链滴

大数据学习 PPT，讲述大数据基础的概念，Hadoop 生态，常用技术。(PPT+ 讲义)

作者：sq8852161

原文链接：https://ld246.com/article/1568271036094

来源网站：链滴

<blockquote>
本文为 PPT 的讲义，请配合 PPT 使用。
因为是对公司同事的大数据科普课程，因此基础思想讲的比较多。技术详情，PPT 中写的比较清楚。
PPT 在此云盘连接中。
链接：大数据学习 PPT　百云盘提取码：ll08
因为是讲义，所以行文比较口语化，大家见谅。
</blockquote>
<h3 id="起始语-">起始语：</h3>
“大家好，今天很荣幸和大家一起来了解一下大数据，让咱们对现在很火热的大数据有一个大概了解。我们今天会讲到大数据的来源，大数据的概念，大数据的相关核心技术与组件，以及大数据的用等。”

首选我们看大数据的来源。
说起大数据，那么就要先说说数据。数据的本质是什么呢？

4：念
在十几年前，可能大家数据还没有现在这么看重，有些时候对一些数据，比如系统日志，操作日志，有一些机器的参数记录，一些视频记录，都认为是杂乱的，没有太大价值的“副产物”，没有认识到据的真正价值，很多时候都是一删了之。但是现在，数据已经被放到了非常重要的战略地位。数据现被认为是最重要的资产和资源之一。国家现在也对数据资产很重视。

5：念。
另一个现实是数据在不断地爆炸性增长。大家自己也能直观的感受到。比如说手机的存储。从几年前的 8 个 G12 个 G，到现在的 64G128G。但是依然很快就不够用了。自己出去玩一次能拍几百多张照片 APP 大小从几 M 到几百 M。更别说我们使用 APP 产生的海量数据了。

6：
但是这就导致了一个新的问题。
数据资产的概念被人们接受了。大家都了解到管理数据资产的重要性。
如何理解呢，曾经，关系型数据库是万能的。我们做任何系统都会去使用关系型数据库。但是现在单的关系型数据库，不管是单机还是集群，已经无法满足现有需求了。看 PPT

7：
现有的数据资产管理的挑战主要来源于哪里。
看 PPT

那对于新时代的 数据资产，对于现有的大数据管理现状，我们都有哪些新的需求呢？

8：
首先是看数据的方式要不同，这个看，怎么理解呢。这个看不仅仅是指 前端看报表，看页面，主要指是看待数据的方式，我们以前看数据，会直观的把数据结构化，关系化。认为有序的，符合范式的数才是好数据，才是可以利用的数据，我们去找数据，利用数据，只看到了这部分数据，却没有看到冰下更多的数据。现在我们要转变看数据的观念，看到隐藏在海面下的冰山。这部分数据是什么呢?PPT

9:
其次，我们需要更高性价比的计算与存储方式。物理上，我们现在单机存储，是很昂贵的。而且是越，越昂贵。计算能力，如果是单纯依靠堆服务器，那成本更加不可想象。数据库层面，超过一定数据后，单机数据库就无法使用了，数据库集群的成本和系统复杂度又过于高昂，现在暴炸性的数据和计，我们需要更加廉价，更高效的处理方式。

10：
再然后，以前的数据管理策略都是基于结构化数据的，在遇到不同的数据结构式，显然已经无法处理。再有，现有的处理逻辑和系统架构，业无法适应大数据时代的需求。想要扩充，只能 scale-up（扩），scale-out（分体扩展。） 不易；scale Up(纵向扩展) 主要是利用现有的存储系统，通过不断增存储容量来满足数据增长的需求。Scale-out 横向扩展架构的升级通常是以节点为单位，每个节点往将包含容量、处理能力和 I / O 带宽。一个节点被添加到存储系统，系统中的三种资源将同时升级。<r>
PPT。

