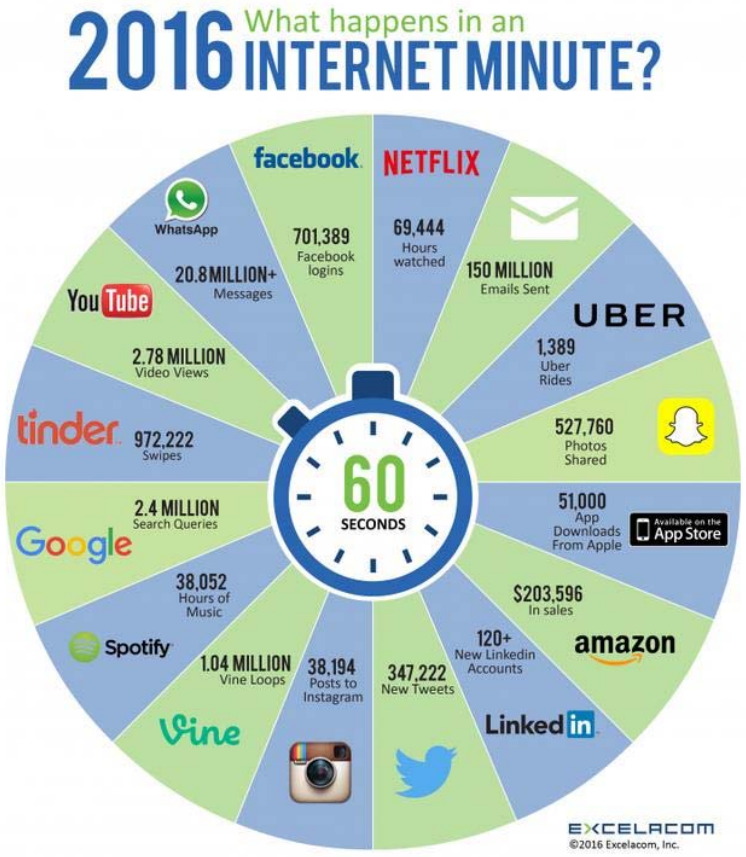
大数据概述

BIG DATA

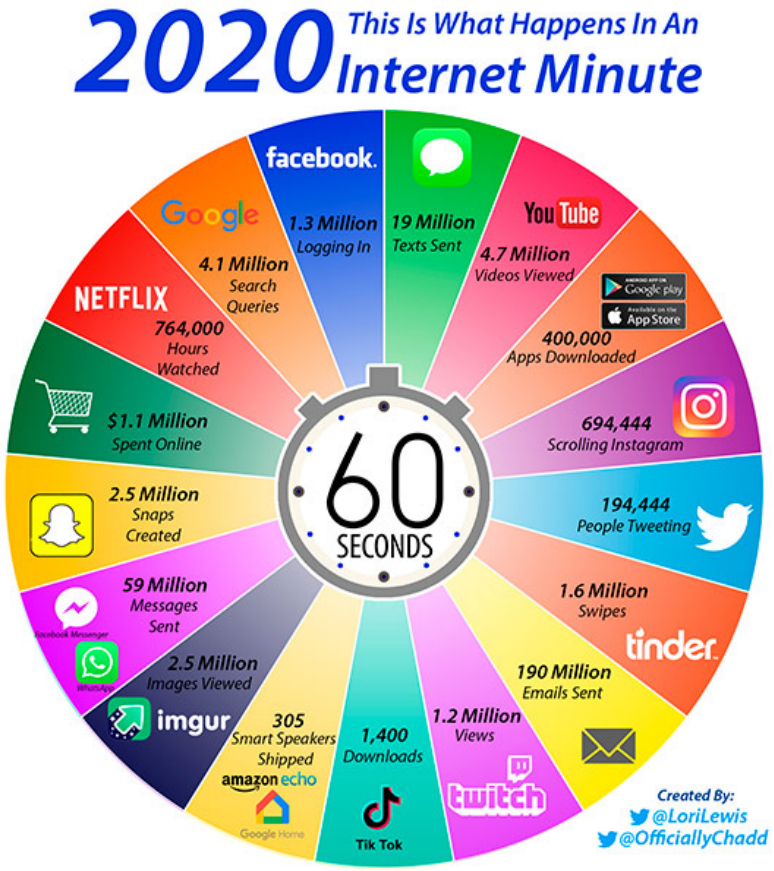
大 数 据 概 述

CONTENTS

1.1 从“数据”到“大数据”



60秒，互联网能发生什么？



1.1.1 大数据发展历程

- 2008年9 月
  - 美国《自然》（Nature）杂志专刊——Big data: The next Google,第一次正式提出“大数据”概念
- 2011年2月1日
  - 《科学》（Science）杂志专刊——Dealing with data，通过社会调查的方式，第一次综合分析了大数据对人们生活造成的影响，详细描述了人类面临的“数据困境”
- 2011年5月
  - 麦肯锡研究院发布报告——Big data: The next frontier for innovation, competition, and productivity,第一次给大数据做出相对清晰的定义：“大数据是指其大小超出了常规数据库工具获取、储存、管理和分析能力的数据集。”

