1-5-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，在本节中，我们将讨论大数据生命周期和大数据处理流程。

2

如图所示，大数据生命周期可分为四个阶段。让我们详细研究它们。

1 收集是大数据生命周期中的关键；它定义了在源头捕获的数据类型。一些示例包括从服务器收集日志、获取用户配置文件、抓取组织评论以进行情绪分析和订单信息。我们提到的示例可能涉及处理本地语言文本、非结构化数据和图像，这些将在我们在大数据生命周期中前进时得到处理。随着自动化数据收集流水平的提高，传统上花费大量精力收集结构化数据以分析和估计业务的关键成功数据点的组织正在发生变化。成熟的组织现在使用通常因其大小或格式而被忽略的数据，在大数据术语中，这通常被称为非结构化数据。这些组织总是试图使用最大量的信息，无论是结构化的还是非结构化的，对于他们来说，数据就是价值。您可以使用数据传输和整合到 HDFS（Hadoop 分布式文件系统）等大数据平台中。借助 Apache Spark 等工具处理数据后，您可以将其加载回 MySQL 数据库，这可以帮助您填充相关数据以显示 MySQL 的组成。

2 存储从各种来源收集的数据。让我们考虑一个抓取组织评论以进行情绪分析的示例，其中每个都从不同站点收集数据，每个站点都有唯一显示的数据。传统上，数据使用 ETL（提取、转换和加载）过程进行处理，该过程用于从各种来源收集数据，根据需要对其进行修改，然后将其上传到存储中进行进一步处理或显示。经常用于此类场景的工具是电子表格、关系数据库、商业智能工具等，有时手动工作也是其中的一部分。大数据平台中最常用的存储是 HDFS。 HDFS 还提供 HQL（Hive 查询语言），它可以帮助我们完成许多传统上在商业智能工具中完成的分析任务。其他一些可以考虑的存储选项是 Apache Spark、Redis 和 MongoDB。每个存储选项在后端都有自己的工作方式；但是，大多数存储提供商都公开了可用于进行进一步数据分析的 SQL API。可能存在我们需要收集实时数据并实时展示案例的情况，这实际上不需要存储数据以供将来使用，并且可以运行实时分析以根据请求生成结果。

3 分析如何用一个常见的问题来分析这些不同的数据类型，如果……？组织与数据一起发展的方式也影响了新的元数据标准，将其组织起来进行初始检测和重新处理，以使结构化方法在所创建数据的价值方面成熟。大多数成熟的组织可靠地提供跨业务部门的可访问性、优势和价值，并通过持续自动化的结构化元数据和结果的过程进行处理以进行分析。成熟的数据驱动型组织的分析引擎通常适用于多种数据源和数据类型，其中还包括实时数据。在分析阶段，处理原始数据，MySQL 在 Hadoop 中有 Map/Reduce 作业，以分析并给出输出。由于 MySQL 数据位于 HDFS，它可以被大数据平台相关工具的生态系统的其余部分访问以进行进一步分析。

4 治理在实践中，如果没有既定的治理政策，就无法期望数据的价值。在缺乏成熟的数据治理策略的情况下，企业可能会遇到错误解释的信息，最终可能对业务造成不可预测的损害。在大数据治理的帮助下，组织可以实现一致、精确和可操作的数据意识。数据治理是关于管理数据以满足合规性、隐私、监管、法律以及根据业务要求特别有义务的任何内容。对于数据治理，持续监控、研究、修订和优化流程质量也应尊重数据安全需求。到目前为止，在大数据方面，数据治理已经很容易了；然而，随着数据快速增长并在各个地方使用，这引起了人们对数据治理的关注。它逐渐成为任何大数据项目必须考虑的因素。

3

大数据分析的目的是从数据中提取知识和智慧，但一开始都是原始数据，DIKW 金字塔模型数据有四层，分别是 Data Information ，Knowledge ，Wisdom。

数据在底部，即个别事实、数字、信号、测量指标值、

而在特定上下文，可以从数据中提取信息，进而达到第二层次；

信息是有组织的、结构化的、分类的、有用的、浓缩的、计算的数据。

比如我们有淘宝网购数据，形成一个客户的购物历史，我们可以总结出这个客户偏好的类别，也就是他的购物兴趣。

在信息之上，是知识，是想法、学习、观念、概念、综合、比较、深思熟虑、讨论。

，从购物兴趣中，我们可以学习到这个顾客的性格，知识就是有意义的信息。
根据知识，我们可以得到洞察力，即理解、整合、应用、反思、可操作、积累、原则、模式、决策的过程。而自下而上，决策风险降低了，因为从处理过的数据中我们得到了信息、知识和智慧，不确定性降低了，我们对决策的了解越来越多。
4
利用大数据获得更多商业智能。 第一阶段，最开始是1968年发明的分层数据库和1970年发明的关系型数据库，有了这些操作型数据库，就可以在静态数据的基础上，对历史数据进行报表和人工分析。这个阶段主要是OLTP：Online Transaction Processing 第二阶段，借助数据仓库，人们可以分析当前数据以改善业务交易，因为数据仓库可以整合组织各个不同方面的综合信息，有助于更好地了解业务和客户，并提供更好的服务.这个阶段主要是OLAP：Online Analytical Processing 第三阶段 随着数据的快速增长，数据来得更快，需要更快地处理，为了解决这个问题，2000 年创建了流计算，它可以在数据到来时快速处理，并支持 RTAP：Real-Time Analytics Processing做出实时决策并改进实时业务响应
5
利用存储在组织系统中的数据，我们可以进行业务智能，即Ad-hoc查询和报告
使用数据挖掘技术，基于结构化数据，中小型数据集中的典型来源。
但是随着新技术的优势，越来越多的数据被收集，我们想要知道更多，在事情没有发生的时候预测它，这样复杂性就会增加，商业价值就会扩大。
基于来自许多来源的所有类型的数据，将它们整合成非常大的数据集，可以进行复杂的统计分析，可以进行优化和预测分析，更多的是实时的方式，反之亦然，这促使人们收集更多的数据。所有这些都将推动大数据的进步。
6
让我们看看商务智能的演变。
在 1990 年代的商业智能报告 OLAP 和数据仓库可以帮助组织生成有用的信息，以使用 Business Objects、SAS、Informatics、Cognos 等 SQL 报告工具等工具来总结组织情况。
在 2000 年，当数据集变大，分析需要更长的时间，我们需要加快分析速度，借助 Tableau、HANA 等交互式商业智能和内存 RDBMS，分析可以更快，当规模变大时，需要大数据处理技术。
大数据：批处理和分布式数据存储 像 Hadoop/Spark、HBase/Cassandra 可以处理这些问题。
2010年对规模和速度都有要求，需要大数据实时和所有综合数据的单一视图，然后需要更先进的技术来应对这些挑战。
7
在本节中，我们了解了大数据生命周期，包括收集、存储、分析和治理。
其中，分析是目的，我们要从原材料中挖掘价值。
我们回顾了传统分析和大数据分析，现在我们需要快速分析海量的大数据。谢谢。