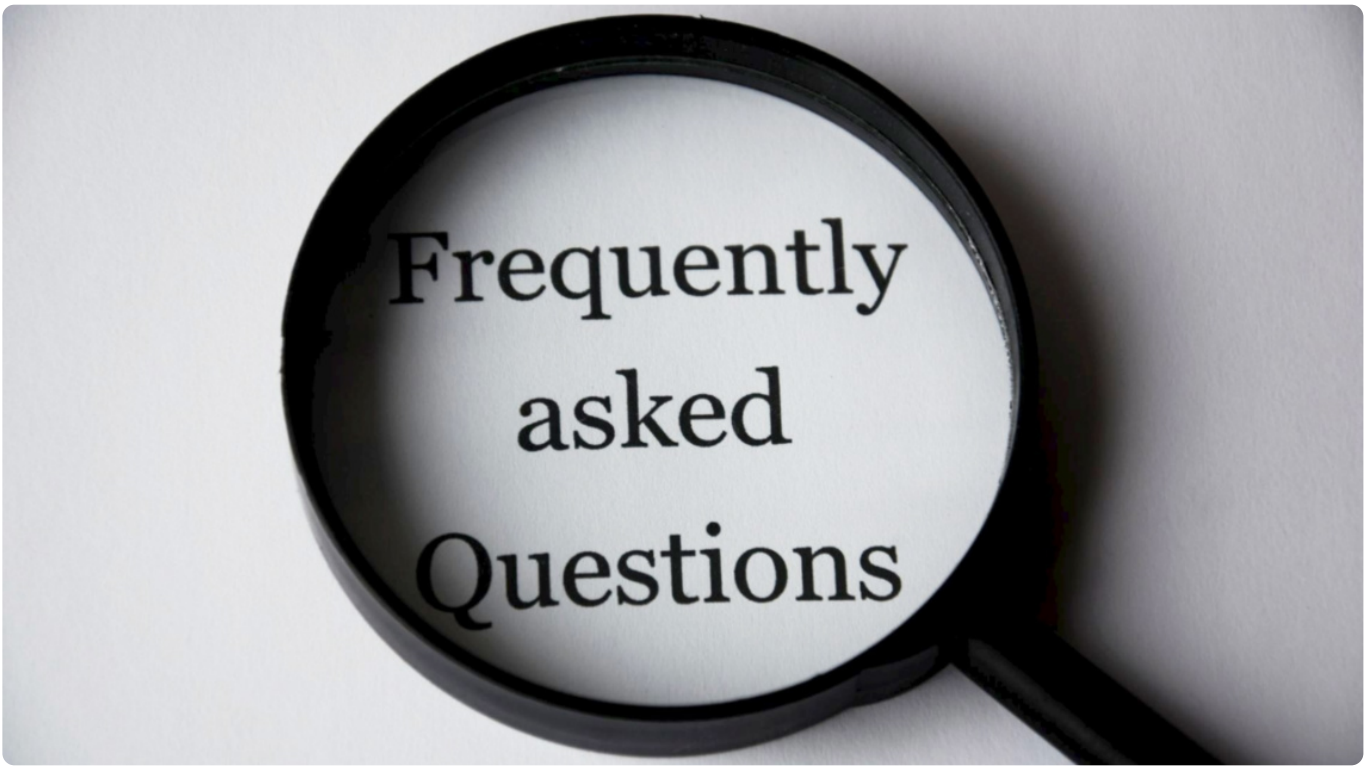
17 | 模块答疑：这么多技术，到底都能用在什么场景里？

2018-12-06 李智慧

从0开始学大数据

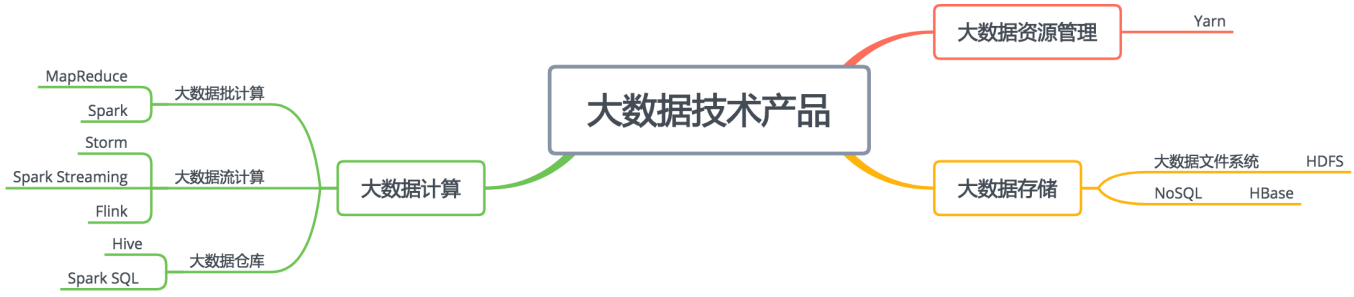
进入课程 >



讲述：李智慧
时长 10:38　大小 9.75M

你好，我是李智慧。

经过前面两个模块，我们学习了大数据最经典、最主流的一些技术和产品，今天我们再回过头来梳理一下这些技术和产品。



从上面这张图来看大数据技术的分类，我们可以分为存储、计算、资源管理三大类。

最基本的存储技术是 HDFS。比如在企业应用中，会把通过各种渠道得到的数据，比如关系数据库的数据、日志数据、应用程序埋点采集的数据、爬虫从外部获取的数据，统统存储到 HDFS 上，供后续的统一使用。

HBase 作为 NoSQL 类非关系数据库的代表性产品，从分类上可以划分到存储类别，它的底层存储也用到了 HDFS。HBase 的主要用途是在某些场景下，代替 MySQL 之类的关系数据库的数据存储访问，利用自己可伸缩的特性，存储比 MySQL 多得多的数据量。比如滴滴的司机每隔几秒就会将当前的 GPS 数据上传，而滴滴上的司机数量号称有上千万，每天会产生数百亿的 GPS 数据，滴滴选择将这样海量的数据存储在 HBase 中，当订单行程结束的时候，会从 HBase 读取订单行程期间的 GPS 轨迹数据，计算路程和车费。

大数据计算框架最早是 MapReduce，目前看来，用的最多的是 Spark。但从应用角度讲，我们直接编写 MapReduce 或者 Spark 程序的机会并不多，通常我们会用 Hive 或者 Spark SQL 这样的大数据仓库工具进行大数据分析和计算。

