Maestría en Finanzas

FI-75301 Macrodatos y Fintech

M.Sc. Walter Jeremías López.

unitec®
LAUREATE INTERNATIONAL UNIVERSITIES®

Generación de conocimiento con Big Data.

Maestría en Finanzas

FI-75301 Macrodatos y Fintech

M.Sc. Walter Jeremías López.

unitec®
LAUREATE INTERNATIONAL UNIVERSITIES®

Objetivos de aprendizaje:

- Describir las principales arquitecturas de Big Data en cuanto a la manera de incorporarlas a sus entornos laborales.
- Explorar las diferentes plataformas disponibles en el mercado para el análisis de Big Data en las empresas.

Maestría en Finanzas

FI-75301 Macrodatos y Fintech

M.Sc. Walter Jeremías López.

unitec®
LAUREATE INTERNATIONAL UNIVERSITIES®

Competencias a desarrollar:

- El alumno conoce la manera en que funcionan las arquitecturas de Big Data para incorporarlas en su trabajo.
- El alumno conoce las plataformas de Big Data disponibles en el mercado para incorporarlas en su empresa.

Maestría en Finanzas

FI-75301 Macrodatos y Fintech.

M.Sc. Walter Jeremías López.

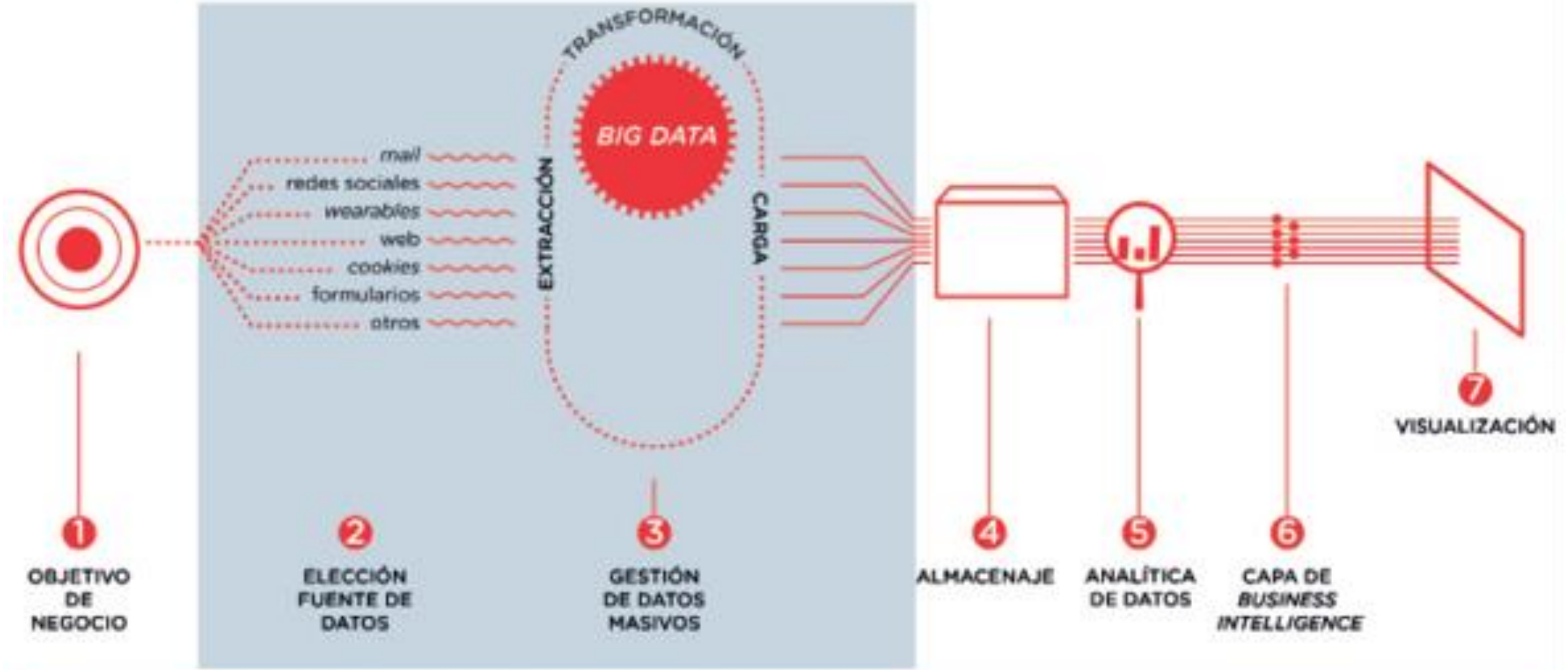
unitec®
LAUREATE INTERNATIONAL UNIVERSITIES®

Agenda:

- Gestión organizativa eficiente mediante la cadena de valor.
- Aplicaciones sectoriales, arquitecturas y almacenes para macrodatos.
- Plataformas, tecnologías y aplicaciones para Big Data.
- Conclusiones.

Cadena de valor

Cómo gestionar eficiente el Big Data mediante la cadena de valor.



FLUJO DE PROCESO DE GESTIÓN DEL *BIG DATA* EN LA EMPRESA

Generación | Adquisición | Almacenamiento | Análisis

Fase 1: Generación

Fases de la cadena de valor de los macrodatos.

1) Los datos se generan de múltiples y diversas fuentes.

2) Según los objetivos estratégicos planteados, se deben elegir las fuentes apropiadas donde se espera obtener los datos que se necesitan para alcanzarlos.

Fase 2: Adquisición

Fases de la cadena de valor de los macrodatos.

1) Recogida de datos – 2 enfoques:
   - Pull.
   - Push.
2) Transmisión.
3) Preprocesamiento:
   - Integración (ETL):
     - Extraer: Extraer los datos.
     - Transformar: con rutinas o scripts.
     - Cargar: a un Datawarehouse.
   - Limieza.
   - Eliminar la redundancia.

Fase 3: Almacenar

Fases de la cadena de valor de los macrodatos.

Los datos se pueden cargar en diferentes repositorios adecuados para datos masivos según su fase, los cuáles pueden ser:

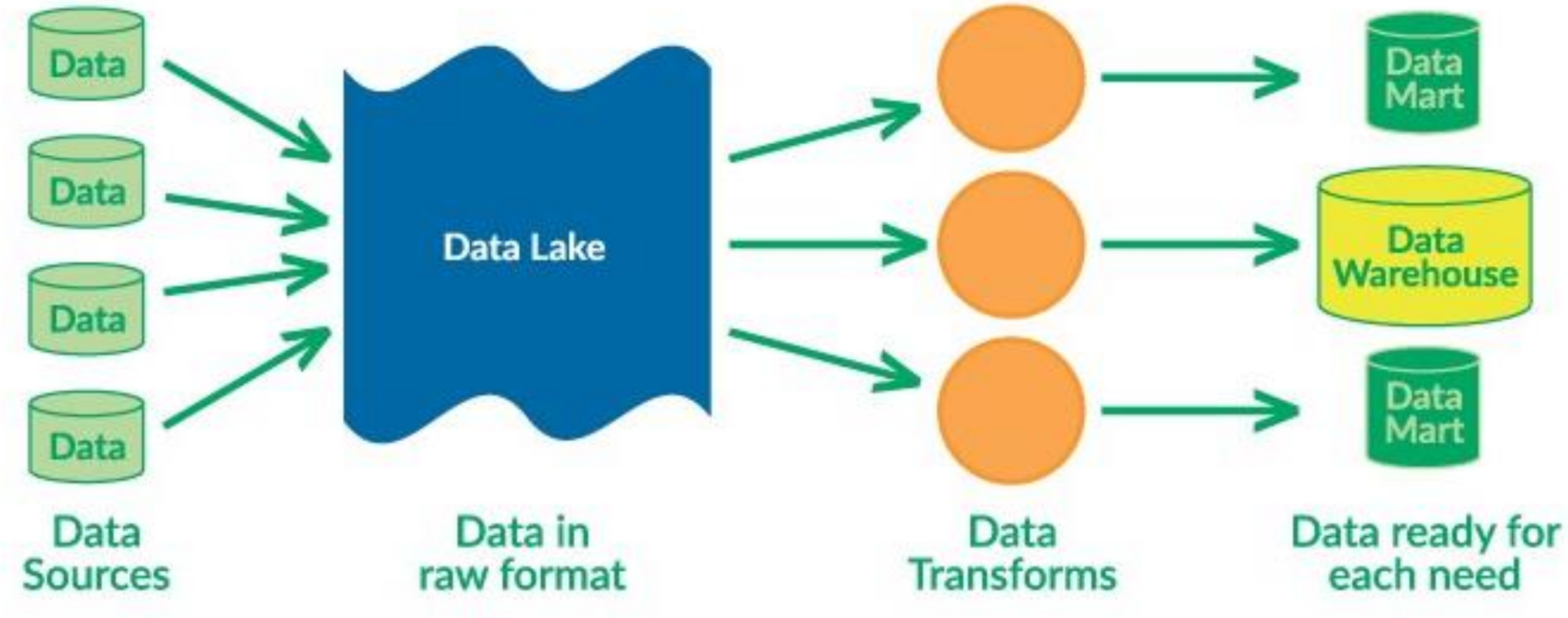
a) Data Lake.

b) Data Warehouse.

c) Data Mart.

