

Transformando

Experiencias de compra

Motor de Recomendaciones y
Propensiones

Auberth Eduardo Hurtado
Data Scientist Sr.
auberth.hurtado@globant.com

<https://www.linkedin.com/in/auberth-eduardo-hurtado-d%C3%ADaz-83746375/>

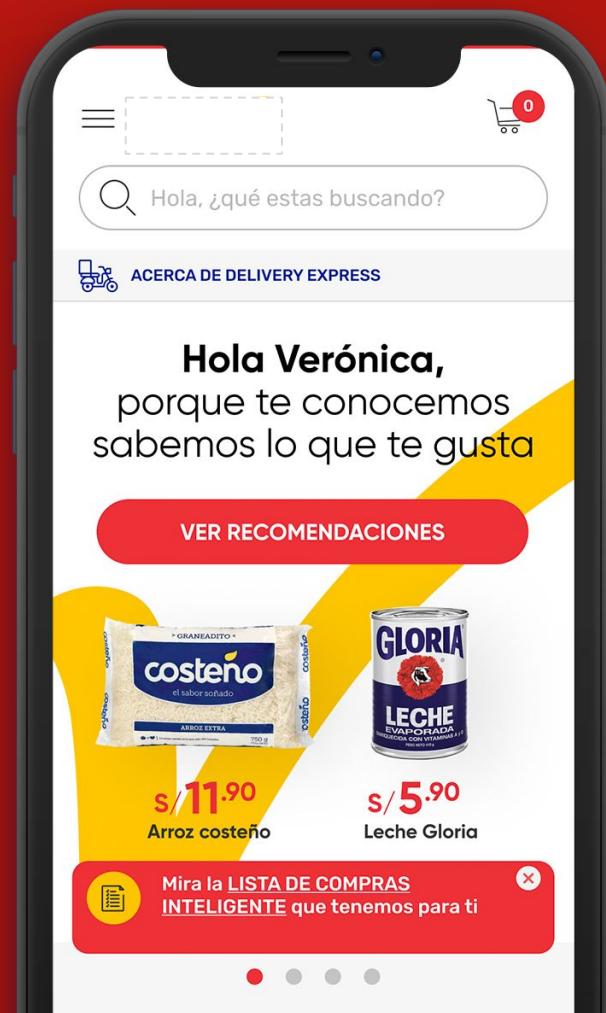


Globant ➤

El objetivo

Presentar un **caso de aplicación** de **sistemas de recomendación** utilizando una aproximación mediante el **descubrimiento de patrones**.

Esta charla brindará la oportunidad de **aprender** sobre métodos de descubrimiento de patrones escalables en **datos transaccionales masivos**.



¿Por qué la personalización?



de los ingresos de Amazon es generado por su **Motor de Recomendación**.

- McKinsey & Co.

Las compañías retail **saben que los productos recomendados pueden generar mayores oportunidades** de cross-selling y up-selling, pero pocos han implementado recomendaciones de productos verdaderamente personalizadas en sus sitios, correos electrónicos u otros canales.

Ser parte del selecto grupo que define
el presente y forja el futuro



Pasar de segmentos...



A hiper-personalizar Las recomendaciones



La personalización Paga



19%

Promedio de **aumento en las ventas** de marcas que personalizan la experiencia en la web.



