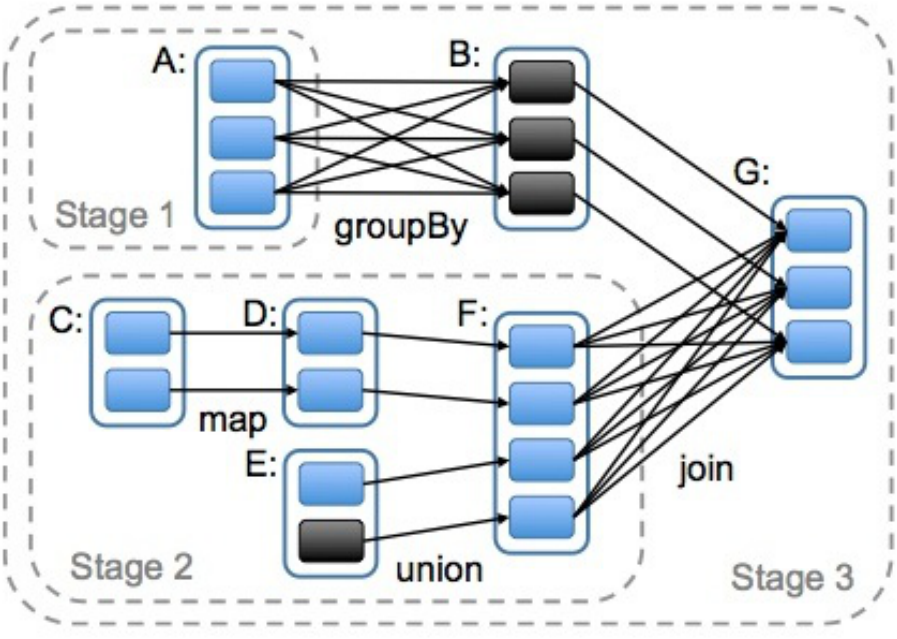
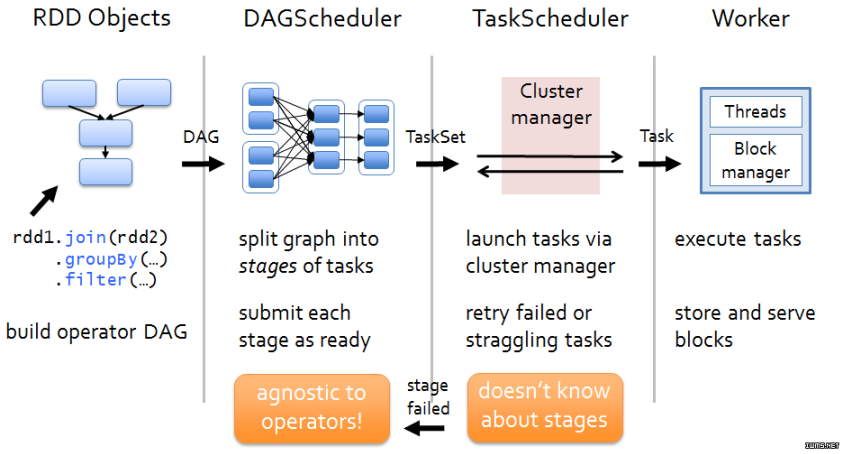
1. DAG 的生成

DAG（Directed Acyclic Graph）有向无环图是由原始 RDD 通过一系列转换而成。根据 RDD 之间依赖关系的不同将 DAG 划分成不同的 Stage。窄依赖是在 Stage 中完成 partition 的转换；而宽依赖由于 Shuffle 的存在，只能在父 RDD 处理完成后开始计算，因此宽依赖是划分 Stage 的依据。



如何划分 Stage：由最后一个 Stage 向前推导，由外向内递归，发生 Shuffle 时划分 Stage。

Spark 执行重要过程：



说明：

1、 调用 Spark 算子生成 RDD，构建 DAG 有向无环图；

2、 DAG Scheduler 依据宽依赖将生成的 DAG 切分成多个 Stage（Task 集合），并将 Stage 以 TaskSet 形式提交到 Task Scheduler；

3、 Task Scheduler 将 Task 提交到 Worker 节点所在的 Executor 进程执行；

4、 Executor 进程有多个线程，可以运行多个 Task，Task 是实现了 Runnable 接口的类。