原文：“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

1.1.2 大数据“4V”特征

体量大（Volume）

从2013年至2020年，人类的数据规模将扩大50倍，每年产生的数据量将增长到44万亿GB，相当于美国国家图书馆数据量的数百万倍，且每18个月翻一番。

速度快（Velocity）

随着现代感测、互联网、计算机技术的发展，数据生成、储存、分析、处理的速度远远超出人们的想象力，这是大数据区别于传统数据或小数据的显著特征。

种类多（Variety）

大数据数据来源广、维度多、类型杂，各种机器仪表在自动产生数据的同时，人自身的生活行为也在不断创造数据；不仅有企业组织内部的业务数据，还有海量相关的外部数据。

价值大密度低（Value）

大数据有巨大的潜在价值，但同其呈几何指数爆发式增长相比，某一对象或模块数据的价值密度较低，这无疑给我们开发海量数据增加了难度和成本。

5V

准确性(Veracity)

6V

动态性(Vitality)

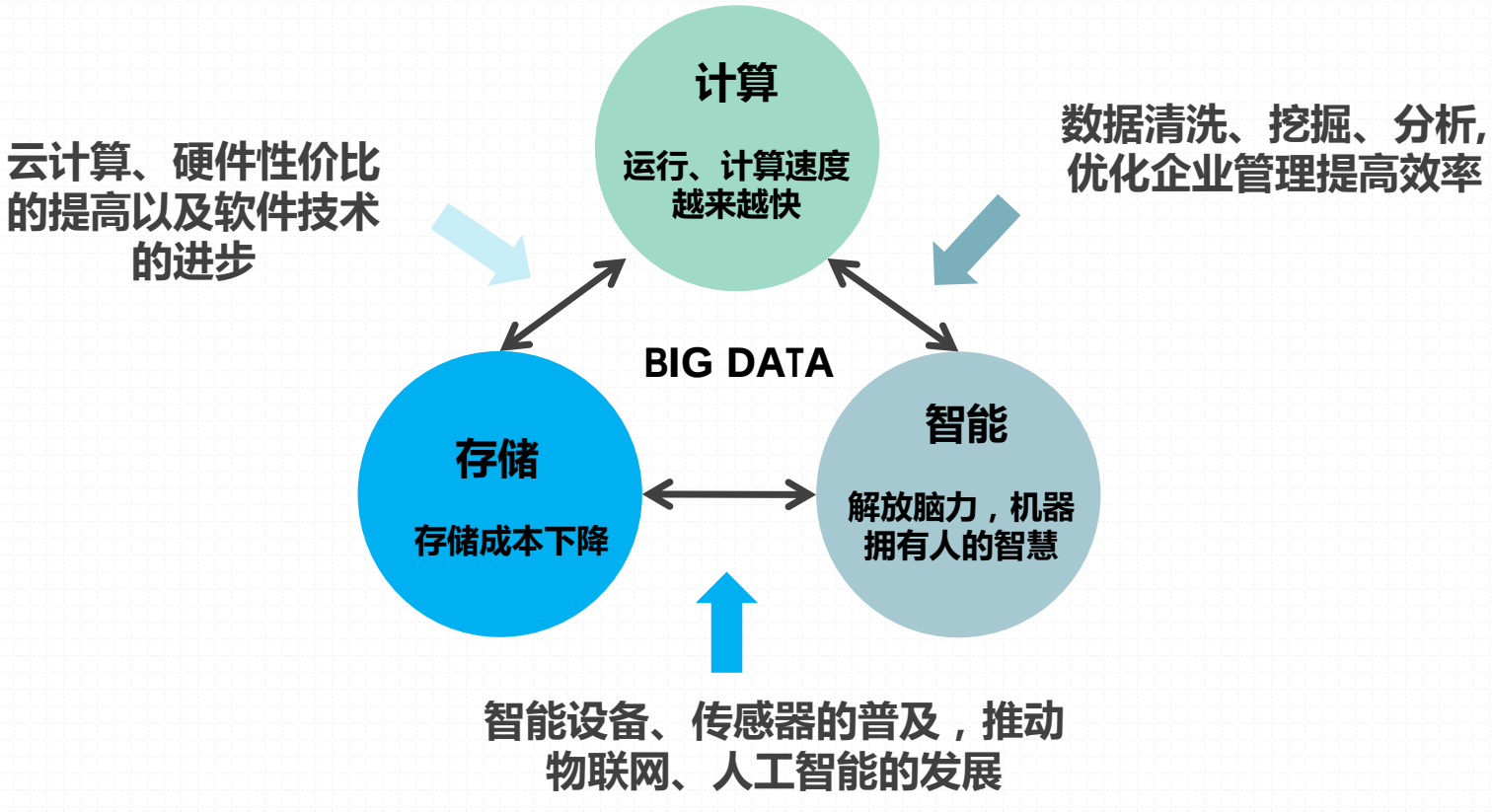
7V

可视化(Visualization)

8V

合法化(Validity)

1.2 大数据的技术支撑


计算
运行、计算速度越来越快
云计算、硬件性价比的提高以及软件技术的进步
数据清洗、挖掘、分析，优化企业管理提高效率
BIG DATA
存储
存储成本下降
智能
解放脑力，机器拥有人的智慧
智能设备、传感器的普及，推动物联网、人工智能的发展

- **存储：存储成本的下降**

云计算出现之前	云计算出现之后
云计算出现之前，数据存储的成本高昂。 例如，公司要建设网站，需购置和部署服务器，安排人员维护，保证数据存储的安全性和数据传输畅通性，定期清理数据，腾出空间以便存储新的数据，机房整体的人力和管理成本都很高。	**云计算出现后，数据中心的出现降低了公司计算和存储成本。** 例如，公司现在要建设网站，不需要去购买服务器，不需要去雇用技术人员维护服务器，可以通过租用硬件设备的方式解决问题。

