

Modern recommender system in large content website

Cyrus Chiu @ PyConTW 2020

Youtube - Landing Page

觀看次數：5.2億次

搜尋

顯示更多

首頁

發燒影片

訂閱內容

媒體庫

觀看記錄

稍後觀看

喜歡的影片

ML

老司机出品【官方...】

极速拍档 Speedste...
曼食慢语 Amanda ...
Fred吃上瘾
Ting's Bistro克里斯...
牛小咖MumaMoo
Hot Drive 热駕/經...
顯示另外 41 個項目

更多 YOUTUBE 功能

[狂人日誌] 小事，小試：愛是唯一的愛快養成日記 Vol.2

Best of Subaru Impreza WRC97-2000 tarmac action - with pure engine...

[熱駕車測] FOCUS LOMMEL賽道特化版 vs ST-LINE / 底盤差異詳解

男生夢寐以求的18坪質感工業風空間！

點煙器換成插座??

為什麼點煙器不直接換成更方便的插座

Cats and Bell

味噌拌炒雞腿定食/Stir fried Chicken with Miso Teishoku|MASAの料理ABC

10 Best Free Mac Apps

穿越格聂神山 上集 (30分钟标准版)

麵有男色

HARIO V60 推粉法

打卤面

13:12 8:00 8:17 18:58

2:53 10:24 8:06

30:13 7:59 8:01 8:37

13:12 8:00 8:17 18:58

2:53 10:24 8:06

30:13 7:59 8:01 8:37

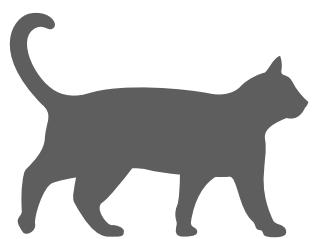
About



a ML engineer at a financial (technology) company

Cyrus Chiu

www.linkedin.com/in/iamcyruschiu



Use case in PIXNET

User2Article (personal)

The screenshot shows the PIXNET homepage with a yellow header bar. Below it, there's a section titled "你可能喜歡的 智慧家電 話題" (Topics you might like: Smart Home Appliances). It features a post by "老虎狗 Tigerdog" from September 28, 2018, at 18:17. The post title is "[分享] 一用就愛上的無線藍牙耳機" (Share: A wireless Bluetooth earphone I fell in love with at first use). The post content includes a photo of a black earphone and some text about its price and availability.

Article2Tag (non-personal)

The screenshot shows the PIXNET search results page for the tag "#iPhone XS". The search bar at the top has "#iPhone XS" entered. Below the search bar, there's a section titled "您可能會有興趣的文章" (Articles you might be interested in). It lists two articles: one from August 24, 2019, titled "「手機選購」2019 旗艦機推薦 - 依預算需求找一隻喜歡的吧!" (Smartphone Purchase Guide: 2019 Flagship Phone Recommendations - Find one that matches your budget), and another from July 29, 2019, titled "「手機選購」2019 萬元左右中階手機推薦" (Smartphone Purchase Guide: 2019 Mid-range Phone Recommendations).

Tag2Tag (non-personal)

The screenshot shows the PIXNET search results page for the tags "#asus zenfone 5 #asus zenfone 5 #MAT". The search bar at the top has "#asus zenfone 5 #asus zenfone 5 #MAT" entered. Below the search bar, there's a section titled "相關文章影音" (Related Article Videos). It lists an article by "旅遊向前走" (Traveling Forward) from September 25, 2018, at 23:17, titled "不止全機包膜、更有頂級的享受的服務 (DEVILCASE 惡魔鋁合金保護框-高" (Not just full-body film protection, but also a premium service (DEVILCASE Devil Alloy Protection Frame - High)).

Tag2Article (personal)

The screenshot shows the PIXNET search results page for the tag "#手機推薦". The search bar at the top has "#手機推薦" entered. Below the search bar, there's a section titled "相關文章影音" (Related Article Videos). It lists an article by "王維 潛藏" (Wang Wei) from October 17, 2018, at 23:17, titled "王維 潛藏1 技術面面俱到的王維 「POCKETIN多" (Wang Wei, Hidden 1, Wang Wei, a multi-faceted technical expert).

