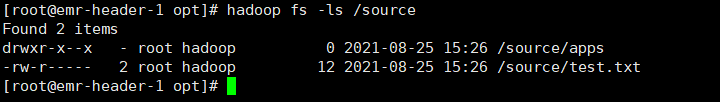
1. **运行步骤**
2. 源文件夹展示

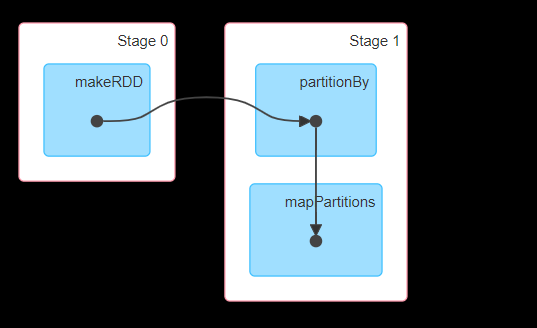


1. 目标文件夹展示



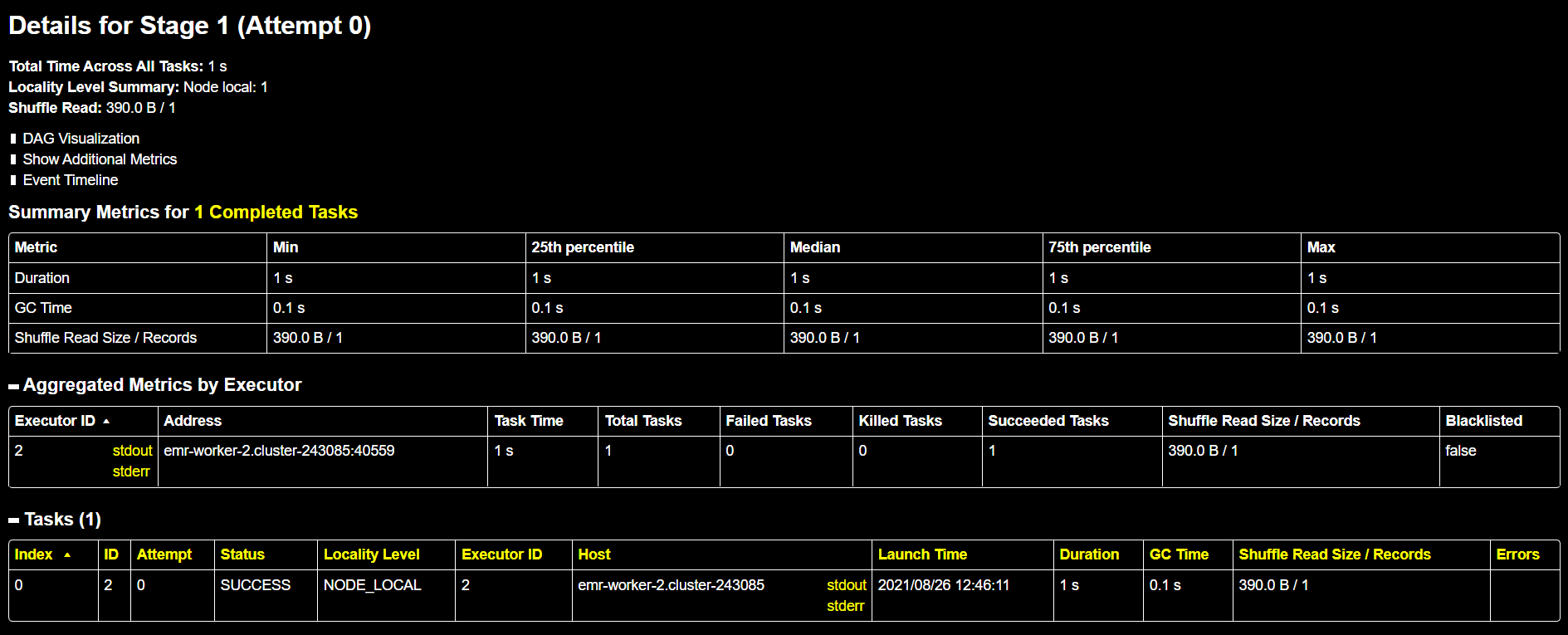
1. 运行结果

从可视化DAG图中可看出运行分为了2个STAGE,其中partitionBy会引起shuffle,所以另起了一个stage



以下进行参数设置验证:

1. 未带参数



可见只有1个task

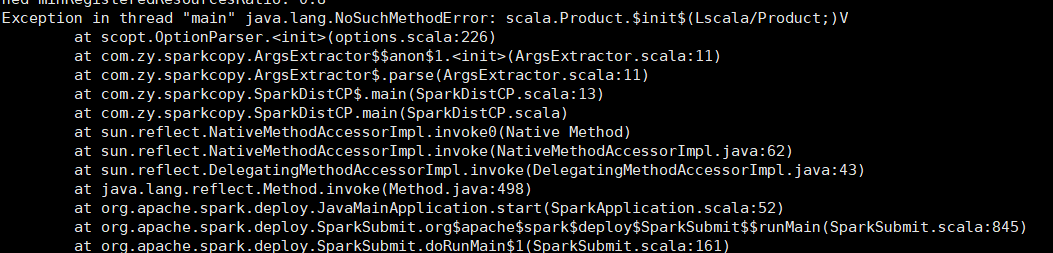
1. 带--m 7参数



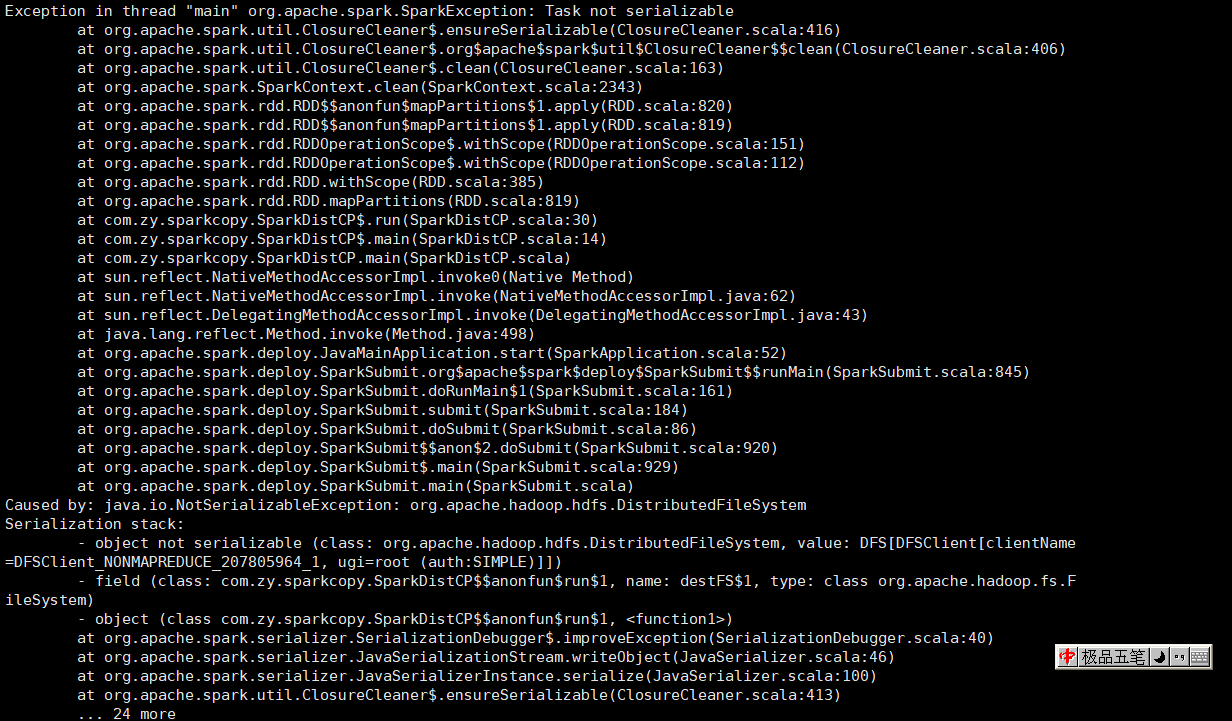
复制的文件及目录数目大于等于7,故使用了7个task进行复制

1. 带--i忽略错误参数

未模拟复制失败的场景,暂时看不出效果

1. **调试中遇到的问题**
2. 

解决方法：sbt assembly时指定的scala版本是2.12，而集群的scala版本是2.11.12，更改build.sbt的scala版本重新打包解决

1. 

解决方法:由于DFS不能被序列化,将sparkContext.hadoopConfiguration转为序列化变量,在mapPartitions时获取该变量,在RDD中根据hadoopConfiguration获取DFS