



# MINERIA DE DATOS

## Arquitectura de Minería de Datos

**M.C. José C Roberto Olvera López**  
Data Science Consultant

[jroberto.olveral@gmail.com](mailto:jroberto.olveral@gmail.com)





# Arquitectura de Minería de Datos

## INDICE

- Repositorios
- Servidores de Datos
- Base de datos de conocimiento
- Proceso de Arquitectura de Minería de Datos
- Evaluación

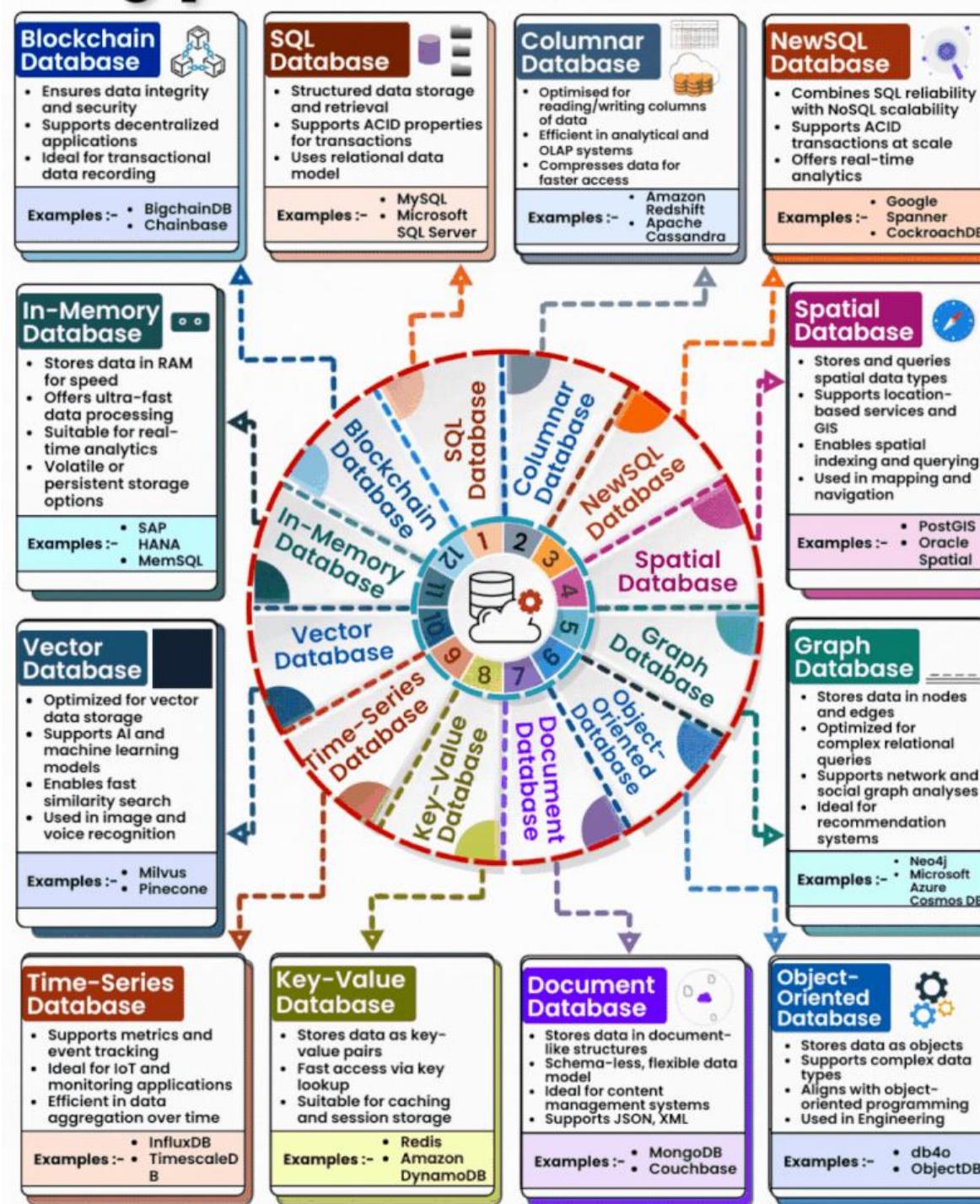


# Las Organizaciones se enfrentan a panoramas de datos complejos y desintegrados



# Arquitectura de Minería de Datos

## Tipos de Base de Datos



# Arquitectura de Minería de Datos

## Opciones de Base de Datos



Repositories



Servidores de  
Datos



Base de Datos  
de Conocimiento

**SAP HANA**

**SAP IQ**

**SAP DWC**

**Microsoft SQL Server**

**ORACLE®**

**Microsoft Azure**

**IBM DB2**

**OPENAPI INITIATIVE**

**MySQL**

**amazon REDSHIFT**

**SAP HANA & HANA Cloud**

**Alibaba Cloud**

**Amazon Redshift**

**amazon web services™ | S3**

**Google Cloud Platform**

**Google BigQuery**

**Hadoop**

**Microsoft Azure**

**DOCUMENT360**

**Helpjuice**

**BookStack**

**DATA INSIDE**  
VERTAALT DATA NAAR INZICHT

**N**



# Arquitectura

- **Repositorios**
- **Servidores de Datos**
- **BD de Conocimiento**



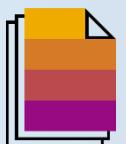
**Servidores de BD**  
SAP HANA, SAP IQ,  
SAP ERP



AWS, Google, Hadoop,  
Azure



**Base de Datos:**  
DB2, Oracle,  
Teradata, SQL



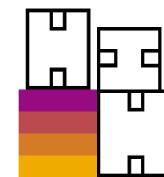
Text and Binary  
Files, XML, Excel,  
CSV & more



Data  
Warehouse

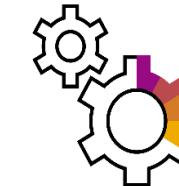
- **Proceso de Minería de Datos**

Prepare  
Data



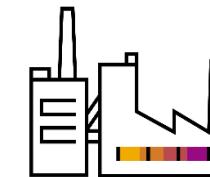
Data Manager

Build  
Model



Automated  
Modeler

Deploy &  
Manage Model



Predictive  
Factory

Evaluación

Operationalize  
& Embed



Predictive Analytics  
Integrator

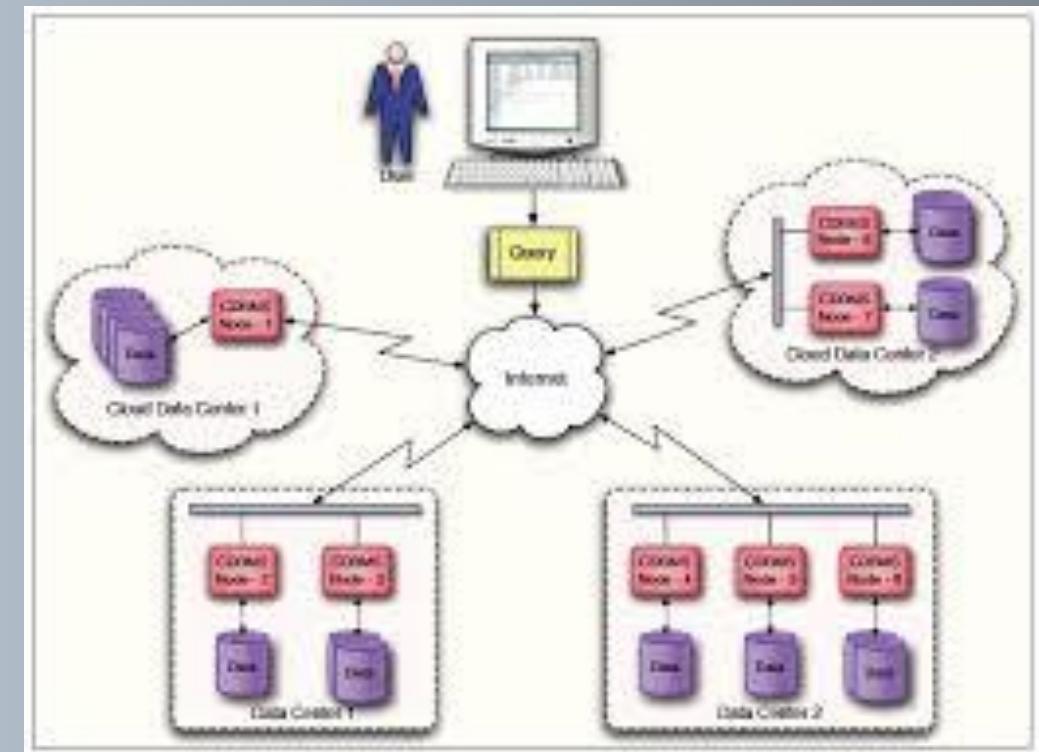
Data Mining

# Arquitectura de Minería de Datos

## Arquitectura de Base de Datos



- Enfoque moderno de arquitectura de datos busca proporcionar una **gestión unificada**, consistente y eficiente de los datos en una organización.
- Una **infraestructura** de datos se conecta de manera inteligente todos los activos de datos de una organización, independientemente de dónde residan.
- Esto permite un **acceso, integración y gestión** de datos más simplificados y automatizados.



## Arquitectura de Base de Datos – Características Clave



- 1. Integración de Datos:** proporciona una vista unificada de los datos, permitiendo que estos sean accesibles y utilizables sin importar su ubicación física o lógica.
- 2. Automatización y Orquestación:** Utiliza tecnologías avanzadas como inteligencia artificial y machine learning para automatizar tareas de integración, limpieza, y preparación de datos.
- 3. Escalabilidad y Flexibilidad:** está diseñado para ser escalable y flexible, adaptándose a las necesidades cambiantes de las organizaciones y permitiendo el crecimiento sin problemas.
- 4. Seguridad y Gobernanza de Datos:** Implementa controles de seguridad robustos y políticas de gobernanza de datos para asegurar que los datos estén protegidos y cumplan con las regulaciones.
- 5. Acceso en Tiempo Real:** Proporciona acceso en tiempo real a los datos, mejorando la toma de decisiones y la agilidad empresarial.





- **Retail:** Integración de datos de ventas, inventarios, y comportamiento de clientes para mejorar la gestión de inventarios y la personalización de ofertas.
- **Finanzas:** Consolidación de datos de transacciones, riesgos, y clientes para mejorar la detección de fraudes y la gestión de riesgos.
- **Salud:** Unificación de datos de pacientes, historiales médicos y dispositivos médicos para proporcionar atención médica más personalizada y efectiva.
- **Manufactura:** Integración de datos de producción, calidad y cadena de suministro para optimizar operaciones y reducir desperdicios.





# Arquitectura de Minería de Datos

## Repositorios de Datos





Un Repositorio es un lugar donde se almacenan los datos.

Es un lugar donde los datos se organizan y almacenan para que se pueda acceder y utilizar fácilmente.

Un repositorio puede ser físico o digital, y puede ser local o global.

Por ejemplo, los repositorios suelen tener :

- Una **estructura y jerarquía** bien definidas.
- Utilizan **metadatos** para describir y organizar su contenido,
- y proporcionan **control de acceso** para garantizar que solo los usuarios autorizados puedan acceder a los datos.

Los repositorios nos ayudan a **administrar** y realizar un seguimiento de la cantidad cada vez mayor de datos que creamos y consumimos.

También nos permiten **compartir datos** con otros para que podamos colaborar en proyectos o simplemente intercambiar información.





## Desafíos para los líderes de datos y análisis

1. Obtener una **vista unificada** y con gobierno de los datos provenientes de entornos complejos y variados.
2. No perder un **valioso contexto** empresarial al extraer datos de las aplicaciones
3. **Aumentar el análisis** con planificación predictiva para tener pronósticos más precisos
4. Aprovechar **las inversiones** en las instalaciones, mientras se traslada a la nube





## Repositorio vs. Base de datos – Diferencia clave

**Base de Datos** es una colección de datos a los que pueden acceder las computadoras para procesos transaccionales.

Los datos se almacenan en un formato que puede ser fácilmente leído por la computadora.

Una base de datos puede ser pequeña, como un solo archivo en su computadora, o puede ser grande, como varios de archivos.

**Repositorio de Datos** es un sistema más amplio que se utiliza para almacenar grandes volúmenes de datos, estructurados y no estructurados, generalmente de diversas fuentes.

Su objetivo principal es el análisis y la minería de datos a largo plazo.

Es ideal para almacenar y analizar datos históricos y para realizar análisis en grandes cantidades de datos, como en proyectos de Big Data o almacenamiento de datos en la nube.





## Repositorio vs. Base de datos – Diferencia clave

	<b>Base de Datos</b>	<b>Repositorio de Datos</b>
<b>Estructura:</b>	<ul style="list-style-type: none"><li>Está organizada en <b>tablas</b> (en el caso de bases de datos relacionales)</li><li>Los datos están estructurados y siguen un esquema definido.</li></ul>	Puede contener datos en diversas formas, incluyendo archivos de texto, imágenes, bases de datos, y más. Los datos no siempre tienen que estar estructurados
<b>Escalabilidad</b>	Las bases de datos pueden ser escaladas, pero están optimizadas para manejar volúmenes de datos que son manejables en tiempo real	Están diseñados para ser altamente escalables, manejando grandes volúmenes de datos provenientes de diversas fuentes
<b>Volumen de Datos</b>	Generalmente, manejan volúmenes de datos más pequeños y enfocados	Manejan grandes cantidades de datos, incluyendo datos históricos y de múltiples fuentes

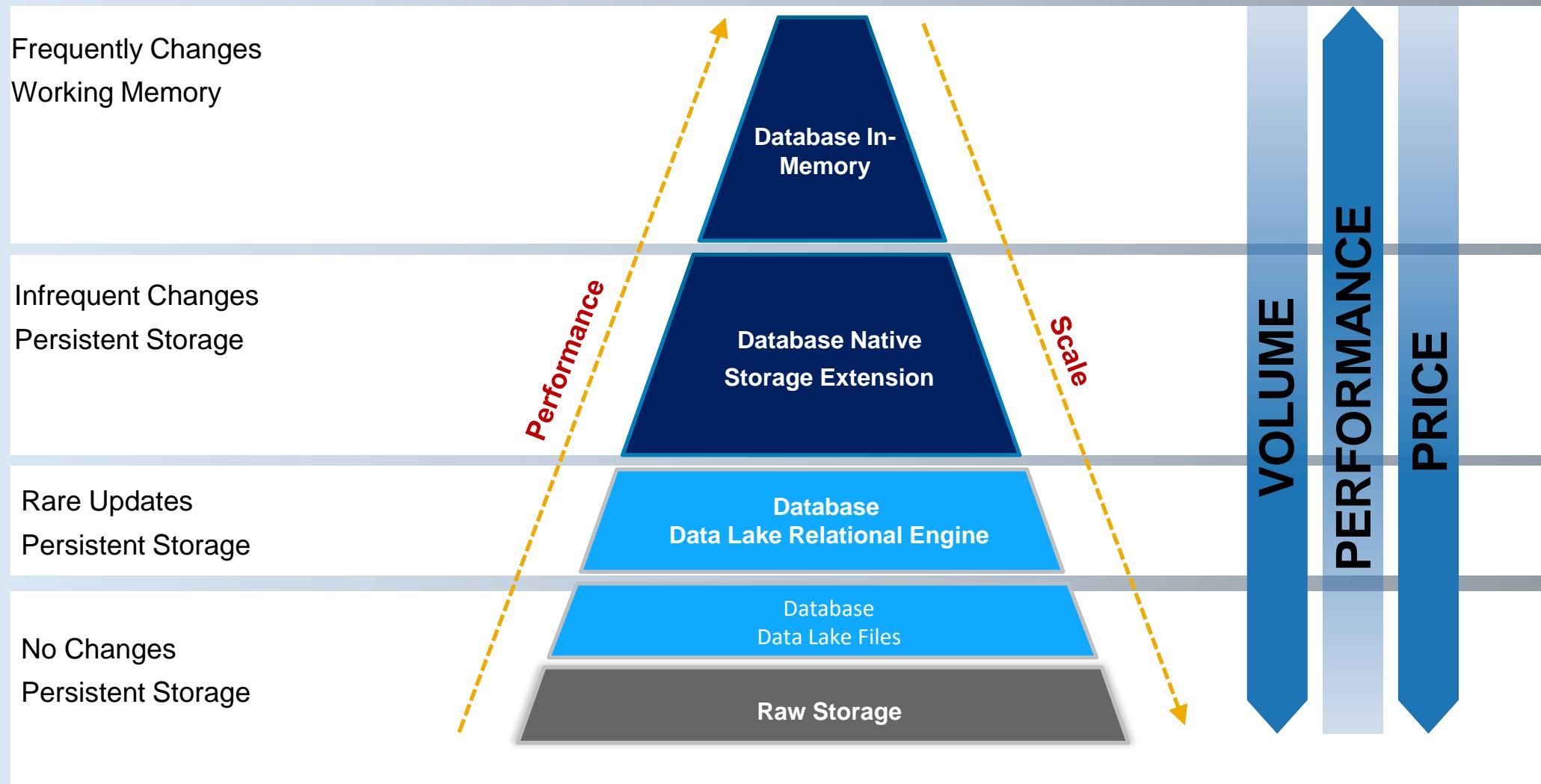


# Arquitectura de Minería de Datos

## Repositories



### Data Pyramid





## Beneficios de los Repositorios de Datos

Las empresas pueden tomar decisiones basadas sobre los datos almacenados.

El uso de repositorios de datos como parte de la gestión de datos es otro **nivel de inversión** que puede mejorar las decisiones empresariales, como:

- Concentrar los datos permite realizar informes o análisis más fáciles y rápidos porque **los datos están organizados**.
- Los administradores de bases de datos tienen más **facilidad para rastrear** problemas porque los repositorios de datos están compartidos
- Los datos se **conservan y archivan** por tiempo para extraer información histórica.





### Desventajas de los Repositorios de Datos

- La vulnerabilidad en los repositorios de datos que las empresas deben administrar de manera efectiva para mitigar los posibles **riesgos de seguridad de datos**, que incluyen:
- El crecimiento de los conjuntos de datos podría **alentar los sistemas**. Por lo que es necesario asegurarse de que los sistemas de **administración de bases de datos** puedan administrar el crecimiento de los datos.
- Un bloqueo del sistema podría afectar todos los datos. Realice **copias de seguridad** de las bases de datos y aíslle las aplicaciones de acceso para que se restrinja el riesgo del sistema.
- Las **políticas para seguridad**, recuperación y respaldo deben ser usadas para todos los datos.
- En algunos casos los repositorios pueden tener **altos costos de soporte y mantenimiento**





## Tipos de repositorios de datos

El término **Repositorio de Datos** se puede utilizar para describir varias formas de recopilar y almacenar datos:

- **Data Mart** son subconjuntos del repositorio de datos, que están orientados a lo que el usuario de datos necesita, se limita a los usuarios autorizados. Esos usuarios no pueden acceder a todos los datos del repositorio de datos.
- **Data Warehouse** es un grande repositorio de datos que agrega datos generalmente de múltiples fuentes o segmentos de una empresa, sin que los datos estén necesariamente relacionados.
- **Cubos de datos** son listas de datos con tres o más dimensiones almacenadas como una tabla, que servirán de consulta
- **Data Lake** es un grande repositorio de datos que almacena datos no estructurados que se clasifican y etiquetan con metadatos.
- **Repositorios de Metadatos** almacenan datos sobre datos y bases de datos. Los metadatos explican dónde se encuentra el origen de datos, cómo se capturó y qué representa.



## Prácticas recomendadas para trabajar con Repositorios de Datos



Al **crear y mantener repositorios de datos**, hay que tomar muchas decisiones de Hardware y Software. Establecer algunas de las mejores prácticas de almacenamiento de datos:

- Involucrar a **todas las partes** durante el desarrollo del proyecto y durante su uso.
- El repositorio de datos tendrá que **crecer**, como un sistema continuo.
- Expertos que puedan construir y **mantener** el repositorio de datos que se necesita.
- Uso de **herramientas ETL** para migrar datos al repositorio de datos.
- Construya primero su Data Warehouse, luego construya los Data Marts.
- Decida con qué **frecuencia** se cargará nuevos datos y considerar el **volumen** de datos.
- Crear **metadatos** para el análisis de datos de calidad y la presentación de informes.
- Los usuarios de datos deben tener **acceso** a educación y apoyo.

