1.Given the dataset “Student.txt” with below schema

studentId,Subject,Marks,School

101,english,60,SMCS

Find out the top 3 students school wise

Loading the data:

Studentdata = load ‘/home/hduser/students.txt’ using PigStorage(‘,’) as(

Studentid:int,

Subject:chararray,

Marks:int,

School:chararray

);

dump;

This will list down the data present in the students.txt file.

grpd = group studendata by studentid;

dump;

(101,{(101,english,60,SMCS),(101,Science,80,SMCS),(101,Math,90,SMCS)})

(102,{(102,english,70,SMCS),(102,science,55,SMCS),(102,Math,42,SMCS)})

(103,{(103,english,81,SMCS),(103,Science,26,SMCS),(103,Math,79,SMCS)})

(104,{(104,english,21,SMCS),(104,science,32,SMCS),(104,Math,78,SMCS)})

(105,{(105,Math,80,SMCS),(105,Science,55,SMCS),(105,english,89,SMCS)})

(201,{(201,english,55,AOG),(201,Math,77,AOG),(201,Science,42,AOG)})

(202,{(202,english,86,AOG),(202,science,68,AOG),(202,Math,59,AOG)})

(203,{(203,english,91,AOG),(203,Science,72,AOG),(203,Math,63,AOG)})

(204,{(204,english,89,AOG),(204,Science,84,AOG),(204,Math,61,AOG)})

(205,{(205,english,73,AOG),(205,Science,92,AOG),(205,Math,91,AOG)})

student\_wise\_marks = foreach grpd generate FLATTEN(studentdata.school),group as studentid,SUM(studentdata.marks) as total\_marks;

grpd1 = group student\_wise\_marks by school;

school\_wise\_top\_marks = foreach grpd1{

sorted = order student\_wise\_marks by total\_marks desc;

lmt = limit sorted 1;

generate FLATTEN(lmt);

};

Output:

(AOG,205,256)

(SMCS,101,230)