DATOS MASIVOS I

UNIDAD IV ALGORITMOS PARA FLUJOS DE DATOS

MODELO DE FLUJO DE DATOS

13 de Marzo de 2023

Tres unidades hasta ahora

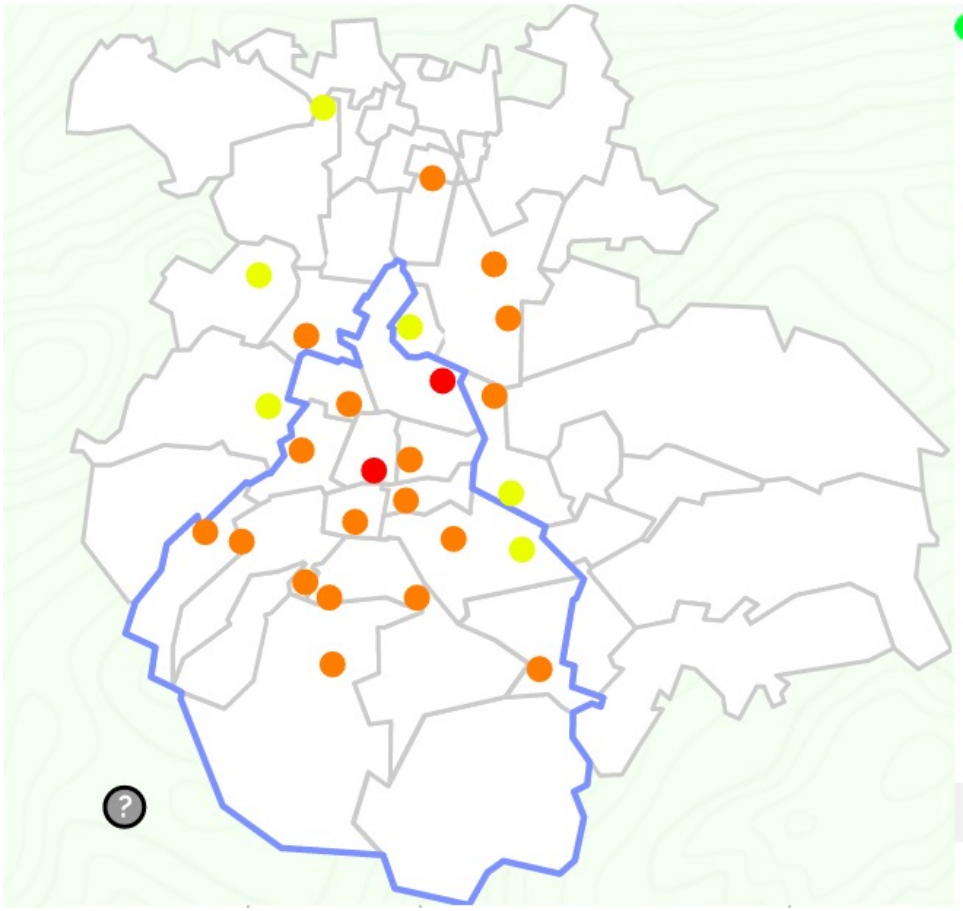
Contenido Temático	
Tema	**Subtemas**
1. Conceptos básicos	
1.1	Definición y características
1.2	Generación, procedencia y preparación de datos
1.3	Consideraciones estadísticas y computacionales de los datos masivos
1.4	El principio de Bonferri
1.5	Privacidad y riesgo
1.6	Modelos de computación para datos masivos
2. Modelo de mapeo y reducción	
2.1	Sistema de almacenamiento y procesamiento distribuido
2.2	Modelo de programación
2.3	Algoritmos con el modelo de mapeo y reducción
2.4	Extensiones
2.5	El modelo costo-comunicación
2.6	Teoría de la complejidad para el modelo de mapeo y reducción
3. Búsqueda de elementos similares	
3.1	Medidas de similitud
3.2	Resúmenes de conjuntos con preservación de similitud
3.3	Funciones hash sensibles a la localidad
3.4	Métodos para altos grados de similitud
3.5	Aplicaciones

Sensores industriales

Los sensores industriales pueden capturar grandes cantidades de datos

Imagen tomada de commons.wikimedia.org

Estaciones de monitoreo de la calidad del aire



BUENA ACEPTABLE MALA MUY MALA EXTREMADAMENTE MALA SIN DATOS O EN MANTENIMIENTO

Ciudad de México, lunes 13 de abril de 2020 19 horas 27 °C

ÍNDICE AIRE Y SALUD ZMVM

Índice de Calidad del Aire CDMX

Índice AIRE Y SALUD: **MUY MALA**
Contaminante(s): O_3-8h
Riesgo: MUY ALTO

Calidad del aire: **REGULAR**
Contaminante: O_3
Índice: 82
Estación: FAC-FES Acatlán
Ver mapa

