SAS FORUM ARGENTINA 2015

5 DE MAYO

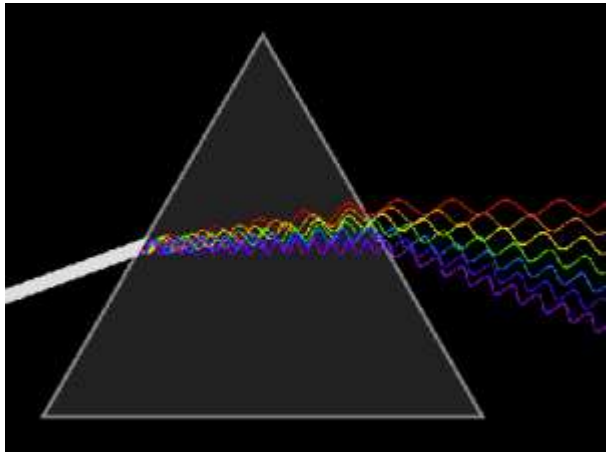
DATA STREAM MINING

Martin Volpacchio ** *
Co-Founder Pi Analytical Solvers
** FI MDM Universidad Austral
** MS Math Economics
* Data Scientist, Mg Sc.

Sergio Uassouf
Líder de Práctica de Gestión de Información e Infraestructura

TEMARIO

0. El tsunami de datos
1. Data Stream Model
2. Modelos Básicos
   - Sampling
   - Filtering
   - Counting
3. Modelos Complejos
4. SAS y Stream Models

