我司CTO马延辉《HBase企业应用开发实战》新书上市

7天+实验掌握大数据技术

7天大容量+大牛授课+实战经验分享+详细实验手册+2套实验环境及实验数据

扫一扫，送新书

2015寒假（第二届）

全国高校大数据Hadoop师资培训班

时间：2015年1月26日–2月1日　地点：北京理工大学

2015 寒假（第二届）

全国高校大数据 Hadoop 师资培训班

邀请函

【首期回顾】

首届全国高校大数据 Hadoop 高校师资班合影

【媒体报道】

http://edu.ifeng.com/a/20140904/40787352_0.shtml 凤凰教育

http://learning.sohu.com/20140911/n404242413.shtml 搜狐教育

http://edu.online.qh.cn/jiaoyudongtai/2014/0912/18146.html 青海热线

http://www.onlinesd.cn/edu/px/2014-09-03/37409.html 山东在线

http://shizheng.xilu.com/20140912/1000150003063370.html 西陆网

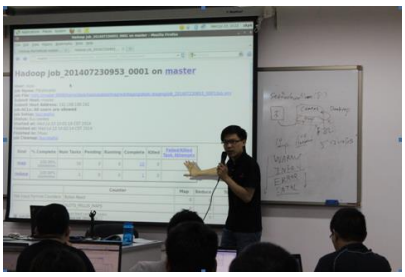
http://www.kaixian.tv/gd/2014/0911/9271485.html 汉丰网

2014 年 7 月 21 日-7 月 27 日，中科普开主办的 2014 年暑期“全国高校大数据 Hadoop 师资培训班”在北京理工大学成功举办。来自全国 50 多所高校包括北邮、重邮、中央财经大学、北京交通大学等的 70 位老师参加了此次培训，共同针对目前高校如何开设大数据专业、师资培养、优秀教材、实验和科研环境等问题进行深度交流。配备了自主研发的权威教材、详细的实验手册、最新的实验数据，旨在培养更多专业的大数据师资，将 Hadoop 的最新技术带入高校课堂。

【权威讲师倾情奉献】

中科普开创始人叶刚主持

向磊老师讲解精彩案例



马延辉老师讲解 MapReduce 编程

陈冠诚老师讲解 HBase

马延辉老师指导 Hadoop 部署实验

每位学员进行 Hadoop 安装实验

【交流晚宴】

普开为老师们精心准备了交流晚宴

【关于普开】www.zkpk.org

中科普开是国内首家致力于 IT 新技术领域的领航者，专注于云计算、大数据、物联网、移动互联网技术的培训，也是国内第一家开展 Hadoop、云计算的培训机构。

中科普开长期致力于企业培训及咨询服务，目前已成功举办 800 多次企业公开课培训、50 多次企业内训，现已成为国内最大的云计算、大数据人才培养基地，拥有全国最多、范围最广的就业保障体系。

中科普开拥有强大的实战专家团队，自主研发的普云云平台，是工信部权威机构的合作单位、是中关村互联网产业联盟单位、是 Hadoop 大会、云计算大会特邀合作伙伴、拥有全国 2000 多家云计算、大数据方向合作企业，每年有数十万学员受益，数千家企业好评。

【普开的优势】

国内最系统的 Hadoop 课程

独家研究，国内首创的系统课程，100%知识产权，30 次改版升级

国内规模最大培训人数最多

多年 Hadoop 培训经验，涵盖 20 多个细分行业，帮助企业搭建大数据平台

国内最人性化的增值服务

首家将 Hadoop 技术和企业需求加教学方案完美结合、集教育、服务于一体

【本期活动详情】

培养对象： 各高等院校计算机科学技术、网络工程、软件工程、信息工程、信息管理、物联网等相关专业教学带头人及骨干教师；各高校教务处、科研处、信息中心、实验中心领导。