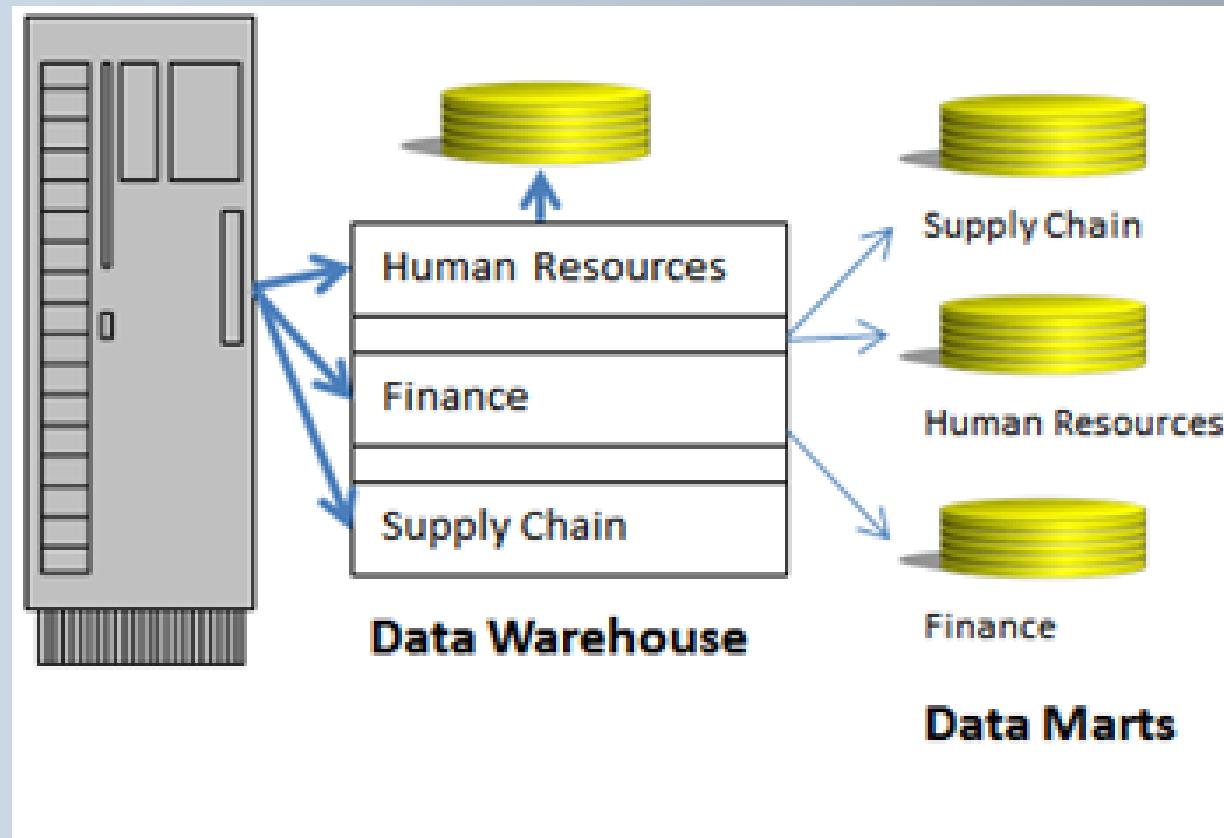


# Arquitectura de Minería de Datos



## Data Marts

Un Data Mart es una forma sencilla de un almacén de datos que se centra en un único tema o línea de negocio, como ventas, finanzas o marketing. Dado su enfoque, los Data Marts obtienen datos de menos orígenes que los Data Warehouse

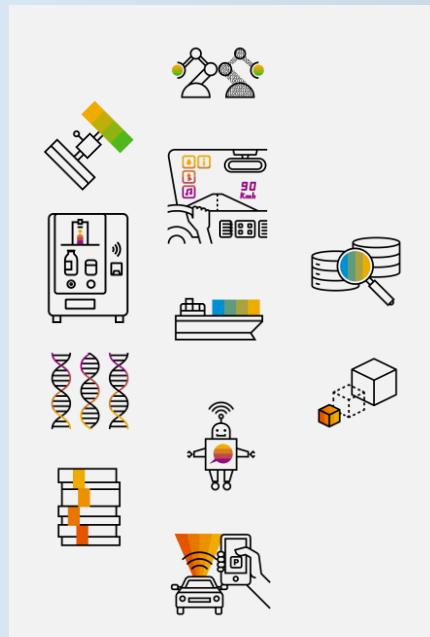


# Arquitectura de Minería de Datos

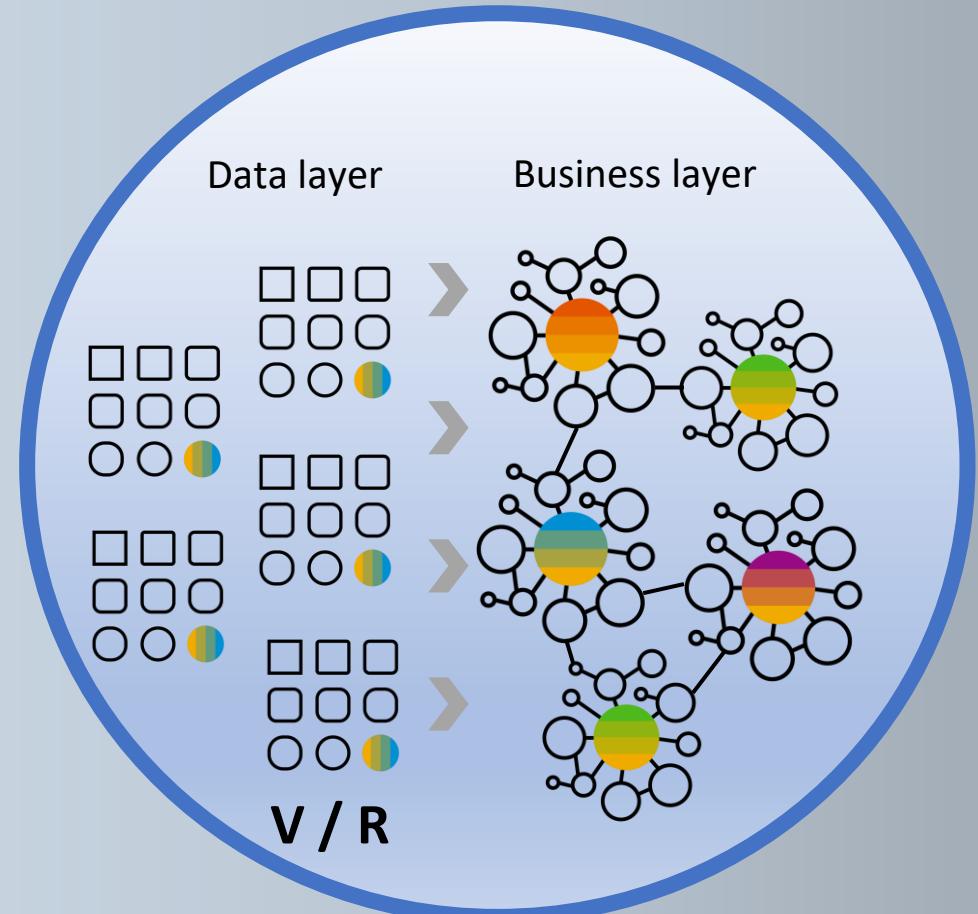


## Data Warehouse

Fuentes de Datos



Data Warehouse Cloud



Analytics Cloud

Ad hoc data exploration



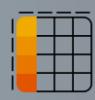
Powerful visualization



Intelligent augmentation



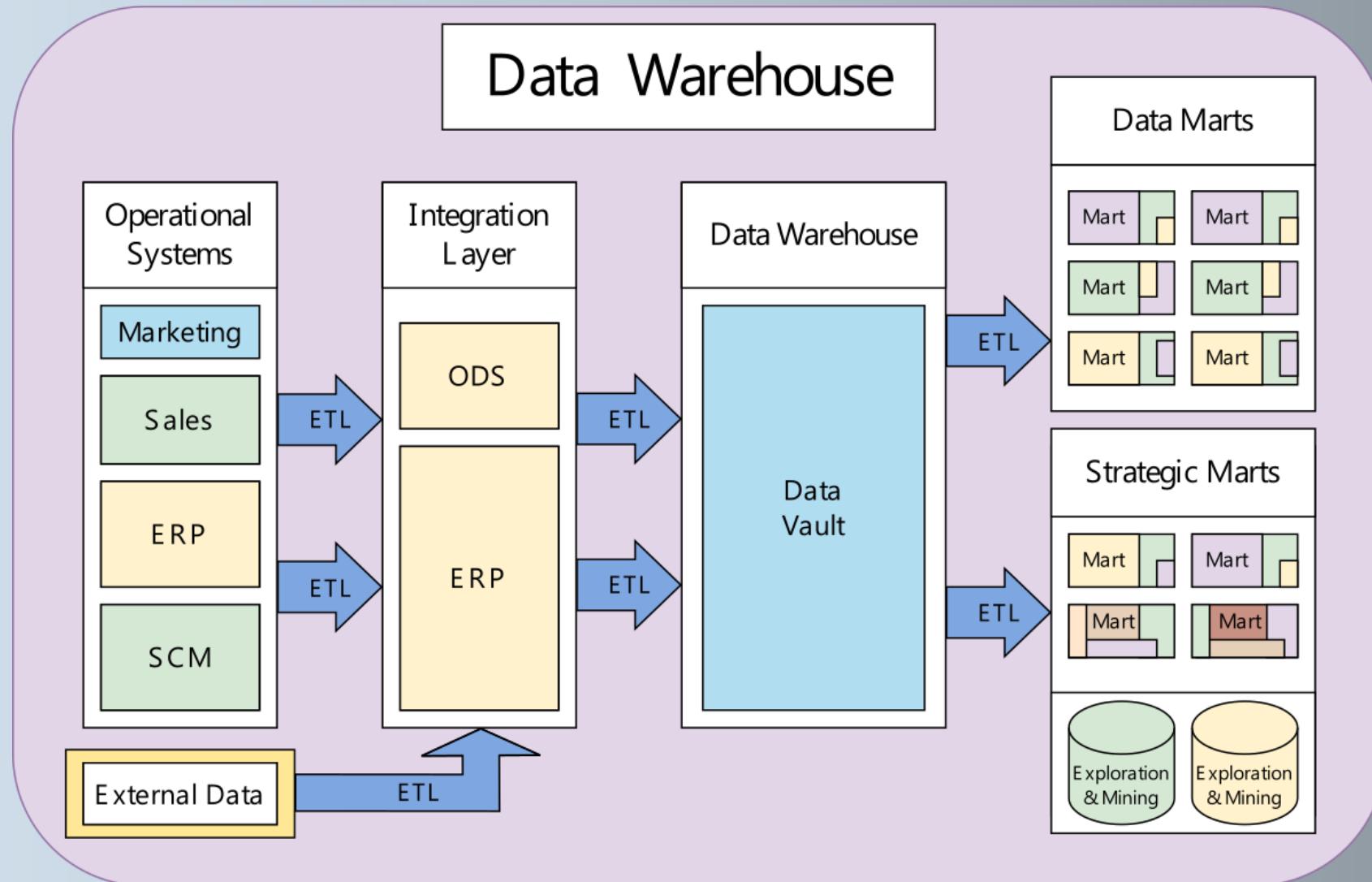
Business planning<sup>1</sup>



# Arquitectura de Minería de Datos



## Data Warehouse / Data Marts



# Arquitectura de Minería de Datos

## Data Warehouse características

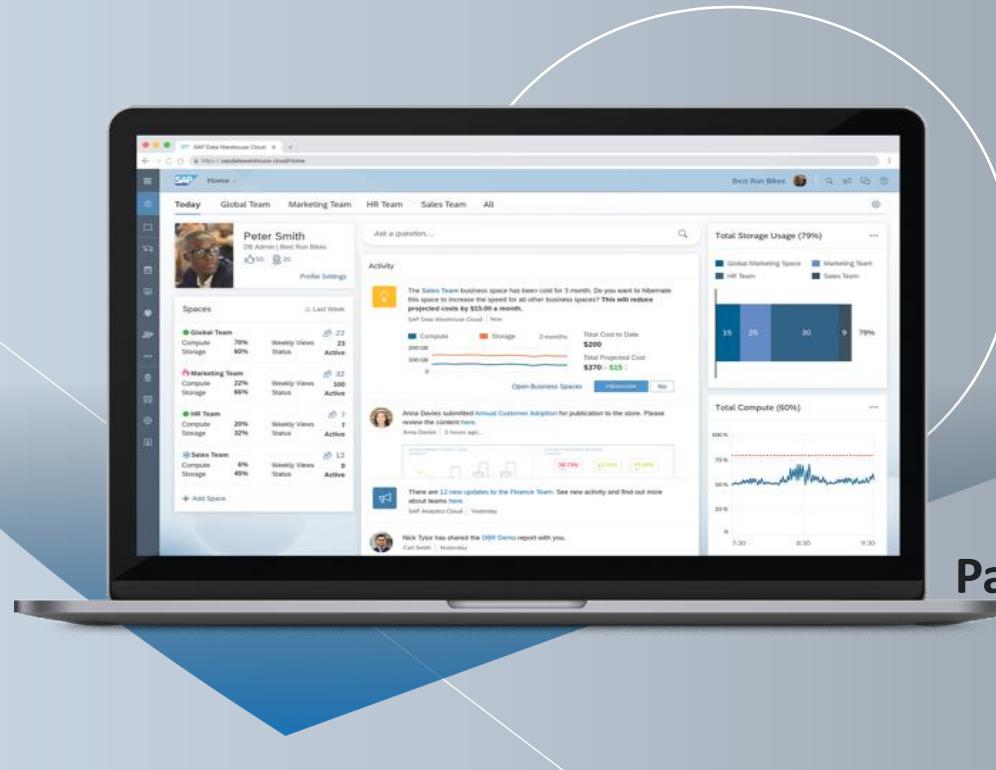


### COLABORATIVO

Diseñado tanto para TI empresarial como para usuarios de línea de negocios para administrar y obtener información valiosa.

### RAPIDO

Aproveche el poder de acceder instantáneamente a los datos que necesita, cuando los necesita, sin costos anticipados.



### END-TO-END

Integra datos de cualquier fuente y consumidos con potentes análisis nativos y semántica empresarial

### ELASTICO

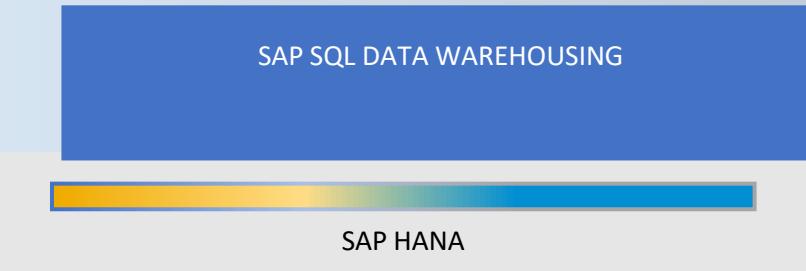
Pagar solo por lo que usa y reasigne los recursos de forma rápida y sencilla entre diferentes casos de uso y líneas de negocio.



# Arquitectura de Minería de Datos



## Data Warehouse Opciones



### ON-PREMISES

The strategic solutions for data warehousing on-premises.



### HYBRID

SAP Data Warehouse Cloud extends your existing data warehousing investment to support **hybrid cloud**.



### CLOUD

The strategic public cloud product offered as a software service managed by SAP.



# Arquitectura de Minería de Datos



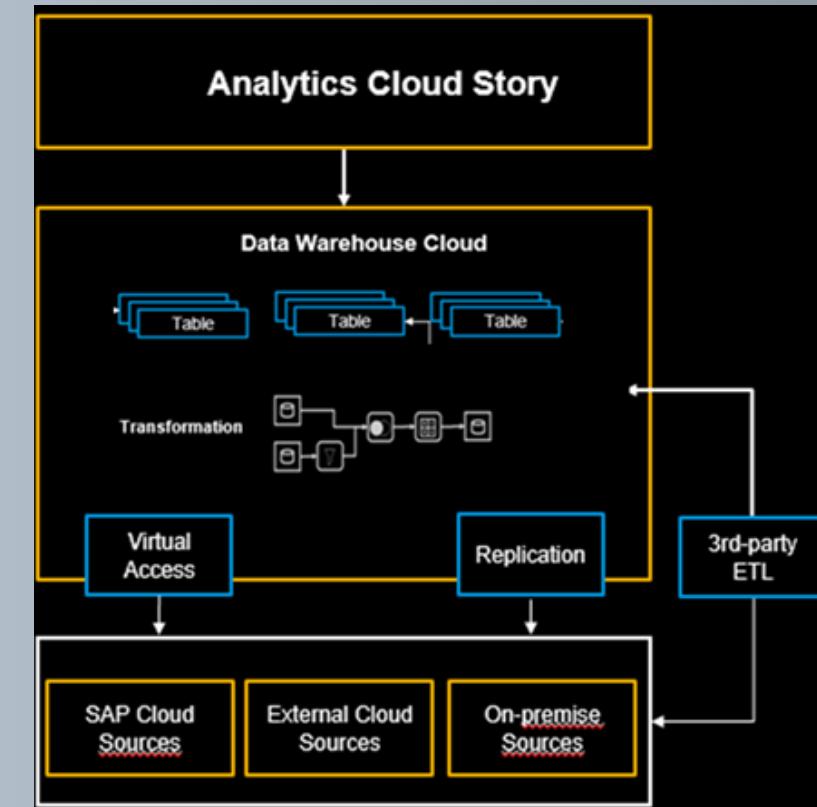
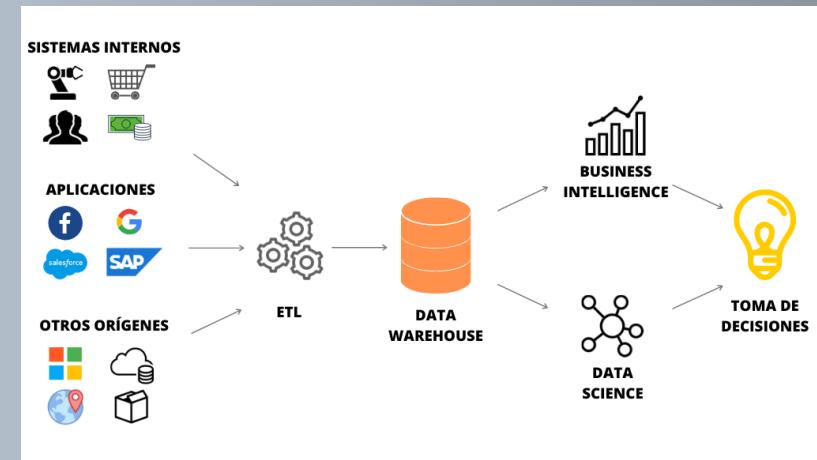
## Data Warehouse

### Alcance

- Lograr una Fuente única confiable a nivel de empresa
- Minimizar la copia de datos a través de un gobierno a los accesos de los datos
- Asegurar la colaboración entre áreas de negocio
- Tener apertura a herramientas de terceros

### Beneficios

- Incrementar la colaboración entre roles basados en seguridad
- Compartir objetos entre departamentos
- Organizarse por si mismos y aislar los ambientes de modelaje de datos
- Contar con capacidades gráficas y modelos basados en scripts para personas de negocio
- Accesos rápidos a la información



# Arquitectura de Minería de Datos



## Data Warehouse - Capacidades

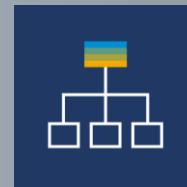
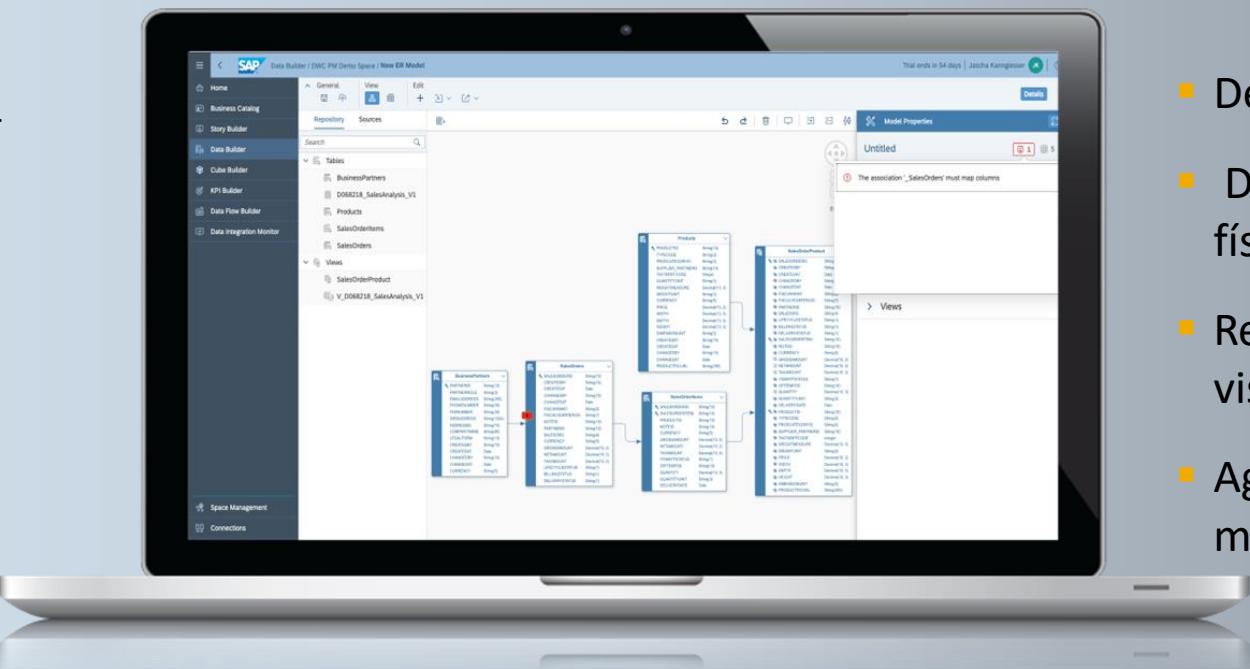


Fuente de Datos e ingestá:  
Acceso a fuentes remotas

- **Adaptadores**

- OData Adapter  
 SAP HANA Adapter

- JDBC / ODBC / CamelJdbc
- Amazon Athena
- Amazon S3
- Google BigQuery
- ...

