大数据笔记

计本 1902-张探探 1931053726

1 大数据概论

1.1 大数据概述

1.1.1 大数据的定义

通常来说，大数据指的是数据量超过一定大小（PB），无法用常规软件工具在规定的时间内进行抓取、管理和处理的数据集合。

1.1.2 大数据的特征

5V + 1C:

1 数据量大: 大数据的起始计量单位最小是 PB 级别以上

2 数据类型多

(a) 结构化数据: 可以使用数据或统一的结构加以表示

(b) 非结构化数据: 无法通过事先定义的数据模型表达或无法存入关系型数据库的数据，picture 等

(c) 半结构化数据: 介于万却结构化数据和非结构化数据之间。

3 价值密度低: 信息无处不在，产生海量数据。需要进行提纯。

4 数据时效性强: 要求能够从数据中及时提取知识和能量。

5 准确性高: 数据处理结果的精确度。通过技术手段分析全部数据，准确度大大提高。

6 复杂度高

(a) 数据本身的复杂性、计算的复杂性和信息系统的复杂性

(b) 数据复杂性：大数据涉及复杂的类型、复杂的结构和复杂的模式，数据本身具有很高的复杂性。

(c) 计算机的复杂性：大数据计算不能像处理小样本数据集那样做全局数据的统计分析和迭代计算。

(d) 系统复杂度：对大数据对计算机系统的运行效率提出了苛刻的要求。

1.2 大数据的分析过程

1.2.1 大数据的采集

1. 对数据进行提取、转换、加载，最终挖掘数据的潜在价值。

2. 针对不同的应用场景，还需要对数据进行转换操作，将数据转换成不同的数据格式

3. 大数据采集系统，主要分为以下 3 类系统：

   (a) 网络数据采集系统。通过网络爬虫和一些网站平台提供的共有 API。抓取网页数据，将网页数据存储在 HDFS 中。

   (b) 系统日志采集系统。手机日志数据提供离线和在线的实时分析数据。

   (c) 数据库采集系统。MySQL、Oracle、HBase、Redis、MongoDB、NoSQL。

1.2.2 大数据的存储方式

1. 分布式系统（HDFS）

   包含多个自主的处理单元，通过计算机网络互联来协作完成分配的任务，分而治之的策略很好的处理大规模的数据分析。

2. NoSQL：关系数据库无法满足需求。HBase 是一个高可靠性、高性能、面向列、可伸缩的分布式数据库系统。使用有以下的前提：

   (a) 使用硬盘

   (b) 把随机存储器作为存储体

3. 云数据库

   给予计算机技术发展的一种共享基础架构的方法。部署在虚拟化云计算环境中的数据库。

4. 大数据库存储技术路线

   (a) 大规模并行处理（MPP）：采用 Shared Nothing 构架，通过列存储、粗粒度索引等多项大数据处理技术，结合 MPP 架构高效的分布式计算模式。

   (b) 基于 Hadoop 技术扩展和封装

   (c) 大数据一体机，专为大数据的分析处理而设计的软硬件结合的产品。由一组集成的服务器、存储设备、操作系统、数据库管理系统以及数据查询、处理、分析用途而特别预先安装的软件构成

5. 数据的分析技术

   分布式处理工具有 Hadoop,MapReduce,Storm、JStorm、Samza 和 Spark。为了能同事进行批处理和流处理，出现了基于内容的

Spark 计算框架。

6. 大数据的展示及应用

可视化是利用计算机图形学及图像处理技术，将数据转换为图形或图像形式显示到屏幕上，并进行交互处理的理论。

1.3 大数据的价值、挑战与风险

1.3.1 商业价值

1. 对顾客群体细分，然后对每个群体量体裁衣般地采取独特的行动：

大数据分析的关键在于如何去搜索消费者相关的信息，如何获得趋势，挖掘出人们头脑中可能会消费的产品概念。

2. 运用大数据模拟实境，发掘新的需求和提高利润。

传感器，汽车和智能家居的普及可搜集到的数据呈现爆炸性增长。

3. 提高大数据成果在各相关部门的分享程度，提高企业决策能力。

(a) 两边到质变。

(b) 决策技术含量、知识含量最大幅度提高

(c) 大数据决策催生了很多过去难以想象的重大解决方案。

4. 进行商业模式、产品和服务的创新：

大数据技术可以有效的帮助企业整合、挖掘、分析其所掌握的庞大数据信息，构建系统的数据体系。从而完善企业自身的结构和管理机制。

1.3.2 社会生活价值

1. 对个人来说，大数据的高透明度和实时性，经过数据的精确挖掘，大数据可以为个人提供个性化的医疗服务。
2. 在大数据的出现使得学生在受教育的过程中的数据得以完整搜集，包括授课的过程，作业的情况，学生的成绩，教师评价等数据。
3. 大数据的诞生让社会安全管理更为井然有序。
4. 大数据带动了社会上各行各业的发展。

1.3.3 大数据的挑战与风险

大数据正在催生以数据资产为核心的多种商业模式，革新生活模式，产生社会价值，引发积极影响。

与传统数据相比，数据量源源不断地增加，容易导致很多不正确的数据写入数据库。大数据技术具有在处理上的风险

大数据逐渐深入人们的生活。

1.4 大数据的应用

大数据技术已经被视为未来经济生活中的基础，这意味着几乎全部行业都能够在大数据分析技术之上获得经济效益的提升。

1. 电商大数据的应用

   大数据技术帮助电子商务行业发现新的商业模式。