中国信息协会信息服务网络委员会

中信网培字[2013]019 号

关于举办“大数据的处理技巧及案例分析”高级工程师研修班的通知

各有关单位:

大数据分析作为数据分析的前沿技术，广泛应用于物联网、云计算、移动互联网等新兴产业。为加强大数据分析，创新发展顶层设计和科学布局，推动相关人员更好理解和掌握大数据分析的关键技术原理和未来发展方向，使各企事业单位利益最大化，中国信息协会信息服务网络委员会决定举办“大数据的处理技巧及案例分析”高级研修班，望各单位收到通知后组织相关人员参加。现将有关事宜通知如下:

一、课程内容:

模块	课程主题	主要内容	案例和演示
模块一	Hadoop 的来源和动机	◆ 传统大规模系统存在的问题 ◆ Hadoop 概述 ◆ Hadoop 分布式文件系统 ◆ MapReduce 工作原理 ◆ Hadoop 集群剖析 ◆ Hadoop 生态系统对一种新的解决方案的需求 ◆ Hadoop 的行业应用案例分析 ◆ Hadoop 在云计算和大数据的位置和关系	◆ Hadoop 在淘宝、支付宝的作用 ◆ 电商眼中的 Hadoop 和推荐系统。 ◆ 移动大云项目(big cloud)中的 Hadoop ◆ 联通使用 Hadoop/Hbase 解决 3G 详单查询问题。
模块二	Hadoop 生态系统介绍和演示	◆ Hadoop HDFS 和 MapReduce ◆ Hadoop 数据库之 HBase ◆ Hadoop 数据仓库之 Hive ◆ Hadoop 数据处理脚本 Pig ◆ Hadoop 数据接口 Sqoop 和 Flume，Scribe DataX ◆ Hadoop 工作流引擎 Oozie	◆ Yahoo 如何使用 Hadoop 构建大规模搜索的应用 ◆ FaceBook 基于 Hadoop 构建数据仓库
模块三	Hadoop 组	◆ Hadoop HDFS 基本结构	◆ Hadoop Mapper 类核心代码

	件详解	◆ Hadoop HDFS 副本存放策略 ◆ Hadoop NameNode 详解 ◆ Hadoop SecondaryNameNode 详解 ◆ Hadoop DataNode 详解 ◆ Hadoop JobTracker 详解 ◆ Hadoop TaskTracker 详解	◆ Hadoop Reduce 类核心代码 ◆ Hadoop 核心代码
模块四	Hadoop 安装和部署	◆ Hadoop 系统模块组件概述 ◆ Hadoop 试验集群的部署结构 ◆ Hadoop 安装依赖关系 ◆ Hadoop 生产环境的部署结构 ◆ Hadoop 集群部署 ◆ Hadoop 高可用配置方法 ◆ Hadoop 集群简单测试方法 ◆ Hadoop 集群异常 Debug 方法	◆ Hadoop 安装部署实验 ◆ Red hat Linux 基础环境搭建 ◆ Hadoop 单机系统版本安装配置 ◆ Hadoop 集群系统版本安装和启动配置 ◆ 使用 Hadoop MapReduce Streaming 快速测试系统 ◆ Hadoopcore-site，hdfs-site，mapred-site 配置详解
模块五	Hadoop 和数据库技术优劣势对比	◆ Hadoop/Hive 对比 Oracle 在构建数据仓库上的优劣势 ◆ Hadoop 如何和传统 IT 系统配合完成原来不可能的任务	◆ Apache 社区版本:Cloudera 版本、MapR 版本、Intel 版本、Oracle、Dell、HP 版本
模块六	编写 MapReduce 高级程序	◆ 使用 Hadoop MapReduce Streaming 编程 ◆ MapReduce 流程 ◆ 剖析一个 MapReduce 程序 ◆ 基本 MapReduceAPI 概念 ◆ 驱动代码 Mapper、Reducer ◆ Hadoop 流 ◆ API 使用 Eclipse 进行快速开发 ◆ 新 MapReduce API ◆ MapReduce 的优化 ◆ MapReduce 的任务调度 ◆ MapReduce 编程实战 ◆ 如何利用其他 Hadoop 相关技术，包括 Apache Hive，Apache Pig，Sqoop 和 Oozie 等 ◆ 满足解决实际数据分析问题的高级 Hadoop API	◆ Hadoop Streaming 和 Java MapReduce Api 差异。 ◆ MapReduce 实现数据库功能 ◆ 利用 Combiners 来减少中间数据 ◆ 数据压缩解压算法 ◆ 基于统计模型的压缩算法 ：Huffman 编码、算数编码、PPM 算法 ◆ 基于字典模型的编码：LZ77 算法、LZ78 算法、LZW 算法 ◆ 面向实时数据的专用压缩算法：矩形波串法、后向斜率法、旋转门压缩算法。 ◆ 其他压缩算法：RLE 文本压缩算法、BWT 算法 ◆ 编写 Partitioner 来优化负载平衡 ◆ 直接访问 Hadoop 分布式文件系统（HDFS） ◆ Hadoop 的 join 操作 ◆ 辅助排序在 Reducer 方的合并 ◆ 定制 Writables 和