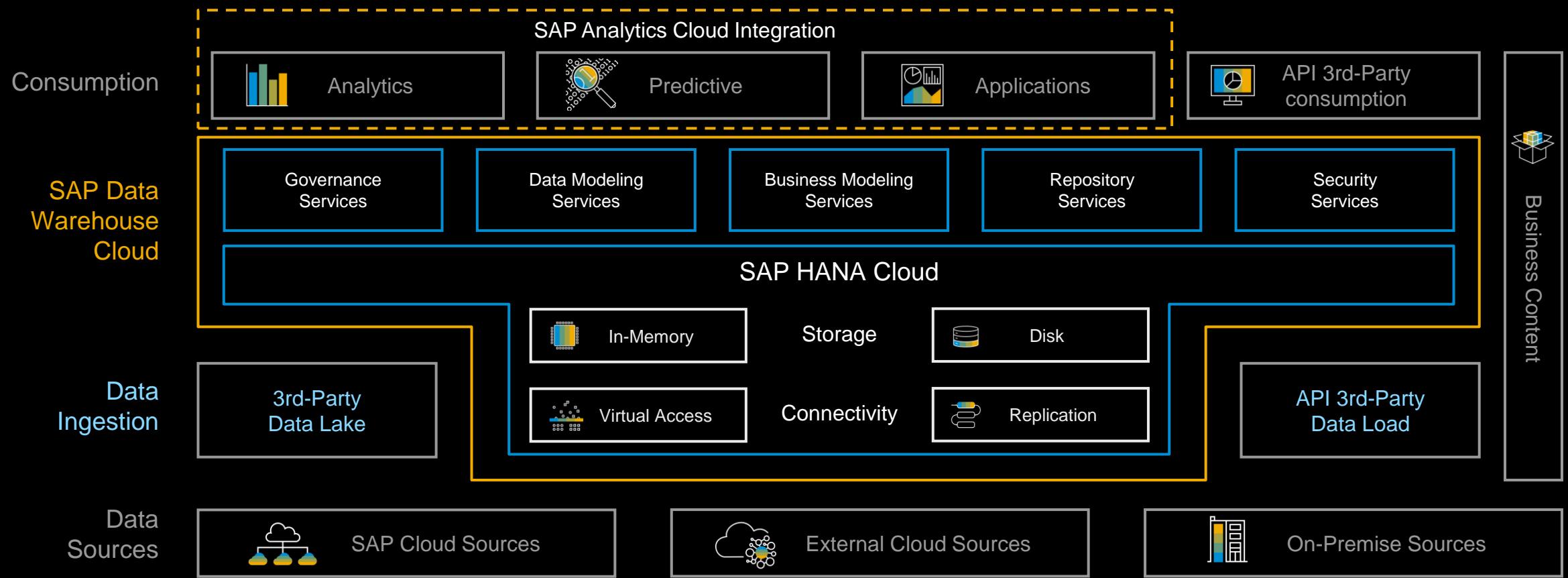


Servicios de Modelaje de Datos:  
Modelo Entidad – Relación

- Definición de modelos entidad-relación
- Diseñe modelos de bases de datos físicos o remotos
- Reutilizar entidades existentes (tabla, vista)
- Agregar nuevas entidades sobre la marcha

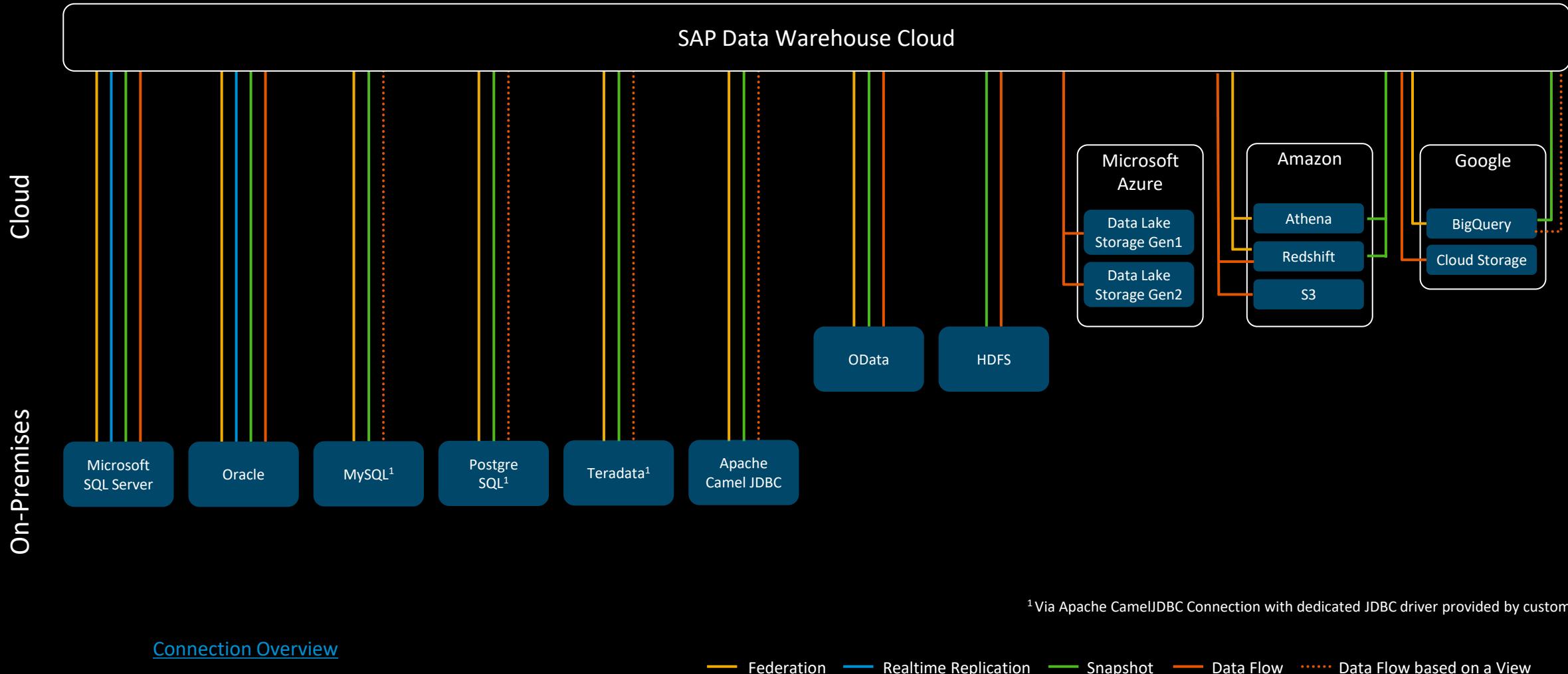


# Architecture



A **comprehensive approach** to data warehousing for **instant data-to-value**

# Integrate. Data Sources & Supported Options for other Systems

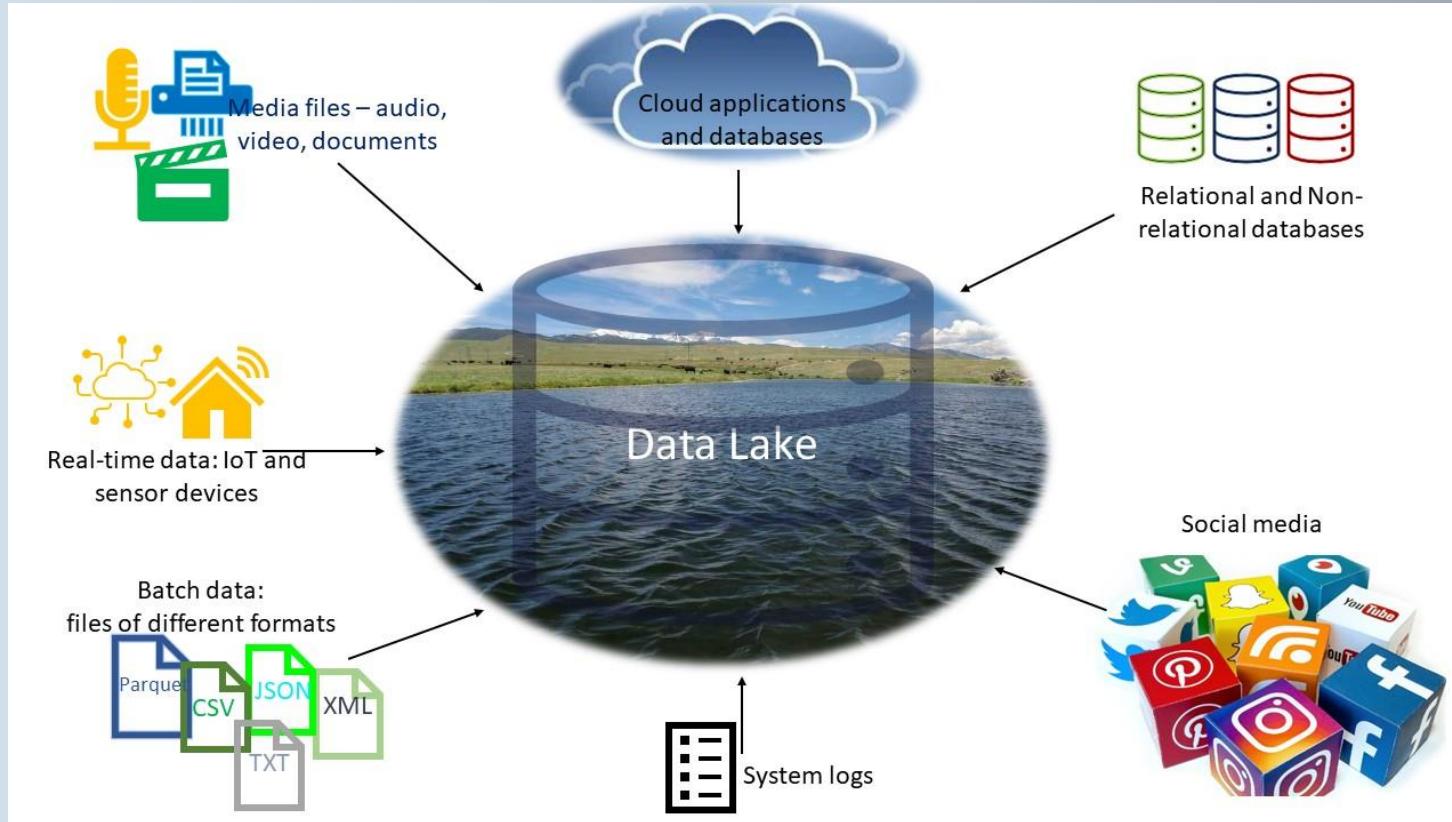


# Arquitectura de Minería de Datos



## Data Lake

Un Data Lake permite realizar nuevos tipos de análisis, como machine learning sobre nuevos orígenes, tales como archivos de transacciones, datos de secuencias de clics, redes sociales y dispositivos conectados a Internet almacenados en lagos de datos.

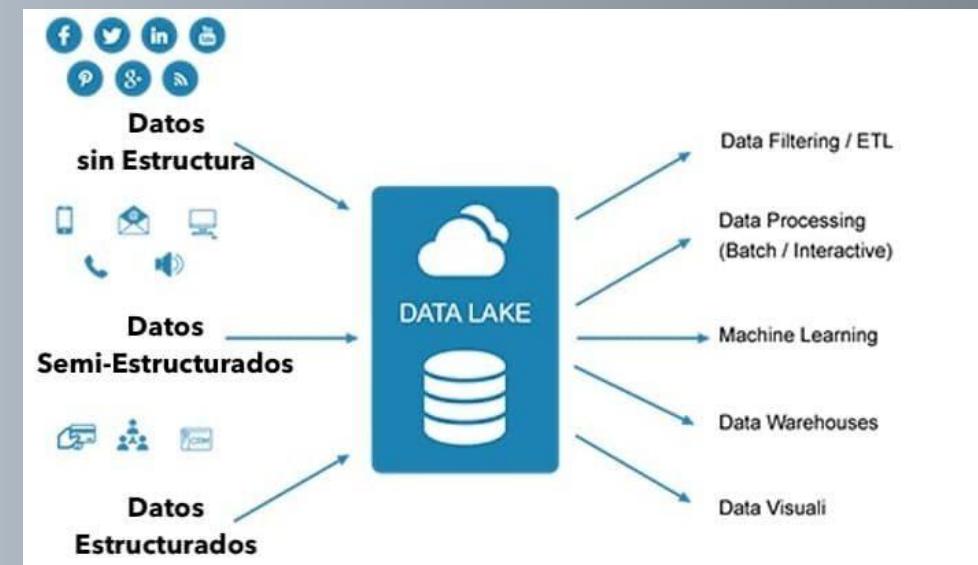


# Arquitectura de Minería de Datos



## Data Lake - Desafío

- El principal desafío de una arquitectura de Data Lake es que los **datos sin procesar** se almacenen sin supervisión de los contenidos.
- Para que un lago de datos habilite el uso de los datos, debe contar con mecanismos definidos para **catalogar y proteger** los datos.
- Sin esos elementos, no se pueden encontrar los datos ni se puede confiar en ellos, lo que resulta en un “pantano de datos”.
- Para satisfacer las necesidades de audiencias más amplias, los data Lake deben tener gobernanza, coherencia semántica y controles de acceso.



## Tipos de Base de Datos

### Distributed Database

It comprises of at least two documents situated in various destinations either on a similar system or on unique systems.

### Centralized Database

A centralized database framework is a framework that keeps the information in one single database at one single place.

### Personal Database

Information is gathered and stored on PCs, which is in small quantity and can easily manageable.

### Relational Database

It is described by a set of tables from where data can be accessed. Relational database can store a large amount of information in a set of tables, which are linked to each other.

### Operational Database

An operational database is utilized to store and manage a huge amount of data in real time.

## Repositorios

### Hierarchical Database

In hierarchical database model, data is organized in a tree structure that links a number of different elements to one parent record.

### Cloud Database

It is deployed and delivered through a cloud platform like Platform-as-a-Service (PaaS) that permits the organizations & their applications to store and mange information from the cloud.

### Object Oriented Database

It is a group of object-oriented programming and relational database, that is organized around object rather than actions and logic.

### NoSQL Database

NoSQL database is used to efficiently manage and analyze a large set of distributed data that may stored at several virtual servers.



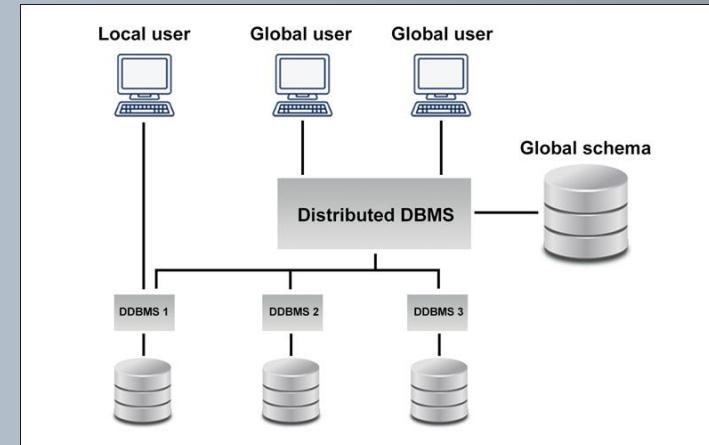
## Tipos de Base de Datos

## Base de Datos Distribuida



Es una Base de Datos que consta de al menos dos documentos situados en varios destinos, ya sea en un sistema similar o en sistemas únicos. Partes de la base de datos se guardan en varios lugares físicos y el manejo se distribuye entre diferentes centros de bases de datos.

- Las bases de datos distribuidas pueden ser homogéneas o heterogéneas. Por lo general, las bases de datos distribuidas pueden incluir las siguientes características:
- Independiente del hardware
- Independiente de la ubicación
- Independiente del sistema operativo
- Independiente de la red
- Transparencia de las transacciones
- Procesamiento de consultas distribuidas
- Gestión de transacciones distribuidas



### Examples of distributed database:

- Apache Cassandra
- Apache HBase
- Apache Ignite
- Couchbase Server
- Amazon SimpleDB
- FoundationDB
- Clusterpoint



## Tipos de Base de Datos

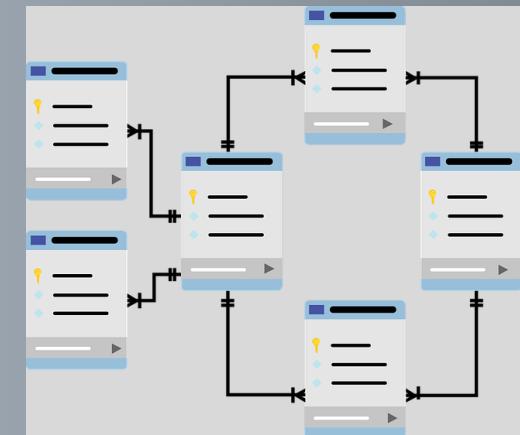
### Base de Datos Relacional



Se describe mediante un conjunto de tablas desde donde se puede acceder a los datos.

La base de datos relacional puede almacenar información en un conjunto de tablas, que están vinculadas entre sí.

Cada tabla se compone de información en filas y columnas en las que cada columna representa un tipo particular de información como nombre, dirección, cada fila contiene información única y cada campo de una tabla tiene su propio tipo de datos.



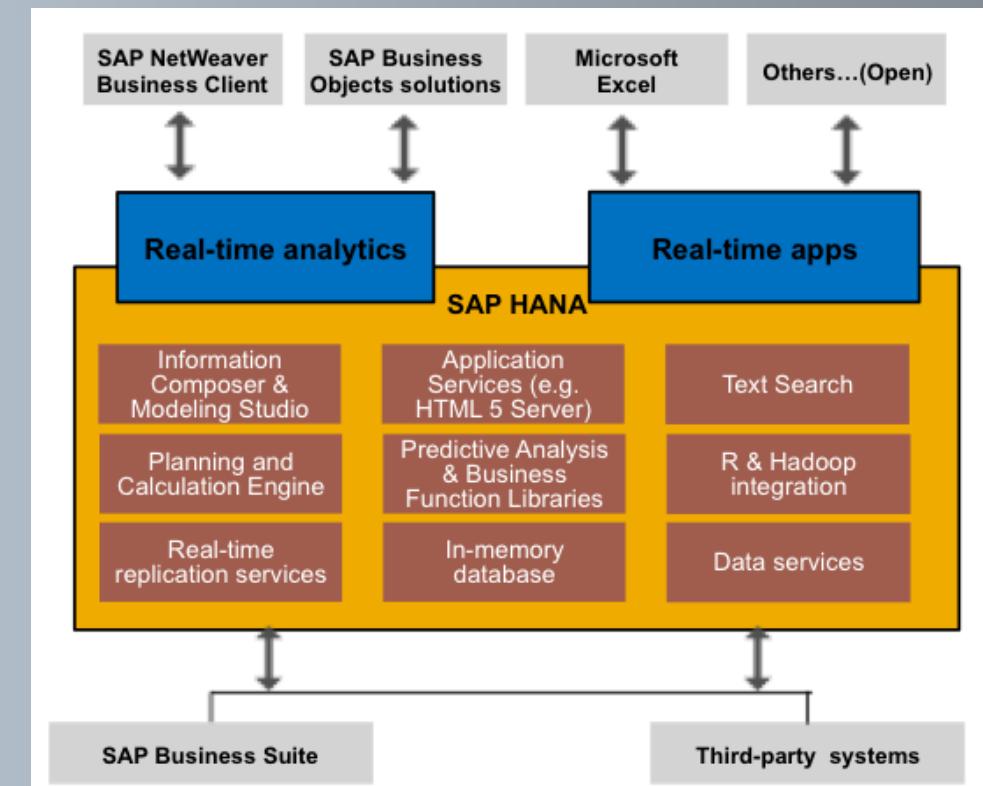
- La base de datos relacional se utiliza para almacenar el registro financiero de cualquier empresa
- Mantiene registros de los envíos de los clientes y sus pedidos.
- La base de datos relacional garantiza la integridad de los datos y un mejor rendimiento.
- Proporciona una mejor seguridad de los datos y permite múltiples usuarios.
- Varios clientes pueden acceder a la misma base de datos.
- Desventajas: Alto costo de configuración y mantenimiento, Se requieren configuraciones sofisticadas de redes y hardware para las bases de datos relacionales.



## Tipos de Base de Datos

- Una base de datos operativa se utiliza para almacenar y gestionar datos en tiempo real. Los datos relativos a las operaciones (marketing, servicios prestados a los clientes y relaciones con ellos) de cualquier proyecto se pueden almacenar dentro de una base de datos operativa.
- Del mismo modo, una empresa permite a sus representantes de ventas en el campo actualizar la información de ventas con el fin de aumentar los ingresos.
- Microsoft, Oracle, Amazon Web Services, SAP e IBM son los actores importantes en el Sistema de Gestión de Bases de Datos Operativas (ODBMS)

## Base de Datos Operativa

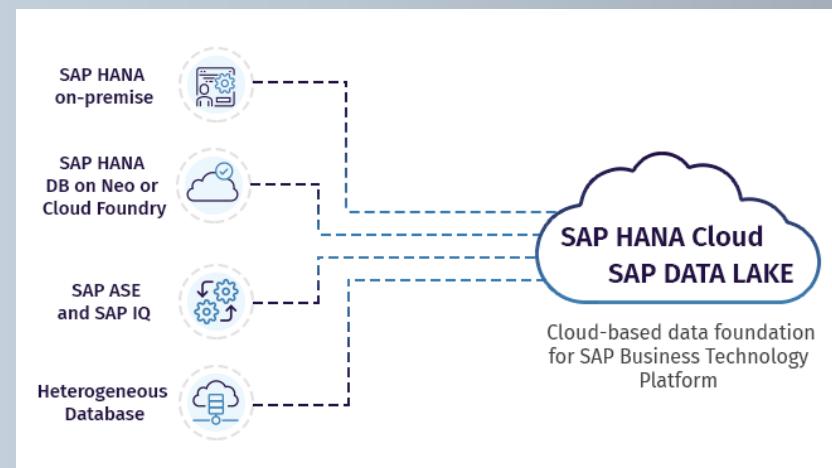


## Tipos de Base de Datos

## Base de Datos Cloud



- La base de datos en la nube es un tipo de administración de bases de datos que se implementa y entrega a través de una plataforma en la nube como plataforma como servicio (PaaS) que permite a las organizaciones y sus aplicaciones almacenar y administrar información desde la nube.
- Es popular debido a varios beneficios, como adquirir más capacidad de almacenamiento, alto ancho de banda, escalabilidad y disponibilidad según la capacidad de pago del usuario.
- Además, una base de datos en la nube también se transmite como una administración, donde el comerciante se ocupa específicamente de los procedimientos de backend de instalación e implementación de bases de datos.



