UdeA
Analytics &
DataScience

DATA STREAMING Y SERVICIOS EN LA NUBE

INTRODUCCIÓN

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

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

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS

AGENDA

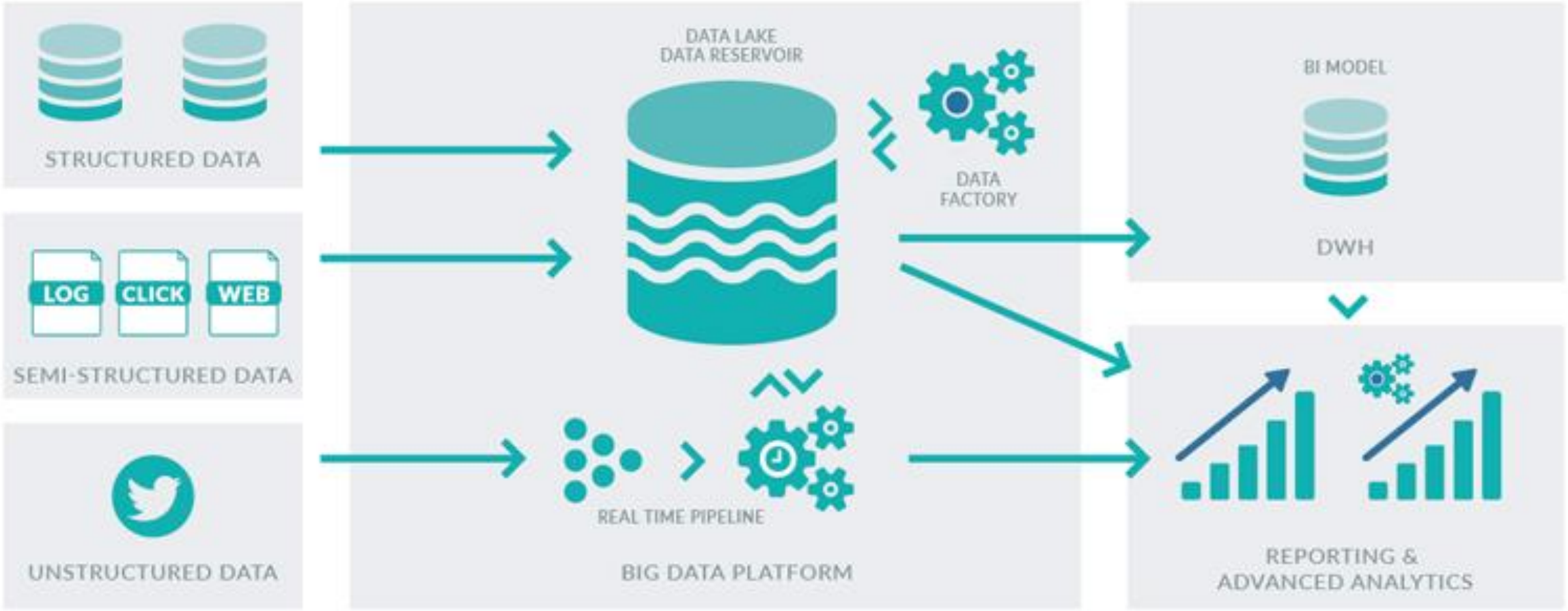
UdeA
Analytics &
DataScience

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

