5.大数据技术

5.1 大数据概述

1、大数据出现背景:
（1）信息技术进步
（2）云计算技术的兴起
（3）数据资源化的趋势
2、大数据
大数据（big data），是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。
3、大数据提出
（1）部分人认为大数据概念由全球知名咨询公司麦肯锡提出。
（2）部分人认为是由 1980 年阿尔文.托夫勒在《第三次浪潮》中提到。
（2）大多数人认为最早在 2008 年 8 月在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》

提出。

4、大数据主要特性：

（1）**数据量大（Volume）**：

单位	换算关系
Byte（字节）	1Byte=8bit
KB（千字节）	1KB=1024Byte
MB（兆字节）	1MB=1024KB
GB（吉字节）	1GB=1024MB
TB（太字节）	1TB=1024GB
PB（拍字节）	1PB=1024TB
EB（艾字节）	1EB=1024PB
ZB（泽字节）	1ZB=1024EB

（2）**数据类型繁多（Variety）**：行业数据多，数据种类多：邮件、视频、音频、微信、微博、位置信息、网络日志等。

（3）**处理速度快（Velocity）**：大数据时代的数据产生速度非常迅速。大数据时代的很多应用都需要基于快速生成的数据给出实时分析结果，用于指导生产和生活实践。

（4）**价值密度低（Value）**：大数据虽然看起来很好，但是价值密度却远远低于传统关系型数据库中已经存在的那些数据。在大数据时代，很多有价值的信息都是分散在海量数据中的。

5、大数据发展历程

（1）**萌芽阶段**（20 世纪 90 年代至 21 世纪）：数据库技术成熟，数据挖掘理论成熟，也称数据挖掘阶段，智能工具、知识管理技术被应用：数据仓库、专家系统、知识管系统等

（2）**突破阶段**（2003 年至 2006 年）：Web2.0 飞速应用，非结构化的数据大量出现，传统的数据库处理难以应对，也称非结构化数据阶段。

（3）**成熟阶段**（2006 年至 2009 年）：解决方案走向成熟，形成了并行计算与分布式系统 两大核心技术，谷歌的 GFS 和 MapReduce 等大数据技术受到追捧，Hadoop 平台开始大行其道。

（4）**应用阶段**（2009 至今）：大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高。2013 年为大数据元年。

6、大数据与云计算技术关系

（1）从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。

（2）大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构。

（3）它的特色在于对海量数据的挖掘，但它必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。

7、大数据与人工智能技术关系

（1）大数据和人工智能虽然关注点并不相同，但是却有密切的联系，一方面人工智能需要大量的数据作为“思考”和“决策”的基础，另一方面大数据也需要人工智能技术进行数据价值化操作，比如机器学习就是数据分析的常用方式。

（2）在大数据价值的两个主要体现当中，数据应用的主要渠道之一就是智能体（人工智能产品），为智能体提供的数据量越大，智能体运行的效果就会越好，因为智能体通常需要大量的数据进行“训练”和“验证”，从而保障运行的可靠性和稳定性。

5.2 大数据的关键技术

1、大数据处理基本流程

①数据采集->②数据归整->③数据存储->④数据处理->⑤数据呈现

2、大数据技术

（1）数据采集与预处理：联机分析处理（OLAP）与实时处理分析

（2）数据存储和管理：对结构、非结构、半结构等海量数据进行存储（关系数据库、非关系数据库、数据仓库、分布式文件系统）

（3）数据处理与分析：利用 MapReduce 等结合着机器学习和数据挖掘算法实现数据分析和处理

（4）数据安全和隐私保护：构建出隐私数据保护体系和数据安全体系，保护个人隐私和数据安全。

3、Hadoop 技术

（1）Hadoop 是用于处理（运算分析）海量数据的技术平台，且是采用分布式集群的方式。

（2）功能

①、存储：提供海量数据的存储服务；

②、计算：提供分析海量数据的编程框架及运行平台；

（3）三大核心组件：

①、HDFS:hadoop 分布式文件系统海量数据的存储(集群服务)