Recomendaciones:

Recomendación UV:
0 NO NECESITA PROTECCIÓN

Hoy no circulan: 5 y 6 Próx. sábado: H1 Impar H2 Todos Este mes verifican:

Estaciones de monitoreo de la calidad del aire

O_3

Fecha de consulta: 2020-04-13

Unidad de Medida: Partes por Millón (ppm)

Hora	AJM	ATI	BJU	CAM	CCA	CHO	CUA	CUT	FAC	FAR	GAM	HGM	IZT	LLA	LPR	MER	MGH	NEZ	PED	SAC	SAG	SFE	TAH	TLA	TLI	UIZ	UAX	VIF	XAL
1																													
2																													
3																													
4	0.043	0.017	0.014	0.010	0.018			0.002	0.019	0.026	0.016		0.013	0.005	0.014	0.004	0.020	0.023	0.025	0.033	0.024	0.039	0.020	0.016		0.016	0.023	0.004	
5	0.042	0.007	0.012	0.004	0.018			0.002	0.014	0.021	0.005		0.010	0.003	0.010	0.005	0.016	0.018	0.022	0.018	0.014	0.035	0.009	0.008		0.005		0.003	
6	0.038	0.001	0.008	0.005	0.007			0.002	0.006	0.011	0.004		0.003	0.003	0.001	0.004	0.004	0.014	0.010	0.007	0.007	0.022	0.016	0.002		0.004	0.012	0.002	
7	0.033	0.000	0.005	0.004	0.004		0.035	0.002	0.003	0.009	0.004	0.008	0.002	0.003	0.004	0.003	0.002	0.005	0.010	0.003	0.006	0.004	0.011	0.003		0.005	0.007	0.003	
8	0.035	0.001	0.006	0.007	0.011		0.034	0.002	0.005	0.011	0.013	0.014	0.008	0.004	0.007	0.008	0.008	0.004	0.016	0.005	0.006	0.007	0.011	0.004		0.009	0.008	0.010	
9	0.041	0.020	0.021	0.021	0.029		0.040	0.010	0.010	0.028	0.027	0.022	0.019	0.013	0.017	0.016	0.021	0.015	0.020	0.014	0.016	0.034	0.026	0.012		0.016	0.027	0.022	
10	0.049	0.031	0.038	0.034	0.050		0.042	0.035	0.022	0.042	0.043	0.035	0.037	0.035	0.022	0.036	0.034	0.034	0.046	0.035	0.027	0.044	0.050	0.026		0.037	0.043	0.044	
11	0.064	0.039	0.052	0.049	0.064		0.054	0.050	0.038	0.058	0.063	0.060	0.050	0.040	0.036	0.060	0.053	0.051	0.060	0.060	0.047	0.056	0.066	0.037		0.060	0.058	0.057	
12	0.079	0.046	0.061	0.062	0.077		0.070	0.058	0.047	0.068	0.076	0.073	0.062	0.048	0.056	0.075	0.069	0.057	0.081	0.066	0.050	0.072	0.077	0.051		0.066	0.068	0.057	
13	0.093	0.052	0.080	0.081	0.093		0.090	0.060	0.066	0.081	0.095	0.089	0.080	0.066	0.076	0.093	0.083	0.063	0.094	0.072	0.065	0.092	0.086	0.059		0.072	0.073	0.055	
14	0.097	0.060	0.086	0.097	0.100		0.099	0.062	0.078	0.093	0.107	0.110	0.091	0.082	0.064	0.103	0.107	0.083	0.103	0.095	0.081	0.118	0.077	0.078		0.095	0.089	0.065	
15	0.100	0.069	0.090	0.116	0.107		0.057	0.069	0.092	0.104	0.114	0.112	0.101	0.088	0.080	0.112	0.116	0.090	0.106	0.074	0.094	0.098	0.053	0.086		0.107	0.099	0.074	
16		0.072	0.094	0.121	0.105		0.073	0.085	0.073	0.111	0.117	0.113	0.100	0.098	0.094	0.112	0.110	0.063	0.100	0.063		0.072	0.054	0.099		0.087	0.096	0.087	
17	0.076	0.068	0.079	0.116	0.084		0.073	0.086	0.071	0.077	0.106	0.101	0.075	0.098	0.086	0.099	0.085	0.060	0.077	0.062	0.093	0.073	0.068	0.092		0.074	0.071	0.092	
18	0.069	0.059	0.062	0.082	0.073		0.061	0.062	0.086	0.059	0.067		0.064	0.062	0.056	0.071	0.069	0.055	0.065	0.063	0.052	0.062	0.079	0.077		0.071	0.070	0.080	
19	0.065	0.036	0.051	0.065	0.065		0.051	0.052	0.050	0.064	0.059		0.060	0.051	0.050	0.063	0.049	0.057	0.057	0.068	0.055	0.056	0.072	0.052		0.065	0.065	0.052	

