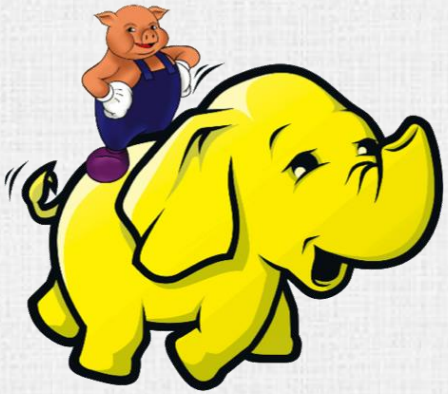


02-Pig Latin

2018金象盃

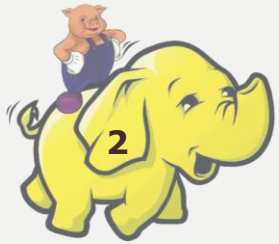
講者: 林葳秦 老師

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

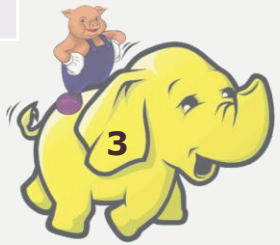
可再細分指令的類型

讀取	LOAD
儲存	STORE
資料處理	FILTER, FOREACH GENERATE, GROUP, inner JOIN, outer JOIN, UNION, SPLIT, …
彙總運算	AVG, COUNT, MAX, MIN, SIZE, …
數學運算	ABS, RANDOM, …
字串處理	INDEXOF, SUBSTRING, REGEX_EXTRACT, …
Debug	DUMP, DESCRIBE
HDFS	cat, ls, cp, mkdir, copyfromlocal, copyToLocal, ……



Pig 基本資料型態

型態	中文	說明
普通資料型態		
int	整數	
long	長整數	
float	浮點數	
double	精確浮點數	
陣列資料型態		
chararray	字串	
bytearray	位元組陣列	



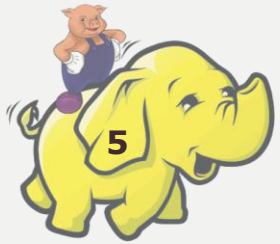
資料集: 美國電影

movies_data.csv

認識資料 - movies_data.csv

欄位名稱	資料型態	欄位說明
id	**int**	電影編號
name	**chararray**	電影名稱
year	**int**	上映年分
rating	**float**	評價星等
duration	**int**	片長(秒)

49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610
49589,Kate Plus Ei8ht,2010,2.7,
49590,Kate Plus Ei8ht: Season 1,2010,2.7,



看到資料 - movies_data.csv

$ hdfs dfs -ls /dataset/movies_data/

```
dsa01@dswk01:~$ hdfs dfs -ls /dataset/movies_data/
Found 1 items
-rw-r--r--   3 bigred alpha    2893177 2018-10-17 05:18 /dataset/movies_data/movies_data.csv
```

$ hdfs dfs -cat /dataset/movies_data/movies_data.csv

:::

電影編號　　電影名稱　　上映年分　評價　片長

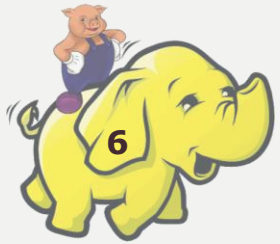
49586,Winter Wonderland,2013,2.8,1812

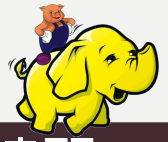
49587,Top Gear: Series 19: Africa Special,2013,,6822

49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610

49589,Kate Plus Ei8ht,2010,2.7,

49590,Kate Plus Ei8ht: Season 1,2010,2.7,





練習1

請研究userprofile.tsv資料集

載入資料集 & 倒出資料

進入Pig交談模式

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

grunt> dump movies;
```

:::

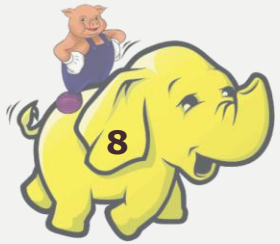
(49586,Winter Wonderland,2013,2.8,1812)

(49587,Top Gear: Series 19: Africa Special,2013,,6822)

(49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610)

(49589,Kate Plus Ei8ht,2010,2.7,)

(49590,Kate Plus Ei8ht: Season 1,2010,2.7,)



載入資料集並顯示前5筆原始資料(定義資料型態)

