

Advanced Topics in Database Activity 1

Assume that you have a Students dataset stored on the HDFS. The dataset contains text files, each text file has the following information (comma separated):

studentID, course_name, grade

each file may contain several students, and for the same student there are multiple lines, each line contains student's grade in a particular course.

Dataset Sample:

19, web , 85

19, Database , 90

20, web, 90

24, Java , 87

23, Java, 88

24, algorithms , 87

23, algorithms, 88

Write a MapReduce job that reads the input files that contains student information, generate student average in the format "student_id, average" for example 19, 87.5