

数据仓库与数据挖掘

(大数据工程技术)

鄂海红 计算机学院 副教授

ehaihong@bupt.edu.cn 微信/QQ : 87837001

2017.10.28

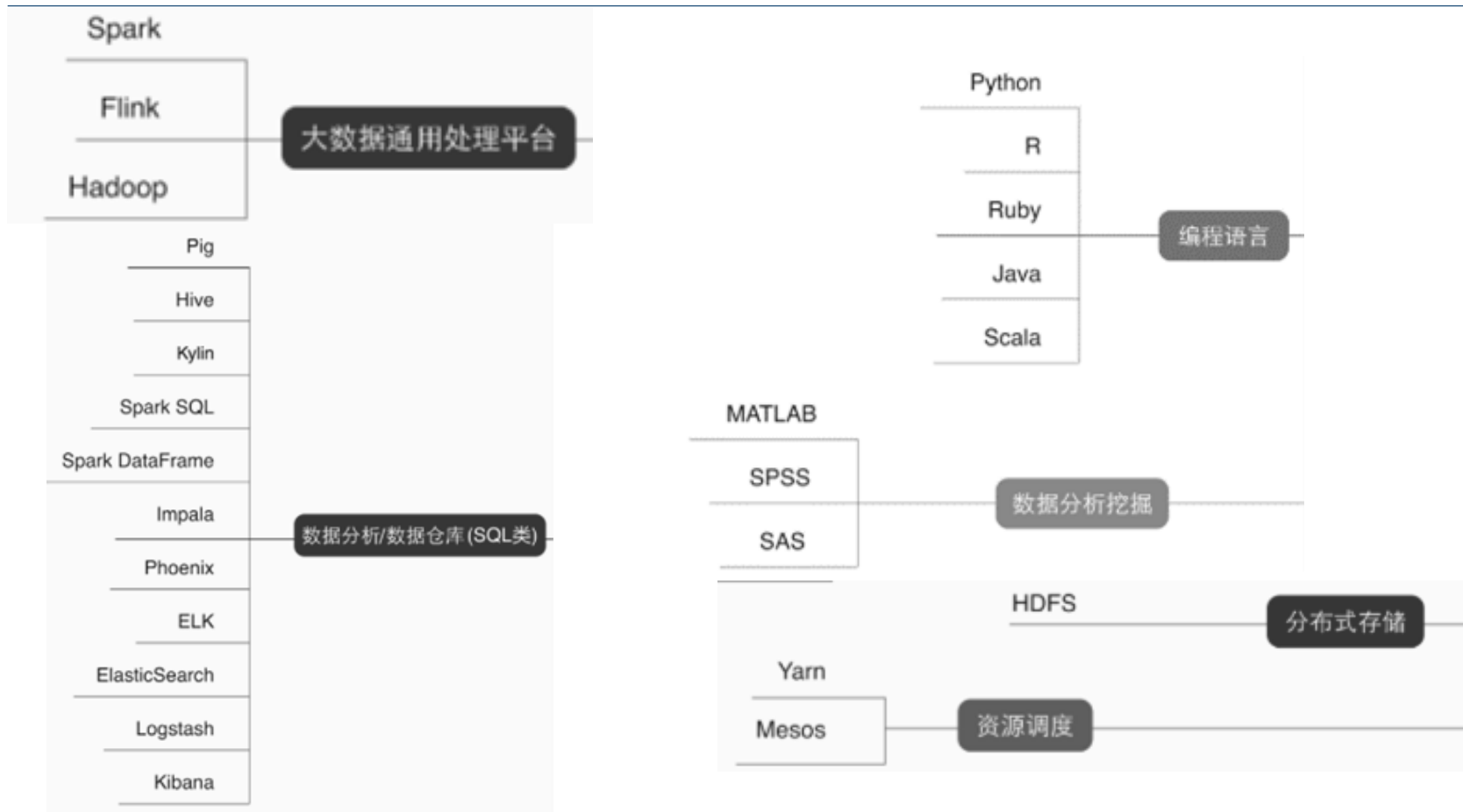
第一个议题

大数据概览

