



MODELADO AVANZADO DE BASE DE DATOS



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

UNIDAD 1

BASES DE DATOS NO RELACIONALES



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Origen y Calidad de Datos

Calidad de datos

La **calidad de datos** son los procesos, metodologías y algoritmos para aumentar el estado cualitativo de los datos que existen en una entidad empresarial u organización de cualquier índole.

Se considera que los **datos son de calidad** cuando estos cumplen con una serie de requisitos previamente estimados para aprovechar su utilización en los diferentes aspectos empresariales.

Poseer datos de calidad no es fácil, y es uno de los mayores inconvenientes que enfrenta una organización a la hora de procesar la información que dispone para mejorar las decisiones empresariales.

Los datos comprenden uno de los activos más importantes de una empresa, pero es aún más trascendental cuando estos son de calidad.



Niveles de calidad de datos

- ✓ **Mala calidad de datos:** Son aquellos que tienen **poca utilidad para utilizarse** en cualquier proceso empresarial. Se caracterizan por no ser fiables en los resultados que presenta, ya que carece de tratamientos, filtrados y mejoras para su aprovechamiento. Las consecuencias de confiar en datos de baja calidad pueden ser dramáticas, como tomar una decisión financiera corporativa incorrecta o provocar el fracaso de un proyecto.
- ✓ **Alta calidad de datos:** Son aquellos que **cumplen con todos los estándares de calidad** y que, por tanto, son utilizables para los diversos aspectos organizacionales. Permite que las tomas de decisiones sean fiables, reduciendo en gran medida los riesgos corporativos.



Atributos o cualidades definen la calidad de datos



Si bien no existe un conjunto universalmente acordado, algunos incluyen:

- ✓ **Precisión:** Mide con qué precisión sus datos reflejan el mundo real que representan. ¿Está seguro de que la edad registrada de un cliente es realmente su edad o podría tratarse de un error tipográfico?
- ✓ **Lo completo:** La integridad mide si falta alguna información esencial en sus datos. ¿Hay campos vacíos en el registro de un cliente o faltan valores?
- ✓ **Consistencia:** Coherencia significa que sus datos se adhieren a reglas y formatos predefinidos en diferentes plataformas y sistemas. ¿Todos los formatos de fecha son consistentes?
- ✓ **Oportunidad:** La puntualidad se refiere a la actualidad y relevancia de sus datos. ¿Están actualizados los datos?
- ✓ **Unicidad:** Verifica que todos los registros de su conjunto de datos sean distintos y no contengan duplicados. ¿Hay varias entradas para el mismo cliente con diferentes direcciones de correo electrónico?
- ✓ **Validez:** La validez verifica si los valores de los datos se encuentran dentro de rangos aceptables y cumplen con las restricciones definidas. ¿Los números de teléfono tienen el formato correcto?

Algunos marcos de calidad de datos también incluyen relevancia, integridad, granularidad y accesibilidad.

Métricas de calidad de datos

Una vez que haya identificado las dimensiones con las que desea medir la calidad de los datos, se debe traducir en métricas específicas y mensurables.

Ejemplos de métricas:

Métricas de precisión: Medir qué tan precisos son los conjuntos de datos.

- ✓ **Tasa de error:** porcentaje de puntos de datos que son incorrectos.
- ✓ **Tasa de coincidencia:** porcentaje de puntos de datos que coinciden con una fuente de verdad conocida.
- ✓ **Error absoluto medio:** diferencia promedio entre los puntos de datos y sus valores verdaderos.

Métricas de integridad: Medir la proporción de datos faltantes dentro de un conjunto de datos.

- ✓ **Porcentaje de valores faltantes:** Porcentaje de campos con valores faltantes.
- ✓ **Tasa de finalización:** porcentaje de registros con todos los campos obligatorios completados.
- ✓ **Proporción de recuento de registros:** Relación entre registros completos y registros totales.



Métricas de coherencia: Si los datos se adhieren a reglas y formatos predefinidos.

- ✓ **Tasa de estandarización:** porcentaje de puntos de datos que se ajustan a un formato específico.
- ✓ **Tasa de valores atípicos:** porcentaje de puntos de datos que se desvían significativamente de la norma.
- ✓ **Tasa de registros duplicados:** Porcentaje de registros que son copias idénticas de otros.

Métricas de puntualidad: Para medir la frescura y relevancia de sus datos.

- ✓ **Antigüedad de los datos:** tiempo promedio transcurrido desde que se capturaron o actualizaron los datos.
- ✓ **Latencia:** Tiempo que tardan los datos en estar disponibles después de su generación.
- ✓ **Tasa de moneda:** Porcentaje de puntos de datos que reflejan la información más reciente.

Métricas de unicidad: Garantizar que todos los registros sean distintos y evitar duplicados.

- ✓ **Tasa de registros únicos:** Porcentaje de registros con identificadores únicos.
- ✓ **Tasa de duplicación:** Porcentaje de registros duplicados identificados y eliminados.



Documentación de los requisitos de calidad de datos

Es conveniente describir, como mínimo, las características de los datos que sean más relevantes para revisar su calidad, incluso establecer métricas “a través de las cuales se define la forma en la que cada dimensión es medida.

Métricas del estándar ISO/IEC 25024:2015 Systems and software engineering

Característica	Descripción	Criterio de evaluación
Compleitud	Nombre: Dato obligatorio Primer apellido: Dato obligatorio Segundo apellido: Dato opcional Número de cuenta: Dato obligatorio	Se evalúa cada uno de los datos obligatorios. Si el dato presenta un valor, obtiene un punto; en caso contrario, su calificación es cero.
Confidencialidad	Número de cuenta, nombre, primer y segundo apellido reciben tratamiento de datos personales.	Si el dato evaluado es tratado como dato personal en el sistema, el dato obtiene un punto; en caso contrario, su calificación es cero.
Exactitud - Semántica	Número de cuenta: El número de cuenta de los alumnos está compuesto por nueve dígitos a partir de la generación 2000. Por ejemplo: 12345678-9 El número de cuenta de las generaciones 1999 y anteriores está conformado por ocho dígitos. Por ejemplo: 1234567-8	Si el número de cuenta está compuesto por 9 dígitos ó, en su caso, por 8 dígitos si corresponde a la generación 1999 o a una anterior, el dato obtiene un punto, en caso contrario su calificación es cero.
Exactitud - Sintáctica	El número de cuenta está compuesto por valores numéricos, sin guion como separador del dígito verificador.	Si el número de cuenta presenta únicamente valores numéricos, obtiene un punto; de lo contrario, su calificación es cero.
Precisión	Entidad federativa de nacimiento, de acuerdo con el catálogo de INEGI.	Si la clave de la entidad federativa corresponde al catálogo de INEGI, obtiene un punto; de lo contrario, su valoración es cero.



Cuestiones de calidad de datos

Estos son algunos de los problemas de calidad de datos más comunes:

Datos inexactos

Suelen deberse a errores tipográficos, ortográficos o información desactualizada. A veces, lo que genera datos inexactos es simplemente el proceso de recopilación de datos defectuoso. Además, si sus datos favorecen a un determinado grupo o excluyen a otros, pueden generar resultados sesgados.

Datos incompletos

Factores como problemas de integración del sistema y errores de entrada de datos con frecuencia provocan registros omitidos y campos vacíos. En ocasiones los usuarios pasan por alto ciertos campos o no proporcionan información completa, especialmente en formularios o encuestas. El análisis de esto conduce a conocimientos deficientes y a una toma de decisiones cuestionable.

Datos obsoletos

Comprometen la confiabilidad y validez de los datos. A medida que los datos envejecen, reflejan menos las circunstancias actuales, lo que potencialmente conduce a análisis y toma de decisiones equivocados. Y en entornos dinámicos donde las condiciones cambian rápidamente, confiar en datos obsoletos puede dar lugar a errores estratégicos y oportunidades perdidas.

Datos duplicados

Este problema suele surgir debido a fallos del sistema o durante la integración de datos de múltiples fuentes. Los errores en la entrada de datos también contribuyen a la duplicación de datos. Las consecuencias son multifacéticas y van desde análisis sesgados hasta ineficiencias operativas. Específicamente, puede llevar a la sobreestimación o subestimación de ciertas métricas, lo que afecta la precisión de los análisis estadísticos y los conocimientos comerciales.

Datos inconsistentes

Generalmente se debe a diferentes formatos, unidades de medida o convenciones de nomenclatura entre los registros. Las causas fundamentales a menudo incluyen diversas fuentes de datos, cambios en los métodos de recopilación de datos o procesos comerciales en evolución. Las consecuencias de la inconsistencia de los datos son sustanciales y generan dificultades en integración de datos y comprometer la fiabilidad de los análisis.

Más allá de estos problemas, a veces demasiados datos también pueden provocar problemas de calidad; de hecho, puede ser un arma de doble filo. Este fenómeno, a menudo denominado **sobrecarga de datos**, ocurre cuando hay un volumen abrumador de información que procesar. Puede sobrecargar los recursos, ralentizar el análisis y aumentar la probabilidad de errores.



Mejores prácticas de calidad de datos

Mantener la calidad de los datos es un proceso continuo que exige un enfoque sistemático. Implica un seguimiento continuo y el perfeccionamiento de las prácticas relacionadas con los datos para mantener la integridad y confiabilidad de los datos. Estas son algunas de las mejores prácticas de calidad de datos:

Estandarizar formatos de datos

Los formatos de datos consistentes son vitales para evitar errores y mejorar la interoperabilidad. Cuando los datos siguen una estructura uniforme, se minimiza el riesgo de malas interpretaciones durante el análisis.

Para implementar esto, establezca un formato estandarizado para varios elementos de datos, incluidos formatos de fecha, representaciones numéricas y convenciones de texto. De esta manera, podrá crear una base para obtener datos precisos y confiables.

Implementar reglas de validación de datos

La implementación de robustas validación de datos Las reglas sirven como defensa de primera línea contra datos inexactos. Estas reglas actúan como controles automatizados que evalúan la precisión, integridad y cumplimiento de los datos entrantes con estándares predefinidos. Al definir y aplicar consistentemente estas reglas, se asegura de que solo datos de alta calidad ingresen al sistema de destino de destino.



Establecer políticas de gobierno de datos

Al crear pautas claras para el uso y el acceso a los datos, proporciona un marco que mitiga el riesgo de cambios no autorizados en los conjuntos de datos. Las auditorías periódicas y la aplicación estricta de estas políticas son esenciales para mantener un ecosistema de datos seguro. De esta manera, se asegura de que siempre se acceda y utilice los datos de acuerdo con los protocolos establecidos.

Priorizar la relevancia de los datos

Priorizar la relevancia de los datos es un enfoque estratégico para mantener un conjunto de datos enfocado e impactante. Las evaluaciones periódicas de la importancia de cada elemento de datos en relación con los objetivos comerciales actuales son cruciales. Identificar y eliminar datos obsoletos o redundantes le permite optimizar su conjunto de datos y hacerlo más eficiente para los análisis y los procesos de toma de decisiones.

Aplicar el seguimiento del linaje de datos

Implementar herramientas y procesos para rastrear el origen y transformaciones de datos durante todo su ciclo de vida es fundamental. Al documentar metadatos, transformaciones y dependencias, se crea un mapa completo de linaje de datos. Este mapa se convierte en un recurso valioso para solucionar problemas, auditar y garantizar la precisión de los conocimientos basados en datos.



Sistemas Gestores de Bases de Datos

Las principales funciones del Sistema Gestor de Bases de datos SGBD (Database Management System, “DBMS” o Relational Database management System, “RDBMS”) son:

- ✓ **Definición de Datos:** (se puede realizar a través del lenguaje de definición de datos o DDL) que provee el DBMS.
- ✓ **Manipulación de Datos:** permite almacenar, modificar y recuperar los datos de la Base de Datos. Esto se logra a través del lenguaje de manipulación de datos o DML provisto por el DBMS
- ✓ **Seguridad de Datos:** el DBMS provee de mecanismos para controlar el acceso y para definir qué operaciones puede realizar cada usuario.

Además, debe proveer de mecanismos de respaldo y recuperación de la Base de Datos, También debe manejar el acceso concurrente a la Base de Datos.

Funciones Generales

- ✓ Permitir a los usuarios almacenar datos, acceder a ellos y actualizarlos, ocultando su estructura física.
- ✓ Proporcionar un catálogo (diccionario de datos) accesible por los usuarios.
- ✓ Proporcionar un mecanismo que garantice el procesamiento de las transacciones.
- ✓ Proporcionar un mecanismo que realice el control de la concurrencia.
- ✓ Proporcionar un mecanismo para recuperación ante fallos.
- ✓ Proporcionar un mecanismo de seguridad.
- ✓ Integrarse con algún software de comunicación.
- ✓ Encargarse de mantener las reglas de integridad.
- ✓ Encargarse de mantener la independencia entre los programas y la estructura de la base de datos.
- ✓ Proporcionar herramientas para administrar la base de datos.



El procesador de consultas

Es el componente principal de un SGBD. Transforma las consultas en un conjunto de instrucciones de bajo nivel que se dirigen al gestor de la base de datos.

El gestor de la base de datos

Es la interface con los programas de aplicación y las consultas de los usuarios. El gestor de la base de datos acepta consultas y examina los esquemas externo y conceptual para determinar qué registros se requieren para satisfacer la petición. Entonces el gestor de la base de datos realiza una llamada al gestor de archivos para ejecutar la petición.

El gestor de archivos

El gestor de archivos o motor de almacenamiento (storage-engine) maneja los archivos en disco en donde se almacena la base de datos y se encarga de almacenar, manejar y recuperar la información contenida en una tabla. Este gestor establece y mantiene la lista de estructuras e índices definidos en el esquema interno. Si se utilizan archivos dispersos, llama a la función de dispersión para generar la dirección de los registros. Pero el gestor de archivos no realiza directamente la entrada y salida de datos. Lo que hace es pasar la petición a los métodos de acceso del sistema operativo que se encargan de leer o escribir los datos en el buffer del Sistema. Sin embargo los hay que gestionan en forma autónoma los archivos relacionados con la base de datos.



MyISAM

En MySQL si se utiliza el que esta por defecto **MyISAM**, cada tabla se almacena en tres (3) archivos distintos (.frm .MYD .MYI). Esta permite una mayor velocidad al recuperar la información de las tablas por lo que se utiliza en aplicaciones donde predomina la lectura mediante el uso de “SELECT” que las escrituras mediante UPDATE o INSERT. No realiza comprobaciones de integridad referencial ni ejecuta bloqueo de filas pero si de tablas. Por ello no admite transacciones.

InnoDB

Por otro lado si se requieren características ACID, soportar transacciones y el bloqueo de registros se utiliza el modelo **InnoDB** para el almacenamiento. Este es el indicado para aplicaciones con preponderancia de escrituras con INSERT o UPDATE.

InnoDB permite un almacenamiento transaccional con capacidad de confirmación (COMMIT) y de cancelación (ROLLBACK) además de recuperación ante fallos.

También efectúa bloqueos a nivel de filas.

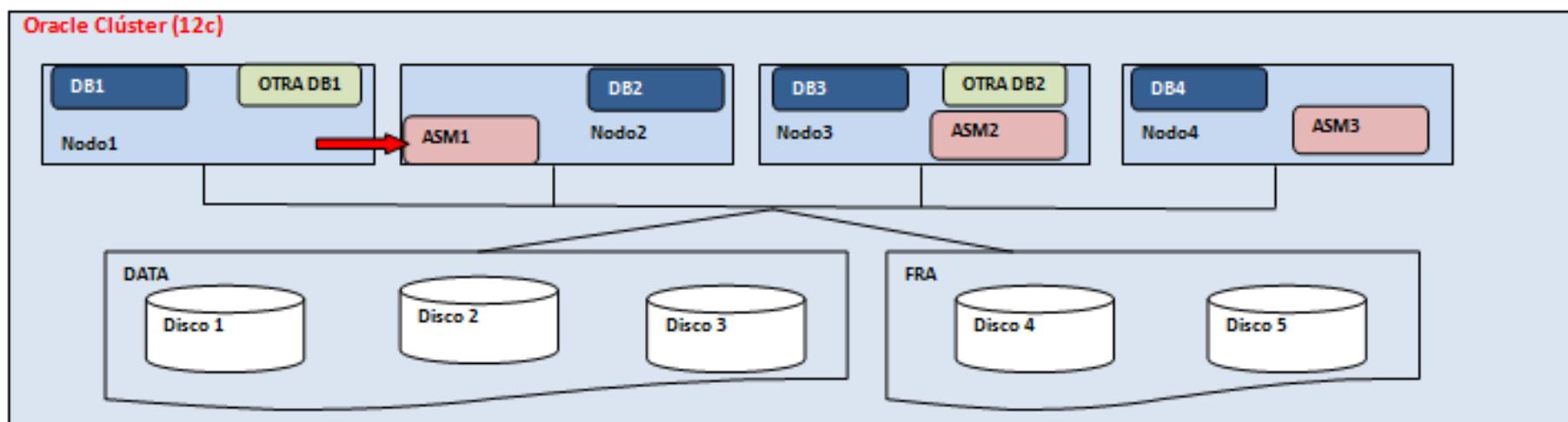
Memory

Crea tablas en memoria por lo que es bastante rápido. No admite transacciones y se utiliza para crear tablas temporales y busquedas rápidas. Como esta basado en memoria los datos se pierden al reiniciar la base de datos.

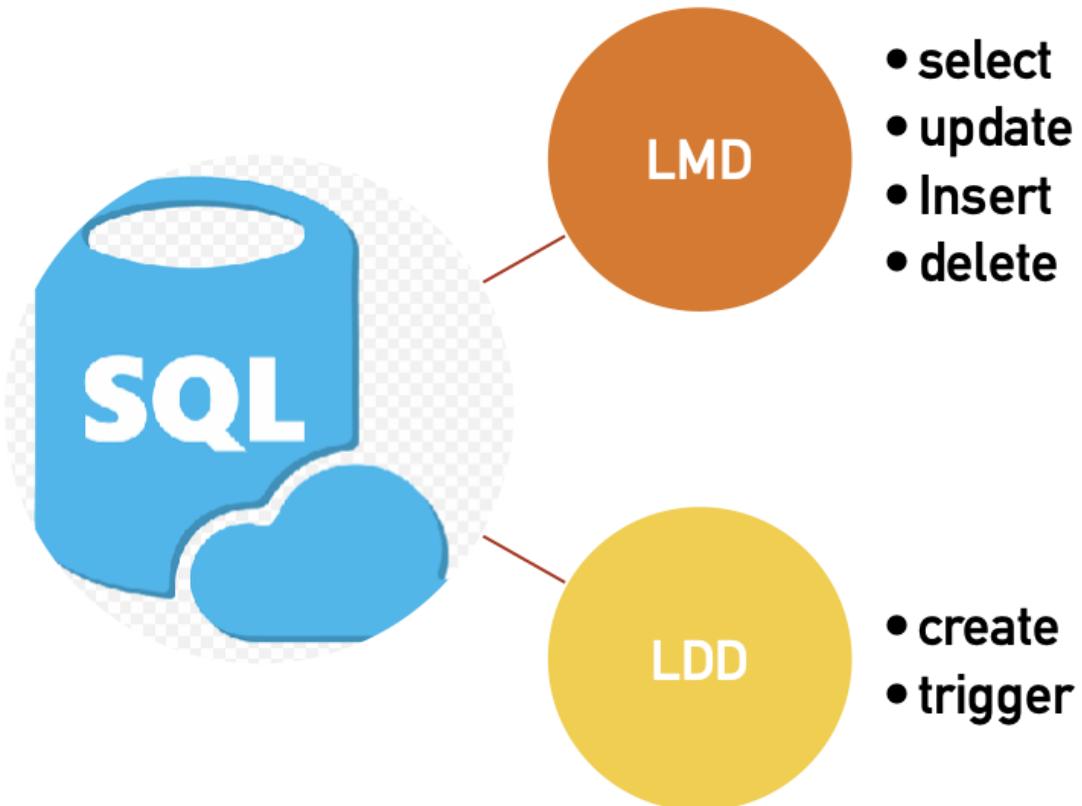


ASM

Oracle utiliza un sistema de gestión de archivos más compleja denominado Automatic Storage Management (ASM)³ pero que facilita la administración. Implementa archivos para diversas funciones (archivos de control, de datos, de parámetro, de contraseñas, de recuperación, de recuperación en línea, de archivado de recuperación en línea, y registros de alertas y archivos de seguimiento) y un gestor ASM que organiza todos los archivos y discos para que trabajen como una sola entidad.



Indices



2.7 ACID

Una **transacción** es una serie de procesos que se aplican dentro de una base de datos en forma secuencial u ordinal y que debe realizarse de una vez y sin alterar la estructura de los datos. Una base de datos es cumplimentaria ACID si presenta estas cuatro propiedades dentro de las transacciones de una BD:

Atomicidad

Referida a la propiedad que determina que la operación se haya realizado o no, pero nunca a medias. Se ejecuta la operación completa con todos sus pasos o no se ejecuta del todo.

Consistencia

Solo se ejecutan las operaciones que no afectan la integridad de la base de datos. Cualquier operación que se lleva a cabo será de un estado válido a otro con datos consistentes.

Aislamiento

(Isolation) Cada operación es única y no afecta a otras aunque se realicen sobre la misma información.

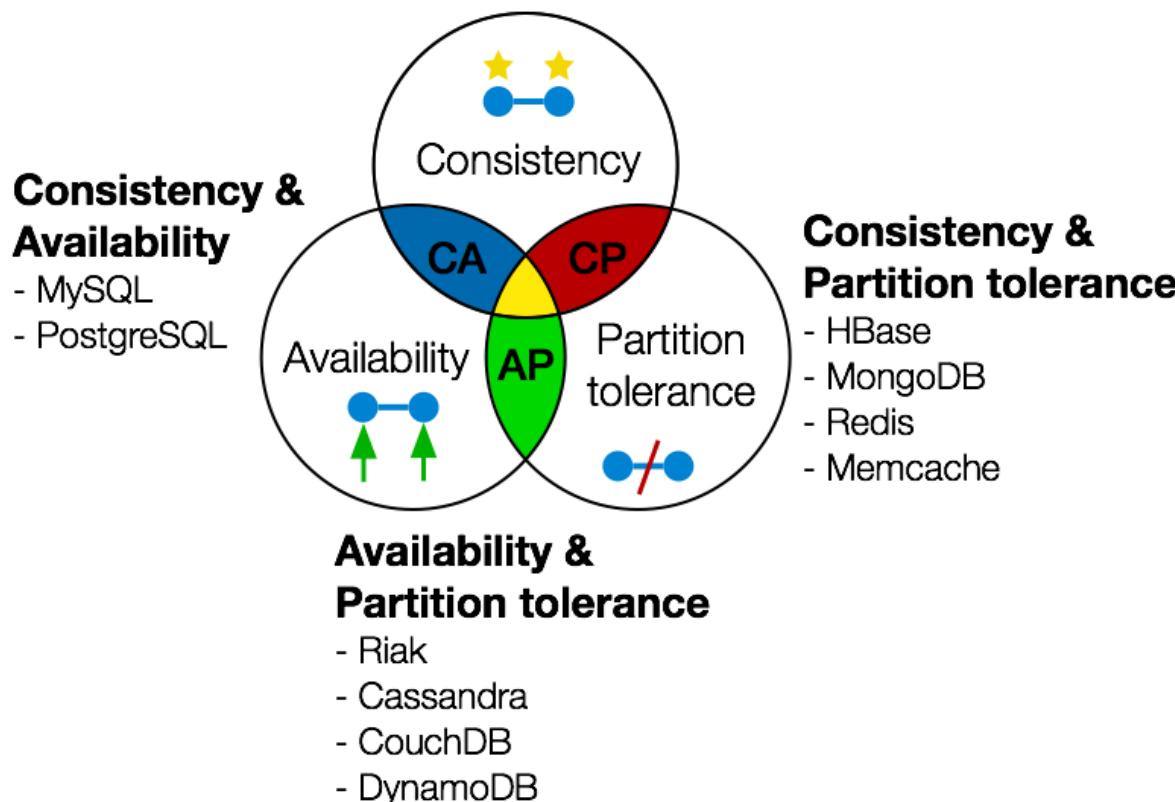
Durabilidad

Asegura que la operación una vez realizada, es persistente y no se podrá deshacer a pesar de fallos en el sistema.



2.8 Teorema CAP

Establece algunos atributos deseados en todo gestor de base de datos que nos permita manejar los datos con cierta seguridad y confianza en ellos. Son tres características que los motores de bases de datos tratan de equilibrar, en el sentido que algunos poseen mayor fortaleza en alguno de los tres términos involucrados;



Por regla general las bases de datos pueden garantizar solo dos de los tres atributos lo que origina las siguientes combinaciones:

CA: *Consistencia y Disponibilidad* - Se garantiza el acceso a la información y el valor del dato es consistente (igual) para todas las peticiones atendidas; de haber cambios, se mostrarán inmediatamente. Sin embargo, la partición de los nodos no es tolerada por el sistema de forma simultánea.

Bases de datos relacionales: MySQL, PostgreSQL, Oracle, SQL Server, etc.

AP: *Disponibilidad y Tolerancia a la partición* - Se garantiza el acceso a los datos y el sistema es capaz de tolerar (gestionar) la partición de los nodos, pero dejando en segundo plano la consistencia de los datos, ya que no se conserva y el valor de dato no estará replicado en los diferentes nodos al instante.

Las bases que adotan este modelo son aquellas construidas para manejar un gran volumen de datos en entornos distribuidos, con múltiples nodos de datos interconectados entre sí. Bases noSQL: MongoDB, DynamoDB, Cassandra

CP: *Consistencia y Tolerancia a la partición* - Se garantiza la consistencia de los datos entre los diferentes nodos y la partición de los nodos se tolera, pero sacrificando la disponibilidad de los datos, con lo cual, el sistema puede fallar o tardar en ofrecer una respuesta a la petición del usuario.

La elección de la base está en relación con el negocio. Ejemplo: en los bancos prevalece **CP**, porque la información debe ser siempre consistente y no admitir fallos, y dejando la disponibilidad en segundo plano.



El Modelo Relacional

- ✓ Es un modelo de datos basado en la lógica de predicados y en la teoría de conjuntos.
- ✓ Su idea fundamental es el uso de relaciones. Estas relaciones podrían considerarse en forma lógica como conjuntos de datos llamados tuplas.
- ✓ Es el modelo más utilizado en la actualidad para modelar problemas reales y administrar datos dinámicamente.
- ✓ El modelo relacional desarrolla un esquema de base de datos (data base schema) a partir del cual se podrá realizar el modelo físico o de implementación en el DBMS.
- ✓ Este modelo esta basado en que todos los datos están almacenados en tablas (entidades/relaciones) y cada una de estas es un conjunto de datos, por tanto una base de datos es un conjunto de relaciones. La agrupación se origina en la tabla: tabla -> fila (tupla) -> campo (atributo)



El Modelo Relacional se ocupa de:

- ✓ La estructura de datos
- ✓ La manipulación de datos
- ✓ La integridad de los datos

Donde las relaciones están formadas por :

- ✓ Atributos (columnas)
- ✓ Tuplas (Conjunto de filas)

Existen dos formas para la construcción de modelos relationales:

- ✓ Creando un conjunto de tablas iniciales y aplicando operaciones de normalización hasta conseguir el esquema más óptimo,
- ✓ O, convertir el modelo entidad relación (ER) en tablas, con una depuración lógica y la aplicación de restricciones de integridad.



