

目 录

第 1 章 大数据概述	(1)
1.1 大数据的概念	(1)
1.2 大数据关键技术	(3)
1.3 大数据采集与数据预处理技术	(6)
1.3.1 大数据采集技术	(7)
1.3.2 数据预处理技术	(8)
1.4 小结	(9)
习题	(9)
第 2 章 数据采集基础	(10)
2.1 传统数据采集技术	(10)
2.1.1 数据采集概述	(10)
2.1.2 数据采集系统架构	(11)
2.1.3 数据采集关键技术	(14)
2.2 大数据采集基础	(18)
2.2.1 数据的发展	(18)
2.2.2 大数据来源	(21)
2.2.3 大数据采集技术	(26)
2.3 小结	(32)
习题	(33)
第 3 章 大数据采集架构	(34)
3.1 概述	(34)
3.2 Chukwa 数据采集	(35)
3.3 Flume 数据采集	(37)
3.4 Scribe 数据采集	(40)
3.5 Kafka 数据采集	(41)
3.6 小结	(45)
习题	(46)
第 4 章 大数据迁移技术	(47)
4.1 数据迁移概念	(47)
4.2 数据迁移相关技术	(48)
4.2.1 基于主机的迁移方式	(48)
4.2.2 基于存储的迁移方式	(48)
4.2.3 备份恢复的方式	(50)
4.2.4 基于主机逻辑卷的数据迁移	(51)
4.2.5 基于数据库的迁移技术	(52)
4.2.6 服务器虚拟化的迁移	(53)
4.2.7 其他数据迁移技术	(55)
4.3 数据迁移工具	(56)
4.3.1 Apache Sqoop	(56)
4.3.2 ETL	(58)

4.4 Kettle 数据迁移实例	(59)
4.5 小结	(65)
习题	(65)
第5章 互联网数据抓取与处理技术	(66)
5.1 网络爬虫概述	(66)
5.1.1 网络爬虫的概念	(66)
5.1.2 网络爬虫的抓取策略	(67)
5.1.3 网页更新策略	(68)
5.2 常用网络爬虫方法	(69)
5.2.1 批量型爬虫	(70)
5.2.2 增量型爬虫	(70)
5.2.3 垂直型爬虫	(70)
5.2.4 通用网络爬虫	(70)
5.2.5 聚焦网络爬虫	(71)
5.2.6 深层网络爬虫	(72)
5.2.7 分布式网络爬虫	(73)
5.3 网络爬虫工具	(75)
5.3.1 Googlebot	(75)
5.3.2 百度蜘蛛	(76)
5.3.3 ApacheNutch	(76)
5.3.4 火车采集器	(77)
5.3.5 集搜客	(77)
5.3.6 八爪鱼采集器	(78)
5.4 Python 爬虫技术	(81)
5.4.1 Python 概述	(81)
5.4.2 Python 爬虫基础	(83)
5.4.3 Python 安装	(88)
5.4.4 Python 爬虫实例	(91)
5.5 文本数据处理	(94)
5.5.1 文本分词概述	(94)
5.5.2 中文分词算法	(96)
5.5.3 MMSEG 分词算法	(97)
5.5.4 常用中文分词工具	(100)
5.5.5 网页分析算法	(101)
5.6 小结	(103)
习题	(103)
第6章 数据预处理技术	(104)
6.1 数据的描述	(104)
6.1.1 数据对象与属性类型	(104)
6.1.2 数据的统计描述	(106)
6.1.3 数据相似性和相异性的度量方法	(109)
6.2 数据预处理概述	(113)
6.2.1 数据质量	(113)
6.2.2 数据预处理的主要任务	(114)

6.3	数据清洗	(115)
6.3.1	缺失值处理	(115)
6.3.2	光滑噪声数据处理	(116)
6.3.3	检测偏差与纠正偏差	(117)
6.4	数据集成	(118)
6.4.1	模式识别和对象匹配	(118)
6.4.2	冗余问题	(119)
6.4.3	元组重复	(121)
6.4.4	数据值冲突的检测与处理	(121)
6.5	数据归约	(122)
6.5.1	小波变换	(122)
6.5.2	主成分分析	(123)
6.5.3	属性子集选择	(123)
6.5.4	回归和对数线性模型	(124)
6.5.5	直方图	(125)
6.5.6	聚类	(126)
6.5.7	抽样	(126)
6.5.8	数据立方体聚集	(127)
6.6	数据变换	(128)
6.6.1	通过规范化变换数据	(129)
6.6.2	通过离散化变换数据	(130)
6.6.3	标称数据的概念分层变换	(131)
6.7	小结	(132)
	习题	(132)
第7章	大数据分析实例	(134)
7.1	Hadoop 相关理论知识	(134)
7.1.1	Hadoop 生态系统	(135)
7.1.2	HDFS	(139)
7.1.3	MapReduce	(143)
7.1.4	HBase	(149)
7.1.5	Hive	(152)
7.1.6	Yarn	(156)
7.1.7	ZooKeeper 和 Sqoop	(159)
7.2	实验内容	(161)
7.2.1	技术方案与实验环境	(161)
7.2.2	实验环境搭建	(161)
7.2.3	实验过程	(167)
7.3	小结	(173)
	习题	(174)
	参考文献	(175)