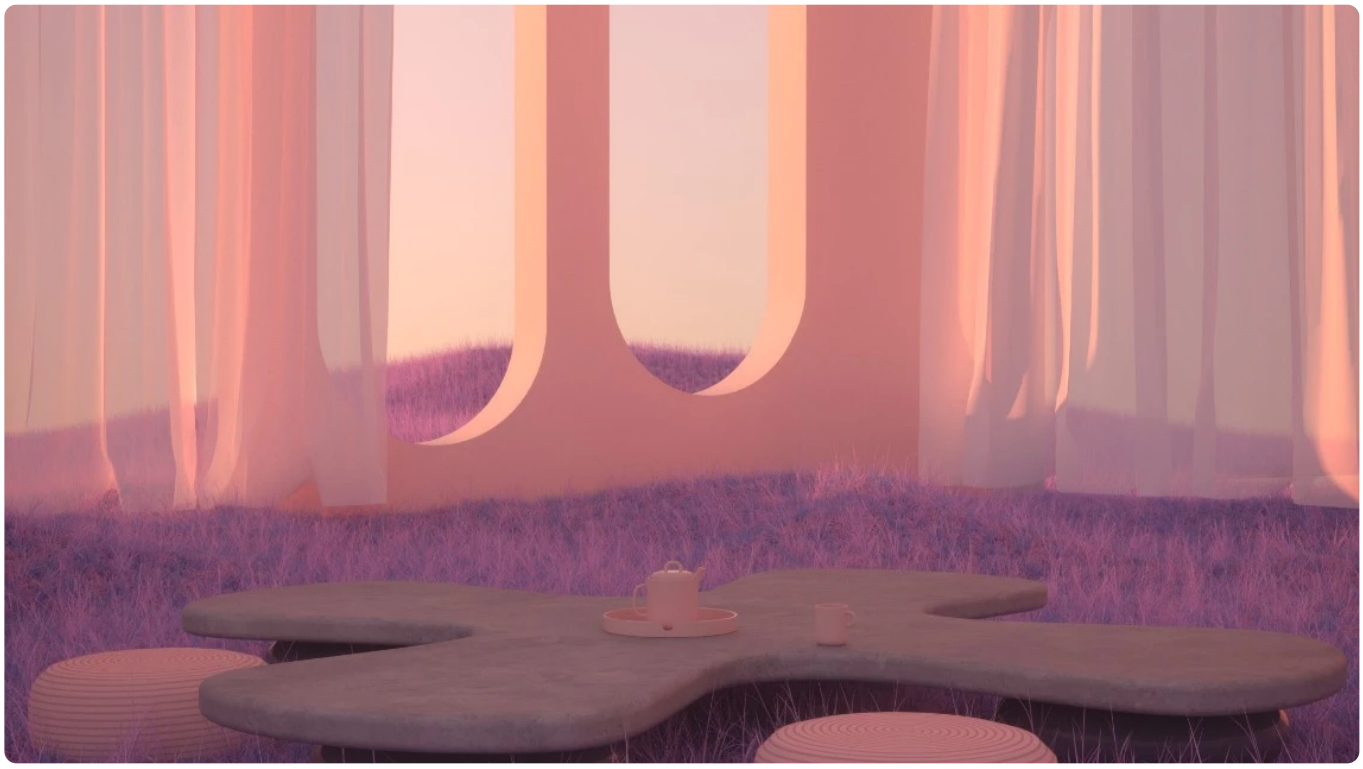
下载APP

加餐3 | 我该使用什么样的大数据系统？

2022-01-03 徐文浩

《大数据经典论文解读》 课程介绍 >



讲述：徐文浩
时长 13:09 大小 12.05M

你好，我是徐文浩。过去几年，不管你处在什么样的行业，收集数据并利用数据做决策已经成为了一个主流趋势。就从我自己的体会来说，2015 年之前，大数据系统在互联网技术团队非常火热，而在 2015~2020 年之间，AI 和机器学习超越了大数据成为最流行的前沿技术方向。

不过，到了最近一两年，大数据系统的需求一下子又激增了起来。这也是因为大数据不是一个不够成熟的前沿技术，而是成为技术基建不可或缺的一部分了，使用大数据的公司不仅仅是互联网公司了。传统行业的 IT 团队、提供 SaaS 的软件公司，也都需要在自己的业务上使用上大数据系统了。

