Data Pipeline Matters
-- 以 Tracking Pixel 為例

Data Pipeline Matters !!
Take Tracking Pixel as an Example

Jazz Yao-Tsung Wang

Data Architect of TenMax.io
Initiator of Taiwan Data Engineering Association
Co-Founder of Taiwan Hadoop User Group

Shared at 2017-11-12 <2017 台灣資料科學年會 >

Hello!

I am Jazz Wang

Co-Founder of **Hadoop.TW**
Initiator of **Taiwan Data Engineering Association (TDEA)**
Hadoop Evangelist since 2008.
Open Source Promoter. System Admin (Ops).

- 11 years (2002/08 ~ 2014/02) **Researcher** in HPC field.
- 2 years (2014/03 ~ 2016/04) Assistant Vice President (AVP), Product Management of 'Big Data Platform Management Product'
- 1.5 years (2016/04 ~ Now) **Data Architect** of Real-Time Bidding

You can find me at `@jazzwang_tw` or
https://fb.com/groups/dataengineering.tw
https://slideshare.net/jazzwang

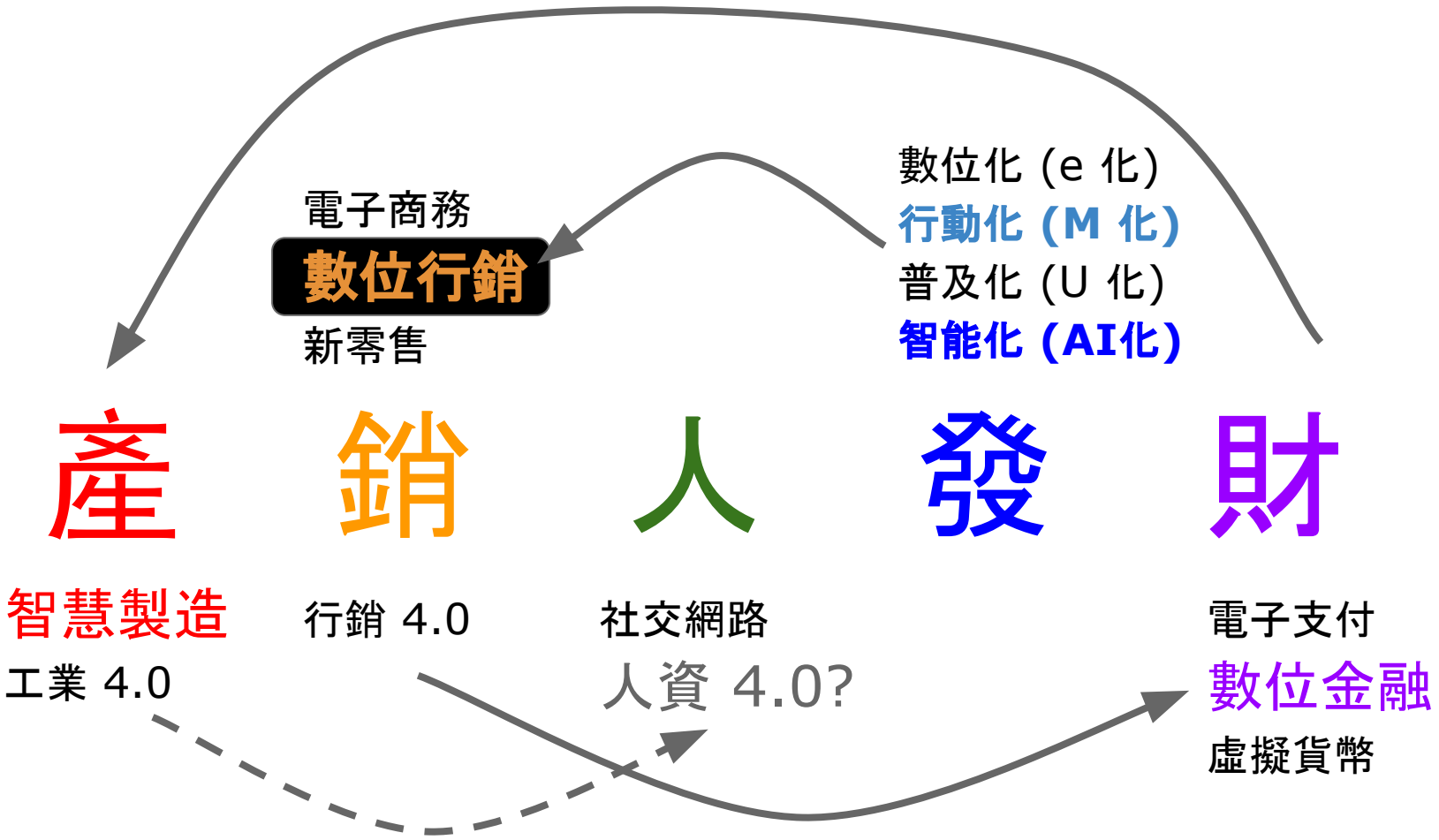
0.0 先暖場一下

畢竟不是每個人都了解線上廣告這個圈子
來點基本背景趨勢簡介

數位轉型

DIGITAL TRANSFORMATION (DX)

