或改进的数据存储与处理系统，并展示了各自的大规模集群。

第一个是腾讯的 VIP 朱会灿博士，灿哥在 Google 有十年的工作经验，貌似以前是科苑数学所的硕士，后来去圣巴巴拉读得博士。由于笔记本和投影仪的问题，实际上第 3 或者第 4 个讲的。主要内容是腾讯的云计算和大数据处理平台 Typhoon 系统，目的是整合公司的集群资源，提供一个同时可以处理在线和离线任务的大数据存储、分析、应用的云平台。讲稿主要介绍了用 C++编写的 Typhoon 的架构，整个架构包括分布式文件系统 XFS，独立的调度器、列存储和关系数据库，还有访问控制模块，架构清晰。为了支持遗留的 Hadoop job，他们修改了 JobTracker，使用 JNA 技术将原有的 Hadoop 的 JobTracker 当做一个 job 运行在他们的系统中。另外他们也自己用 C++搞了一个分布式文件系统，在腾讯的师兄也说他们正在完善这个系统。目前在 Typhoon 上已经有应用在跑，包括在线的新闻服务。经常看到水木上有

腾讯的发招聘分布式系统工程师的帖子，原来他们一直在搞这个系统。

第二个是人民搜索的技术大牛何鹏，他们的搜索平台基本上基于 HDFS 和 MapReduce，一个 HDFS 来用来存放网络爬虫抓回来的网页，一个 HDFS 用来存放建好的分布式索引。分布式索引和网页以自己开发的 sorted string table 格式存储，支持随机查找，性能比 Hadoop 的 SequenceFile 好。与其他企业使用 Hadoop Streaming 来支持多语言的 MapReduce job 不同，他们使用 pipeline 的方式来支持多语言，streaming 用的是 Unix/Linux 的标准输入输出、pipeline 用的是 socket。他也讲了一些实际使用 Hadoop 遇到的问题，很有价值，他们也针对问题，对调度器、输入输出类型和格式、资源分配和序列化部分做了优化，目前用于搜索的集群有 500+台，数据量有 300 亿的网页。

第三个是风趣幽默又不失技术水准的人人网的白伯纯和张叶银。白伯纯介绍了人人网的 Hadoop 平台使用情况，他们负责管理和优化 Hadoop 的只有三个人，但是其他部分都在使用这个平台。张叶银主要介绍了好友推荐部分的算法和 MapReduce 实现，用了 DM 里常见的层次聚类和 Mean-shift 聚类方法，也重点抱怨了 MapReduce 在处理迭代型 job 的不足，以及他的一些使用技巧，比较有价值。这种既有 ML/DM 背景也有分布式经验的人是目前各大公司最抢手的了。问了一下旁边的文沛（目前在人人实习）得知他来自自动化所的模式识别实验室。

然后出场的是 FreeWheel 的技术大牛兰向荣，FreeWheel 在欧美的广告市场上很牛，这个算是上商务智能领域了吧。他诟病了 MySQL 在存储和分析数据方面的不足，逐步将数据迁移到了列存储的 InfoBright。迁移后，压缩比大大提高，统计分析也超方便，他们在使用过程中也发现不少 bug，并做了修正，是 BI 厂商从传统数据存储分析方式开始转变的例子，也是数据交换应用的实际例子。

最后出场的是阿里巴巴的技术大牛强琦，大块头有大智慧。报告主题是关于火的朝天的数据流计算和实时计算平台，首先介绍了 Yahoo 的 S4、Twitter 的 Storm、Google 的 Pregel 的优缺点，他们都存在计算中间结果不可见的弱点，而这点恰恰是实时计算所诟病的地方。另外这些系统用户需要判断是否达到设定的条件，容错、checkpoint 也需要自己负责。他们提出并实现一种新的流数据和实时计算平台叫做 IProcess，采用完整的事件驱动、树存储模型、并由微内核+组件模型构造。能够显示出中间结果，并有良好的可扩展性。听起来很靠谱，能够解决的问题很多，值得进一步研究。