11：
再有，从以下四个方面 PPT。这四方面都提出了巨大的考验，这些已经超出了现有企业 It 能独自解决能量范围了。我们需要一种新的，适应爆炸性的数据增长，能解决我们之前提出的问题的解决方案。

12：
再从政策层面说，现在的中央政府对大数据很支持，PPT。
郑州已经被设立为八个国家大数据综合试验区之一。
郑州是地级市中唯一设立大数据管理局的。其他都为政务与大数据管理局。
从这些国家领导人的话和各项落实的政策里，可以看到，大数据时代已经到来。我们也需要新的技术解决我们遇到的数据问题，而这个选择就是大数据。

13：
我们接下来来了解一下大数据的概念。
什么是大数据?
以及大数据技术所带来的的思维模式与之前有什么不同？有什么特点，又能带来怎样的变化。

14：
首选，什么是数据？
PPT
从数据结构上来说，有结构化，半结构化，非结构化。
从感受上来说，万物皆为数据，我们所能看到的，所能感受到的，甚至无法感受的，都可以说是数据能看到的，结构化的文档，非结构化的视频，等等。

15：
那什么是大数据呢？
这个大作何解释，首选大，是体量上的大。PPT
1K 就已经是 2 的十次方 bit 了。一个 bit 是 一个 0 或 1。
1M 是 2 的 20 次方，1G 是 2 的 30 次方，1T 是 2 的 40 次方，1PB 是 2 的 50 次方
银河系星球数量是 4*10 的 11 次方。

16：
数据是如此之多，以至于，已经没有办法在可容忍的时间下使用常规软件方法完成存储、管理和处理务。
从 PPT 可以看到，
2010 年产生的新数据就可以抵得上 52000 个美国国会图书馆。那么到如今 2018 年呢。每年的数据长量都是百分之几十的递增。如今的数据存储量更是暴涨。

17：
综合以上所述的种种情况，我们来看一下各个机构对大数据的定义是什么？
PPT

18：
我们综合这些定义，可以得出一个什么结论呢？我们看 PPT，念定义。
这个图展示了系统和数据量级，以及数据的复杂性增长。
ERP 业务是最复杂的，但是数据结构是最清洗，数据量是最小的。
最后的日志之类的，甚至没有业务逻辑，但是数据是最复杂的，量是最大的。

19：
定义了大数据，我们来看一下大数据的 4v 特性。
PPT
有人呢，把大数据 4V 特性，扩展为了十个字。我们接下来再讲

20：
首选讲讲体量特性，就是指大数据体量极大。这个大家从之前的 PPT 都能看到。大数据首先要解决就是数据体量大的问题。

21：
速度特性，一方面是指数据增长的速度极快，另一方面讲是说数据处理的速度极快

22：
年 PPT

23：
价值密度，大数据并不是说数据量多了，数据价值就更高。可能数据量增大后，数据平均价值是下降。也就是说数据的价值密度比较低。

24:
以上四个特性是外国人总结的，下面我们讲讲中国人总结的，这是华为的大数据专家傅一航总结的，数据的十字特性。PPT

25：
念 PPT

26：
念 PPT

27：
念 PPT

28：
念 PPT
这十个字，是对大数据 4v 特性的深化。

29：
大数据时代，大数据技术的到来，不光带来了技术上，数据存储上的革新。同时也带来了新的思维模
来处理数据，主要是从以下三个方面来。PPT

30：
如何理解呢，我们看 PPT，如何理解更多。？
以前，我们无法收集尽可能多的数据，只能尽量收集关键数据，也无法存储所有的数据，没有能力去
算所有的数据。所以，我们进行的计算，分析都是基于样本数据的。但是现在不一样了，我们能够获
全部数据，可以用比较廉价的算力和存储来进行数据的收集和分析，我们就可以对，所有的数据进行
析。以 PPT 所示的人口调查的例子来说。

