

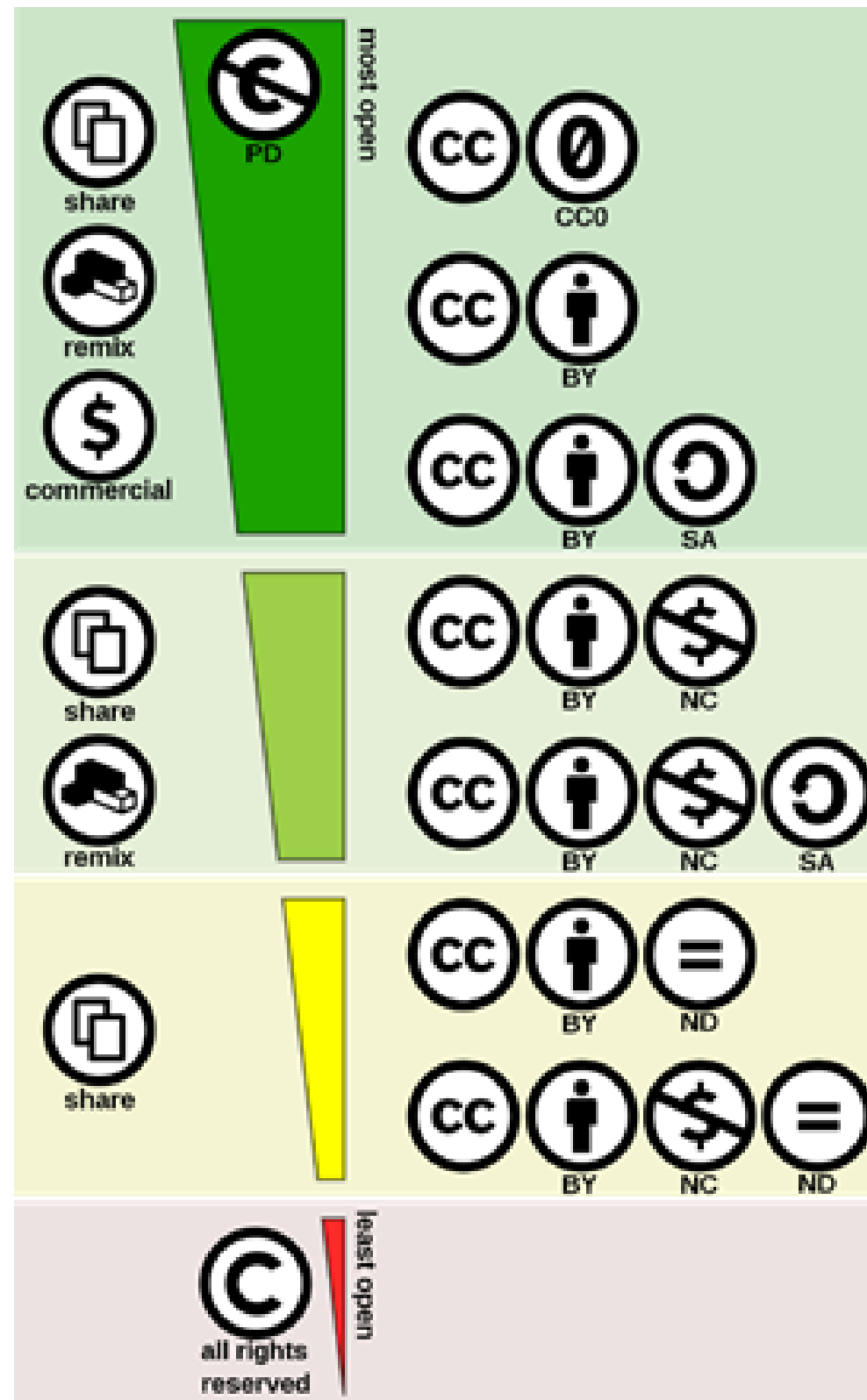
Analítica de Datos

OPEN DATA, Arquitectura BIG DATA
Y Desarrollo de Soluciones para
Inteligencia de Negocios



Los Datos Abiertos

Es una iniciativa para permitir que los datos estén disponibles para todos aquellos que los necesitan.