Objetivos

Los objetivos que este modelo persigue son:

Independencia Física: La forma de almacenar los datos no debe influir en su manipulación. Si el almacenamiento físico cambia, los usuarios que acceden a esos datos no tienen que modificar sus aplicaciones.

Independencia Lógica: Las aplicaciones que utilizan la base de datos no deben ser modificadas por que se inserten, actualicen y eliminen datos.

Flexibilidad: En el sentido de poder presentar a cada usuario los datos de la forma en que éste prefiera

Uniformidad: Las estructuras lógicas de los datos siempre tienen una única forma conceptual (las tablas), lo que facilita la creación y manipulación de la base de datos por parte de los usuarios.

Sencillas: Las características anteriores hacen que este Modelo sea fácil de comprender y de utilizar por parte del usuario final



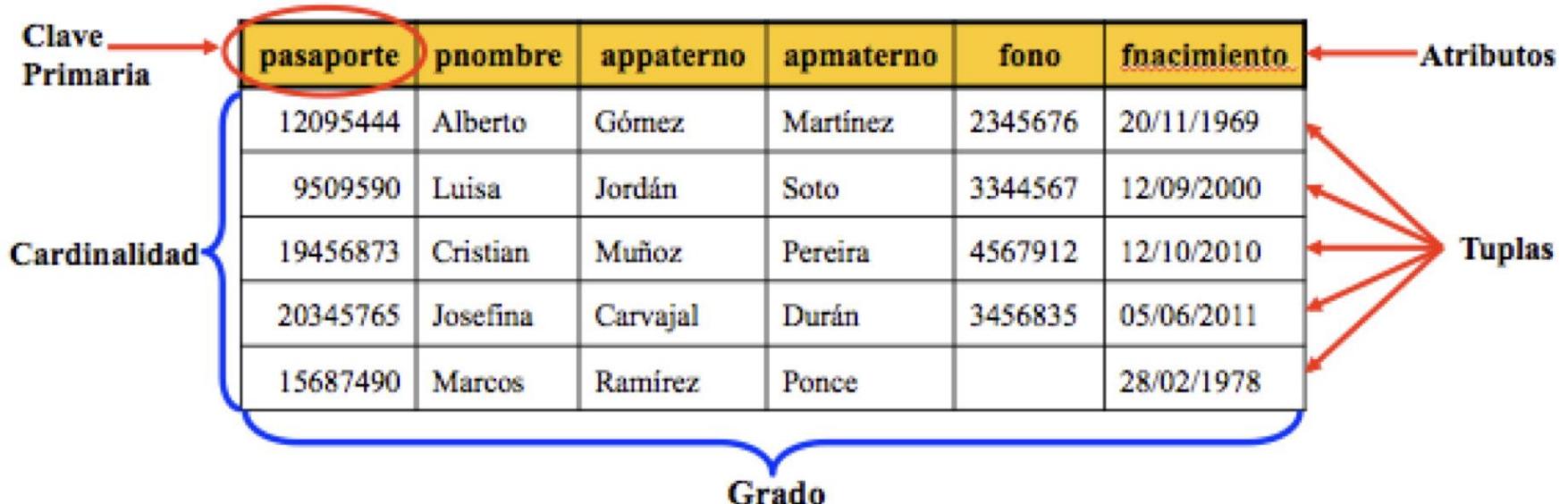
Características

- ✓ Los datos son atómicos ó monovaluados;
- ✓ Los datos de cualquier columna son de un solo tipo.
- ✓ Cada columna posee un nombre único.
- ✓ El orden de las columnas no es de importancia para la tabla.
- ✓ Las columnas de una relación se conocen como atributos.
- ✓ Cada atributo tiene un dominio,
- ✓ No existen 2 filas en la tabla que sean idénticas.
- ✓ La información en las bases de datos son representados como datos explícitos.
- ✓ Cada relación tiene un nombre específico y diferente al resto de las relaciones.
- ✓ Los valores de los atributos son atómicos: en cada tupla, cada atributo (columna) toma un solo valor. Se dice que las relaciones están normalizadas.
- ✓ El orden de los atributos no importa: los atributos no están ordenados.
- ✓ Cada tupla es distinta de las demás: no hay tuplas duplicadas
- ✓ El orden de las tuplas no importa: las tuplas no están ordenadas.
- ✓ Los atributos son atómicos: en cada tupla, cada atributo (columna) toma un solo valor. Se dice que las relaciones están normalizadas.



Terminología Relacional		Terminología de Tablas		Terminología de Archivo
Relación	=	Tabla	=	Archivo
Tupla	=	Fila	=	Registro
Atributo	=	Columna	=	Campo
Grado	=	Número de columnas	=	Número de campos
Cardinalidad	=	Número de filas	=	Número de registros

EMPLEADO ← Nombre de la Relación



Reglas de Integridad

1

Si dos tablas tienen una relación entre ellas 1:1, entonces el campo clave de una de las tablas debe aparecer en la otra tabla.

2

Si dos tablas tienen una relación entre ellas 1:M, entonces el campo clave de la tabla (1) debe aparecer en la otra tabla (M).

3

Si dos tablas tienen una relación entre ellas M:M, entonces debe crearse una nueva tabla que contenga los campos clave de las dos tablas.



Tupla: es cada una de las filas de la relación. Representa por tanto el conjunto de cada elemento individual (ejemplar ó ocurrencia) de esa tabla. En la relación OFICINA, cada tupla tiene cinco valores, uno para cada atributo. Las tuplas de una relación no siguen ningún orden.

Grado: número de columnas de la relación (número de atributos). La relación OFICINA es de grado seis porque tiene seis atributos. Esto quiere decir que cada fila de la tabla es una tupla con seis valores.

Cardinalidad: número de tuplas de una relación (número de filas). Ya que en las relaciones se van insertando y borrando tuplas a menudo, la cardinalidad de las mismas varía constantemente.

pasaporte	pnombre	appaterno	apmaterno	fono	fnacimiento
12095444	Alberto	Gómez	Martínez	2345676	20/11/1969
9509590	Luisa	Jordán	Soto	3344567	12/09/2000
19456873	Cristian	Muñoz	Pereira	4567912	12/10/2010
20345765	Josefina	Carvajal	Durán	3456835	05/06/2011
15687490	Marcos	Ramírez	Ponce		28/02/1978

Cardinalidad

Grado

Tuplas

Claves

Ya que en una relación no hay tuplas repetidas, éstas se pueden distinguir unas de otras, es decir, se pueden identificar de modo único. La forma de identificarlas es mediante los valores de sus atributos.

Clave candidata

Conjunto de atributos que permiten identificar en forma única cada tupla de la relación. Es decir columnas cuyos valores no se repiten para esa tabla. Los atributos candidatos para una tabla de individuos (clientes, pacientes, etc.) es el ‘rut’, un número de seguro social, un ‘id’ de cliente (numérico o de carácter).

Clave primaria

Clave candidata que se escoge como identificador de las tuplas. Se elige como primaria la candidata que identifique mejor a cada tupla en el contexto de la base de datos..

Clave alternativa

Cualquier clave candidata que no sea primaria y que también puede identificar de manera única una tupla. Al momento de crear la relación como tabla en la Base de Datos se debe definir una constraint de tipo UNIQUE.

Clave externa, ajena o foránea

Atributo cuyos valores coinciden con una clave candidata (normalmente primaria) de otra tabla

NoSQL.

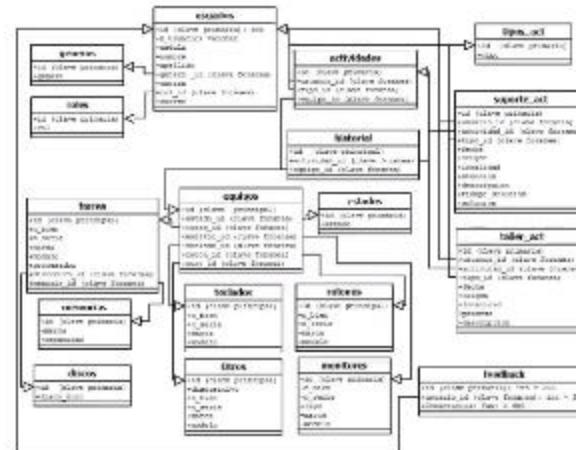
Structure Query Language(SQL)

Read Insert Update Delete

Relational Database Management System(RDBMS)



ORACLE®



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

- ✓ NoSQL, también conocido como "no solo SQL", "no SQL", es un enfoque de diseño de base de datos que permite almacenar y consultar datos fuera de las estructuras tradicionales que se encuentran en las bases de datos relacionales.
- ✓ Aunque puede almacenar los datos que se encuentran dentro de los sistemas de gestión de bases de datos, los almacena de manera diferente a un RDBMS.
- ✓ Las bases de datos NoSQL alojan datos dentro de una estructura de datos, como un documento JSON.
- ✓ Dado que este diseño de base de datos no relacional no requiere un esquema, ofrece una rápida escalabilidad para gestionar grandes conjuntos de datos normalmente no estructurados.
- ✓ NoSQL es también un tipo de base de datos distribuida, lo que significa que la información se copia y almacena en varios servidores, que pueden ser remotos o locales.
- ✓ Garantizan la disponibilidad y la fiabilidad de los datos. Si alguno de los datos queda fuera de línea, el resto de la base de datos puede seguir ejecutándose.

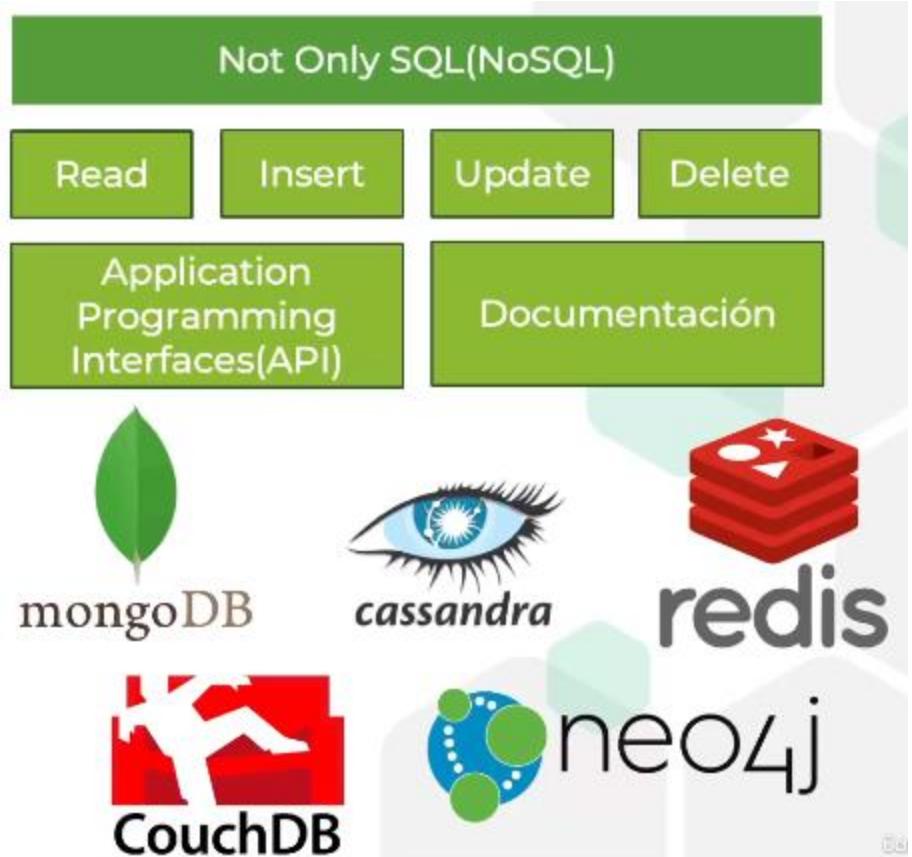
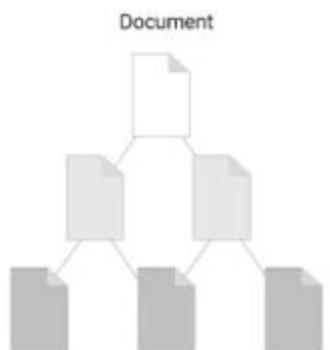
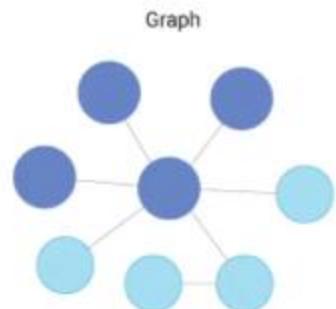
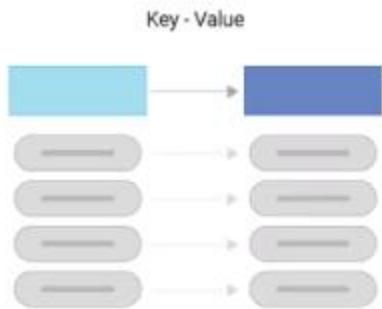


Ventajas de los sistemas NoSQL

- ✓ Se ejecutan en máquinas con pocos recursos: Estos sistemas, a diferencia de los sistemas basados en SQL, no requieren de apenas computación, por lo que se pueden montar en máquinas de un coste más reducido.
- ✓ Escalabilidad horizontal: Para mejorar el rendimiento de estos sistemas simplemente se consigue añadiendo más nodos, con la única operación de indicar al sistema cuáles son los nodos que están disponibles.
- ✓ Pueden manejar gran cantidad de datos: Esto es debido a que utiliza una estructura distribuida, en muchos casos mediante tablas Hash.
- ✓ No genera cuellos de botella: El principal problema de los sistemas SQL es que necesitan transcribir cada sentencia para poder ser ejecutada, y cada sentencia compleja requiere además de un nivel de ejecución aún más complejo, lo que constituye un punto de entrada en común, que ante muchas peticiones puede ralentizar el sistema.

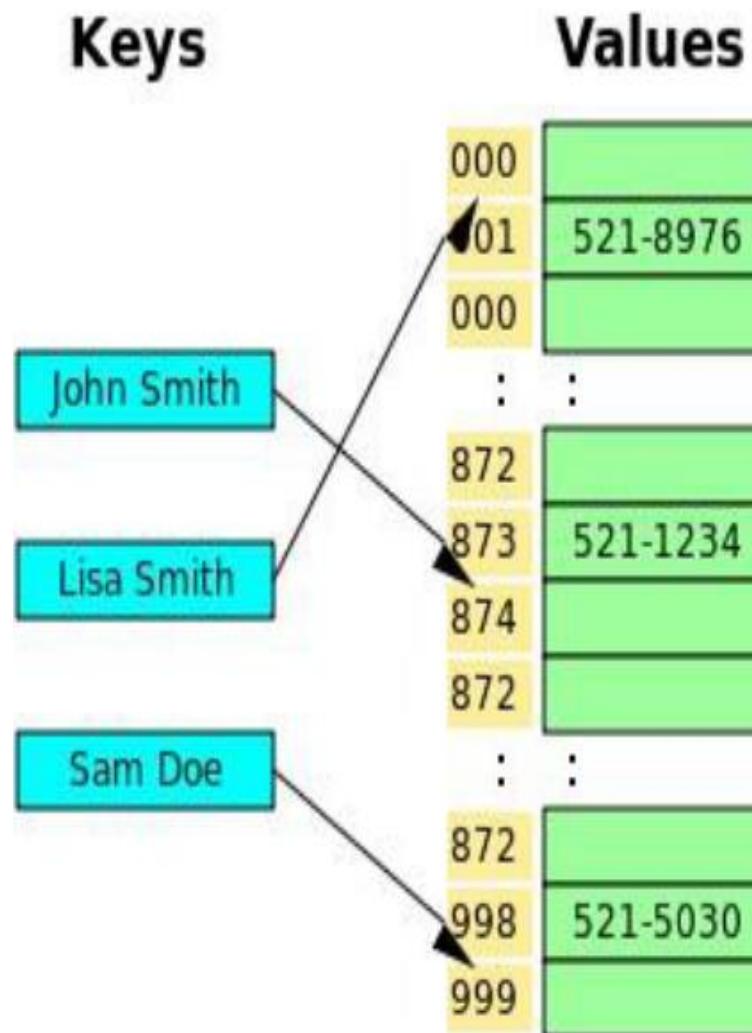
Principales diferencias con las bases de datos SQL

- ✓ **No utilizan SQL como lenguaje de consultas.** La mayoría de las bases de datos NoSQL evitan utilizar este tipo de lenguaje o lo utilizan como un lenguaje de apoyo. Por poner algunos ejemplos, Cassandra utiliza el lenguaje CQL, MongoDB utiliza JSON o BigTable hace uso de GQL.
- ✓ **No utilizan estructuras fijas como tablas para el almacenamiento de los datos.** Permiten hacer uso de otros tipos de modelos de almacenamiento de información como sistemas de clave–valor, objetos o grafos.
- ✓ **No suelen permitir operaciones JOIN.** Al disponer de un volumen de datos tan extremadamente grande suele resultar deseable evitar los JOIN. Esto se debe a que, cuando la operación no es la búsqueda de una clave, la sobrecarga puede llegar a ser muy costosa. Las soluciones más directas consisten en desnormalizar los datos, o bien realizar el JOIN mediante software, en la capa de aplicación.
- ✓ **Arquitectura distribuida.** Las bases de datos relacionales suelen estar centralizadas en una única máquina o bien en una estructura máster–esclavo, sin embargo en los casos NoSQL la información puede estar compartida en varias máquinas mediante mecanismos de tablas Hash distribuidas



Bases de datos clave – valor

- ✓ Son las bases de datos NoSQL más populares, y son las más sencillas en cuanto a funcionalidad.
- ✓ Cada elemento está identificado por una llave única, lo que permite la recuperación de la información de forma muy rápida, información que habitualmente está almacenada como un objeto binario (BLOB).
- ✓ Son muy eficientes tanto para las lecturas como para las escrituras.
- ✓ Algunos ejemplos de este tipo son Cassandra, BigTable o HBase.



Bases de datos clave – valor



Bases de datos clave – valor

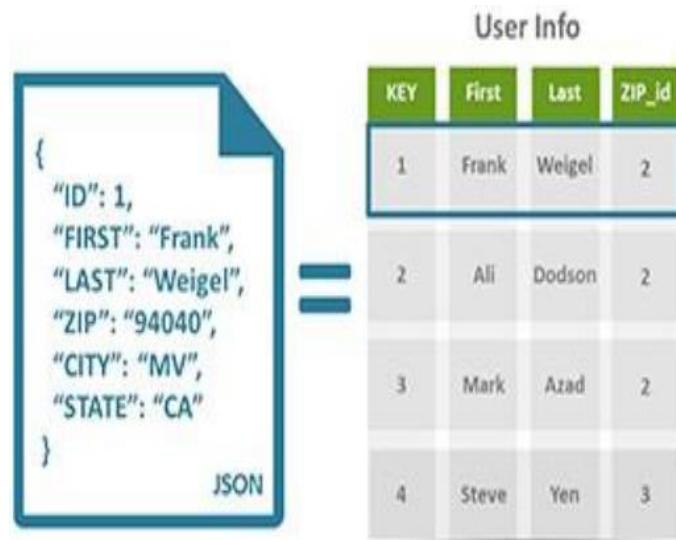
Redis

- Está apoyado por VMWare.
- Se puede imaginar como un array gigante en memoria para almacenar datos, datos que pueden ser cadenas, hashes, conjuntos de datos o listas.
- Tiene la ventaja de que sus operaciones son atómicas y persistentes.
- No permite realizar consultas, sólo se puede insertar y obtener datos, además de las operaciones comunes sobre conjuntos (diferencia, unión e inserción).



Bases de datos documentales

- ✓ Almacena la información como un documento, generalmente utilizando para ello una estructura simple como JSON o XML y donde se utiliza una clave única para cada registro.
- ✓ Permite realizar búsquedas por clave–valor y consultas más avanzadas sobre el contenido del documento.
- ✓ Se pueden utilizar en gran cantidad de proyectos, incluyendo muchos que tradicionalmente funcionarían sobre bases de datos relacionales.
- ✓ Algunos ejemplos de este tipo son MongoDB o CouchDB.



Bases de datos documentales

```
{  
    "empid": "SJ011MS",  
    "personal": {  
        "name": "Smith Jones",  
        "gender": "Male",  
        "age": 28,  
        "address": {  
            "streetaddress": "7 24th Street",  
            "city": "New York",  
            "state": "NY",  
            "postalcode": "10038"  
        }  
    },  
    "profile": {  
        "designation": "Deputy General",  
        "department": "Finance"  
    }  
}
```

Casos de Uso

Facebook

twitter

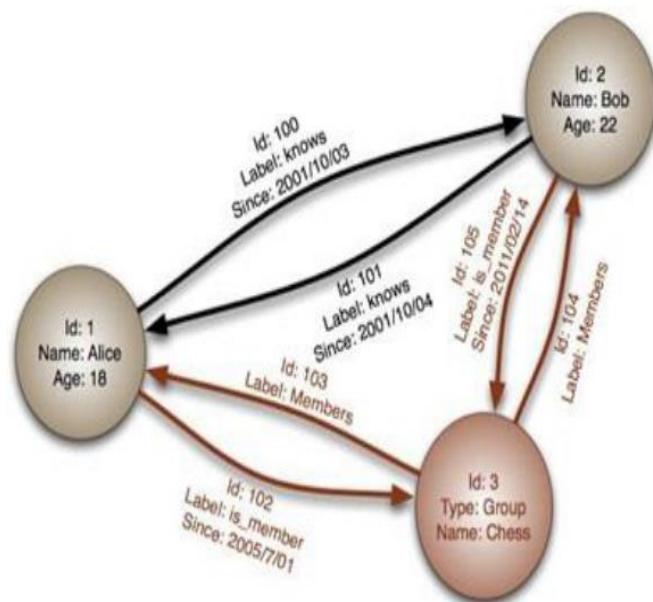
Reddit



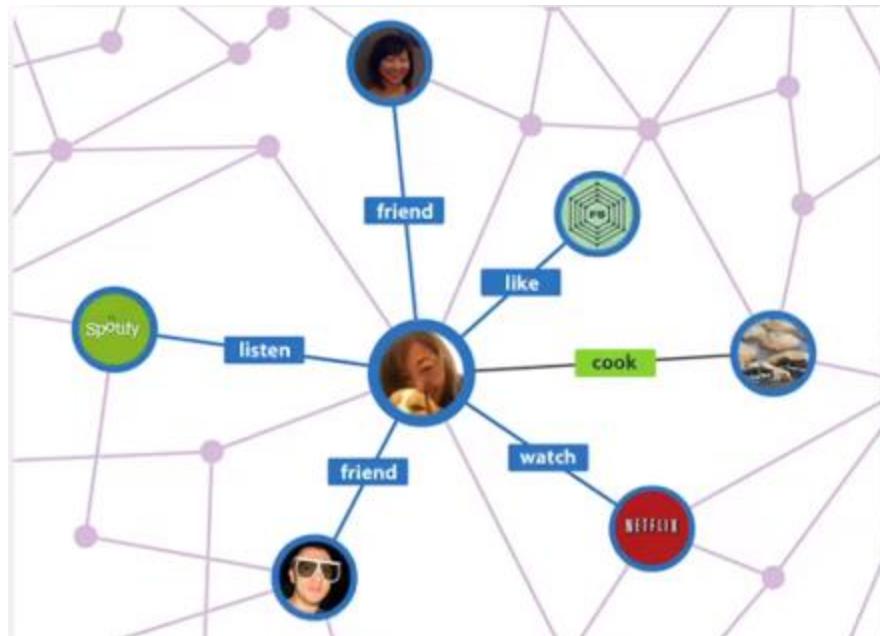
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Bases de datos en grafo

- ✓ La información se representa como nodos de un grafo y sus relaciones con las aristas del mismo, de manera que se puede hacer uso de la teoría de grafos para recorrerla.
 - ✓ Para sacar el máximo rendimiento a este tipo de bases de datos, su estructura debe estar totalmente normalizada, de forma que cada tabla tenga una sola columna y cada relación dos.
-
- Ofrece una navegación más eficiente entre relaciones que en un modelo relacional.
 - Algunos ejemplos de este tipo son Neo4j, InfoGrid o Virtuoso



Bases de datos en grafo



Base de datos columnar

- ✓ Es un modelo que organiza los datos en columnas en lugar de filas.
- ✓ En este enfoque, cada columna contiene datos de un solo tipo, y las filas contienen un conjunto de valores correspondientes a esas columnas.
- ✓ La idea principal es almacenar y procesar los datos columnarmente en lugar de hacerlo en forma de filas.

Características de las bases de datos columnares

- ✓ **Compresión eficiente:** Almacenar los datos por columnas permite una mayor compresión, ya que los valores en una columna tienden a repetirse o ser similares, lo que reduce el tamaño de almacenamiento.
- ✓ **Procesamiento optimizado:** Al realizar operaciones sobre columnas en lugar de filas, las bases de datos columnares pueden acelerar ciertas consultas y análisis que implican grandes volúmenes de datos.
- ✓ **Análisis selectivo:** Permite que las consultas seleccionen solo las columnas necesarias para una operación específica, lo que reduce la búsqueda de datos que deben ser leídos desde el almacenamiento.
- ✓ **Análisis analítico:** Las bases de datos columnares se utilizan principalmente para análisis analítico

Base de datos columnar

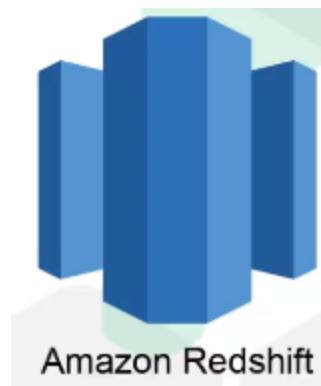
Casos de Uso

Información de los
productos

Reseñas de los
productos



cassandra



Amazon Redshift



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Bases de datos orientadas a objetos

- La información se representa mediante objetos, de la misma forma que son representados en los lenguajes de programación orientada a objetos (POO) como ocurre en JAVA, C# o Visual Basic .NET.
- Algunos ejemplos de este tipo de bases de datos son Zope, Gemstone

db4objects
BY VERSANT



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Usos

- ✓ Cuando se necesita una BBDD para una aplicación que hace una **consulta/lectura** intensiva de grandes cantidades de datos.
- ✓ Cuando no hay la necesidad de que los **datos sean consistentes**.
- ✓ Si los datos a almacenar no tienen una **estructura fija**.
- ✓ Una misma aplicación puede usar una BBDD relacional y una BBDD NoSQL y guardar cosas diferentes en cada una de ellas.
- ✓ Algunos **ejemplos** de uso de este tipo de BBDD:
 - Amazon.
 - Facebook.
 - Google.



	NoSQL o no relacional	SQL o relacional
LA MEJO R OPCI ÓN PARA:	<ul style="list-style-type: none"> • Administrar datos de gran volumen, no relacionados, indeterminados o que cambian rápidamente. • Datos independientes del esquema o esquema dictados por la aplicación. • Aplicaciones en las que el rendimiento y la disponibilidad son más importantes que una coherencia alta. • Aplicaciones siempre activas que dan servicios a usuarios de todo el mundo. 	<ul style="list-style-type: none"> • Administrar datos relacionales con requisitos lógicos y discretos que se puedan identificar con antelación. • Esquema que se debe mantener sincronizado entre la aplicación y la base de datos. • Sistemas heredados creados para estructuras relacionales. • Aplicaciones que requieren transacciones de varias filas o consultas complejas.
ESCE NARIO S:	<ul style="list-style-type: none"> • Aplicaciones celulares. • Análisis en tiempo real. • Administración de contenido. • Aplicaciones de IoT. • Migración de bases de datos. 	<ul style="list-style-type: none"> • Sistemas de contabilidad, finanzas y bancarios. • Sistemas de administración de inventario. • Sistemas de administración de transacciones.
ESCA LA:	<ul style="list-style-type: none"> • Escala los datos horizontalmente mediante el particionamiento entre los servidores. 	<ul style="list-style-type: none"> • Escala los datos verticalmente al aumentar la carga del servidor.
MODE LO DE DATO S:	<ul style="list-style-type: none"> • Tipos de base de datos: bases de datos de pares clave-valor, documentos, en columnas y de grafos. • Almacena los datos en función del tipo de base de datos. 	<ul style="list-style-type: none"> • Tipo de base de datos: tablas de filas, agrupadas en relaciones. • Usa el Lenguaje de consulta (SQL). • Almacena datos como filas en tablas; datos relacionados almacenados por separado y unidos para consultas complejas.



