大数据基础与实务

项目三　大数据采集与清洗

项目三　大数据采集与清洗

职业能力

能运用大数据采集与清洗的知识，做好大数据清洗与采集的全面准备工作

能使用大数据采集工具采集所需数据

能准确把握数据清洗的内容和目的

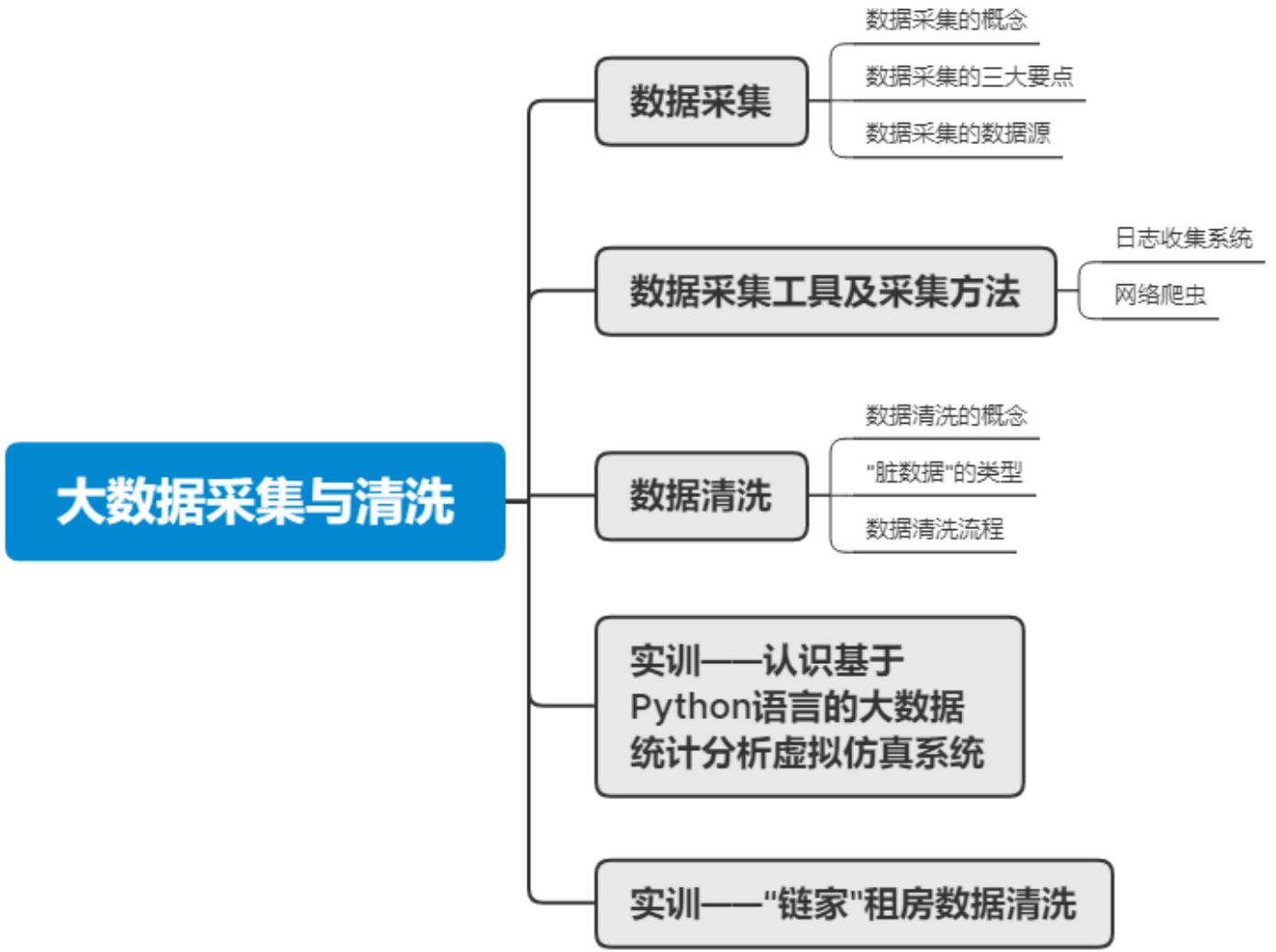
职业素养

能分析数据并定义清洗规则、搜寻并标识错误实例、纠正发现的错误

熟练使用数据清洗工具

具备大数据平台实践能力

知识图谱


大数据采集与清洗
数据采集
数据采集的概念
数据采集的三大要点
数据采集的数据源
数据采集工具及采集方法
日志收集系统
网络爬虫
数据清洗
数据清洗的概念
"脏数据"的类型
数据清洗流程
实训——认识基于Python语言的大数据统计分析虚拟仿真系统
实训——"链家"租房数据清洗

