大数据方面学习

制作人：PPt创作者
时间：2024年X月

目录

第1章 简介
第2章 大数据存储技术
第3章 大数据处理技术
第4章 大数据分析技术
第5章 大数据应用场景
第6章 总结

01

第一章 简介

01

大数据量级

PB级或以上

02

数据类型

结构化与非结构化数据

03

数据处理速度

实时或批处理

分布式计算

概念

多台计算机协同工作

技术

MapReduce，Spark

优势

高性能，可靠性强

大数据处理流程

数据采集

传感器数据采集

日志数据采集

数据存储

HDFS

NoSQL数据库

数据处理

ETL过程

数据清洗

数据分析

数据挖掘

数据建模

大数据挑战

随着大数据规模的不断增长，数据安全性、隐私保护、可靠性和一致性等问题逐渐凸显。如何解决这些挑战成为大数据研究的重要内容。

大数据的应用领域

金融

风险控制、信贷评分

电商

推荐系统、用户行为分析

医疗

疾病诊断、基因研究

02

第2章 大数据存储技术

分布式文件系统

分布式文件系统是大数据存储技术中的重要组成部分，主要用于存储大规模数据。其中，HDFS是Apache Hadoop生态系统的一部分，提供高可靠性、高吞吐量的数据访问，Ceph是一个自由软件存储平台，支持对象存储、块存储、文件系统等。GlusterFS则是一个开源的分布式文件系统，可扩展到数PB级别的数据规模。

分布式文件系统

HDFS

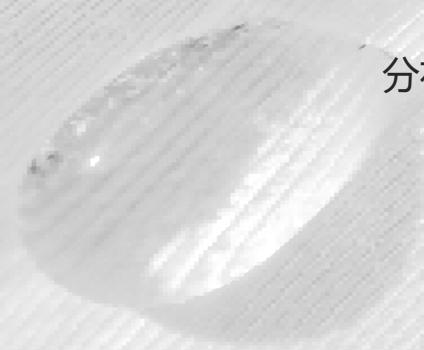
Hadoop分布式文件系统

GlusterFS

可扩展的分布式文件系统

Ceph

分布式存储平台



列式存储

列式存储是一种针对列而非行进行数据存储的技术，适用于读取单个列或列子集的查询。Cassandra是一款高可用的分布式数据库系统，HBase是建立在Hadoop之上的面向列的NoSQL数据库。这些系统能够实现海量数据的高效存储和查询。

列式存储

Cassandra

分布式数据库系统

HBase

面向列的NoSQL

数据库

内存数据库

内存数据库是将数据存储在内存中，提高数据读写性能的一种数据库技术。Redis是一个开源的内存数据库，支持多种数据结构的存储和操作；Memcached则是一个高性能的分布式内存对象缓存系统，用于加速动态Web应用程序。

内存数据库

Redis

开源内存数据库

Memcached

分布式内存缓存系统

数据仓库

数据仓库是用于集中存储和管理企业数据的系统，Amazon Redshift是一种快速、可扩展的数据仓库服务，适用于大规模数据分析；Google BigQuery则是一种云数据仓库，可实现大规模数据的快速查询和分析。

数据仓库

Amazon Redshift

快速可扩展的数据仓库服务

Google BigQuery

云数据仓库系统

● 03

第三章 大数据处理技术

批处理技术

批处理是大数据处理的一种常见方式，常用的技术包括MapReduce、Apache Spark和Apache Flink。MapReduce是Google提出的一种分布式计算框架，适用于大规模数据处理；Apache Spark是基于内存计算的大数据处理框架，具有高性能；Apache Flink是一个分布式流式数据处理引擎，支持事件时间处理。

流式处理技术

Apache Kafka

分布式流式数据传输系统

Samza

LinkedIn开发的流处理框架

Storm

开源流式计算系统