首頁

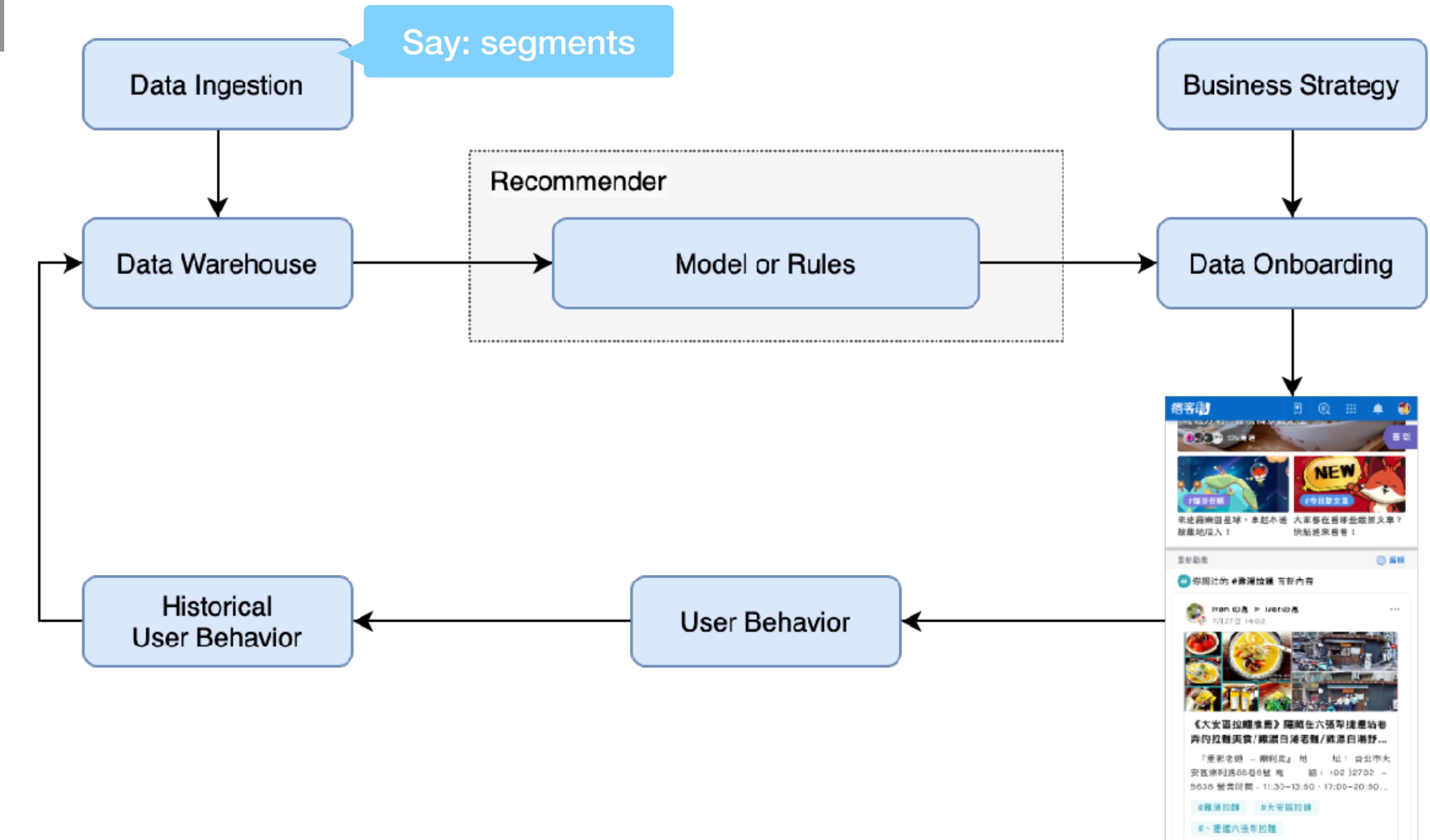
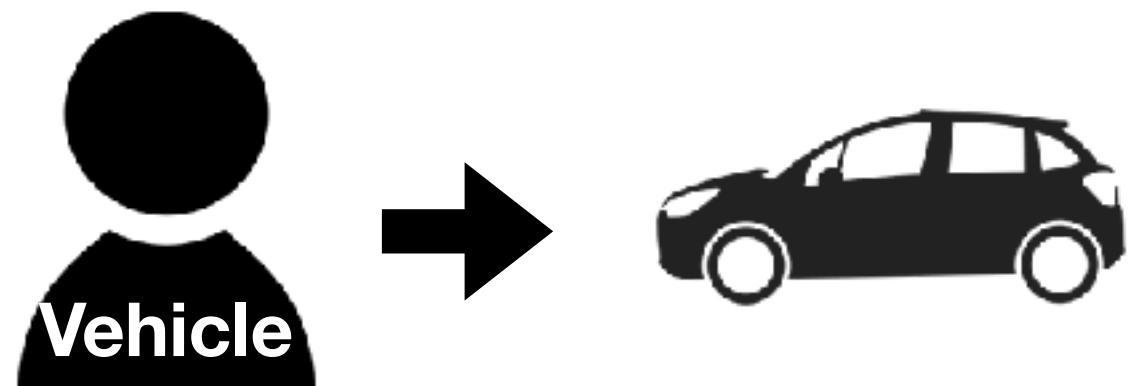
部落格頁面

搜尋結果頁

How we building a recommender system

Basic

- Data infrastructure
- Data collection
- User tracking
- Some rules or model
- Position on the webpage