MapReduce、Spark、Hive、Spark SQL 这些技术主要用来解决离线大数据的计算，也就是针对历史数据进行计算分析，比如针对一天的历史数据计算，一天的数据是一批数据，所以也叫批处理计算。而 Storm、Spark Streaming、Flink 这类的大数据技术是针对实时的数据进行计算，比如摄像头实时采集的数据、实时的订单数据等，数据实时流动进来，所以也叫流处理大数据技术。

不管是批处理计算还是流处理计算，都需要庞大的计算资源，需要将计算任务分布到一个大规模的服务器集群上。那么如何管理这些服务器集群的计算资源，如何对一个计算请求进行资源分配，这就是大数据集群资源管理框架 Yarn 的主要作用。各种大数据计算引擎，不管是批处理还是流处理，都可以通过 Yarn 进行资源分配，运行在一个集群中。

所以上面所有这些技术在实际部署的时候，通常会部署在同一个集群中，也就是说，在由很多台服务器组成的服务器集群中，某台服务器可能运行着 HDFS 的 DataNode 进程，负责 HDFS 的数据存储；同时也运行着 Yarn 的 NodeManager，负责计算资源的调度管理；而 MapReduce、Spark、Storm、Flink 这些批处理或者流处理大数据计算引擎则通过 Yarn 的调度，运行在 NodeManager 的容器（container）里面。至于 Hive、Spark SQL 这些运行在 MapReduce 或者 Spark 基础上的大数据仓库引擎，在经过自身的执行引擎将 SQL 语句解析成 MapReduce 或者 Spark 的执行计划以后，一样提交给 Yarn 去调度执行。

这里相对比较特殊的是 HBase，作为一个 NoSQL 存储系统，HBase 的应用场景是满足在线业务数据存储访问需求，通常是 OLTP（在线事务处理）系统的一部分，为了保证在线业务的高可用和资源独占性，一般是独立部署自己的集群，和前面的 Hadoop 大数据集群分离部署。

