

DATA STREAMING Y SERVICIOS EN LA NUBE

INTRODUCCIÓN

Magister - Efraín Alberto Oviedo
alberto.oviedo@udea.edu.co

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA

ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS



**UNIVERSIDAD
DE ANTIOQUIA**

1 8 0 3

AGENDA

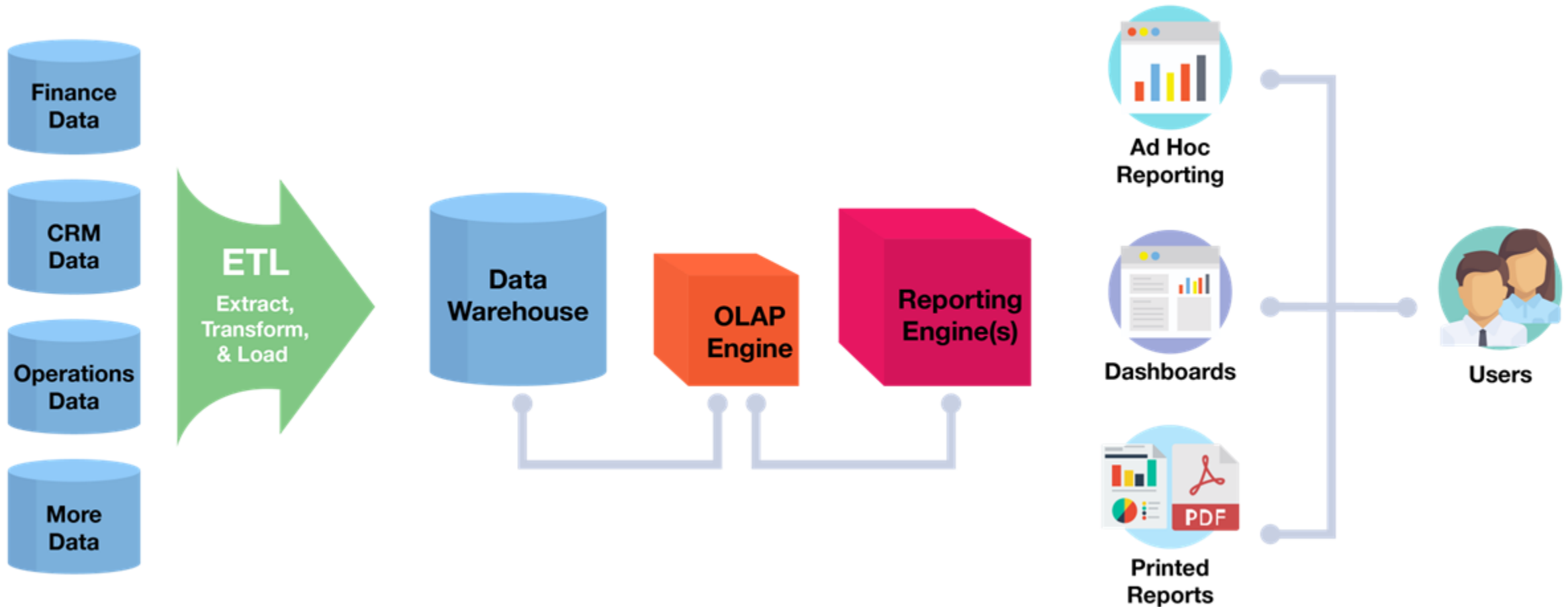
1. Contexto

- 2. Procesamiento de Datos
- 3. Cloud Computing
- 4. Aplicaciones

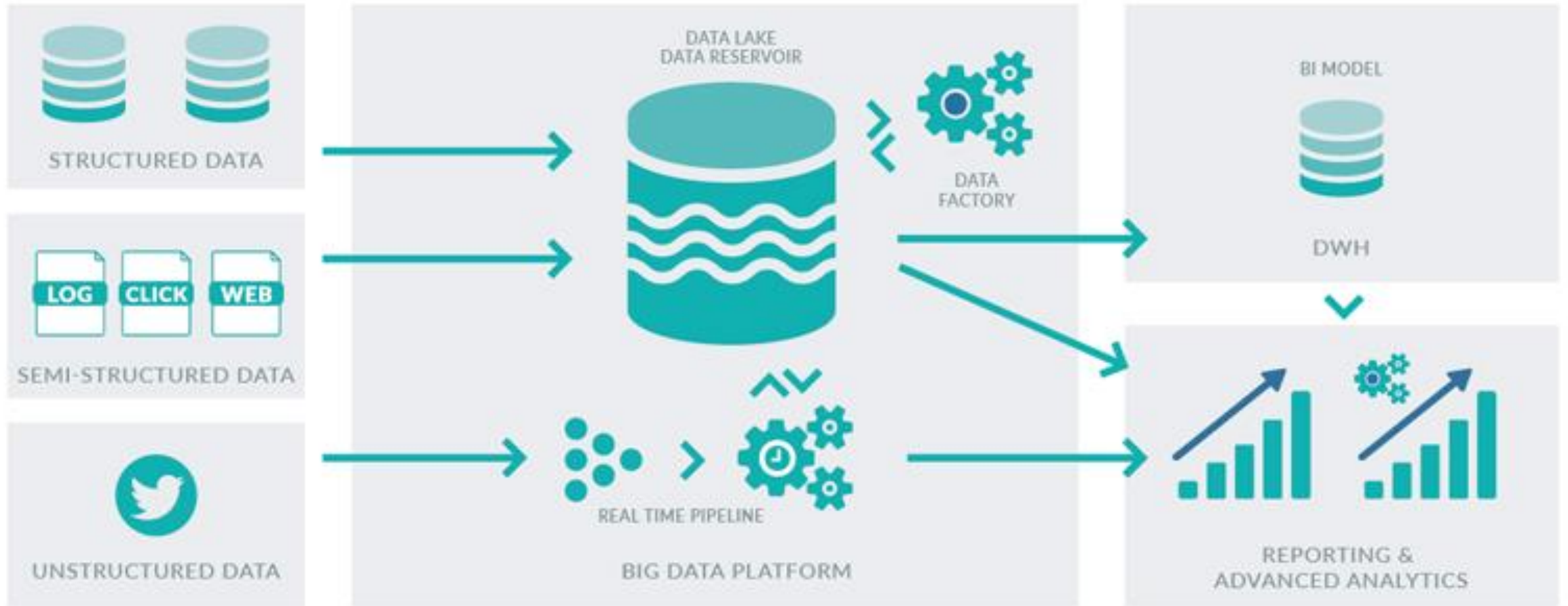
Inteligencia de Negocios



Inteligencia de Negocios Tradicional



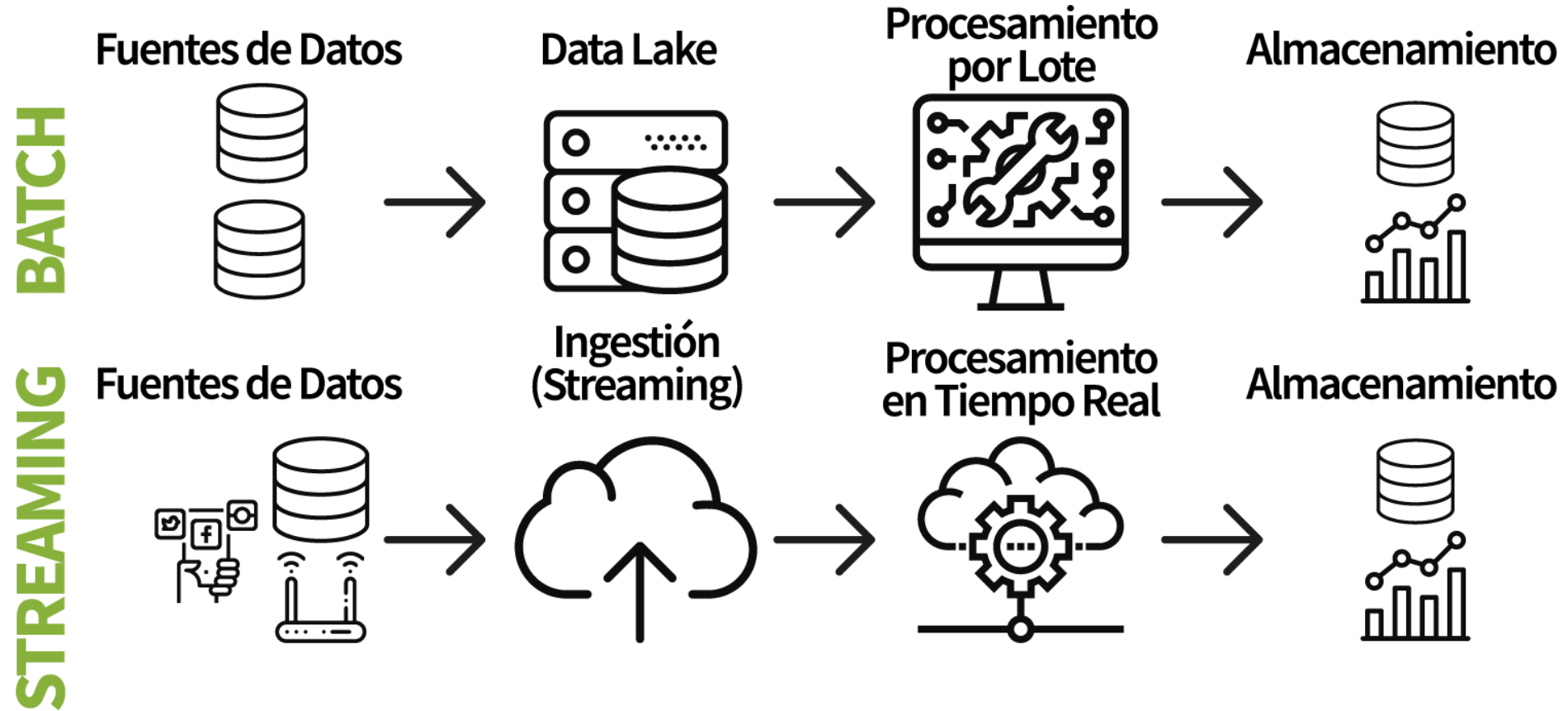
Inteligencia de Negocios con Big Data



AGENDA

1. Contexto
- 2. Procesamiento de Datos**
3. Cloud Computing
4. Aplicaciones

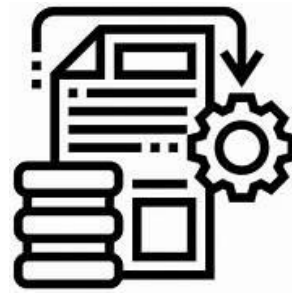
Tipos de procesamiento



Procesamiento en Batch



- Un lote (batch) es una colección de datos que ha sido agrupada durante un **intervalo de tiempo**
- Realiza periódicamente trabajos repetitivos de grandes volúmenes de datos (**millones de registros**)
- Requiere un **gran esfuerzo computacional**
- Se suelen ejecutar en **horas de menor actividad**
- Las tareas se pueden ejecutar de forma secuencial o **simultánea**



Procesamiento en Batch

- Requieren una **interacción humana mínima** ejecutando de manera eficiente tareas repetitivas
- Se recomienda utilizarlos en **proceso que tomen mucho tiempo en ser realizados** sea por el gran volumen de los datos o por la complejidad misma del proceso

