# 十大国内外知名大数据专家探讨：Hadoop是生是死？

原创

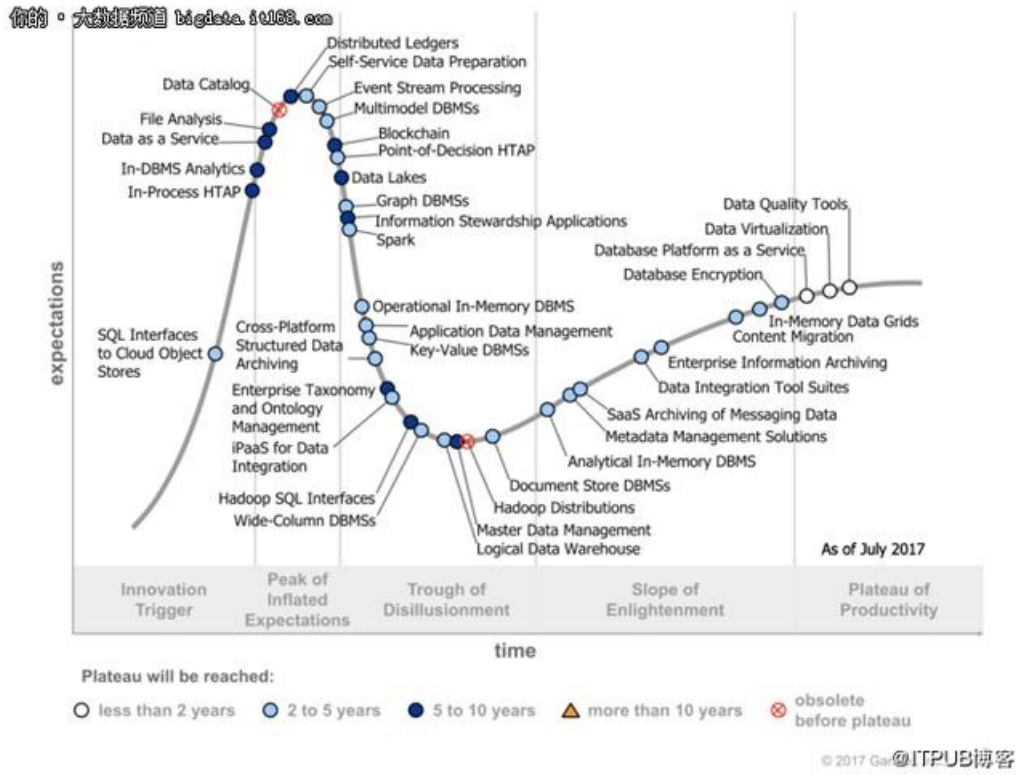
qunqun8889 于 2019-12-21 14:10:51 发布 286 收藏
分类专栏： 大数据 文章标签： 大数据 大数据开发 大数据分析 Hadoop 大数据入门

本文链接：https://blog.csdn.net/qunqun8889/article/details/103643909




2017年，Gartner发布的《2017年数据管理技术成熟度曲线》将Hadoop掀上舆论巅峰，报告极其明显的标识出Hadoop即将在到达生产成熟期之前进入淘汰席。

Gartner预测，到2018年，70%的Hadoop部署无法实现节约成本和收入增长的目标。在今年年初，Hadoop被列为2018年大数据领域的“渐冻”趋势之一，不少人将Hadoop称作“倒下的大象”，比如Lucidworks首席执行官Will Hayes。




Gartner认为，Hadoop到达生产成熟期前即被淘汰

当然，国内很多人将这种现象归结于国内外大数据领域发展状况不平衡造成的，因此笔者耗费了半年时间，走访了国内数家大数据厂商及技术专家，以下是10位技术专家的观点汇总，这些技术专家涵盖了国外的大数据厂商、银行、国内互联网公司以及国内大数据厂商，“Hadoop是生是死”一目了然。

**1、任何IT技术发展到一定阶段都会被挑战，Hadoop也不例外！**

采访对象：王苹，荣之联解决方案架构师。曾就职于IBM大数据团队，具有多年大数据平台研发经验。目前专注于大数据企业级应用的方案设计及技术选型，同时带领团队研发荣之联大数据产品。

王苹认为，企业之所以愿意使用Hadoop，是因为其足以解决现阶段企业用户在大数据方面存在的问题，并且其开源社区成熟完善。企业用户没有互联网公司乐于冒险，他们更愿意选择成熟稳定的解决方案，因此Hadoop的需求量还是很大。