Hasta un 20%

Aumento del ROI para las marcas que realizan personalización basada en datos

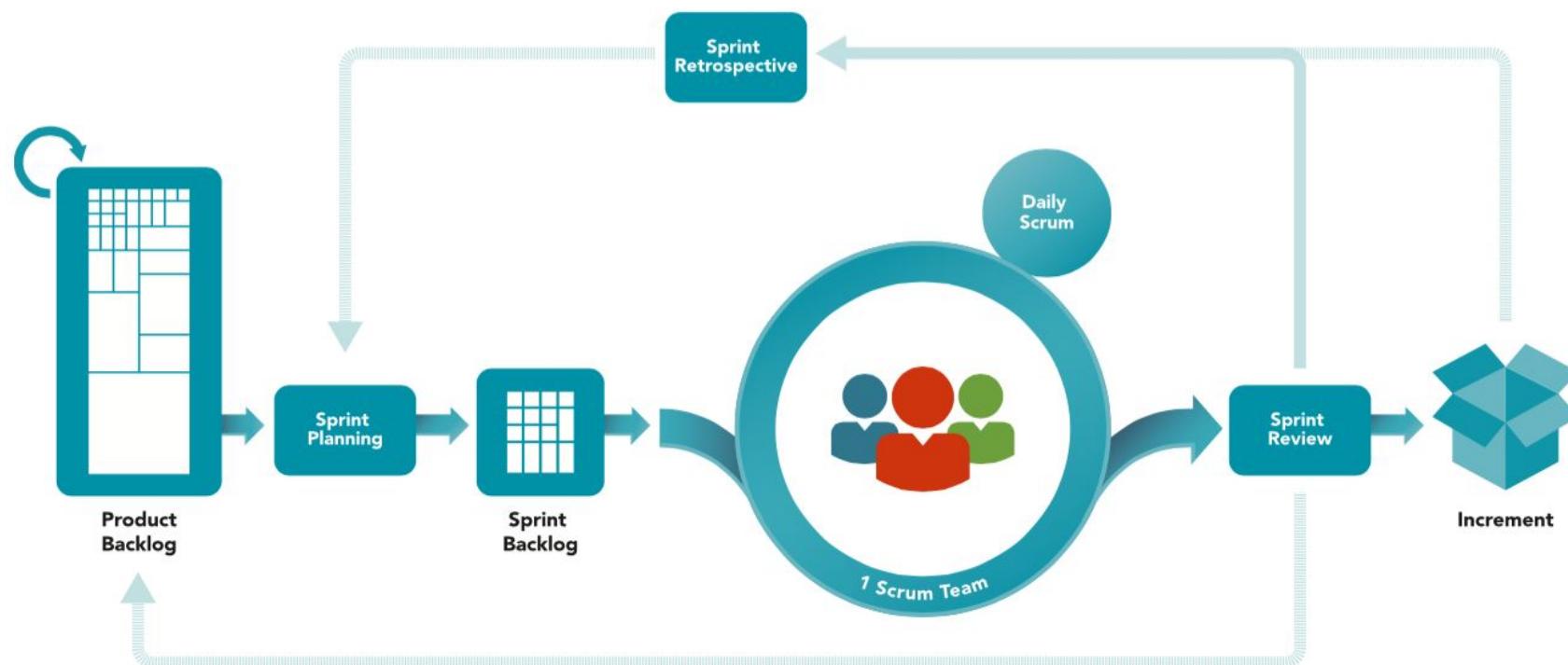


78%

Más probable que los consumidores sean **clientes habituales** si una marca proporciona ofertas personalizadas y específicas.

¿Cómo lo hicimos?

SCRUM FRAMEWORK



Discovery
(2 semanas iniciales)

Visibilidad diaria

Planificaciones semanales

Revisiones de incremento cada sprint

Constante comunicación

Levantamiento rápido de bloqueos o impedimentos

12 sprints: 2 semanas de duración



Product Owner
Intermediario entre Globant y el cliente.



Project manager



Technical director



Business intelligence



Data architect

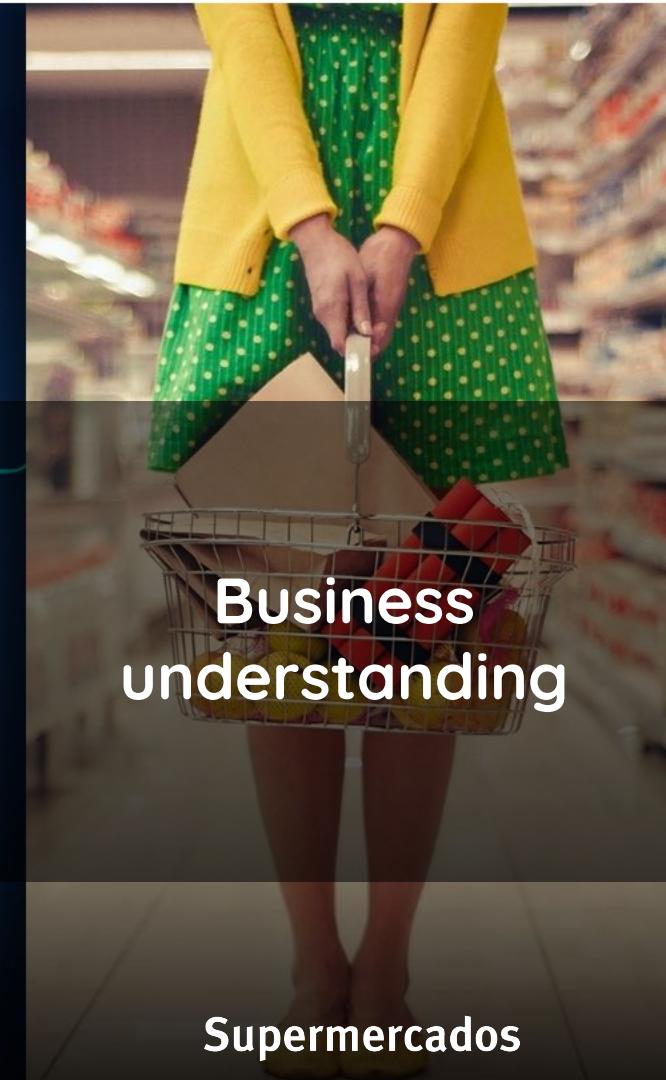
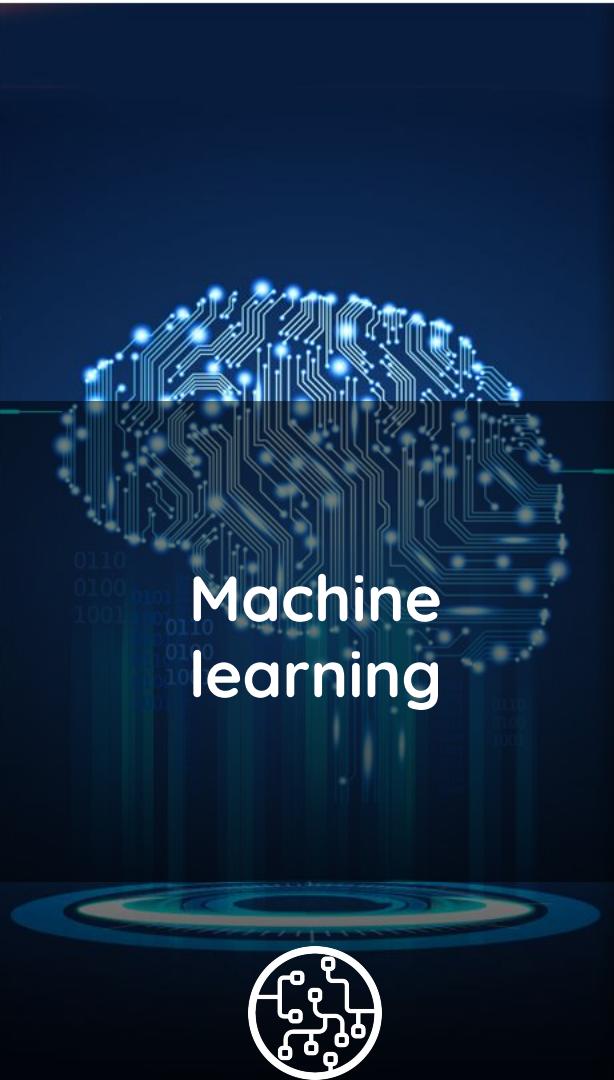


