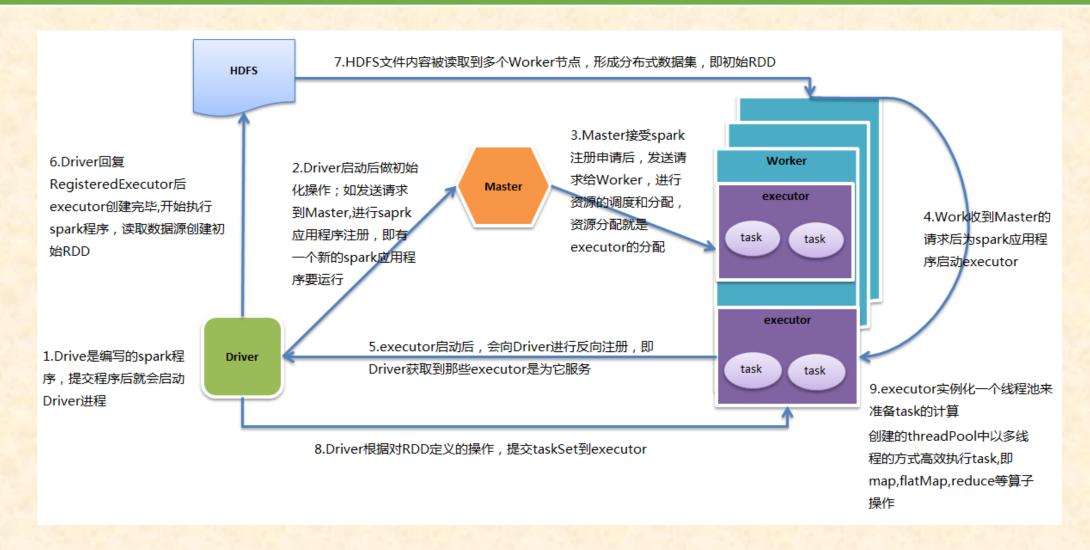


专注于商业智能BI和大数据的垂直社区平台

Spark初级课程

讲师: 平常心

Spark工作原理





RDD

1.RDD(Resilient Distributed Datasets):弹性分布式数据集,是只读,可分区,容错的,并行的数据结构。可以缓存在内存中,在多次计算阶段重用,并能够控制数据的分区。

引入RDD是为了在并行计算阶段之间高效地数据共享。

弹性是指在内存不够时可以与磁盘进行交换。

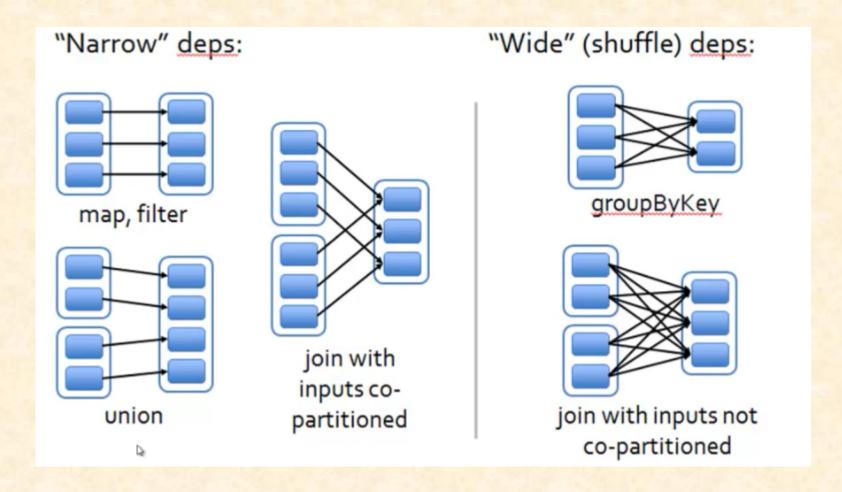
2.RDD的特征:

- 分区列表, 即能被切分, 能够切分的数据才能并行计算;
- compute函数对每个Partition进行计算
- 对其他RDD的依赖列表,依赖分为宽依赖和窄依赖,并不是所有的RDD都有依赖
- key-value型RDD默认使用的是HashPartitioner(可选)
- 分区的优先计算位置,如hdfs的block所在的位置应该是优先计算的位置(可选)



RDD依赖

RDD分为窄依赖和宽依赖





RDD的创建

1. Parallelized Collections 并行集合

val data = Array(1, 2, 3, 4, 5)

val distData = sc.parallelize(data)

2. External Datasets 外部数据集,如本地文件, hdfs文件, hbase等

val distFile = sc.textFile("data.txt")



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)