## Tipos de Base de Datos



### Amazon

- Amazon Relational Database Service (RDS) es un servicio de base de datos completamente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos relacional en la nube. Es compatible con una variedad de motores de bases de datos
- Amazon DynamoDB: Amazon DynamoDB es un servicio de base de datos NoSQL completamente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos no relacional en la nube. Ofrece alto rendimiento, escalabilidad y fiabilidad.

### Google

- Google Cloud SQL: Google Cloud SQL es un servicio de base de datos totalmente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos MySQL en la nube. Ofrece alto rendimiento, escalabilidad y seguridad.
- Google Cloud Firestore: Google Cloud Firestore es un servicio de base de datos NoSQL totalmente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos no relacional en la nube. Ofrece actualizaciones en tiempo real y soporte fuera de línea.



## Tipos de Base de Datos

Popular Cloud Databases



### Azure

- Base de datos SQL de Azure: Base de datos SQL de Azure es un servicio de base de datos totalmente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos relacional en la nube. Admite una variedad de motores de bases de datos
- Azure Cosmos DB: Azure Cosmos DB es un servicio de base de datos totalmente administrado que facilita la configuración, el funcionamiento y el escalado de una base de datos NoSQL en la nube. Ofrece una variedad de modelos de datos y API, incluida la compatibilidad con MongoDB, Cassandra y Azure Table Storage.

### MongoDB

MongoDB Atlas: MongoDB Atlas es un servicio de base de datos totalmente administrado que facilita la configuración, operación y escalado de una base de datos MongoDB en la nube. Ofrece alto rendimiento, escalabilidad y seguridad.



Características en Minería de Datos	Amazon	Google Cloud	Azure	SAP HANA	IBM InfoSphere	Oracle
Soporte para Algoritmos Minería	Limitado, se apoya en servicios externos	Limitado, se apoya en BigQuery ML	Integrado con Azure Machine Learning	Soporte nativo de algoritmos de ML y minería de datos	Soporte con IBM SPSS, Modeler	Oracle Data Mining (ODM) integrado
Herramientas Integradas de Minería	Amazon Machine Learning, SageMaker	BigQuery ML, AI Platform	Azure Machine Learning, Synapse Analytics	Predictive Analysis Library Automated Predictive Library	InfoSphere Warehouse con minería	Oracle Advanced Analytics, Oracle Data Mining (ODM)
Integración con Big Data para Minería	Integración con EMR, Redshift, Kinesis	Integración con BigQuery, Dataflow	Integración con Azure Synapse, Databricks	Integración nativa con SAP BW y Big Data Services	Integración con BigInsights, Hadoop	Integración con Oracle Big Data SQL, Hadoop
Soporte para Procesamiento en Tiempo Real	Soporte limitado, mejorado con Amazon Kinesis	Soporte a través de Spanner para transacciones globales	Azure Stream Analytics para procesamiento en tiempo real	Procesamiento en tiempo real con HANA Streaming Analytics	InfoSphere Streams para análisis en tiempo real	Oracle Stream Analytics, Real-Time Decision Server
Capacidades Predictivas y Prescriptivas	Mediante servicios adicionales como SageMaker	A través de BigQuery ML y AI Platform	A través de Azure ML y Power BI	Integrado en SAP HANA con análisis predictivo	InfoSphere y SPSS con capacidades predictivas	Oracle Advanced Analytics, Oracle Data Mining (ODM)
Escalabilidad de Procesos de Minería de Datos	Depende del escalado de los servicios externos	Escalabilidad global con Spanner y BigQuery	Escalabilidad con Synapse y Machine Learning	Alto rendimiento y escalabilidad in-memory	Escalabilidad en procesamiento distribuido	Alto rendimiento y escalabilidad a través de RAC
Capacidades de Visualización de Datos	Amazon QuickSight, Tableau (integrado)	Google Data Studio, Looker	Power BI, Tableau (integrado)	SAP Analytics Cloud, Lumira	IBM Cognos, Tableau (integrado)	Oracle BI, Oracle Analytics Cloud
Facilidad de Implementación de Modelos	Integra modelos entrenados desde SageMaker	Fácil implementación con BigQuery ML	Fácil implementación con Azure ML	Directamente desde HANA con modelos predictivos	Integración con SPSS para despliegue rápido	Integración directa con Oracle Data Mining (ODM)
Automatización de Modelos de Minería de Datos	Automatización limitada, depende de SageMaker	Automatización a través de AutoML	Azure AutoML para automatización de modelos	APL y PAL para automatización en HANA	Soporte a través de Modeler	Oracle AutoML y ODM para automatización

## Capacidades para Minería de Datos



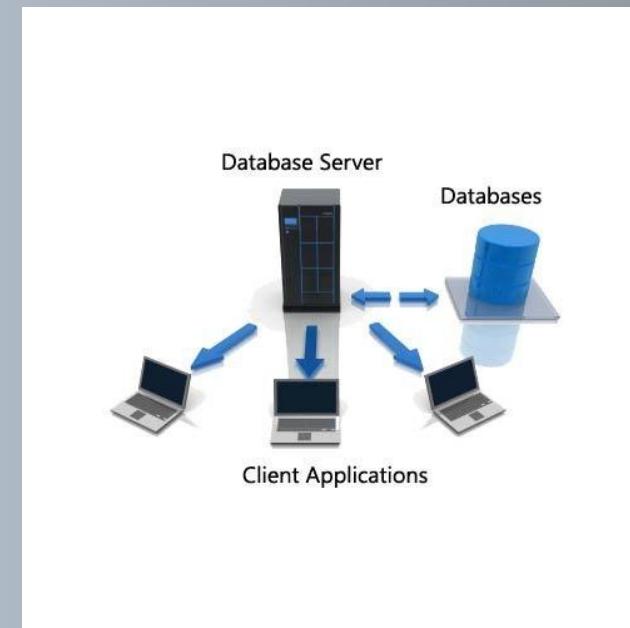
- **Amazon RDS/Aurora:** Fuerte integración con AWS SageMaker y otros servicios, pero requiere de herramientas adicionales para capacidades avanzadas de minería de datos.
- **Google Cloud SQL/Spanner:** Soporte a través de BigQuery ML y AI Platform, con opciones para escalar y analizar grandes volúmenes de datos.
- **Azure SQL Database/Cosmos DB:** Integración completa con Azure Machine Learning y Synapse para minería de datos con herramientas avanzadas de modelado y visualización.
- **SAP HANA:** Ofrece un soporte robusto nativo para minería de datos, con bibliotecas de análisis predictivo y automatizado.
- **IBM InfoSphere:** Potente en integración de datos y análisis, soportado por herramientas como SPSS y Modeler.
- **Oracle Real Application Clusters (RAC):** Fuerte en capacidades integradas de minería de datos con Oracle Data Mining y herramientas de análisis predictivo avanzadas.

Cada una de estas bases de datos ofrece diferentes niveles de soporte y capacidades para la minería de datos, dependiendo del enfoque y necesidades específicas.





# Arquitectura de Minería de Datos Servidores de Datos



# Arquitectura de Minería de Datos

## Servidores de Base de Datos

### Definición



- Un servidor de base de datos es una plataforma de hardware y software de base de datos dedicado a proporcionar servicios de base de datos.
- Es un componente crucial en el entorno informático **cliente-servidor** donde proporciona información crítica para el negocio solicitada por las aplicaciones cliente.
- Del **lado del software** es la aplicación de base de datos. La aplicación tiene un conjunto de estructuras en memoria y procesos que acceden a un conjunto de archivos de la base de datos.
- Del **lado del hardware** es el sistema de servidor utilizado para el almacenamiento y la recuperación de bases de datos.
- Las cargas de trabajo de bases de datos requieren una gran capacidad de almacenamiento y una alta densidad de memoria para procesar los datos de manera eficiente lo que significa que la máquina suele ser una computadora dedicada de gama alta.





# Arquitectura de Minería de Datos

## Servidores de Base de Datos ¿Para qué se utiliza?

Los servidores de bases de datos tienen varios casos de uso.

Algunos de ellos son:

- Tratar con grandes cantidades de datos regularmente.
- Los servidores de bases de datos brillan en una **arquitectura cliente-servidor**, donde los clientes procesan datos con frecuencia.
- Gestión de la **recuperación y seguridad** del DBMS.
- Los servidores de bases de datos llevan a cabo las **restricciones** especificadas dentro del DBMS.
- El servidor controla y **administra** todos los clientes conectados y maneja las solicitudes de acceso y control de la base de datos.
- Proporcionar control de **acceso simultáneo**, un entorno multiusuario donde muchos usuarios pueden acceder a la base de datos simultáneamente manteniendo la seguridad



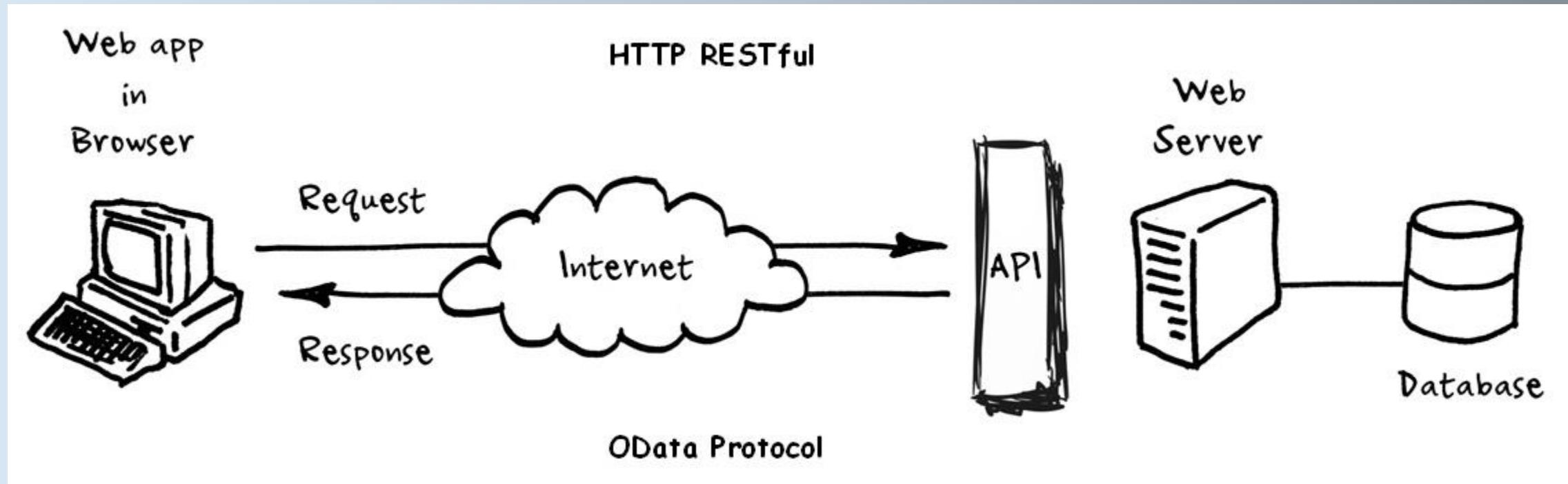


# Arquitectura de Minería de Datos

## Servidores de Base de Datos ¿Cómo funciona?

- El DBMS proporciona funcionalidad de servidor de base de datos, y algunos DBMS proporcionan acceso a la base de datos solo a través del modelo cliente-servidor.
- Su función principal es recibir solicitudes de los equipos cliente, buscar los datos requeridos y devolver los resultados.
- Los clientes tienen acceso a un servidor de bases de datos a través de una **aplicación front-end** que muestra los datos solicitados en el equipo cliente o a través de una **aplicación back-end** que se ejecuta en el servidor y administra la base de datos.
- El **estándar ODBC** (Open Database Connectivity) proporciona la API que permite a los clientes llamar al DBMS. ODBC requiere el software necesario tanto en el lado del cliente como del servidor.
- En un **modelo maestro-esclavo**, el servidor maestro de base de datos es la ubicación de datos principal. Los servidores esclavos de base de datos son réplicas del servidor maestro que actúan como servidores proxy.





REST es cualquier interfaz entre sistemas que use HTTP para obtener datos o generar operaciones sobre esos datos en todos los formatos posibles, como XML y JSON.



# Server Computer

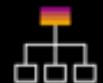
## Application Development



JavaScript



Fiori UX



Graphic Modeler

## Advanced Analytical Processing



Spatial



Graph



Predictive



Json & Python

## Data Integration and Quality



Data Virtualization



ELT & Replication



Data Quality



Hadoop & Spark Integration

## Database Management



Columnar  
OLTP + OLAP



Multi-Core &  
Parallelization



Advanced  
Compression



Multi-Tier Storage



Data Modeling



Openness



Admin &  
Security



## Servidores de Base de Datos - Ejemplos



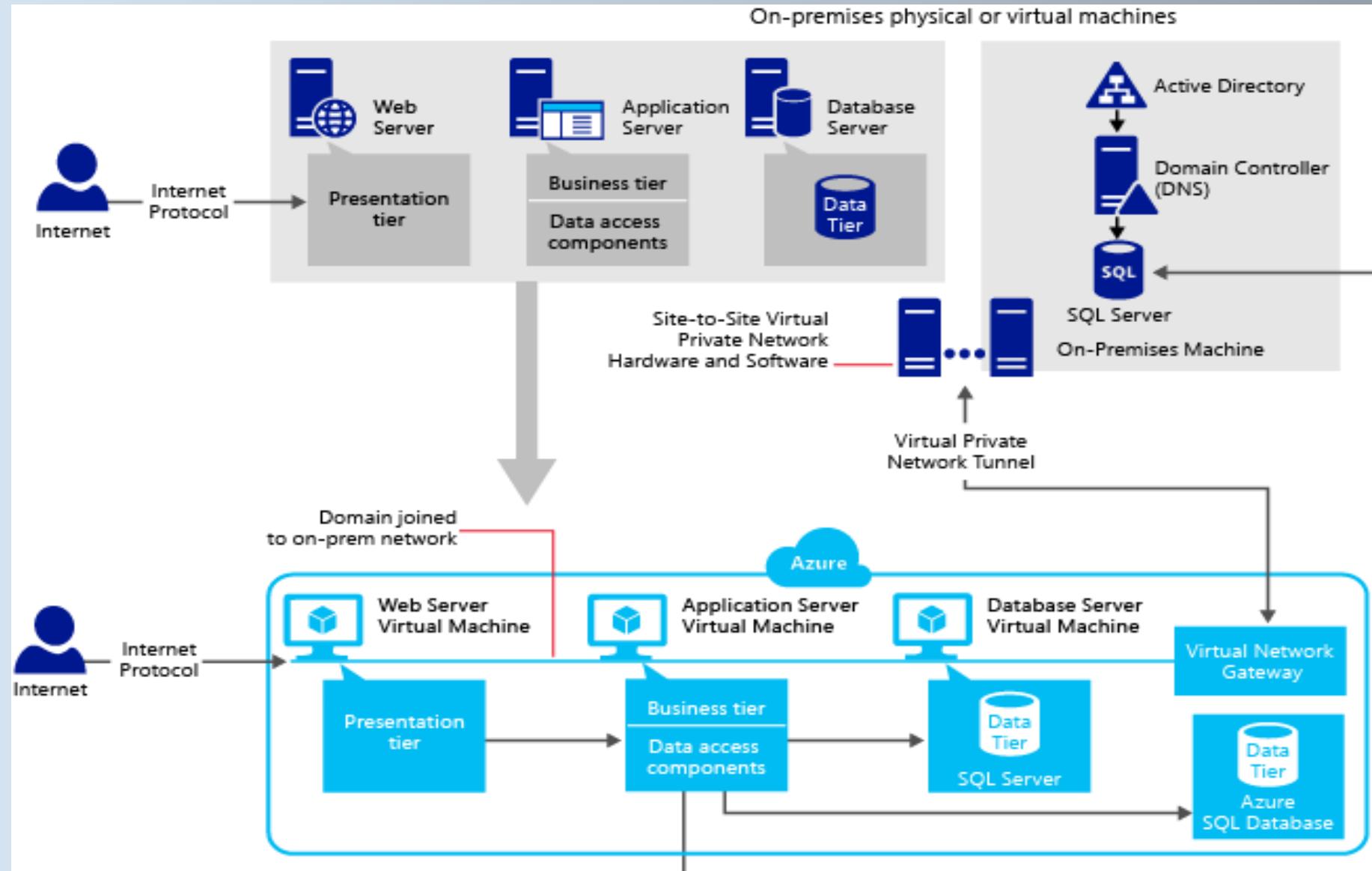
La siguiente es una lista de algunos servidores de bases de datos conocidos y ampliamente utilizados. La lista no es exhaustiva y hay muchas otras soluciones en el mercado.

- **MySQL** es un sistema de gestión de bases de datos relacionales (RDBMS) de código abierto basado en SQL. MySQL viene en una versión gratuita y de pago, y es compatible con Linux y Windows, ofrece análisis nativos en tiempo real y un servicio unificado para bases de datos OLAP y OLTP.
- **PostgreSQL** es un DBMS relacional de objetos avanzado. PostgreSQL es compatible con Windows y Linux, y sus características son una excelente seguridad de datos y una rápida recuperación de datos.
- **Microsoft SQL Server** es un sistema de administración de bases de datos relacionales utilizado principalmente para almacenar y recuperar datos solicitados por otras aplicaciones. MSSQL Server permite a muchos usuarios tener acceso a la misma base de datos al mismo tiempo, admite varios lenguajes de programación, como Assembly, C / C ++, Linux y opera en Windows y Linux.



# Arquitectura de Minería de Datos

## Servidores de Base de Datos - Microsoft





# Arquitectura de Minería de Datos

## Servidores de Base de Datos

**SAP HANA** es un RDBMS orientado a columnas desarrollado por SAP SE. La función principal del sistema es almacenar y recuperar datos según lo solicitado por las aplicaciones cliente. Es compatible con muchos tipos diferentes de aplicaciones. SAP HANA admite OLTP, OLAP y SQL, y puede administrar datos SAP y no SAP.

**IBM Db2** es un RDBMS que entrega datos a sus clientes de servidor de datos IBM. Db2 está escrito en C/C++ y Assembly. Está basado en NoSQL y admite tipos de archivo JSON y XML. Db2 da soporte a plataformas Linux, UNIX y Windows.

**Oracle** ofrece uno de los DBMS relacionales de objetos más populares, admite JSON binario y ofrece escaneos de datos diez veces más rápidos en comparación con las versiones anteriores. Windows, Linux y muchas versiones de sistemas operativos UNIX son compatibles.

**MongoDB Atlas** versión gratuita y comercial. MongoDB está desarrollado para aplicaciones que utilizan datos estructurados y no estructurados, y su motor admite documentos JSON y NoSQL.



# Arquitectura de Minería de Datos

## Servidores de Base de Datos Comparativo On-premise vs Cloud



	On-premises	Cloud	
	Application	Application	
• Manage users • Manage catalog/content	Data Layer	Data Layer	
• Install DBMS software • Update/patch • Backup, tune • Monitor, restart	DBMS Platform + Databases	Database Platform	• Provision capacity • Select service options
• Select OS • Configure OS • Update OS	OS	Managed Service	
• Select, purchase, provision individual machines and storage	Hardware		



# Arquitectura de Minería de Datos

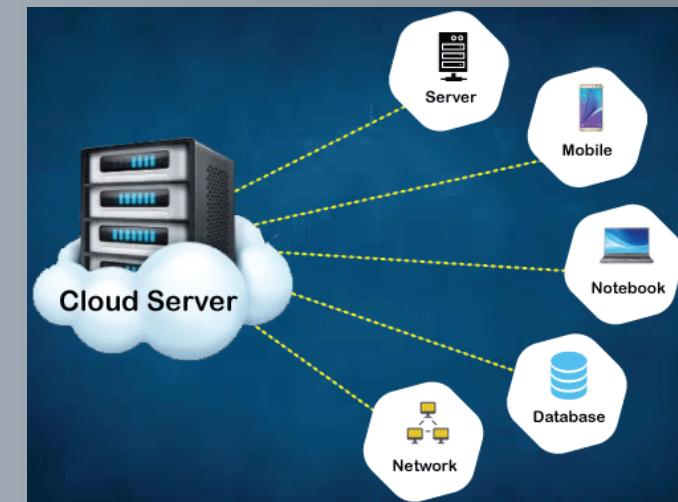
## Servidores de Datos      Cloud Databases



Una base de datos en Cloud es independiente creada, implementada y para que se acceda a través de un entorno de nube. Una base de datos en la nube tiene todas las funcionalidades de una base de datos tradicional, junto con la flexibilidad de la computación en la nube.