QC analyst



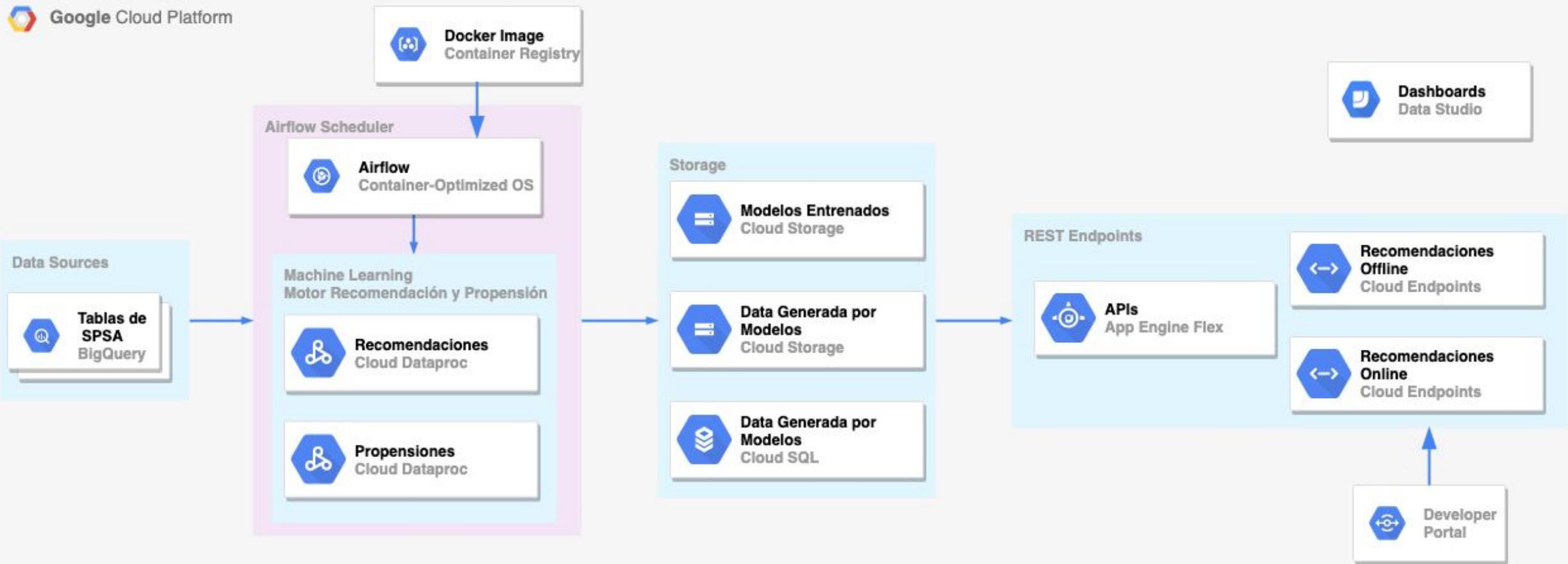
Data scientist

¿Qué utilizamos en la construcción?



Arquitectura Desarrollada

Arquitectura: Motor de recomendación y Propensión



Objetivo de negocio

- Identificar la subfamilia que **detonan la compra de más subfamilias** asociadas a la subfamilia ancla con el objetivo de incrementar la probabilidad de compra de subfamilias asociadas.



Elaboramos un modelo de recomendación de Subfamilias Ancla:

“

Estamos interesados en aquellas subfamilias que detonan la compra de otras subfamilias

Con este modelo el supermercado podrá saber cuáles son las subfamilias que detonan la compra de otras subfamilias originadas a partir del hábito de consumo de sus clientes.