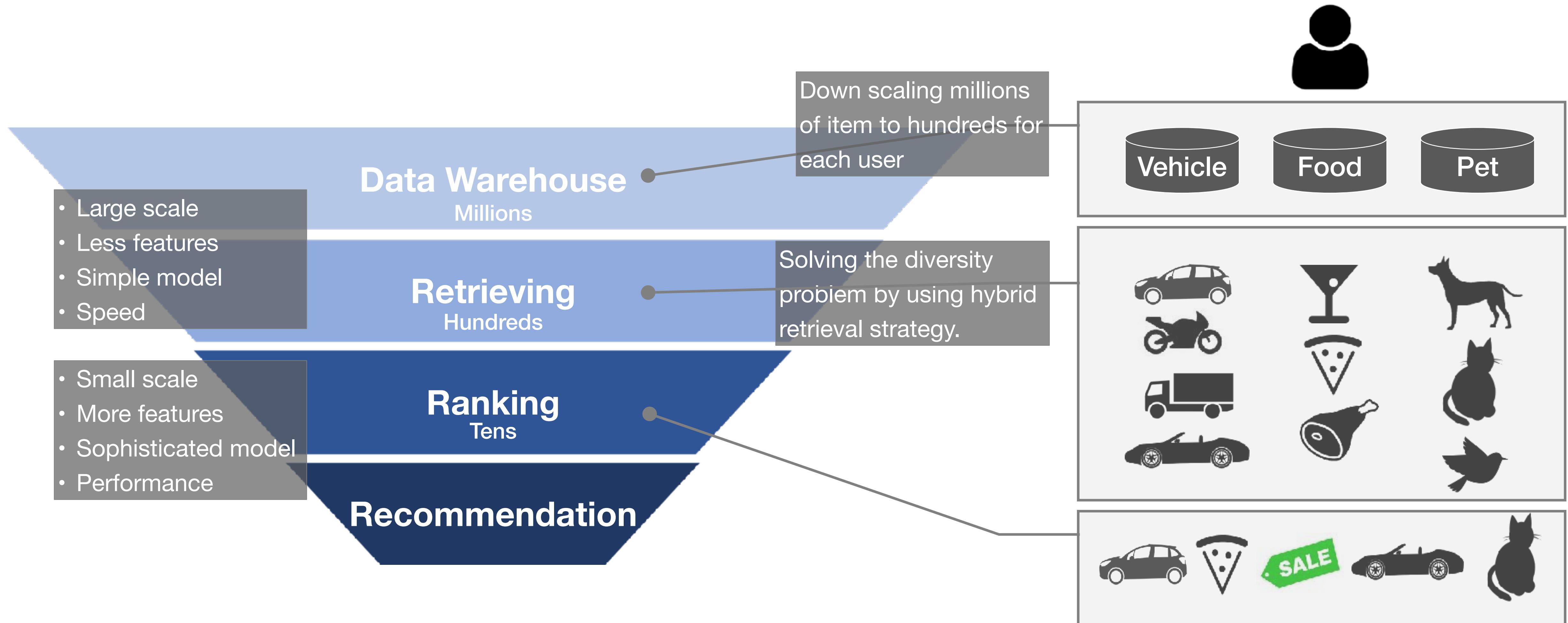
How we building a recommender system

Advanced

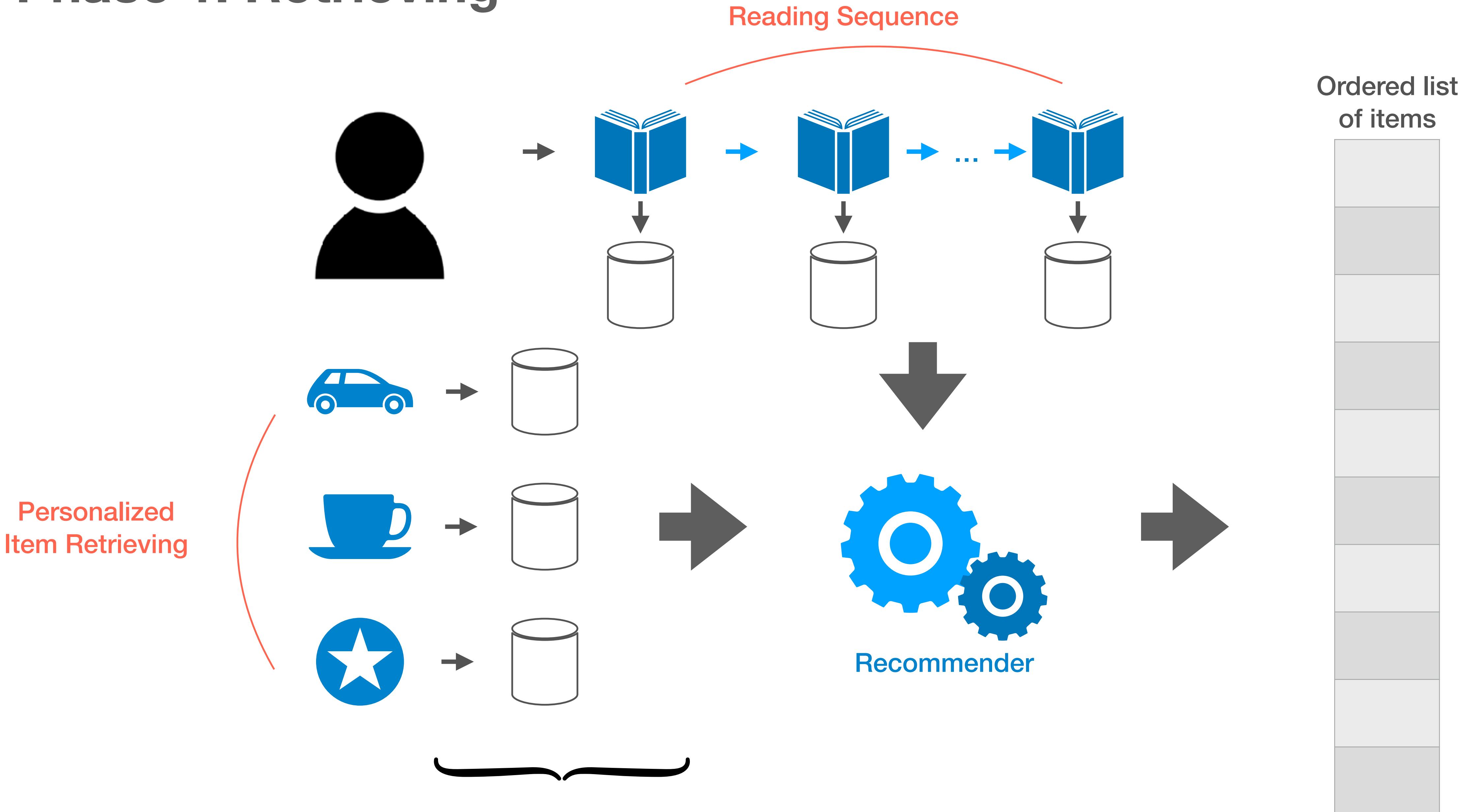
- Data lifecycle
- Scalability
- User profiling
- Item profiling
- Item retrieving
- Item ranking



