1-7-cn

1

大家好，北京理工大学计算机学院，我是数据科学与知识工程研究所的车海莺，

在本节中，我们讨论大数据通用架构

2

当数据量越来越大时，任何一台传统的高性能服务器都不能满足需求，需要更多的服务器。

但是传统的高性能服务器非常昂贵，可能要上亿，因为它非常可靠，像IBM Z系列服务器，

可靠性可以达到99.9999%，号称零宕机，所以叫Z系列。高性能和高可靠性意味着高价格。

但没有任何一家公司能够负担得起如此昂贵的服务器来处理大数据问题。

所以分布式计算被用来解决这个问题。

分布式计算的思想是用一组廉价的服务器组成一个服务器组，来代替昂贵的高性能服务器，这些服务器可以是分布式处理、分布式存储。并使用高冗余来实现高性能。

关键的硬件和软件突破彻底改变了数据管理产业。

首先，创新和需求增加了数据处理能力，降低了硬件的价格。

能够通过自动化流程来优化使用服务器群的软件诞生了，比如可以跨庞大节点集群的负载平衡和优化硬件性能。

这种软件包含内置规则，可以知道某些工作负载需要达到一定的性能水平。

该软件将所有节点简单的当做由计算、存储和网络资产等资源组成的一个资源池，如果一个节点发生故障，使用虚拟化技术不中断的将进程转移到另一个节点。

3

数据需要存储，然后处理，中间和最终结果都需要进行存储。

在大数据的范围内，当单个节点很难完成超大数据规模的数据存储和数据处理时，需要分布式计算，多个节点协同完成大规模数据存储和处理任务。

大数据有两个主要组成部分，分布式存储和分布式处理。这两个组件协同完成大数据分析的任务。

4

一般大数据架构包括3层，从下到上，

1) 数据存储系统包括数据收集和建模，负责根据预定义的模型收集数据，分布式文件系统和数据分布式数据库/Datawarehouse存储非结构化数据，将结构化数据存储在关系数据库中。

2) 数据处理系统包括计算引擎、计算平台、计算模型和算法。基于存储的数据，我们需要设计计算模型、算法，还需要MapReduce，这样的大数据分布式计算引擎，以及Hadoop、spark、Storm等计算平台，提供数据处理相关功能。

3) 基于数据存储和数据处理结果，可以构建大数据应用系统，包括大数据应用、数据产品和服务，以及数据可视化。在处理系统的帮助下，所有需要的数据都已经被计算和建模，上层应用程序可以将数据可视化结果呈现或给出决策建议，或支持做出明智的决策。