培训时间：2015 年月 1 月 26 日—2 月 1 日，共 7 天（25 日全天报到）

培训地点：北京理工大学，名额有限，请提前预约！

食宿安排：食宿统一协助安排，费用自理。

培训费用：培训费：3900 元/人（含教材费、资料费、午餐费）、**考试及证书费用**（可选）：500 元/人。

颁发证书：参加相关培训并通过考试的学员，可以获得：

1．工业和信息化部颁发《**Hadoop 开发者**》培训证书。该证书可在工信部相关网站查询，可作为能力评价、考核和任职的重要依据。

2.中科普开培训中心颁发的《**Hadoop 高级开发工程师**》培训证书。

赠送礼包：X-Hadoop：参加培训获得免费赠送的最新 X-Hadoop 软件管理平台。

【增值服务】

- 协助高校大数据专业共建和课程置换
- 建立大数据联合实验室，协助高校搭建大数据实验平台
- 培养大数据专业讲师，为高校大数据课程储备人才
- 培养大数据应用型人才，面向就业提高学生就业率
- 共同撰写大数据相关书籍供高校教学使用，为高校大数据课程提供技术支持
- 免费提供大数据相关咨询服务
- 免费参加中科普开举办的各种会议和沙龙，第一时间了解业内最新技术动态

【培训大纲】（7 天大容量）

（本课程属于动手实验课程，提供 55GB 数据供实验使用；有完整实验数据和 2 套实验环境）

第一天

模块一、高校如何开设大数据教学课程

- 大数据相关专业人才需求缺口
- 大数据相关招聘岗位需求分析
- 大数据技术演进与变革
- 应对 IT 新技术变革，教师知识的储备与提升
- 在哪个层面进行教学

精彩案例

- ✧ 开展大数据分析挖掘教学的形式
- ✧ 高校开设大数据的教学可在多个层面上进行
- ✧ 有条件的高校可以开设云计算专业
- ✧ 不成熟的可在在计算机相关专业上增设数据挖掘和分析方向
- ✧ 不具备条件的高校可以开设大数据方面的课程，介绍大数据的知识，引导学生向大数据方面发展

模块二、大数据的发展现状

- 大数据起源和产生、大数据概念的发展与解析
- 大数据在国内外发展现状、大数据在互联网发展现状
- 大数据四个特点分析

模块三、大数据带来的机遇和挑战

- 大数据能带来什么 、引领社会进入“大数据时代”
- 大数据对国家、社会的作用 、大数据将推动经济发展
- 大数据将推动科技发展进程、开启商业智能新阶段
- 数据分析的发展——从数据到知识
- 大数据如何让商业更智能、大数据应用案例
- 带来数据处理新变革 、大数据的关键技术
- 大数据与云计算 、大数据技术的发展趋势

精彩案例

- ✧ 电信手机上网日志分析
- ✧ 移动 GPRS 上网日志查询系统
- ✧ 某省份联通网络不良信息检测系统
- ✧ 国土资源部门下属单位非结构离线网格分析平台
- ✧ 某银行海量数据统一分析平台
- ✧ 某视频公司用户属性精分系统
- ✧ 交通某下属单位实时计算平台
- ✧ 湖南某知名电台电视节目推荐系统

模块四、Hadoop 在云计算技术的作用和地位

- Hadoop 概述
- Hadoop 与分布式文件系统
- Hadoop 生态系统

✧ Apache 社区版本：Cloudera 版本、MapR 版本、Oracle、Dell、HP 版本

第二天

模块二、安装 HadoopV2

- Hadoop V2 核心组件简单介绍
- Hadoop V2 部署角色简单介绍
- Hadoop V2 试验集群的部署结构
- Hadoop V2 安装依赖关系
- Hadoop V2 生产环境的部署结构
- Hadoop V2 集群部署
- Hadoop V2 高可用配置方法
- Hadoop V2 集群简单测试方法
- Hadoop V2 集群异常 Debug 方法

