BIG DATA TECHNOLOGIES LAB

(ACADEMIC YEAR : 2017-2018)

I SEMESTER

ASSIGNMENT 3

TOPIC : PIG PRACTICE SESSION I

ASHISH CHANDRAKANT DUSANE

M. TECH. (ACDS)

COMPUTER ENGG. DEPARTMENT

{ PRN : 170101261004 }

~: PIG Practice Session I :~

## To Extract data from HDFS

```
student1 = LOAD '/home/student/Documents/Pig/student_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray);
```

```
grunt> dump student1;
```

# Output :

```
(1,Rajiv,Reddy,21,9848022337,Hyderabad)

(2,siddarth,Battacharya,22,9848022338,Kolkata)

(3,Rajesh,Khanna,22,9848022339,Delhi )

(4,Preethi,Agarwal,21,9848022330,Pune )

(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar )

(6,Archana,Mishra,23,9848022335,Chennai )

(7,Komal,Nayak,24,9848022334,trivendram )

(8,Bharathi,Nambiayar,24,9848022333,Chennai)
```

## To Load data into HDFS

```
STORE student1 INTO '/home/student/Documents/Pig/pig_Output/' USING PigStorage (',');
```

# Output :

```
1,Rajiv,Reddy,21,9848022337,Hyderabad

2,siddarth,Battacharya,22,9848022338,Kolkata

3,Rajesh,Khanna,22,9848022339,Delhi

4,Preethi,Agarwal,21,9848022330,Pune

5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar
```

```
6,Archana,Mishra,23,9848022335,Chennai

7,Komal,Nayak,24,9848022334,trivendram

8,Bharathi,Nambiayar,24,9848022333,Chennai
```

```
## Group Operator 1

## Student Details

hdfs dfs -put student_details.txt /pig

hdfs dfs -ls /pig

hdfs dfs -cat /pig/student_details.txt

student_details = LOAD 'hdfs://master:8020/pig/student_details.txt' USING
PigStorage(',') as

(id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,
city:chararray);

group_data = GROUP student_details by age;

Dump group_data;
```

```
## Group Operator 2

group_multiple = GROUP student1 by (age, city);

dump group_multiple;

# Output :
```

```
((21,Pune ),{(4,Preethi,Agarwal,21,9848022330,Pune )})

((21,Hyderabad),{(1,Rajiv,Reddy,21,9848022337,Hyderabad)})

((22,Delhi ),{(3,Rajesh,Khanna,22,9848022339,Delhi )})

((22,Kolkata),{(2,siddarth,Battacharya,22,9848022338,Kolkata)})

((23,Chennai ),{(6,Archana,Mishra,23,9848022335,Chennai )})
```

```
((23,Bhuwaneshwar ),{(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar )})

((24,Chennai),{(8,Bharathi,Nambiayar,24,9848022333,Chennai)})

((24,trivendram ),{(7,Komal,Nayak,24,9848022334,trivendram )})
```

## Group Operator 3

```
group_all = GROUP student1  All;

dump group_all;
```

# Output :

```
(all,{(8,Bharathi,Nambiayar,24,9848022333,Chennai),(7,Komal,Nayak,24,9848022
334,trivendram ),(6,Archana,Mishra,23,9848022335,Chennai
),(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar
),(4,Preethi,Agarwal,21,9848022330,Pune
),(3,Rajesh,Khanna,22,9848022339,Delhi
),(2,siddarth,Battacharya,22,9848022338,Kolkata),(1,Rajiv,Reddy,21,9848022337,
Hyderabad)})
```

## Cogroup

```
hdfs dfs -put employee_details.txt /pig

hdfs dfs -ls /pig/employee_details.txt

hdfs dfs -cat /pig/employee_details.txt
```

```
employee_details = LOAD '/home/student/Documents/Pig/employee_details.txt'
USING PigStorage(',') as (id:int, name:chararray, age:int, city:chararray);

dump employee_details;
```