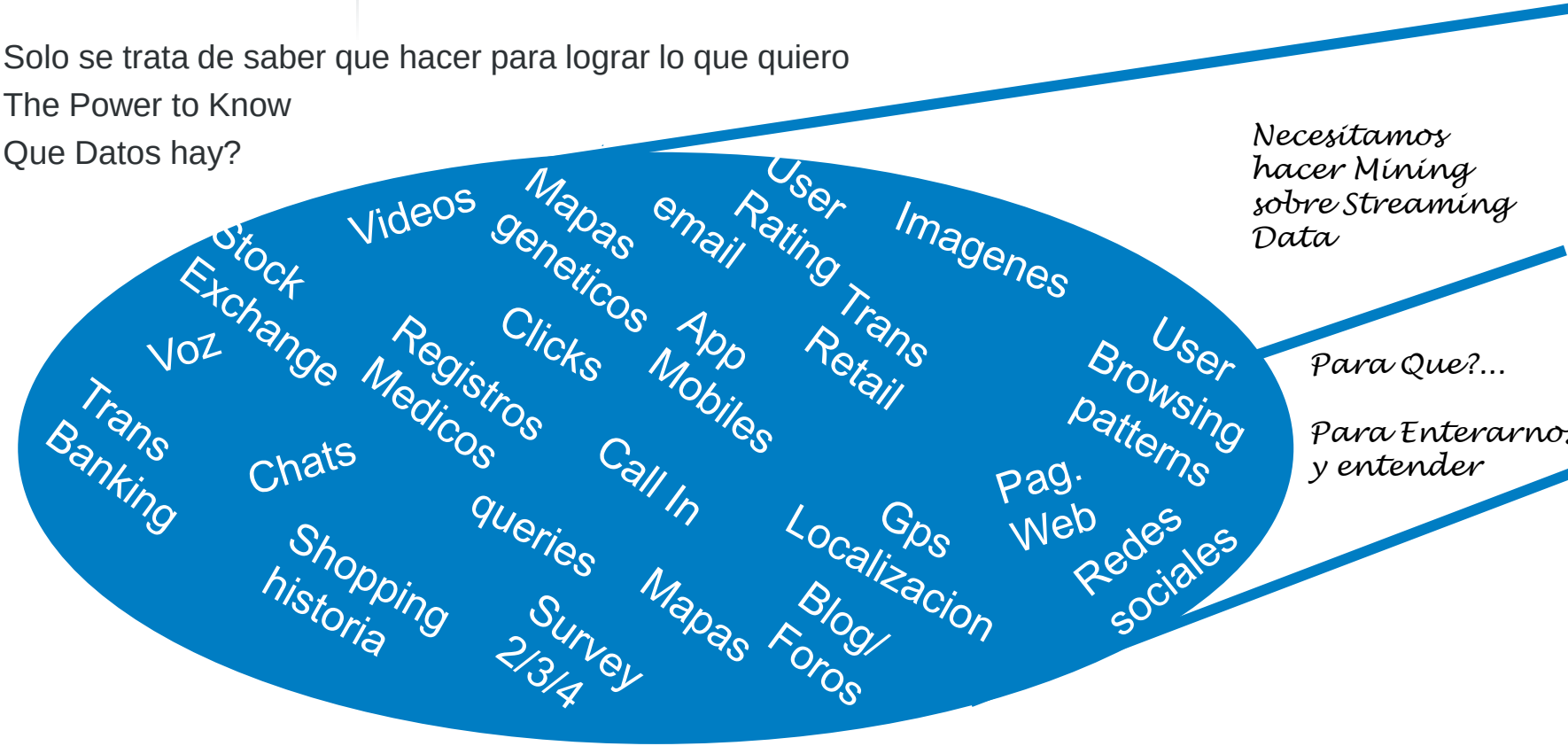


0. EL TSUNAMI DE DATOS

…Y OPORTUNIDADES DE NUEVOS NEGOCIOS

TSUNAMI | NUEVAS OPORTUNIDADES DE ENTERARNOS

- Solo se trata de saber que hacer para lograr lo que quiero
- The Power to Know
- Que Datos hay?



1. DATA STREAM MODEL

SE CORRE LA FRONTERA DEL MINING

DATA STREAM UNA TENDENCIA ACTUAL

- La alta tasa de arribo es relative al poder computacional de la maquina
- Tipicos, Trafico de Red, Extracciones en ATM, web serachs, datos de sensors, bid requests, etc
- Data Stream Mining:

"Extraer conocimiento a traves de modelos y patrones en un No Stopping Stream de Información"

- Principales restricciones para hacer query sobre data stream son
  - Requerimiento de Ilimitada Memoria no Controlada
  - Velocidad de llegada de los datos comparada a velocidad de Procesarla

- Generalmente hay dos objetivos:
  - Modelos Basicos
  - Predecir la clase o el valor nuevo de un elemento de un stream basado en ventanas anteriores

DATA STREAM | UNA TENDENCIA ACTUAL

- Lo usual…hacer Batch Data Mining sobre una base de datos. Esto supone disposicion off-Line para manejarla cuando queramos, Al correr la frontera de lo que deseamos resolver…esto ya no es asi,

- Lo Nuevo…los datos arrivan en flujo/os en Real Time de forma NO controlada , si no los procesamos inmediatamente, perdemos la oportunidad de Ganar o de No perder.

- Supondremos,

  - El flujo es inmenso y debe ser rapidamente procesado para tomar decisiones en RT o NRT y almacenado filtrando el ruido,
  - El flujo llega a altisima velocidad que no puede ser almacenado en una base de datos convencional e interactuar normalmente
  - El tiempo de respuesta de la accion puede involucrar Milisegundos

THINGS USEFUL TO REMENBER…

- Re-visited Hash Function
  - $Hash - Function$: $Hash - key \rightarrow Bucket\ Number$
  - $(numero\ entero\ en\ el\ rango\ 0\ a\ b, donde\ b\ es\ el\ numero\ de\ bucket)$
  - Ej: $Hash - Function$ (x) = x Mod B (OK si positivos enteros)
  - Factibilidad de dominio no numerico…String o structuras mas complejas

- Storage Secundario
  - Windows divide el disco en block de 64 Kb y tarde 10 milisegundos en mover los datos al Ram,
  - Como minimo, 5 orden de magnitud (10^5) en tiempo de computacion si los datos estan en Disco con repecto a Main Memory (Mover el head del disco al trach del block y esperar que el block rote sobre la cabeza)
  - Optimizacion de como distribuir los datos en el disco..cilindros de blocks de radio fijo a la cabeza del head. Se mejora sensiblemente los 10 ms.
  - Un disco no puede transferir mas de 100 Mill de bytes por Segundo al Ram, independientemente de como se organice los datos en los blocks,
  - Si el dataset es en Mb..No Problem…pero si son cientos de Gb o Tb….Tenemos un problema