Consulta el índice por zonas

Interpretación del Índice AIRE Y SALUD	
Concentraciones	Condición
0-0.051	Buena
>0.051 y 0.095	Aceptable
>0.095 y 0.135	Mala
>0.135 y 0.175	Muy Mala
>0.175	Extremadamente Mala
M	Mantenimiento

Los datos presentados en esta sección son preliminares y podrían sufrir modificaciones durante las siguientes etapas de validación.

Datos de imágenes: Otro ejemplo

Aproximadamente existen en órbita **5,000 satélites** que captan imágenes multiespectrales de la Tierra de **resolución media y alta**.
Aproximadamente capturan y envían: millón y medio de imágenes diarias

Datos de imágenes: Otro ejemplo

Cámaras de vigilancia generalmente producen imágenes de baja resolución (en comparación con los satélites), sin embargo el intervalo de envío es de 1 segundo.
Londres tiene alrededor de 6 millones de cámaras.

Datos de sensores

Un sensor en el océano envía cada hora la temperatura del agua a una estación hidrológica (tasa de envío baja 4kb)

Internet y tráfico web

- Correos electrónicos.
- Mensajes en mensajería instantánea.
- Envío de imágenes.

[1]Fuente: https://www.internetlivestats.com/google-search-statistics/

Datos de sensores

Un problema interesante sería:

- Un millón de sensores
- Cada uno enviando sus datos en una tasa de 10kb/seg

Un problema interesante sería:

- Esto replicado cada 150 millas
- El océano Pacífico tiene 9,320.6 millas de norte a sur

Internet y tráfico web

- Google procesa 81,226 búsquedas por segundo
- 3,500 millones de búsquedas por día

[1]Fuente: https://www.internetlivestats.com/google-search-statistics/

Internet y tráfico web

- Cada pregunta viaja 1,500 millas (hacia el centro de datos y de regreso)
- La respuesta a una consulta tarda 2 segundos (usando 1,000 computadoras)

[1]Fuente: https://www.internetlivestats.com/google-search-statistics/

Internet y tráfico web

- En 1999, a Google le tomó un mes indexar ≈ 50 millones de páginas.
- En el 2012 le tomó un minuto.

[1]Fuente: https://www.internetlivestats.com/google-search-statistics/

Flujo de datos

- Son datos que se generan constantemente (tiempo real) a partir de cientos/miles/millones de fuentes de datos.

Flujo de datos

- Normalmente los datos son enviados simultáneamente en conjuntos de tamaño pequeño (*kbs*)
- Dinámicos

Flujo de datos

- Normalmente los datos son enviados simultáneamente en conjuntos de tamaño pequeño (*kbs*)
- Dinámicos
- Si los datos no se almacenan o se procesan rápido, estos se perderán

Flujo de datos: Contenido

- Atributos:

Cada atributo representa un tipo de dato (segmento, geo-localización, ID, etecétera).

Flujo de datos

- Marca de tiempo:

Indica hora y fecha de los datos generados

Flujo de datos

- Dato crudo:

Contiene la información original generada por la fuente de datos

Ejemplo de fuentes de flujos de datos

- Monitoreo
- Dispositivos IoT
- Internet y tráfico web (por ej. secuencias de páginas visitadas (*clickstream*))

Ejemplo de fuentes de flujos de datos

- Transacciones financieras
- Video juegos en línea
- Videos

Retos en el procesamiento de flujo de datos

- Memoria limitada para almacenar los datos
- Debido a la vasta cantidad de datos, no es siempre posible generar respuestas exactas

Retos en el procesamiento de flujo de datos

- Se espera que la calidad de la respuesta sea confiable

- ¿Cómo trabajar con los datos:
  - Selección aleatoria,
  - tomar los últimos, … ?

¿Cómo se procesan los flujos de datos?

Un procesador de flujos de datos es un tipo de Sistema de Administración de Datos (DSMS).

