5-1

1 大家好，我是来自北京理工大学计算机学院，数据科学与知识工程研究所的车海莺，在这节课我们讨论 数据处理架构。

2 数据处理系统提供大数据计算处理能力和应用开发平台。从计算架构的角度，将数据处理系统分为算法层、计算模型层、计算平台和引擎层等。

与大数据相关的计算算法包括机器学习算法和数据挖掘算法。计算模型是指不同类型的大数据在不同场景下的处理方式，包括批处理、流计算、结构化数据的大规模并发处理(MPP)模型、内存计算模型和数据流图模型。

3 大数据处理算法基本上包括数据挖掘算法和机器学习算法。 数据挖掘算法可分为 4 组:分类、聚类、相关分析和异常检测。 机器学习算法包括监督学习、无监督学习、半监督学习和强化学习。 监督学习的范围包括回归、分类和深度学习。 在无监督学习的范围内，有聚类、降维和异常检测。 半监督学习包括自我训练和低密度分离模型。 强化学习包括动态编程和蒙特卡罗方法。

4 **针对不同类型数据的计算模型**

基本的大数据处理从批处理开始，以 MapReduce 为代表。

批处理是一次处理大量数据。 数据很容易由每天数百万条记录组成，并且可以以各种方式存储(文件、记录等)。 这些工作通常以不间断的顺序同时完成。

批处理作业的一个常用示例是金融公司在一周内可能提交的所有事务。批处理还可以用于：工资流程；产品的发票；供应链与实施；

批处理数据是处理在一段时间内收集的大量数据的一种极为有效的方法。它还有助于降低企业可能花费在劳动力上的运营成本，因为它不需要专门的数据输入职员来支持其功能。

它可以离线使用，经理可以完全控制何时开始处理，无论是在夜间还是在一周或发薪期结束时。

流处理是能够几乎瞬间分析从一个设备流到另一个设备的数据的过程。 这种连续计算方法在数据流过系统时进行，对输出没有强制的时间限制。 由于流量几乎是即时的，系统不需要存储大量的数据。 如果您希望跟踪的事件频繁发生，则流处理非常有用。 如果需要立即检测事件并快速响应，最好使用此方法。 因此，流处理对于欺诈检测和网络安全等任务非常有用。 如果交易数据是流处理的，欺诈交易甚至可以在完成之前就被识别和停止。

5 **大规模并行处理-MPP** DB 架构

MPP 采用完全并行的 MPP + Shared Nothing 的分布式扁平架构，这种架构中的每一个节点（node）都是独立的、自给的、节点之间对等，而且整个系统中不存在单点瓶颈，具有非常强的扩展性。

MPPDB

MPPDB 是一款 Shared Nothing 架构的分布式并行结构化数据库集群，具备高性能、高可用、高扩

展特性，可以为超大规模数据管理提供高性价比的通用计算平台，并广泛地用于支撑各类数据仓库系统 BI 系统和决策支持系统

大规模并行处理（MPP，massively parallel processing）是多个处理器（processor）处理同一程序的不同部分时该程序的协调过程，工作的各处理器运用自身的操作系统（Operating System）和内存。大规模并行处理器一般运用通讯接口交流。在一些执行过程中，高达两百甚至更多的处理器为同一应用程序工作。数据通路的互连设置允许各处理器相互传递信息。一般来说，大规模并行处理（MPP）的建设很复杂，这需要掌握在各处理器间区分共同数据库和给各数据库分派工作的方法。大规模并行处理系统也叫做"松散耦合"或"无共享"系统。

一般认为，对于允许平行搜索大量数据库的应用程序，大规模并行处理（MPP，massively parallel processing）系统比对称式并行处理系统（SMP Symmetric Multiprocessing ）更好。这些包括决策支持系统（decision support system）和数据仓库（data warehouse）应用程序。

6 服务器架构类型

服务器架构有三种类型，1 对称多处理器结构（SMP：Symmetric Multi-Processor） UMA (Uniform Memory Access) 2 非一致存储访问结构（NUMA：Non-Uniform Memory Access），3 大规模并行处理结构（MPP：Massive Parallel Processing)

对称多处理器结构 SMP，Symmetric Multiprocessing， 相对应还有 asymmetric multiprocessing (AMP)；一致性内存访问，均匀存储器存取 UMA (Uniform Memory Access)系统是一种多处理器共享的内存架构。Nonuniform-Memory-Access，简称 NUMA 非均匀存储器存取