DATA STREAM | DSMS vs DBMS

- Relaciones Permanentes
- Queries Discretos
- Acceso aleatorio
- Storage Planificable
- Research Off-Line
- Baja Tasa de Actualizxacion
- Cualquier granularidad
- Data Consistida

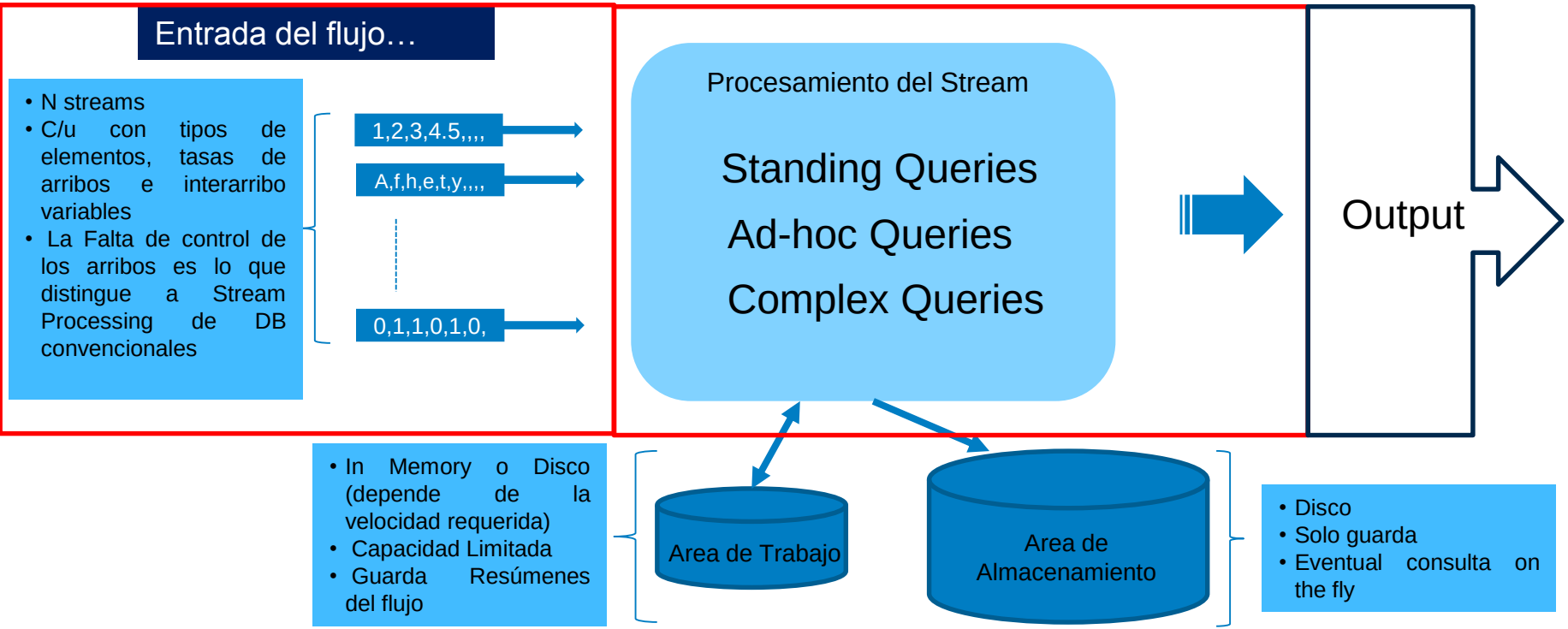
DBMS

- Relaciones transitorias
- Queries Continuos
- Acceso secuencial
- Ram Acotado
- Research On the Fly RT o NRT
- No controlada Tasa de arribo
- Granularidad muy fina
- Data No Consistida

DSMS

DATA STREAM MANAGEMENT SYSTEM

PROCESANDO STREAM…



PREVIAMENTE…EJEMPLOS…

Sensores en el Océano para medir Temperatura y altura

1) Solo midamos Temperatura. Baja variación. No desafiante
2) Agreguemos medición de altura con una unidad GPS. Alta variación. Si envía un numero real de 4 bytes cada 1/10 seg= 3,5 MB/dia
3) Instalemos 1 cada 250 km cuadrados de Oceano, 1 Millón de sensores=3,5 TB/dia