今天我帮你把专栏前面讲过的大数据技术串联起来，并放到了比较具体的应用场景中，后面在专栏模块四，我们还会讨论如何将这些技术产品集成为一个大数据平台，希望能帮你更进一步了解大数据技术的应用方法和使用场景，请你一定坚持把专栏学完。

在专栏文章里，“蜗牛”同学问了我这样一个问题，我回顾了下自己的过往，也想和更多同学分享一下我的人生观。

蜗牛

写于 2018/11/29

老师、那我代表同学们也问一个宽泛的问题可以吗：

请问您觉得，是哪一本书、或者哪一件事、或者哪一句话，对你的人生产生过很大的影响？或者说、您觉得您人生的转折点在哪里。

谢谢老师！ 蜗牛🐌 给您致敬！

> 引自：从0开始学大数据
> 14 | BigTable的开源实现：HBase

识别二维码打开原文
「极客时间」 App

我在评论回复里，讲到王小波的《我的精神家园》。读这本书，大概是在我大学快毕业的时候，当时面临种种困惑，努力思考自己这一生应该如何度过，自己应该做一个什么样的人。

当时我就读于一所不入流的大学，在大学里面我又是个不入流的学生，从当时的趋势看，我未来的人生大概率也是不入流的人生，浑浑噩噩、蝇营狗苟度过一生。

虽然有的时候踌躇满志，也想要改变自己，将来有一天能出人头地。但是更多的时候想想自己的天分、背景，就不由得万念俱灰。进一步想，如果自己注定要平庸一生，活着又有什么意义。现在想来，当时可能是有一点抑郁的。

也就在那个时候读到了王小波的杂文集，关于人应该如何度过自己的一生，有了不一样的理解。王小波说：“**我活在世上，无非想要明白些道理，遇见些有趣的人，做一些有趣的事。倘能如我所愿，我的一生就算成功。**”

王小波的书当时给了我一种全新的认知世界的视角，我不一定要出人头地才叫成功，我能把自己的一生过得有趣、好玩，我也就算没白活一生。

但是如果简单的把好玩、有趣理解成自得其乐、不思进取，这样的生活肯定是有问题的。看王小波的其他文章就会明白，这个好玩、有趣也不是一件容易的事，没有一定的知识、见识，没有有深度的思考，没有经历过足够的困难、挫折，就根本不能理解哪些是真正好玩、有趣的人和事。

所以你必须还是要努力拼搏、锐意进取，然后在这个过程中才能真正明白一些道理，并且会遇到一些有趣的人。“**我只愿蓬勃生活在此时此刻，无所谓去哪，无所谓见谁。那些我将要去的地方，都是我从未谋面的故乡。以前是以前，现在是现在。我不能选择怎么生，怎么死；但我能决定怎么爱，怎么活。**”

想通了这一点后，我就不再纠结自己是不是足够的优秀，能够成就什么样的事业。我只要每天都有一点点进步，明白一点点道理，生活就是值得的。所以毕业以后我觉得编程好玩，就去自学编程；觉得自己学得差不多了，就去找了一份程序员的工作；觉得缺乏创造、不断重复的程序员工作并不好玩，就去考计算机专业的研究生；后来又去北京、杭州、上海不同的城市生活；去阿里巴巴、Intel、创业公司等不同的公司去工作；期间遇到过很多有趣的人，跟很多聪明的人合作，明白了一些道理，也做过一些有趣的事。

再说几本对我影响比较大的技术书籍。我大学读得不是计算机专业，后来偶尔在图书馆里看到一本 C 语言编程的书，讲图形编程和游戏编程的，当时觉得特别好玩，就开始自学编程。但是后来做了程序员以后，却发现天天按着需求写代码特别无聊，既没有挑战也没有创

新，偶然看了一本名为《Effective Java》的书，发现这些常规的程序也有很多有意思的写法。

这本书讲了很多有趣的编程技巧，当时我在北京的中关村上班，每天上下班的地铁上，刚好可以看完一个技巧，很多技巧可以直接用在工作中，特别好玩。同时我也意识到，很多时候不是事情本身无趣，而是做事情的方式无趣。循规蹈矩、反复重复可以把事情做完，但是这样就会很无聊，如果去寻找更巧妙的解决之道，事情就变得有趣多了。