MPP 服务器架构

它由多个 SMP 服务器通过一定的节点互联网络进行连接，协同工作，完成相同的任务，从用户的角度来看是一个服务器系统。其基本特征是由多个 SMP 服务器(每个 SMP 服务器称节点)通过节点互联网络连接而成，每个节点只访问自己的本地资源(内存、存储等)，是一种完全无共享(Share Nothing)结构，因而扩展能力最好，理论上其扩展无限制。

7 在该模型中，所有处理器都使用一个内存，并通过互联网络实现多处理器系统。每个处理器具有相同的内存访问时间(延迟)和访问速度。它可以采用单总线、多总线或交叉开关中的任意一种。由于它提供平衡的共享内存访问，它也被称为 SMP(对称多处理器)系统。

SMP 服务器的主要特征是共享，系统中所有资源（CPU 、内存、 I/O 等 ）都是共享的。也正是由于这种特征，导致了 SMP 服务器的主要问题，那就是它的扩展能力非常有限。对于 SMP 服务器而言，每一个共享的环节都可能造成 SMP 服务器扩展时的瓶颈，而最受限制的则是内存。由于每个 CPU 必须通过相同的内存总线访问相同的内存资源，因此随着 CPU 数量的增加，内存访问冲突将迅速增加，最终会造成 CPU 资源的浪费，使 CPU 性能的有效性大大降低。一些博客上关于 cpu 个数和利用率方面有相关说明：实验证明， SMP 服务器 CPU 利用率最好的情况是 2 至 4 个 CPU 。

8 "UMA"是"Uniform Memory Access 一致性内存访问，均匀存储器存取

传统的多核运算是使用 SMP(Symmetric Multi-Processor)模式：

将多个处理器与一个集中的存储器和 I/O 总线相连。所有处理器只能访问同一个物理存储器，因此 *SMP 系统有时也被称为一致存储器访问（UMA）结构体系*，一致性指无论在什么时候，处理器只能**为内存的每个数据保持或共享唯一一个数值**。 物理存储器被所有 CPU 均匀共享。所有处理机对所有存储器具有相同的存取时间，这就是为什么称它为均匀存储器存取的原因。 每台处理机可以有私用高速缓存 Cache, 外围设备也以一定形式共享。显然，SMP 的缺点是可伸缩性有限，因为在存储器和 I/O 接口达到饱和的时候，增加处理器并不能获得更高的性能。

9 非均匀访问存储模型 NUMA （ Non-Uniform Memory Access)

非均匀访问存储模型，这种模型的是为了解决 SMP 扩容性很差而提出的技术方案，如果说 SMP 相当于多

个 CPU 连接一个内存池导致请求经常发生冲突的话，

NUMA 就是将 CPU 的资源分开，以 node 为单位进行切割，每个 node 里有着独有的 core ，memory 等资源，这也就导致了 CPU 在性能使用上的提升，设计原理就是访问本地资源（本地内存、I/O 槽口）的速度远远高于访问远地资源（其他 node 的资源）的速度，**多个 node 之间的资源交互非常慢**，当 CPU 增多的情况下，性能提升的幅度并不是很高，无法实现性能的线性增加。很多 core 的服务器却只有 2～4 个 node 区。

10 非均匀内存访问，NUMA (Non-uniform Memory Access)

NUMA(非均匀内存访问)也是一个多处理器模型，其中每个处理器都与专用内存连接。它允许通过使用物理地址访问任何内存位置。非统一内存访问(NUMA)是当今多处理系统中使用的一种共享内存体系结构。每个 CPU 被分配给自己的本地内存，并且可以访问系统中其他 CPU 的内存。本地内存访问提供了低延迟-高带宽的性能。而访问另一个 CPU 拥有的内存有较高的延迟和较低的带宽性能。但是，这些内存的小部分组合在一起形成一个地址空间。这里要考虑的重点是，与 UMA 不同的是，内存的访问时间依赖于放置处理器的距离，这意味着改变内存访问时间。

11 **大规模并行处理 MPP (Massive Parallel Processing)**

MPP (Massive Parallel Processing) ，大规模并行处理系统，这样的系统是由许多松耦合的**处理单元**组成的，**处理单元而不是处理器**。每个单元内的 CPU 都有自己私有的资源，如总线，内存，硬盘等。在每个单元内都有操作系统和管理数据库的实例复本。这种结构最大的特点在于**不共享资源。**