大数据工程师的技术图谱

来自

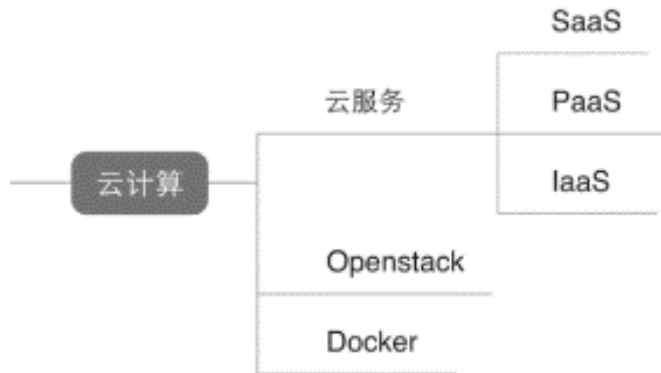
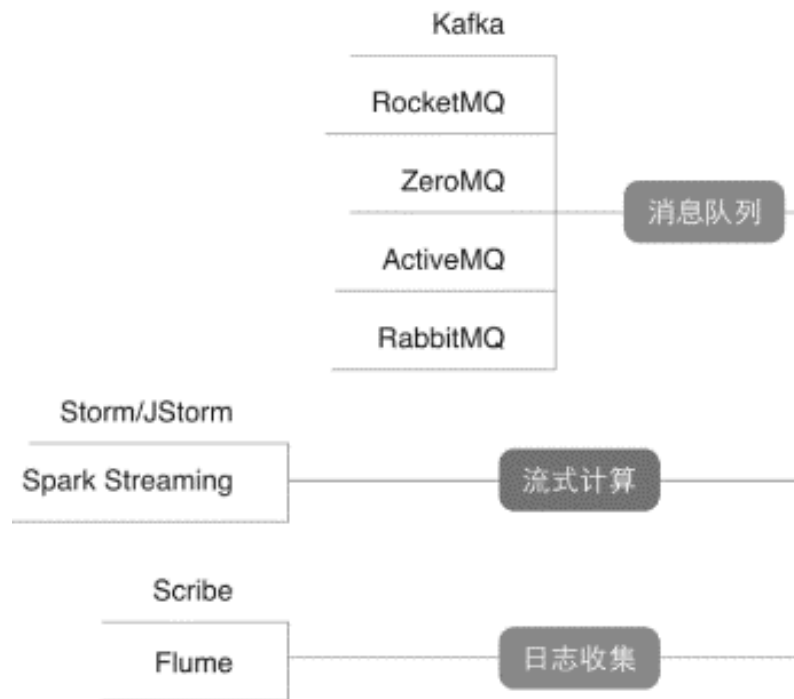
<https://github.com/TeamStuQ/skill-map>