What is 2-phase recommender system



Phase 1: Retrieving



Define the similarity between item A and B

According to the content



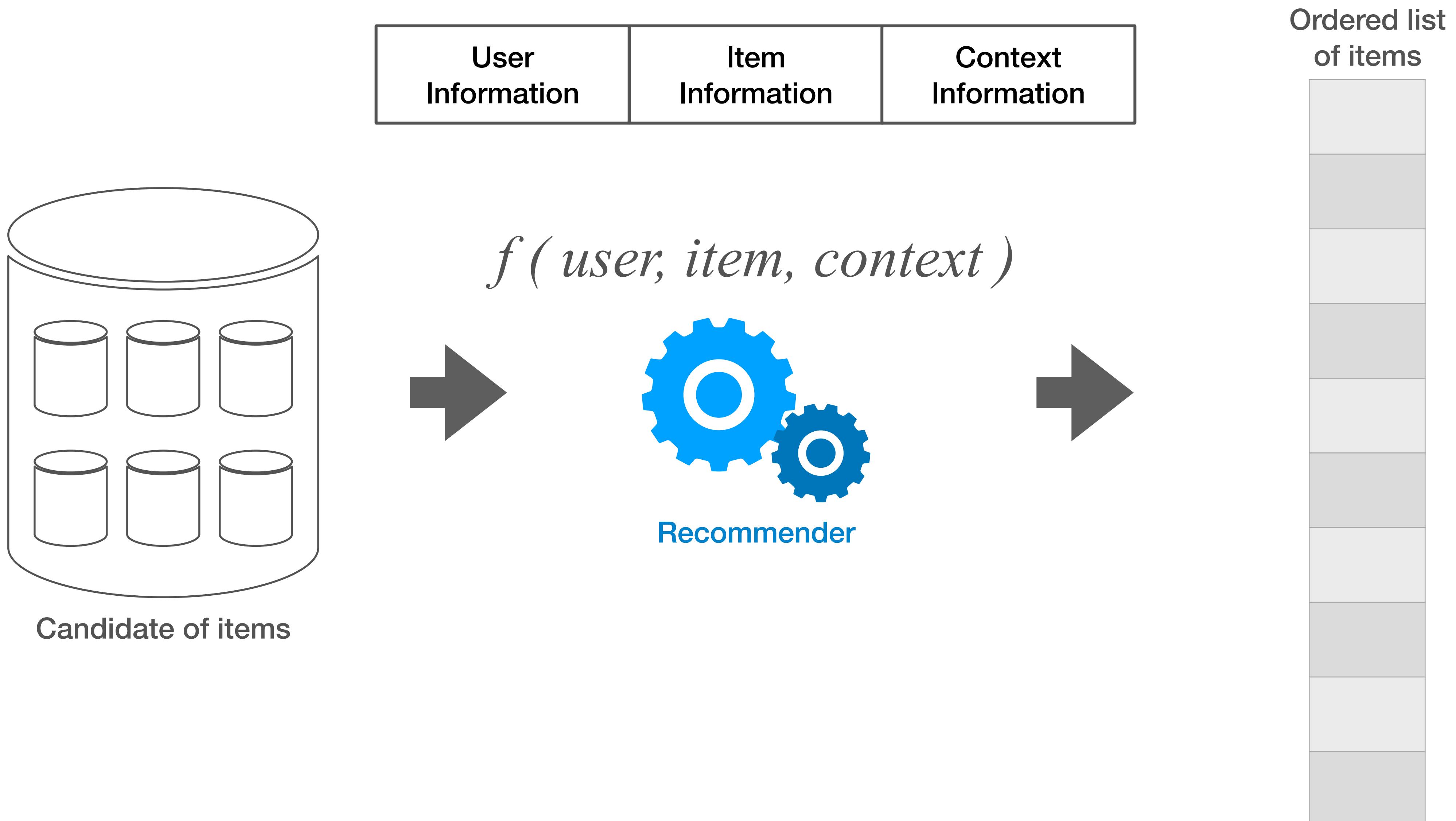
According to the user behavior



Some Retrieving Methods

Strategy	Search Engine	Word Embedding	Item Embedding	Frequently Occurring Together	Hot
Features	<ul style="list-style-type: none">• Easy to use• Multiple query type• Another cost to maintain your Elasticsearch cluster	<ul style="list-style-type: none">• Unexplainable• Vector space• High computational cost	<ul style="list-style-type: none">• Cold start problem• Data sparsity issue• Vector space• High computational cost	<ul style="list-style-type: none">• Cold start problem• Data sparsity issue• Explainable• Easy to maintain	<ul style="list-style-type: none">• Explainable• Easy to use• Suitable for any situations
Recommended					

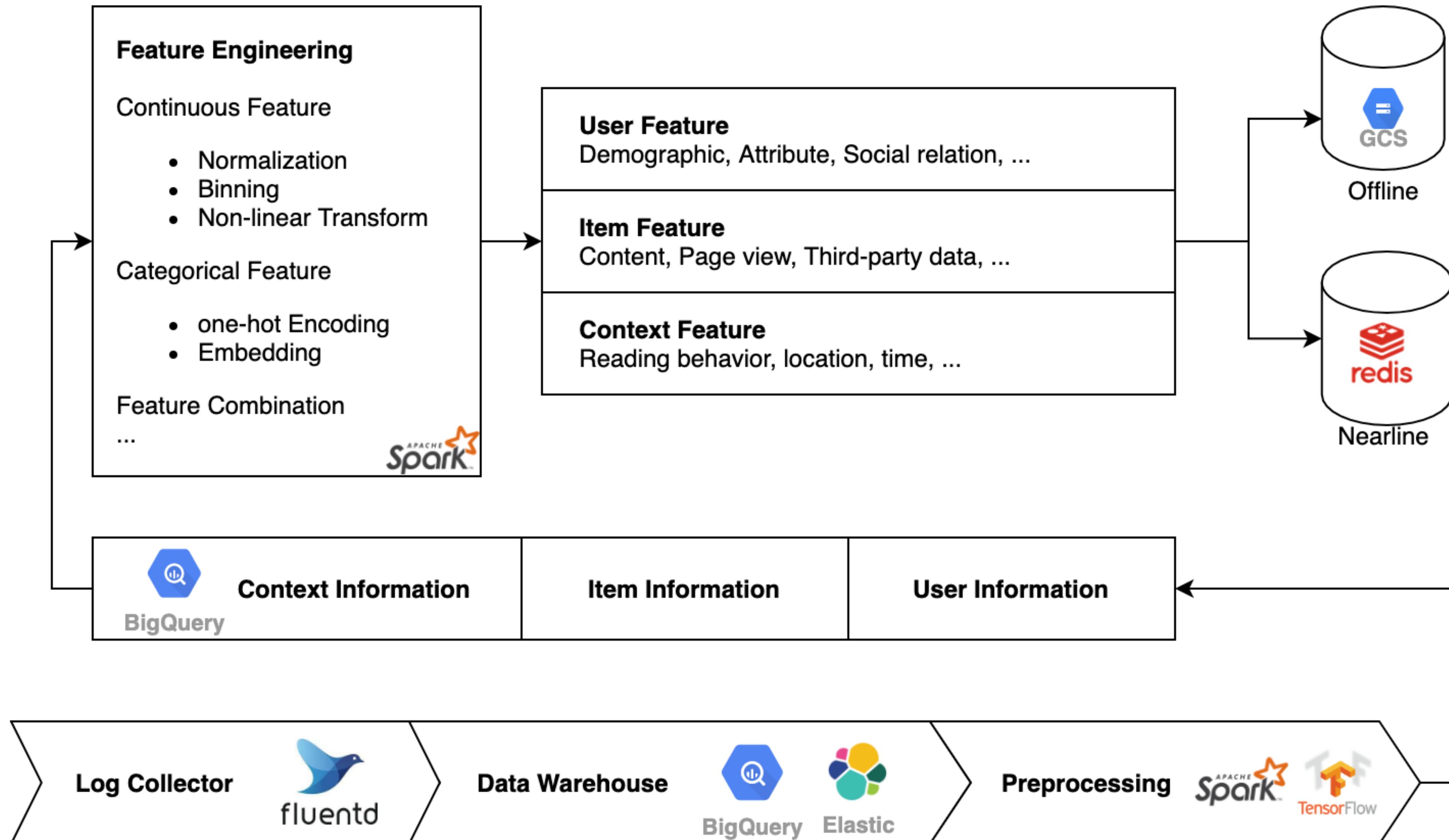
Phase 2: Ranking



Evolution of the prediction model

Model	Logistic Regression	GBDT + LR	Factorization Machine	DNN
Features	<ul style="list-style-type: none">• Fast• Highly explainable• Manually crafting features• Limited effective	<ul style="list-style-type: none">• Fast• Highly explainable• Capability to generate feature interaction	<ul style="list-style-type: none">• Slow• Explainable• Good at learning feature interactions• Exhausting resources	<ul style="list-style-type: none">• More slowly• Unexplainable• Learning sophisticated feature interactions automatically• Exhausting more resources• Wide & Deep, Deep & Cross, DeepFM, xDeepFM, ...

