

## <<Hadoop基础教程>>

### 图书基本信息

书名：<<Hadoop基础教程>>

13位ISBN编号：9787115341334

10位ISBN编号：7115341338

出版时间：2014-1

出版时间：人民邮电出版社

作者：[英] Garry Turkington

译者：张治起

版权说明：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：<http://www.tushu007.com>

## <<Hadoop基础教程>>

### 内容概要

Hadoop和云服务出现的历史背景，以及何时适用Hadoop的背景知识

安装并配置Hadoop集群的最佳方式，根据手头的问题调整系统配置

用Java和Ruby示例程序讲解如何编写运行在Hadoop上的程序

Amazon网络服务提供的托管Hadoop集群的运行方式，以及它与用户直接管理的Hadoop集群有何区别

Hadoop与关系数据库的融合，使用Hive执行SQL查询，使用Sqoop迁移数据

组成Hadoop生态系统的其他项目和工具，以及Hadoop的发展方向

针对初学者的有效方法

通过清晰操作步骤讲解最有用的任务

边干边学——立刻动手实践

摆脱枯燥的二进制

有启发意义的理想的案例，能够带给读者灵感，从而解决面临的问题

促进读者动手练习的作业和习题

## <<Hadoop基础教程>>

### 作者简介

#### 作者简介：

Garry Turkington

拥有14年行业经验，其大部分时间都专注于大型分布式系统的设计与实现。

目前，他在Improve Digital公司担任数据工程部副总裁和公司的首席架构师。

他主要负责实现可以存储、处理并从公司海量数据中挖掘潜在价值的系统。

在加入 Improve Digital公司之前，他曾在Amazon 英国公司领导着几个软件开发团队，他们开发的系统用于处理Amazon为全世界所有对象创建的目录数据。

在此之前，他还曾在英国和美国政府机关任职十年。

他在北爱尔兰的贝尔法斯特女王大学获得了计算机学士和博士学位，并在美国斯蒂文斯理工学院获得系统工程的工程硕士学位。

#### 译者简介：

张治起

Hadoop技术爱好者和研究者，对Hadoop技术有非常深刻的认识和理解，热切关注Hadoop和相关大数据处理技术。

有着丰富的实践经验，热衷于技术分享，致力于不断探索揭开Hadoop的神秘面纱，帮助更多初学者接触和理解Hadoop。

## &lt;&lt;Hadoop基础教程&gt;&gt;

## 书籍目录

第1章 绪论	1
1.1 大数据处理	1
1.1.1 数据的价值	2
1.1.2 受众较少	2
1.1.3 一种不同的方法	4
1.1.4 Hadoop	7
1.2 基于Amazon Web Services的云计算	12
1.2.1 云太多了	12
1.2.2 第三种方法	12
1.2.3 不同类型的成本	12
1.2.4 AWS: Amazon的弹性架构	13
1.2.5 本书内容	14
1.3 小结	15
第2章 安装并运行Hadoop	16
2.1 基于本地Ubuntu主机的Hadoop系统	16
2.2 实践环节: 检查是否已安装JDK	17
2.3 实践环节: 下载Hadoop	18
2.4 实践环节: 安装SSH	19
2.5 实践环节: 使用Hadoop计算圆周率	20
2.6 实践环节: 配置伪分布式模式	22
2.7 实践环节: 修改HDFS的根目录	24
2.8 实践环节: 格式化NameNode	25
2.9 实践环节: 启动Hadoop	26
2.10 实践环节: 使用HDFS	27
2.11 实践环节: MapReduce的经典入门程序——字数统计	28
2.12 使用弹性MapReduce	33
2.13 实践环节: 使用管理控制台在EMR运行WordCount	34
2.13.1 使用EMR的其他方式	41
2.13.2 AWS生态系统	42
2.14 本地Hadoop与EMR Hadoop的对比	42
2.15 小结	43
第3章 理解MapReduce	44
3.1 键值对	44
3.1.1 具体含义	44
3.1.2 为什么采用键/值数据	45
3.1.3 MapReduce作为一系列键/值变换	46
3.2 MapReduce的Hadoop Java API	47
3.3 编写MapReduce程序	50
3.4 实践环节: 设置classpath	50
3.5 实践环节: 实现WordCount	51
3.6 实践环节: 构建JAR文件	53
3.7 实践环节: 在本地Hadoop集群运行WordCount	54
3.8 实践环节: 在EMR上运行WordCount	54
3.8.1 0.20之前版本的Java MapReduce API	56
3.8.2 Hadoop提供的mapper和reducer实现	57

