

《基于Hadoop的大数据分析和处理》

书籍信息

版次：1

页数：

字数：

印刷时间：2017年06月01日

开本：16开

纸张：胶版纸

包装：平装-胶订

是否套装：否

国际标准书号ISBN：9787121317392

内容简介

本书基于云计算和大数据，介绍大数据处理和分析的技术，分为两部分。*部分介绍Hadoop基础知识，内容包括：Hadoop的介绍和集群构建、Hadoop的分部式系统架构、Map Reduce及其应用、Hadoop的版本特征及进化。第二部分以云计算为主题，详细论述利用Hadoop的大数据分析和处理工具，以及NoSQL技术，内容包括：云计算和Hadoop、Amazon服务中的MapReduce应用、Hadoop应用下的大数据分析、NoSQL、HBase。本书不单纯地讲述理论和概念，而是基于具体的工具和技术(Hadoop和NoSQL)，利用大量实际案例，通过实际的操作和应用来组织大数据处理和分析技术，有利于读者从工程应用的角度进行实际掌握和利用。适合相关专业的本科生、研究生和软件工程师学习。

作者简介

魏祖宽，男，电子科技大学教授，博士，韩国科技协会、中国计算机学会、日本电子电器协会会员。承担计算机以及软件学院的本科和研究生的数据库课程教学和实验教学，及云计算和大数据方面的新课。主持数据库应用、GIS应用等方面的应用课题10多项(国家自然科学基金委，省/市级科技局等科研项目，以及企业横向项目)，现专注于云存储方面的应用科研项目。

目录

目录

- 第1章 Hadoop的介绍和集群构建 2
 - 1.1 Hadoop介绍 2
 - 1.1.1 云计算和Hadoop 2
 - 1.1.2 Hadoop的历史 4
 - 1.2 Hadoop构建案例 6
 - 1.2.1 欧美构建案例 6
 - 1.2.2 韩国构建案例 7
 - 1.3 构建Hadoop集群 8
 - 1.3.1 分布式文件系统 8
 - 1.3.2 构建Hadoop集群的准备事项 12
 - 1.3.3 构建伪分布式 17
 - 1.3.4 分布式集群 (Cluster) 构建 29
 - 1.4 Hadoop界面 36
 - 1.4.1 Hadoop分布式文件系统指令界面 36
 - 1.5 总结 40

第2章 Hadoop分布式处理文件系统	41
2.1 Hadoop分布式文件系统的设计	42
2.2 概观Hadoop分布式文件系统的整体构造	43
2.3 Namenode的角色	44
2.3.1 元数据管理	44
2.3.2 元数据的安全保管——Edits和FsImage文件及Secondary Namenode	49
2.3.3 Datanode管理	52
2.4 Datanode的角色	59
2.4.1 block管理	59
2.4.2 数据的复制和过程	61
2.4.3 Datanode添加	63
2.5 总结	65
第3章 大数据和MapReduce	67
3.1 大数据的概要	68
3.1.1 大数据的概念	69
3.1.2 大数据的价值创造	69
3.2 MapReduce	71
3.2.1 MapReduce 示例：词频统计（Word Count）	71
3.2.2 MapReduce开源代码：词频统计（Word Count）——Java基础	75
3.2.3 MapReduce 开源代码：词频统计（Word Count）——Ruby语言基础	76
3.3 MapReduce的结构	78
3.3.1 通过案例了解MapReduce结构	79
3.3.2 从结构性角度进行的MapReduce最优化方案	81
3.4 MapReduce的容错性（Fault Tolerance）	85
3.5 MapReduce的编程	86
3.5.1 搜索	86
3.5.2 排序	87
3.5.3 倒排索引	87
3.5.4 查找热门词	88
3.5.5 合算数字	89
3.6 构建Hadoop：通过MapReduce的案例介绍	90
3.6.1 单词频率统计MapReduce的编程	91
3.6.2 MapReduce—用户界面	95
3.7 总结	99
第4章 Hadoop版本特征及进化	101
4.1 Hadoop 0.1x版本的API	103
4.2 Hadoop附加功能（append）	107
4.3 Hadoop安全相关功能	109
4.4 Hadoop 2.0.0 alpha	111
4.4.1 安装Hadoop 2.0.0	112
4.4.2 Hadoop分布式文件系统的更改	120
4.4.3 跨时代MapReduce框架：YARN	128

4.5 总结	135
第5章 云计算和Hadoop	137
5.1 大规模Hadoop集群的构建和案例	138
5.2 云基础设施服务的登场	139
5.2.1 Amazon云服务	141
5.3 在Amazon EC2中构建Hadoop集群	156
5.3.1 Apache Whirr	156
5.3.2 构建Hadoop 集群	157
5.4 总结	160
第6章 Amazon Elastic MapReduce的倍增利用	161
6.1 Amazon EMR的活用	162
6.1.1 Amazon EMR的概念	162
6.1.2 Amazon EMR的构造	162
6.1.3 Amazon EMR的特征	163
6.1.4 Amazon EMR的 Job Flow和Step	164
6.1.5 使用Amazon EMR前需要了解的事项	165
6.1.6 Amazon EMR的实战运用	170
6.2 总结	178
第7章 Hadoop应用下的大数据分析	179
7.1 Hadoop应用下的机器学习 (Mahout)	180
7.1.1 设置及编译	181
7.1.2 K-means 聚类算法	183
7.1.3 基于矢量相似度的协同过滤	188
7.1.4 小结	194
7.2 基于Hadoop的统计分析Rhive (R and Hive)	195
7.2.1 R的设置及灵活运用	195
7.2.2 Hive的设置及灵活运用	198
7.2.3 RHive的设置及灵活运用	201
7.2.4 小结	207
7.3 利用Hadoop的图形数据处理Giraph	207
7.4 总结	216
第8章 数据中的DBMS , NoSQL	217
8.1 NoSQL出现背景 : 大数据和Web 2.0	218
8.1.1 基于Web 2.0的大数据的登场	218
8.1.2 基于大数据的NoSQL的登场	221
8.1.3 适合大数据和Web 2.0的数据库NoSQL	222
8.2 NoSQL的定义和类别特征	226
8.3 NoSQL数据模型概要和分类	229
8.4 NoSQL数据模型化	231
8.4.1 NoSQL数据模型化基本概念	232
8.4.2 一般的NoSQL建模方法	234
8.5 主要NoSQL的比较和选择	239