后来我就想，能不能把这些技巧提炼出来，让大家一起用，所以在看到《敏捷软件开发：原则、模式与实践》这本书的时候，我非常激动。在读这本书之前，我也看过设计模式的书，不过感觉这些书都把设计模式当做编程技巧来讲，虽然也有意思，但是技巧嘛，看得多了也就那么回事。

但是《敏捷软件开发》这本书把设计模式上升到设计思想的高度，书中说“**软件设计不应该是面向需求设计，而应该是面向需求变更设计**”，也就是说在设计的时候，主要要考虑的是当需求变更的时候，如何用最小的代价实现变更。**优秀的工程师不应该害怕需求变更，而应该欢迎需求变革，因为优秀的工程师已经为需求变更做好了设计，如果没有需求变更，那就显示不出自己和只会重复的平庸工程师的区别。**这就非常有意思了不是吗。

因为比较关注设计，并且后来又做了一些架构设计、框架和工具开发的工作，也因此我看到《企业应用架构模式》这本书的时候特别震撼。当时自己觉得能够做框架、工具的开发很了不起，能够做架构设计、指导其他工程师开发挺厉害，但是看了《企业应用架构模式》这本书，才发现我做的这些事情只不过是在更大的领域解决方案、架构模式里非常小的一个部分，同类的东西还有很多。当时我感觉自己真的是坐井观天、夜郎自大，也非常羞愧。

如果感觉《敏捷软件开发》的作者 Bob 大叔、《企业应用架构模式》的作者 Martin Fowler 比自己牛太多还没有太多感觉，因为毕竟隔得太远、没有交集，所以触动还不是特别大，那么后来在工作中遇到被高手全方位碾压的时候，就真正的感受到：生活还真是有意思呢。

如果你也有和我一样有过类似的困惑，不知该如何面对理想和现实之间的差距，希望我的经验可以给你一些启发。人类的进步是因为人们对美的不懈追求，而有趣也是美的一种，追逐有趣就是在追求进步，而一点一滴的进步终会引领我们实现自己的人生价值和目标。

在专栏第 15 期，我邀请了淘宝的高级技术专家李鼎来聊聊流式业务架构重构的心得体会，我把他的留言也贴在下面，感兴趣的同学可以返回第 15 期，也和我们聊聊你对这种架构的思考与理解。

李鼎(哲良)

写于 2018/12/06

数据流是久经考验的典型思路，在网络协议（如 TCP）、数据平台这样场景，早就应用多年习以为常了。淘宝业务的应用架构升级可以认为是把这样思路应用到了业务系统开发中，把『流』作为业务表达上的一等概念和手段，并在业务架构 / 系统能力优化提升。

简单地说，因为业务面向数据流来编写，一方面业务逻辑表达可以自然接近业务流程；另一方面逻辑运行可以是全异步有很好的性能提升，一核心后端应用在双 11 线上，单机 QPS 提升 30%，RT 下降 40%。流程的表达与异步 / 同步执行方式是分离的（如果了解过像 RxJava，这句会容易理解：）。

另外，『流』也为业务系统的保护提供新的一些方法，在思路上其实和流计算平台是一样的，这对业务大型系统的稳定性来非常重要

的，这对业务大型系统的稳定性来非常重要。

当然，业务的流式架构，在业务编写上有些 FP 风格（简单地说比如充分使用了 Lambda），平时我们大家业务上主要是用命令式顺序平铺方式来表达，会有要个熟悉过程，虽然不见得有多难 :)

引自：从0开始学大数据
15 | 流式计算的代表：Storm、Flink、Spark Streaming

识别二维码打开原文
「极客时间」 App

最后还是老规矩，我精选了三木子、Lambda、hunterlodge、老男孩这几位同学的留言，贴在今天的文稿里分享给你，希望同学们的思考也能对你有所启发。

三木子

写于 2018/11/24

比如学习机器学习，可能有很多人和我有同感，基本上是从入门到放弃。我自己也思考了原因。主要是恐惧心态，因为数学差，恐惧那些数学公式，而现在又崇尚几十天学会 xxx，这会让人更加焦虑，更不能静下心学习。所以我认为解决问题主要根本也就是调整心态，想象学数学公式就像谈恋爱，从陌生到熟悉，再到走入婚姻的殿堂，不是一蹴而就，罗马不是一天建成的。所以公式一遍看不懂就看两遍，三遍，刻意练习，逃离舒适区。念念不忘，必有回响！

