被仰望和遗忘过的Cloudera，能否王者归来?

钛媒体APP
发布时间: 2020-04-09 15:56

> 文 | 郭华

先想象这样一个场景：

> 你家有个天台，上面什么都没有。你一筐筐的把土背上去，铺了一小块地，然后又挑水施肥，种了几垄黄瓜、几棵西红柿和一小簇青菜。你照顾的很用心，他们长势也很不错，红红绿绿，晨曦中沾满露水，散发出泥土的气息。然后等到秋天，你兴冲冲跑上天台，结果门一开，噗的惊飞一片麻雀，噗噗啦啦之后只见黄瓜断了，西红柿也满是洞，他们绿的红的汁，滴在被爪子踩的不成样子的青菜里，一片狼藉。而且这还是个温暖又明媚的午后。于是你闻着别处的稻香，在金黄的秋风里，感到一阵凄凉。

这时你再看Cloudera的故事，大概才能感同身受。

01 被仰望的

Cloudera由来自Facebook、谷歌和雅虎的前工程师杰夫·哈默巴切(Jeff Hammerbacher)、克里斯托弗·比塞格利亚(Christophe Bisciglia)、埃姆·阿瓦达拉(Amr Awadallah)以及现任CEO、甲骨文前高管迈克·奥尔森(Mike Olson)在2008年创建。

我们先从Hadoop开始说，而说Hadoop就离不开Doug Cutting。

Doug Cutting现在是Apache基金会的主席，标准的大神。在我心目中，他和谷歌AI负责人Jeff Dean、Linux之父Linus并称三大天王，高山仰止，令我自惭形秽，最终放弃编程。

2004年，Doug Cutting正在捣鼓Nutch，Nutch是一个开源搜索引擎，关于它和Lucene的故事又是一个传奇，可以再开一篇单独讲。总之，Doug Cutting碰到了一些大规模索引和分布式计算的问题，恰好这时看到谷歌发表的两篇论文里有相似场景及解决方案。仔细研究之后，他觉得非常有道理，于是很快就把论文里的技术实现了，包括一个分布式计算框架MapReduce和一个分布式存储系统HDFS，然后放进了自己的Nutch里。

2006年，Doug Cutting预感到这种数据处理技术有着更大的潜力，便把MapReduce和HDFS从Nutch中独立出来，合成一个后开源了，取名为Hadoop。那会他儿子刚好两岁，不大会讲话，总管自己的玩具象叫Hadopp，Doug Cutting灵机一动，便把这个名字拿了过来。

同年，为了进一步发展Hadoop，Doug Cutting决定加入互联网公司里最大的雅虎。

大概他也没想到，大数据时代的序幕就这样被拉开了。

在雅虎，Hadoop的集群规模很快过千，Doug Cutting也认识了雅虎副总裁Amr Awadallah。

为了对抗日渐崛起的谷歌，Amr当时正在研究如何让雅虎搜索更智能，也碰到很多性能、成本与弹性的问题。在Doug Cutting的建议下，他开始尝试Hadoop。在随后的两年里，Amr基于Hadoop改造了之前的数据处理系统，结果可以说是惊人的好——完成相同的工作，新系统成本不过是之前的十分一，更重要的，他们还能做之前根本无法想象的事情，比如全量分析以PB记的数据。

这让Amr大为惊喜，他进一步想，这些问题应该不止雅虎会碰到，Hadoop这种革命性的数据处理能力里蕴含着巨大商机。于是他萌生了一种想法，创业。这并不是Amr的第一次创业，实际上他在很早之前就创立了一家叫做VivaSmart的公司，然后2000年公司被雅虎收购，他才随之加入雅虎。

Amr召集起几个志同道合的人，包括两位分别来自谷歌和Facebook的工程师和一位来自Oracle的经理人Mike Olson，很快在硅谷成立了一家公司，自己担任CTO。

公司的名字叫Cloudera，CEO是Mike Olson。

Mike Olson何许人也？其实他和Amr一样，也是自己的公司被收购后加入大公司的。在Oracle之前，他曾是Sleepycat的CEO，而在Sleepycat之前，他又参与过IIIustra的创业。这两家都是在开源软件上创业的商业公司，Sleepycat基于Berkeley DB，IIIustra基于PostgreSQL。这几乎和他们要做的Cloudera一模一样。

