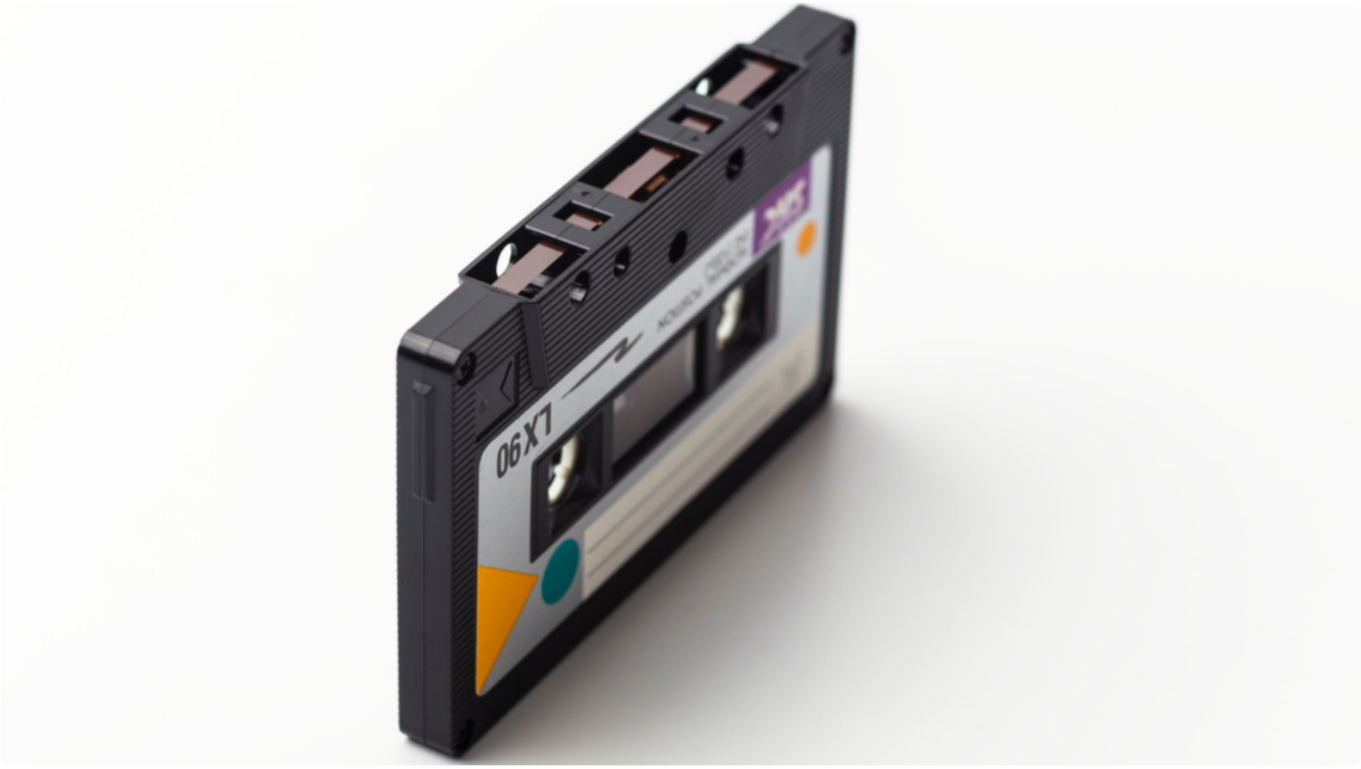
预习 01 | 大数据技术发展史：大数据的前世今生

2018-10-30 李智慧

从0开始学大数据 进入课程 >



讲述：李智慧
时长 10:54 大小 5.00M

在正式落地谈技术之前，我先花一些篇幅给你讲讲大数据技术的发展史，因为这对于你理解技术来说至关重要。

从我的角度而言，不管是学习某门技术，还是讨论某个事情，最好的方式一定不是一头扎到具体细节里，而是应该从时空的角度先了解它的来龙去脉，以及它为什么会演进成为现在的状态。当你深刻理解了这些前因后果之后，再去看现状，就会明朗很多，也能更直接地看到现状背后的本质。说实话，这对于我们理解技术、学习技术而言，同等重要。

今天我们常说的大数据技术，其实起源于 Google 在 2004 年前后发表的三篇论文，也就是我们经常听到的“三驾马车”，分别是分布式文件系统 GFS、大数据分布式计算框架 MapReduce 和 NoSQL 数据库系统 BigTable。