Algunos de los principales **beneficios** de las bases de datos en la nube son:

- **Implementación rápida.** Las bases de datos en la nube eliminan la necesidad de comprar e instalar hardware y configurar una red.
- **Accesibilidad.** Los usuarios tienen acceso rápido a las bases de datos en la nube de forma remota a través de la API o la interfaz web del proveedor.
- **Escalabilidad.** Puede ampliar la capacidad de almacenamiento de la base de datos en la nube sin interrupciones y cumplir con los requisitos
- **Recuperación ante desastres.** Las copias de seguridad de datos se realizan regularmente en bases de datos en la nube y se mantienen en servidores remotos.
- **Menores costos de hardware.** Los proveedores suministran la infraestructura y realizan el mantenimiento de la base de datos.
- **Relación calidad-precio.** permite a las empresas pagar solo por lo que usan y desactivar los servicios cuando no los necesitan.
- **Seguridad.** La mayoría de los proveedores de bases de datos en la nube cifran los datos en las soluciones de seguridad en la nube para mantenerlas seguras.



# Arquitectura de Minería de Datos

## Servidores de Datos      Cloud Databases



**SAP HANA Cloud** es una solución de base de datos como servicio (DBaaS) en la nube totalmente administrada, escalable y en memoria. La base de datos se puede implementar en la nube o en un entorno híbrido. La solución de base de datos proporciona un alto rendimiento de procesamiento debido a las transacciones híbridas multimodelo. Los usuarios reciben parches de software regularmente, las copias de seguridad están automatizadas y no hay necesidad de preocuparse por las instalaciones de software requeridas. La desventaja es que SAP HANA Cloud requiere ingenieros de bases de datos experimentados, y la capacitación necesaria en la nube es costosa.

**IBM Db2** en la nube es una base de datos SQL totalmente gestionada con un SLA de tiempo de actividad del 99,99%, almacenamiento independiente y escalado de computación a través de UI y API, varias opciones de recuperación ante desastres, cifrado de datos y otras características. La base de datos relacional de IBM ofrece gestión avanzada de datos y capacidades analíticas para cargas de trabajo transaccionales y de almacenamiento. Esta base de datos ofrece un alto rendimiento, cuenta con grandes conocimientos, disponibilidad de datos, confiabilidad y amplia compatibilidad con el sistema operativo. La desventaja de IBM Db2 es que tiene menos opciones regionales, lo que afecta el rendimiento en algunos casos.





**Google Cloud** ofrece varios servicios que utilizan el mismo hardware e infraestructura que otros productos de Google. La oferta de GCP incluye una amplia gama de servicios alojados para computación en la nube, almacenamiento, redes, big data, aprendizaje automático, IoT, gestión de la nube, etc.

GCP proporciona IaaS, PaaS y entornos informáticos sin servidor. Uno de los productos de Google Cloud Platform es Cloud Datastore, una solución de almacenamiento de bases de datos para almacenamiento no relacional NoSQL. Otros productos de Google Cloud son Cloud SQL para el almacenamiento totalmente relacional MySQL y la base de datos nativa Cloud Bigtable de Google.

La desventaja es la falta de servicios administrados y los altos precios, incluida una costosa tarifa de soporte.

**Oracle** ofrece tecnología de base de datos en la nube a escala empresarial a sus usuarios. La solución de base de datos utiliza el aprendizaje automático para automatizar la administración de bases de datos, lo que garantiza un alto rendimiento, confiabilidad y seguridad. La base de datos en la nube de Oracle cubre cargas de trabajo de Big Data y Streaming a hiperescala, incluidos OLTP, almacenamiento de datos, Spark, búsqueda de texto, análisis de imágenes y catálogo de datos. Las diferentes soluciones ofrecidas son Infraestructura como Servicio (IaaS), Plataforma como Servicio (PaaS), Software como Servicio (SaaS) y Datos como Servicio (DaaS). La desventaja es la falta de integración con otras soluciones en la nube.



# Arquitectura de Minería de Datos

## Servidores de Base de Datos      Cloud Databases



**Amazon Web Service (AWS)** es uno de los líderes del mercado cuando se trata de DBaaS. Amazon ofrece varios servicios para la gestión e integración de datos. Algunas de las soluciones de bases de datos de AWS son: Amazon RDS. Amazon Relational Database Service se ejecuta en instancias de servidor Oracle, SQL o MySQL. Amazon SimpleDB. Diseñado para cargas de trabajo más pequeñas, SimpleDB es principalmente una base de datos sin esquema. Amazon DynamoDB. DynamoDB es una base de datos NoSQL capaz de replicar automáticamente las cargas de trabajo en tres zonas de disponibilidad. La desventaja es que las operaciones de escalado y aplicación de parches requieren tiempo de inactividad.

**Microsoft Azure** La base de datos en la nube de Microsoft Azure es una de las plataformas en la nube más populares y extendidas a nivel mundial. Ofrece servicios de computación, redes, bases de datos, análisis, IA e IoT. La plataforma de computación en la nube pública de Microsoft ofrece varias soluciones, incluida la infraestructura como servicio (IaaS), la plataforma como servicio (PaaS) y el software como servicio (SaaS). Microsoft Azure ofrece una amplia gama de soluciones de software que permiten a los usuarios crear un vasto ecosistema con la misma base, lo que hace que cualquier problema sea fácil de resolver. La desventaja es que Azure debe administrarse y mantenerse de manera experta, incluida la aplicación de revisiones y la supervisión del servidor.





Sin embargo, hay algunas características clave que debe buscar al elegir una base de datos en la nube:

- **Rendimiento.** Una base de datos con escalabilidad en línea e independiente garantiza que la carga de trabajo y las necesidades de su empresa se satisfagan en todo momento, la disponibilidad ininterrumpida de datos durante el crecimiento es un factor importante.
- **Servicios automatizados** y la optimización del rendimiento en línea son características necesarias que garantizan que todo funcione sin problemas. La indexación automática es uno de esos servicios, que proporciona una recuperación rápida de datos mediante el mantenimiento y el uso automáticos de índices.
- **Seguridad.** El cifrado de datos y las actualizaciones de seguridad automatizadas son imprescindibles al elegir una base de datos en la nube.
- **Compatibilidad.** Una base de datos debe tener una amplia compatibilidad con aplicaciones de terceros para garantizar que todo funcione correctamente.
- **Aislamiento de hardware.** Para las aplicaciones críticas para el negocio, se recomienda tener una *infraestructura de nube dedicada* con hardware aislado de otros inquilinos.
- **Copia de seguridad.** El proveedor de la base de datos en la nube debe ofrecer copias de seguridad periódicas de los datos almacenados en múltiples ubicaciones geográficamente dispersas para evitar la pérdida de datos en caso de desastre.





# Arquitectura de Minería de Datos

## Servidores de Base de Datos

### Big Data DB

Hoy en día, casi más de 7.000 millones de dispositivos comparten información por Internet. Se estima que esta cifra se elevará hasta los 20.000 millones en 2025. En este sentido, el Big Data se encarga de analizar este océano de datos para convertirlo en la información que está transformando el mundo.

Las necesidades de gigantes como Google fueron incrementándose con el paso del tiempo. En un momento dado, se tuvieron que plantear **qué hacer con tanta cantidad de datos y cómo sacar provecho de los mismos**. Esto les llevó a comprender que, si analizaban toda la información que recopilaban, podían llegar a entender mejor el mercado y a **crear estrategias personalizadas** en base a esos datos con el objetivo de satisfacer mejor las necesidades de los consumidores.

Google Cloud dispone de otras soluciones de Big Data que te permiten desarrollar aplicaciones adaptadas al contexto, incorporar la inteligencia artificial y convertir los datos en métricas útiles.





# Arquitectura de Minería de Datos

## Servidores de Base de Datos

### Big Data DB

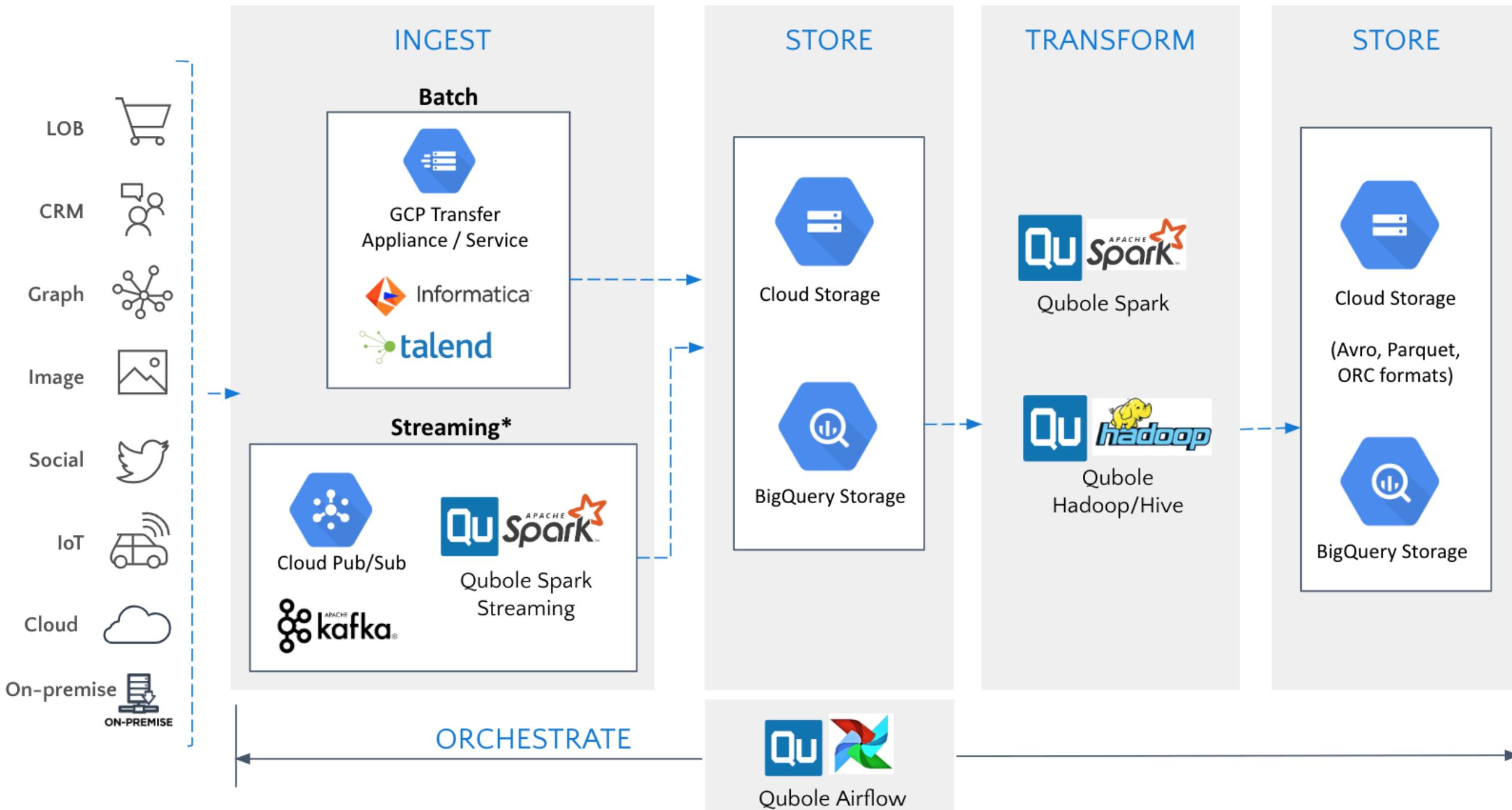
Big data puede hacer referencia tanto a un conjunto de datos grandes como complejos, así como los métodos usados para procesar este tipo de datos. Big data tiene las siguientes características:

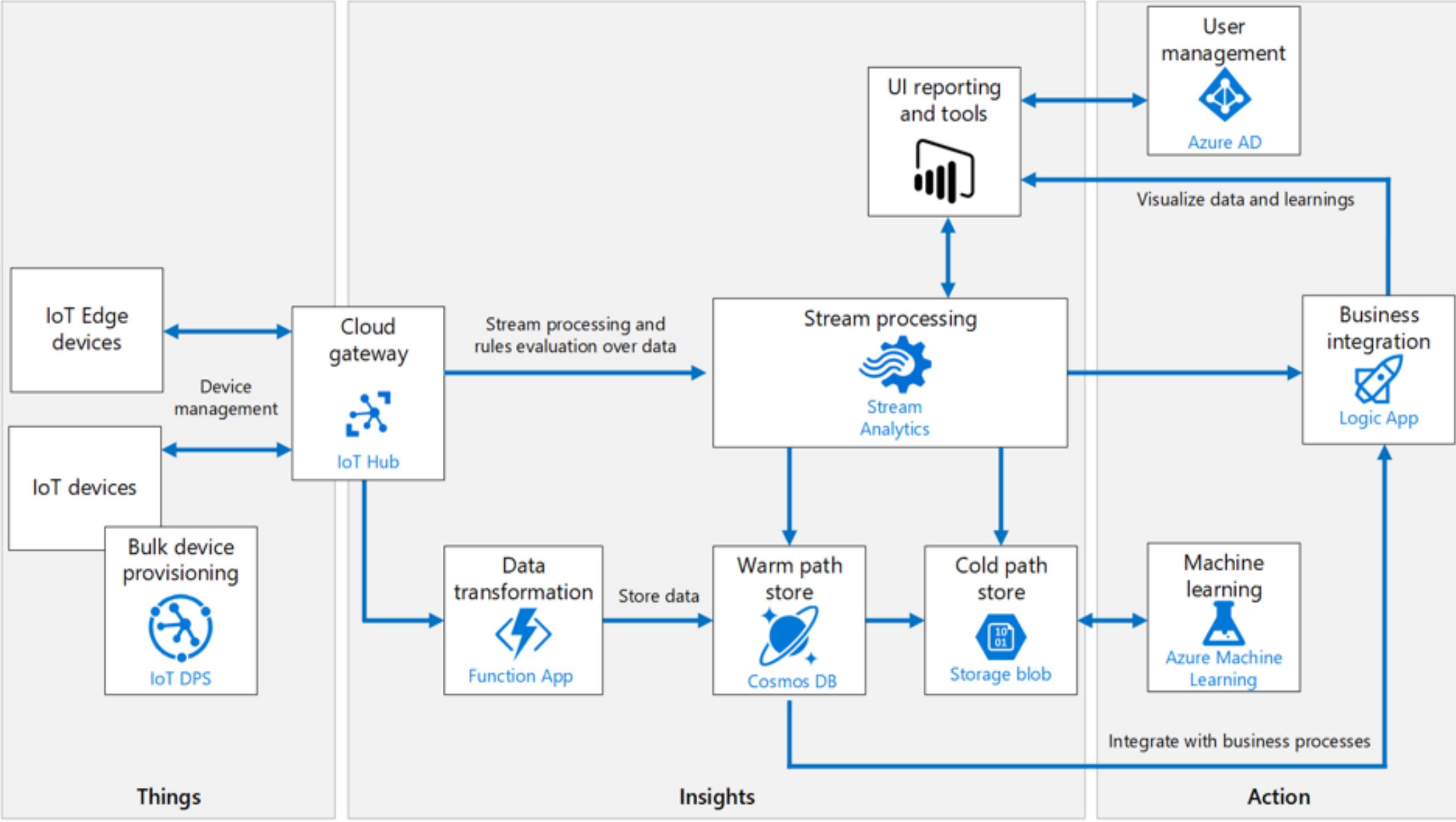
- **Volumen:** Big data maneja grandes volúmenes
- **Variedad:** contiene datos estructurados, semiestructurados y no estructurados.
- **Velocidad:** se generan rápidamente y por lo general se procesan en tiempo real.
- **Veracidad:** su veracidad (precisión) es extremadamente importante. Las anomalías, los sesgos y los ruidos pueden impactar en la calidad.
- **Valor:** saber cómo tratar la data que se recopila para sacarle un valor a la misma que ayude a tomar decisiones acertadas.
- **Variabilidad:** las diferentes interpretaciones que pueden resultar en el proceso.
- **Visión:** el poder tener una visión clara de cómo proceder en base a los diferentes patrones e interpretaciones de comportamiento del consumidor.

#### Diferencias entre Big Data y datos tradicionales

Se usan varias características para distinguir entre big data y datos tradicionales. Entre ellas se incluyen: El tamaño de los datos, Cómo se organizan los datos, La arquitectura requerida para administrar los datos, Las fuentes desde las cuales derivan los datos, Los métodos utilizados para analizar los datos.









# Arquitectura de Minería de Datos Base de Datos de Conocimiento



## Base de Datos de Conocimiento – Descripción



Una base de conocimientos es una biblioteca en línea de autoservicio de información sobre un producto, servicio o tema.

Por lo general, los colaboradores en los temas relevantes suman y amplían la base de conocimientos. El contenido puede variar desde los tópicos de su departamento legal o de recursos humanos hasta una explicación de cómo funciona un producto.

- La **base de conocimientos** puede incluir preguntas frecuentes, manuales, guías de solución de problemas y otra información que su equipo desee o necesite saber.
- Muchas **bases de conocimiento** están estructuradas en torno a la inteligencia artificial que puede interactuar y responder a la entrada del usuario.
- O ser simplemente una enciclopedia indexada.
- También hay **bases de conocimiento** que se basan en lo que llamamos razonamiento deductivo automatizado. Cuando un usuario introduce una consulta, el software ayuda a proponer una solución.
- Una **base de conocimientos** es la gestión del conocimiento que permite crear, curar, compartir, utilizar y administrar el conocimiento en toda su empresa y en todas las industrias.



# Arquitectura de Minería de Datos

## Base de Datos de Conocimiento

Por qué necesita una base de conocimientos



- En el mundo conectado de hoy, las personas esperan y exigen un **fácil acceso** a información precisa.
- En ocasiones las personas no están dispuestas a recibir una llamada telefónica. O se envíe un correo electrónico. O se presente un ticket de servicio. Quieren la respuesta que necesitan de inmediato. Es por eso que necesita una base de conocimientos rica y profunda.
- Las organizaciones están encontrando más usos. La forma en que utiliza una base de conocimientos depende, por supuesto, de lo que hace su organización y a quién sirve. Pero estas son algunas de las formas en que las bases de conocimiento están demostrando ser invaluables para varios equipos como, por ejemplo:
  - **TI:** simplifica todo, desde la solución de problemas hasta la capacitación / incorporación y las preguntas generales de cómo hacerlo y soporte
  - **HR:** Una vez más, ideal para todo, desde capacitación / incorporación hasta distribución de políticas de la empresa y horarios de pago.
  - **Legal:** Ayuda con contratos y otros procesos de aprobación, políticas, marcas comerciales y registros



# Arquitectura de Minería de Datos

## Base de Datos de Conocimiento

Estas son algunas de las formas en que una base de conocimientos puede marcar la diferencia para su organización:

- Menores costos de capacitación.
- Una base de conocimientos, respaldada por un sólido programa de gestión del conocimiento, garantiza que los nuevos empleados reciban capacitación con la información más reciente y obtengan una orientación consistente.
- Una base de conocimientos también puede:
  - Organizar todo lo que la gente necesita saber en un solo lugar
  - Estandarizar las respuestas
  - Lograr que la empresa sea inteligente, actualizada y profesional

## Beneficios de una base de conocimientos





### Como construir y mantener una base de conocimientos

#### 1. Determinar que necesitas

Comience por preguntarse cuánto tiempo ahorraría si los empleados no tuvieran que responder las mismas preguntas una y otra vez. En base a los objetivos de satisfacción y productividad la organización podría hacerlo mejor, una base de conocimientos es un excelente lugar para comenzar.

#### 2. Reunir su contenido.

Esta no es una tarea fácil: el contenido está en todas partes. Recopile preguntas frecuentes (y respuestas) de cualquier departamento que brinde servicio. Los equipos de toda la organización pueden y deben contribuir a su base de conocimientos. Y deben ser parte del proceso de gestión del conocimiento que lo mantiene.

#### 3. Personalizar y estandarizar las páginas

Cree una guía de estilo, de modo que toda la información que vierta en su base de conocimientos se vea y suene igual. Esto cubre toda la presentación visual, incluida la fuente, el tamaño del tipo, los colores e incluso las imágenes.

#### 4. Encontrar el lenguaje estandar.

Averigüe cómo habla su empresa u organización. Ya sea que esté pulido o tenga un ambiente relajado, úselo en la presentación de su base de conocimientos. Las personas en marketing pueden ayudar.





### Como construir y mantener una base de conocimientos

#### 5. Obtener las herramientas adecuadas para administrarlo.

Asegúrese de tener las herramientas adecuadas para alojar y administrar su base de conocimientos.

Todo, desde la frecuencia de los cambios de contenido hasta la forma en que los clientes suelen interactuar con su información, debe ser parte de su decisión.

#### 6. Hacerlo fácil. y mantenlo.

Una vez que esté en funcionamiento, recuerde que su base de conocimientos es una operación de autoservicio. Deberá asegurarse de que su base de conocimientos sea fácil de navegar. Y fácil de usar. Permitir a los colaboradores usar plantillas de velocidad para cargar datos. Use etiquetas y términos de búsqueda para categorizar la información y hacer que los artículos sean más fáciles de encontrar.

Organice el contenido para que se ajuste a su organización y luego tenga cuidado de mantenerlo.

#### 7. Manténgalo relevante. Y al día.

Aquí es donde la creación de su base de conocimientos fluye hacia la tarea continua de la gestión del conocimiento. Implementar un sistema de análisis, para que comprenda cómo las personas usan su contenido. Permite a los usuarios dejar comentarios y calificaciones. Asegúrese de que su personal de administración y marketing tenga voz en la administración de la base de conocimientos. Establecer Administradores y déjelos actuar cuando la información deba eliminarse, agregarse o cambiarse.