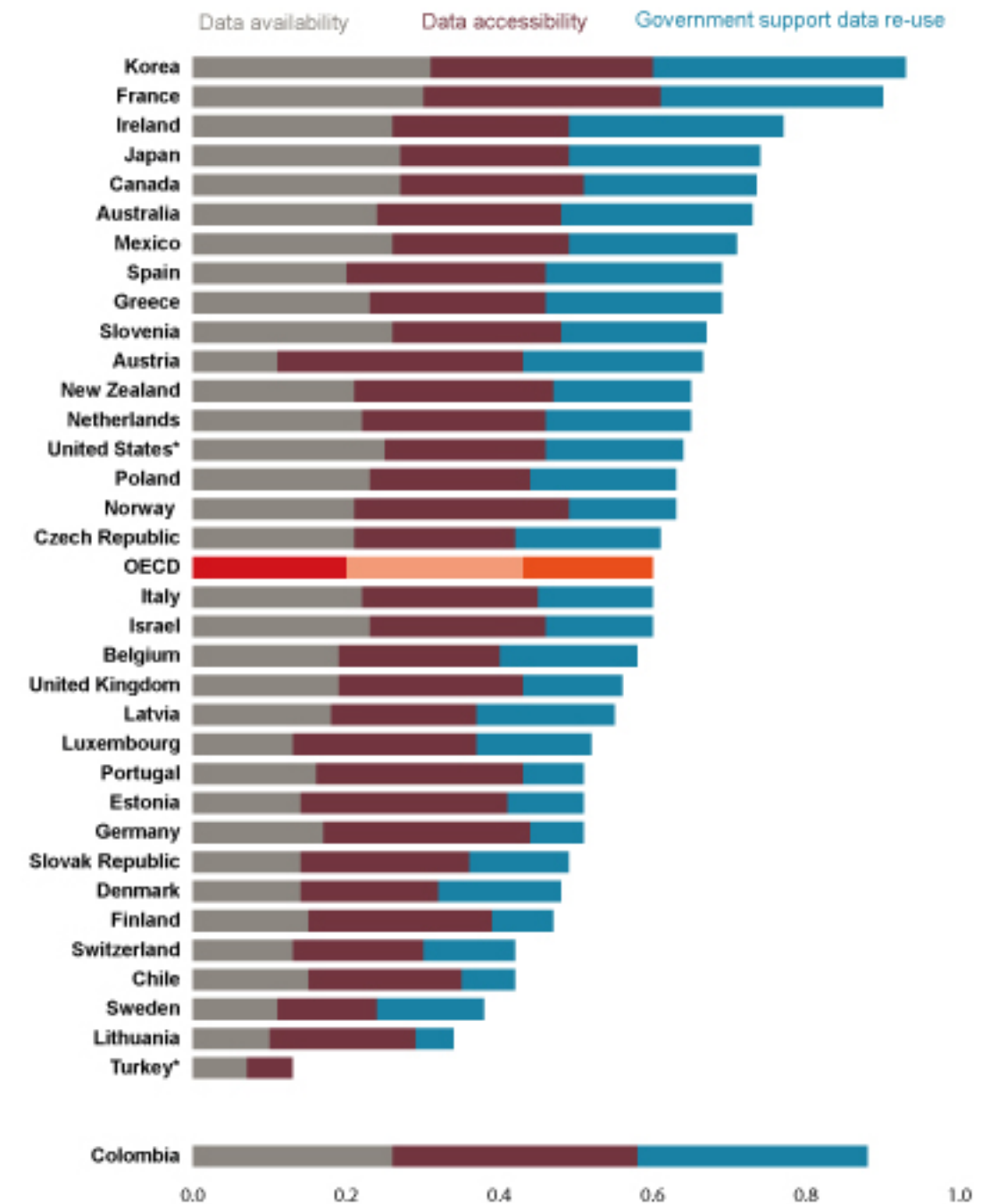
Espectro CC - Shaddim; original CC license symbols by Creative Commons

Datos Gubernamentales

Es una iniciativa para permitir que los datos recogidos y almacenados por los estados estén disponibles para sus ciudadanos, bajo la premisa de que dichos datos fueron obtenidos de los mismos ciudadanos y pagados a través de los impuestos.

OURdata Index: Open-Useful-Reusable Government Data 2019

Composite index: 0 lowest to 1 highest



*Data for United States and Turkey is from 2017
Source: OECD Government at a Glance, 2019



Data index 2019- OECD.org

Actividad

Entre a <https://www.datos.gov.co/> y busque datos sobre su ciudad que hablen de:

- Niñez
- Seguridad
- Cobertura de servicios
- Discotecas

La pagina de Datos Abiertos de Colombia tiene 5421 conjuntos de datos a Marzo 2021.