你知道，搜索引擎主要就做两件事情，一个是网页抓取，一个是索引构建，而在这个过程中，有大量的数据需要存储和计算。这“三驾马车”其实就是用来解决这个问题的，你从介绍中也能看出来，一个文件系统、一个计算框架、一个数据库系统。

现在你听到分布式、大数据之类的词，肯定一点儿也不陌生。但你要知道，在 2004 年那会儿，整个互联网还处于懵懂时代，Google 发布的论文实在是让业界为之一振，大家恍然大悟，原来还可以这么玩。

因为那个时间段，大多数公司的关注点其实还是聚焦在单机上，在思考如何提升单机的性能，寻找更贵更好的服务器。而 Google 的思路是部署一个大规模的服务器集群，通过分布式的方式将海量数据存储在这个集群上，然后利用集群上的所有机器进行数据计算。 这样，Google 其实不需要买很多很贵的服务器，它只要把这些普通的机器组织到一起，就非常厉害了。

当时的天才程序员，也是 Lucene 开源项目的创始人 Doug Cutting 正在开发开源搜索引擎 Nutch，阅读了 Google 的论文后，他非常兴奋，紧接着就根据论文原理初步实现了类似 GFS 和 MapReduce 的功能。

两年后的 2006 年，Doug Cutting 将这些大数据相关的功能从 Nutch 中分离了出来，然后启动了一个独立的项目专门开发维护大数据技术，这就是后来赫赫有名的 Hadoop，主要包括 Hadoop 分布式文件系统 HDFS 和大数据计算引擎 MapReduce。

当我们回顾软件开发的历史，包括我们自己开发的软件，你会发现，有的软件在开发出来以后无人问津或者寥寥数人使用，这样的软件其实在所有开发出来的软件中占大多数。而有的软件则可能会开创一个行业，每年创造数百亿美元的价值，创造百万计的就业岗位，这些软件曾经是 Windows、Linux、Java，而现在这个名单要加上 Hadoop 的名字。

如果有时间，你可以简单浏览下 Hadoop 的代码，这个纯用 Java 编写的软件其实并没有什么高深的技术难点，使用的也都是一些最基础的编程技巧，也没有什么出奇之处，但是它却给社会带来巨大的影响，甚至带动一场深刻的科技革命，推动了人工智能的发展与进步。

我觉得，我们在做软件开发的时候，也可以**多思考一下，我们所开发软件的价值点在哪里**？真正需要使用软件实现价值的地方在哪里？你应该关注业务、理解业务，有价值导向，用自己的技术为公司创造真正的价值，进而实现自己的人生价值。而不是整天埋头在需求说明文档里，做一个**没有思考的代码机器人**。

Hadoop 发布之后，Yahoo 很快就用了起来。大概又过了一年到了 2007 年，百度和阿里巴巴也开始使用 Hadoop 进行大数据存储与计算。

2008 年，Hadoop 正式成为 Apache 的顶级项目，后来 Doug Cutting 本人也成为了 Apache 基金会的主席。自此，Hadoop 作为软件开发领域的一颗明星冉冉升起。

同年，专门运营 Hadoop 的商业公司 Cloudera 成立，Hadoop 得到进一步的商业支持。

这个时候，Yahoo 的一些人觉得用 MapReduce 进行大数据编程太麻烦了，于是便开发了 Pig。Pig 是一种脚本语言，使用类 SQL 的语法，开发者可以用 Pig 脚本描述要对大数据集上进行的操作，Pig 经过编译后会生成 MapReduce 程序，然后在 Hadoop 上运行。

编写 Pig 脚本虽然比直接 MapReduce 编程容易，但是依然需要学习新的脚本语法。于是 Facebook 又发布了 Hive。Hive 支持使用 SQL 语法来进行大数据计算，比如说你可以写个 Select 语句进行数据查询，然后 Hive 会把 SQL 语句转化成 MapReduce 的计算程序。

这样，**熟悉数据库的数据分析师和工程师便可以无门槛地使用大数据进行数据分析和处理了**。Hive 出现后极大程度地降低了 Hadoop 的使用难度，迅速得到开发者和企业的追捧。据说，2011 年的时候，Facebook 大数据平台上运行的作业 90% 都来源于 Hive。

随后，众多 Hadoop 周边产品开始出现，**大数据生态体系**逐渐形成，其中包括：专门将关系数据库中的数据导入导出到 Hadoop 平台的 Sqoop；针对大规模日志进行分布式收集、聚合和传输的 Flume；MapReduce 工作流调度引擎 Oozie 等。