5

存储结构上：数据库提供数据的逻辑存储结构；分布式文件系统提供数据的物理存储结构。数据存储系统可以包括几个部分。

1) 数据采集层：　　系统日志、网络爬虫、无线传感器网络、物联网和各种数据资源，我们可以从中收集数据。

2.) 数据清洗、提取和建模然后我们可以将来自不同来源的各类结构化、非结构化和异构数据转换为标准存储格式的数据，并定义数据属性和取值范围，为数据分析做准备。

3)、数据存储架构；　集中式或分布式文件系统、关系型数据库或分布式数据库、基于行的存储数据结构或基于列的存储数据结构、键值对结构、哈希表检索等。

数据科学家可以选择合适的数据存储架构来存储数据，使数据存储和数据检索更加方便。

4).统一数据访问接口等应用程序对数据库的访问和数据交换是分布式计算系统中的一个重要问题。

6
应用程序对数据库的访问及数据交换是分布式计算系统的一个重要问题。业界较早使用的是微软公司提供的数据库访问应用程序编程接口，ODBC Open DataBase Connectivity，它采用 X/open 和 ISO/IEC 的调用接口（CLI call-level interface）标准为基础，使用结构化查询语言，SQL 作为数据库访问语言。ODBC 本质上是一组数据库访问 API， 有一组函数调用组成，核心是 SQL 语句。一个基于 ODBC 的应用程序对数据库进行操作时， 用户直接将 SQL 语句传送给 ODBC， 同时 ODBC 对数据库的操作也不依赖于任何的 DBMS， 不直接和 DBMS 打交道， 所有数据库操作由对应的 ODBC 驱动程序完成。
而 ODBC、JDBC 等数据库连接编程接口虽然可以支持应用程序对数据库的 SQL 访问，但无法在分布式计算环境中提供事务管理、并发调度、缓冲区管理、异构数据库转换和继承等复杂功能。
这 j 就需要引入了数据访问层。DAL 是在数据库之上提供数据交换功能的软件层。它的功能主要是实现应用程序数据的持久化存储，即将数据写入数据库，从数据库中读取数据传给应用程序，实现数据交换。
具体来说，它提供支持数据库的 CRUD（创建检索更新和删除）的基本操作。
交易管理，
并发处理
异构数据转换。
当系统扩展为需要访问跨平台的异构数据库时，系统可能是 UNIX、Linux 或 Windows，
数据可以是表单、邮件、XML 文档、EJB 组件、Web 服务、图像、音视频文件或其他非结构化数据，
DAL 很难支持这种跨平台的异构数据库访问。
而大数据应用层的技术也是多元化的，遵循着各种标准。数据访问层 DAL 的设计需要兼容各种标准技术和产品，因此需要统一数据访问接口
7
不同类型数据的计算模型，比如海量数据的 MapReduce 批处理模型，动态数据流的流计算模型，结构化数据的大规模并发处理（MPP-Massively Parallel Processing）模型，以及大规模物理内存内存计算模型；
支持机器学习算法的数据流图模型；各种分析算法的实现，以及提供各种开发包和运行环境的计算平台如 Hadoop、Spark、Storm 等。
计算引擎为基于计算平台为特定计算模型而设计和封装的服务器端程序，用于支撑特定计算模式下的后端的大数据处理、计算和分析任务，
比如， MapReduce 计算引擎提供大数据的划分，节点分配，作业调度及计算结果融汇等功能， 直接支持上层应用的开发。
Google 的交互式计算引擎采用 Dremel，PowerDrill 技术，提供了对大规模数据集的快速计算分析；
开源的 apache drill 项目基于列存储结构、数据本地化、内存存贮等技术力图实现对大规模数据的快速查询访问。
图并行计算引擎提供对网络图数据（社交网络、电信网络、脑功能连接网络这一类数据常常可用权重有向图来表征）的高效计算处理（google 搜索引擎处理的数据量中有 20%是用图计算引擎来处理），
这方面技术包括 Google 的 Pregel，开源技术的 Hama、GraphLab 等
S4(Simple, Scalable Streaming System) 是 Yahoo！提供的一个分布式流计算引擎，最初目标是提高 cost-per-click 广告点击率问题，通过实时数据计算预测用户对广告的可能的点击行为

8
基于上述计算架构和处理平台，提供了各行业、各领域的大数据应用技术解决方案。
目前，互联网、电子商务、电子政务、金融、电信、医疗卫生等行业是大数据应用最热门的领域，
而制造、教育、能源、环保、智能交通是大数据技术将会或已经开始拓展产业的领域。

9看完大数据的总体架构，我们再从另一个逻辑的角度来看大数据层次。
在底层，收集网络数据、收集探测器数据、收集服务数据、收集业务数据、收集终端数据。
在中间层，分析用户行为，监控关键指标，利用模型和算法对其他数据进行整合分析。 并将处理后的数

据通过统一的数据平台提供给上层数据服务。
在顶层，基于统一数据平台提供的分析数据，用户可以进行不同类型的分析，业务流程分析，用户分析数据分析和数据输出可视化。

10
我告诉过你大数据不等于 Hadoop。 Hadoop 只是它的子集。有更多区域和特定项目是该区域的一部分。
你看到黄色大象了吗？这是一个 Hadoop 徽标。你知道它的历史吗？
1 许多人认为大数据与 Hadoop 技术有关。它是，也不是。它比 Hadoop 多得多。
关键要求之一是理解和导航大数据的联合来源——以发现数据。
发现、索引、搜索和导航各种大数据源的新技术已经出现。
2 当然，大数据也与 Hadoop 有关。 Hadoop 是开源功能的集合。其中最突出的两个是用于存储各种信息的 Hadoop 文件系统，和 MapReduce – 一个并行处理引擎。
3 数据仓库还管理大数据——结构化数据量正在快速增长。对大量结构化数据运行深度分析查询的能力是一个大数据问题。它需要大量并行处理数据仓库和专用设备来进行深度分析。
4 大数据不仅处于静止状态，而且还在运动中。流数据代表了一个完全不同的大数据问题-在数据仍在移动的同时快速分析和处理数据的能力。这项新技术开启了一个充满可能性的世界-从处理无法存储的大量数据，到检测洞察力和快速响应。
5 由于世界上大部分的大数据都是非结构化的并且是文本内容，因此文本分析是分析文本并从文本中获取意义的关键组成部分。
6 最后，集成和治理技术——ETL、数据质量、安全性、MDM-主数据管理和生命周期管理。
集成和治理技术确立了大数据的真实性，对于确定信息是否可信至关重要。
大数据目前成为热门话题的部分原因在于这项技术——现在成为可能

11
在本节中，我们讨论了大数据 3 层架构 ， 数据存储系统；数据处理系统；数据应用系统；感谢您的关注如果您有任何问题，请随时与我联系。