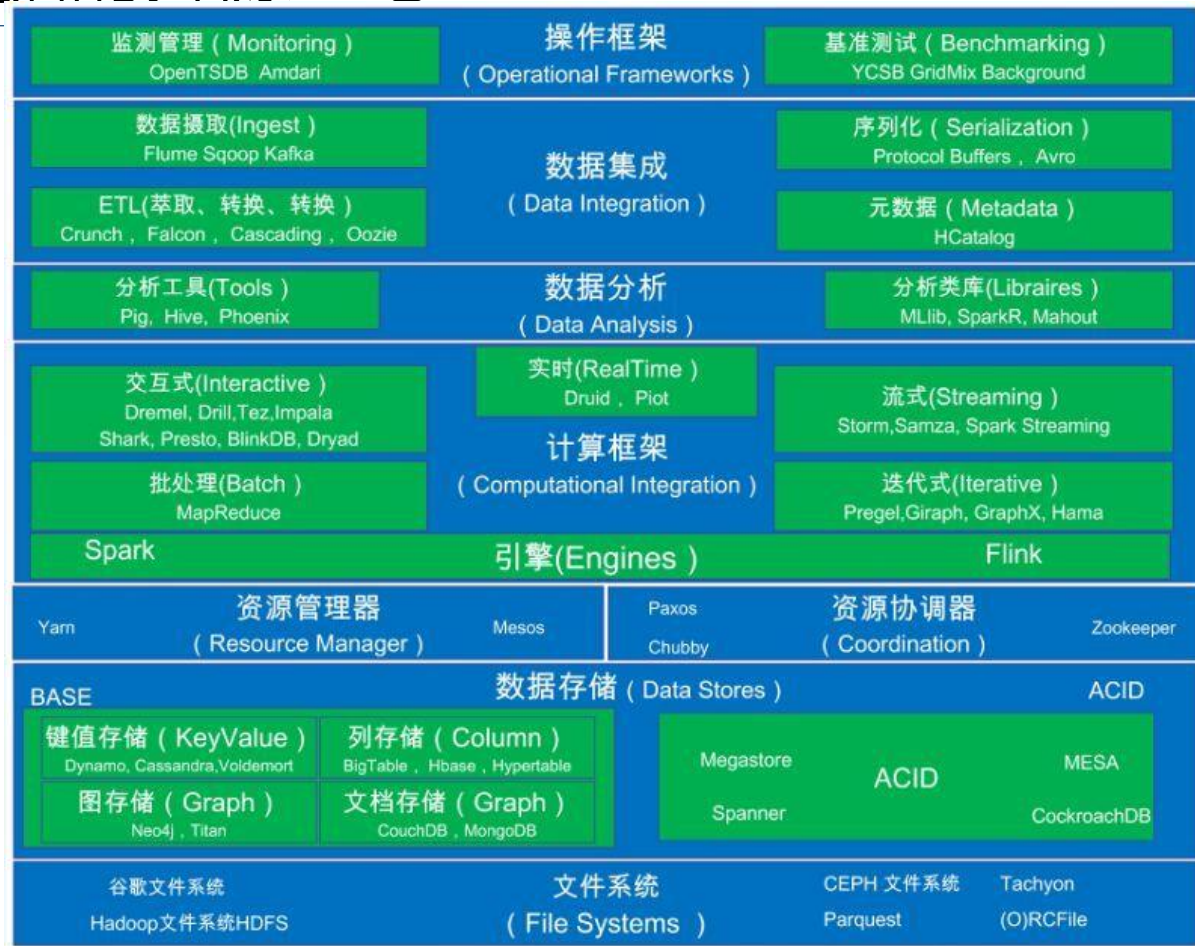
大数据工程师的技术图谱



大数据工程师技术图谱



大数据的开源生态



大数据工程实战从哪入手

- **数据科学的数据分析实战**

- Python/R等语言
- 基于Python的机器学习实践
- 数据加工全流程：数据导入、清洗、预处理、建模、分析、可视化

- **大数据开源工具的操作实战**

- Hadoop、HDFS、Hive、HBase、Spark、Yarn、Sqoop.....

- **云计算的操作实战**

- 分布式、微服务架构、容器技术.....

大数据离我们很近

情人节第一分钟
谁用微信红包说了“我爱你”

情人节
微信520表白红包 再度开启

数据统计时间: 2017.2.14 00:00~18:00

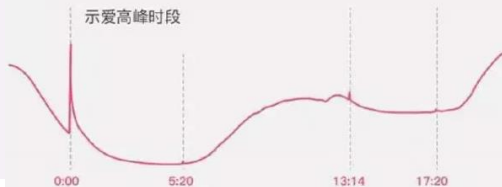
表白红包指以下金额的个人红包

5.2 13.14 52 5.21 520
131.4 52.1 0.52 52.13 52.14

我要你最先听到我的爱



全世界都在说我爱你



男女有别, 爱无别



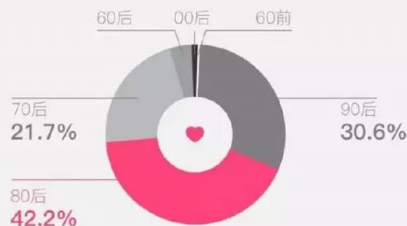
最受欢迎的红包金额



这些是最有爱的城市



从少年到白头, 爱一直都在



大数据的概念与意义——从“数据”到“大数据”

