1-6-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，本节我们讨论大数据处理流程

2

所有的大数据分析和处理都打算从数据中获取价值。这就是大数据分析的最终目标。

为了实现它，我们需要做几个步骤，这张图从分析的角度展示了大数据的流程。

1）根据分析需求找到合适的数据源，可以是结构化数据、半结构化数据或非结构化数据。

2） 然后从选定的数据源收集数据，可以使用 ETL（Extract Transform load）从不同的数据库或数据仓库中提取数据或使用数据爬虫来爬取数据。

3 ）从不同数据源获取数据后，有时数据格式不同，需要进行转换，以方便进一步的数据存储或处理。我们需要规划合适的数据存储方式，由于大数据的性质，数据量很大，可能需要分布式数据存储解决方案。

4） 然后需要处理数据以支持分析。在数据处理中，数据科学家可能首先清理数据，删除错误数据、冗余数据和其他脏数据。

之后，数据科学家需要进行特征工程，提取或构建具有代表性的关键特征，以支持进一步的模型或算法

然后数据科学家可以应用统计、数学模型或机器学习算法来找到模式、相关性、分类等。

5） 将模式、相关性、分类等可视化，使结果易于理解。例如，不同省份的流行病数据以地图的形式比表格更能说明分布。

6 ）还可以生成一些报告以进行进一步分析。或者结果可用于监控业务。

这些是分析角度的大数据流程。

3

所有的大数据分析和处理都是要从数据中获取价值。这就是大数据分析的最终目标。为了实现它，我们需要做几个步骤，

这张图从分析的角度展示了大数据的流程。最左边是数据资产，包括事务、OLTP、在线事务处理和 OLAP 在线分析处理，

不同类型的文档、社交媒体内容和机器设备生成的物联网数据。所有这些数据都需要集成到集群或大数据存储中，以支持进一步的分析。

1） 数据库或数据仓库中的事务数据需要卸载 和通过 ETL 处理导入到 Hadoop，这主要是旅程的第一步。交易数据主要是结构化数据。

2） 一些文档数据也可以用于数据分析，这些数据也需要卸载到 Hadoop 集群进行进一步的处理和分析。

3） 社交媒体数据可以帮助了解客户的意见和偏好，从而实现更好的定制服务。

4） 在物联网设备的帮助下，我们可以从不同类型的传感器、摄像机等收集物联网数据，这些数据可以帮助我们了解设备的状态及其内容。

所有这些原始数据都可以集成到集群中，更新的元数据也需要更新到集群中，此外，还需要及时收集实时数据，以支持进一步的实时分析。

5） 在集群中收集到数据后，数据科学家将根据分析目的尝试对数据进行发现和画像

6） 并使用定义的元数据来管理和丰富数据集。

7） 然后为业务分析解析和准备数据。

8）数据科学家在分析数据时，应屏蔽姓名、身份证等敏感数据，以保护隐私。基于集群中的数据，数据科学家可以生成商业智能报告或其他可视化图表。

有时，为了进行有效的分析，数据科学家希望建立一个企业数据仓库并将整理的数据移动到企业数据仓库，仅存储商业智能的相关数据。并在企业数据仓库的基础上，分析可以生成商业智能报告以支持决策。

4

举个例子，京东大数据分析。

京东有大量来自在线购物历史的交易数据。

而且他们还有大量的文档数据，比如客户邮件、行业报告、用户协议等等。

以及顾客对购物后感受的评价。
他们也有大量的物联网数据，比如物流卡车的 GPS、产品包装的 RFID 等。
将所有这些数据集成到 Hadoop 集群中，京东数据科学家可以发现和分析客户、产品、快递以及所有预定义的模型、元数据，
他们可以解析原始原始数据，并生成京东销售趋势 BI 报告， 快递效率 BI 报告、产品类别 BI 报告等。
但当然，京东数据科学家应该屏蔽客户敏感数据以保护客户隐私。

5
在本节中，我们从分析角度和技术角度讨论了大数据处理流程。 我们了解大数据分析的一般步骤，包括数据源的选择、数据的收集、数据存储前的数据清洗。基于存储在大数据分布式存储中的数据，数据科学家可以对数据进行处理以进行进一步分析，在所有数据处理完毕后，数据科学家可以使用算法或设计模型对数据进行分析，生成商业智能报告或做 可视化分析以挖掘数据的洞察力并支持决策制定。