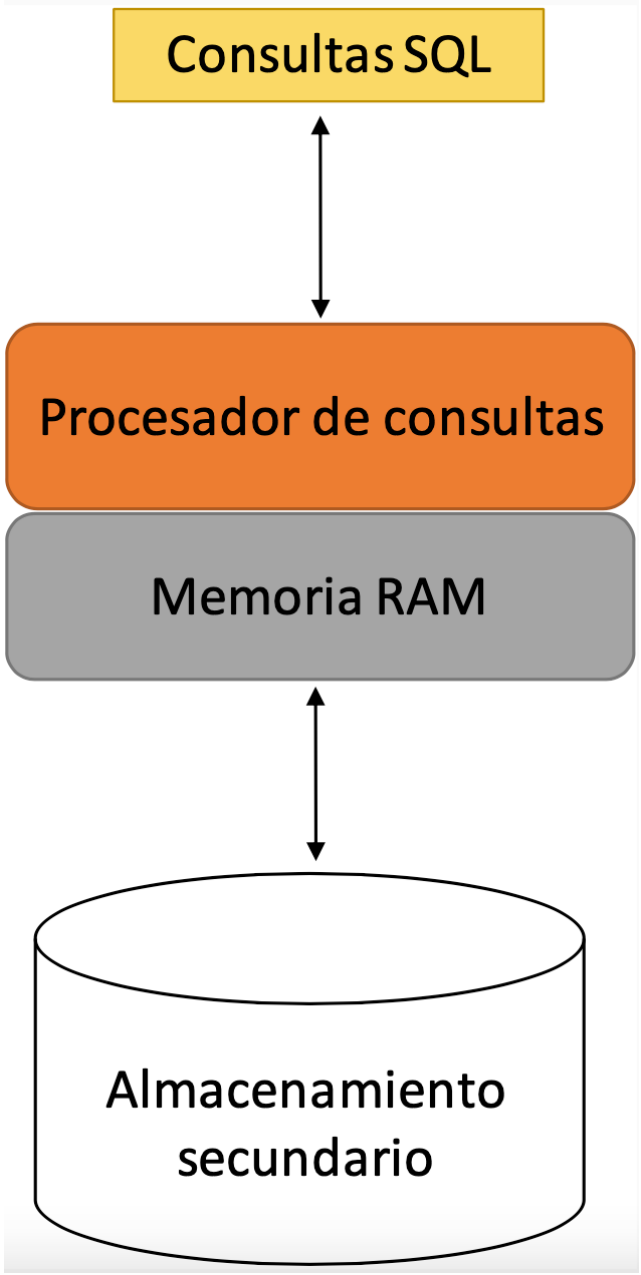
- Cualquier número de flujos puede ingresar al DSMS.
- Los flujos que se reciben no necesariamente deben tener la misma tasa de datos o tipo de datos.

¿Cómo se procesan los flujos de datos?

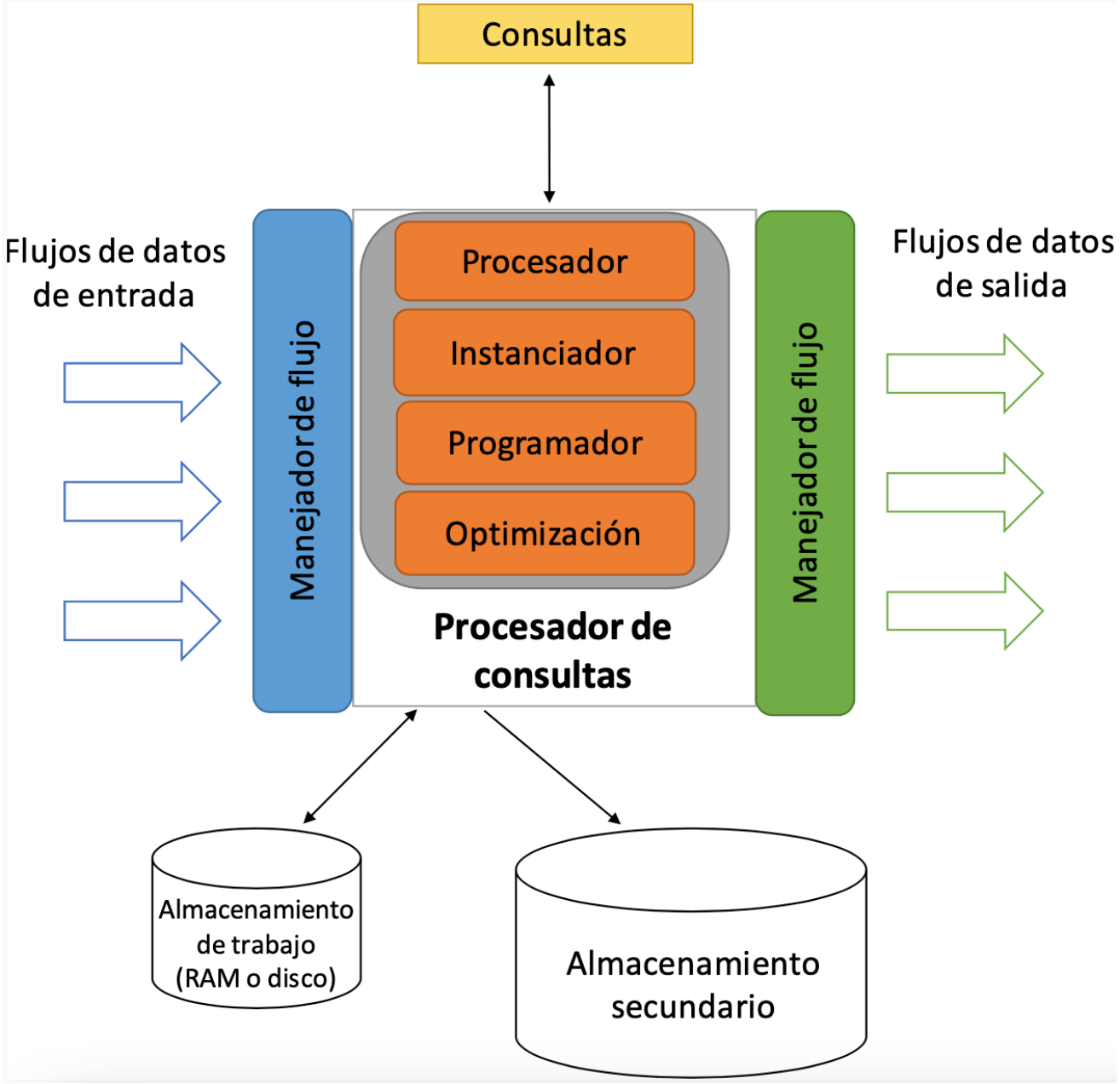
Un procesador de flujos de datos es un tipo de Sistema de Administración de Datos (DSMS).

