

<<典型Hadoop云计算>>

图书基本信息

书名：<<典型Hadoop云计算>>

13位ISBN编号：9787121188077

10位ISBN编号：7121188074

出版时间：2013-1

出版时间：电子工业出版社

作者：赵书兰

页数：475

字数：774000

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>

<<典型Hadoop云计算>>

前言

什么是云计算？

云计算是一种基于互联网的超级计算模式，在远程数据中心，几万台甚至几千万台计算机和服务器连接成一片。

云计算甚至可以让人们体验每秒超过10万亿次的运算能力，如此强大的运算能力几乎无所不能。

用户通过计算机、笔记本电脑、手机等方式接入数据中心，按各自的需求进行存储和运算。

云计算是传统计算机技术和网络技术发展融合的产物，也是引领未来信息产业创新的关键战略性技术和手段。

云计算的广泛普及与应用，也将催生信息技术的第三次生产变革浪潮，引发未来新一代信息技术变革、IT应用方式的核心变革，同时也将带来工作方式和商业模式的根本性变革。

作为21世纪IT业界乃至全社会关注的焦点和热点，未来云计算应用可以作为一种IT基础设施服务模式、一种信息服务的交付模式、一种基于互联网的新型商业模式，一种像供水、输电一样的创建性资源服务模式，将为人类社会提供更加方便、快捷、廉价的信息服务，为人们的工作、生活提供更多便利。

云计算作为新一代的技术变革，其发展受到了各行各业的关注，无论是政府部门，还是企事业单位，他们所签的如雪片一般的订单昭示了云计算的巨大市场。

中国把云计算提高到了战略发展项目的高度，2010年10月，发改委、工信部联合发文通知，加强中国云计算创新发展的总体布局，并开展了云计算服务创新发展的一系列试点示范工作。

“十二五”规划纲要中云计算也占据了引人注目的位置。

有了广阔的前景和市场，加之政策的扶持，云计算项目在中国大有遍地开花之势。

与以往的高新技术不同，云计算不光给大型的企业和单位带来了实惠，对于中小型企业来说，更是一种莫大的机遇。

过去，中小型企业人力资源不足，IT预算吃紧，动辄数百万元的IT设备和设备所带来的生产力提升离他们相距甚远，而现在，云计算为它们提供了大企业级的技术，前期成本低，升级方便。

这一新兴技术的重要性毋庸置疑，不过，它尚处于起步阶段，接下来会有一系列的变革。

云计算甚至可以抹平企业规模所导致的优劣差距，而且极有可能让小企业占据优势。

简言之，目前世界上功能最强、最具创新能力的技术已不再为大型企业所独享。

事实上，每个人都已经开始不同程度地享受到这一技术了。

人们前期已经热衷于对云计算技术概念的专注研讨，而这个术语被广泛地定义为一个“计算”的本身。

关于这个主题的文献已经被很多权威专家想象、分析描述如下。

一个毁灭性结构的架构。

一个21世纪的巨大潮流。

大的经济平均主义者。

一个全球社会和经济网络。

一个国家工业的毁灭。

媒体的民主时代。

一个国家终结的开始。

一个地球村的崛起。

全球恐怖组织扩展的加速器。

权力从西方向东方转移的加速器。

计算历史上的下一次革命。

数字和物理世界的融合。

有专业人士预测，到2020年，全球云计算市场规模将比现在增长5倍多，从406亿美元增长到2410亿美元以上，中小企业对于云计算的需求会逐年上升。

可以预见的是，随着云计算日趋走向成熟，小型企业出于越来越多的技术需求会更加依赖于云计算，

<<典型Hadoop云计算>>

而其成本和复杂性也会逐渐降低。

Hadoop云是一个实现了MapReduce计算模型的开源分布式并行编程框架。MapReduce的概念来源于Google实验室，它是一个简化并行计算的编程模型，适用于大规模集群上的海量数据处理。

随着2007年底该模式Java开源实现项目ApacheHadoop的出现，使得程序员可以轻松地编写分布式并程序，并将其运行于计算机集群上，完成海量数据的计算。

近两年，尤其是2012年，国内外采用MapReduce模型的应用也逐渐丰富起来，如像NTTKDDI和中国移动这类公司采用该模型分析用户信息，优化网络配置；美国供电局采用该模型来分析电网现状；包括VISA和JP摩根在内的金融公司采用该模型来分析股票数据；包括Amazon和ebay在内的零售商和电子商务公司也开始采用该模型；甚至部分生物公司也采用该模型来进行DNA测序和分析。

经过几年的发展，在所有的开源云计算系统里，Hadoop一直稳居第一。而本书也因此诞生。

本书共分为9章，各章主要介绍内容如下。

第1章：介绍云计算背景与Hadoop，主要包括云的基本概念、云计算的优缺点、云计算存在的问题及Hadoop的概述及使用等内容。

第2章：介绍Hadoop的安装与配置，主要包括Hadoop的安装、Hadoop的组件、Hadoop常用命令及Hadoop的I/O操作等内容。

第3章：介绍Hadoop云计算的MapReduce详解，主要介绍MapReduce的总体结构、MapReduce工作原理、MapReduce基础模板及MapReduce应用实例等内容。

