Pig 資料分析工具

1. Programming Pig
http://chimera.labs.oreilly.com/books/1234000001811/index.html

2. Apache Pig Tutorial – Part 1
http://www.rohitmenon.com/index.php/apache-pig-tutorial-part-1/

3. Apache Pig Tutorial – Part 2
http://www.rohitmenon.com/index.php/apache-pig-tutorial-part-2/

資料科學家上工

資料科學家登入 Hadoop Client

```
$ ssh dsa01@cla01
dsa01@cla01's password: dsa01
Welcome to Ubuntu 14.04.5 LTS (GNU/Linux 4.4.0-31-generic x86_64)

 * Documentation:  https://help.ubuntu.com/
```

下載美國電影資料集 (Dataset)

$ wget https://raw.githubusercontent.com/rohitsden/pig-tutorial/master/movies_data.csv

$ head -n 10 movies_data.csv

1,The Nightmare Before Christmas,1993,3.9,4568
2,The Mummy,1932,3.5,4388
3,Orphans of the Storm,1921,3.2,9062
4,The Object of Beauty,1991,2.8,6150
5,Night Tide,1963,2.8,5126
6,One Magic Christmas,1985,3.8,5333

啟動 Pig 分析工具, 並上載 Dataset

```
$ pig
grunt> pwd
hdfs://nna:8020/user/dsa01

grunt> copyfromlocal movies_data.csv   .

grunt> ls
hdfs://nna:8020/user/dsa01/movies_data.csv<r 2>          2893177
hdfs://nna:8020/user/dsa01/school.txt<r 2>       20807

grunt> quit
```

Pig Latin 首部曲

兒童黑話（Pig Latin）是一種英語語言遊戲，形式是在英語上加上一點規則使發音改變。據說是由在德國的英國戰俘發明來瞞混德軍守衛的。兒童黑話於1950年代和1960年代在英國利物浦達到顛峰，各種年紀和職業的人都有使用。兒童黑話多半被兒童用來瞞著大人秘密溝通，有時則只是說著好玩。雖然是起源於英語的遊戲，但是規則適用很多其他語言。

Pig Latin 基本資料型態

Int: An integer. Ints are represented in interfaces by java.lang.Integer. They store a four byte signed integer. Constant integers are expressed as integer numbers, **for example 12**.

Long: A long integer. Long are represented in interfaces by java.lang.Long. They store a eight byte signed integer. Constants are expressed as integer numbers with a L appended, **for example 34L**.

Float: A floating point number. Floats are represented in interfaces by java.lang.Float. They store a four byte floating point number. Constants are represented as floating point numbers with f appended, **for example, 2.18f**.

Double: A double precision floating point number. Doubles are represented in interfaces by java.lang.Double. They store a eight byte floating point number. Constants are represented either as floating point numbers or in exponent notation, **for example, 32.12567 or 3e-17.**

Chararray: A string or array of characters. Represented in interfaces by java.lang.String. Constant chararrays are represented by single quotes, for example, 'constant chararray'.

Bytearray: A blob or array of bytes. Represented by java class DataByteArray which wraps a java byte[]. There is no way to specify a bytearray constant.

Pig 命令類型

Pig 所使用的指令稱為 Pig Latin Statements，執行可以簡單分成三個步驟

1. 使用 LOAD 讀取資料
2. 一連串操作資料的指令
3. 使用 DUMP 來看結果或用 STORE 把結果存起來。如果不執行 DUMP 或 STORE 是不會產生任何 MapReduce job 的

可再細分指令的類型

讀取 ： LOAD
儲存 ： STORE
資料處理 ： FILTER, FOREACH, GROUP, COGROUP, inner JOIN, outer JOIN, UNION, SPLIT, …
彙總運算 ： AVG, COUNT, MAX, MIN, SIZE, …
數學運算 ： ABS, RANDOM, ROUND, …
字串處理 ： INDEXOF, SUBSTRING, REGEX EXTRACT, …
Debug：DUMP, DESCRIBE, EXPLAIN, ILLUSTRATE
HDFS 或本機的檔案操作 ： cat, ls, cp, mkdir, copyfromlocal, copyToLocal, ……