- El tiempo entre flujos no necesita ser uniforme.
- Los algoritmos para procesar los flujos pueden involucrar resumen, filtrado o uso de ventanas.

Modelo general de un DBMS



Modelo general del procesamiento de flujos de datos



Comparación DBMS y DSMS

DBMS	DSMS
Almacenamiento persistente	Almacenamiento transitorio
Acceso aletorio	Acceso secuencial
Baja tasa de actualización	Tasas de múltiples Gbs
Servicios no de tiempo real	Servicios de tiempo real
Almacenamiento en disco ilimitada*	Memoria principal limitada

Consultas sobre flujos de datos: Características

Las consultas son frecuentes

- Los flujos son evaluados a medida que se van recibiendo.
- Actualizaciones constantes.

Consultas sobre flujos de datos: Características

Las consultas son complejas

- Pre-procesamiento de atributos y extracción de datos crudos

Consultas sobre flujos de datos: Características

Existen dos formas generales para hacer consultas sobre los flujos de datos:

- Consultas permanentes: están almacenadas dentro del procesador, son ejecutadas permanentemente y producen salidas en momentos apropiados.

Consultas sobre flujos de datos: Características

Existen dos formas generales para hacer consultas sobre los flujos de datos:

- Consultas Ad-hoc: se realiza una sola vez sobre el flujo o flujos actuales.

Consultas permanentes

- Supongamos un sensor de temperatura en el océano, la consulta permanente sería “si la temperatura excede los 25 grados, emite una alerta”.

- Esta consulta solo depende del último flujo recibido.

Consultas permanentes

- Otro ejemplo de consulta permanente sería: cada vez que llegue una nueva lectura (temp) genera el promedio de las últimas 24 lecturas.

- Aquí almacenados las últimas 24 lecturas, cuando un nuevo valor llega se hace el cálculo y se borra la primera lectura.

Consultas permanentes

Otro ejemplo de consulta permanente sería: obtén la temperatura máxima:

- ¿Cómo lo haríamos?
- Y si la consulta es obtener el promedio, ¿cómo lo haríamos?

Consultas Ad-hoc

- Son consultas hechas una sola vez sobre los flujos actuales.
- Un enfoque común es almacenar una *ventana deslizante* de cada flujo en el *working storage*.

Ventanas deslizantes

Técnica para el procesamiento de flujos de datos el cual divide dicho flujo en grupos de datos basándose en 2 parámetros:

- Longitud de la ventana (*window length*): indica el tiempo que se tendrá en cuenta para el cálculo (desde t_{actual} hasta t_{actual} - longitud de ventana).

Ventanas deslizantes

Técnica para el procesamiento de flujos de datos el cual divide dicho flujo en grupos de datos basándose en 2 parámetros:

- Intervalo (*sliding interval*): cada cuánto tiempo se vuelve hacer los cálculos sobre los datos de la ventana.

Ventanas deslizantes: Ejemplo

Ejemplo: Actualizar cada segundo (*intervalo*) con el valor de la mayor compra de los últimos 2 segundos (*longitud de la ventana*).