第4章：介绍Hadoop云计算的ZooKeeper详解，主要介绍ZooKeeper的基本概念、ZooKeeper的安装、ZooKeeper的配置及ZooKeeper的典型应用等内容。

第5章：介绍Hadoop云计算的Mahout详解，主要介绍Mahout的安装与配置、Mahout的相关算法、NaiveBayes分类器及Mahout的应用等内容。

第6章：介绍Hadoop云计算的Avro详解，主要介绍Avro模式、Avro数据、Avro协议等内容。

第7章：介绍Hadoop云计算的Chukwa详解，主要介绍Chukwa的架构、Chukwa的安装与配置、Chukwa源代码分析及Chukwa的应用实践等内容。

第8章：介绍Hadoop云计算的其他相关项目，主要介绍Hadoop的HBase详解、Hadoop的Hive详解、Hadoop的Pig详解及Hadoop的Cassandra详解等内容。

第9章：介绍Hadoop云计算的综合实例，主要介绍Hadoop云计算在移动通信信令监控与查询方面的综合实例及Hadoop在Last.fm的应用这两个应用实例。

本书由赵书兰编著，此外参加编写的还有李晓东、丁伟雄、雷晓平、李娅、杨文茵、何正风、赵书梅、栾颖、刘志为、周灵、周品、张德丰、余智豪和赵新芬。

由于作者水平有限，加之时间仓促，书中难免存在不足之处，敬请广大读者批评指正。

编著者

<<典型Hadoop云计算>>

内容概要

本书系统地阐述了当今IT业界最热门的话题——云计算，全书共分为9章。第1章介绍云计算背景与Hadoop；第2章介绍Hadoop的安装与配置；第3~8章系统、详细地介绍了Hadoop的子项目及相关项目的基本概念和实例分析，主要包括Hadoop的MapReduce、ZooKeeper、Mahout、Avro、Chukwa、HBase、Hive、Pig及Cassandra等项目；第9章总结了Hadoop云计算的综合实例。

<<典型Hadoop云计算>>

书籍目录

目 录

第1章 云计算背景与Hadoop

1.1 云计算概述

1.1.1 云计算的定义

1.1.2 云计算的特点

1.1.3 云计算的发展史

1.1.4 云计算的发展趋势

1.1.5 云计算的基本特征

1.2 云计算的优缺点

1.3 云计算存在的问题

1.4 Hadoop使用

1.4.1 国外Hadoop使用

1.4.2 国内Hadoop使用

1.5 Hadoop概述

1.5.1 Hadoop的起源及简介

1.5.2 Hadoop的历史

1.5.3 Hadoop的功能与作用

1.5.4 Hadoop的优势

1.5.5 Hadoop的现状与趋势

1.6 Hadoop的总体结构与模块简介

1.7 Hadoop的数据管理

1.7.1 HDFS的数据管理

1.7.2 HBase的数据管理

1.7.3 Hive的数据管理

第2章 Hadoop的安装与配置

2.1 在Linux上安装Hadoop

2.1.1 安装单节点Hadoop

2.1.2 在Linux系统上安装多节点Hadoop

2.2 Windows安装Hadoop

2.2.1 单机模式下安装Hadoop

2.2.2 分布模式下安装Hadoop

2.3 Hadoop的组件

2.3.1 基本文件命令

2.3.2 剖析MapReduce

2.4 Hadoop的InputFormat与OutputFormat

2.4.1 InputFormat

2.4.2 OutputFormat

2.5 Hadoop的常用命令

2.5.1 常规选项

2.5.2 HDFS Shell命令

2.5.3 用户命令

2.5.4 管理命令

2.6 Hadoop的I/O操作

<<典型Hadoop云计算>>

- 2.6.1 I/O操作数据检查
- 2.6.2 压缩数据
- 2.6.3 数据I/O中的序列化操作
- 2.6.4 MapReduce文件类
- 2.7 Hadoop的权限管理
 - 2.7.1 概述
 - 2.7.2 用户分组管理
 - 2.7.3 作业管理
 - 2.7.4 提交作业
- 第3章 Hadoop云计算的MapReduce详解
 - 3.1 总体结构
 - 3.2 MapReduce的工作原理
 - 3.3 MapReduce的基础模板
 - 3.4 一个气象数据集
 - 3.4.1 数据格式
 - 3.4.2 使用UNIX工具进行数据分析
 - 3.4.3 使用Hadoop进行数据分析
 - 3.4.4 新的Java MapReduce API
 - 3.4.5 Combiner
 - 3.4.6 运行分布式MapReduce作业
 - 3.5 MapReduce的应用实例
 - 3.5.1 数据去重
 - 3.5.2 排序
 - 3.5.3 二次排序
 - 3.5.4 单表关联
 - 3.5.5 多表关联
 - 3.6 复合键值对的使用
 - 3.6.1 合并键值
 - 3.6.2 用复合键排序
 - 3.7 定制数据类型及格式
 - 3.7.1 定制数据类型
 - 3.7.2 定义数据输入格式与RecordReader
 - 3.7.3 定制数据输出格式与RecordWriter
 - 3.7.4 定制Partitioner
 - 3.7.5 定制Combiner
 - 3.8 shuffle与排序
 - 3.8.1 map端
 - 3.8.2 reduce端
 - 3.8.3 shuffle过程
 - 3.9 组合式MapReduce作业
 - 3.9.1 迭代MapReduce计算任务
 - 3.9.2 顺序组合式MapReduce