## &lt;&lt;Hadoop基础教程&gt;&gt;

3.9	实践环节：WordCount的简易方法	58
3.10	查看WordCount的运行全貌	59
3.10.1	启动	59
3.10.2	将输入分块	59
3.10.3	任务分配	60
3.10.4	任务启动	60
3.10.5	不断监视JobTracker	60
3.10.6	mapper的输入	61
3.10.7	mapper的执行	61
3.10.8	mapper的输出和reducer的输入	61
3.10.9	分块	62
3.10.10	可选分块函数	62
3.10.11	reducer类的输入	62
3.10.12	reducer类的执行	63
3.10.13	reducer类的输出	63
3.10.14	关机	63
3.10.15	这就是MapReduce的全部	64
3.10.16	也许缺了combiner	64
3.11	实践环节：使用combiner编写WordCount	64
3.12	实践环节：更正使用combiner的WordCount	65
3.13	Hadoop专有数据类型	67
3.13.1	Writable和Writable-Comparable接口	67
3.13.2	wrapper类介绍	68
3.14	实践环节：使用Writable包装类	69
3.15	输入/输出	71
3.15.1	文件、split和记录	71
3.15.2	InputFormat和RecordReader	71
3.15.3	Hadoop提供的InputFormat	72
3.15.4	Hadoop提供的RecordReader	73
3.15.5	OutputFormat和Record-Writer	73
3.15.6	Hadoop提供的OutputFormat	73
3.15.7	别忘了Sequence files	74
3.16	小结	74
第4章	开发MapReduce程序	75
4.1	使用非Java语言操作Hadoop	75
4.1.1	Hadoop Streaming工作原理	76
4.1.2	使用Hadoop Streaming的原因	76
4.2	实践环节：使用Streaming实现Word-Count	76
4.3	分析大数据集	79
4.3.1	获取UFO目击事件数据集	79
4.3.2	了解数据集	80
4.4	实践环节：统计汇总UFO数据	80
4.5	实践环节：统计形状数据	82
4.6	实践环节：找出目击事件的持续时间与UFO形状的关系	84
4.7	实践环节：在命令行中执行形状/时间分析	87
4.8	实践环节：使用ChainMapper进行字段验证/分析	88
4.9	实践环节：使用Distributed Cache改进地点输出	93

