## 查看spark job

software/hadoop-2.6.0/bin/yarn application –list

## spark学习要点

Spark 集群 的构建、Spark 架构设计、RDD、Shark/SparkSQL、机器学 习、图计算、实时流处理、Spark on Yarn、JobServer、Spark 测试、Spark 优化

## spark学习6阶段

**第一阶段:熟练的掌握Scala语言**

1. Spark 框架是采用 Scala 语言编写的,精致而优雅。要想成为 Spark 高手,你就必须

阅读 Spark 的源代码,就必须掌握 Scala,;

2. 虽然说现在的 Spark 可以采用多语言 Java、Python 等进行应用程序开发,但是最快

速的和支持最好的开发 API 依然并将永远是 Scala 方式的 API,所以你必须掌握 Scala

来编写复杂的和高性能的 Spark 分布式程序;

3. 尤其要熟练掌握 Scala 的 trait、apply、函数式编程、泛型、逆变与协变等;

推荐课程:”精通Spark的开发语言:Scala最佳实践”

**第二阶段:精通Spark平台本身提供给开发者API**

1. 掌握 Spark 中面向 RDD 的开发模式,掌握各种 transformation 和 action 函数的使用;

2. 掌握 Spark 中的宽依赖和窄依赖以及 lineage 机制;

3. 掌握 RDD 的计算流程,例如 Stage 的划分、Spark 应用程序提交给集群的基本过程和

Worker 节点基础的工作原理等

推荐课程:“18 小时内掌握Spark:把云计算大数据速度提高 100 倍以上!”

**第三阶段:深入Spark内核**

此阶段主要是通过 Spark 框架的源码研读来深入 Spark 内核部分:

1. 通过源码掌握 Spark 的任务提交过程;

2. 通过源码掌握 Spark 集群的任务调度;

3. 尤其要精通 DAGScheduler、TaskScheduler 和 Worker 节点内部的工作的每一步的

细节;

推荐课程:“Spark 1.0.0 企业级开发动手:实战世界上第一个Spark 1.0.0 课程,涵 盖Spark 1.0.0 所有的企业级开发技术”

**￼第四阶级:掌握基于Spark上的核心框架的使用**

Spark 作为云计算大数据时代的集大成者,在实时流处理、图技术、机器学习、NoSQL 查 询等方面具有显著的优势,我们使用 Spark 的时候大部分时间都是在使用其上的框架例如 Shark、Spark Streaming 等:

1. Spark Streaming 是非常出色的实时流处理框架,要掌握其 DStream、transformation

和 checkpoint 等;

2. Spark 的离线统计分析功能,Spark 1.0.0 版本在 Shark 的基础上推出了 Spark SQL,

离线统计分析的功能的效率有显著的提升,需要重点掌握; 3. 对于 Spark 的机器学习和 GraphX 等要掌握其原理和用法; 推荐课程:“Spark企业级开发最佳实践”

**第五阶级:做商业级别的Spark项目**

通过一个完整的具有代表性的 Spark 项目来贯穿 Spark 的方方面面,包括项目的架构设计、 用到的技术的剖析、开发实现、运维等,完整掌握其中的每一个阶段和细节,这样就可以让 您以后可以从容面对绝大多数 Spark 项目。 推荐课程:“Spark架构案例鉴赏:Conviva、Yahoo!、优酷土豆、网易、腾讯、淘 宝等公司的实际Spark案例”

**第六阶级:提供Spark解决方案**

1. 彻底掌握 Spark 框架源码的每一个细节;

2. 根据不同的业务场景的需要提供 Spark 在不同场景的下的解决方案;

3. 根据实际需要,在 Spark 框架基础上进行二次开发,打造自己的 Spark 框架;

推荐课程:“精通Spark:Spark内核剖析、源码解读、性能优化和商业案例实战”

函数式编程的核心特色之一是把函数作为参数传递给函数、在函数内部可以定义函数等。

# Scala

### Scala高阶函数