精彩案例

- ✧ Hadoop V2 安装部署实验
- ✧ Red hat Linux 基础环境搭建
- ✧ Hadoop V2 单机系统版本安装配置
- ✧ Hadoop V2 集群系统版本安装和启动配置
- ✧ 使用 Hadoop MapReduceV2 样例代码快速测试系统
- ✧ Hadoop V2 配置文件 core-site.xml，hdfs-site.xml，mapred-site.xml 和 yarn-site.xml 详解
- ✧ Hadoop V2 环境变量 Hadoop-env.sh 和 yarn-env.sh 详解

实战内容（现场实验环节）

- ✧ 单机环境&模拟线上环境，Hadoop V2 集群安装
- ✧ Hadoop V2 集群初始化、启动和停止操作
- ✧ 结合修改 Hadoop V2 配置文件，讲解 Hadoop 运维和优化

真实集群下验证 Hadoop 启动是否成功及实际启动部署问题现场解决

第三天

模块三、Hadoop YARN 架构设计和核心概念

- YARN 出现的背景
- YARN 的设计思想和基本概念
- YARN 的基础架构
- YARN 的工作流程
- YARN 基础类库详解
- YARN 核心模块 ApplicationMaster 详解
- YARN 核心模块 ResourceManager 剖析
- YARN 核心模块 NodeManager 详解
- Hadoop V2 资源调度器详细介绍

精彩案例

- ✧ 基于 YARN 的应用程序设计和简单实现案例
- ✧ YARN 使用第三方类库，包含通信、服务、时间、状态等
- ✧ 从资源管理角度理解 YARN 框架
- ✧ ApplicationMaster 核心源码
- ✧ ResourceManager 核心源码
- ✧ NodeManager 核心源码

模块四、MapReduce V2 高级编程

- MapReduce V2 基本架构、原理和相关角色介绍
- MapReduce V1 和 MapReduce V2
- MapReduce V2 中 MRAppMaster 的工作流程
- MapReduce V2 作业生命周期理解
- MapReduce V2 资源调度理解
- MapReduce V2 作业恢复和推测执行机制介绍
- 剖析 MapReduce V2 样例程序代码流程
- 基本 MapReduce V2 API 概念
- 驱动代码 Mapper、Reducer
- Hadoop 流在 MapReduce V2 中的使用
- 使用 Eclipse 进行快速开发
- MapReduce V2 的编程优化
- 满足解决实际数据分析问题的高级 MapReduce V2 多语言编程实践

精彩案例

- ✧ Hadoop Streaming 和 Java MapReduce V2 编程的差异
- ✧ 利用 Combiner 来减少中间数据
- ✧ 编写 Partitioner 来优化负载平衡
- ✧ 直接访问 Hadoop 分布式文件系统 HDFS
- ✧ MapReduce V2 的 join 操作
- ✧ 辅助排序在 Reducer 方的合并
- ✧ 定制开发 Writables 和 WritableComparable 高级类
- ✧ 使用 SequenceFiles 和 Avro 文件保存二进制数据
- ✧ 定制开发 InputFormats 和 OutputFormat
- ✧ 基于 MapReduce V2 的海量日志分析

实战内容（现场实验环节）

- ✧ 熟悉 Java 编程 IDE Eclipse
- ✧ MapReduce V2 wordcount 实例测试
- ✧ MapReduce V2 wordcount 源代码查看讲解
- ✧ MapReduce V2 实现单词统计扩展代码逻辑并测试结果
- ✧ 基于 MapReduce V2 编写海量日志的简单分析需求

第四天

模块五、基于 Hadoop V2 的 Hive/Pig 开发技巧

- Hive 和 Pig 架构和理论基础
- Hive 的作用和原理说明
- Hadoop 仓库和传统数据仓库的协作关系
- Hadoop/Hive 仓库数据数据流
- 基于 Hadoop V1 和 V2 使用 Hive 和 Pig 等工具的异同
- Hive 部署和安装
- Hive Cli 的基本用法
- HQL 基本语法
- 使用 Oozie 的动机
- Oozie 工作流定义格式