存储成本的下降，也改变了大家对数据的看法，更加愿意把1年、2年甚至更久远的历史数据保存下来，有了历史数据的沉淀，才可以通过对比，发现数据之间的关联和价值。正是由于存储成本的下降，才能为大数据搭建最好的基础设施。

- **计算：运算速度越来越快**

  - 海量数据从原始数据源到产生价值，期间会经过存储、清洗、挖掘、分析等多个环节，如果计算速度不够快，很多事情是无法实现的。
    - 分布式系统基础架构Hadoop的出现，为大数据带来了新的曙光
    - HDFS为海量的数据提供了存储
    - MapReduce则为海量的数据提供了并行计算，从而大大提高了计算效率
    - Spark、Storm、Impala等各种各样的技术进入人们的视野

- **智能：机器拥有理解数据的能力**

  - 大数据带来的最大价值就是“智慧”，大数据让机器变得有智慧，同时人工智能进一步提升了处理和理解数据的能力

1	**谷歌AlphaGo大胜世界围棋冠军李世石**
2	**阿里云小Ai成功预测出《我是歌手》的总决赛歌王**
3	**iPhone上智能化语音机器人Siri**
4	**微信上与大家聊天的微软小冰**

1.3 大数据的意义

美国著名管理学家爱德华·戴明所言：“我们相信上帝。除了上帝，任何人都必须用数据来说话。”

“In God we Trust, all others bring data!”

（1）有数据可说

在大数据时代，“万物皆数”，“量化一切”，“一切都将被数据化”。人类生活在一个海量、动态、多样的数据世界中，数据无处不在、无时不有、无人不用，数据就像阳光、空气、水分一样常见，好比放大镜、望远镜、显微镜那般重要。

（2）说数据可靠

大数据中的“数据”真实可靠，它实质上是表征事物现象的一种符号语言和逻辑关系，其可靠性的数理哲学基础是世界同构原理。世界具有物质统一性，统一的世界中的一切事物都存在着时空一致性的同构关系。这意味着任何事物的属性和规律，只要通过适当编码，均可以通过统一的数字信号表达出来。

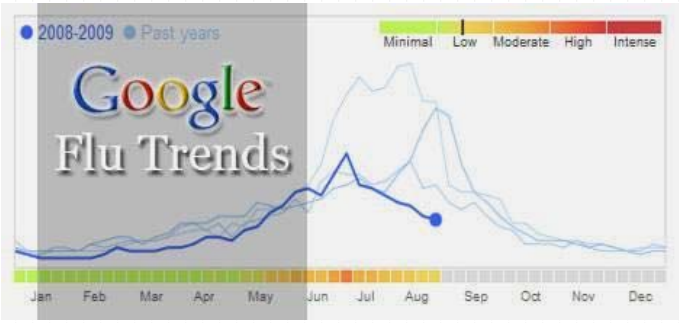
风马牛可相及

在大数据背景下，因海量无限、包罗万象的数据存在，让许多看似毫不相干的现象之间发生一定的关联，使人们能够更简捷、更清晰地认知事物和把握局势。大数据的巨大潜能与作用现在难以进行估量，但揭示事物的相关关系无疑是其真正的价值所在。

经典案例

（1）啤酒与尿布

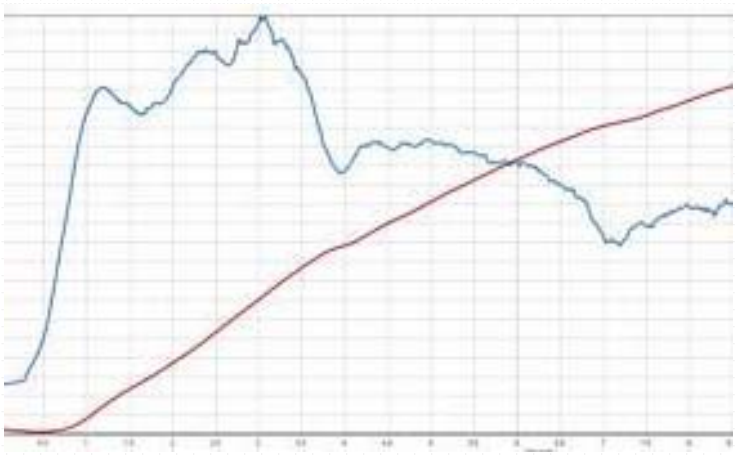
（2）谷歌与流感



沃尔玛（WALMART ）啤酒与尿布

- 沃尔玛通过分析海量的购物小票数据发现，啤酒与尿布经常同时被购买。
- 背后的消费心理学：在美国很多有婴儿的家庭中，通常是母亲在家照顾婴儿，年轻的父亲下班后去超市买尿布。父亲在买尿布的同时，也会顺手为自己买些啤酒作为犒劳
- 沃尔玛如何应对？通过将啤酒与尿布并排摆放在一起使二者的销量双双增长

谷歌流感趋势监测系统(Google Flu Trends)

