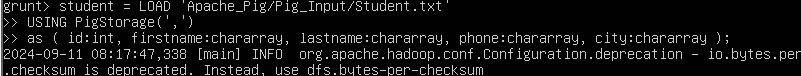
Apache Tutorials

# 1. Apache Pig - Reading Data

**- Chạy trên local**

|  |
| --- |
| grunt> student = LOAD 'Apache\_Pig/Pig\_Input/Student.txt'  USING PigStorage(',')  as ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray ); |

****

# 2. Apache Pig - Storing Data

**- Chạy trên local**

|  |
| --- |
| grunt> STORE student INTO ' Apache\_Pig/Pig\_Output/ ' USING PigStorage (','); |

****

**A screenshot of a computer screen

Description automatically generated**

**- Kiểm tra**

A screen shot of a computer

Description automatically generated

# 3. Apache Pig - Diagnostic Operators

|  |
| --- |
| grunt> Dump student |

A screenshot of a computer program

Description automatically generated

# 4. Apache Pig - Describe Operator

|  |
| --- |
| grunt> describe student; |

A screen shot of a computer

Description automatically generated

# 5. Apache Pig - Explain Operator

|  |
| --- |
| grunt> explain student; |

A screenshot of a computer

Description automatically generated

# 6. Apache Pig - Illustrate Operator

|  |
| --- |
| grunt> illustrate student; |

A screenshot of a computer

Description automatically generated

# 7. Apache Pig - Group Operator

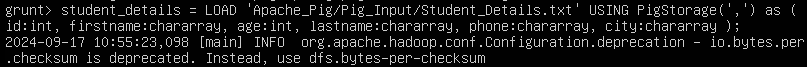
Tạo file Student\_Details.txt chứa nội dung sauls



A screen shot of a computer

Description automatically generated

|  |
| --- |
| grunt> student\_details = LOAD 'Apache\_Pig/Pig\_Input/Student\_Details.txt'  USING PigStorage(',')  as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray ); |
| grunt> group\_data = GROUP student\_details BY age; |
| grunt> Dump group\_data |
| grunt> descirbe group\_data |
| grunt> illustrate group\_data |





A screenshot of a computer

Description automatically generated

|  |
| --- |
| Group\_multiple = GROUP student\_details by (age,city); |
| Dump group\_multiple; |



A screenshot of a computer

Description automatically generated

|  |
| --- |
| group\_all = GROUP student\_details All; |
| dump group\_all |

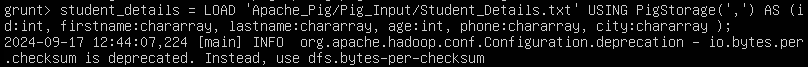


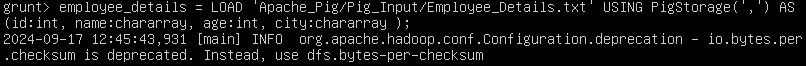
A computer screen with a black screen and white text

Description automatically generated

# 8. Apache Pig - Cogroup Operator

|  |
| --- |
| grunt> student\_detail= LOAD 'Apache\_Pig/Pig\_Input/Student\_Details.txt'  USING PigStorage(',')  as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray ); |
| grunt> employee\_details = LOAD 'Apache\_Pig/Pig\_Input/Employee\_Details.txt'  USING PigStorage(',')  as ( id:int, name:chararray, age:int, city:chararray ); |
| cogroup\_data = COGROUP student\_details by age, employee\_details by age; |
| dump cogroup\_data; |





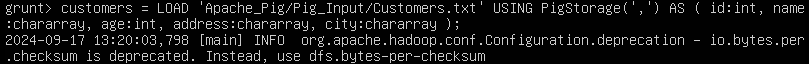


A screenshot of a computer

Description automatically generated

# 9. Apache Pig – Join Operator

|  |
| --- |
| grunt> customers = LOAD 'Apache\_Pig/Pig\_Input/Customers.txt'  USING PigStorage(',')  as ( id:int, name:chararray, age:int, address:chararray, city:chararray ); |
| grunt> orders = LOAD 'Apache\_Pig/Pig\_Input/Orders.txt'  USING PigStorage(',')  as ( oid:int, date:chararray, customer\_id:int, amount:int ); |





## 9.1 Self-join

|  |
| --- |
| grunt> customers1 = LOAD 'Apache\_Pig/Pig\_Input/Customers.txt' USING PigStorage(',')  as ( id:int, name:chararray, age:int, address:chararray, salary:chararray ); |
| grunt> customers2 = LOAD 'Apache\_Pig/Pig\_Input/Customers.txt' USING PigStorage(',')  as ( id:int, name:chararray, age:int, address:chararray, salary:chararray ); |
| grunt> customers3 = JOIN customers1 BY id, customers2 BY id; |
| grunt> dump customers3; |

A screen shot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

## 9.2 Inner Join

|  |
| --- |
| grunt> customer\_order = JOIN customers BY id, orders BY customer\_id; |
| grunt> dump customer\_order; |