任务一

数据采集

目录 | CONTENTS

ONE 任务描述

TWO 知识准备

THREE 课堂研讨

FOUR 拓展训练

任务描述

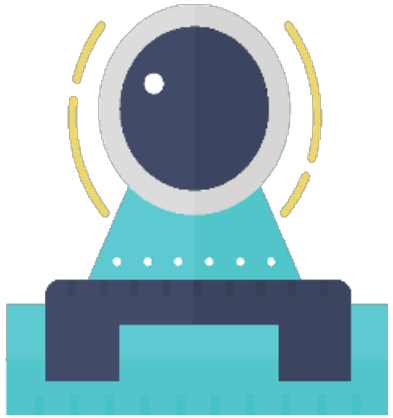
大数据开启了一个大规模生产、分享和应用数据的时代，它给技术和商业带来了巨大的变化。麦肯锡研究表明，在医疗、零售和制造业领域，大数据每年可以提高劳动生产率0.5%～1%。大数据在核心领域的渗透速度有目共睹，然而调查显示，未被使用的信息比例高达99.4%，很大程度都是由于高价值的信息无法获取采集。因此在大数据时代背景下，如何从大数据中采集出有用的信息已经是大数据发展的关键因素之一。

勤奋好学的张明找到老师并提问：什么是数据采集呢？

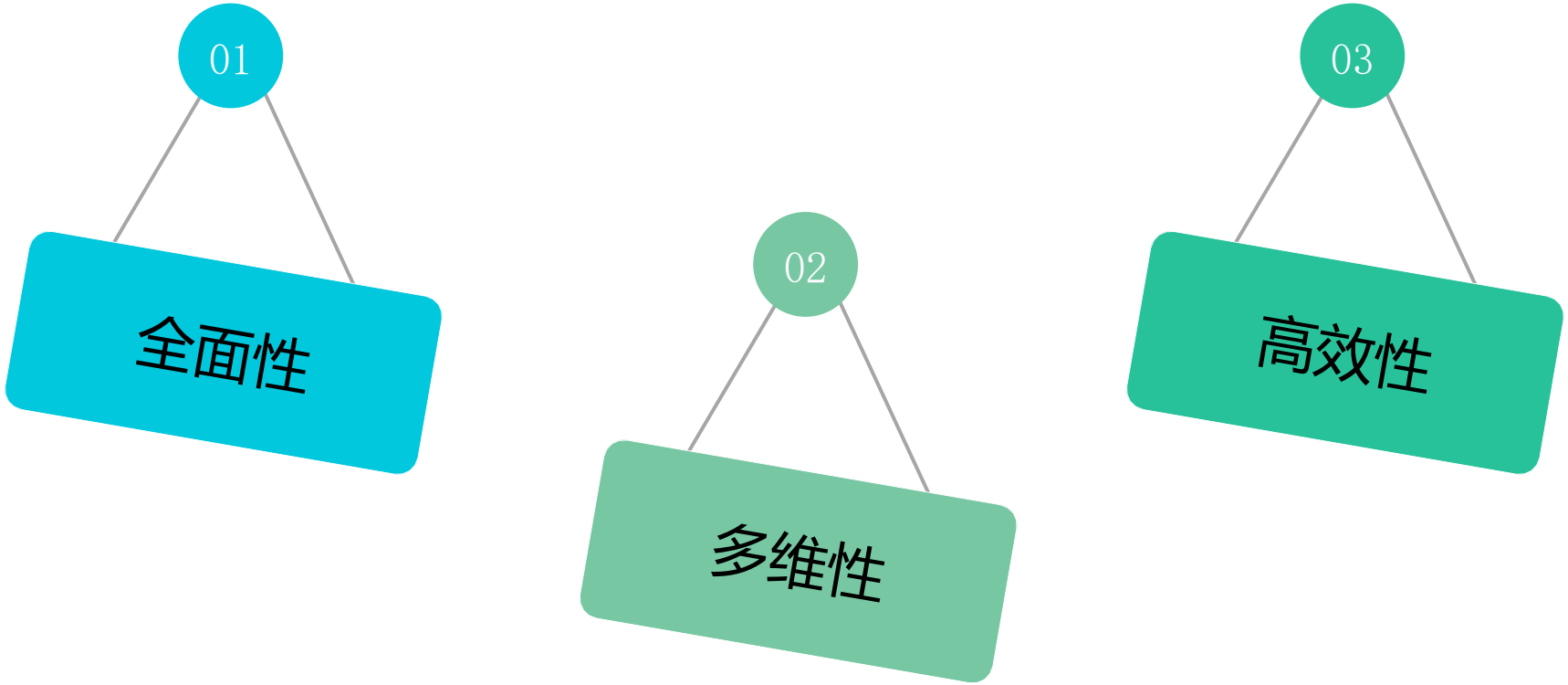
一 数据采集的概念

定义

数据采集就是使用某种技术或手段，将数据收集起来并存储在某种设备上。数据采集处于大数据生命周期中的第一个环节，之后的分析挖掘都建立在数据采集的基础上。数据采集技术广泛应用在各个领域，比如摄像头和麦克风，都是数据采集工具。