```
$ pig

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

grunt> limit5 = LIMIT movies 5;

grunt> dump limit5;
```

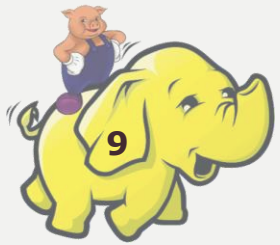
(1,The Nightmare Before Christmas,1993,3.9,4568)

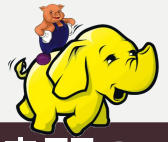
(2,The Mummy,1932,3.5,4388)

(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)





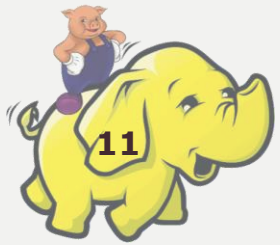
練習2

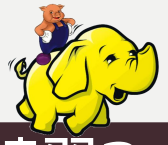
如何顯示全部的資料?

1920年共有幾部電影?

```
$ pig

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

grunt> a = FILTER movies by (year == 1920);

grunt> dump a;
```

(62,Dr. Jekyll and Mr. Hyde,**1920**,3.2,4679)
(79,The Mark of Zorro,**1920**,3.1,6433)
(189,Way Down East,**1920**,3.0,8985)
(2744,One Arabian Night,**1920**,2.5,6203)
(2745,Anna Boleyn,**1920**,2.4,7118)
(42676,Headin' Home,**1920**,2.7,4390)





練習3

1917年共有幾部電影?

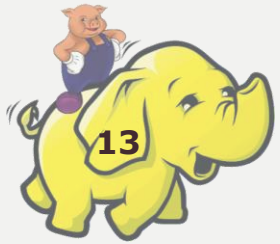
列出片長大於2小時的電影資料

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

grunt> long_movies = FILTER movies by (duration>7200);
```

:::

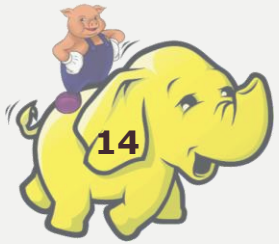
2018-10-13 14:06:49,933 [main] ERROR org.apache.pig.tools.grunt.Grunt - **ERROR** 1025: <line 3, column 31> Invalid field projection. Projected field **[daration] does not exist**.

Details at logfile: /home/dsa03/pig_1539439566328.log



列出片長大於2小時的電影資料

```
grunt> long_movies = FILTER movies by (duration>7200);

grunt> dump long_movies;

      :::

(45108,Settai,2013,2.8,7390)

(46832,WWE for All Mankind: The Life &#38; Career of Mick Foley,2013,4.2,8066)

(46949,Example Short XII 23976 2 Hour,2010,2.6,7370)

(48714,WWE: The Top 25 Rivalries in Wrestling History,2013,3.7,10228)

(49055,Theeya Velai Seyyanum Kumaru,2013,3.3,8239)


grunt> quit
```



「只」列出電影名稱

```
grunt> movies = LOAD '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> movies_name = FOREACH movies GENERATE $1;

grunt> dump movies_name;
```

:::

(Silver Bells)

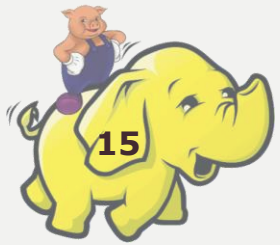
(Winter Wonderland)

(Top Gear: Series 19: Africa Special)

(Fireplace For Your Home: Crackling Fireplace with Music)

(Kate Plus Ei8ht)

(Kate Plus Ei8ht: Season 1)



列出電影資料集的評價、電影名稱欄位

```
grunt> movies = LOAD '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> a= FOREACH movies GENERATE $3,$1;

grunt> dump a;
```

:::

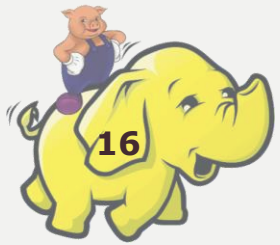
(2.8,Winter Wonderland)

(,Top Gear: Series 19: Africa Special)

(,Fireplace For Your Home: Crackling Fireplace with Music)

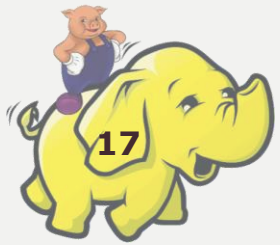
(2.7,Kate Plus Ei8ht)