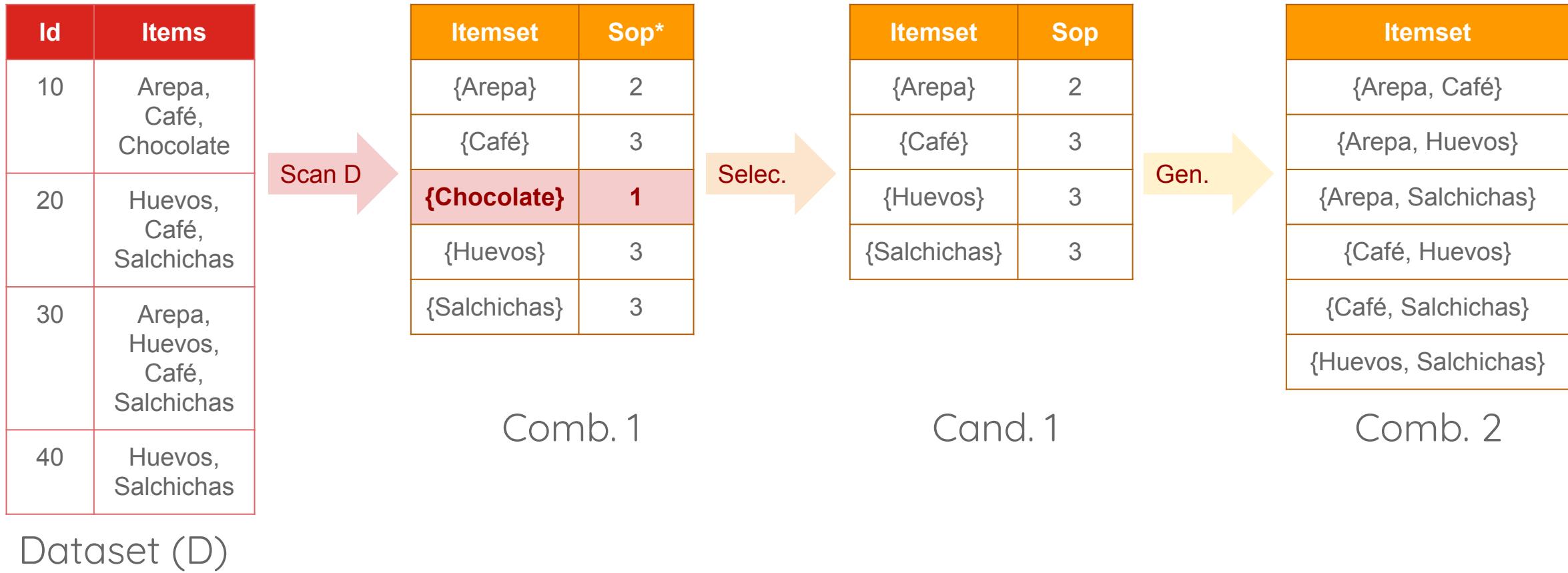
Qué es FP-Growth

El patrón frecuente de crecimiento es una mejora del algoritmo Apriori diseñada para superar alguna de sus debilidades.



Algoritmo Apriori

Ejemplo: Min. Soporte = 2



*Soporte: Cantidad de veces que el ítem fue adquirido

Algoritmo Apriori

Ejemplo: Min. Soporte = 2

Itemset
{Arepas, Café}
{Arepas, Huevos}
{Arepas, Salchichas}
{Café, Huevos}
{Café, Salchichas}
{Huevos, Salchichas}

Scan D

Itemset	Sop
{Arepas, Café}	2
{Arepas, Huevos}	1
{Arepas, Salchichas}	1
{Café, Huevos}	2
{Café, Salchichas}	2
{Huevos, Salchichas}	3

Selec.

Itemset	Sop
{Arepas, Café}	2
{Café, Huevos}	2
{Café, Salchichas}	2
{Huevos, Salchichas}	3

Gen.

Itemset	Sop
{Arepas, Café, Huevos}	1
{Arepas, Café, Salchichas}	1
{Café, Huevos, Salchichas}	2

Scan D

Itemset
{Café, Huevos, Salchichas}

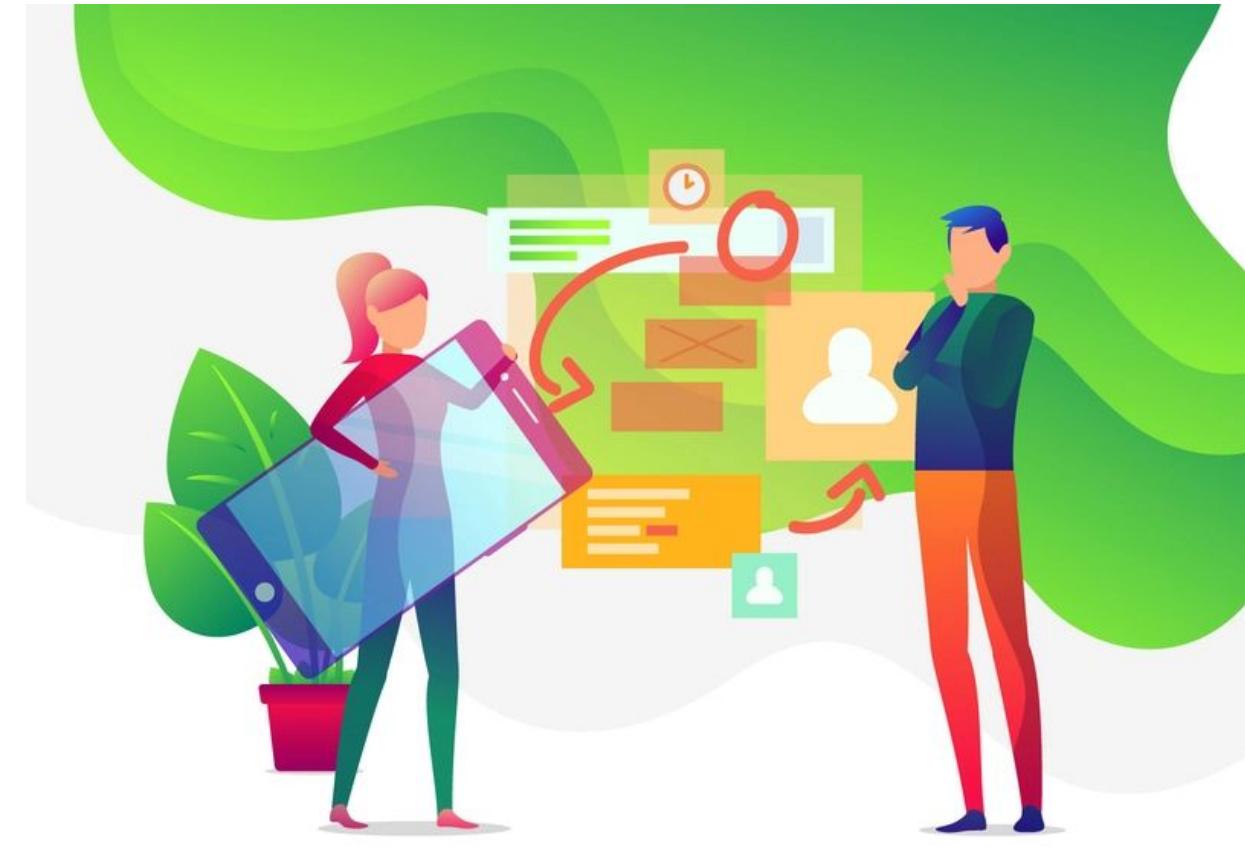
Comb. 2

Comb. 2

Cand. 2

Desafíos

- **Múltiples lecturas** a la base de datos de transacciones
- Generación de una **gran cantidad de candidatos**
 - Catálogo de 250.000 productos
 - En la 4a iteración asumiendo que se elimina en cada iteración la mitad: $2.54 \cdot 10^{109}$.
 - Se estima que la cantidad de granos de arena en el mundo es de 10^{95} .
- Tediosa carga de **trabajo auxiliar**.
 - Base de clientes de 5 millones.