领资料

所以，最近一两年，周围会有很多朋友来咨询我关于大数据系统的各类问题。比如，“我们想要搭建一个大数据系统，应该选择 Hadoop 还是 Spark？” “现在 MySQL 性能不行

了，我们想要搭建数据仓库，是 ClickHouse 更好还是 Apache Doris 更合适”。而在咱们的课程中，也有同学希望我来讲一讲，搭建大数据体系，从搭建数据湖、数据批处理、流式处理以及交互式分析，分别应该什么样的产品。

那么，今天这节课，我们就利用加餐来讲讲，对于大数据系统的选择方法问题。

以 TCO 和 ROI 为中心的决策方法

作为技术人，我们往往在内心对于技术先进性有着深深地向往。不过，在进行系统和产品选择的时候，“技术先进性”往往不是我们的第一原则。

作为一个需要对团队、公司、业务长期负责的人来说，我们选择什么样的产品，第一原则应该是考虑**总体拥有成本（TCO，Total Cost of Ownership）**，以及对应的**投资回报（ROI）**，也就是我们需要考虑选择了特定的数据产品之后，需要花多少钱，以及会带来什么样的回报。

那么，在选择使用什么样的大数据产品的时候，你至少需要考虑以下这些成本和回报。

首先是**硬件成本**，你要考虑对应的数据产品需要什么样的硬件配置，成本大概是多少？更重要的是，你需要选择，是自建或者租用的数据中心，去采购硬件？还是使用云平台的虚拟机，自己搭建一个集群？或者是直接使用云平台原生提供的数据产品，比如 AWS 的 Redshift 或者 Google Cloud 里的 BigQuery。

你的数据量越大，团队规模越大，越有可能选择自建，这样可以通过规模效应压缩成本。反过来，你越应该考虑使用云平台的原生方案，特别是像 Google Cloud 的 BigQuery 这样的方案，可以直接按照你运行的 SQL 扫描的数据量收费，如果不使用，只有极低的存储成本，连虚拟机租赁费用都不需要。

其次是**人力成本**，一旦开始使用大数据产品，我们通常会有三种角色。

> 第一种，是一个**基础设施团队**。如果我们选择自己建设数据中心，那么就需要有人负责采购、维护硬件，维护使用的开源系统的最新版本，这些都是我们“拥有”大数据系统的成本。如果我们选择使用云自带的 BigQuery 这样的系统，那么这类成本其实已经包含在我们使用系统的费用里了。

第二种，通常是一个**数据处理团队**，一般进行的是 ETL 类型的开发工作，也就是把系统里散落在各处的日志统一收集起来，清洗处理落地到我们的数据湖或者数据仓库里。

最后一种，则是我们的**数据分析团队**，也就是拿着数据处理团队生成的数据湖或者数据仓库，进一步进行业务层面的分析。

通常来说，我们不希望我们的数据团队是“头重脚轻”的。数据分析团队的人数应该要比数据处理团队的人多，数据处理团队的人数应该要比基础设施团队的人数多。因为我们不是为了使用大数据系统而使用大数据系统，最终的目的还是**利用好数据进行决策**。

最理想的情况是整个公司所有的人，都能充当数据分析人员，而这个就需要我们有一套体系化的数据系统，让任何人都可以很容易地获取、分析、展示数据。这个，我们稍后会进一步讲解。

第三点，要考虑的则是**系统的锁定成本和沉没成本**。我们一旦选择了一个方案或者一个产品，就会面临未来可能需要替换这个系统的问题。如果我们选择了自建数据中心，那么考虑到采购的硬件成本，我们是很难在短期内把这些硬件都报废掉，然后迁移到使用云平台的。

而一旦选择了使用 HBase 作为我们的 KV 数据库，过了一年再要迁移到 Cassandra，也会困难重重。一方面，我们已经有了大量的数据在 HBase 里，迁移系统意味着迁移海量的数据，以及在迁移过程中需要双倍的硬件冗余；另一方面，之前的大量系统的代码，可能都是依赖于 HBase 的，迁移数据产品的同时意味着要有大量的系统代码的修改。