Normalmente no están normalizados y son diferentes a una base de datos tradicional.

Esquema de los repositorios de Big Data:



Comparación de repositorios de datos:

	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	Weeks - Months	$$$$$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	Hours - Days	$$$$$		
Data Mart	Line of Business Specific Reporting & Analytics	Minutes - Hours	$$$$$		

Comparación de repositorios de datos:

Características	Base de datos relacional	Data Warehouse	Data Lake	Data Mart	Operational Data Store
Tipos de datos.	Estructurados, numéricos, texto, fechas, organizados en un modelo relacional.	Relacional, datos de sistemas transaccionales, bases operacionales y aplicaciones.	Estructurados y no estructurados de sensores, sitios web, apps móviles y de negocios, etc.	Datos relacionales, subconjuntos para funciones específicas.	Datos transaccioneales de distintas bases de datos de la empresa.
Propósito.	Procesamiento de transacciones.	Almacenar datos para inteligencia de negocios, reportes por lotes y visualización de datos.	Análisis de Big Data, machine learning, análisis predictivo y descubrimiento de datos.	Datos usados para análisis por una comunidad específica de usuarios.	Ingestar, integrar, guardar y prerar datos para operaciones o análisis, alimenta el data warehouse.

Comparación de repositorios de datos:

Características	Base de datos relacional	Data Warehouse	Data Lake	Data Mart	Operational Data Store
Captura de datos.	Desde una sola fuente, como un TPS.	Desde múltiples fuentes relacionales.	Desde múltiples fuentes y varios tipos de datos.	Típicamente del data warehouse, pero puede venir de sistemas operacionales o fuentes externas	Múltiples bases de datos de aplicaciones empresariales y fuentes.
Normalización de datos.	Esquemas estáticos y normalizados.	Esquemas desnormalizados, sobre escritura.	Desnormalizado, esquema sobre lectura.	Normalizado o desnormalizado.	Desnormalizado.

Comparación de repositorios de datos:

Características	Base de datos relacional	Data Warehouse	Data Lake	Data Mart	Operational Data Store
Beneficios	Provee datos consistentes para aplicaciones críticas de negocios.	Datos históricos de diferentes fuentes en un solo lugar para accesibilidad.	Datos en su formato nativo de diversas fuentes. Flexibilidad para análisis y desarrollo de modelos.	Fácil, rápido acceso a datos relevantes para aplicaciones específicos y tipos de usuarios.	Consultas rápidas en tiempo real (o casi) para reportes y decisiones operativas.
Calidad de los datos.	Organizados y consistentes.	Datos curados, centralizados y listos para usarlos en BI y análisis.	Datos crudos que pueden tener errores y redundancia para su uso.	Datos con alto nivel de curación.	Datos limpios y conformes, pero no tan consistentes como en data warehouse.

Fase 4: Análisis

Fases de la cadena de valor de los macrodatos.

Engloba un conjunto de procedimientos y modelos estadísticos. Existen 3 líneas de investigación:

a) Diseño de tecnologías y SW para análisis según el tipo de datos.

b) Diseño de métodos de análisis.

c) Visualización.

Algunos métodos de análisis son:

Métodos de extracción y análisis de datos:

Tipos de datos	Extracción	Métodos
Estructurados.	Detección de anomalías.	Algoritmos.
	Descubrimiento de estructuras mediante la explotación de características, tiempos y espacio.	Minería de datos (Data Mining).
Desestructurados: datos en texto.	Datos de texto	Sistemas de minería de texto basados en: expresiones, procesamiento de lenguaje natural: resumen de texto, clasificación, agrupación, minería de opinión, etc:

Métodos de extracción y análisis de datos:

Tipos de datos	Extracción	Métodos
Datos de Web.	Texto, multimedia, foros.	Minería de contenido web, análisis multimedia, minería de hipertexto.
	Estructuras de los enlaces dentro de una web entre varias web.	Minería de estructura web.
	Logs de servidores y proxies, historiales de navegación, perfiles de usuarios, sesiones de usuario, preguntas, bookmarks, clics, etc.	Minería de uso web: para cualquier dato generado en la interacción con la web.

Métodos de extracción y análisis de datos:

Tipos de datos	Extracción	Métodos
Datos multimedia	Video, música, imágenes.	Resumen, anotación, indexación, recuperación, detección de acontecimientos.
Redes sociales	Datos masivos enlazables, datos de contenidos	Análisis de redes sociales. Análisis de la estructura basada en enlaces. Análisis basado en contenidos.

Análisis estadístico

Tipos de estadística y análisis a realizar.

La estadística se divide en 2: descriptiva e inferencia, puede ser paramétrica o no paramétrica. En su mayoría se basa en la teoría de la probabilidad, creando modelos estocásticos.

Los tipos de análisis pueden ser:

a) Univariante.

b) Bivariante.

c) Multivariante.

Minería de Datos

Modelos predictivos, de clasificación o segmentación.

Engloba un conjunto de metodologías, procesos de modelización y técnicas matemáticas para analizar datos de distintas fuentes con el objetivo de extraer información previamente desconocida.

Analiza estructuras de datos de las que emergen patrones de comportamiento y tendencias.

Minería de Datos

Técnicas de data mining empleadas con datos masivos.

Utiliza técnicas basadas en estadística e Inteligencia Artificial:

Aprendizaje de Máquina (Machine Learning).

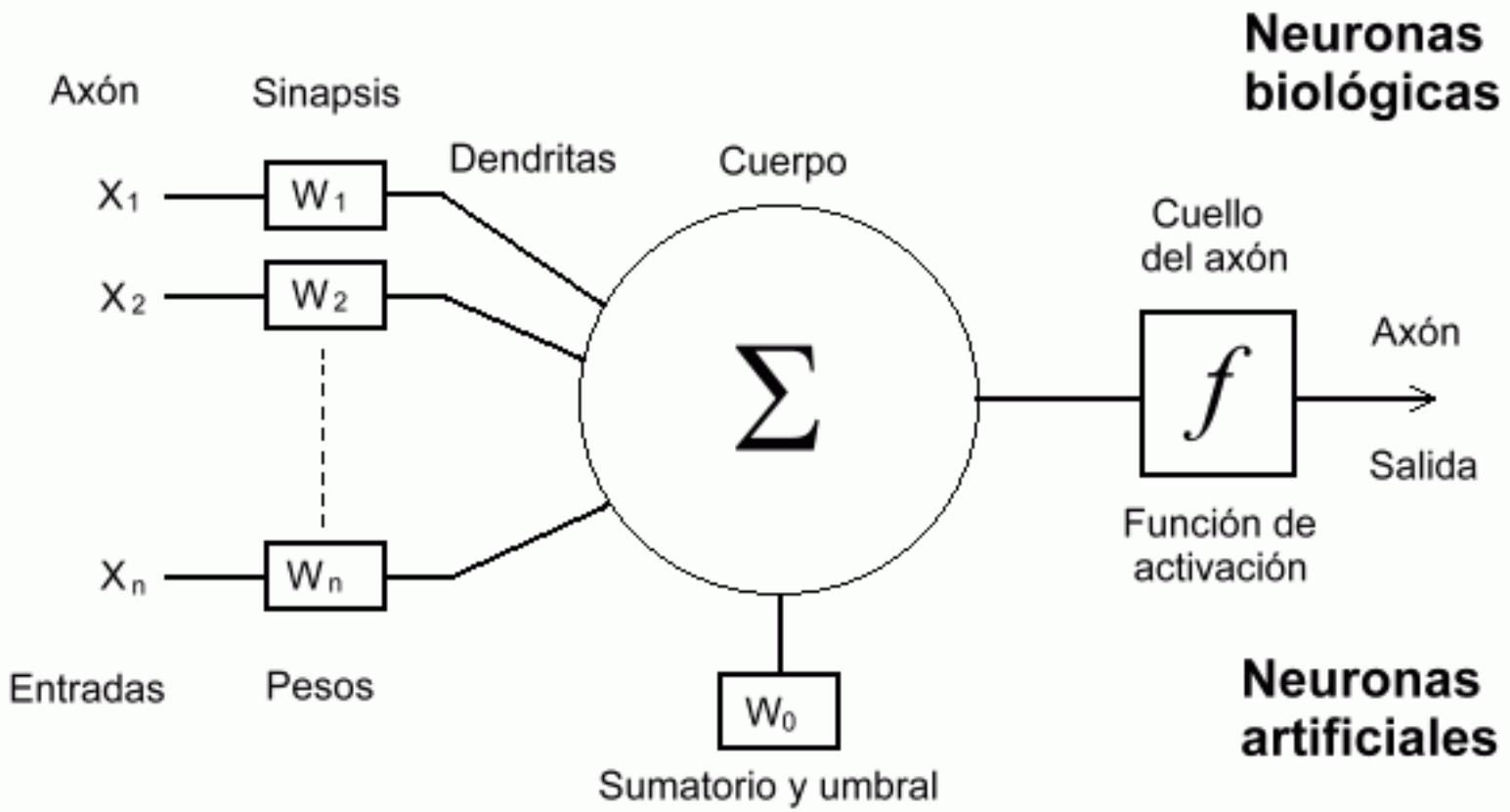
- Algoritmos supervisados.
- Algoritmos no supervisados.