2020 *This Is What Happens In An Internet Minute*



BIG DATA

Los datos se dividen en:

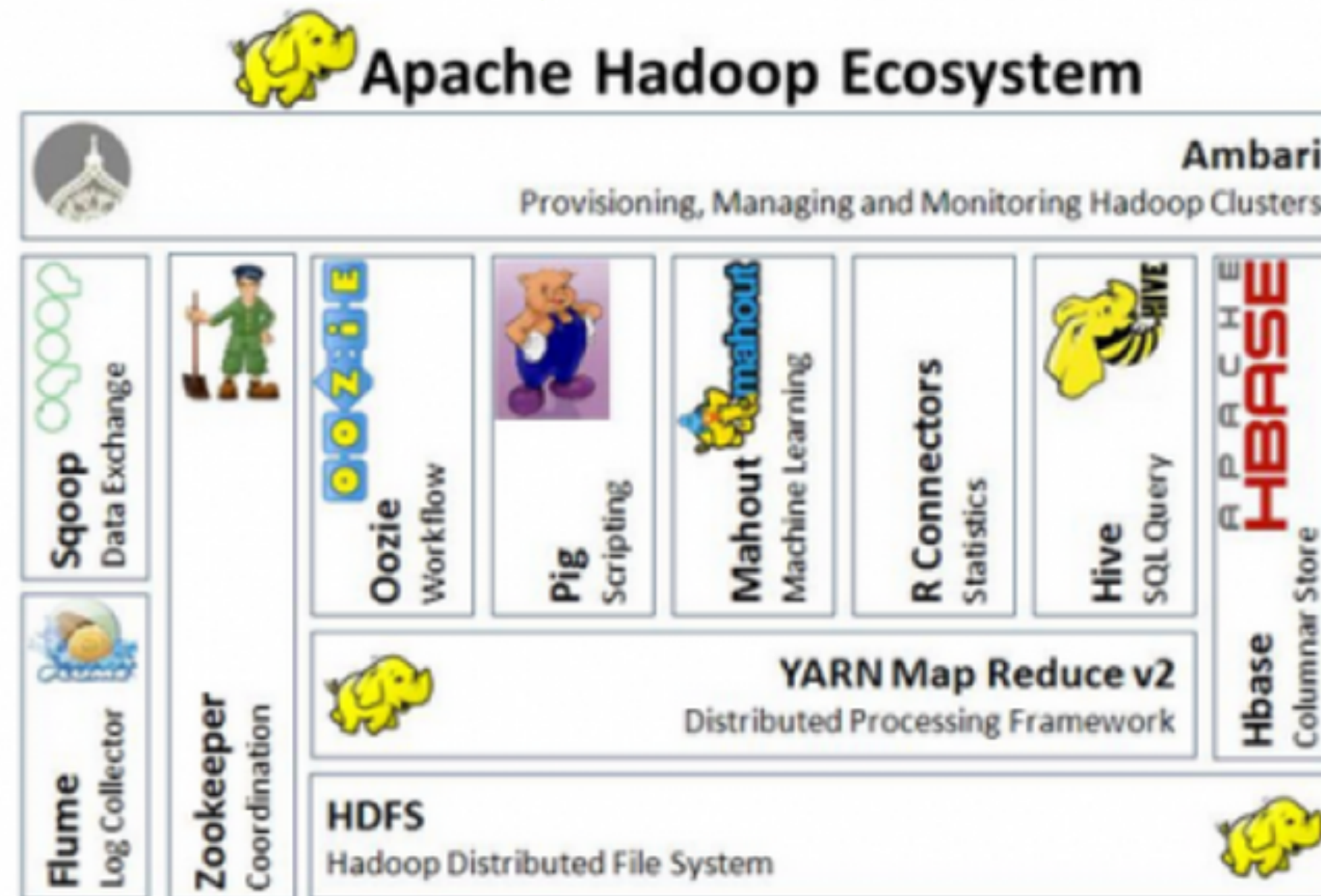
- Estructurados
- No Estructurados
 - Repetitivos (semiestructurados)
 - No Repetitivos