早就悄悄地進行著



廣告一直是支撐免費服務的基石

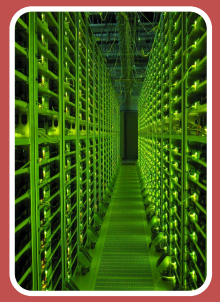
“

羊毛出在狗身上，

豬來買單！

線上廣告的五大技術特點

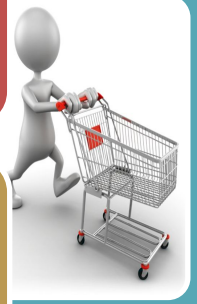
預估



技術和計算導向

- 較精細的受眾定向 – **更精準的廣告成效預估**
- 計算驅動的廣告決策與競價交易 – 最佳化能力
- 數位行銷：成本低，可高度客製化

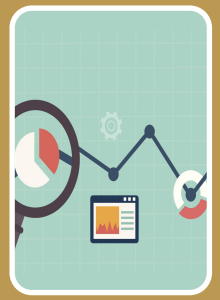
導購 vs 導流



媒體概念的多樣性

- 入口網站、垂直網站、搜尋引擎、電子商務網、優惠代碼網：越來越靠近使用者轉化的特徵
- 反思：ROI 越高，引導潜在客戶的能力卻降低

追蹤



效果的可量測性

- 可忠實呈現展示(Impression)與點擊(Click)
- 比較不同時期不同產品的點擊率絕對值沒意義
- 特定時期同類產品的點擊率差異比較才有意義

大數據

數據驅動的投放決策

- 機械化：電力 → 互聯網+：數據力
- DMP：對受眾貼標籤 Tagging
- DSP：依歷史投放結果，調整投放策略

素材與投放方式的標準化

- 標準化的驅動力：**受眾定向**與**程序化購買**
- 影音廣告的 **VAST 標準**
- 即時競價的 **OpenRTB 標準**

Tracking Pixel 是讓這一切發生的根本

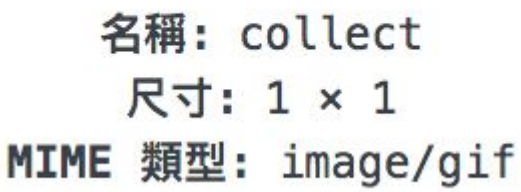
1.
Tracking Pixel
如何產生資料？

什麼是 Tracking Pixel ？

有哪些應用場景？

Tracking Pixel 會在哪裡產生哪些資料呢？

What is Tracking Pixel ？



- 追蹤像素(Tracking Pixel)
  - 一個大小 1 x 1 像素，通常是透明的圖片
  - 也稱為 web bugs, beacons, tracking bugs, page tags
  - 埋在網站或 e-mail 中，用來追蹤使用者行為跟線上廣告的成效
- 基本應用場景
  - E-mail (EDM) 開信率
  - 網站到訪率、跳出率、哪些頁面最熱門
  - 線上廣告的曝光率(Impression)、可視曝光率(Viewable Impression)、點擊率(Click)、到達率(Landing)、轉化率(Conversion)
- 進階應用場景
  - 搭配 cookie 可以做到更精準的受眾定向(Targeting)
  - 個人化瀏覽體驗(Personalized web experiences)
  - 跨網域 cookie syncing / matching (跨螢用戶追蹤的基礎)

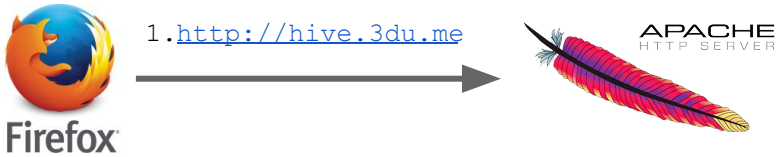
Tracking Pixel 的種類

- 實作上有兩種 Tracking Pixel
  - **Client Based Tracking** : 使用 cookie
  - **Server Based Tracking** : 不用 cookie

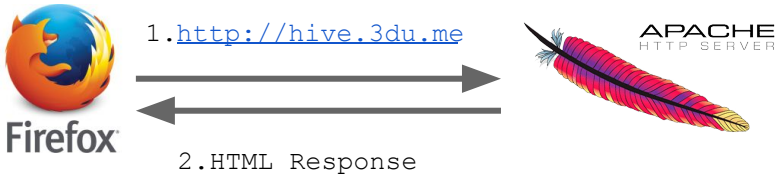
```
<img src="http://hive.3du.me/images/hive_labs.png?uid=XXXXXXXXXX">
```

- 常見實作
  - 使用第三方服務
    - Google Analytics
      https://analytics.google.com
    - Facebook Pixel
      https://www.facebook.com/business/help/651294705016616
    - Doubleclick Pixel Loader
      https://support.google.com/richmedia/answer/6187378
  - 自建服務
    - Piwik
      https://piwik.org/docs/tracking-api/

Google Analytics 的 Tracking Pixel 資料產生流程

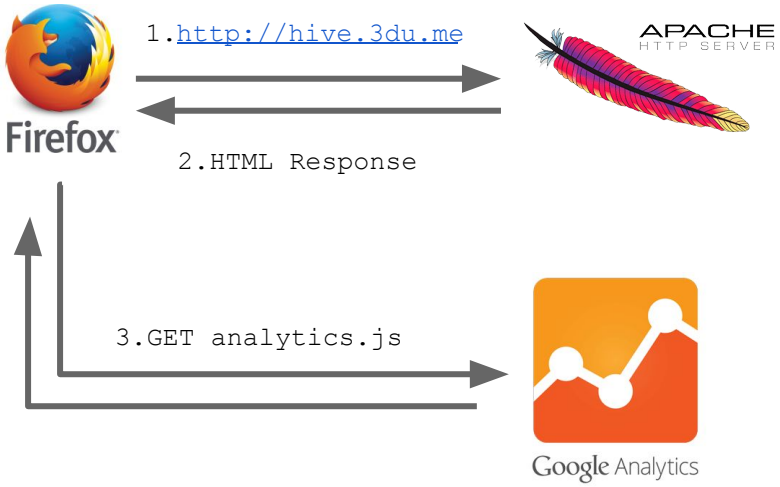


Google Analytics 的 Tracking Pixel 資料產生流程




```
51 <script>
52   (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
53   (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
54   m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
55   })(window,document,'script','//www.google-analytics.com/analytics.js','ga');
56
57   ga('create', '[redacted]', 'auto');
58   ga('send', 'pageview');
59
60 </script>
```