https://nexusintegra.io/es/business-intelligence-industria-inteligencia-negocio/

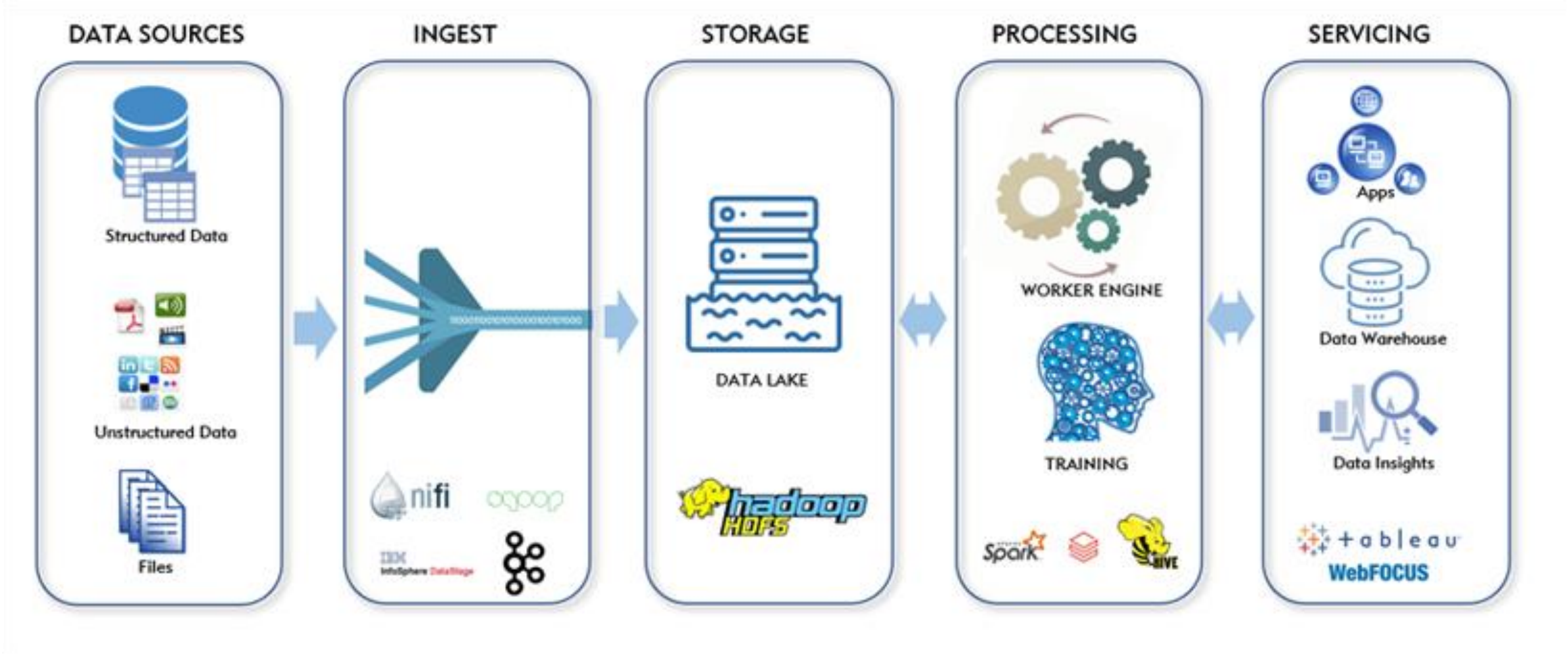
Inteligencia de Negocios Tradicional

Inteligencia de Negocios con Big Data

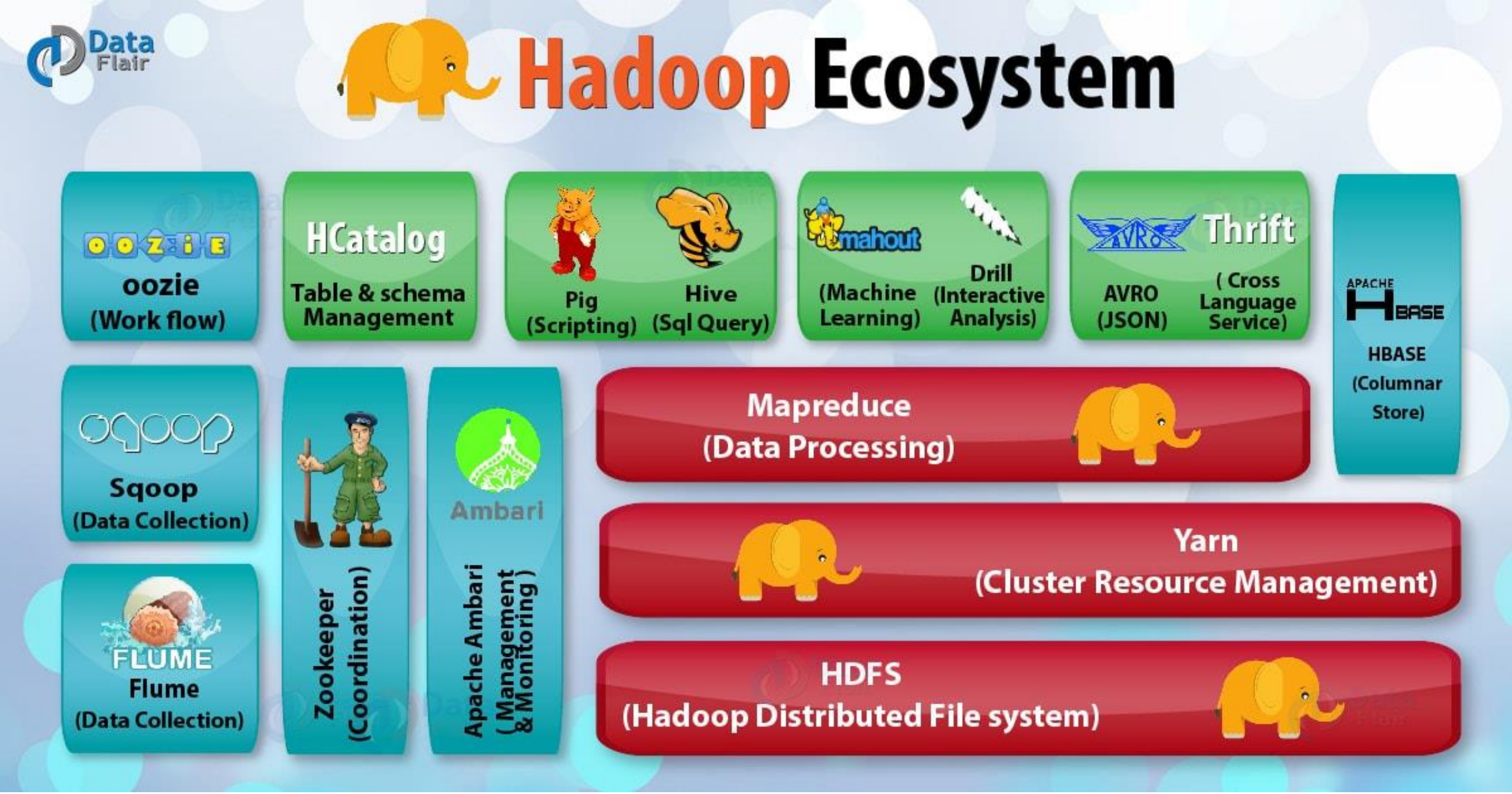


https://www.clearpeaks.com/what-can-big-data-do-for-bi/

Big Data & Analytics

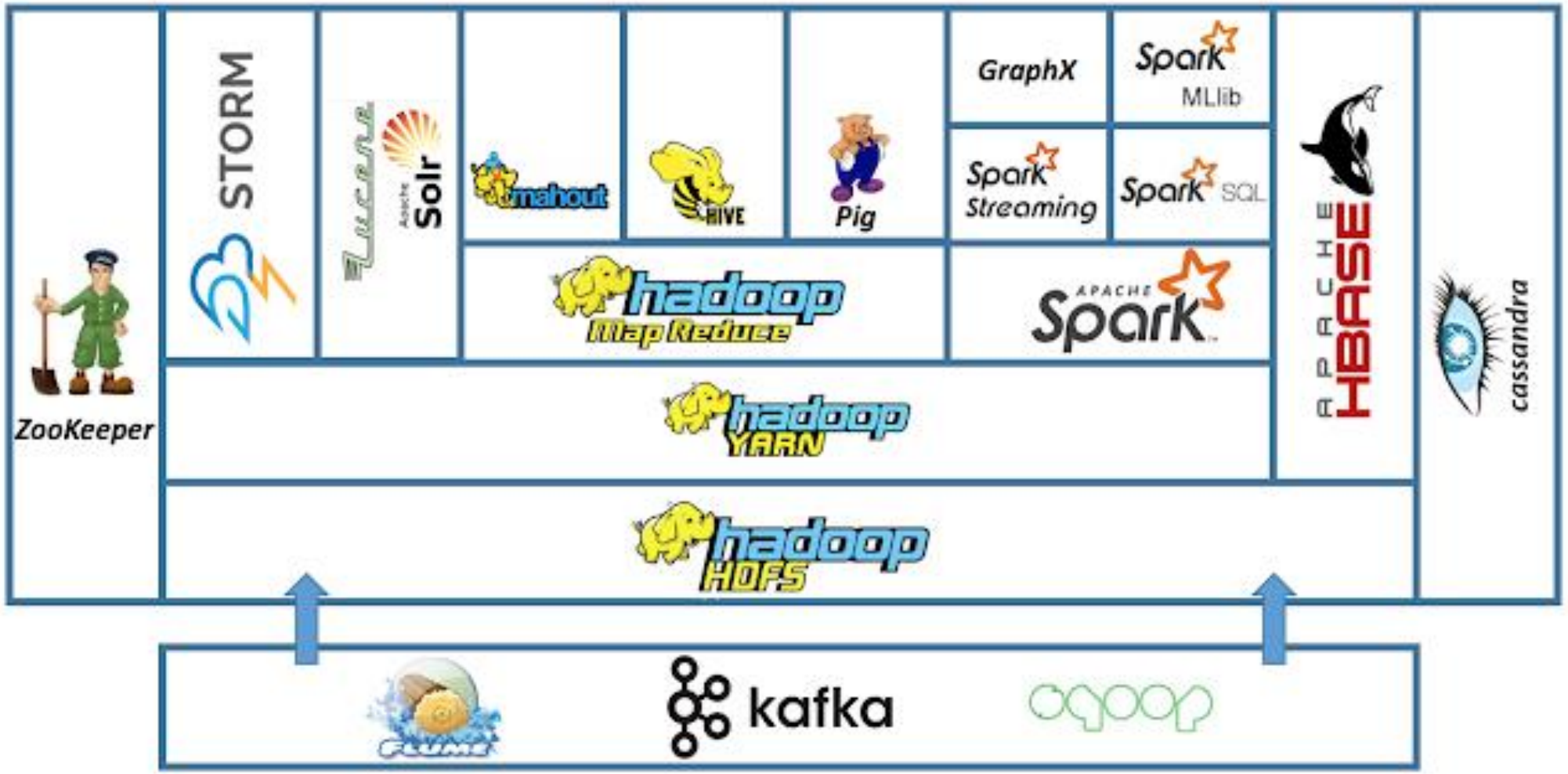