RecSys on DMP



Example: Find the pair-wise distance



Solution 1

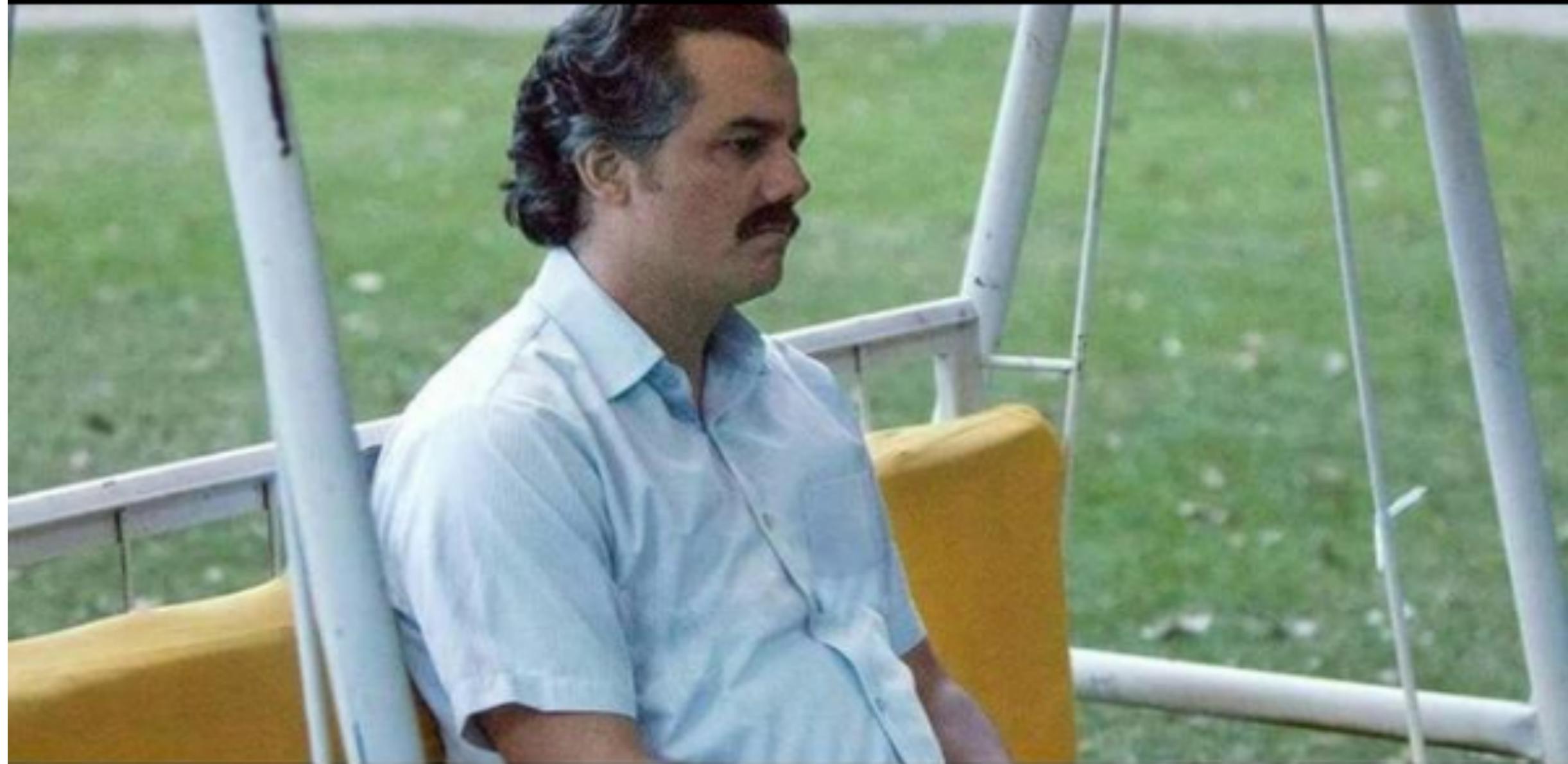
```
from sklearn.metrics.pairwise import cosine_similarity  
cosine_similarity(X)
```

```
array([[0.34723291, 0.80343977, 0.60506014, 0.23067618, 0.79965735],  
       [0.43726941, 0.04598019, 0.19007782, 0.85376095, 0.25351579],  
       [0.24234896, 0.22349041, 0.63439333, 0.40484165, 0.57103437],  
       [0.38264449, 0.93276329, 0.73085271, 0.98900326, 0.52048102],  
       [0.99288529, 0.31834113, 0.97249914, 0.08776034, 0.90290914],  
       [0.90180205, 0.63893861, 0.22106869, 0.09200385, 0.86880432],  
       [0.04775797, 0.44545171, 0.9665554 , 0.27558683, 0.13279217],  
       [0.40203503, 0.53100031, 0.67645917, 0.17551579, 0.37494095],  
       [0.77076934, 0.12681486, 0.1971458 , 0.79065758, 0.4628704 ],  
       [0.20363516, 0.72497933, 0.92195548, 0.96016578, 0.340993 ]])
```

Solution 2

```
from sklearn.metrics.pairwise import pairwise_distances  
pairwise_distances(X, metric="cosine")
```

**When you submit the job
without parallel computing..**



OUT OF MEMORY ...



import pyspark (1)