Es prácticamente
imposible usar
Apriori en Retail.

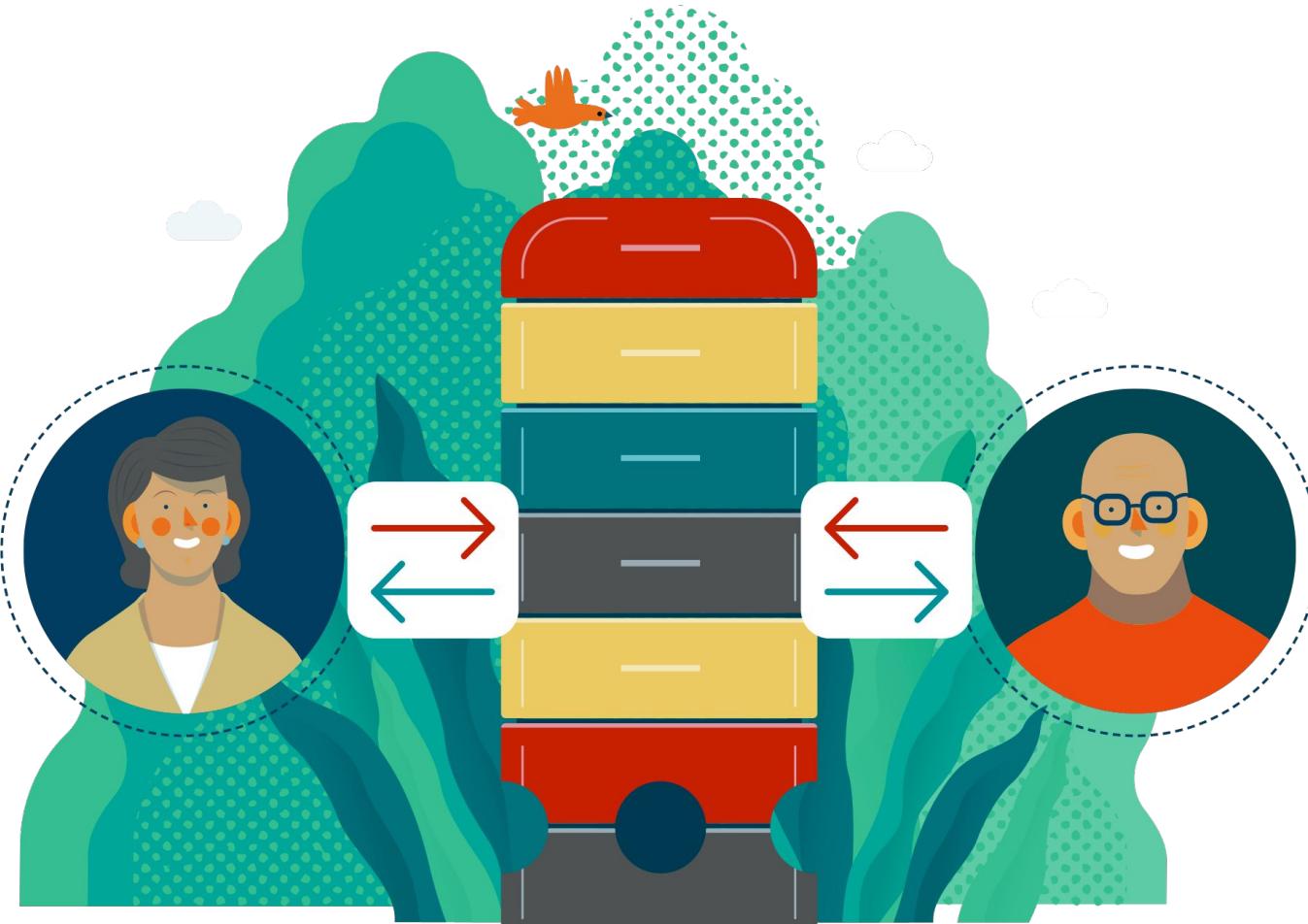
FP-Growth

Objetivo

Encontrar regularidades inherentes en los datos.

Aplicaciones

Análisis de canasta, cross-selling, diseño de catálogo, diseño de campañas, web log (secuencia de clics) y análisis de secuencia de ADN.



Algoritmo FP-Growth

Ejemplo: Min. Soporte = 2

Id	Items
10	Arepa, Café, Chocolate
20	Huevos, Café, Salchichas
30	Arepa, Huevos, Café, Salchichas
40	Huevos, Salchichas