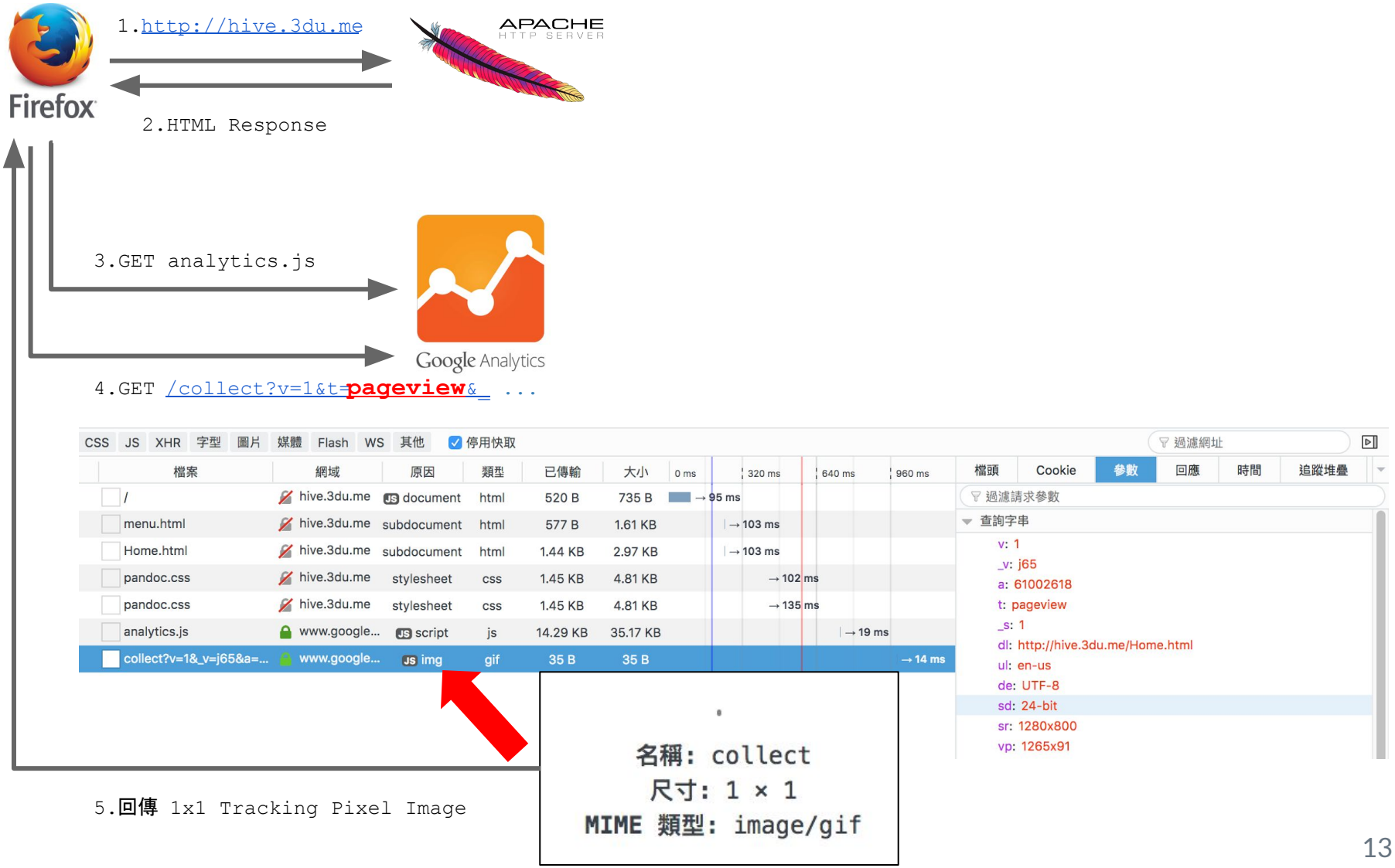
Google Analytics 的 Tracking Pixel 資料產生流程



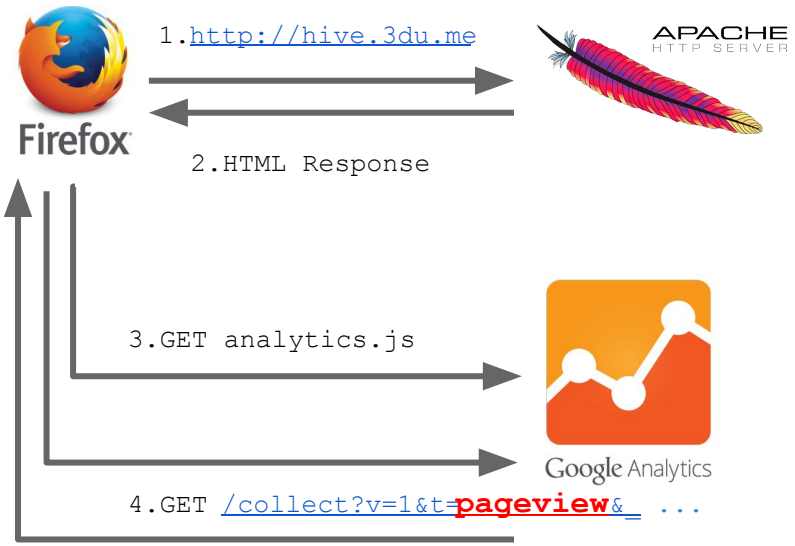

```
ga('create', '[redacted]', 'auto');
ga('send', 'pageview');

</script>
</div>
```