Lote Día
22/2/2020



24 horas
después

Lote Día
23/2/2020



24 horas
después

Lote Día
24/2/2020

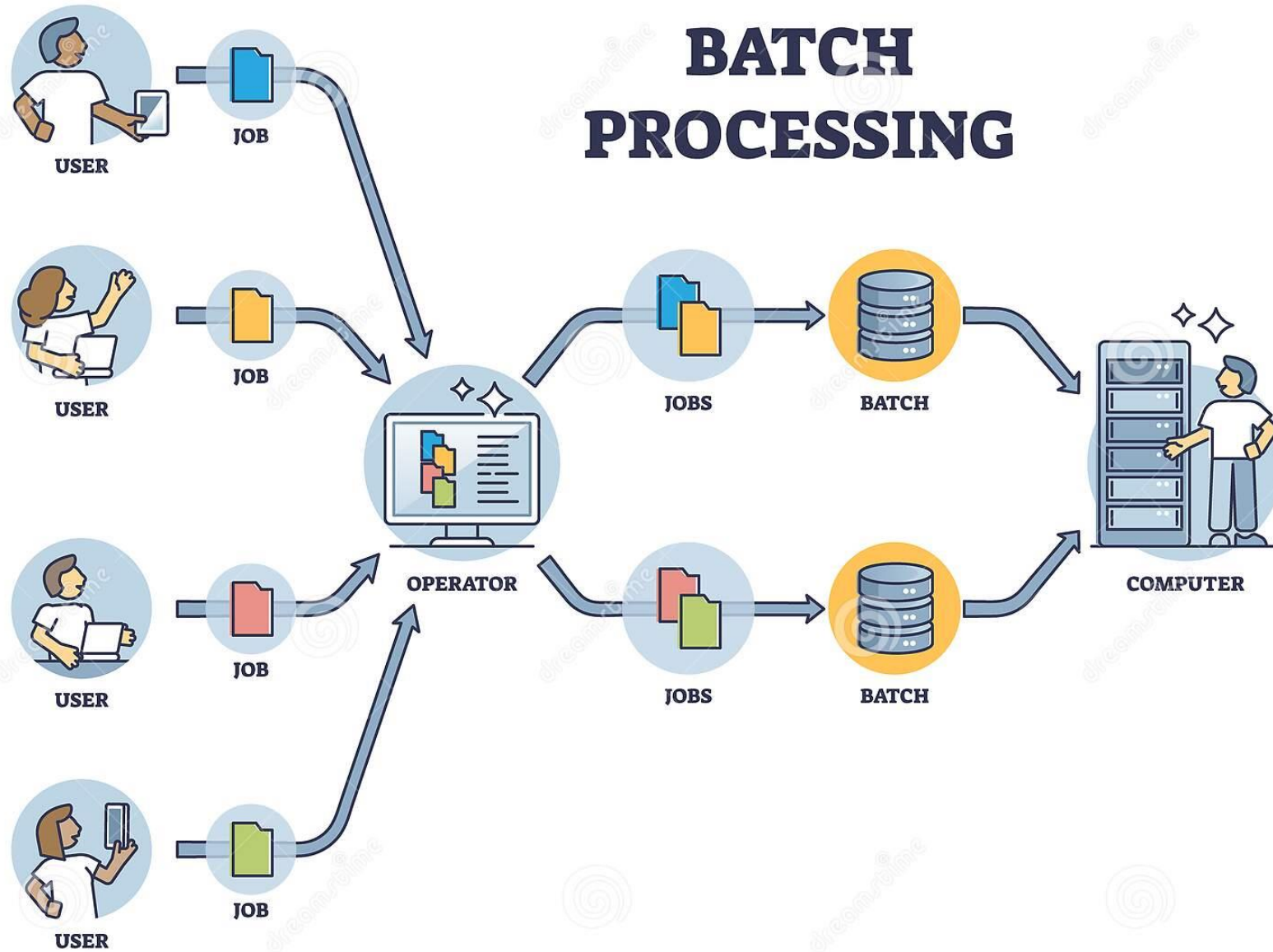


24 horas
después

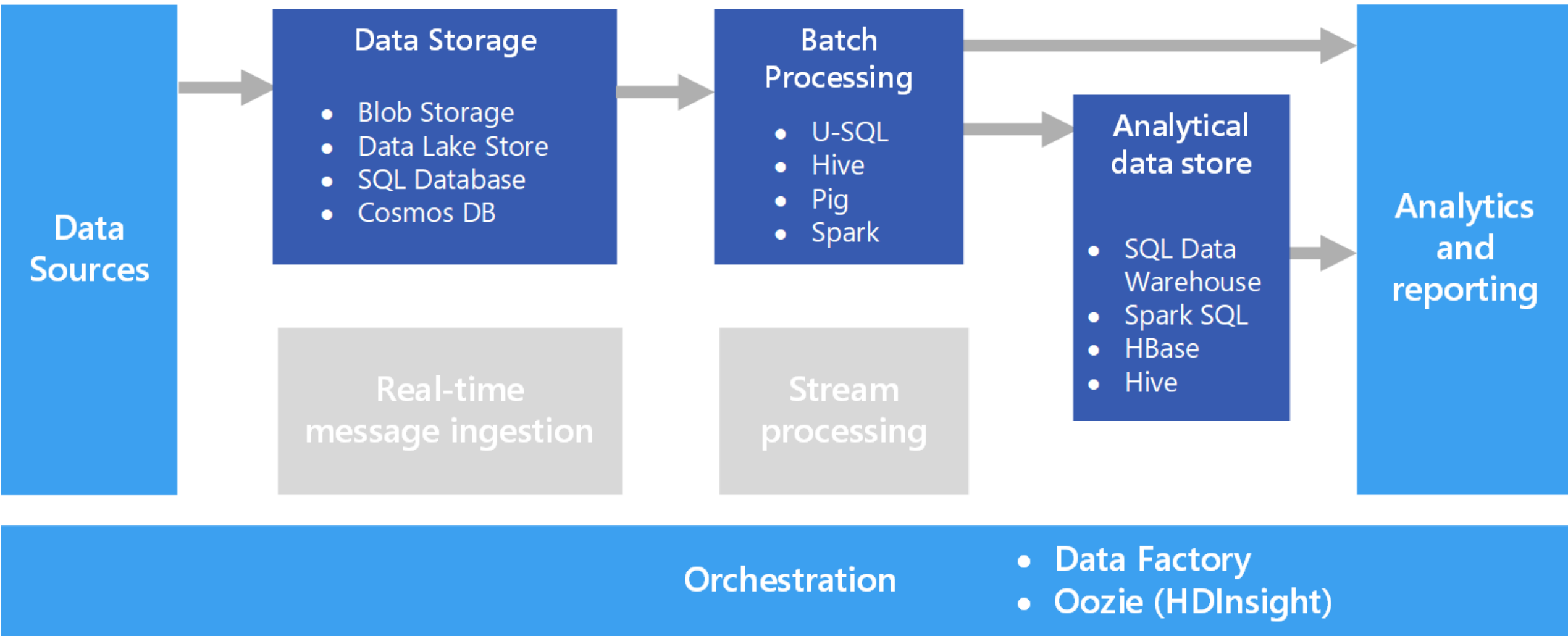
Lote Día
25/2/2020



BATCH PROCESSING



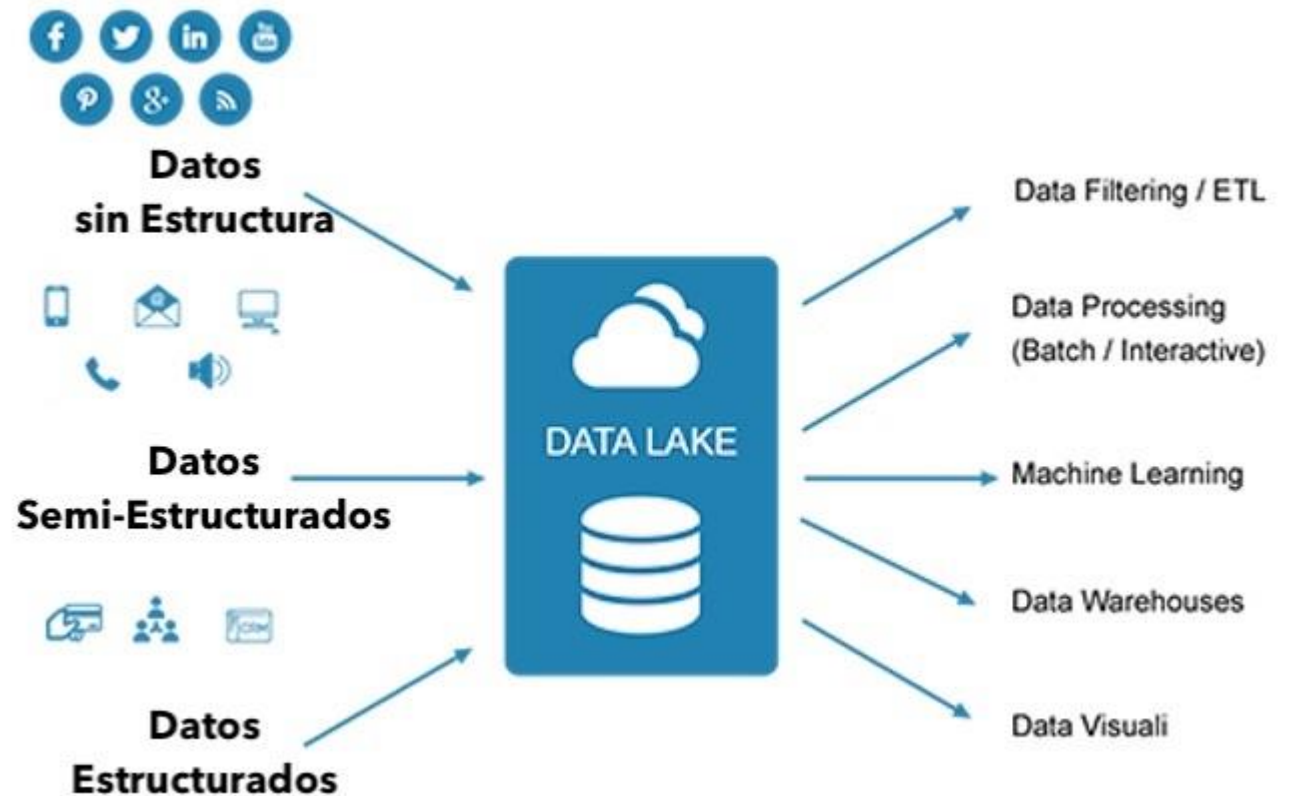
Procesamiento en Batch



Data Lake

Data Lake (Lago de datos) consiste en un repositorio centralizado donde es posible almacenar grandes cantidades de **datos estructurados y no estructurados**

Permite **almacenar los datos tal y como están, y procesarlos posteriormente bajo demanda**, es decir que no es necesario almacenarlos con una estructura de datos definida



Data Lake

Características	Almacenamiento de datos	Lago de datos
Datos	Relacionales provenientes de sistemas transaccionales, bases de datos operativas y aplicaciones de línea de negocio	No relacionales y relacionales provenientes de dispositivos de IoT, sitios web, aplicaciones móviles, redes sociales y aplicaciones corporativas