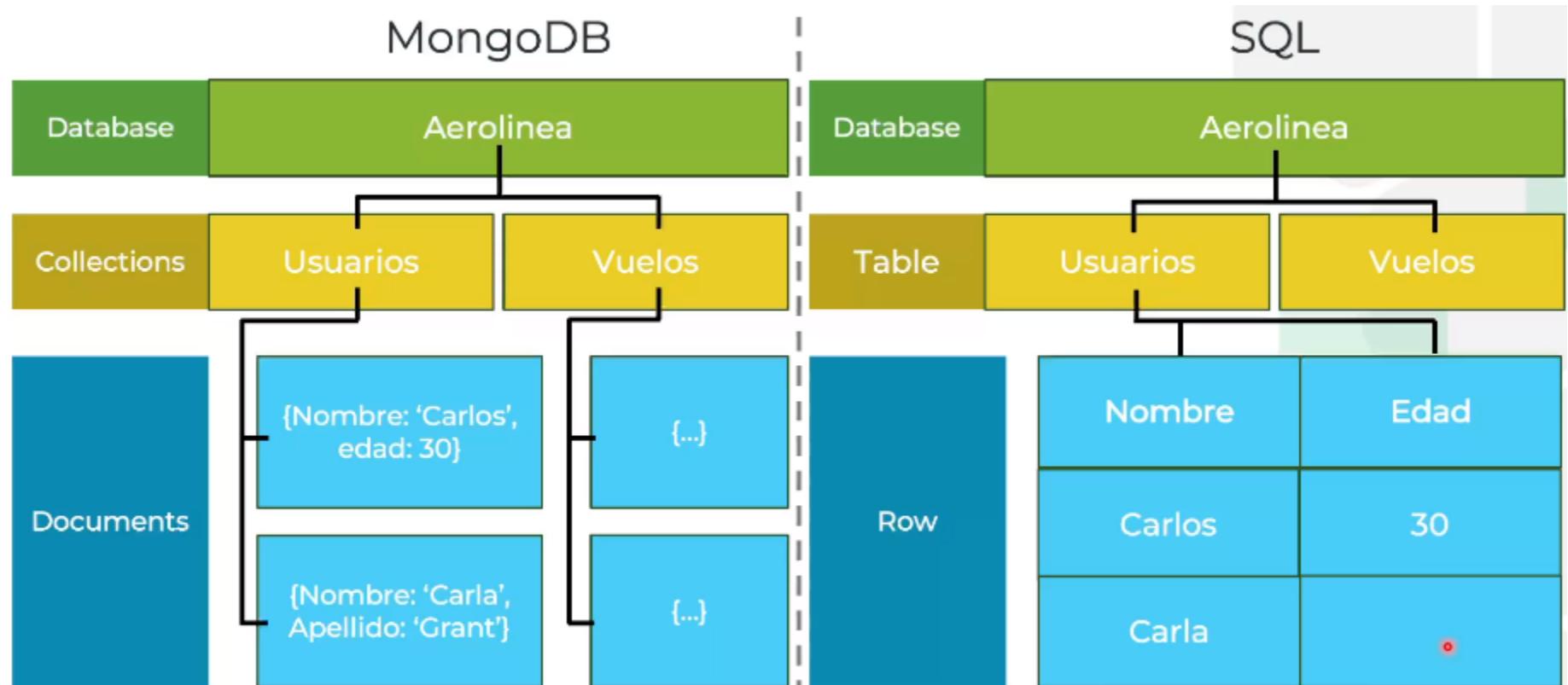
MongoDB

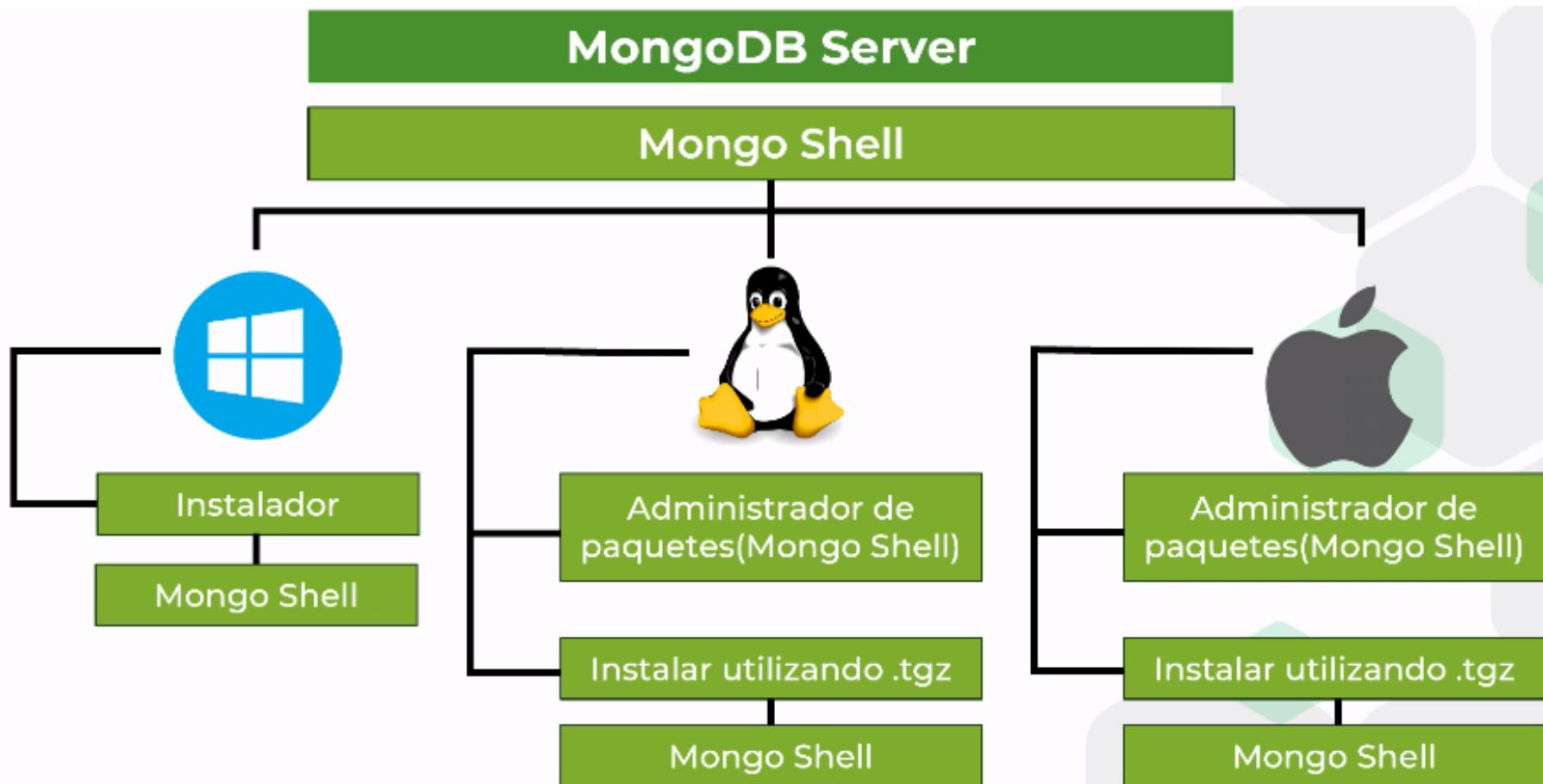
- ✓ Se trata de una base de datos creada por 10gen del tipo orientada a documentos, de esquema libre, es decir, que cada entrada puede tener un esquema de datos diferente que nada tenga que ver con el resto de registros almacenados.
- ✓ Es bastante rápido a la hora de ejecutar sus operaciones ya que está escrito en lenguaje C++.
- ✓ Para el almacenamiento de la información, utiliza un sistema propio de documento conocido con el nombre BSON, que es una evolución del conocido JSON pero con la peculiaridad de que puede almacenar datos binarios.



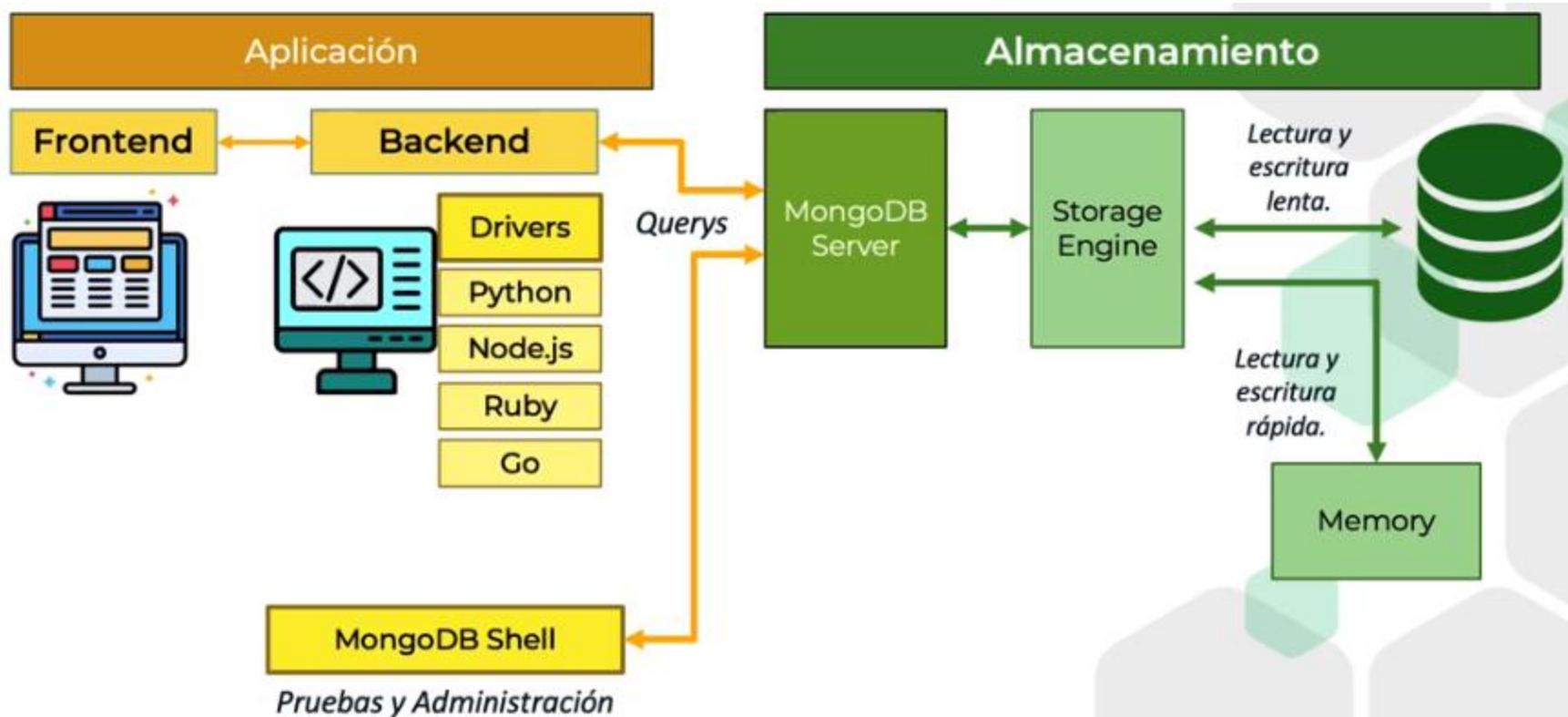
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

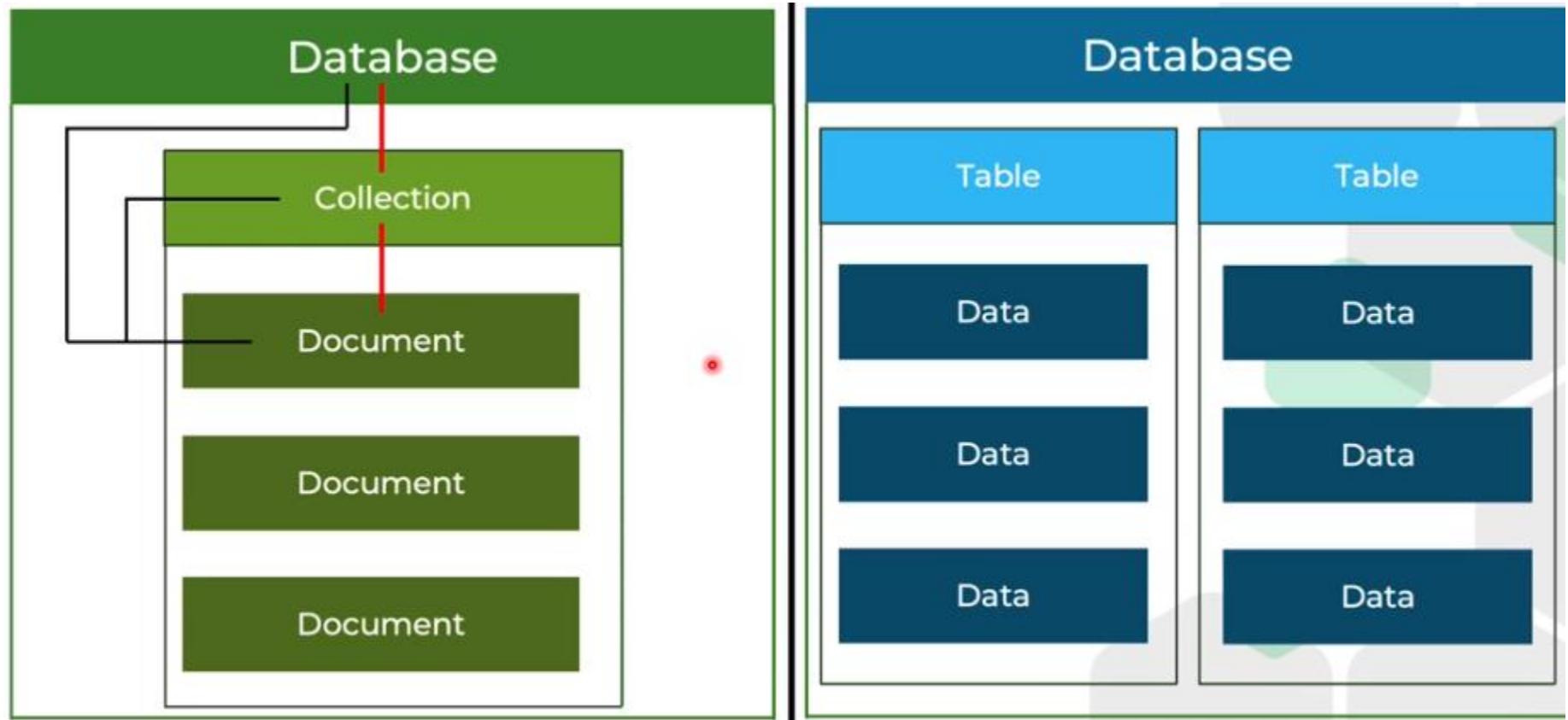
MongoDB

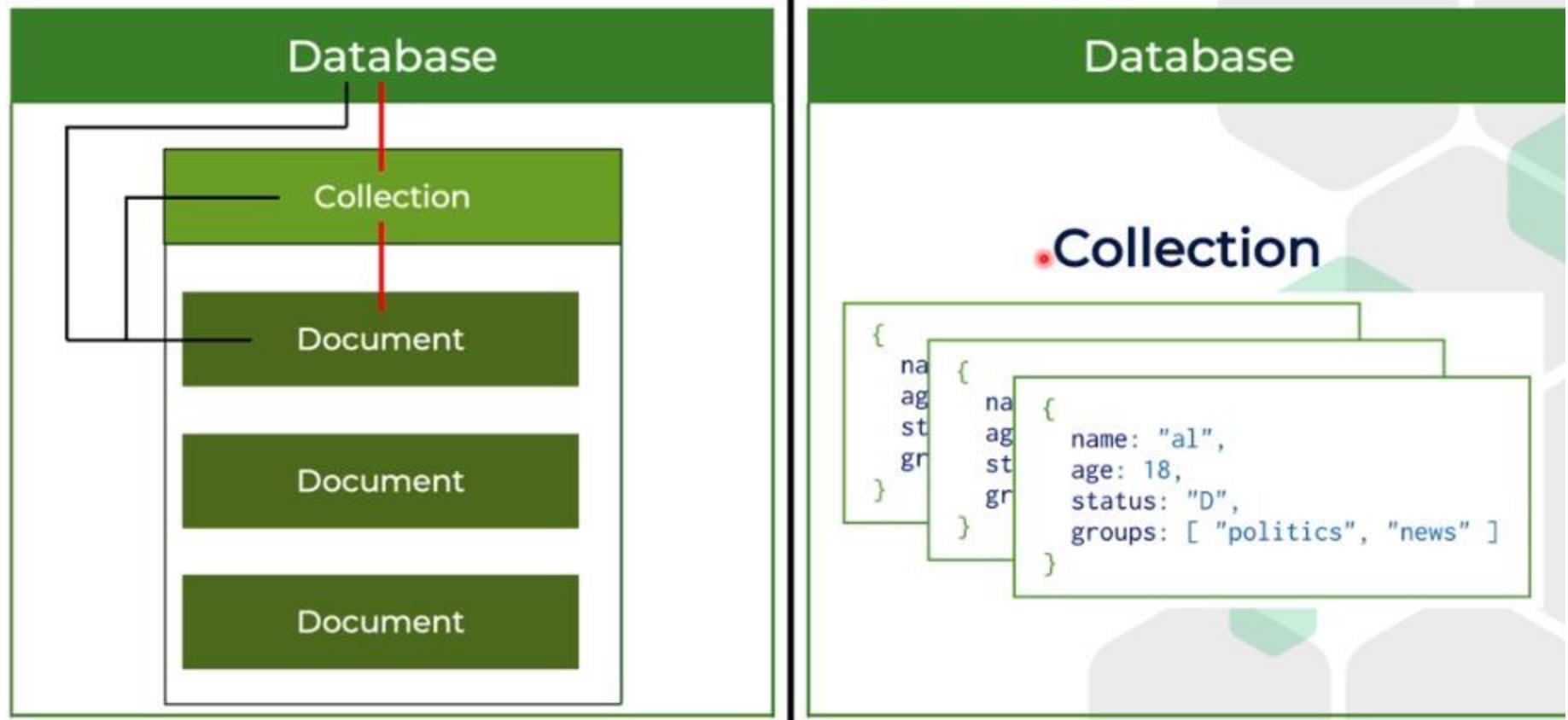




Funcionamiento de mongo







- ✓ En MongoDB se utiliza otro formato para almacenar los documentos, **BSON**, que no es más que una extensión de **JSON** que se representa de manera binaria.
- ✓ JSON (JavaScript Object Notation) es un formato estándar utilizado principalmente para transmitir datos entre un servidor web y una aplicación web.

Json (Bson)

```
{
  "nombre": "Hugo",
  "edad": "40",
  "direccion": {
    "pais": "Perú",
    "ciudad": "Lima"
  },
  "deporte": [
    {"nombre": "Fútbol"},
    {"nombre": "Natación"},
    {"nombre": "Voley"}
  ]
}
```

- ✓ Si MongoDB utilizase JSON para almacenar la información estaríamos **limitados** a seis tipos de datos,
- ✓ Por ejemplo, no dispone de un tipo de datos específico para la fechas, aunque podemos utilizar una cadena de texto. Tampoco dispone de un tipo de datos para almacenar el contenido de un fichero.
- ✓ BSON **proporciona** tipos de datos que no existen en JSON, como Date para fechas, BinData para información binaria, ObjectId para valores únicos (generalmente usado para el campo _id de las colecciones),...

Json vs bson

JSON - JavaScript Object Notation

Década de los 2000

MongoDB - Json

JSON solo admite una cantidad limitada de tipos de datos básicos.

Los objetos y propiedades JSON no tienen una longitud fija, lo que hace que el recorrido sea mas lento

JSON - JavaScript Object Notation

API

Archivos de configuración

Registro de mensajes

Almacenamiento en BBOD



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

BSON – JSON binario

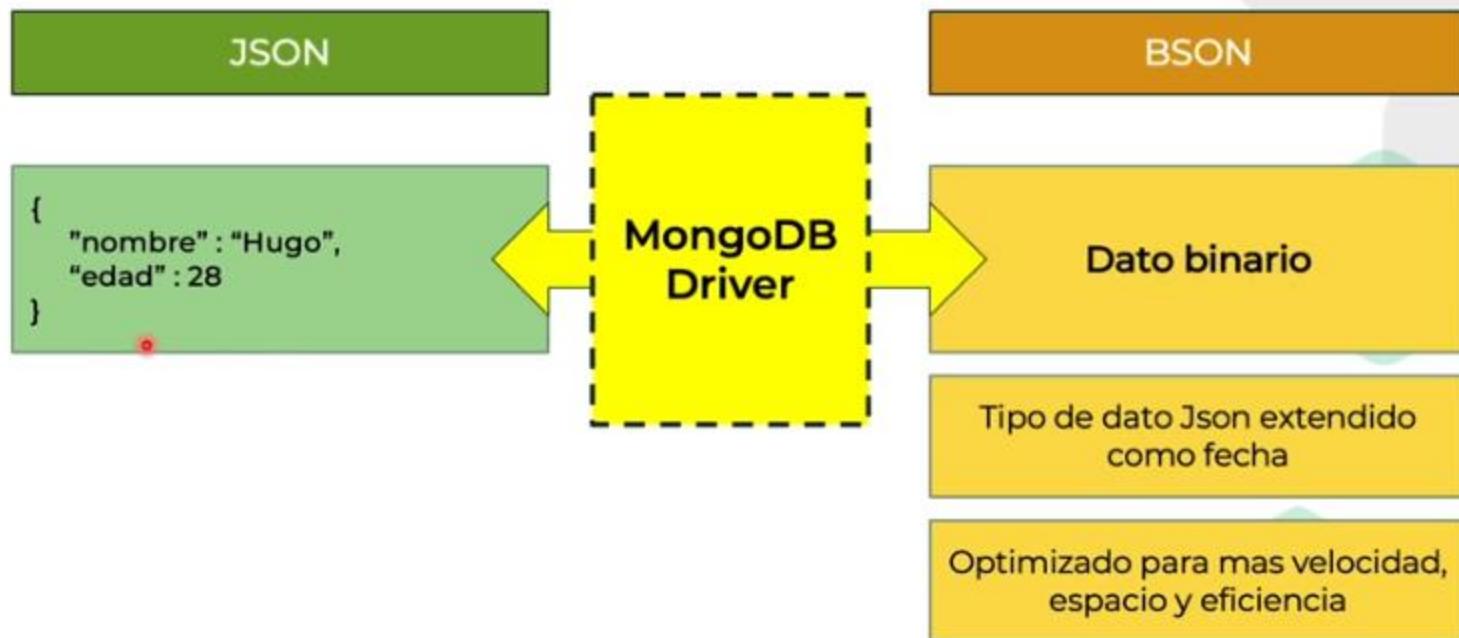
La estructura binaria, de BSON codifica información de tipo y longitud lo que permite recorrer mucho mas rápido en comparación con JSON.

BSON agrega algunos tipos de datos que no son nativos de JSON, como fechas y datos binarios

```
{"hello": "world"} →  
\x16\x00\x00\x00          // total document size  
\x02                      // 0x02 = type String  
hello\x00                  // field name  
\x06\x00\x00\x00world\x00  // field value  
\x00                      // 0x00 = type E00 ('end of object')  
  
{"BSON": ["awesome", 5.05, 1986]} →  
\x31\x00\x00\x00          // 0x31 = type Object  
\x04BSON\x00  
\x26\x00\x00\x00          // 0x26 = type String  
\x02\x30\x00\x08\x00\x00\x00awesome\x00  
\x01\x31\x00\x33\x33\x33\x33\x33\x33\x33\x14\x40  
\x10\x32\x00\xc2\x07\x00\x00  
\x00  
\x00
```



¿MongoDB usa JSON o BSON?



	JSON	BSON
Encoding	UTF-8 String	Binary
Data Support	String, Boolean, Number, Array, Object, null	String, Boolean, Number (Integer, Float, Long, Decimal128...), Array, null, Date, BinData
Readability	Human and Machine	Machine Only



Campo `_id`

El campo `_id` tiene el siguiente comportamiento

De forma predeterminada, mongoDB crea un índice único en el campo `_id`

Si el servidor recibe un documento que no tiene el campo `_id` primero, entonces el servidor moverá el campo al principio

```
cafeteria01> db.tipoCafe.find()  
[  
  {  
    _id: ObjectId("647e14ebf7acc1e45899664c"),  
    nombre: 'Carla',  
    bebida: 'Café Expreso',  
    cantidad: 2,  
    stock: 80,  
    disponible: true  
  }  
]
```

ObjectID (value)

Esta conformado por 12 bytes

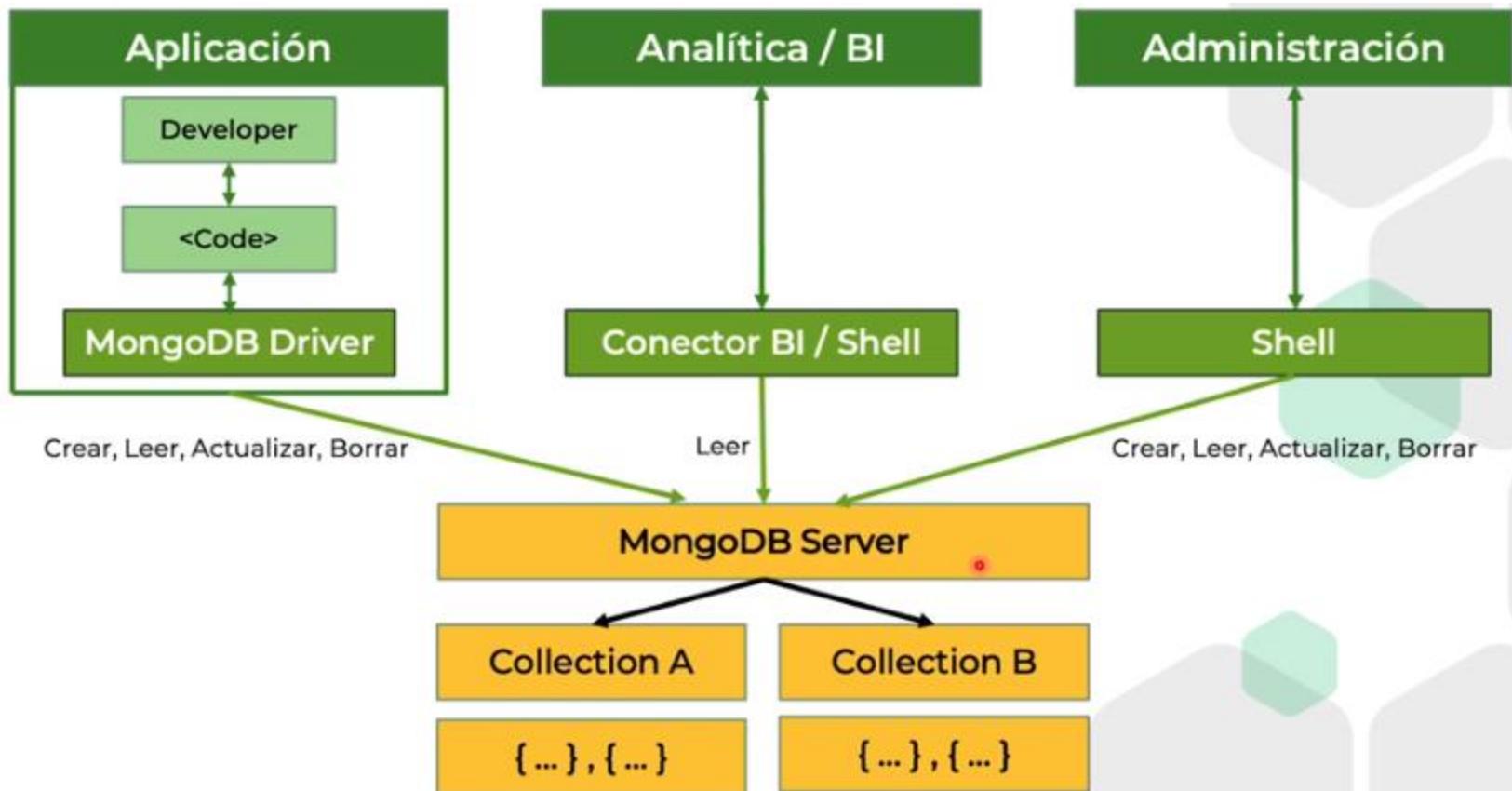
Una marca de tiempo de 4 bytes que representa la creación de Objectid

Un valor aleatorio de 5 bytes generado una vez por proceso. Este valor aleatorio es único para la maquina y el proceso

Un contador incremental de 3 bytes inicializados a un valor aleatorio.