那时是2008年，Hadoop正以燎原之势蔓延，不仅席卷了硅谷，也燃烧到了大洋彼岸的淘宝和百度等，于是很快就成了Apache的顶级项目。

一年后，Doug Cutting加入Cloudera，职位是首席架构师，而作为Hadoop的创始人，他也很快被选为Apache基金会主席。

天时、地利、人和，独角兽的羽翼鼓涨满满，只等风来。

2009年，Cloudera拿到了500万美金的第一笔投资，2011年，拿到了4000万美金的第二笔，三年之后，它又拿到了高达9亿美金的第三笔。

如果那时你搜索Hadoop is，输入栏会自动补齐 future。

我记得那时我大学快毕业。有次看到班里一个同学正坐在电脑前贱兮兮的笑，我问他在干啥，他说在改简历，我更加好奇便凑了过去，只见他正把一段网上复制的内容贴到“技能“那一栏里去。我不解，他咔咔使劲按了几下Control+S后嘻嘻道，现在流行云计算，只要描述里出现Hadoop，肯定能过简历关。

而且经过实践，这是真的。Hadoop受欢迎的夸张程度，可见一斑。

另外，估计那会大家都分不清什么是云计算哪个是大数据，不仅我分不清，你看阿里云早期的飞天系统，其实也是一个大数据处理工具，而且可能Cloudera也分不清，不然怎么他一个搞Hadoop的公司，起名叫Cloud - era呢?

这边Hadoop野蛮生长，那边Cloudera合纵连横，先是和Oracle达成战略合作，接着戴尔、Intel、埃森哲、德勤、MasterCard、SAP、TeraData、微软等也纷纷入局。

所有人都关注着它，生怕错过什么。

2013年的时候，Mike Olson信心十足的写下了《The Cloudera Model》一文，表示Cloudera已经找到了Hadoop上成功的商业模式。

那时的Cloudera，可以说是大数据领域最耀眼的星。2015年华尔街日报做了一个独角兽排名，它是唯一上榜的大数据公司，排名21，比大众点评还高。

2017年，Cloudera成功上市。

02 被遗忘的

随着时间发展，Hadoop的概念逐渐泛化。一开始只有HDFS和Mapreduce，然后是一个以HDFS和YARN为基础的平台，再之后是一个包含Spark、Hive、Hbase等几十个项目和子项目的生态，最后，甚至又带上了以Hadoop为基础的商业公司，如Cloudera、Hortonworks、MapR等。

这种泛化有个坏处，那就是一旦出现负面新闻，大家往往分不清该怪谁。

比如作为一种技术，Hadoop肯定有其时效性。就像Mapreduce，虽然计算能力强大，但一切都是先Map再Reduce的抽象程度实在太粗鲁，以至于理念上很快就被Spark、Flink等这种更先进的技术打败了。与之相似的，还有HDFS和YARN，从技术上说，前者不如云存储方便，后者不如K8S灵活，都有被取代的风险。

于是有人便开始宣称Hadoop已死，然后又说，因为Hadoop已死，Cloudera也不行了。典型的用狭义概念做总结，用广义概念做推导，就像偶尔看到汽车超过了一辆绿皮车，就立马得出铁路运输已死，公路运输是未来一样。

然而这种暴力论断却非常有市场。

尤其是2019年，这年Hadoop三个主要的独立供应商过的都不大好。MapR裁员，苦寻几个月金主后卖给了HPE。Cloudera合并了Hortonworks，合并后Q1财报略不及预期，然后股价暴跌，CEO离职。往日的Hadoop三巨头，似乎已是英雄末路。

这时不少人纷纷站了出来，用Mapreduce的问题论断Cloudera，表示它将不出所料的要玩完。

如果这时你搜索Hadoop is，输入栏会自动补齐 dead。

于是一个魔幻的现象出现了，一边是Hadoop已死，独立供应商要完，一边各大云厂商却在拿Hadoop疯狂赚钱。据分析师测算，2018年单AWS的EMR就产生了2.5亿美金的营收，而该产品介绍就是“Hosted Hadoop framework“。而这并不是孤例，除AWS的EMR外，谷歌云有Dataproc，Azure有HDInsight，阿里云有E-MapReduce，云计算四巨头，全都把托管Hadoop放到了自己大数据产品的首页，这显然不是已死的技术该享受的待遇。

Cloudera的心情，大概就像开头说的种一年菜最后都被鸟收割了一样。

它当然解释过，但没什么效果。另外其实它很早就开始淡化自己是Hadoop供应商的概念了，比如和O'Reilly合办的Strata大会，以前叫Strata+Hadoop，2017年之后便把Hadoop字眼拿掉，改叫了Strata Data Conference。

只不过这一切几乎没有人听，就像自己已经被大家遗忘掉一样。

当然，Cloudera无论如何肯定面临着一些问题，而且远比技术问题复杂。

在合并之前，Cloudera和Hortonworks有各自不同的产品线，有各自不同的思路，Cloudera主打开源引擎加商业周边，Hortonworks主打全开源。所以合并后的第一个问题便是产品线怎么整合，原有客户怎么迁移。它Q1的财报里提到不少客户推迟了续费，主要就是这个原因，大家都在等。

