## 面試前題目：

感謝您對 LINETV 的 Data Engineer 有興趣，若對下述內容有任何問題或建議，請隨時聯繫我們。

## 一、事件與資料庫

請考慮下述需求，請您回答後續的問題

* 首頁戲劇版位每天的曝光數、點擊數以及點擊率
* 使用者的基本資料（只有性別和年齡）
* 想知道不同性別的使用者點擊率是否有差別？

1. App 需要在「什麼時機」、傳送「什麼事件及內容」或資料到後端伺服器或資料庫？每個事件請分別舉一個例子。
2. 曝光數，可以在該影片被點擊時，發送一筆資料，該資料內容包含  
   {  
   “UUID for user” (使用UUID的原因是點擊者有可能為訪客，不一定是會員),

“使用者所在頁面停留時間”,  
“使用者所在頁面URL”,  
}

1. 若為網站會員，則在使用者登入後，任何該使用者的活動紀錄之回傳資料都包含該使用者的ID number，之後若要做數據分析可直接透過該ID number從資料庫中得知使者性別、年齡。

(c.) 若為網站會員，則在使用者登入後，紀錄該使用者(ID number)訪問的每部戲劇的紀錄。之後在以使用者的ID number查詢資料庫中該名使用者的活動紀錄加以分析

1. 請問存放上述資料的表格和欄位會如何設計？  
   uuid, user\_id, gender, stay\_url, stay\_time
2. 上述資料適合存放在哪（舉您熟悉的元件即可）？為什麼？  
   適合儲存在像是MySQL, PostgreSQL之類的關聯式資料庫
3. 承上題，請以您規劃的表格以及欄位，製造出 10 筆假資料（以表格為單位即可，不用每個事件都產 10 筆資料）
4. 承上題，根據上述假資料，請您使用 Standard SQL Query 比較不同性別每天的點擊率？
5. 承上題，您會如何用視覺化的方式呈現上述問題？（簡單畫圖表示即可）

## 二、程式與系統架構

假設我們今天某個收到的資料的正常內容如下：

{"ad\_network": "FOO",

"date": "2019-06-05",

"app\_name": "LINETV",

"unit\_id": "55665201314",

"request": "100",

"revenue": "0.00365325",

"imp": "23"}

但是很不幸的，我們系統常常收到奇怪的資訊，可能包括缺漏值、也有可能格式不對。

1. 請寫一個 Python Function 來驗證收的上述資料格式是否正確。（程式碼請提供 github 連結）
2. 承上題，請寫一個或多個（有需要的話） Unittests 來驗證上述驗證程式是否正確。（程式碼請提供 github 連結）
3. 當您寫完程式後，要怎樣透過 git 指令推送到公司的 repo？（當這隻程式為多人開發時）
4. 您會設計怎樣的 Data Pipeline 來接受 App 傳來的資料、檢查資料格式、以及整理到資料庫/資料倉儲？（舉您熟悉的架構元件即可）
5. 在上述架構中，您會將這隻驗證程式放在哪裡來執行？
6. 如果您需要負責監控系統，會觀察哪些關鍵指標來判斷這套系統是否正常運行？
7. 當流量突然暴增 10 倍時，您上述的系統設計撐得住嗎？如果撐不住您會如何調整架構？

## 三、Docker

假設今天要上架一個服務，需要使用container的服務

1. 當寫好了repo 要如何把它變成docker 的container?
2. 要進入container的語法?
3. 當遇到container發生異常會怎麼處理?

## 四、專案管理

回到第一個問題（事件與資料庫設計）：

今天當您是開發負責人時：

1. 您會如何安排進度與時間？（在正常合理的工作時間下）
2. 如何定義專案完成？
3. 以目前的規劃來說：
   1. 您覺得自己可以做得不錯的地方有哪些？
   2. 您覺得需要加強/或需要他人協助的地方有哪些？

**感謝您的時間與耐心，請問我們的題目或面試流程有沒有什麼可以做得更好的地方？**