大数据技术之大数据概论

版本：V3.3

第 1 章 大数据概念

大数据概念

大数据（Big Data）：指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的**信息资产**。

大数据主要解决，海量数据的采集、存储和分析计算问题。

按顺序给出数据存储单位：bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。

1Byte = 8bit 1K = 1024Byte 1MB = 1024K

1G = 1024M 1T = 1024G 1P = 1024T

第 2 章 大数据特点（4V）

大数据特点

1、Volume（大量）

截至目前，人类生产的所有印刷材料的数据量是200PB，而历史上全人类总共说过的话的数据量大约是5EB。当前，典型个人计算机硬盘的容量为TB量级，而 一些大企业的数据量已经接近EB量级。

大数据特点

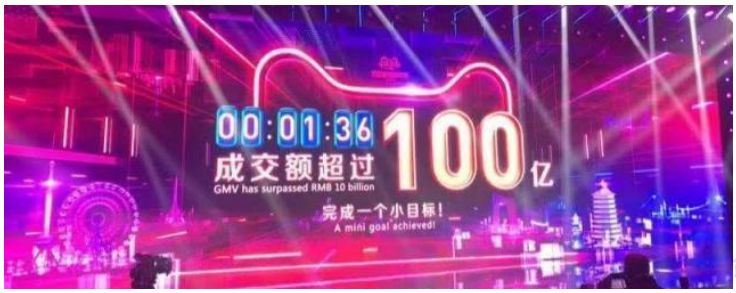
尚硅谷

2、Velocity（高速）

这是大数据区分于传统数据挖掘的最显著特征。根据IDC的“数字宇宙”的报告，预计到2025年，全球数据使用量将达到163ZB。在如此海量的数据面前，处理数据的效率就是企业的生命。

天猫双十一：2017年3分01秒，天猫交易额超过100亿

2020年96秒，天猫交易额超过100亿



大数据特点

尚硅谷

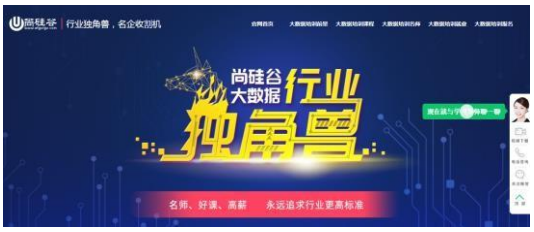
3、Variety（多样）

这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以数据库/文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

订单数据

id	用户	日期	购买商品	购买数量
1001	canglaoshi	20200710-9:10:10	面膜	2
1002	xiaozelaoshi	20200710-9:11:20	化妆品	3
1003	boduolaoshi	20200710-9:22:50	内衣	4
1004	sslaoshi	20200710-10:12:20	海狗人参丸	100