Operaciones crud



Operaciones crud

Create

`insertOne(data, options)`

`insertMany(data, options)`

Update

`updateOne(filter, data, options)`

`updateMany(filter, data, options)`

`replaceOne(filter, data, options)`

Read

`find(filter, options)`

`findOne(filter, options)`

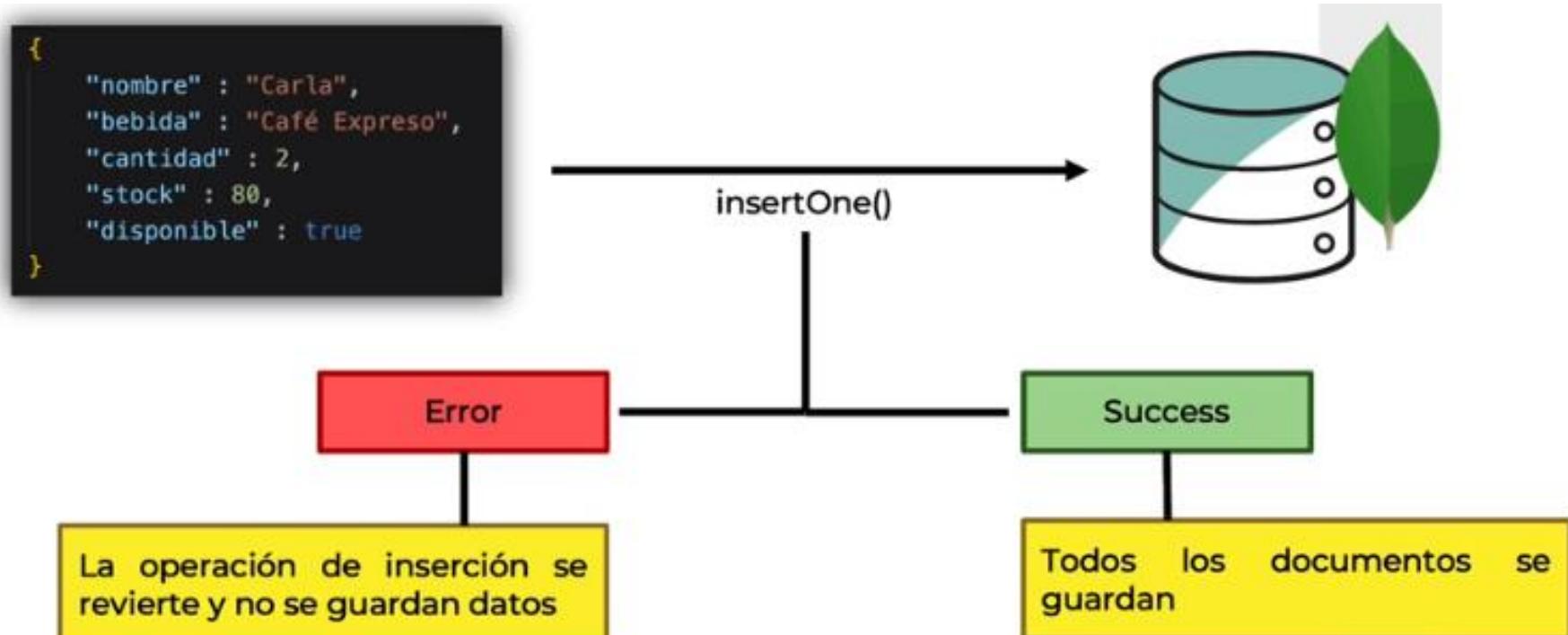
Delete

`deleteOne(filter, options)`

`deleteMany(filter, options)`



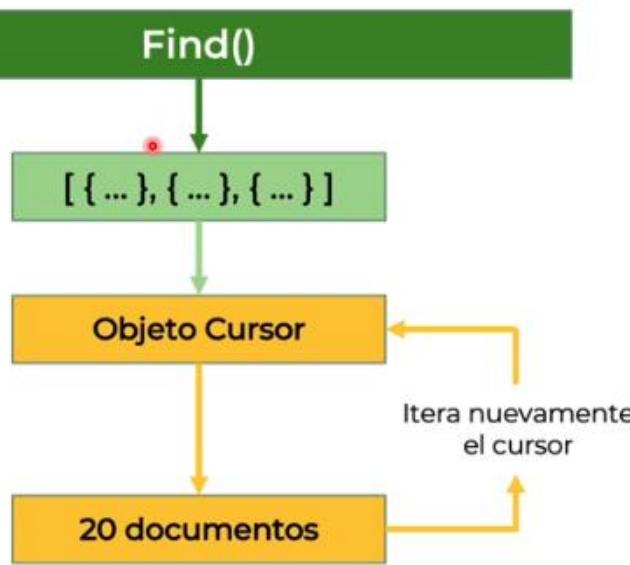
Atomicidad



Las Operaciones CRUD en MongoDB son Atómicas en los niveles de Documentos (incluidos documentos embebidos)



Cursor



El tamaño máximo del documento es de 16 megabytes.

Cursores

Es la forma de la que podemos modificar o leer un documento, el método que hemos utilizado desde un principio, **find**, siempre nos ha devuelto un cursor. Pero al estar utilizando menos de 20 registros no hemos tenido problemas, así que al introducir 100 registros veremos que nos hará falta el uso de **cursores**.

Documentos enbebidos

```
{  
  _id: ObjectId(...),  
  firstName: "John",  
  lastName: "King",  
  email: "john.king@abc.com",  
  salary: 3000,  
  additionalInfo: {  
    age: 30,  
    gender: male  
  },  
  address: {  
    street: "Upper Street",  
    house: "No 1",  
    city: "New York",  
    country: "USA",  
    phone: {  
      type: "Home",  
      number: 111-000-000  
    }  
  }  
}
```

Hasta 100 niveles de anidamiento

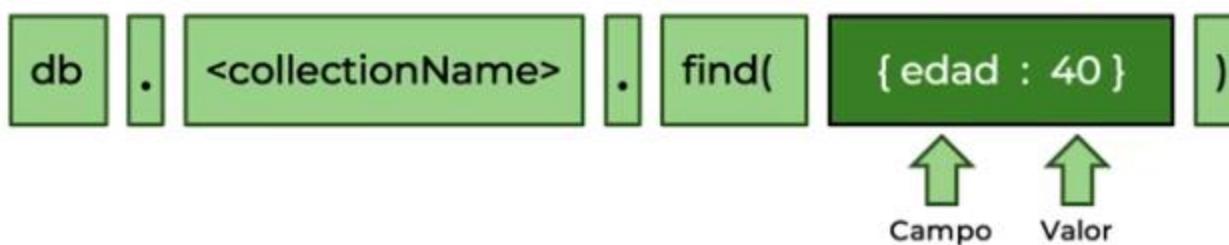
Buscar

Base
de
Datos
actual

Acceso a la colección

Metodo

Filtro



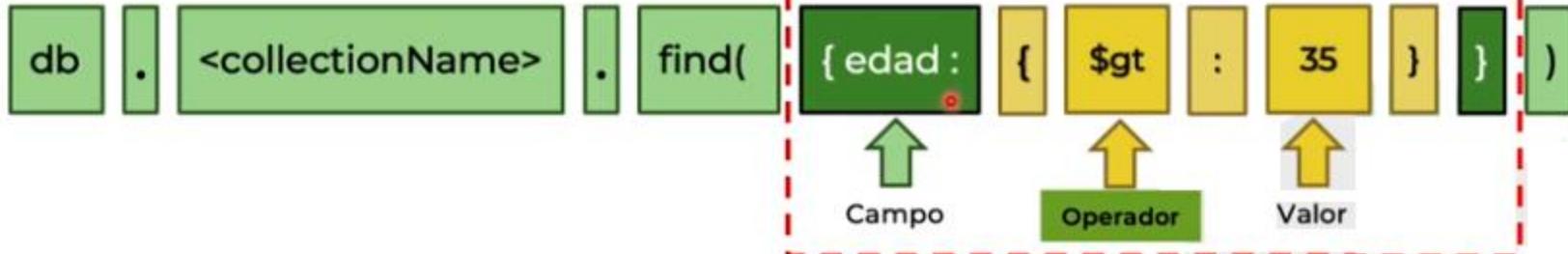
Base
de
Datos
actual

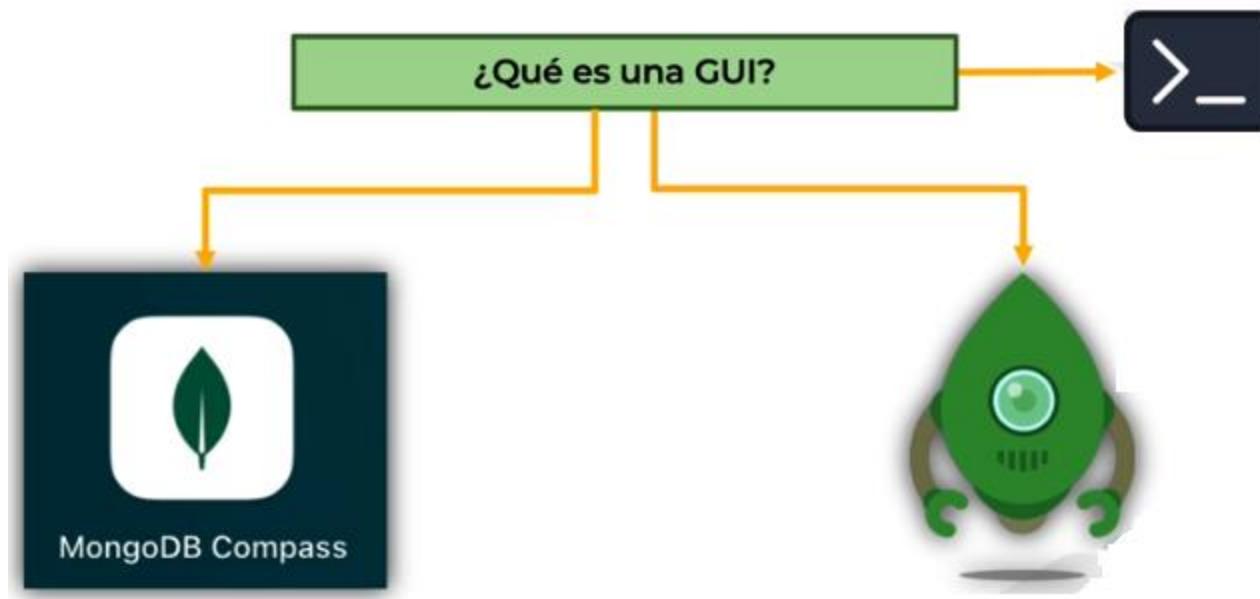
Acceso a la colección

Metodo

Rango

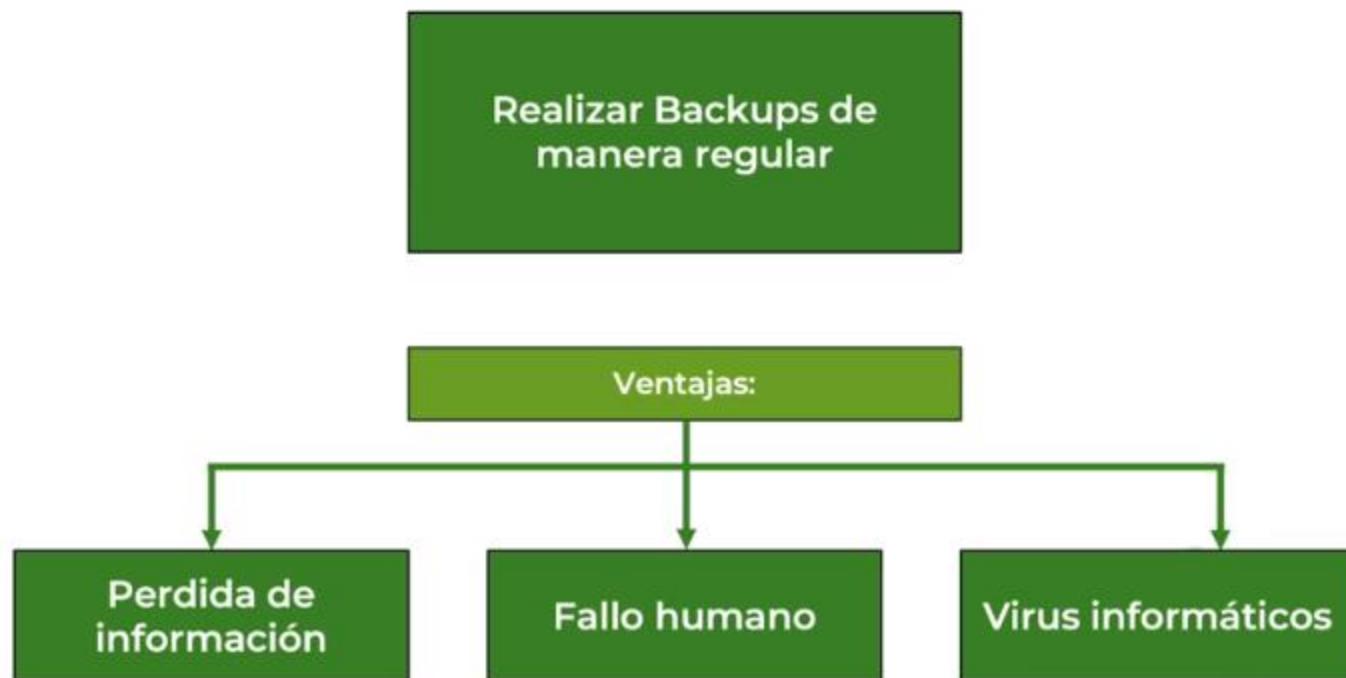
Filtro





MongoDB Compass es una herramienta de administración de base de datos que te permitirá utilizar una interfaz para interactuar con bases de datos de MongoDB. Con esta app podrás visualizar y manipular datos con facilidad, ya que te facilitará la gestión de estas bases de datos aunque no tengas experiencia.

Las copias de seguridad y la recuperación de datos implican el proceso de hacer un *respaldo de los datos* en caso de pérdida y configurar sistemas seguros con los que sea posible recuperar los datos como resultado.



Mongoimport

Se trata de una herramienta nativa de MongoDB. El hecho de que sea nativa no indica que venga por defecto en la instalación. Para ello, **deberemos instalar mongodb-org-tools**, que es un paquete de herramientas entre las que viene entre otros, mongoimport

Opciones:

--host=<hostname><:port>, -h=<hostname><:port> Default: localhost:27017

--port=<port> Default: 27017

--db=<database>, -d=<database>

--collection=<collection>, -c=<collection>

--jsonArray

--drop

```
mongoimport <fileName>.json -d <databaseName> -c <collectionName> --jsonArray --drop
```



UNIDAD 2

PROCESOS ETL.

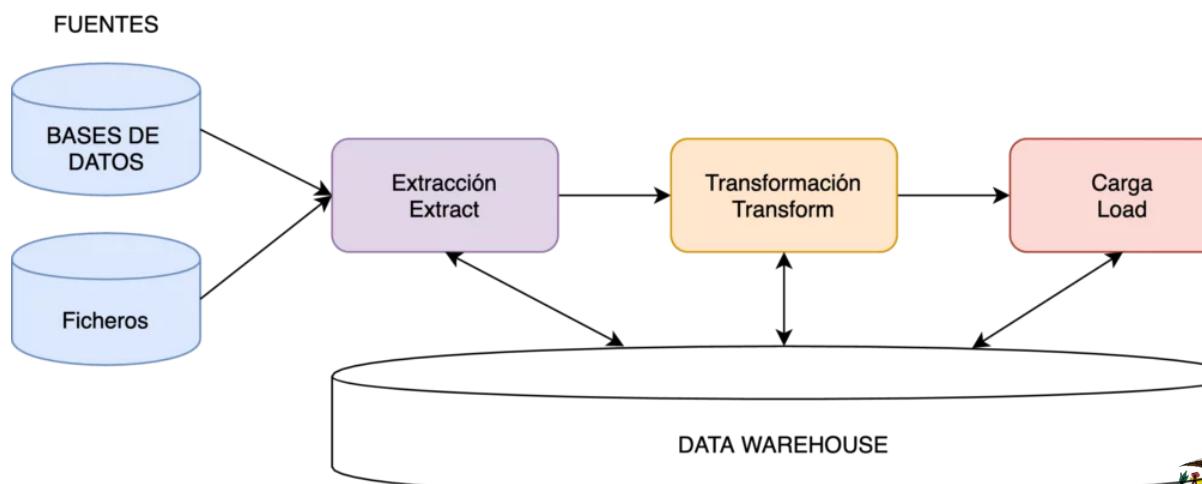


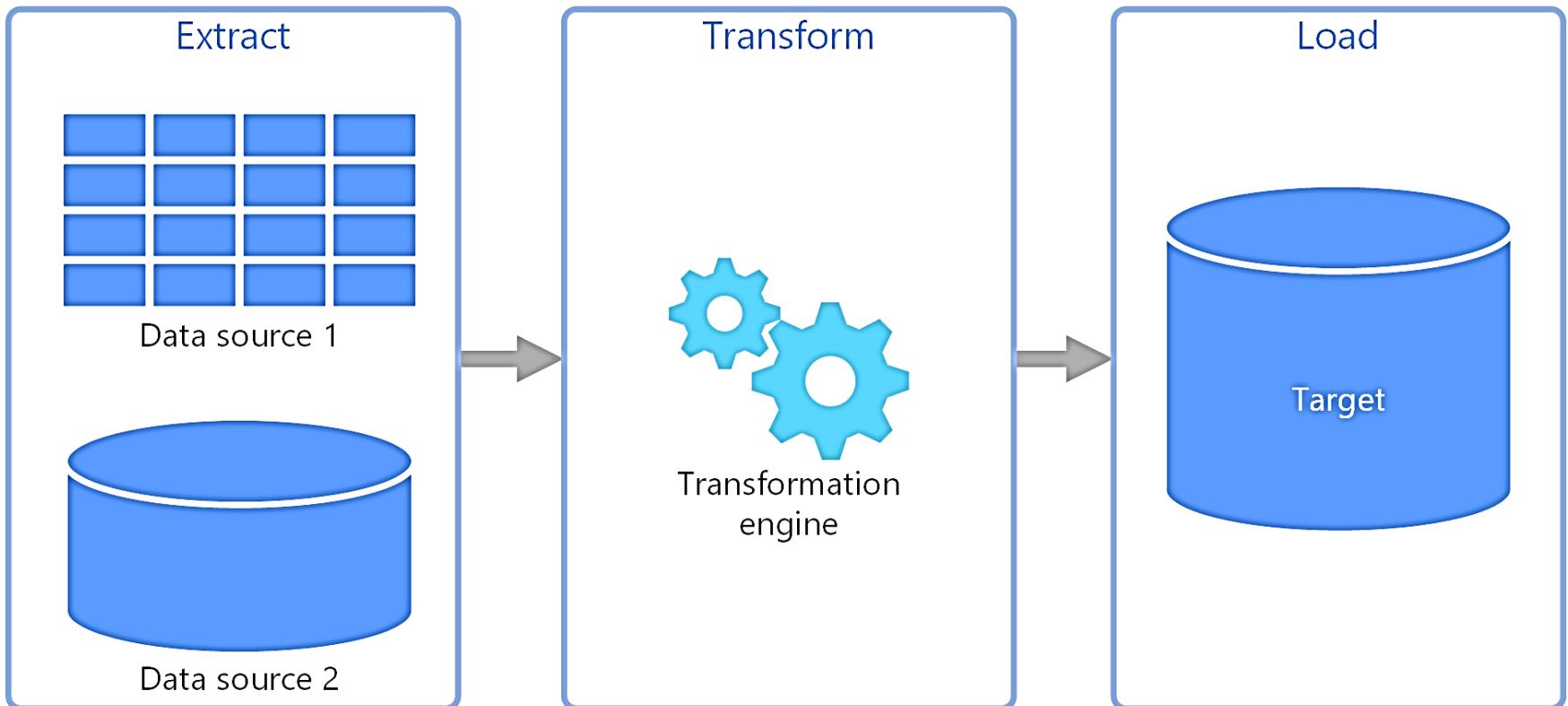
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

PROCESO ETL.

ETL -Extracción, transformación y carga (ETL)

- Es el proceso consistente en combinar datos de diferentes orígenes un gran repositorio central llamado almacenamiento de datos.
- Utiliza un conjunto de reglas comerciales para limpiar y organizar datos en bruto y prepararlos para el almacenamiento, el análisis de datos y el machine learning (ML).
- Puede abordar necesidades de inteligencia empresarial específicas mediante análisis de datos (como la predicción del resultado de decisiones empresariales, la generación de informes y paneles, la reducción de la ineficacia operativa y más).





Ventajas

- Analizan grandes cantidades de datos de empresariales con más sencillez que con procesos manuales.
- Aumentan la productividad en la recopilación y uso de datos, que se recopilan desde varias fuentes con más facilidad.
- Al mismo tiempo, al automatizar procesos, reducen los posibles fallos humanos.
- Unifican distintos orígenes de datos bajo un modelo capaz de proveer información de alta calidad que facilite la toma de decisiones de negocio
- Algunas soluciones no requieren contar con conocimientos técnicos, como saber escribir código, para ponerlas en funcionamiento. De esta forma, su manejo es más sencillo para algunos trabajadores.



ANÁLISIS DE HERRAMIENTAS

¿Cómo desarrollar un proceso ETL?

Programación de la ETL,

- El desarrollar un ETL desde cero conlleva la gran ventaja de la flexibilidad y las capacidades casi ilimitadas de la ETL final.
- Como contraparte esto conlleva tiempos de desarrollo elevados y una depuración compleja en caso de errores.
- El uso de lenguajes de programación como Python ayudan mucho a la consecución de logros por la gran cantidad de librerías existentes.

Utilizar herramientas de terceros.

- Las ventajas entre otras son: la simplicidad de realizar las transformaciones a través de interfaces visuales y un sistema de depuración mucho más ágil.
- Como contraparte el coste es mayor y la flexibilidad del proyecto es menor.
- **Herramientas ETL** existen muchas, pero entre las más utilizadas encontramos a Pentaho, Talend, **AWS Data Pipeline** o Alteryx.

En el diseño y **desarrollo de una ETL** la documentación es importante. Hay que decir que no existe un estándar para documentar este tipo de procesos, pero sí que existen ciertas recomendaciones como, por ejemplo, el desarrollo de gráficos que muestren el camino que siguen los datos.



En el pasado, las organizaciones escribían su propio código ETL. Ahora hay muchas herramientas ETL comerciales y de código abierto y servicios en la nube entre los que elegir.

Entre las capacidades típicas de estos productos se incluyen las siguientes:

- **Automatización completa y facilidad de uso:** las principales herramientas ETL automatizan todo el flujo de datos, desde las fuentes de datos hasta el almacén de datos de destino. Muchas herramientas recomiendan reglas para extraer, transformar y cargar los datos.
- **Una interfaz visual de arrastrar y soltar:** esta funcionalidad se puede utilizar para especificar reglas y flujos de datos.
- **Soporte para la gestión de datos complejos:** esto incluye asistencia con cálculos complejos, integraciones de datos y manipulaciones de cadenas.
- **Seguridad y cumplimiento:** las mejores herramientas ETL cifran los datos tanto en movimiento como en reposo y están certificadas conforme a las regulaciones de la industria o el gobierno, como HIPAA y GDPR.



Figure 1: Magic Quadrant for Data Integration Tools



Herramientas ETL

- 1.Talend Data Integration
- 2.Integrate.io
- 3.Fivetran
- 4.Skyvia
- 5.IRI Voracity
- 6.Sprinkle data
- 7.AWS Glue
- 8.Azure



Apache NiFi

Es una herramienta gratuita y open source mantenida y desarrollada por la Apache Software Foundation. Permite definir flujos o topologías de una forma visual, sencilla e intuitiva. Las unidades de procesamiento o carga de datos se denominan processors y se pueden extender con funcionalidad personalizada.

Pros:

- Licencia Apache 2.0
- Concepto de programación de flujo de datos
- Integración con Data Provenance y auditoría
- Posibilidad de manejar datos binarios
- Componentes disponibles
- Interfaz de usuario con grafos visuales
- Política de Usuarios (LDAP)

Contras:

- Falta de estadísticas por registro procesado
- Consumo de recursos elevado

AWS Data Pipeline

AWS Data Pipeline es la solución propuesta por la cloud de Amazon Web Services para transferir y transformar datos en la nube. Esta solución no es gratuita y Amazon cobra por uso. Permite realizar transformaciones de datos sencillas y se integra con buena parte de las tecnologías y servicios de Amazon en la nube.

Pros:

- Facilidad de uso
- Flexibilidad
- Precio razonable

Contras:

- Falta integración de funciones



Talend

Solución open source, con integraciones listas para usar con numerosas herramientas y tecnologías en cloud y on-premise. Su versión de pago ofrece componentes adicionales alrededor del gobierno del dato y gestión e incorpora la monitorización de los procesos de integración del dato y ETL.

Pros:

- Gran cantidad de integración con tecnologías externas listas para usar
- Versión open source gratuita disponible
- Interfaz sencilla con funcionalidad de arrastrar y soltar
- Extensible fácilmente con scripts y librerías en Java

Contras:

- Es posible que sean necesarios perfiles expertos en java para crear elementos ad-hoc

Pentaho

Pentaho es la herramienta ofrecida por la empresa Hitachi, también llamada Kettle para realizar transformaciones y migraciones de datos entre aplicaciones. Tiene versiones enterprise y open source (community edition). En la versión empresarial, la herramienta añade componentes adicionales a su catálogo.

