4-1-cn

1

大家好，我是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，从本节开始，我们学习数据存储系统，本节我们讨论数据存储系统中的数据建模.

2

大数据计算系统可以概括为三类：数据存储系统、数据处理系统、数据应用系统

现在让我们看看基础 系统数据存储系统。数据存储架构是大数据计算的基础。

上层分析算法， 计算模型和计算性能都取决于数据存储系统的性能。

因此，数据存储系统是大数据研究的一个重要领域。

在数据存储系统中，有 4 个部分完成不同的任务，即数据收集和数据建模、分布式文件系统、分布式数据库、数据仓库和统一数据访问接口。

数据采集从多种数据资源中采集数据，如系统日志、网络爬虫、无线传感器网络、物联网等各种数据资源。

数据采集完毕后，我们需要清理数据，删除重复数据、空数据、错误数据等脏数据。

将各类结构化、非结构化、异构数据转换为标准存储格式数据，定义数据属性和取值范围。

我们建立数据模型。在数据收集之上，是文件系统，物理上实现数据的存储，可以是集中式或分布式文件系统，

数据库包括结构化的 RDBM 和非结构化的 No SQL 数据库来设计存储数据的逻辑结构。

以上所有的文件系统和数据库都是通过 UDAI 向数据处理系统提供数据，它是一个数据访问接口，数据处理系统可以从数据存储系统中检索数据。

1） 统一数据访问接口

2） 将各种类型的结构化、非结构化、异构数据转化为标准存储格式数据，并定义数据属性及值域

3） 集中式或分布式文件系统，关系型数据库或分布式数据库、行存储数据结构或者列存储数据结构，键值对结构，哈希表检索等。

4） 数据统一接口

数据存储架构是大数据计算的基础，上层各种分析算法、计算模型及计算性能都依赖于数据存储系统的表现，因此数据存储系统是大数据研究的一个重要的领域

3

数据存储主要提供数据采集、清洗建模。大规模数据存储管理、数据操作（添加，删除，查询，更新。数据同步等）功能

由于大数据处理的多重数据源/数据异构性/非结构化数据，分布式计算环境等特点，大数据存储系统的设计比关系型数据库系统复杂。

目前的大数据存储架构主要由数据层，分布式文件系统/非关系型数据库（NoSQL）以及统一数据库读取界面组成， 有些设计还会再 NoSQL 数据库之上加一个提供数据挖掘和分析功能的数据仓库层

可以从 CRM、ERP、财务、社交媒体、排气数据、日志、文件中收集数据。

（data exhaust 或 exhaust data 是互联网或其他计算机系统用户在其在线活动、行为和交易期间的活动留下的数据痕迹。这是更广泛的非常规数据类别的一部分，包括地理空间、网络和时间序列数据，可能对预测分析有用。