https://www.narwalinc.com/services-big-data-and-analytics

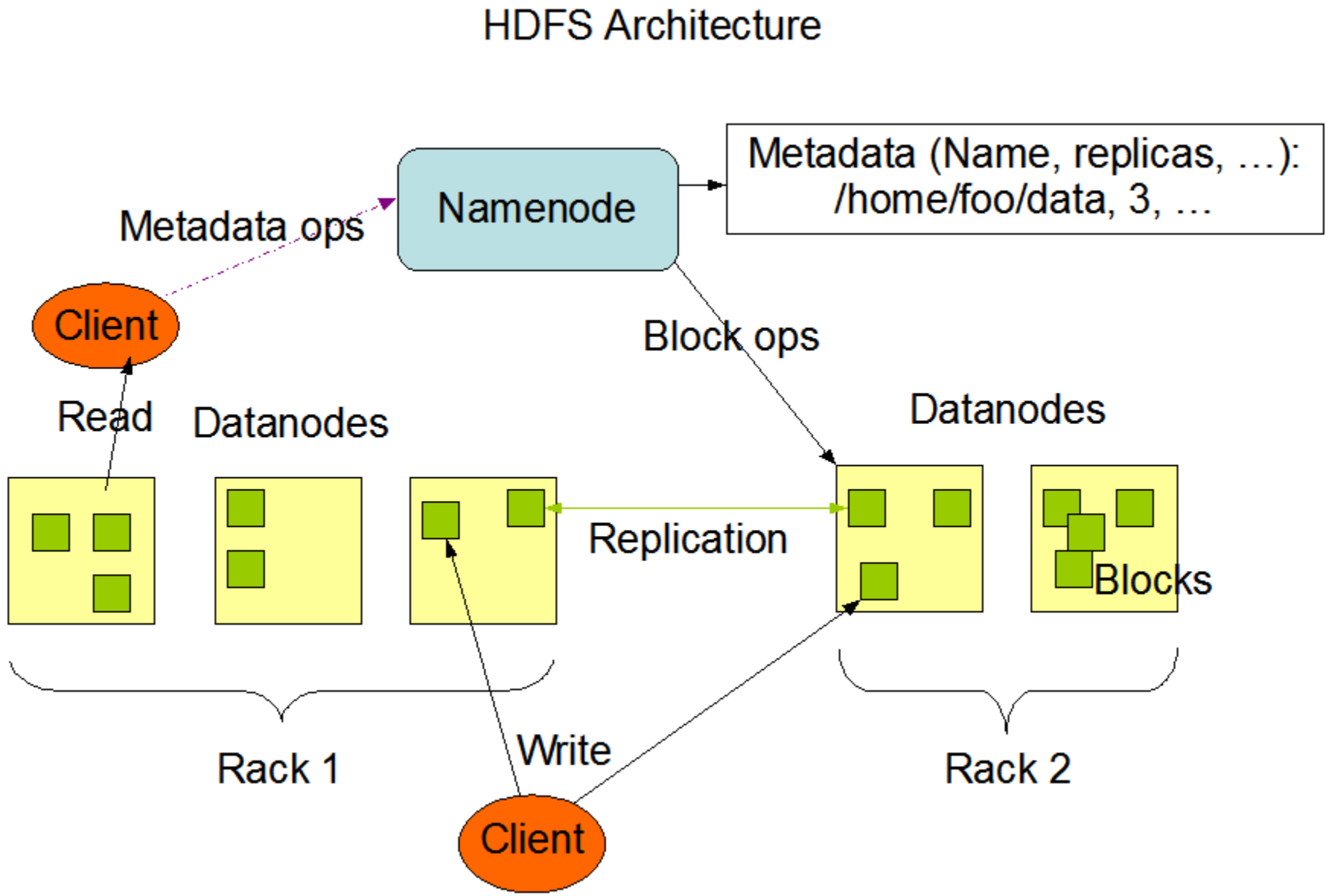


https://data-flair.training/blogs/hadoop-ecosystem-components/

Ecosistema de Hadoop con Spark

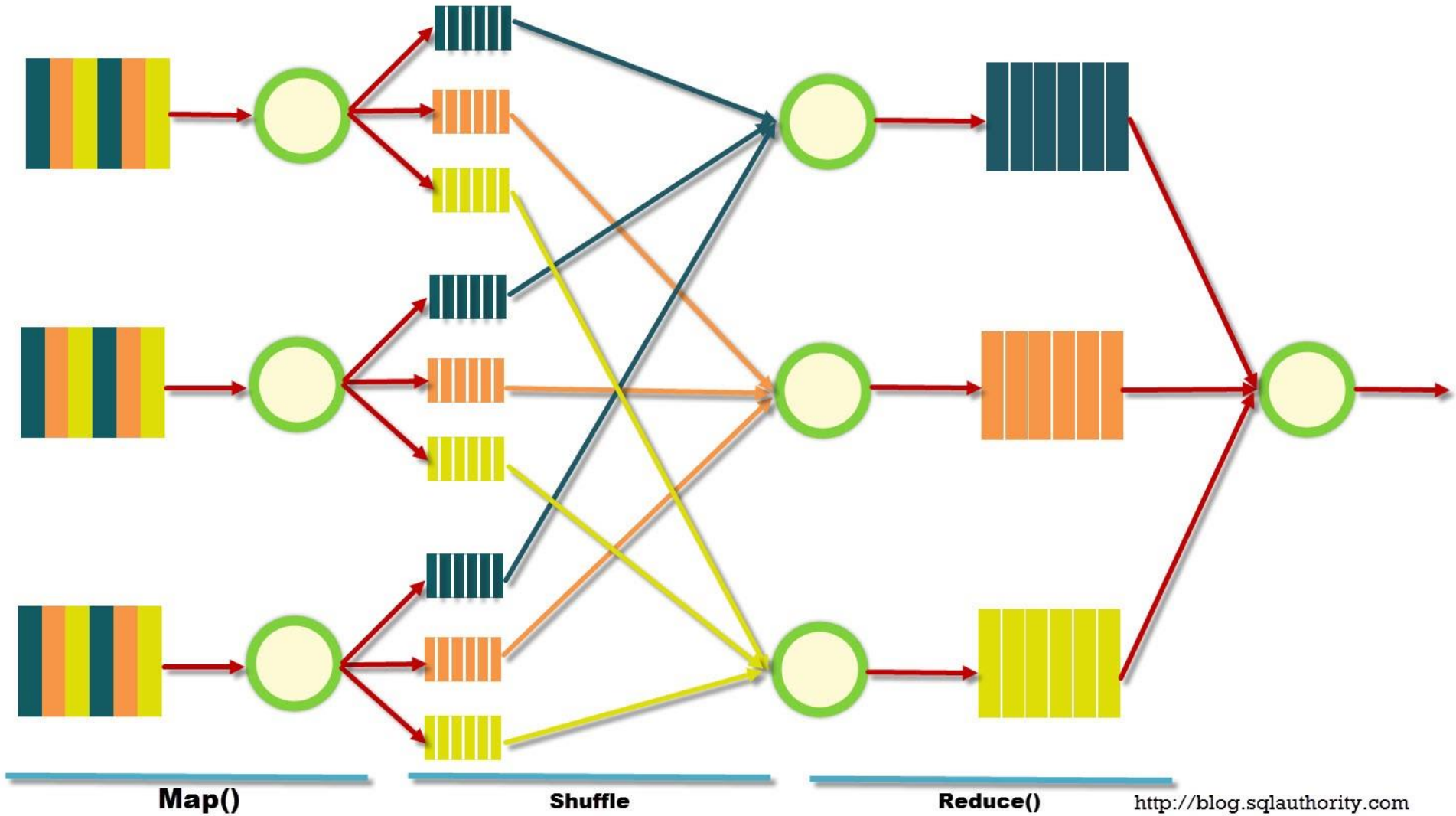


http://blog.newtechways.com/2017/10/apache-hadoop-ecosystem.html



http://hadoop.apache.org/docs/stable1/hdfs_design.pdf

How MapReduce Works?