Pros:

- Interfaz gráfica intuitiva y fácil de usar (arrastrar y soltar)
- Versión gratuita (community edition)

Contras:

- Plantillas limitadas



Azure Data Factory

Es el servicio cloud para ETLs en la nube de Azure. Tiene una interfaz de usuario que permite implementar flujos de datos, ETL y ELT sin la necesidad de usar código.

Pros:

- Integración con servicios de Azure
- Evita mantener infraestructura y su sobrecoste

Otras herramientas a considerar:

También hay alternativas menos populares como:

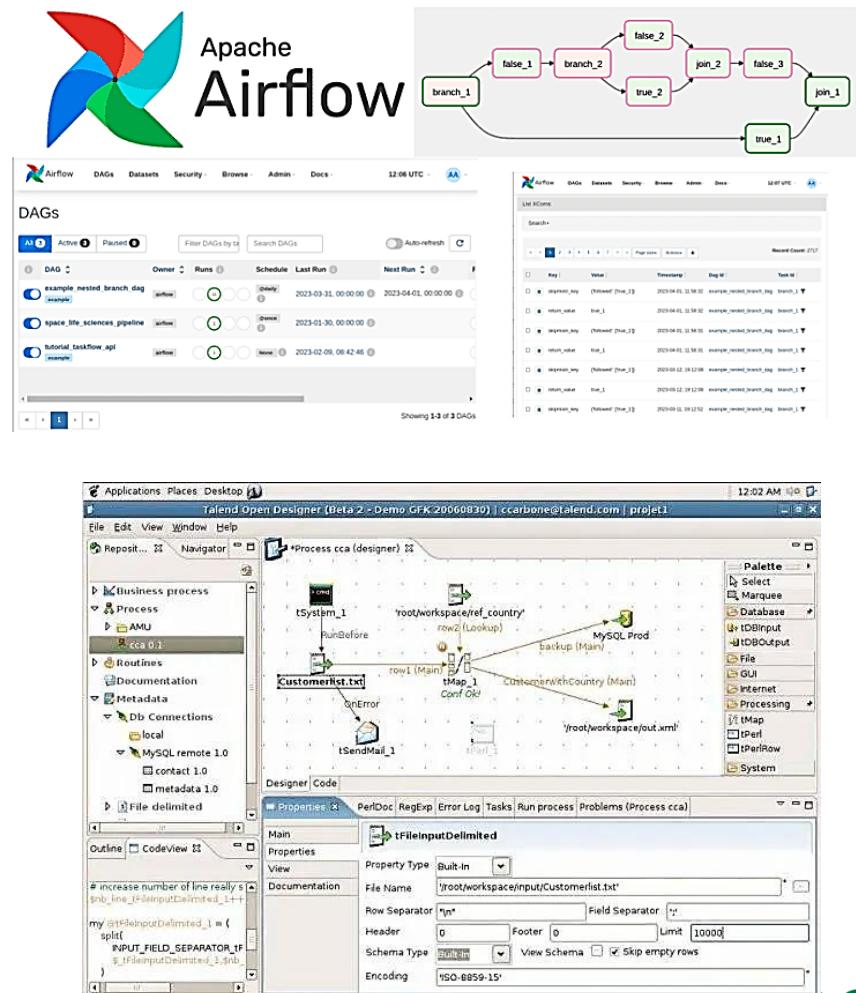
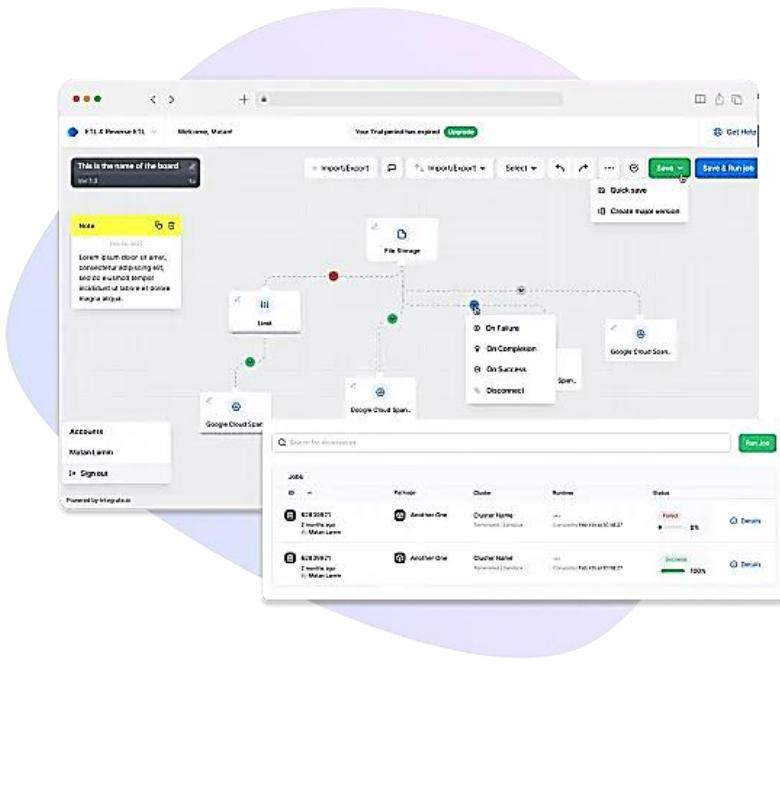
Xplenty, Striim, Fivetran, Stitch, Alooma, Skyvia, Panoply, Hevo Data, Matillion, FlyData (Amazon Redshift), Bluemetrix, Apache Hop.

Integrador de datos de Oracle

- Es parte del ecosistema de gestión de datos de Oracle.
- Es una opción para quienes ya utilizan otras aplicaciones de Oracle, como Hyperion Financial Management u Oracle E-Business Suite (EBS).
- Ofrece versiones locales y en la nube.
- Uno de los aspectos más exclusivos de ODI es que admite cargas de trabajo ETL, lo que puede resultar útil para muchos usuarios.
- Admite un amplio espectro de solicitudes de integración de datos, como cargas por lotes de gran volumen y servicios de datos de arquitectura orientada a servicios.
- La herramienta también admite la ejecución de tareas en paralelo, lo que ayuda a lograr un procesamiento de datos más rápido.



ENTORNO DE TRABAJO



EXTRACCIÓN

- Las herramientas de extracción, extracción, transformación y carga (ETL) de datos extraen o copian datos en bruto de múltiples fuentes y los almacenan en un área de ensayo.
- Un área de ensayo (o zona de aterrizaje) es un área de almacenamiento intermedio para almacenar temporalmente los datos extraídos.
- Las áreas de ensayo de datos suelen ser transitorias, lo que significa que su contenido se borra una vez que se completa la extracción de datos.
- El área de ensayo también puede conservar un archivo de datos para fines de resolución de problemas.
- La frecuencia con la que el sistema envía datos desde el origen de datos al almacenamiento de datos de destino depende del mecanismo subyacente de captura de datos modificados.



La extracción de datos comúnmente ocurre en una de las tres formas siguientes.

Notificación de actualización

- El sistema de origen le notifica cuando cambia un registro de datos.
- La mayoría de las bases de datos y aplicaciones web proporcionan mecanismos de actualización para admitir este método de integración de datos.

Extracción progresiva

- Algunos orígenes de datos no pueden proporcionar notificaciones de actualización, pero pueden identificar y extraer datos que se han modificado durante un período de tiempo determinado.
- El sistema busca cambios a intervalos periódicos, como una vez a la semana, una vez al mes o al final de una campaña. Sólo necesita extraer los datos que han cambiado.

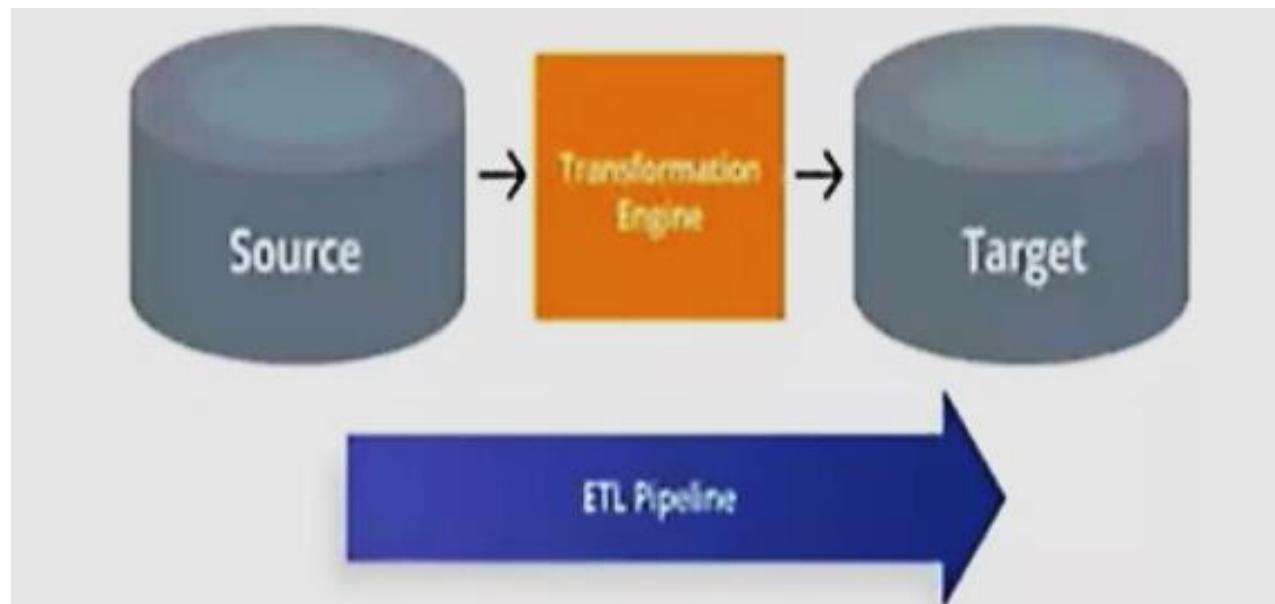
Extracción completa

- Algunos sistemas no pueden identificar los cambios de datos ni enviar notificaciones, por lo que recargar todos los datos es la única opción.
- Este método de extracción requiere que conserve una copia del último extracto para verificar qué registros son nuevos.
- Debido a que este enfoque implica grandes volúmenes de transferencia de datos, se recomienda solo para tablas pequeñas.



PROCESO

En definitiva, lo que permite este **proceso ETL (extraer, transformar, cargar)** es leer un gran volumen de datos, cargarlos para trabajarlos y, finalmente, **convertirlo en información valiosa**.



¿Cómo funciona la Fase de Extracción?

El objetivo de un proceso ETL es **producir datos limpios y accesibles que pueden utilizarse para analíticas u operaciones comerciales**. Esta primera fase, consiste en la ingesta de datos desde una o distintas fuentes de datos.

Tareas que se desempeñan en la fase de extracción:

- **Analizar el origen de los datos**: este es el primer paso de la fase de extracción. Los datos en bruto pueden extraerse de una gran variedad de fuentes.
- **Extraer los datos**: es la función principal y consiste en extraer la información desde los sistemas de origen.
- **Analizar los datos extraídos**: en el desarrollo de esta, se estudian las propiedades de los datos.
- **Verificar los datos extraídos**: en esta fase también se supervisa si los datos cumplen los requisitos establecidos en calidad y forma.
- **Convertir los datos a un formato preparado**: si es necesario, en este proceso se convierten los datos a un formato preparado para iniciar un proceso de transformación.



Limpieza de datos

- Siempre esta asociado a los objetivos determinados en el proceso de captura de datos.
- Esta definido por las características del negocio de la organización: ¿para qué se estan capturando datos? ¿qué tipo de datos se requieren? según las respuestas se requiere definir diversos procesos de limpieza o ninguno.
- Un buen diseño inicial de captura de datos que esta alineado con los requerimientos establecidos, hará prácticamente innecesario un proceso que es bastante costoso en recursos humanos.

La limpieza de los datos incluye 5 actividades principales:

- Depurar
- Corregir
- Estandarizar
- Relacionar
- Consolidar



¿POR QUÉ HACER LIMPIEZA DE DATOS?

1. Los algoritmos, métodos, modelos, análisis, etc. no tienen capacidad de intuición como los humanos, por lo que el éxito depende de los datos de entrada.
2. Si los datos no son de calidad, incluso el mejor algoritmo nos dará unas predicciones de mala calidad.
3. El guardar muchos datos puede ralentizar la toma de decisiones, puede hacer que se tenga una visión distorsionada o puede generar el estrés de tener demasiados datos y no saber qué hacer con ellos
4. Muchas veces los datos están “sucios”, es decir, tienen errores o discrepancias entre los diferentes campos que están en bases de datos distintas.

La limpieza de datos comprende:

- Igualar formatos
- Descartar campos
- Corregir errores ortográficos
- Dar formato a fechas
- Eliminar columnas duplicadas
- Borrar registros no útiles

“Las organizaciones actúan bajo la suposición de que la información de la que disponen es precisa y válida. Si la información no es válida, entonces no pueden responder de las decisiones basadas en ella.”

Bill Inmon (Padre del almacén de datos)



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Otras razones por las cuales hacer limpieza:

- Los consumidores mienten
 - No toda la información que se acumula es veraz.
 - Esta información desvirtúa la realidad.
 - Ocupa espacio.
- Los datos son redundantes
 - EJ: Una empresa puede tener los datos de un consumidor porque este usa su tarjeta de fidelidad y también porque participó en una promoción.
 - Hay que unir la información y eliminar toda aquella que está duplicada.
- La información se queda obsoleta
 - La información se mueve a gran velocidad y más cuando está ligada a seres vivos
- No toda la información es valiosa
 - Las empresas se empeñan que acumular más y más datos, con el objetivo de tener una base de datos brutal. Lo cierto es que no todos los datos “valen” igual y que no todos tienen la misma utilidad



TRANSFORMACIÓN

¿Qué es la transformación de datos?

En la transformación de datos, las herramientas de extracción, transformación y carga (ETL) transforman y consolidan los datos en bruto en el área de preparación para prepararlos para el almacenamiento de datos de destino. Puede implicar los siguientes tipos de cambios de datos.

Transformación básica de datos

Las transformaciones básicas mejoran la calidad de los datos eliminando errores, vaciando campos de datos o simplificando datos. Ejemplos de estas transformaciones.

- **Limpieza de datos:** Elimina errores y asigna datos de origen al formato de datos de destino. Por ejemplo, puede asignar campos de datos vacíos al número 0, asignar el valor de datos “Principal” a “P” o asignar “Secundario” a “S”.
- **Duplicación de datos:** Identifica y elimina los registros duplicados.
- **Revisión del formato de datos:** Convierte datos, como conjuntos de caracteres, unidades de medida y valores de fecha/hora, en un formato coherente. Por ejemplo, una empresa de alimentos puede tener diferentes bases de datos de recetas con ingredientes medidos en kilogramos y libras. ETL convertirá todo a libras.



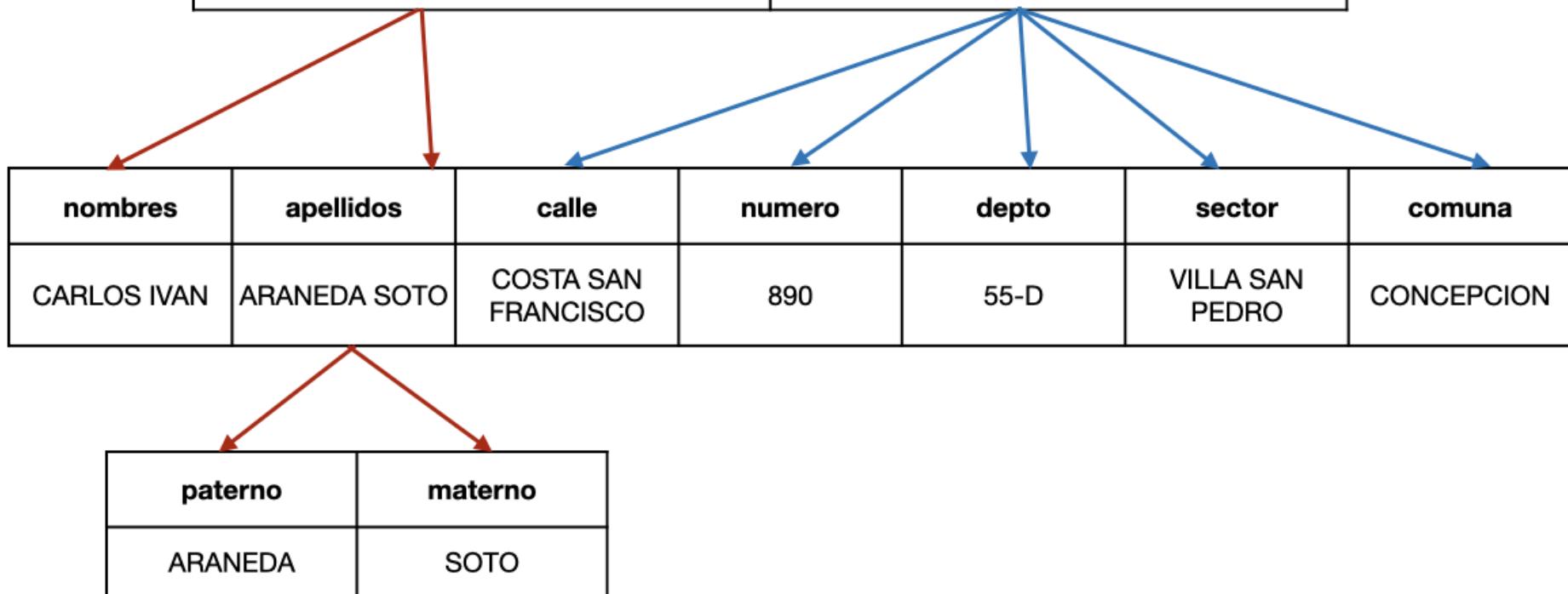
TRANSFORMACIÓN

Transformación avanzada de datos: Utilizan reglas comerciales para optimizar los datos y facilitar el análisis. Ejemplos de estas transformaciones.

- **Derivación:** Aplica reglas comerciales a sus datos para calcular nuevos valores a partir de valores existentes. Por ejemplo, puede convertir los ingresos en ganancias restando los gastos o calculando el costo total de una compra multiplicando el precio de cada artículo por la cantidad de artículos pedidos.
- **Vinculación:** Conecta los mismos datos de diferentes orígenes de datos. Por ejemplo, puede encontrar el costo total de compra de un artículo sumando el valor de compra de diferentes proveedores y almacenando solo el total final en el sistema de destino.
- **División:** Puede dividir una columna o un atributo de datos en varias columnas en el sistema de destino. Por ejemplo, si el origen de datos guarda el nombre del cliente como “María Isabel Pérez”, puede dividirlo en nombre, segundo nombre y apellido.
- **Integración:** Mejora la calidad de los datos al reducir una gran cantidad de valores de datos en un conjunto de datos más pequeño. Por ejemplo, los valores de las facturas de los pedidos de los clientes pueden tener muchos importes pequeños diferentes. Puede integrar los datos sumándolos durante un período determinado para crear una métrica de valor de vida útil del cliente (CLV).
- **Cifrado:** Puede proteger los datos confidenciales para cumplir con las leyes de datos o la privacidad de los datos agregando cifrado antes de que los datos se transmitan a la base de datos de destino.



nombre	direccion
CARLOS IVAN ARANEDA SOTO	COSTA SAN FRANCISCO 890 DEPARTAMENTO 55-D VILLA SAN PEDRO CONCEPCION



El lado más práctico del proceso de transformación

Dependiendo de las fuentes de datos, se aplicará:

- Seleccionar sólo ciertas columnas para su carga (pe: que las columnas con valores nulos no se carguen).
- Traducir códigos (pe: si la fuente almacena una “H” para Hombre y “M” para Mujer pero el destino tiene que guardar “1” para Hombre y “2” para Mujer).
- Codificar valores libres (pe: convertir “Hombre” en “H” o “Sr” en “1”).
- Obtener nuevos valores calculados (pe: total_venta = cantidad * precio).
- Unir datos de múltiples fuentes (pe: búsquedas, combinaciones, etc.).
- Calcular totales de múltiples filas de datos (pe: ventas totales de cada región).
- Generar campos clave en el destino.
- Transponer o pivotar (girando múltiples columnas en filas o viceversa).
- Dividir una columna en varias (pe: columna “Nombre: García, Miguel”; pasar a dos columnas “Nombre: Miguel” y “Apellido: García”).
- Aplicar para formas simples o complejas, la acción que en cada caso se requiera, por ejemplo:
 - Datos OK: entregar datos a la siguiente etapa (fase de carga).
 - Datos erróneos: ejecutar políticas de tratamiento de excepciones.



Depurar

- ✓ Según las necesidades es requisito depurar los datos obtenidos sobre todo si estos difieren de la forma a ser tratados posteriormente.
- ✓ Esta depuración siempre será necesaria cuando no haya congruencia entre el método de captura y el dato definido a capturar.
- ✓ Este proceso consiste en localizar e identificar los elementos individuales de información en las fuentes de datos y los aísla en las estructuras de destino.
- ✓ Por ejemplo: el caso más común es la captura de los nombres de personas y direcciones; separar el nombre completo en nombre, primer apellido, segundo apellido, o la dirección en: calle, número, piso, etc.

nombres, apellidos

nombres, apellido paterno, apellido materno

dirección

calle, numero, departamento, sector, comuna, ciudad, región, país



Corregir

Este proceso corrige los valores individuales de los atributos usando algoritmos de corrección y fuentes de datos externas. Por ejemplo: comprueba una dirección y el código postal correspondiente.

nombres	apellidos	genero	calle	numero	sector	comuna
CARLOS IVAN	ERNANDEZ	F	COSTA SAN FRANCISCO	890	VILLA SAN PEDRO	CONCEPCION

nombres	apellidos	genero	calle	numero	sector	comuna
CARLOS IVAN	HERNANDEZ	M	COSTA SAN FRANCISCO	890	VILLA SAN PEDRO	SAN PEDRO DE LA PAZ



Estandarizar

Este proceso aplica rutinas de conversión para transformar valores en formatos definidos (y consistentes) aplicando procedimientos de estandarización y definidos por las reglas del negocio. Por ejemplo: trato de Sr., Sra., etc. o sustituyendo los diminutivos de nombres por los nombres correspondientes.

nombres	apellidos	genero	calle	numero	sector	comuna
CARLOS IVAN	ERNANDEZ	H	O`HIGGINS	No. 890	CENTRO	SANTIAGO
JUAN	VERA	M	BERNARDO O`HIGGINS	700	MAIPU	STGO
MARCELA	VARAS SOLIS	F	ALAMEDA BERNARDO O`HIGGINS	#660	CENTRO	SANTIAGO

nombres	apellidos	genero	calle	numero	sector	comuna
			ALAMEDA BERNARDO O`HIGGINS	890		SANTIAGO



CARGA

En la carga de datos, las herramientas de extracción, transformación y carga (ETL) mueven los datos transformados desde el área de ensayo al almacenamiento de datos de destino. Para la mayoría de las organizaciones que usan ETL, el proceso está automatizado, bien definido, continuo y por lotes. A continuación se presentan dos métodos para cargar datos:

Carga completa

Todos los datos de la fuente se transforman y se mueven al almacenamiento de datos. La carga completa suele tener lugar la primera vez que carga datos de un sistema de origen en el almacenamiento de datos.

Carga progresiva

La herramienta ETL carga el delta (o la diferencia) entre los sistemas de origen y destino a intervalos regulares. Almacena la fecha del último extracto para que sólo se carguen los registros agregados después de esta fecha. Hay dos formas de implementar.

- **Transmisión de carga progresiva:** Si tiene volúmenes de datos pequeños, puede transmitir cambios continuos mediante canales de datos al almacenamiento de datos de destino. Cuando la velocidad de los datos aumenta a millones de eventos por segundo, puede usar el procesamiento de flujo de eventos para monitorear y procesar las secuencias de datos para tomar decisiones más oportunas.
- **Carga progresiva por lotes:** Si tiene grandes volúmenes de datos, puede recopilar cambios de datos de carga en lotes periódicamente. Durante este período de tiempo establecido, no se pueden realizar acciones ni en el sistema de origen ni en el de destino a medida que se sincronizan los datos.





ETL PROCESS

ANALYZE



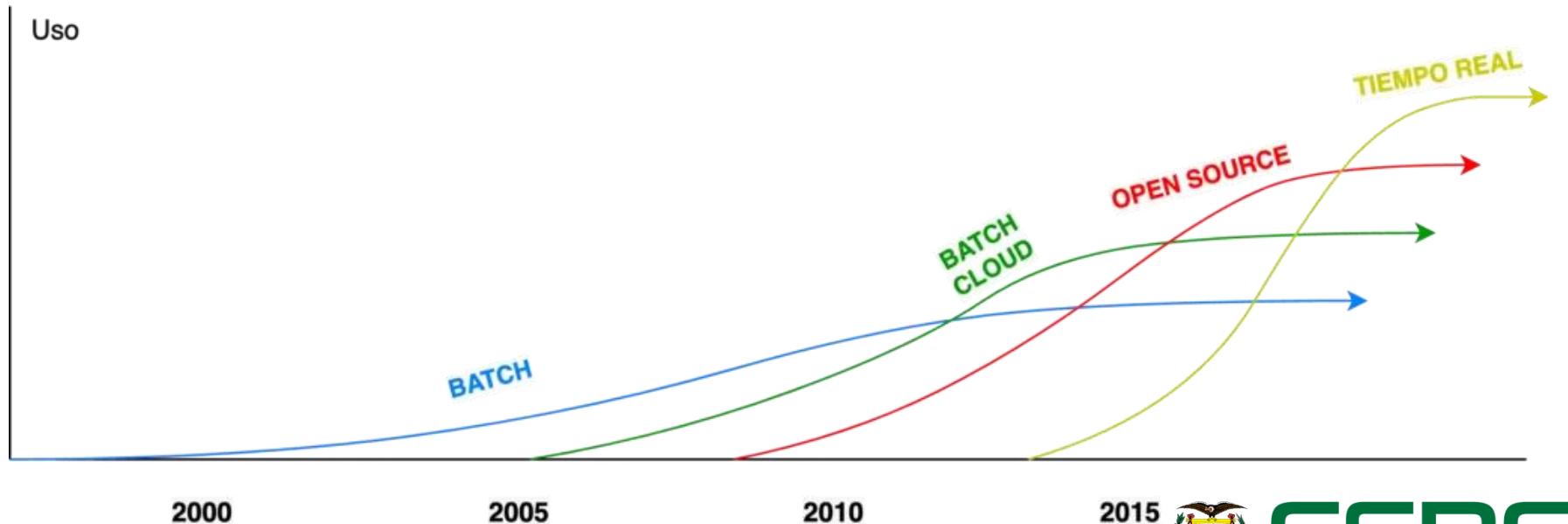
Astera
Enabling Data-Driven Innovation



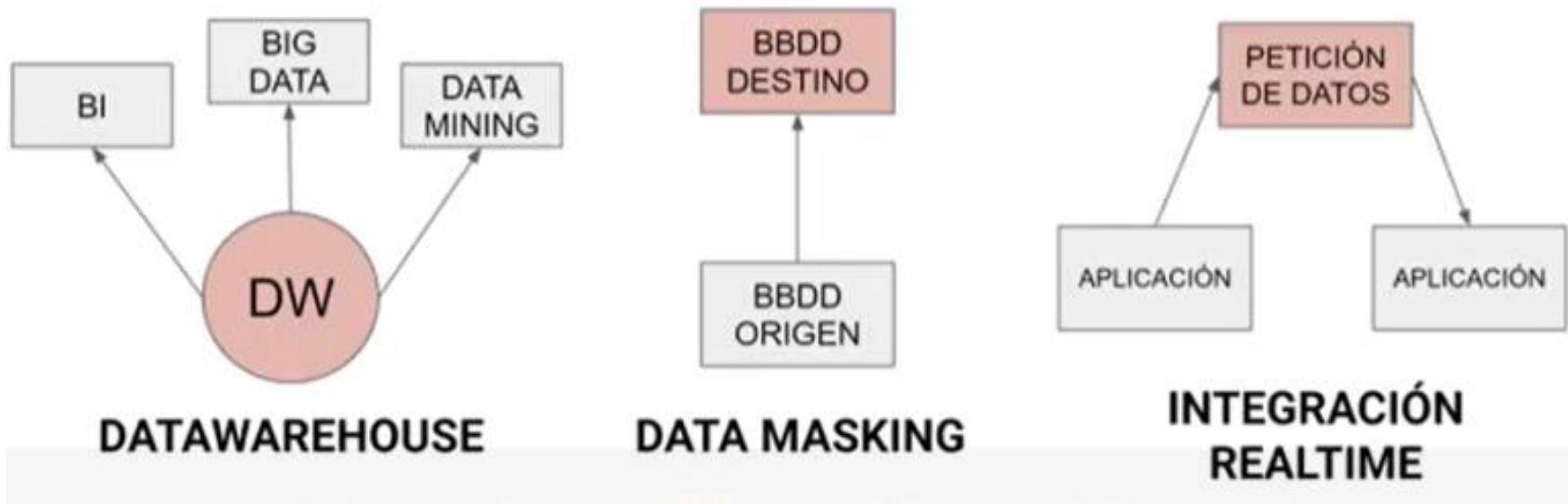
ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Una ETL comienza extrayendo datos de bases o ficheros donde pueden incluirse **bases de datos relacionales** como SQL Server, DB2 u **Oracle**, así como ficheros planos.