时至今日，“数据”变身“大数据”，“开启了一次重大的时代转型”。

“大数据”这一概念的形成，有三个标志性事件：

1

- 2008年9月，美国《自然》（Nature）杂志专刊——The next google,第一次正式提出“大数据”概念。

2

- 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”。

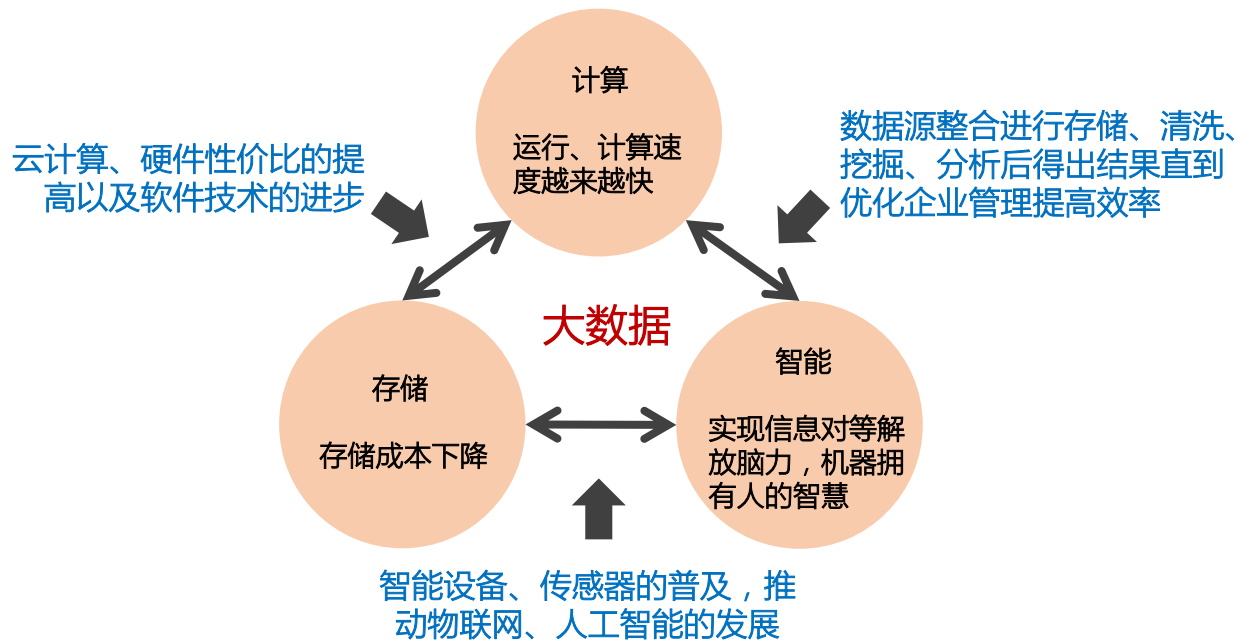
3

- 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。”

大数据的4V特征



大数据的技术支撑



1) 存储：存储成本的下降

云计算出现之前

在云计算出现之前，数据存储的成本是非常高的。

例如，公司要建设网站，需要购置和部署服务器，安排技术人员维护服务器，保证数据存储的安全性和数据传输的畅通性，还会定期清理数据，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。

云计算出现之后

云计算出现后，数据存储服务衍生出了新的商业模式，数据中心的出现降低了公司的计算和存储成本。

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。

2) 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。所以，在大数据的发展过程中，计算速度是非常关键的因素。

- 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。

3) 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。例如：

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

大数据的意义



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

(1) 有数据可说

在大数据时代，“万物皆数”，“量化一切”，“一切都将数据化”。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。