二 数据采集的三大要点



三 数据采集的数据源

新数据源的归纳与分类

1 线上行为数据

页面数据、交互数据、表单数据、会话数据等

2 内容数据

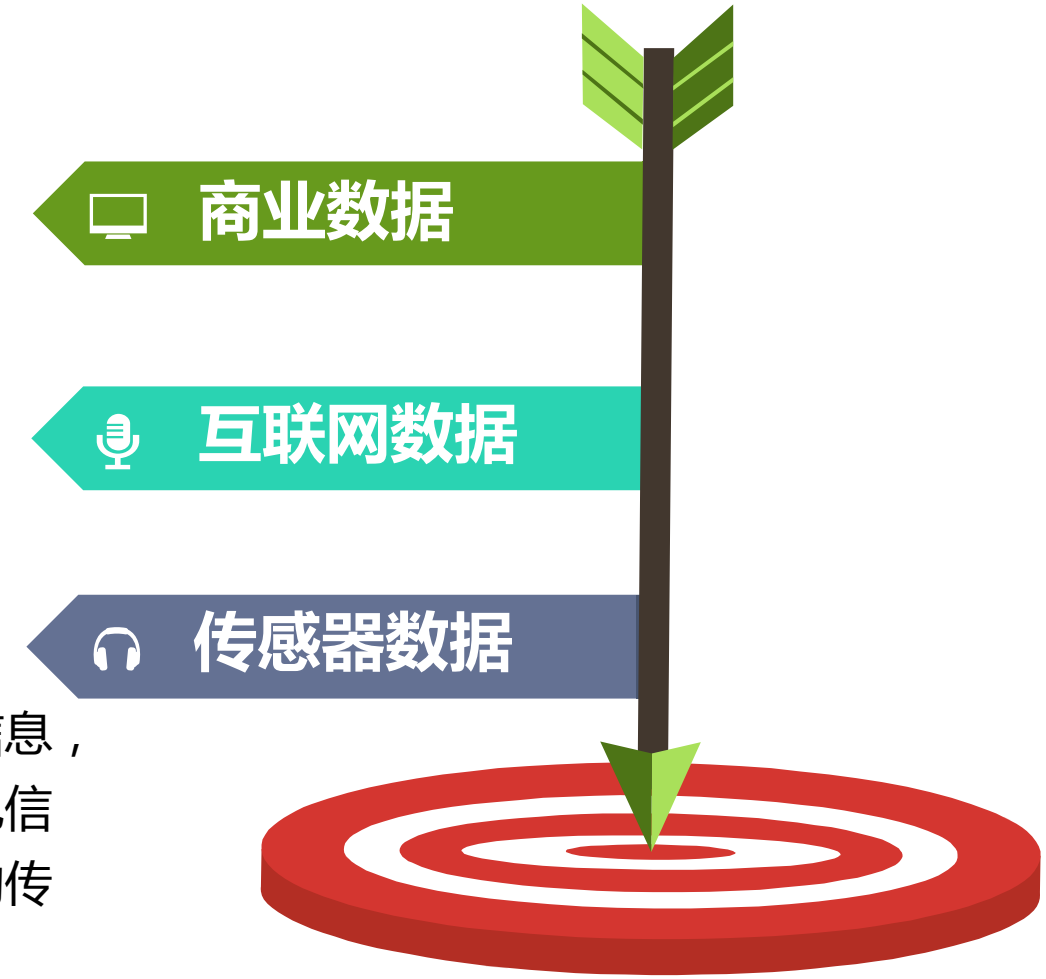
应用日志、电子文档、机器数据、语音数据、社交媒体数据等

三 数据采集的数据源

- 商业数据主要来源于公司业务平台的日志文件以及业务处理系统

- 互联网数据的采集通常是借助于网络爬虫来完成的。所谓“网络爬虫”，就是一个在网上到处或定向抓取网页数据的程序。

- 传感器是一种检测装置，能感受到被测量的信息，并能将感受到的信息，按一定规律变换成为电信号或其他所需形式的信息输出，以满足信息的传输、处理、存储、显示、记录和控制等要求。



课堂研讨

Q1

在一些专业二手平台上，网售大数据采集和定制业务颇为盛行。有些从事信息贩卖的“商家”，正大肆兜售着覆盖诸多行业的用户信息，内容颇为庞杂，可谓五花八门，无所不包。有的还以行业明码标价，成行成市。这些人打着“专业定制”的旗号，无论需要哪类信息，只要客户提出要求，其都能从网上为你采集到。这些数据商的背后隐藏着一条非法获取用户数据的产业链。他们通过专业的“爬虫软件”，侵入搜索引擎、企业网页、公众号及微信朋友圈等，采集各类个人信息及实时数据，经过汇总、整理然后生成所谓大数据产品出售。