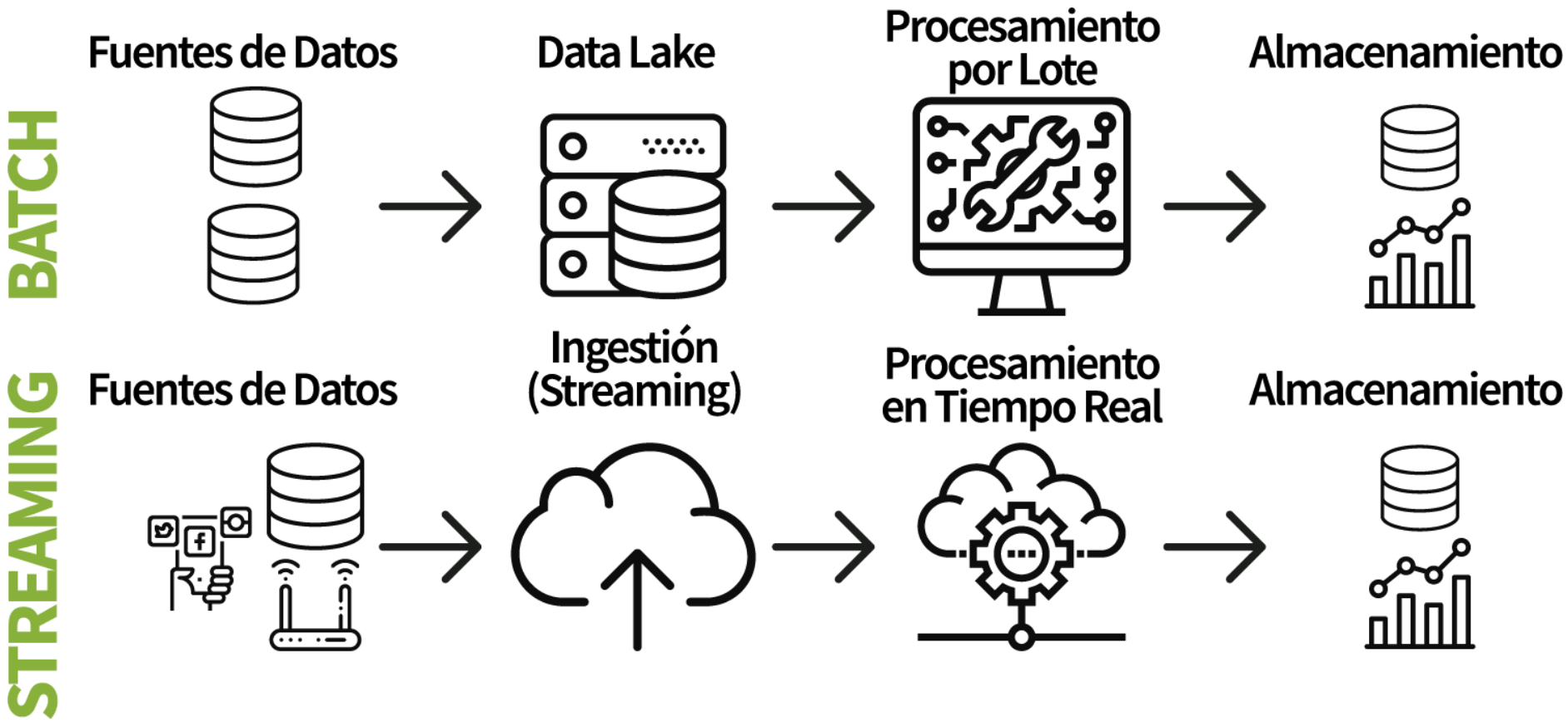


AGENDA

UdeA
Analytics &
DataScience

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

Tipos de procesamiento



https://www.pandaid.com/big-data-procesos-en-tiempo-real/

Procesamiento en Batch

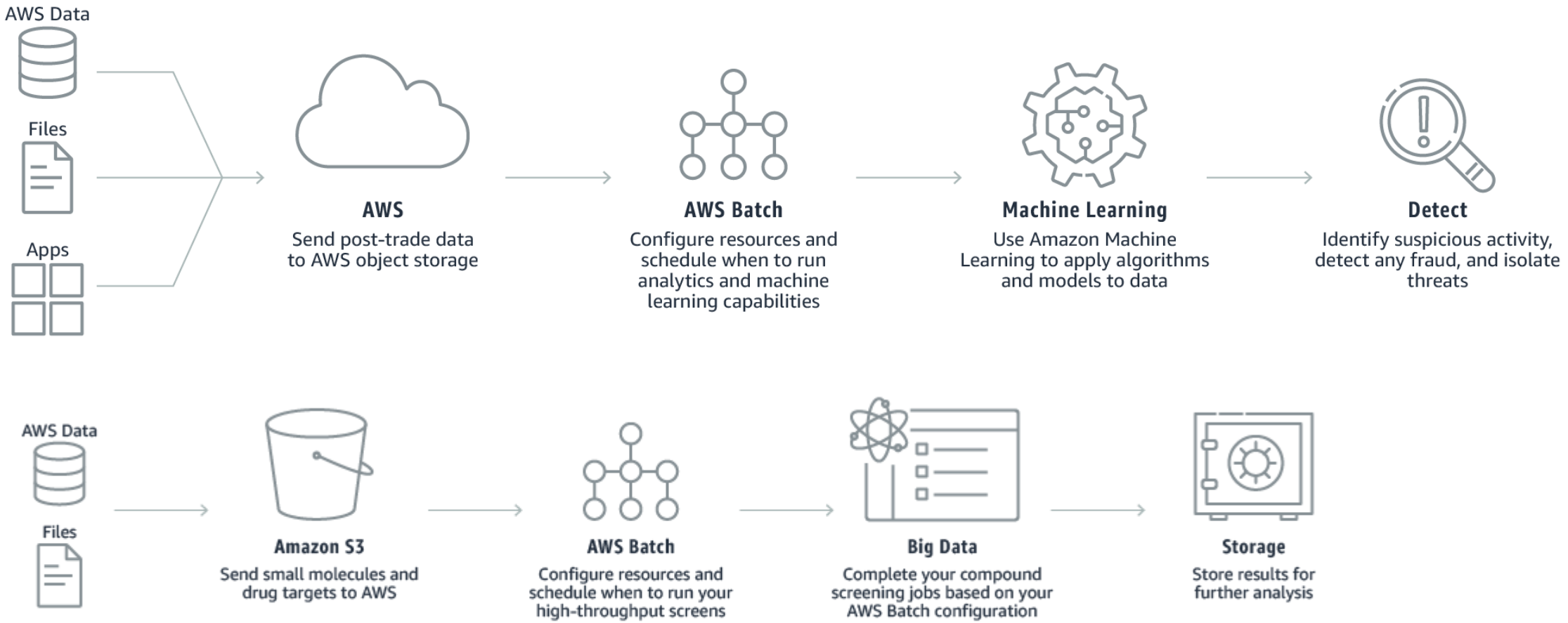
Un lote (batch) es una colección de datos que ha sido agrupada durante un intervalo de tiempo.