在 Hadoop 早期，MapReduce 既是一个执行引擎，又是一个资源调度框架，服务器集群的资源调度管理由 MapReduce 自己完成。但是这样不利于资源复用，也使得 MapReduce 非常臃肿。于是一个新项目启动了，将 MapReduce 执行引擎和资源调度分离开来，这就是 Yarn。**2012 年，Yarn 成为一个独立的项目开始运营，随后被各类大数据产品支持，成为大数据平台上最主流的资源调度系统。**

同样是在 2012 年，UC 伯克利 AMP 实验室（Algorithms、Machine 和 People 的缩写）开发的 Spark 开始崭露头角。当时 AMP 实验室的马铁博士发现使用 MapReduce 进行机器学习计算的时候性能非常差，因为机器学习算法通常需要进行很多次的迭代计算，而 MapReduce 每执行一次 Map 和 Reduce 计算都需要重新启动一次作业，带来大量的无谓消耗。还有一点就是 MapReduce 主要使用磁盘作为存储介质，而 2012 年的时候，内存

已经突破容量和成本限制，成为数据运行过程中主要的存储介质。Spark 一经推出，立即受到业界的追捧，并逐步替代 MapReduce 在企业应用中的地位。

一般说来，像 MapReduce、Spark 这类计算框架处理的业务场景都被称作**批处理计算**，因为它们通常针对以“天”为单位产生的数据进行一次计算，然后得到需要的结果，这中间计算需要花费的时间大概是几十分钟甚至更长的时间。因为计算的数据是非在线得到的实时数据，而是历史数据，所以这类计算也被称为**大数据离线计算**。

而在大数据领域，还有另外一类应用场景，它们需要对实时产生的大量数据进行即时计算，比如对于遍布城市的监控摄像头进行人脸识别和嫌犯追踪。这类计算称为**大数据流计算**，相应地，有 Storm、Flink、Spark Streaming 等流计算框架来满足此类大数据应用的场景。流式计算要处理的数据是实时在线产生的数据，所以这类计算也被称为**大数据实时计算**。

在典型的大数据的业务场景下，数据业务最通用的做法是，采用批处理的技术处理历史全量数据，采用流式计算处理实时新增数据。而像 Flink 这样的计算引擎，可以同时支持流式计算和批处理计算。

除了大数据批处理和流处理，NoSQL 系统处理的主要也是大规模海量数据的存储与访问，所以也被归为大数据技术。 NoSQL 曾经在 2011 年左右非常火爆，涌现出 HBase、Cassandra 等许多优秀的产品，其中 HBase 是从 Hadoop 中分离出来的、基于 HDFS 的 NoSQL 系统。

我们回顾软件发展的历史会发现，差不多类似功能的软件，它们出现的时间都非常接近，比如 Linux 和 Windows 都是在 90 年代初出现，Java 开发中的各类 MVC 框架也基本都是同期出现，Android 和 iOS 也是前脚后脚问世。2011 年前后，各种 NoSQL 数据库也是层出不穷，我也是在那个时候参与开发了阿里巴巴自己的 NoSQL 系统。