Google Analytics 的 Tracking Pixel 資料產生流程

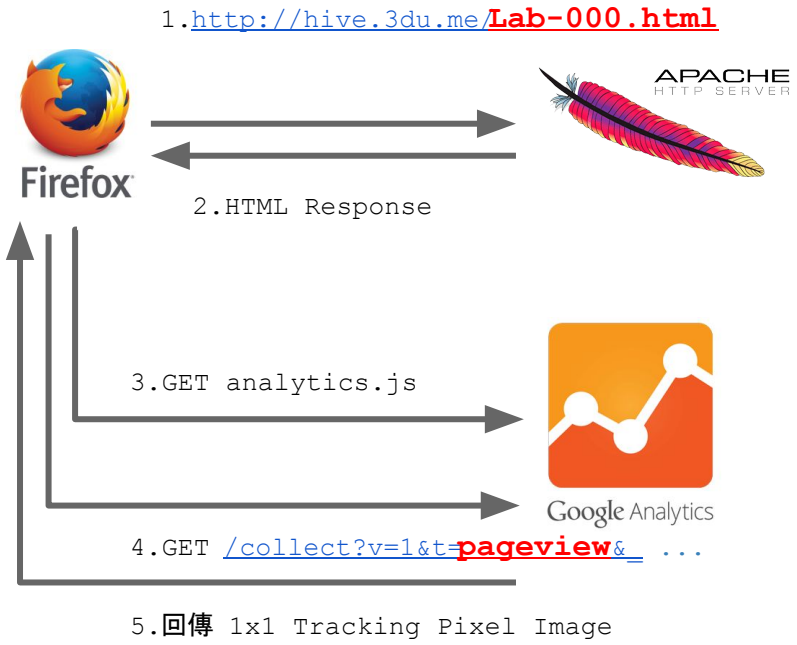


Google Analytics 的 Tracking Pixel 資料產生流程



以上過程中也寫入了一個 3du.me 網域的 cookie 用來識別這是同一個裝置的同一個瀏覽器

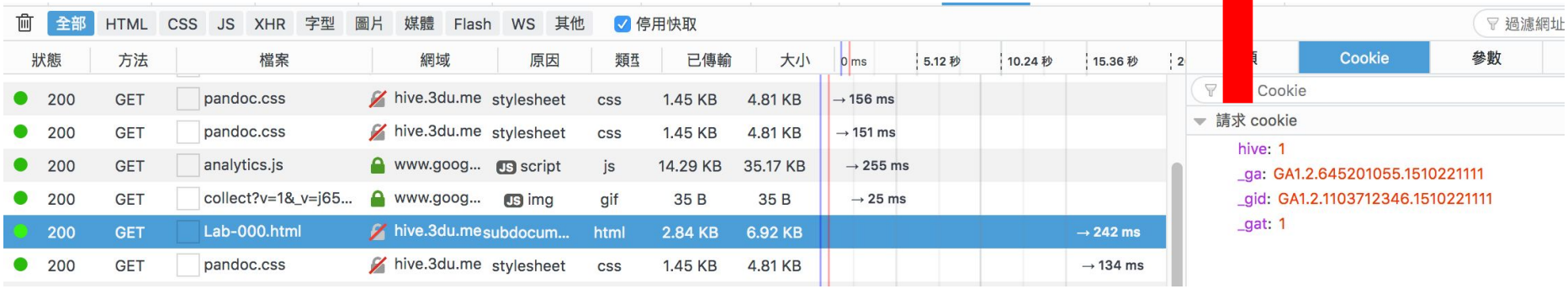
Google Analytics 的 Tracking Pixel 資料產生流程



若**網站主**有打開 access log 印 cookie 的參數，就會在日檔裡看到這些 cookie

```
來源 IP - - [日期:時間 +時區] "GET /Lab-000.html HTTP/1.1" 200 2910 "來源頁面 http://hive.3du.me/Home.html" 瀏覽器 user-agent Firefox/56.0" "hive=1; _ga=GA1.2.645201055.1510221111; _gid=GA1.2.1103712346.1510221111; _gat=1"
```

當**訪客**瀏覽其他頁面時(如:Lab-000.html)就會連同這些 cookie 一起告訴**網站主**



對 GA 來說，這些 cookie 是判斷**首次造訪**或**回訪**的依據。

Tracking Pixel 產生的紀錄分散於不同角色

網站主、分析服務、廣告網路，互相拿不到對方的資料

網站主用自身資料來跟**廣告網路**要錢

廣告網路也用自身資料跟**廣告主**要錢

廣告主想從**訪客**身上賺到錢

小結

- Tracking Pixel 會在線上廣告生態系的不同角色，
  留下不同的足跡。這些角色因為無法取得對方的資料，
  必須透過對方產生的報表來「付費 / 收費」。

> ⇨ 這些 web service 的 access log 跟收錢有關，不能漏記～
> ⇨ 快速回應(Response Quickly)、儘早落地(Write Early)

- 能產生商業價值的是由大量訪客足跡所拼湊出的「訪客特徵(Audience Profile)」。誰能同時掌握真實客戶資料與匿名的訪客足跡，就能組出更精準的「客戶洞察(Custom Insight)」

