[Examen 2 - Base de datos 1](#_Toc182402120)

[Data Warehouse 1](#_Toc182402121)

[OLTP (Online Transaction Processing) 2](#_Toc182402122)

[OLAP (Online Analytical Processing) 2](#_Toc182402123)

[Data Marts: 4](#_Toc182402124)

[Inmon y Kimball 4](#_Toc182402125)

[Diseño de Data Warehouse 8](#_Toc182402126)

[Modelado 8](#_Toc182402127)

[Dimensiones/ atributos: 8](#_Toc182402128)

[Hechos/ métricas: 8](#_Toc182402129)

[Esquema Estrella 9](#_Toc182402130)

[Esquema Snowflake (Copo de nieve) 9](#_Toc182402131)

[ETL 10](#_Toc182402132)

[Carga inicial: 10](#_Toc182402133)

[Refresh 11](#_Toc182402134)

[Problemas a tratar en proceso ETL 11](#_Toc182402135)

[Beneficios 11](#_Toc182402136)

[Cloud Computing 11](#_Toc182402137)

[Tipo de nubes 12](#_Toc182402138)

[Arquitectura de la nube 14](#_Toc182402139)

[Preguntas de examen 14](#_Toc182402140)

# Examen 2 - Base de datos

## Data Warehouse

Es una BD centralizado que organiza y gestiona grandes volúmenes de diversas fuentes para análisis. Estructura datos provenientes de diferentes sistemas operativos (Como venta, inventario, marketing, etc.), para ser consultado de forma eficiente

**Características**

* **Integración de datos**: Agrupa datos de distintas fuentes, estandarizando y eliminando duplicados
* **Orientado a temas**: Los datos se organizan en torno a temáticas (Venta, cliente, producto, etc.), facilitando el análisis para cada área de la empresa
* **No volátil**: No suelen cambiar después de ser cargados. Se agregan datos nuevos, pero los existentes no se modifican en general
* **Variantes de tiempo**: Datos a largo plazo para analizarse las tendencias en el tiempo, para las decisiones estratégicas
* **Versionado**: Se guardan versiones nuevas para cada cambio del sujeto de análisis

### OLTP (Online Transaction Processing)

Es un procesamiento en transacciones en línea**. Son para aplicaciones operacionales** (Ventas, Bancos, RRHH)

* **Operaciones** de **actualización** (**INSERT, UPDATE, DELETE**)
* **Alta** **ocurrencia** y **velocidad**: Soporta gran número de transacciones seguidas y de manera eficiente. Son específicos en las tareas que desean ejecutar
* **Datos actualizados:** Los datos son recientes y cambian con frecuencia para ver la actividad en tiempo real
* **Estructura** **normalizada**: Minimiza redundancia con respuesta rápida

***Ejemplo de OLTP:***

*Un sistema de punto de venta en un supermercado donde se registran las compras y el inventario se actualiza en tiempo real.*

### OLAP (Online Analytical Processing)

Es un procesamiento analítico en línea. **Son para análisis y somas de decisiones**. Este permite hacer **consultas complejas sobre grandes volúmenes** de datos históricos en el data warehouse (Clientes, Ventas, Stock, Costos)

* **Orientado a análisis**: Analiza grupos grandes de datos para tener información útil
* **Datos históricos**: Están organizados de manera que incluyan información para el análisis de tendencias y patrones. Bajo volumen transaccional
* **Consultas complejas**
* **Desnormalización**: Esta desnormalizada, para acceder rápido a grandes conjuntos de datos
* **Velocidad dependiente**: Depende de la cantidad de datos involucrados su velocidad

Tabla

Descripción generada automáticamente

Captura de pantalla de un celular

Descripción generada automáticamente

Arquitectura

Gráfico

Descripción generada automáticamente

**S0/S1**: **Área de copia de datos transaccionales**. Esta la latencia, no hace transformaciones ni modelado lógico

**S2**: **Operational data store**. Tareas de reporting de los sistemas operacionales. No perjudica el performance y puede ser omitido

**S3**: **DW**. Se hace la normalización elegida en el modelado lógico y físico, integrando datos de distintas fuentes, es como el repositorio central de la empresa

**S4/S5**: **Reporting y Exploración**: n este nivel, se crean vistas y reportes personalizados para diferentes departamentos. Los datos se agregan y resumen según la granularidad requerida para análisis específicos. También se generan data sets y se aplican técnicas de OLAP y Data Marts