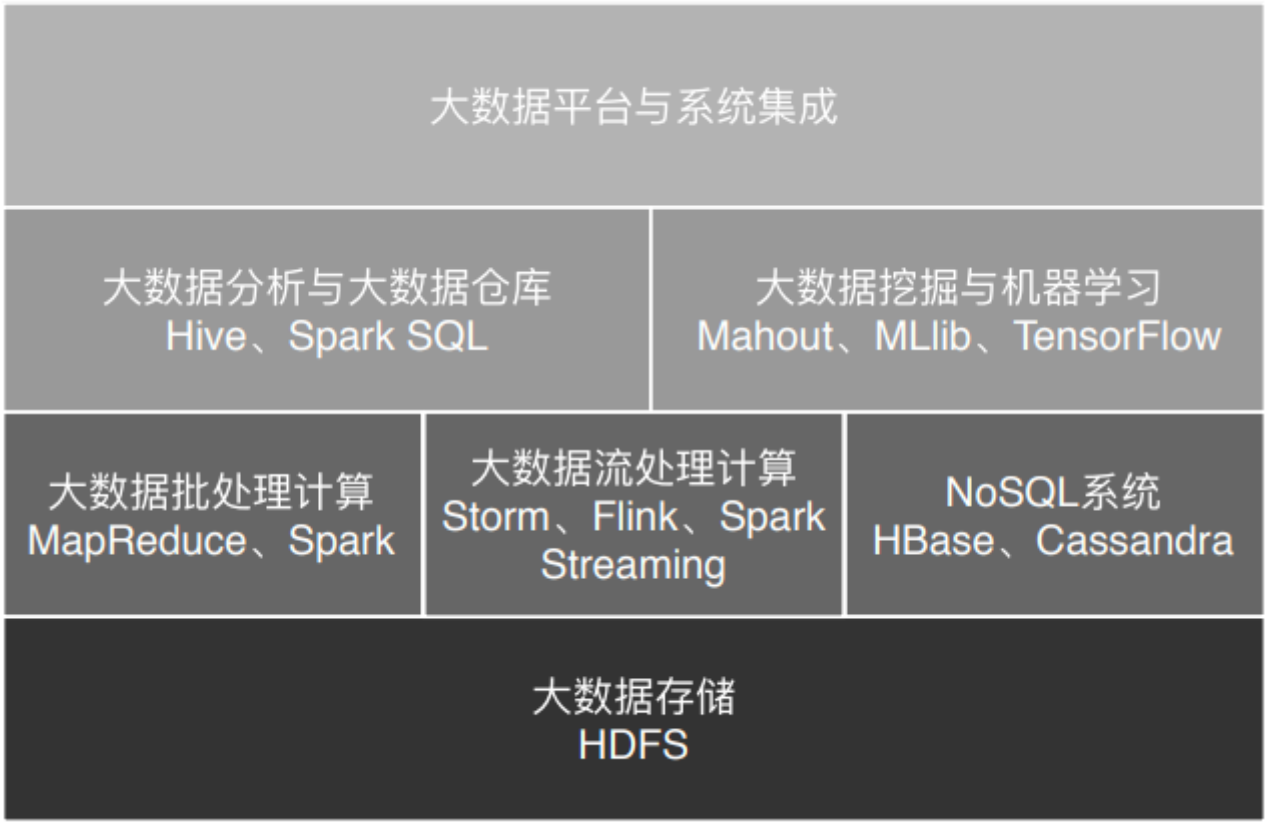
事物发展有自己的潮流和规律，当你身处潮流之中的时候，要紧紧抓住潮流的机会，想办法脱颖而出，即使没有成功，也会更加洞悉时代的脉搏，收获珍贵的知识和经验。而如果潮流已经退去，这个时候再去往这个方向上努力，只会收获迷茫与压抑，对时代、对自己都没有什么帮助。

但是时代的浪潮犹如海滩上的浪花，总是一浪接着一浪，只要你站在海边，身处这个行业之中，下一个浪潮很快又会到来。你需要敏感而又深刻地去观察，略去那些浮躁的泡沫，抓住真正潮流的机会，奋力一搏，不管成败，都不会遗憾。

正所谓在历史前进的逻辑中前进，在时代发展的潮流中发展。通俗的说，就是要在风口中飞翔。

上面我讲的这些基本上都可以归类为大数据引擎或者大数据框架。而**大数据处理的主要应用场景包括数据分析、数据挖掘与机器学习。**数据分析主要使用 Hive、Spark SQL 等 SQL 引擎完成；数据挖掘与机器学习则有专门的机器学习框架 TensorFlow、Mahout 以及 MLlib 等，内置了主要的机器学习和数据挖掘算法。

此外，大数据要存入分布式文件系统（HDFS），要有序调度 MapReduce 和 Spark 作业执行，并能把执行结果写入到各个应用系统的数据库中，还需要有一个**大数据平台**整合所有这些大数据组件和企业应用系统。



图中的所有这些框架、平台以及相关的算法共同构成了大数据的技术体系，我将会在专栏后面逐个分析，帮你能够对大数据技术原理和应用算法构建起完整的知识体系，进可以专职从事大数据开发，退可以在自己的应用开发中更好地和大数据集成，掌控自己的项目。

思考题

你从大数据生态的发展史中，能得出什么样的结论？又有怎样的思考？

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　开篇词 | 为什么说每个软件工程师都应该懂大数据技术?