Los algoritmos más representativos son: regresiones, árboles de decisión, redes neuronales, clustering, segmentación y reglas de asociación.

Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

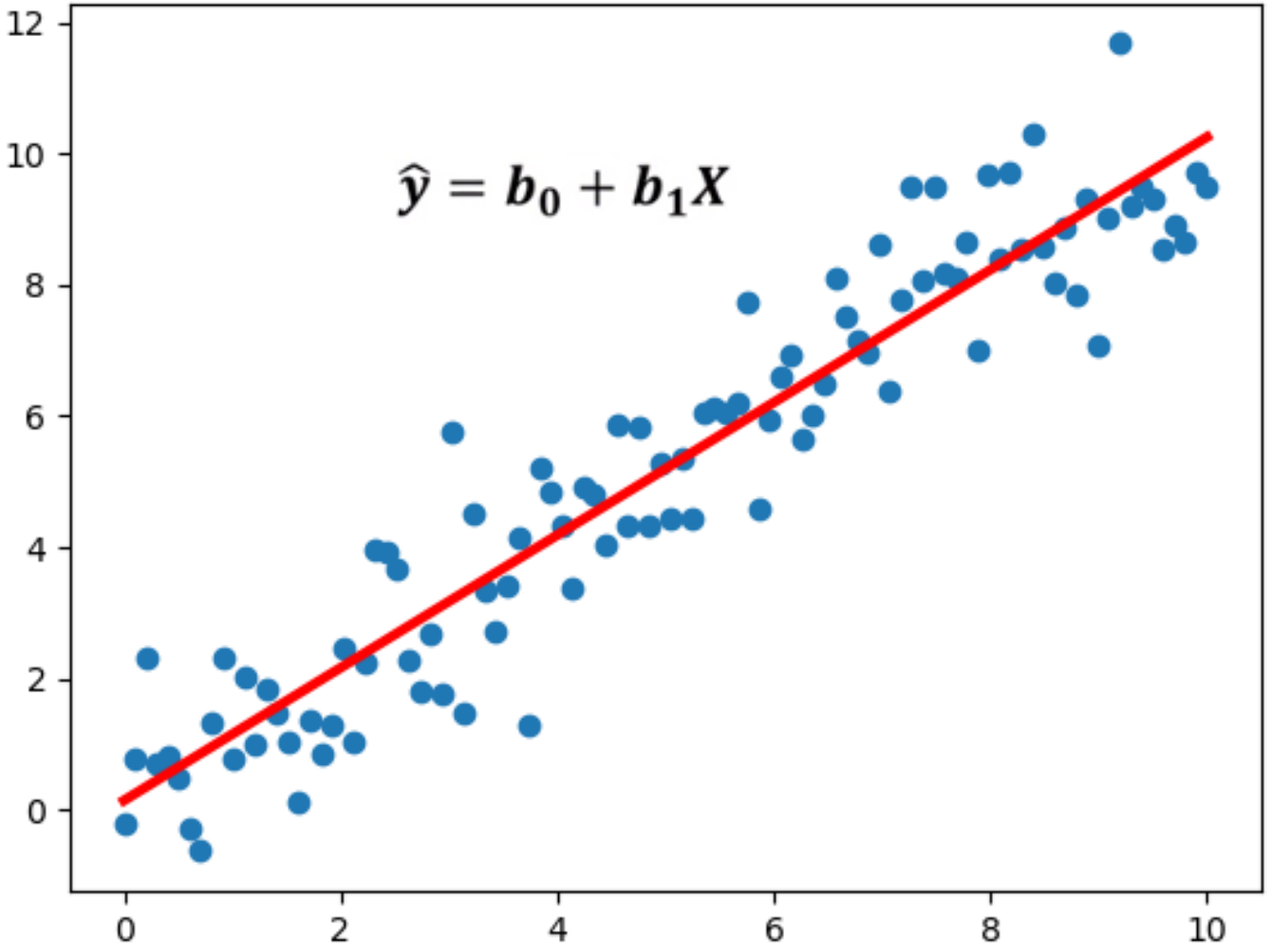
Redes Neuronales (ANN – Artificial Neural Networks):



Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

Regresión lineal simple:

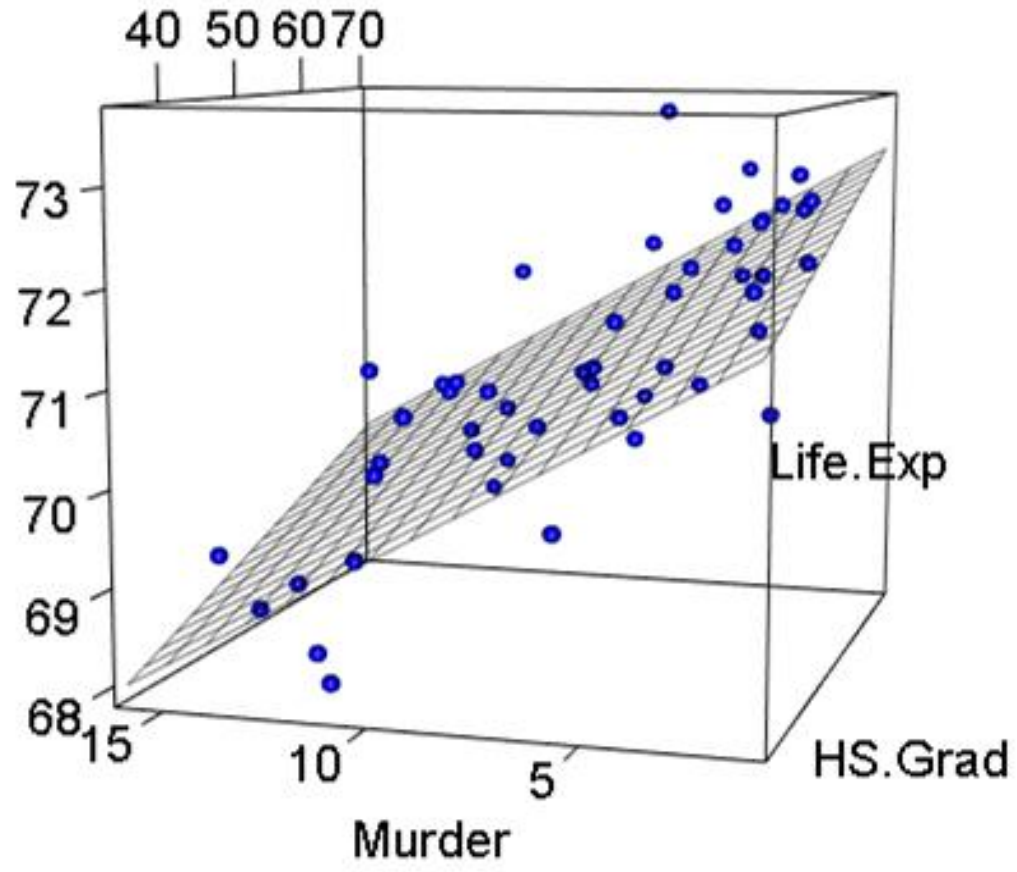


Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

Regresión lineal múltiple:

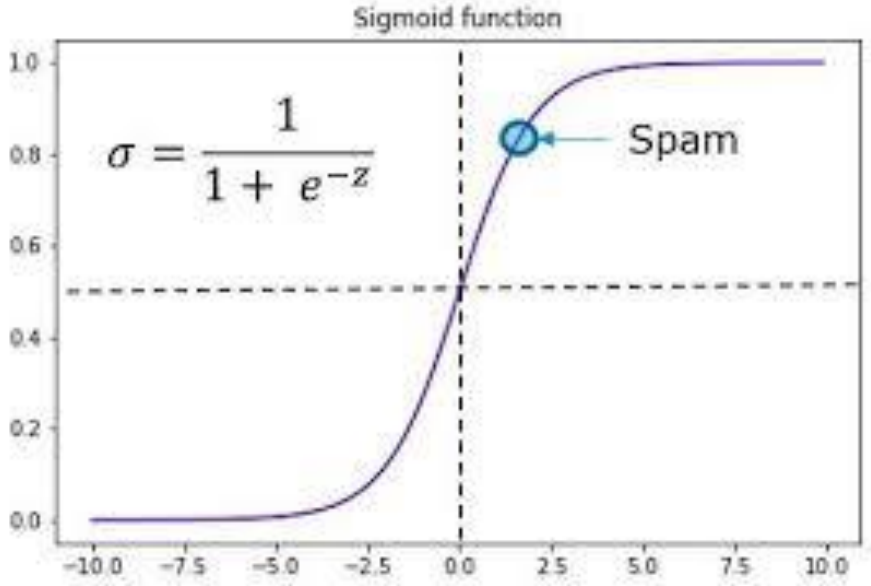
Multiple Regression → $y = m_1x_1 + m_2x_2 + b$



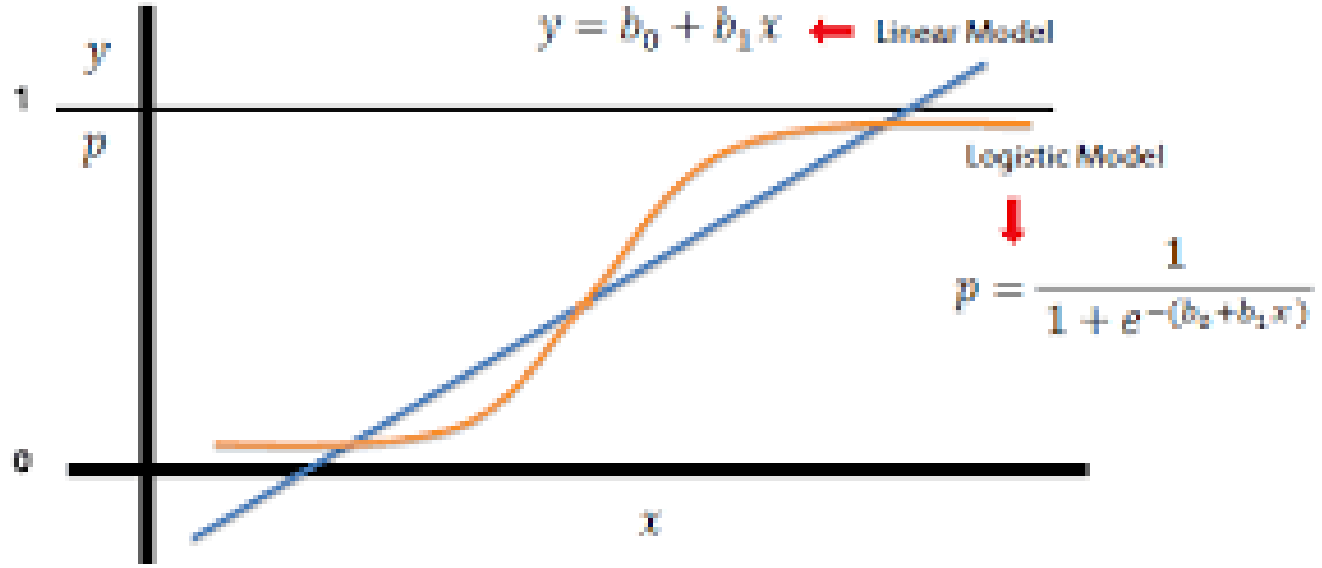
Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

Regresión logística:



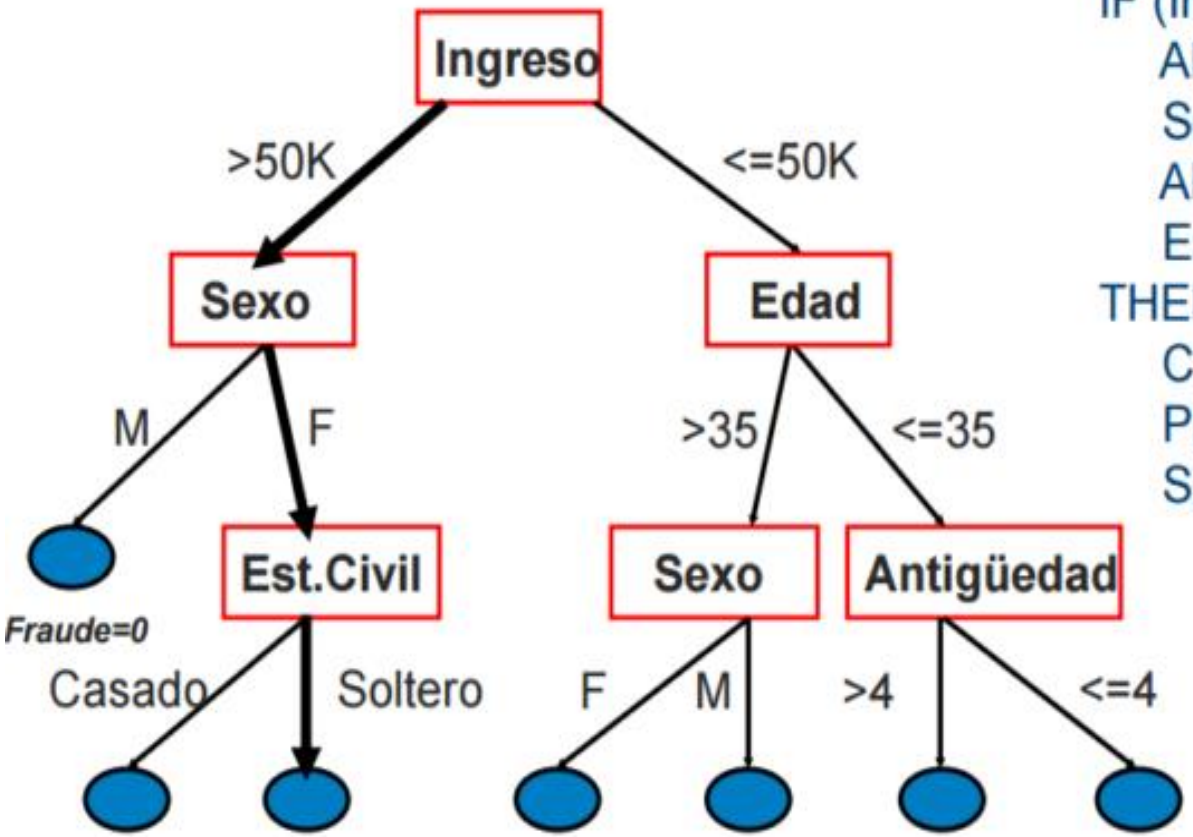
Comparación:



Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

Árboles de decisión:

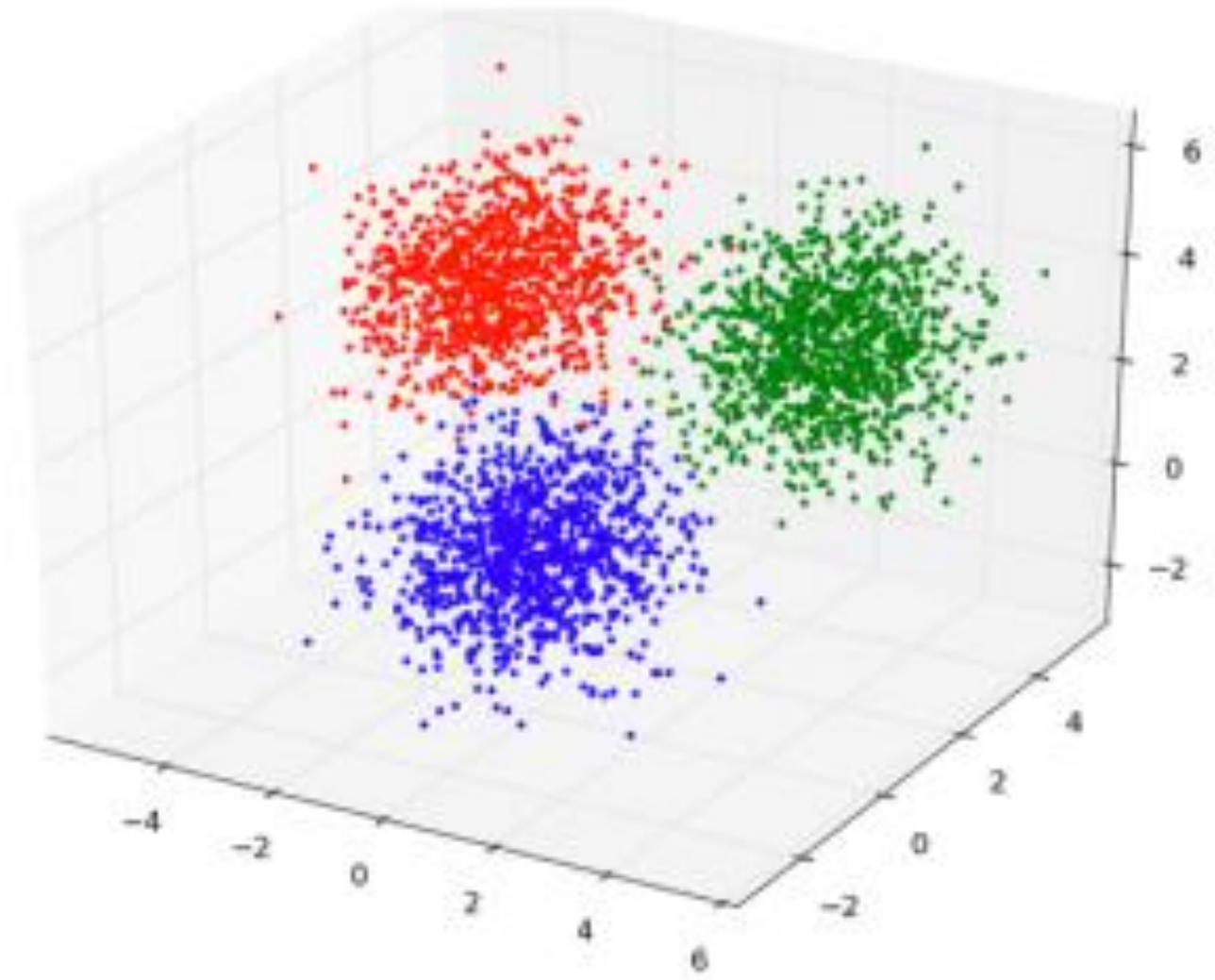



```
IF (Ingreso >50K
    AND
    Sexo = F
    AND
    EstCivil = Soltero)
THEN
    Churn = 1
    Prob = .77
    Soporte = 250
```

Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

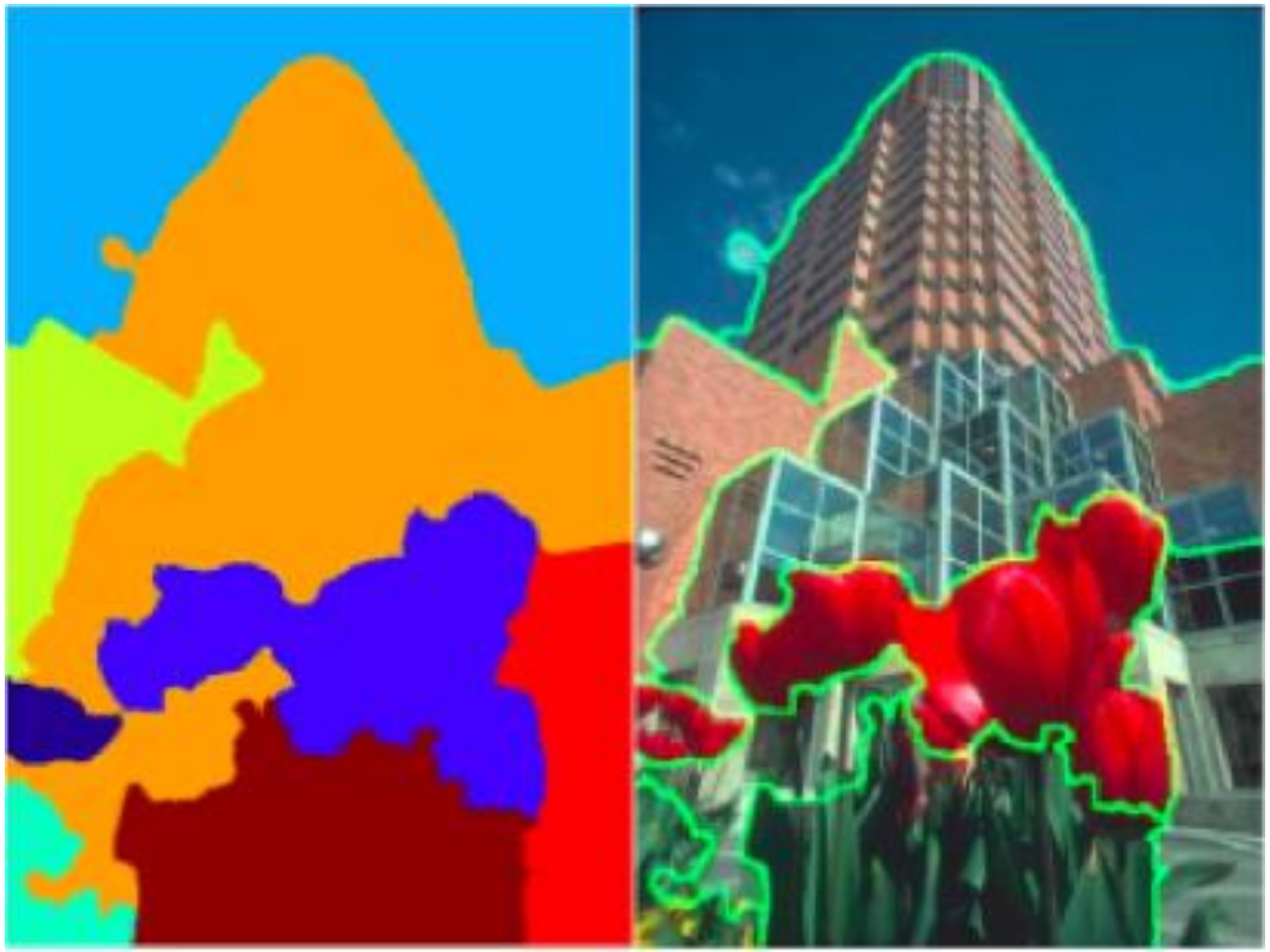
Agrupamiento (clustering):



Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

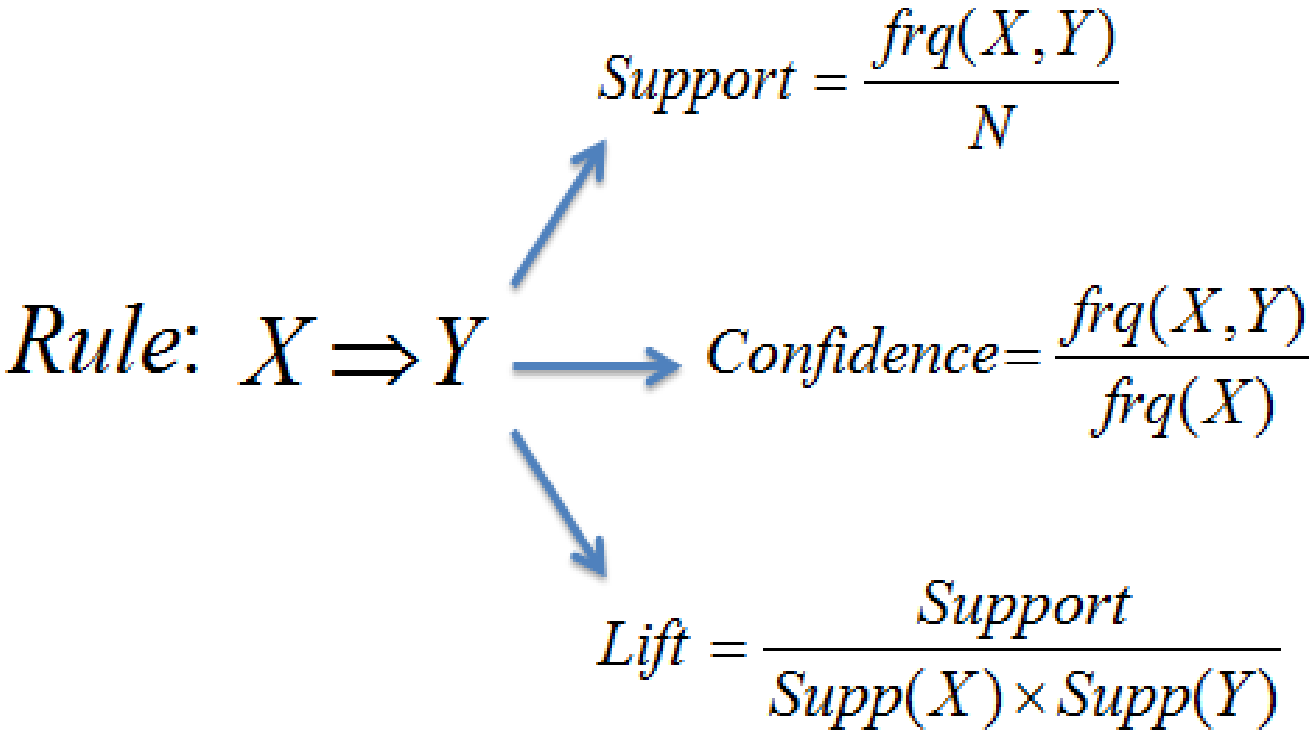
Segmentación:



Minería de Datos

Algoritmos de data mining más utilizados con datos masivos.