MAS EJEMPLOS…

Imágenes y video

- Satélites, cámaras de seguridad. Autenticación
- Solo Londres se cree que posee 6 Millones, Cada una genera un stream

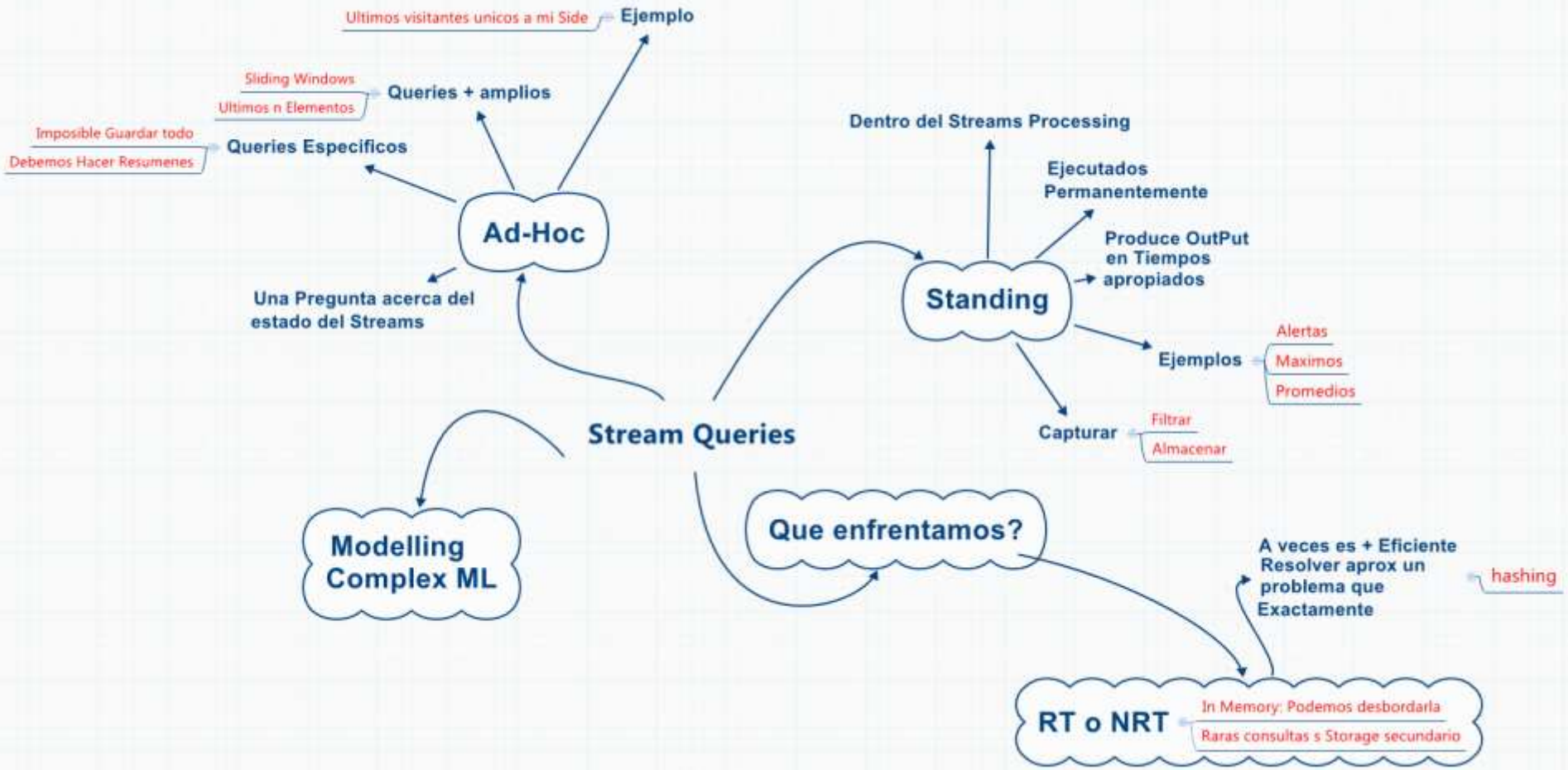
Trafico de Internet

- Google/Yahoo/Telco/Redes Sociales
- Ad-Exchange, DSP, Bit-Request…

Banca/Retail/Bolsa/Gobierno/….