Un minuto en Internet – @LoriLewis, @OfficiallyChadd

Tabla 2. Tipos de Datos en el Paradigma BIG Data

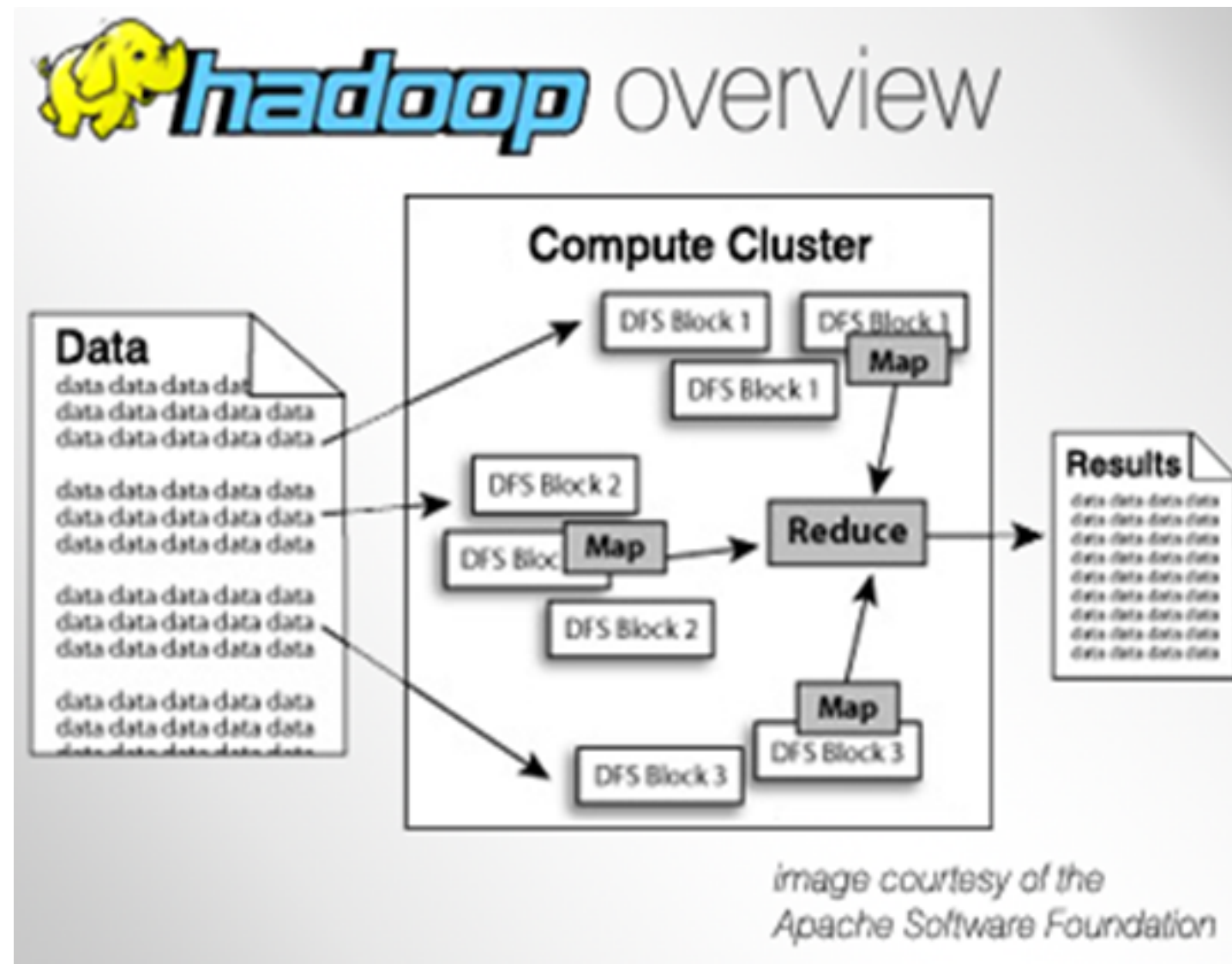
Datos estructurados	Datos semiestructurados	Datos no estructurados
Fichas de clientes Fecha de nacimiento Nombre Dirección Transacciones en un mes Puntos de compra	Correos electrónicos Parte estructurada: destinatario, receptores, tema Parte no estructurada: cuerpo del mensaje	Persona a persona Comunicaciones en las redes sociales Persona a máquina Dispositivos médicos Comercio electrónico Ordenadores, móviles Máquina a máquina Sensores, dispositivos GPS Cámaras de seguridad

