

<<Hadoop云计算实战>>

图书基本信息

书名：<<Hadoop云计算实战>>

13位ISBN编号：9787302296737

10位ISBN编号：7302296731

出版时间：2012-10

出版时间：清华大学出版社

作者：周 品

页数：411

字数：612000

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云计算实战>>

内容概要

《hadoop云计算实战》全面介绍了云计算的基本概念、google（谷歌）云计算的关键技术，以及hadoop云计算的相关配套项目及其实战，包括hadoop的hdfs、mapreduce、hbase、hive、pig、cassandra、chukwa及zookeeper等配套项目的实现机制、用法及应用。

《hadoop云计算实战》可作为高等院校本科生和研究生的教材，也可作为广大科研人员、学者、工程技术人员的参考用书。

<<Hadoop云计算实战>>

书籍目录

《hadoop云计算实战》

第1章 云计算概论

1. 云计算概述

1.1. 云计算的定义

1.1. 云计算产生的背景

1.1. 云时代谁是主角

1.1. 云计算的特征

1.1. 云计算的发展史

1.1. 云计算的服务层次

1.1. 云计算的服务形式

1.1. 云计算的实现机制

1.1. 云计算研究方向

1.1. 云计算发展趋势

1. 云计算关键技术研究

1.2. 虚拟化技术

1.2. 数据存储技术

1.2. 资源管理技术

1.2. 能耗管理技术

1.2. 云监测技术

1. 云计算应用研究

1.3. 语义分析应用

1.3. it企业应用

1.3. 生物学应用

1.3. 电信企业应用

1.3. 数据库的应用

1.3. 地理信息应用

1.3. 医学应用

1. 云安全

1.4. 云安全发展趋势

1.4. 云安全与网络安全的差别

1.4. 云安全研究的方向

1.4. 云安全难点问题

1.4. 云安全新增及增强功能

1. 云计算生命周期

1. 云计算存在的问题

1. 云计算的优缺点

第2章 hadoop相关项目介绍

2. hadoop简介

2.1. hadoop的基本架构

2.1. hadoop文件系统结构

2.1. hadoop文件读操作

2.1. hadoop文件写操作

2. hadoop系统性质

2.2. 可靠存储性

2.2. 数据均衡

<<Hadoop云计算实战>>

2. 比较sql数据库与hadoop

2. mapreduce概述

2.4. mapreduce实现机制

2.4. mapreduce执行流程

2.4. mapreduce映射和化简

2.4. mapreduce输入格式

2.4. mapreduce输出格式

2.4. mapreduce运行速度

2. hbase概述

2.5. hbase的系统框架

2.5. hbase访问接口

2.5. hbase的存储格式

2.5. hbase的读写流程

2.5. hbase的优缺点

2. zookeeper概述

2.6. 为什么需要zookeeper

2.6. zookeeper设计目标

2.6. zookeeper数据模型

2.6. zookeeper工作原理

2.6. zookeeper实现机制

2.6. zookeeper的特性

2. hive概述

2.7. hive的组成

2.7. hive结构解析

2. pig概述

2. cassandra概述

2.9. cassandra主要功能

2.9. cassandra的体系结构

2.9. cassandra存储机制

2.9. cassandra的写过程

2.9. cassandra的读过程

2.9. cassandra的删除

2. chukwa概述

2.10. 使用chukwa的原因

2.10. chukwa的不是

2.10. chukwa的定义

2.10. chukwa架构与设计

第3章 hadoop配置与实战

3. hadoop的安装

3.1. 在linux下安装hadoop

3.1. 运行模式

3.1. 在windows下安装hadoop

3. 运行hadoop

3. hadoop的avatar机制

3.3. 系统架构

3.3. 元数据同步机制

3.3. 切换故障过程

<<Hadoop云计算实战>>

- 3.3. 运行流程
- 3.3. 切换故障流程
- 3. hadoop实战
- 3.4. 使用hadoop运行wordcount实例
- 3.4. 使用eclipse编写hadoop程序
- 第4章 hadoop的分布式数据hdfs
- 4. hdfs的操作
- 4.1. 文件操作
- 4.1. 管理与更新
- 4. fs shell使用指南
- 4. api使用
- 4.3. 文件系统的常见操作
- 4.3. api的java操作实例
- 第5章 hadoop编程模型mapreduce
- 5. mapreduce基础
- 5.1. mapreduce编程模型
- 5.1. mapreduce实现机制
- 5.1. java mapreduce
- 5. mapreduce的容错性
- 5. mapreduce实例分析