Esquema	Diseñado con anterioridad a la implementación del almacenamiento de datos (esquema en escritura)	Escrito al momento del análisis (esquema en lectura)
Precio/desempeño	Resultados de búsqueda más rápidos con almacenamiento de mayor costo	Resultados de consultas que se tornan más rápidos con almacenamiento de bajo costo
Calidad de los datos	Datos seleccionados cuidadosamente que funcionan como la versión central de la verdad	Cualquier dato seleccionado o no (es decir, datos sin procesar)
Usuarios	Analistas de negocios	Científicos de datos, desarrolladores de datos y analistas de negocios (con datos seleccionados)
Análisis	Generación de informes en lotes, inteligencia empresarial y visualizaciones	<i>Machine learning</i> , análisis predictivo, detección de datos y creación de perfiles

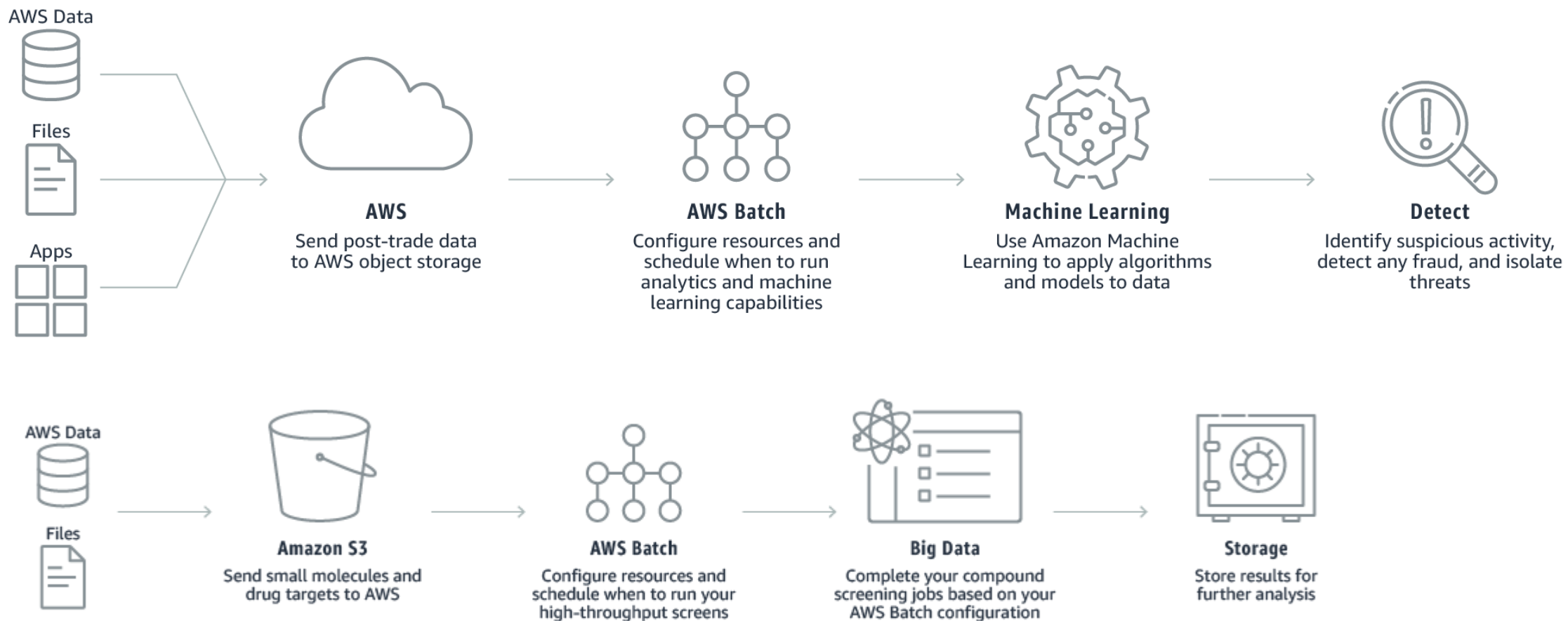
<https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/>