网络日志



让天下没有难学的技术

4、Value（低价值密度）

价值密度的高低与数据总量的大小成反比。

比如，在一天监控视频中，我们只关心宋宋老师晚上在床上健身那一分钟，如何快速对有价值数据“提纯”成为目前大数据背景下待解决的难题。

第 3 章 大数据应用场景

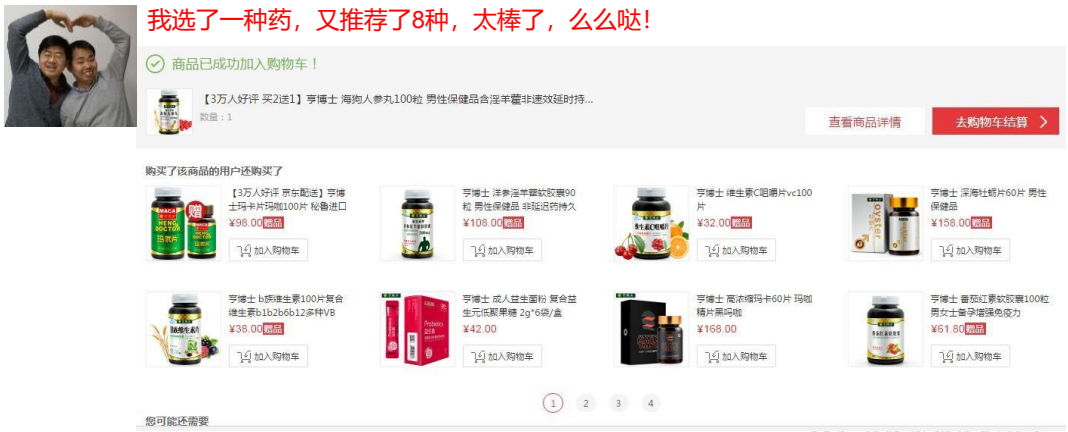
1、抖音：推荐的都是你喜欢的视频

我抖音里面的视频

ss抖音里面的视频

大数据应用场景

2、电商站内广告推荐：给用户推荐可能喜欢的商品

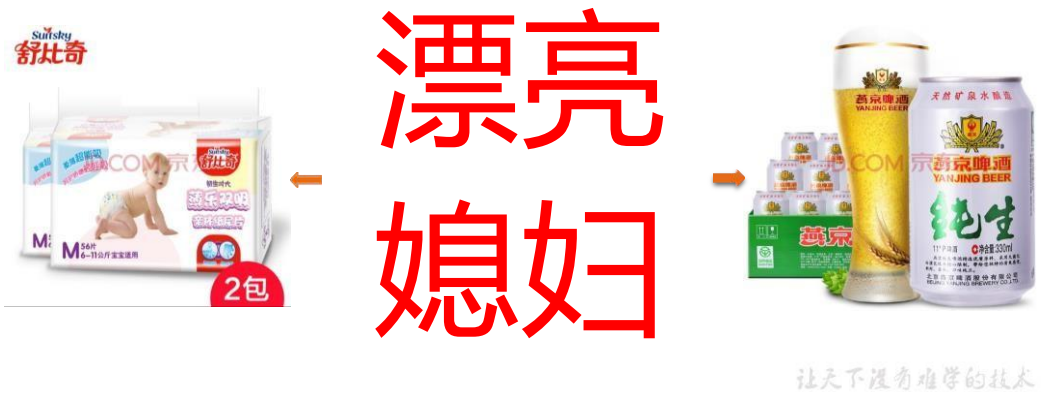


大数据应用场景

尚硅谷

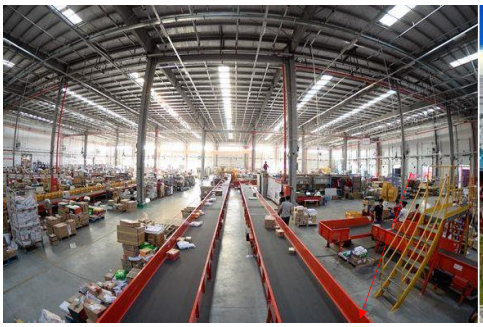
3、零售：分析用户消费习惯，为用户购买商品提供方便，从而提升商品销量。

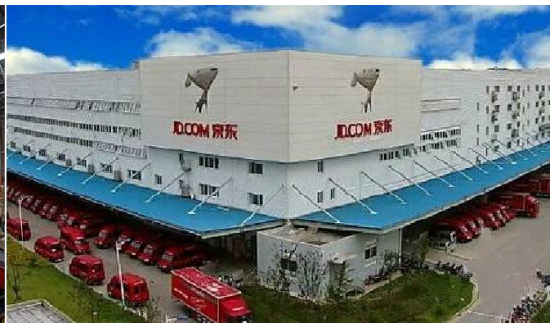
经典案例，纸尿布+啤酒。



大数据应用场景

4、物流仓储：京东物流，上午下单下午送达、下午下单次日上午送达





大数据应用场景

5、保险：海量数据挖掘及风险预测，助力保险行业精准营销，提升精细化定价能力。

6、金融：多维度体现用户特征，帮助金融机构推荐优质客户，防范欺诈风险。

7、房产：大数据全面助力房地产行业，打造精准投策与营销，选出更合适的地，建造更合适的楼，卖给更合适的人。

8、人工智能+ 5G + 物联网 + 虚拟与现实





第 4 章 大数据发展前景

1、党的十九大提出“推动互联网、大数据、人工智能和实体经济深度融合”。

2、2020年初，中央推出34万亿“新基建”投资计划

"新基建"投资规模拆分	
项目	2020年投资规模（亿元）
5G	3000
特高压	600
轨道交通	5000
充电桩	100
数据中心	1000
人工智能	350
工业互联网	100
合计	10150