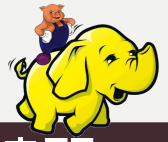
(2.7,Kate Plus Ei8ht: Season 1)



列出有評價的資料

```
grunt> b= filter a by ($0 is not null);
grunt> dump b;
      :::
(3.0,Sunset Strip)
(3.5,Silver Bells)
(2.8,Winter Wonderland)
(2.7,Kate Plus Ei8ht)
(2.7,Kate Plus Ei8ht: Season 1)
```





練習4

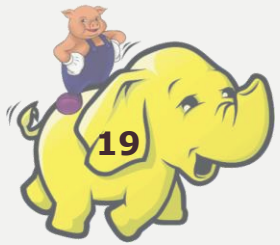
列出電影資料集的片長、

評價欄位並過濾掉所有空值

ex. (片長,評價)

列出電影評價大於4顆星的資料

```
$ pig
grunt> movies = LOAD '/dataset/movies_data/movies_data.csv'
USING PigStorage(',');
grunt> best_movies = FILTER movies BY $3>4;
grunt> dump best_movies;
      :::
???(沒有資料)???
```



列出電影評價大於4.0的資料

```
grunt> best_movies = FILTER movies BY (float)$3>4;

grunt> dump best_movies;
```

:::

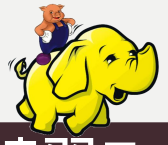
(49549,Life With Boys: Season 1,2011,**4.1**,)

(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)



練習5

列出片長大於2.5小時且
評價界於3.5~4.5之間
的電影資料

將電影由年份遠到近排序

```
grunt> movies = LOAD '/dataset/movies_data/movies_data.csv'
USING PigStorage(',');

grunt> orderby_year = ORDER movies BY $2 ASC;

grunt> dump orderby_year;
```

:::

(45575,Caillou: Season 5: Big Time Caillou & Other Stories: Blast Off to Space,**2013**,,1551)

(44561,InAPPropriate Comedy,**2013**,2.6,5020)

(49561,The Square (Trailer),**2014**,3.6,154)

將電影由年份近到遠、評價低到高排序

```
grunt> movies = load '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> orderby_year = order movies by $2 DESC, $3 ASC;

grunt> dump orderby_year;
```

:::

(610,Cabiria,1914,2.9,7684)

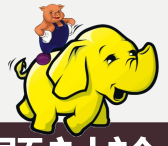
(29801,Charlie Chaplin Collection,1914,3.8,)

(20067,Charlie Chaplin Collection: Shorts,1914,3.8,)

(14328,Fant?mas III: The Murderous Corpse,1913,2.6,5432)

(42671,Fant?mas II: Juve vs. Fant?mas,1913,2.7,3718)

(42665,Fant?mas I: In the Shadow of the Guillotine,1913,2.9,3268)



問題討論

a. **ORDER BY** 命令可以不指定排序方法嗎？

b. 預設是如何排序？

最早上映電影的年分

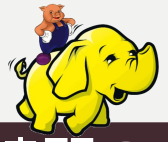
```
grunt> movies = load '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> a1 = ORDER movies by $2 ASC;

grunt> a2 = LIMIT a1 1;

grunt> a3 = FOREACH a2 GENERATE $2;

grunt> dump a3;
```

(1913)

最晚上映的電影年分

```
grunt> movies = load '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> b1 = ORDER movies by $2 DESC;

grunt> b2 = LIMIT b1 1;

grunt> b3 = FOREACH b2 GENERATE $2;

grunt> dump b3;
```

(2014)



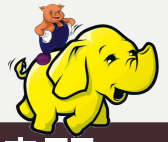
練習6

a. 哪一部電影片長最長？

b. 哪一部電影片長最短？

此資料集中有哪幾年的電影資料

```
grunt> movies = load '/dataset/movies_data/movies_data.csv'
using PigStorage(',');

grunt> c1 = ORDER movies by $2;

grunt> c2 = FOREACH c1 GENERATE $2;

grunt> c3 = DISTINCT c2;

grunt> dump c3;
  :::
(1916)
(1918)
  :::
(2013)
(2014)
```



練習7

列出此資料集中評價最高的

電影名稱與評價

ex. (電影名稱,評價)