Procesamiento en Batch - Ejemplos

- Gestión de Inventario
- Informes Automatizados
- Facturación como un proceso periódico (semanal o mensual)
- Actualización de modelos de Machine Learning



Procesamiento en Batch - Ejemplos





Procesamiento en Streaming

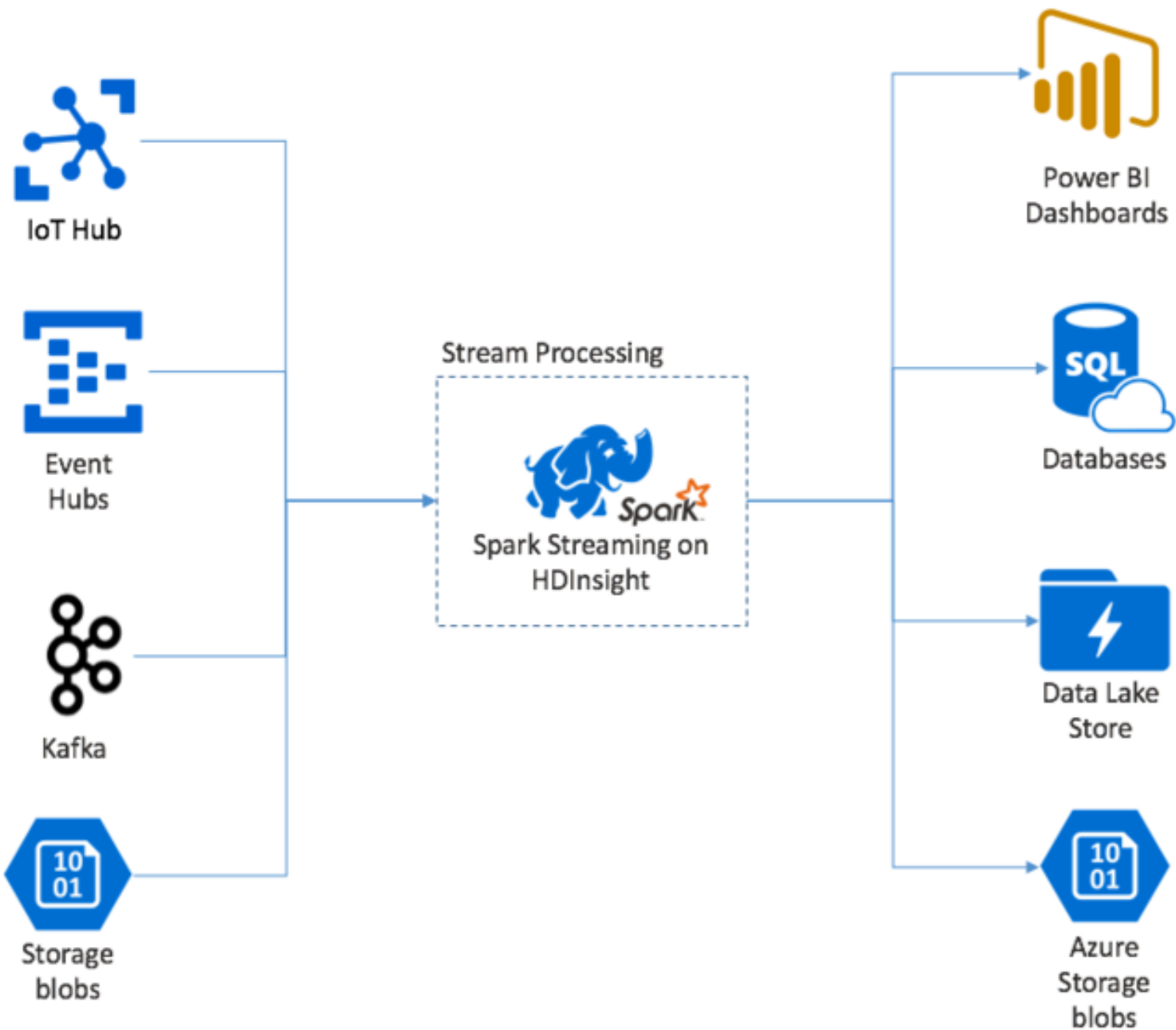
- Los datos se generan y transmiten en tiempo real en **pequeños paquetes (kb)**
- Al recibir los datos, se debe procesar registro por registro de **forma secuencial**
- Se requiere una latencia muy baja del orden de segundos o incluso **milisegundos**
- Se requiere de una capa de **almacenamiento** y otra de **procesamiento**



Procesamiento en Streaming

- Permite **procesar los datos de forma continua en el mismo instante en que llegan** sin esperar ningún otro evento
- Cada lectura de datos se concibe como un **flujo de datos** que pueden representarse como un conjunto de lista de valores o como un conjunto de pares clave-valor





El procesamiento de datos en Streaming usualmente es realizado en **cluster de Spark** que garantizan que cada uno de los eventos sea procesado de manera correcta incluyendo el concepto de la tolerancia a fallas

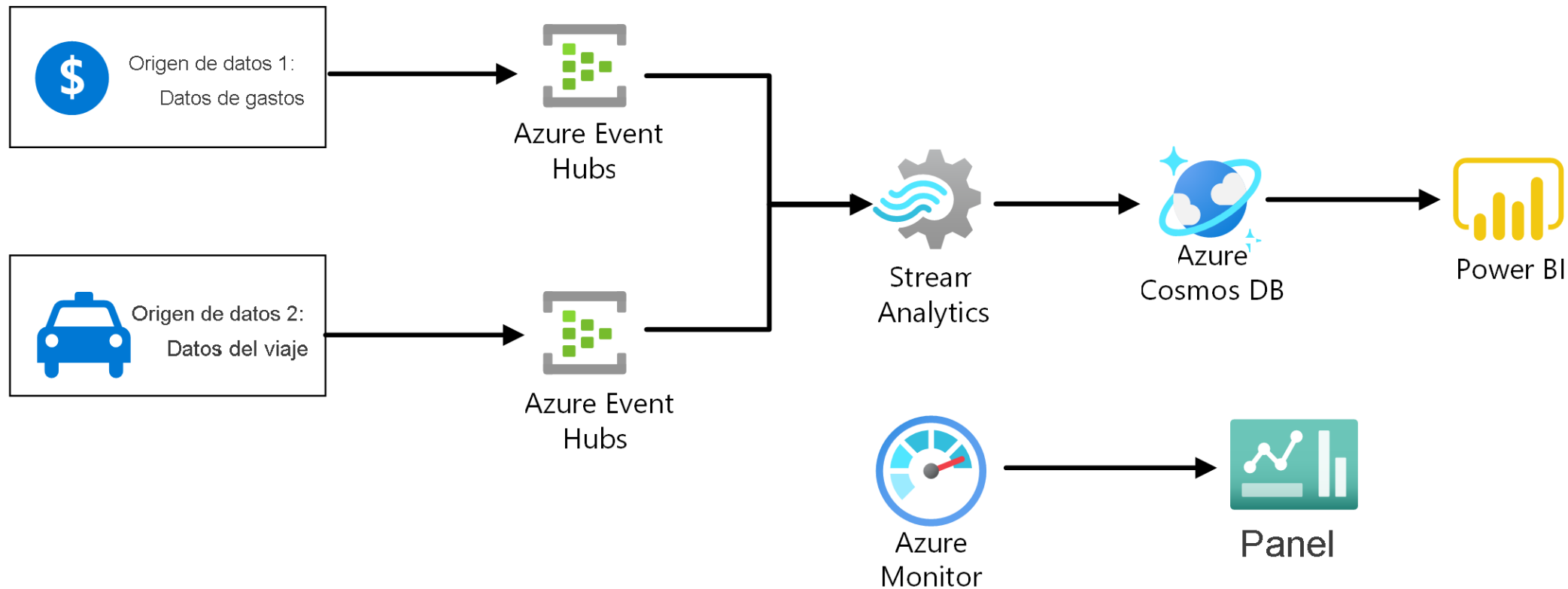
Data Stream

- Data Stream o flujo de datos se refiere a la transmisión de una secuencia de señales codificada digitalmente en paquetes para transmitir información
- Data Stream tiene la necesidad de **procesar y realizar análisis en tiempo real**
- En el Data Stream cuenta con una naturaleza continua de generación de datos a gran velocidad, que implica que los datos sean procesados sin necesidad de ser previamente almacenados