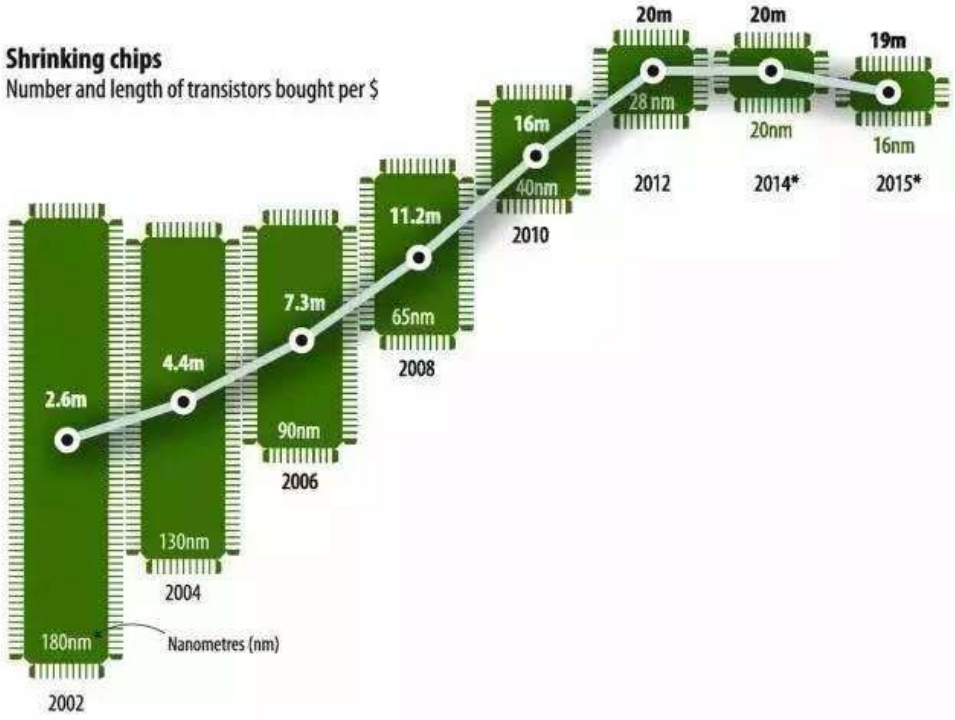


- 美国人在去医院前，喜欢在谷歌搜索类似“流感症状”的词汇
- 谷歌据此预测流感趋势，比美国疾病控制和预防中心(CDC)的流感通报提早一周到10天
- 政府由此可提前准备应对措施

大 数 据 概 述

CONTENTS

2.1 摩尔定律


Shrinking chips
Number and length of transistors bought per $
2.6m
180nm
2002
4.4m
130nm
2004
7.3m
90nm
2006
11.2m
65nm
2008
16m
40nm
2010
20m
28 nm
2012
20m
20nm
2014*
19m
16nm
2015*
Nanometres (nm)


“芯片上的晶体管数量每隔 24 个月
将增加一倍”
-戈登·摩尔
两层含义:
1. 每平方毫米的成本日渐上升
2. 晶体管数量翻倍 =“微缩”
性能提升
每个晶体管的成本下降

大数据到底有多大？

互联网每天产生的全部内容可以刻满6.4亿张DVD

全球每秒发送290万封电子邮件，一分钟读一篇的话，足够一个人昼夜不停地读5.5年

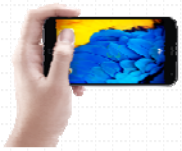
Google每天需要处理24PB的数据

每天会有2.88万个小时的视频上传到YouTube，足够一个人昼夜不停地观看3.3年

网民每天在Facebook上要花费234亿分钟，被移动互联网使用者发送和接收的数据高达44PB

Twitter上每天发布5000万条消息，假设10秒就浏览一条消息，足够一个人昼夜不停地浏览16年

2 大数据来源



智能终端拍照、拍视频

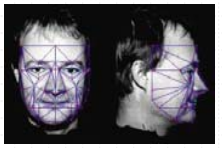
发微博、发微信

淘宝网 Taobao.com

其他互联网数据

来自“大人群”泛互联网数据

来自大量传感器的机器数据



科学研究及行业多结构专业数据

随着人类活动的进一步扩展，数据规模会急剧膨胀，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的各行业累积的数据量越来越大，数据类型也越来越多、越来越复杂，已经超越了传统数据管理系统、处理模式的能力范围，于是“大数据”这样一个概念才会应运而生

2.2 大数据来源分类

01 按产生数据的主体划分

- **少量企业应用产生的数据**

  如关系型数据库中的数据和数据仓库中的数据等

- **大量人产生的数据**

  如推特、微博、通信软件、移动通信数据、电子商务在线交易日志数据、企业应用的相关评论数据等

- **巨量机器产生的数据**

  如应用服务器日志、各类传感器数据、图像和视频监控数据、二维码和条形码（条码）扫描数据等

2.2 大数据来源分类

02 按数据来源的行业划分

- **以 BAT 为代表的互联网公司**
  - 百度公司数据总量超过了千PB级别，阿里巴巴公司保存的数据量超过了百PB级别，腾讯公司总存储数据量经压缩处理仍超过了百PB级别，数据量月增加达到10%
- **电信、金融、保险、电力、石化系统**
  - 电信行业年度用户数据增长超过10%，金融每年产生的数据超过数十PB，保险系统的数据量也超过了PB级别，电力与石化方面，仅国家电网采集获得的数据总量就达到了数十PB，石油化工领域每年产生和保存下来的数据量也将近百PB级别

2.2 大数据来源分类

02 按数据来源的行业划分

■ **气象、地理、政务等领域**

□ 中国气象局保存的数据将近10PB，每年约增数百TB；各种地图和地理位置信息每年约数十PB；政务数据则涵盖了旅游、教育、交通、医疗等多个门类，且多为结构化数据

■ **制造业和其他传统行业**

□ 制造业的大数据类型以产品设计数据、企业生产环节的业务数据和生产监控数据为主。其中产品设计数据以文件为主，非结构化，共享要求较高，保存时间较长；企业生产环节的业务数据主要是数据库结构化数据，而生产监控数据则数据量非常大。