Debido a la disparidad de estos datos, la siguiente fase es la de transformación, en la cual se realizan las operaciones necesarias para homogeneizar los datos y prepararlos para almacenarlos en un Data Mart. Para realizar estas transformaciones, es frecuente apoyarse en un **Data Warehouse** que almacena datos con diversas características.



De hecho, en el proceso de la fase de carga de los datos podrás encontrar una migración de datos que puede variar entre **Data Warehouse**, **Data Masking** y la **Integración Realtime**.



Desafíos comunes en materia de ETL

- **Latencia de la red:** El experto en inteligencia de negocios Paul Mponzi menciona que los cuellos de botella en la transferencia de datos son comunes cuando hablamos de grandes volúmenes. Por ello, verifica que la plataforma que elijas tenga la capacidad necesaria y que tu negocio cuente con la infraestructura adecuada.
- **Calidad de los datos:** este es un problema común en la captura de datos. Procura que desde la entrada haya un proceso estandarizado que asegure la fiabilidad de los mismos y que haya protocolos de limpieza activos.
- **Mantenimiento a largo plazo:** con el tiempo, es probable que tu empresa deba escalar las soluciones, así que necesitas monitorizar los requerimientos a lo largo del tiempo. Implementar ETL suele requerir una inversión importante, así que no querrás que este proceso sea desperdiciado.



Desafíos comunes en materia de ETL

- **Pérdida de datos:** así como es vital llevar un seguimiento del análisis y procesamiento de la información, también es necesario encontrar aquellos momentos en que puedan darse «fugas» debido a filtros innecesarios, falta de capacidad de procesamiento y otros asuntos técnicos.
- **Falta de un banco de pruebas:** si tu organización requiere una solución a medida, es casi seguro que requerirás un banco de pruebas para revisar que la extracción, transformación y almacenamiento sean correctos. Verifica si tu proveedor te permite hacer este tipo de pruebas o por cuáles otros medios podrías realizarlas, y con qué regularidad.
- **Recursos insuficientes:** monitoriza la capacidad de almacenamiento, transferencia e incluso la salud de tus archivos, para evitar que el sistema ETL se vea ralentizado o abruptamente interrumpido, incluso durante cortos periodos de tiempo. De ello dependerá que la información saliente sea completa.

UNIDAD 3

**BASES DE DATOS MULTIDIMENSIONALES,
DESNORMALIZACIÓN, MODELO
MULTIDIMENSIONAL, METADATA.**



BUSINESS INTELLIGENCE - INTELIGENCIA DE NEGOCIOS

- Business Intelligence (BI) es un software que ingiere datos empresariales y los presenta en vistas fáciles de usar, como informes, cuadros de mando, tablas y gráficos. El análisis de estos datos ayuda a las empresas a obtener información procesable e informar la toma de decisiones
- Combina análisis de negocios, minería, visualización, herramientas e infraestructura de datos, para ayudar a las empresas a tomar decisiones basadas en los datos.
- Implica contar con una vista integral de todos los datos de la organización.
- Consiste en usar estos datos para impulsar el cambio, eliminar las ineficiencias y adaptarse rápidamente a los cambios del mercado o la demanda.
- Surgió por los 1960 como un sistema para compartir información entre organizaciones
- En la década de 1980 se desarrolló aún más junto con los modelos informáticos. Se utilizó para tomar decisiones y transformar datos en información antes de convertirse en un producto específico de los equipos de BI con soluciones de servicio basadas en TI.
- En las soluciones de BI modernas, se priorizan factores como el análisis de autoservicio flexible, los datos gobernados en plataformas confiables, la capacitación de los usuarios corporativos y la rapidez para obtener información.



Entre los procesos de BI se incluyen los siguientes:

- **Minería de datos:** usar bases de datos, estadísticas y aprendizaje automático para descubrir tendencias en conjuntos de datos más grandes.
- **Generación de informes:** compartir análisis de datos con las partes interesadas para que todos puedan sacar sus propias conclusiones y tomar decisiones.
- **Métricas de rendimiento y valores de referencia:** comparar los datos del rendimiento actual con los datos históricos para hacer un seguimiento del rendimiento frente a los objetivos.
- **Análisis descriptivos:** usar un análisis de datos preliminar para descubrir qué ocurrió.
- **Consultas:** el usuario realiza preguntas específicas relacionadas con los datos y la BI extrae las respuestas de los conjuntos de datos.
- **Análisis estadístico:** a partir de los resultados de análisis descriptivos, se exploran aún más los datos a través de estadísticas, por ejemplo, para determinar cómo ocurrió una tendencia y por qué.
- **Visualización de datos:** transformar el análisis de datos en representaciones visuales, como gráficos e histogramas, a fin de consumir más fácilmente los datos.
- **Análisis visual:** explorar los datos a través de la narración visual de historias para compartir información sobre la marcha y permanecer en el flujo de análisis.
- **Preparación de datos:** recopilar varias fuentes de datos, identificar las dimensiones y las medidas y preparar los datos para el análisis.



¿Por qué es importante la inteligencia de negocios?

Los analistas pueden aprovechar las innumerables ventajas del Business Intelligence para establecer valores de referencia de rendimiento y de la competencia. De esta manera, la organización podrá operar de manera más ágil y eficiente.

Algunas de las alternativas que ofrece el Business Intelligence para tomar decisiones basadas en los datos:

- Identificar maneras de aumentar los beneficios
- Analizar el comportamiento de los clientes
- Comparar los datos con la información de la competencia
- Hacer un seguimiento del rendimiento
- Optimizar las operaciones
- Predecir el éxito
- Identificar las tendencias del mercado
- Detectar los inconvenientes o problemas



Business Analytics (Análisis de Negocios)

- Se refiere a un conjunto de **métodos y herramientas** utilizados para **analizar datos empresariales** con el fin de obtener información relevante y tomar decisiones estratégicas.
- Emplea técnicas estadísticas y matemáticas avanzadas, así como tecnologías de análisis de datos, para examinar grandes conjuntos de información empresarial.
- El **objetivo principal es mejorar el rendimiento de una organización mediante el análisis exhaustivo de datos**. Para lograr esto, se recopilan y analizan diversos tipos de datos, como información financiera, datos de ventas, datos de clientes/as y datos de producción.
- Mediante el uso de técnicas y modelos estadísticos, como análisis descriptivos, análisis predictivos y análisis prescriptivos, el BA permite descubrir patrones, tendencias y relaciones en los datos.
- Se aplica en distintas áreas para obtener una comprensión profunda de las operaciones y tomar decisiones informadas.
- Para llevar a cabo el análisis de negocios, se utilizan herramientas como software de visualización de datos, técnicas de minería de datos y algoritmos de aprendizaje automático para generar predicciones y modelos.
- Ofrece un conjunto enormemente diverso de aplicaciones para mejorar el rendimiento empresarial.



Usos del Business Analytics

- 1.Optimización de procesos y operaciones empresariales:** ayuda a identificar áreas de mejora y optimizar los procesos y operaciones internas.
- 2.Análisis de mercado y segmentación de clientes:** se puede analizar el mercado en detalle, identificar tendencias y segmentar a los clientes de manera efectiva.
- 3.Predicción de demanda y comportamiento de consumidores/as:** con el análisis de datos, es posible predecir la demanda futura y comprender el comportamiento de los/as consumidores/as.
- 4.Mejora de la eficiencia en la cadena de suministro:** al analizar los datos de la cadena de suministro, es posible identificar cuellos de botella, optimizar la gestión de inventarios y mejorar la planificación de la producción.
- 5.Detección y prevención de fraudes y riesgos:** utilizando técnicas de análisis, el BA ayuda a identificar patrones y anomalías que pueden indicar fraudes o riesgos.
- 6.Toma de decisiones basada en datos y análisis:** proporciona información precisa y oportuna para respaldar la toma de decisiones empresariales.
- 7.Identificación de oportunidades de crecimiento y expansión:** el análisis de datos permite identificar nuevas oportunidades de mercado, evaluar el potencial de crecimiento y apoyar la expansión de la empresa.
- 8.Evaluación y seguimiento del desempeño empresarial:** el BA proporciona métricas y KPIs relevantes para evaluar el desempeño de la empresa.



BI VS. BA: Descriptivo vs. Predictivo

- Cuál modelo es mejor no hay una respuesta definitiva. **Escoger entre BI y BA depende de las necesidades de la compañía y de las respuestas que quiera obtener a partir de sus datos.**
- El objetivo de las dos metodologías es tomar decisiones de negocios basada en datos
- BI se preocupa por el *qué* y el *cómo*, no por el *por qué*.
- BA sí se pregunta *por qué* pasan las cosas para poder predecir futuros comportamientos o tendencias.
- John Meyers, Director Ejecutivo de Investigación de EMA, dice que para tomar la decisión entre BI y BA hay que hacerse varias preguntas
 - ¿Qué necesita que resuelva el sistema?
 - ¿Quién va a utilizarlo?
 - ¿Qué tanto control y visibilidad se necesita en el proceso?
 - ¿Está más interesado en entender cómo llegó al punto en el que está o hacia dónde debería dirigir su compañía?
 - ¿Qué quiere saber?:
 - ¿Quiénes fueron sus 10 mejores clientes el año pasado?
 - ¿Quiénes serán sus 10 mejores clientes el próximo año?



Diferencias y similitudes

- **BI** se encarga de recopilar y analizar los datos en tiempo real con el fin de proporcionar información valiosa sobre posibles mejoras; así como de prever posibles problemáticas generando una toma de decisiones más segura y oportuna.
- En **Business Analytics** su intención es el de predecir las **tendencias** manteniendo un enfoque hacia la **mejora** y la **preparación** para el cambio de la empresa.

	Business Intelligence	Business Analytics
Recopila, Analiza y Visualiza: (Minería, paneles y capacidad de análisis varios)	si	si
Puntos Débiles: Identifica puntos débiles de la empresa y ofrece soluciones para optimizar	si	si
Reportes: Presenta los datos de manera organizada en reportes	si	si
Análisis Descriptivo: Crea resúmenes de datos históricos visuales	si	no
Analítica de diagnóstico: Identifica la fuente de los problemas descubiertos por la analítica descriptiva	si	no
Analítica Predictiva: Realiza predicciones basadas en datos recopilados	no	si
Analítica Prescriptiva: Ofrece soluciones para problemas encontrados por análisis prescriptivos y descubiertos en los datos	no	si



Aplicación de la BI y el BA en el mundo real

Ejemplo: de aplicaciones del Business Intelligence y Business Analytics.

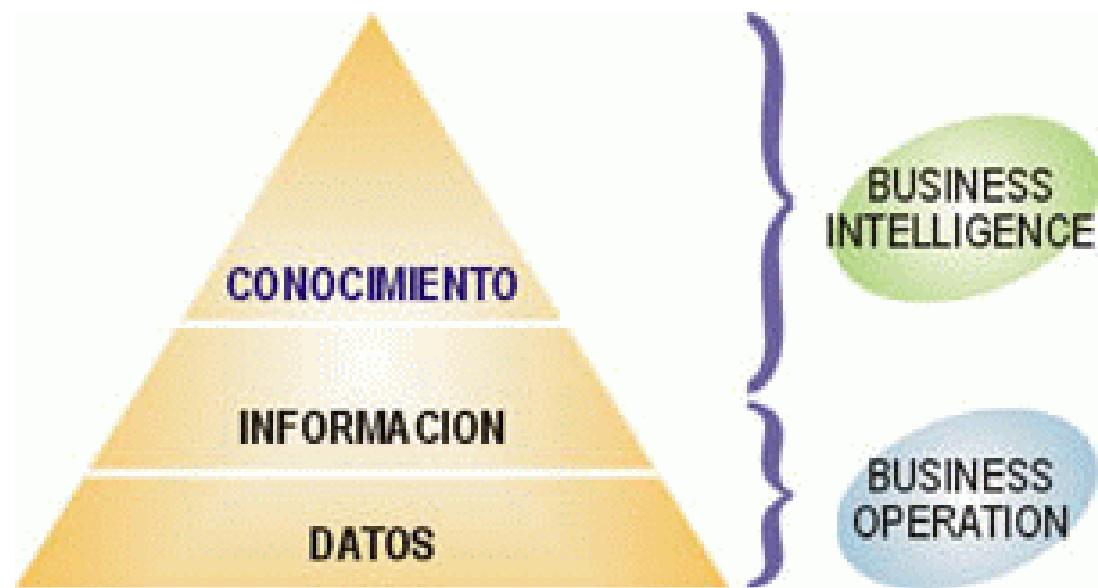
Imagine que vende joyas que fabrica en su casa a través de una tienda en línea.

- La **inteligencia de negocios** proporciona informes útiles de la situación pasada y actual de su empresa.
 - Esta le indica que las ventas de sus pendientes de plumas azules han aumentado en Utah las últimas tres semanas.
 - Como resultado, decide fabricar más pendientes de plumas azules para satisfacer la demanda.
- El **análisis de negocios o Business Analytics** se pregunta,
 - “¿por qué aumentaron las ventas de pendientes de plumas azules en Utah?”.
 - Al extraer los datos, descubre que la mayor parte del tráfico proviene de una publicación de una blogger de moda de Salt Lake City que usó los pendientes.
 - A partir de esta información, decide enviar pendientes de regalo a otros bloggers de moda conocidos en todo Estados Unidos.
 - Consulta la información de ventas anterior para anticipar cuántos pendientes deberá fabricar y cuántos suministros tendrá que encargar para mantenerse al día con la demanda si los bloggers realizan una publicación acerca de los pendientes.

Datos, información, conocimiento

¿En qué se diferencia el conocimiento de los datos y de la información? En una conversación informal, los tres términos suelen utilizarse indistintamente y esto puede llevar a una interpretación libre del concepto de conocimiento.

Quizás la forma más sencilla de diferenciar los términos sea pensar que los datos están localizados en el mundo y el conocimiento está localizado en agentes de cualquier tipo (personas, empresas, máquinas...), mientras que la información adopta un papel mediador entre ambos.



Datos

- Los datos son la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones.
- También se pueden ver como un conjunto discreto de valores, que no dicen nada sobre el por qué de las cosas y no son orientativos para la acción.
- Un número telefónico o un nombre de una persona, por ejemplo, son datos que, sin un propósito, una utilidad o un contexto no sirven como base para apoyar la toma de una decisión.
- Los datos pueden ser una colección de hechos almacenados en algún lugar físico como un papel, un dispositivo electrónico (CD, DVD, disco duro...), o la mente de una persona.
- Como cabe suponer, los datos pueden provenir de fuentes externas o internas a la organización, pudiendo ser de carácter objetivo o subjetivo, o de tipo cualitativo o cuantitativo, etc.



Información

- Se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre. Los datos se pueden transformar en información añadiéndoles valor:
 - **Contextualizando:** se sabe en qué contexto y para qué propósito se generaron.
 - **Categorizando:** se conocen las unidades de medida que ayudan a interpretarlos.
 - **Calculando:** los datos pueden haber sido procesados matemática o estadísticamente.
 - **Corrigiendo:** se han eliminado errores e inconsistencias de los datos.
 - **Condensando:** los datos se han podido resumir de forma más concisa (agregación).
- La información es la comunicación de conocimientos o inteligencia, y es capaz de cambiar la forma en que el receptor percibe algo, impactando sobre sus juicios de valor y sus comportamientos.
- Información = Datos + Contexto (añadir valor) + Utilidad (disminuir la incertidumbre)



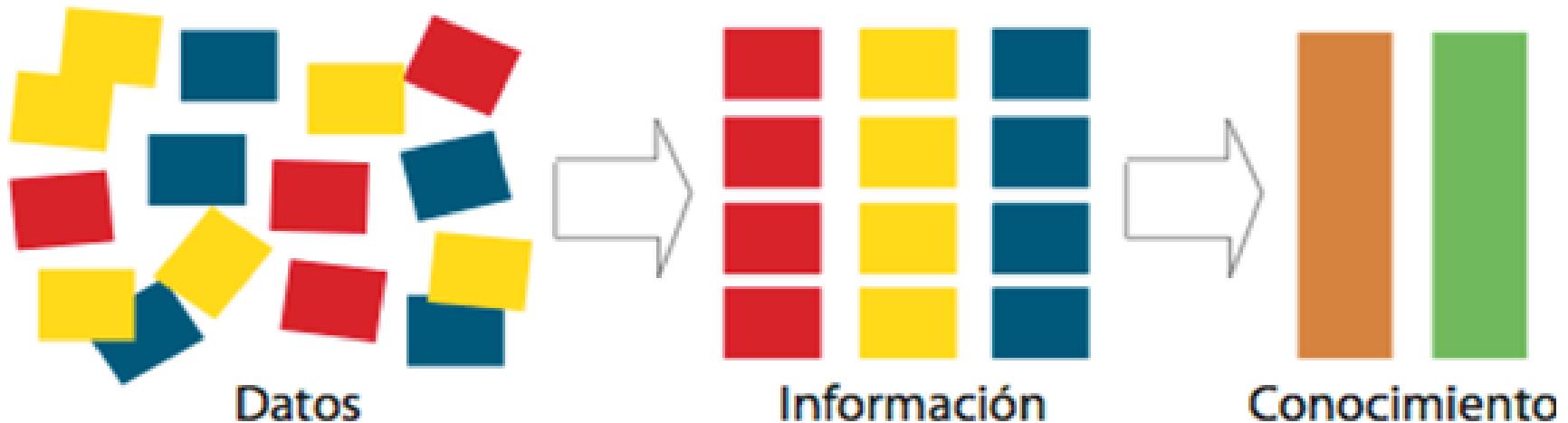
Conocimiento

- El conocimiento es una mezcla de experiencia, valores, información y know-how que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción.
- Se origina y aplica en la mente de los conoecedores. En las organizaciones con frecuencia no sólo se encuentra dentro de documentos o almacenes de datos, sino que también está en rutinas organizativas, procesos, prácticas, y normas.
- El conocimiento se deriva de la información, así como la información se deriva de los datos.
- Para que la información se convierta en conocimiento es necesario realizar acciones como:
 - Comparación con otros elementos.
 - Predicción de consecuencias.
 - Búsqueda de conexiones.
 - Conversación con otros portadores de conocimiento.



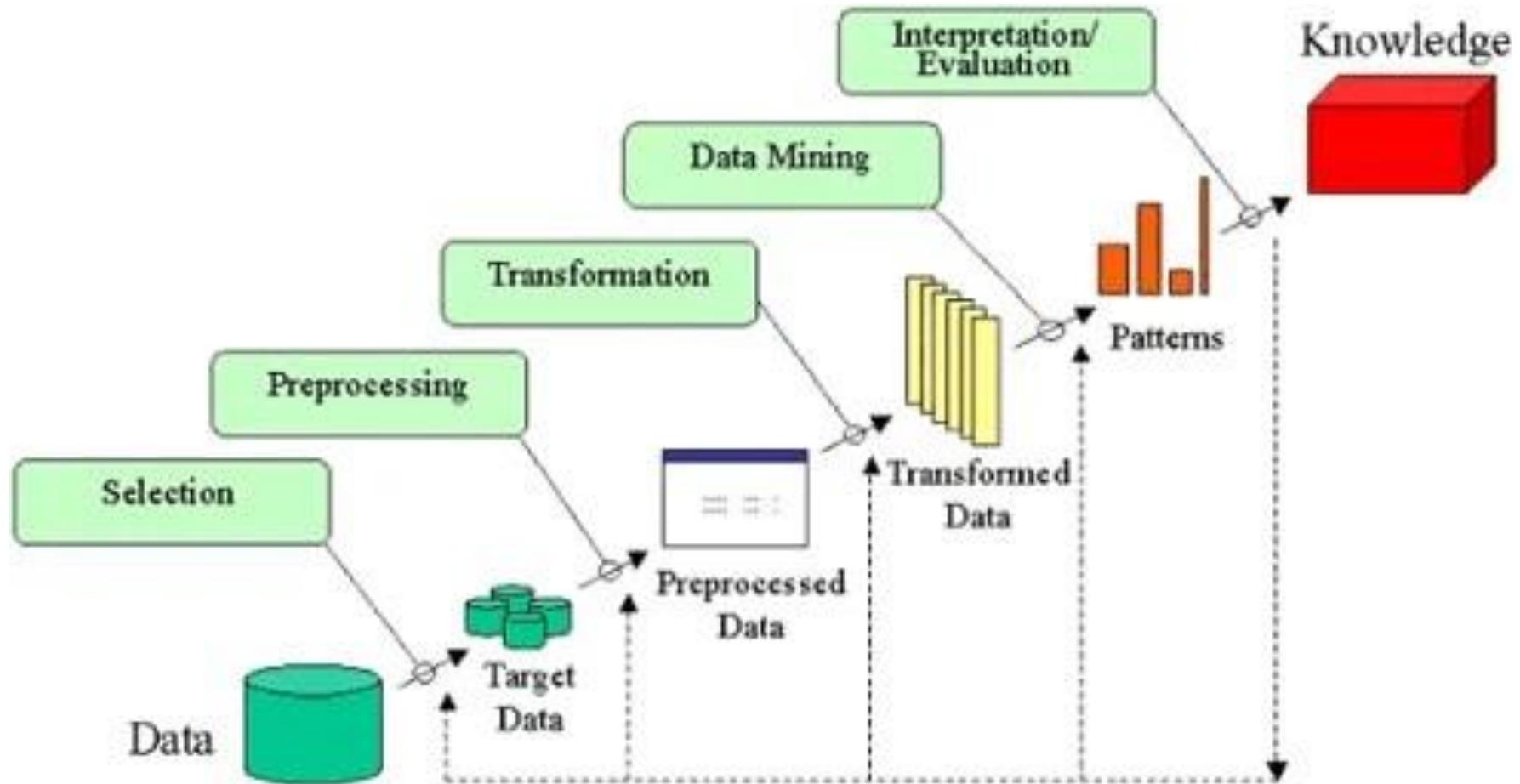
- Información y conocimiento no son sinónimos.
- La información es el término general y el conocimiento es información bajo determinadas condiciones.
- Cuando la información se encuentra en las condiciones adecuadas para propagarse, la llamamos conocimiento

Tipos de contenidos y su evolución



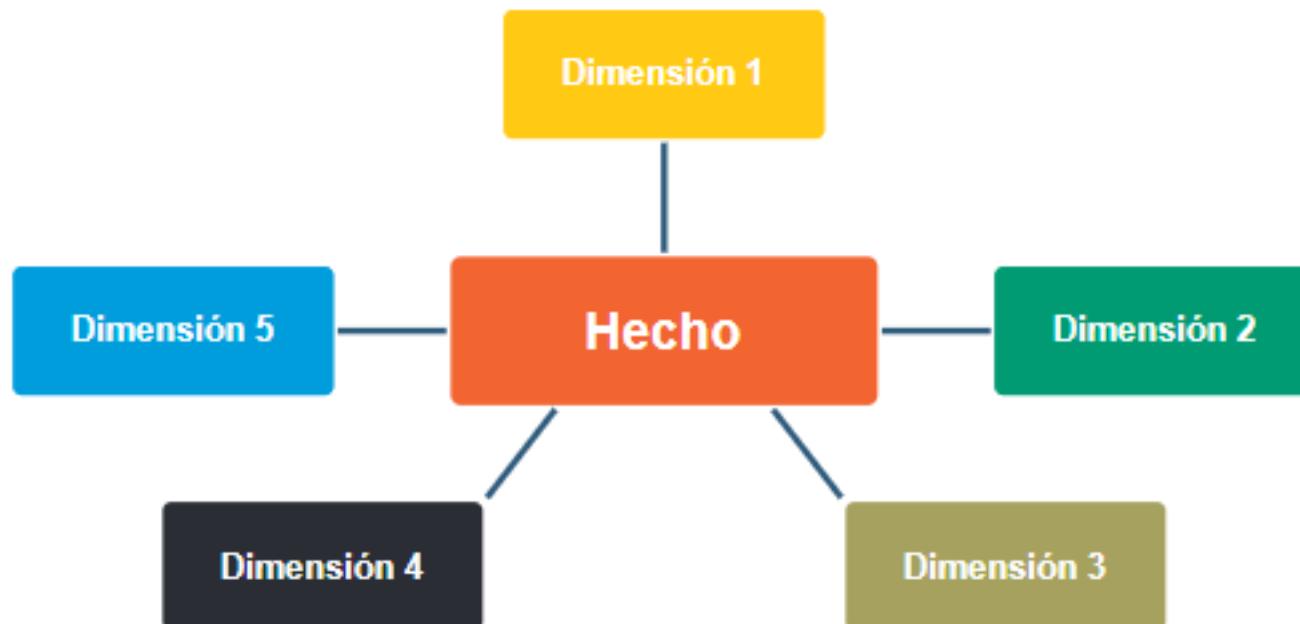
Breve explicación del proceso KDD

- Cuando se habla de extraer conocimientos de una ingente cantidad de datos, instantáneamente nos viene a la cabeza el término minería de datos.
- Aunque realmente la minería de datos es una fase más de un proceso de descubrimiento de conocimiento en base de datos conocido como **KDD** (Knowledge Discovery in Databases).
- El **proceso KDD** es un proceso utilizado para llevar a cabo la extracción automatizada de conocimiento partiendo de grandes volúmenes de datos, el cual es de naturaleza iterativa, por lo tanto, es aplicable tantas veces como sea necesario hasta obtener la información necesaria.
- Normalmente el proceso KDD tiene como motivación la detección de información que permita resolver los problemas o necesidades que surgen en las empresas y es a menudo solicitado por directivos y/o *stakeholders*.
- El conocimiento que se pretende extraer con el proceso KDD debe ser no trivial, implícito, previamente desconocido y potencialmente útil.