- 5. 不带map()、reduce()的mapreduce
- 5. shuffle过程
- 5. 新增hadoop api
- 5. hadoop的streaming
- 5.7. 通过unix命令使用streaming
- 5.7. 通过ruby版本使用streaming
- 5.7. 通过python版本使用streaming
- 5. mapreduce实战
- 5.8. mapreduce排序
- 5.8. mapreduce二次排序
- 5. mapreduce作业分析
- 5. 定制mapreduce数据类型
- 5.10. 内置的数据输入格式和recordreader
- 5.10. 定制输入数据格式与recordreader
- 5.10. 定制数据输出格式实现多集合文件输出
- 5. 链接mapreduce作业
- 5.11. 顺序链接mapreduce作业
- 5.11. 复杂的mapreduce链接
- 5.11. 前后处理的链接
- 5.11. 链接不同的数据
- 5. hadoop的pipes
- 5. 创建bloom filter
- 5.13. bloom filter作用
- 5.13. bloom filter实现
- 第6章 hadoop的数据库hbase
- 6. hbase数据模型
- 6.1. 数据模型

<<Hadoop云计算实战>>

- 6.1. 概念视图
- 6.1. 物理视图
- 6. hbase与rdbms对比
- 6. bigtable的应用实例
- 6. hbase的安装与配置
- 6. java api
- 6. hbase实例分析
- 6.6. rowlock
- 6.6. hbase的hfileoutputformat
- 6.6. hbase的tableoutputformat
- 6.6. 在hbase中使用mapreduce
- 6.6. hbase分布式模式
- 第7章 hadoop的数据仓库hive
- 7. hive的安装
- 7.1. 准备的软件包
- 7.1. 内嵌模式安装
- 7.1. 安装独立模式
- 7.1. 远程模式安装
- 7.1. 查看数据信息
- 7. hive的入口
- 7.2. 类clidriver
- 7.2. 类clisessionstate
- 7.2. 类commandprocessor
- 7. hive ql详解
- 7.3. hive的数据类型
- 7.3. hive与数据库比较
- 7.3. ddl操作
- 7.3. join查询
- 7.3. dml操作
- 7.3. sql操作
- 7.3. hive ql的应用实例
- 7. hive的服务
- 7.4. jdbc/odbc服务
- 7.4. thrift服务
- 7.4. web接口
- 7. hive sql的优化
- 7.5. hive sql优化选项
- 7.5. hive sql优化应用实例
- 7. hive的扩展性
- 7.6. serde
- 7.6. map/reduce脚本
- 7.6. udf
- 7.6. udaf
- 7. hive实战
- 第8章 hadoop的大规模数据平台pig
- 8. pig的安装与运行
- 8.1. pig的安装

<<Hadoop云计算实战>>

- 8.1. pig的运行
- 8. pig实现
- 8. pig latin语言
- 8.3. pig latin语言概述
- 8.3. pig latin数据类型
- 8.3. pig latin运算符
- 8.3. pig latin关键字
- 8.3. pig内置函数
- 8. 自定义函数
- 8.4. udf的编写
- 8.4. udfs的使用
- 8. jaql和pig查询语言的比较
- 8.5. pig和jaql运行环境和执行形式的比较
- 8.5. pig和jaql支持数据类型的比较
- 8.5. pig和jaql操作符和内建函数以及自定义函数的比较
- 8.5. 其他
- 8. pig实战
- 第9章 hadoop的非关系型数据cassandra
- 9. cassandra的安装
- 9.1. 在windows 7中安装
- 9.1. 在linux中安装
- 9. cassandra的数据模型
- 9.2. column
- 9.2. supercolumn
- 9.2. columnfamily
- 9.2. row
- 9.2. 排序
- 9. cassandra的实例分析
- 9.3. cassandra的数据存储结构
- 9.3. 跟踪客户端代码
- 9. cassandra常用的编程语言
- 9.4. java使用cassandra
- 9.4. php使用cassandra
- 9.4. python使用cassandra
- 9.4. c#使用cassandra
- 9.4. ruby使用cassandra
- 9. cassandra与mapreduce结合
- 9.5. 需求分析
- 9.5. 代码分析
- 9.5. mapreduce代码
- 9. cassandra实战
- 9.6. buyerdao功能验证
- 9.6. sellerdao功能验证
- 9.6. productdao功能验证
- 9.6. 新建schema在线功能
- 9.6. 功能验证
- 第10章 hadoop的收集数据chukwa