Procesamiento en Streaming - Ejemplos

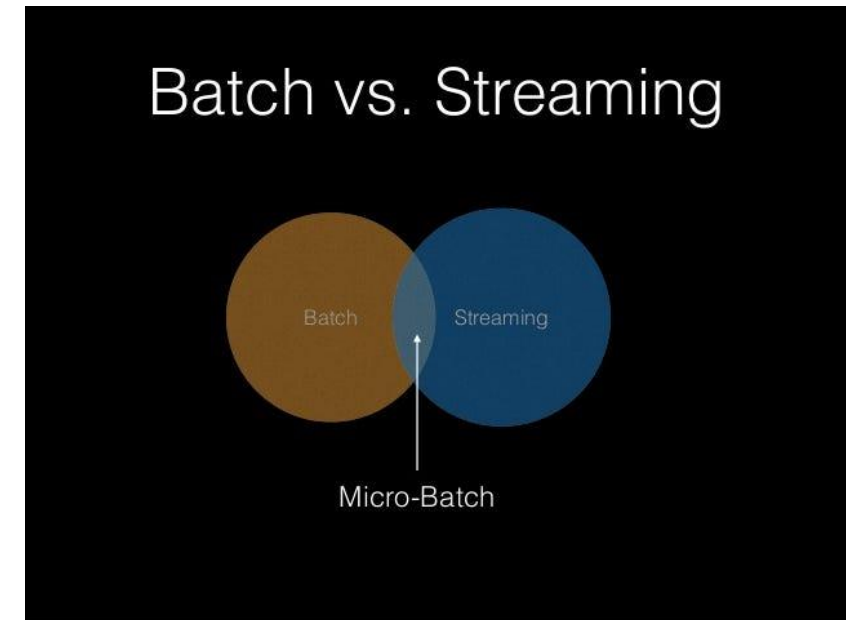
- En **videojuegos**, la interacción del jugador se puede transmitir para ser analizada en tiempo real y **ofrecer experiencias dinámicas**
- Una **página web** puede almacenar los **registros de clics** de cada usuario para aprender sobre su comportamiento y **ofrecer contenido adecuado**
- Los comentarios en **redes sociales** se pueden analizar para **gestionar las publicaciones** de una marca de manera oportuna

Procesamiento en Streaming - Ejemplos

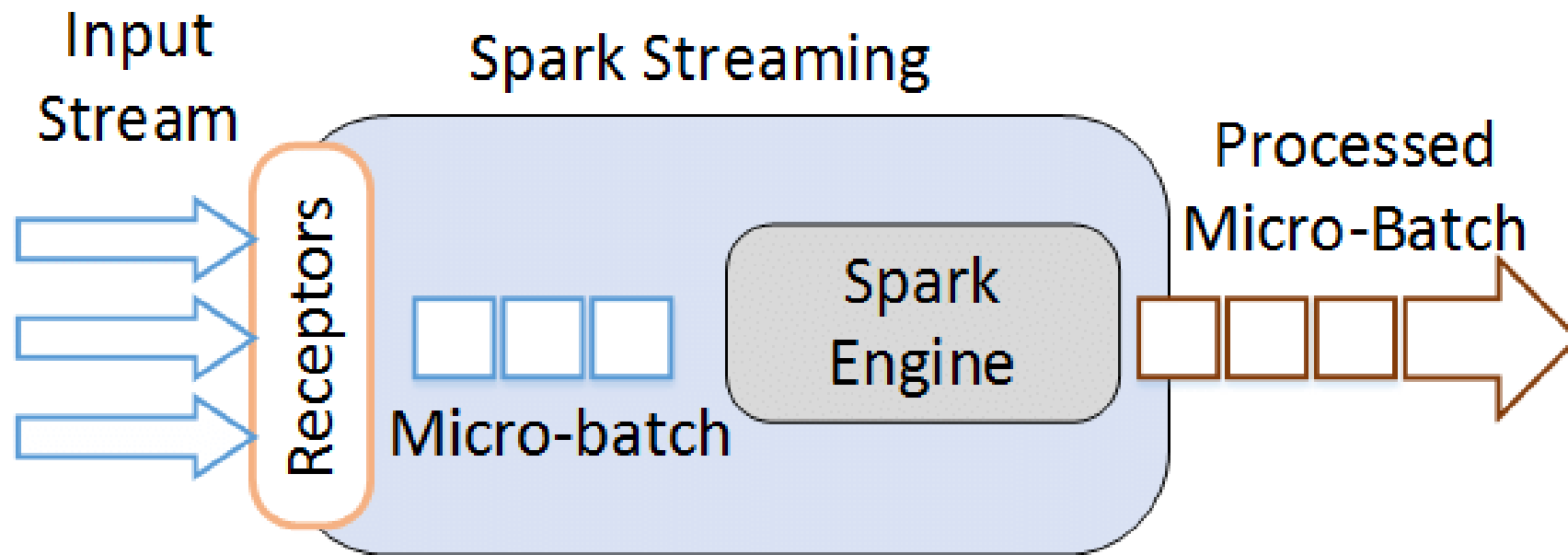


Procesamiento en micro-batch

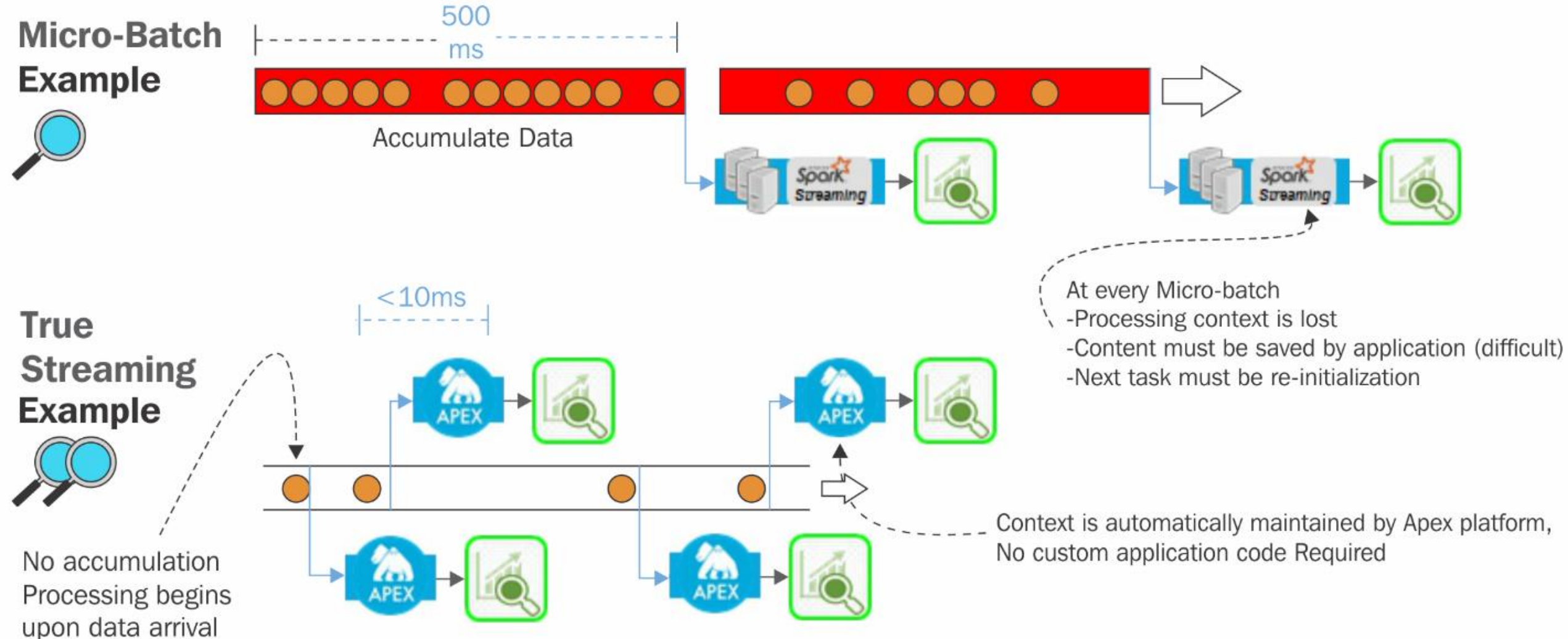
- El procesamiento en micro lotes consiste en **agrupar los datos en pequeños grupos o lotes** para luego ser procesados en conjunto.
- Es una variante del procesamiento por lotes, donde **se procesan grupos pequeños de nuevos datos** y se debe determinar una frecuencia de procesamiento de datos.
- A pesar de no ser un verdadero procesamiento en tiempo real, es suficiente para un gran número de aplicaciones



Procesamiento en micro-batch



micro-batch vs Native Stream



AGENDA

1. Contexto
2. Procesamiento de Datos
- 3. Cloud Computing**
4. Aplicaciones

Definición Cloud Computing

Plataforma **altamente escalable** que promete un acceso rápido al recurso **hardware** o **software** y donde el usuario **no necesita ser experto** para su manejo y acceso.

http://www.innovacion.gob.pa/descargas/FAQ_CloudComputing.pdf

Modelo para habilitar el **acceso** a un conjunto de **servicios computacionales** de manera conveniente y **por demanda**, que pueden ser rápidamente aprovisionados y liberados con un esfuerzo administrativo y una interacción con el proveedor del servicio mínimos.