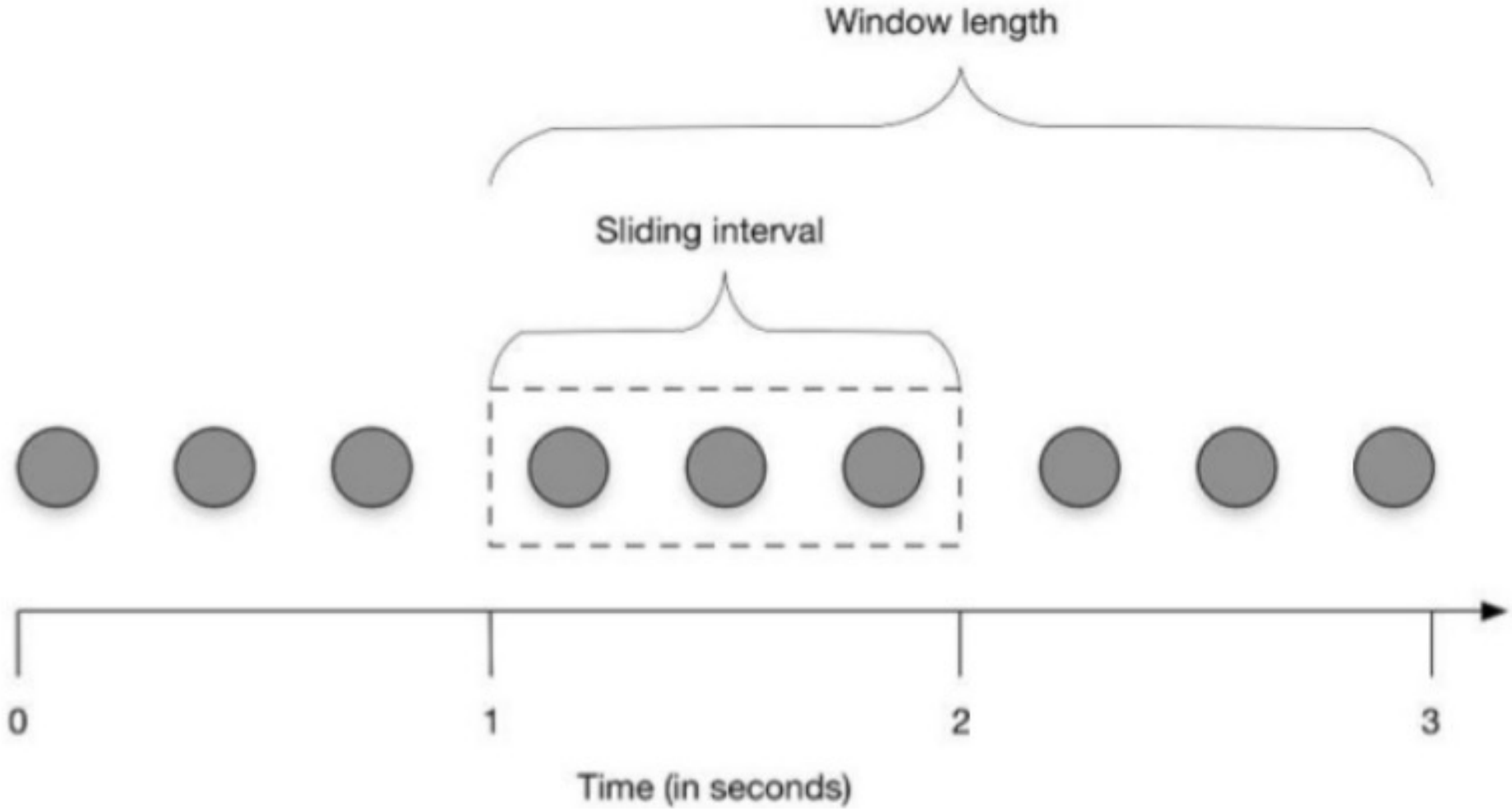
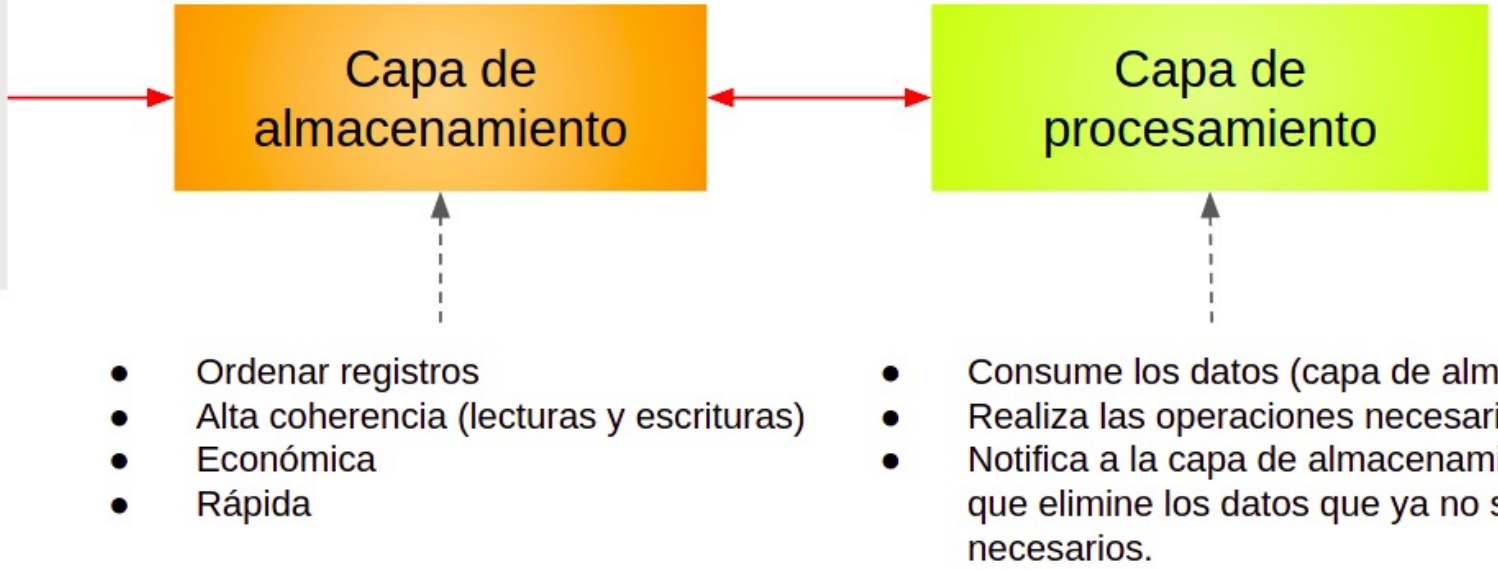