很多时候，为了迁移数据产品又不中断服务，花上一个季度乃至更长时间进行系统迁移，都是经常发生的情况。而这些硬件投入、迁移过程中的开发投入，也是很大一块成本。

最后，需要考虑的则是**使用各种数据产品带来的回报**。我们上面讲的都是成本，那么该不该用某个数据产品，或者是不是在特定情况下需要进行数据迁移，我们最终都是要从投资回报率来看的。

比如，我们可能原先只有批处理的数据分析系统，那要不要上实时的流式数据分析系统呢？当然，从技术先进性的角度来看，我们确实应该这么做。但是，这也意味着要投入新集群的硬件成本，招募有对应开发经验的工程师，以及要培训现有的数据分析团队，如何对于实时数据处理进行分析。

并不是所有的时候，技术先进都是最佳答案。

比如，即使是 2022 年的今天，像 Spanner 这样能做到全球部署、外部强一致性的分布式数据库，也只是在极少数公司的少数产品和团队中使用到。因为全球部署对于大部分的公司来说，能够带来的收益是极其有限的，但是成本却比只在一个数据中心里的 Hive 数据仓库，或者 HBase 数据库要高得多。在选择数据产品的时候，我们需要去评估对应这些新系统，能给我们带来什么样的帮助。

所以，这三类成本以及最终能够得到的收益，成为了我们评估和选择使用什么样的数据系统的思考点。而且，这个基于 TCO 和 ROI 的原则，我们也可以用来去评估使用其他系统的情况。

在这里，我只是做了简单的原则性的介绍，而在实际应用中，你应该尽可能地把这里的 TCO 和 ROI **量化**出来。

比如，你应该预估一下，使用云平台提供的 Hadoop 集群，每月的服务器费用是多少，而使用 BigQuery 的话，预计扫描的数据量会需要多少成本；比如，在你决定使用流式系统，提升广告点击率带来的收益的时候，虽然很难事先衡量回报会是什么，但是也可以在系统上线后，通过 A/B 测试，看看带来额外广告收益，是否可以覆盖新增流式系统的硬件和维护成本。

搭建数据产品和系统的两个核心经验

光有这么一个原则，可能对很多同学来说，还是很难立刻运用到实践中去。那么，我再来分享一下我自己的经验。这些经验，既来自我自己过去工作中，搭建的大大小小的数据平台的直接经验，也有来自为其他公司提供咨询，从零搭建数据团队和数据产品的切身体会。我把这些经验总结为了两个核心要点。

从可变成本起步

第一个核心要点，是在数据产品的选择和数据系统的搭建的初期，尽量让你的成本都变成是一个**“可变成本”，尽量避免一次性的大额投入**。

和我们做互联网产品一样，我们很难在一开始，就做出一个百分百正确的决策。团队本身也需要在实际的数据产品的使用、开发中获得经验，才能更好地做出正确的决策。而如果

一开始就采购上 100 台硬件服务器，那就真的是“开弓没有回头箭”了。即使后续发现决策是错误的，也很难纠正。

因为我们面临着大量的沉没成本，我们不可能就把 100 台服务器都报废了，很多时候不得不咬着牙将错就错，为了这些硬件，再去招募运维基础设施的团队，最后我们的 TCO 会越来越高。

所以，如果你之前只有 MySQL 这样的关系型数据库，现在随着业务需要开始搭建大数据系统。你可以先从**云平台**本身提供的 BigQuery 这样的，“按使用量付费”的产品开始。这样，你就避开了一开始就要做出重大决策，进行硬件采购的坑。

你可以先使用这些云平台本身提供的产品和功能开始“试错”，并且在内部磨合基础设施、数据处理和数据分析团队之间的流程和职责划分。你的数据系统，在一开始跑一个分析任务，只有几分钱乃至几毛钱的成本，而当你真的每天都有大量的数据分析任务，需要花上个几百几千块钱的时候，再来重新做一次产品选择的决策也完全来得及。

不要忽视数据处理流程管理和数据可视化

第二个核心要点，是**尽早重视数据处理流程（Data Pipeline）的管理，以及数据可视化（Data Visualization）的功能**。在我们这个课程里，没有任何一篇论文是关于 ETL 流程，以及数据可视化的。的确，这两个都是纯粹的工程应用问题，而不是新鲜的大数据处理的前沿技术。但是，在现实世界里，往往这两点，才是你的数据系统是不是真的能被好好使用起来的关键点。