思考：如果任由此类行业继续发展，将会带来怎样的后果？

拓展训练

01

请在网上查找有关数据采集的企业应用实例。

任务二

数据采集工具及采集方法

目录 | CONTENTS

ONE 任务描述

TWO 知识准备

THREE 课堂研讨

FOUR 拓展训练

任务描述

近年来，由于互联网大数据技术的快速发展，以及消费者需求不断发生变化，对企业的营销方式也提出了更高的要求，以“产品为中心”的营销观念和手段无法适应目前市场和消费者需求多样化发展的形势。某烟草企业就面临这样的问题，想要找到基于大数据采集技术的企业营销的创新模式，能够实现对消费者的需求变化及时把控，真正做到以消费者为导向，从而进行有针对性的市场营销活动。

任务描述

作为一名普通大学生，张明也想为该烟草公司筹谋划策，并准备从寻找合适的数据采集工具与方法开始入手。在上一任务中，张明了解到数据采集的数据源主要分为商业数据、互联网数据、传感器数据三大类，根据烟草公司的特性，张明想知道，可以采集到商业数据、互联网数据的工具及方法有哪些呢？

一 日志收集系统

（一）Scribe

Scribe是Facebook开源的日志收集系统，在Facebook内部已经得到的应用，其体系架构如图3-1所示。它能够从各种日志源上收集日志，存储到一个中央存储系统（可以是NFS，分布式文件系统等）上，以便于进行集中统计分析处理。