引自：从0开始学大数据
12 | 我们并没有觉得MapReduce 速度慢，直到Spark出现

识别二维码打开原文

「极客时间」 App 

Lambda

写于 2018/11/26

看了一些留言，感觉大家还是”面向工具“学习，对层出不穷的”工具“，感到困惑。但是归根结底，这些工具本身还是计算机科学中很多基础概念的具象化，因此，”面向思想“学习应该是更好的一种做法。先对一种最原始的实现透彻的研究，理解其背后的思想和设计理念，然后再逐步学习后期更为先进的技术，这种学习路径应该更为有效。

> 引自：从0开始学大数据
> 12 | 我们并没有觉得MapReduce 速度慢，直到Spark出现

hunterlodge

写于 2018/11/26

工作中，一个新的方案出现的时候，如果它在某个或某些方面优于当前最好方案，我一定会去思考它的 catch(另一面) 是什么？比如新方案更快，我就大概会看看它的空间使用率、可维护度、全面度。一般都会发现一些问题。生活里也是如此，对表面上只有好处而无需付出或者代价很低的东西永远保持警惕。说白了，世上没有免费的午餐，一些都是权衡利弊的结果

引自：从0开始学大数据
12 | 我们并没有觉得MapReduce 速度慢，直到Spark出现

识别二维码打开原文
「极客时间」 App

老男孩

写于 2018/11/25

惭愧，我遇到的产品经理或者需求人员，基本上分为两类。一类经常说，这是客户的要求必须马上改，用客户来压制研发。一类比较以自我为中心，把自己的观点等同于用户的观点。常常想当然，结果用户一看不是我想要的。结果就是开发人员一次次的从坑里刚爬上来，又被产品一脚踹下去。有几次我真的无法克制，有一种想套麻袋然后一顿打的冲动。🤔 非常赞同老师的观点，不管解决技术问题，还是设计产品都需要深刻的洞察力。想起前面您说的抽象是事物本质的洞察，遇到问题先猫在后面（虽然这种方式比较猥琐），冷静思考，暗中观察，从别人的方案或者错误中总结发现规律，然后顺势而为。

> 引自：从0开始学大数据
> 12 | 我们并没有觉得MapReduce 速度慢，直到Spark出现

识别二维码打开原文
「极客时间」 App

如果你身边也有感到迷茫困惑的朋友，欢迎你点击“请朋友读”，把今天的文章分享给好友。也欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇 16 | ZooKeeper是如何保证数据一致性的?

下一篇 18 | 如何自己开发一个大数据SQL引擎?

精选留言 (24)

写留言

itzzy
2018-12-06

18

文中老师提到是工作一段时间读研的，我自己也有读研的想法，从现在开始准备，也要1-2年才能考的上，读完研出来33，34了，担心没公司要，互联网还是看重年龄的，特别迷茫，希望老师给些建议，感谢！

展开

老男孩
2018-12-08

6

看到这篇文章，我想起了一首歌《you raise me up》。因为我的背景过往不入流。想想周边工作的同事大多都是好学校毕业的，或者学历很高，感觉到很自卑。就怕别人问我是哪个学校毕业的😑，总是纠结于过往，高中三年为什么不好好读书？有时候想放弃，算了吧我这样的不适合混程序员这行业，不如去做一个歌手。2014年几乎半年在家里待着，不想上班也找不到合适的工作。直到看到《大型网站技术架构》那本书，我犹豫了片刻，因...

展开

在路上
2019-01-08

5

我也是非计算机专业出身，不入流学校毕业，然后通过自己努力并不比别人差，和清华北大毕业的一起工作也不觉得比我强呢，哈哈，跟着老师一起加油。

展开

Zach_
2018-12-06

5

因为脑子想的有点多而失眠……
看完python的课程已经1点了……
洗漱完，处理完厨房的事宜已经2点了……
失眠睡不着觉，看完大数据专栏还是睡不着觉，已经3点了……
…

展开

杰之7
2018-12-06

4

通过过去接近一个月的大数据技术学习，真把老师所授的每一篇文章都认真读过，也写了一些阅读记录，今天我先用自己的话梳理一下自己对大数据架构结构的理解。首先因为有海量数据储存的需求，Google发表的大数据技术的三驾马车的论文，分布式文件系统、分

布式计算框架、分布式数据库。然后人们开发了相关的产品。在储存上开发了HDFS、Hbase，在分布式计算框架上开发了基于批处理的Mapreduce、spark的计算框架，数...