每个访问过的网站、点击过的链接，甚至是鼠标悬停都会被收集起来，留下一连串的数据。）

4

数据清洗与建模：

数据层主要包含数据采集系统并提供数据抽取，清洗，与转换 ETL，数据建模功能，

大数据应用面对多种数据源

（企业数据、商务数据、个人社交数据、政府统计数据，互联网数据，物联网数据，

系统日志数据，基因测序数据，大气物理监测数据，地球卫星观测数据等）

异构数据（文本，图片，音频和视频）、
非结构化数据（医学影像资料，银行凭证扫描件，碎片化通信记录，截屏等）的特点使得原始数据很难直接存入数据库，
经常遇到问题，原始数据格式不能被数据平台识别和处理，很多情况先原始数据还存在记录缺失，值域缺损，数据质量参差不齐等问题
这就要求在构建数据库或数据仓库之前对原始数据完成清洗，（合并或者去除重复数据项，消除数据错误）、抽取（从多个数据源的
数据项中抽取不同值域构成目标数据库的数据结构，或从一个数据源抽取数据项分解成多个结构装载入目标数据库）、
转换（将不同格式的原始数据项转换为统一标准的目标数据库格式）等步骤。数据抽取、清洗和转换可以是人工或采用软件工具的方式完成。
5
在大数据采集中，存在一些问题：
数据平台无法识别和处理原始数据格式。
在很多情况下，原始数据仍然存在缺失记录、缺失值范围、不同级别数据质量等问题。
清洗
这需要在构建数据库或数据仓库之前清理原始数据，通过合并或删除重复数据项，消除数据错误，
提取
可以从多个数据源中提取数据，从数据项中提取不同的取值范围，形成目标数据库的数据结构，
或者可以从一个数据源中提取数据项并将它们分解为多个结构并加载到目标数据库中，
转换是指将不同格式的原始数据项转换成统一的标准目标数据库格式。
数据提取、清理和转换可以手动完成，也可以使用软件工具完成。
6
数据建模是数据层工作的一个重要内容，数据建模是对实体数据（或用户对数据功能的描述）建立一个抽象模型，包括元数据，数据结构，属性，值域，关联关系，一致性，时效性等元素。
数据模型为进一步的数据存储结构设计，数据库设计和计算模型提供了参考依据。
7
业务模型通常包括业务流程模型和数据模型，
业务流程模型描述了业务是如何进行的。
数据模型描述了业务流程中产生了哪些数据，需要存储哪些数据，即支持业务流程的数据。
数据模型从抽象到具体分为三个层次，
即概念模型、逻辑模型和物理模型。
概念模型和逻辑模型是数据组织的逻辑模型。
物理模型是关于数据如何物理存储的。
8
在这个流程图中。 首先根据业务流程模型和数据需求进行逻辑建模，然后生成逻辑数据模型作为输出。
综合逻辑数据模型、技术要求和性能要求，进行物理数据建模，生成物理数据模型。
在我们有了逻辑数据模型并实现为对应的物理模型之后，
业务系统运行过程中产生的业务数据就可以通过在逻辑和物理模型中创建、更新操作来存储。
9
数据建模过程让我们一一看数据建模。根据用户的数据功能需求。得到函数和关联关系，我们可以找到对应业务元素和功能的 Entity Class。
1）概念模型模式描述了域的语义（模型的范围）。例如，它可能是一个组织或一个行业的兴趣领域的模型。
这包括实体类，代表领域中的各种重要事物，以及关于实体类对之间关联的关系断言。概念模式指定可以使用模型表达的事实或命题原告的种类。

从这个意义上说，它以模型有限范围的人工“语言”定义了允许的表达式。简单地说，概念模式是组织数据需求的第一步。
2）在逻辑模型设计中，数据实体的更多细节，包括主键、外键、属性、索引、关系、约束、甚至视图，用数据表、数据列、值域、面向对象的类、XML 标签等形式来描述。
逻辑模型模式描述了某些信息域的结构。这包括（例如）表、列、面向对象的类和 XML 标记的描述。逻辑模式和概念模式有时被实现为一体。
3） 物理模型（也称为存储模型）描述了数据的存储实现，包括数据分区、数据表空间和数据集成。物理中间模式：描述用于存储数据的物理方式。这与分区、CPU、表空间等有关。

10
ANSI 美国国家标准学会（AMERICAN NATIONAL STANDARDS INSTITUTE: ANSI）
强调，上述三个层次模型之间是相对独立的，即物理模型的改变（数据存储方式的改变，数据划分的调整等）不影响逻辑模型和概念模型的内容；
逻辑模型的改变（数据表修改、属性的增减，值域的调整等）不影响概念模型的定义。
在进行数据集成和数据库实现时，要注意三个层次数据模型描述和定义的不一致性。
UML 是最常用的数据模型建模语言，常见的数据建模工具 Power designer、ER/studio.CA Erwin, IBM Infosphere Data Architect 等
11
让我们总结一下，在这节课中，我们学习了数据存储系统的第一层，包括数据采集、提取、转换和建模。

在数据建模中，有概念模型、逻辑模型和物理模型三个层次。
12
这节课我们就学习到这里，谢谢大家