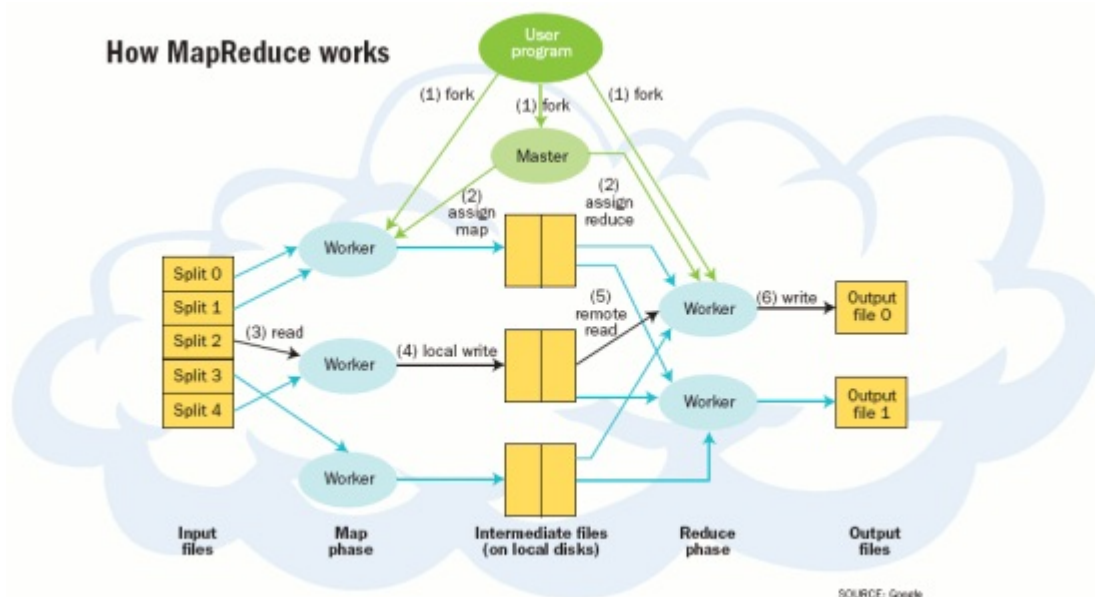




云计算 hadoop

课程定位与课程目标

现今企业数据仓库和关系型数据库擅长处理结构化数据，并且可以存储大量的数据。但成本上有些昂贵。这种对数据的要求限制了可处理的数据种类，同时这种惯性所带的缺点还影响到数据仓库在面对海量异构数据时对于敏捷的探索。这通常意味着有价值的数据源在组织内从未被挖掘。这就是 Hadoop 与传统数据处理方式最大的不同。Hadoop 带来了廉价的处理大数据（大数据的数据容量通常是 10-100GB 或更多，同时数据种类多种多样，包括结构化、非结构化等）的能力。本课程就重点探讨云计算与 Hadoop 系统的组成部分，并解释各个组成部分的功能。



适用学员：从 IT 行业和通信行业的构架师、数据分析师、程序员、工程师、产品经理

课程设计：

授课课时：	12 学时
授课条件：	学员必须具有基本的计算机知识

内容摘要：

第一讲：Hadoop 分布式系统

1. 什么是 Hadoop

- 发展简史
- 体系构架
- 厂家支持度与兼容

2. 为什么要选择 Hadoop

- 竞争环境
- 性能比拼

3. Hadoop 关键词和术语

第二讲：Hadoop 平台安装

1. 部署条件

- ◆ 支持平台
- ◆ 所需软件
- ◆ 安装软件

2. 下载源

3. 运行 Hadoop 集群的准备工作

4. Hadoop 单机操作方法

5. 伪分布式模式的操作

- ◆ 配置文件
- ◆ 免密码 Shh 设置
- ◆ 执行程序

第三讲：Hadoop 集群搭建与规划

1. Hadoop 软件安装

2. Master 配置

3. Slave 配置

4. 初始化和启动 Hadoop 集群

- ◆ 初始文件系统
- ◆ 启动 Hadoop

◆ 停止 Hadoop

5. 测试用例
6. 管理界面与命令

第四讲：Hadoop 资源调度

1. HDFS 特点与角色
2. MapReduce 介绍
3. 综合架构分析
4. MapReduce 应用

第五讲：Hadoop 集群性能评估测试

1. Hadoop 的测试方法
2. Mapred-site.xml 测试结果
3. 调度算法和压缩算法
4. Core-site.xml 测试结果
5. 整体系统测试工具

第六讲：Hadoop 集群监控和故障排除

1. Hadoop 的系统监控
2. Hadoop 中的命令总结
3. NameNode 与 JobTracker 单点故障说明
4. 经验总结

第七讲：Hadoop 运维与管理服务

1. 解决 tasktracker 存在，但无法 reduce 的问题
2. 解决单个 datanode 无法启动的问题
3. 性能查询常用命令
4. 状态获取常用命令
5. 其它日常问题说明

第八讲：Hadoop 应用案例

1. 全球最大超市业者 Wal-Mart
2. 全球最大拍卖网站 eBay
3. 全球最大信用卡公司 Visa
4. 全球最多用户社交圈 Facebook
5. 经验总结

授课语言：

中文