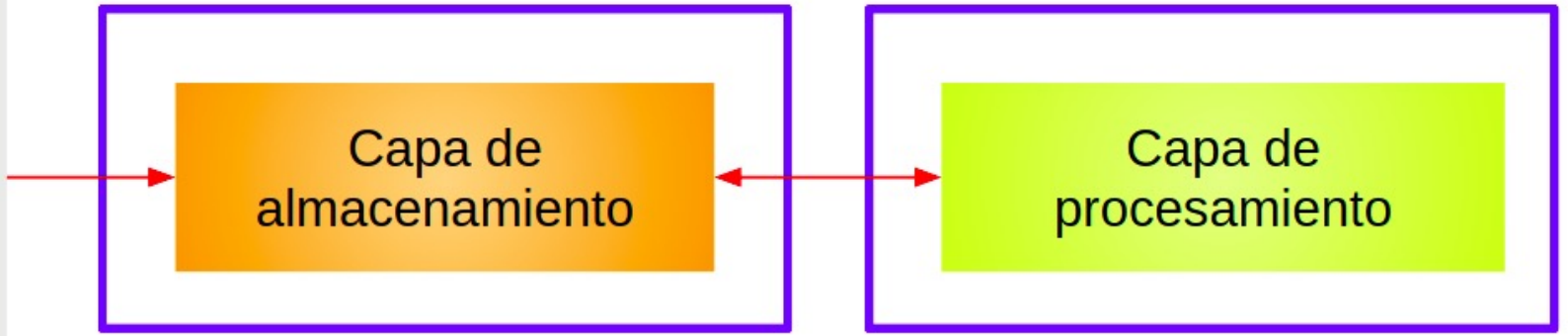