31：
如何理解更杂?
我们以前整理数据，获取数据都是追求越精确越好，数据越干净越纯净越好。这一方面是因为我们人
对于真理，对于精准的追求，另一方面，也是迫于现实情况。我们只能从混乱中提取出精准，才能处
。我们有限的资源只能处理精准的数据。但是大数据技术，大数据时代改变了这个形式。根据 4v 特
，大数据的价值密度是很低的。体量是很大的。必然，这些混杂数据质量没有精准数据高。但是，这
是缺陷，而是另一种价值。我们从中可以获取更多的信息，更多的价值。看 ppt

32：
如何理解更好，也就是因果关系与相关关系呢。
我们看 PPT
因果很好理解，从 A 推出 B，从 B 推出 C。那什么是相关呢。我们同属于一个公司，我们是相关的
我是男性，所有男性和我都是相关的。所以，大数据时代，分析问题的逻辑不再是因为 A，所以 B。
是有无数个相关条件，那么得出一个结论。这个结论只是可能性。数据越多，样本越多，相关性越多
这个可能性越高。但是，我们不能说，因为我是男的，所以我就要怎么样怎么样。所以，这是相关，
不是因果。

33：
好，我们都看了大数据的来源，大数据概念，对大数据的一些特性和一些思维上的改变。那么我们接
来看一下大数据相关的技术。
我们会看一下，大数据的生态，相关大数据技术的主流厂商，从六个方向分析大数据所用的技术。一
组件。

34：
首先，我们来看大数据核心相关信息。

我们要明确，广义上，大数据是一种概念，一种理论。同时，在狭义上，大家也习惯将大数据指为一技术，也就是以 hadoop 为核心的生态。
那么我们来看看这个生态系统。PPT

35：
我们看一下大数据生态的整体架构，我们从五个方面说。
数据采集，就是获取数据，包括从数据库，日志，摄像头，一些数据流信息，甚至物联网设备，这些是数据源。
。。。。。。
PPT
可以看到，大数据的整个生态，是从实，到虚，又到实。是一个闭环。对，从用户的角度来说，看起像是一个黑盒系统，和之前的关系型数据库，传统的处理，好像没有什么不同。但是实际上，我们做比以前要多的多。

36：
接下来，我们看一下，现在大数据生态中，比较主流的大数据软件集成商是那些。
主要有以下三家：PPT
我们使用的是 Cloudera，点击看一下。

37：
我们看一下三家主流厂商的系统对比。
今年，Hortonworks 的 Ambari 系统。 和 Cloudera 合并了。。。
所以，以后 ambari 和 clouderaManager 可能就是一个了。。

38：
接下来，我们看一下，我们公司现有的大数据平台的架构。
我们采用了 Cloudera 公司的方案，使用了 ClouderaManager 平台。
大体架构如下：
PPT

39：
接下来，我们从大数据获取的方向来看一下大数据的组件
第一个是 sqoop，sqoop 是类似于 ETL 的一个工具组件。
PPT

40：
接下来，我们看一下另一个数据获取组件 Flume，水槽。。
它是由 Cloudra 贡献的。

41：
接下来，我们来看一个 hadoop 里非常重要的组件 HDFS，谷歌他们用的是 GFS 系统。
PPT
可以理解为 hadoop 是一个架子，你往里放什么砖，就会做出什么功能的建筑。

42：
念 PPT

43：
念 PPT

44：
关于无法存储小文件，HDFS 对大文件处理是很优秀的。但是大量小文件会导致运行效率变差。

45：
文件是被拆开，分别存储在不同的服务器硬盘上的。

46：
namenode 存的是目录，datanode 存的是文件。

47：
namenode 负责什么呢？

**中间部分 PPT 都为技术详细情况介绍，可以直接看 PPT

92：
京东的仓储体系，早几年前就使用了大数据的分析。会分析某个地区，将来可能会购买的物品，将其前调配到最近的仓库。以达到快速配送的目的。所以京东敢喊次日达。当日达。

93：
阿里的智慧工厂系统，通过大数据 +AI，曾经给某个光伏长提高了 1% 的生产效率。提高了一个多亿效益。

