

<<Pig编程指南>>

图书基本信息

书名：<<Pig编程指南>>

13位ISBN编号：9787115301116

10位ISBN编号：7115301115

出版时间：2013-2

出版时间：人民邮电出版社

作者：盖茨

译者：曹坤

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>

## <<Pig编程指南>>

### 内容概要

《Pig编程指南》不仅为初学者讲解ApachePig的基础知识，同时也向有一定使用经验的高级用户介绍更加综合全面的Pig重要特性，如PigLatin脚本语言、控制台shell交互命令以及用于对Pig进行拓展的用户自定义函数（UDF）等。

当读者有大数据处理需求时，提供了如何更高效地使用Pig来完成需求的方法。

## <<Pig编程指南>>

### 作者简介

alan gates 是将PIG从雅虎的研究项目转化成一个成功的Apache开源项目的工程师团队中最早的成员。他负责监督Pig的实现、编程接口和总体设计。

## 书籍目录

## 第1章初识Pig 1.1Pig是什么？

1.1.1Pig是基于Hadoop的 1.1.2PigLatin，一种并行数据流语言 1.1.3Pig的用途 1.1.4Pig的设计思想 1.2Pig发展简史 第2章安装和运行Pig 2.1下载和安装Pig 2.1.1从Apache下载Pig软件包 2.1.2从Cloudera下载Pig 2.1.3使用Maven下载Pig 2.1.4下载Pig源码 2.2运行Pig 2.2.1本地单机运行Pig 2.2.2在Hadoop集群上运行Pig 2.2.3在云服务上运行Pig 2.2.4命令行使用以及配置选项介绍 2.2.5返回码 第3章命令行交互工具Grunt 3.1在Grunt中输入Pig Latin脚本 3.2在Grunt中使用HDFS命令 3.3在Grunt中控制Pig 第4章Pig数据模型 4.1数据类型 4.1.1基本类型 4.1.2复杂类型 4.1.3NULL值 4.2模式 第5章PigLatin介绍 5.1基础知识 5.1.1大小写敏感 5.1.2注释 5.2输入和输出 5.2.1加载 5.2.2存储 5.2.3输出 5.3关系操作 5.3.1foreach 5.3.2Filter 5.3.3Group 5.3.4Orderby 5.3.5Distinct 5.3.6Join 5.3.7Limit 5.3.8Sample 5.3.9Parallel 5.4用户自定义函数UDF 5.4.1注册UDF 5.4.2define命令和UDF 5.4.3调用静态Java函数 第6章PigLatin高级应用 6.1高级关系操作 6.1.1foreach的高级功能 6.1.2使用不同的Join实现方法 6.1.3cogroup 6.1.4union 6.1.5cross 6.2在Pig中集成遗留代码和Map Reduce程序 6.2.1stream 6.2.2mapreduce 6.3非线性数据流 6.4执行过程控制 6.4.1set 6.4.2设置分割器 6.5PigLatin预处理器 6.5.1参数传入 6.5.2宏 6.5.3包含其他的Pig Latin脚本 第7章开发和测试Pig Latin脚本 7.1开发工具 7.1.1语法高亮和语法检查 7.1.2describe 7.1.3explain 7.1.4illustrate 7.1.5Pig统计信息 7.1.6Map Reduce任务运行状态信息 7.1.7调试技巧 7.2使用Pig Unit测试用户的脚本 第8章让Pig飞起来 8.1编写优质的脚本 8.1.1尽早地并经常地进行过滤 8.1.2尽早地并经常地进行映射 8.1.3正确并合理使用join 8.1.4适当的情况下使用multiquery 8.1.5选择正确的数据类型 8.1.6选择合适的并行值 8.2编写优质的UDF 8.3调整Pig和Hadoop 8.4对计算中间结果进行压缩 8.5数据层优化 8.6垃圾数据处理 第9章在Python中嵌入Pig Latin脚本 9.1编译 9.2绑定 9.3运行 9.4工具方法 第10章编写评估函数和过滤函数 10.1使用Java编写评估函数 10.1.1UDF将在哪里执行 10.1.2求值函数基本概念 10.1.3输入和输出模式 10.1.4错误处理和过程信息报告 10.1.5构造器和将数据从前端传送到后端 10.1.6重载UDF 10.1.7运算函数的内存问题 10.2代数运算接口 10.3累加器接口 10.4使用Python写UDF 10.5书写过滤器函数 第11章编写加载函数和存储函数 11.1加载函数 11.1.1前端执行计划函数 11.1.2从前端调用传递信息到后端调用 11.1.3后端数据读取 11.1.4可扩展的加载函数接口 11.2存储函数 11.2.1存储函数前端执行计划 11.2.2存储函数和UDF Context 11.2.3写数据 11.2.4任务失败后数据的清理 11.2.5存储元数据信息 第12章Pig和其他Hadoop社区的成员 12.1Pig和Hive 12.2Cascading 12.3NoSQL数据库 12.3.1HBase 12.3.2Cassandra 12.4Hadoop中的元数据 附录A内置的用户自定义函数和Piggybank 内置UDF 内置加载函数和存储函数 内置求值函数和过滤函数 Piggybank 附录BHadoop综述 Map Reduce Map阶段 Combiner阶段 Shuffle阶段 Reduce阶段 输出阶段 分布式缓存 故障处理 HDFS 作者介绍 书末说明