Almacenamiento y Manejo de Datos



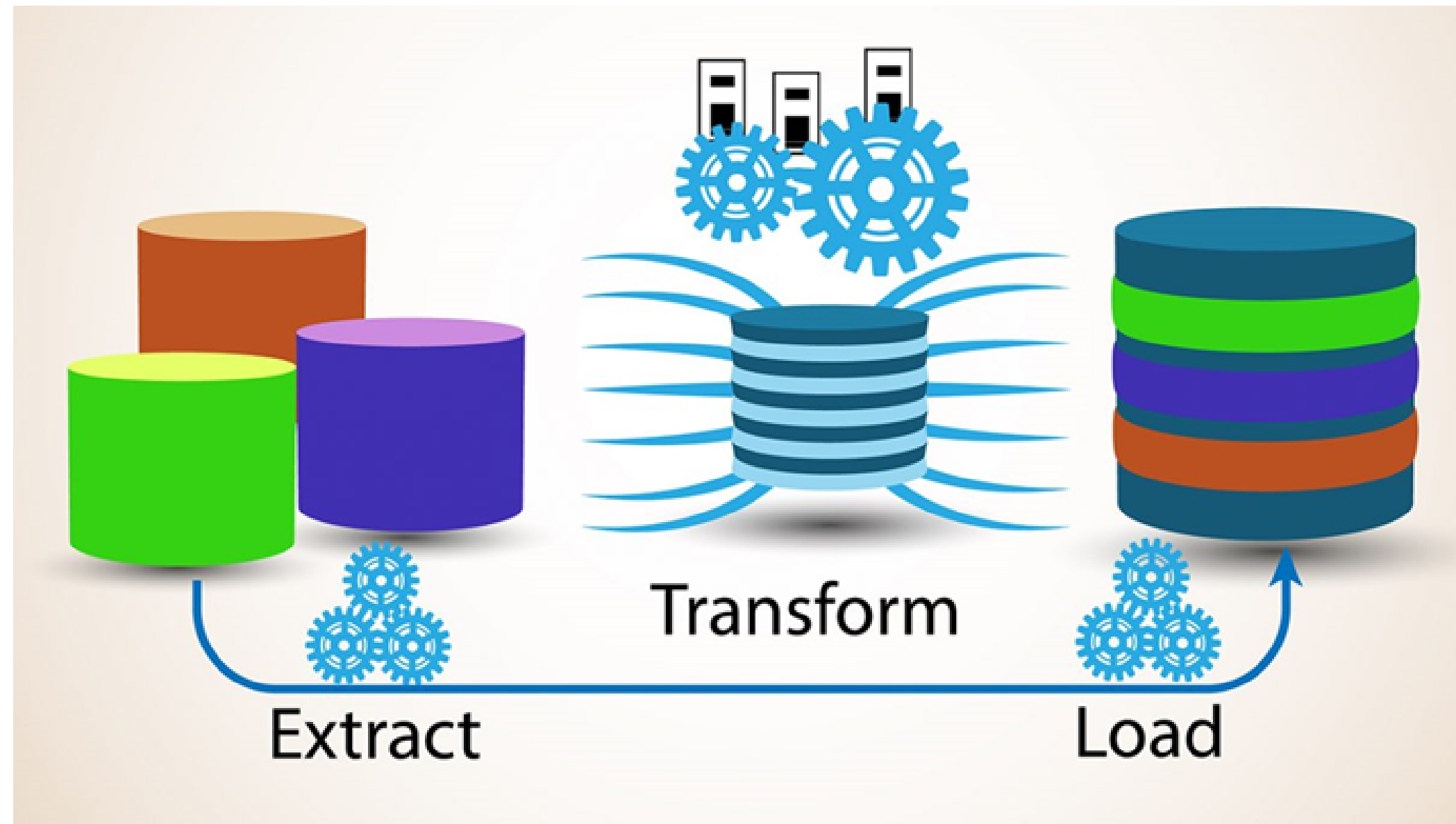
Ecosistema Hadoop – Apache Hadoop, DataPrix

Almacenamiento y Manejo de Datos



HDFS,YARN- Apache Software Foundation

Procesos ETL



ETL - <https://dataintegrationinfo.com/>

Entre el proceso de extracción y carga se puede realizar un proceso de transformación que le de estructura a los datos o saque información relevante de estos

Emails: Extracción

- Recuperar los datos y metadatos del correo.