Reglas de asociación:



Minería Web

Definición y tipos de Web Mining.

Extrae información de los enlaces, contenidos de páginas web y logs de los recursos de Internet:

1. Estructura web (Web structure mining).
2. Contenido web (Web content mining).
3. Uso de la web (Web usage mining).

Minería de texto

Extraer conocimiento de texto no estructurado.

Se fundamenta en lingüística computacional, procesamiento de texto y aprendizaje automático:

1. Usan diccionarios para eliminar conjunciones, preposiciones, lematización, signos, tags, etc.
2. Dispone de modelos de representación de documentos como modelo vectorial de Salton (esquema tf.idf).
3. Usan glosarios, tesauros, taxonomías y ontologías para relaciones semánticas.

Minería de texto

Otras técnicas de minería de texto.

1. Opinion mining y Sentiment Analysis: valorar opiniones positivas o negativas del público.
2. Social Network Analysis (SNA): estudia las interacciones y relaciones en redes sociales (centralidad, proximidad, intermediación, visibilidad y exposición, notoriedad, influencia, engagement, popularidad).
3. Reputation Management.

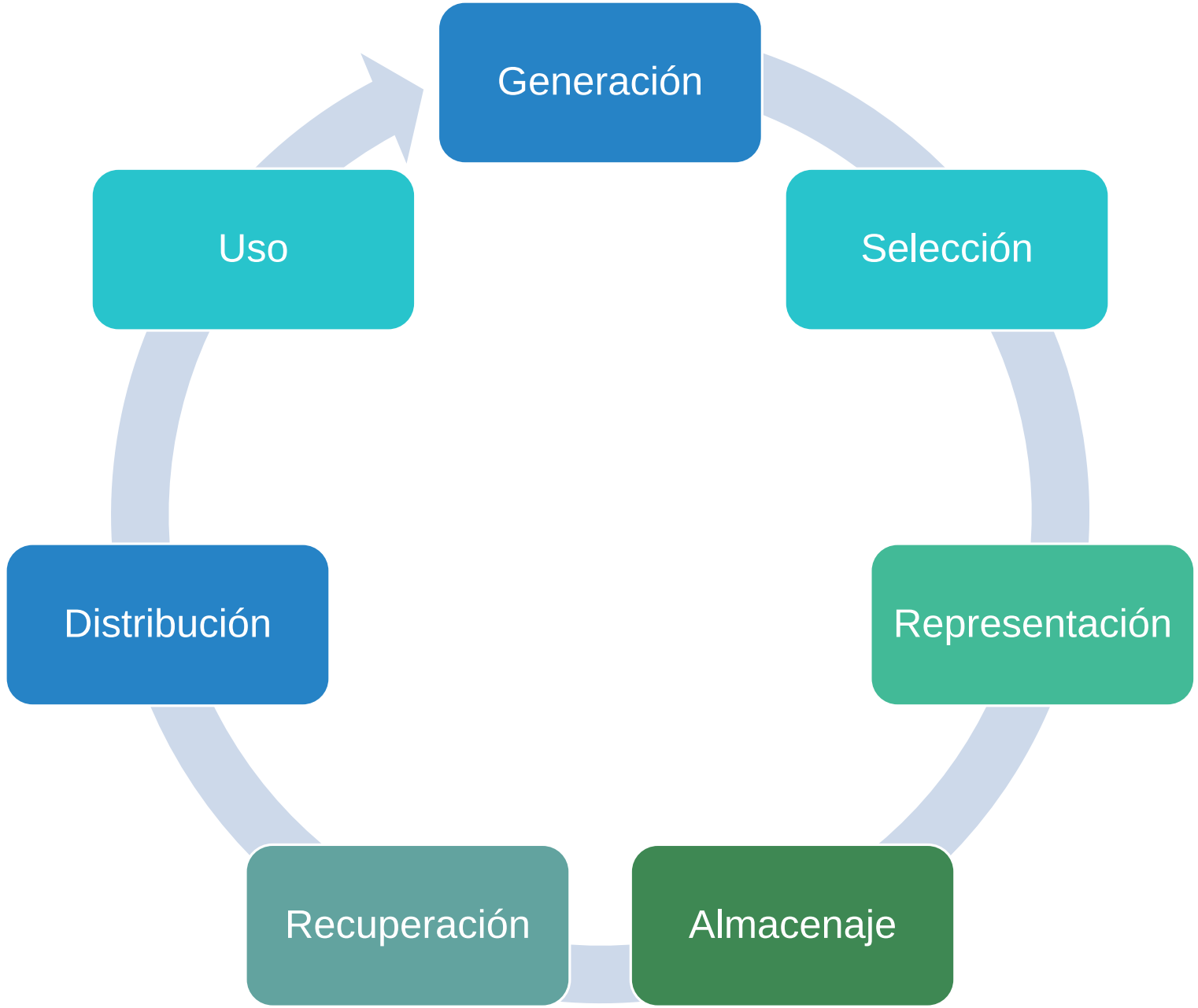
Data Governance

La gobernanza de los datos y de la información.

1. En la cadena de valor es importante velar por la calidad de los datos: La norma técnica es la ISO 8000.
2. Los roles identificados son: gestor de datos, administrador de dato y técnico de datos.
3. La gobernanza de la información:
   A. Gestión y cultura organizativa.
   B. Control y uso de la información de forma holística y transversal.

Data Governance

Ciclo de vida de la gestión de la información.