- 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 - Ejemplos

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

Procesamiento en Batch



https://aws.amazon.com/es/batch/use-cases/

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**

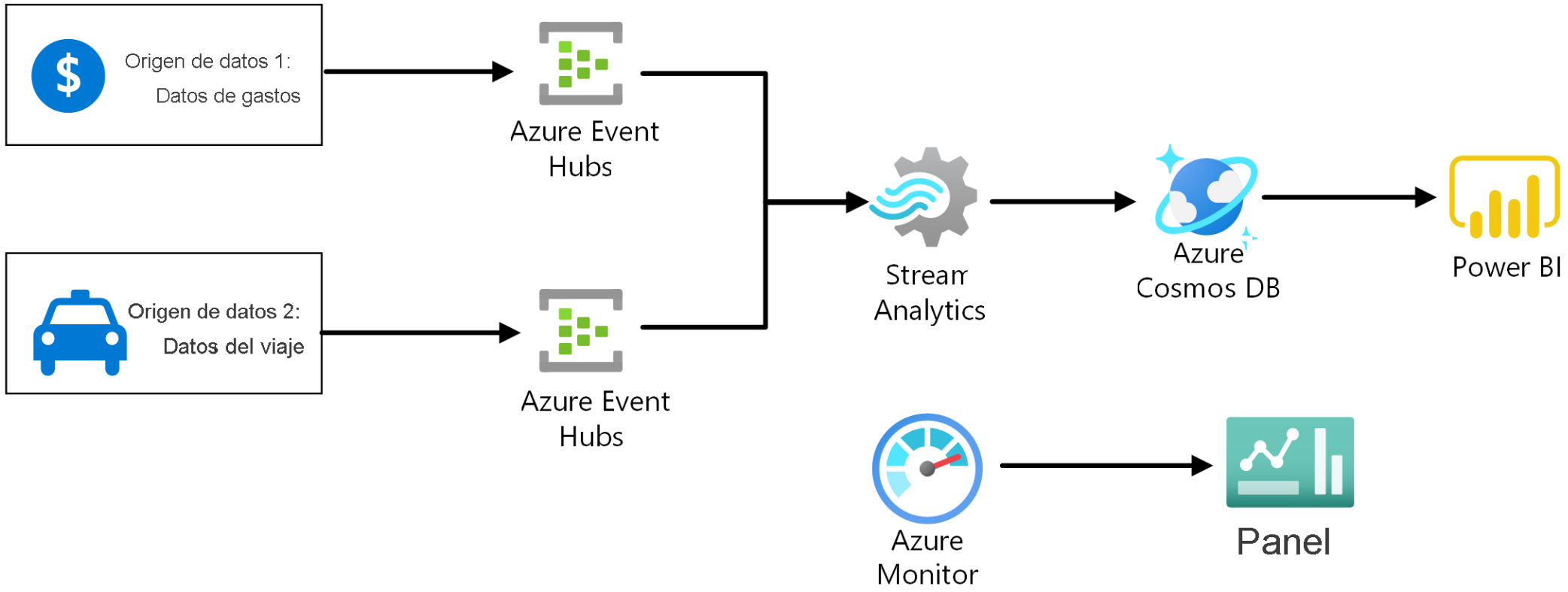
Procesamiento en Streaming - Ejemplos

- 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



Microsoft Azure

https://learn.microsoft.com/es-es/azure/architecture/reference-architectures/data/stream-processing-stream-analytics

AGENDA

UdeA
Analytics &
DataScience

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**

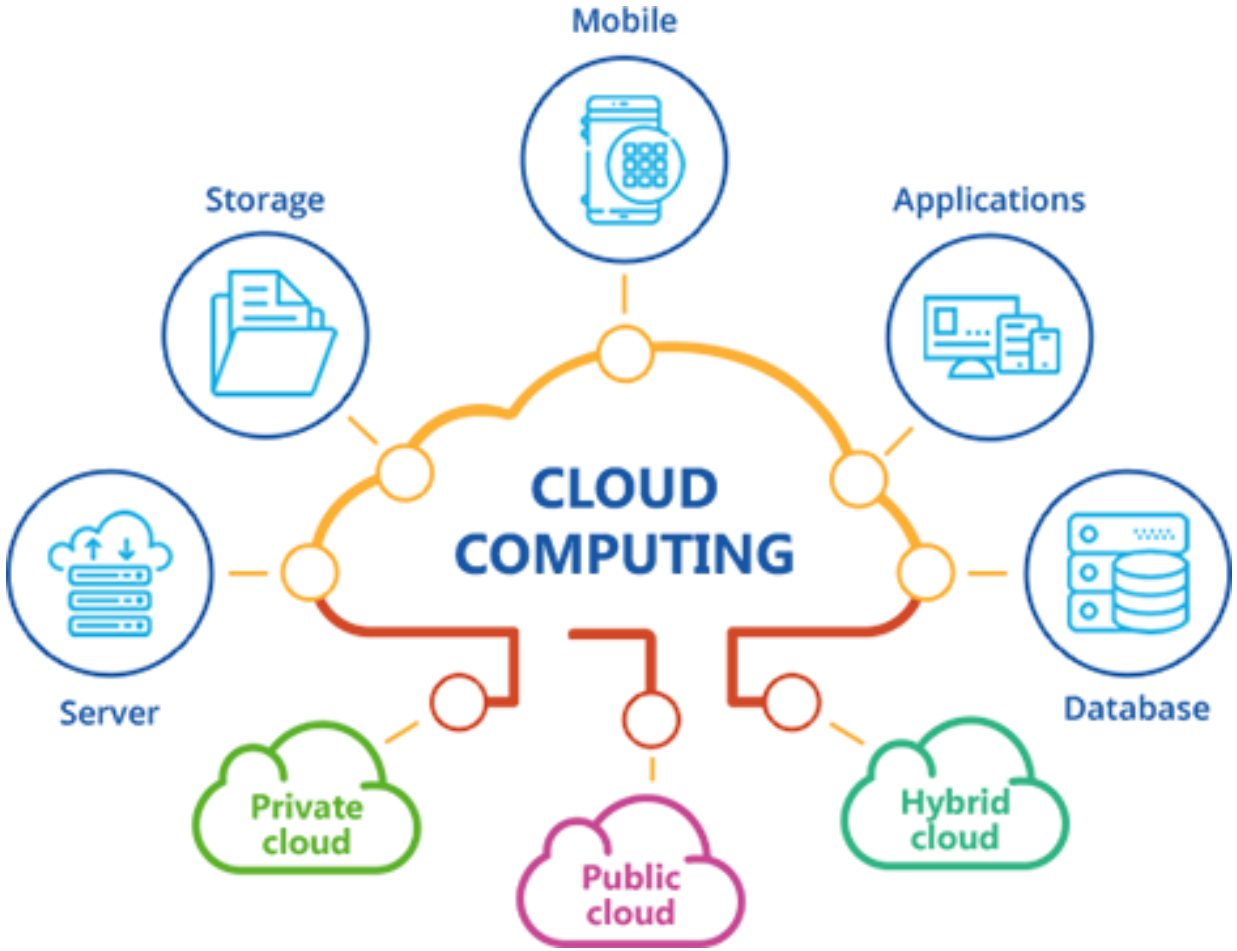
  Vende sus servicios a cualquier usuario

- **Privada**

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

- **Híbrida**

  Combinación entre nube pública y privada



https://www.ittravelservices.com/cloud-computing.html

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, ente 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

