Hadoop 大数据解决方案平台技术培训

【课程目标】

Hadoop 作为开源的云计算平台，为大数据处理提供了一整套解决方案，应用非常广泛。Hadoop 作为一个平台框架，包括了如何存储海量数据，如何处理海量数据，以及相应的数据库、数据仓库、数据流处理、数据分析和挖掘算法库，等等。本课程主要介绍 Hadoop 的思想、原理，以及重要技术等相关知识。

通过本课程的学习，达到如下目的：

1、全面了解大数据处理技术的相关知识。

2、学习 Hadoop 的核心技术以及应用。

3、深入掌握 Hadoop 的相关工具在大数据中的使用。

4、掌握 Hadoop 的常用模块的工作原理及开发应用技术。

5、掌握传统数据中心向大数据中心转换的关键技术。

6、掌握海量数据处理的性能优化及维护技巧。

【授课时间】

2 天时间

【授课对象】

网络部、大数据系统开发部、大数据中心、网络运维部等相关技术人员。

【授课方式】

原理精讲+案例演练+开发实践+系统优化

【课程大纲】

第一部分： Hadoop 的基本框架

1、大数据时代面临的问题

2、当前解决大数据的技术方案

3、Hadoop 架构和云计算

4、Hadoop 简史及安装部署

5、Hadoop 设计理念和生态系统

第二部分： HDFS 分布式文件系统--海量数据存储的摇篮

1、HDFS 的设计目标

2、HDFS 的基本架构

- NameNode 名称节点
- SecondaryNameNode 第二名称节点
- DataNode 数据节点

3、HDFS 的存储模型

- 数据块存储
- 元数据存储（空间镜像与编辑日志）
- 多副本存储

4、多副本放置策略

5、多数据节点管理机制与交互过程

6、文件系统操作与管理

- 读文件过程
- 写文件过程（数据流管道）

7、数据完整性机制

- 数据校验和

- 数据完整性扫描线程
- 元数据备份与合并

8、数据可靠性设计

- 安全模式（数据块与节点映射关系管理）
- 心跳检测机制（节点失效管理）
- 租约机制（多线程并发控制）

9、其它

- HDFS 的安全机制
- 负载均衡
- 文件压缩

10、 操作接口与编程接口

- HDFS Shell
- HDFS Commands
- WebHDFS REST API
- HDFS Java API

演练：HDFS 文件操作命令

演练：HDFS 编程示例

第三部分： MapReduce 分布式计算系统--海量数据处理的利器

1、 MapReduce 的三层设计理念

- 分布治之的设计思想（Map 与 Reduce）
- 数据处理引擎（编程模型）
- 运行时环境（任务调度与执行）

2、 MapReduce 的基本架构

- JobTracker 作业跟踪器
- TaskTracker 任务跟踪器
- MapReduce 与 HDFS 的部署关系

3、 MapReduce 编程模型概述

- 编程接口介绍
- Hadoop 工作流实现原理

4、 MapReduce 作业调度机制

- MapReduce 作业生命周期
- 作业调度策略
- 静态资源管理方案

5、 数据并行处理机制（五步骤）

- Input 阶段实现
- Map 阶段实现
- Shuffle 阶段实现
- Reduce 阶段实现
- Output 阶段

6、MapReduce 容错机制

- 任务失败与重新尝试
- 节点失效与重调度
- 单点故障

7、MapReduce 性能优化

- 优化方向与思路
- 磁盘 IO 性能优化
- 分片优化
- 线程数量优化
- 内存优化
- 压缩优化

8、MapReduce 操作接口

- Job Shell

- Web UI

案例演练：MapReduce 编程示例

9、YARN：下一代通用资源管理系统

- MRv1 的局限性
- YARN 基本框架
- NN HA：解决单点故障
- HDFS Federation：解决扩展性问题

第四部分： HBase 非关系型数据库--海量数据的黎明

1、HBase 的使用场景

2、HBase 的基本架构

- Zookeeper 分布式协调服务器
- Master 主控服务器
- Region Server 区域服务器

3、HBase 的数据模型

- HBase 的表结构
- 行键、列键、时间戳

4、HBase 的存储模型

- 基本单位 Region
- 存储格式 HFile

5、数据分裂机制 Split

6、数据合并机制 Compaction

- minor compaction
- major compaction

7、HLog 写前日志

8、数据库读写操作

- 数据库写入
- 数据库读取
- 三次寻址

9、HBase 操作接口

- Native Java API
- HBase Shell
- 批量加载工具
- HiveQL 操作

10、 HBase 性能优化

- 写速度优化

- 读速度优化

11、 HBase 集群监控与管理

案例演练：HBase 命令操作实例

第五部分： Hive 分布式数据仓库--高级的编程语言

1、Hive 是什么

2、Hive 与关系数据库的区别

3、Hive 系统架构

- 用户接口层
- 元数据存储层
- 驱动层

4、Hive 常用服务

5、Hive 元数据的三种部署模式

6、Hive 的命名空间

7、Hive 数据类型与存储格式

- 数据类型
- TextFile/SequenceFile/RCFile

8、Hive 的数据模型

- 管理表
- 外部表
- 分区表
- 桶表

9、HQL 语言命令实例

- DDL 数据定义语言
- DML 数据操作语言
- QUERY 数据查询语言

10、 Hive 自定义函数

- 基本函数（UDF）
- 聚合函数（UDAF）
- 表生成函数（UDTF）

11、 Hive 性能优化

- 动态分区
- 压缩
- 索引

- JVM 重用

案例演练：Hive 命令操作实例

第六部分： Sqoop 数据交互工具--与传统数据库的桥梁

1、 Sqoop 是什么

2、 Sqoop 的架构和功能

- Sqoop1 架构
- Sqoop2 架构

3、 数据双向交换

- 数据导入过程
- 数据导出过程

4、 数据导入工具与命令介绍

案例演练：Sqoop 数据导入/导出实际操作

第七部分： Pig 数据流处理引擎--数据脚本语言

1、 Pig 介绍

2、 命令行交互工具 Grunt

3、 Pig 数据类型

4、 Pig Latin 脚本语言介绍

- 基础知识
- 输入和输出
- 关系操作
- 调用静态 Java 函数

5、 Pig Latin 高级应用

6、 开发与测试 Pig Latin 脚本

- 开发工具
- 任务状态监控
- 调试技巧

7、 脚本性能优化

8、 用户自定义函数 UDF

案例演练：Pig Latin 脚本编写、测试与运行操作

结束：课程总结与问题答疑。