和 NUMA 不同， MPP 提供了另外一种进行系统扩展的方式，它由多个 SMP 服务器通过一定的节点互联网络进行连接，协同工作，完成相同的任务，从用户的角度来看是一个服务器系统。其基本特征是由多个 SMP 服务器（ 每个 SMP 服务器称节点 ）通过节点互联网络连接而成，每个节点只访问自己的本地资源（ 内存、存储等 ），是一种完全无共享 (Share Nothing) 结构，因而扩展能力最好，理论上其扩展无限制，目前的技术可实现 512 个节点互联，数千个 CPU 。目前业界对节点互联网络暂无标准，如 NCR 的 Bynet ， IBM 的 SPSwitch ，它们都采用了不同的内部实现机制。但节点互联网仅供 MPP 服务器内部使用，对用户而言是透明的。

在 MPP 系统中，每个 SMP 节点也可以运行自己的操作系统、数据库等。但和 NUMA 不同的是，它不存在异地内存访问的问题。换言之，每个节点内的 CPU 不能访问另一个节点的内存。节点之间的信息交互是通过节点互联网络实现的，这个过程一般称为数据重分配 (Data Redistribution) 。

但是 MPP 服务器需要一种复杂的机制来调度和平衡各个节点的负载和并行处理过程。目前一些基于 MPP 技术的服务器往往通过系统级软件（ 如数据库 ）来屏蔽这种复杂性。举例来说， NCR 的 Teradata 就是基于 MPP 技术的一个关系数据库软件，基于此数据库来开发应用时，不管后台服务器由多少个节点组成，开发人员所面对的都是同一个数据库系统，而不需要考虑如何调度其中某几个节点的负载。

12 **SMP、NUMA 和 MPP 比较**

从架构来看，NUMA 与 MPP 具有许多相似之处：它们都由多个节点组成，每个节点都具有自己的 CPU、内存、I/O，节点之间都可以通过节点互联机制进行信息交互。

那么它们的区别在哪里？

首先是节点互联机制不同，NUMA 的节点互联机制是在同一个物理服务器内部实现的，当某个 CPU 需要进行远地内存访问时，它必须等待，这也是 NUMA 服务器无法实现 CPU 增加时性能线性扩展的主要原因。

而 MPP 的节点互联机制是在不同的 SMP 服务器外部通过 I/O 实现的，每个节点只访问本地内存和存储，节点之间的信息交互与节点本身的处理是并行进行的。因此 MPP 在增加节点时性能基本上可以实现线性扩展。

其次是内存访问机制不同。在 NUMA 服务器内部，任何一个 CPU 可以访问整个系统的内存，但远地访问的性能远远低于本地内存访问，因此在开发应用程序时应该尽量避免远地内存访问。在 MPP 服务器中每个节点只访问本地内存，不存在远地内存访问的问题。

数据仓库的选择哪种服务器更加适应数据仓库环境？这需要从数据仓库环境本身的负载特征入手。众所周知，典型的数据仓库环境具有大量复杂的数据处理和综合分析，要求系统具有很高的 I/O 处理能力，

并且存储系统需要提供足够的 I/O 带宽与之匹配。而一个典型的 OLTP 系统则以联机事务处理为主，每个交易所涉及的数据不多，要求系统具有很高的事务处理能力，能够在单位时间里处理尽量多的交易。显然这两种应用环境的负载特征完全不同。从 NUMA 架构来看，它可以在一个物理服务器内集成许多 CPU，使系统具有较高的事务处理能力，由于远地内存访问时延远长于本地内存访问，因此需要尽量减少不同 CPU 模块之间的数据交互。显然，NUMA 架构更适用于 OLTP 事务处理环境，当用于数据仓库环境时，由于大量复杂的数据处理必然导致大量的数据交互，将使 CPU 的利用率大大降低。相对而言，MPP 服务器架构的并行处理能力更优越，更适合于复杂的数据综合分析与处理环境。当然，它需要借助于支持 MPP 技术的关系数据库系统来屏蔽节点之间负载平衡与调度的复杂性。另外，这种并行处理能力也与节点互联网络有很大的关系。显然，适应于数据仓库环境的 MPP 服务器，其节点互联网络的 I/O 性能应该非常突出，才能充分发挥整个系统的性能。但这并不是绝对的，性能的好坏由很多因素组成，比如 Oracle Exadata， 它没有使用 MPP 架构，但性能是相当的优越了。所以单从服务器的一个方面分析性能有一定的片面性，而现在的趋势是整体的从多方面（包括软件层面）优化服务器的性能。

13 **各个体系结构区别**

性能方面：

NUMA 的节点互联机制是在同一个物理服务器内部实现的，当某个 CPU 需要进行远地内存访问时，它必须等待，这也是 NUMA 服务器无法实现 CPU 增加时性能线性扩展。