- Miles de extracciones/transferencias/pagos/compras
- Localización

STREAM QUERIES…



2. BASICS MODEL

LOS DE MAYOR IMPACTO

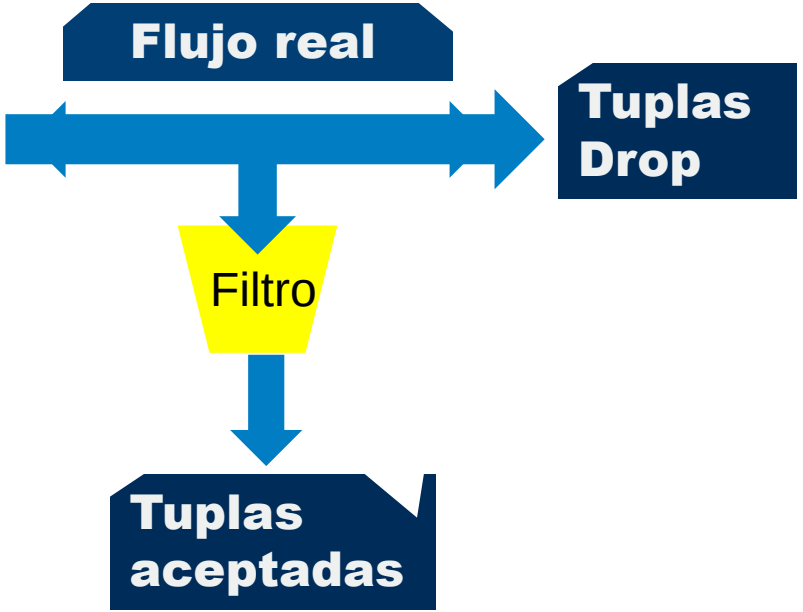
SAMPLING DATA…

- Objetivo: Obtener muestras de flujo que Representan al proceso

  - Si quiero aplicar Queries de cierta complejidad en RT
  - Ej: Que fracción de los query son repetidos en la ultima hora
  - Supongamos que se desea retener 1/10 elementos del flujo
  - Usamos Memoria y funciones Hash
  - Si deseamos a/b user, debemos “Hashear” a los elementos en b buckets
  - Stream= Tupla de n elementos,
  - Un Sub-set de la Tupla es la Key, en la que se basa el sampling

FILTERING DATA…

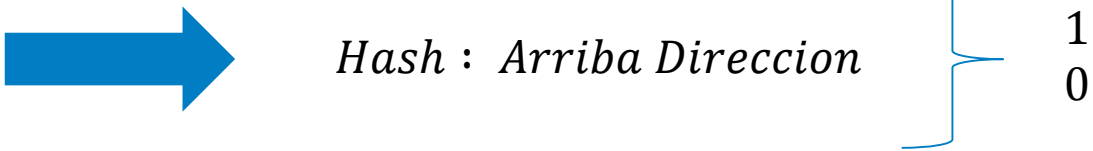
- Objetivo: Obtener muestras de flujo que Cumplen cierta condición
  - 2 casos:
    - Simple: Un elemento cumple una condición
    - Complejo: si requiere un query sobre el flujo Y el set es demasiado grande para estar in memory…Bloom Filtering



Ejemplo: Se quiere filtrar 100 millones de direcciones de e-mail (No Spam)
Tupla=(Direccion, contenido)
Direccion aprx 20 Bytres → No razonable guardar todo in memory
2 Opciones: Disco o usamos método que no requiera mas memoria que disponibles