【MapReduce 已经沦为白菜技术】

就像苹果的产品已经成为街头货一样，MapReduce 已经成为目前及以后程序员的必备技能。Hadoop 项目也与 LAMP 技术一样成为白菜技术。

从问问题的观众来看，目前会写 MapReduce 程序的人不在少数。目前工业界还是主要关注 Hadoop 的高可用性（HA），包括 Namenode Cluster 和下一代的 HDFS，另外有实力的企业希望拥有 C++版本的 Hadoop 并维护属于自己的分支。

我纳闷的为啥是 04 年 MapReduce 就在 OSDI 上正式发表，到现在才得到广泛应用？原因可能是大家都想坐享其成。08 年 Hadoop 赢得了 Terabyte Sort Benchmark 的冠军，受到了广泛关注，到了 0.20 的稳定版本后，开始能够真正在生产线上发挥作用。学习资料的变多也加快了普及速度。另一股不可忽视力量就是来自企业内部开发人员的推动，不想再受容错、多线程程序、MPI 程序的困扰，对于大数据的处理，只想使用简单粗暴的方式。

另一方面，关系型数据库在大数据面前越来越扛不住，scale-out 能力弱，压缩比小、行存储、统计性能差、难以忍受的多表 Join 速度，各种设计范式都慢慢让程序员从 SQL 转向 NoSQL。小到处理简历的招聘网站，大到处理网页索引的搜索引擎，都开始采用灵活高效的 NoSQL 技术，与 MapReduce 技术结合更紧密的 HBase、Hive、Pig 等也加快了 MapReduce 的

普及速度。

以后的技术发展也不是我这个小罗罗能染指的，我只是觉得对程序员来说简单粗暴、灵活高效的技术将成为主流。

【对比工业界，学术界表示压力很大】

第二天的《云计算研究》没有听到，从第一天的情况来看，研究人员还是专注自己擅长的领域，尽量使用自己掌握的技术来往大数据上靠。除非直接参与工业界的真实项目，不然很难与身经百战的工业界程序员对拼解决方案。

近年来，数据库和分布式系统的几个顶级会议对大数据存储、分析都密切关注，从实验性系统到各种工具、算法应有尽有，做数据库的擅长应用传统数据库中的分析优化技术到大数据上，分布式系统的研究人员也擅长应用分布式算法、共享内存、多核、网络优化、虚拟化、性能建模技术等对大数据系统进行分析研究。相关的机器学习、数据挖掘、信息检索领域也都看准了 large-scale 的方向，从单机的 Weka 到 MapReduce 上的 Mahout，从 Lucene 到 Cloud9，都在慢慢蜕变。

然而真正学术界应该研究哪些问题？这个也不是我这个小罗罗能够染指的，我试试隐隐感觉到学术界应该把小问题做到极致，由点到面，利用自己擅长的研究和技术方法来尝试一些新的实验性的系统或者对问题细粒度、理论上的阐述。工业界毕竟关心的是能够满足自己功能需求的系统，在学术界看来毕竟是粗的。

【以后会怎样】

以后会怎样谁也说不准，反正 IT 就是这样，技术革新每天都在进行，只是以后招聘程序员要求会越来越高，要熟悉这个那个。对研究人员来说，需要关注更多的系统、更多的模型、更多的方法。非常具有实际应用价值的系统，工业界肯定是做的最好的，但新的或者不成熟的东西还需要研究人员的推进。

【另外】

在会场上也碰到了阮老师，她们正在做本体知识库，已经在用 Hadoop，买了很多硬盘来存储数据，也已经使用了 NoSQL 的 MongoDB 技术。由于她们第二天就赶回去了，只是口头交流，下次过来的话在深入交流一下看有什么可以具体合作的地方。