至于是否会失宠，**已经为大家精心准备了大数据的系统学习资料，从Linux-Hadoop-spark-.....，需要的小伙伴可以点击**在快节奏的IT圈，任何一种技术发展到一定阶段都会被挑战，Hadoop也不例外。当然，Hadoop自身确实存在一定的问题，也有很多新技术足以弥补其缺陷。但是，新技术如果不与已经在企业中站稳脚跟的Hadoop打配合，又何谈市场呢？

**2、Hadoop或衰落，但核心组件生命力旺盛！**

采访对象：刘译璟，百分点集团技术副总裁兼首席架构师。

刘译璟认为，单就Gartner报告，我们很难对Hadoop判死刑。毕竟，事实上，它已经存在于国内很多企业的大数据架构中，每天都会有成千上万的任务运行在Hadoop之上，这其中不免核心任务。

但是，Hadoop生态中的各组件生命力有很大差异，一旦其中的大部分组件都被替换掉，整个生态也很难称之为“Hadoop生态”。

谈到组件，他认为HBase、HDFS以及ZooKeeper这类组件的生命力还是挺长的，短期内不会消失。

毕竟，类似HDFS这样的基础组件消失是很困难的，无论是Spark还是Flink，底层的文件系统都是HDFS，很少有第三方厂商基于开源再造一个文件系统，HDFS在某种程度上奠定了大数据的基础。

但是，MapReduce、Hive这类组件确实可能被Spark等替换掉，随着硬件越来越成熟，Spark的优化工作越来越好，企业很可能倾向于在内存中计算。

此外，Hadoop在机器学习方面确实不太擅长，Mahout等组件表现不佳，成为不少企业选择Spark的原因之一。

最后，资源管理器Yarn与Hadoop的绑定过于紧，而实际上，我们有很多资源调度管理方法可供选择，比如Kubernetes等，对各种应用的支持某种程度上比Yarn更完善，无论是外部类型应用，大数据应用还是机器学习应用均可处理。

**3、Hadoop确实有问题，但不能成为“看衰”论断的主要原因！**

采访对象：星环科技，星环Transwarp Data Hub是Gartner认可的Hadoop国际主流发行版。

有不少人认为Gartner报告中提到的Hadoop是指“Hadoop发行版”，如果是这样，那么星环科技相当有话语权，因为其创业团队很早之前就在做Hadoop发行版的工作。

在实际的使用中，星环也承认Hadoop有一些缺点，比如使用门槛略高，技术迭代快导致学习成本和运维成本升高。不过，这些缺点并不是致命的。

至于Gartner的这一言论，星环科技认为这与Hadoop自身存在的问题以及国内外大数据环境的差异有关，一方面，Hadoop的使用有一定门槛，虽然过去几年人才供应数量在不断增加，但是企业对人才的需求增加速度更快，所以企业构建Hadoop团队的人才成本较高，初次构建成本偏高。

另一方面，Gartner的调查客户主要集中在国外，而国外Hadoop厂商给客户提供的功能无法完全取代传统数据库的地位，未能将Hadoop的优势全部体现，导致国外用户对Hadoop的应用比较简单，未能充分体现新技术带来的优势，故容易得出Hadoop投入产出比较低、能力局限较大的结论。

国内用户对Hadoop的认可度偏高，是因为国内使用Hadoop技术的数据量和应用场景的复杂度都远超国外用户，新技术替换旧技术的过程给国内用户带来的价值显著，例如整体成本降低，性能提升，扩展方便，基于新技术进行的业务场景创新等，这些都让国内用户切实感受到Hadoop生态的强大。

**4、Hadoop地位稳固，其他竞争者尚不具备叫板能力！**

采访对象：天云大数据，天云大数据是国内为数不多的大数据PaaS层组件研发厂商，其BDP(Beagledata Platform)平台是一款基于Hadoop生态体系的企业级大数据中间件平台。

天云大数据认为，Hadoop未来发展还是泛生态的发展，它会是整个通用计算框架演进迭代的一个过程。企业与其花心思研究其组件级别的优劣，不如将更多精力放在Hadoop生态演进以及自我大数据架构的优化上。

至于可能的竞争对手——Spark和Flink，天云认为二者尚且不具备与Hadoop叫板的实力，未来更倾向于合作共赢的方式。

**5、Gartner看衰结论正确解读：此“Hadoop”非彼“Hadoop”!**

采访对象：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。

封神认为，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，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；

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。

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。

**6、Hadoop失宠前提是出现更强大的替代品！**

采访对象：苏宁易购，其大数据平台基于Hadoop构建。

对于Gartner的唱衰论调，苏宁易购认为，Hadoop就好比日常生活中的水电煤，因为太普遍反而引不起特别关注，或者，Gartner报告中所说的Hadoop是指狭义上的Hadoop，也就是原始的HDFS和MapReduce组合。

