大数据的起源

我今天给大家讲的主题是大数据，大数据是前几年特别火的一个话题，但是近两年被人工智能取代了。今天，我就跟你聊聊大数据的前生今世，大数据是怎么诞生的、怎么发展的，将来会是什么样。

人们通常认为，大数据起源于谷歌的“三驾马车”：谷歌文件系统、MapReduce和 BigTable，这三篇论文分别发表于 2003年、2004年和 2007年。2007年亚马逊也发表了一篇关于 Dynamo系统的论文。这几篇论文奠定了大数据时代的基础。

为什么因为 Google发表了三篇论文，就有了“大数据”的概念呢？原因众说纷纭。我个人的理解是这样的：Google作为互联网泡沫破灭后第一次上市的大型 IT公司，它的市值在上市之后飞速增长。原因在于，Google的广告业务做得非常成功，而广告业务成功的很大一部分原因是它用了大数据技术。当时，很多相关的互联网企业因此都认为大数据是改变自己命运的机会，因此纷纷加入大数据圈子，入局的有微软、阿里巴巴、雅虎、Facebook、LinkedIn、Twitter等公司。

抱团取暖的 Hadoop 生态圈

2008年的时候，大部分公司围绕在一个叫作 Hadoop的项目周围，这个项目最初开始于 2006年 1月，是“大数据之父”道格 · 卡丁（Doug Cutting）把他对谷歌文件系统、MapReduce的实现，从爬虫项目 Nutch里独立出来形成的。Hadoop的主要贡献者是雅虎，Facebook、LinkedIn、Twitter等公司也都贡献了一些影响深远的项目。

Hadoop系统，在 2008年 6月就比较稳定了。当时 Hadoop并不是很有名，我之所以知道这个事情，是因为我那个时候正在 IBM研究院实习，而我实习的组是 IBM内部两个最早开始使用 Hadoop做研发的组之一。当时我参加了一次 Hadoop在硅谷的活动，这个活动上雅虎的 Hadoop开发人员非常兴奋，因为他们第一次拿到了全球排序算法的冠军。那个会场设置在雅虎，Facebook、LinkedIn和 Twitter的人也都出席了。这些公司除了对 Hadoop的贡献以外，雅虎贡献了 Pig、Facebook贡献了 Hive、LinkedIn贡献了 Kafka、Twitter贡献了 Storm。

今天去看雅虎，你可能觉得雅虎这个公司已经算是倒闭了，一无是处。但是在 2008年到 2009年的时候，雅虎看起来是非常成功的。雅虎曾经是“互联网第一股”，引领了互联网门户时代的发展。所以那个时候，雅虎自然成了其他互联网初创公司的中心，由它主导大家一起来“造一个轮子”，让这些互联网公司有一套可以和 Google的“三驾马车”相媲美的系统。

为什么这些公司没有信心单干，自己独立做一套呢？我想主要原因是，这些公司都意识到自己的技术实力和 Google相比差距巨大。

自己造轮子的微软和阿里巴巴

Hadoop以外的系统主要有两个：一个是微软自己研发的 Cosmos，中文叫作“宇宙”；另外一个系统是阿里巴巴的 ODPS。

微软自己研发 Cosmos的原因很简单：一方面，微软当时和开源社区关系不好，无法和雅虎一起合作；另一方面，微软当时虽然是“帝国没落”，但仍感觉自己很厉害，有信心凭借一己之力“造轮子”。

我当时正在 Cosmos组，整个 Cosmos的技术架构有很多方面像谷歌。当时我们每个新人入门的时候，组内资深老人总是把谷歌的几篇论文给新人，帮助他们入门。可见微软虽然觉得自己厉害，但是没能改变参考谷歌系统的本质。

阿里巴巴在决定做大数据的时候，实际上成立了两个团队：一个是在 Hadoop基础上做开源系统，另一个是自研一套叫作 ODPS的系统。两边都投入很多人，但是最后 ODPS得到了大力支持，Hadoop团队衰弱了。

一场大论战

2008年的大数据圈里发生了一次非常引人瞩目的事件。这次事件的一方是数据库领域的元老级人物迈克尔 · 斯通布雷克（Michael Stonebraker）和大卫 · 德威特（David Dewitt），另外一方是主导了谷歌技术发展的杰夫 · 迪恩（Jeff Dean）。这两群人就谷歌“三驾马车”之一的 MapReduce是创新还是倒退，争得不可开交。

