CSSE 490

Introduction to the Hadoop Ecosystem

Exam 1 – Take Home Part (50 Points) Due: 1:30 PM Tuesday October 6th 2015

Task 1: Let us assume you have two input files titled grades.txt and courses.txt. The file grades contains the following information (Name, Course Number, Score)

Bill Watterson,CS100, 92.5
Randall Munroe,CS100, 90
Bill Watterson,AH255, 82.5
Randall Munore,AH255, 85
Bill Watterson,CSSE255, 87.5
Randall Munroe,CSSE255, 89.5
Bill Watterson,CHEM100, 65
Bill Watterson,PH201, 70
Bill Watterson,BIO355, 72.75
Randall Munroe,BIO355, 74
Randall Munroe,CHEM100, 75
Randall Munroe,PH201, 76
Sriram Mohan,CSSE255, 95
Sriram Mohan,PH201, 88.5
Sriram Mohan,CHEM100, 84.5
Sriram Mohan,AH255, 79
Sriram Mohan,BIO355, 80

Courses.txt contains the following information (Course Number, Course Name)

CS100,Advanced Comic Strips
AH255,History of Modern Comic Strip Art
CSSE255,Digital Comic Strip Design
CHEM100,Essential Chemistry for Artists
PH201,Time and Space
BIO355,Genetics and Molecular Biology for Artists
CAP410,Senior Project- Design your own Comic Strip

Please implement a reduce side join that accepts the following input arguments

1. Path to grades.txt
2. Path to courses.txt
3. Path of output directory

and produces the following output.

Randall Munore	AH255	History of Modern Comic Strip Art	85
Sriram Mohan	AH255	History of Modern Comic Strip Art	79
Bill Watterson	AH255	History of Modern Comic Strip Art	82.5
Sriram Mohan	BIO355	Genetics and Molecular Biology for Artists	80
Randall Munroe	BIO355	Genetics and Molecular Biology for Artists	74
Bill Watterson	BIO355	Genetics and Molecular Biology for Artists	72.75
Randall Munroe	CHEM100	Essential Chemistry for Artists	75
Bill Watterson	CHEM100	Essential Chemistry for Artists	65
Sriram Mohan	CHEM100	Essential Chemistry for Artists	84.5
Randall Munroe	CS100	Advanced Comic Strips	90
Bill Watterson	CS100	Advanced Comic Strips	92.5
Randall Munroe	CSSE255	Digital Comic Strip Design	89.5
Bill Watterson	CSSE255	Digital Comic Strip Design	87.5
Sriram Mohan	CSSE255	Digital Comic Strip Design	95
Bill Watterson	PH201	Time and Space	70
Sriram Mohan	PH201	Time and Space	88.5
Randall Munroe	PH201	Time and Space	76

Turn in

1. Create a folder called Exam1Join
2. This folder will contain all the java files needed to accomplish the task and the .jar file that you built using Maven. The folder should also contain a text file that includes the command for executing your code on a test cluster.

Task 2: Please write a simple Map Reduce job that accepts the output of the previous question as input and calculates the average grade in each course.

Turn in

1. Create a folder called Exam1Average
2. This folder will contain all the java files needed to accomplish the task and the .jar file that you built using Maven. The folder should also contain a text file that includes the command for executing your code on a test cluster.

Things to Include in your Submission:

Your submission should be a zip file named username_Exam1.zip that comprises of 2 folders:

1. Exam1Join - This folder will contain all the java files needed to accomplish Task 1 and the .jar file that you built using Maven. The folder should also contain a text file that includes the command for executing your code on a test cluster.

2. Exam1Average - This folder will contain all the java files needed to accomplish Task 2 and the .jar file that you built using Maven. The folder should also contain a text file that includes the command for executing your code on a test cluster.

3. Please zip up the file and upload to Moodle → Exam Drop Box – Take Home Exam 1