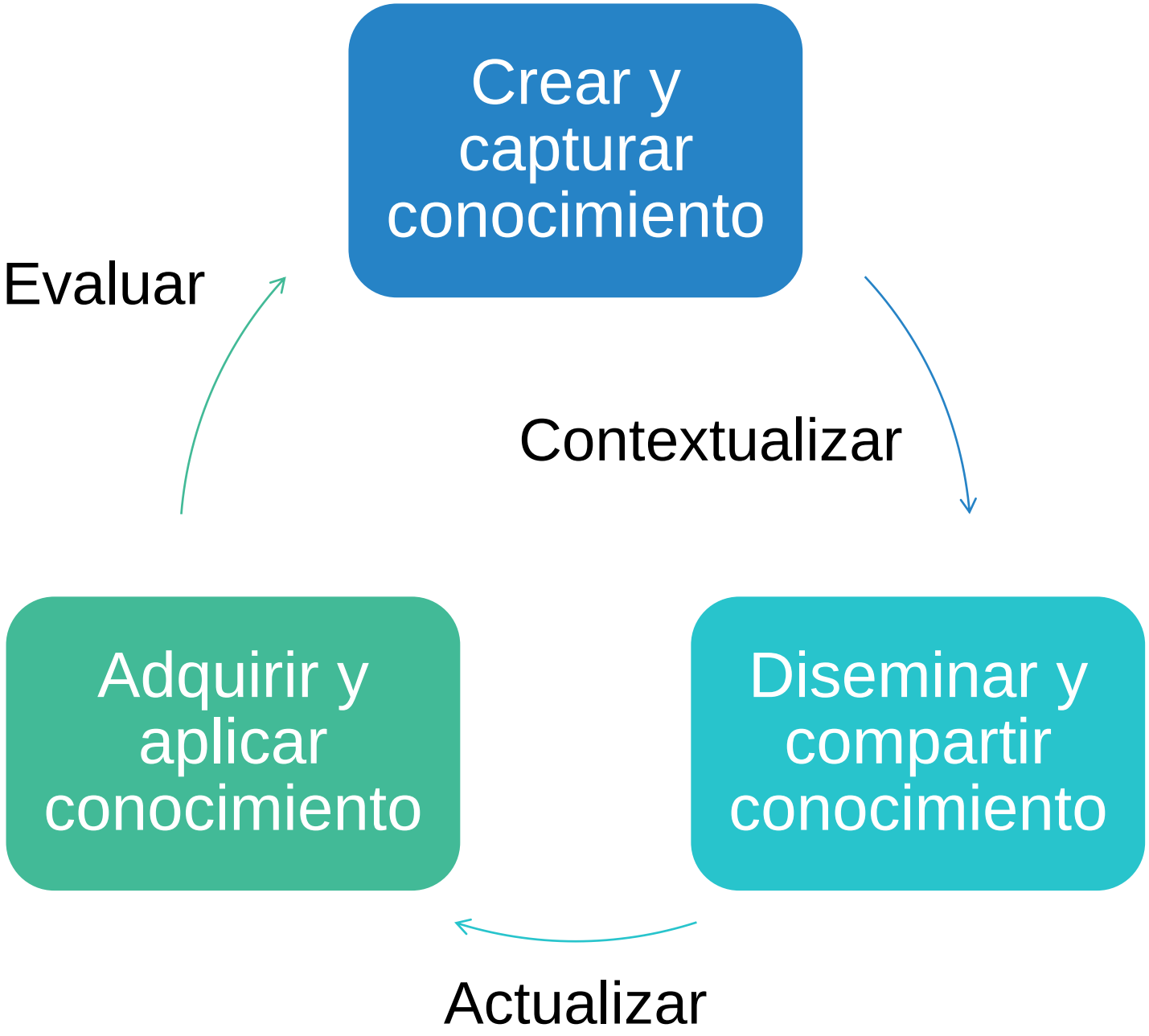


Data Governance

Ciclo de vida de la inteligencia competitiva.

Data Governance

Ciclo de la gestión del conocimiento.



Gestión organizativa

Generación de conocimiento para la acción.

El análisis se puede hacer desde 3 visiones:

Gestión organizativa

Aplicaciones sectoriales del Big Data o por funciones de negocio.

1. Alta gerencia: BSC con Big Data.
2. Mercadeo: clientes, fuerza de ventas.
3. Producción: cadena de suministro, líneas de producción.
4. Contabilidad, auditoría y finanzas: prevención de fraude y lavado, análisis de transacciones, riesgos.

Vean Open Data: ODI

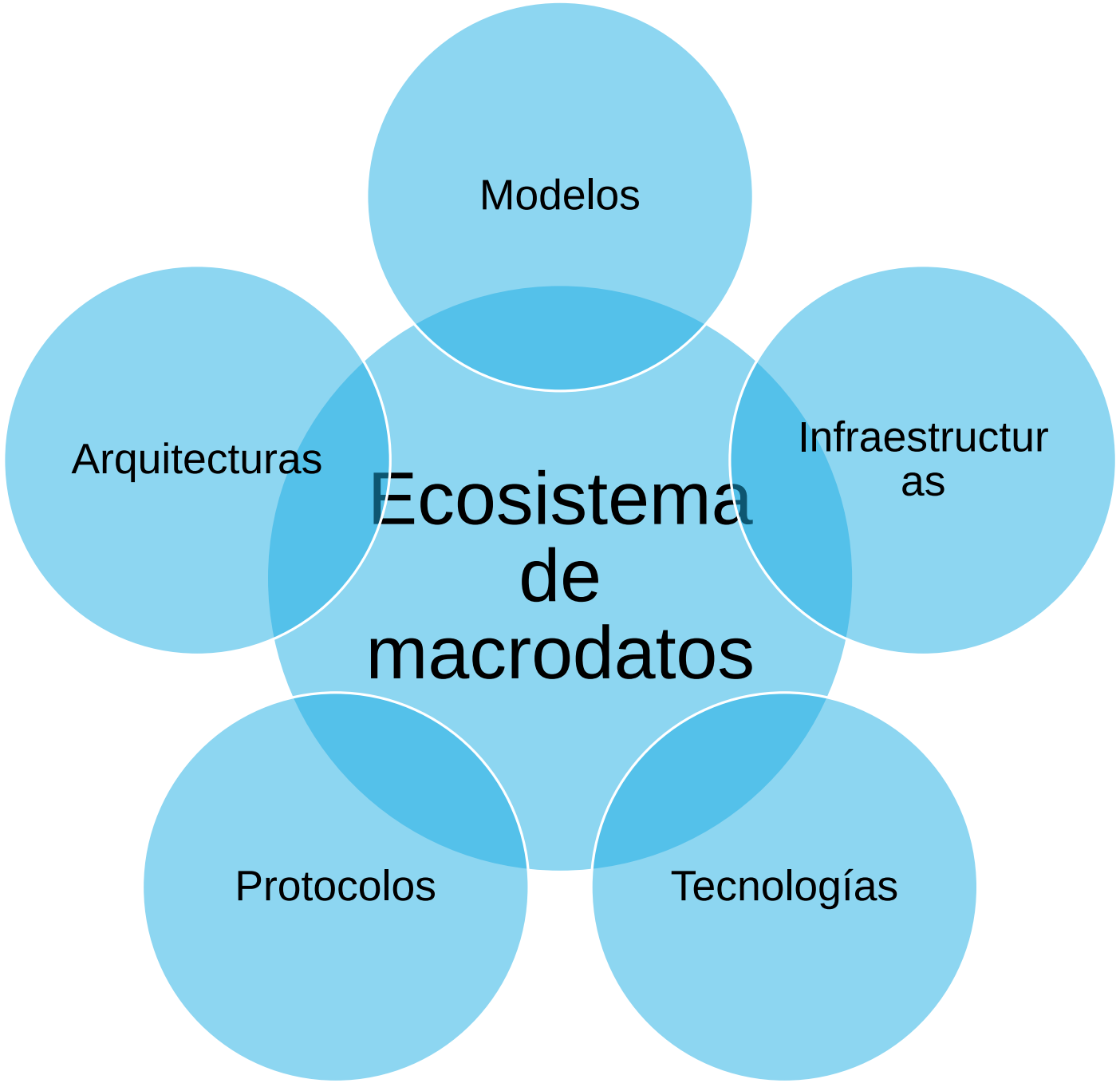
Ejemplos: data.gov data.gov.uk

Gestión organizativa

Los 5 retos de la gestión de datos masivos en la empresa.

Gestión de Macrodatos

Herramientas para la gestión de datos masivos en la empresa.



Gestión de Macrodatos

Sistemas de archivo para datos masivos.

Deben proporcionar rendimiento de R/W, acceso a datos simultáneos, creación de FS bajo demanda y sincronizar archivos:

1. Acceso distribuido y transparencia en la localización.
2. Gestión de fallos (Fault Tolerance).
3. Heterogeneidad.
4. Distribución definida de datos.
5. Tolerancia a la partición de la red.

Gestión de Macrodatos

Tecnologías de bases de datos para datos masivos.

No son relacionales ni utilizan SQL, tienen las siguientes cualidades:

1. No siguen el esquema E-R (Entidad – Relación). Por lo tanto carecen de estructura prefijada en tablas y relaciones.
2. Usan Lenguaje NoSQL, que significa Not only SQL.
3. Implementaciones BigTable y orientadas a grafos.

Gestión de Macrodatos

Modelos de programación para datos masivos.

Los Macrodatos se almacenan en cientos o miles de servidores, que operan con modelos de programación paralelos (PPM):

1. Los modelos paralelos tradicionales como MPI u OpenMP pueden ser inadecuados para gran escala.
2. Los nuevos modelos son: MapReduce, Dyrad, Ajo-Pairs y Pregel.

Gestión de Macrodatos

Modelos de programación para datos masivos.

MapReduce:

Modelo de programación creado por aplicaciones que deben grandes cantidades de datos de forma paralela, dividiéndolos en grupos para procesarlos distribuidos en diferente HW, para luego combinar el resultado.

Permite varios lenguajes: Java, Ruby, Python y C++.

