零基础学习Spark 1.x应用开发系列课程

Spark内核分析

讲师-梦琪

版权声明

课程详情访问云帆大数据官方培训网站

http://www.cloudyhadoop.com

Initializing Spark

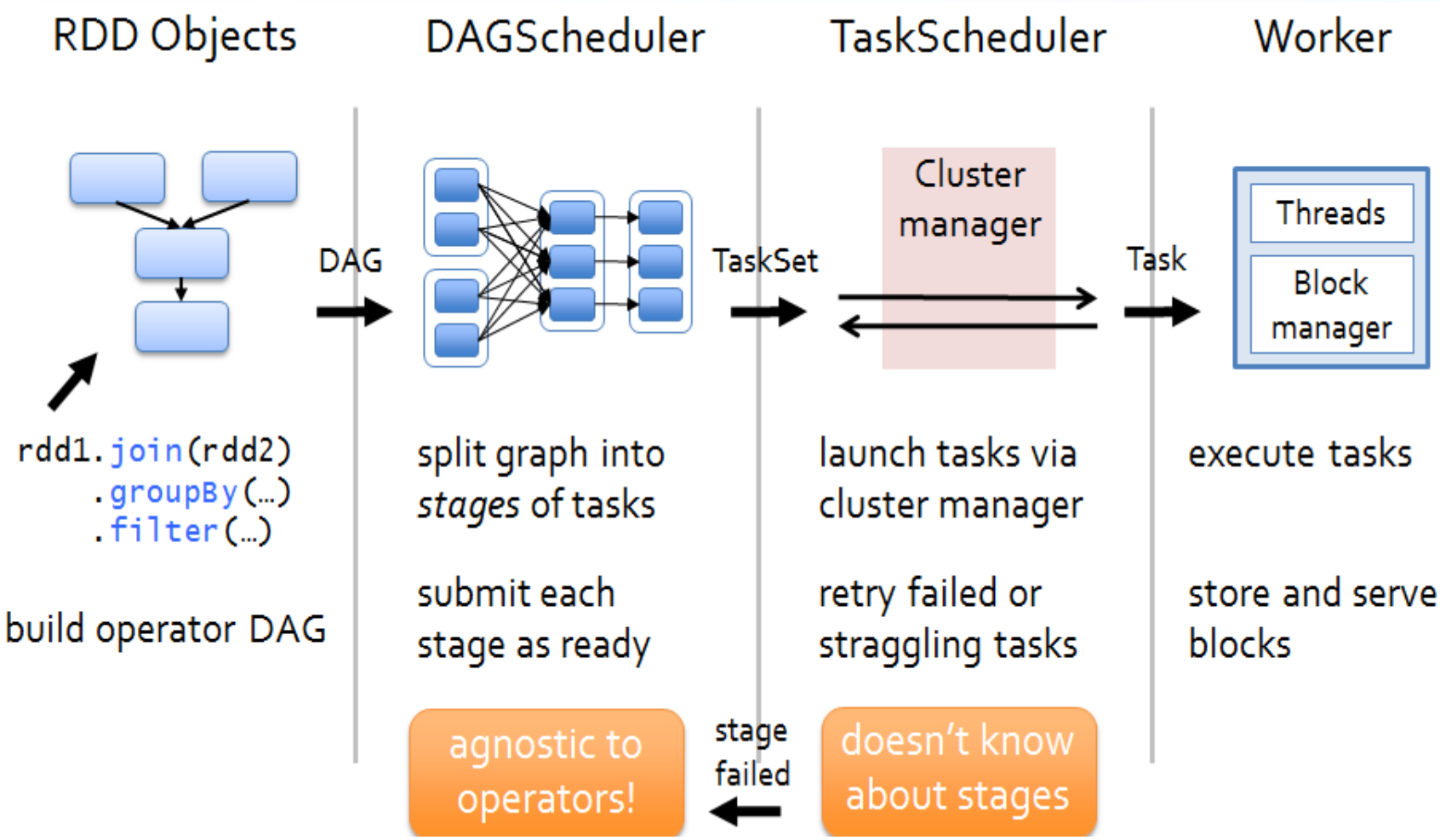
The first thing a Spark program must do is to create a SparkContext object, which tells Spark how to access a cluster. To create a `SparkContext` you first need to build a SparkConf object that contains information about your application.

