Etu Hadoop Competition 2015

那兩年，我們一起追的Hadoop

詹景逸、薛元揆、祿仲倫

2015/06/27

Outline

- 資料總覽
- 資料前處理
- 預測模型建立
- 系統架構與參數
- 結語

資料總覽

	View	Search	Cart	Order	Total
Train	4,696,645 (86%)	337,306 (6%)	368,450 (7%)	58,922 (1%)	5,461,323 (100%)
Test	4,662,896 (93%)	374,964 (7%)	-	-	5,037,860 (100%)

資料前處理

Missing Values

商品缺值統計

資料集	缺少價格資訊	缺少類別資訊
Train	188,059 (77%)	21,038 (9%)
Test	203,508 (87%)	0 (0%)
Train + Test	251,732 (82%)	21,038 (7%)

產品總數為 308,682

商品id轉換

- 資料集中商品id的格式不一致，可以轉換成一致的格式

Digits	Product Id	Uniform Product Id	Comment
10	1234567890	U123456789	
13	1234567890ABC	U123456789	product variant

內插法填補
產品價格與產品類別缺值

使用內插法前

```
000001015,188,A_004_017_003,
000001017,188,A_004_017_003,
000001018,188,A_004_017_003,
000001022,188,A_004_017_003,
000001023,198,A_004_017_003,
000001033,188,A_004_017_003,
000001034,188,A_004_017_003,
000001070,0,A_004_017_001,
000001072,220,A_002_017_015,
000001097,1790,F_017_011_002
000001105,1490,F_017_008_005
000001108,1730,F_017_008_005
000001110,1299,F_017_008_003
```

使用內插法後

```
000001015,188,A_004_017_003,
000001017,188,A_004_017_003,
000001018,188,A_004_017_003,
000001022,188,A_004_017_003,
000001023,198,A_004_017_003,
000001033,188,A_004_017_003,
000001034,188,A_004_017_003,
000001070 220,A_004_017_001,
000001072,220,A_002_017_015,
000001097,1790,F_017_011_002
000001105,1490,F_017_008_005
000001108,1730,F_017_008_005
000001110,1299,F_017_008_003
```

商品缺值統計

資料集	缺少價錢	缺少類別
Train + Test (原始資料集)	251,732 (82%)	21,038 (7%)
Train + Test (產品id轉換)	240,594 (78%)	3,444 (1%)
Train + Test (爬商品網站)	3,827 (0.1%)	501 (0.2%)
Train + Test (使用內插法填值)	0 (0%)	0 (0%)

產品總數為 308,682

Data Cleaning

統一類別資訊

- 部分資料只有小類資訊，缺少大類資訊

Category type	Log	Log after extraction
multiple cids	cat=J,J_007,J_007_009, J_007_009_016	cat=J,J_007,J_007_009, J_007_009_016
one cids	cat=H_004_017_004	cat=H,H_004,H_004_017, H_004_017_004

搜尋記錄解碼

```
36.230.39.90,2015-02-01 00:10:51,,３ｍ　隱形,ff0ff75f-e8ac-9ddb-bf1b-aa449822b085
1.34.131.167,2015-02-01 00:10:57,,gucci|包,dc945994-2472-2cf5-2fdc-eb85defc5465
220.137.3.34,2015-02-01 00:10:58,U234579365,mp3,92e720da-17be-2b67-3383-9e5ccbd9499f
218.166.6.240,2015-02-01 00:11:05,,落健,3227e323-71df-70fd-efec-dc03e856ad07
118.161.204.147,2015-02-01 00:11:07,U398804258,電腦桌,8e253a92-1e43-480-b644-79441bf03b8a
61.58.168.22,2015-02-01 00:11:19,,奧利佛,189c0c8-6509-776d-d3f7-d4fb7208b200
1.162.43.249,2015-02-01 00:11:20,,創見,66ed4b7-bfd4-8432-2d85-ee0880326f07
1.34.131.167,2015-02-01 00:11:26,,gucci|夾,dc945994-2472-2cf5-2fdc-eb85defc5465
1.169.33.96,2015-02-01 00:11:33,,6632,eab7f756-4ace-a67e-b59c-44aeefb72c16
175.180.94.248,2015-02-01 00:11:40,,好吃滷味,853c463d-4996-31e8-1fe2-f43c9b52a9bb
36.225.164.46,2015-02-01 00:11:50,,包鐘包,84689b9b-afaa-2b47-1886-4b4fbbdd91d5
36.236.115.77,2015-02-01 00:11:50,,espon|141,57fc3a0d-11dd-912b-2be6-5deb3aadebd
112.104.97.194,2015-02-01 00:11:51,,空白,e18fc7c7-490-b470-1301-9d552daf8b
123.204.126.78,2015-02-01 00:11:55,,彩虹馬,d33c7bc8-182c-dd0d-d256-89d05d97550f
36.225.164.46,2015-02-01 00:11:58,,包中包,84689b9b-afaa-2b47-1886-4b4fbbdd91d5
1.162.43.249,2015-02-01 00:12:01,,創見隨身碟,66ed4b7-bfd4-8432-2d85-ee0880326f07
218.166.6.240,2015-02-01 00:12:01,,首爾,3227e323-71df-70fd-efec-dc03e856ad07
61.58.168.22,2015-02-01 00:12:04,,奧利佛,189c0c8-6509-776d-d3f7-d4fb7208b200
106.1.53.220,2015-02-01 00:12:08,,耳麥,fb4180e4-d875-f34d-fbdb-6773ee72a199
49.158.208.247,2015-02-01 00:12:11,,Hello|Kitty,47cac58b-3311-31a-b4f-26b6af67248d
114.32.66.251,2015-02-01 00:12:11,,分裝,2fd022c8-480c-ff30-5b74-39823f4ee41f
125.230.67.137,2015-02-01 00:12:12,U46488849,eyah,d8d2d9f6-52d8-cbb5-2aef-baafd1472995
106.1.53.220,2015-02-01 00:12:15,,耳麥,fb4180e4-d875-f34d-fbdb-6773ee72a199
1.162.43.249,2015-02-01 00:12:15,,創見隨身碟,66ed4b7-bfd4-8432-2d85-ee0880326f07
114.46.139.167,2015-02-01 00:12:17,,Sandia|POLO,d07514e5-595b-d987-cd79-e2d935a2111
219.85.255.122,2015-02-01 00:12:27,U440888317,U002282725,bb96ea56-41a-75-7b32-75e88f8737
123.193.123.32,2015-02-01 00:12:32,U448908360,夏慕尼,9c48c11e-1a93-73fd-9a47-991ec83a94fc
36.233.145.46,2015-02-01 00:12:33,,傳真機,f9eb1fb0-6fca-989b-ea0f-bacaa31d6b6
```