# Arquitectura de Minería de Datos

## Base de Datos de Conocimiento

### Datos que se incluyen en una base de conocimiento

Todas las bases de conocimiento comparten información valiosa con clientes y prospectos, pero el tipo de datos e información que incluya en ella depende del propósito de su negocio para crear una. Los tipos comunes de datos incluidos en una base de conocimiento son:

- Instrucciones y consejos para usar sus productos y servicios,
- Respuestas a preguntas frecuentes,
- Contenido que pueda proporcionar soluciones en mayor detalle,
- Demostraciones en vídeo,
- Información de la empresa,
- Conocimiento de diferentes departamentos de negocio.

Algunas empresas también crean contenido que es útil para los consumidores generales de la industria, no solo para clientes específicos. Esto proporciona información útil y expone su contenido a diferentes grupos de audiencia que pueden convertirse en clientes si se benefician de la información que se proporciona.



# Arquitectura de Minería de Datos

## Base de Datos de Conocimiento

### Importancia del diseño de la base de conocimiento



Si bien puede ser tentador comenzar a escribir y publicar artículos de inmediato, pensar un poco más en cómo presentar la información a sus clientes o empleados ayudará a garantizar que aprovechen al máximo el recurso.

Estas son algunas de las prácticas recomendadas que debe tener en cuenta al diseñar :

- Mantén la coherencia de tu marca. Utiliza los mismos colores, fuentes y logotipos en tu sitio web
- Diseña pensando en el usuario. Organice su contenido de manera lógica, priorice los recursos.
- Elabora una documentación clara y concisa en el lenguaje y el tono utilizados por la organización.
- Utiliza encabezados, listas, espacios en blanco, imágenes, videos y rótulos para mejorar la legibilidad.
- Construya su contenido a partir de datos. Comience su base de conocimientos con artículos que solo respondan a preguntas frecuentes conocidas.
- Luego, a medida que los lectores interactúan con su contenido determinar qué nuevos conocimientos deben agregarse para satisfacer mejor las necesidades de su lector.
- Construya una base de conocimientos teniendo en cuenta la accesibilidad.
- Agregar texto alternativo a las imágenes y un texto de enlace significativo a sus artículos



## Base de Datos de Conocimiento      Opciones



Centro de ayuda o wiki o simplemente una manera de organizar sus notas de trabajo, aquí hay 12 plataformas de software de base de conocimiento.

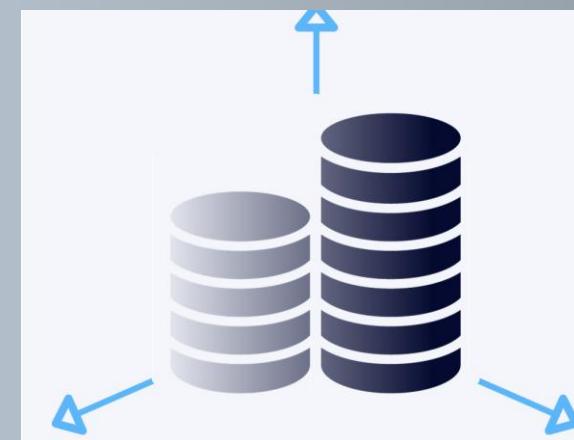
1. Help Scout
2. Guru
3. Document360
4. Obsidian
5. HelpJuice
6. Zendesk
7. BookStack
8. inSided
9. Notion
10. Confluence
11. Bloomfire
12. MediaWiki

- **flexible y fácil** de usar que le permite crear artículos de la base de conocimientos en minutos utilizando un editor de texto que ofrece múltiples opciones de formato tanto en WYSIWYG como en HTML.
- **pueda cargar rápidamente** imágenes y videos a los artículos para mejorar aún más el contenido,
- **pueda incrustar un widget de estilo chat** en cualquier página de su sitio o aplicación para que los visitantes
- **tenga acceso directo a su biblioteca** de base de
- **integra perfectamente** con nuestra bandeja de entrada compartida.
- **pueda compartir enlaces al contenido** sin necesidad de copiar y pegar, lo que reduce los tiempos de respuesta y las molestias.
- **que facilite crear contenido** que ocupa un lugar destacado en el motor de búsqueda de Google,





# Arquitectura de Minería de Datos Base de Datos Vectoriales



# Arquitectura de Minería de Datos

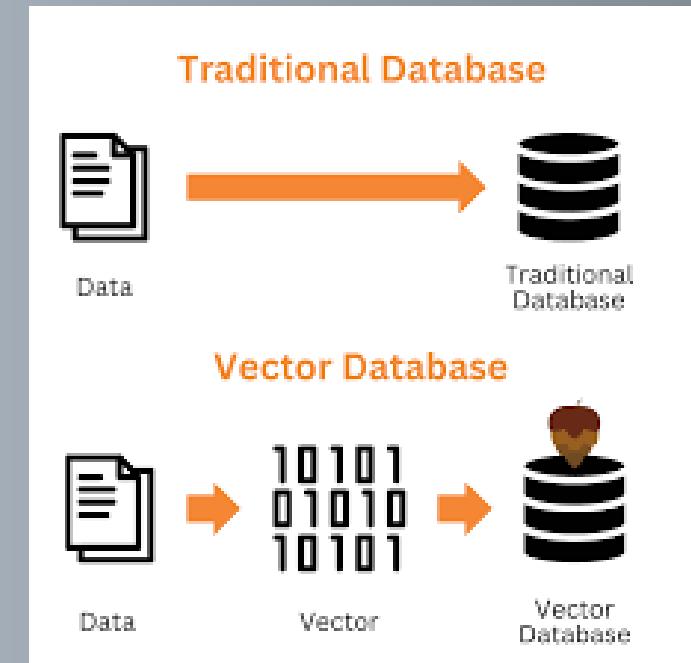
## Base de Datos Vectoriales



Una base de datos vectorial esta diseñada para simplificar la implementación de búsquedas vectoriales a gran escala y en tiempo real.

Es una solución completamente gestionada, lo que significa que los usuarios no tienen que preocuparse por la infraestructura, el mantenimiento, o la escalabilidad.

Permite a través de vectores relacionar imágenes generalmente implican la búsqueda de imágenes similares, clasificación de imágenes o recuperación de imágenes basadas en contenido visual.



# Arquitectura de Minería de Datos

## Base de Datos Vectoriales



### Características Principales:

- **Servicio Gestiónado:** en la nube eliminando la necesidad de que los usuarios gestionen la infraestructura subyacente.
- **Alta Escalabilidad:** Capaz de manejar miles de millones de vectores, lo que es ideal para aplicaciones de gran escala.
- **Búsqueda en Tiempo Real:** Optimizada para realizar consultas rápidas para aplicaciones que requieren respuestas inmediatas.
- **Soporte para Múltiples Dimensiones:** Maneja vectores de diferentes tamaños y dimensiones, adaptándose a diversas aplicaciones.
- **Integración con Pipelines de Machine Learning:** facilitando la implementación de modelos de IA que requieren búsquedas vectoriales.
- **Actualizaciones y Mantenimiento Automático:** garantizando que la plataforma esté siempre optimizada.



# Arquitectura de Minería de Datos

## Base de Datos Vectoriales



- 1. Búsqueda de Imágenes Similares:** Dado un conjunto de imágenes en una base de datos, un usuario puede subir una imagen nueva, y el sistema busca en la base de datos las imágenes más similares. **Convierte la imagen en un vector** mediante un modelo de aprendizaje profundo (como un modelo CNN) y luego compara este vector con los vectores de las imágenes almacenadas para encontrar las más similares.
- 2. Clasificación de Imágenes:** En un sistema de reconocimiento de imágenes, las imágenes se pueden clasificar en diferentes categorías (como "gato", "perro", "paisaje"). Se puede almacenar los vectores de características de las imágenes y luego, basándose en nuevas entradas, determinar a qué categoría pertenece una nueva imagen.
- 3. Recuperación de Imágenes Basada en Texto:** Un usuario puede realizar una búsqueda textual, como "paisajes con montañas", y el sistema devuelve imágenes que coincidan con esa descripción. Se puede almacenar vectores tanto de imágenes como de descripciones textuales y encontrar coincidencias entre ellos utilizando un espacio vectorial común.
- 4. Análisis de Imágenes para Detección de Anomalías:** En aplicaciones industriales, se pueden analizar imágenes de productos para detectar defectos. Se puede comparar las características de una imagen nueva con las de imágenes "normales" y detectar cualquier desviación significativa que indique un defecto.



# Arquitectura de Minería de Datos

## Base de Datos Vectoriales



Carga de vectores de imágenes implica varios pasos, desde la conversión de la imagen en un vector numérico hasta su inserción en la base de datos vectorial.

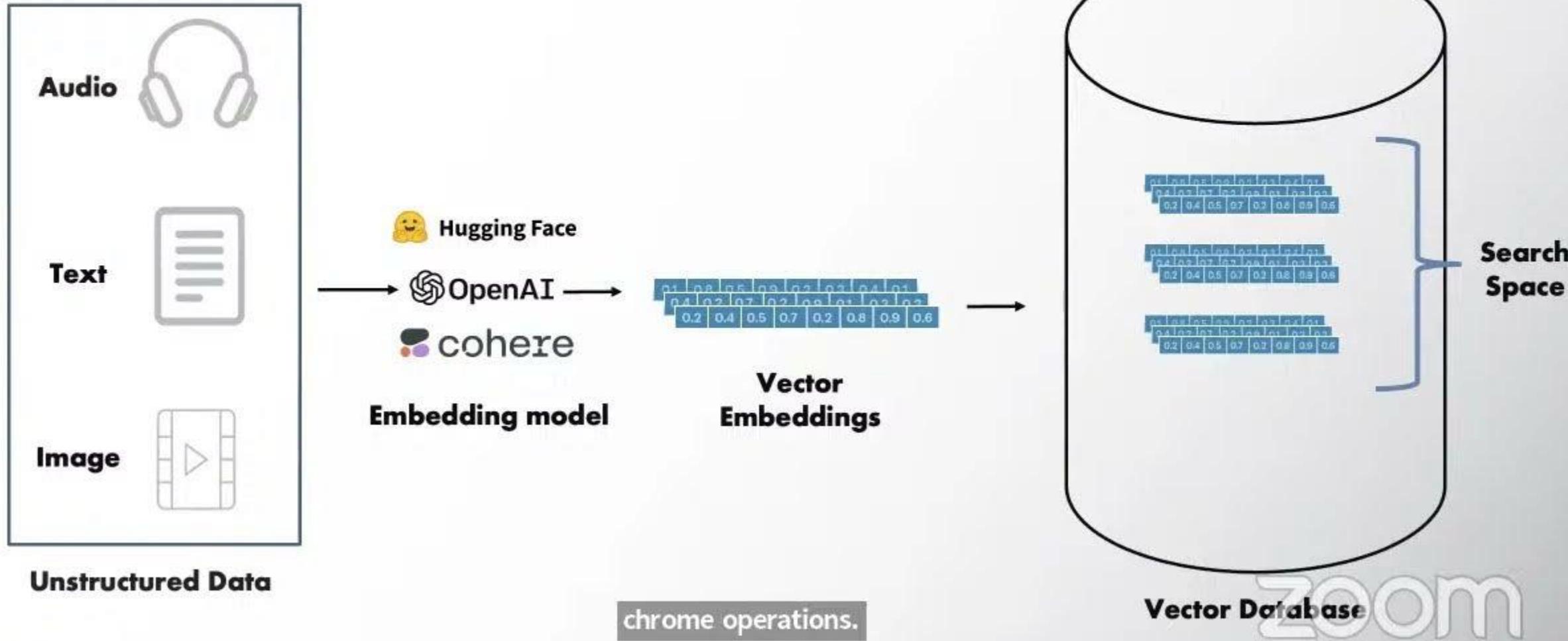
### Conversión de la Imagen a un Vector:

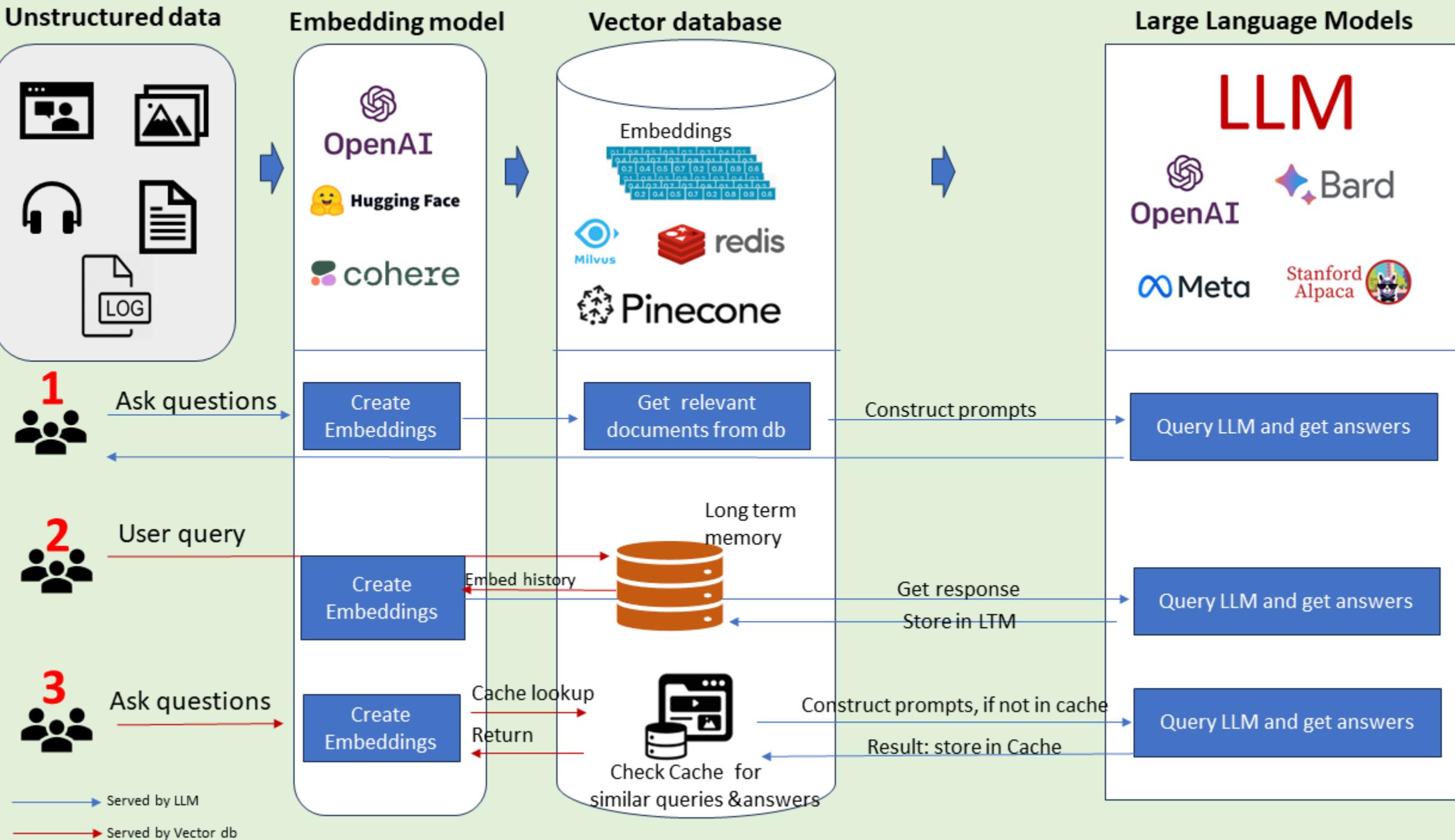
- Primero, necesitas convertir la imagen a un vector utilizando un modelo de aprendizaje profundo (Deep Learning).
- Los modelos más comunes para este propósito son las Convolutional Neural Networks (CNN), como **ResNet**, **Inception**, o **VGG**. Estos modelos están pre-entrenados en grandes conjuntos de datos y son capaces de extraer características visuales importantes de una imagen.
- La imagen se pasa a través de la red neuronal, y una de las capas finales, que contiene una representación numérica de la imagen, se toma como el vector característico (embedding).



# Vector Database

How are vector embeddings used in production?







# Arquitectura de Minería de Datos

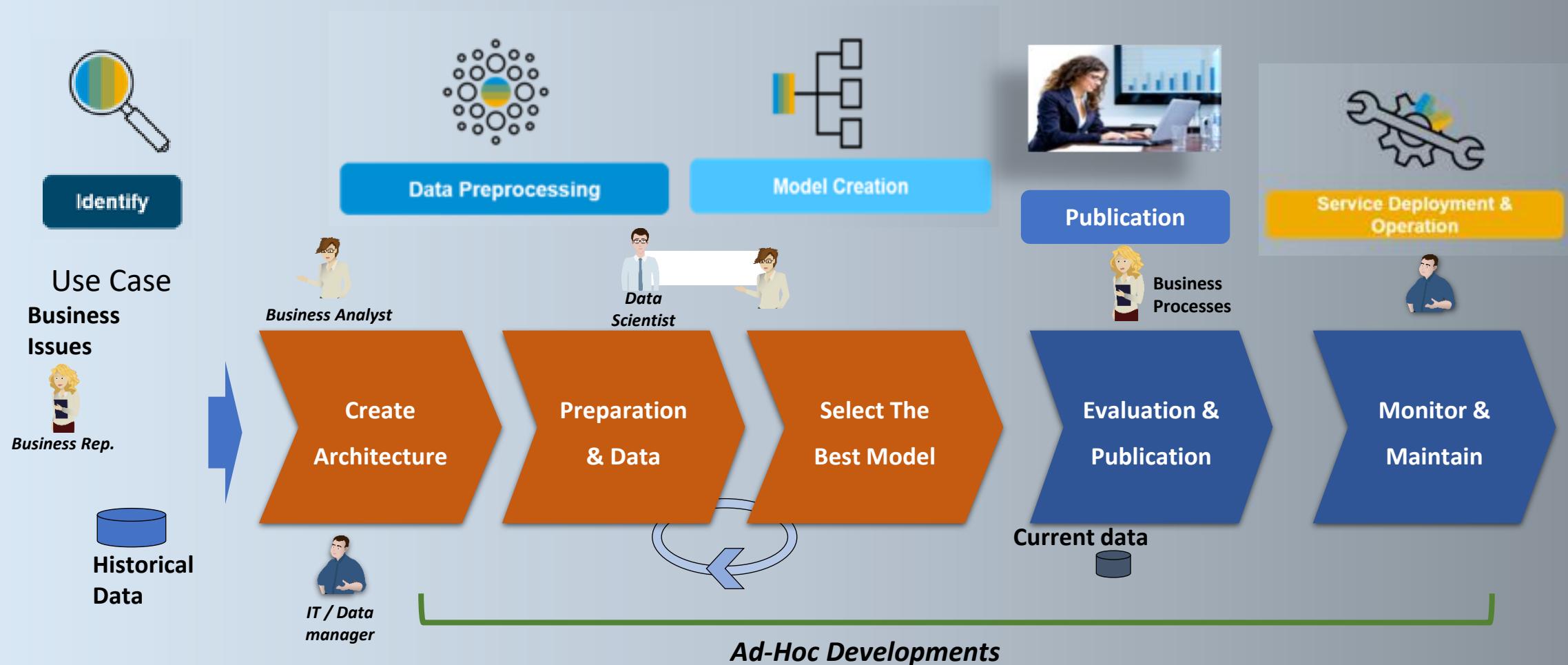
# Proceso de Minería de

# Datos

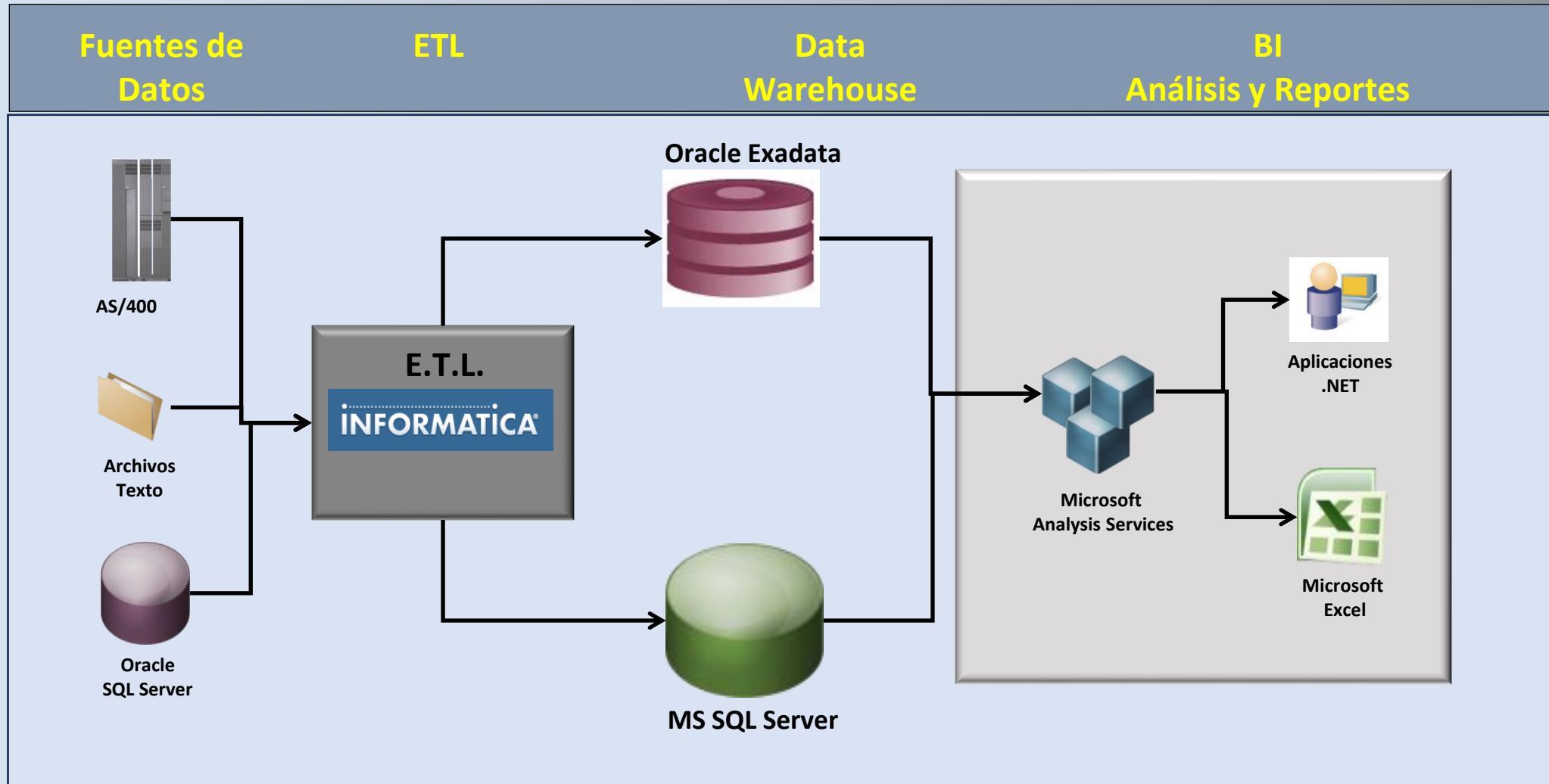


# Arquitectura de Minería de Datos

# Proceso de Minería de Datos



# Arquitectura Actual Banco



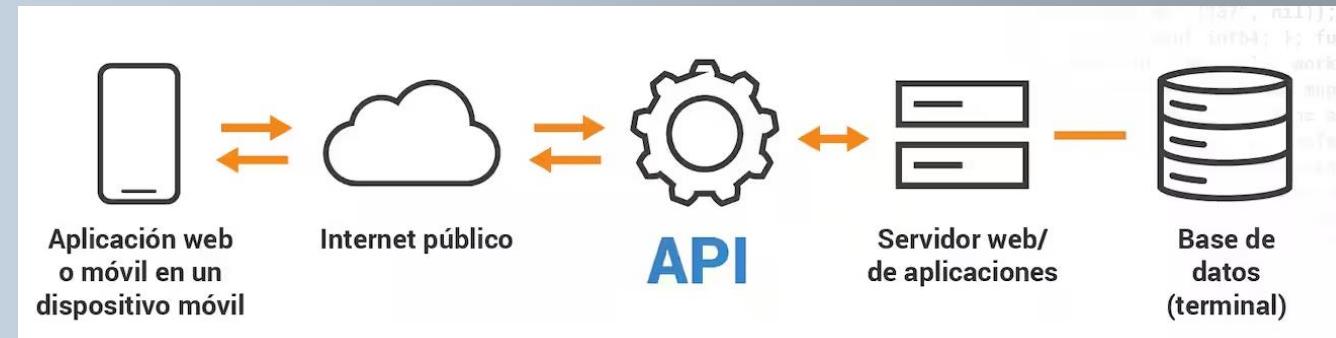
# Arquitectura de Minería de Datos Arquitectura – A.P.I.