精彩案例

- ✧ 使用 JDBC 连接 Hive 进行查询和分析
- ✧ 使用正则表达式加载数据
- ✧ HQL 高级语法
- ✧ 编写 UDF 函数
- ✧ 编写 UDAF 自定义函数
- ✧ Pig Latin 编程实例
- ✧ Pig Latin 编程注意事项
- ✧ 使用 Sqoop 进行数据分析
- ✧ 使用 oozie 配置工作流
- ✧ HUE 介绍

实战内容（现场实验环节）

- ✧ 动手安装 Hive、Pig，并验证是否安装成功
- ✧ 现场解决启动、部署问题，并总结解决问题的方法
- ✧ 使用 Hive 创建自己的数据仓库
- ✧ 使用 Hive 在数据仓库上进行增删改查操作
- ✧ 使用 Pig Latin 进行编程
- ✧ 使用 Pig 上进行增删改查操作

第五天

模块六、HBase（Hadoop V2）海量实时处理实战技巧

- HBase 简介和架构
- HBase 核心知识点
- HBase 高级应用
- HBase 应用场景
- HBase（For Hadoop V2）安装、部署、启动
- HBase 常用接口和 SQL 引擎层实战
- 新进展、新技术
- 基于 Hadoop V1 和 V2 使用 HBase 的异同
- HBase 调优

精彩案例

- ✧ NoSQL 数据库与关系数据库的对比
- ✧ 基于列簇的 NoSQL
- ✧ HBase 存储逻辑结构介绍
- ✧ HBase 与其他 NoSQL 的比较
- ✧ 列式存储核心：LSM
- ✧ 日志系统：WAL
- ✧ 集群下安装部署 HBase
- ✧ 启动 HBase 和启动顺序
- ✧ 如何通过命令和 WebUI 验证启动是否成功
- ✧ 通过 HBase Shell 增删改查数据
- ✧ 通过 HBase Shell 进行管理表和 Region 操作
- ✧ 通过 HBase Shell 执行 Java 类中方法

实战内容（现场实验环节）

- ✧ 动手安装 HBase，并验证是否安装成功
- ✧ 现场解决启动、部署问题，总结解决问题的方法
- ✧ 使用 HBase 设计特定需求下的表结构
- ✧ 使用 Java API 对 HBase 进行增删改查操作

精彩案例

- ✧ 解决 Master 单点问题
- ✧ Native Java 增删改查实战
- ✧ Thrift&Thrift2 增删改查实战
- ✧ MapReduce 访问 HBase 编程实战
- ✧ 数据批量导入（Bulk Load）实战
- ✧ 读、写优化中需要参数的调整
- ✧ hbase-site.xml 参数调优
- ✧ JVM 优化中需要调整的参数
- ✧ split & compact 优化相关参数
- ✧ 表设计优化相关参数
- ✧ HBase 客户端优化相关参数
- ✧ 监控工具使用方法及注意事项

第六天

模块七、实时流框架 Storm on YARN（Hadoop V2）实战技巧

- 什么是实时流计算
- Storm 是什么
- Storm 核心组件
- Storm 特性
- Storm 应用于什么场景
- Storm 核心概念和数据流模型
- 在 YARN 上安装 Storm
- 运行基于 Storm 的编程实例

精彩案例

- ✧ 拓扑结构 Topology 介绍
- ✧ 数据源 Spout、Bolt
- ✧ 流概念 Stream 和 Stream Grouping
- ✧ 运行任务 Task、Work
- ✧ 消息和事务
- ✧ 数据流模型重点讲解
- ✧ 安装 Storm 过程中注意的问题
- ✧ Storm 监控工具介绍

实战内容（现场实验环节）

- ✧ 动手安装 Spark on YARN，并验证是否安装成功
- ✧ 现场解决启动、部署问题，并总结解决问题的方法
- ✧ 导入数据，并使用 Java 语言编程进行查询操作