<https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>

El navegador que antes solo servía para navegar en Internet, se está convirtiendo en nuestro sistema operativo

Tipos de Nube

- **Pública**

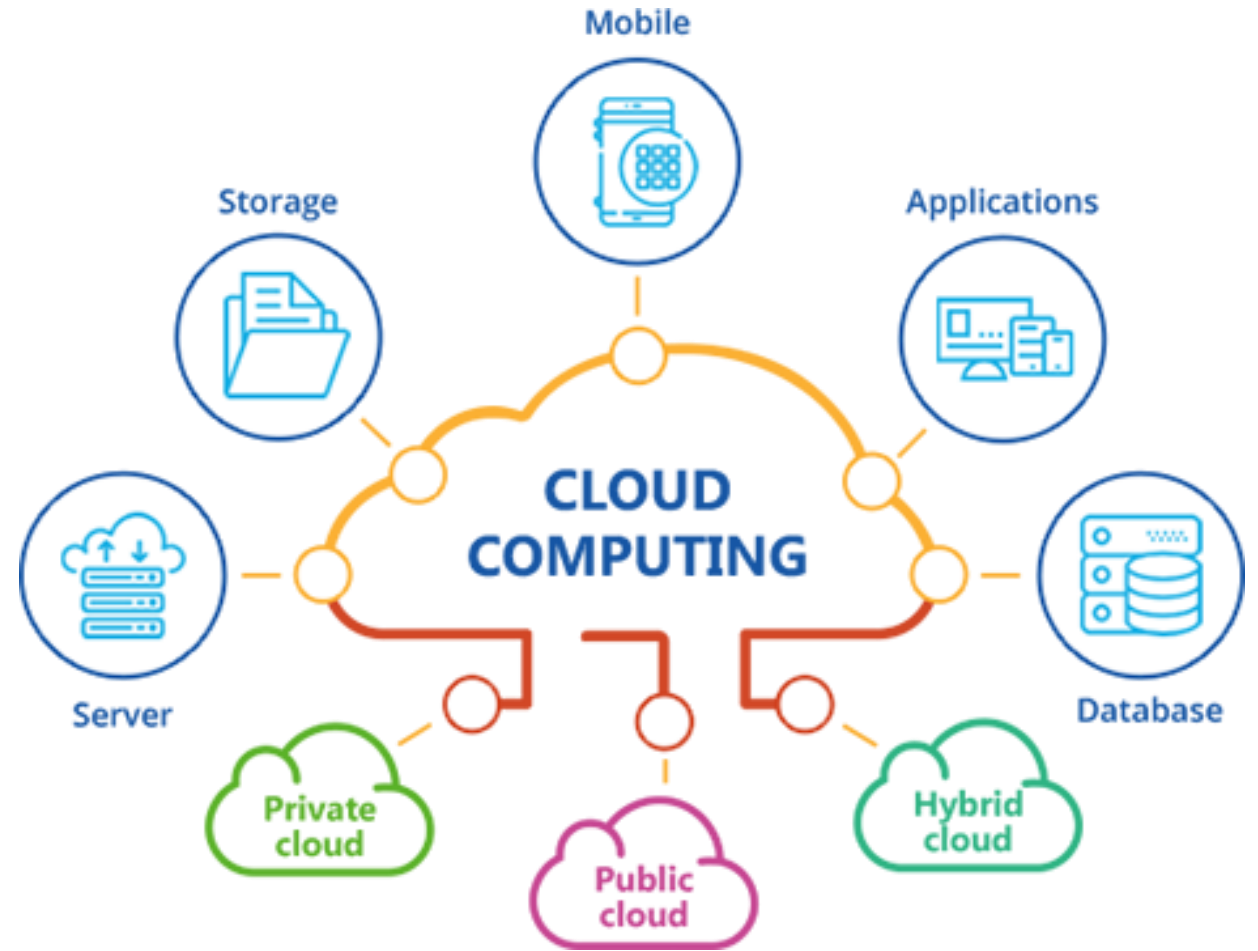
Vende sus servicios a **cualquier usuario**

- **Privada**

Se ofrecen los servicios a un número limitado de **usuarios previamente seleccionados**

- **Híbrida**

Combinación entre nube pública y privada



Características

- **Auto-servicio por demanda (On-demand self-service)**

Disponer de las capacidades de computo, de acuerdo a la necesidad sin la intervención del proveedor del servicio

- **Acceso ubicuo a la red (Broad network access.)**

Servicios disponibles para todo tipo de clientes y dispositivos simultáneamente

- **Agrupación de Recursos (Resource pooling)**

Servicios disponibles para múltiples usuarios con una independencia de la ubicación de los recursos

Características

- **Rápida elasticidad (Rapid elasticity)**

Recursos dinámicos, escalables y elásticos. Pueden variar en función de las necesidades

- **Medición del Servicio (Measured Service)**

El uso de los recursos es monitoreado, medido e informado al usuario

Ventajas

Reducción de costos

No hay inversión en Hardware, mantenimiento y licencias. Se paga solo por los recursos utilizados

Optimización de recursos

Recursos dinámicos que están disponibles solo cuando son necesarios

Fácil recuperación

Los recursos están en la nube generalmente en distintas ubicaciones

Administración

El proveedor se puede encargar de tareas de mantenimiento, actualización, seguridad, entre otras