②、MapReduce:分布式运算框架（编程框架）（导 jar 包程序）

③、Yarn:资源调度管理集群

4、MapReduce 技术

（1）MapReduce 是 Hadoop 核心技术之一。

（2）MapReduce 框架的核心步骤主要分两部分：Map 和 Reduce。

（3）为分布式计算的程序设计提供了良好的编程接口，并且屏蔽了底层通信原理，使得程序员只需关心业务逻辑本事，就可轻易的编写出基于集群的分布式并行程序。

（4）“Map”就是将一个任务分解成为多个子任务并行的执行；

（5）“Reduce”就是将分解后多任务处理的结果汇总起来，得出最后的分析结果并输出。

（6）MapReduce 的功能：

①、数据划分和计算任务调度：将 job 分成多个数据块来计算，并自动调度计算节点来处理这些数据块。

②、数据/代码互定位：减少数据通信，从数据所在的本地机架上寻找可用节点以减少通信延迟。

③、系统优化：为了减少数据通信开销，中间结果数据进入 Reduce 节点前会进行一定的合并处理

④、出错检测和恢复： MapReduce 需要能检测并隔离出错节点，并调度分配新的节点接管出错节点的计算任务，维护数据存储的可靠性。

5、NoSQL 技术

（1）NoSQL 数据库是非关系型数据库，它主要是用来解决半结构化数据和非结构化数据的存储问题。（mongoDB、redis、hbase 等）

（2）NoSQL 是一种非关系型 DMS，不需要固定的架构，可以避免 joins 链接，并且易于扩展。

（3）NoSQL 技术功能：

①、数据管理：提供查询窗口和命令窗口功能。

②、结构管理：提供库、文档和索引等对象管理功能。

③、实时性能展示：提供核心性能指标的实时展示。

6、爬虫技术

（1）网络爬虫是一种按照一定的规则、自动的抓取万维网信息的脚本或者程序，简单点就是爬虫是用事先写好的程序去抓取网络上的数据，这样的程序叫爬虫。是数据采集的一种手段。

（2）爬虫的分类

①、按照使用场景来分，可以分为两类：

- 通用爬虫：搜索引擎爬虫（百度）
- 聚焦爬虫：获取想要的数据

②、爬虫软件可以分为两类：

- 云爬虫（不需要安装软件）
- 采集器（下载安装）

（3）常用工具：

①、神箭手云爬虫

②、八爪鱼

③、集搜客 GooSeeker

④、WebMagic

⑤、DenseSpider

7、**清洗技术**

（1）**数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。**

（2）我们要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。

（3）需要清洗数据的主要类型:残缺数据、错误数据、重复数据。

（4）数据清洗的内容：一致性检查、无效值和缺失值的处理。

（5）常用工具

①、DataWrangler

②、Google Refine

8、大数据分析

（1）**数据分析是指用适当的统计方法对收集来的大量第一手资料和第二手资料进行分析，以求最大化地开发数据资料的功能，发挥数据的作用。是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。**

（2）数据分析与数据挖掘密切相关，但数据挖掘往往倾向于关注较大型的数据集，较少侧重于推理，且常常采用的是最初为另外一种不同目的而采集的数据。

（3）**数据分析的目的是把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来，以找出所研究对象的内在规律。**

（4）互联网行为大数据分析的方法：

①、漏斗分析法:营销中转化率问题

②、对比分析法：多个指标进行对比

③、用户分析法：活跃分析，用户分群等

④、细分分析法：深入和精细化

⑤、指标分析法：运用统计学来分析

9、大数据挖掘

（1）**数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们不知道的、但又是潜在有用的信息和知识的过程。**

（2）**数据源是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的伙要可接爱、可理解、可运用；**

（3）数据挖掘的方法：

①、机器学习：神经网络、决策树、SVM（支持向量机）、深度学习

②、统计方法：回归分析（多元回归）、判别分析（贝叶斯判别）、聚类分析（动态聚类）等

③、数据库方法：SQL、OLAP(联机分析处理)

