

# UNIDAD 4: ALGORITMOS PARA FLUJOS DE DATOS

## BÚSQUEDA DE LOS ELEMENTOS MÁS RECIENTES

---

Blanca Vázquez

Abril 2020

# INTRODUCCIÓN



# INTRODUCCIÓN: BÚSQUEDA DE ELEMENTOS

**Número total de entradas de cine vendidas**

94.2	92.3	88	52.3
Avengers endgame	Frozen II	Star Wars: The Force Awakens (2015)	Gladiator

los datos están en millones

Pregunta: ¿Cuáles son las películas más populares (tomando como base el número tickets)?

- Supongamos que la película '*Star Wars: episodio IV*' vendió más de 200 millones de entradas.
- Podríamos afirmar que, por el número de entradas, es una película popular.
- Acaso podríamos afirmar que es *¿popular y reciente?*

- Supongamos que la película '*Star Wars: episodio IV*' vendió más de 200 millones de entradas.
- Podríamos afirmar que, por el número de entradas, es una película popular.
- Acaso podríamos afirmar que es *¿popular y reciente?*

## ¿CUÁL DE LAS 2 PELÍCULAS ES MÁS POPULAR?

Flujo de datos para: Frozen II

1	1	1	0	0	1	1
---	---	---	---	---	---	---

Flujo de datos para: Avengers endgame

0	0	0	1	1	0	0
---	---	---	---	---	---	---

- Desarrollado por Datar-Gionis-Indyk-Motwan
- Diseñado para encontrar el número de 1s en una ventana binaria
- La pregunta a resolver es: ¿cuántos 1s hay en los últimos  $k$  bits? donde  $k < N$
- Se basa en la construcción de cubetas\*

...101011000101110110010110...

N = 24 (tamaño de la ventana)

- Para iniciar, cada bit en la ventana tiene una marca de tiempo (la posición en la que llega).
- Siempre empezamos del lado derecho y debe empezar con 1.
- Cada cubeta debe tener al menos un 1.
- Todas las cubetas deben estar en potencias de 2
- Las cubetas no pueden disminuir de tamaño, a medida que nos movemos en el tiempo




# EJEMPLO: ALGORITMO DGIM

Flujo de datos:

...101011000101110110010110...

Paso 1: Crear cubetas

...101011000101110110010110...  


Siempre empezamos del lado derecho y en 1.

...**101011** 000 **10111** 0 **11** 00 **101 1** 0...  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$

Todas las cubetas están en potencia de 2.

# EJEMPLO: ALGORITMO DGIM

...101011000101110110010110...

Siempre empezamos del lado derecho y en 1.

Paso 1: Crear cubetas

...**101011** 000 **10111** 0 **11** 00 **101 1** 0... Todas las cubetas están en potencia de 2.  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$

...<sup>87</sup>**101011** 000 <sup>92</sup>**10111** 0 <sup>95</sup>**11** 00 <sup>98 100</sup>**101 1** 0...

Añadiendo un nuevo bit

...<sup>87</sup>**101011** 000 <sup>92</sup>**10111** 0 <sup>95</sup>**11** 00 <sup>98 100</sup>**101 1** 0... ← <sup>101</sup>**0** Nuevo bit

...<sup>87</sup>**101011** 000 <sup>92</sup>**10111** 0 <sup>95</sup>**11** 00 <sup>98 100 101</sup>**101 1 0** **0**...

No hay cambios en la cubeta!

# EJEMPLO: ALGORITMO DGIM

Añadiendo un nuevo bit

... <sup>87</sup>101011 000 <sup>92</sup>10111 0 <sup>95</sup>11 00 <sup>98</sup>101 <sup>100</sup>1 <sup>101</sup>0 0... ← <sup>102</sup>1 Nuevo bit

... <sup>87</sup>101011 000 <sup>92</sup>10111 0 <sup>95</sup>11 00 <sup>98</sup>101 <sup>100</sup>1 <sup>101</sup>0 0 1.. Ingresamos el nuevo bit

... <sup>87</sup>101011 000 <sup>92</sup>10111 0 <sup>95</sup>11 00 <sup>98</sup>101 <sup>100</sup>1 <sup>101</sup>00 <sup>102</sup>1.. Creamos una nueva cubeta de tamaño 1

$2^2$                        $2^2$                        $2^1$                        $2^1$      $2^0$                        $2^0$

# EJEMPLO: ALGORITMO DGIM

Añadiendo un nuevo bit

$87$   $92$   $95$   $98$   $100$   $101$   $102$   $103$  Nuevo bit  
 $\dots$  **101011** 000 **10111** 0 **11** 00 **101** **1** 00 1..  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$   $2^0$

$87$   $92$   $95$   $98$   $100$   $101$   $102$   $103$   
 $\dots$  **101011** 000 **10111** 0 **11** 00 **101** **1** 00 1 **1**.  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$   $2^0$  Ingresamos el nuevo bit

