1 对大数据领域的理解和观点

大数据不是本人专业方向，但它和搜索引擎关系很大，此处尽我所能结合搜索引擎来阐述一下我对大数据历史的观点。

作为大数据原点那个抛砖引玉的公司，Google 三驾马车 [1−3] 中的 “前两驾” GFS, MapReduce 无不在论文中透露解决搜索瓶颈的需求。首先，GFS 只能 append 不优化或支持 overwrite，迎合了索引构建的特点：新文档 ID 只进行 append，旧文档的删除也是靠创见一个 append-only 的 posting list。借此优化发挥了 HDD 硬盘的优势。因为搜索和构建索引都需要大量的 merging 操作，减少顺序读写时间精准地瞄准了这个痛点。即使 2001 年 Google 不再使用 on-disk 索引处理搜索请求 [4]，优化对于搜索时的帮助可能不大，但是 2000 年 Google 索引文档量级已达到 billion[5]，对顺序读写的一点优化至少能节省相当的索引时间。另一方面，MapReduce 也可以看作针对分布式搜索必要任务（例如文档词频统计、分发 crawler 任务）的编程范式，他的出发点还是解决搜索引擎的问题，效仿者 Hadoop 的创始人 Doug Cutting 一路的创作从 Apache Lucene 到 Nutch 也验证了这条清晰的历史发展。作为有趣的副作用，搜索引擎和大数据的工业界领域被 Java 统治，即使 C/C++ 系可能有更大的性能优势。

在 MapReduce 范式不适用的问题上（比如 iterative algorithm），诞生了 streaming 范式的系统（Spark, JStorm, Flink 等）作为补充。从业务角度，本质上都是解决如何更方便管理分布式系统，更快更专注地处理分布式问题/业务。在 DT 时代，数据处理能力不但能服务于技术（云计算、机器学习）、还能更好的挖掘商业消费趋势、金融趋势、甚至可以服务于军事（Plantir）。

2 对 Cloudera 的观点

2009 年 Hadoop 生态圈的精神领袖 Doug Cutting 加入了 Cloudera，Cloudera 就曾被认为开源大数据领域的核心公司。虽然论贡献雅虎是 Hadoop 生态圈前期主要的参与者 [6]，但两年后裁掉了它的 Hadoop 团队，标志着逐渐落末的雅虎失去了对 Hadoop 生态圈的控制。于是 Cloudera 在 2012 年以后盛极一时，与 Oracle, Dell, Intel 建立合作和投资关系。

技术上，Cloudera 依赖对 Hadoop 生态技术积累提供 Apache 开源基金下的 Hadoop、Spark、Impala、Kudu、HBase 等工具集合的发行和技术支持，以 Cloudera Manager 为代表的产品，主要产生的价值是给企业节约开发、部署成本。本质上，Cloudera 提供的是技术的服务。它的业务也是技术导向的。

然而，我认为 Cloudera 的服务本质决定了它天花板很低（市值也可以印证，17 年上市以来股价基本就是平缓、很少超过 20 的三角函数 [7]）。首先从市场角度，所基于的技术核心壁垒并不高（这个判断基于两点，1、大多数核心产品开源; 2、单 streaming process 就仅仅在开源界有多套解决方案，例如 Flink、Storm、samza 和 Spark ）。所以 Cloudera 面向的市场面比较窄：给技术不高不低的公司提供大数据技术服务。相对于大公司（比如

亚马逊和阿里）不但有能力掌控大数据技术的核心，而且本身产生大量高价值数据（比如分析买家需求），也能使能对外提供的云计算服务（带来更为广阔的技术服务市场）。相比之下，Cloudera 没有生产数据，依靠成熟的开源生态输出技术服务，极大限制了它市场价值。

即使在技术方面，Hadoop 开源社区曾因为和 Hortonworks（从被裁的雅虎 Hadoop 团队成立的）打嘴仗，使得 Hadoop 有大概近两年时间没有大的 feature release，给了类似 Databricks 这样基于新平台公司可乘之机，削弱了 Cloudera 对技术核心的覆盖。

宏观上，我认为大数据是互联网发展的必然趋势，其配套的工具的成熟也是历史的大方向，将来能发挥商业价值的大数据公司，必然要结合云计算或者本身的商业、金融、行政平台，才能提供高的价值（结合云结算能让大数据技术输出更易于客户企业使用、拥有大平台也才轮得到大数据）。单纯以技术作为商业优势的企业，很容易被那些体量价值大的公司透过对技术和开源的持续投入而被淹没在历史里。

钟威
2018 年 4 月 2 号

3 Reference

[1] The Google File System; http://labs.google.com/papers/gfs-sosp2003.pdf
[2] MapReduce: Simplifed Data Processing on Large Clusters; http://labs.google.com/papers/ma
osdi04.pdf
[3] Bigtable: A Distributed Storage System for Structured Data
[4] https://en.wikipedia.org/wiki/Google_Data_Centers
[5] http://infolab.stanford.edu/~backrub/google.html
[6] https://zhuanlan.zhihu.com/p/25206116
[7] 来自 Robinhood 数据