COMP5349 Assignment 1 Tuto r Name: Heming Ni

Name: Shanshan Luo SID: 470349580

|  |  |  |
| --- | --- | --- |
| Workload | Implementation | Programming Language |
| Category and Trending Correlation | MapReduce | Java |
| Controversial Trending Videos Identification | Spark | Python |

Workload: Category and Trending Correlation

The computation graph is illustrated in Figure 1.

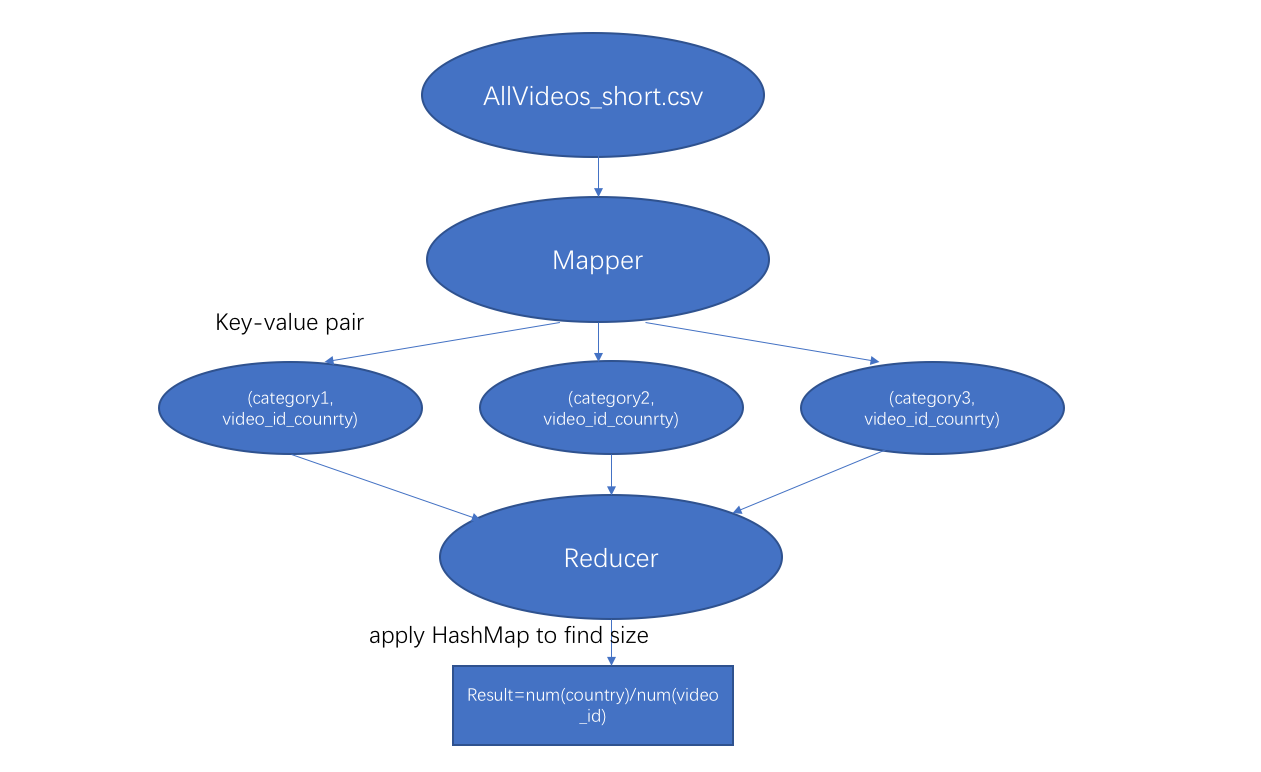


Figure1: MapReduce computation graph for workload 1.

In map phase, the AllVideos\_short.csv is read in split by comma to get all attributes we need. Video\_id, country and category information are extracted after ensuring the length of data row is correct. Then extracted information can be stored in key-value pair. Category is regarded as key and the combination of video\_id and country are regarded as value, which can be written to context to do the reduce operation

In reduce phase, key-value pairs are transformed to reducer. Keys are text format and values are iterable text format. two HashMap are applied to help count the number. The iterable value are putted into one HashMap and the split video\_id information is putted into another one. So, the first HashMap contains corrected number of categories which is also the country number and the second HashMap contains corrected number of different video\_id. after figuring out the size of each HashMap the average country number for videos in each category can be calculated by division method. The final result is converted to text format to be written into context.

Workload: Controversial Trending Videos Identification

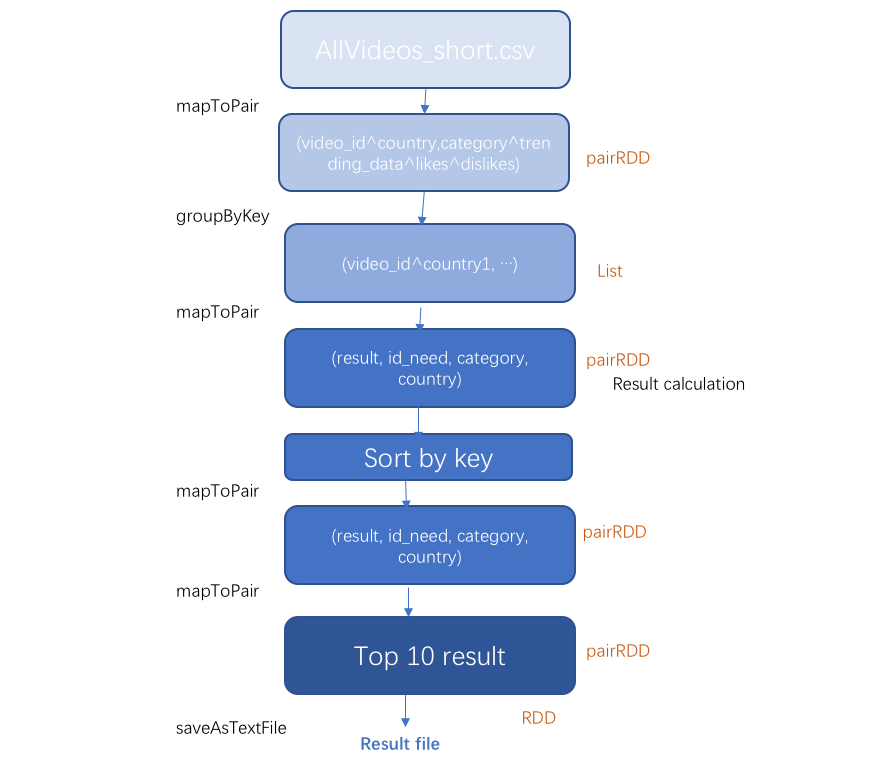


Figure1: Spark computation graph for workload 2.

This workload applied three functions in Spark: map, groupByKey and sortByKey. At the first stage, AllVideos\_short.csv are extracted to get information. Video\_id, country, category, trending\_date, likes and dislikes are stored in binary-tuple and the first element in this tuple is the combination of video\_id and country. Then the binary-tuple are grouped by key (the first element) and converted to list for the further operation. Filter is applied for filtering out information which appears once and cannot compare with others. then processed list can be mapped again to store information back to tuple. The growth number of dislikes between the first and second trending appearance is also calculated at this stage. Then tuple is sorted by key and stored in the correct format. At the final stage, the first 10 sorted tuple are taken and converted to rdd format for saving in text file.