Solution: RDD

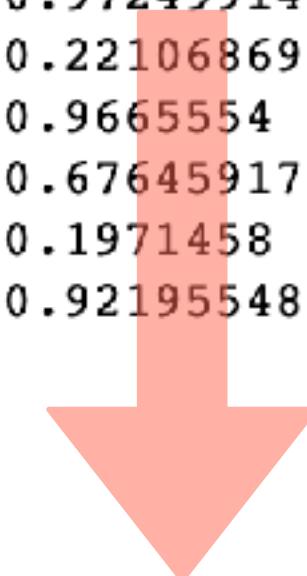
```
X_bc = sc.broadcast(X)
n_part = 100
batch_size = 64
data_size = X.shape[0]

batch_vectors_rdd = sc.parallelize(
    [X[i: i+batch_size]
        for i in range(0, data_size, batch_size)]
    ],
n_part)

batch_vectors_rdd.mapPartitions(
    lambda x: cosine_similarity(x, X_bc.value))
```

array([[0.34723291, 0.80343977, 0.60506014, 0.23067618, 0.79965735],
 [0.43726941, 0.04598019, 0.19007782, 0.85376095, 0.25351579],
 [0.24234896, 0.22349041, 0.63439333, 0.40484165, 0.57103437],
 [0.38264449, 0.93276329, 0.73085271, 0.98900326, 0.52048102],
 [0.99288529, 0.31834113, 0.97249914, 0.08776034, 0.90290914],
 [0.90180205, 0.63893861, 0.22106869, 0.09200385, 0.86880432],
 [0.04775797, 0.44545171, 0.9665554 , 0.27558683, 0.13279217],
 [0.40203503, 0.53100031, 0.67645917, 0.17551579, 0.37494095],
 [0.77076934, 0.12681486, 0.1971458 , 0.79065758, 0.4628704],
 [0.20363516, 0.72497933, 0.92195548, 0.96016578, 0.340993]])

} batch



```
array([[0.34723291, 0.80343977, 0.60506014, 0.23067618, 0.79965735],
       [0.43726941, 0.04598019, 0.19007782, 0.85376095, 0.25351579],
       [0.24234896, 0.22349041, 0.63439333, 0.40484165, 0.57103437],
       [0.38264449, 0.93276329, 0.73085271, 0.98900326, 0.52048102],
       [0.99288529, 0.31834113, 0.97249914, 0.08776034, 0.90290914],
       [0.90180205, 0.63893861, 0.22106869, 0.09200385, 0.86880432],
       [0.04775797, 0.44545171, 0.9665554 , 0.27558683, 0.13279217],
       [0.40203503, 0.53100031, 0.67645917, 0.17551579, 0.37494095],
       [0.77076934, 0.12681486, 0.1971458 , 0.79065758, 0.4628704 ],
       [0.20363516, 0.72497933, 0.92195548, 0.96016578, 0.340993 ]])
```

X_bc

import pyspark (2)

Solution: DataFrame

```
@F.udf(returnType=DoubleType())
def get_distance(i, j):
    sim = cosine_similarity(i, j)
    return sim

vector_sdf.alias("i").join(
    vector_sdf.alias("j"),
    col("i.id") < col("j.id")).select(
        F.col("i.id").alias("i"),
        F.col("j.id").alias("j"),
        get_distance("i.vector", "j.vector").alias(
            "cosine_similarity")
    )
```

	vector	id
0	[0.01227569580078125, -0.07232666015625, 0.036...	59358
1	[-0.00469970703125, -0.0190277099609375, 0.011...	59427
2	[0.04229736328125, -0.0195465087890625, -0.058...	59389
3	[0.0528564453125, -0.0261688232421875, 0.00719...	59373
4	[0.043609619140625, -0.041107177734375, -0.022...	59330
...
83118	[0.034881591796875, -0.0396728515625, 0.002809...	38903
83119	[0.08929443359375, -0.0239410400390625, 0.0203...	38873
83120	[1.0478515625, 0.6865234375, -1.03125, 0.82080...	38651
83121	[-0.01203155517578125, 0.0203857421875, -0.020...	38807
83122	[0.025238037109375, -0.056640625, 0.0049400329...	38987

J = I

	vector	id
0	[0.01227569580078125, -0.07232666015625, 0.036...	59358
1	[-0.00469970703125, -0.0190277099609375, 0.011...	59427
2	[0.04229736328125, -0.0195465087890625, -0.058...	59389
3	[0.0528564453125, -0.0261688232421875, 0.00719...	59373
4	[0.043609619140625, -0.041107177734375, -0.022...	59330
...
83118	[0.034881591796875, -0.0396728515625, 0.002809...	38903
83119	[0.08929443359375, -0.0239410400390625, 0.0203...	38873
83120	[1.0478515625, 0.6865234375, -1.03125, 0.82080...	38651
83121	[-0.01203155517578125, 0.0203857421875, -0.020...	38807
83122	[0.025238037109375, -0.056640625, 0.0049400329...	38987



CROSS JOIN



CosineSimilarity (
Column I.vector,
Column J.vector
)

Comparison

Calculating distances between coordinates from a 32*1k/10k/100k matrix
(32 core, 32 GB memory)

	1K	10K	100K
sklearn	0.1	1.5	OOM
pyspark	2	5.1	70
Faiss	0.05	0.6	21

Workflows of the data project

Data Management

- ETL pipeline design
- Data infra design
- Metalayer design
- Data toolkit

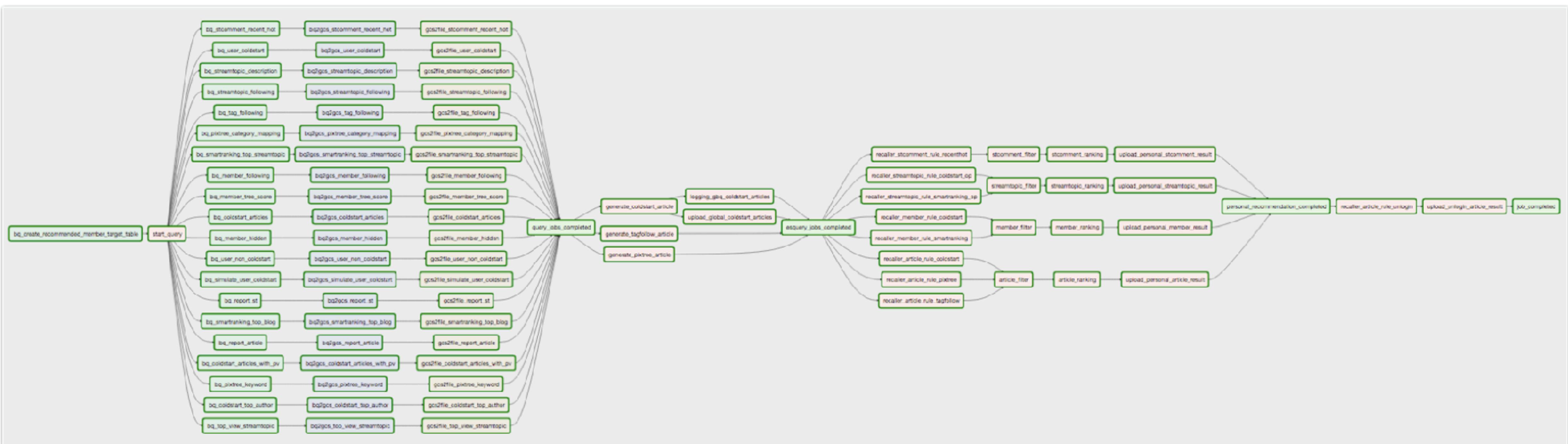
Programming

- SQL
- Framework design
- Library development
- Parallel computing
- Machine learning