## &lt;&lt;Hadoop基础教程&gt;&gt;

4.10	计数器、状态和其他输出	96
4.11	实践环节：创建计数器、任务状态和写入日志	96
4.12	小结	102
第5章	高级MapReduce技术	103
5.1	初级、高级还是中级	103
5.2	多数据源联结	103
5.2.1	不适合执行联结操作的情况	104
5.2.2	map端联结与reduce端联结的对比	104
5.2.3	匹配账户与销售信息	105
5.3	实践环节：使用MultipleInputs实现reduce端联结	105
5.3.1	实现map端联结	109
5.3.2	是否进行联结	112
5.4	图算法	112
5.4.1	Graph 101	112
5.4.2	图和MapReduce	112
5.4.3	图的表示方法	113
5.5	实践环节：图的表示	114
5.6	实践环节：创建源代码	115
5.7	实践环节：第一次运行作业	119
5.8	实践环节：第二次运行作业	120
5.9	实践环节：第三次运行作业	121
5.10	实践环节：第四次也是最后一次运行作业	122
5.10.1	运行多个作业	124
5.10.2	关于图的终极思考	124
5.11	使用语言无关的数据结构	124
5.11.1	候选技术	124
5.11.2	Avro简介	125
5.12	实践环节：获取并安装Avro	125
5.13	实践环节：定义模式	126
5.14	实践环节：使用Ruby创建Avro源数据	127
5.15	实践环节：使用Java语言编程操作Avro数据	128
5.16	实践环节：在MapReduce中统计UFO形状	130
5.17	实践环节：使用Ruby检查输出数据	134
5.18	实践环节：使用Java检查输出数据	135
5.19	小结	137
第6章	故障处理	138
6.1	故障	138
6.1.1	拥抱故障	138
6.1.2	至少不怕出现故障	139
6.1.3	严禁模仿	139
6.1.4	故障类型	139
6.1.5	Hadoop节点故障	139
6.2	实践环节：杀死DataNode进程	141
6.3	实践环节：复制因子的作用	144
6.4	实践环节：故意造成数据块丢失	146
6.5	实践环节：杀死TaskTracker进程	149
6.6	实践环节：杀死JobTracker	153

## &lt;&lt;Hadoop基础教程&gt;&gt;

6.7	实践环节：杀死NameNode进程	154
6.8	实践环节：引发任务故障	160
6.9	数据原因造成的任务故障	163
6.10	实践环节：使用skip模式处理异常数据	164
6.11	小结	169
第7章 系统运行与维护		170
7.1	关于EMR的说明	170
7.2	Hadoop配置属性	171
7.3	实践环节：浏览默认属性	171
7.3.1	附加的属性元素	172
7.3.2	默认存储位置	172
7.3.3	设置Hadoop属性的几种方式	173
7.4	集群设置	174
7.4.1	为集群配备多少台主机	174
7.4.2	特殊节点的需求	176
7.4.3	不同类型的存储系统	177
7.4.4	Hadoop的网络配置	178
7.5	实践环节：查看默认的机柜配置	180
7.6	实践环节：报告每台主机所在机柜	180
7.7	集群访问控制	183
7.8	实践环节：展示Hadoop的默认安全机制	183
7.9	管理NameNode	187
7.10	实践环节：为fsimage文件新增一个存储路径	188
7.11	实践环节：迁移到新的NameNode主机	190
7.12	管理HDFS	192
7.12.1	数据写入位置	192
7.12.2	使用平衡器	193
7.13	MapReduce管理	193
7.13.1	通过命令行管理作业	193
7.13.2	作业优先级和作业调度	194
7.14	实践环节：修改作业优先级并结束作业运行	194
7.15	扩展集群规模	197
7.15.1	提升本地Hadoop集群的计算能力	197
7.15.2	提升EMR作业流的计算能力	198
7.16	小结	198
第8章 Hive：数据的关系视图		200
8.1	Hive概述	200
8.1.1	为什么使用Hive	200
8.1.2	感谢Facebook	201
8.2	设置Hive	201
8.2.1	准备工作	201
8.2.2	下载Hive	202
8.3	实践环节：安装Hive	202
8.4	使用Hive	203
8.5	实践环节：创建UFO数据表	204
8.6	实践环节：在表中插入数据	206
8.7	实践环节：验证表	208

## &lt;&lt;Hadoop基础教程&gt;&gt;

8.8	实践环节：用正确的列分隔符重定义表	210
8.9	实践环节：基于现有文件创建表	212
8.10	实践环节：执行联结操作	214
8.11	实践环节：使用视图	216
8.12	实践环节：导出查询结果	219
8.13	实践环节：制作UFO目击事件分区表	221
8.13.1	分桶、归并和排序	224
8.13.2	用户自定义函数	225
8.14	实践环节：新增用户自定义函数	225
8.14.1	是否进行预处理	228
8.14.2	Hive和Pig的对比	229
8.14.3	未提到的内容	229
8.15	基于Amazon Web Services的Hive	230
8.16	实践环节：在EMR上分析UFO数据	230
8.16.1	在开发过程中使用交互式作业流	235
8.16.2	与其他AWS产品的集成	236
8.17	小结	236
第9章	与关系数据库协同工作	238
9.1	常见数据路径	238
9.1.1	Hadoop用于存储档案	238
9.1.2	使用Hadoop进行数据预处理	239
9.1.3	使用Hadoop作为数据输入工具	239
9.1.4	数据循环	240
9.2	配置MySQL	240
9.3	实践环节：安装并设置MySQL	240
9.4	实践环节：配置MySQL允许远程连接	243
9.5	实践环节：建立员工数据库	245
9.6	把数据导入Hadoop	246
9.6.1	使用MySQL工具手工导入	246
9.6.2	在mapper中访问数据库	246
9.6.3	更好的方法：使用Sqoop	247
9.7	实践环节：下载并配置Sqoop	247
9.8	实践环节：把MySQL的数据导入HDFS	249
9.9	实践环节：把MySQL数据导出到Hive	253
9.10	实践环节：有选择性的导入数据	255
9.11	实践环节：使用数据类型映射	257
9.12	实践环节：通过原始查询导入数据	258
9.13	从Hadoop导出数据	261
9.13.1	在reducer中把数据写入关系数据库	261
9.13.2	利用reducer输出SQL数据文件	262
9.13.3	仍是最好的方法	262
9.14	实践环节：把Hadoop数据导入MySQL	262
9.15	实践环节：把Hive数据导入MySQL	265
9.16	实践环节：改进mapper并重新运行数据导出命令	267
9.17	在AWS上使用Sqoop	269
9.18	小结	270



## &lt;&lt;Hadoop基础教程&gt;&gt;

第10章 使用Flume收集数据	271
10.1 关于AWS的说明	271
10.2 无处不在的数据	271
10.2.1 数据类别	272
10.2.2 把网络流量导入Hadoop	272
10.3 实践环节：把网络服务器数据导入Hadoop	272
10.3.1 把文件导入Hadoop	273
10.3.2 潜在的问题	273
10.4 Apache Flume简介	274
10.5 实践环节：安装并配置Flume	275
10.6 实践环节：把网络流量存入日志文件	277
10.7 实践环节：把日志输出到控制台	279
10.8 实践环节：把命令的执行结果写入平面文件	281
10.9 实践环节：把远程文件数据写入本地平面文件	283
10.9.1 信源、信宿和信道	284
10.9.2 Flume配置文件	286
10.9.3 一切都以事件为核心	287
10.10 实践环节：把网络数据写入HDFS	287
10.11 实践环节：加入时间戳	289
10.12 实践环节：多层Flume网络	292
10.13 实践环节：把事件写入多个信宿	294
10.13.1 选择器的类型	295
10.13.2 信宿故障处理	295
10.13.3 使用简单元件搭建复杂系统	296
10.14 更高的视角	297
10.14.1 数据的生命周期	297
10.14.2 集结数据	297
10.14.3 调度	297
10.15 小结	298
第11章 展望未来	299
11.1 全书回顾	299
11.2 即将到来的Hadoop变革	300
11.3 其他版本的Hadoop软件包	300
11.4 其他Apache项目	303
11.4.1 HBase	303
11.4.2 Oozie	303
11.4.3 Whir	304
11.4.4 Mahout	304
11.4.5 MRUnit	305
11.5 其他程序设计模式	305
11.5.1 Pig	305
11.5.2 Cascading	305
11.6 AWS资源	306
11.6.1 在EMR上使用HBase	306
11.6.2 SimpleDB	306
11.6.3 DynamoDB	306
11.7 获取信息的渠道	307

## <<Hadoop基础教程>>

11.7.1	源代码	307
11.7.2	邮件列表和论坛	307
11.7.3	LinkedIn群组	307
11.7.4	Hadoop用户群	307
11.7.5	会议	308
11.8	小结	308
	随堂测验答案	309

## <<Hadoop基础教程>>

### 版权说明

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:<http://www.tushu007.com>