Only one SparkContext may be active per JVM. You must `stop()` the active SparkContext before creating a new one.

```
val conf = new SparkConf().setAppName(appName).setMaster(master)
new SparkContext(conf)
```

The `appName` parameter is a name for your application to show on the cluster UI. `master` is a Spark, Mesos or YARN cluster URL, or a special "local" string to run in local mode. In practice, when running on a cluster, you will not want to hardcode `master` in the program, but rather launch the application with `spark-submit` and receive it there. However, for local testing and unit tests, you can pass "local" to run Spark in-process.

Spark Scheduler



DAGScheduler

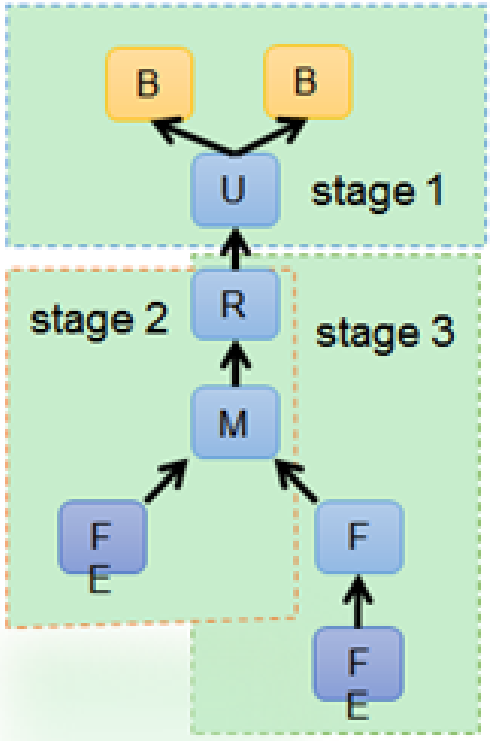
Spark program

```
val lines1 = sc.textFile(inputPath1)
val lines2 = sc.textFile(inputPath2)

t = t1.union(t2).map(…).reduce(…)

t.saveAsHadoopFiles(…)
t.filter(…).foreach(…)
```