Pig Latin 命令(一)

```
$ pig
grunt> movies = LOAD 'movies_data.csv' USING
PigStorage(',') as (id,name,year,rating,duration);

grunt> describe movies;
movies: {id: bytearray,name: bytearray,year: bytearray,rating:
bytearray,duration: bytearray}


grunt> movies_greater_than_four = FILTER movies BY
(float)rating>4.0;

grunt> dump movies_greater_than_four;
                    ::
(49546,Bo Burnham: what.,2013,4.1,3614)
(49549,Life With Boys: Season 1,2011,4.1,)
(49554,Max Steel,2013,4.1,)
(49556,Lilyhammer: Season 1 (Recap),2013,4.2,194)
(49571,The Short Game (Trailer),2013,4.1,156)
(49579,Transformers Prime Beast Hunters: Predacons Rising,2013,4.2,3950)
```

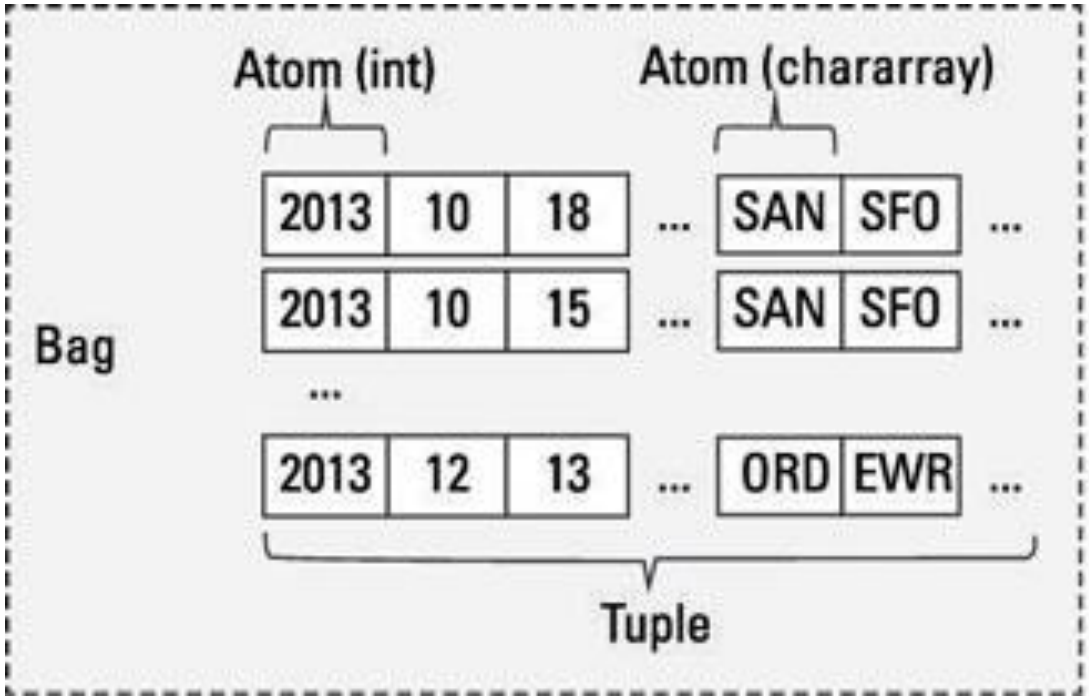
Pig Latin 命令(二)

```
grunt> store movies_greater_than_four into
'movies_greater_than_four.csv';
                    ::
Output(s):
Successfully stored 897 records (35853 bytes) in:
"hdfs://nna:8020/user/dsa01/movies_greater_than_four.csv"

grunt> ls
hdfs://nna:8020/user/dsa01/movies_data.csv<r 2> 2893177
hdfs://nna:8020/user/dsa01/movies_greater_than_four.csv <dir>
hdfs://nna:8020/user/dsa01/pigdata.txt<r 2>     324

grunt> cat movies_greater_than_four.csv;
                    ::
49554   Max Steel       2013    4.1
49556   Lilyhammer: Season 1 (Recap)    2013    4.2     194
49571   The Short Game (Trailer)        2013    4.1     156
49579   Transformers Prime Beast Hunters: Predacons Rising      2013    4.2     3950
```