日期校正

Dataset	Date	Date after shift
Train	2015/2	2013/9
Test	2015/3	2013/10

- 訓練資料集有 iPhone 5, iPad 4 與 iPad mini等產品資訊，但缺少 iPhone 5s

- 測試資料集在 2013/10/29~2013/10/31 出現 iPhone 5s 的瀏覽紀錄

預測模型建立

Problem#1

預測 erUid 是否有購買

設計思路

- 商品統計數據能對總體有個概念，但無法預測過去數據不顯著的商品
- 對使用者行為歸類，能套用在不同商品
  - view log by session -> 買 or 不買
  - view log by pid -> 買得多 or 買得少

Input: view log by session

access log 依據 erUid 合併

Features	Description
erUid	使用者id
viewCount	使用者在Session看過幾次商品
uniqeViewCount	使用者在Session看過幾次不重複商品
cat[01ABCDEFGHIJKLOV]	Session中看過某類別的次數
maxCat	Session中看過多次的類別
buy	Session中是否有發生購買

Output: erUid是否有購買

erUid	buy
A	Yes
B	No
C	No

Algorithm
Random Forest

- 最初是想用 view log 的使用者行為，利用決策樹簡單的分類「買」或「沒買」

- 而 Random Forest 可以給我們較單一決策樹更好的建議

- ntree: 30

Problem#2

預測商品的購買權重

Input: 商品與使用者的購物歷程

access log 依據 pid 合併，並整合erUid購買預測結果

Features	Description
pid	商品id
viewCount	商品被看過的次數
viewBySession	商品被幾個Session瀏覽過的次數
price	商品價格
cat	商品類別
buyWeight	商品購買權重

Output: 商品的購買權重

pid	buyWeight
A	6.1
B	1.7
C	8.8

Algorithm
SVM

- SVM 是相當適合做數據預測的工具，使用先前介紹的欄位去預測出 buyWeight。
- Kernel: Radial

Problem#3

估算購買數

設計思路

- 已有先前的二個 model：判斷某個 session 會不會買、判斷某個商品可能被買的權重，對於最終的結果還差商品數量
- 針對所有商品，需找出推算實際購買數量
  - {pidN} * {countN} * {price}
  - 其中 pidN 與 price 是已知的，採用 Genetic Algorithms 推算 countN

Algorithm
Genetic Algorithms

- initial Population：以 Product Model (svm) 獲得的 {pid, buyWeight} 集合為主，並加已知在榜內的 {Intop20Pid, 1.0} 作為第一代資料集

- fitness function：

  - 已知答案離 top 20 越近分數越高

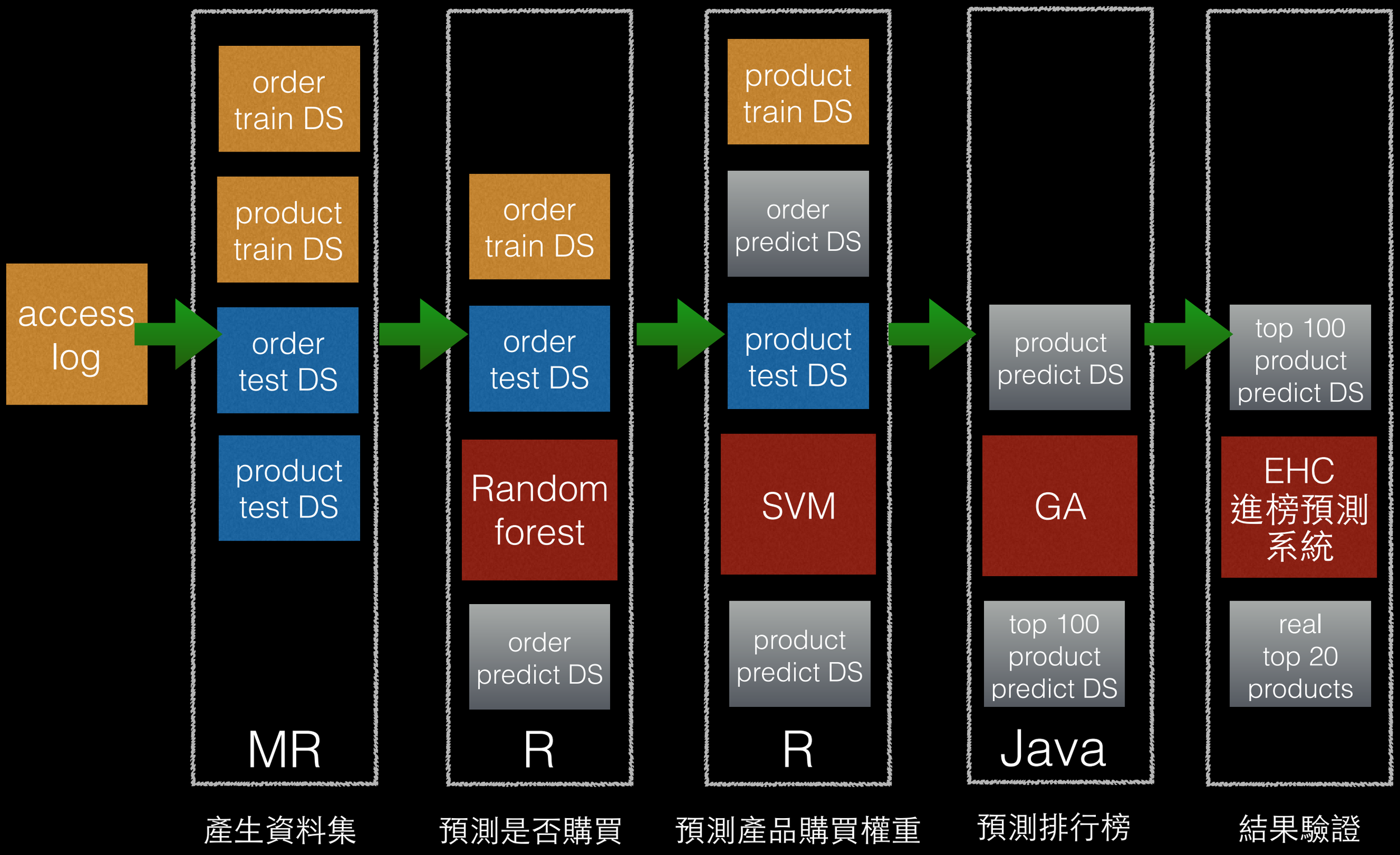
  - 整體參數 (representation) 分佈與 {pid, buyWeight} 越一致分數越高

