2-1-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，本节我们讨论大数据的数据来源

2

根据组织边界，我们可以将数据分为内部数据和外部数据。
顺便说一下，我们之所以要将大数据资源分为内部和外部，
是因为内部数据和外部数据使用的获取数据的方法和工具不同。
1 内部数据资源是指组织创建并存储在组织内部的所有数据，
包括组织自营系统、业务交易系统、制造系统、财务系统、
人力资源系统、会计系统等，这些都是结构化数据。
组织内部数据还包括一些历史遗留数据，例如一些文档数据、电子邮件等。
另外还有一些内部的物联网设备可以产生一些数据，比如制造自动化生产线中有很多传感器，
这些传感器可以收集很多生产状态数据，便于分析生产制造的状况。
2 外部数据是指特定组织以外的数据。外部数据包括
2.1 其他组织业务运营平台数据，如来自其他组织的业务交易系统、制造系统、财务系统、人力资源系统、会计系统等数据。
2.2 其他组织的物联网设备产生的数据
2.3 政府收集和公布的数据
2.4 互联网或移动互联网数据，由所有互联网节点发布，如社交媒体、博客数据、YouTube 数据、维基百科数据等。
所有这些内部和外部数据都可以作为专门分析的数据资源。

3

1 ）政府数据，出于社会管理目的而设置的各种部门，比如公检法，财政部，发改委，工商、税务、海关、人社、医疗等，

这些组织出于有效完成部门职能的目的，会构建很多业务系统，这些系统产生的数据主要以特定的结构存储在相应的数据中心中，
其数据内部蕴含着巨大的价值，能够为政府宏观政策的制定，国家安全防控、社会有效管理等提供数据支撑。
政府数据具有可信度高，完整性好，实时性强，实体描述指向性明确等特点。

2 ）各利益主体的 IT 系统，ERP，SCM，在线办公，在线交易等

3 ）将物联网数据搜集纳入数据富集的考虑范畴，根据物联网终端或相应 App,物联网数据以企业自营数据库的形式存放在企业内部数据库中，或者存放在互联网中

4

在组织内部，所有的业务数据都是从不同的业务系统收集的，
如 CRM、ERP、营销系统、财务系统、HR 系统、供应链系统等数据。
所有这些数据都被提取并存储在数据存储中，可以是 SQL Server 这样的关系数据库，
Hive 这样的数据仓库，或者 Mongo DB 这样的文档数据库，或者 neo4j 这样的图形数据库，
然后使用 Hadoop、Spark 等数据处理平台处理数据。并使用 TensorFlow、R、
IBM Watson 等工具分析数据。数据经过处理和分析后，我们还可以使用 tableau 等可视化工具将数据可视化。
通过大数据分析，我们可以获得业务改进、业务方案，收益分析、收入分析、客户概况、合适的定价、一些业务模式等，
所有这些都可以帮助组织更好地了解客户和市场，提供更好的产品和服务，做出更好的决策并改进他们的业务战略。

5
除了内部数据，还有外部数据可供分析。
外部数据主要是互联网数据，其渠道包括门户网站、政府公开信息、社交媒体、电子商务公共数据和一些针对特定主题的专题论坛。
6
互联网有海量数据，但是当您从互联网收集数据时，您应该注意以下问题。
1）不同网站的 IT 水平和结构不同，所有网站没有统一的收集方法。
2）不同网站对爬虫的控制策略不同，为什么？因为网络爬虫方便了网络信息的收集和查询，但也带来了以下负面影响：
2.1 网络爬虫总是消耗过多的服务器带宽并增加服务器负载，因为它们使用特定的策略来浏览网站上尽可能多的高价值信息。
2.2 不良行为者可能使用网络爬虫对网站发起 DoS 攻击。因此，网站可能会因资源枯竭而无法提供正常服务。
2.3 不良行为者可能会使用网络爬虫窃取您网站上的关键任务数据，这将损害网站的经济利益。
一些 WAF-web 应用防火墙提供了三种反爬虫策略，
2.3.1 通过识别 User-Agent 进行机器人检测，
2.3.2 网站反爬虫通过检查浏览器的有效性，
2.3.3 通过限制访问频率进行 CC-Challenge Collapsar 挑战黑洞攻击防护，全面缓解爬虫对您网站的攻击。
2.3.3.1 所以有些网站会启用机器人检测来识别用户代理如果启用机器人检测，WAF 可以检测和阻止恶意爬虫、扫描程序和 Web Shell 等威胁。
2.3.3.2 开启反爬虫保护验证浏览器有效性如果开启反爬虫防护，WAF 会动态分析网站服务模型，根据数据风控和爬虫识别手段，准确识别爬虫行为。
2.3.3.3 配置 CC-Challenge Collapsar 挑战黑洞攻击保护限制访问频率
CC 攻击 -挑战黑洞攻击是一种 DDoS 攻击，使用代理服务器向受害服务器发送大量貌似合法的请求
保护规则使用特定的 IP 地址、cookie 或引用来限制对特定路径（URL）的访问，从而减轻 CC 攻击对 Web 服务的影响。
3） 互联网数据可以是文本、表格、音频和视频的形式，不同的形式增加了收集的难度。因为你需要设计不同的方式来收集它们。
4） 其真实性和数据质量不如其他数据由于任何人都可以在互联网上发布数据，因此没有人对数据的真实性和数据质量负责，这使得其真实性和数据质量不如其他数据，例如内部数据。

7
让我们总结一下数据获取通道
如果我们要做大数据分析，我们需要有足够的数据，数据可以从组织内部或外部收集，包括内部数据和外部数据。
内部数据包括业务系统数据、归档数据（如文档、电子邮件等）和组织物联网数据。
外部数据包括政府公共数据、其他组织数据、互联网数据和外部物联网数据。

8
在我们学习了大数据资源、内部和外部数据。
基于多维度的数据，我们可以全面了解组织或业务。
感谢您的关注，如果您有任何问题，请随时与我联系。