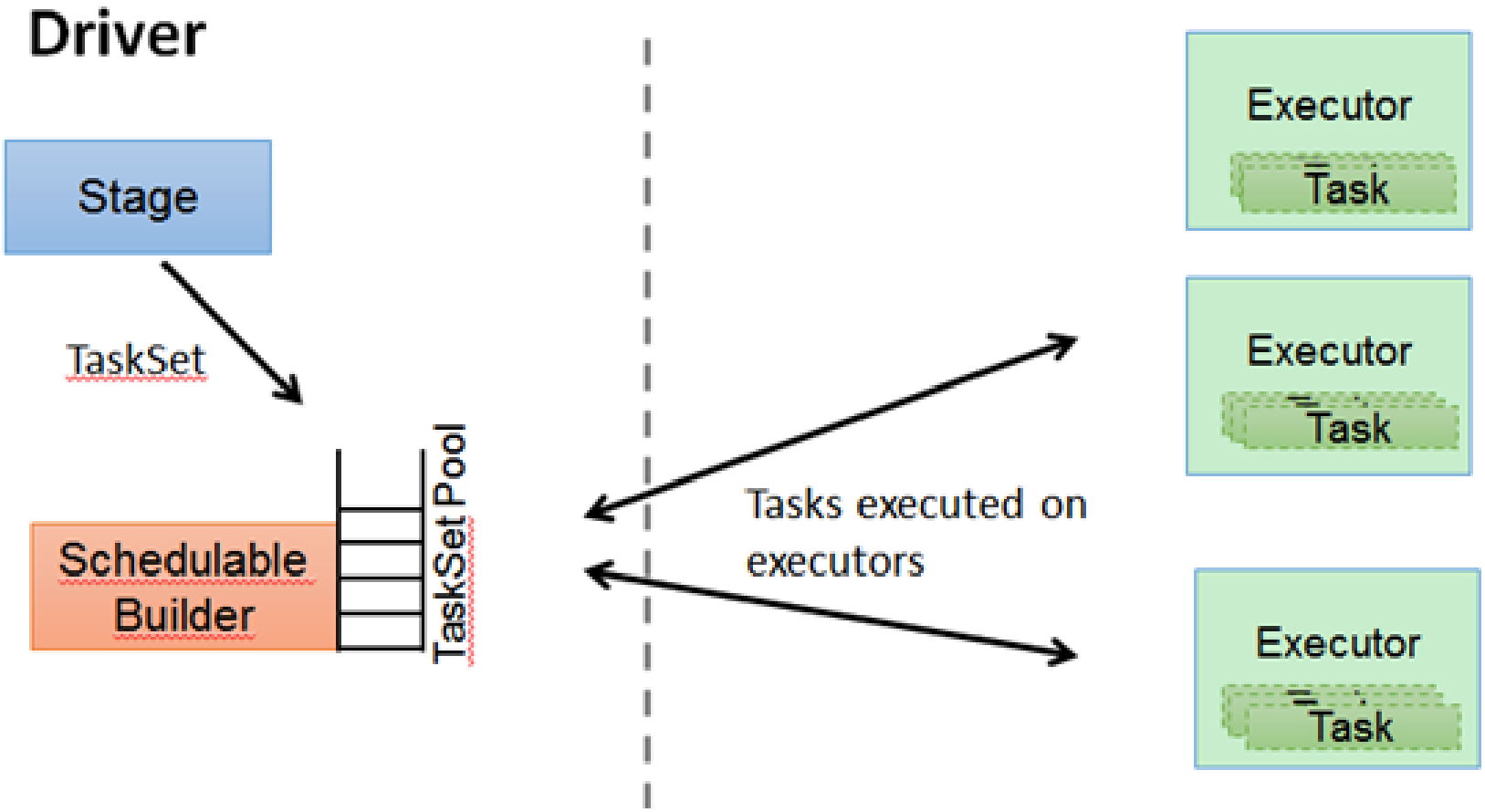
RDD Graph



DAGScheduler

- 接收用户提交的job
- 构建 Stage，记录哪个 RDD 或者 Stage 输出被物化
- 重新提交 shuffle 输出丢失的 stage
- 将 Taskset 传给底层调度器