Tuning Genetic Algorithm

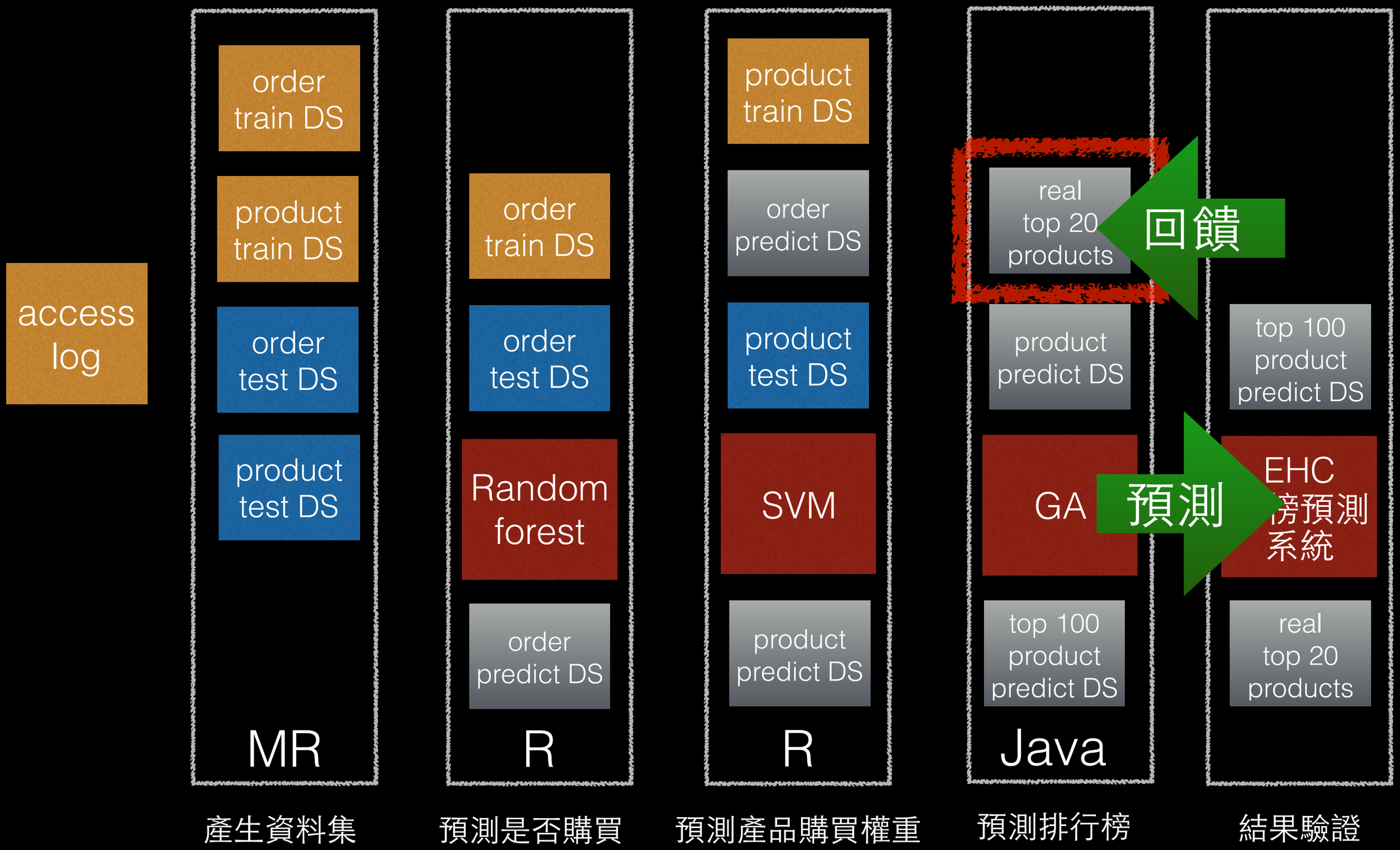
- 運用 GA 產生的最終 buyCount 去產生 top 100 清單
- 利用新的 top 100 剔除測試過的 pid 後，用 web 驗證工具找出新的解，配合商品統計數據比較，並再次加入 GA 初始參數產生新的清單

系統架構設計

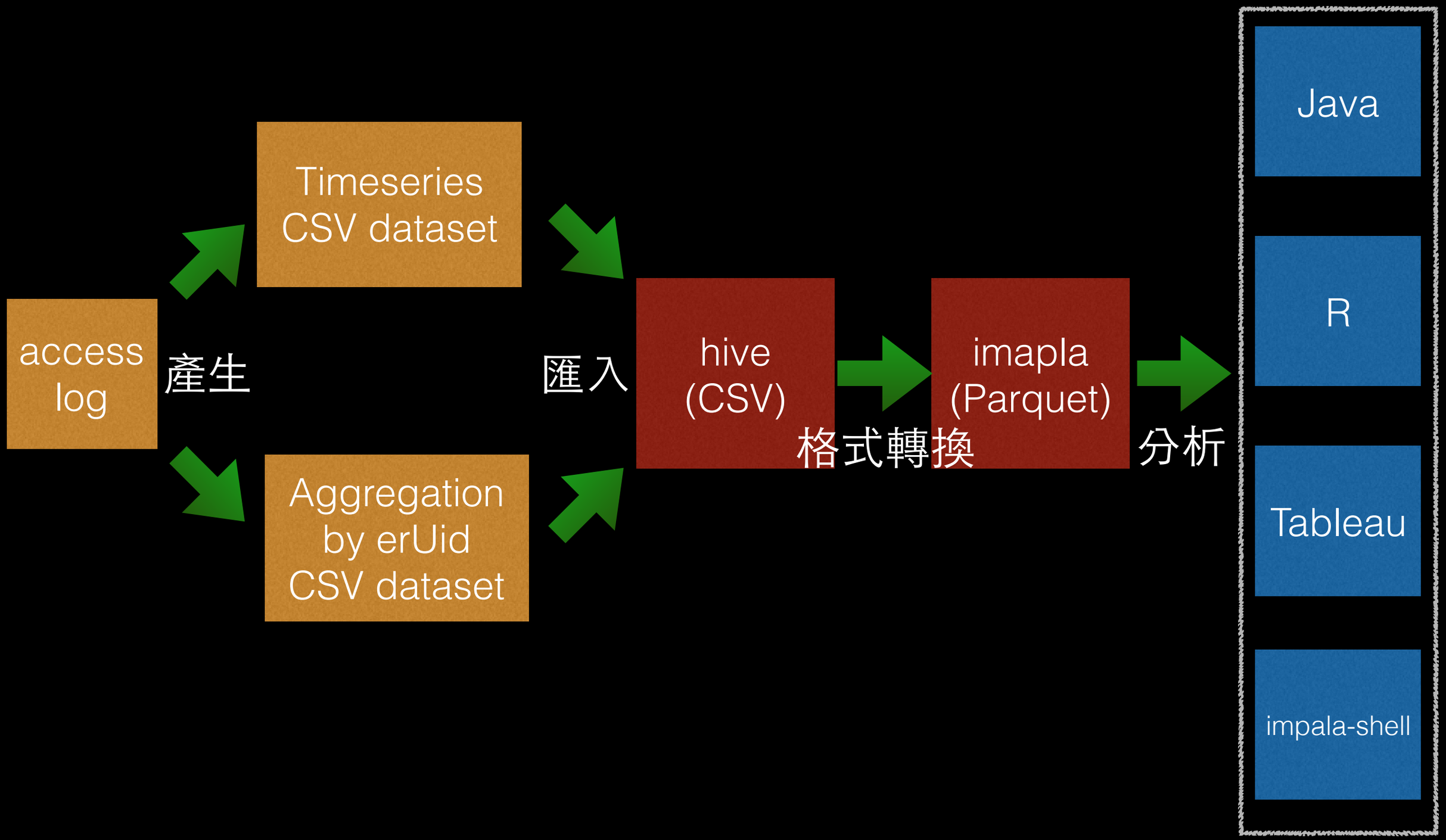
Machine Learning flow



Machine Learning flow



Analysis flow



效能參數說明

硬體規格與資料

- 硬體規格
  - Instance: m3.xlarge
  - vCPU: 4 cores
  - Compute Units (ECU): 13 units
  - Mem (GiB): 15.0 GB
- 前處理Job數量
  - MapReduce Jobs: 4次
- 資料大小
  - Train Dataset: 1.66 GB
  - Test Dataset: 1.55 GB

參數調整

- HDFS
  - Replica factor: 3 -> 1
  - Block Size: 128MB -> 256MB
- MapReduce
  - Mapper數量: 6
  - Reducer數量: 2
  - Mapper記憶體: -Xmx1152m
  - Reducer記憶體: -Xmx2304m

