

阿里云封神:Gartner看衰的并不是Hadoop生态

转载

coqomndr954803650 于 2018-06-06 10:04:23 发布 251 收藏

文章标签: [大数据](#) [数据库](#) [区块链](#)

在过去几个月的走访调研中,针对Gartner《2017年数据管理技术成熟度曲线》做出的Hadoop“即将在到达生产成熟期之前衰落”的结论,笔者询问了数十位大数据领域技术专家的观点,Hadoop在国内大数据市场的地位正如笔者所预料的一样稳固。既然如此,那么是Gartner的结论有误吗?我们应该如何正确解读这份报告?抛开报告,Hadoop生态各组件到底表现如何呢?



本期采访嘉宾:封神,09年加入阿里,9年来专注在分布式计算、存储、数据库领域。曾研发集团超过1w台Hadoop集群,万台规模的跨机房建设,并负责其中分布式调度及内存计算引擎Spark。曾研发集团超过5w台的ODPS集群,自主研发分布式内存引擎,处理超过PB级别数据的迭代计算,并把开源体系融合到内部ODPS体系之中。16年起,开始做云产品,构建大数据数据库服务,研究分布式检索与原始数据的融合技术、研究行列混合、存储计算分离、新硬件加速、统一SQL、弹性调度计算、时序、时空、图数据等非结构化等技术,并负责计算与存储产品化,目前为千百客户提供专业存储计算云HBase服务。

Gartner看衰结论正确解读:此“Hadoop”非彼“Hadoop”!

Gartner所做出的的看衰结论不假,但其主体范围需要界定清楚。根据封神的介绍,Gartner所提及的Hadoop更多是狭义上的Hadoop一体化平台,但我们通常意义上讨论的是广义Hadoop生态,整个生态包含了众多组件,这个范围与前者相差很大。

对于Hadoop生态的发展状态,我们可以分层逐级解析。首先是HDFS分布式文件系统层,封神表示,目前尚没有任何一款开源产品足以完整替代HDFS,因此其生命力必定是旺盛的。目前看起来,有可能挑战的是业务模式的变化,用户上云后,直接用对象存储,做到存储与计算分离,节约成本。

其次是Yarn所在的分布式调度层。作为大数据核心调度组件,Yarn的使用覆盖率非常高。虽然在离线与在线数据混合方面表现欠缺,但Yarn一直在不断改进。这一点也与Yarn的发展背景有关,支持Yarn的几家公司主要做离线系统,对在线系统部署问题关注不够。当然,目前市场已经有了存在一定竞争关系的产品,比如Mesos,但这两大调度系统的设计目标并不完全相同,并且Yarn也在朝着Mesos的领域进军。此外,从某种意义上讲,Yarn与Hadoop生态体系中的一些组件包都可共享,贸然更换势必面临着适配问题。

在分布式文件系统和分布式调度系统的基础之上，各类组件的加入让Hadoop生态更加丰富。在绝大多数用户的认知中，Hive、MapReduce、热议的Spark以及Flink的定位都只是Hadoop生态中的一个计算引擎，并不存在替代Hadoop生态的关系，Hadoop生态的整体生命力非常强，这一点也在连续几个月的调研中得到了证实，Hadoop在大数据领域的标准地位远比我们想象得要稳固。



Hadoop生态圈组件生命力解析：替不替代取决于需求！

Hadoop生态圈组件繁多，组件替换是大多数企业对Hadoop生态圈进行改良的重要一步，但很多企业和技术人员对如何替换和挑选组件完全没有头绪。根据封神的介绍，虽然短期内Spark和Flink没有能力替代整个Hadoop生态，但是它们有能力替代个别计算引擎。Spark的生态日益完善，其在机器学习、SQL以及Streaming方面的先天优势明显；Flink主攻流式处理，延迟较低，这两大引擎目前的整体表现还是不错的。虽然部分企业会尝试使用Spark替换MapReduce，但在封神看来，这样的组件替换需要从实际业务需求出发，如果看中稳定性，MapReduce略胜一筹；如果看中生态所提供的能力，自然是Spark更加完善。

此外，Hive也非常有竞争力，阿里内部PB级数据量的处理非Hive不可，因为其稳定性非常高，当数据量达到PB级别，一个任务要运行两到三天时，没有一个组件可以挑战Hive的稳定性。作为开源分布式数据库，HBase也是大多数企业的首选，本月发布的HBase 2.0版本被称为迄今为止最大的版本，共包含4551个Issues，对可用性和延迟均做出了优化。对于现在很多数据库都在做的一写多读，HBase在1.X版本早期就具备该能力。其次，HBase非常适合云计算环境。不管是AWS还是阿里主推的关系型数据库，其本质都是存储计算分离，HBase诞生的第一天就是存储计算分离的，现如今经过了多年在各大公司的磨炼，HBase的存储计算分离已经相当成熟。

Hadoop渐步成熟期，其可能替代品的生存现状如何？

封神表示，就个人参加Hadoop大会的感受而言，Hadoop不是衰退了而是成熟了，现在的Hadoop大会更多的是在谈人工智能、区块链和知识图谱等新兴技术，大家更关注的不再是Hadoop生态的底层实现，而是其上的应用，这就造成了用户对Hadoop关注度降低的假象。

当然，任何一项技术发展了十余年，总会出现一些挑战者。封神谈及，NewSQL确实是分布式，也能解决部分存储相关的事情，具备存储和并发能力，但它主要解决的是核心场景，比如金融级场景下的核心问题，并不擅长海量数据存储且成本偏高。此外，Cassandra和Elasticsearch也能变相解决部分Hadoop问题，在工具的支撑下，用户非常容易上手，但随着规模的不断扩大，最终还是要回到整个Hadoop生态。

综上，封神认为Gartner所做出的结论更偏向于Hadoop本身而非Hadoop生态，但在大多数用户的认知中，我们在谈论Hadoop时已经默认是Hadoop生态，因此我们没必要对Gartner的结论太过苛责，毕竟目前国内的一线互联网企业都在使用Hadoop生态，短期内我们并没有看到改朝换代的迹象，反倒是其核心组件越来越成熟。如果不用Hadoop，你还可以做出什么选择呢？



▲ 点击参与大话IT讨论活动，有机会获得精美礼品

来自“ITPUB博客”，链接：<http://blog.itpub.net/31077337/viewspace-2155682/>，如需转载，请注明出处，否则将追究法律责任。

转载于：<http://blog.itpub.net/31077337/viewspace-2155682/>