我们先来看看**数据处理流程管理**，在使用数据系统的时候，我们往往会有来自多个不同数据源的原始数据（Raw Data）。我们需要通过 ETL 这样的数据处理程序，进行清洗和处理，才能得到我们设计好的数据仓库的各类报表。

不同的 ETL 程序之间会有依赖关系，并且因为海量日志的存在，我们总会面对一些脏数据。这个就会导致我们会遇到 ETL 程序执行失败，或者计算错误等情况。在这个情况下，我们就需要修改程序，重新运行，并清理掉过去的错误数据。

所以，我们需要有一个机制，一方面，确保所有的数据处理程序，执行起来都是幂等的，也就是反复多次执行同一个 ETL 程序，最后生成的报表数据是永不变的。

另一方面，我们还需要让各个依赖关系能够通过一个系统管理起来。任何一个中间计算节点失败，或者需要重新运行，我们都可以很容易地重新运行后续的所有任务。最后，我们也不能让依赖关系过于复杂，避免数据处理流程长期维护会越来越困难。

一般来说，你可以选用像Apache DolphinScheduler这样的调度系统，来作为你的数据处理的流程管理。做好了数据处理的流程管理，可以大大减少你的团队在日常 Bug 修复、系统维护上的时间。这个，可以加速你的数据产品和系统的迭代，你也就自然有了更高的 ROI。

然后是**数据可视化**，我和很多周围做数据系统的朋友说过“**千言万表，不如一图**”。我们搭建数据系统的最终目标，是为了帮助我们进行业务决策。那么，最容易帮助我们决策的方式，是把对应的数据以图的形式展现出来。无论是文字解释，还是复杂的表格，都不如一个简单的图能够说明问题。

很多时候，数据产品没有在业务中体现出价值，并不是我们系统搭建得不好，产品选得不对，而是数据团队的产出，没有正确地和业务团队“沟通”。而把数据可视化，变成一个个看板，是比起任何“沟通技巧”有效的沟通方式。

而且，好的数据可视化系统，可以进一步让业务团队，**直接基于你的数据仓库进行数据分析**。让业务团队人人学会用 SQL 或者 MapReduce 并不现实，但是一个设计良好的数据仓库，配上一个可以拖拽的数据可视化系统，却是人人都可以很快上手使用的。

所以，尽早用上一个好的数据可视化系统，其实是最快能够体现数据产品的业务价值的办法。而这样的正反馈，也会使得我们有信心进一步加大对于大数据产品和系统的投入。无论是 Google Data Studio 这样免费的系统，还是 Tableau 这样收费的商用系统，现在的价格都已经不贵了，用起来也很方便，绝对是值得回票价的。

小结

好了，希望这节课的内容，能为你选择大数据产品、搭建大数据系统提供一些有益的参考。

在开始搭建大数据系统和团队的时候，你的核心决策原则，应该是围绕着 **TCO 和 ROI** 这样两个可量化的数字进行。技术先进性，也只是围绕这两个指标中思考的一个角度而已。

而在实际选择产品、搭建系统的时候，尽可能让所有成本变成**可变成本**，会降低你试错的成本。

进一步地，你应该重点关注数据之间的依赖关系，尽早让你的**数据流程可维护**；反而，具体选用哪一个数据产品，很多时候并没有那么重要，因为主流的产品都已经很完善了。而**数据可视化**，是让你的数据能够真正体现在决策上的重要环节，很多时候会被大家忽视，希望你千万不要踩到这个坑里。

希望这一讲，能对你选择数据产品、搭建数据系统的决策有所帮助。之后，我会再加入一讲加餐，给出我自己使用过的产品之间的具体对比，以及实战案例。

推荐阅读

除了选择产品、搭建系统之外，更重要的一个事情是搭建你的**数据团队**。LinkedIn 原先的数据产品团队的负责人，也担任过白宫首席数据科学家的 DJ Patil，为这个问题也专门写过一本书，叫做《Building Data Science Teams》，作为他山之石，你也不妨拿来读一下。

思考题

你能在留言区聊聊，之前你所在的团队选择某一个大数据产品的决策原因么？为什么你们选择了当前使用的大数据产品，而没有选择另一个？

欢迎把今天的内容分享给更多的朋友，咱们下节课再见。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 1 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 加餐2 | 设置你的学习“母题”：如何选择阅读材料？

更多课程推荐

精选留言

写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。