<<Hadoop云计算实战>>

10. chukwa的安装与配置

10.1. 配置要求

10.1. chukwa的安装

10.1. 基本命令

10. chukwa数据流处理

10.2. 支持数据类型

10.2. 数据处理

10.2. 自定义数据模块

10. chukwa源代码分析

10.3. chukwa适配器

10.3. chukwa连接器

10.3. chukwa收集器

10. chukwa实例分析

10.4. 生成数据

10.4. 收集数据

10.4. 处理数据

10.4. 析取数据

10.4. 稀释数据

第11章 hadoop的分布式系统zookeeper

11. zookeeper的安装与配置

11.1. zookeeper的安装

11.1. zookeeper的配置

11.1. zookeeper数据模型

11.1. zookeeper的api接口

11.1. zookeeper编程实现

11. zookeeper的leader流程

11. zookeeper锁服务

11.3. zookeeper中的锁机制

11.3. zookeeper的写锁实现

11.3. zookeeper锁服务实现例子

11. 创建zookeeper应用程序

11. zookeeper的应用开发

11. zookeeper的典型应用

11.6. 统一命名服务

11.6. 配置管理

11.6. 集群管理

11.6. 共享锁

11.6. 队列管理

11. 实现namenode自动切换

网上参考资源

参考文献

章节摘录

版权页：插图：3.收集器 Chukwa的收集器弥补了Hadoop集群不利于存储大量小文件的缺点。收集器先是把收集到的小文件数据进行部分合并，然后写入集群，大幅减少了Chukwa产生的HDFS文件数量。

具体来说，通过HTTP数据被传送到收集器，每个收集器接收来自数百台主机的数据，并将所有数据写入到一个Sink文件中，MapReduce作业定期将Sink中记录的信息整合为日志收集文件。

Sink文件是一个由连续的Chunks组成的Hadoop序列文件，其是由大量的数据块和描述每一个数据块来源和格式的元数据组成的。

在收集数据期间，收集器会定期关闭Sink文件，更改文件名（便于保存及整理），重新创建一个新的文件，新文件仍被命名为“Sink”，接着再用新的Sink文件存储收集的信息，这就是所谓的“文件循环”。

收集器位于数据源和数据存储间，其屏蔽了HDFS文件系统的一些细节，方便于使用HDFS。

在某种意义上，收集器缓解了大量低速率数据源和文件系统间“步调”不协调的矛盾，优化了少量高速率数据源的写入。

为了防止收集器出现单点，Chukwa允许设置多台收集器，代理可以从收集器列表中随机地选择一个收集器传输数据。

当某个收集器失败或繁忙时，就选择其他收集器，以免影响代理的正常工作。

随机选择的节点使收集器的载入可能会极不均匀。

在实际应用中，收集器的任务负载很轻的情况很少出现；为了防止过载，系统设置了代理重试限制机制，如果数据写入收集器失败，收集器把待写入数据标记为“坏”数据，在重新写入数据前代理需要等待一段配置时间。

在实际应用中，多收集器的负载几乎是平均的，从而实现了负载的均衡化。

4.MapReduce作业 收集器顺序写入数据文件，方便于快速获取数据和稳定存储，但是，不便于数据分析和查找特征数据。

因此，Chukwa利用MapReduce作业实现数据分析和处理。

在MapReduce阶段，Chukwa提供了复用和存档任务两种内置的作业类型。

（1）demux作业 demux作业负责对数据的分类、排序和去重。

由收集器写入集群中的数据，都有自己的类型。

demux作业在执行过程中，通过数据类型和配置文件中指定的数据处理类，执行相应的数据分析工作，一般是把非结构化的数据结构化，抽取其中的数据属性。

由于demux的本质是一个MapReduce作业，所以用户可以根据自己的需求制定自己的demux作业，进行各种复杂的逻辑分析。

Chukwa提供的demux接口可以用Java语言来方便地扩展。

<<Hadoop云计算实战>>

编辑推荐

《Hadoop云计算实战》可作为高等院校本科生和研究生的教材，也可作为广大科研人员、学者、工程技术人员参考用书。

<<Hadoop云计算实战>>

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>