

# 尚硅谷大数据技术之大数据概论

(作者：尚硅谷大数据研发部)

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

### 2、Velocity ( 高速 )

这是大数据区别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



让天下没有难学的技术

## 大数据特点

### 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



让天下没有难学的技术

## 2、电商站内广告推荐：给用户推荐可能喜欢的商品



我选了一种药，又推荐了8种，太棒了，么么哒！

商品已成功加入购物车！

【3万人好评 买2送1】 罗博士 海狗人参丸100粒 男性保健品含淫羊藿非速效延时特...

数量：1

查看商品详情 去购物车结算 >

购买了该商品的用户还购买了

 <p>【3万人好评 京东配送】 罗博士 海狗人参丸100粒 包装进口</p> <p>¥98.00 商品</p> <p>加入购物车</p>	 <p>罗博士 洋参温补鹿茸胶囊90粒 男性保健品 非速效药持久</p> <p>¥108.00 商品</p> <p>加入购物车</p>	 <p>罗博士 维生素C咀嚼片vc100片</p> <p>¥32.00 商品</p> <p>加入购物车</p>	 <p>罗博士 深海牡蛎片60片 男性保健品</p> <p>¥158.00 商品</p> <p>加入购物车</p>
 <p>罗博士 b族维生素100片复合维生素b1b2b6b12多种VB8</p> <p>¥38.00 商品</p> <p>加入购物车</p>	 <p>罗博士 成人益生面粉 复合益生元低聚果糖 2g*50g/盒</p> <p>¥42.00</p> <p>加入购物车</p>	 <p>罗博士 高浓缩玛卡60片 玛咖精片黑玛咖</p> <p>¥168.00</p> <p>加入购物车</p>	 <p>罗博士 鲁源红景天胶囊100粒 男女士常备增强免疫力</p> <p>¥61.80 商品</p> <p>加入购物车</p>