Imagen tomada de Workshop Apache Flink,2016

Capas en el procesamiento del flujo de datos



Capas en el procesamiento del flujo de datos



Una capa adicional se añade para cada una de las capas:

- Planificar la escalabilidad
- Durabilidad de los datos
- Tolerancia a fallos

Aprendizaje en línea

En Aprendizaje máquina ha surgido el aprendizaje en línea:

- Nos permite modelar problemas en donde la entrada son flujos continuos de datos.
- Se busca encontrar un algoritmo que aprenda a partir de los datos y que pueda adaptarse a pequeños cambios.

Aprendizaje en línea

En Aprendizaje máquina ha surgido el aprendizaje en línea:

- Ejemplos: Descenso del gradiente estocástico (SGD) permite pequeñas actualizaciones.

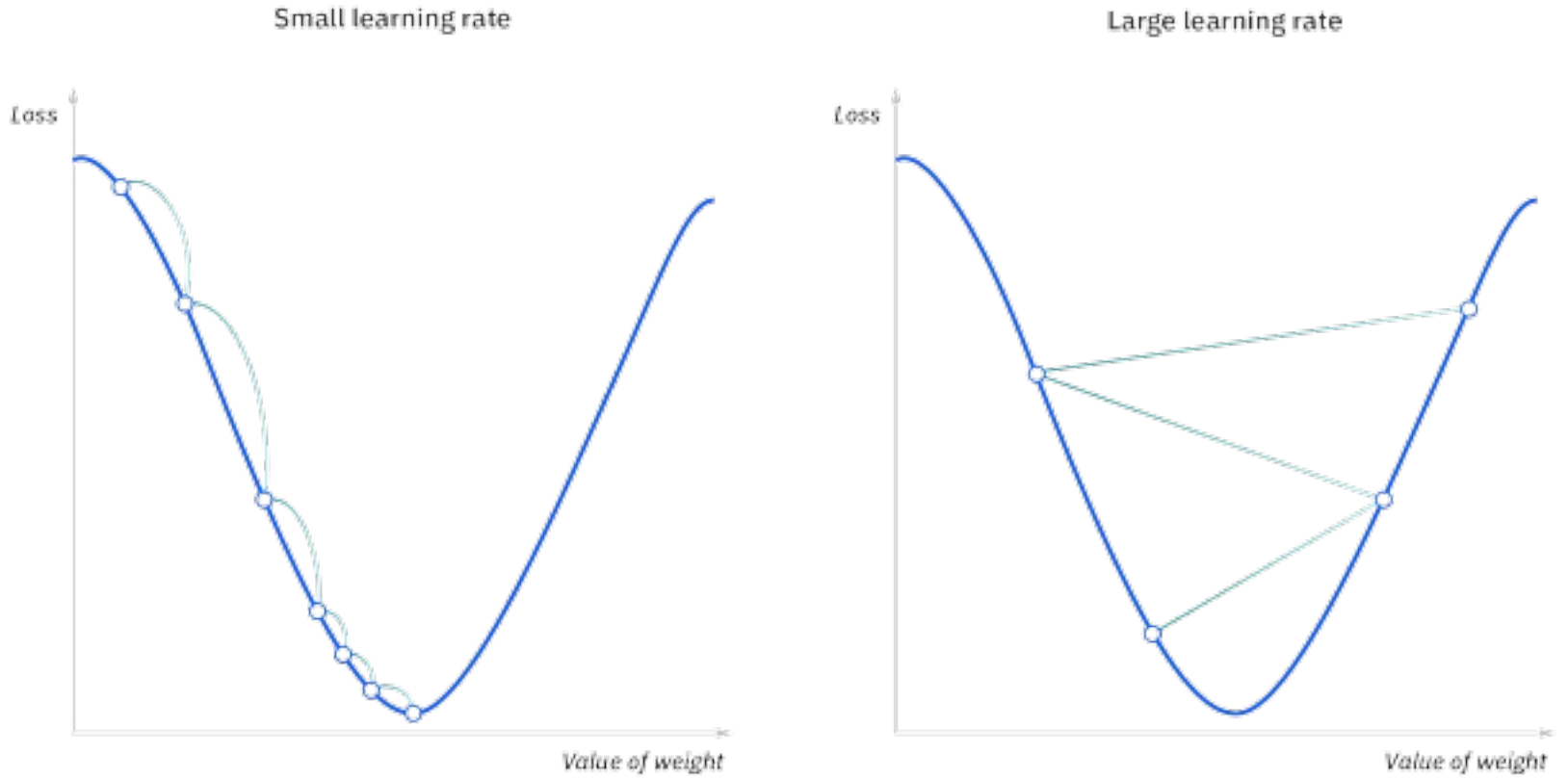
Aprendizaje en línea

En Aprendizaje máquina ha surgido el aprendizaje en línea:
Ejemplos: Descenso del gradiente estocástico (SGD) permite pequeñas actualizaciones.

- Choose an initial vector of parameters w and learning rate η.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle samples in the training set.
  - For $i = 1, 2, \ldots, n$, do:
    - $w := w - \eta \nabla Q_i(w)$.

Aprendizaje en línea

En Aprendizaje máquina ha surgido el aprendizaje en línea:



Plataformas *Open-source*

Actualmente existen numerosas plataformas que soportan el procesamiento de flujos de datos.

- Amazon Kinesis Streams
- Amazon Kinesis Firehose
- Apache Kafka
- Apache Flume
- Apache Spark Streaming
- Apache Storm

Aplicaciones sencillas

- Implementación de mínimo - máximo
- Generación de informes básicos
- Emitir alertas

Aplicaciones complejas

- Uso de aprendizaje máquina
- Procesamiento de eventos y transmisiones