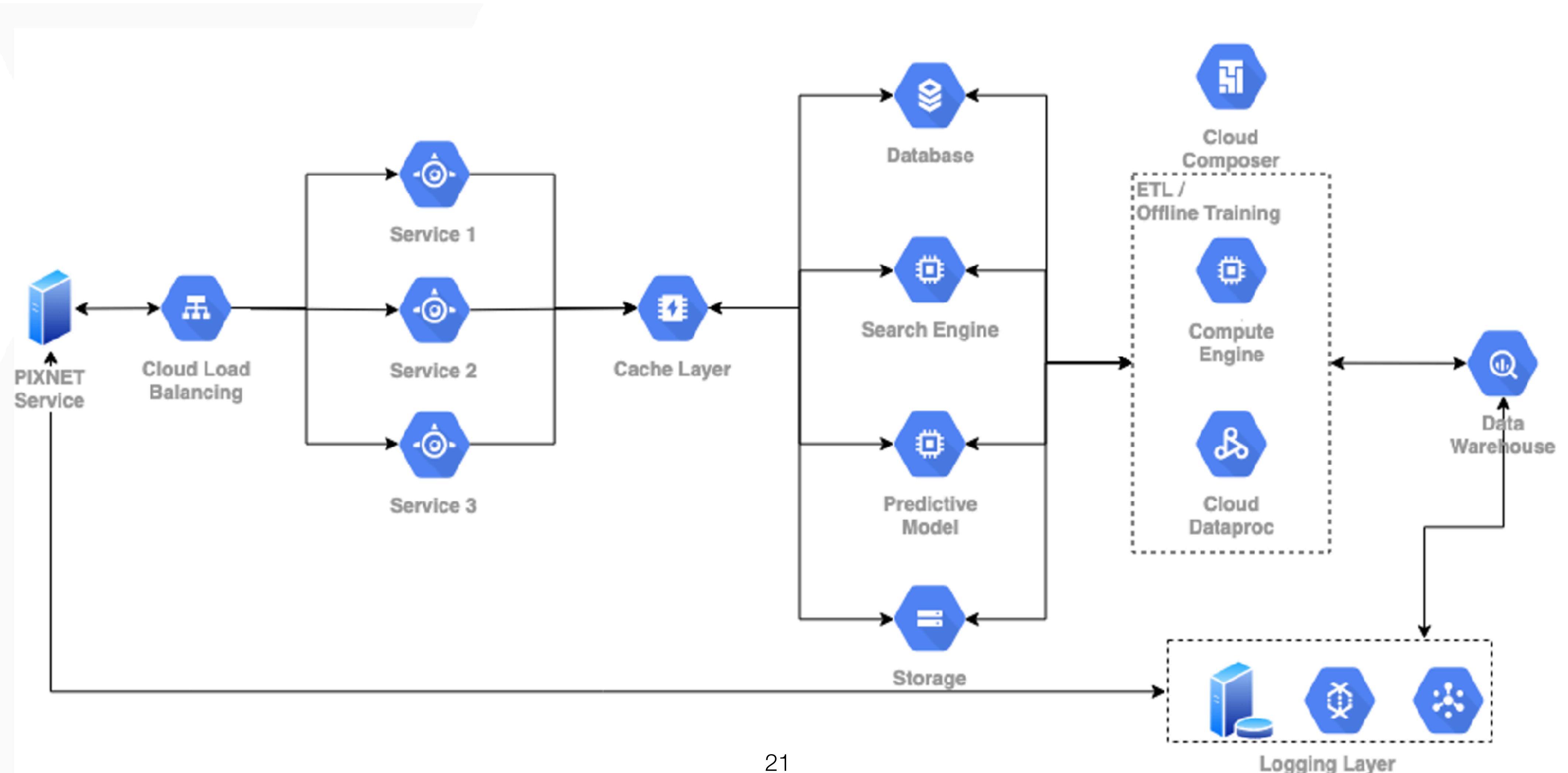
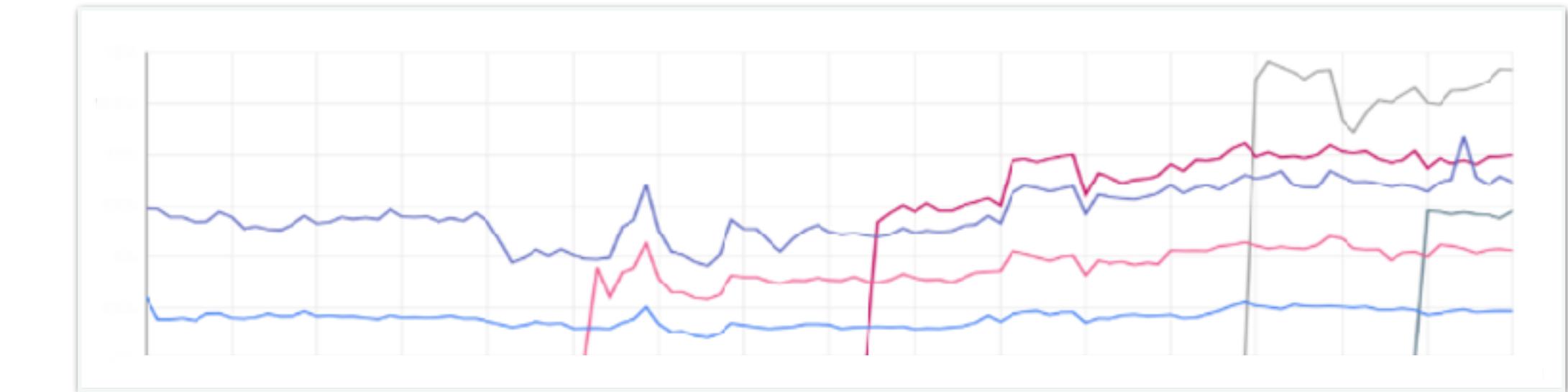
Project Management