如果单看这两大组件的发展，MapReduce确实在逐渐退出舞台，被Spark/Flink所取代。苏宁易购认为，Hadoop失宠前提一定是出现更强大的可替代大数据解决方案，现在来看，并没有这样的方案出现。

**7、Hadoop已经展现出极强的年代感，并且其在机器学习方面是有欠缺的！**

采访对象：Ness SES的CTO Moshe Kranc

Ness SES的CTO Moshe Kranc认为，Hadoop已经展示出了其年代感，不管是Hadoop的HDFS、MapReduce还是它的机器学习组件Mahout。

在这一方面，Spark似乎表现更加优异，Spark不断从Hadoop的经历中学习，具有更通用和可扩展的编程模型，易于分析且拥有强大的图形数据库(Graphx)和全功能数据科学库(MLib)。当然，如果企业自己具备生态整合的能力，那么这个问题可能就不存在了。

**8、很多企业都低估了部署Hadoop的复杂度！**

采访对象：Silicon Valley Data Science的CTO John Akred

Silicon Valley Data Science的CTO John Akred表示，在国外，无论是医疗保健、制造业还是金融领域，公司在部署Hadoop这样的分布式系统时一般会选择从初始用例也就是简单用例开始，以便了解整个Hadoop的工作流程。

公司可能会开始尝试将部分数据收集并运行到Hadoop之上，通过简单的测试证明，确实可以使用Hadoop来存储大量非结构化数据，到这里所有步骤似乎都没有问题，但这真的对业务产生价值了吗?如果企业并没有通过部署Hadoop而对业务产生价值，那么这一决策的意义是什么呢?

其次，很多企业会低估Hadoop的操作复杂性，无法清晰认知习惯了使用IBM Db2和Oracle等传统数据技术的人在使用Hadoop方面会面临多少转型问题。

**9、企业用户对数据湖需求旺盛，但对Hadoop接受意愿较低！**

采访对象：Teradata天睿公司策略性产品管理高级副总裁Tim Henry

Tim Henry认为，Hadoop更直接的使用者是企业用户而不是大数据厂商，虽然数据湖或Hub的概念最初由大数据厂商提出，但真正的大规模应用还是在企业内部。这些企业很可能并不会选择Hadoop，因为Hadoop的管理相当困难，尤其是技术层面。

要想使用Hadoop进行数据治理，企业员工必须对Hadoop的整体运作流程以及各大组件非常熟悉，否则无法从众多组件中挑选出符合业务需求的组合，导致无法发挥Hadoop的真正价值。

企业并不是对数据湖没有需求，而是对Hadoop的接受意愿较低，这也同样契合了Gartner的结论。

**10、我们非常看好Hadoop的未来发展，不知道Gartner的这一结论从何说起！**

采访对象：Cloudera创始人Mike Olson

Cloudera创始人Mike Olson在国外接受采访时，对Gartne报告中关于Hadoop的观点进行了驳斥，并谈到了他的看法。Mike Olson表示并不认同Gartner对Hadoop的结论，有很多客户在其平台上执行关键业务，他不清楚Gartner到底跟谁讨论得出的这一结论。

他表示，Cloudera不仅是看好，更为重要的是已经在一些方面已经取得了很大成功。例如：通过使用Impala等工具进行高性能分析查询，企业可以在扩展平台上为其传统关系工作负载的某些部分提供替代方案。

他坦言，不得不怀疑Gartner是否看到的是10年前的Hadoop，而非现在。早期的Hadoop只有MapReduce和HDFS，确实非常有限，但它并不是Hadoop的全部，现在有26个不同的开源项目，包括Spark，其中有18种是Cloudera创建的，这是一个比过去更广阔、更有能力的生态系统。

**结论**

从上述多位技术专家的言论中不难看出，Hadoop在国内之所以流行是因为国内很多大数据厂商为企业省去了部署Hadoop解决方案的麻烦，而国内的互联网企业具备自己搭建并改进的技术实力，如果这些问题都得到了解决，那我们自然愿意享有Hadoop带来的优势。

但是，国外的技术专家却是非常一致的不看好，除了Mike Olson，这与国外的大数据环境也有关，国外很多企业所拥有的数据量可能尚未达到使用Hadoop的级别，国外大数据厂商所提供的服务可能没有国内厂商深入，这或许就是Gartner这一报告的症结所在。

当然，对于Gartner报告中所提及的Hadoop到底是Hadoop发行版，Hadoop一体化商业模式还是Hadoop生态似乎各种说法都有，但其报告中（如头图）使用的“Hadoop Distributions”似乎更倾向于Hadoop发行版。