Scan

Item	Sop*
Café	3
Huevos	3
Salchichas	3
Arepa	2

Scan

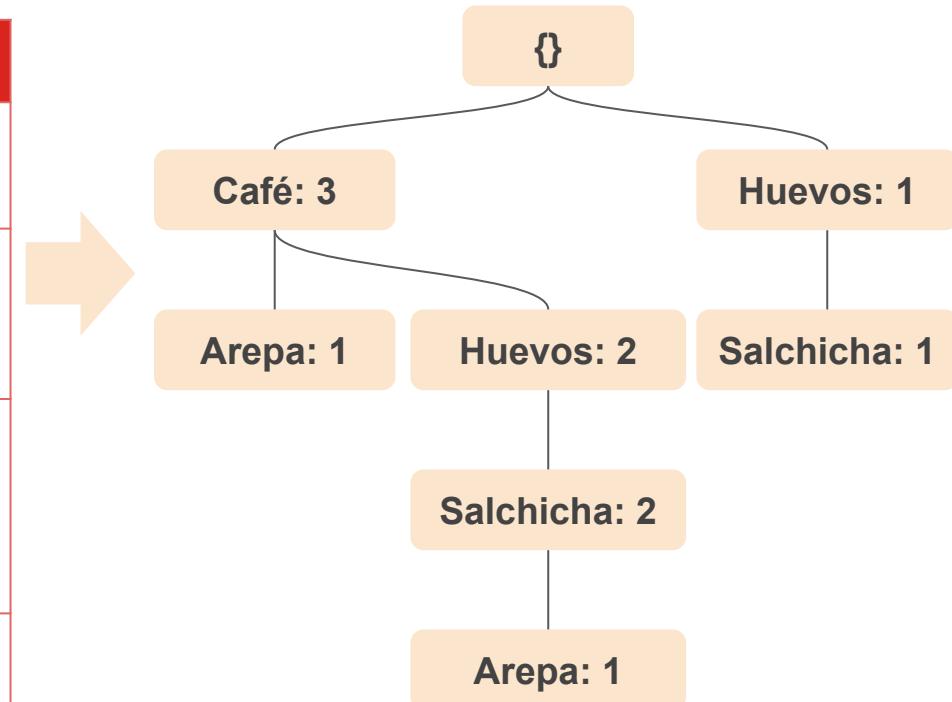
Id	Items ord.
10	Café, Arepa
20	Café, Huevos, Salchichas
30	Café, Huevos, Salchichas, Arepa
40	Huevos, Salchichas

Dataset (D)

Ordenar desc.

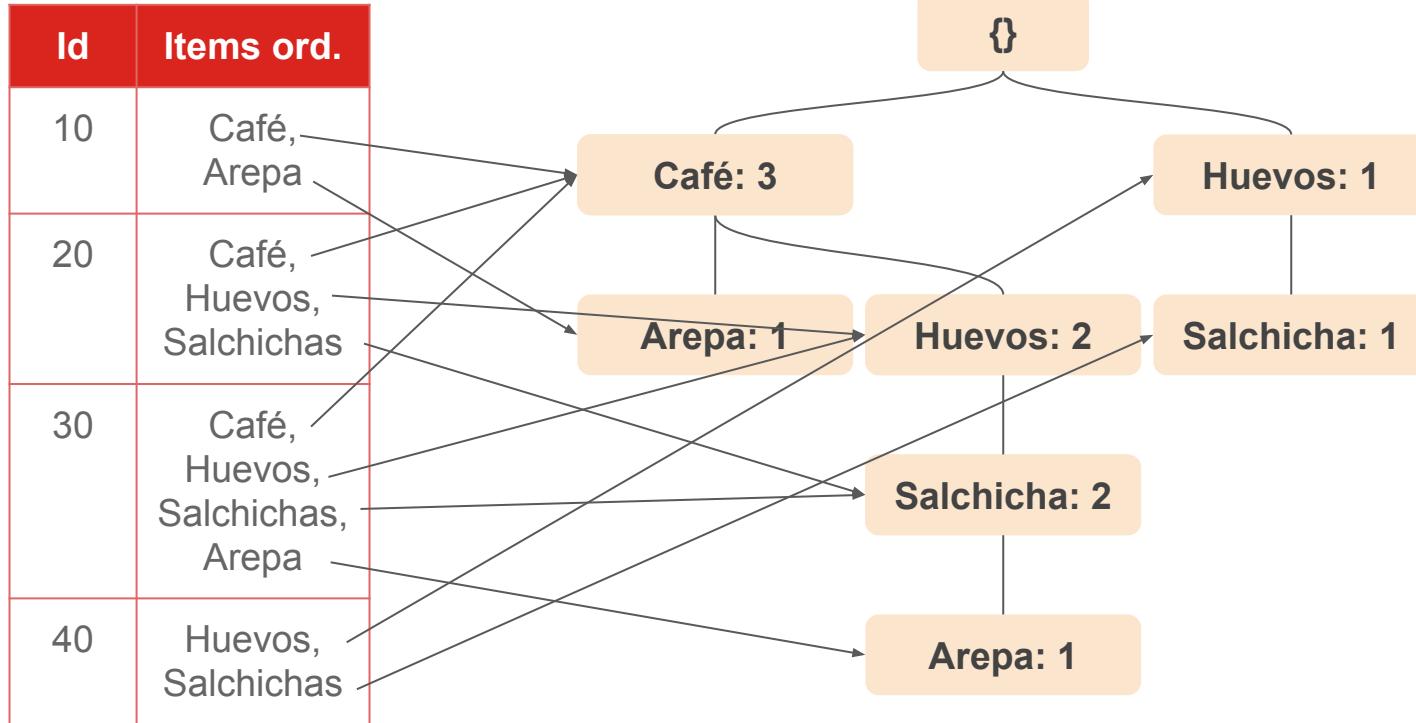
Items ord.