> ⇨ 這些 access log 要靠後續的**離線分析**來產生商業價值～

2.
分析 Tracking Pixel 數據的 Data Pipeline 設計攸關營運成本

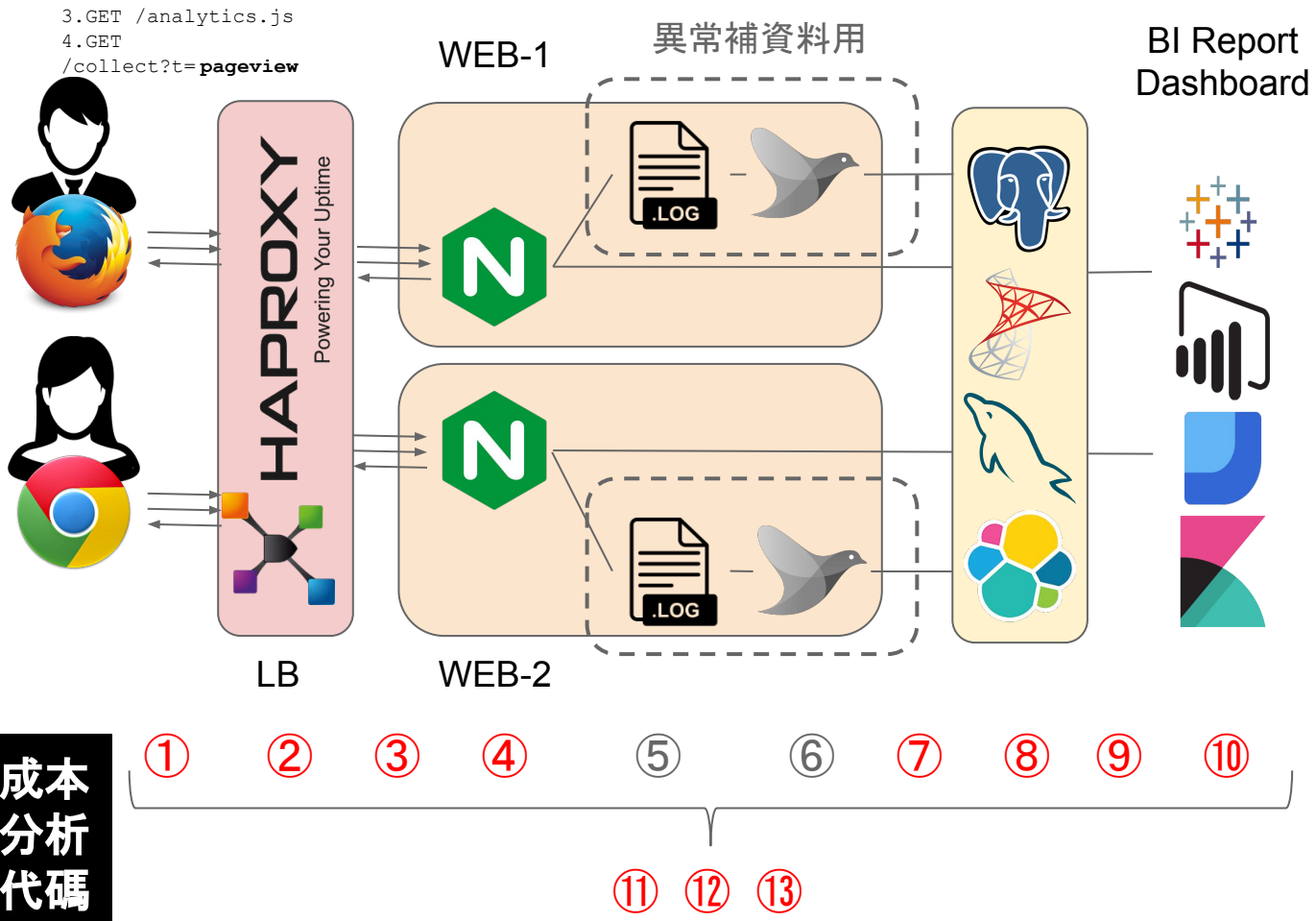
比較不同 Data Pipeline 的設計與營運成本

Lambda Architecture

Kapa Architecture

Serverless / Microservice

小規模 Tracking Pixel 的日誌分析 Data Pipeline

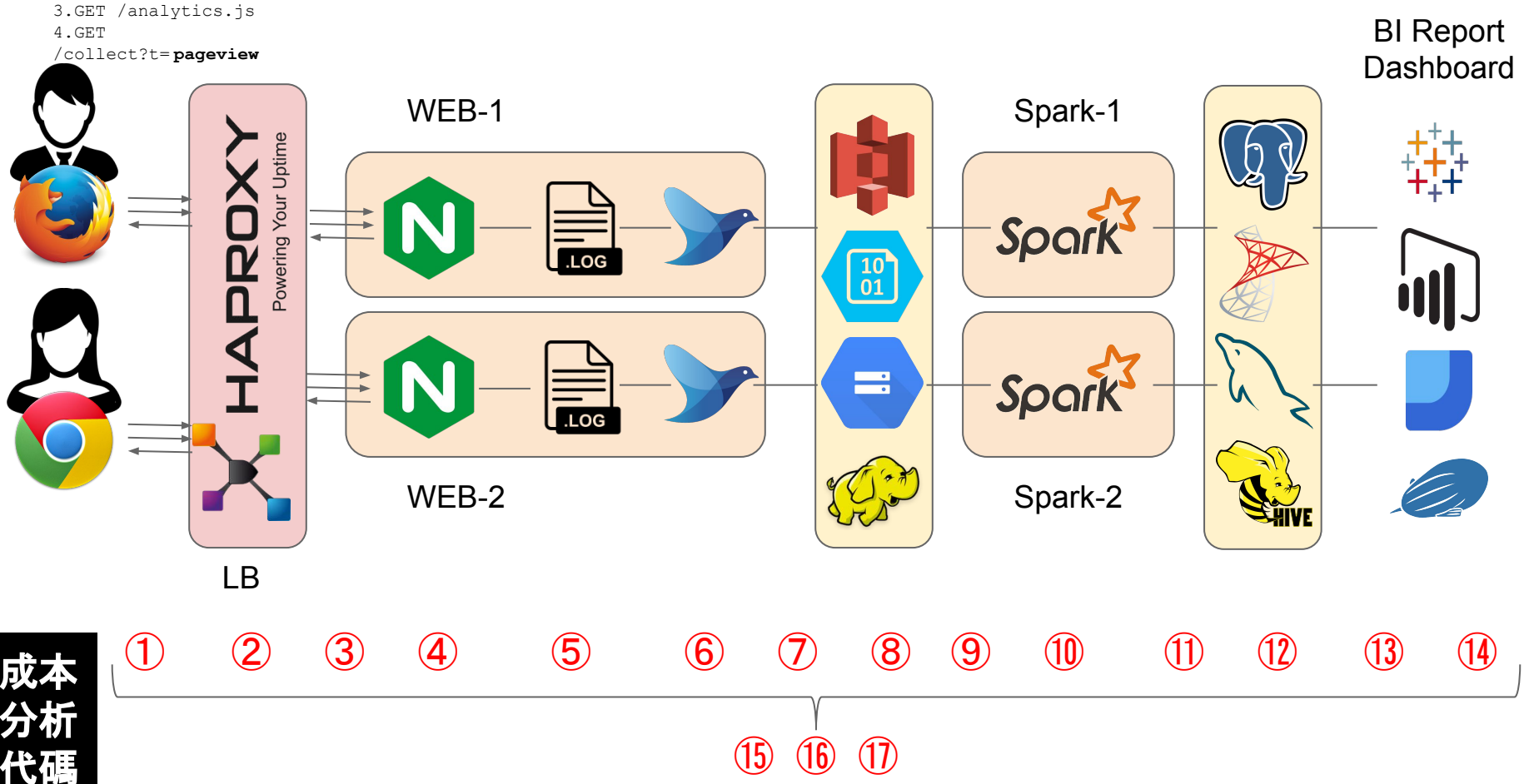


成本分析代碼

優點:資料流短而單純，易於維護與除錯

缺點:當 log 資料量過於龐大時，資料庫的同時寫入速度跟資料量會是效能瓶頸

批次 Tracking Pixel 的日誌分析 Data Pipeline

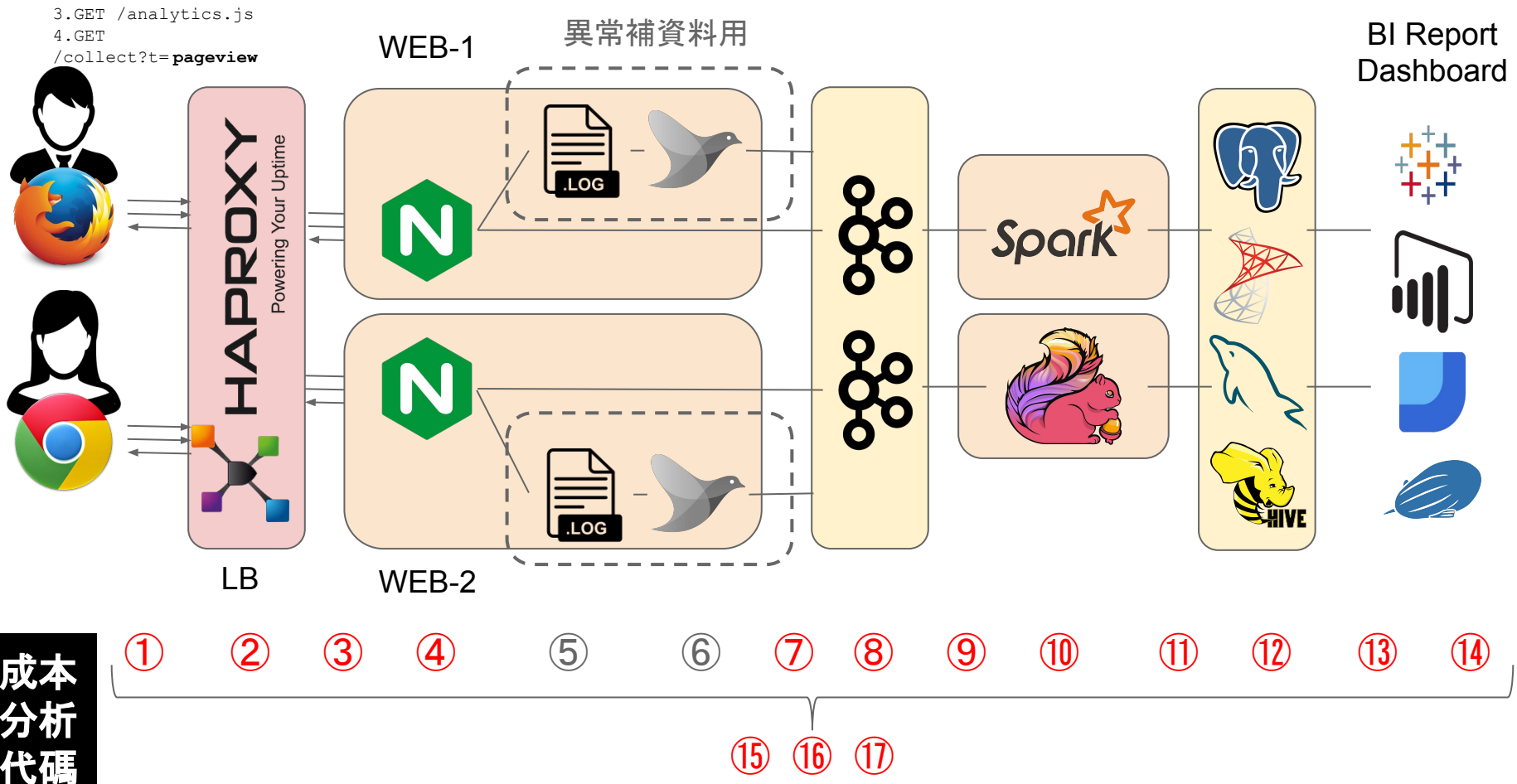


成本分析代碼

優點:適用 log 資料量龐大、需要複雜分析的情境

缺點:資料流長，不易維護與除錯，

串流 Tracking Pixel 的日誌分析 Data Pipeline

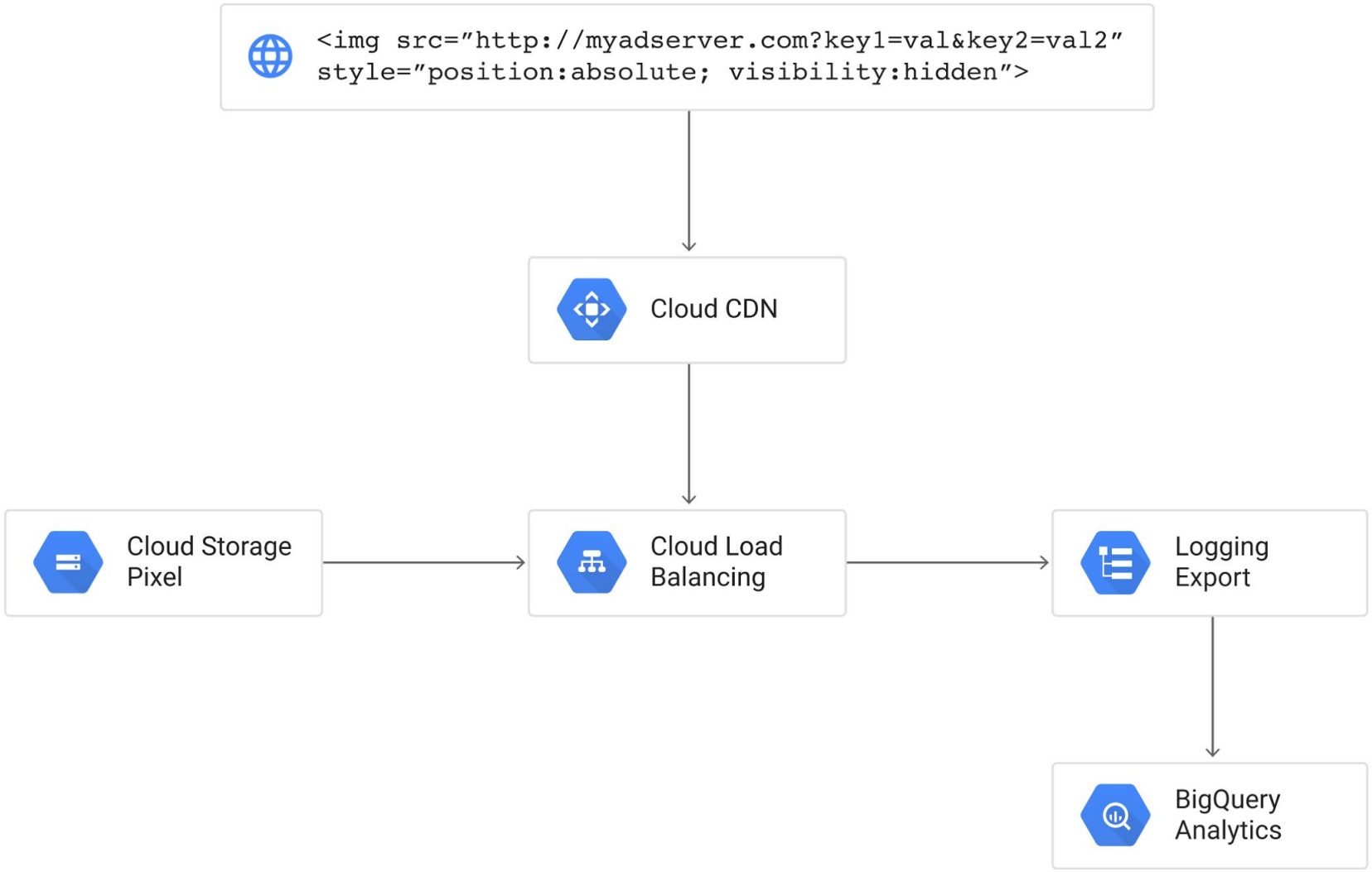


成本分析代碼

優點:適用 log 資料量龐大、需要複雜分析的情境