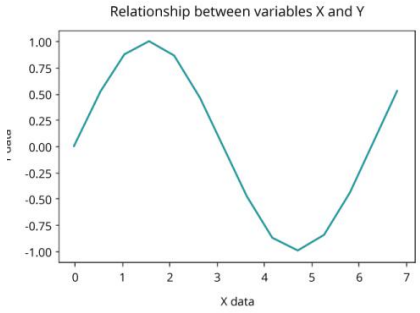
10、**大数据可视化**

（1）**大部分学习的信息来源于视觉，所以数据的可视化可以帮助我们更好的理解和学习信息。**

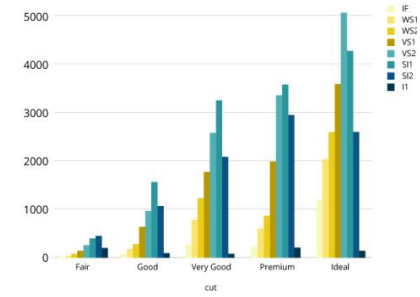
（2）**数据可视化提供了一套重要的工具和技术，可用于定性理解。**

11、可视化的图形表示：

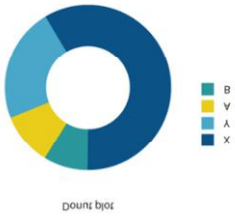
(1)**线图**:线图是最简单的技术，用于绘制一个变量与另一个变量之间的关系或依存关系。



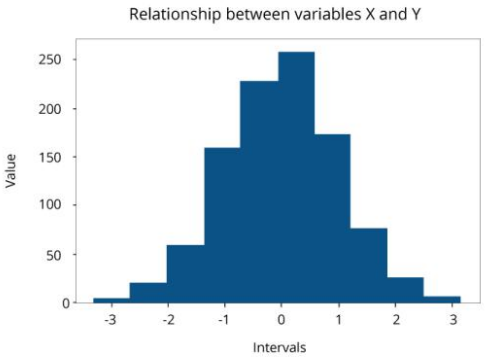
(2)**条形图**:条形图用于比较不同类别或组的数量。



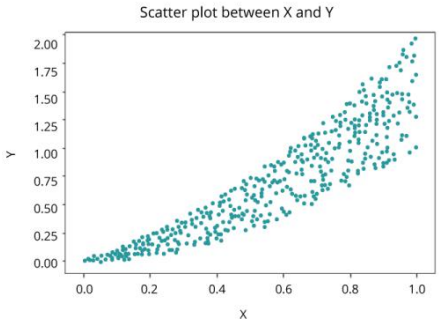
(3)**饼图和甜甜圈图**:用于比较整体的各个部分，并且在组成部分有限以及包含文本和百分比来描述内容时最有效。