Una API (Application Programming Interface) interfaz de programación de aplicaciones es una pieza de código que permite a diferentes aplicaciones comunicarse entre sí y compartir información y funcionalidades.

APIs utilizan el lenguaje JSON (JavaScript Object Notation) que es usado para pasar datos

Una API web o API de servicios web es una interfaz de procesamiento de aplicaciones entre un servidor web y un navegador web.



# Arquitectura de Minería de Datos Arquitectura – E.T.L.

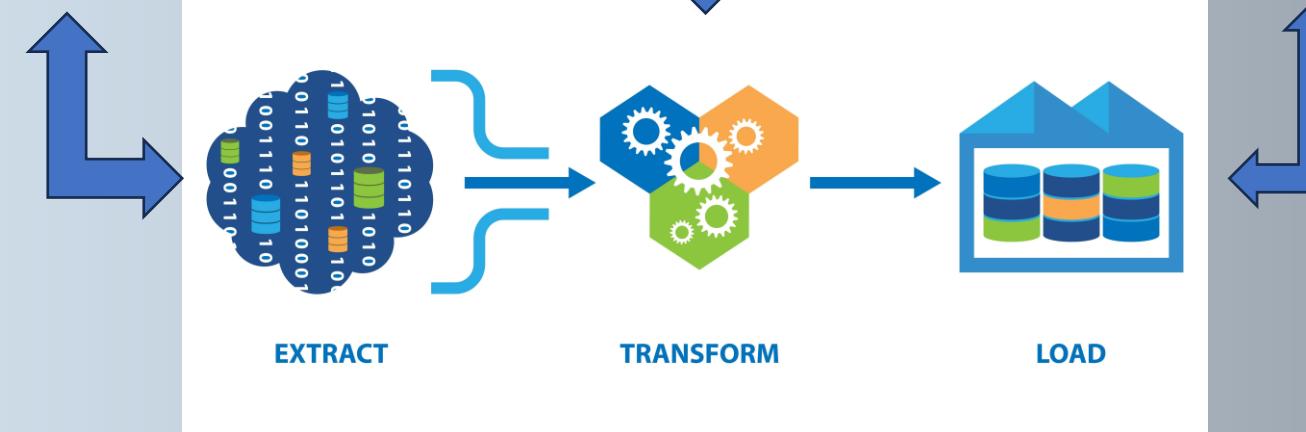


**Extract, Transform and Load** («extraer, transformar y cargar», frecuentemente abreviado **ETL**) es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra **base de datos**, **data mart**, o **data warehouse** para analizar, o en otro sistema operacional para apoyar un **proceso de negocio**.

Un requisito importante que se debe exigir a la **tarea de extracción** es que ésta cause un impacto mínimo en el sistema origen.

La **fase de transformación** aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados

La **fase de carga** es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino.



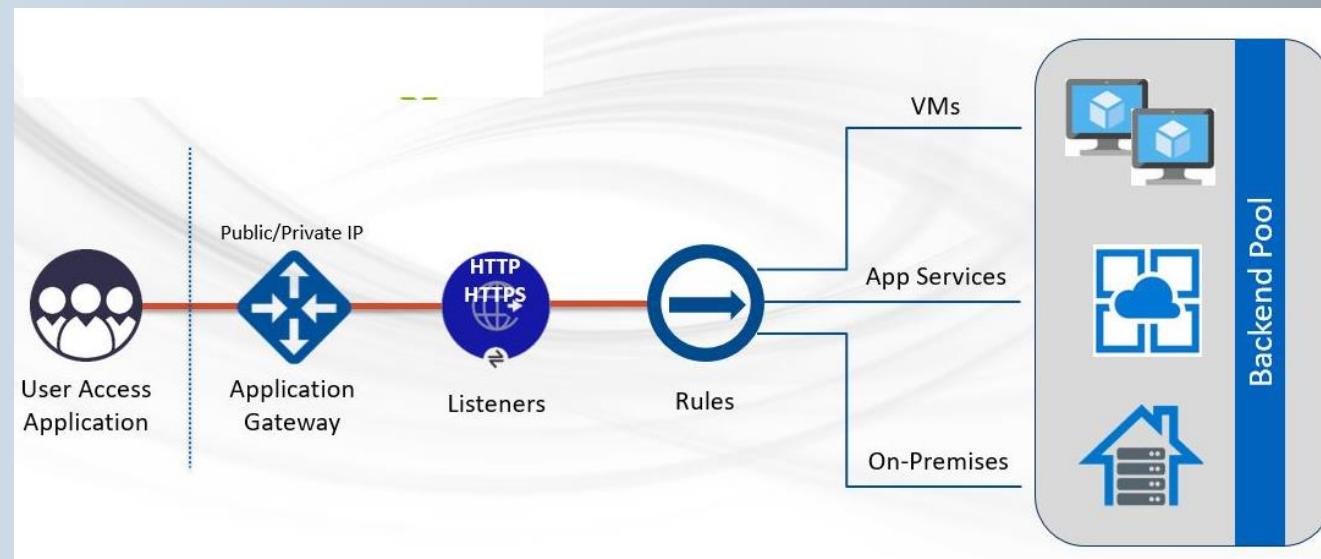
# Arquitectura de Minería de Datos Arquitectura – Gateway



Gateway es un tipo de enrutador que funciona como un punto de parada para los datos en su camino hacia otras redes. Gracias a los Gateway es posible la comunicación y envío de datos de un lado a otro.

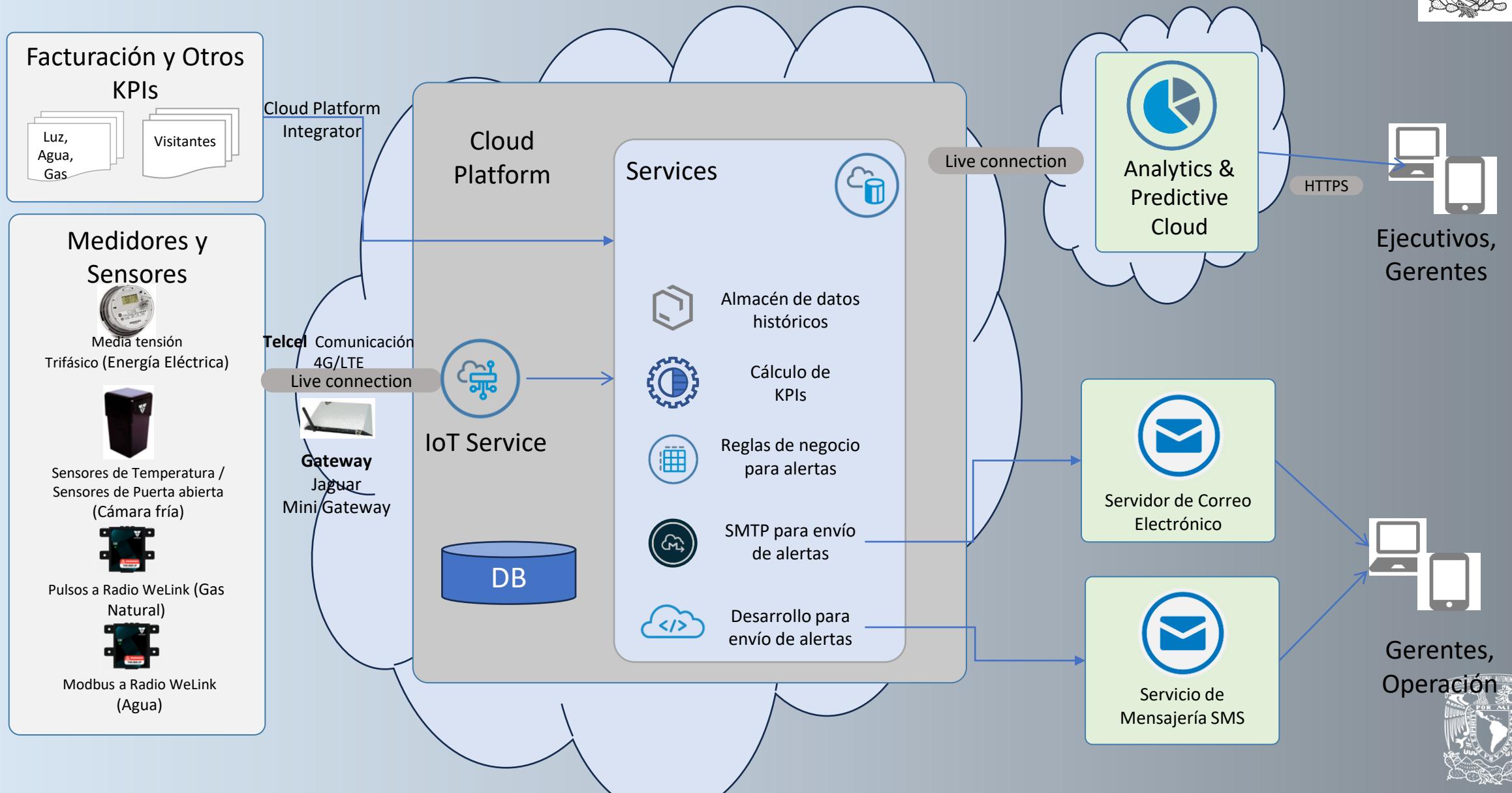
Un servicio Gateway es la vista de un servicio de destino que se otorga a un solicitante de servicios. Un servicio Gateway individual puede tener más de un servicio de destino asociado,

La gateway (puerta de enlace) es la ip del router que te da la conexión a internet y la tiene guardada tu pc para saber que ruta es la que debe seguir para ir a internet.



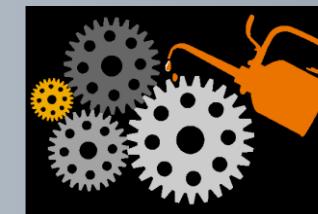


# Minería de Datos - Arquitectura de Solución





# Arquitectura de Minería de Datos Evaluación



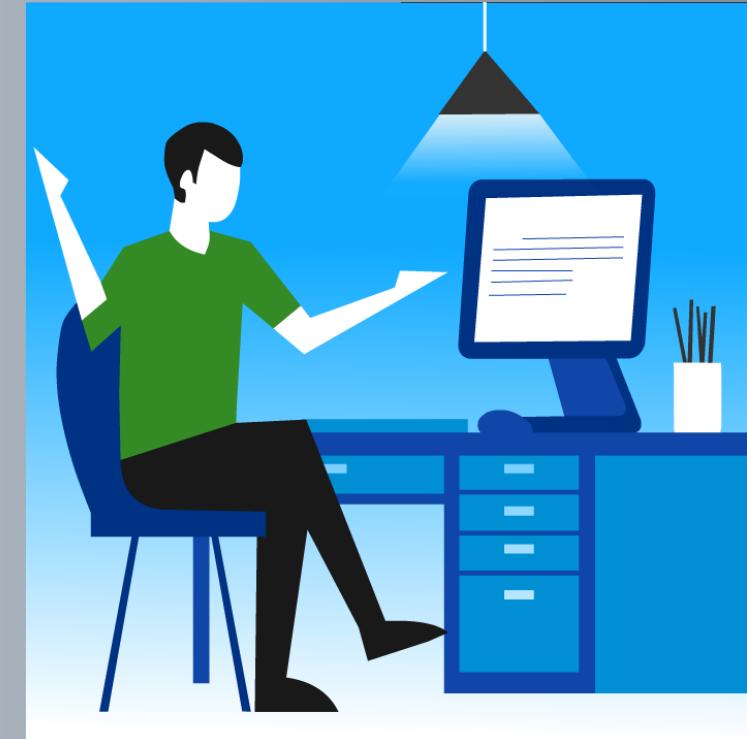
## Métricas de rendimiento de un modelo Predictivo



La elección de la métrica de rendimiento debe ser la que más se ajusta a los objetivos empresariales definidos en el inicio del proyecto durante la fase de entendimiento del negocio.

La métrica utilizada para la selección del modelo es de importancia crítica.

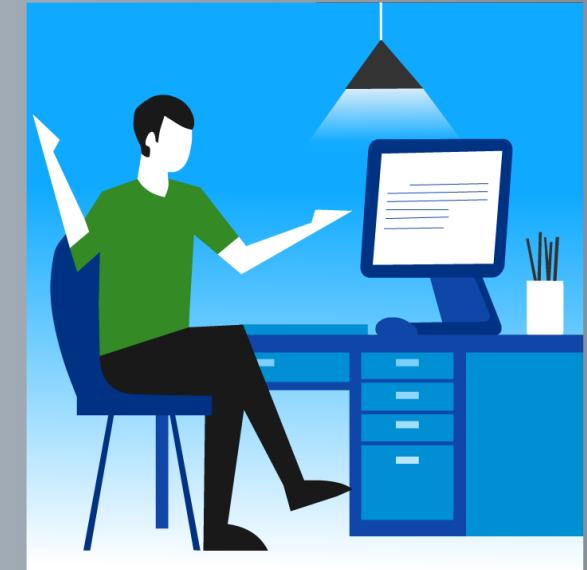
Porque el modelo seleccionado debe estar en función de una métrica, puede ser un buen modelo para una métrica diferente.





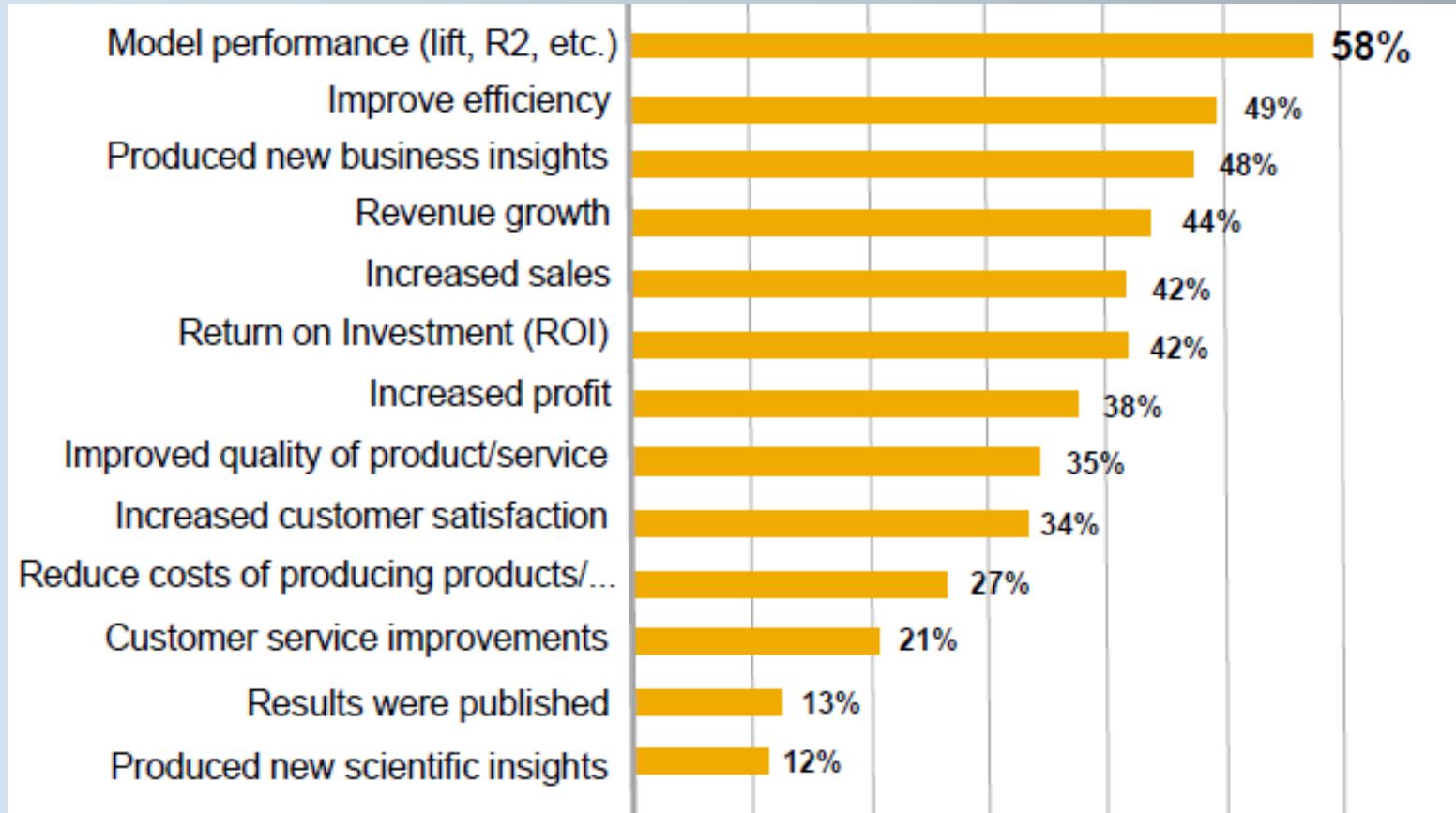
### Formas de probar la precisión de sus modelos predictivos

- Compare el rendimiento del modelo predictivo con resultados aleatorios con gráficos de Lift
- Evaluar la validez de su descubrimiento del modelo con la combinación de objetivos.
- Probar la coherencia del modelo predictivo con muestreo Bootstrap (prueba que utiliza muestreo aleatorio con reemplazo)





### Métricas de rendimiento de un modelo Predictivo



## Métricas que se pueden utilizar para evaluar un modelo de clasificación



Clasificación de corrección porcentual (PCC): mide la precisión general. Cada error tiene el mismo peso.

**Matriz de confusión:** mide la precisión pero distingue entre errores, es decir, falsos positivos, falsos negativos y predicciones correctas.

**Área bajo la curva ROC (ACC – ROC):** es una de las métricas más utilizadas para la evaluación. Popular porque clasifica las predicciones positivas más altas que las negativas. Además, la curva ROC es independiente del cambio en la proporción de respondedores.

**Gráficos de elevación y ganancia:** ambos gráficos miden la efectividad de un modelo calculando la relación entre los resultados obtenidos con y sin el modelo de evaluación del desempeño.