BLOOM FILTERING…

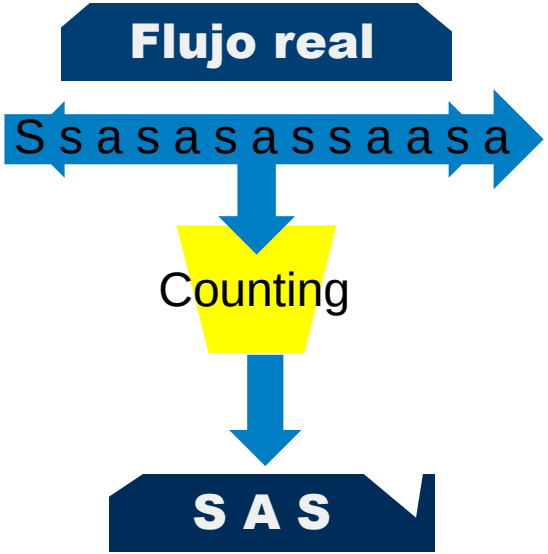
- S= 1000 millones de direcciones de email
- Simplifiquemos: 1 Gb de Ram implica un vector de 8000 millones de bit de largo,
- Desarrollemos una colección de función Hash Hi para 8000 millones de Buckets,
- Cada elemento de S Hash a un Bit, q encendemos a 1, el resto 0
- 1/8 bit = 1 (en realidad algo menor)



DATA STREAM MANAGEMENT SYSTEM

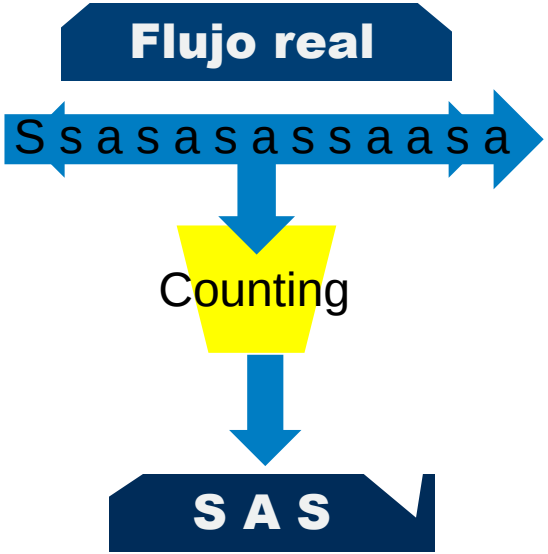
COUNTING DATA…

- Objetivo: Count Distinct Problem. Cuantos elementos diferentes en un lapso de tiempo…Ej: visitantes distintos en un web site.
  - 2 casos:
    - C/Log in: Amazon
    - S/Log In: Uso de IP Address. Conjunto Universal: 4,000 Mill, secuencias de 4 bytes de 8-bit
    - SN Obvia Si Size ajusta a memoria : Mantener en Main Memory todos los elementos vistos hasta el momento en el stream Hash Table o árbol de búsqueda: Rapida agregación de nuevos y detectar pre-existencia



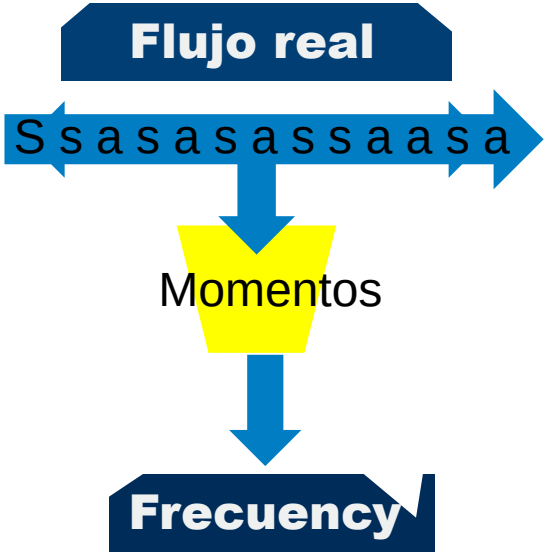
COUNTING DATA…

- SN NO Obvia Si Size NO ajusta a memoria
  - Varias Maquinas
  - Uso secuencial de Memoria secundaria
  - Usar Menos Memoria que números de elementos distintos: Algoritmo Flajolet-Martin. Usa muchas Hash Function y análisis de Probabilidades
    - No es necesario Almacenar los elementos
    - Solo almacenamos un entero por F Hash