<<典型Hadoop云计算>>

作业的执行

3.9.3 前、后处理的链接

3.9.4 复杂依赖关系的组合式

3.10 使用DataJoin包实现Join

3.11 参数/数据文件的传递与使用

3.11.1 传递全局作业参数

3.11.2 查询全局MapReduce

作业属性

3.11.3 全局数据文件的传递

第4章 Hadoop云计算的ZooKeeper详解

4.1 ZooKeeper的基本概念

4.1.1 ZooKeeper的角色

4.1.2 设计目的

4.1.3 工作原理

4.1.4 工作流程

4.1.5 数据模型

4.2 ZooKeeper的安装

4.2.1 软件及环境要求

4.2.2 单机模式

4.2.3 集群模式

4.2.4 集群伪分布

4.2.5 启动并测试ZooKeeper

4.3 ZooKeeper的配置

4.3.1 Client和ZK集群的连接及
session的建立过程

4.3.2 关于ACL

4.3.3 关于Watcher

4.3.4 关于Log文件和snapshot

4.4 ZooKeeper的使用

4.4.1 ZooKeeper的简单操作

4.4.2 ZooKeeper的四字命令

4.4.3 ZooKeeper的命令行工具

4.4.4 常用接口列表

4.4.5 基本操作

4.4.6 ZooKeeper API的使用

4.5 ZooKeeper的特性

4.5.1 ZooKeeper的会话及状态

4.5.2 读写性能测试

4.5.3 可靠性测试

4.6 ZooKeeper的典型应用

4.6.1 统一命名服务 (Name
Service)

4.6.2 配置管理 (Configuration
Management)

4.6.3 集群管理 (Group
Membership)

4.6.4 分布式锁

<<典型Hadoop云计算>>

4.6.5 共享锁 (Locks)

4.6.6 队列管理

4.7 ZooKeeper实例详解

4.8 进程调度系统

4.8.1 设计方案

4.8.2 设计实现

第5章 Hadoop云计算的Mahout详解

5.1 Mahout概述

5.2 Mahout的安装与配置

5.3 Mahout API简介

5.4 Mahout的相关算法

5.4.1 分类算法

5.4.2 聚类算法

5.4.3 Slope one算法

5.5 应用Mahout建立一个推荐引擎

5.5.1 推荐引擎概述

5.5.2 推荐引擎的实现代码

5.6 运行Naive Bayes分类器

5.6.1 生成分类结果的前提准备

5.6.2 实现生成分类结果代码

5.7 基于Mahout的应用

5.7.1 应用构架

5.7.2 应用实例

第6章 Hadoop云计算的Avro详解

6.1 Avro概述

6.2 Avro模式

6.2.1 Avro模式声明

6.2.2 组件化Avro模式

6.3 Avro数据

6.3.1 数据序列化

6.3.2 数据排序顺序

6.4 Avro协议

6.4.1 Avro协议声明

6.4.2 协议传输格式

6.5 使用Avro实现继承

6.6 使用Avro实现多态性

6.7 使用Avro的向后兼容性

6.8 Avro的C/C++实现

6.9 Avro的Java实现

6.10 Avro IDL语言

第7章 Hadoop云计算的Chukwa详解

7.1 初识Chukwa

7.1.1 Chukwa不是什么

7.1.2 Chukwa是什么

7.1.3 Chukwa流水线数据处理

7.2 Chukwa架构

7.2.1 adaptors和agents

<<典型Hadoop云计算>>

7.2.2 collector与Demux

7.2.3 HICC

7.2.4 dbadmin

7.2.5 数据接口与

<<典型Hadoop云计算>>

章节摘录

版权页： 插图：

<<典型Hadoop云计算>>

编辑推荐

《典型Hadoop云计算》由赵书兰编写，共分为9章，内容包括：云计算背景与Hadoop，Hadoop的安装与配置，Hadoop云计算的MapReduce、ZooKeeper、Mahout、Avro、Chukwa详解，Hadoop云计算的其他相关项目和综合实例。

可以作为高等学校教材，也可供广大科研人员、工程技术人员自学或参考。

<<典型Hadoop云计算>>

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>