传统数据库一方以一篇“MapReduce：一个巨大的倒退”（MapReduce：a Step Backward）博文掀起了这场论战。他们认为 MapReduce是数据库领域的人早就淘汰了的，不值一提。但是谷歌那批人觉得 MapReduce是一个伟大的发明。

此次事件影响深远，但是当时双方的观点都太绝对了。在我看来，两方面都有道理，但是两方面谁都没有看到对方正确的地方。业界有些人则看明白了，并经过论证和思考开发了新系统，比如 Spark。

Hadoop的生意经

当一个开源的产品做得越来越好以后，就有人希望从开源产品里面赚钱。在 Hadoop开源中赚钱的标志是 Cloudera公司的诞生。

2009年，几个人有感于现在的 Hadoop难用，觉得如果能够让 Hadoop这套系统可以变得更容易安装、更好用，可以卖给其他企业赚钱。所以，这几个人成立了一家叫作 Cloudera的公司来做这个生意，并推出了一个 Hadoop版本。

有关 Cloudera的详细情况，在 Cloudera上市之前，我写了一篇很详细的分析文章，你可以在公众号“飞总聊 IT” 里面找到这篇文章。

之后有了另外一家公司，叫 MapR。它的印度 CTO原来在谷歌文件系统组工作，他觉得 Hadoop的文件系统太差，就自己在印度召集了一群哥们儿，重新写了一个和 Hadoop兼容的文件系统，做了一个叫作 MapR的 Hadoop版本。

之后雅虎里面做 Hadoop的团队从雅虎分离出来，成立了 Hortonworks。

最终整个 Hadoop市场变成了三家分销商的天下。

交互式查询的风波

大数据发展史中，谷歌经常扮演“搅局”的角色，这次谷歌发表了一篇论文，介绍了一个叫作 Dremel的系统，也就是现在的 BigQuery。简单来说，BigQuery可以帮助用户很快查到所要的结果。

BigQuery推出后大家都觉得 Hive太慢了，三大 Hadoop分销商都对交互式查询提出了自己的做法。Cloudera做了一个叫作 Impala的项目，这是一个相对蛮成功的产品，到今天还有比较大的影响力；MapR提出了一个叫作 Drill的项目；Hortonworks的做法是干脆努力提高 Hadoop上的查询语言 Hive的查询效率。

说到三家公司最后的成长和发展，MapR比较微妙，创始人兼 CTO离开 MapR，成了 Uber的高管；Cloudera和 Hortonworks都已经上市了。

后 MapReduce时代

Hadoop发展过程中一个标志性的公司是 Facebook。Facebook开始是开源系统里面非常重要的一个角色，它在 Hive上投入了大量的资源。但是，Facebook在 2012年转做了 Presto。Presto查询速度非常快，也没有用到 MapReduce，它的势头很猛，很快就战胜了 Hive。

另外一个是伯克利大学 AMP实验室开发的 Spark，Spark的兴起算是 Hadoop生态圈一个比较关键的转折点。Spark的战略非常牛，最开始他们没有跟 Hive一争高下，而是选择了机器学习。Spark兴起以后，慢慢地花了很多力气做 SQL，最近 Spark又在努力做流计算。

除了 Spark以外，还有另外一个产品叫 Flink。在国内，阿里巴巴用 Flink用得比较多。Flink是德国柏林理工大学团队在 2014年重新开始的一个项目。他们 2008年做了一个项目，但是失败了。这几年的流计算领域，主要就是 Spark和 Flink在竞争。

大数据的未来

大数据时代的有些未来是可以预见的。

第一是数据库能力的提升。谷歌的 Spanner和亚马逊的 Redshift都体现了这种变化：数据库的能力越来越强，它可以解决很多大数据的问题。

第二点是大数据平台的发展非常强调实时性。流计算现在变得非常重要，我个人很看好流计算的发展。

第三点是 AI给大数据准备了什么。谷歌的一篇论文中说到，我们可以通过 AI技术给大数据建立更好的索引。我认为，AI促进大数据发展和大数据融合将来是个很重要的方向。

总 结

听完我的分享，你可以发现，在商业飞速发展的今天，作为一个技术人员“如何顺势而为”是一个非常值得深思的问题，希望我的《技术与商业案例解读》专栏可以给你启发。