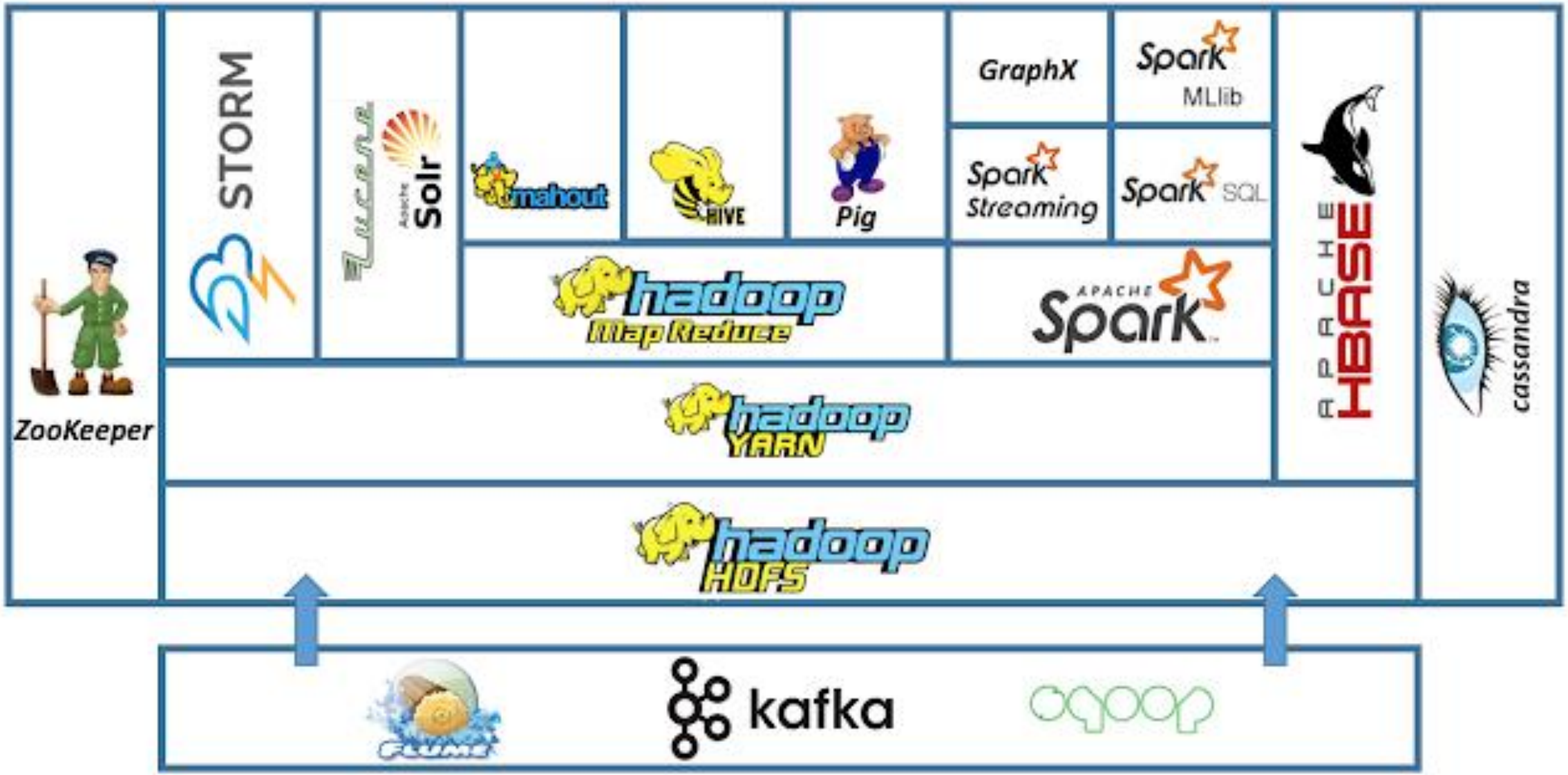
Gestión de Macrodatos

Modelos de programación para datos masivos

Es una estructura de SW Open Source para almacenar datos y ejecutar aplicaciones en clústeres de HW.

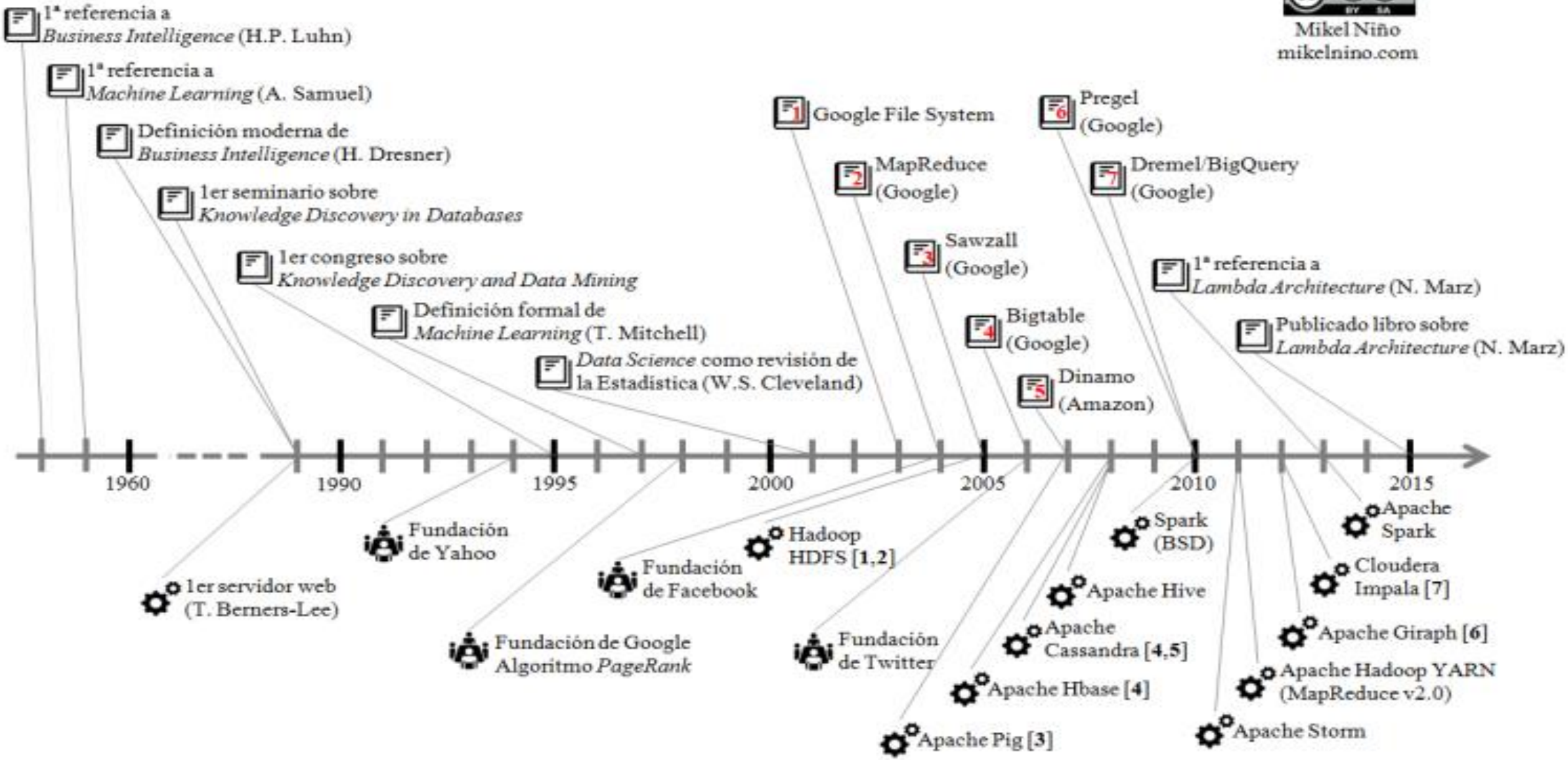
Surge de las ideas de datos distribuidos de Google, así como del proyecto Nutch dividido en dos partes: el rastreador web que se mantuvo como tal y el motor de cómputo y procesamiento distribuido se convirtió en Hadoop.

Ecosistema Hadoop:



CRONOLOGÍA DE ANTECEDENTES, ORIGEN Y DESARROLLO DEL BIG DATA

Mikel Niño
mikelnino.com



v1.0 (150924)

La cadena de valor para la gestión de macrodatos consta de 4 fases principales: Generación, adquisición, almacenamiento y análisis. Las cuales a su vez tienen subfases o subprocesos para gestionar la calidad de los datos.

Conclusiones

Para obtener datos, se pueden extraer automáticamente mediante crawlers o enviar bajo demanda (pull o push).

El preprocesamiento se hace con el método ETL (Extract, Transform, Load).

Conclusiones

Los datos procesados se deben limpiar, eliminar la redundancia y luego almacenar.

Los tipos de repositorios para guardarlos pueden ser: Data Lakes, Data Warehouse o Data Mart.

Conclusiones

Para analizar y visualizar datos se usan diferentes técnicas estadísticas, minería de datos y algoritmos de aprendizaje automático lo cual se conoce como ciencia de datos.

Las principal tecnología para gestión de macrodatos son MapReduce y el ecosistema Hadoop.

Conclusiones

Maestría en Finanzas

FI-75301 Macrodatos y Fintech.

M.Sc. Walter Jeremías López.

unitec®
LAUREATE INTERNATIONAL UNIVERSITIES®

¡Gracias por su atención!

¿Preguntas o comentarios?