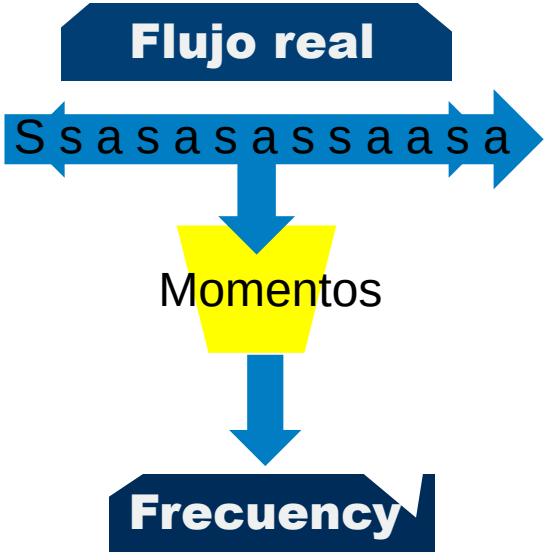
GENERALIZACION DE COUNTING DATA…ESTIMAR MOMENTOS DE UN STREAM

- Computar momentos requiere estimar densidades del flujo
- Nuevamente,
  - Trivial si hay exceso de Main Memory
  - Caso No trivial. Para el segundo momento, Algoritmo de Alon-Matias-Szegedy
  - Cuando los flujos involucran Big Data element…se deben crear métodos diferentes a los de la Estadistica Clásica,,,



DERIVACIONES….

- Contar en forma exacta e l numero de elementos en una ventana del tiempo del flujo, permanentemente
- Los elementos mas comunes o populares (Redes sociales o comprar de películas)
- Tasas de variaciones de flujo (Acciones, Rate Tweet/Twitter-Users, bit request en DSP, etc)

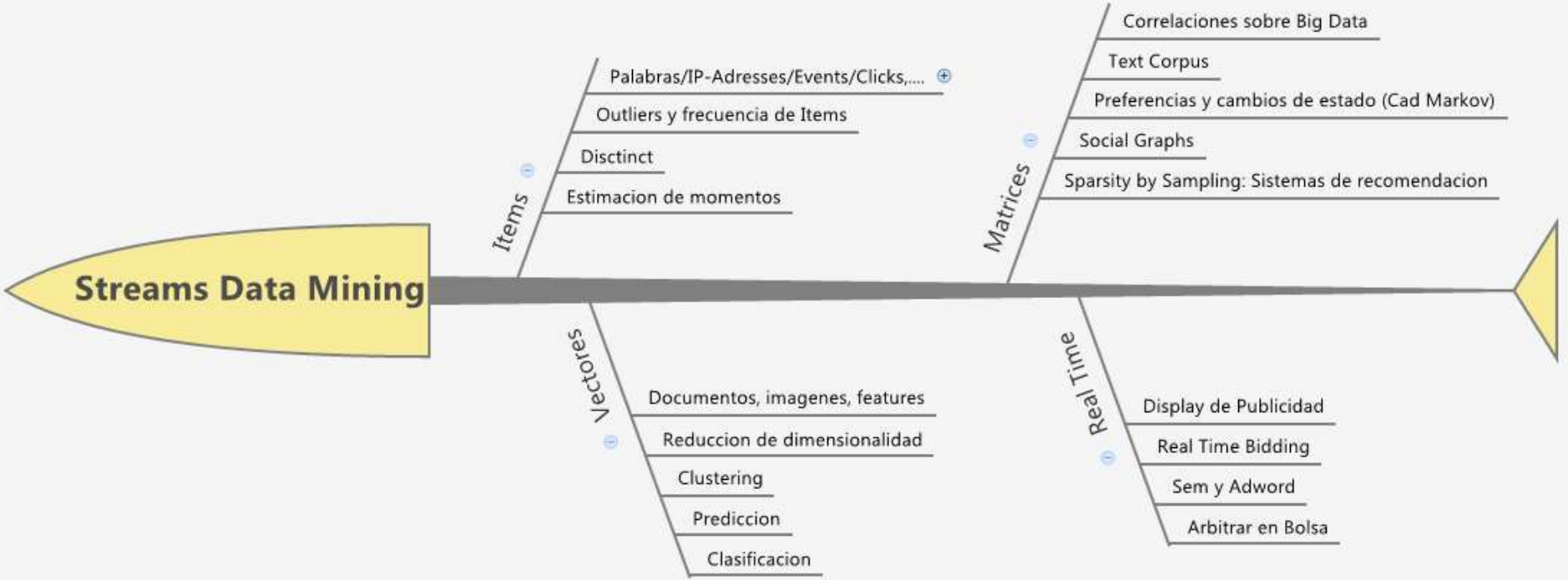


3. MODELOS COMPLEJOS

QUÉ DESEAMOS RESOLVER?