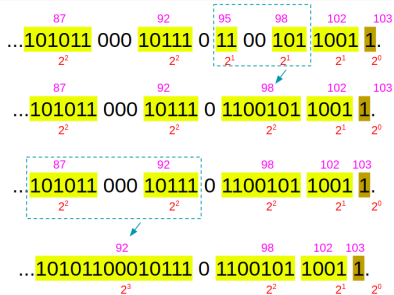
$87$   $92$   $95$   $98$   $100$   $101$   $102$   $103$   
 $\dots$  **101011** 000 **10111** 0 **11** 00 **101** **1** 00 1 **1**.  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$   $2^0$   $2^0$  Creamos una nueva cubeta de tamaño 1

$87$   $92$   $95$   $98$   $100$   $101$   $102$   $103$   
 $\dots$  **101011** 000 **10111** 0 **11** 00 **101** **1** 00 1 **1**.  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^0$   $2^0$   $2^0$  Observamos que tenemos 3 cubetas de tamaño 1: 100,102,103

$87$   $92$   $95$   $98$   $102$   $103$   
 $\dots$  **101011** 000 **10111** 0 **11** 00 **101** **1001** **1**.  
 $2^2$   $2^2$   $2^1$   $2^1$   $2^1$   $2^0$  Combinamos las 2 cubetas más antiguas (más a la izquierda) y la marca de tiempo será la más reciente (la más a la derecha)

# EJEMPLO: ALGORITMO DGIM

Añadiendo un nuevo bit



Observamos que tenemos 3 cubetas de tamaño 2: 95, 98, 102

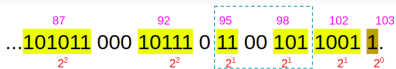
Combinamos las 2 cubetas más antiguas (más a la izquierda) y la marca de tiempo será la más reciente (la más a la derecha)

Observamos que tenemos 3 cubetas de tamaño 4: 98, 92, 87

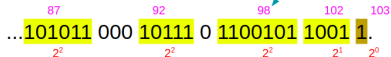
Combinamos las 2 cubetas más antiguas (más a la izquierda) y la marca de tiempo será la más reciente (la más a la derecha)

# EJEMPLO: ALGORITMO DGIM

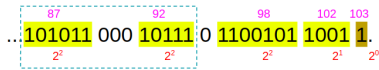
Añadiendo un nuevo bit



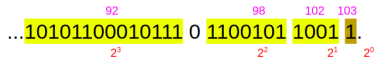
Observamos que tenemos 3 cubetas de tamaño 2: 95, 98, 102



Combinamos las 2 cubetas más antiguas (más a la izquierda) y la marca de tiempo será la más reciente (la más a la derecha)



Observamos que tenemos 3 cubetas de tamaño 4: 98, 92, 87



Combinamos las 2 cubetas más antiguas (más a la izquierda) y la marca de tiempo será la más reciente (la más a la derecha)

Añadiendo un nuevo bit

Repetimos el proceso, ya sea 0 o 1

.

.

.

.

¿En qué momento se detiene el algoritmo?

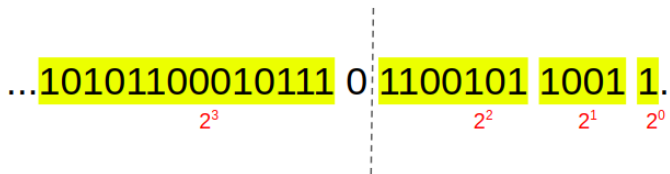
- Continuamos cuando el tiempo actual menos el intervalo de tiempo más a la izquierda sea menor que  $N$  (tamaño de la venta)

Tiempo actual = 103, intervalo más a la izquierda = 92,  $N = 24$

$103 - 92 = 11 < 24$ , **por lo tanto continuamos**

Cuando el resultado es mayor o igual, nos detenemos.

¿Cuántos 1s hay en los últimos 12 bits?

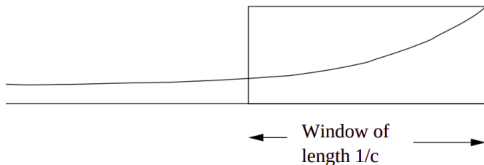


¿Cuántos 1s hay en los últimos 12 bits?  $2^0 + 2^1 + 2^2 = 7$



- En el caso de películas, este algoritmo puede aplicarse (miles)
- Sin embargo, falla cuando estamos hablando de millones de registros como Amazon o Twitter.
- Otros algoritmos, como el desvanecimientos de ventanas puede ayudar definiendo un factor como  $10^{-6}$  o  $10^{-9}$

## Desvanecimiento de ventanas



$$\sum_{i=0}^{t-1} a_{t-i} (1-c)^i$$

Imagen tomada de Jure Leskovec, Anand Rajaraman, Jeff Ullman

En esta unidad estudiamos:

- Modelo general para el procesamiento de flujos de datos
- Muestreo (tamaño fijo, aleatorio, ventanas deslizantes)
- Filtrado (Algoritmo de Bloom)
- Conteo (Algoritmo de Flajolet–Martin)
- Estimación de momentos (Algoritmo de AMS)
- Búsqueda de elementos más recientes (Algoritmo de DGIM)

- DGIM Algorithm, Madhuragj 2019,  
<https://medium.com/fnplus/dgim-algorithm-169af6bb3b0c>
- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman, Stanford University,  
<http://www.mmds.org>