實作練習

如何將 **HDFS** 檔案系統中的 **movies_greater_than_four.csv** 下載至資料科學家的工作主機

Pig Latin 複雜資料型態



網址：http://morebigdata.blogspot.in/2012/09/pignalytics-pigs-eat-anything-reading.html

取出 10 筆 Tuple 資料

```
grunt> movies = LOAD 'movies_data.csv' USING
PigStorage(',') as
(id:int,name:chararray,year:int,rating:float,duration:int);

grunt> ten = limit movies 10;

grunt> dump ten;
(1,The Nightmare Before Christmas,1993,3.9,4568)
(2,The Mummy,1932,3.5,4388)
(3,Orphans of the Storm,1921,3.2,9062)
(4,The Object of Beauty,1991,2.8,6150)
(5,Night Tide,1963,2.8,5126)
(6,One Magic Christmas,1985,3.8,5333)
(7,Muriel's Wedding,1994,3.5,6323)
(8,Mother's Boys,1994,3.4,5733)
(9,Nosferatu: Original Version,1929,3.5,5651)
(10,Nick of Time,1995,3.4,5333)
```

轉換 Tuple 資料格式

```
grunt> ten_trans = foreach ten generate name,year,duration;

grunt> dump ten_trans;
(The Nightmare Before Christmas,1993,4568)
(The Mummy,1932,4388)
(Orphans of the Storm,1921,9062)
(The Object of Beauty,1991,6150)
(Night Tide,1963,5126)
(One Magic Christmas,1985,5333)
(Muriel's Wedding,1994,6323)
(Mother's Boys,1994,5733)
(Nosferatu: Original Version,1929,5651)
(Nick of Time,1995,5333)
```

轉換 Tuple 資料為 Bag 格式

```
grunt> ten_group = group ten by year;

grunt> dump ten_group;
(1921,{(3,Orphans of the Storm,1921,3.2,9062)})
(1929,{(9,Nosferatu: Original Version,1929,3.5,5651)})
(1932,{(2,The Mummy,1932,3.5,4388)})
(1963,{(5,Night Tide,1963,2.8,5126)})
(1985,{(6,One Magic Christmas,1985,3.8,5333)})
(1991,{(4,The Object of Beauty,1991,2.8,6150)})
(1993,{(1,The Nightmare Before Christmas,1993,3.9,4568)})
(1994,{(8,Mother's Boys,1994,3.4,5733),(7,Muriel's Wedding,1994,3.5,6323)})
(1995,{(10,Nick of Time,1995,3.4,5333)})
```

儲存 Bag 資料

```
grunt> store ten_group into 'ten_group.csv';
```

ten_group.csv 資料是以 Tab 作為欄位分隔字元

```
grunt> cat ten_group.csv
1921    {(3,Orphans of the Storm,1921,3.2,9062)}
1929    {(9,Nosferatu: Original Version,1929,3.5,5651)}
1932    {(2,The Mummy,1932,3.5,4388)}
1963    {(5,Night Tide,1963,2.8,5126)}
1985    {(6,One Magic Christmas,1985,3.8,5333)}
1991    {(4,The Object of Beauty,1991,2.8,6150)}
1993    {(1,The Nightmare Before Christmas,1993,3.9,4568)}
1994    {(8,Mother's Boys,1994,3.4,5733),(7,Muriel's Wedding,1994,3.5,6323)}
1995    {(10,Nick of Time,1995,3.4,5333)}

grunt> quit;
```