			WritableComparables ◆ 使用 SequenceFiles 和 Avro 文件保存二进制数据 ◆ 创建 InputFormats OutputFormats ◆ Hadoop 的二次排序 ◆ Hadoop 的海量日志分析 ◆ 在 Map 方的合并
模块七	集成 Hadoop 到现有工作流及 Hadoop API 深入探讨	◆ 存储系统 ◆ 利用 Sqoop 从关系型数据库系统中导入数据到 Hadoop ◆ 利用 Flume 导入实时数据到 Hadoop ◆ ToolRunner 介绍、使用 MRUnit 进行测试 ◆ 使用 Configure 和 Close 方法来进行 Map/Reduce 设置和关闭	◆ 使用 FuseDFS 和 Hadoop 访问 HDFS ◆ 使用分布式缓存（Distributed Cache） ◆ 直接访问 Hadoop 分布式文件系统（HDFS） ◆ 利用 Combiners 来减少中间数据 ◆ 编写 Partitioner 来优化负载平衡
模块八	使用 Hive 和 Pig 开发及技巧	◆ Hive 和 Pig 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 ◆ 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	◆ 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 ◆ 使用 Sqoop 进行数据分析 ◆ 使用 oozie 配置工作流 ◆ phpHiveAdmin 安装和使用
模块九	实用开发技巧	◆ 排序和搜索索引 ◆ 用 Mahout 进行机器学习 ◆ Term Frequency - Inverse Document Frequency ◆ 图论简介	◆ Word Co-Occurrenc ◆ 用 Hadoop 表示图 ◆ 一个图算法的实现：单源最短路径

二、时间及地点：

2013 年 12 月 13 日—12 月 16 日 （13 日全天报到） **深圳**

2014 年 01 月 14 日—01 月 17 日 （14 日全天报到） **北京**

三、培训对象：

各地政府云计算、物联网产业相关负责人，各企业 CIO、信息中心负责人、技术总监，云计算产业投资团队，云计算应用开发商，云计算硬件设备供应商，云服务提供商，高校、科研院所云计算项目负责人等。

四、培训特色：

注重应用：分析国内实际情况，结合国际、国内成功经验。Hadoop 采用实战的项目，让学员在短时间内掌握 Hadoop 基本运维思路和方法；对 Hadoop 集群进行管理和优化。并进行高效的大数据清洗和分析。

五、师资力量：

王宝会：主要从事对物联网、云计算相关技术、应用架构及实施有深入的研究。先后参与国家科技支撑项目《增强型搜索引擎研究及示范应用》、《中国移动 POC 项目设计与开发》、《华为公司网络高级应用协议测试》《汽车制造工艺开发平台》、《工商系统食品在线监管》项目。在国内外期刊多次发表学术论文及学术著作。

白　硕：系统架构设计师；高级软件工程师；项目总监。做 hdfs 相关的产品。基于 hadoop2.0 源代码做了修改，修改的功能主要包括了文件的读写,安全模式，添加 RPC 调用，FileStatus，装载 image，FsEditLog，

六、培训费用及颁发证书

培训费:3900 元/人（含培训、教材、专家、场地、证书、费等费用），

食宿统一安排，费用自理。往返交通费自理。

通过考试的学员可以获得：《大数据分析高级工程师》证书。并且可通过国家信息技术人才服务网（www.ciso.net.cn）查询。

该证书可作为专业技术人员职业能力考核的证明，以及专业技术人员岗位聘用、任职、定级和晋升职务的重要依据。

注：报到时携带二寸免冠蓝底彩照 2 张，身份证和最高学历证书复印件各 1 张。

七、联系方式

联系人：张 磊

电话：010-52830300　　010-52833988

传真：010-80300265（自动）

E-mail：chinazbpx@126.com

附　件：报名回执表

中国信息协会信息服务网络委员会

附件:

“大数据的处理技巧及案例分析”高级研修班

报名回执表

单位名称						行业类别	
通讯地址	-------省/市-------县/区----------						
审 批 人		职务		电话		手机	
联 系 人		部门		职务		手机	
电 话				传真		E-mail	

代表姓名	部 门	性别	职务	电话	手机	电子信箱

参会地点			
住宿安排	单住□ 标间拼住□ 订房数量__间 ； 自行安排□; 其他说明:_____		
费用总额	万 仟 佰 拾 元整	小写	￥:
付款方式	□通过银行汇款 □现金		
相关事宜	1、以上内容培训均可提供内训， 2、内训部电话：18611082369	单位印章 年 月 日	

备注：1.此表格可复印使用，传真件有效，请用正楷字填写。

2.报名传真:（010）80300265（自动）

3.报名邮箱：chinazbpx@126.com 联系人：张 磊