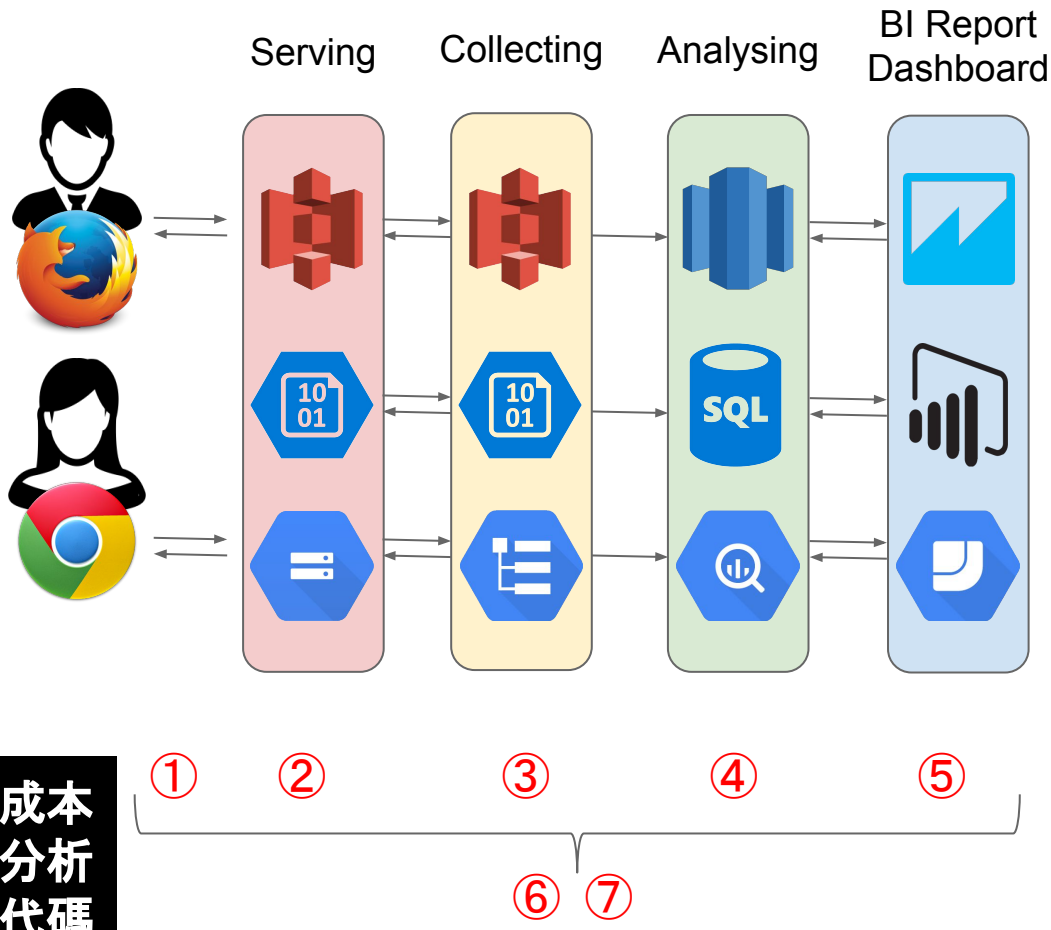
缺點:資料流長，不易維護與除錯，

Serverless Tracking Pixel Architecture



https://cloud.google.com/solutions/serverless-pixel-tracking

Serverless Tracking Pixel Data Pipeline



AWS Documentation » Amazon Simple Storage Service (S3) » Developer Guide » Hosting a Static Website on Amazon S3

Hosting a Static Website on Amazon S3

You can host a static website on Amazon Simple Storage Service (Amazon S3). On a static website, individual webpages include static content. They might also contain client-side scripts. By contrast, a dynamic website relies on server-side processing, including server-side