問題討論

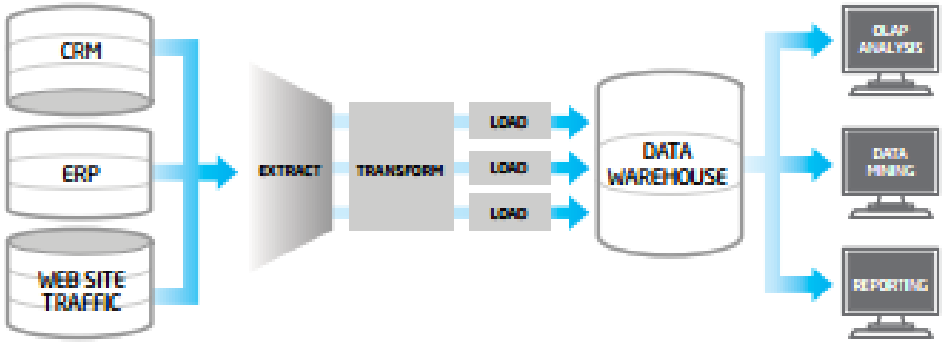
如儲存 ten_group.csv 以 "," 作為分隔字元, 後續如再讀入此檔案內容, 是無法使用 "," 作為分隔字元, 因 bag 資料中有 "," 字元 詳細內容請看備註

group by 命令應用 - Word count

```
$ nano wordcount.pig
A = LOAD 'input.txt';
B = FOREACH A GENERATE flatten(TOKENIZE((chararray)$0)) AS word;
C = GROUP B BY word;
D = FOREACH C GENERATE group, COUNT(B);
STORE D INTO './output.txt';

$ pig -f wordcount.pig
```

Pig Latin 二部曲

ETL (Extract-Transform-Load)



Pig Latin 資料抽取（一）

1. 重新定義 Schema

```
grunt> movies = LOAD 'movies_data.csv' USING
PigStorage(',') as
(id:int,name:chararray,year:int,rating:double,duration:int);
```

2. List the movies that were released between 1950 and 1960

```
grunt> movies_50_60 = FILTER movies by year>1950 and
year<1960;
```

3. List the movies that have duration greater that 2 hours

```
grunt> movies_50_60_2 = FILTER movies_50_60 by duration
> 7200;
```

Pig Latin 資料抽取（二）

4. List the movies that have rating between 3 and 4

grunt> **mymovies = FILTER movies_50_60_2 BY rating>3.0 and rating<4.0;**

grunt> **dump mymovies;**

::

(719,War and Peace,1956,3.6,12500)
(2437,Daddy Long Legs,1955,3.8,7593)

Pig Latin 資料轉換 – 改變欄位結構與資料轉換

```
grunt> movie_duration = FOREACH movies GENERATE name,
(double)(duration/60);
                                    ::
(Lady Gaga &#38; The Muppets' Holiday Spectacular,58.0)
(Sunset Strip,96.0)
(Silver Bells,88.0)
(Winter Wonderland,30.0)
(Top Gear: Series 19: Africa Special,113.0)
(Fireplace For Your Home: Crackling Fireplace with Music,60.0)
(Kate Plus Ei8ht,)
(Kate Plus Ei8ht: Season 1,)


grunt> mymovies = filter movie_duration by $1 is not null;
grunt> dump mymovies;
                              ::
(Winter Wonderland,30.0)
(Top Gear: Series 19: Africa Special,113.0)
(Fireplace For Your Home: Crackling Fireplace with Music,60.0)
```

Pig Latin 資料轉換 – 總計分析

```
grunt> desc_movies_by_year = ORDER movies BY year ASC;

grunt> grouped_by_year = group desc_movies_by_year by
year;

grunt> count_by_year = FOREACH grouped_by_year
GENERATE group, COUNT(desc_movies_by_year);

grunt> dump count_by_year;
                          ::
(2007,2892)
(2008,3358)
(2009,4451)
(2010,5107)
(2011,5511)
(2012,4339)
(2013,981)
(2014,1)
```

Pig Latin 資料裝載

```
grunt> store count_by_year into 'count_by_year.csv';

grunt> ls
hdfs://nna:8020/user/ds01/count_by_year.csv    <dir>
hdfs://nna:8020/user/ds01/movies_data.csv<r 2>  2893177
hdfs://nna:8020/user/ds01/movies_greater_than_four.csv  <dir>
hdfs://nna:8020/user/ds01/movies_with_duplicates.csv<r 2>      539

grunt> ls count_by_year.csv
hdfs://nna:8020/user/ds01/count_by_year.csv/_SUCCESS<r 2>      0
hdfs://nna:8020/user/ds01/count_by_year.csv/part-r-00000<r 2>   841

grunt> cat count_by_year.csv
                    ::
2008    3358
2009    4451
2010    5107
2011    5511
2012    4339
2013    981
2014    1
```

