**Pig and PigLatin**

**What is Pig in Hadoop?**

Pig is a scripting platform that runs on Hadoop clusters designed to process and analyze large datasets. Pig is extensible, self-optimizing, and easily programmed.

Programmers can use Pig to write data transformations without knowing Java. Pig uses both structured and unstructured data as input to perform analytics and uses HDFS to store the results.

**Components of Pig**

There are two major components of the Pig:

● Pig Latin script language

● A runtime engine

**Pig Latin script language:**

The Pig Latin script is a procedural data flow language. It contains syntax and commands that can be applied to implement business logic. Examples of Pig Latin are LOAD and STORE.

**A runtime engine:**

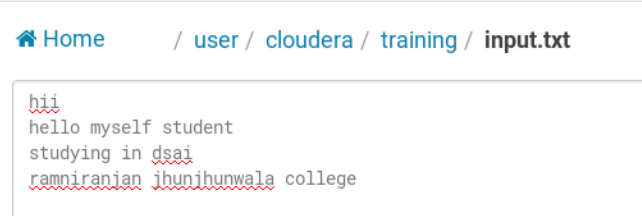
The runtime engine is a compiler that produces sequences of MapReduce programs. It uses HDFS to store and retrieve data. It is also used to interact with the Hadoop system (HDFS and MapReduce).

The runtime engine parses, validates, and compiles the script operations into a sequence of MapReduce jobs.

Pig and Pig Latin

Ref: <https://docs.cloudera.com/HDPDocuments/HDP1/HDP-Win-1.3.0/ds_Pig/basic.html#Data+Types+and+More>

1. Open cloudera browser
2. Locate Hue and get logged in with username cloudera and password cloudera.
3. Locate file Browser and check for directory user/cloudera.
4. Create a new directory named ‘training’.
5. Create the pig directory inside the training directory.
6. Create an input.txt file in the pig directory.
7. Add some text into the input.txt and save the file.



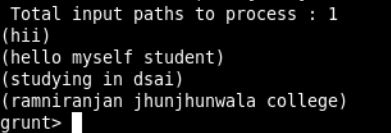
1. Open terminal and start pig tool by executing pig command.
2. Load input.txt file in input variable.

Input = LOAD ‘/usr/cloudera/Training/pig/input.txt’ AS (f1:chararray);



1. Display the contents of the input variable.

DUMP input

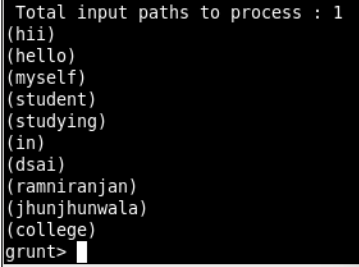


1. Tokenize the text that is stored in variable input.

wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;

dump wordsInEachLine;



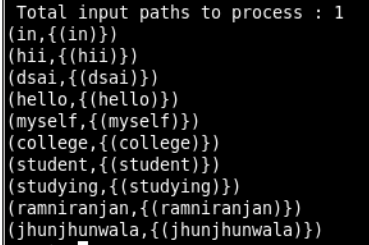


1. Group all similar words.

groupedWords = group wordsInEachLine by word;

dump groupedWords;



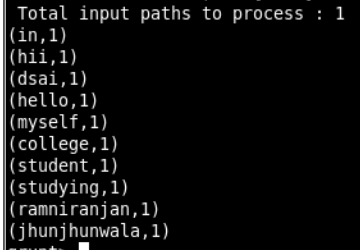


1. Count the number of occurrences of each word.

countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);

dump countedWords;





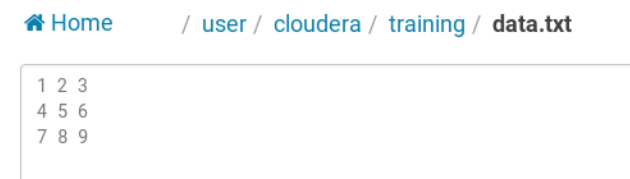
Ex 2 : Loading matrix in a pig variable.