A screenshot of a computer

Description automatically generated

## 9.3 Left Outer Join

|  |
| --- |
| grunt> customer\_order\_LO = JOIN customers BY id LEFT OUTER, orders BY customer\_id; |
| grunt> dump customer\_order\_LO; |



A screenshot of a computer

Description automatically generated

## 9.4 Right Outer Join

|  |
| --- |
| grunt> customer\_order\_R = JOIN customers BY id RIGHT, orders BY customer\_id; |
| grunt> dump customer\_order\_R; |



A screenshot of a computer

Description automatically generated

## 9.5 Full Outer Join

|  |
| --- |
| grunt> customer\_order\_FO = JOIN customers BY id FULL OUTER, orders BY customer\_id; |
| grunt> dump customer\_order\_FO; |



A screenshot of a computer

Description automatically generated

# 10. Apache Pig – Cross Operator

|  |
| --- |
| grunt> customers = LOAD 'Apache\_Pig/Pig\_Input/Customers.txt' USING PigStorage(',')  as ( id:int, name:chararray, age:int, address:chararray, salary:int ); |
| grunt> orders = LOAD 'Apache\_Pig/Pig\_Input/Orders.txt' USING PigStorage(',')  as ( oid:int, date:chararray, customer\_id:int, amount:int ); |
| grunt> cross\_data = CROSS customers, orders; |
| grunt> dump cross\_data; |

A screenshot of a computer

Description automatically generated

# 11. Apache Pig – Union Operator

|  |
| --- |
| grunt> students\_data1 = LOAD 'Apache\_Pig/Pig\_Input/Students\_Data1.txt' USING PigStorage(',') AS ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray ); |
| grunt> students\_data2 = LOAD 'Apache\_Pig/Pig\_Input/Students\_Data2.txt' USING PigStorage(',') AS ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray ); |
| grunt> students\_union = UNION students\_data, students\_data2; |
| grunt> dump students\_union |

A screen shot of a computer

Description automatically generated



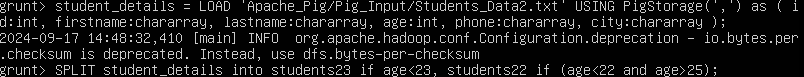
A screenshot of a computer screen

Description automatically generated

# 12. Apache Pig – Split Operator

|  |
| --- |
| grunt> student\_details = LOAD 'Apache\_Pig/Pig\_Input/Student\_Details.txt'  USING PigStorage(',') AS ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray, city:chararray ); |
| grunt> SPLIT student\_details into student\_details23 if age<23, student\_details22 if age<22 and age>25); |
| grunt> student\_details |





A screenshot of a computer

Description automatically generated

# 13. Apache Pig – Filter Operator

|  |
| --- |
| grunt> filter\_data = FILTER student\_details BY city == 'Hyderabad'; |
| grunt> dump filter\_data; |



A screenshot of a computer

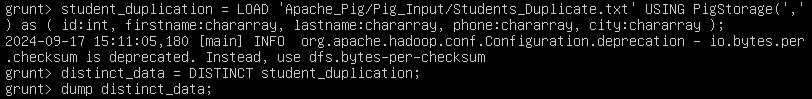
Description automatically generated

# 14. Apache Pig – Distinct Operator

A screenshot of a computer

Description automatically generated

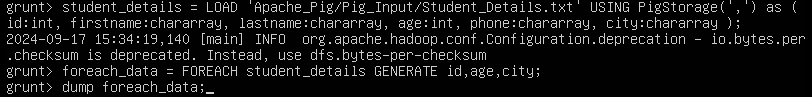
|  |
| --- |
| grunt> student\_duplication = LOAD 'Apache\_Pig/Pig\_Input/Students\_ Duplicate.txt'  USING PigStorage(',') AS ( id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray ); |
| grunt> distinct\_data = DISTINCT student\_duplication; |
| grunt> dump distinct\_data; |



A screenshot of a computer

Description automatically generated

# 15. Apache Pig - Foreach Operator



A screenshot of a computer

Description automatically generated

# 16. Apache Pig - Order By



A computer screen shot of a black and white computer screen

Description automatically generated

# 17. Apache Pig - Limit Operator



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# 17. Apache Pig – User Defined Function

Tạo fie chứa dữ liệu và file đó và file jar pig

A screenshot of a computer

Description automatically generated

|  |
| --- |
| grunt> REGISTER /home/hadoopngocphung/Apache\_Pig/Pig\_Input/sample\_udf.jar; |
| grunt> DEFINE sample\_eval mypack.Sample\_Eval(); |
| grunt> emp\_data = LOAD '/home/hadoopngocphung/Apache\_Pig/Pig\_Input/Data\_JAR\_Ex.txt USING PigStorage (',') AS (id:int, name: chararray, age:int, city:chararray); |
| grunt> Upper\_case = FOREACH emp\_data GENERATE sample\_eval (name); |
| grunt> dump Upper\_case; |

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated