智慧型租屋物件決策支援系統

指導老師：蔡智勇老師、黃登揚老師

組員：陳俊邦、任樹菁、許博喻、吳文中、蔡忠良、黃詰凱、林幸怡、蘇筱涵、張景翔

租屋需求
逐年提升

內政部統計，
2017年
租賃市場規模
約 98.5萬戶

相較 2010年的
80.8萬戶，成長
21%；乘上每戶
平均人口 2.69，
等於每年265萬
人在外租屋生活

加上2017年教
育部統計的外
宿學生人數 30
萬，現今有租
屋需求的人口
數約300萬人

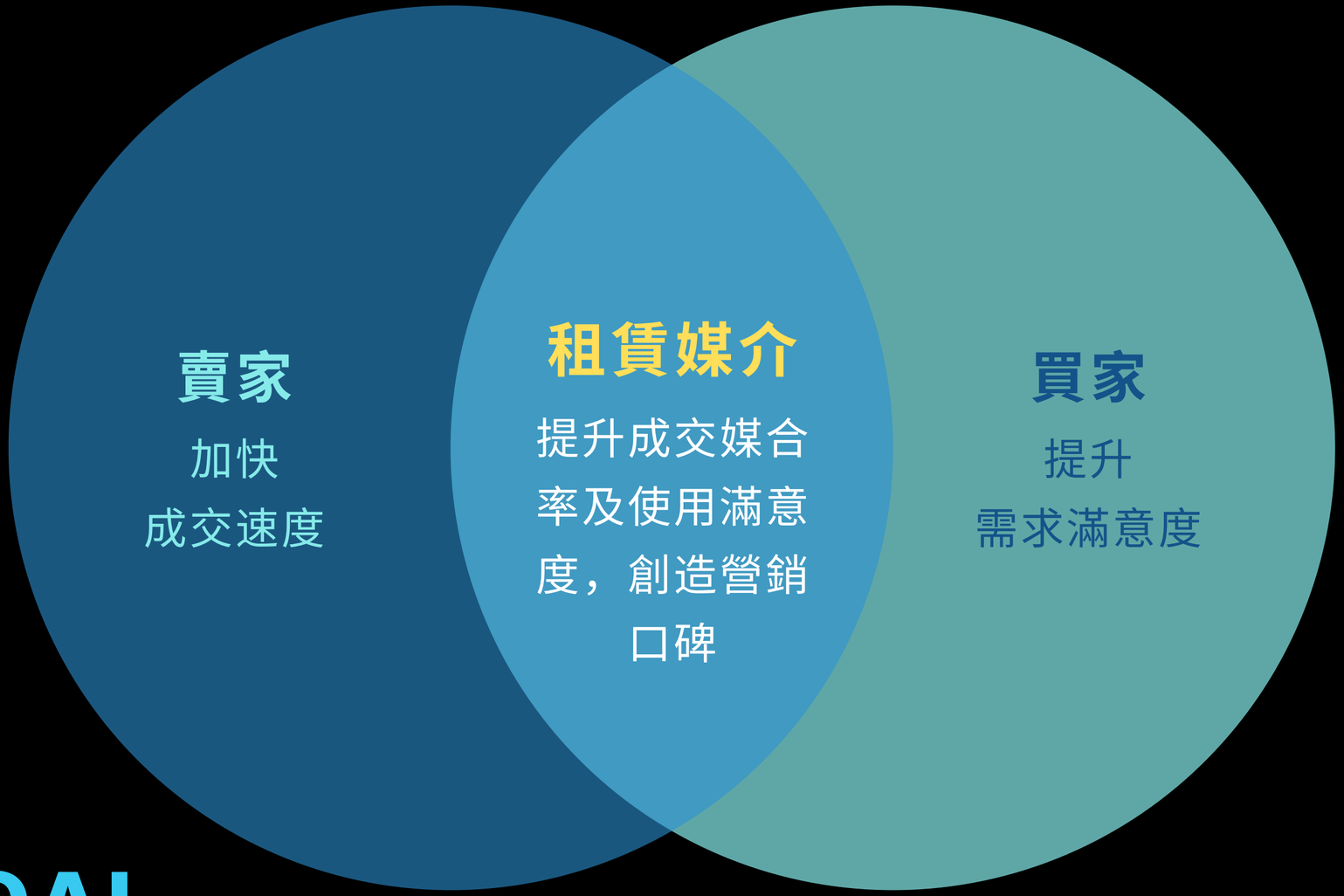
佔全台人口
約八分之一

研究動機 RESEARCH MOTIVATION

房價高漲的年代，租屋已成為越來越多人的居住型態，尤以都市人口居多。希望透過大數據的相關分析技術進行系統化的研究，讓租屋市場的使用者獲得更清晰有效的資訊，改善出租物件的競爭力！

「591租屋網，
出租就是
快！」

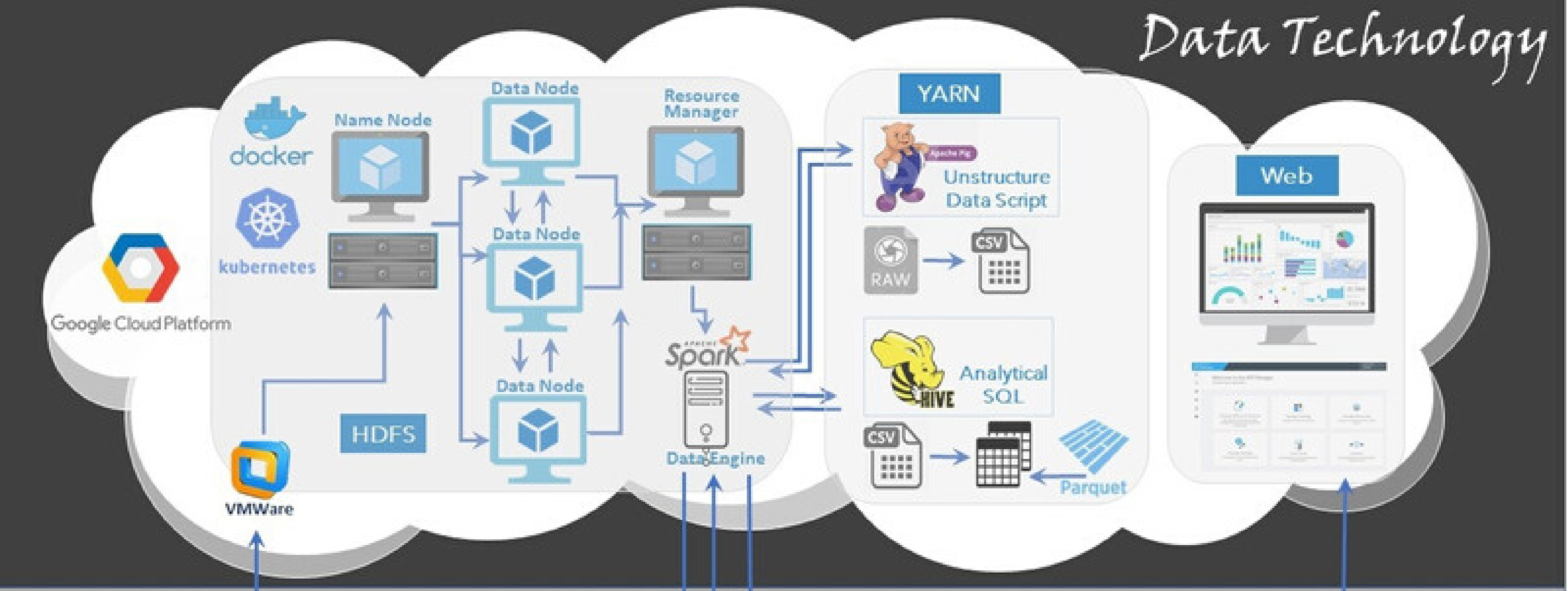
專案目標
PROJECT GOAL

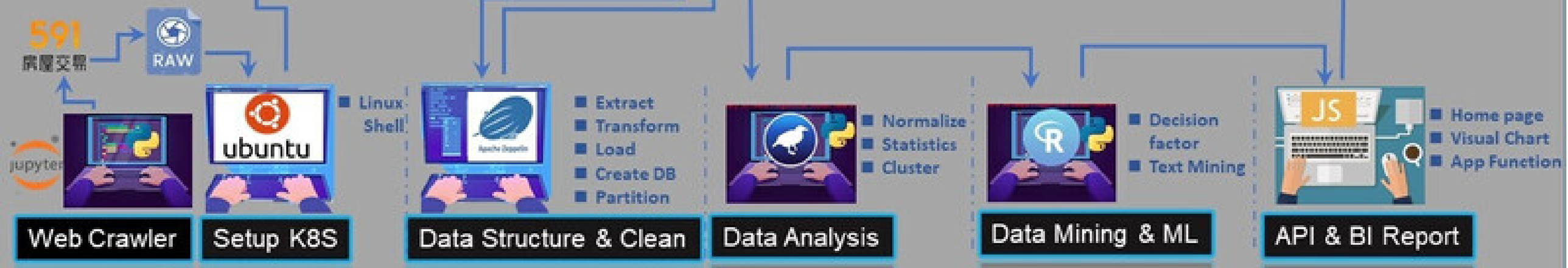


系統架構

System Architecture Diagram

Data Technology





雲端運算
CLOUD COMPUTING

PIG

資料清理，簡化清理資料的繁複指令

HIVE

建構在Linux介面下的SQL程式

PARQUET

新一代的儲存資料格式

ZEPPELIN

多功能網頁式開發工具

開放資料
Open Source

台灣最大網路租屋平台
591房屋交易網
2018年7月~2019年10月
共橫跨16個月份交易資料

- 基本交易資訊
- 房東提供之租屋現況資訊
- 租屋格局
- 周邊交通設施、周邊生活機能
- 額外收費欄目

☆ 資料集來源：開放台灣民間租屋資料
https://g0v.hackpad.tw/Ih7Jp4pUD5y

爬蟲資料
Web Crawler

使用python Beautiful Soup 套件
運用爬蟲技術取得下列資料

- 物件收藏數
- 手機瀏覽量、電腦瀏覽量
- 照片總數
- 標題內文與文章內文

資料清洗過程

DATA PROCESSING

1,723,253筆

85欄

262,460筆

原始資料
爬蟲資料
結合

73欄

262,460筆

無法分析
欄位剔除

67欄

256,932筆

空值處理

67欄

236,968筆

數值資料
常態化處理

68欄

234,432筆

異常值處理

79欄

234,432筆

欄位擴充

台中市資料

57欄

98,292筆

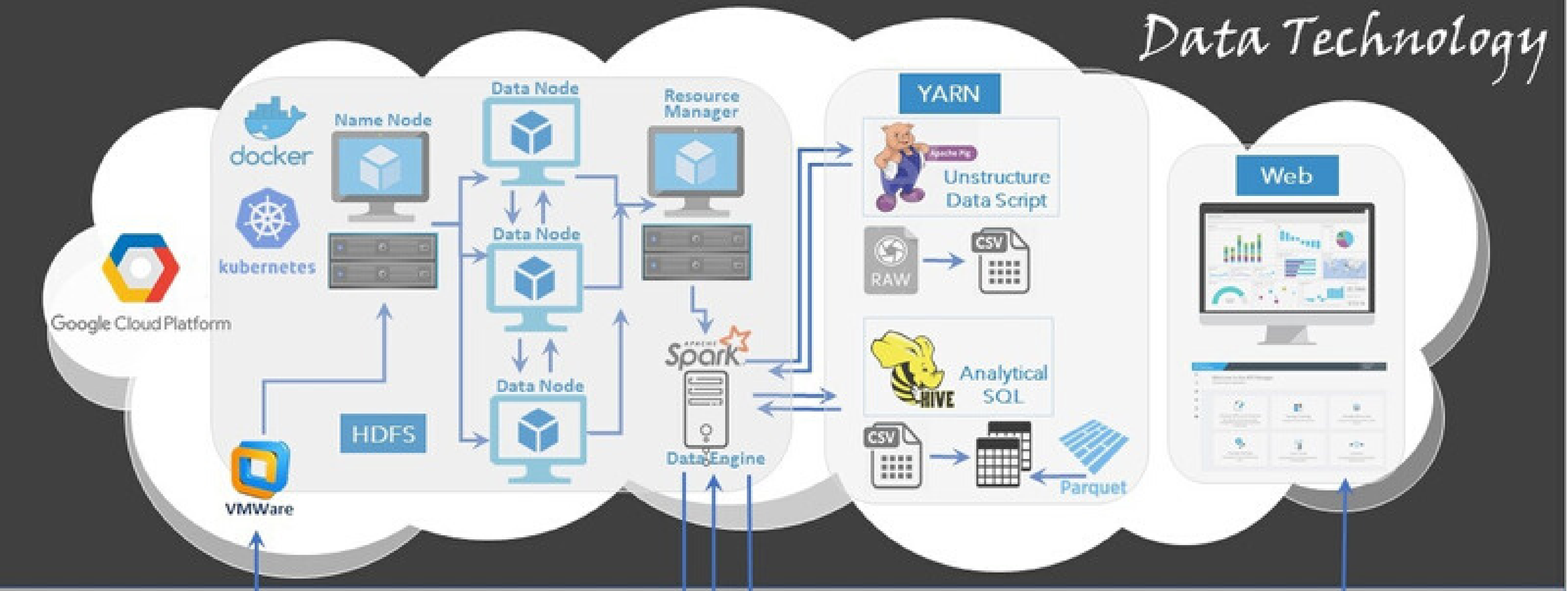
鎖定分析範圍

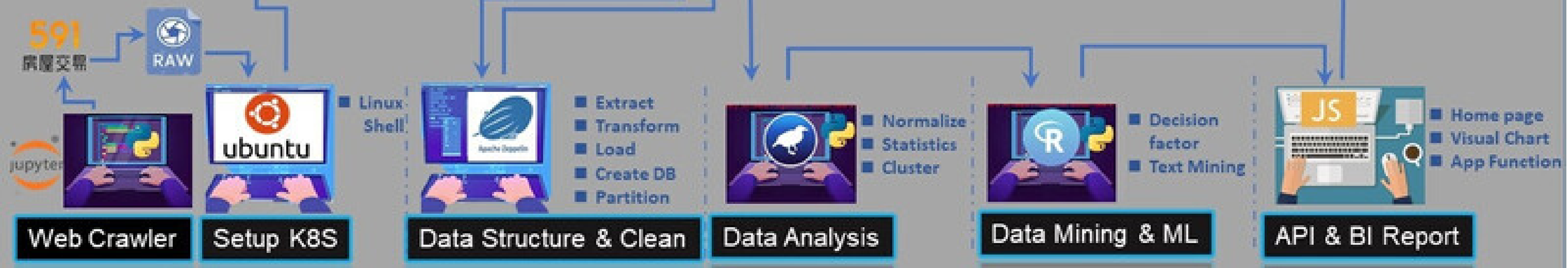
SQL

PIG

Pandas

Data Technology





預處理

- 將公開資料與爬蟲資料合併→串聯物件資訊及網頁資訊
- 全部皆為已出租資料→分析出租天數

Zeppelin Notebook Job

591_raw_pig

```
f19 = LOAD '/dataset/591/raw/dealData/19.csv' USING PigStorage(',');
f20 = LOAD '/dataset/591/raw/dealData/20.csv' USING PigStorage(',');
f21 = LOAD '/dataset/591/raw/dealData/21.csv' USING PigStorage(',');
f22 = LOAD '/dataset/591/raw/dealData/22.csv' USING PigStorage(',');
f23 = LOAD '/dataset/591/raw/dealData/23.csv' USING PigStorage(',');
f24 = LOAD '/dataset/591/raw/dealData/24.csv' USING PigStorage(',');
f25 = LOAD '/dataset/591/raw/dealData/25.csv' USING PigStorage(',');
f26 = LOAD '/dataset/591/raw/dealData/26.csv' USING PigStorage(',');

all_f = UNION f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11,f12,f13,f14,f15,f16,f17,f18,f19,f20,f21,f22,f23,f24,f

--刪除 網址為空直 及 成交天數非數值
all_n = FILTER all_f BY ($0 IS NOT NULL) AND ($3 >=0);
--刪除網址重複紀錄 利用 UniqueID, GROUP及 MAX 處理
all_id = FOREACH all_n GENERATE UniqueID() as id,$0,$1,$2,$3,$4,$5,$6,$7,$8,$9;
all_g = GROUP all_id BY $1;
all_c = FOREACH all_g {
    all_c1 = FOREACH all_id GENERATE id;
    GENERATE group,MAX(all_c1);
}
--只保留每一網址 UniqueID最大值的紀錄
all_d = JOIN all_id BY id,all_c BY $1;
--匯出整理後紀錄
RMF /dataset/591/out/allDD.csv;
```

Zeppelin Notebook Job

591_Q_pig

```
logs = LOAD '/dataset/591/out/allPD.csv';
logs_grouped = GROUP logs ALL;
number = FOREACH logs_grouped GENERATE COUNT_STAR(logs);
dump number;
```

(1723253)

Took 1 min 20 sec. Last updated by bigred at November 28 2019, 11:20:47 AM.

```
log  = LOAD '/dataset/591/out/allID.csv';
lo  grouped = GROUP logs ALL;
nu er = FOREACH logs_grouped GENERATE COUNT_STAR(logs);
dump number;
```

(262460)

資料觀察

DATA OBSERVATION

共85欄

時間地點

物件編號
租屋平台
物件網址
上架時間
最後更新
所在縣市
鄉鎮市區
經緯度
出租狀態
出租時間
出租天數

租房支出

月租金
押金類型
押金月數
押金金額
需管理費
月管理費
提供車位
需停車費
月停車費
每坪租金

物件型態

建築類型
物件類型
自報頂加
所在樓層
建物樓高
距頂樓數
坪數
陽台數
衛浴數
房數
客廳數
格局編碼

額外費用

電費
水費
瓦斯
網路
第四台

生活機能

學校
公園
百貨公司
超商
傳統市場
夜市
醫療機構
捷運站數
公車站數
火車站數
高鐵站數
公共自行車數

限制&資訊

有身份限制
有性別限制
性別限制
可以開伙
可養寵物
有產權登記
刊登者類型
刊登者編碼
仲介資訊

家具提供

床架
書桌
椅子
電視
熱水器
冷氣
沙發
洗衣機
衣櫃
冰箱
網路
第四台
天然瓦斯

網頁訊息

資料來源
id
網址
地址
標題
物件成交天數
物件收藏數
總瀏覽人數
電腦瀏覽人數
手機瀏覽人數
照片總數
內文

無效欄位的剔除

根據欄位類型及分析目的，找出哪些欄位在本次研究標的中不具有分析價值

剔除12欄

欄位類型	不符合分析目的
物件編號 租屋平台 物件網址 爬蟲來源平台 編號 網址 刊登者編碼 格局編碼	物件最後更新時間 房屋出租狀態 出租大約時間 出租所費天數

空值欄位的處理

根據空值的數量多寡及佔比，在觀察網站填寫狀況後，刪除或調整空值。

01 **空值太多：佔比>5%**

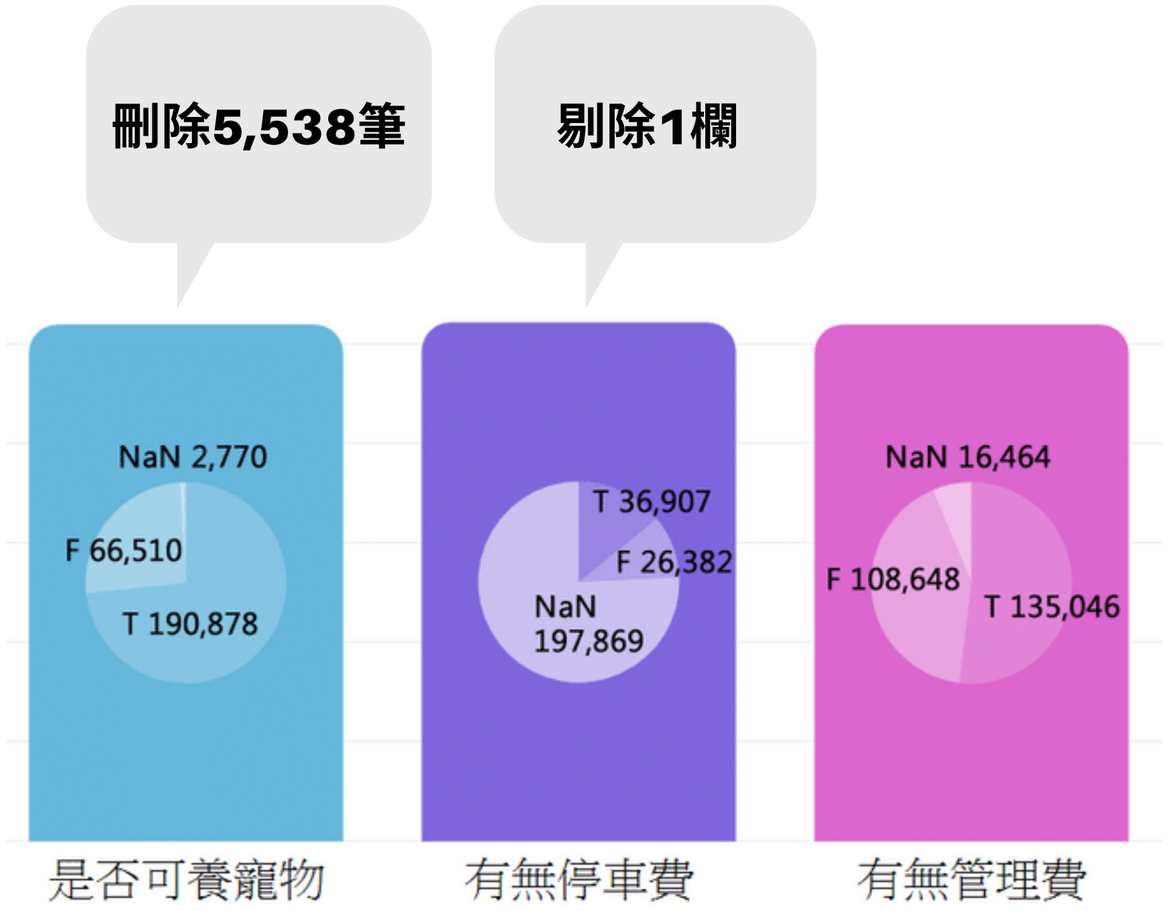
- 需停車費

02 **空值偏多，經觀察後將空值意義合理化**

- 需管理費
  - →"T"：要管理費
  - →"F"：不需管理費
  - →"NULL"：面議

03 **空值少，空值資料整筆刪除**

- 可養寵物→刪除2,770筆
- 可開伙→刪除2,758筆



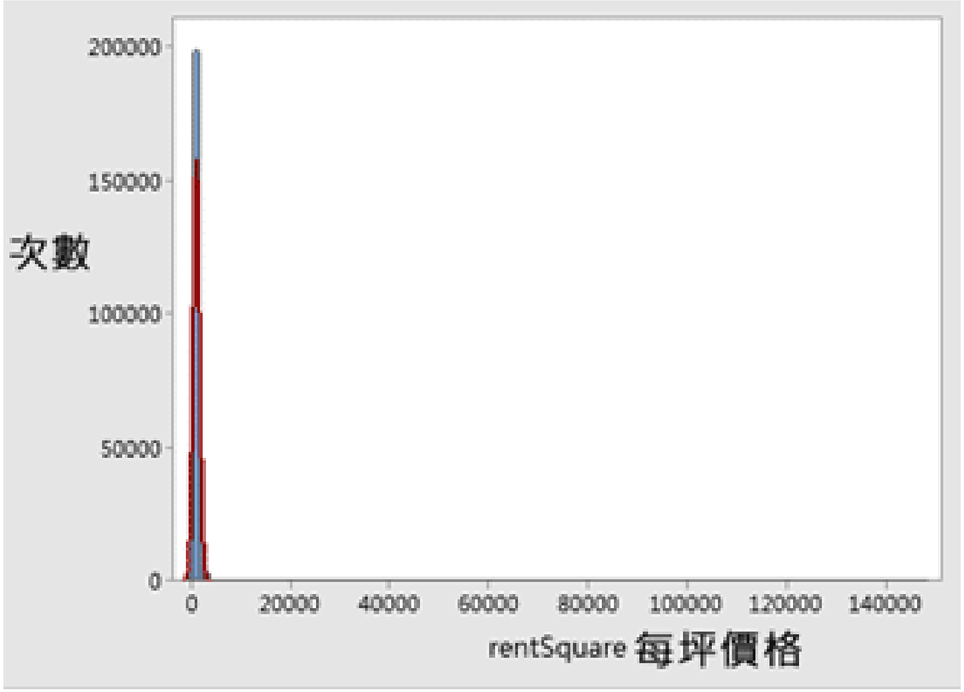
數值資料正規化處理

針對數值差異大的欄位 → 取 ln 後取平均數 ±2 * 標準差

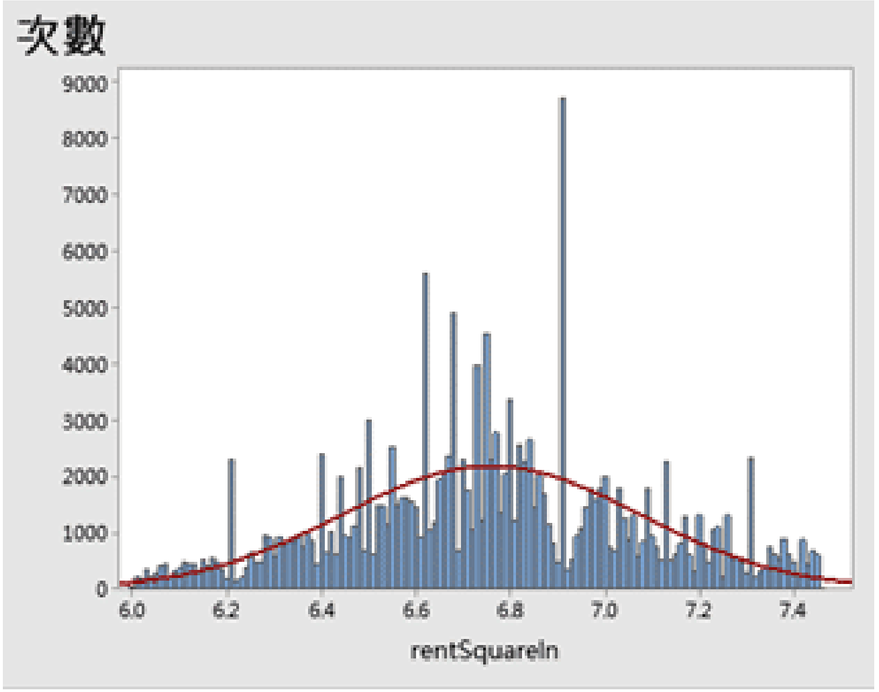
- 避免單位、數字大小差異影響
- 刪除離群值

刪除19,964筆

Before



After



異常值處理

排除因爬蟲的網站、原物件資訊有問題的資料刪除

刪除2,545筆

01

成交天數

因591網站有部分舊網頁資料有誤，使爬蟲內容出現異常
→成交天數 > 60 刪除
共2,535筆

02

成交天數

依591分群規則分群
→建立預測標的
3天內→快速成交
7天內→較快成交
15天內→正常成交
15天以上→慢速成交

03

瀏覽數

瀏覽數 = 0
卻是已出租物件
不合哩，故刪除
共10筆

欄位擴充

將原有資料中的欄位，以整合、數學計算等方式，合併產生新欄位

增加7項欄位

迎毗設施

學校
公園
百貨公司
超商
傳統市場
夜市
醫療機構

是否有填格局房數

陽台數
衛浴數
房數
客廳數

有幾種額外費用

電費
水費
瓦斯
網路
第四台

是否有填大眾交通

捷運站數
公車站數
火車站數
高鐵站數
公共自行車數

提供家具數

床架
書桌
椅子
電視
熱水器
冷氣
沙發
洗衣機
衣櫃
冰箱
網路
第四台
天然瓦斯

物件刊登月份

物件刊登年份

物件刊登日期

欄位擴充

將原有資料中的欄位，以整合、數學計算等方式，產生新欄位

增加5項欄位

每日瀏覽數

瀏覽數/交易天數
=每日瀏覽數

總瀏覽人數
電腦瀏覽人數
手機瀏覽人數

文字長度

標題長度
介紹文長度

縮小研究範圍

在時間及成果上，縮小地區標的→台中

並選出可以進決策樹的因子

刪除136,140筆

剔除22欄

成交天數

3天內→快速成交　7天內→較快成交

15天內→正常成交　15天以上→慢速成交

供網頁應用資料	留作其他分析標的 (拿來預測天數會有影響)	文字探勘用資料	針對可以細項分析的先行拉掉	已運用新欄位的刪去舊欄位
地址 縣市 鄉鎮市區 東西經 南北緯 仲介資訊	總瀏覽人數 電腦瀏覽人數 手機瀏覽人數 每日總瀏覽人數 每日電腦瀏覽人數 每日手機瀏覽人數 成交天數 收藏數	標題 內文	有幾種額外費用 提供家具數	陽台數 衛浴數 房數 客廳數

各因子對出租天數的影響

運用統計檢定，證實各因子皆與出租天數有半結構化上的關係，
但因為因子眾多，所以使用決策樹抓出決策因子

Rows: Cluster Columns: publisherType

	0	1	2	All
cluster0	613	5952	18875	25440
	2319	5928	17193	
cluster1	880	2899	8258	12037
	1097	2805	8135	
cluster2	1842	3080	7925	12847
	1171	2994	8683	
cluster3	1797	1190	2998	5985
	545	1395	4045	
All	5132	13121	38056	56309

Cell Contents: Count
Expected count

Pearson Chi-Square = 5092.935, DF = 6, P-Value = 0.000

Rows: Cluster Columns: manageFee

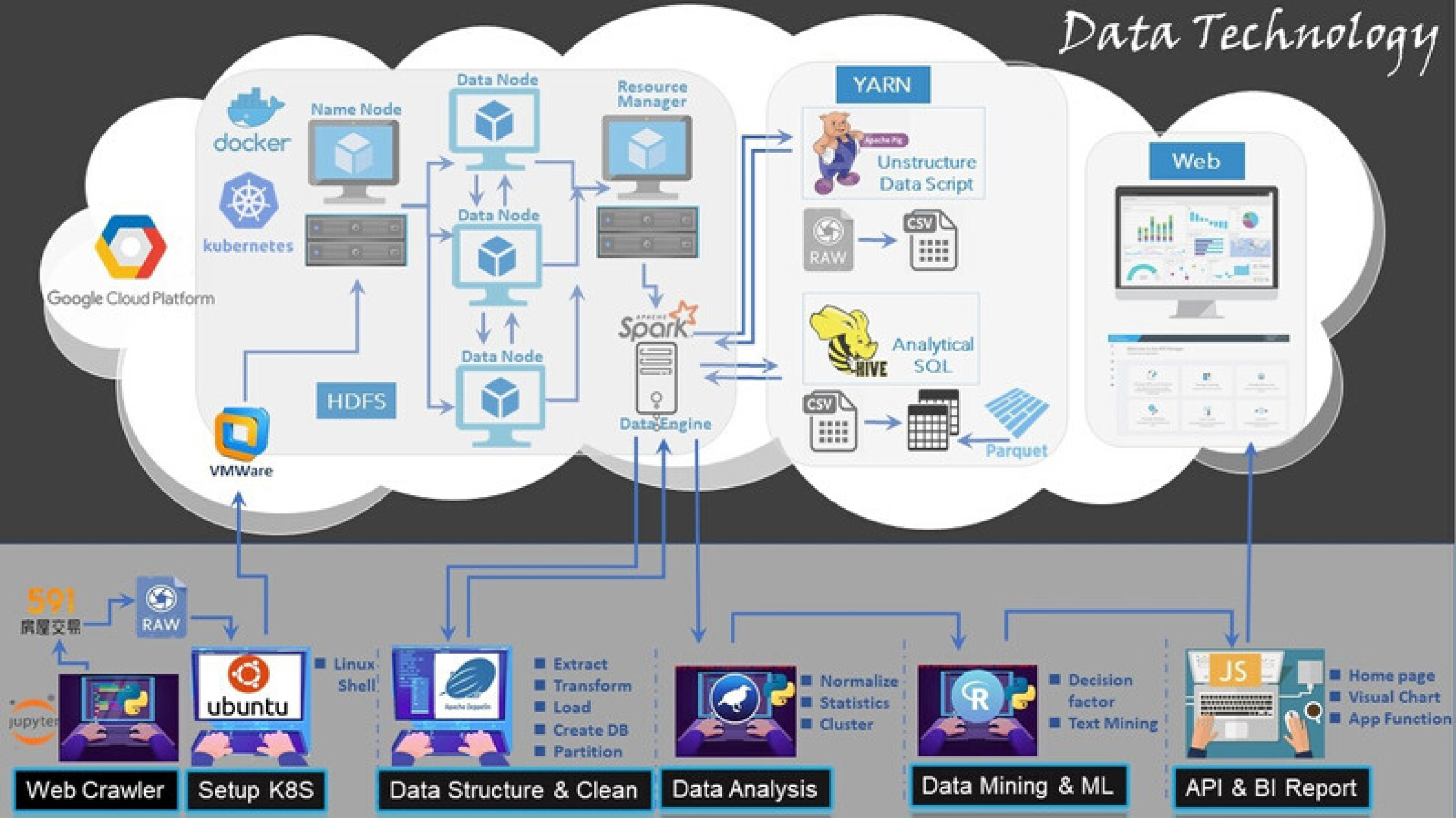
	'yMF '	nego	nMF	All
cluster0	17450	213	7777	25440
	15839	260	9341	
cluster1	6760	209	5068	12037
	7494	123	4420	
cluster2	8035	76	4736	12847
	7999	131	4717	
cluster3	2813	77	3095	5985
	3726	61	2198	
All	35058	575	20676	56309

Cell Contents: Count
Expected count

Pearson Chi-Square = 1279.366, DF = 6, P-Value = 0.000

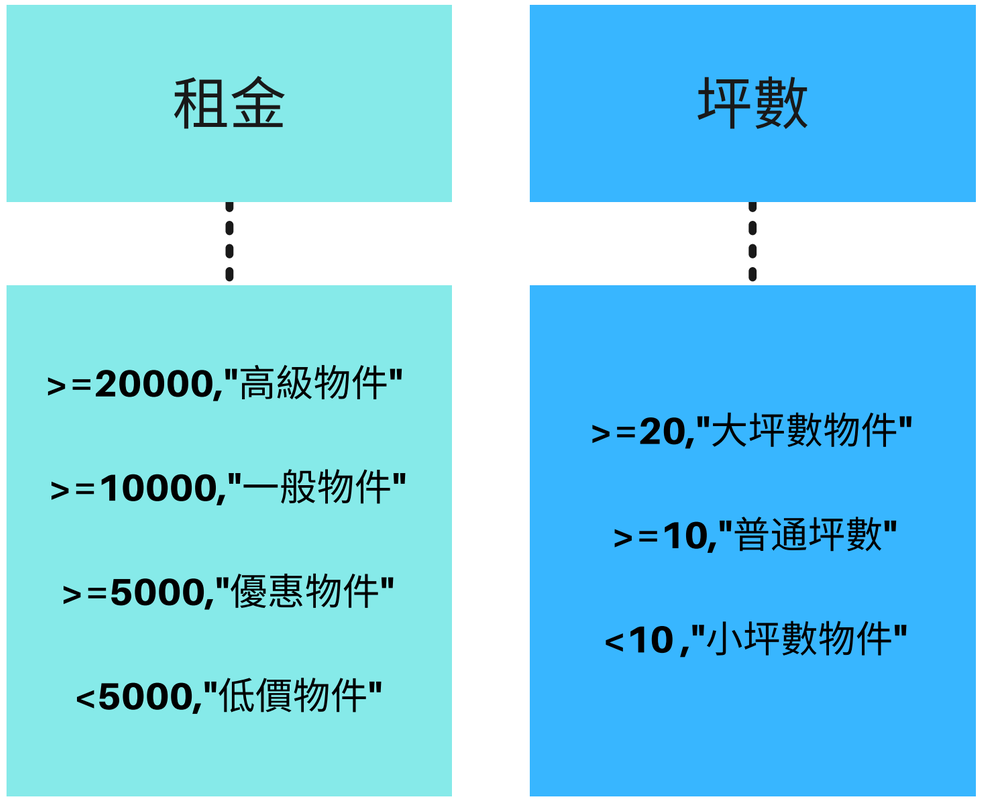
P-value < 0.05 ，由卡方檢定中可以得知，上述兩種因子分別與Y之間有顯著關係

Data Technology



重點數值欄位分群

針對在管理意涵影響力大的欄位：租金、坪數，將其分群類別化，以利決策樹的建立



決策樹與決策因子-C4.5

IF internetBill = F AND cableBill = F
Then nD
IF bed =T AND airCondition=T AND chair=T AND internet=T AND TV=T AND refrigerator=T AND washer=T AND cable=T Then qD

IF cable = T AND refrigerator=T AND sofa=T
Then nD
IF Rent = Expensive AND Square = Normal Place Then sD

IF Rent = Expensive AND Square = Big
Then sD

決策樹與決策因子-RandomTree

IF Rent = Cheap AND Square = Big
Then qD
IF cableBill = F
Then nD
IF cableBill = T AND gasBill = T Then sD

IF refrigerator=F AND waterHeater=T AND sofa=F
Then sD
IF depStore =T AND hospital=T AND cooking=T AND market=T AND nightmarket=T AND trafficYN=T AND pet=T AND parkeing=T Then qD

IF depStore =T AND hospital=T AND cooking=T AND market=T AND nightmarket=T AND trafficYN=T AND pet=T AND parkeing=F Then nD

IF depStore =F then nD

IF depStore =T AND hospital=F then nD

IF depStore =T AND hospital=T AND cooking=F then nD

決策樹與決策因子

C4.5&RandomTree共同因子 & 說明

從決策樹中觀察到的趨勢→

●資料填寫越完整，越早出租出去的機會越大。
例如**1**：床、冷氣、椅子、網路、電視、冰箱、洗衣機、第四台都附的話，有機會在**3**天內出租出去
例如**2**：有提供附近 學校**&**公園**&**百貨公司**&**便利商店**&**市場**&**醫院、診所**&**大眾交通工具 的資訊，且可以 開伙**&**養寵物 的話，有機會在**3**天內出租出去

●在額外費用中→
水費大部分物件都有附，所以沒附的話有 機會要**15**天以上才能出租出去。
另外如果收了第四台費用，建議不要再收天然氣費用，否則相較於沒收天然氣費用的，會花較多時間才能出租出去

●如果租金昂貴的物件，無論坪數大小，均會比較慢成交，所以有較高機會在大於**15**天後成交
如果是比較便宜的物件而且坪數又大，會有**4**成的機率快速成交 **-->3**天內成交

決策樹與決策因子

C4.5&RandomTree共同因子 & 說明

租金便宜&坪數大

理想物件十分搶手，有都
在短時間內出租的趨勢

租金昂貴

只要租金貴，
不論坪數大小成交天數都
相對耗時。

額外費用慎選

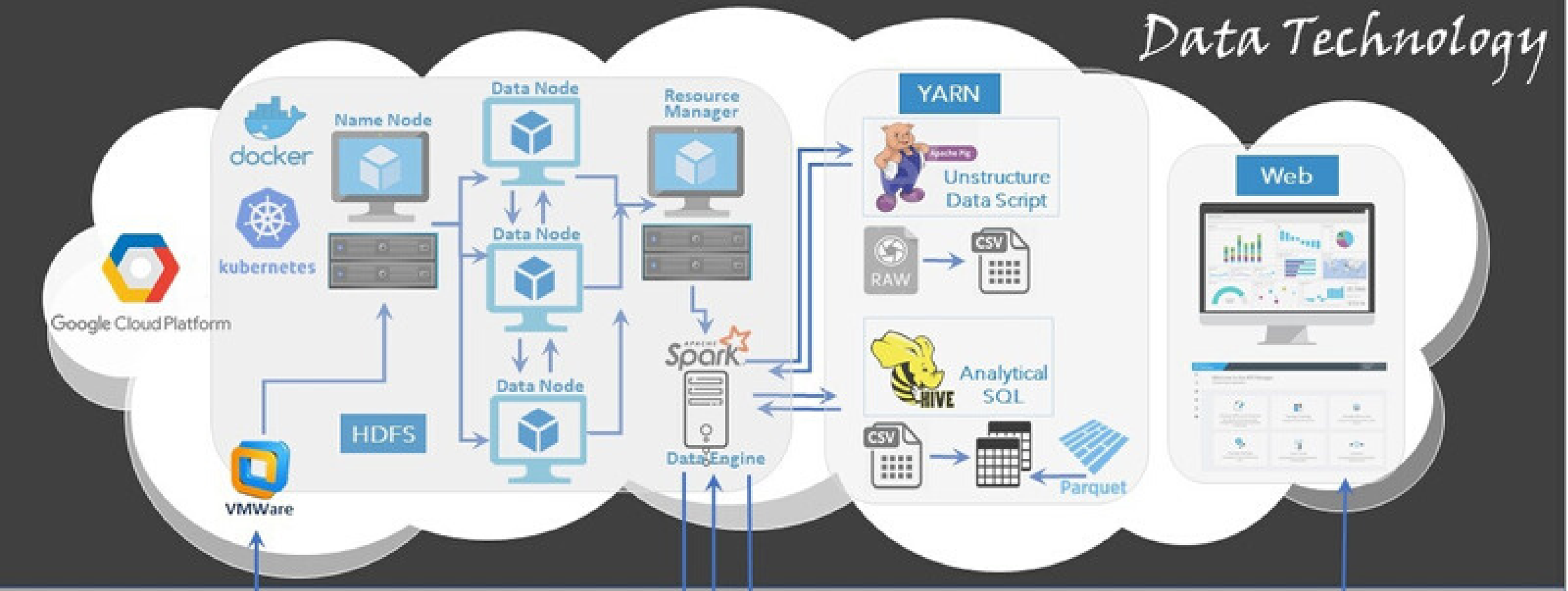
水費
第四台費
天然氣費
顯著影響成交所需天數

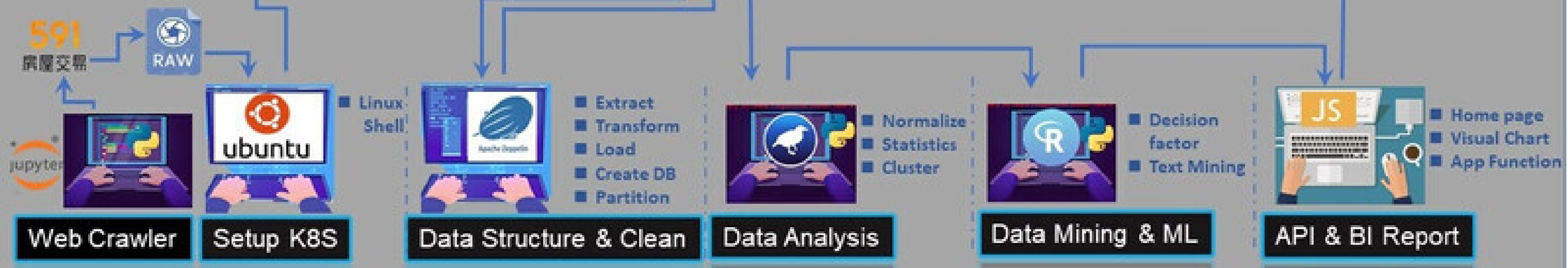
刊登資料完整性

資料填寫越完整，越早出
租出去的機會越大。

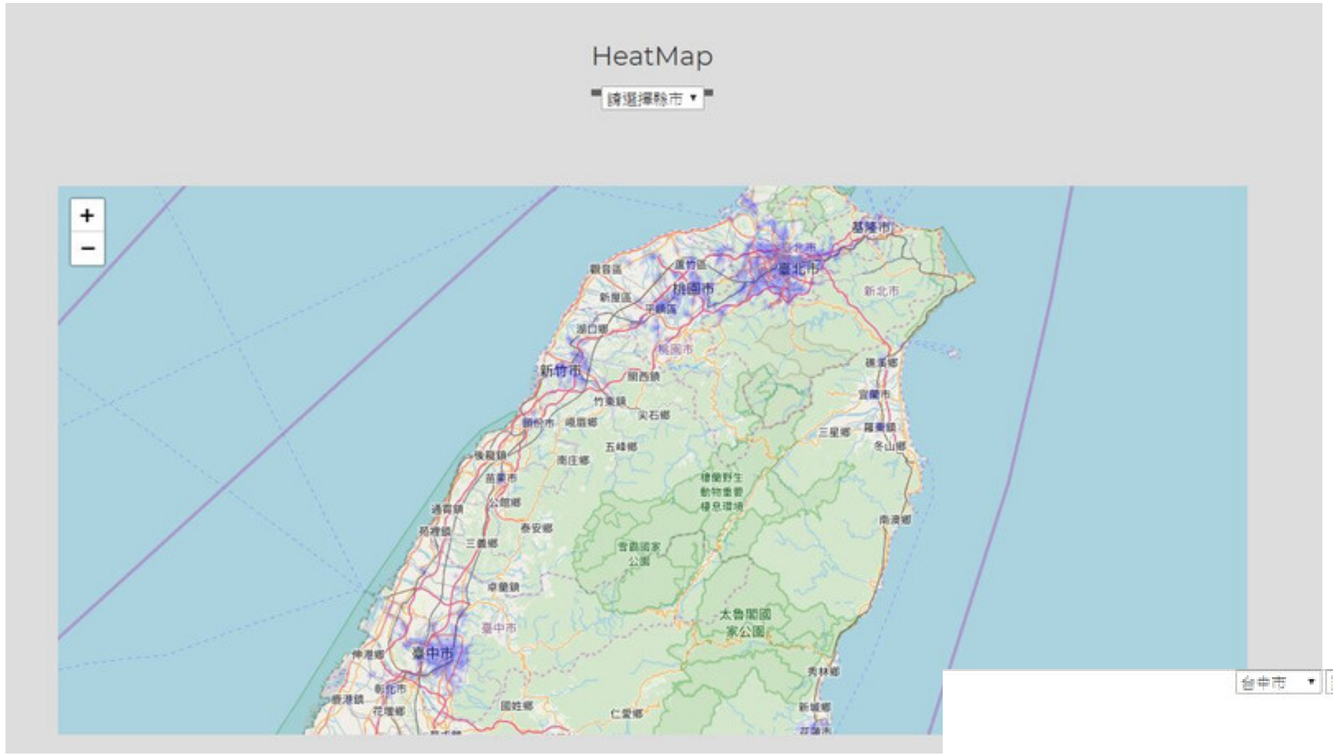
文字探勘
TEXT MINING

Data Technology



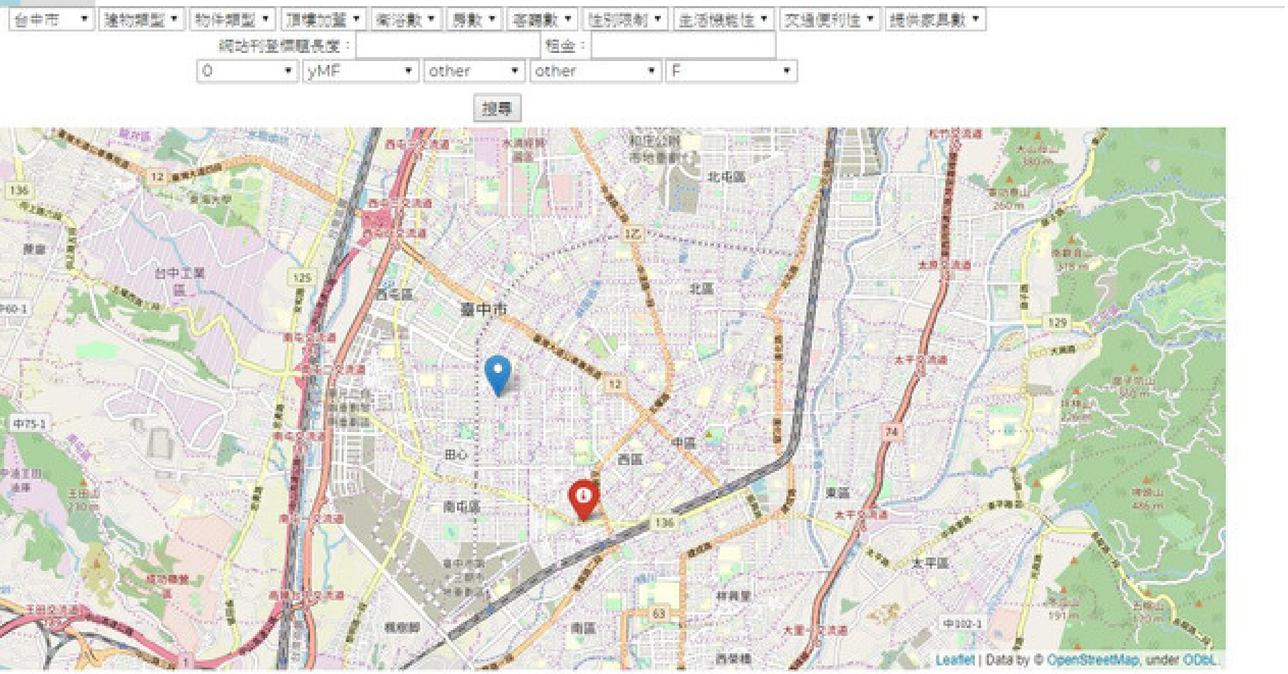


網頁連結 WEB LINK



透過網頁呈現台灣租屋市場的熱度分佈圖

依類型分類後輸入物件特徵，可提供物件快速成交的關鍵因子



建議和挑戰
SUGGESTIONS & CHALLENGES

智慧決策系統

建立智慧型租屋物件決策系統，讓房東查詢房屋物件分析結果，根據物件周邊競爭程度差異，輔助房東進行判斷是否該提供設備或減低租金等方式，加速成交速度。

優化上架資訊

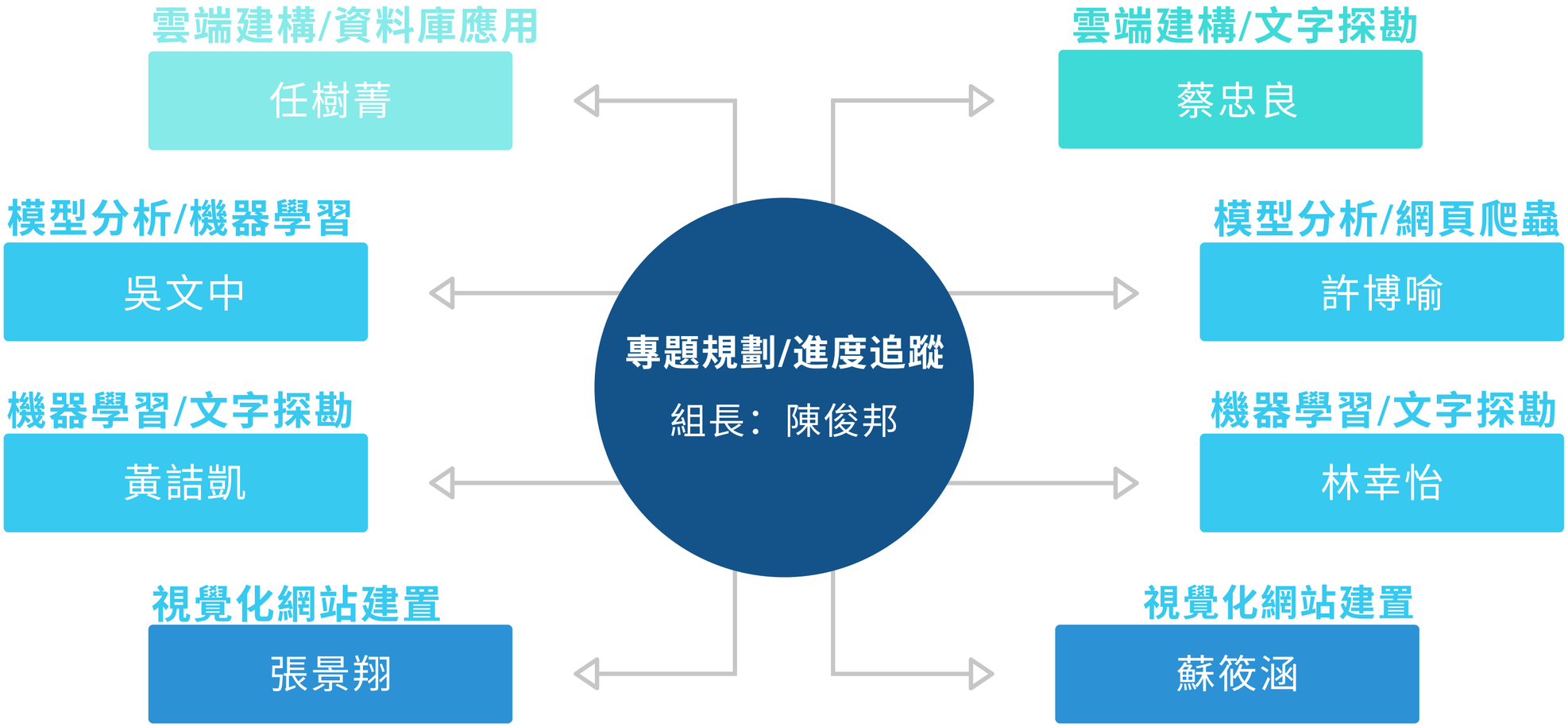
針對房東租屋條件，有10個關鍵變數可提供房東改進。建立租屋介紹時的標題與內文高吸引力字詞推薦，可做為房東刊登物件時的參考。

未來相關應用

本研究目前僅針對台中地區做分析，在未來可以根據不同區域進行更深入的研究。

- 租金預測
- 瀏覽數預測
- 手機/PC瀏覽量比較.

成員介紹
MEMBER INTRODUCTION



團隊分工
TEAM DIVISION

專題操作	陳俊邦	任樹菁	許博喻	吳文中	蔡忠良	黃詰凱	林幸怡	蘇筱涵	張景翔
網頁爬蟲	★		★	★	★	★	★	★	★
資料清洗	★	★	★	★	★	★	★	★	★
雲端系統建構		★			★				
機器學習			★	★		★	★		
文字探勘	★				★	★	★		
網站製作	★							★	★

感謝您的聆聽與指教!