Disponibilidad

Acceso a los recursos desde cualquier lugar

Desventajas

Percepción de inseguridad

La información se encuentra por fuera de la empresa

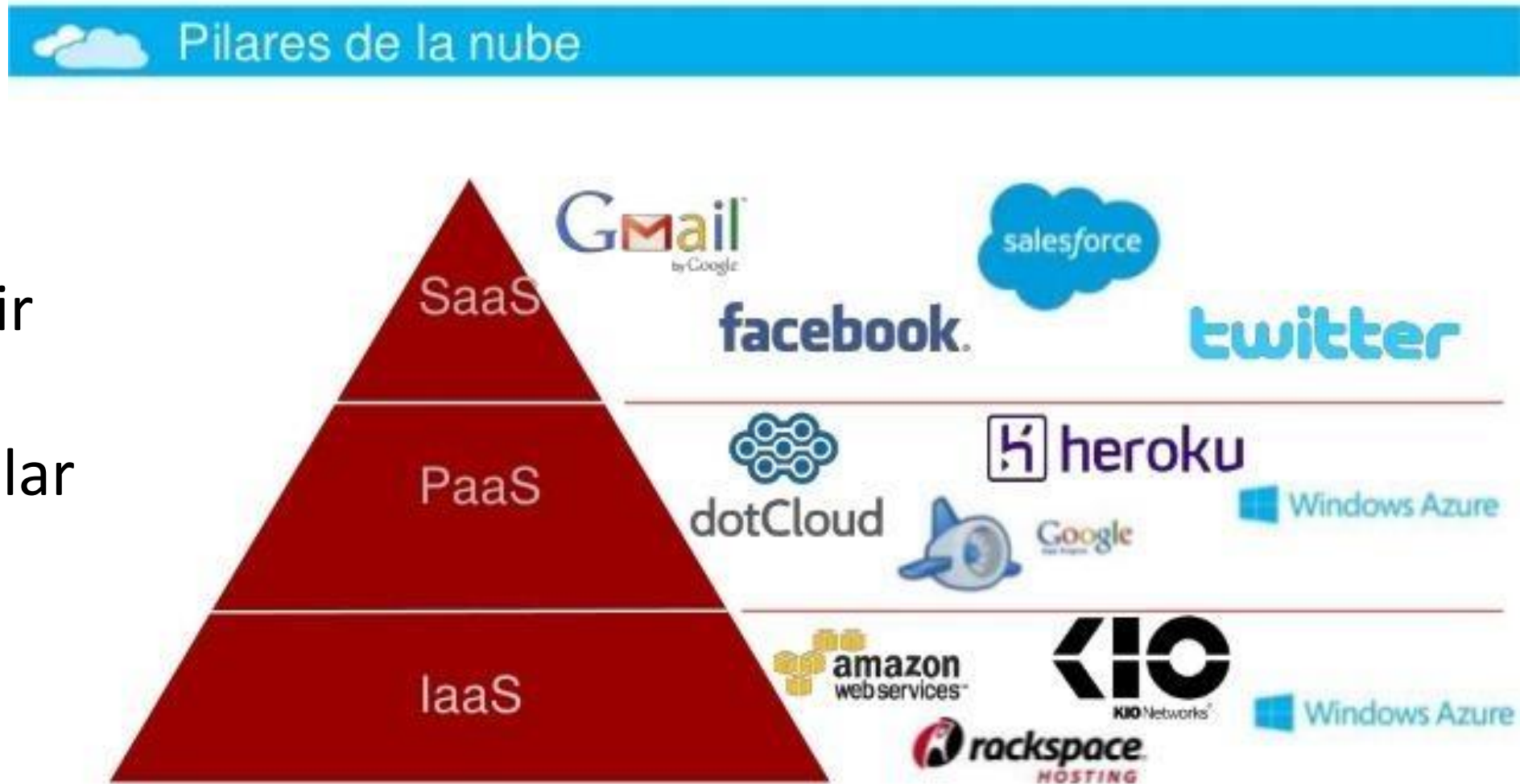
Pérdida de control

No tenemos acceso físico al sitio donde están ubicados los recursos

Acceso a Internet

Si no tenemos Internet, no podemos usar nuestros recursos

Servicios



Cloud Computing - SaaS

Software como servicio (SaaS, Software As A Service)

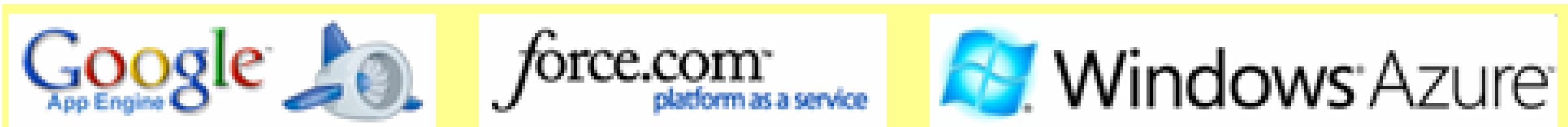
- El usuario accede al software que está alojado en infraestructura de nube
- La forma de acceder al Software es através de un navegador
- No hay control de la infraestructura



Cloud Computing - PaaS

Plataforma como servicio (PaaS, Platform As A Service)

- Solución para la construcción y puesta en marcha de aplicaciones y servicios Web que estarán completamente disponibles a través de Internet.
- Se utiliza la infraestructura de nube para que el usuario publique aplicaciones propias o de terceros
- No hay control sobre la infraestructura pero si sobre la aplicaciones



Cloud Computing - IaaS

Infraestructura como servicio (IaaS, Infrastructure As A Service)

- Se dispone de infraestructura de computación como un servicio, usando virtualización
- El cliente compra recursos para hosting, capacidad de cómputo, redes, entre otras
- No hay control directo sobre la infraestructura, pero si se puede controlar el sistema operativo



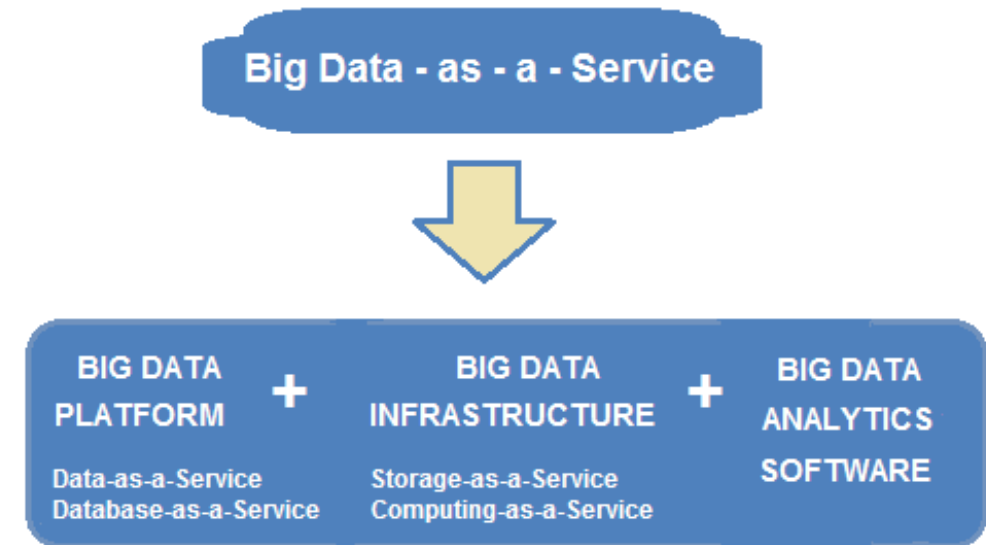
Cloud Computing – Otros Servicios

Escritorio como servicio (Daas, Desktop as a Service)

- Escritorios virtuales alojados en la nube por un proveedor cloud.

Big Data como servicio (BDaaS, Big Data as a Services)

- Servicios que ofrecen análisis de conjuntos de datos grandes o complejos, utilizando los servicios alojados en la nube.



Cuadrante Mágico de Gartner: Servicios estratégicos de plataforma en la nube

https://www.gartner.com/technology/media-products/reprints/amazon/1-2FTPMD8S-MX.html?trk=44f67619-4f3b-42e8-93b9-32ad8a123845&sc_channel=el&refid=5b366b4c-27fd-4b41-93cb-3195213bc9f6.