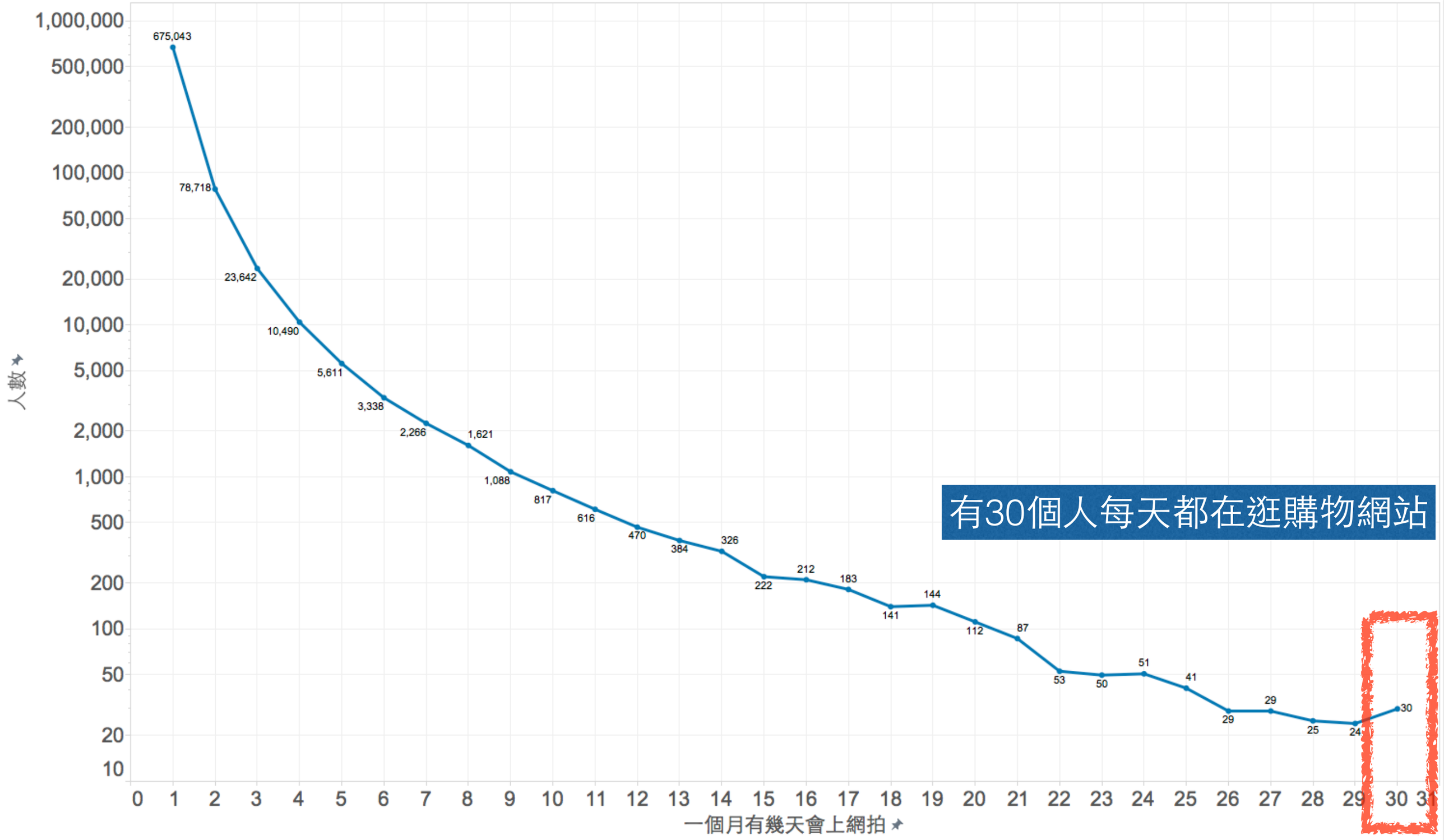
調整前後效能差異

調整階段	MapReduce耗時
調整前	5m12.175s
調整後	3m29.987s

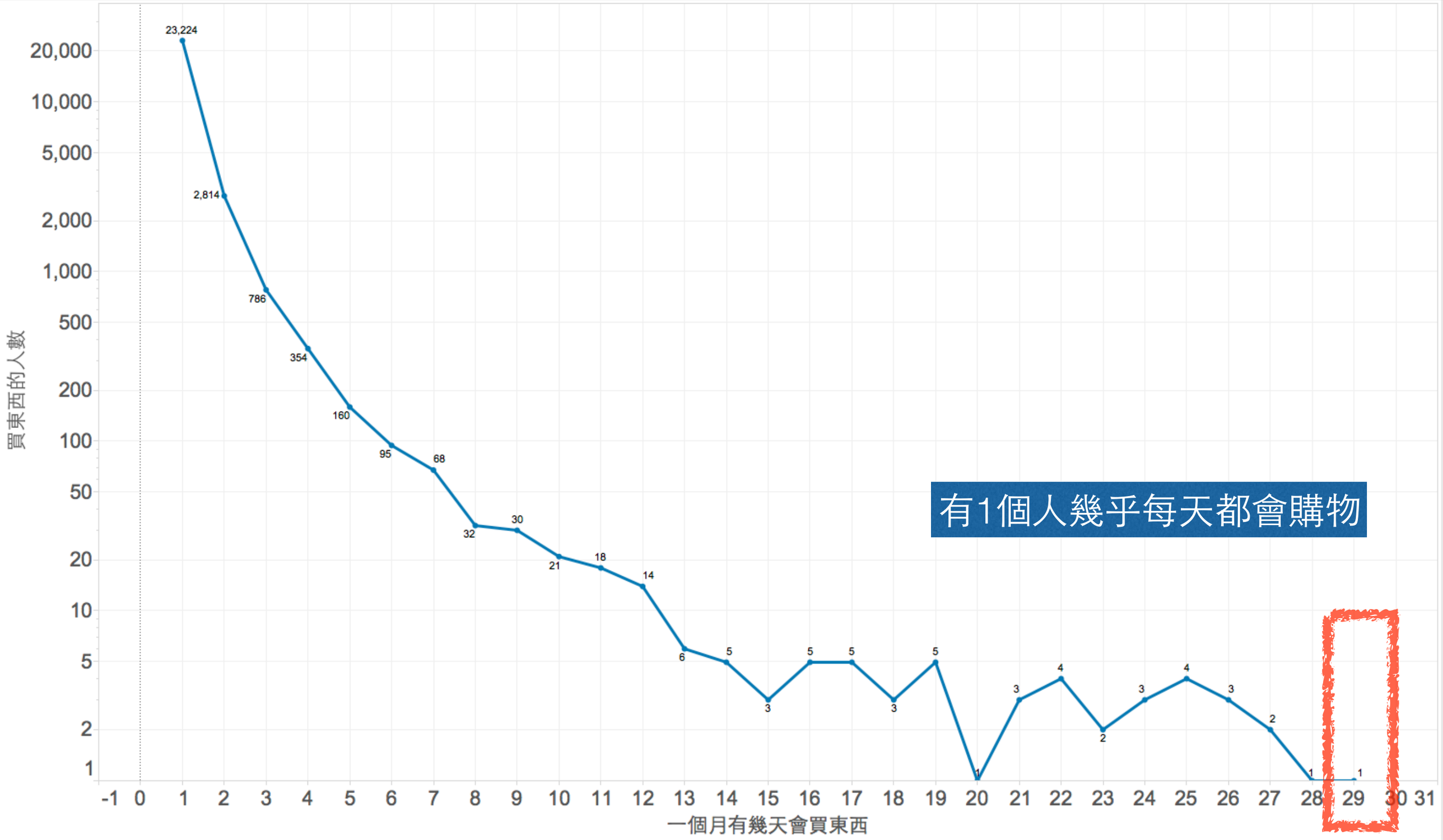
多個nodes時，應會有更好的結果

創意加分

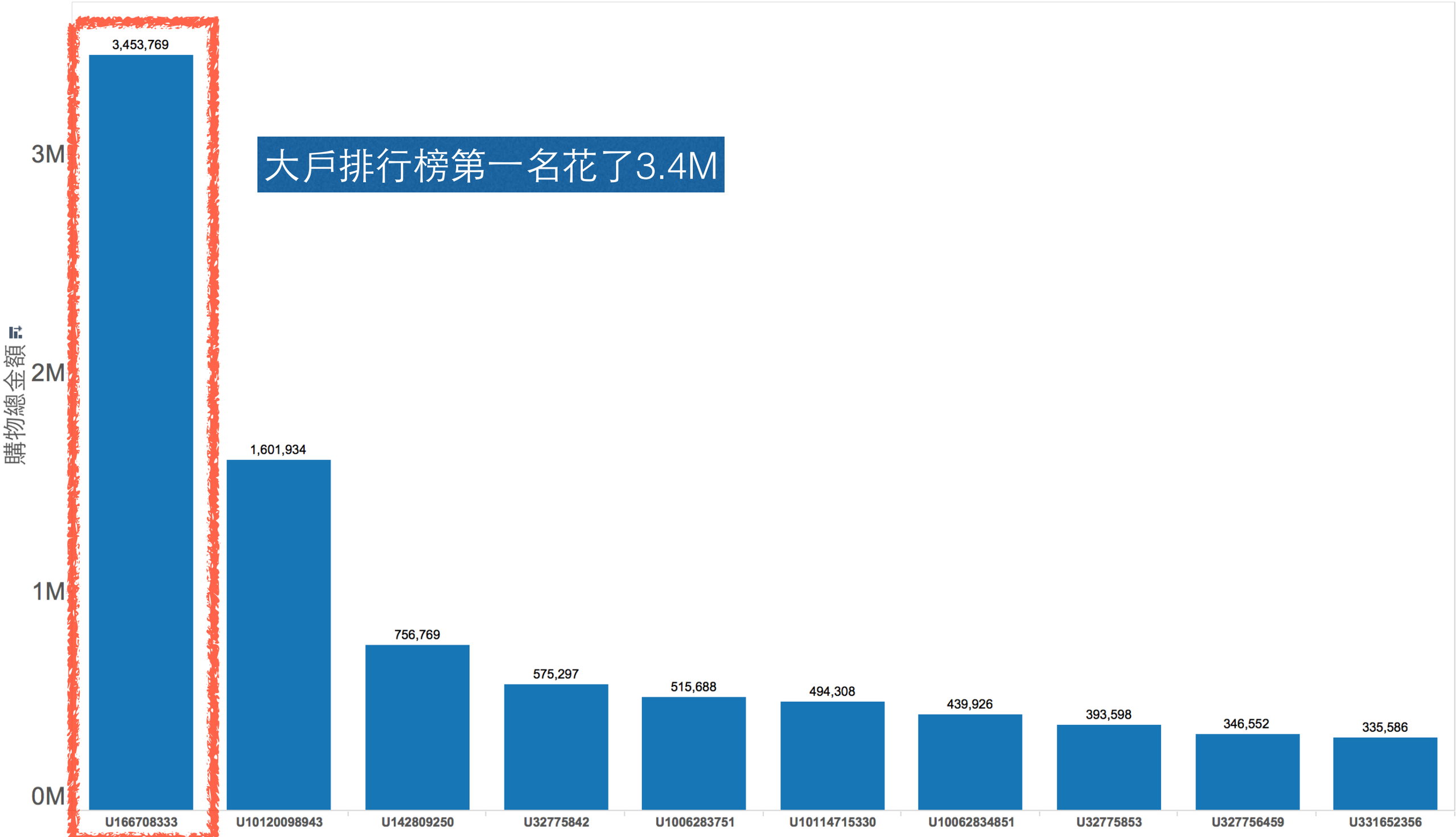
愛逛街


1,000,000
500,000
200,000
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
50
20
10
人數
675,043
78,718
23,642
10,490
5,611
3,338
2,266
1,621
1,088
817
616
470
384
326
222
212
183
141
144
112
87
53
50
51
41
29
29
25
24
30
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
一個月有幾天會上網拍
有30個人每天都在逛購物網站

愛購物


買東西的人數
20,000
10,000
5,000
2,000
1,000
500
200
100
50
20
10
5
2
1
23,224
2,814
786
354
160
95
68
32
30
21
18
14
6
5
3
5
5
3
5
1
3
4
2
3
4
3
2
1
1
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
一個月有幾天會買東西
有1個人幾乎每天都會購物