但这个问题最多只算近忧，Cloudera真正的远虑，则是前面提到的公共云厂商。

云在吞噬一切，包括大数据，它们不止有托管的Hadoop，还有自研的替代产品。所以逻辑会变成这样，它们会用托管Hadoop鲸吞开源市场，然后用自研替代品蚕食Hadoop。比如AWS里的Redshift，从场景上基本可以看成是Hadoop+Hive的替代方案，但2018年营收约4亿美金，远超Cloudera。而且这个逻辑对所有开源厂商都成立，几乎成了开源软件的公地悲剧。2018年，在怒斥云厂商为吸血鬼而收效甚微之后，Redis和Mongo两家公司直接修改了开源协议，不再允许云厂商提供托管服务。

不过Hadoop用的是Apache协议，修改起来比较困难。但针对近忧远虑，Cloudera也给出了自己的答案。

那就是CDP。

03 王者归来

CDP，全称Cloudera Data Platform，是Cloudera和Hortonworks合并后的统一产品线，做了诸多技术升级，更重要的是其部署形态发生了根本性改变——CDP是基于云的，而且是混合云。

有人戏称，Cloudera终于迎来了Cloud Era。

并且，Cloudera还宣布2022年后停止对原来两条老产品线的支持，全统一到CDP上。很显然，这种大刀阔斧的革新，表明Cloudera孤注一掷想借CDP王者归来。它说CDP是一种新的数据方法，是世界上第一个企业数据云产品，对应的市场规模高达260亿美金，并将在三年后翻倍。

能行吗?

我们一层一层来看。

技术上，首先被大家诟病已久的Mapreduce在Cloudera的产品里早有了很多替代品，比如Spark和Flink；其次，CDP整合了云存储，这意味着HDFS的争议也能得到解决；最后，CDP在调度上对接了K8S，先不说可能性很小，哪怕最后K8S完全替换了YARN，CDP也能做到几乎不受影响。

所以技术层面，按照Cloudera CPO的说法，这叫“Hadoop已死，Hadoop万岁“。Hadoop里几十个项目，是一个生态，甚至一种哲学，早就超越了十几年前Mapreduce的范畴，正波浪式的向前蓬勃发展。

商业上，Cloudera一方面在2019年7月宣布所有代码全部开源，向红帽的商业模式靠拢。另一方面又在部署形态上做了大幅调整，改成了混合云。

这又可以分两层来说。

先说第一层，为什么要学习红帽。

实际上业界一直有一种说法，那就是开源软件的商业公司里，真正称得上成功的只有一家，那就是红帽。红帽自Linux起家，营收一度高达30多亿美金，而且长期盈利，直到2019年以340亿美金的天价卖给了IBM。

所以Cloudera学习红帽的商业模式很容易理解，而且红帽商业模式里的三个要点，Cloudera也基本都能满足。

- 深度参与开源社区——Cloudera有一百多位Apache committer，在大数据方面的技术实力无可争议。
- 代码全部开源，社区版激进，企业版稳定——全部开源是Cloudera发表的《我们对开源的承诺》一文的主要内容。
- 靠企业版订阅产生营收，并提供咨询、支持等服务——CDP，也包括CDH和HDP。

红帽基于Linux，Linux和Hadoop都是基础软件。虽然层次不同，但按照红帽CEO的说法，他们的商业模式比较适合于“复杂、流行、社区驱动的基础软件上“。而大家对Hadoop最大的指责就是太复杂，所以，没准这种复杂性恰好有其商业价值，毕竟太简单的也没必要找个商业公司来兜底。

然而，学习红帽是否就够了呢?

红帽成立于1993年，那会可没有云计算的威胁。

这就说到了商业上的第二层，也就是CDP所指的混合云。

云在吞噬一切，只不过这种吞噬是从互联网创业的增量市场开始的，但随着网络应用的逐渐饱和，这部分市场越来越小，于是云巨头不得不把眼光放到传统IT的存量市场中去。但这些企业跟要么增长要么死亡的互联网创业公司不同，他们更关心稳定性、更关心数据安全，他们有自己的机房，拒绝被云厂商锁定。

于是混合云应运而生。

所谓混合云，就是搭建在自建机房和不同云厂商资源之上的云平台，这种情况下，不管是自建机房还是云厂商，提供的仅仅是底层计算资源，可以根据使用者的意愿随便切换，就像水和电一样，即插即用。