Modelo Estrella

- El Modelo Estrella es una técnica de modelado de datos que se utiliza para **diseñar y optimizar almacenes de datos y data marts**.
- Su nombre se debe a la forma que tiene el esquema lógico, que consta de una tabla central llamada tabla de hechos y varias tablas periféricas llamadas tablas de dimensiones.
- Estas tablas se relacionan entre sí mediante claves primarias y foráneas, formando una estructura en forma de estrella.



DESNORMALIZACIÓN BD

Normalización de bases de datos

- La normalización de la base de datos es el proceso de organizar los datos en tablas y columnas que siguen ciertas reglas o formularios normales.
- El objetivo de la normalización es eliminar la redundancia de datos, las anomalías y las inconsistencias.
- Por ejemplo, la normalización puede ayudarle a evitar insertar, actualizar o eliminar datos en varios lugares, lo que puede provocar errores y conflictos.
- La normalización también puede hacer que sus consultas sean más rápidas y sencillas, ya que puede unir tablas relacionadas con claves primarias y externas.

Desnormalización de bases de datos

- La desnormalización de la base de datos es lo opuesto a la normalización.
- Es el proceso de combinar o consolidar datos de varias tablas en menos tablas o más grandes.
- El objetivo es mejorar el rendimiento y la facilidad de uso de la base de datos. Al hacerlo, puede reducir el número de combinaciones, agregaciones y cálculos que deben realizar las consultas.
- La desnormalización también puede hacer que sus datos sean más accesibles y comprensibles, ya que puede almacenar datos en un formato que coincida con su lógica empresarial o sus necesidades de informes.



¿Cuándo desnormalizar su base de datos?

- Cuando deseé mejorar el rendimiento y la facilidad de uso de los datos.
- La desnormalización puede ayudarle a optimizar sus consultas, informes y análisis que dependen de sus datos.
- Por ejemplo, si tiene un gran volumen de operaciones de lectura que necesitan tener acceso a datos de varias tablas, puede beneficiarse de la desnormalización de la base de datos para reducir el número de combinaciones y mejorar la velocidad de consulta.

¿Cómo desnormalizar tu base de datos?

- Se debe invertir los pasos o niveles de normalización y combinar o combinar datos de varias tablas en menos tablas o más grandes.
- Los métodos más comunes son agregar columnas redundantes, crear tablas de resumen y usar estructuras anidadas o jerárquicas. Para aplicar estos métodos, debe analizar los patrones de uso de datos, los requisitos de consulta y las compensaciones de rendimiento.
- Por ejemplo, para agregar columnas redundantes, debe identificar las columnas a las que se accede con frecuencia o se unen desde otras tablas y copiarlas en la tabla principal. Para crear tablas de resumen, debe identificar los cálculos o agregaciones que se realizan con frecuencia en los datos y almacenar los resultados en una tabla independiente.



Modelo de datos dimensionales

- Un modelo de datos dimensional es una forma de organizar y estructurar datos en una base de datos o almacén de datos para facilitar a las empresas el análisis y la obtención de información a partir de sus datos.
- Son útiles cuando se trata de grandes volúmenes de datos y cuando los usuarios necesitan explorar datos desde diferentes ángulos o dimensiones.
- Diferentes aplicaciones requieren diferentes técnicas de modelado dimensional.
- Existen dos técnicas de modelado:
- 1. Los **modelos normalizados** de entidad-relación (modelos ER), están diseñados para eliminar la redundancia de datos, realizar rápidamente las operaciones de inserción, actualización y eliminación y obtener los datos dentro de una base de datos.
- 2. Los **modelos dimensionales** son estructuras desnormalizadas diseñadas para recuperar datos de un almacén de datos. Utilizan tablas de hechos y dimensiones para mantener un registro de datos históricos en almacenes de datos. Están optimizados para realizar las *select* y se utilizan en el marco de diseño básico para construir almacenes de datos altamente optimizados y funcionales.

Una tabla de hechos

- Es una tabla que almacena los datos numéricos o cuantitativos que se quieren analizar.
- Estos datos se llaman medidas y suelen ser valores agregados, como sumas, promedios, conteos, etc.
- Por ejemplo, el importe de una venta, la cantidad de un producto o el número de clientes.
- Es la tabla principal del modelo dimensional
- Contienen campos claves que se unen a las tablas de dimensión
- Contiene métricas o también llamadas medidas y es aquello que queremos medir o analizar.
- Generalmente son valores numéricos que se suelen agregar
- Evitan la redundancia de atributos por estas estos en las tablas de dimensiones
- normalmente tienen muchos (millones) registros
- por ejemplo: ventas, compras, movimientos de contabilidad



Una tabla de dimensiones

- es una tabla que almacena los datos cualitativos o descriptivos que se quieren usar para filtrar, agrupar o clasificar las medidas.
- Estos datos se llaman atributos y suelen ser valores categóricos, como nombres, textos, fechas, etc. Por ejemplo, el nombre de un producto, la categoría de un cliente o el mes de una venta.
- La relación entre una tabla de hechos y una tabla de dimensiones se establece mediante claves.
- Son tablas simples desnormalizadas
- se unen a las tablas de hechos a través de un campo clave
- los atributos de la tabla de dimensión ofrecen información característica de las tablas de hechos
- no hay límite de tablas de dimensión
- las dimensiones pueden contener una o varias relaciones jerárquicas
- normalmente tiene pocos (miles) registros
- por ejemplo: clientes, productos, almacenes, proveedores, calendario..



Métricas

- Las métricas son los indicadores de negocio de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio.
- Por ejemplo, en una venta tenemos el importe de la misma y la cantidad vendida.
- Existen métricas derivadas, como el precio unitario, que se obtiene al dividir el importe total por las unidades vendidas.



Ventajas del Modelo Estrella

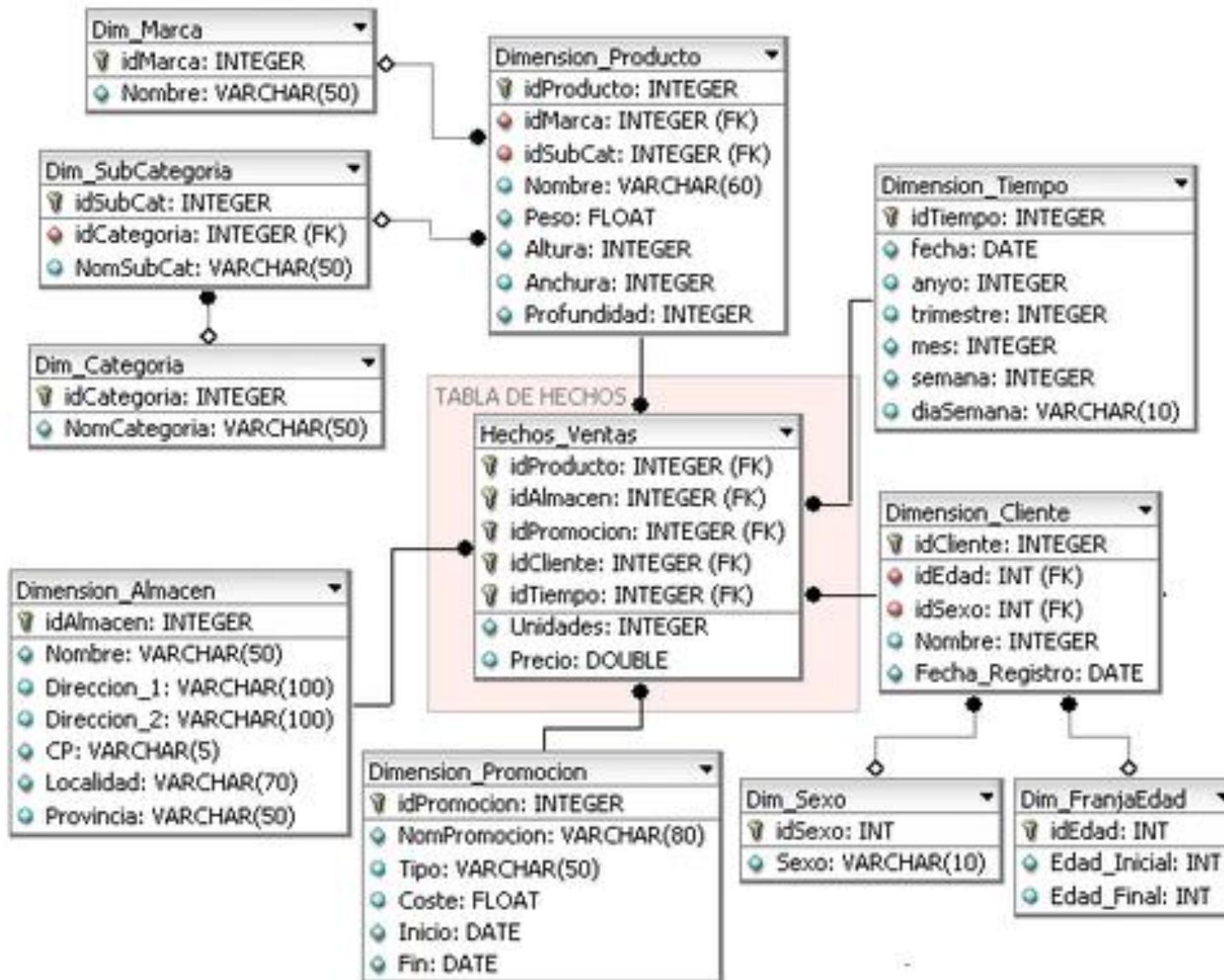
- **Facilita la comprensión del negocio y los requisitos de los usuarios,**
 - Las tablas de dimensiones contienen atributos descriptivos que definen las entidades del dominio (productos, clientes, tiempos, etc.).
 - Estos atributos permiten expresar las consultas en un lenguaje natural y cercano al usuario final.
 - Las tablas de dimensiones pueden contener jerarquías y niveles que reflejen la organización lógica de los datos.
- **Mejora el rendimiento de las consultas,**
 - Reduce el número de uniones necesarias para obtener la información deseada.
 - Las tablas de hechos suelen estar indexadas por las claves de dimensiones, lo que acelera la recuperación de los datos.
 - Al tener un esquema simple y desnormalizado, se evita la redundancia y la inconsistencia de los datos.
 - También se pueden aplicar técnicas como el particionamiento o la compresión para optimizar el almacenamiento y la consulta de los datos.
 -

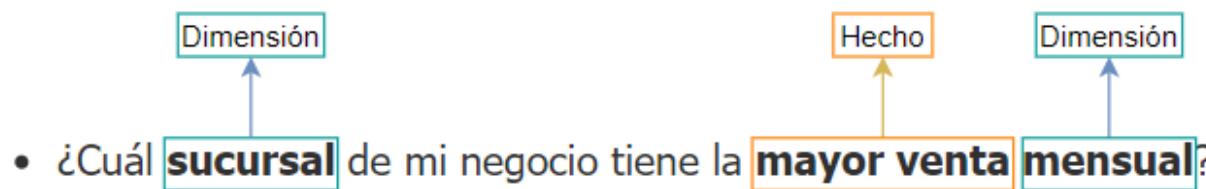
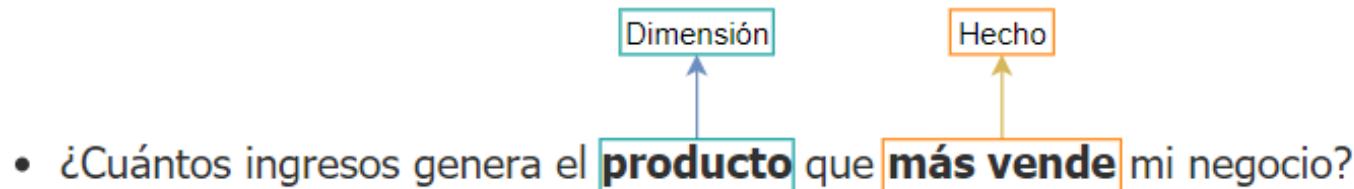


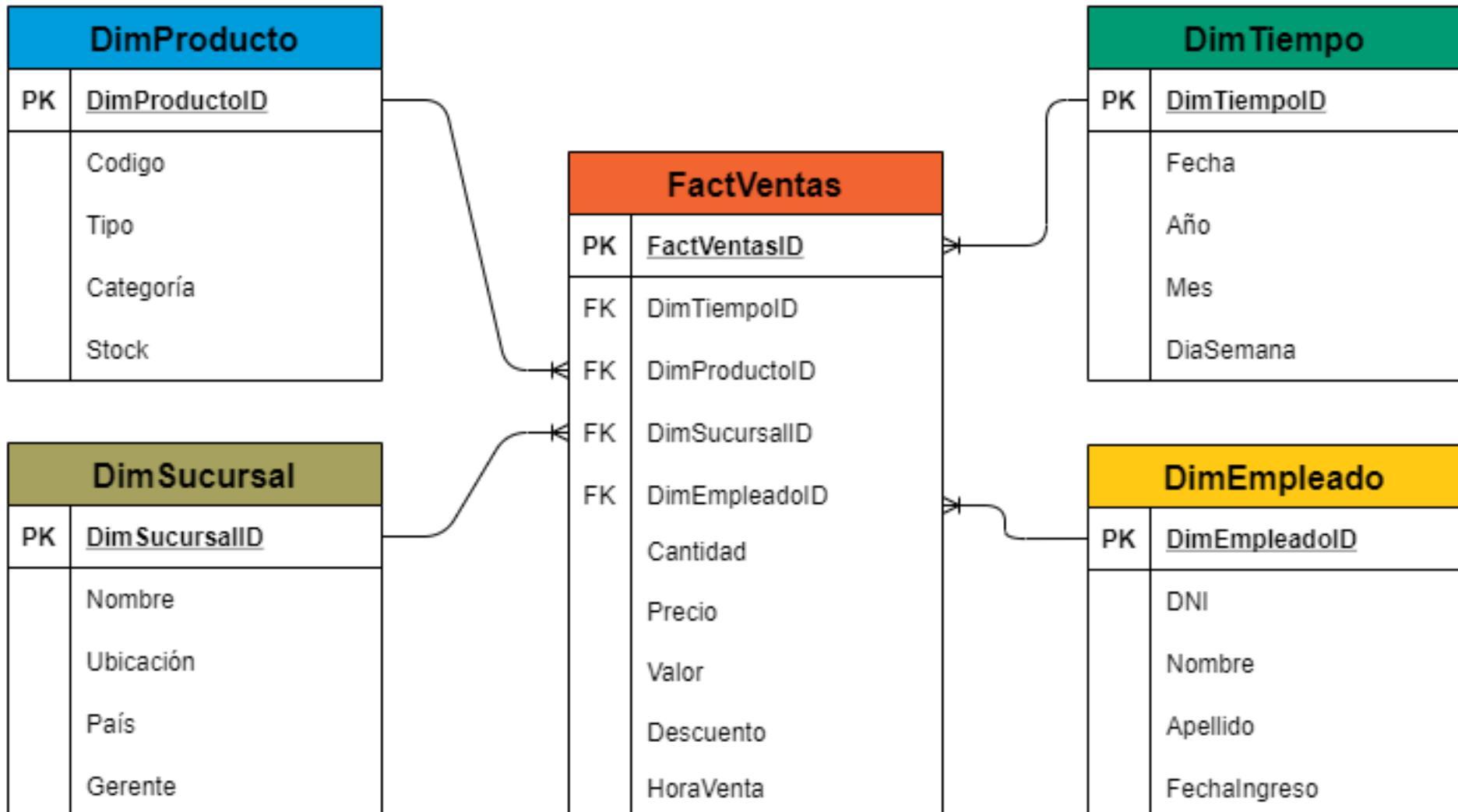
Ventajas del Modelo Estrella

- **Permite la flexibilidad y la escalabilidad del diseño,**
 - Se puede agregar o modificar dimensiones y medidas sin afectar a las demás tablas
 - También se pueden crear vistas o agregaciones para simplificar o resumir los datos.
 - El Modelo Estrella facilita el crecimiento incremental del almacén de datos, ya que se pueden añadir nuevas tablas de hechos o dimensiones según las necesidades del negocio.
 - Pueden integrar diferentes fuentes de datos con distintos niveles de detalle o granularidad.
- **Favorece la implementación de herramientas de Business Intelligence (BI),**
 - Power BI, se basa en el Modelo Estrella para crear modelos de datos interactivos y visuales.
 - Estas herramientas permiten explorar los datos desde diferentes perspectivas y niveles de detalle, aplicando filtros, agrupaciones y cálculos sobre las medidas y los atributos.
 - Permiten crear informes y paneles personalizados que faciliten la toma de decisiones basada en los datos.



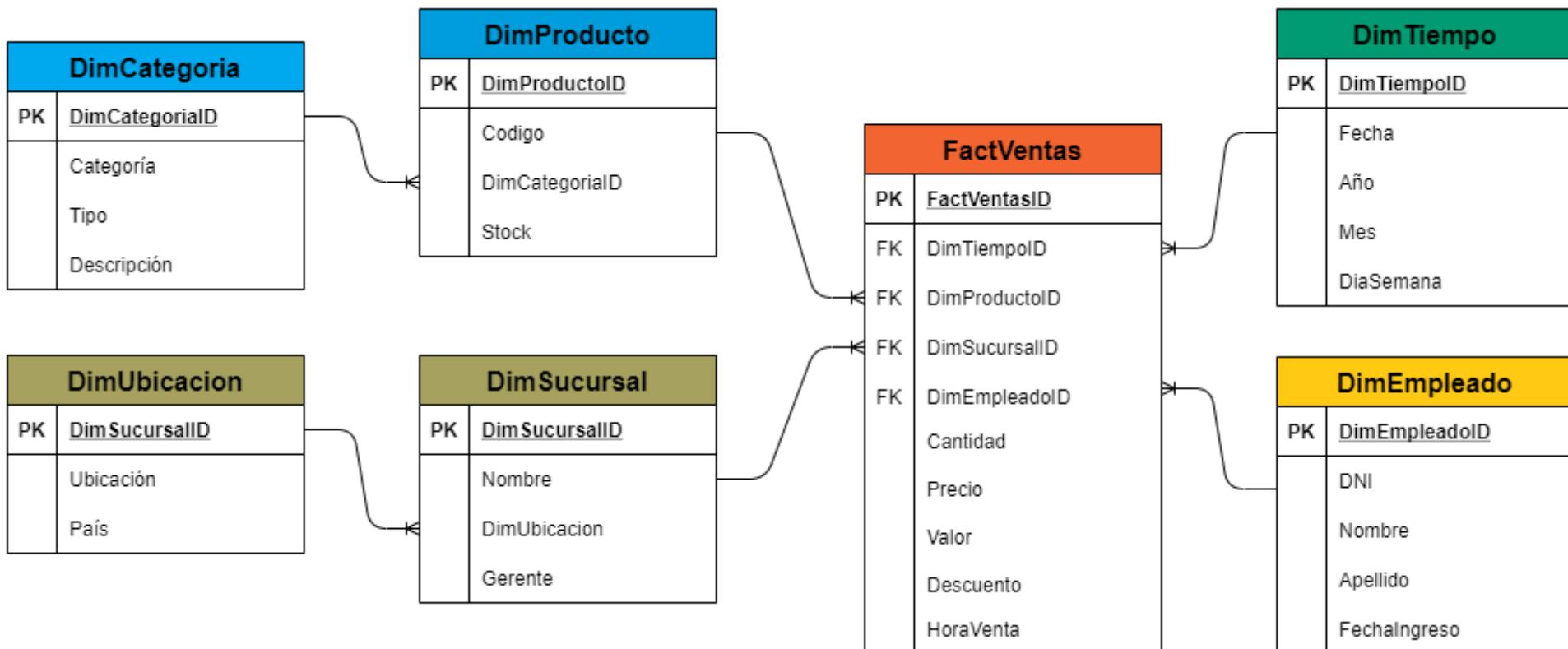






Modelo de datos en copo de nieve

- Este tipo de modelo es más complejo que un modelo en estrella, pero el concepto de creación parte de la misma base de análisis.
- Se presenta cuando las **dimensiones tienen más de una tabla** para conformarla, y el objetivo es básicamente **disminuir el almacenamiento** con la normalización de estas tablas.
- Bajo este modelo, **la tabla de hecho no está relacionada directamente** a todas las tablas que componen el modelo de datos.

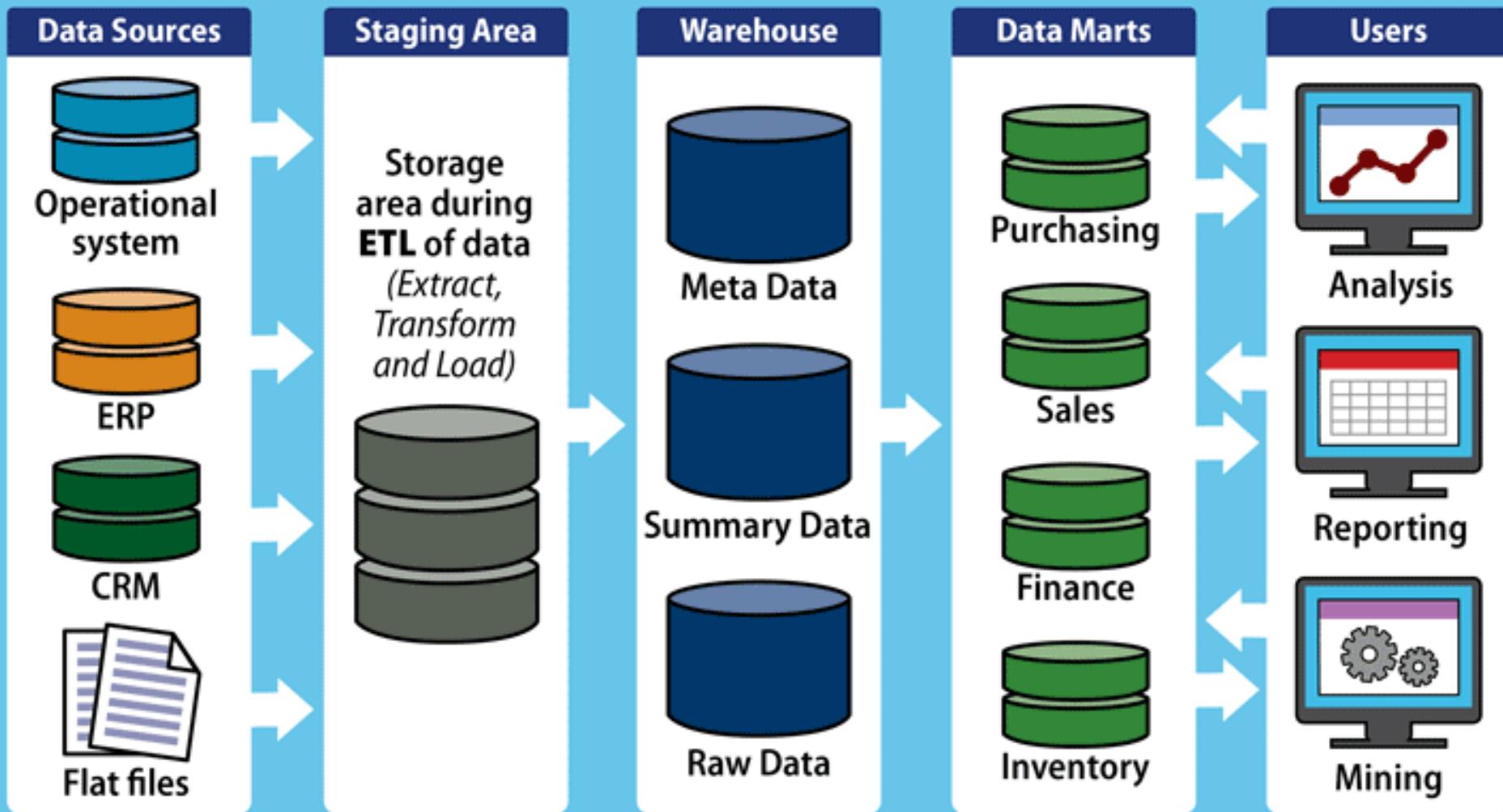


CONCEPTOS DE UNA DATAMARK Y DATAWAREHOUSE

DATAWAREHOUSE

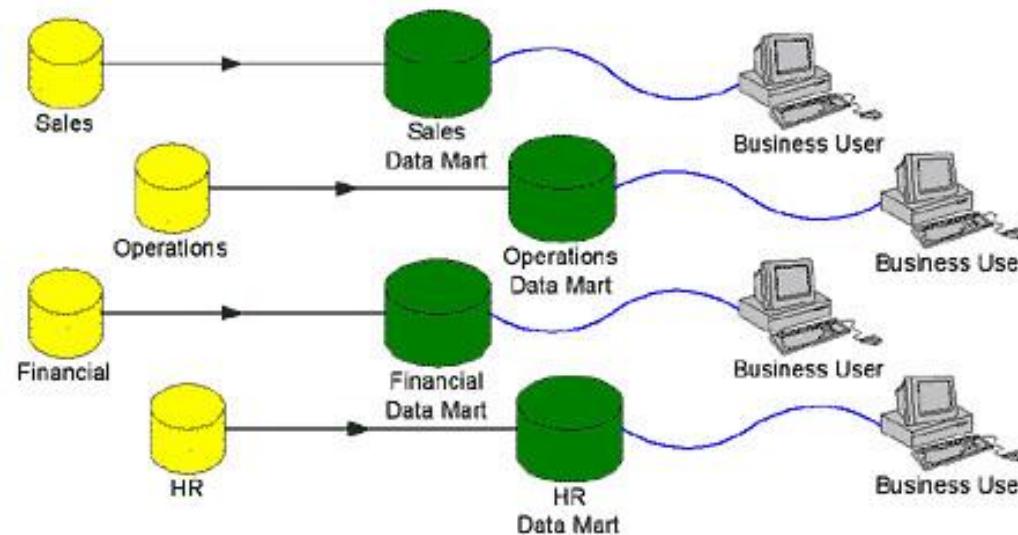
- En esencia, se trata de una base de datos relacional que integra datos de múltiples fuentes dentro de una empresa.
- La creación de un data warehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.
- La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales... etc).
- Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales)
- Para la creación de un Data Warehouse tradicional es necesario la ejecución de procesos ETL (Extracción, Transformación y Carga) a partir de los sistemas operaciones de una compañía

Data Warehouse Architecture



DATA MART:

- Es un almacén de datos orientado a un área específica, como por ejemplo, Ventas, Recursos Humanos u otros sectores en una organización. Por ello, también se le conoce como una base de información departamental.
- Este almacén permite que una empresa pueda acceder a datos claves de un área de forma sencilla, además de realizar diversas funciones, tales como:
 - La organización de información para su posterior análisis.
 - La elaboración de indicadores clave de rendimiento (KPI).
 - La creación de informes para un aprendizaje automático.
 - La evaluación de datos sobre el cumplimiento de objetivos de un sector.