AWS Services

Deployment & Management

Application Services



Amazon
SQS



Amazon
ElasticTranscoder



Amazon
SES



Amazon
AppStream



Amazon
CloudSearch

Mobile Services



Amazon
Cognito



Amazon
Mobile Analytics



Amazon
SNS

Enterprise Applications



Amazon
WorkDocs



Amazon
WorkSpaces



Amazon
WorkMail

Application Services

Administration & Security



AWS
DirectoryService



AWS
IAM



AWS
Trusted Advisor



AWS
Config



AWS
CloudTrail



Amazon
CloudWatch

Deployment & Management



Amazon
CloudFormation



AWS
OpsWorks



AWS
CodeDeploy

Analytics



Amazon
Kinesis



AWS
Data Pipeline



Amazon
EMR

Foundation Services

Compute



Amazon
EC2



AWS
Lambda

Storage & Content Delivery



Amazon
CloudFront



Amazon
Glacier



AWS
Storage Gateway



Amazon
Content Delivery

Database



Amazon
DynamoDB



Amazon
RDS



Amazon
Redshift



Amazon
Elastic Cache

Networking



Amazon
Route 53

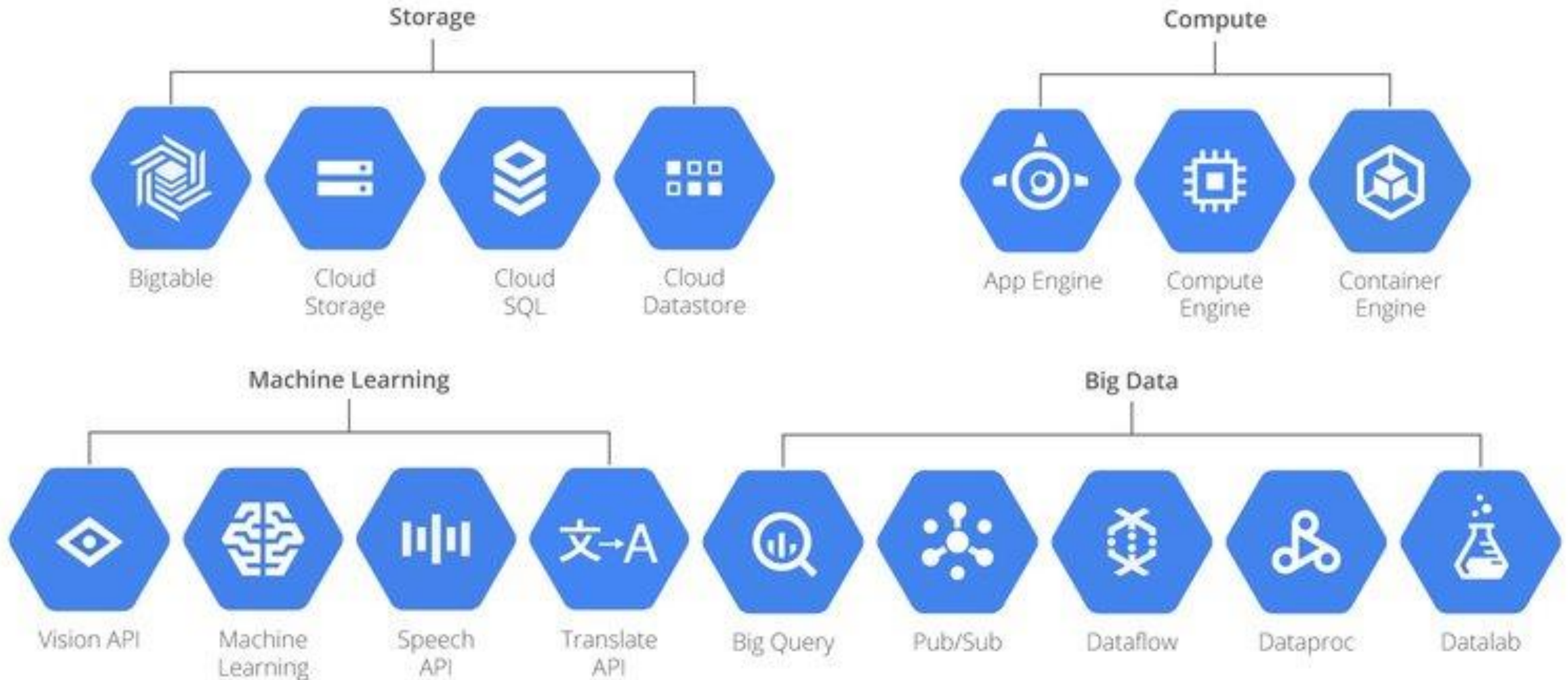


Amazon
VPC



AWS
Direct Connect

Google Cloud Platform



Microsoft Azure

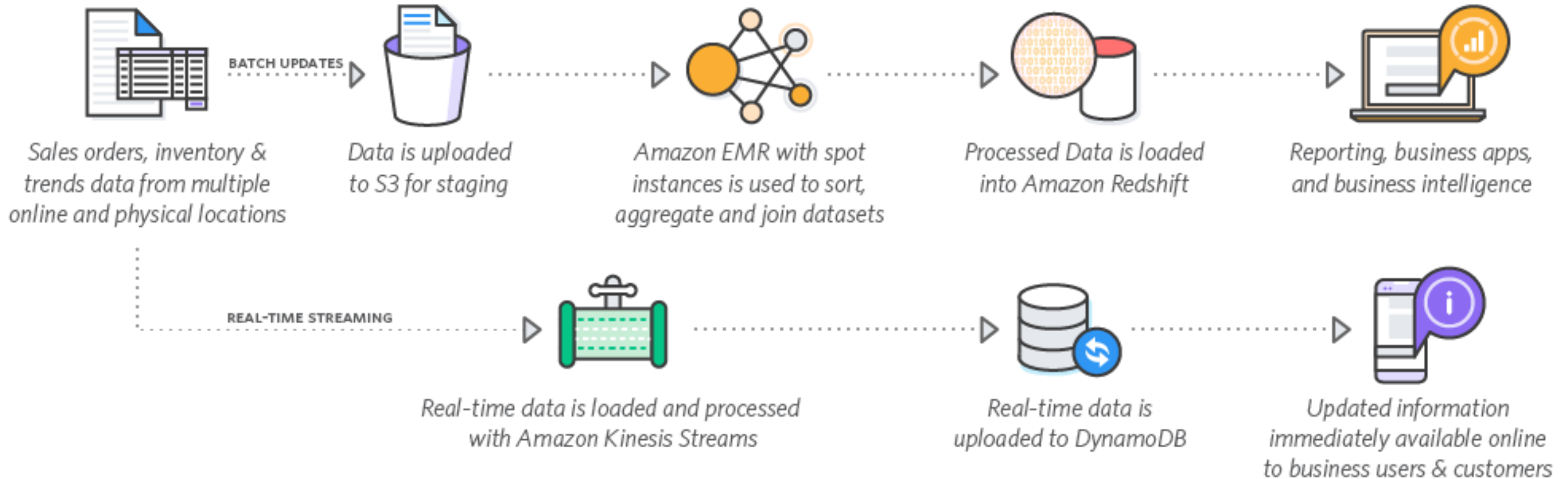


AGENDA

1. Contexto
2. Procesamiento de Datos
3. Cloud Computing
- 4. Aplicaciones**

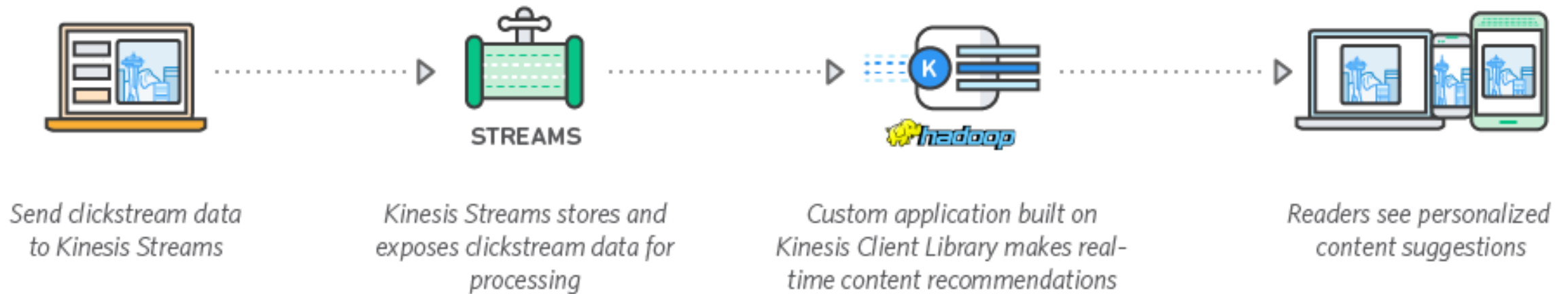
Aplicaciones

REDFIN.



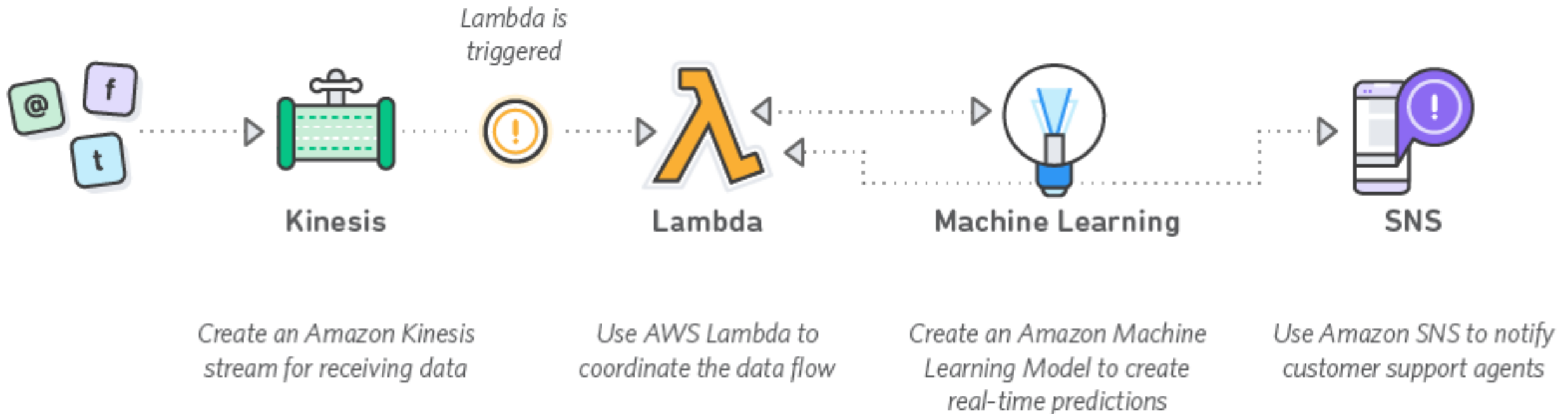
Aplicaciones

HEARST *corporation*



<https://aws.amazon.com/es/solutions/case-studies/hearst/>

Aplicaciones



<https://aws.amazon.com/es/solutions/case-studies/buildfax/>

Aplicaciones



<https://customers.microsoft.com/es-es/story/1423193863644293457-nba-media-entertainment-azure-es-xl>



<https://customers.microsoft.com/es-es/story/1473578443276306172-iberia-express-other-azure-es-spain>



<https://cloud.google.com/customers/ach-colombia/?hl=es-419>



<https://cloud.google.com/customers/auteco-mobility/?hl=es-419>



<https://cloud.google.com/customers/sodimac/?hl=es-419>



<https://cloud.google.com/customers/globo/?hl=es-419>

Contenido

Tema	Detalle
Introducción	Procesamiento de datos en batch, streaming y micro-batch. Cloud Computing
Procesamiento de Datos	Serialización de Datos (JSON, XML, YAML) Protocol Bufer, Apache Thrift
Databricks	Cluster Spark en lenguajes: SQL, Python, R, Scala
AWS	VPC, EC2, S3, EMR, CloudWatch