大戶排行榜

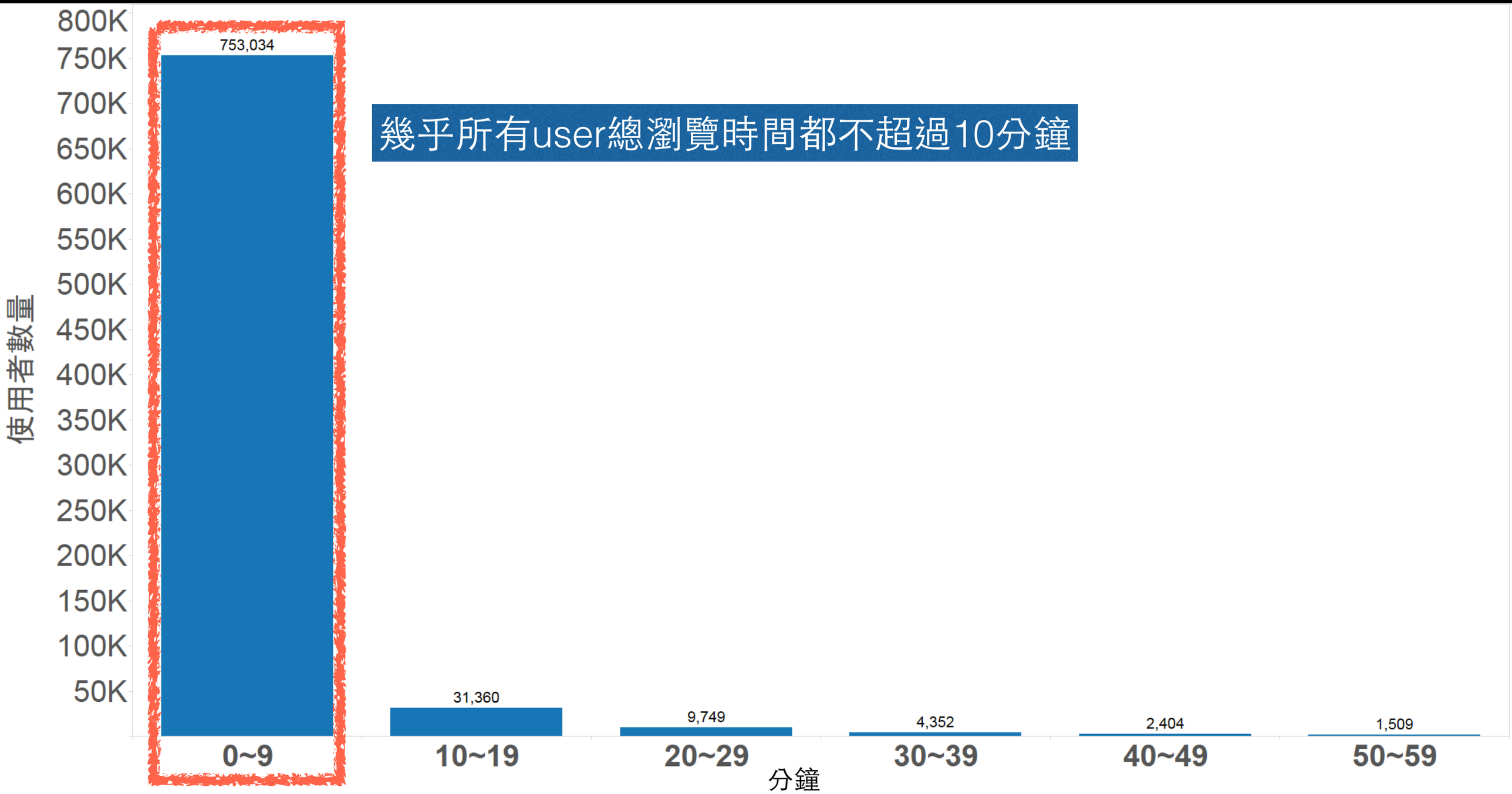

3,453,769
3M
2M
1M
0M
購物總金額
大戶排行榜第一名花了3.4M
1,601,934
756,769
575,297
515,688
494,308
439,926
393,598
346,552
335,586
U166708333
U10120098943
U142809250
U32775842
U1006283751
U10114715330
U10062834851
U32775853
U327756459
U331652356

瀏覽時間

項目	時間
Training Dataset 總瀏覽時間	2171天
平均單筆商品瀏覽時間	24.186秒

濾掉爬蟲後的數據 (eruid view count > 1000)

使用者總瀏覽時間分析



幾乎所有user總瀏覽時間都不超過10分鐘

價格異動影響銷售量

```
Query: select eruid, pid, price, day from ts_train where pid = '00068
+--------------------------------------+------------+-------+-----+
| eruid                                | pid        | price | day |
+--------------------------------------+------------+-------+-----+
| bbef31c-4fe3-10d-f381-80b662431d5    | 0006800651 | 11490 | 5   |
| e8728d18-5b8f-6962-da42-a1b7121cf005 | 0006800651 | 11490 | 7   |
| e6359516-282a-9c69-edfc-34b9d904e0ee | 0006800651 | 10490 | 11  |
| 34473f42-c5d3-cb98-eda1-fa6b17faed41 | 0006800651 | 10490 | 11  |
| fddea533-8a4a-d5d0-79c5-4559904d1e9b | 0006800651 | 10340 | 11  |
| a5b798d1-de76-1efd-54a6-d184e89ba38f | 0006800651 | 10340 | 11  |
| 37a73b79-5a71-4129-7577-aa2b7f07c9b0 | 0006800651 | 10340 | 11  |
| 37a73b79-5a71-4129-7577-aa2b7f07c9b0 | 0006800651 | 10340 | 11  |
| d5dceec3-9d75-d33c-d244-8d9b82ec686  | 0006800651 | 10340 | 11  |
| d5dceec3-9d75-d33c-d244-8d9b82ec686  | 0006800651 | 10340 | 11  |
| d5dceec3-9d75-d33c-d244-8d9b82ec686  | 0006800651 | 10340 | 11  |
| c1e8bc56-94ca-f6bf-65d8-45e58d598ec2 | 0006800651 | 10340 | 11  |
| ecbfb29-f94-3a62-8d43-9c89f00f225a   | 0006800651 | 10690 | 18  |
| 4eada449-504e-d9ac-3d32-3efa6da76778 | 0006800651 | 10690 | 18  |
| 4eada449-504e-d9ac-3d32-3efa6da76778 | 0006800651 | 10690 | 18  |
| 2e3e825c-879b-3ecc-9d22-c51bd3c3997c | 0006800651 | 10690 | 19  |
+--------------------------------------+------------+-------+-----+
Fetched 16 row(s) in 0.95s
```