大数据发展前景

3、下一个风口

2020年是**5G**的元年，国家在大力铺设5G设备，2021年就是5G手机应用的开始，也是大数据要爆发的1年。5G带来的是每秒钟10g的数据，会给每家公司都带来海量的数据。那么传统的Java工具根本解决不了海量数据的存储。就更不用说海量数据的计算了。如果你对5G的感触不够深，可以回忆一下3G和4G的区别。3G时只能打电话、发短信，当时还觉得很好，觉得3G不错。但是4G来了后，大家很少打电话和发短信了，都改为语音、视频、直播、网上购物等生活方式，带火了淘宝、京东、美团、字节跳动等企业。没有跟上节奏的百度，有点摇摇欲坠。

自古不变的真理：先入行者吃肉，后入行者喝汤，最后到的买单！

大数据发展前景

尚硅谷

4、人才紧缺、竞争压力小

有句话叫：“选择大于努力”选择一个好的方向，少奋斗十年。是否记得国家在2017年才开设大数据课程，当时是北京大学、人民大学等25所高校开设第一批大数据课程。今年才2021年。也就是今年才毕业，那么像Java、前端大学已经开设多少年了，包括培训班都加在一起，10多年，可想而知目前市场上，Java和前端的人才有多少。

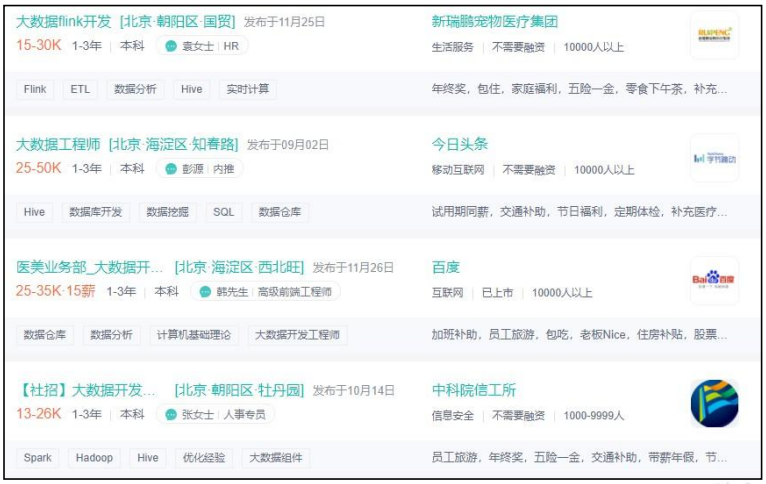
大数据的人才目前除了培训机构培养的，没有真正的科班毕业，而且真正能培养好大数据人才的培训机构又有几个。 所以目前选择大数据是最佳选择。

如果担心自己不是科班，其实也大可不必，因为大学真的学不了啥。只要是能考上本科，说明你不笨，那学大数据就没问题。

让天下没有难学的技术

大数据发展前景

5、Boss直聘网站上的部分大数据工程师薪水如下



第 5 章 大数据部门间业务流程分析

大数据部门间业务流程分析

产品人员提需求（统计双11实时交易额、各地区销售排行TopN等）

数据部门搭建数据平台、分析数据指标

数据可视化（报表展示、邮件发送、大屏幕展示等）

第 6 章 大数据部门内组织结构