下一篇　预习 02 | 大数据应用发展史：从搜索引擎到人工智能

精选留言 (102)

写留言

江
2018-10-31

167

1.论文奠定技术发展基石;
2.业务催生技术不断突破;
3.效率倒逼技术迭代更新;

展开

作者回复: 总结精炼，赞

今天也是爆...
2018-10-30

62

只要你站在海边，身处这个行业之中，下一个浪潮很快又会到来。你需要敏感而又深刻地去观察，略去那些浮躁的泡沫，抓住真正潮流的机会，奋力一搏，不管成败，都不会遗憾。

这句话真好

展开

hua168
2018-10-30

16

大神，从0学习大数据需要哪些基础呀？后面能分享一下大数据入门的顺序和相关书籍吗，好让我们这些菜鸟能有个系统的学习

作者回复: 这个专栏就是从零学习大数据，而且很系统，希望你坚持下来。

猫头鹰爱拿...
2018-10-30

13

昨天刚订阅专栏 今天就得知消息 公司要上大数据项目了 我是项目组成员 好开心 真的好巧啊 感谢下订阅专栏带来的运气 同时也要好好学习 哈哈

暴风雪
2018-11-03

10

看了两篇专栏，感觉作者用文字描述的效果，胜过大多数视频教程

展开

作者回复: 谢谢

公号-代码...
2018-10-30

5

大数据生态的发展还是遵循着不断演进的过程，出现新问题、解决新问题、更加容易的解决新问题、然后再次出现新问题，以此不断螺旋上升。
大数据技术本身就是为业务而生，而不是脱离业务而产生的新技术；正是由于将业务刻在

大数据本身的基因里面，所以很多大厂商都对大数据的发展、应用、推广、普及起到了很重要的促进作用。...

展开

hashmap

2018-11-09

3

如果潮流已经退去，这个时候再去往这个方向上努力，只会收获迷忙于压抑
真的是这样，有感触，
要站在时代的前列腺，顺势而为

展开

y y

2018-11-15

2

怎么没看见zookeeper啊

展开

作者回复: 下个模块，敬请期待

龙华强

2018-10-30

2

时代的发展就是科技的发展，我们紧跟科技发展的潮流，现在和将来，都不会迷茫，我们不是码农，我们是时代的开创者和见证者

Droices

2018-10-30

2

读过李老师的大型网络技术架构，在公众号看到了这个就直接买了。

展开

韩程

2019-04-19

1

老师你好，能不能系统的讲解一下数据仓库和大数据有什么区别和联系呢？以及应用场景有什么不同。

作者回复: 数据仓库是解决数据问题的方案和方法，大数据是具体实现技术。大数据和关系数据库都能实现数据仓库。

eldon 1
2019-01-16

我们要顺应潮流，也要众争勿往。

展开

chenssy 1
2019-01-08

从 0 开始学习大数据，现在正好在数据平台组，从 0 开始搭建大数据平台，希望跟着这个专栏一起成长

Luckiness 1
2018-12-14

有没有什么关于大数据学习技巧或者方式方法提高我们学习大数据的途径，让我们少走弯路，提高效率的学习？只需要按照智慧哥的脚步就能学好吗？

CHEN 1
2018-11-08

从大数据的发展史想到
一是思维方式的重要性，在多数企业在提高单机性能与更高更大更全的大型服务器死磕时，谷歌的思路是部署分布式服务器集群，少花钱还多办事。有人固步自封满足与现有知识，有人则紧跟时代脉搏不断学习前进，比如来定智慧老师的大数据专栏😁
二是hadoop用纯java语言编写，没什么技术难点，但是它价值巨大。我们许多码农总是...

展开

作者回复: 👍

小太白ding... 1
2018-11-06

大数据很早就有行业和学科一直在研究和处理，比如金融和天文学。只是这一波浪潮涌来，被推到了浪尖。应用层面，人工智能和大数据交叉甚多不分彼此，很多机构和媒体把二者完全分离，叫人唏嘘。初学者云雾缭绕，幸有极客邦专栏，拨乱反正，指明方向。谢谢！

作者回复: 谢谢你

JIY 1
2018-11-05

每个时代在进行的过程中都会遇到自己的瓶颈 在冲破瓶颈的时候 大量的技术会涌现出来。在海边风来了 就肯定有一批海鸥还迎着风飞翔，剩下的一批会去躲避风浪。

展开

---

aspire 1
2018-10-30

数据为信息基础原，大数据为大规模数据信息，而如何将大规模数据信息进行处理则是关键事宜。公司一般会根据业务场景高效的的计算处理大规模数据信息，但各公司业务逻辑又不同，所以应有一套基础技术数据处理框架～

展开

sophie
2019-05-27

大数据平台和系统集成没有看到有相关的技术工具推荐，是这块还没有主流成熟应用软件还是因为与业务结合所以需要自行设计和实现呢？

作者回复: 请参考
29 | 盘点可供中小企业参考的商业大数据平台

Eazow
2019-05-25

请问大数据技术都是以hdfs为存储基础的吗？

展开

作者回复: 都支持HDFS，但不是非HDFS不可，各种非HDFS存储方案也得到越来越多的支持