Emails: Transformación

- Separar cada una de las partes del correo.
Identificar a que corresponde.

Transformación de los datos

¿Cómo...

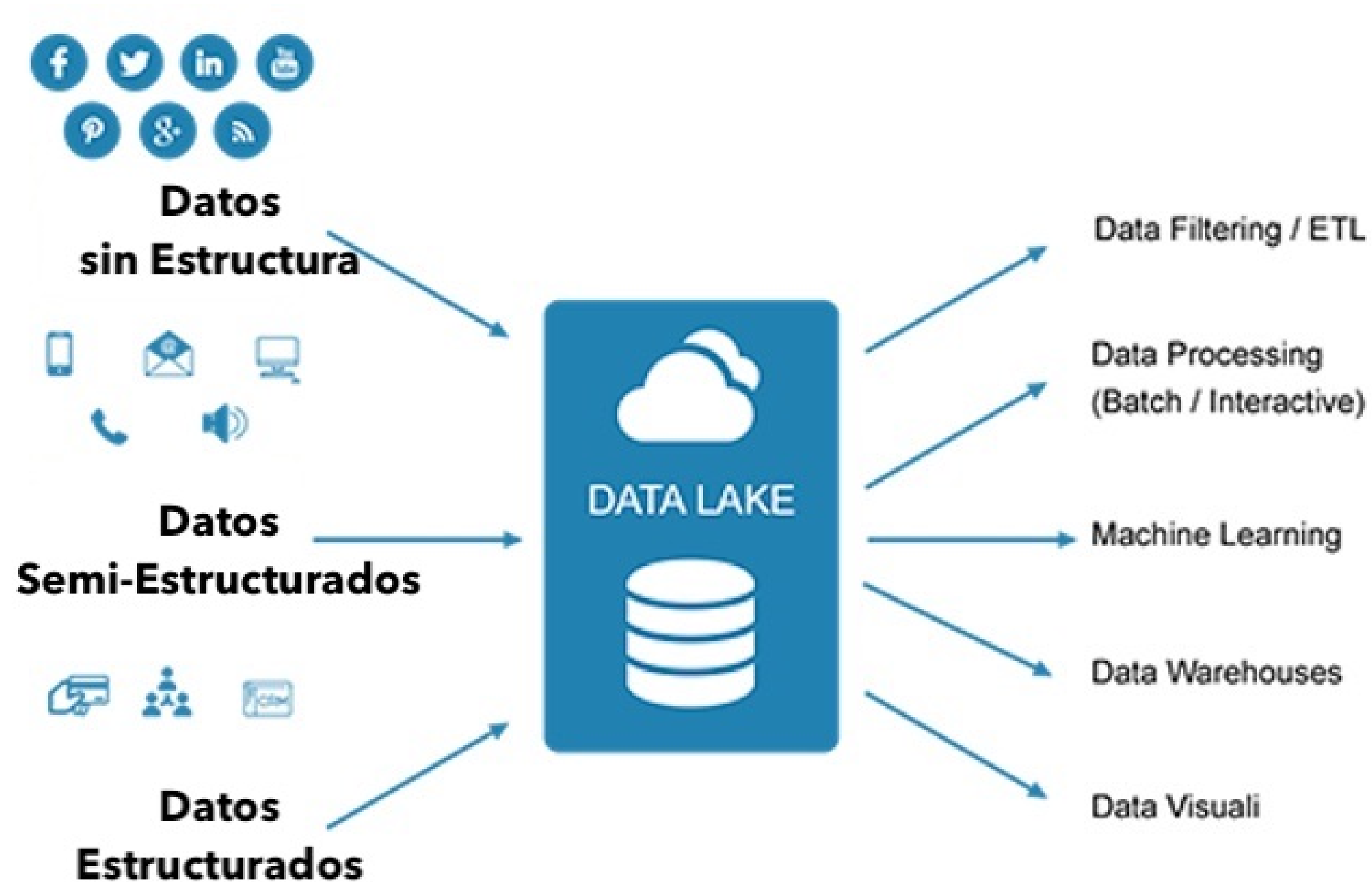
...Extraer datos...

- En una institución educativa?

...Transformar datos...

