Unidad 4: Algoritmos para flujos de datos

Conteo

Blanca Vázquez

Abril 2020

Introducción

- Hoy en día los flujos de datos se generan por múltiples fuentes: satélites, radares, celulares, redes sociales, sitios web, sensores, etc.

Introducción

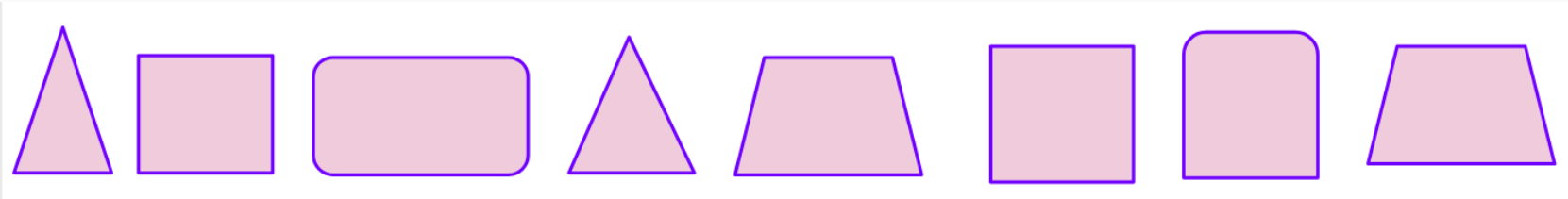
- Con estas fuentes de datos, podríamos estar interesados en el *número de visitas o número de clics hacia un producto o página web, o en el número de lecturas de un sensor, el número tweets sobre un tema, etc.*

Explicando con manzanas

Productos	Piezas vendidas
Manzanas	20
Piñas	5
Plátanos	12
Manzanas	9
Uvas	15
Plátanos	8

¿Cuántas tipos de frutas diferentes se vendieron en un día?

Introducción: conteo de elementos distintos



Ejemplo de flujos de datos

Problema del conteo de elementos distintos

- Encontrar el número de elementos distintos en un flujo de datos con elementos repetidos
- También conocido como problema de estimación de la cardinalidad

Dado un strem s de elementos $x_1, x_2, ..., x_n$, encontrar el número de elementos distintos n, donde $n = |\{x_1, x_2, ..., x_n\}|$

Problema del conteo de elementos distintos

Dado un strem s de elementos $x_1, x_2, ..., x_n$, encontrar el número de elementos distintos n, donde $n = |\{x_1, x_2, ..., x_n\}|$

Ejemplo:
Dado $s = \{a, b, a, c, d, b, d\}$ entonces n es igual a
$n = |\{a, b, c, d\}| = 4$

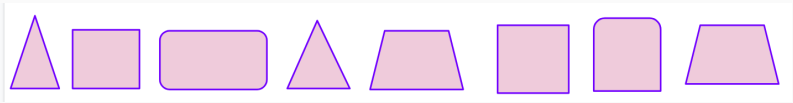
¿Elementos únicos?

El conteo de elementos únicos puede representar:

- Direcciones IP que pasan a través de un router
- Número de visitantes únicos a un sitio web
- Secuencias de ADN
- Dispositivos IoT

Ejemplo 2: conteo de elementos distintos

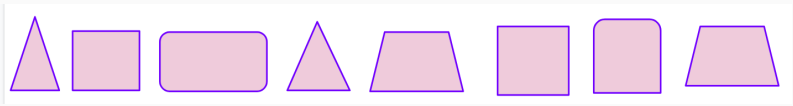
Supongamos que tenemos un flujo de datos m con x elementos



¿Cuántos elementos distintos tenemos?

Ejemplo 2: conteo de elementos distintos

Supongamos que tenemos un flujo de datos *m* con *x* elementos

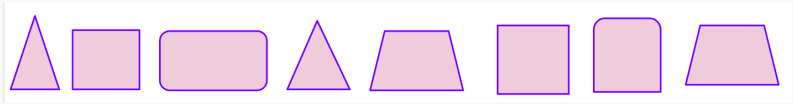


¿Cuántos elementos distintos tenemos?

- Si *m* es pequeña:
  - Solución: generar un diccionario
  - Memoria: O(m)
  - Costo computacional: O log(m) para almacenamiento y para búsqueda

Ejemplo 2: conteo de elementos distintos

Supongamos que tenemos un flujo de datos m con x elementos



¿Cuántos elementos distintos tenemos?

- Si m es grande:
  - Almacenamiento de todos los elementos: imposible!
  - Memoria: muy alta
  - Costo computacional: doblemente alta!

