大数据应用
培训方案

一、培训的目的及意义

随着我国智能电网的发展，电力系统发、输、变、配、用电各个环节的信息化进程不断推进。在用电侧，利用电力大数据分析可以了解产业结构、经济走势、房屋空置率、区域消费能力等情况，从而可以更好地为经济服务。

伴随着智能电网的全面建设，以物联网和云计算为代表的新一代信息通信技术在电力行业中的广泛应用，电力数据资源开始急剧增长并形成了一定的规模。电力与社会经济的发展密切相关，电力需求变化是经济运行的“晴雨表”和“风向标”，能够真实、客观地反映国民经济的发展状况与态势。因此，发展电力大数据是电力行业革新的必然过程。国家电网公司正在制定以云计算和大数据为技术支撑的信息系统方案，以提高国网公司信息系统的安全性、稳定性、可用性，为公司决策、生产、运营、管理提供更好的支撑平台，助

力公司发展方式的转变。

二、培训对象

公司专兼职信息通信运维及管理人员。

三、培训时间

培训共五天

四、培训内容

1．大数据（Hadoop）介绍，分布式文件系统应用
2．MapReduce 应用及调优
3．hadoop 集群及管理
4．hadoop 子项 zookeeper、hbase、pig、hive、sqoop、rdbms 应用
5．大数据实战

五、课程安排

日程	授课主题	课程安排
第一～二天	Hadoop 入门，了解什么是 hadoop	Hadoop 产生背景 Hadoop 在大数据、云计算中的位置和关系 国内外 Hadoop 应用案例介绍 国内 Hadoop 的就业情况分析及课程大纲介绍 分布式系统概述 Hadoop 生态圈以及各组成部分的简介 Hadoop 核心 MapReduce 例子说明
	分布式文件系统 HDFS，是数据库管理员的基础课程	分布式文件系统 DFS 简介 HDFS 的系统组成介绍 HDFS 的组成部分详解 副本存放策略及路由规则 NameNode Federation 命令行接口 Java 接口

		客户端与 HDFS 的数据流讲解 HDFS 的可用性（HA）
	初级 MapReduce，成为 Hadoop 开发人员的基础课程	如何理解 map、reduce 计算模型 剖析伪分布式下 MapReduce 作业的执行过程 Yarn 模型 序列化 MapReduce 的类型与格式 MapReduce 开发环境搭建 MapReduce 应用开发 更多示例讲解，熟悉 MapReduce 算法原理
第三~四天	高级 MapReduce，高级 Hadoop 开发人员的关键课程	使用压缩分隔减少输入规模 利用 Combiner 减少中间数据 编写 Partitioner 优化负载均衡 如何自定义排序规则 如何自定义分组规则 MapReduce 优化 编程实战
	Hadoop 集群与管理，是数据库管理员的高级课程	Hadoop 集群的搭建 Hadoop 集群的监控 Hadoop 集群的管理 集群下运行 MapReduce 程序
	ZooKeeper 基础知识，构建分布式系统的基础框架	ZooKeeper 体现结构 ZooKeeper 集群的安装 操作 ZooKeeper
	HBase 基础知识，面向列的实时分布式数据库	HBase 定义、HBase 与 RDBMS 的对比 数据模型、系统架构 HBase 上的 MapReduce、表的设计
	HBase 集群及其管理 HBase 客户端	集群的搭建过程讲解 集群的监控 集群的管理
		HBase Shell 以及演示 Java 客户端以及代码演示
第五天	Pig 基础知识，进行 hadoop 计算的另一种框架	Pig 概述 安装 Pig 使用 Pig 完成手机流量统计业务
	Hive，使用 sql 进行计算的 hadoop 框架	数据仓库基础知识 Hive 定义、Hive 体系结构简介、Hive 集群 客户端简介 HiveQL 定义、HiveQL 与 SQL 的比较 数据类型 表与表分区概念、表的操作与 CLI 客户端演示

		数据导入与 CLI 客户端演示、查询数据与 CLI 客户端演示 数据的连接与 CLI 客户端演示、用户自定义函数（UDF）的开发与演示
	Sqoop，hadoop 与 rdbms 进行数据转换的框架	配置 Sqoop 使用 Sqoop 把数据从 mysql 导入到 HDFS 中 使用 Sqoop 把数据从 HDFS 导出到 mysql 中
	结训考试	结训考试