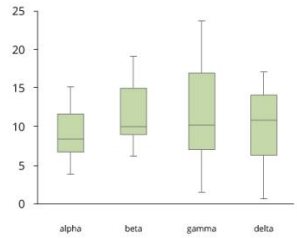
(4)**直方图**:直方图表示连续变量在给定间隔或时间段内的分布



(5)**散点图**:散布图是表示两个数据项的联合变化的二维图

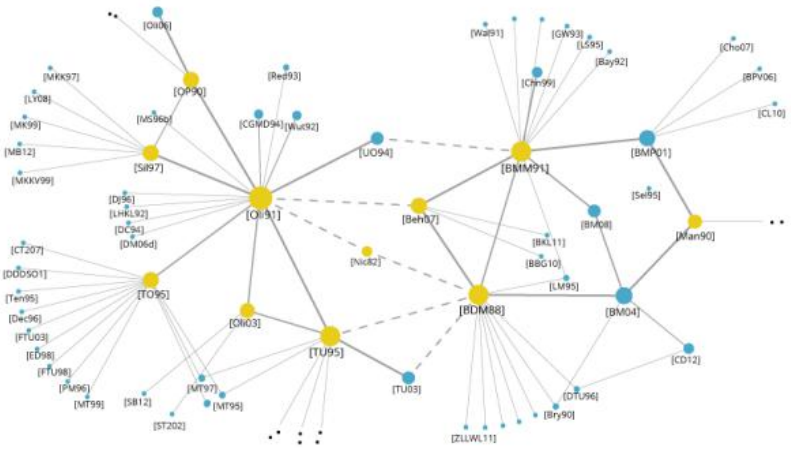


(6)**大数据的盒须图**:带须状图的装箱图显示了大数据的分布，并且很容易看到异常值。

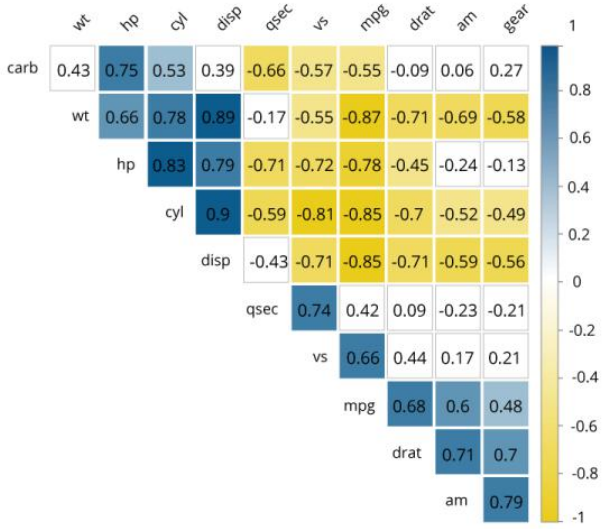


(7)**非结构化数据的词云和网络图**:作为显示高频或低频单词的一种方式.

网络图将关系表示为节点（网络内的各个参与者）和关系（关系在个人之间）.



(8) 相关矩阵：相关矩阵可以通过结合大数据和快速响应时间来快速识别变量之间的关系。



5.3 大数据主要应用领域

1、**大数据在农业领域的典型应用**

（1）**大数据理论和技术在农业上的应用和实践，是指运用大数据理念、技术和方法，解决农业或涉农领域数据的采集、存储、分析与应用等一系列问题，然后以此来指导农业生产经营。这里所讲的农业大数据就是指以大数据分析为基础去解决一系列问题。**

（2）典型的一些应用

①、**监测农情**：对自然灾害监测、作物估产及生长动态监测

②、**监测预警农产品**：大数据的技术给农产品信息的全面收集提供了技术基础，使农产品质量能够进行全方位比对，增加农产品质量监测的准确性。

③、**精准农业决策**：精准农业决策是指根据各个方面的农业信息制定出一整套有可实施性的精准管理措施。

④、**搭建农村综合信息服务系统**：搭建农村综合信息服务系统 是为了帮助农业信息的快速和有效的传播，使用全国范围实现信息资源的共享。

2、大数据有工业互联网领域的典型应用

（1）工业互联网中产生的数据，叫工业大数据库

（2）五大典型要求：超低时延控制、高清视频回传、数据不出场、安全保护与隔离、定制化的网络资源。

（3）在国家工业互联网大数据中心对大数据中心整个体系架构包括设备层、边缘层、企业区域层、产业层。

（4）典型应用：

①、**对工业设备的实时监控**。

②、**设备故障识别与预警的场景**。

③、**智能化的工艺流程优化。**

3、大数据有服务业领域的典型应用

（1）大数据时代的来临给服务行业的推进带来了新的机遇和挑战。

（2）典型应用：

①、**医疗服务**：电子病历、实时的健康状况告警、医学影像分析等

②、**旅游服务**：旅游个性化定制（线路、景区）、利用 GPS 定位来完善景区的用户体验、实现景区的无人购物系统、客流分析等

③、**金融服务**：客户画像、精准营销、风险管控、运营优化等

4、大数据未来发展趋势

（1）物联网

（2）智慧城市

（3）增强现实(AR)与虚拟现实(VR)

（4）区块链技术

（5）语音识别技术

（6）人工智能(AI)

（7）数字汇流