Algoritmos que abordan el conteo de elementos distintos

- Algoritmo de Flajolet–Martin
- Algoritmo de HyperLogLog

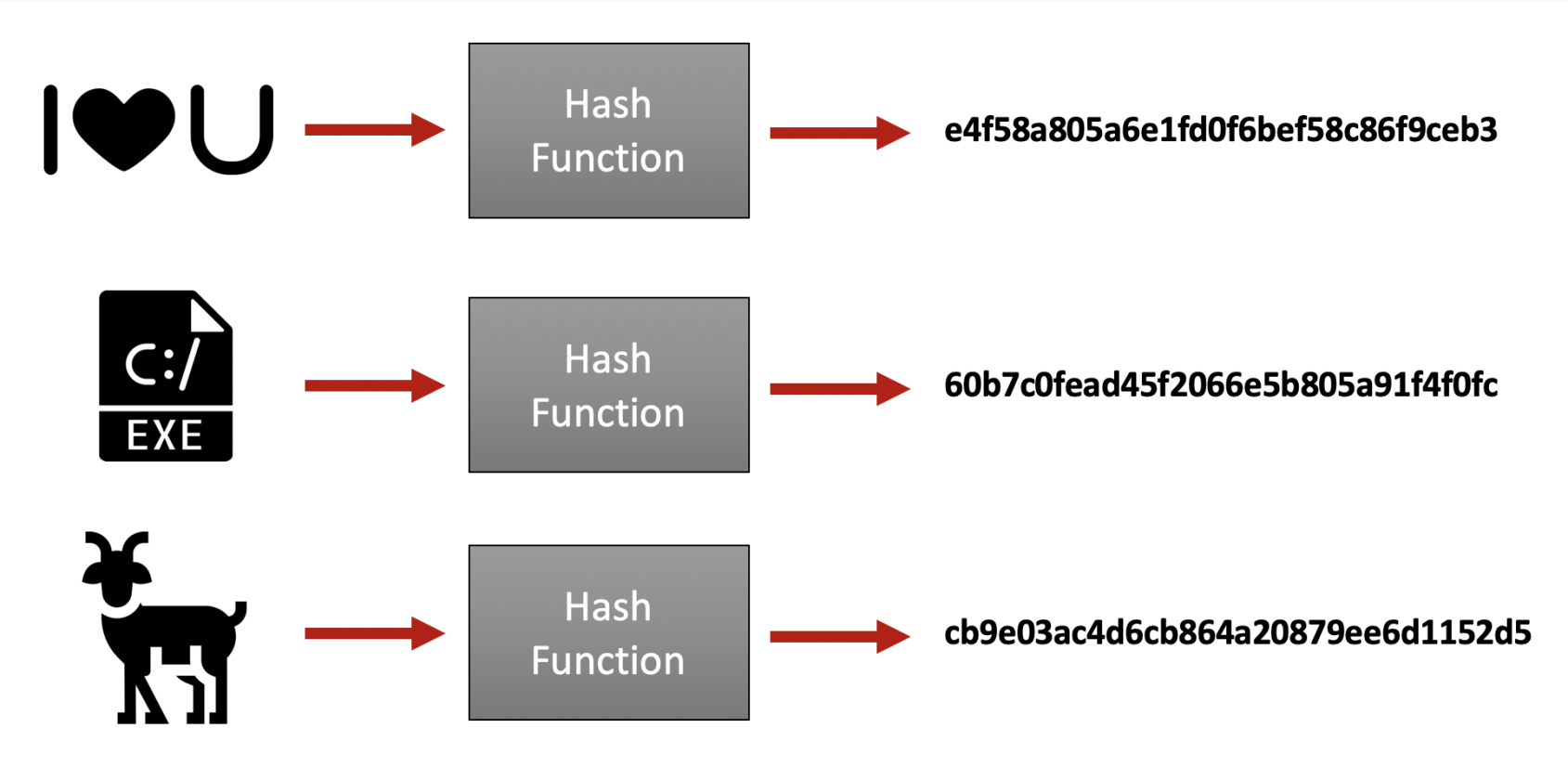
Algoritmo de Flajolet–Martin

Es un algoritmo para aproximar el número de elementos distintos en un flujo de datos

- Este algoritmo fue creado por Philippe Flajolet y G. Nigel Martin
- *Probabilistic Counting Algorithms for Data Base Applications, 1985*
- Recordemos: no buscamos cuántas veces apareció un elemento, sino el número de elementos distintos