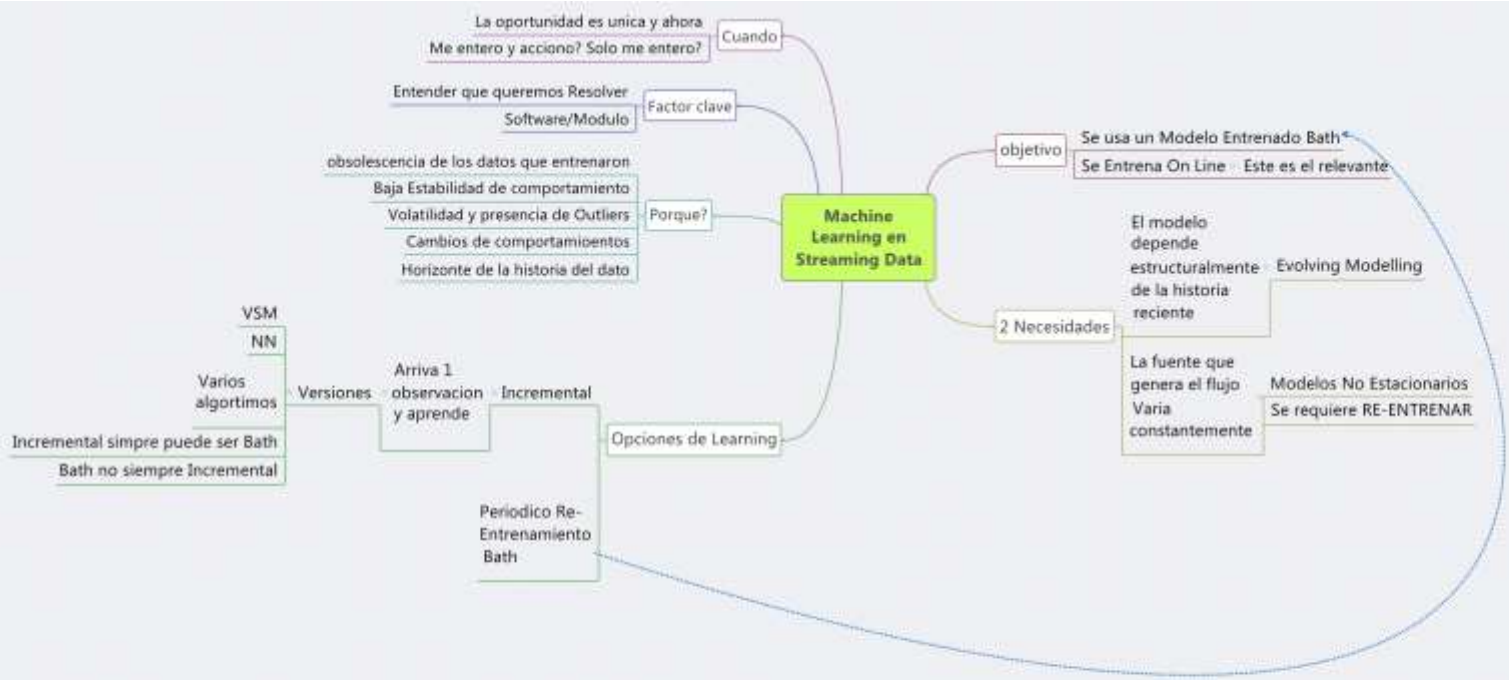
DATA STREAM MANAGEMENT SYSTEM

MODELOS MAS COMPLEJOS...



DATA STREAM MANAGEMENT SYSTEM

MODELOS MAS COMPLEJOS...



4. SERGIO ACA VAS VOS

A. ENMARCAR EL PROBLEMA

DATA STREAM MANAGEMENT SYSTEM

MODELOS MAS COMPLEJOS...

$$\sum_{k=1}^{Ustedes} Muchas\ Gracias!!_{\binom{Martin}{Sergio},k}$$

Mev.volpacchio@gmail.com

sergio.uassouf@sas.com