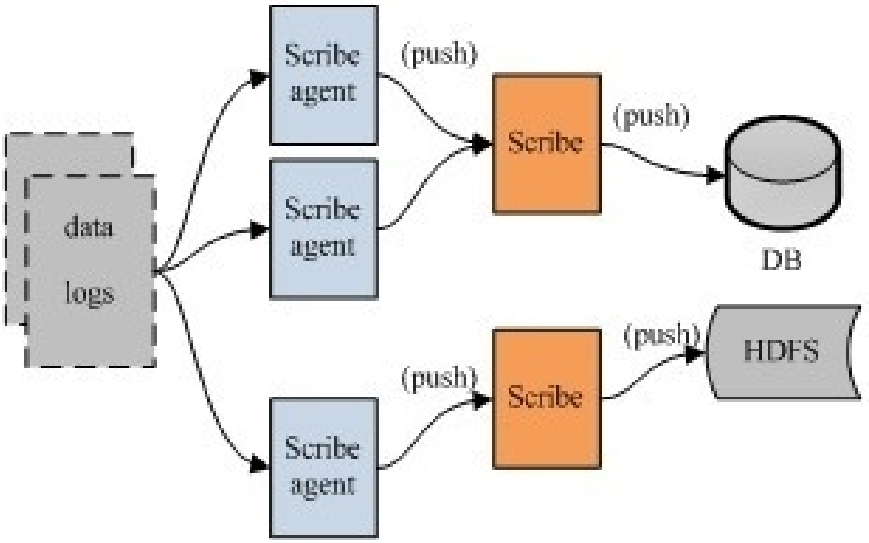


图3-1 Scribe体系架构图

一 日志收集系统

（二）Flume

Flume是Cloudera提供的一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

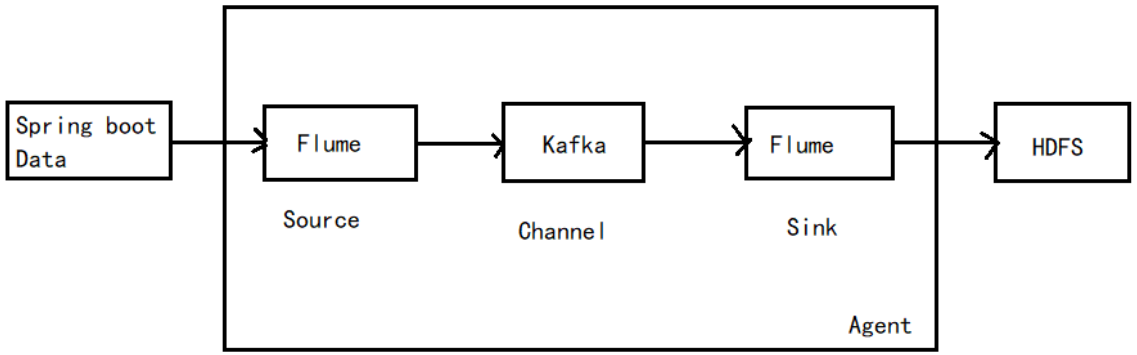


图3-2 Flume体系架构图

一 日志收集系统

（三）Chukwa

Chukwa是一个开源的用于监控大型分布式系统的数据收集系统。这是构建在Hadoop 的HDFS和Mapreduce框架之上的，继承了Hadoop的可伸缩性和健壮性。Chukwa还包含了一个强大和灵活的工具集，可用于展示、监控和分析已收集的数据。

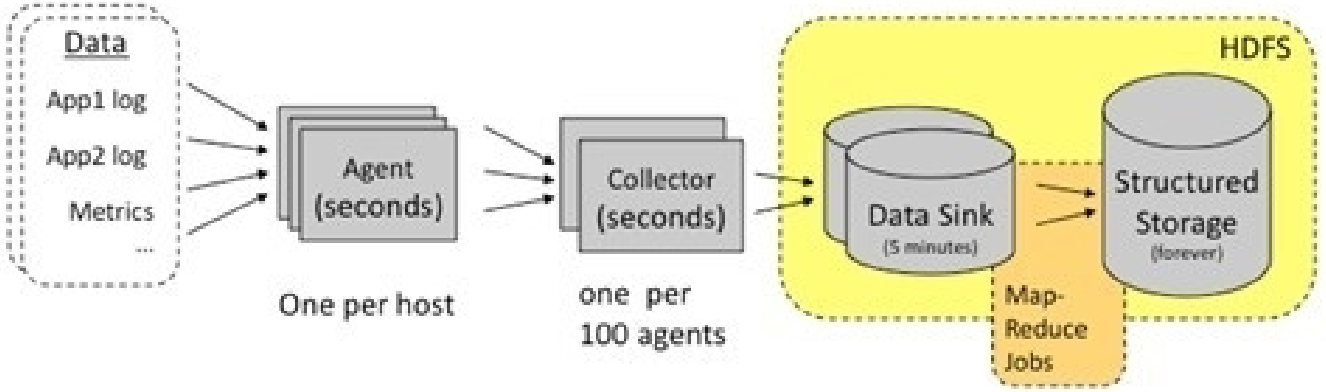


图3-3 Chukwa结构图