Intuición: Algoritmo de Flajolet–Martin

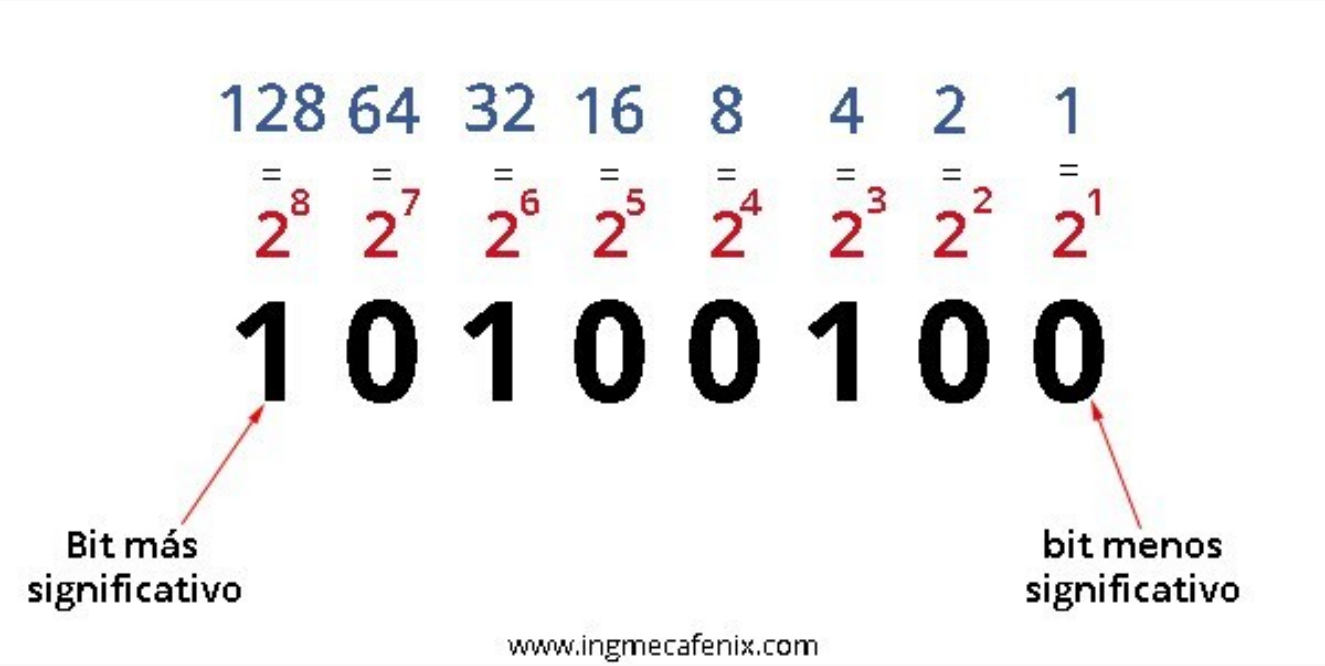
- Transformar los elementos de entrada sobre una función hash binaria con distribución uniforme e independiente de probabilidad.



Intuición: Algoritmo de Flajolet–Martin

- La propiedad de distribución uniforme (de la función hash) permite entonces prever que *la mitad de los elementos tendrán un 1 en el bit menos significativo*, que una cuarta parte de los elementos tendrán un 1 en el segundo bit menos significativo y así sucesivamente

Bits menos significativo (LSB)



LSB es la posición de bit en un número binario que tiene el menor valor (el situado más a la derecha).

Algoritmo de Flajolet–Martin

A partir de la idea de identificar los bits menos significativos es posible realizar una aproximación probabilista del número de elementos distintos en el flujo de datos

Input Stream= 1,3,2,1,2,3,4,3,1

Let's h(x)= 3X+1mod5

Solve:

Calculation	Reminder	Binary Conversion
h(1)=3(1)+1mod5=	4	100
h(3)=3(3)+1mod5 =	0	000
h(2)=3(2)+1mod5 =	2	010
h(1)=3(1)+1mod5 =	4	100
h(2)=3(2)+1mod5 =	2	010
h(3)=3(3)+1mod5 =	0	000
h(4)=3(4)+1mod5 =	3	011
h(3)=3(3)+1mod5 =	0	000
h(1)=3(1)+1mod5 =	4	100

Número de ceros: 2,0,1,2,1,0,0,0,2

r = 2

$2^r = 2^r = 4$

En otras palabras...

- La idea detrás del algoritmo de Flajolet-Martin es que cuantos más elementos diferentes veamos en el flujo de datos, **más valores hash diferentes veremos**.
- A medida que vemos valores hash más diferentes, es más probable que uno de estos valores sea **'inusual'**.
- Un valor inusual será *aquel que termine en muchos ceros***

Ventajas - desventajas

- Usa menos cantidad de memoria para aproximar el número de elementos únicos
- Una de las desventajas del algoritmo de Flajolet–Martin es la suposición de la generación de claves hash totalmente aleatorias

Referencias

Victoria López , Monitorización y Análisis del Cambio Social a partir de Big Data, 2014 en https://es.slideshare.net/vlopezlo/present-federico-castanedo