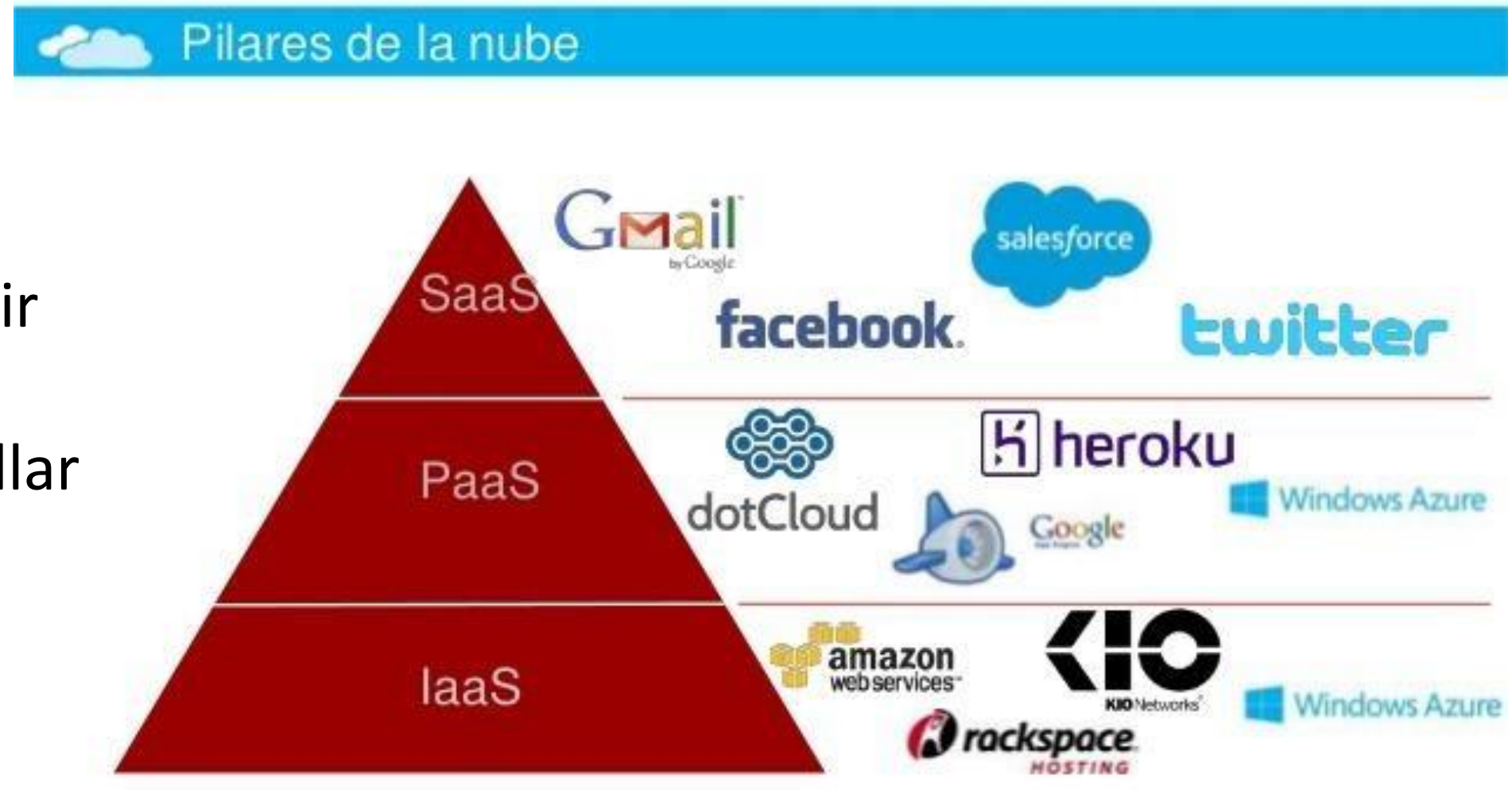
Si no tenemos Internet, no podemos usar nuestros recursos

Servicios

Consumir

Desarrollar

Migrar



https://sites.google.com/site/siclou003/conceptos-previos/capas-de-cloud-computing

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

https://sreeninet.files.wordpress.com/2014/02/cloud_service4.png

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

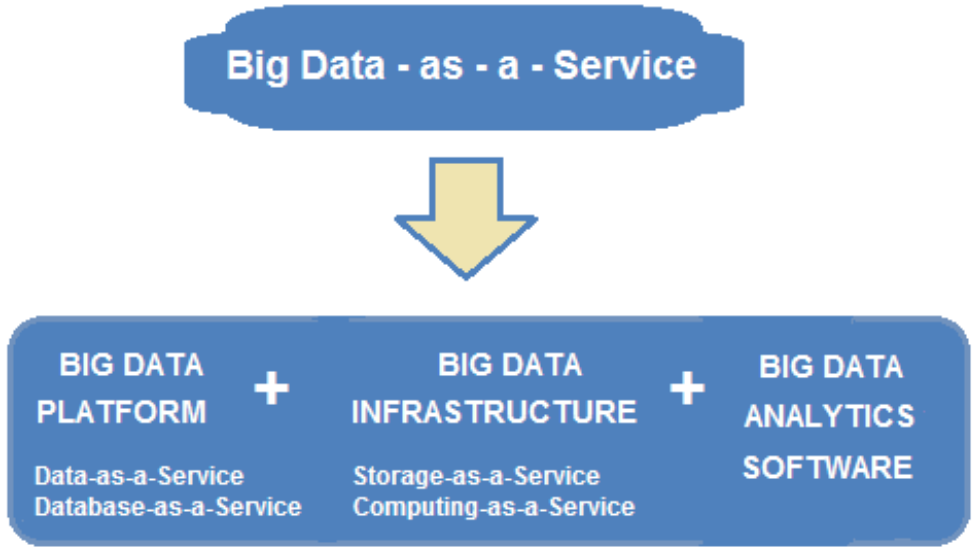
Cloud Computing – Otros Servicios

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.

