第二章 大数据技术基础

2.1 云计算和 hoop 简介

大数据技术分类（1）

1 基础架构支持

2 数据采集

3 数据储存

4 数据计算

5 展现交互

基础结构支持

涉及概念：云计算和 Hadoop

云计算（cloud computing）是分布式计算的一种，指的是通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后，通过多部服务器组成的系统进行处理和分析这些小程序得到结果并返回给用户。

云计算与大数据的关系：从某种角度来说，没有计算机的云计算技术就不会有大数据的被分析和利用

云计算特点：1 超大规模 2 虚拟化 3 高可靠性 4 通用性 5 高扩展性 6 按需服务

云计算服务形式：IaaS 基础设施即服务

基础设施资源：计算资源；储存资源；网络资源；

SaaS 软件即服务

应用资源：办公服务；测试环境；

PaaS 平台即服务

平台资源：数据库；操作系统资源；

Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力进行高速运算和存储。

特点：1 高可靠性 2 高效性 3 高扩展性 4 高容错性 5 成本低 6 运行在 Linux 平台上 7 支持多种编程语言

其核心是分布式文件系统 HDFS 和 MapReduce

2.2 大数据采集与云处理

一、数据采集和预处理

（1）数据采集：是指从真实世界对象中获得原始数据的过程。

（2）数据预处理：可以使残缺的数据完整，并将错误的数据纠正，多余的数据去除，将所需的数据挑选出来，进行数据整合。

常见预处理的方法：数据变换、不一致检测和修复等

（3）数据集成：是把不同的数据在逻辑上或物理上有机地集中，通过一种表示法，对同一种实体对象的不同数据做整合的过程。

二、大数据的存储

大数据存储的特性：高度分散的，结构松散，并且体积越来越大。

大数据与传统数据对比

传统数据	大数据
千兆字节~百万兆字节	拍字节（PB）~ 艾字节（EB）
集中化	分布式
结构化	半结构化和无结构化
稳定的数据模型	平面模型
已知的复杂的内部关系	不复杂的内部关系

分布式储存系统

（1）分布式文件系统（DFS）：用户数据没有直接连接到本地主机，而是储存到远程服务器上。

（2）集群文件系统（CFS）：将若干个普通性能的储存系统联合起来，来组成

“储存的集群”，可以提供高性能、高可用的文件系统，可以消除单点故障，解决性能问题。

（3）并行文件系统（GFPS）：支持并行应用，所有客户端可以同一时间并发读写同一文件。

云储存

是通过集群应用、网格技术、分布式文件系统等将大量不同的储存设备通过应用软件集合起来协作，对外提供数据存储和业务访问

数据库

定义：按照数据结构来组织、存储和管理数据的建立在计算储存设备上的仓库。

云数据库：部署在云计算环境中的数据库，具有高可扩展性，高可用性。

数据仓库（DW 或 DWH）

是对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的。

特点：反应历史、面向主题、相对稳定、集成的、

2.3 大数据计算、分析与可视化

一、大数据计算

大数据计算框架:面向大数据处理的数据查询、统计、分析、挖掘等需求，促生了大数据计算的不同计算模式，常用的大数据计算框架有两种：离线批处理计算和实时流式处理计算

（1）离线批处理适用于数据在计算之前已经完全到位，不会发生变化，数据量巨大且保存时间长，在大量数据上进行复杂的批量运算。

（2）在很多实时应用场景中，比如实时交易系统、实时诈骗分析、实时广告推送、实时监控、社交网络实时分析等，实时性要求高，而且数据源是实时不间断的。新到的数据必须马上处理完，不然后续的数据就会堆积起来，永远也处理不完。反应时间经常要求在秒级以下，甚至是毫秒级，这就需要一个高度可扩展的流式计算解决方案。

两种算框架的对比

批量计算　　流式计算

数据到达　计算开始前数据已准备好　　计算进行中数据持续到来

计算周期　计算完成后会结束计算　一般会作为服务持续运行

使用场景　时效性要求低的场景　　时效性要求高的场景

二、大数据分析

(1)数据描述性分析

数据描述性分析关注于描述一组数据的统计特征，帮助我们了解数据分布特征、以及分散性和关联性等数字特征。

典型的统计指标:均值、方差、中位数、分位数等。

(2)数据挖掘和机器学习算法.

分类算法、聚类算法、关联规则算法、PageRank 算法、人工神经网络和深度学习方法、统计机器学习方法等.....。

(3) 预测分析

预测分析法是根据客观对象的已知信息而对事物在将来的某些特征、发展状况的一种估计、测算活动。

典型的算法:回归分析、时间序列预测法和因果关系预测。

(4)推荐系统

推荐系统根据用户的兴趣特点和购买行为，向用户推荐感兴趣的信息和商品。

典型的算法:协同过滤算法、基于内容的过滤算法和基于关联规则推荐算法。

(5)社会网络分析(Social Network Analysis)

社会网络分析被用来建立社会关系的模型，发现群体内行动者之间的社会关系，描述社会关系的结构，研究这种结构对群体功能或者群体内部个体的影响。

典型的应用:社会舆情分析、网络社区发现、情感.分析等。

三. 数据可视化

1.定义：利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。

可视化是理解、探索、分析大数据的重要手段。

2.数据可视化的工具和实例

数据可视化工具的类型包括图表生成工具、可视化报表、商业智能分

析、地图类和数据挖掘编程语言等。

3.常用的数据可视化工具

（1）纯可视化图表生成工具(适合开发，工程师) : Echart 和 AntV。

（2）可视化报表类(适合报表开发、BI 工程师) : FineReport。

（3）商业智能分析(适合 BI 工程师、数据分析师) : Tableau、FineBI 和 PowerBI。

（4）数据地图类: Power Map、Modest Maps 和地图慧。

（5）数据挖掘编程语言(适合技术性数据分析师、数据科学家) : R 和 P

整理人： 2.1 王佳伟 19760102

2.2 童效凯 19760101

2.3 周沛 19760104

语音文本 曾天煦 19760109