邀请函 | 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：

- 数据是如何在 Hadoop 机群里进行分布式存储及处理的
- 如何使用 Sqoop 和 Flume 导入数据
- 如何使用 Apache Spark 处理分布式数据
- 如何使用 Impala 及 Hive 将结构化数据建模成表并进行分析查询
- 如何根据数据使用场景来确定最佳存储格式
- 数据存储最佳实践

培训对象及学员基础

本课程适合准备报考 CCA Spark 及 Hadoop 开发员认证考试的技术人员。虽然通过该认证考试，考生仍然需要做进一步的学习和准备，但是本课程涵盖了在该认证考试中考核的很多主题和知识点。

在参加完本培训后，我们建议学员参加此课程的一个后继课程：“设计和创建大数据应用”。

会议日程

课程介绍

Hadoop 及生态系统介绍

- 传统大规模系统的问题
- Hadoop ！
- Hadoop 生态系统

Hadoop 体系结构及 HDFS

- 机群环境下的分布式处理
- 存储：HDFS 体系结构
- 存储：使用 HDFS
- 资源管理：YARN 体系结构
- 资源管理：使用 YARN

使用 Apache Sqoop 导入关系数据

- Sqoop 简介
- 数据的基本导入导出
- 减少传输的数据量
- 改善 Sqoop 性能
- Sqoop 2

Impala 及 Hive 介绍

- 简介
- 为什么使用 Impala 及 Hive
- Hive 和传统数据库的比较
- Hive 应用场景

使用 Impala 及 Hive 管理数据及建模

- 数据存储
- 创建数据库及表
- 表数据导入
- HCatalog
- Impala 元数据缓存

数据格式

- 选择文件格式
- 支持不同文件格式的工具
- Avro 数据格式定义模式
- 在 Hive 及 Sqoop 里使用 Avro
- Avro 格式数据模式变更
- 压缩

数据分区

- 分区概述
- Impala 及 Hive 里的数据分区

Apache Flume 实时数据采集

- 什么是 Apache Flume
- Flume 基本体系结构
- Flume 源
- Flume 槽
- Flume 通道
- Flume 配置

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

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

编写和部署 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 Spark 属性
- 运行日志

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 缓存和持久化

- RDD 演变
- 缓存
- 分布式持久化

Spark 数据处理的常见模式

- 常见 Spark 应用案例
- 迭代式算法
- 图处理及分析
- 机器学习
- 例子：K - Means

预览：Spark SQL

- Spark SQL 和 SQL Context
- 创建 DataFrames
- 变更及查询 DataFrames
- 保存 DataFrames
- Spark SQL 对比 Impala

总结

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