Mydata.txt

1 2 3

4 5 6

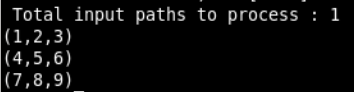
7 8 9



A=LOAD ‘/user/cloudera/Training/pig/mydata.txt’ AS (c1:int,c2:int,c3:int);

dump A;





**Creating pig script**

/\*

Wordcountex.pig

Counting the occurrences of word

\*/

–Execute this script in mapreduce mode

Input = LOAD ‘/usr/cloudera/Training/pig/input.txt’ AS (f1:chararray);

wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;

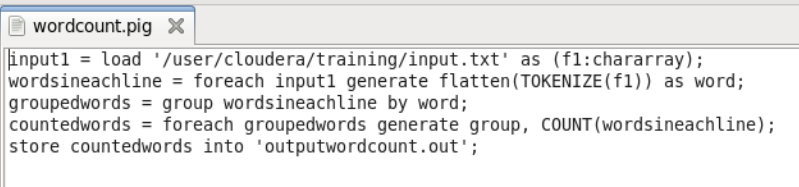
groupedWords = group wordsInEachLine by word;

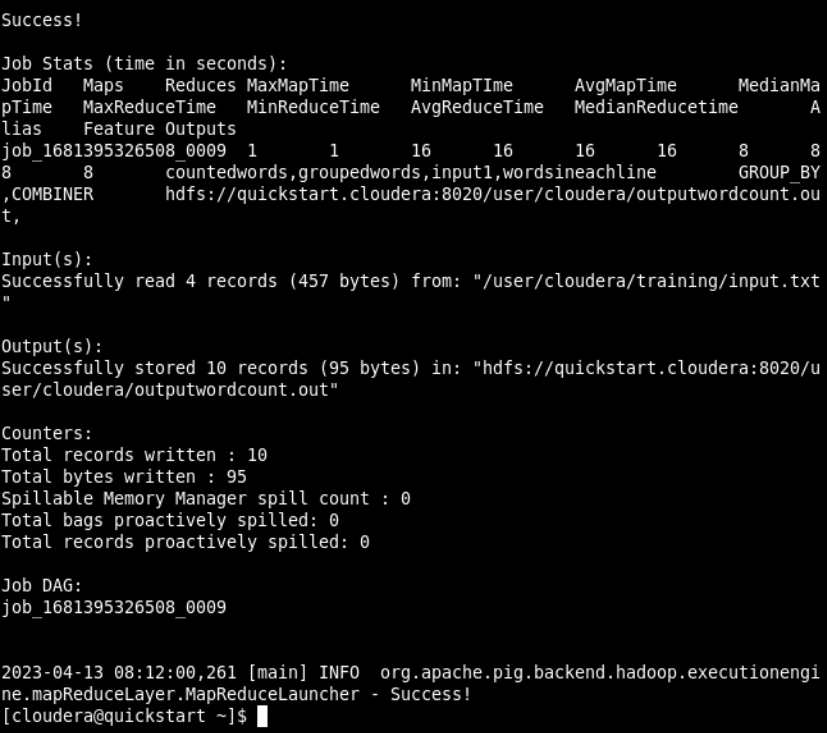
countedWords = foreach groupedWords generate group, COUNT(wordsInEachLine);

store countedWords into ‘outputwordcount.out’;

~~dump countedWords;~~**Running .pig script**

****

****

****

$pig -x local Wordcountex.pig

Or

$pig Wordcountex.pig

or

$pig -x mapreduce Wordcountex.pig

Example 2: Creating pig script

1. Create file named Mydata.txt

1 2 3

4 5 6

7 8 9

1. Create a matA.pig file with the following commands.

A=LOAD ‘/user/cloudera/Training/pig/mydata.txt’ AS (c1:int,c2:int,c3:int);

dump A;

Or

$pig -x local

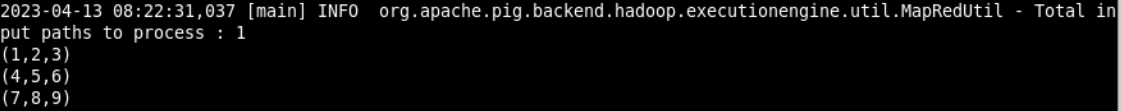
A=LOAD ‘/home/cloudera/Desktop/Data/mydata.txt’ AS (c1:int,c2:int,c3:int);

dump A;

A=LOAD ‘/home/cloudera/Desktop/Data/mydata.txt’ USING PigStorage (‘ ‘) AS (c1:int,c2:int,c3:int);

dump A;

pig mata.pig



1. Execute matA.pig file in mapreduce mode.

$pig -x mapreduce matA.pig



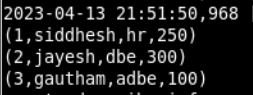
Example 3: Creating pig script

1. Create a pig script for the following tasks:
   1. Load employees data into emp variable. Also specify the proper schema for the same.