## 章节摘录

版权页： bytearray 一团或一组字节。

bytearray是通过封装了Java的byte[]的DataByteArray Java类来实现的。

没有办法去定义一个bytearray常量。

4.1.2复杂类型 Pig有3个复杂数据类型：map、mple和bag。

这3种类型都可以包含任意类型的数据，包括其他复杂类型的数据。

所以如果有一个map，它的值字段是bag类型，这个bag包含了一个tuple，而该mple的字段是map，这种情况是可以存在的。

Map Pig中的map是一种chararray和数据元素之间的键值对映射，其中数据元素可以是任意的Pig类型，包括复杂类型。

其中的chararray被称为键（key），它作为查找对应数据元素的索引，相应的数据元素被称为值（value）。

因为Pig不知道值的类型，那么它就会假设值为bytearray类型，尽管实际的值可能为其他类型。

如果用户想知道真实的数据类型是什么（或者用户想让它成为什么数据类型），用户可以对它进行类型转换，相关信息请查看4.2.1节“类型转换”。

用户如果没有显式地对值进行类型转换，那么Pig将会根据用户在脚本中如何使用这个值将其转换成一个最有可能的类型。

如果值是bytearray外的其他类型，那么Pig会在运行时获得数据类型然后进行处理。

关于Pig如何处理未知数据类型的更多信息请查看4.2节“模式”。

默认情况下并不要求一个map中的所有值具有相同的数据类型。

一个map包含两个键：name和age，其中name对应的值是chararray类型，而age对应的值是int类型，像这种情况是合法的。

从Pig 0.9版本开始，map可以将它的值声明为具有相同的数据类型。

这个新功能是非常有帮助的，因为如果用户事先就知道map中所有的值都是具有相同的数据类型的话，那么就可以避免进行类型转换，而且Pig也就无需在执行阶段运行时对数据类型进行控制。

map常量通过方括号来划定map结构，键和值间是一个#号，键值对之间使用逗号分隔。

例如：['name'#'bob', 'age'#55] 将创建一个包含“name”和“age”两个键的map。

第一个值是chararray类型的，第二个值是一个整数。

## <<Pig编程指南>>

### 编辑推荐

Apache Pig 是一个高级过程语言，适合于使用 Hadoop 和 MapReduce 平台来查询大型半结构化数据集。

通过允许对分布式数据集进行类似 SQL 的查询，Pig 可以简化 Hadoop 的使用。

本文不仅为初学者讲授，Pig 的基础知识，同时还向有经验的用户更加全面的介绍Pig的重点特性。

通过学习本书，你将能够身日了解数据模型，包括基本数据和复杂数据类型。

掌握更高效的在Hadoop集群中运行脚本的方法和技巧。

#### 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>