Pig 常用指令

LOAD

代號 = LOAD '[路徑]' using PigStorage('[, ; .]') AS ([名稱]:[類別], [名稱]:[類別], ...);

- LOAD '[檔案路徑]'
  載入資料的時候要使用的敘述 ,'[檔案路徑]' 這邊要打上存在 **HDFS** 的檔案路徑
- using PigStorage(',')
  敘述中其中的這一段就是設定要 Pig 系統在讀取資料的時候要用什麼符號來辨識欄位的分隔 , 舉例來說如果你讀取的是 Excel(.csv) 的檔案在 Excel 的系統中是使用 , 號來做欄位的分隔 , 所以這邊的程式碼就要這樣寫 using PigStorage(',') , 注意要要依照自己要讀取的檔案來設定分隔符號 , 不要傻傻的都填上 , 這樣你會發現讀取進來的資料全部通通都擠在一個欄位裡面
- AS ([欄位名稱]:[資料型別], [欄位名稱]:[資料型別], ...)
  這裡就是要分別給每一個欄位設定「欄位的名稱」與「資料的型態」,詳細的說明於範例中做加以說明

DUMP

Dump [代號]

- 將下達的指令 撰寫 -> 編譯 -> 執行 -> 顯示 ,可以在一份程式你呼叫 DUMP 無限次 , 而唯一需要遵守的規則指有兩條：
  1.必須擺在你要看的敘述的後面
  2.必須是存在的 [代號]

FOREACH

[代號] = FOREACH [代號] GENERATE [欄位], [欄位], [欄位], ...;

```
Ex:
B = FOREACH A GENERATE a, b;（只列出 a,b）
```

- 對資料中的每一筆做操作 , 可以做 增加欄位、刪除欄位 、 資料型態轉換、欄位命名、欄位運算

FILTER

[代號] = FILTER [代號] BY ([表達式]);

Ex:

```
B = FILTER A BY (d == 8);指留下欄位 b 等於 8 的資料
C = FILTER A BY (d > 4);指留下欄位 b 大於 4 的資料
D = FILTER A BY (e < 5);指留下欄位 e 小於 5 的
```

可以使用 正規表示式 或是一般的 表達式 作過濾

GROUP

[代號] = GROUP [代號] BY [欄位];

依照指定的 [欄位] 將資料做分類

ORDER

[代號] = ORDER [代號] BY [欄位] [ASC|DESC];

依照指定的 [欄位] 做排序

```
B = ORDER A BY a ASC, d DESC;
```

可以只排列一個欄位，也可以排列兩個欄位甚至是更多個欄位，並且可以分別指定排序方式。

Ex:a ASC 將欄位 a 由小排到大,d DESC 將欄位 d 由大排到小

DISTINCT

[代號] = DISTINCT [代號];

移除重複的資料

```
Ex:E = DISTINCT D;
```

LIMIT

[代號]＝limit [代號] 筆數；

看前？筆數

```
Ex:E = limit D 5;
```

步驟

一、首先將資料從「開放平台」下載至作業系統上

```
$ wget 網址
```

二、檢視資料集

```
$ head -n 6 檔名
$ tail -n 1 檔名
```

三、簡易過濾不要的資料，萃取必需的原始資料

```
$ cat  檔名 | grep '條件' > 檔名
$ cat  檔名 | grep '條件' > 檔名
```

四、將過濾後的數據資料放入 HDFS

```
grunt> copyfromlocal 檔名 .
grunt> ls
```

五、使用分析工具定義數據資料型態

```
grunt> movies = LOAD '檔名'USING PigStorage(',') AS
(id:int,name:chararray,year:int,rating:float,duration:int);
```

六、顯示要看的資料筆數「五筆」

```
grunt> limit5 = LIMIT  檔名  5;
grunt> dump limit5;
```