- Git flow
- Review

Using Apache Airflow to schedule and monitor workflows



Online A/B Testing



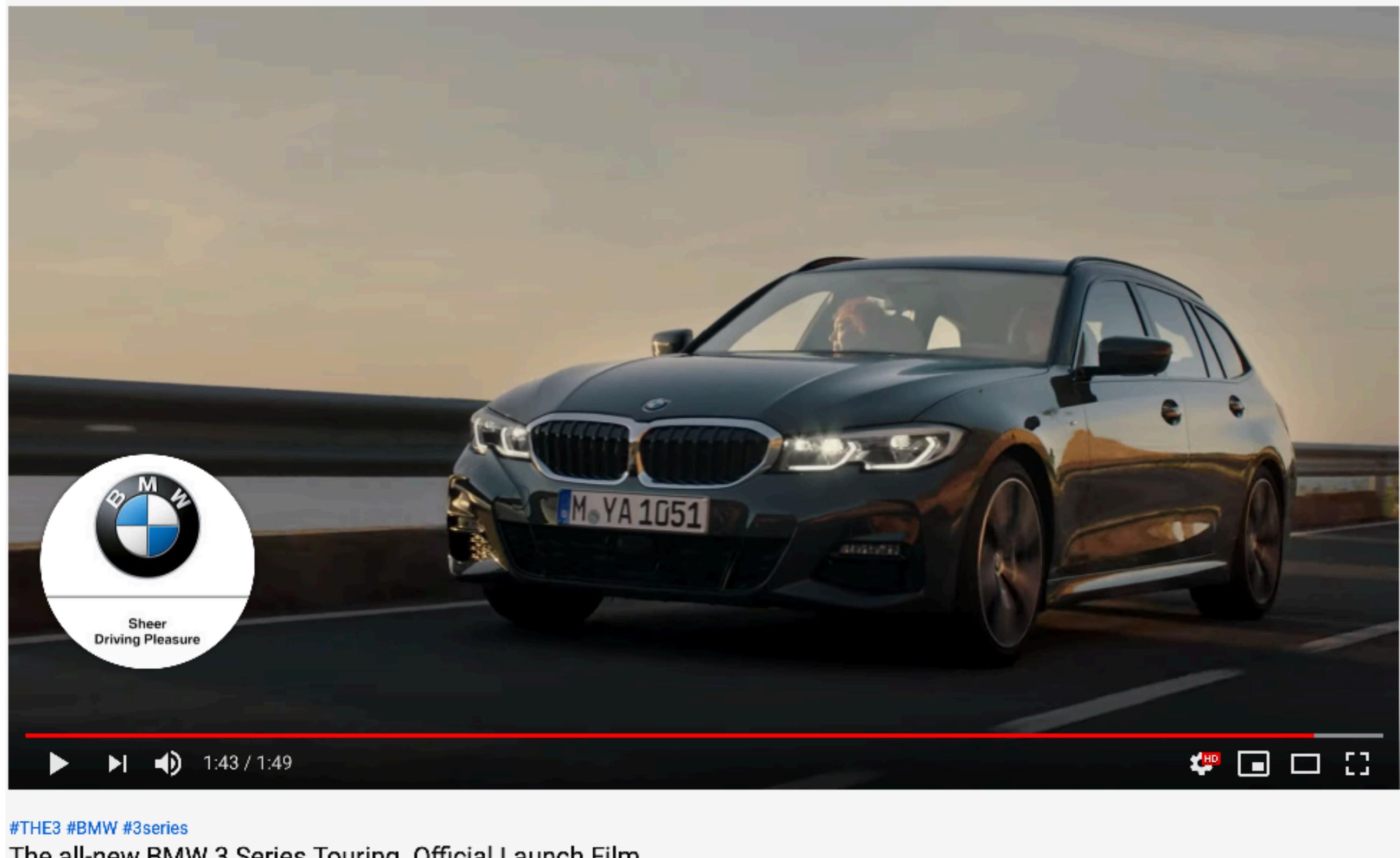
Youtube - Landing Page

The screenshot shows the YouTube homepage in Chinese. The left sidebar includes links for Home, Popular Videos, Subscriptions, Library, Watch History, Later, and Favorites. It also lists channels like 老司机出品【官方...】 and 命运共同体。 A search bar at the top right contains the text '熱駕車測'.

The main content area displays a grid of video thumbnails:

- [狂人日誌] 小事，小試：愛是唯一的愛快養成日記 Vol.2 (13:12) - MADVNZ 狂人日誌
- Best of Subaru Impreza WRC97-2000 tarmac action - with pure engine... (8:00) - amjayes2
- [熱駕車測] FOCUS LOMMEL賽道特化版 vs ST-LINE / 底盤差異詳解 (8:17) - Hot Drive 热驾/经典90
- 男生夢寐以求的18坪質感工業風空間！獨享酒吧、世界地圖全部放進家中！... (18:58) - Lo-Fi House
- 為什麼點煙器不直接換成更方便的插座？？ (4:41) - 备胎说车
- Cats and Bell (2:53) - Cats and Bell
- 味噌拌炒雞腿定食 (10:24) - MASAの料理ABC
- 10 Best Free Mac Apps (8:06) - Mac
- 穿越格聂神山 上集 (30分钟标准版) (30:13) - Ting's Bistro克里斯...
- 麵有男色 (7:59) - 牛小咖MumaMoo
- HARIO V60 推粉法 (8:01) - Hot Drive 热驾/经典90
- 打卤面 (8:37) - MASAの料理ABC

Youtube - Video Page



即將播放

自動播放

New BMW 3 Series Touring 2020 - see why it's the best ca...

carwow 觀看次數：57萬次

New BMW 3 Series Tou 6:15

The all-new BMW 3 Series. Official Launch Film. (G20,...

BMW 觀看次數：34萬次

BMW 2:05

【全民瘋學堂】換檔撥片的使用時機與教學 - 廖怡塵 【全民瘋...

全民瘋車Bar Recommended for you

怎麼用 13:11

真男人！开本田！

极速拍档 Speedsters Recommended for you 新影片

19:11

Découverte de la BMW Série 3 Touring (2019)

autojournalfr 觀看次數：2823次

1:55

The all-new BMW X6. Official Launch Film.

BMW 觀看次數：82萬次

BMW 2:31

Thank You