2.2 大数据来源分类

03 按数据存储的形式划分

- 大数据不仅仅体现在数据量大，还体现在数据类型多。仅有20%左右属于结构化的数据，80%的数据属于社交网络、电子商务等领域等非结构化数据

  - **结构化数据** ： 简单来说就是数据库，如企业ERP、财务系统、医疗HIS数据库、教育一卡通、政府行政审批、其他核心数据库等数据

  - **非结构化数据**：包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频、视频信息等数据

2.2 大数据来源分类

04 常用的大数据获取途径

■系统日志采集

- 使用海量数据采集工具，用于系统日志采集，如Hadoop的Chukwa、Cloudera等，采用分布式架构，能满足大数据的日志数据采集和传输需求

■互联网数据采集

- 通过爬虫等获取数据信息，可以数据从网页中抽取出来并存储。除网站中包含的内容之外，还可以使用DPI或DFI等带宽管理技术实现对数据的采集

■与数据服务机构进行合作

- 数据服务机构通常具备规范的数据共享和交易渠道，人们可以在平台上快速、明确地获取自己所需要的数据。

■APP移动端数据采集

- APP中的SDK插件可以将用户使用APP的信息汇总给指定服务器。单个APP用户规模有限，数据量有限；但数十万APP用户，获取的用户终端数据和部分行为数据也会达到数亿的量级

大 数 据 概 述

CONTENTS

重庆大学 CHONGQING UNIVERSITY

3.1 大数据应用场景

3.1 大数据应用场景

3 大数据应用

零售行业

零售行业大数据应用有两个层面，一个层面是零售行业可以了解客户的消费喜好和趋势，进行商品的精准营销，降低营销成本。另一个层面是依据客户购买的产品，为客户提供可能购买的其他产品，扩大销售额，也属于精准营销范畴。

未来考验零售企业的是如何挖掘消费者需求，以及高效整合供应链满足其需求的能力，因此，信息技术水平的高低成为获得竞争优势的关键要素。

- 淘宝上的买家在购买商品前，会比较多家供应商的产品，进而反映到淘宝网站统计数据中，通过用户比选、购买行为进行贸易分析预测

- 2008年初，阿里巴巴平台上买家询盘数急剧下滑，淘宝网预测到欧美对中国的采购在下滑，最后推断出世界贸易即将发生变化

3 大数据应用

金融行业

1）银行数据应用场景

利用数据挖掘来分析出一些交易数据背后的商业价值。

2）保险数据应用场景

用数据来提升保险产品的精算水平，提高利润水平和投资收益。

3）证券数据应用场景

对客户交易习惯和行为分析可以帮助证券公司获得更多的收益。

- 汇丰银行在防范信用卡和借记卡欺诈的基础上，利用SAS（统计分析系统）构建了一套全球业务网络的防欺诈管理系统，为多种业务线和渠道提供完善的欺诈防范
- 该系统通过收集和分析大数据，以更快的信息获取速度挖掘交易的不正当行为，并迅速启动紧急告警

3 大数据应用

智慧城市

大数据技术可以了解经济发展情况、各产业发展情况、消费支出和产品销售情况等，依据分析结果，科学地制定宏观政策，平衡各产业发展，避免产能过剩，有效利用自然资源和社会资源，提高社会生产效率。

大数据技术也能帮助政府进行支出管理，透明合理的财政支出将有利于提高公信力和监督财政支出。

- 美国佛罗里达州，通过对高速口收费站数据（约为110万条）比对分析发现当地的3900辆警车在13个月的时间里共发生了5100多次的超速行驶记录。进一步的筛选分析发现，警车超速时间竟然大部分都发生在上下班时间

- 翔实的纪录与可信的分析结果引起了当地民众广泛关注，牵扯到超速案件的12个部门近800名警察受到处理，“警察开快车”事件被有效治理和纠正

3 大数据应用

农业行业

借助于大数据提供的消费能力和趋势报告，政府可为农业生产进行合理引导，依据需求进行生产，避免产能过剩造成不必要的资源和社会财富浪费。

通过大数据的分析将会更精确地预测未来的天气，帮助农民做好自然灾害的预防工作，帮助政府实现农业的精细化管理和科学决策。

- 2013年美国孟山都公司以9.3亿美元收购了加州气候公司Climate Corporation。它通过分析气象、天气、降雨、地质土壤等海量数据，预测未来可能对农业生产造成破坏的各种情况，通过农业大数据信息帮助农民预测作物产量
- 农业大数据的有效利用，既节省了肥料资源，也避免了过量化肥的污染，还可通过收集分析降水、温度、土壤种类和作物生长周期信息，帮助农民及时发现和解决农田存在问题

教育行业

信息技术已在教育领域有了越来越广泛的应用，教学、考试、师生互动、校园安全、家校关系等，只要技术达到的地方，各个环节都被数据包裹。

通过大数据的分析来优化教育机制，也可以作出更科学的决策，这将带来潜在的教育革命。

- 南京理工大学教育基金会通过大数据分析，每个月在食堂吃饭超过 60顿、一个月总消费不足420元的，被列为受资助对象。南京理工大学还采取直接将补贴款打入学生饭卡的方式，学生无需填表申请，不用审核

