

001,56,CSE

002,67,DSE

003,78,CCE

004,89,DSE

005,40,CSE

006,51,CCE

007,58,DSE

008,63,CCE

this is 1st dataset and another dataset is:

001,Rajiv,Reddy,21,9848022337,Hyderabad

002,siddarth,Battacharya,22,9848022338,Kolkata

003,Rajesh,Khanna,22,9848022339,Delhi

004,Preethi,Agarwal,21,9848022330,Pune

005,Trupthi,Mohanthi,23,9848022336,Bhuwaneshwar

006,Archana,Mishra,23,9848022335,Chennai

007,Komal,Nayak,24,9848022334,trivendram

008,Bharathi,Nambiayar,24,9848022333,Chennai

CODE:

-- Load the first dataset

```
data1 = LOAD 'dataset1.csv' USING PigStorage(',') AS (id1:chararray, marks:int,
department:chararray);
```

-- Load the second dataset

```
data2 = LOAD 'dataset2.csv' USING PigStorage(',') AS (id2:chararray, firstname:chararray,
lastname:chararray, age:int, phone:chararray, city:chararray);
```

-- Join the datasets based on the common ID

```
joined_data = JOIN data1 BY id1, data2 BY id2;
```

-- Filter out records for CSE students

```
cse_students = FILTER joined_data BY department == 'CSE';
```

-- Calculate average marks for CSE students

```
cse_avg_marks = FOREACH (GROUP cse_students BY department) GENERATE
AVG(cse_students.marks) AS avg_marks;
```

-- Display CSE student names and marks along with the average marks

```
cse_student_details = FOREACH cse_students GENERATE data2::firstname, data2::lastname,
data1::marks;
```

```
cse_student_avg_marks = CROSS 'cse_avg_marks' ALL, cse_student_details;
```

```
cse_student_avg_marks = FOREACH cse_student_avg_marks GENERATE data2::firstname,
data2::lastname, data1::marks, $0 AS avg_marks;
```

-- Store the output

```
STORE cse_student_avg_marks INTO 'output' USING PigStorage(',');
```