

UNIDAD 4: ALGORITMOS PARA FLUJOS DE DATOS

MUESTREO

Blanca Vázquez

Abril 2020



🔍 Search Google or type a URL

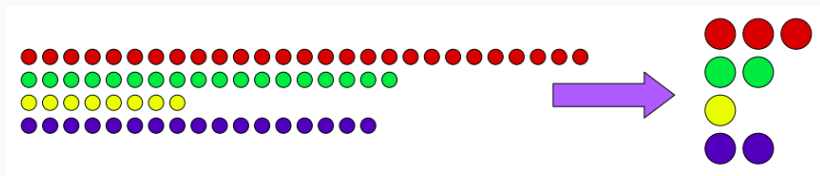


Google procesa **81,226 búsquedas** por segundo

Uno de los principales retos del procesamiento del flujo de datos parte del **almacenamiento** y las futuras consultas que podemos hacer.

MUESTREO

Debido a que en la mayoría de los casos no es posible almacenar todos los flujo de datos que se reciben, uno de las soluciones es hacer **muestreo**.



Muestrear consiste en almacenar una porción de los datos que se reciben

- **Ventaja:** el costo computacional es bajo, debido a que estamos solo usando una porción de los datos recibidos.
- **Desventaja:** ¿cómo saber que tan largo es el flujo de datos?, ¿cada cuánto tiempo debemos muestrear?, ¿cómo hacemos el muestreo, por tipo de proveedor?

Existen algunas soluciones que aplican muestreo:

- Ventanas deslizantes
- Muestreo de tamaño fijo
- Muestreo aleatorio

VENTANAS DESLIZANTES

Es una técnica útil para el procesamiento de flujos de datos, en dónde las consultas se realizan sobre una **ventana** de tamaño w . Cuando un nuevo elemento llega en el tiempo t este expira en el momento $t + w$.

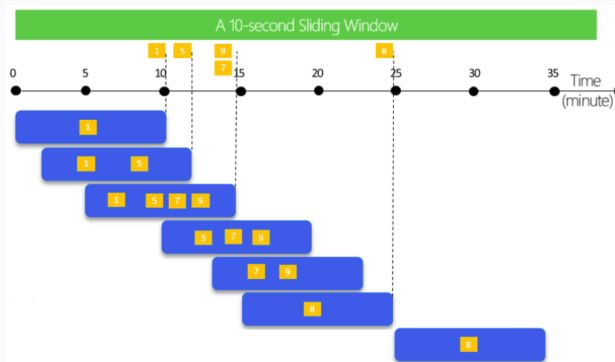


Imagen tomada de Azure Stream Analytics

VENTANAS DESLIZANTES

En este ejemplo el tamaño de la ventana deslizante es 6, observamos el traslape entre datos.

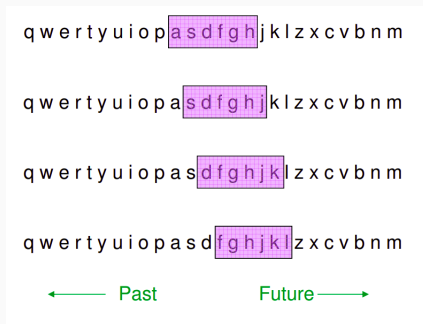


Imagen tomada de J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

Ventana de promedio deslizante

Calificaciones

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

Ventana de promedio deslizante

Calificaciones

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

Definir:

- Tamaño de la ventana: 3

Ventana de promedio deslizante

Calificaciones

Tamaño de la ventana: 3

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

$\text{suma_ventana} = (10 + 7.8 + 6.8)$

$\text{suma_ventana} = 24.6$

$\text{resultado} = 24.6 / 3$

$\text{resultado} = 8.2$

Secuencia resultante: [8.2]

Ventana de promedio deslizante

Calificaciones

Tamaño de la ventana: 3

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

$\text{suma_ventana} = 24.3 - 10 + 8.0$

$\text{suma_ventana} = 22.5$

$\text{resultado} = 22.5/3$

$\text{resultado} = 7.43$

Secuencia resultante: [8.2, 7.43]

Ventana de promedio deslizante

Calificaciones

Tamaño de la ventana: 3

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

$\text{suma_ventana} = 22.5 - 7.8 + 9.2$

$\text{suma_ventana} = 23.9$

$\text{resultado} = 23.9/3$

$\text{resultado} = 7.96$

Secuencia resultante: [8.2, 7.43, 7.96,]

Ventana de promedio deslizante

Calificaciones

Tamaño de la ventana: 3

10.0	7.8	6.8	8.0	9.2	9.0
------	-----	-----	-----	-----	-----

$\text{suma_ventana} = 23.9 - 6.8 + 9.0$

$\text{suma_ventana} = 26.1$

$\text{resultado} = 26.1/3$

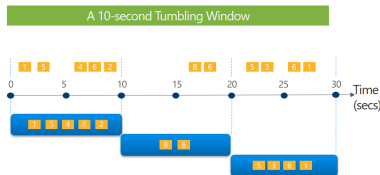
$\text{resultado} = 8.7$

Secuencia resultante: [8.2, 7.43, 7.96, 8.7]

VENTANAS DE SALTOS DE TAMAÑO CONSTANTE

Se usan para segmentar una transmisión de datos en segmentos de tiempo distintos y realizar una función con ellos. Una de las características principales es que no existe traslape entre los datos.