* Soporte: Cantidad de veces que el ítem fue adquirido



Algoritmo FP-Growth

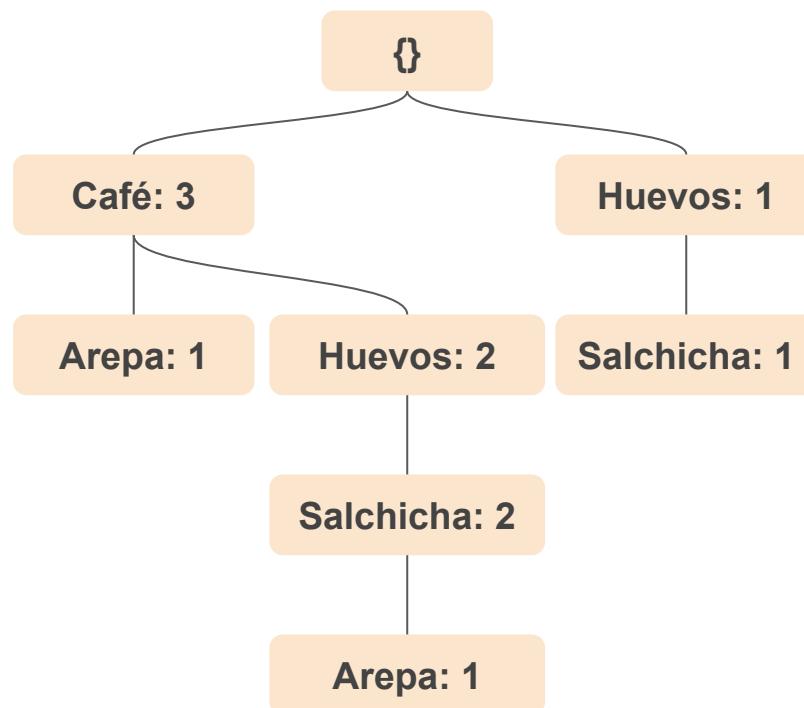
Funcionamiento



Algoritmo FP-Growth

Reglas

Item	Sop*
Café	3
Huevos	3
Salchichas	3
Arepas	2



Item	Patrón condicionante
Huevos	Café: 2
Salchichas	Cafe-Huevos: 2, Huevos: 1
Arepas	Cafe: 1, Cafe-Huevos-Salchichas: 1

Patrones condicionales

* Soporte: Cantidad de veces que el ítem fue adquirido

Ventajas

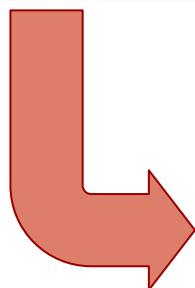


- **Preserva la información** de patrones completa, no rompe los patrones largos de cualquier transacción
- **Elimina la información irrelevante** (los elementos poco frecuentes se van).
- Más **probabilidades tienen de compartirse** en cuanto más frecuentes son.
- **Siempre en menor** que la base de datos original.
- **Se puede particionar y paralelizar.**

Siguientes pasos

Cálculo de lift como medida de asociación

	antecedent	consequent	lift
0	A01-S01-L001-F0001-SF00001	A01-S01-L003-F0005-SF00014	3.078796
1	A01-S01-L001-F0001-SF00001	A01-S01-L004-F0009-SF00023	3.147651
2	A01-S01-L001-F0001-SF00001	A04-S08-L039-F0160-SF00484	2.545631
3	A01-S01-L001-F0001-SF00001	A04-S08-L043-F0170-SF00504	2.054499
4	A01-S01-L001-F0001-SF00001	A06-S13-L058-F0206-SF00619	2.203242
5	A01-S01-L001-F0001-SF00001	A06-S13-L058-F0207-SF00623	2.263044
6	A01-S01-L001-F0001-SF00001	A10-S23-L082-F0254-SF00733	1.361845



$$lift_{AC} = \frac{P(C|A)}{P(C)P(A)} = \frac{P(C|A)}{P(C)} = \frac{Confidence}{P(C)}$$

donde,

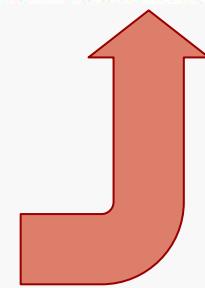
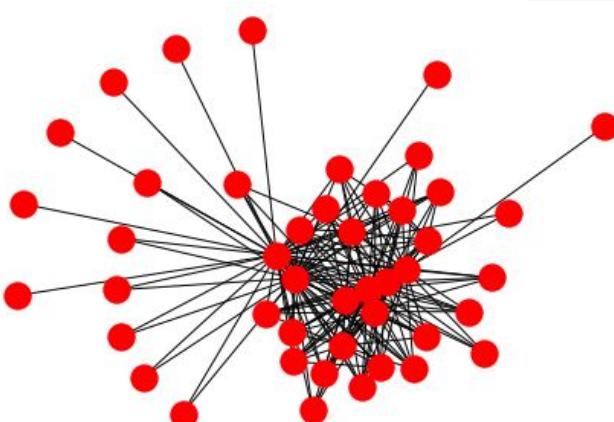
A: Subfamilia Antecedente