3 大数据应用

环境行业

借助于大数据技术，天气预报的准确性和实效性将会大大提高，预报的及时性将会大大提升。

同时对于重大自然灾害如龙卷风，通过大数据计算平台，人们将会更加精确地了解其运动轨迹和危害的等级，有利于帮助人们提高应对自然灾害的能力。

- 纽约曼哈顿哈德森河(Hudson River)，在过去20年里，居民造成的下水道污物的沉积，以及近代大型工厂倾倒的有毒化学物质，致使这条生态系统敏感的河流受到严重污染
- 纽约州政府发起了一个“新一代的水资源管理计划”，他们在河流的全程都安装了传感器，这些传感器把水的不同层面、各种物理、化学、生物数据，实时地通过网络传递到后台的计算中心区。各种数据汇成了一条虚拟的哈德森河，流水何时被污染，化学、物理、生物成分发生了什么变化，一看便知
- 经过多年努力，哈德森河流已恢复其清澈的水质和优美的环境

大 数 据 概 述

CONTENTS

4.1 大数据处理

- 大数据正带来一场信息社会的变革。大量的结构化数据和非结构化数据的广泛应用，致使人们需要重新思考已有的IT模式
- 与此同时，大数据将推动进行又一次基于信息革命的业务转型，使社会能够借助大数据获取更多的社会效益和发展机会
- 庞大的数据需要我们进行剥离、整理、归类、建模、分析等操作，通过这些动作后，我们开始建立数据分析的维度，通过对不同的维度数据进行分析，最终才能得到想到的数据和信息

因此，如何进行大数据的采集、导入/预处理、统计/分析和大数据挖掘，是“做”好大数据的关键基础

4.1.1 大数据采集

- 大数据的采集通常采用多个数据库来接收终端数据，包括智能硬件端、多种传感器端、网页端、移动APP应用端等，并使用数据库进行简单处理

- 常用的数据采集的方式主要包括以下几种：

01 数据抓取

02 数据导入

03 物联网传感设备自动信息采集

4.1.2 导入/预处理

- 采集端本身有很多数据库，如果要对这些海量数据进行有的分析，应该将这些数据导入到一个集中大型分布式数据库或者分布式存储集群中，同时，在导入的基础上完成数据清洗和预处理工作。

- 现实世界中数据大体上都是不完整、不一致的“脏”数据，无法直接进行数据挖掘，或挖掘结果差强人意，为了提高数据挖掘的质量，产生了数据预处理技术

数据清洗	**主要是达到数据格式标准化、异常数据清除、数据错误纠正、重复数据的清除等目标。**
数据集成	**是将多个数据源中的数据结合起来并统一存储，建立数据仓库。**
数据变换	**通过平滑聚集、数据概化、规范化等方式将数据转换成适用于数据挖掘的形式。**
数据归约	**寻找依赖于发现目标的数据的有用特征，缩减数据规模，最大限度地精简数据量。**

4.1.3 统计和分析

- 统计与分析主要是利用分布式数据库，或分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总，以满足大多数常见的分析需求。

常见的统计分析方法

- 描述性统计：平均数、中位数、众数、方差、标准差、协方差等
- 假设检验：正态检验、t检验、卡方检验等
- 相关分析：单相关、复相关等
- 方差分析：单因素、多因素、协方差分析

