**ACD\_BDD\_Session\_9\_Assignment\_2\_Main**

**Problem Statement**

**1. Introduction**

In this assignment you will write the Pig code for analyzing the Student data set.

**2. Objective**

This assignment will help you to understand the implementation of the Pig commands in real time scenario.

**3. Prerequisites**

Acadgild’s VM, or Linux operating system with Hadoop and Pig installed in it.

**4. Associated Data Files**

https://github.com/prateekATacadgild/mAPrEDuce/blob/master/studentDS

**Data Set Description:**

**Data Dictionary**

1. Name : chararray

2. Col.Undertaken : chararray

3. Dob : chararray

4. Stream : chararray

5. Grade : float

6. State : chararray

7. City : chararray

**5. Problem Statements**

1. Find the number of students who scored less than 5(50%).

2. List the name of students who are from Alaska.

3. How many government collages are there in Alabama?

4. List the name of students who are from Oregon and persuing BE.

**Solution**

**Problem 1:**

grunt> stud = load '/home/acadgild/Downloads/Student\_DS' USING PigStorage(',') AS (Name:chararray, Col:chararray, Dob:chararray, Stream:chararray, Grade:float, State:chararray, City:chararray);

grunt> grade\_5 = FILTER stud BY Grade <= 5;

grunt> grp = GROUP grade\_5 ALL;

grunt> Count = FOREACH grp\_all GENERATE group, COUNT(filt.$0);

grunt> DUMP Count;

(all,188)

**Problem 2:**

grunt> stud = load '/home/acadgild/Downloads/Student\_DS' USING PigStorage(',') AS (Name:chararray, Col:chararray, Dob:chararray, Stream:chararray, Grade:float, State:chararray, City:chararray);

grunt> filt = FILTER stud BY State == 'alaska';

grunt> listname = FOREACH filt GENERATE Name;

grunt> DUMP listname;

(BELL)

(DAVIS)

(LOGAN)

(YOUNG)

(COWEN)

(DISHONG)

(GIBSON)

(IMMILE)

(KEYSER)

(POLK)

(ROBTS&CMRN)

(STEPHENS)

(WEIRS)

(JACOB)

(BARBARA)

(GEORGE)

(SARAH)

(WILLIAM)

(JOHN)

(JOHN)

(STILLWEL)

(Albright)

(Bowman)

(Brunner)

(Carson)

(Diller)

(Denyes)

(Ellmaker)

(Fisher)

(Gumpf)

(Hartman)

(Halbach)

(Huber)

(Houser)

(Jordan)

(Keubler)

(Kendig)

(Long)

(McGinnis)

(Miller)

(Nourse)

(Quinn)

(Rotharmel)

(Sherer)

(Sherwood)

(Voight)

(Waters)

(Wiley)

(Zahm)

(Matthew)

(Peter)

(Christian)

(George)

(Alexander)

(Frederick)

(Michael)

(Richard)

(Jacob)

(Joseph)

(John W)

(Michael O)

(Daniel)

(William)

(William)

(Joseph)

(Henry)

(Samuel)

(John)

**Problem 3:**

grunt> stud = load '/home/acadgild/Downloads/Student\_DS' USING PigStorage(',') AS (Name:chararray, Col:chararray, Dob:chararray, Stream:chararray, Grade:float, State:chararray, City:chararray);

grunt> coll = FILTER stud BY (Col == 'goverenment') AND (State == 'alabama');

grunt> coll\_grp = GROUP coll ALL;

grunt> Num = FOREACH coll\_grp GENERATE group, COUNT(coll.$1);

grunt> DUMP Num;

(all,2198)

**Problem 4:**

grunt> stud = load '/home/acadgild/Downloads/Student\_DS' USING PigStorage(',') AS (Name:chararray, Col:chararray, Dob:chararray, Stream:chararray, Grade:float, State:chararray, City:chararray);

grunt> filt1 = FILTER stud BY ($3 == 'BE') AND ($5 == 'oregon');

grunt> stu = FOREACH filt1 GENERATE $0;

grunt> DUMP stu;

(BLOOM)

(PATTERSON)

(WAGGONER)

(IMMILE)

(SHOCK)

(WHITE)

(JOHN)

(RICHARD)

(JACOB)

(Albright)

(Buckius)

(Bowman)

(Cordes)

(Danner)

(Donas)

(Ehrisman)

(Frick)

(Grimler)

(Heinitsh)

(Himmelsbach)

(Hubley)

(Kitch)

(Lambert)

(McNair)

(Naumann)

(Royal)

(Shindle)

(Wayne)

(Reformed)

(P.K.)

(Charles H)

(Jacob)

(Daniel)

(George)

(John)

(Peter)

(George B)

(Philip)

(Dennis)

(George)

(Edwin)

(Frederick)