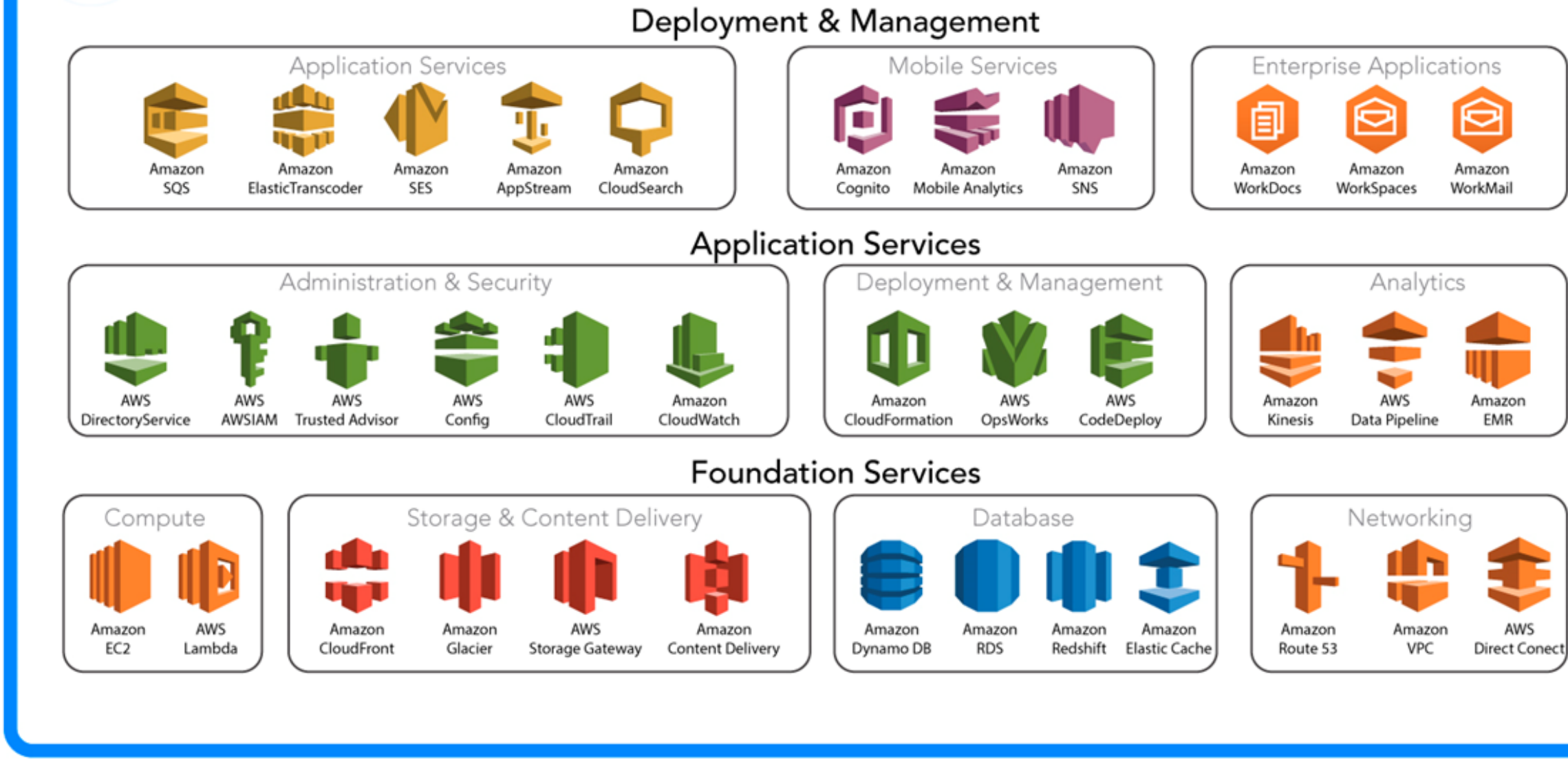


http://ivyproschool.com/blog/2015/01/14/big-data-as-a-service-bdaas-an-emerging-trend-of-2015/

Cuadrante Mágico de Gartner: Infraestructura en la nube

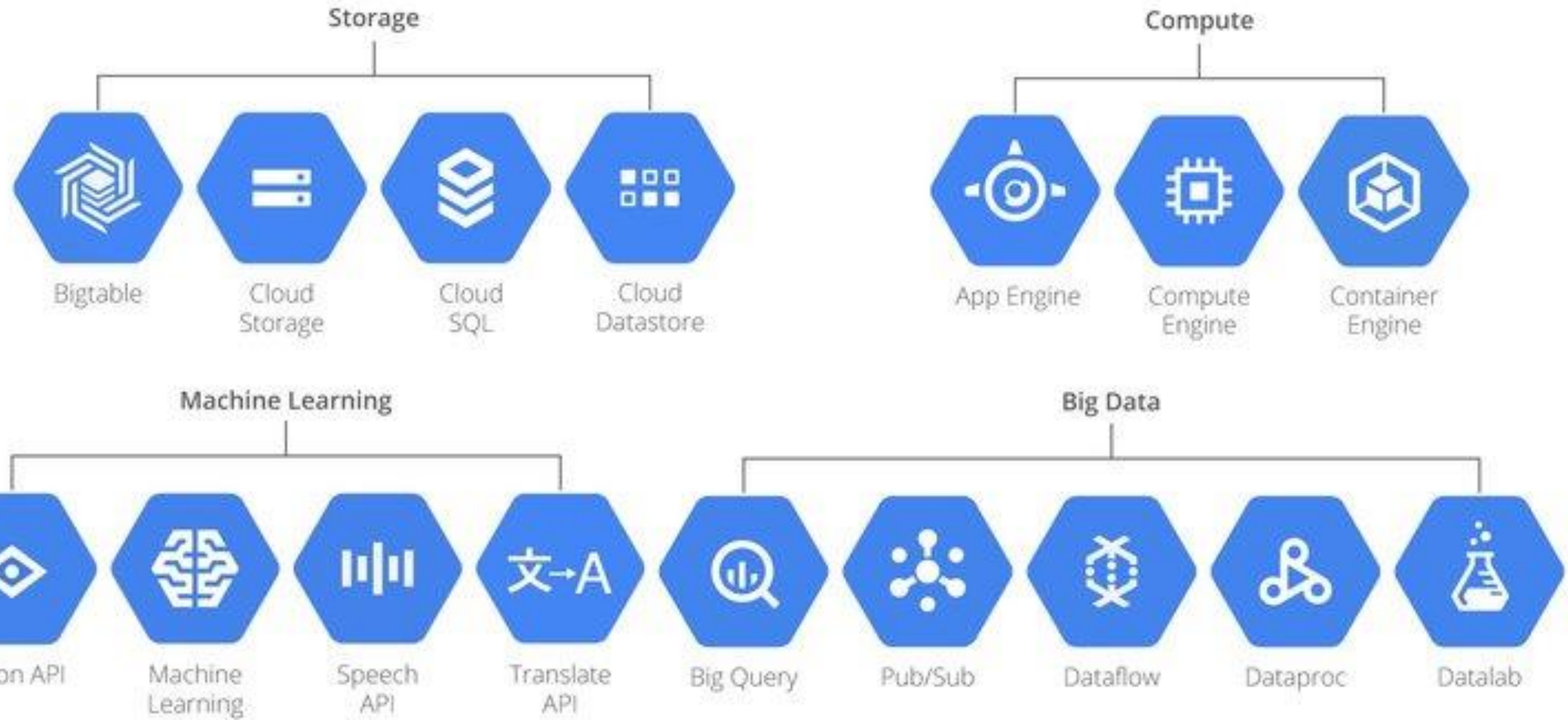
Gartner, [Magic Quadrant for Cloud Infrastructure and Platform Services](), 19 October 2022, Raj Bala, et. al.

