# **mapreduce spark short video**

[ant.zip](https://www.yuque.com/attachments/yuque/0/2019/zip/166647/1554786923460-8d39d6a6-67c6-402d-a7a3-885dbd0b20fe.zip?_lake_card=%7B%22uid%22%3A%22rc-upload-1554786913453-7%22%2C%22src%22%3A%22https%3A%2F%2Fwww.yuque.com%2Fattachments%2Fyuque%2F0%2F2019%2Fzip%2F166647%2F1554786923460-8d39d6a6-67c6-402d-a7a3-885dbd0b20fe.zip%22%2C%22name%22%3A%22ant.zip%22%2C%22size%22%3A4897421%2C%22type%22%3A%22application%2Fzip%22%2C%22ext%22%3A%22zip%22%2C%22progress%22%3A%7B%22percent%22%3A0%7D%2C%22status%22%3A%22done%22%2C%22percent%22%3A0%2C%22card%22%3A%22file%22%7D)

### **Category and Trending Correlation**

if there is any correlation between video category and trending popularity among countries

* we may expect to see a common set of trending **music** videos among many countries
* we may expect to see each country with a distinctive set of trending **political** videos

the average country number for videos in each category.

hadoop jar mr-video.jar video.trending.mr.AvgCountryNumber /tmp/video /tmp/vo

需要自定义分区，因为需要对每一个类目对视频ID以及国家去重，  
a/b: a是有多少不重复的国家，b是有多少不重复的视频

##### **Map阶段**

读取一行数据，变成一个KV（catalog + ":#" + video\_id, country)，即以类目和视频ID为key，以country为Value

##### **自定义分区**

为了让同一个reduce处理同一个类目下的数据，需要从K=catalog + ":#" + video\_id中截取出catalog，根据它返回一个固定的reduce id

##### **Reducer阶段**

Reduce输入的Key是（catalog + ":#" + video\_id），Value是对 country的迭代器，这样遍历Value就可以拿到同一个视频下有多少不重复的国家。  
由于Key有序，那么同一个catalog下的video\_id是有序的，依次处理，处理一条去重的video的计数器就可以加一，这样就能拿到同一个catalog下的去重的video\_id

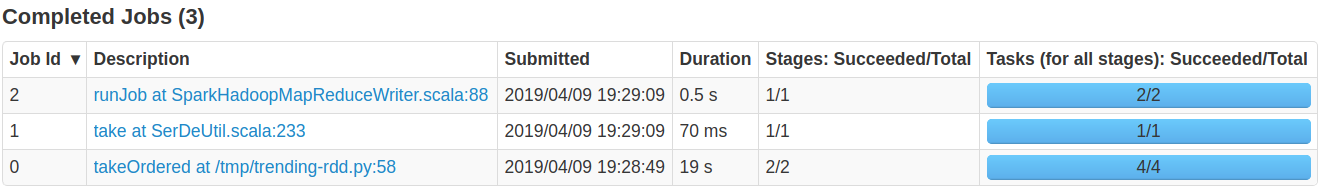
### **Controversial Trending Videos Identification**

