Session 8: Pig

Assignment 2

# Problem Statement

**Create a sample dataset and implement the below Pig commands on the same dataset.**

**File system:** Dataset: student.tsv

A01 John 45.0

B02 James 85.0

C03 John 45.0

A02 James 85.0

A03 Smith 60.0

A04 Scott 70.0

A07 Shoba 80.0

B08 Taanu 90.0

C02 Anbu 95.0

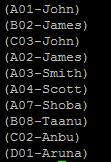
D01 Aruna 85.0

1. **Concat : Script:**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

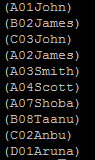
Y = FOREACH A GENERATE CONCAT((chararray)rollno,CONCAT('-',(chararray)name));

DUMP Y;



X = FOREACH A GENERATE CONCAT (rollno, name) AS new\_id;

DUMP X;



1. **Tokenize : Script :**

lines = LOAD 'assignment1.txt' AS (line:chararray);

w1 = FOREACH lines GENERATE TOKENIZE(line) as w2;

DUMP w1;



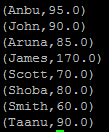
1. **Sum : Script :**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

Y = GROUP A BY name;

X = FOREACH Y GENERATE group, SUM(A.gpa);

DUMP X;



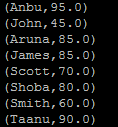
1. **Min : Script :**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

B = GROUP A BY name;

C = FOREACH B GENERATE group, MIN(A.gpa);

DUMP C;

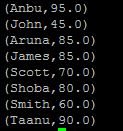


1. **Max : Script :**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

Y = GROUP A BY name;

X = FOREACH Y GENERATE group, MAX(A.gpa);



1. **Limit : Script :**

X = load 'student.tsv' as (rollno: chararray, name:chararray, gpa:float);

Y = LIMIT X 3;

DUMP Y;

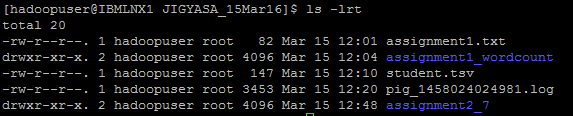


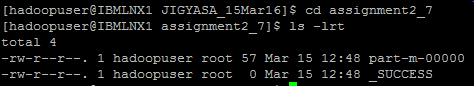
1. **Store : Script :**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

Y = FOREACH A GENERATE name;

store Y into '/home/hadoopuser/JIGYASA\_15Mar16/assignment2\_7';





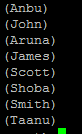
1. **Distinct : Script :**

A = load 'student.tsv' as (rollno:chararray, name:chararray, gpa:float);

Y = FOREACH A GENERATE name;

Z = DISTINCT Y;

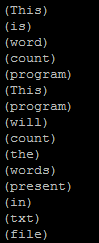
DUMP Z;



1. **Flatten : Script :**

lines =LOAD '/home/hadoopuser/JIGYASA\_15Mar16/assignment1.txt' AS (line:chararray);

words= FOREACH lines GENERATE FLATTEN(TOKENIZE(line)) as word;



1. **IsEmpty: Script :**

rollno.tsv

1001

1002

1003

1004

1005

assinment2.tsv

1001 John

1003 James

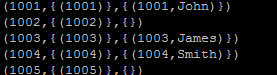
1004 Smith

StudId = LOAD 'rollno.tsv' AS (StudNum:int);

StudName = LOAD 'assinment2.tsv' AS (StudId:int,StudName:chararray);

Stud = COGROUP StudId BY StudNum INNER,StudName BY StudId;

DUMP Stud;



EmptyData = FILTER Stud BY IsEmpty(StudName);

DUMP EmptyData;