C: Subfamilia Consecuente

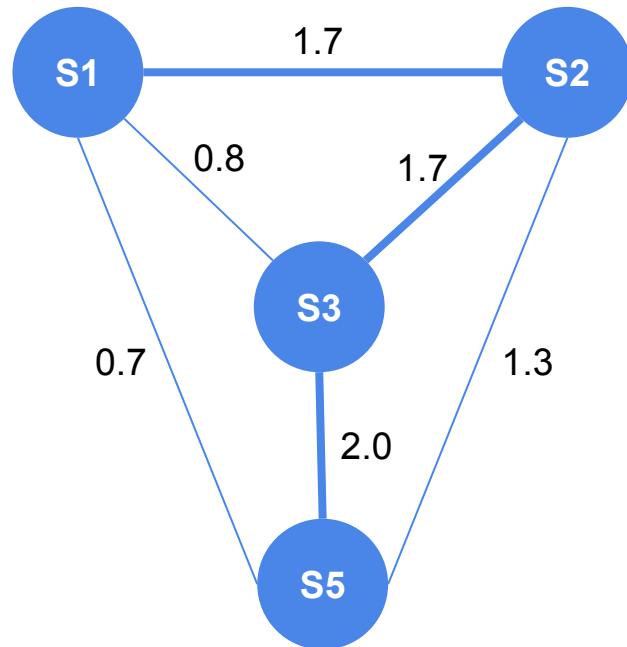
Cálculo medida de centralidad (Degree centrality)

```
'A01-S01-L001-F0001-SF00001': 0.1590909090909091,  
'A01-S01-L003-F0005-SF00014': 0.25,  
'A01-S01-L004-F0009-SF00023': 0.25,  
'A04-S08-L039-F0160-SF00484': 0.34090909090909094,  
'A04-S08-L043-F0170-SF00504': 0.9545454545454546,  
'A06-S13-L058-F0206-SF00619': 0.6136363636363636,  
'A06-S13-L058-F0207-SF00623': 0.6363636363636364,  
'A10-S23-L082-F0254-SF00733': 0.72727272727273,  
'A01-S01-L002-F0003-SF00007': 0.20454545454545456,  
'A06-S13-L058-F0209-SF00627': 0.5454545454545454,  
'A06-S13-L058-F0209-SF00628': 0.6136363636363636,  
'A06-S12-L057-F0203-SF00609': 0.5227272727272727,  
'A01-S01-L004-F0009-SF00022': 0.1590909090909091,  
'A01-S02-L011-F0046-SF00139': 0.045454545454545456,
```

Elaboración red de interacciones



Siguientes pasos



Degree centrality: representa el número de enlaces incidentes en un nodo (número de vínculos).

El grado puede interpretarse en términos de la probabilidad que un **nodo atrape lo que fluye** (lift) **a través de la red** y con esto determinar el nivel de influencia que tiene sobre la red.

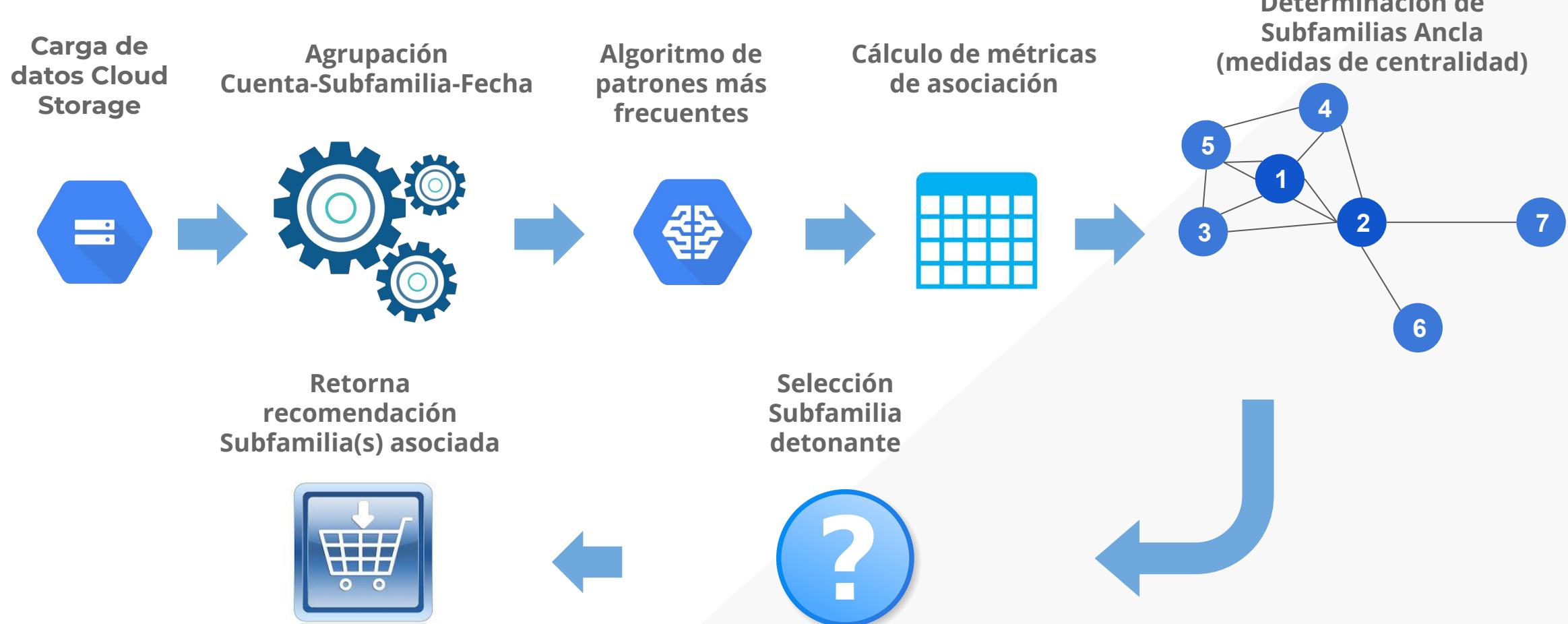
Para la subfamilia 3

$$DC = d_i = \sum_j L_{ij}$$

$$DC' = \frac{d_i}{(L-1)}$$

- El degree centrality (DC) es:
 $0.8+1.7+2.0=4.5$.
- El degree centrality normalizado es:
 $4.5/(8.2-1)=0.625$

Flujo Final



Resultados

- El **entrenamiento** del modelo toma cerca de **2 horas** para procesar **2.500 millones de transacciones**.
- La **respuesta a requests** tarda en promedio menos de **0,5 segundos** para más de **500 usuarios** conectados concurrentemente.



Para Más
información
jhon.garces@globant.com

¡Muchas **Gracias!**

auberth.hurtado@globant.com



Globant ➤