實作練習

1. 那一年最多電影產出？
2. 列出沒有分等級的電影

撰寫 pig 程式 - 轉換 Bag 資料為 Tuple 格式

```
$ nano bag2tuple.pig
a = LOAD 'ten_group.csv' USING PigStorage('\t') as (year:int,
movie:bag{item:tuple(id:int,name:chararray,year:int,rating:float,du
ration:int)});
b = FOREACH a GENERATE $0, BagToString($1, ',');
dump b;

$ pig -f bag2tuple.pig
                    :::
(1921,3,Orphans of the Storm,1921,3.2,9062)
(1929,9,Nosferatu: Original Version,1929,3.5,5651)
(1932,2,The Mummy,1932,3.5,4388)
(1963,5,Night Tide,1963,2.8,5126)
(1985,6,One Magic Christmas,1985,3.8,5333)
(1991,4,The Object of Beauty,1991,2.8,6150)
(1993,1,The Nightmare Before Christmas,1993,3.9,4568)
(1994,8,Mother's Boys,1994,3.4,5733,7,Muriel's Wedding,1994,3.5,6323)
(1995,10,Nick of Time,1995,3.4,5333)
2016-07-28 23:00:49,585 [main] INFO  org.apache.pig.Main - Pig script
completed in 58 seconds and 976 milliseconds (58976 ms)
```

撰寫 pig 程式 - 排序 Bag 資料

```
$ nano sortbag.pig
a = LOAD 'movies_data.csv' USING PigStorage(',');
b = limit a 20;
c = group b by $2;
d = FOREACH c {
    d1 = foreach b generate $1,$3,$4;
    d2 = order d1 by $1 desc;
    generate group, d2;
}
dump d;

$ pig -f sortbag.pig
                    :::
(1981,{(Bustin' Loose,3.7,5598)})
(1985,{(The Breakfast Club,4.0,5823),(One Magic Christmas,3.8,5333)})
(1991,{(The Object of Beauty,2.8,6150)})
(1993,{(The Nightmare Before Christmas,3.9,4568)})
(1994,{(Muriel's Wedding,3.5,6323),(Mother's Boys,3.4,5733)})
(1995,{(Nick of Time,3.4,5333)})
(1996,{(Big Night,3.6,6561),(Beautiful Girls,3.5,6755)})
2016-07-28 21:48:20,380 [main] INFO  org.apache.pig.Main - Pig script completed in 4
minutes, 4 seconds and 827 milliseconds (244827 ms)
```

資料科學家收工

離開 cla01 貨櫃主機

```
$ exit
logout
Connection to cla01 closed.
```

課後練習

大專校院名錄分析

1. 104 年大專校院總數
2. 各縣市大學總數

下載 104 年大專校院名錄

```
$ wget --no-check-certificate
https://stats.moe.gov.tw/files/school/104/u1_new.txt
$ iconv -f UCS-2 -t utf8 u1_new.txt -o u104.txt

$ head -n 5 u104.txt
104學年度大專校院名錄

代碼    學校名稱        縣市名稱        地址    電話    網址    體系別
0001    國立政治大學    [38]臺北市      [116]臺北市文山區指南路二段64號
(02)29393091    http://www.nccu.edu.tw  [1]一般
0002    國立清華大學    [18]新竹市      [300]新竹市東區光復路二段101號
(03)5715131     http://www.nthu.edu.tw  [1]一般

$ hdfs dfs –put u104.txt
$ pig
grunt> a = load 'u104.txt';
grunt> b = foreach a generate REGEX_EXTRACT($2,'(.*)](.*)',2),$3;
grunt> dump b;
                    ::
(宜蘭縣,[266]宜蘭縣三星鄉三星路二段265巷100號)
(桃園市,[325]桃園市龍潭區中豐路高平段418號)
(,)
(,)
```