

专注于商业智能BI和大数据的垂直社区平台

Spark初级课程

讲师: 平常心

目录





大数据技术之Hadoop

+Hadoop:

高效、可靠、可伸缩,能够为你的数据存储项目提供所需要的HDFS、MapReduce、YARN基础架构,并且运行主要的大数据服务和应用程序。

+Hadoop大事记:

- 2004年: Doug Cutting实现了HDFS和MapReduce的初版
- 2006年: Doug Cutting加入雅虎, Apache Hadoop项目正式启动
- 2008年:雅虎在900个节点上运行1TB排序测试集仅需要209秒,成为全球最快
- 2009年: 雅虎在1400个节点排序500GB数据59秒内, 在3400个节点排序100TB数据173分钟内
- ...
- 2016年: Apache Hadoop 3.0-alpha正式发布



大数据技术之Spark

Spark:

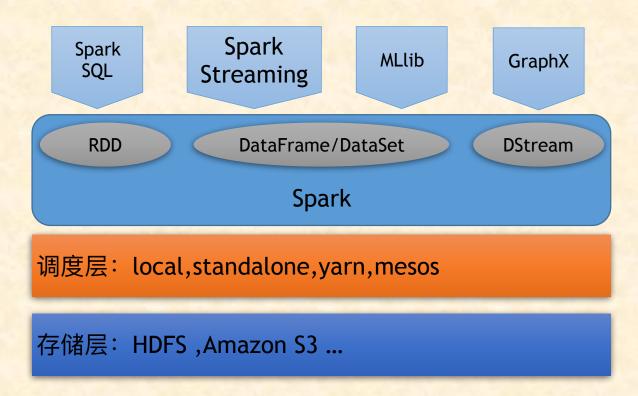
计算快速、使用简单、通用大数据处理引擎(One Stack To Rule Them All),Spark能更好地适用于需要迭代式计算。

Spark大事记:

- 2009年: Spark诞生于伯克利大学的AMPLab实验室
- 2010年: 伯克利大学正式开源Spark项目
- 2013年: Spark成为Apache的项目, 进入高速发展期
- 2014年: Spark成为Apache的顶级项目
- 2016年: Spark2.0正式发布



Spark应用架构





Spark和Hadoop MapReduce对比

★性能:

通常Spark将中间结果保存到内存中而不是将其写入磁盘,当数据大小适于读入内存,尤其是在专用集群上时, Spark 表现更好。

Hadoop MapReduce适用于那些数据不能全部读入内存的情况,同时它还可以与其它服务同时运行。

兼容性:

Spark 和 Hadoop MapReduce 具有相同的数据类型和数据源的兼容性。

下容错

Spark 和 Hadoop MapReduce 都有着较好的容错能力,但是 Hadoop MapReduce 要稍微更好一点。

成本

Spark基于高的CPU和内存配置,而MapReduce面向的是普通PC或者廉价PC,如果真的需要处理非常大的数据,Hadoop MapReduce绝对是合适之选,毕竟硬盘的费用要远远低于内存的费用。



更多商业智能BI和大数据精品视频尽在 www.hellobi.com

数据分析 与挖掘案例分析与实践

包括数据分析与挖掘的步骤,数据分析理 论方法及案例实现



特邀讲师: 彭远权 腾讯高级数据分析师

爬虫架构演变、爬虫安装配置、爬虫对比、爬虫技能 介绍、爬虫案例--以大众占评网站为案例进解整个过程

Python 网络爬虫实战

案例示范、定义数据产品、网络萃取、数据分析 本课程将会进行Python爬取新浪微博实战演示 BI、商业智能 数据挖掘 大数据 数据分析

R Python

机器学习

Tableau

QLIKVIEW

Hive Hadoop

BIWORK

BAO胖子 seng

曹浩贝克汉姆





R&Python机器学习入门

通俗易懂的讲解原理 结构、逐句的讲解使用



Power Bl可视化分析

快速玩转

3分析(小下) ^{职场课程}

从规划思维到技巧套路 打造数据分析师中的offer收割机

统计学基础

高效整合知识体系,奠定数据分析基础

讲师:菜鸟数据(Olivia)