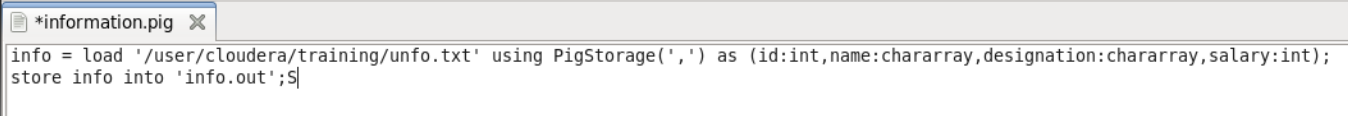
Ex:

empID, empName,designation,experience,salary





* 1. Send or store emp data in empout.out file.



* 1. Check the contents of the empout.out file.



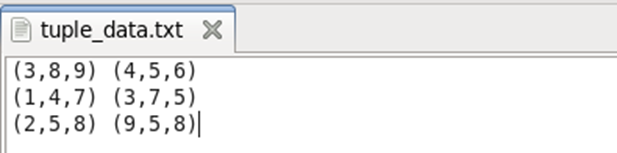
Example 4: Creating pig script and loading tuple data.

1. Create a relation named ‘tupledata’ and enter the following data.

(3,8,9) (4,5,6)

(1,4,7) (3,7,5)

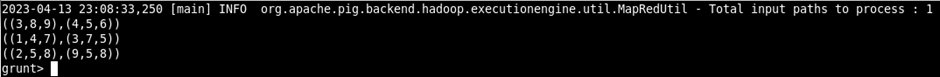
(2,5,8) (9,5,8)



* 1. Load tupledata into a pig variable named ‘inputtuple’.

inputtuple = LOAD 'tupledata' AS (t1:tuple(t1a:int, t1b:int,t1c:int),t2:tuple(t2a:int,t2b:int,t2c:int));

DUMP inputtuple;

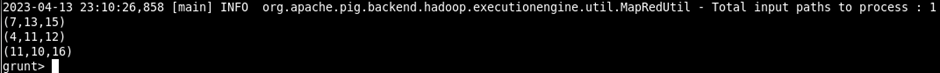


1. Add the columns with a similar index of each tuple and store the result in addout.out file.

//try following cmd

addition = FOREACH A GENERATE t1a+t2a,t1b+t2b,t1c+t2c;

store addition into ‘addout.out’;



1. Display the content of addout.out file.

(7,13,15)

(4,11,12)

(11,10,16)

9/3/2023

A = LOAD 'data1' AS (a1:int,a2:int,a3:int);

DUMP A;

(1,2,3)

(4,2,1)

B = LOAD 'data2' AS (b1:int,b2:int);

DUMP B;

(2,4)

(8,9)

(1,3)

X = CROSS A, B;

DUMP X;

(1,2,3,2,4)

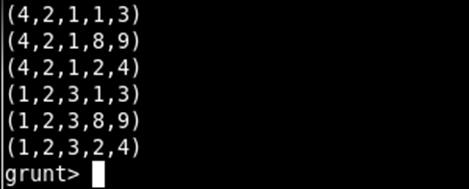
(1,2,3,8,9)

(1,2,3,1,3)

(4,2,1,2,4)

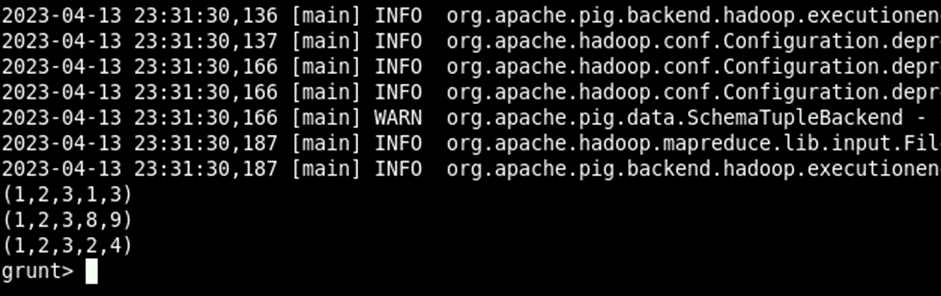
(4,2,1,8,9)

(4,2,1,1,3)



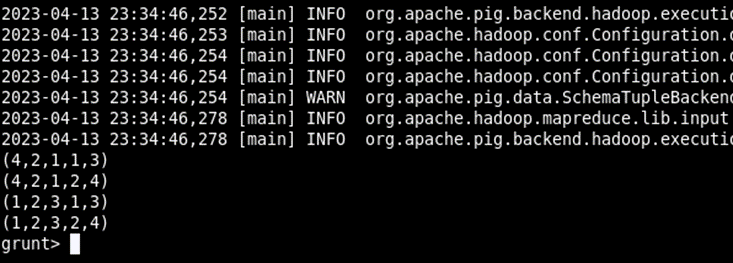
d= DISTINCT A;

f1 = FILTER X BY (f1 == 1);