第七天

模块八、内存计算框架 Spark on YARN(Hadoop V2) 实战技巧

- 深入 Spark 核心架构
- 在 YARN 上安装 Spark
- Spark 集群配置介绍
- Spark 多语言编程
- Spark 案例介绍

精彩案例

- ✧ RDD
- ✧ 缓存策略介绍
- ✧ transformation
- ✧ action
- ✧ lineage
- ✧ 容错处理

实战内容（现场实验环节）

- ✧ 动手安装 Storm on YARN，并验证是否安装成功
- ✧ 现场解决启动、部署问题，总结解决问题的方法
- ✧ 根据特定需求设计并用 Java 语言实现 Topology，并部署验证编程逻辑

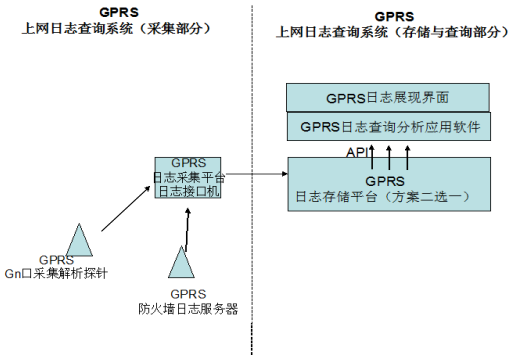
精彩案例

- ✧ 宽依赖与窄依赖
- ✧ Spark on Yarn 原理
- ✧ Spark on Yarn 实践
- ✧ Python 简介
- ✧ PySpark API
- ✧ 使用 Python 编写 Spark 程序 Spark with Java

项目实战一

手机上网日志分析系统将海量日志数据进行分布式存储，并通过分布式算法和网络爬虫技术形成标签化的用户模型，最终实现人与内容、人与行为、人与商品的智能配对系统通过对分词、机器学习等技术将 URL 进行分类，通过网络爬虫爬取用户访问的 URL 内容，并对内容进行分词，与 URL 类别分词进行权重比对，最终决定用户访问 URL 的类别，通过大量数据的统计决定用过的行为偏向、内容喜好，从而实现“用户画像”。

项目实战二



此项目采用业界领先的 Hadoop 分布式文件系统技术框架，采用 X86 集群，利用本地硬盘存储，极大节省了总体投资；且由于负荷分担，数据入库、即席查询的效率是对于传统单节点（或少数几个节点）小机+关系型数据库+磁盘阵列方式有数倍乃至数十倍的提高（取决于集群规模和数据量比值）；利用多副本拷贝保证数据安全；低成本的 X86 硬件可节省总体投资 60—70%。（无需再购置昂贵的磁盘阵列和数据库 License）。此技术架构是如 Goole、Facebook、Amazon、Baidu、Alibaba 等大型互联网公司当前普遍采用的技术，成本低，效率高，安全性好。

【讲师介绍】

向　磊

国内某影音公司数据部门高级软件工程师，拥有10年以上开发和运维工作经验。 负责公司整体Hadoop数据集群平台的部署、运维、性能优化工作，以及基于Python和Hive的Map/Reduce程序研发。对Hadoop及其周边生态系统如Hive ,HBase ,Mahout ,Zookeeper等有较深入研究，对集群系统的监控，高负载，高可用有丰富的相关经验。

个人成就

- ✧ 世界首款 GPL 协议开源 Hive web 管理查询工具 phpHiveAdmin 唯一作者
- ✧ 精通 Linux，FreeBSD，OpenBSD，AIX 等*NIX 类操作系统。
- ✧ 精通 PHP，Python，Mysql 等 web 相关技术，熟悉 shell，R 语言。
- ✧ 世界首款 GPL 协议开源 Hadoop web 部署管理监控系统 EasyHadoop 作者，
- ✧ 国内最大的 Hadoop 在线社区 EasyHadoop 开源社区创办者。
- ✧ 目前开发的软件用户众多，不乏各大知名互联网公司，包括迅雷，优酷，360 安全卫士，窝窝团，趣游，风行，即刻搜索，江西电信，网秦安全，新浪微博等，国外公司下载使用同样广泛。
- ✧ 源码在 Github.com 上被 follow 和 fork 数量是 Apache 基金会开源项目 Ambari 的数倍。

马延辉

原Answers.com搜索部门架构师，目前担任中科普开CTO，曾先后在淘宝、Answers.com从事垂直搜索、大数据分析挖掘和平台建设等方向的研发，对Hadoop生态系统，特别是Hive、HBase、Mahout等开源框架的业务应用、可靠性、基础架构和高级应用等方面有着丰富经验。项目大部分属于互联网领域，项目对性能，特别是实时性、稳定性、可用性的要求非常高，参与咨询和实施的项目涉及地质、交通、气象等诸多领域。此外开源了若干项目深受业界认可，如Ella、Hbase-secondary-index。

项目经验

- ✧ 用户精分系统
- ✧ 海量数据实时计算系统
- ✧ HBase 监控项目——Ella
- ✧ HBase 二级索引项目
- ✧ 数据平台整体迁移
- ✧ 手机端综合推荐系统
- ✧ 某视频指数项目
- ✧ 数据魔方产品
- ✧ 化妆品个性化推荐项目
- ✧ 问答类网站主题搜索系统
- ✧ 中文全文搜索引擎系统
- ✧ 地调网格大数据平台

陈冠诚

IBM中国研究院研究员，主要从事大数据系统性能分析与优化方面的技术研发工作。参与项目主要包括IBM BigInsights软件在Power服务器上软硬件协同优化，MapReduce性能分析与调优工具，高性能FPGA加速器在大数据平台上的应用等。在国际顶级会议和期刊上发表过多篇论文，并拥有六项大数据领域的技术专利。他同时还是《程序员》杂志的技术作者，分享过多篇分布式计算，大数据处理技术等方面的技术文章。

项目经验

- ✧ 某大型银行的 HBase 海量数据加载性能分析与调优
- ✧ 海量数据排序的性能分析与优化
- ✧ 基于 FPGA 加速的大数据平台设计与实现
- ✧ MapReduce 性能分析与调优工具的设计与实现

【实验环境】

提供 2 套实验环境：Hadoop 安装实验和安装部署好 Hadoop 实验平台

中科普开 www.zkpk.org

普开同创 云领未来| www.zkpk.org

软件包介绍

名称： Hadoop In Action Experiment

所含子目录：

- ide-code：包含所有开发用 IDE 和部分代码，如 eclipse、JAR 包、shell 脚本等
- web-ui：数据报表结果展示用 J2EE 代码包
- sogou-data：搜狗搜索日志数据
- software：实验过程中所需的软件，如 Hadoop 生态系统组件、JDK、VMWare 等
- doc：课程教材，用于学员参考使用
- vmare-centos-bare：基于 VMWare 构建的两个 CentOS 虚拟机，用于搭建 Hadoop 集群。属于比较干净的环境，没有安装 Hadoop 系统所需要的一些软件
- vmware-centos-all-in-one :基于 VMWare 构建的两个 CentOS 虚拟机，用于搭建 Hadoop 集群。Hadoop、HBase、Hive、Mahout 以及实验数据都已经在两个虚拟中。属于 All in One 的虚拟机。

如何使用？

需要将该软件包全部复制到磁盘空间足够的机器上（可以是个人笔记本、台式机、服务器），安装实验手册中的内容，逐步复制不同的目录进行操作即可。

本次师资培训班设置了大量动手实验，并提供 55GB 数据供实验使用，有完整实验数据和实验环境，为使实验具有连贯性，我们设计了 Hadoop 安装实验和安装部署好 Hadoop 的 2 套实验场景。

整个教学过程中采用理论结合实践的方式，以帮助高校老师快速掌握大数据知识体系及开发方法。丰富、使用的课程体系，可帮助高校老师快速的开设大数据课程。

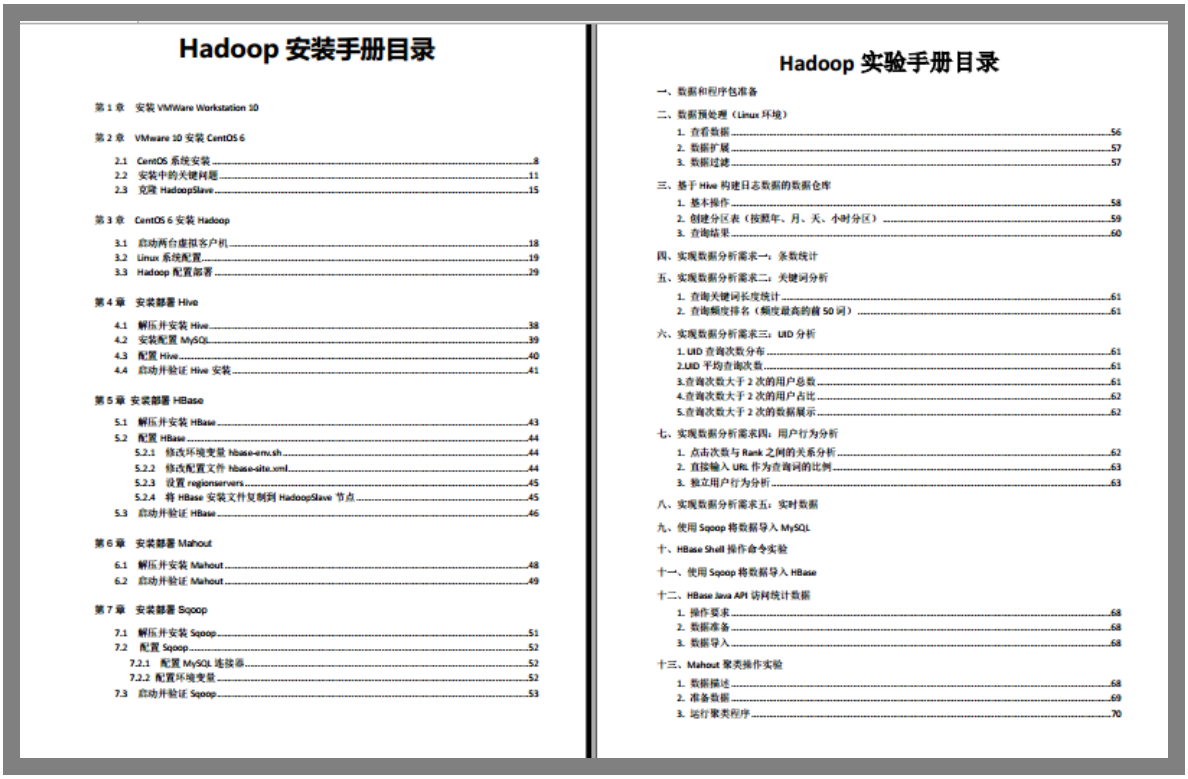
【配套资料】

1、《大数据运维与开发实战》教材一本

2、配套实验手册一本

3、本次大数据培训全套教学课件

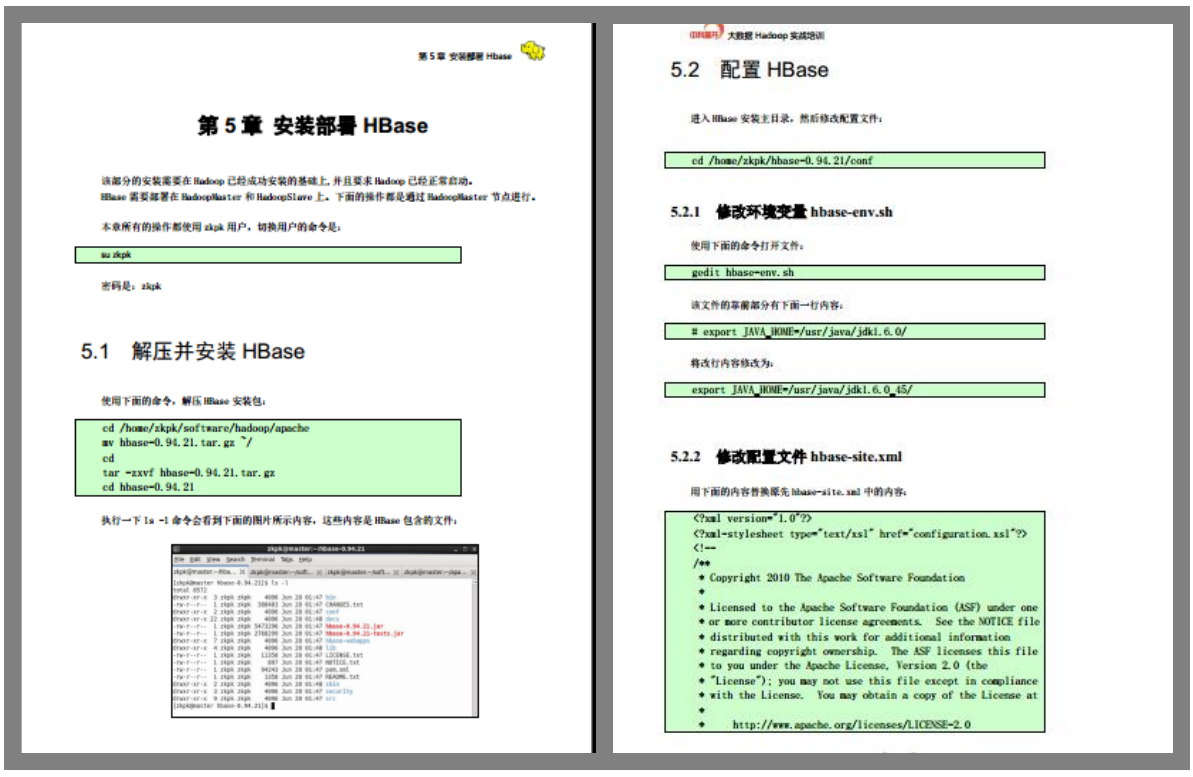
4、赠送最新 X-Hadoop 软件管理平台



Hadoop 安装手册目录

Hadoop 实验手册目录

实验手册目录



第 5 章 安装部署 HBase

第 5 章 安装部署 HBase

该部分的安装需要在 Hadoop 已经成功安装的基础上，并且要求 Hadoop 已经正常启动。
HBase 需要部署在 HadoopMaster 和 HadoopSlave 上。下面的操作都是通过 HadoopMaster 节点进行。

本章所有的操作都使用 zkpk 用户，切换用户的命令是：

```
su zkpk
```

密码是：zkpk

5.1　解压并安装 HBase

使用下面的命令，解压 HBase 安装包：

```
cd /home/zkpk/software/hadoop/apache
mv hbase-0.94.21.tar.gz ~/
cd
tar -zxvf hbase-0.94.21.tar.gz
cd hbase-0.94.21
```

执行一下 ls -l 命令会看到下面的图片所示内容，这些内容是 HBase 包含的文件：

大数据 Hadoop 实战培训

5.2　配置 HBase

进入 HBase 安装主目录，然后修改配置文件：

```
cd /home/zkpk/hbase-0.94.21/conf
```

5.2.1　修改环境变量 hbase-env.sh

使用下面的命令打开文件：

```
gedit hbase-env.sh
```

该文件的靠前部分有下面一行内容：

```
# export JAVA_HOME=/usr/java/jdk1.6.0/
```

将改行内容修改为：

```
export JAVA_HOME=/usr/java/jdk1.6.0_45/
```

5.2.2　修改配置文件 hbase-site.xml

用下面的内容替换原先 hbase-site.xml 中的内容：

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
/**
 * Copyright 2010 The Apache Software Foundation
 *
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements.  See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership.  The ASF licenses this file
 * to you under the Apache License, Version 2.0 (the
 * "License"); you may not use this file except in compliance
 * with the License.  You may obtain a copy of the License at
 *
 *     http://www.apache.org/licenses/LICENSE-2.0
```

实验手册详尽的操作步骤