Cloudera Developer training for Spark and Hadoop(CCA-175)

尊敬的_____先生/女士，您好！

Cloudera Developer training for Spark and Hadoop(CCA-175)将于2017年08月在上海召开。

会议内容

Spark 及 Hadoop 开发员培训

学习如何将数据导入到 Apache Hadoop 机群并使用 Spark、Hive、Flume、Sqoop、Impala 及其他 Hadoop 生态系统工具对数据进行各种操作和处理分析

在为期四天的培训中，学员将学习关键概念和掌握使用最新技术和工具将数据采集到 Hadoop 机群并进行处理。通过学习掌握诸如 Spark、Hive、Flume、Sqoop 和 Impala 这样的 Hadoop 生态系统工具和技术，Hadoop 开发员将具备解决实际大数据问题和挑战的能力。本课程包含了大量的实操及编程练习来帮助学员熟悉并掌握各种工具，并最终获得在实际工作中针对特定的问题或场景来选取最佳解决工具或技术的能力。

“通过 Cloudera 的培训，让我们在使用大数据核心平台 Hadoop 方面，能把握现在、更能信心百倍地在未来面对和赢得更多的大数据挑战。”

——Persado

培训内容

通过讲师在课堂上的讲解，以及实操练习，学员将学习 Apache Spark 及如何将其集成到整个 Hadoop 生态系统中去，包括以下内容：

- 在 Hadoop 机群上进行分布式存储和处理数据。
- 通在 Hadoop 机群上编写、配置和部署 Apache Spark 应用。
- 使用 Spark shell 进行交互式数据分析。
- 使用 Spark SQL 查询处理结构化数据。
- 使用 Spark Streaming 处理流式数据。
- 使用 Flume 和 Kafka 为 Spark Streaming 采集流式数据。

培训对象及学员基础

本课程适合于具有编程经验的开发员及工程师。无需 Apache Hadoop 基础

- 培训内容中对 Apache Spark 的介绍所涉及的代码及练习使用 Scala 和 Python，因此需至少掌握这两个编程语言中的一种。

- 需熟练掌握 Linux 命令行。
- 对 SQL 有基本了解。

会议日程

cloudera

培训

课程大纲

课程介绍

Hadoop 及生态系统介绍
- Apache Hadoop 概述
- 数据存储和摄取
- 数据处理
- 数据分析和探索
- 其他生态系统工具
- 练习环境及分析应用场景介绍

Apache Hadoop 文件存储
- 传统大规模系统的问题
- HDFS 体系结构
- 使用 HDFS
- Apache Hadoop 文件格式

Apache Hadoop 机群上的数据处理
- YARN 体系结构
- 使用 YARN

使用 Apache Sqoop 导入关系数据
- Sqoop 简介
- 数据导入
- 导入的文件选项
- 数据导出

Apache Spark 基础
- 什么是 Apache Spark
- 使用 Spark Shell
- RDDs(可恢复的分布式数据集)
- Spark 里的函数式编程

Spark RDD
- 创建 RDD
- 其他一般性 RDD 操作

使用键值对 RDD
- 键值对 RDD
- MapReduce
- 其他键值对 RDD 操作

编写和运行 Apache Spark 应用
- Spark 应用对比 Spark Shell
- 创建 SparkContext
- 创建 Spark 应用（Scala 和 Java）
- 运行 Spark 应用
- Spark 应用 WebUI

配置 Apache Spark 应用
- 配置 Spark 属性
- 运行日志

Apache Spark 的并行处理
- 回顾：机群环境里的 Spark
- RDD 分区
- 基于文件 RDD 的分区
- HDFS 和本地化数据
- 执行并行操作
- 执行阶段及任务

Spark 持久化
- RDD 演变族谱
- RDD 持久化简介
- 分布式持久化

Apache Spark 数据处理的常见模式
- 常见 Spark 应用案例
- 迭代式算法
- 机器学习
- 例子：K – Means

DataFrames 和 Spark SQL
- Apache Spark SQL 和 SQL Context
- 创建 DataFrames
- 变更及查询 DataFrames
- 保存 DataFrames
- DataFrames 和 RDD
- Spark SQL 对比 Impala 和 Hive-on-Spark
- Spark 2.x 版本上的 Apache Spark SQL

Apache Kafka
- 什么是 Apache Kafka
- Apache Kafka 概述
- 如何扩展 Apache Kafka
- Apache Kafka 机群架构
- Apache Kafka 命令行工具

使用 Apache Flume 采集实时数据
- 什么是 Apache Flume
- Flume 基本体系结构
- Flume 源
- Flume 槽
- Flume 通道
- Flume 配置

集成 Apache Flume 和 Apache Kafka
- 概要
- 应用案例
- 配置

Apache Spark Streaming：DStreams 介绍
- Apache Spark Streaming 概述
- 例子：Streaming 访问计数
- DStreams
- 开发 Streaming 应用

Apache Spark Streaming：批处理
- 批处理操作
- 时间分片
- 状态操作
- 滑动窗口操作

Apache Spark Streaming：数据源
- Streaming 数据源概述
- Apache Flume 和 Apache Kafka 数据源
- 例子：使用 Direct 模式连接 Kafka 数据源

结论

会议门票

8500元/人次，包含一次对应的考试（如果来参加培训的人不考试，仅参加培训的价格为6500/人次）

费用包含：教材、实验手册、虚拟机、税票费用（8500是含考试的）

教室设施：投影、WiFi、排插、饮水（三餐及住宿请自理）

PS:上课时需携带笔记本电脑，虚拟机及课件由讲师发放

CCA Spark and Hadoop Developer (CCA175) 开发者认证

认证准备建议：Spark and Hadoop开发者培训

考试形式：120分钟；70%通过；解决10~12基于CDH5机群上需通过实际操作的问题

活动家

成都云数海量智能科技有限公司

2017年7月31日