您可能还需要

1 2 3 4

让人工智能有难学的技术

## 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，纸尿裤+啤酒。



漂亮媳妇

让天下没有难学的技术



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



让天下没有难学的技术

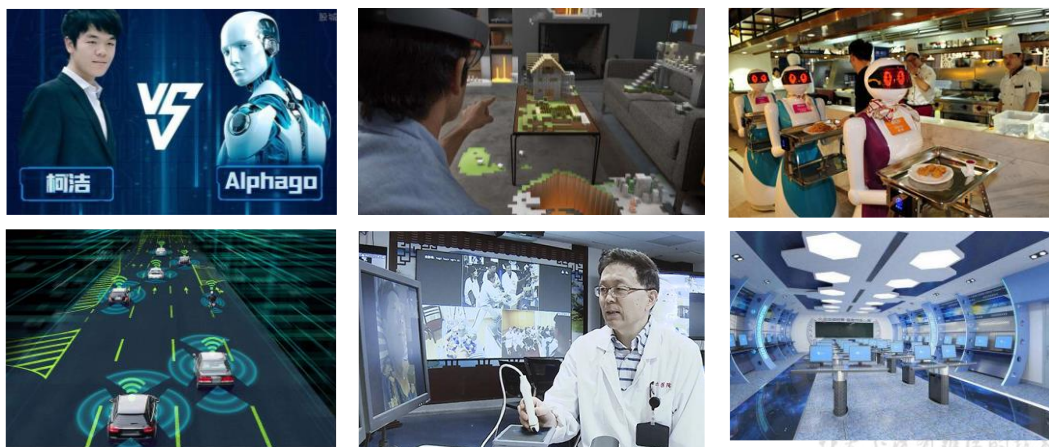
5、保险：海量数据挖掘及风险预测，助力保险行业精准营销，提升精细化定价能力。

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

让天下没有难学的技术

### 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！

让天下没有难学的技术

### 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。

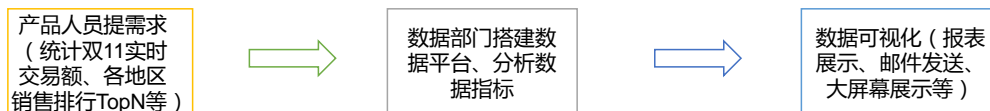
让天下没有难学的技术

### 5、Boss直聘网站上的部分大数据工程师薪水如下

<b>大数据flink开发</b> [北京·朝阳区·国贸] 发布于11月25日 15-30K 1-3年   本科 ● 袁女士   HR Flink ETL 数据分析 Hive 实时计算 年终奖, 包住, 家庭福利, 五险一金, 零食下午茶, 补充...	<b>新瑞鹏宠物医疗集团</b> 生活服务   不需要融资   10000人以上 试用期月薪, 交通补助, 节日福利, 定期体检, 补充医疗...
<b>大数据工程师</b> [北京·海淀区·知春路] 发布于09月02日 25-50K 1-3年   本科 ● 彭源   内推 Hive 数据库开发 数据挖掘 SQL 数据仓库 试用期月薪, 交通补助, 节日福利, 定期体检, 补充医疗...	<b>今日头条</b> 移动互联网   不需要融资   10000人以上 试用期月薪, 交通补助, 节日福利, 定期体检, 补充医疗...
<b>医美业务部_大数据开...</b> [北京·海淀区·西北旺] 发布于11月26日 25-35K 15薪 1-3年   本科 ● 韩先生   高级前端工程师 数据库 数据分析 计算机基础理论 大数据开发工程师 加班补助, 员工旅游, 包吃, 老板Nice, 住房补贴, 股票...	<b>百度</b> 互联网   已上市   10000人以上 加班补助, 员工旅游, 包吃, 老板Nice, 住房补贴, 股票...
<b>【社招】大数据开发...</b> [北京·朝阳区·牡丹园] 发布于10月14日 13-26K 1-3年   本科 ● 张女士   人事专员 Spark Hadoop Hive 优化经验 大数据组件 员工旅游, 年终奖, 五险一金, 交通补助, 带薪年假, 节...	<b>中科院信工所</b> 信息安全   不需要融资   1000-9999人 员工旅游, 年终奖, 五险一金, 交通补助, 带薪年假, 节...

让天下没有难学的技术

## 第5章 大数据部门间业务流程分析



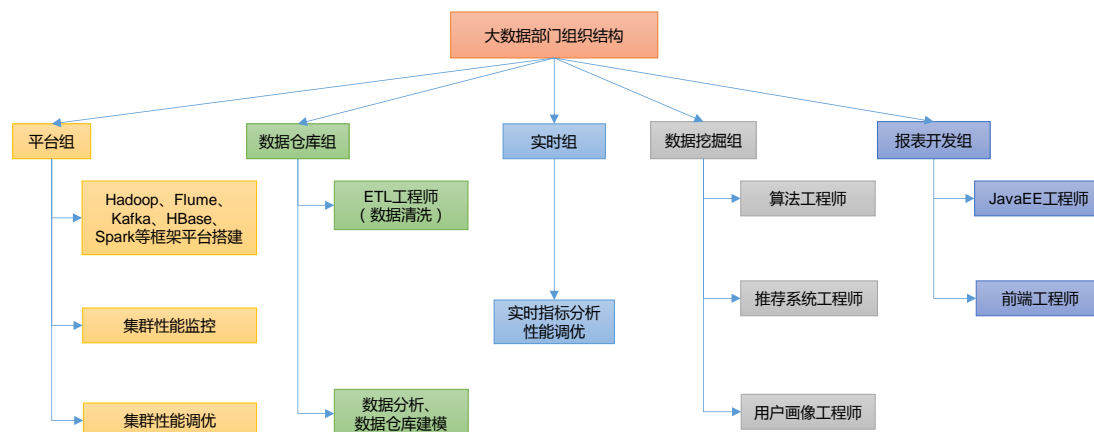
让天下没有难学的技术



## 第 6 章 大数据部门内组织结构



### 大数据部门内组织结构



让天下没有难学的技术