4.1.4 大数据挖掘

分类

根据重要数据类的特征向量值及其他约束条件，构造分类函数或模型，目的是根据数据集的特点把未知类别的样本映射到给定类别中。

朴素贝叶斯算法　支持向量机SVM算法　AdaBoost算法　C4.5算法　CART算法

聚类

目的将数据集内具有相似特征属性的数据聚集在一起，同一个数据群中的数据特征要尽可能相似，不同的数据群中的数据特征要有明显的区别。

BIRCH算法　K-Means算法　期望最大化算法（EM算法）　K近邻算法

关联规则

找出所有能把一组事件或数据项与另一组事件或数据项联系起来的规则，以获得预先未知的和被隐藏的、不能通过逻辑操作或统计得出的信息。

Apriori算法　FP-Growth算法

预测模型

一种统计或数据挖掘的方法，包括可以在结构化与非结构化数据中使用以确定未来结果的算法和技术，可为预测、优化、预报和模拟等许多业务系统所使用。

序贯模式挖掘SPMGC算法

4.2 大数据挖掘工具——Spark MLlib

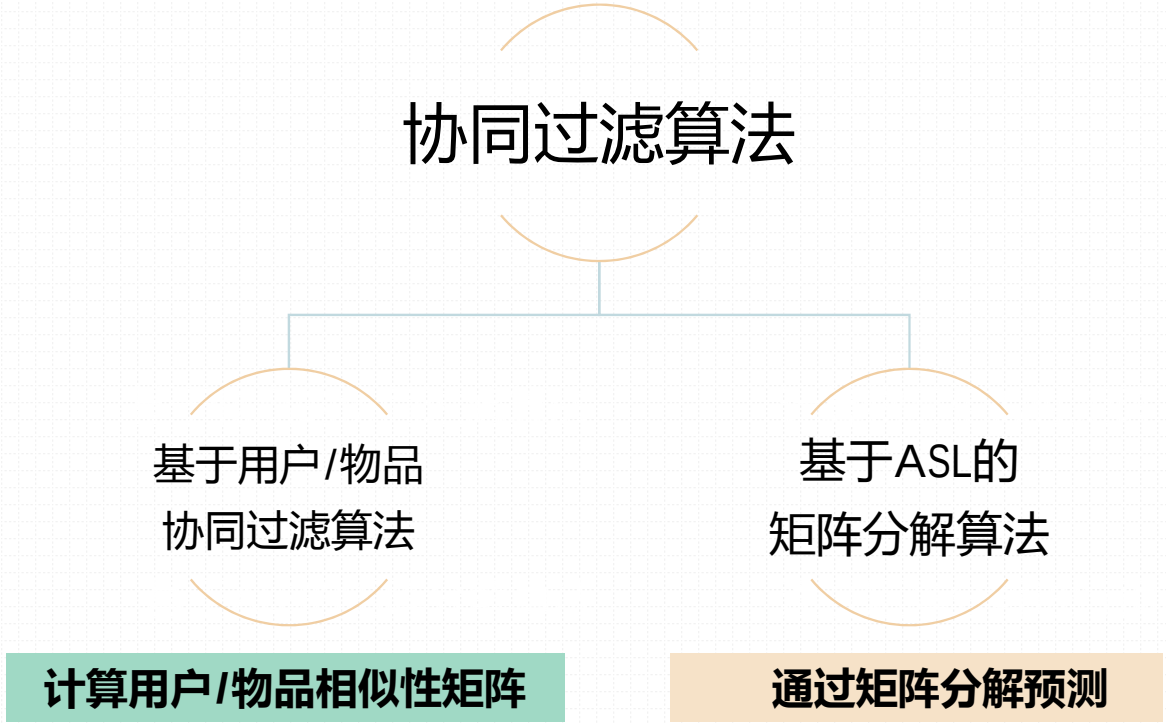
- 运行在Spark平台上专为在集群上并行运行而设计
- 内存中更快地实现多次迭代，适用于大规模数据集

Spark MLlib
The Machine Learning Library

	离散型	连续型
有监督机器学习	分类 逻辑回归 支持向量机(SVM) 朴素贝叶斯 决策树 随机森林 梯度提升决策树 (GBT)	回归 线性回归 决策树 随机森林 梯度提升决策树 (GBT) 保序回归
无监督机器学习	聚类 k-means 高斯混合 快速迭代聚类(PIC) 隐含狄利克雷分布(LDA) 二分k-means 流k-means	协同过滤、降维 交替最小二乘(ALS) 奇异值分解(SVD) 主成分分析(PCA)

4.2.1 协同过滤

通过收集大量用户（协同）的喜好信息，以自动预测（过滤）用户感兴趣的商品。



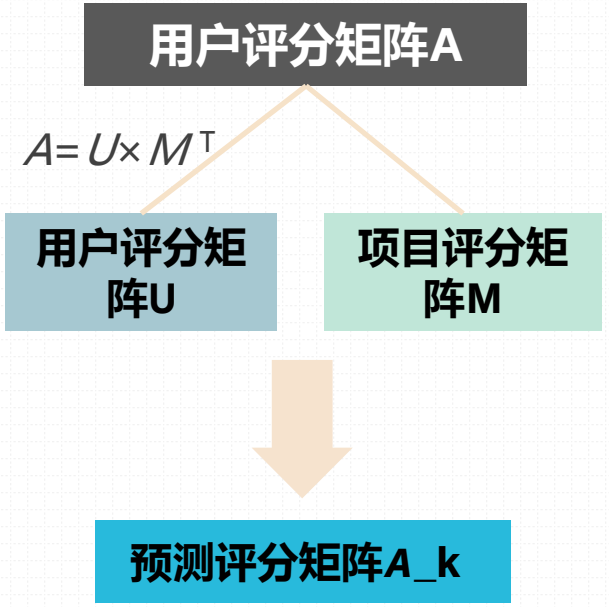
重庆大学 CHONGQING UNIVERSITY

◆ 基于ALS的矩阵分解算法

□ 算法介绍

- **基本思想：**是对稀疏矩阵进行模型分解，评估出缺失项的值，得到一个基本训练模型。然后依照此模型可以针对新的用户和物品数据进行评估。
- **缺失项：**ALS采用交替的最小二乘法计算缺失项，交替的最小二乘法是在最小二乘法的基础上发展而来的。

□ 算法流程



◆ 基于ALS的矩阵分解算法

□ 算法实例（1/3）

协同过滤的分类来说，ALS算法属于User-Item CF，也叫做混合CF，它同时考虑了User和Item两个方面。

Table1：听众对每首歌的评分矩阵（原始数据）

	痴心绝对	小酒窝	红豆	明天你好	浮夸
听众1	5			4	
听众2		6			3
听众3	3		7		
听众4				4	
听众5		4			6

◆ 基于ALS的矩阵分解算法

□ 算法实例（2/3）

ALS矩阵分解会把矩阵A分解成两个矩阵的相乘，分别是X矩阵和Y矩阵：

$$A = X * Y$$

x列&Y行表示：称为ALS中的因子

Table2：矩阵X

	性格	教育程度	兴趣爱好
听众1	X_{11}	X_{12}	X_{13}
听众2	X_{21}	X_{22}	X_{23}
听众3	X_{31}	X_{32}	X_{33}
听众4	X_{41}	X_{42}	X_{43}
听众5	X_{51}	X_{52}	X_{53}

Table3：矩阵Y

	痴心绝对	小酒窝	红豆	明天你好	浮夸
性格	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
教育程度	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
兴趣爱好	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}

◆ 基于ALS的矩阵分解算法

□ 算法实例（3/3）

音乐评分预测：如听众5，他从没听过“红豆“这首歌，拿到听众5在 矩阵分解中X矩阵的向量M，把向量M和”红豆“在 Y矩阵中的对应向量N **相乘**，预测出听众5对于”红豆“评分。

Table2：矩阵X

	性格	教育程度	兴趣爱好
听众1	X_{11}	X_{12}	X_{13}
听众2	X_{21}	X_{22}	X_{23}
听众3	X_{31}	X_{32}	X_{33}
听众4	X_{41}	X_{42}	X_{43}
听众5	X_{51}	X_{52}	X_{53}

Table3：矩阵Y

	痴心绝对	小酒窝	红豆	明天你好	浮夸
性格	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}
教育程度	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
兴趣爱好	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}

◆ 基于用户/物品协同过滤算法

□ 算法介绍

□ 基本思想

- **UserCF思想**：是根据用户行为数据，找到与目标用户有 相似兴趣的其他用户，给用户推荐其他用户喜欢的物品。
- **ItemCF思想**：是根据用户行为数据，计算物品间的相似度，基于用户以往的喜好记录，推荐给用户相似的物品。

重庆大学 CHONGQING UNIVERSITY

◆ 基于物品的协同过滤算法

◘ 算法流程

构建评分矩阵X

↓

构建物品相似度矩阵Y

↓

分值：$A = X * Y$

1. 根据原始数据计算用户对物品的评分

用户对物品的评分						
	101	102	103	104	105	106
A	5	3	2.5	0	0	0
B	2	2.5	5	2	0	0
C	2	0	0	4	4.5	6
D	5	0	3	4.5	0	4
E	4	3	2	4	3.5	4

重庆大学 CHONGQING UNIVERSITY

◆ 基于物品协同过滤算法

◘ 算法流程

构建评分矩阵X

↓

构建物品相似度矩阵Y

↓

分值：$A = X * Y$

2.计算物品之间的相似度，可使用余弦相似度，如下：

$$w_{ij} = \frac{A_i \cdot A_j}{\| A_i \| \| A_j \|}$$

物品与物品的相似度

	101	102	103	104	105	106
101	5	3	4	4	2	3
102	3	3	3	2	1	1
103	4	3	4	3	1	2
104	4	2	3	4	2	3
105	2	1	1	2	2	2
106	3	1	2	3	2	3

重庆大学 CHONGQING UNIVERSITY

◆ 基于物品协同过滤算法

◘ 算法流程

构建评分矩阵X

↓

构建物品相似度矩阵Y

↓

分值：$A = X * Y$

3.用户对物品的评分矩阵 × 物品相似矩阵 = 推荐列表

$$A = X * Y^T$$

	A	B	C	D	E
101	5	2	2	5	4
102	3	2.5	0	0	3
103	2.5	5	0	3	2
104	0	2	4	4.5	4
105	0	0	4.5	0	3.5
106	0	0	6	4	4

□4.2.2 聚类算法

Scale 代码

```
import org.apache.spark.mllib.clustering.{KMeans,KMeansModel}
import org.apache.spark.mllib.linalg.Vectors

// Load and parse the data
val data = sc.textFile("data/mllib/points.txt")
val parsedData = data.map(s =>
Vectors.dense(s.split("\\s+").map(_.toDouble))).cache()

// Cluster the data into three classes using KMeans
val k = 3
val numIterations = 20
val clusters = KMeans.train(parsedData, k, numIterations)

for(c <- clusters.clusterCenters){ println(c) }
clusters.predict(Vectors.dense(10,10))

// Evaluate clustering by computing Within Set Sum of Squared Errors
val WSSSE = clusters.computeCost(parsedData)
println("Within Set Sum of Squared Errors = " + WSSSE)
```

输出结果

```
[1.5,10.5]
[10.5,1.5]
[10.5,10.5]
2
Within Set Sum of Squared Errors =
6.000000000000057
```

重庆大学 CHONGQING UNIVERSITY

□4.2.3 回归算法

回归算法和分类算法都是有监督的学习，分类算法预测的结果是离散的类别，而回归算法预测的结果是连续的数值。

- 线性回归——最常用的算法之一，使用输入值的线性组合来预测输出值
- 类LinearRegressionWithSGD——MLlib实现线性回归算法的常用类之一，基于随机梯度下降实现线性回归

输入数据

```
-0.801696864,-1.497953299 -0.263601072
0.407654716,0.796247055 0.047655941
-0.979828061,-1.622338485 -0.843294092
-0.403657629,-0.990720665 0.458513517
-0.183790121,-0.171901282 -0.489197399
-0.92193133,-1.607582523 -0.59070034
0.100333967,0.366273919 -0.414014963
-0.312807305,-0.710307385 0.211731938
-0.364812555,-0.262791728 -1.167083456
0.331381491,0.899043117 -0.59070034
-0.236406401,-0.903451691 1.07659722
-0.307844837,-0.06333379 -1.380889709
-0.769339565,-1.1539379 -0.961853075
0.044169664,0.062020372 0.065797389
-0.96409108,-0.757310278 -2.927179705
0.769104799,1.112269933 1.064849162
```

输入函数

y=0.5*x1+0.2*x2

输出结果

weights:
[0.5000000000539042,0.19999999999
89402], intercept:0.0

training Mean Squared Error =
9.576567731363342E-20

□4.2.4 分类算法

输入数据

lable	x1	x2	x3
0	1	0	0
0	2	0	0
0	3	0	0
0	4	0	0
1	0	1	0
1	0	2	0
1	0	3	0
1	0	4	0
2	0	0	1
2	0	0	2
2	0	0	3
2	0	0	4

Scale代码

加载训练数据文件

解析每行数据

训练模型

预测分类

测试案例

Vector(0 0 9) 's label is 2.0

Accuracy: 1.0

Thank You!