Tell me the count of tweets per time zone every 10 seconds



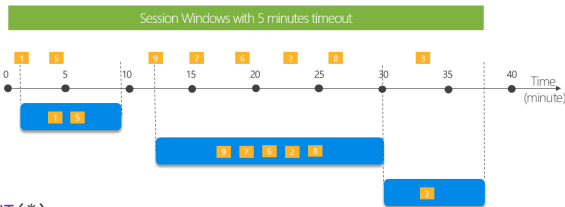
```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Imagen tomada de Azure Stream Analytics

VENTANAS DE SESIÓN

Agrupan eventos que llegan a la misma hora, filtrando los periodos en dónde no se recibe ningún dato. En este caso se deben fijar los parámetros de tiempo de espera y duración máxima.

Tell me the count of tweets that occur within 5 minutes to each other.



```
SELECT Topic, COUNT(*)  
FROM TwitterStream TIMESTAMP BY CreatedAt  
GROUP BY Topic, SessionWindow(minute, 5, 10)
```

Imagen tomada de Azure Stream Analytics

Existen algunas soluciones que aplican muestreo:

- Muestreo de tamaño fijo
- Muestreo aleatorio
- Ventanas deslizantes
- Estadístico
- Determinístico

Todos los elementos deben tener la misma probabilidad de ser seleccionados.

Ejemplo: Cobertura de la vacuna anti-sarampión entre 1,200 niños de la escuela 'Juan Escutia':

- Muestra: 60 niños
- Hacer una lista de todos los niños
- Numerarlos del 1 al 1,200
- Selección aleatoria de 60 números (probabilidad igual)

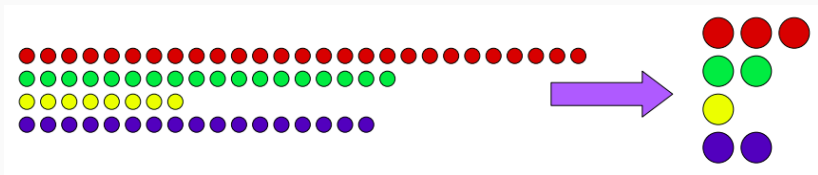
La selección de los elementos se basa en criterios de selección, es decir, los elementos seleccionados no son aleatorios.

Ejemplo: únicamente se registraran a los pacientes que acudan a la clínica en cierto día u horario particular.

- Es posible que los elementos seleccionados sean poco representativos de todos los conjuntos generados

MUESTREO DE TAMAÑO FIJO

Consiste en muestrear una porción fija de los elementos recibidos (digamos 1 de cada 10 recibidos)



MUESTREO DE TAMAÑO FIJO (EJEMPLO)

- Motor de búsqueda: Google
- Entrada: flujos de datos en forma de tupla

User (IP)	Query	Time
-----------	-------	------

- Pregunta: ¿Qué fracción de las consultas de un usuario son duplicadas?

¿Qué fracción de las consultas de un usuario son duplicadas?

- Supongamos que cada usuario realiza X número de consultas únicas y D número de consultas repetidas
- El total de consultas que el usuario hace es: $X + 2D$
- Usando el muestreo, mantendríamos $1/10$ de todas las consultas
 - $x/10$ (de todas las consultas únicas)
 - $2D/10$ (de las consultas duplicadas)

¿Qué fracción de las consultas de un usuario son duplicadas?

- De las D consultas duplicadas, en la muestra solo tendríamos $D/100$
 - $D/100 = 1/10 * 1/10 * D$
- De todo el conjunto de preguntas repetidas, $18D/100$ aparecería realmente una vez.
 - $18D/100 =$ es la probabilidad de que una de las ocurrencias estará en el $1/10$ del stream seleccionado, mientras que el otro estará en el $9/10$ (que no fue seleccionado)
 - $18D/100 = ((1/10 * 9/10) + (9/10 * 1/10)) * D$

¿Qué fracción de las consultas de un usuario son duplicadas?

- La respuesta basada en muestreo fijo sería:

$$\frac{D}{\frac{X}{10} + 19D}$$

- Por lo tanto:

$$\frac{\frac{D}{100}}{\frac{X}{10} + \frac{D}{100} + \frac{18D}{100}} = \frac{D}{10X + 19D}$$

¿Qué fracción de las consultas de un usuario son duplicadas?

- Como observamos hacer el muestreo tomando una muestra de cada usuario, puede arrojar resultados pocos confiables.
- Y si en lugar de tomar 1/10 de las búsquedas de cada usuario, tomamos 1/10 de todos los usuarios.
 - Es decir, vamos a almacenar TODAS las búsquedas
 - Descartando las consultas del resto de usuarios 9/10
 - Tomando la IP del usuario (como ID), usamos una función hash que almacene las consultas hacia cubetas
 - Como resultado, tendríamos una muestra más representativa.

Una de las técnicas más comunes en el muestreo aleatorio es el 'Reservoir Sampling'

- Consiste en muestrear los primeros M elementos recibidos y los mantiene en memoria (reserva)

- Almacena los primeros **s** elementos del stream en la muestra **S**.
- Supongamos que hemos visto **n-1** elementos, y ahora recibimos el n^{th} elemento ($n > s$)
- Con probabilidad **s/n**, mantenemos el elemento n^{th} , reemplazando uno de los **s** elementos en la muestra S

Apache Parquet

- Initial effort by Twitter & Cloudera
- Open source storage format
 - Hybrid storage model (PAX)
- Widely used in Spark/Hadoop ecosystem
- One of the primary formats used by Databricks customers



 databricks

Parquet is a binary data storage format that, in combination with Spark, enables fast queries by getting you **just the data you need**, **getting it efficiently**, and keeping much of the **work out of Spark**.



PARQUET

	Parquet
Usability	Good!
Administration	None!
Spark Integration	FANTASTIC!!
Resource Efficiency	WONDERFUL!! (Storage, I/O, Data cardinality)
Scalability	FANTASTIC!!
CO\$\$\$\$T	¢ ¢ ¢
QUERY TIME	GOOD!!