展开

Zach_

2018-12-06

4

看到我的昵称的时候，还有点小害怕呢！ 加油💪！

展开

追梦小乐

2018-12-09

1

不甘目前重复业务的工作，工作之余一直在看大数据这一块，决心进入这个领悟去看看有着什么不一样的风景！所以在看到与大数据相关的专栏出来毫不犹豫就买了，一路坚持看下来，之前不太理解的一些知识点有了更进一步的理解，我是通过这个专栏第一次认识了老师，感觉老师不但技术深度高，对人生哲学认知这一块也理解很深刻！

展开

风轻扬

2018-12-07

1

同问关于在职考研，烦请解惑，谢谢

展开

往事随风，...

2018-12-06

1

能不能多讲些原理性工作

展开

🐱您的好...

2018-12-06

1

《沉默的大多数》也不错，一次在飞机场买的，坐飞机的过程中看完的，非常好，据说和《我的精神家园》有重复的文章，也可以看一看，开启独立思考新篇章。

展开

纯洁的憎恶
2018-12-06

1

我的硕士方向就是图形学，我当时觉得图形学的成果看起来很好玩，但要做研究都是枯燥复杂的数学公式，我很佩服老师通过兴趣自己学通了计算机和图形学。与老师相比，我可能要算是比较入流大学的不入流的学生，工作已经好几年了，工作内容早已和计算机无关。看着当初一起在实验室通宵的同学们，一个一个都在财富自由的道路上高歌猛进时，我既羡慕又彷徨。不知是不是自己的选择出来问题。

展开

linazi
2018-12-06

1

跟着老师 听完 读完 思考完
终于把之前的大数据碎片概念可以拉成一条线了
计算 存储 资源管理
围绕着它们衍生出各种各样的产品 技术 框架
不再杂乱了

展开

周小桥
2019-05-24

木子的话很受用。

展开

路平
2019-05-19

《企业应用架构模式》，已下单。

展开

盖饭
2019-03-26

老师这期答疑，恰如其分的为快要看不下去的人打了个鸡血。

展开

cwx0220
2019-03-04

我只愿蓬勃生活在此时此刻，无所谓去哪，无所谓见谁。那些我将要去的地方，都是我从未谋面的故乡。以前是以前，现在是现在。我不能选择怎么生，怎么死；但我能决定怎么爱，怎么活。终极一生，到最后不因虚度光阴而悔恨，不因碌碌无为而遗憾。有人和我一样是从帅张那过来的么？😀

live

2019-01-28

李老师，请教下，现在比较火的阿里相互宝，目前有2000多万的用户，每个月需要离线批处理代扣，这样的业务适合用大数据技术，还是关系型数据库？请解惑谢谢

展开

> 作者回复: 关系数据库和大数据都需要，大数据平台的数据来自关系数据库。
> 在线业务，包括离线的批处理代扣，都有事务要求，用关系数据库。而对这些数据的分析和挖掘在大数据平台进行。详细参考专栏第四个模块。

小美

2018-12-10

老师好，我作为大数据外行，想了解下 HBase 和 HDFS 有啥区别呢，都是做数据存储

REAL_MADIR...

2018-12-10

老师对人生的理解跟我很像，曾经我也想着人生就要干大事，可是慢慢的就被生活一次次的打回原形，渐渐的接受了自己是个普通人的事实。可是那有怎么样呢，只要我每天都在学习，每天都有进步，我的每一天都生活的有意义，那我连起来的人生也就有了意义，况且我还年轻，要走的路还很长，说不准什么时候，积累的量变就转化成了质变，这也许就是我活着的意义。

展开

足迹

2018-12-09

老师，现在的大数据技术大多都基于HDFS的存储，有没有哪种技术可以取代HDFS的？比如有人说KUDU是第一个动了HDFS奶酪的，你怎么看？

作者回复: 还有alluxio，也就是tachyon，从spark分离出来的内存存储，挑战者会不断出现，但是代替hdfs的路肯定会很漫长