<<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

<<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。

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。

<<Pig编程指南>>

书籍目录

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，像这种隋况是合法的。
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