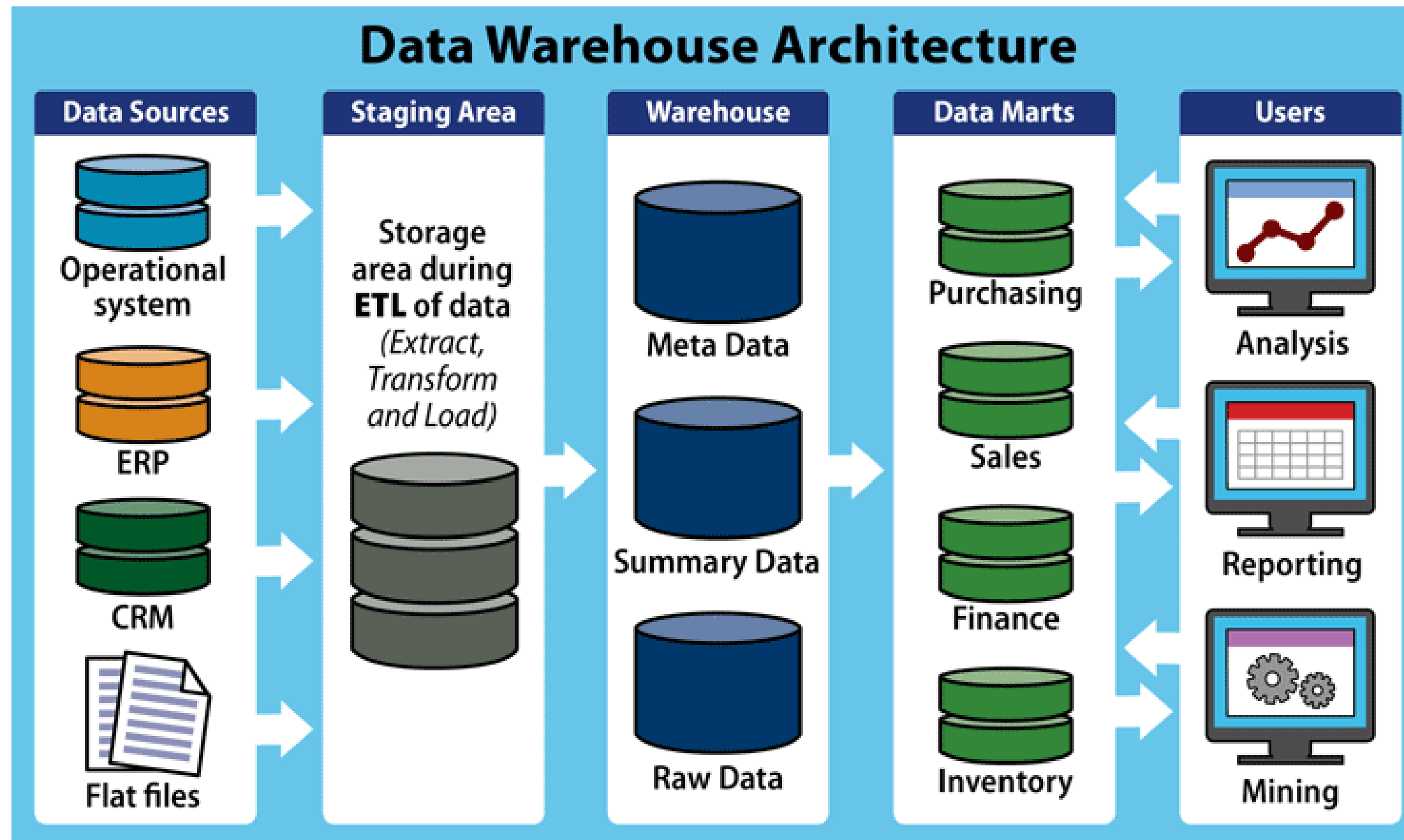
- Para almacenarlos de manera optima?

Proceso ETL - Data Lake



Data Lake -<https://www.tecnologias-informacion.com/>

Data Warehouse



Data Marts

Subestructuras de la data warehouse orientadas a requerimientos específicos.