# Output :

```
(1,Robin,22,newyork)

(2,BOB,23,Kolkata)

(3,Maya,23,Tokyo)
```

```
(4,Sara,25,London)

(5,David,23,Bhuwaneshwar)

(6,Maggy,22,Chennai)
```

```
cogroup_data = COGROUP student1 by age, employee_details by age;

Dump cogroup_data;

# Output :

(21,{(4,Preethi,Agarwal,21,9848022330,Pune ),(1,Rajiv,Reddy,21,9848022337,Hyderabad)},{})

(22,{(3,Rajesh,Khanna,22,9848022339,Delhi ),(2,siddarth,Battacharya,22,9848022338,Kolkata)},{(6,Maggy,22,Chennai),(1,Robin,22,newyork)})

(23,{(6,Archana,Mishra,23,9848022335,Chennai ),(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar )},{(5,David,23,Bhuwaneshwar),(3,Maya,23,Tokyo),(2,BOB,23,Kolkata)})

(24,{(8,Bharathi,Nambiayar,24,9848022333,Chennai),(7,Komal,Nayak,24,9848022334,trivendram )},{})

(25,{},{(4,Sara,25,London)})
```

```
## Join Operator

hdfs dfs -put customers.txt /pig

hdfs dfs -put orders.txt /pig

customers = LOAD '/home/student/Documents/Pig/customers.txt' USING PigStorage(',') as (id:int, name:chararray, age:int, address:chararray, salary:int);

dump customers;

# Output :
```

```
(1,Ramesh,32,Ahmedabad,2000)

(2,Khilan,25,Delhi,1500)

(3,kaushik,23,Kota,2000)

(4,Chaitali,25,Mumbai,6500)

(5,Hardik,27,Bhopal,8500)

(6,Komal,22,MP,4500)

(7,Muffy,24,Indore,10000)
```

```
orders = LOAD '/home/student/Documents/Pig/orders.txt' USING PigStorage(',')
as (oid:int, date:chararray, customer_id:int, amount:int);

dump orders;

# Output :

(102,2009-10-08 00:00:00,3,3000)

(100,2009-10-08 00:00:00,3,1500)

(101,2009-11-20 00:00:00,2,1560)

(103,2008-05-20 00:00:00,4,2060)
```

```
## Self-Join :

customers1 = LOAD '/home/student/Documents/Pig/customers.txt' USING
PigStorage(',') as (id:int,name:chararray, age:int, address:chararray, salary:int);

dump customers1;

# Output :

(1,Ramesh,32,Ahmedabad,2000)

(2,Khilan,25,Delhi,1500)

(3,kaushik,23,Kota,2000)
```

```
(4,Chaitali,25,Mumbai,6500)

(5,Hardik,27,Bhopal,8500)

(6,Komal,22,MP,4500)

(7,Muffy,24,Indore,10000)
```

```
customers2 = LOAD '/home/student/Documents/Pig/customers.txt' USING
PigStorage(',') as (id:int, name:chararray, age:int, address:chararray, salary:int);

# Output :

(1,Ramesh,32,Ahmedabad,2000)

(2,Khilan,25,Delhi,1500)

(3,kaushik,23,Kota,2000)

(4,Chaitali,25,Mumbai,6500)

(5,Hardik,27,Bhopal,8500)

(6,Komal,22,MP,4500)

(7,Muffy,24,Indore,10000)
```

```
customers3 = JOIN customers1 BY id, customers2 BY id;

Dump customers3;

# Output :

(1,Ramesh,32,Ahmedabad,2000,1,Ramesh,32,Ahmedabad,2000)

(2,Khilan,25,Delhi,1500,2,Khilan,25,Delhi,1500)

(3,kaushik,23,Kota,2000,3,kaushik,23,Kota,2000)

(4,Chaitali,25,Mumbai,6500,4,Chaitali,25,Mumbai,6500)

(5,Hardik,27,Bhopal,8500,5,Hardik,27,Bhopal,8500)

(6,Komal,22,MP,4500,6,Komal,22,MP,4500)
```

```
(7,Muffy,24,Indore,10000,7,Muffy,24,Indore,10000)
```

```
## Inner Join :
customer_orders = JOIN customers BY id, orders BY customer_id;
Dump customer_orders;
# Output :
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
```

```
## Outer Joins
outer_left = JOIN customers BY id LEFT OUTER, orders BY customer_id;
Dump outer_left;
# Output :
(1,Ramesh,32,Ahmedabad,2000,,,,)
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
(5,Hardik,27,Bhopal,8500,,,,)
(6,Komal,22,MP,4500,,,,)
(7,Muffy,24,Indore,10000,,,,)
```

```
outer_right = JOIN customers BY id RIGHT, orders BY customer_id;

Dump outer_right;

# Output :

(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)

(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)

(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)

(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
```

```
outer_full = JOIN customers BY id FULL OUTER, orders BY customer_id;

Dump outer_full;

# Output :

(1,Ramesh,32,Ahmedabad,2000,,,,)

(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)

(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)

(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)

(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)

(5,Hardik,27,Bhopal,8500,,,,)

(6,Komal,22,MP,4500,,,,)

(7,Muffy,24,Indore,10000,,,,)
```

```
## Joins by Multiple-Keys

employee = LOAD 'hdfs://master:8020/pig/employee.txt' USING PigStorage(',')

as (id:int, firstname:chararray, lastname:chararray, age:int, designation:chararray,
jobid:int);
```

```
employee_contact = LOAD 'hdfs://master:8020/pig/employee_contact.txt' USING
PigStorage(',')as (id:int, phone:chararray, email:chararray, city:chararray, jobid:int);
```

```
## Cross operator

cross_data = CROSS customers, orders;

Dump cross_data;

# Output :
```

```
(7,Muffy,24,Indore,10000,103,2008-05-20 00:00:00,4,2060)
(7,Muffy,24,Indore,10000,101,2009-11-20 00:00:00,2,1560)
(7,Muffy,24,Indore,10000,100,2009-10-08 00:00:00,3,1500)
(7,Muffy,24,Indore,10000,102,2009-10-08 00:00:00,3,3000)
(6,Komal,22,MP,4500,103,2008-05-20 00:00:00,4,2060)
(6,Komal,22,MP,4500,101,2009-11-20 00:00:00,2,1560)
(6,Komal,22,MP,4500,100,2009-10-08 00:00:00,3,1500)
(6,Komal,22,MP,4500,102,2009-10-08 00:00:00,3,3000)
(5,Hardik,27,Bhopal,8500,103,2008-05-20 00:00:00,4,2060)
(5,Hardik,27,Bhopal,8500,101,2009-11-20 00:00:00,2,1560)
(5,Hardik,27,Bhopal,8500,100,2009-10-08 00:00:00,3,1500)
(5,Hardik,27,Bhopal,8500,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
(4,Chaitali,25,Mumbai,6500,101,2009-11-20 00:00:00,2,1560)
(4,Chaitali,25,Mumbai,6500,100,2009-10-08 00:00:00,3,1500)
(4,Chaitali,25,Mumbai,6500,102,2009-10-08 00:00:00,3,3000)
(3,kaushik,23,Kota,2000,103,2008-05-20 00:00:00,4,2060)
(3,kaushik,23,Kota,2000,101,2009-11-20 00:00:00,2,1560)
```

```
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)

(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)

(2,Khilan,25,Delhi,1500,103,2008-05-20 00:00:00,4,2060)

(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)

(2,Khilan,25,Delhi,1500,100,2009-10-08 00:00:00,3,1500)

(2,Khilan,25,Delhi,1500,102,2009-10-08 00:00:00,3,3000)

(1,Ramesh,32,Ahmedabad,2000,103,2008-05-20 00:00:00,4,2060)

(1,Ramesh,32,Ahmedabad,2000,101,2009-11-20 00:00:00,2,1560)

(1,Ramesh,32,Ahmedabad,2000,100,2009-10-08 00:00:00,3,1500)

(1,Ramesh,32,Ahmedabad,2000,102,2009-10-08 00:00:00,3,3000)
```