AWS Services



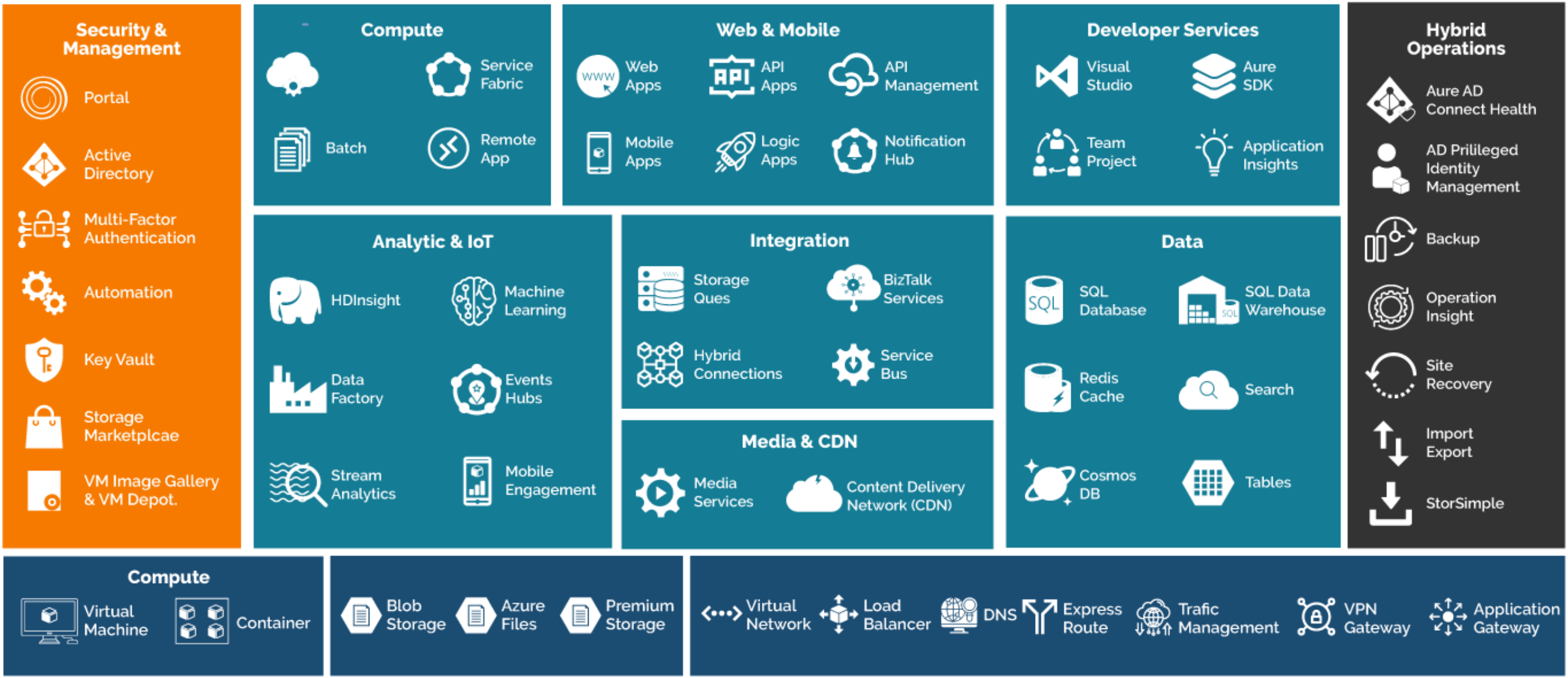
https://aws.amazon.com/es/

Google Cloud Platform



https://cloud.google.com/

Microsoft Azure



https://azure.microsoft.com/es-es/

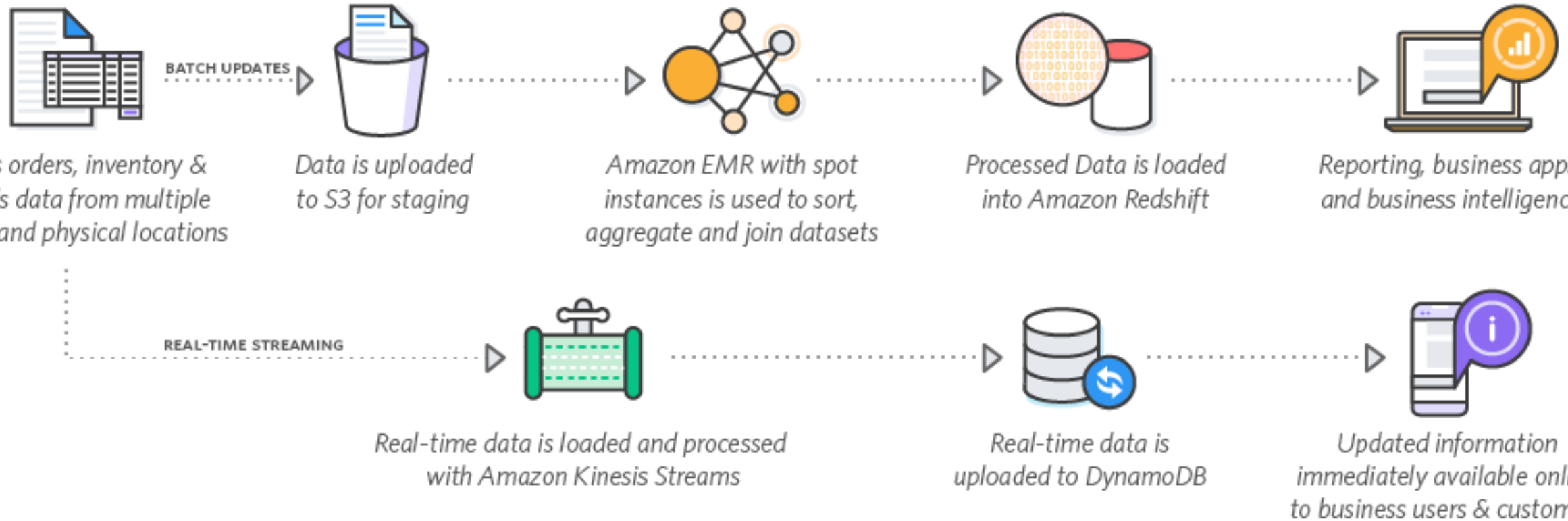
AGENDA

UdeA
Analytics &
DataScience

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

Aplicaciones

REDFIN.



https://aws.amazon.com/es/solutions/case-studies/redfin/

Aplicaciones

HEARST *corporation*

Send clickstream data to Kinesis Streams

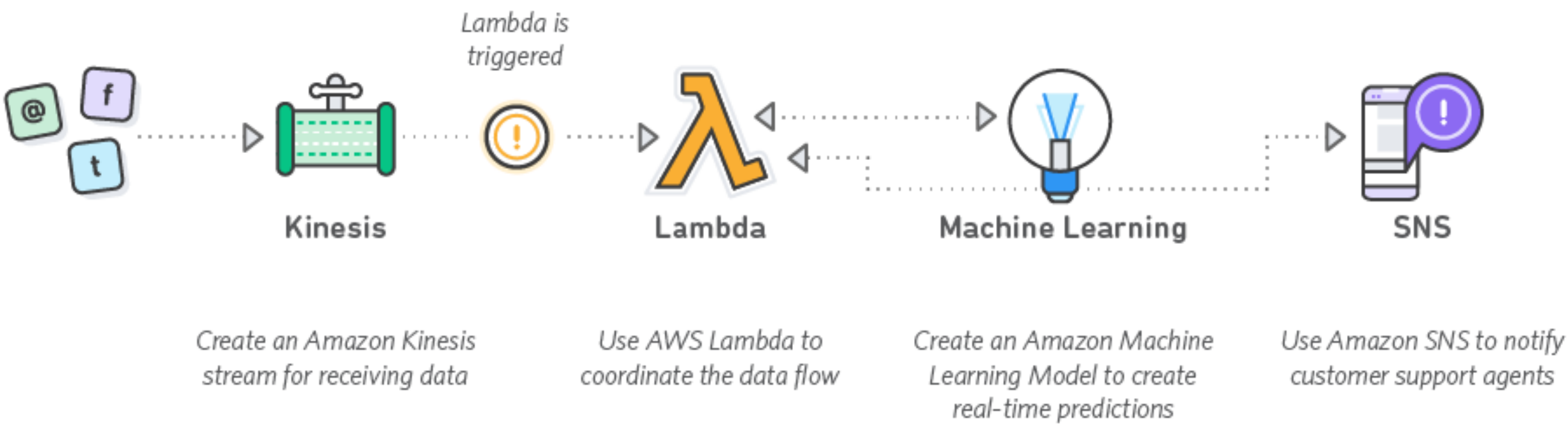
Kinesis Streams stores and exposes clickstream data for processing

Custom application built on Kinesis Client Library makes real-time content recommendations

Readers see personalized content suggestions

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/14231938636442934 57-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

UdeA
Analytics &
DataScience

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