原價

降價

調漲

Ad campaign轉換率

Conversation rate = buy count / clicks

- 訓練資料集中來源網址有utm_campaign欄位即為click
- Clicks: 967
- Buy count: 4

轉換率約0.4%

Etu推薦準不準?

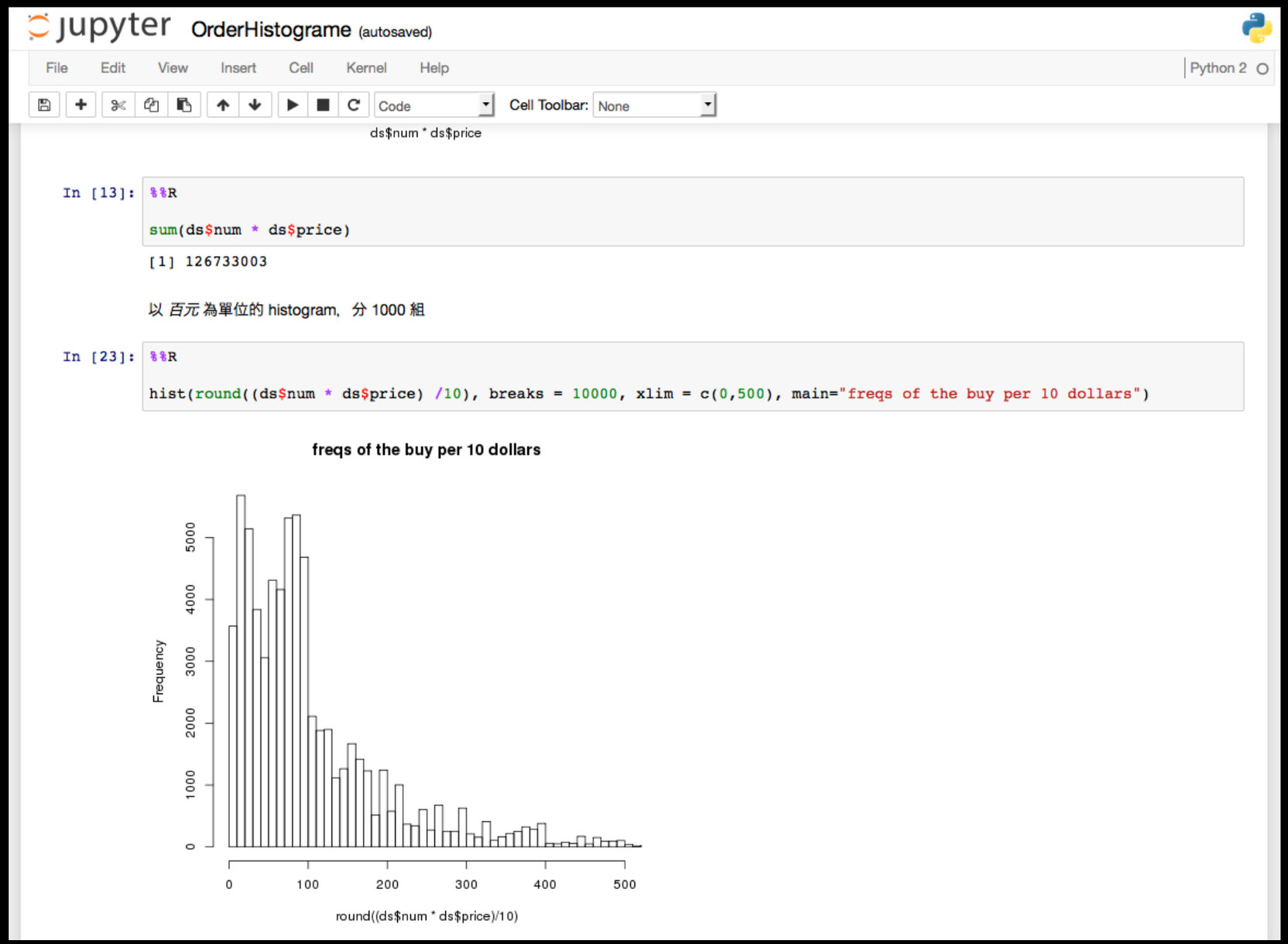
Buy rate = buy count / clicks

- 訓練資料集中有eturec欄位應為etu推薦商品
  - Eturec count: 894,339
  - Buy count: 10,593

購買率約1.14%

協作開發環境

- iPython notebook + rmagic



Conclusions

- 拋棄成見
  - 常拿著統計數據與預測結果在猶豫「真的有人買這東西嗎？」
  - Machine Learning 的黑盒子，推薦了個上個月 800 名以外的品項，主觀上不敢放入名單，但它卻是在名單之內。

Conclusions

- 實務 v.s. 比賽?
  - 高達82%的產品沒有價格, 導致
    - 沒有金額的產品不會進top 20 (含測試資料集才出現的產品)
    - 13碼的產品不會進top 20
- 部分沒有金額與13碼的產品有極高的點閱率/不重複點閱率，且價錢適中 (合理懷疑會進top 20清單)

Thanks :)