TaskScheduler

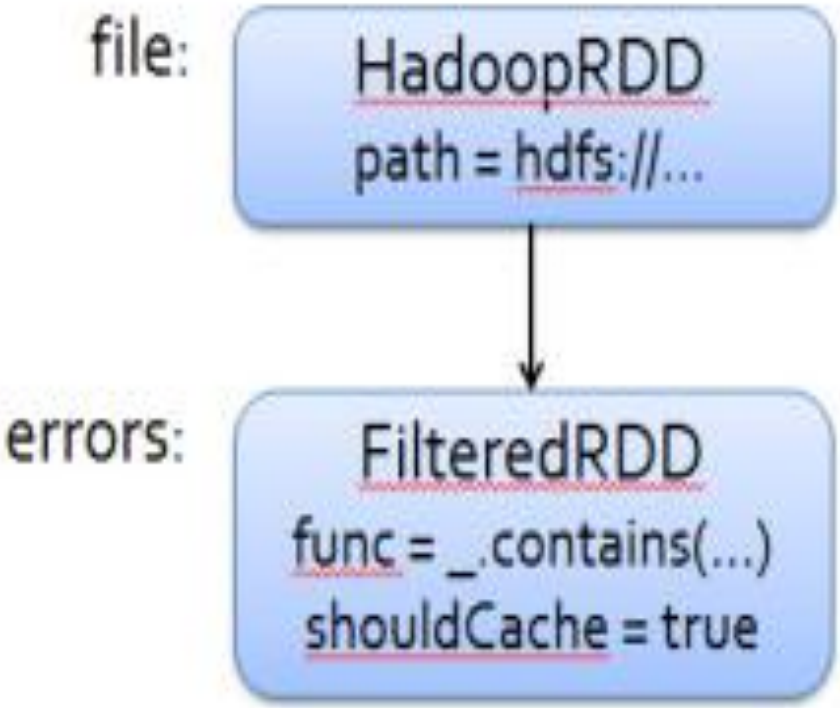


TaskScheduler

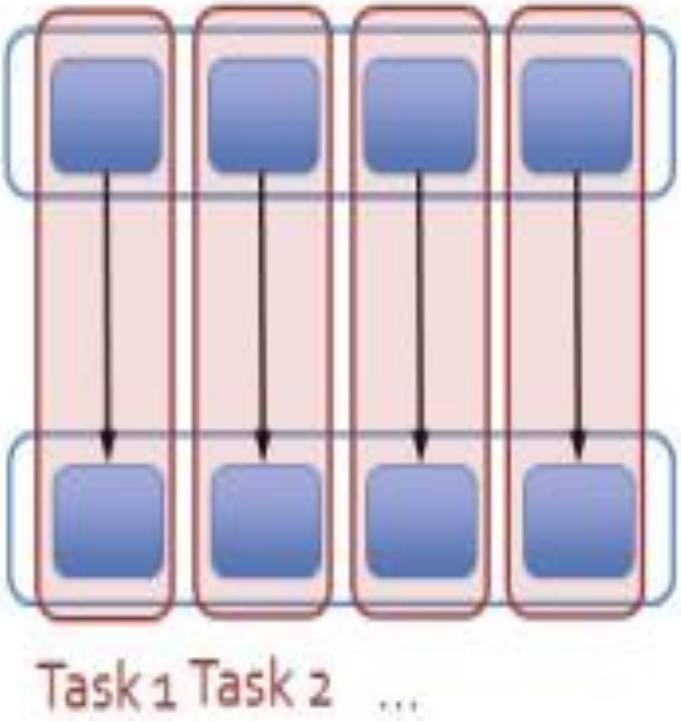
- 提交 taskset(一组 task) 到集群运行并监控
- 为每一个 TaskSet 构建一个 TaskSetManager 实例管理这个 TaskSet 的生命周期
- 数据本地性决定每个 Task 最佳位置 (process-local, node-local, rack-local and then any)
- 推测执行，碰到 straggle 任务需要放到别的节点上重试出现 shuffle 输出 lost 要报告 fetch failed 错误

Partition and Task

Dataset-level view:



Partition-level view:



Task

- Task是Executor中的执行单元
- Task处理数据常见的两个来源：外部存储以及shuffle数据
- Task可以运行在集群中的任意一个节点上
- 为了容错，会将shuffle输出写到磁盘或者内存中

云帆大数据网络课程

- 云帆大数据是国内首家坚持实时在线授课、提供高端开发课程网络培训机构。采用新兴的互联网教育模式，坚持实时在线授课模式，既继承传统教育的学习交流特点，又发挥互联网的无处不在的时空特性，将天南地北有志向的人才组织在一起学习交流，使原先孤立的个体学习，组合成有组织的学习探讨，并且把原先的学费用降低到十分之一左右，使更多的人能学习到最新的高端课程技术。云帆大数据同时是一个平台，如果你是一个学员，可以尽情的学习和交流；如果你是一个有梦想有才华的人，可以联系我们，给你提供一片驰骋的原野。

- 云帆大数据已推出国内首家《企业级Hadoop 2.x 应用开发课程》、《企业级Hadoop 2.x 项目实战课程》和《企业级Spark 1.x 应用开发课程》，更多其他详情，请登录我们的培训网站http://www.cloudyhadoop.com。

实时在线授课，专业课程辅导