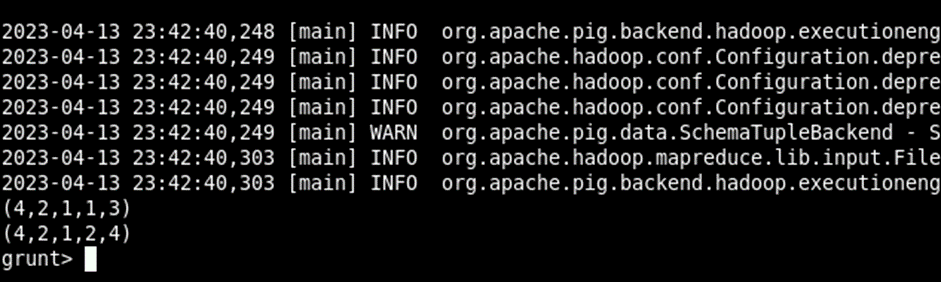
Q. Filter all rows for 2nd column val+4th column val=5th column val

f1 = FILTER X BY (f2+f4 == f5));

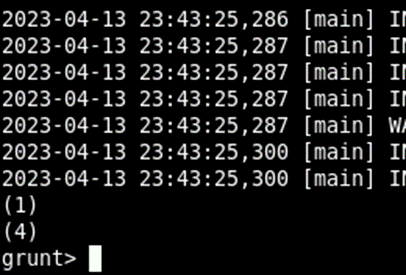


Q. Filter all rows for 2nd column val+4th column val=5th column val and also check for 1st column val as 4

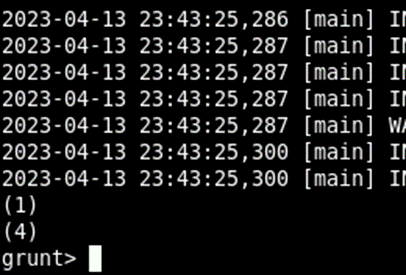
f1 = FILTER X BY (f2+f4 == f5) AND (f1==4);



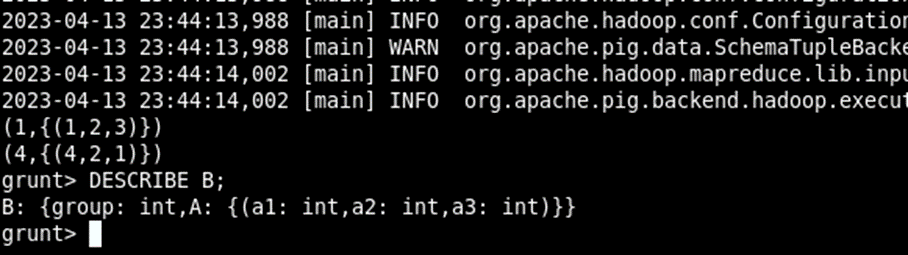
XX = FOREACH A GENERATE f1;



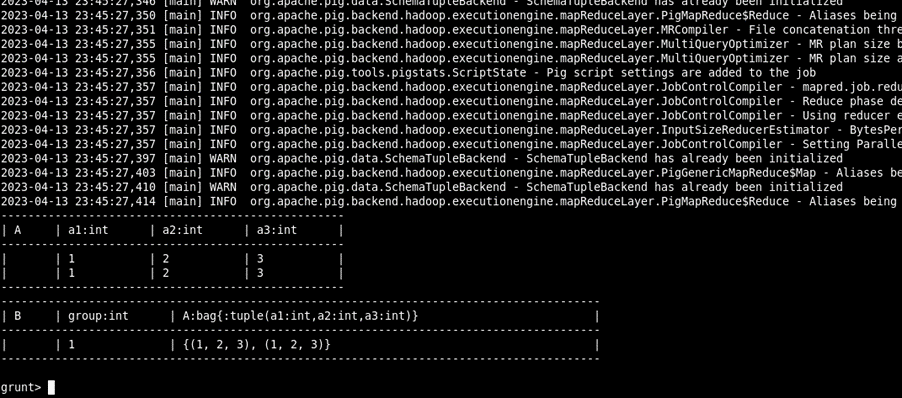
B = GROUP A BY f1



DESCRIBE B



ILLUSTRATE B



CREATE FILE TO STORE THE FOLLOWING DATA in stud.txt

1 AMIT 45 50 56 C

2 SUMIT 44 46 67 C

3 jOHN 45 56 67 D

Load

Students data into a pig relation named ‘studInfo’.