8.6 总结	241
第9章 Hbase : Hadoop中的NoSQL	243
9.1 Hadoop生态界中的HBase	244
9.2 HBase介绍	248
9.3 HBase数据模型	250
9.3.1 map	250
9.3.2 持续性 (persistent)	250
9.3.3 分布性 (distributed)	250
9.3.4 排序性 (sorted)	250
9.3.5 多维性 (multidimensional)	251
9.3.6 稀疏性 (sparse)	254
9.4 HBase的数据库模式	255
9.5 HBase构造	259
9.6 HBase的构建及运行	261
9.7 HBase的扩展——DuoBase中的HBase	264
9.8 HBase的用户定义索引	266
9.8.1 HBase用户定义索引—HFile格式的扩展	267
9.8.2 HBase用户定义索引—Region的扩展	267
9.9 总结	270

前言

前言

欧盟的“INFO2000计划”中对内容产业的定义是：那些制造、开发、包装和销售信息产品及其服务的企业，其中包括在各种媒介上的印刷品（报纸、书籍、杂志等）；电子出版物（联机数据库、音像制品服务，以传真及光盘为基础的服务以及电子游戏等）；音像传播（电视、录像、广播和影院），还有一些定义把部分软件业（包括课程软件）也放进去了。

“在不久未来，信息服务内容的质量高低将取决于如何加工大数据”。

很久以前就已经感觉到，内容（contents），在大部分的服务和产品中，已经成为最重要的决定要素。最初由谷歌出世、最近各家厂商纷纷推出的互联网电视，就是这样一个例子，虽然产品硬件各有特色地优异，但其中最核心的内容提供才是吸引顾客的关键。

问题是，随着互联网技术的急速发展，构建信息内容的数据量也在急速增加。这类量级巨大、急速增加的数据信息我们称为“大数据”。一般来讲，当我们说“信息内容的质量高低取决于如何加工信息大数据”的时候，就意味着优质高效地加工这些信息大数据所对应的软件技术是必需的。

欧盟的“INFO2000计划”中对内容产业的定义是：那些制造、开发、包装和销售信息产品及其服务的企业，其中包括在各种媒介上的印刷品（报纸、书籍、杂志等）；电子出

出版物（联机数据库、音像制品服务，以传真及光盘为基础的服务以及电子游戏等）；音像传播（电视、录像、广播和影院），还有一些定义把部分软件业（包括课程软件）也放进去了。

很久以前就已经感觉到，内容（contents），在大部分的服务和产品中，已经成为最重要的决定要素。最初由谷歌出世、最近各家厂商纷纷推出的互联网电视，就是这样一个例子，虽然产品硬件各有特色地优异，但其中最核心的内容提供才是吸引顾客的关键。

我们通过本书试图和读者们分享和思考“如何存储和处理这类信息大数据”。我们看到的YouTube或别的视频网站已经在多年前就在思考这些问题：适应不同的服务平台，从成千上万个视频中，根据顾客的兴趣，精心地经过推荐和过滤等环节，向顾客提供高质量的内容视频。本书中，正是要介绍可以简单地完成这些数据加工任务的开源软件Hadoop及其关联工具。特别的，对和Hadoop一起用于实际大数据分析的专用工具进行了有深度的探讨，并基于图表和案例进行了形象的说明。通过本书，比起对Hadoop的相关开源代码的理解来说，作者更着重于读者在实战中对实际大数据分析平台的理解和见识。特别是，在数据分析处理、平台架构构建时针对大数据处理所遇见的共通性必需技术进行了详细的介绍。

第二部分包括第5章 云计算和Hadoop、第6章 Amazon Elastic MapReduce的倍增利用、第7章 Hadoop应用下的大数据分析、第8章 数据中的DBMS、NoSQL和第9章 HBase：Hadoop中的NoSQL。该部分从云计算的基本概念讲起，通过介绍Amazon的主要服务内容，详细了解将云计算和大数据有效结合的典型云服务——Amazon Hadoop服务，对Hive、Pig、EC2等可供应用的技术进行了说明；通过了解Mahout、R RHive和Giraph Framework等工具的设置方法和应用实例，进一步了解大数据分析的具体方法；最后介绍了高度综合大数据存储、实时查询及分析功能为一体的NoSQL技术，并详细讲解了Hadoop生态界中的NoSQL——HBase技术。

2017年春 于西南

[显示全部信息](#)

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

[更多资源请访问www.tushupdf.com](http://www.tushupdf.com)