scripts such as PHP, JSP, or ASP.NET. Amazon S3 does not support server-side scripting. Amazon Web Services (AWS) also has resources for hosting dynamic websites. To learn more about website hosting on AWS, go to Websites and Website Hosting.

http://docs.aws.amazon.com/AmazonS3/latest/dev/WebsiteHosting.html

將「靜態網頁」存放在「雲儲存」服務
是運用雲服務的 Best Practice!!

成本分析代碼

優點：技術門檻略低，不需自架網頁服務，不怕流量龐大

缺點：**僅適用 Server Based Tracking。**雲服務元件是黑盒子，不易除錯。

不同雲儲存服務的 Log 格式

- Azure Blob Storage
  - Storage Analytics Log Format
  - https://docs.microsoft.com/en-us/rest/api/storageservices/storage-analytics-log-format
- Google Cloud Storage
  - Access and storage log format
  - https://cloud.google.com/storage/docs/access-logs#format
- Amazon S3
  - Server Access Log Format
  - http://docs.aws.amazon.com/AmazonS3/latest/dev/LogFormat.html

結語

- 某些行業的 access log 跟收錢有關 ⇨ 數據是 21 世紀的黑金
- 但這些 access log 要靠後續的**離線分析**來產生商業價值～
- 練習設定 Tracking Pixel 的 Data Pipeline 是個瞭解資料工程相關技術的好開始。
- 選擇 Data Pipeline 時，請根據開發成員的技能樹、應用需求（Ex. QPS、HA、Scalability）、資料生成量與資料運算量等進行評估。
- 也請別忘了營運成本不單純只有 Data Pipeline 的運算成本，還包括網路傳輸成本、備份機制、監控機制、維運人力成本（網路管理員、系統管理員、DataOps / Data Engineer）。
- 若資料不敏感，可存放於雲服務，可考慮用 Serverless 架構
- 若有 Client based Tracking 種 cookie 需求，可用 Lambda、Functions 服務來實作。

Thank You!

Q & A