Comparación con un Data Warehouse:

- Es inevitable la comparación de data marts con los data warehouse y al final se acaba diciendo que son como estos, pero *en pequeño*, y en cierto modo esto es así, pero esta idea suele hacer caer en los siguientes errores sobre la implementación y funcionamiento de los data marts:
 - **Son más simples de implementar que un Data Warehouse:** *Falso*, la implementación es muy similar, ya que debe proporcionar las mismas funcionalidades.
 - **Son pequeños conjuntos de datos y, en consecuencia, tienen menor necesidad de recursos:** *Falso*, una aplicación corriendo sobre un data mart necesita los mismos recursos que si corriera sobre un data warehouse.
 - **Las consultas son más rápidas, dado el menor volumen de datos:** *Falso*, el menor volumen de datos se debe a que no se tienen todos los datos de toda la empresa, pero sí se tienen todos los datos de un determinado sector de la empresa, por lo que una consulta sobre dicho sector tarda lo mismo si se hace sobre el data mart que si se hace sobre el data warehouse.
 - **En algunos casos añade tiempo al proceso de actualización:** *Falso*, actualizar el data mart desde el data warehouse cuesta menos que actualizar el data warehouse desde sus fuentes de datos primarias, donde es necesario realizar operaciones de transformación



METODOLOGÍAS DE DISEÑO DE UN DATAMARK DATAWAREHOUSE

Kimball Model:

- Sigue un enfoque de abajo hacia arriba para almacenamiento de datos (DW) diseño de arquitectura en el que los mercados de datos se forman primero en función de los requisitos comerciales.
- Es una metodología empleada para la construcción de un almacén de datos que no es mas que, una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.
- Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos:
 - Centrarse en el negocio
 - Construir una infraestructura de información adecuada
 - Realizar entregas en incrementos significativos (este se parece a las metodologías ágiles de construcción de software)
 - Ofrecer la solución completa (En este se punto proporcionan todos los elementos necesarios para entregar valor a los usuarios de negocios).
- La construcción de una solución de Datawarehouse/Business Intelligence es sumamente compleja, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad



Ventajas del método Kimball

- Es rápido de construir ya que no implica normalización, lo que significa una ejecución rápida de la fase inicial del almacenamiento de datos de procesos.
- La mayoría de los operadores de datos pueden comprenderlo fácilmente debido a su estructura desnormalizada.
- La huella del sistema de almacenamiento de datos es trivial porque se centra en áreas y procesos comerciales individuales en lugar de en toda la empresa.
- Permite la recuperación rápida de datos del almacén de datos, ya que los datos se segregan en tablas de hechos y dimensiones.
- Un equipo pequeño de diseñadores es suficiente para la gestión del DWH porque los sistemas de origen de datos son estables.
- Permite que las herramientas profundicen en varios esquemas en estrella y genere información confiable.

Desventajas

- Los datos no están completamente integrados antes de la presentación de informes; la idea de una "fuente única de verdad se pierde".
- Pueden ocurrir irregularidades cuando los datos se actualizan. Esto se debe a que, en el almacén de datos de técnicas de desnormalización, se agregan datos redundantes.
- La adición de columnas puede expandir las dimensiones de la tabla de hechos, afectando su rendimiento.
- El modelo DWH se vuelve difícil de modificar con cualquier cambio.



Inmon Model:

- Bill Inmon, el "padre del Data Warehousing", define un Data Warehouse (DW) como "una colección de datos orientada a un tema, integrada, variable en el tiempo y no volátil, para apoyar el proceso de toma de decisiones de la dirección".
- El diseño Inmon utiliza la forma normalizada para construir la estructura de la entidad, evitando la redundancia de datos tanto como sea posible.
- El modelo da como resultado una identificación clara de los requisitos comerciales y la prevención de irregularidades en la actualización de datos.
- La ventaja de este enfoque de arriba hacia abajo en el diseño de bases de datos es que es robusto a los cambios comerciales y contiene una perspectiva dimensional de los datos en el mercado de datos.



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

- Este modelo de Inmon crea una única fuente de verdad para todo el negocio. La carga de datos se vuelve menos compleja debido a la estructura normalizada del modelo.
- Utilizar esta disposición para realizar consultas es un desafío, ya que incluye numerosas tablas y enlaces.
- Esta metodología de Inmon propone la construcción de data marts por separado para cada división, como finanzas, marketing, ventas, etc.



- Todos los datos que ingresan al data warehouse están integrados.
- El almacén de datos actúa como una única fuente de datos para varios data marts a fin de garantizar la integridad y coherencia en toda la empresa.

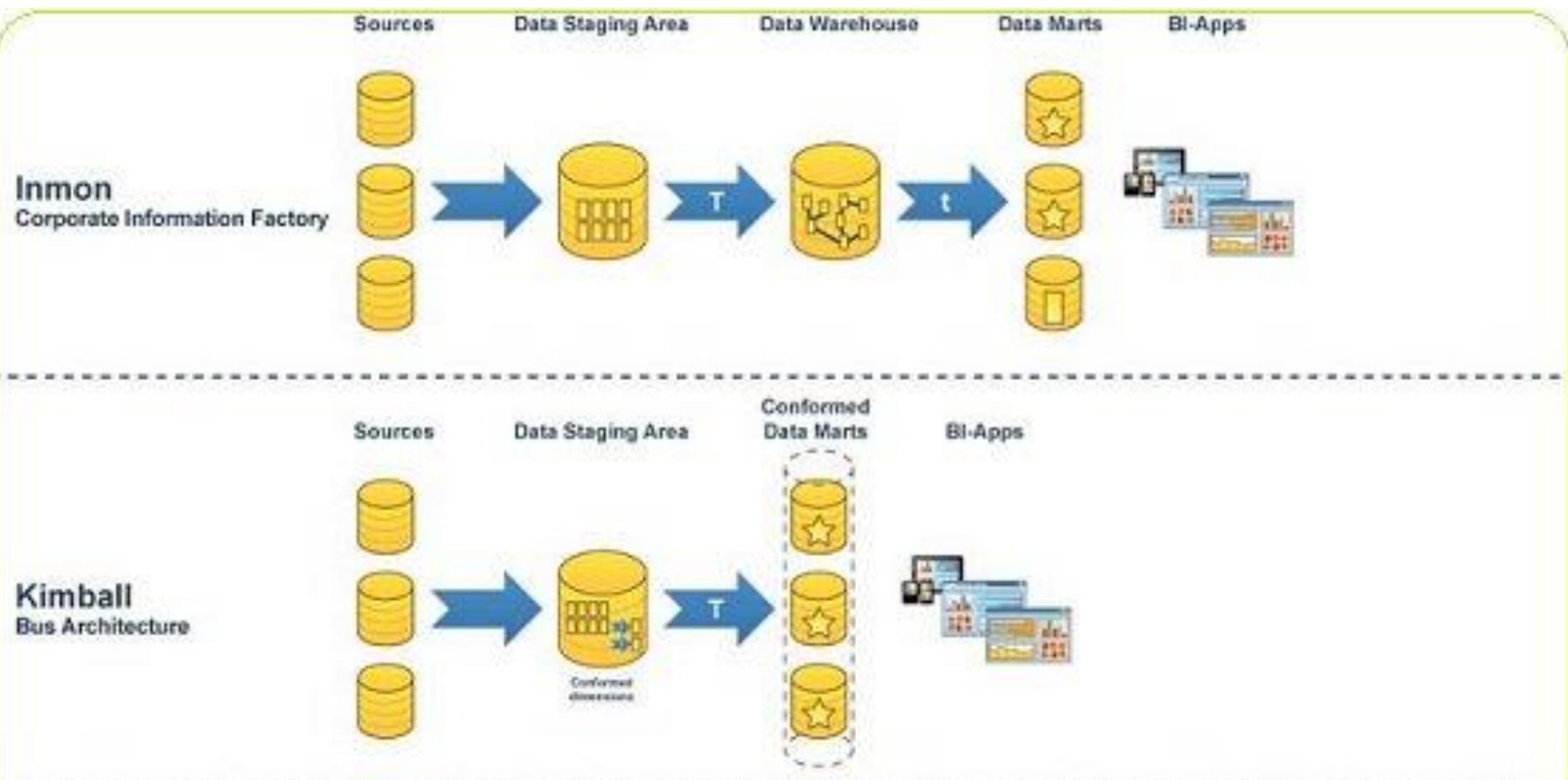


Ventajas del método Inmon

- El almacén de datos actúa como una fuente de verdad unificada para todo el negocio, donde todos los datos están integrados.
- Este enfoque tiene una redundancia de datos muy baja. Hay menos posibilidad de irregularidades en la actualización de datos, lo que hace que el proceso de almacenamiento de datos ETL sea más sencillo y menos susceptible a fallas.
- Simplifica los procesos comerciales, ya que el modelo lógico representa objetos comerciales detallados.
- Este enfoque ofrece una mayor flexibilidad, ya que es más fácil actualizar el almacén de datos en caso de que haya algún cambio en los requisitos.
- Puede manejar diversos requisitos de informes en toda la empresa.

Desventajas

- La complejidad aumenta a medida que se agregan varias tablas al modelo de datos con el tiempo.
- La configuración preliminar y la entrega requieren mucho tiempo.
- Se requiere una operación ETL adicional, ya que los data marts se crean después de la creación del almacén de datos.
- Este enfoque requiere que los expertos administren un almacén de datos de manera efectiva.



Data Lake:

- Un Data Lake es un sistema o repositorio de datos almacenados en su formato natural/bruto, normalmente blobs de objetos o archivos.
- Suele ser un almacén único de datos que incluye copias en bruto de los datos del sistema de origen, datos de sensores, datos sociales, etc., y datos transformados que se utilizan para tareas como la elaboración de informes, la visualización, el análisis avanzado y el aprendizaje automático.
- Un Data Lake puede incluir datos estructurados de bases de datos relacionales (filas y columnas), datos semiestructurados (CSV, registros, XML, JSON), datos no estructurados (correos electrónicos, documentos, PDF) y datos binarios (imágenes, audio, vídeo).
- Un Data Lake puede ser "on premise" (dentro de los centros de datos de una organización) o "en la nube"



Data Mining

- Es el proceso de extraer información útil de una acumulación de datos, a menudo de un almacén de datos o de la recopilación de conjuntos de datos vinculados.
- Las herramientas de minería de datos incluyen potentes capacidades estadísticas, matemáticas y analíticas cuyo propósito principal es examinar grandes conjuntos de datos para identificar tendencias, patrones y relaciones para respaldar la toma de decisiones y la planificación fundamentada



¿Por qué utilizar Data Mining?

- El beneficio de la minería de datos es su poder para identificar patrones y relaciones en grandes volúmenes de datos de múltiples fuentes.
- La minería de datos ofrece las herramientas para explotar Big Data y convertirlo en inteligencia accionable. Además, puede actuar como un mecanismo para “pensar fuera de la caja”.
- El proceso de minería de datos puede detectar relaciones y patrones sorprendentes en bits de información aparentemente no relacionados.
- Debido a que la información tiende a ser compartimentada, históricamente ha sido difícil o imposible de analizar en su conjunto.
- Los reportes señalan “lo que pasó” pero hace poco para descubrir el “por qué sucedió de esta manera”. La minería de datos puede llenar esta brecha.
- Data Mining puede buscar correlaciones con factores externos; mientras que la correlación no siempre indica causalidad, estas tendencias pueden ser indicadores valiosos para guiar las decisiones de producto, canal y producción.



Asociación

- La técnica de asociación implica buscar ciertas ocurrencias con atributos conectados.
- La idea es buscar variables vinculadas en función de atributos o eventos específicos.
- Las reglas de asociación pueden ser particularmente útiles para estudiar el comportamiento del consumidor.
- Por ejemplo, una tienda en línea podría saber que los clientes que compran un determinado producto probablemente comprarán un artículo complementario.

Seguimiento de patrones

- Implica reconocer y monitorear tendencias en conjuntos de datos para realizar análisis inteligentes con respecto a los resultados comerciales.
- Por ejemplo, el patrón en los datos de ventas puede mostrar que un determinado producto es más popular entre grupos demográficos específicos o una disminución en el volumen total de ventas después de la temporada navideña. Luego, la empresa puede usar esta información para dirigirse a mercados específicos y optimizar la cadena de suministro.

técnicas de minería de datos

Clasificación

- Es el proceso de dividir grandes conjuntos de datos en categorías objetivo.
- Esta categorización también está determinada por el marco de datos, por ejemplo, base de datos relacional, base de datos orientada a objetos, etc.
- Suponga que su empresa quiere pronosticar el cambio en los ingresos de los clientes a los que se les otorga una membresía de lealtad. Puede crear una categoría que contenga los datos demográficos de los clientes con una membresía de lealtad para diseñar un modelo de clasificación binaria para predecir un aumento o disminución en el gasto.

Detección de valores atípicos

- Hay instancias en las que el patrón de datos no proporciona una comprensión clara de los datos. En tales situaciones, la técnica de detección de valores atípicos resulta útil.
- Implica identificar anomalías o "valores atípicos" en su conjunto de datos para comprender las causas específicas o derivar predicciones más precisas.

Técnicas de minería de datos

Clustering

- Al igual que la clasificación, consiste en agrupar datos en función de las similitudes.
- Ayuda en el descubrimiento de conocimientos, la detección de anomalías y la obtención de información sobre la estructura interna de los datos.
- Por ejemplo, puede agrupar audiencias de diferentes regiones en paquetes según su grupo de edad, sexo e ingresos disponibles, de modo que pueda adaptar su campaña de marketing para maximizar su alcance.
- Los resultados del análisis de datos generalmente se muestran mediante gráficos para ayudar a los usuarios a visualizar la distribución de datos e identificar tendencias en sus conjuntos de datos.

Patrones Secuenciales

- Se enfoca en descubrir patrones o una serie de eventos que tienen lugar en una secuencia.
- Por ejemplo, puede ayudar a las empresas a recomendar artículos relevantes a los clientes para maximizar las ventas.
- Ejemplo: si una tendencia secuencial en una tienda donde es probable que los clientes que compran un iPhone compren una MacBook dentro de seis meses. Se puede utilizar esto para crear campañas dirigidas a los compradores de iPhone.

Técnicas de minería de datos

Árbol de decisión

- Es una técnica de minería de datos en el aprendizaje automático (ML) que se centra en las relaciones de modelado de entrada y salida mediante reglas si/entonces.
- Puede aprender cómo las entradas de datos influyen en las salidas. Los árboles suelen estar diseñados en una estructura similar a un diagrama de flujo de arriba hacia abajo.
- Un modelo de análisis predictivo con varios modelos de árboles de decisión facilita análisis de datos más complejos.
- Los árboles de decisión se utilizan para modelos de clasificación y regresión.

Análisis De Regresión

- Es una de las técnicas de minería en el aprendizaje automático que utiliza la relación lineal entre variables.
- Le ayuda a predecir el valor futuro de las variables. La técnica tiene numerosas aplicaciones en pronósticos financieros, planificación de recursos, toma de decisiones estratégicas y más.
- Por ejemplo, se puede utilizar para comprender la correlación entre la educación, los ingresos y los hábitos de gasto. La complejidad de la predicción aumenta a medida que agrega más variables.

Procesamiento de memoria a largo plazo

- Se utiliza para analizar datos durante períodos prolongados. Le permite identificar patrones de datos basados en el tiempo, como datos climáticos, de manera más efectiva.
- Su objetivo es escalar los datos en la memoria del sistema y utilizar información adicional en el análisis.
- Por ejemplo, puede diseñar un modelo predictivo para identificar transacciones fraudulentas mediante la asignación de probabilidades. Puede usar este modelo para transacciones existentes y luego, después de un tiempo, actualizar el modelo con los datos derivados de nuevas transacciones, lo que resulta en una mejor toma de decisiones.

Redes Neuronales

- Es usado en los modelos de aprendizaje automático utilizados con Inteligencia Artificial (IA). Al igual que las neuronas en el cerebro, busca identificar relaciones en los datos.
- Tienen diferentes capas que trabajan juntas para producir resultados de análisis de datos con gran precisión.
- Buscan patrones en una gran cantidad de datos. Si bien pueden ser muy complejos, el resultado generado puede proporcionar información valiosa para las organizaciones.



HERRAMIENTAS DE VISUALIZACIÓN DE DATOS.

Visualización de datos

- Es el proceso de expresar visualmente datos e información utilizando gráficos, mapas y otros elementos visuales.
- Visualiza conjuntos de datos complejos, patrones y relaciones para facilitar la comprensión, el análisis y la comunicación.
- El objetivo es transformar datos en bruto en conocimientos útiles y accionables.

Herramientas de visualización de datos

- Son plataformas de software o aplicaciones que permiten a los usuarios generar, diseñar y mostrar representaciones visuales de datos.
- Estos sistemas incluyen herramientas y funcionalidades para transformar datos en bruto en representaciones relevantes e interactivas.
- Ofrecen una variedad de formatos gráficos y otros elementos visuales para representar datos de manera visual.
- Proporcionan integración de datos, diseño de visualizaciones, interactividad y exploración, creación de paneles y generación de informes, colaboración y uso compartido, conectividad, rendimiento y escalabilidad, seguridad y gobernanza de datos, exploración, análisis y presentación de datos.
- Estas herramientas ayudan a los usuarios a comprender los datos, tomar decisiones inteligentes y comunicar eficazmente sus hallazgos.



FUNDAMENTOS

Importancia de las herramientas de visualización de datos

- **Mejorar la comprensión de los datos:** Simplifican datos complejos, lo que facilita a las personas identificar patrones, tendencias y relaciones. Las visualizaciones simplifican el análisis de datos.
- **Mejorar el análisis de datos:** Los usuarios pueden interactuar con las visualizaciones, filtrar y profundizar en puntos de datos específicos y analizar datos en tiempo real. Facilitan la exploración de datos y la identificación de valores atípicos, correlaciones y otros conocimientos útiles.
- **Mejorar la comunicación y la colaboración:** Facilitan la compartición de hallazgos de datos entre equipos y departamentos. Permiten a múltiples usuarios trabajar juntos en visualizaciones, compartir conocimientos y fomentar una cultura de toma de decisiones basada en datos.
- **Apojar conocimientos accionables:** Ayudan a identificar conocimientos accionables. Estas plataformas permiten a los usuarios supervisar el rendimiento, seguir objetivos y tomar medidas inmediatas basadas en datos en tiempo real mediante funciones interactivas y actualizaciones de datos en tiempo real. Las visualizaciones resaltan áreas que necesitan mejoras, lo que ayuda a las organizaciones a mejorar su estrategia y acciones.



Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms



© Gartner, Inc



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

ANÁLISIS DE HERRAMIENTAS

Tableau

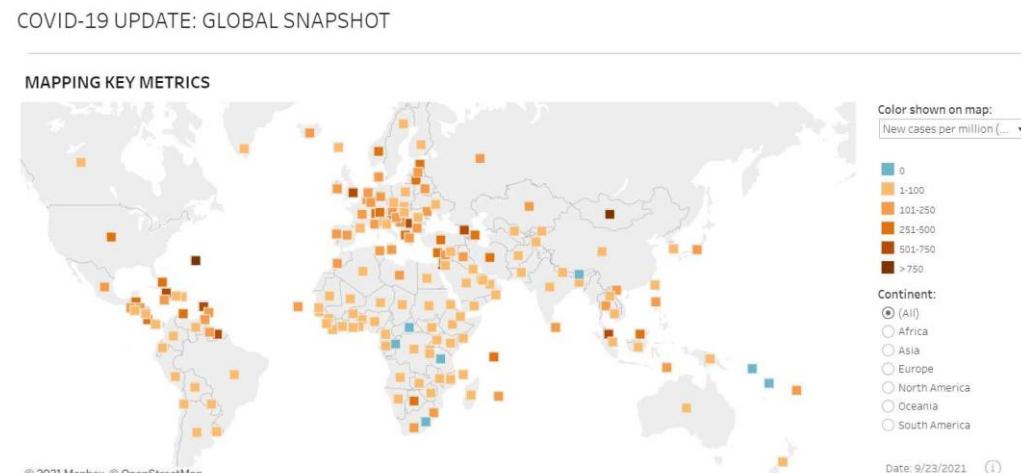
- Conocido por su interfaz fácil de usar y su amplia gama de opciones de visualización. Ofrece un entorno de arrastrar y soltar fácil de usar que permite a los usuarios crear visualizaciones interactivas
- Cuenta con una amplia selección de estilos de gráficos y opciones de personalización para satisfacer diferentes objetivos analíticos.

Desventajas:

- Costoso, especialmente para características a nivel empresarial.
- Capacidades avanzadas con una curva de aprendizaje más pronunciada.

Ventajas:

- Interfaz sencilla de arrastrar y soltar.
- Amplio apoyo y recursos de la comunidad.
- Variedad de opciones de personalización.



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Power BI

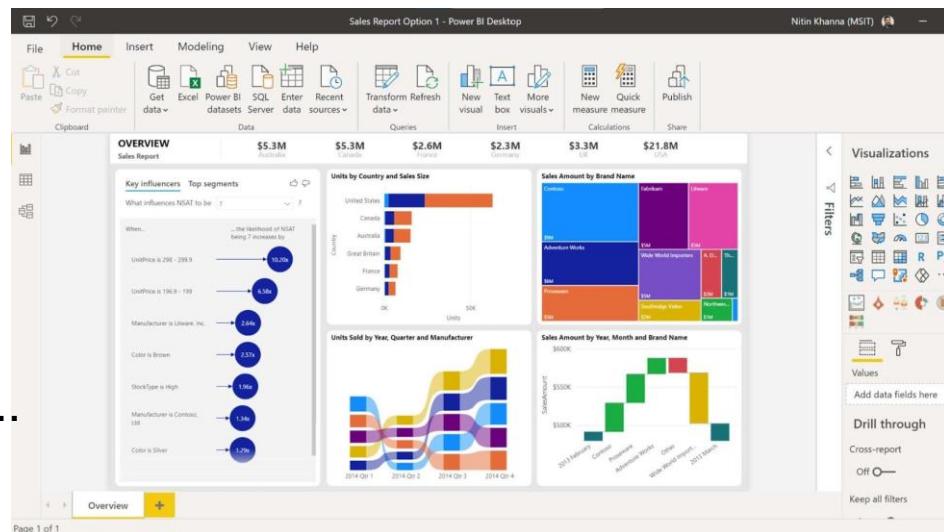
- Power BI, una herramienta de Microsoft, se integra con otros servicios de Microsoft y cuenta con una interfaz fácil de usar.
- Ofrece funciones de arrastrar y soltar que simplifican la creación de visualizaciones atractivas.
- Incluye características avanzadas de creación de paneles e informes que permiten a los usuarios supervisar y analizar datos de manera efectiva.

Ventajas:

- Integración con el ecosistema de Microsoft.
- Paneles e informes con muchas características.

Desventajas:

- Relativamente pocos tipos de gráficos..
- Se necesita conocimiento de Power Query y DAX para funcionalidad avanzada.



ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Qlik Sense

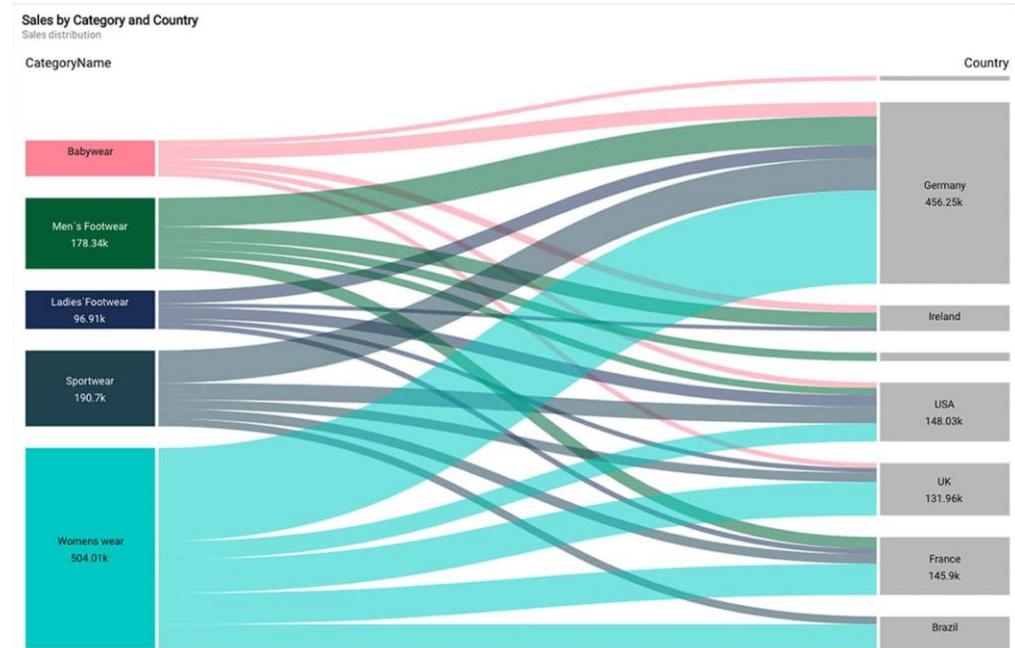
- Es una herramienta de visualización de datos diseñada para grandes empresas y personas que quieren utilizar la analítica aumentada para analizar datos.
- Qlik Sense es el sucesor de «QlikView», una herramienta de análisis visual similar (pero más pequeña).

Pros

- Funciona online y offline en dispositivos móviles
- Perfecto para equipos
- Escalable para grandes empresas

Contras

- Más adecuado para personas con experiencia en análisis de



Google Data Studio

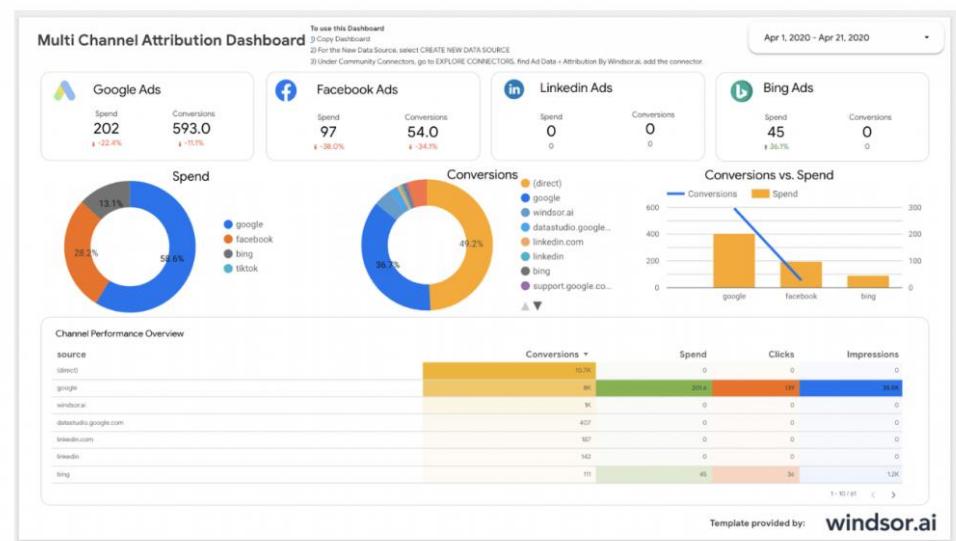
- Es una plataforma gratuita que ofrece una variedad de conexiones de datos, así como opciones de uso compartido simples.
- Es una interfaz fácil de usar que permite a los usuarios conectarse a diversas fuentes de datos.
- Facilita la creación de paneles interactivos y la compartición de resultados con otros.

Ventajas:

- Es gratuito
- La interfaz es fácil de usar.
- La integración con otros productos de Google es fluida.

Desventajas:

- Personalización comparativamente limitada.
- Falta de análisis sofisticado.



IBM Cognos Analytics

- Es un conjunto completo de herramientas para visualizar e informar sobre datos a nivel empresarial.
- Dispone de un completo conjunto de herramientas para visualizar e informar sobre datos a nivel empresarial. Se integra bien con el entorno de datos de IBM y cuenta con sólidas funciones de seguridad y gestión.

Ventajas:

- Funciones de nivel empresarial.
- Fuerte integración con la comunidad de servicios de datos de IBM.
- Potentes herramientas de protección y gestión.

Desventajas:

- La elaboración de informes complejos requiere una curva de aprendizaje pronunciada.
- Los despliegues a gran escala requieren recursos.