## Union Operators

```
student2 = LOAD '/home/student/Documents/Pig/student_details.txt' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int,phone:chararray, city:chararray);


dump student2;
```

# Output :

```
(1,Rajiv,Reddy,21,9848022337,Hyderabad)

(2,siddarth,Battacharya,22,9848022338,Kolkata)

(3,Rajesh,Khanna,22,9848022339,Delhi )

(4,Preethi,Agarwal,21,9848022330,Pune )

(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar )

(6,Archana,Mishra,23,9848022335,Chennai )

(7,Komal,Nayak,24,9848022334,trivendram )
```

```
(8,Bharathi,Nambiayar,24,9848022333,Chennai)
```

```
student3 = LOAD '/home/student/Documents/Pig/student_details1.txt' USING
PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, age:int,
phone:chararray, city:chararray);
```

# Output :

```
(1,Rajiv,Reddy,21,9848022337,Hyderabad)

(2,siddarth,Battacharya,22,9848022338,Kolkata)

(7,Komal,Nayak,24,9848022334,trivendram )

(8,Bharathi,Nambiayar,24,9848022333,Chennai)
```

```
student = UNION student2, student3;

Dump student;
```

# Output :

```
(1,Rajiv,Reddy,21,9848022337,Hyderabad)

(2,siddarth,Battacharya,22,9848022338,Kolkata)

(3,Rajesh,Khanna,22,9848022339,Delhi )

(4,Preethi,Agarwal,21,9848022330,Pune )

(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar )

(6,Archana,Mishra,23,9848022335,Chennai )

(7,Komal,Nayak,24,9848022334,trivendram )

(8,Bharathi,Nambiayar,24,9848022333,Chennai)

(1,Rajiv,Reddy,21,9848022337,Hyderabad)

(2,siddarth,Battacharya,22,9848022338,Kolkata)

(7,Komal,Nayak,24,9848022334,trivendram )
```

```
(8,Bharathi,Nambiayar,24,9848022333,Chennai)
```

## Split Command

SPLIT student into student_details if age<23, student_details1 if (21<age and age>23);

Dump student_details;

# Output :

```
(1,Rajiv,Reddy,21,9848022337,Hyderabad)
(2,siddarth,Battacharya,22,9848022338,Kolkata)
(3,Rajesh,Khanna,22,9848022339,Delhi )
(4,Preethi,Agarwal,21,9848022330,Pune )
(1,Rajiv,Reddy,21,9848022337,Hyderabad)
(2,siddarth,Battacharya,22,9848022338,Kolkata)
```

Dump student_details1;

# Output :

```
(7,Komal,Nayak,24,9848022334,trivendram )
(8,Bharathi,Nambiayar,24,9848022333,Chennai)
(7,Komal,Nayak,24,9848022334,trivendram )
(8,Bharathi,Nambiayar,24,9848022333,Chennai)
```