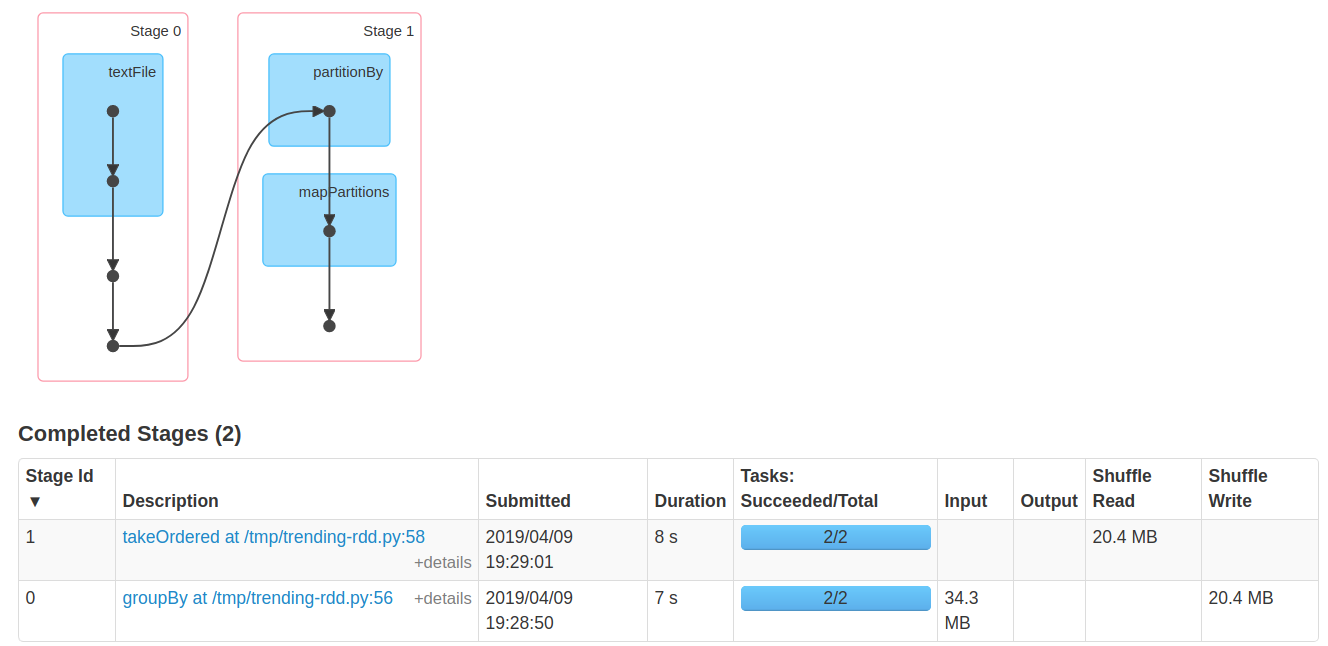
For some video, listing it as trending would increase its dislikes number more than the increase of its likes number. This workload aims to identify such videos.  
find out the top 10 videos with fastest growth of dislikes number between its first and second trending appearances.  
If a video has changed its category name over time, you can use the category of the first appearance.  
某一视频在某一个国家，以不同的类目出现时选取首次，意味着按照video\_id / country分区后，catalog要选最开始的2条

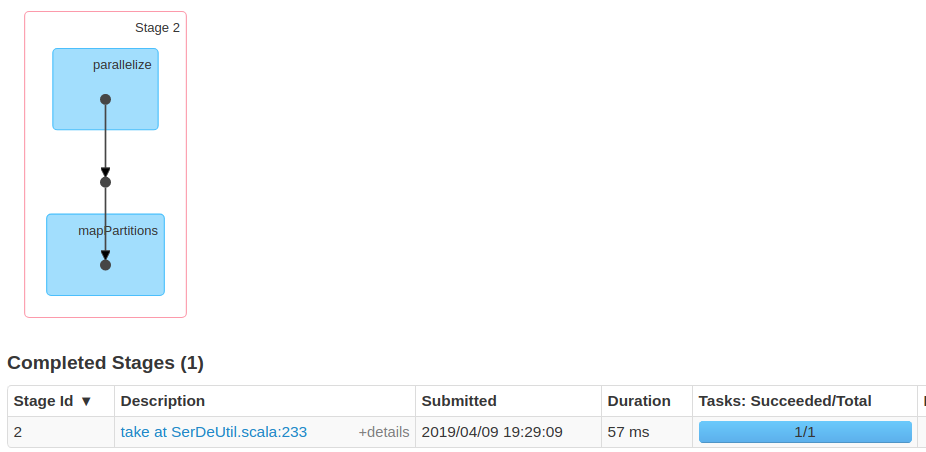
spark-submit --num-executors 4 --executor-cores 2 --executor-memory 4g --driver-memory 4g --master yarn trending-rdd.py /tmp/video /tmp/vo2



##### **RDD API**

分3个Job运行，为Spark根据调用的transform和action进行优化，生成的DAG。  


第一个Job完成了主要的逻辑，即读取数据，按照指定字段进行分区，分区内进行过滤、求和，最终生成一个有序的List。  


第二个Job和第三个Job如下，主要负责将这个List持久化到HDFS，并为了输出成规定的格式，将KV中各个字段的顺序进行了调整。  
  
第三个Job  


##### **Dataframe API**

[trending.zip](https://www.yuque.com/attachments/yuque/0/2019/zip/166647/1554547866699-76eb56d0-065a-47f1-b04c-411e40436c2d.zip?_lake_card=%7B%22uid%22%3A%22rc-upload-1554547796902-9%22%2C%22src%22%3A%22https%3A%2F%2Fwww.yuque.com%2Fattachments%2Fyuque%2F0%2F2019%2Fzip%2F166647%2F1554547866699-76eb56d0-065a-47f1-b04c-411e40436c2d.zip%22%2C%22name%22%3A%22trending.zip%22%2C%22size%22%3A1561%2C%22type%22%3A%22application%2Fzip%22%2C%22ext%22%3A%22zip%22%2C%22progress%22%3A%7B%22percent%22%3A0%7D%2C%22status%22%3A%22done%22%2C%22percent%22%3A0%2C%22card%22%3A%22file%22%7D)