1-2

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，本节我们讨论大数据基本概念中的结构化数据和非结构化数据

2

据 Gartner 估计，组织的 20% 数据是结构化数据，其余大部分是非结构化数据。

结构化数据通常是有行有列的表格形式，如左上角的表格，是一个在不同列中包含属性 id、名字、姓氏和年龄的表格。

结构化数据以预定义的格式组织数据，因此很容易处理，因为我们知道数据结构，我们可以设计相应的算法来处理它。

但是对于非结构化数据，所有的数据都没有统一的格式，处理起来比较困难。

我们先来谈谈结构化数据的特点。

结构化数据可以显示在行、列和关系数据库中。

数据类型可以是数字、日期和字符串。

结构化数据需要更少的存储空间。并且使用遗留解决方案更易于管理和保护。因为我们在处理结构化数据方面积累了丰富的经验。

非结构化数据不能在行、列和关系数据库中显示。它们通常是图像、音频、视频、文字处理文件、电子邮件、电子表格。

根据 Gartner 统计，它们占企业数据的 80%。非结构化数据需要更多存储空间，因为它们数量庞大且组织不善。而且它们很难以传统方式使用遗留解决方案进行管理和保护。

3

在该图中，数据分为四组，结构化、半结构化、准结构化和非结构化数据。

从下往上，

1）结构化数据具有预定义的数据模型、格式、结构，例如数据库中的数据。

2） 这类数据有些结构化，但并不完整。 起初这似乎是非结构化的，并且不遵循数据模型（如 RDBMS）的任何正式结构。

例如，NoSQL 文档具有用于处理文档的关键字。

半结构化数据，例如具有明显模式的文本数据文件，可以进行分析。例如，电子表格和 XML 文件。CSV 文件也被视为半结构化数据。

3）格式不稳定的文本数据，可以通过努力和软件工具进行格式化，例如点击流数据。经过一些处理，它可以被格式化。

4）非结构化数据没有固有结构，通常存储在不同类型的文件中，如文本文档、PDF、图像和视频。

4

结构化数据就像冰山的表面部分，非结构化的数据就像冰山的水下隐藏部分，需要探索。

NPS（Net Promoter Score）口碑、CSAT-Client Satisfaction 客户满意度等结构化数据，以及 CRM 系统、销售系统、财务系统和 Excel 中的记录。非结构化数据是生活变得有趣的地方。 根据定义，它缺乏标准化，并且通常具有有限的预设边界。 它是从文档、社交媒体、电子邮件、音频/视频文件、开放式响应字段、注释字段和其他形式的内容中收集的点点滴滴，使用我们的任何标准数据分析工具都不容易装箱和分析。它的格式可以从文字或数字到图像和音频。 每次客户、成员、潜在客户或利益相关者与您的组织或品牌互动或提及您的组织或品牌时，他们都会生成非结构化数据。 每个组织都有它，但许多人在分析它时不知道从哪里开始。

5

了解了结构化数据和非结构化数据之后，我们再来看看传统数据库和大数据的区别，在这张表中，我们从规模、种类、模式、数据和工具等方面，

用隐喻的方式将数据库与大数据进行了比较。
传统的数据库就像在水池里钓鱼，大数据就像在海里钓鱼。
数据库的规模是 MB，2 的 10 的幂 KB，2 的 20 的幂 Bytes，但是大数据的规模可能是 GB TB PB，2 的 30 次方，40 ,50 Bytes,
在种类方面，数据库就像一些种类的鱼；这是不同类型的结构化数据。大数据就像成千上万鱼；结构化半结构化和非结构化，
所有这些。数据库有自己的预定义模式，所以它是 First schema，然后是 data，
但是对于大数据，不可能有预定义的 schema，所以它是 First data，然后尝试为数据找到合适的 schema。
在数据库中数据只是要处理的对象，在大数据中我们可以通过一些数据感知其他数据。
至于工具，在数据库中我们使用一个工具，传统的数据库管理系统可以解决所有问题，
在大数据中，没有工具可以解决所有问题，大数据需要不同的工具来解决问题的不同方面。
6
本节我们学习了什么是结构化数据，什么是非结构化数据，感谢您的关注，如果您有任何问题，请随时与我联系。