Display the list of students rollno and name.

Display the total marks of each student.Display distinct names of students

Check if the total marks of the sumit is 157 or not.

Display distaint marks of subject1.

20/3/2023

1. Create the file named ‘data’ for the following data.

1,2,3

4,2,1

8,3,4

4,3,3

7,2,5

8,4,3

1. Load the contents of ‘data’ into relation A.

A = LOAD 'data' AS (a1:int,a2:int,a3:int);

DUMP A;

(1,2,3)

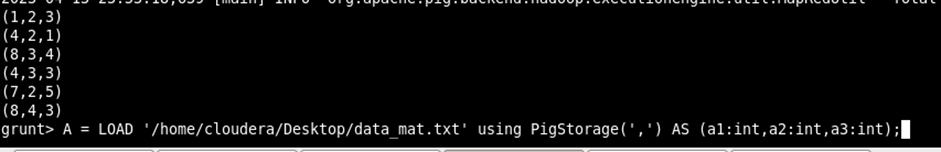
(4,2,1)

(8,3,4)

(4,3,3)

(7,2,5)

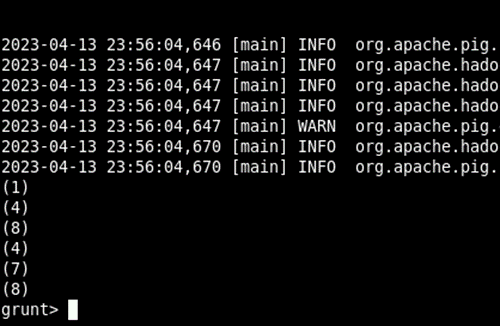
(8,4,3)



1. Display the elements of the first column.

X = FOREACH A GENERATE f1;

DUMP X



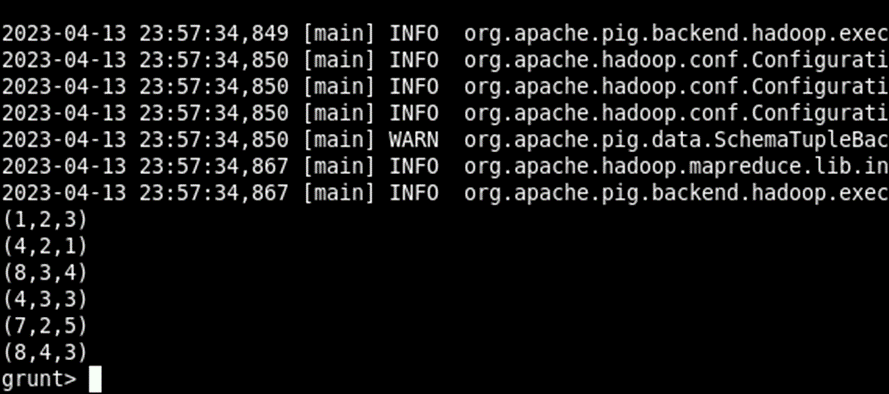
X = FOREACH A GENERATE a1;

DUMP X

1. Display all tuples of relation A.

X = FOREACH A GENERATE \*;

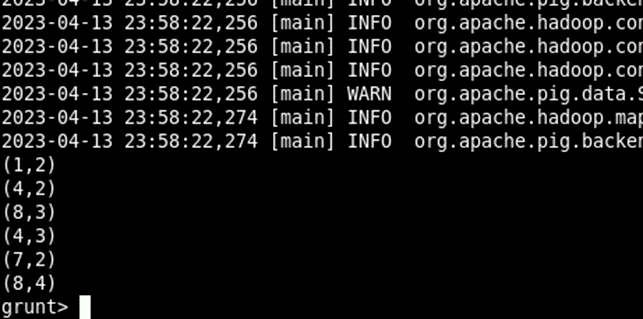
DUMP X;



1. Display the first two columns of relation A.

X = FOREACH A GENERATE a1,a2;

DUMP X;



1. Create a file named ‘data1’ to store the following matrix.

1

4

8

1. Load data of ‘data1’ into relation C.
2. Sum column a1

X = FOREACH A GENERATE group, SUM (A.a1);

DUMP X;

1. Nested projection:

X = FOREACH A GENERATE group, A.a1;

DUMP X;

1,{1}

2,{3}

2,{4}

Reference

<https://docs.cloudera.com/HDPDocuments/HDP1/HDP-Win-1.3.0/ds_Pig/basic.html#data-types>

.