因此，“用数据说话”、“让数据发声”，已成为人类认知世界的一种全新方法。

大数据的意义

风马牛可相及

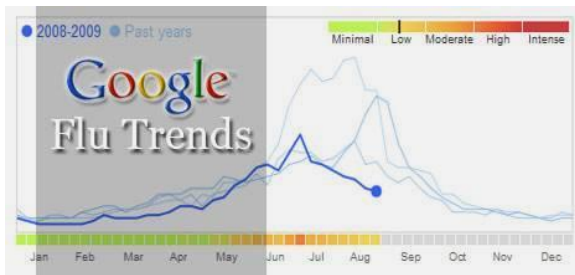
- ✓ 在大数据背景下，因海量无限、包罗万象的数据存在，让许多看似毫不相干的现象之间发生一定的关联，使人们能够更简捷、更清晰地认知事物和把握局势。
- ✓ 大数据的巨大潜能与作用现在难以进行估量，但揭示事物的相关关系无疑是其真正的价值所在。

经典案例：

(1) 啤酒与尿布

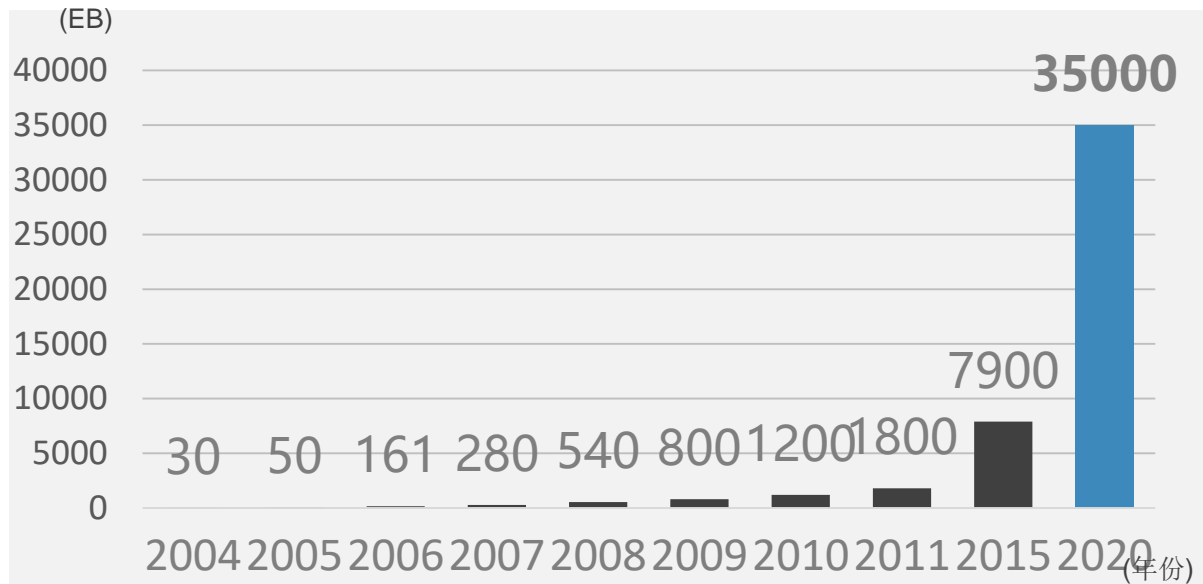


(2) 谷歌与流感



大数据的新摩尔定律

杰姆·格雷（Jim Gray）提出著名的“新摩尔定律”，即人类有史以来的数据总量，每过18个月就会翻一番。



全球数据总量图



为什么全球数据量
增长如此之快？

大数据的疯狂增长

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

大数据到底有多大？

以上一组互联网数据

海量数据的产生

海量的数据的产生



智能终端拍照、拍视频



发微博、发微信





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。

海量数据产生的来源主体

01

按产生数据的主体划分

1) 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。

2) 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。

3) 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。

海量数据产生的来源行业

02

按数据来源的行业划分

1) 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，拥有90%以上的电商数据，腾讯公司总存储数据量经压缩处理以后仍然超过了百PB级别，数据量月增加达到10%。

2) 电信、金融、保险、电力、石化系统

电信行业数据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。

3) 公共安全、医疗、交通领域

一个中、大型城市，一个月的交通卡口记录数可以达到3亿条；整个医疗卫生行业一年能够保存下来的数据就可达到数百PB级别；航班往返一次产生的数据就达到TB级别；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十PB。

4) 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。

5) 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。在其他传统行业，虽然线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业数据量剧增，但是数据量还处于积累期，整体体量都不算大，多则达到PB级别，少则数十TB或数百TB级别。

海量数据的存储形式

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。如此海量的数据中，仅有20%左右属于结构化的数据，80%的数据属于广泛存在于社交网络、物联网、电子商务等领域的非结构化数据。

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。

海量数据的获取途径

04

常用的大数据获取途径

1) 系统日志采集

可以使用海量数据采集工具，用于系统日志采集，如Hadoop的Chukwa、Cloudera的Flume、Facebook的Scribe等，这些工具均采用分布式架构，能满足大数据的日志数据采集和传输需求。

2) 互联网数据采集

通过网络爬虫或网站公开API等方式从网站上获取数据信息，该方法可以从网页中抽取出来，将其存储为统一的本地数据文件，它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。除了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网络流量的采集。

3) APP移动端数据采集

APP是获取用户移动端数据的一种有效方法，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，即使用户在没有访问时，也能获知用户终端的相关信息，包括安装应用的数量和类型等。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。


4) 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。而对于企业生产经营数据或学科研究数据等保密性要求较高的数据，也可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

大数据应用场景



大数据处理方法



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。

大数据处理方法-1-数据采集

1

大数据的采集

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，并且可以使用数据库进行简单的处理工作。

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

大数据处理方法-2-数据导入/预处理

2

导入/预处理

虽然采集端本身有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些数据导入到一个集中的大型分布式数据库或者分布式存储集群当中，同时，在导入的基础上完成数据清洗和预处理工作。也有一些用户会在导入时使用来自Twitter的Storm来对数据进行流式计算，来满足部分业务的实时计算需求。

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。

数据清理

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

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。

大数据处理方法-3-数据统计与分析

3

统计与分析

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，在这些方面可以使用Python语言。

一、实验环境的安装

1、Python版本选择

Python2.7版本目前使用较为广泛，建议安装2.7的版本，

2、解释器安装

通过miniconda 安装Python环境。

二、开发环境搭建

1、开发环境选择PyCharm

2、在线调试工具 jupyter notebook

三、Python 基本语法学习

学习Python的基本参考操作语法及基本数据结构

廖雪峰Python2.7教程

菜鸟教程

四、数据探索分析相关工具包

数据探索与分析工具包：numpy、pandas、matplotlib、seaborn三个工具包的学习。

大数据处理方法-4-数据挖掘

4

大数据挖掘

数据挖掘是创建数据挖掘模型的一组试探法和计算方法，通过对提供的数据进行分析，查找特定类型的模式和趋势，最终形成创建模型。

分类

一种重要的数据分析形式，根据重要数据类的特征向量值及其他约束条件，构造分类函数或分类模型，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

目的在于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

索系统中的所有数据，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，不能通过数据库的逻辑操作或统计的方法得出的信息。

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

第三个议题

大数据工程课程内容

课程内容

- 大数据概览
- Hadoop及HDFS
- Hive及HBase
- Spark基于内存的并行计算框架
- Kylin大数据OLAP框架
-



本门课程的考试方式

- 开卷考试/期末大作业（待和教务确认）
- 平时作业40%、期末考试60%（待和教务确认）
- 平时作业：上机实验报告

第一次作业

- **调研报告：**

- 云计算IaaS/PaaS产品调研（3个）
- 大数据生态组件调研（3个）

- **实验报告：**

- 第一次实验报告书

感谢聆听