嗯，至少理论上是这样。

但实际上混合云市场还处在混战之中，参战者至少有三类：一类是公共云厂商，如AWS的Outposts、Azure的Azure Stack和谷歌的Anthos等，他们的混合云往往为了线下资源，终点在云，不在混合；另一类是独立混合云供应商，如红帽的Openshift，他们试图在各大公共云厂商基础上搭建一个通用混合云平台，终点在混合，不在云；还有一类，就是各开源应用厂商自己搭建的混合云，如Confluent的Confluent Cloud、Cloudera的CDP、Elastic的Elasticsearch Service等，他们的目标也是混合，但更纯粹，就是要反过来屏蔽云厂商提供的特定开源托管产品。

现在讨论混合云的最终格局显得有点太不自量力，我们不妨把问题稍微缩小一点，那就是在这种混合云的状态下，Cloudera的混合“企业数据云“，是否能做成？

先说市场，关于市场分析师有诸多测算，少的也有几百亿，但我们不妨说的简单点——只要大家还要做大数据，Hadoop就一直有市场。实际上我至今都没发现有谁在做大数据而不用Hadoop的。说Hadoop有问题，大家都承认，但说他要完，这可有点早。可见范围之内，它还看不到有威胁的整体竞争对手。

既然市场成立，那就看Cloudera的竞争对手情况了。

在MapR被收购并且Cloudera合并了Hortonworks之后，应该没必要再讨论独立供应商里谁最强的问题了，因为答案是显然的。

至于像Openshift这样的独立混合云，很像云计算版的聚合平台，可能长期存在，但我不认为能做大，主要原因还是他们无法通过网络效应增强自己的竞争力，因为可选的供应商太少。而且，他们瞄准的是PaaS这层，即通过K8S屏蔽IaaS，所谋甚大，志不在Hadoop。

所以Cloudera的竞争对手只有公共云厂商。

不过这并不好对比，从技术先进性上来说，Cloudera显然有优势，如果客户对产品有更高要求，那选CDP的可能性要大一些，但从产品模式和市场策略来说，云厂商以IaaS高频打低频，在搞定了客户的基础资源之后，再给客户推一个大数据平台，也算顺理成章。

于是似乎只能草草得出一个要看情况的结论。

但我们不妨跳出来再看一下。

毛爷爷曾写过一篇文章《中国的红色政权为什么能够存在？》，里边这样说道：“我们只须知道中国白色政权的分裂和战争是继续不断的，则红色政权的发生、存在并且日益发展，便是无疑的了。“ 那时中国处于国民党的统治之下，并且共产党刚刚遭受了重大打击，很多人开始产生悲观情绪。但毛爷爷指出，国民党的统治貌合神离，蒋桂冯阎四大军阀的背后是不同的帝国主义诉求，他们之间的斗争是不可调和的。所以在他们的斗争之间，红色政权便能产生和发展。

我发现这段精彩论述，完全可以套用在CDP要做的混合云身上。

第一，公共云厂商之间互相斗争，只要世界上不止有一个云厂商，那被锁定的担忧就一直存在，混合云的需求也就一直存在。又因为这种担忧是针对云厂商的，所以云厂商提供的混合云天生说服力不足。

第二，CDP的数据混合云是一种“地方经济“，可以脱离统一的“大资本主义经济“而独立存在。即限定到大数据领域，客户可以只用CDP而不必依赖特定云厂商的某些特定功能。

第三，CDP的混合云将会先诞生于经过“民主革命“训练的地方。也就是说，那些曾经习惯使用Cloudera产品的用户，将会率先迁移到CDP的混合云上来，而Cloudera长期以来的客户都集中在财富2000里，大多是传统客户，正是云计算10%渗透率以外的地方。是的，云计算高歌猛进了这么多年，渗透率依然只有10%。

第四，CDP这种混合云的诞生和长期发展，需要一支相当力量的“正式武装“。在开源混合云的场景下，这支“正式武装“可以理解为商业公司，言下之意是纯社区建立的混合云无法长期存在。所幸，目前开源应用混合云都是由商业公司建立的，不管是Confluent、Elastic，还是Cloudera，均不例外。而且，Cloudera这支武装相当有力量，100多名Apache Committer，3000多名员工，看似和云巨头动辄几万人的规模差距很大，但限定到大数据领域，恐怕很少有公司能达到这个规模和质量。

所以，CDP这种企业数据混合云的长期存在和发展，“便是无疑的了“。

但要问在这种情况下Cloudera是否真的能王者归来，我无法下断言。

至于原因，不妨引用一下著名史学家史华兹的观点：否认客观环境先验的重要性是绝对愚蠢的行为，但我的确反对那种主张“形势“自动引起结果的万物有灵论，任务完成与否，不仅取决于所用的方法和客观环境，也取决于承担任务的那些人的思想、意图和抱负。

所以我能做的，只有拭目以待。