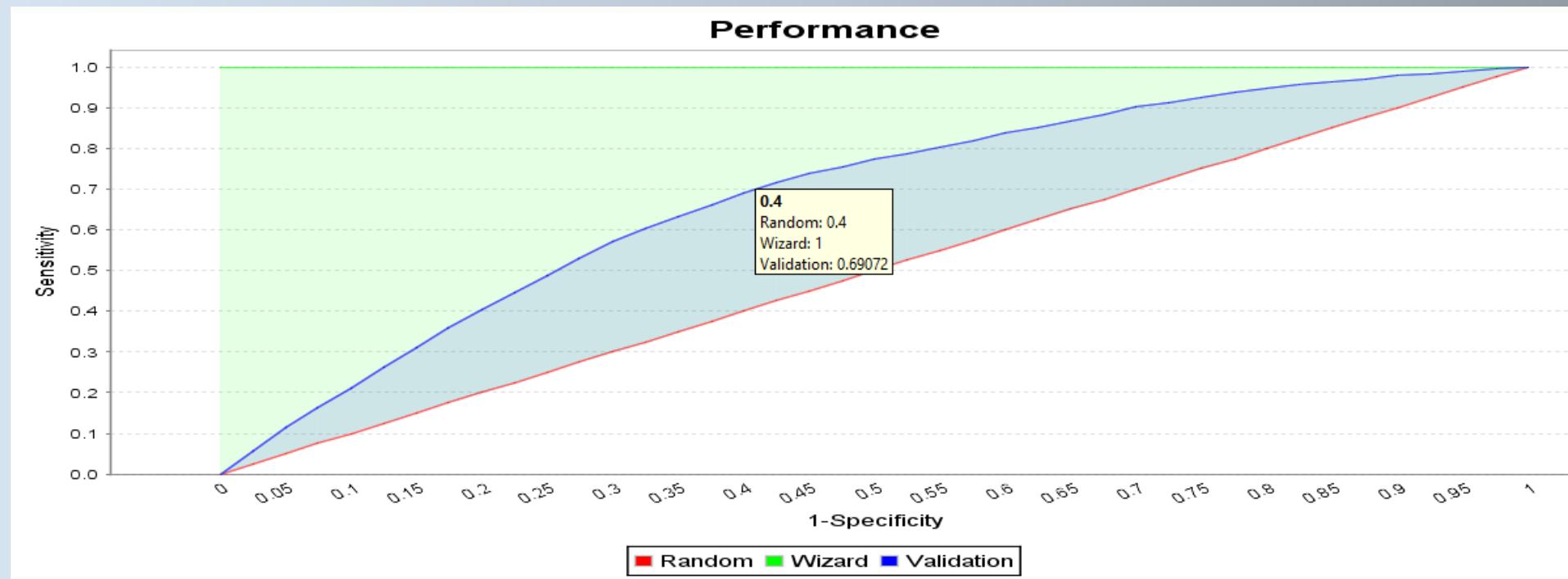
Estas métricas examinan si el uso de modelos predictivos tiene algún efecto positivo o no.



### ROC



Las curvas ROC (característica operativa del receptor) constituyen una herramienta importante para evaluar el rendimiento de un modelo de machine learning. Por lo general, se utilizan en problemas de clasificación binaria.

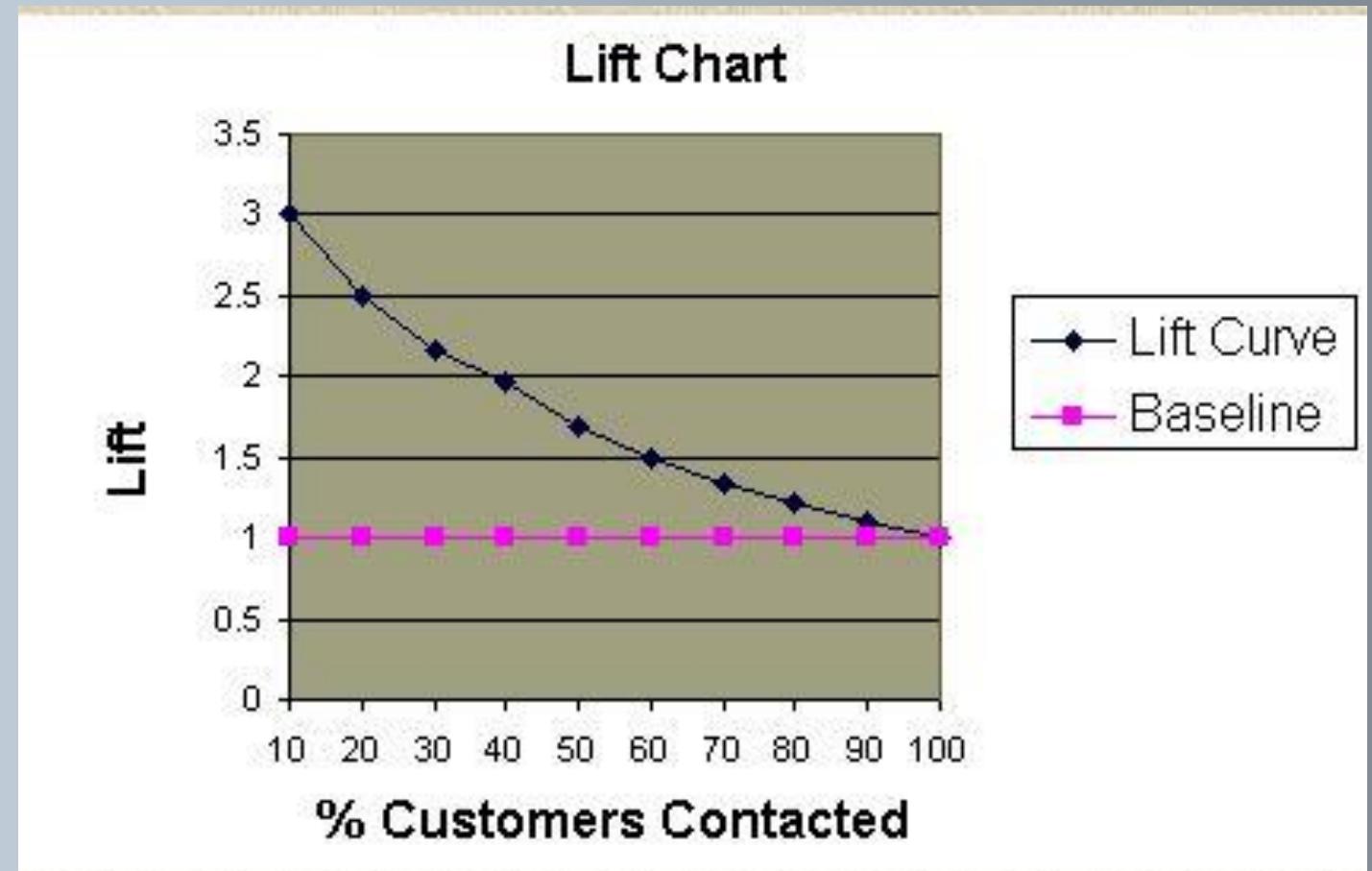


## LIFT

**Lift** es simplemente la proporción de estos valores: la respuesta objetivo dividida por la respuesta promedio.

Por ejemplo, supongamos que una población tiene una tasa de respuesta promedio del 5%, pero un determinado modelo ha identificado un segmento con una tasa de respuesta del 20%.

Entonces ese segmento tendría un aumento de 4,0 ( $20\% / 5\%$ ).



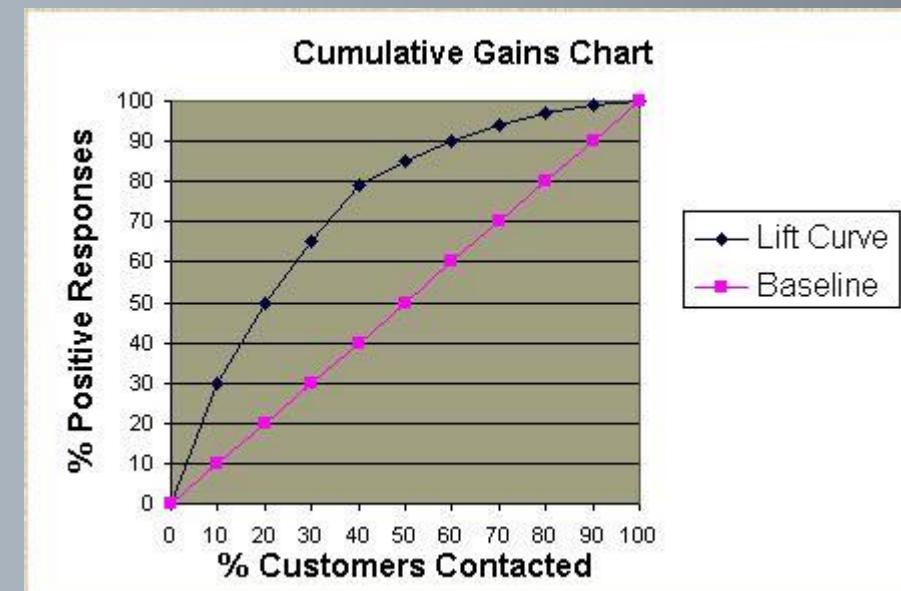
Lift es una medida de la efectividad de un modelo predictivo calculada como la relación entre los resultados obtenidos con y sin el modelo predictivo

### LIFT

Un gráfico de Lift elevación representa gráficamente la mejora que proporciona un modelo de minería de datos cuando **se compara** con una estimación aleatoria y mide el cambio en términos de una puntuación de elevación.

Al comparar las puntuaciones de Lift de diferentes modelos, puede determinar qué modelo es el mejor. También puede determinar el punto en el que las predicciones del modelo se vuelven **menos útiles**.

Por ejemplo, al revisar el gráfico de Lift, es probable que una campaña promocional **sea efectiva solo contra el 30%** de tus clientes y uses esa cifra para limitar el alcance de la campaña.



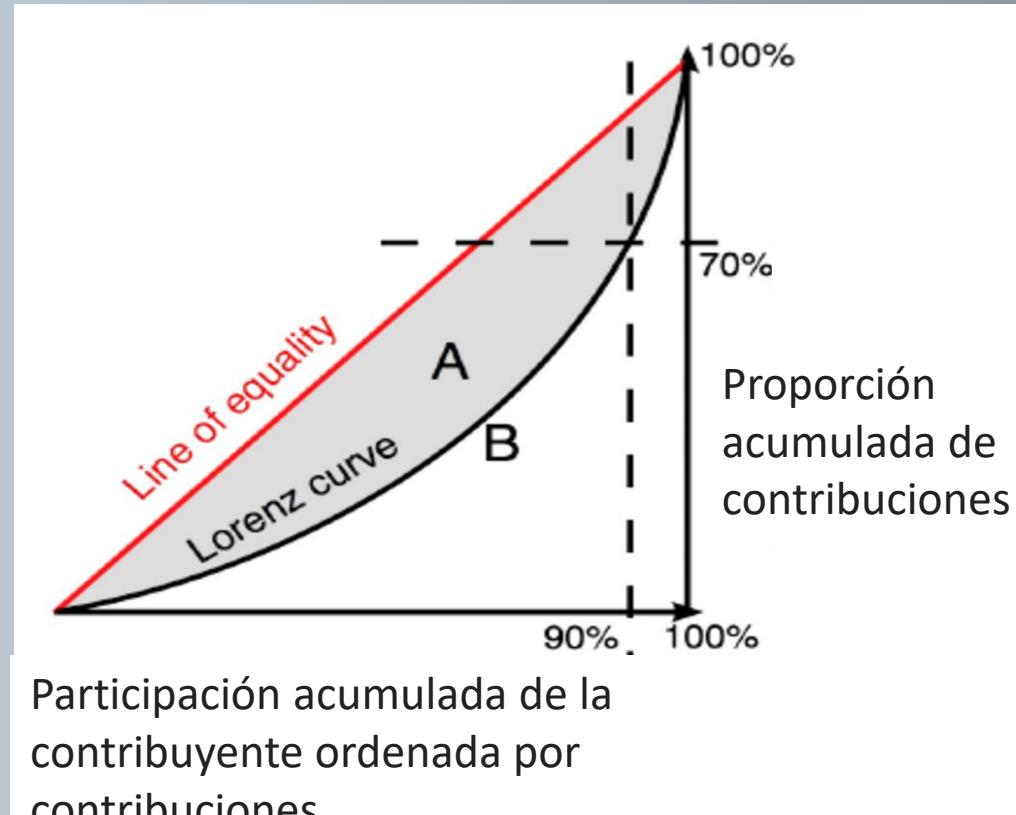


La curva de Lorenz se encuentra por debajo de la línea de igualdad de 45°. Cuanto más se aleje la curva de la línea de 45°, mayor será la desigualdad de renta o riqueza en una economía.

El primer paso para hacer la curva de Lorenz es ordenar la información de los ingresos de una determinada muestra de la población en forma ascendente. Se calcula el promedio de los ingresos, haciendo la sumatoria total y luego dividiendo entre la cantidad de la muestra.

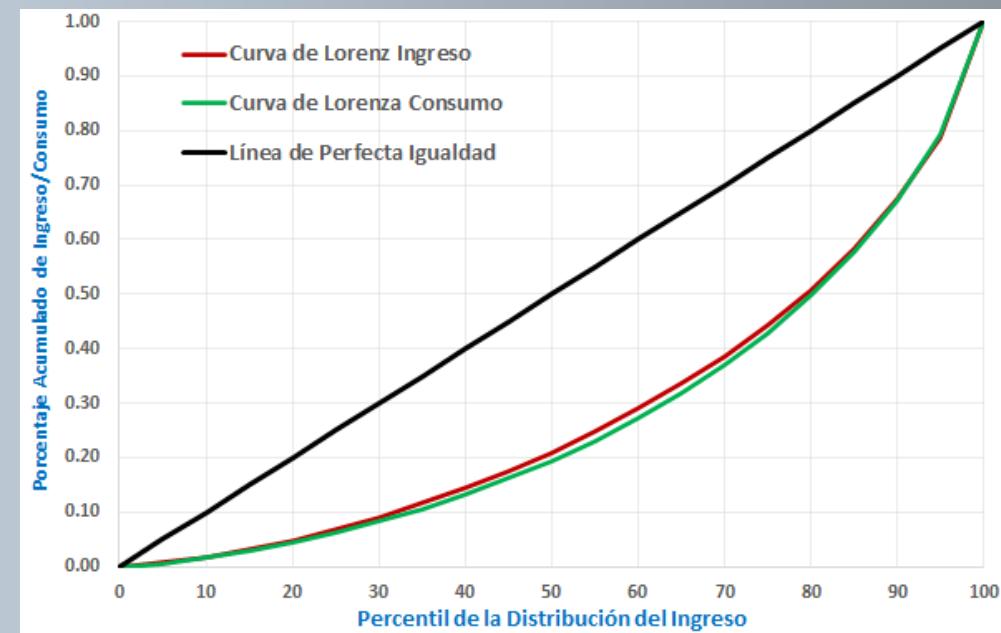
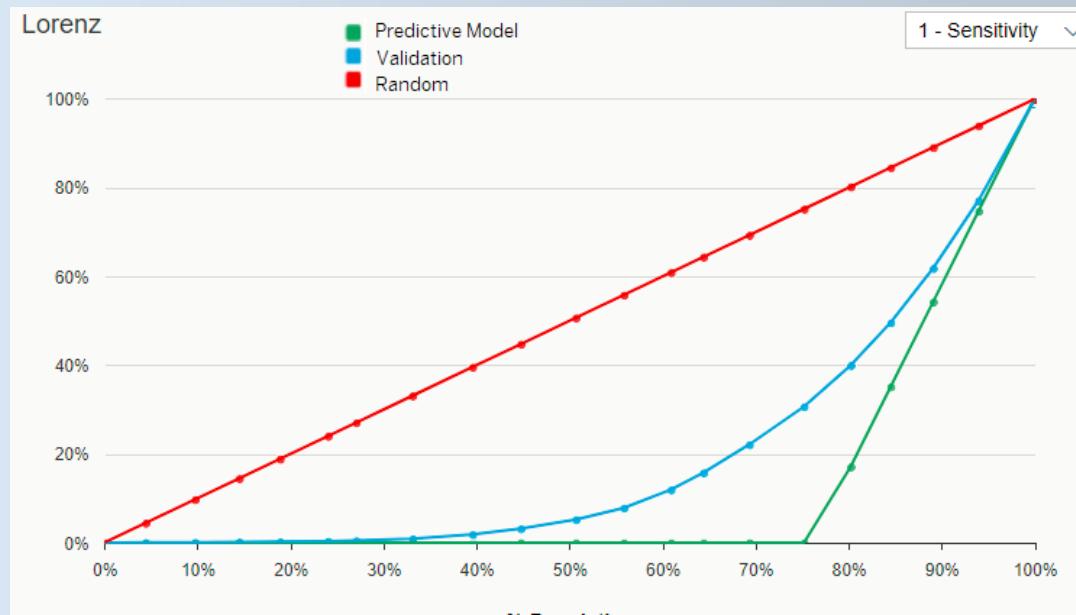


### LORENZ





Una curva de Lorenz para la riqueza de una población indica, por ejemplo, que el 50% menos rico de la población posee el 10% de la riqueza. (Si la riqueza se distribuyera equitativamente, la curva de Lorenz sería una línea recta).



## Matriz de confusión: métricas de uso común



		Predicted Class		Total
		1	0	
Actual Class	1	TP	FN	P
	0	FP	TN	N

<b>True Positive Rate, Hit Rate, Recall, Sensitivity</b>	TP/P	The proportion of positive instances that are correctly classified as positive
<b>False Positive Rate, False Alarm Rate</b>	FP/N	The proportion of negative instances that are erroneously classified as positive
<b>False Negative Rate</b>	FN/P	The proportion of positive instances that are erroneously classified as negative = 1 - True Positive Rate
<b>True Negative Rate</b>	TN/N	The proportion of negative instances that are correctly classified as negative



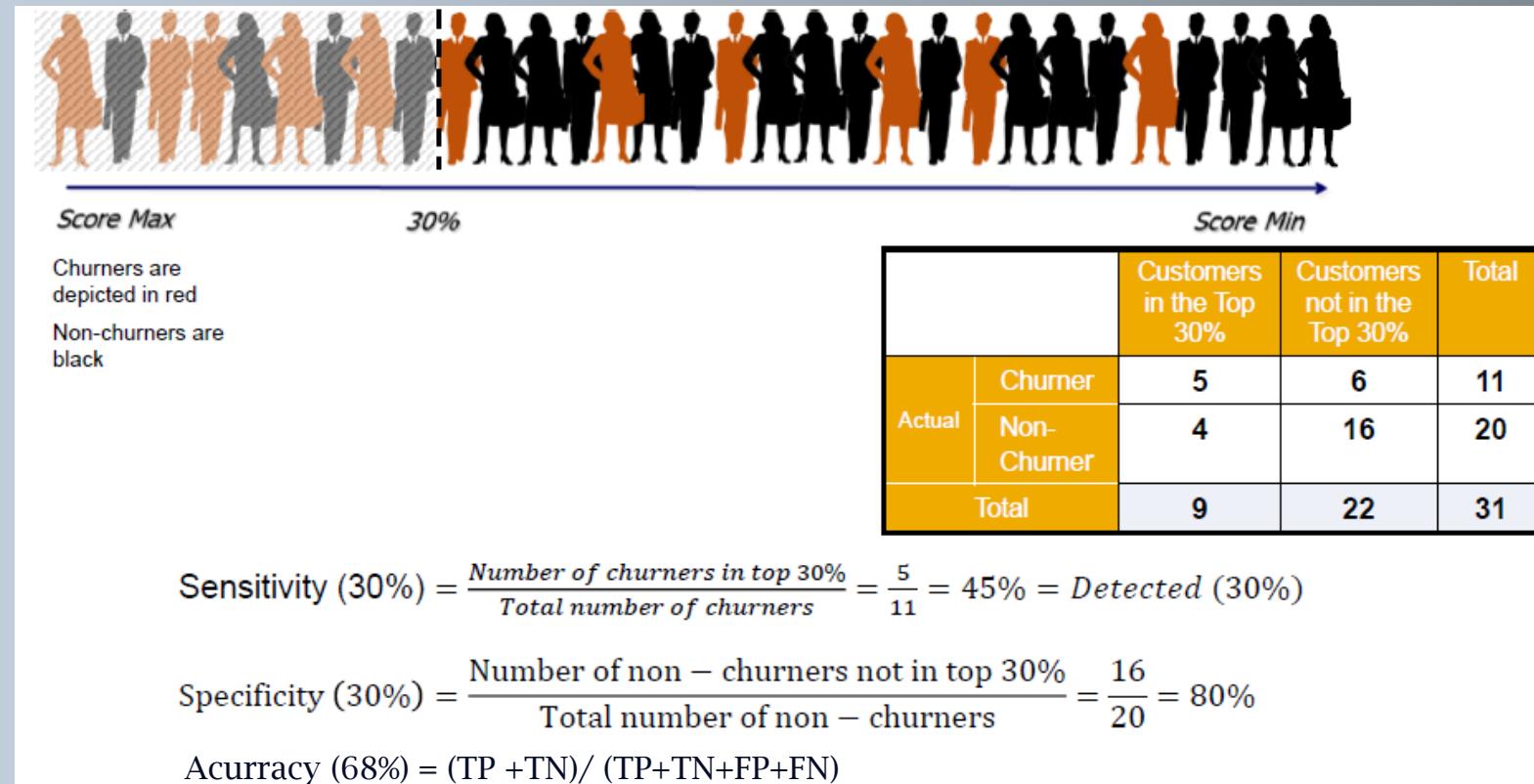
## Matriz de Confusión, sensibilidad, especificidad y Exactitud



La **sensibilidad** se refiere a designar a un individuo de churn como positivo.

La **especificidad** de una prueba es su capacidad para designar a un individuo que no es churn como negativo.

La **Exactitud** (Accuracy ) se mide en el total de ocurrencias entre el total de muestra .





# MINERIA DE DATOS

## Arquitectura de Minería de Datos

**José C Roberto Olvera López**  
Data Science Consultant

[jroberto.olveral@gmail.com](mailto:jroberto.olveral@gmail.com)