MPP 的节点互联机制是在**不同的 SMP 服务器外部通过 I/O 实现的**，每个节点只访问本地内存和存储，**节点之间的信息交互与节点本身的处理是并行进行的**。因此 MPP 在增加节点时性能基本上可以实现线性扩展。

SMP 所有的 CPU 资源是共享的，随着 CPU 的增加，内存访问冲突加剧，造成 CPU 资源浪费，使 CPU 的性能有效性打打降低。因此**没办法**完全实现线性扩展。

扩展方面：

NUMA 理论上可以无限扩展，目前技术比较成熟的能够支持上百个 CPU 进行扩展。如 HP 的 SUPERDOME。 MPP 理论上也可以实现无限扩展，目前技术比较成熟的能够支持 512 个节点，数千个 CPU 进行扩展。

SMP 扩展能力很差，目前 2 个到 4 个 CPU 的利用率最好，但是 IBM 的 BOOK 技术，能够将 CPU 扩展到 8 个。 MPP 是由多个 SMP 构成，多个 SMP 服务器通过一定的节点互联网络进行连接，协同工作，完成相同的任务。

SMP 的优势：

MPP 系统因为要在不同处理单元之间传送信息，所以它的效率要比 SMP 要差一点。在通讯时间多的时候，那 MPP 系统可以充分发挥资源的优势。因此当前使用的 OLTP 程序中，用户访问一个中心数据库，如果采用 SMP 系统结构，它的效率要比采用 MPP 结构要快得多。

NUMA 架构的优势：一个物理服务器内集成许多 CPU，使系统具有较高的事务处理能力，由于远地内存访问时延远长于本地内存访问，因此需要尽量减少不同 CPU 模块之间的数据交互。显然，NUMA 架构更适用于 OLTP 事务处理环境，当用于数据仓库环境时，由于大量复杂的数据处理必然导致大量的数据交互，将使 CPU 的利用率大大降低。

MPP 的优势：MPP 系统不共享资源，因此对它而言，资源比 SMP 要多，当需要处理的事务达到一定规模时，MPP 的效率要比 SMP 好。由于 MPP 系统因为要在不同处理单元之间传送信息，在通讯时间少的时候，那 MPP 系统可以充分发挥资源的优势，达到高效率。也就是说：操作相互之间没有什么关系，处理单元之间需要进行的通信比较少，那采用 MPP 系统就要好。因此，MPP 系统在决策支持和数据挖掘方面显示了优势。

14 内存计算

内存计算是一种完全在计算机内存(如 RAM)中运行计算的技术。这个术语通常意味着大规模、复杂的计算，需要专门的系统软件在集群中一起工作的计算机上运行计算。

作为一个集群，计算机将它们的 RAM 聚集在一起，因此计算基本上是在计算机之间运行的，并一起利用所有计算机的集体 RAM 空间。

内存计算通过消除所有缓慢的数据访问并完全依赖存储在 RAM 中的数据来工作。
通过消除访问硬盘驱动器或 SSD 时常见的延迟，整体计算性能大大提高。
运行在一台或多台计算机上的软件管理计算和内存中的数据，在多台计算机的情况下，软件将计算分成更小的任务，这些任务分布到每台计算机上并行运行。

15 **并行图计算模型**

针对大量数据计算，需要并行，针对大规模图的计算也需要并行模型
图是实体群之间联系的有用的理论表示，在数据科学中被用于各种目的，从根据受欢迎程度对网页进行排名，绘制社交网络，到辅助导航。
在许多情况下，这类应用程序需要处理包含数千亿条边的图，这些边太大了，无法在一台普通机器上处理。
缩放图算法的典型方法是在分布式环境中运行，也就是说，在多台计算机中划分数据(和算法)以并行地执行计算。 虽然这种方法允许处理具有数万亿条边的图，但它也带来了新的挑战。也就是说，因为每台计算机每次只能看到输入图的一小部分，所以需要处理机间通信，并设计可以跨多台计算机分割的算法。

16 **数据处理系统**

应用数据算法层算法，结合批处理、流处理、大规模并行处理、内存计算或图计算模型，我们可以处理大数据问题，但我们很难自己实现所有的算法和计算模型， 计算平台和计算引擎层可以为我们提供平台和引擎，包括所需的工具，库，以帮助我们实现复杂的算法和计算模型

17 **计算平台和计算引擎**

计算平台与引擎指为大数据计算分析提供可技术标准、计算架构，以及一系列开发技术和工具的开发集成环境。目前有代表性的计算平台有：Hadoop，Cloudera，Spark，Storm 以及 Google 基于其一系列大数据计算技术的商业平台。

18 本节我们概述了数据处理架构，今天内容就到这里，谢谢大家