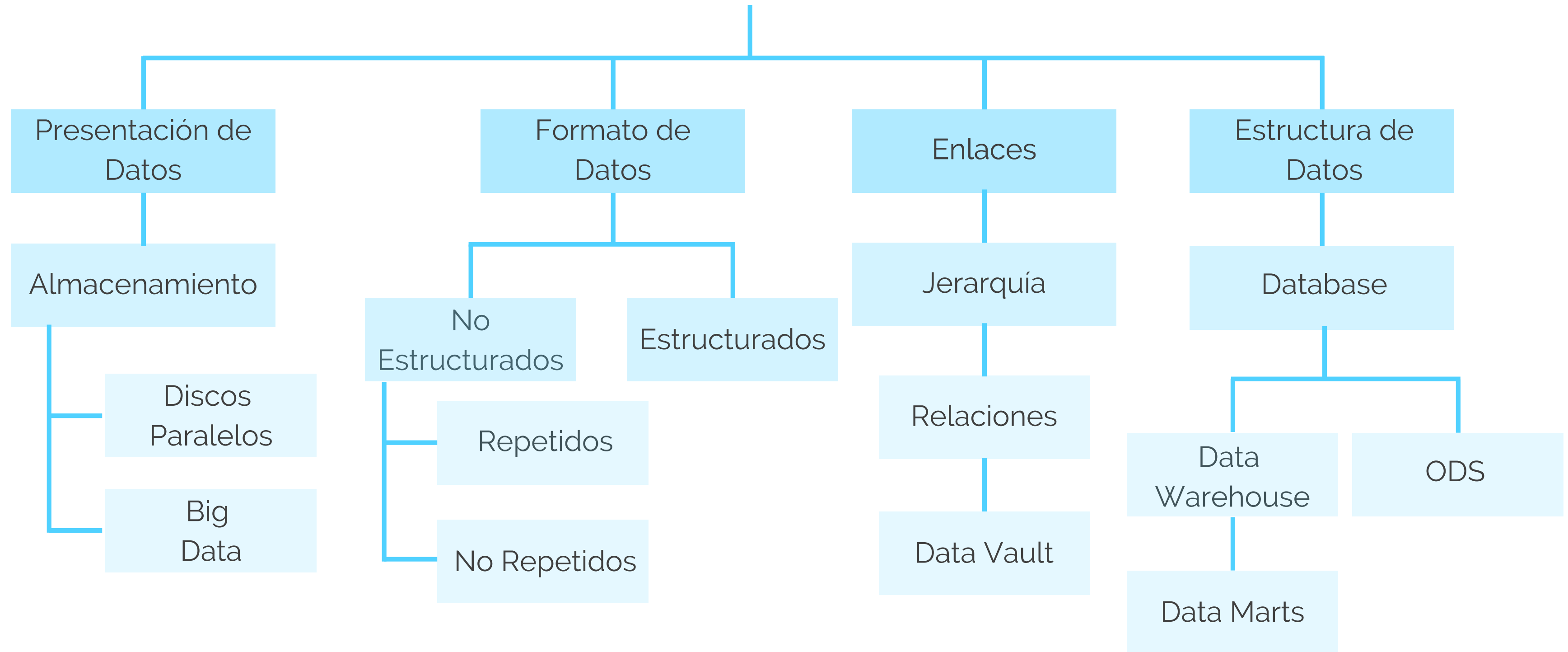
DW Empresa – DM Recursos humanos

- Reportes de rendimiento, horarios, entrenamientos.

DW Empresa – DM Tesorería

- Reportes de nómina.

Data Architecture



Arquitectura de datos - Data architecture: A primer for the data scientist ,Inmon

Necesito...

...un data lake...

- Para procesar tweets?
- Para hacer un repositorio de imágenes?

...data marts...

- Para una universidad?
- Para una gran superficie?

...Hadoop...

- Para los viajes reportados por una aplicación de transporte?
- Para una empresa de importaciones de maquinaria pesada?

¿Preguntas, comentarios?

Actividad

- Enumere cinco fuentes o tipos de datos no estructurados no repetitivos.
- Considere los registros de velocidad de las cámaras de seguridad vial en una ciudad. ¿Son datos estructurados o no estructurados? ¿Son repetitivos o no repetitivos? ¿Se considera Big Data? ¿Valdría la pena, pensando en la estructura gubernamental de una ciudad, construir un data warehouse para trabajar dichos registros?

Actividad

- ¿Sería necesario un proceso de transformación para los datos que se extraen de redes sociales para extraer información?
- Se ha diseñado un data warehouse sobre los datos de los estudiantes de una institución educativa. El objetivo es almacenar bajo un mismo formato los datos de los estudiantes y usarlos para realizar reportes para las diferentes dependencias de la institución. ¿Qué datos considera deberían buscar y almacenarse? ¿Qué tipo de reporte podrían pedir las diferentes dependencias?