#### Data Marts:

Son **subconjuntos** del Data Warehouse diseñados para **departamentos específicos o áreas temáticas**, como ventas o finanzas. Los Data Marts facilitan el acceso a datos relevantes para cada departamento y optimizan la rapidez en los reportes y consultas. Siendo **vistas diseñadas para sectores**. El grado de **granularidad, latencia e información depende de los requerimientos** del departamento

### Inmon y Kimball

* **Enfoque de Inmon (Top-Down)**: Propone construir un Data Warehouse centralizado y normalizado que luego alimenta Data Marts departamentales, asegurando una alta integridad y un diseño centrado en la empresa. Normalizado en 3FN, con datos atómicos. Desventaja con esto son las consultas complejas, dificultando el análisis directo de la información y el uso de herramientas de reporting

Diagrama

Descripción generada automáticamente

* **Enfoque de Kimball (Bottom-Up)**: Inicia con Data Marts para departamentos específicos y los integra en un Data Warehouse. Este método es más ágil y permite a cada departamento construir y consultar datos según sus necesidades. **Es la unión de diferentes data marts de una organización**

Diagrama

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

## Diseño de Data Warehouse

* **Proceso** **de** **negocio**: Se selecciona la oferta en los datos disponibles y cuales de ellos son claves en las decisiones. Algunos requerimientos los puede seleccionar el usuario
* **Granularidad**: Que nivel de detalle necesita el usuario final
* **Latencia**: frecuencia o rapidez con la que los datos son actualizados/cargados
* **Detallar dimensiones y hechos**

### Modelado

Diagrama

Descripción generada automáticamente

#### Dimensiones/ atributos:

* **Son de tipo texto**
* **Agrupables**
* **Cualitativos en transacciones (Dia, local, vendedor, producto)**

#### Hechos/ métricas:

* **Son de tipo numérico**
* **Agregables (SUM, AVG, COUNT, etc.)**
* **Cuantitativos en transacción (Venta neta, cantidad, saldo, promedio, etc.)**

### Esquema Estrella

El **esquema estrella** es un modelo dimensional simple, caracterizado por una estructura de **tablas de hechos** en el centro conectada a varias **tablas de dimensiones** que la rodean, formando una forma similar a una estrella.

**Características:**

* **Tabla de hechos**: Contiene datos numéricos o métricas, como ventas, ingresos o cantidades. Está en el centro del esquema.
* **Tablas de dimensiones**: Son tablas descriptivas que rodean la tabla de hechos y contienen atributos cualitativos (ej., nombre del producto, fecha, región). Las tablas de dimensiones son desnormalizadas, por lo que tienen todos los datos necesarios sin depender de otras tablas.
* **Facilidad de consulta**: Al estar desnormalizado, el esquema estrella permite consultas rápidas, ya que evita la unión de muchas tablas.

**Diagrama

Descripción generada automáticamente**

### Esquema Snowflake (Copo de nieve)

El **esquema snowflake** es una versión más compleja del esquema estrella, en la cual las **tablas de dimensiones** están normalizadas (divididas en múltiples tablas), formando una estructura que se asemeja a un copo de nieve.

**Características:**

* **Normalización de dimensiones**: Las tablas de dimensiones están divididas en subtablas para reducir la redundancia. Por ejemplo, en lugar de tener una sola tabla de dimensión de ubicación, puede tener tablas separadas para país, estado y ciudad.
* **Estructura más compleja**: La normalización de las tablas de dimensiones aumenta el número de tablas y las relaciones entre ellas, creando una estructura ramificada.

Diagrama

Descripción generada automáticamente

## ETL

Para crear un data warehouse, los datos se someten a un proceso llamado **ETL** (Extract, Transform, Load), que consta de:

1. **Extracción**: Los datos se extraen de distintas fuentes de datos (sistemas operacionales, CRM, ERP, etc.). Almacén intermedio llamado STAGING (Pasarela entre sistema fuente y destino), consiste en evitar saturación de servidores
2. **Transformación**: Los datos se limpian, se estructuran y se adaptan a un formato estándar. Unas series de reglas del negocio
3. **Carga (Load)**: Finalmente, los datos se cargan en el data warehouse para su almacenamiento y consulta. A veces se sobrescribe información antigua con la nueva

Es fundamental la Metadata

#### Carga inicial:

* Genera el inicio del DW
* Define la cantidad de datos históricos disponibles (Informa al usuario cual NO está disponible)
* Tiene impacto en sistemas fuentes por el volumen de datos alto

#### Refresh

* Se definen que entidades serán actualizadas por hora, día, semana, mes, etc.
* Como tratar los cambios en las dimensiones
  + **Hay dimensiones que cambian lentamente**: Ej. Estado Civil

**Estrategias**: Actualiza la forma, nueva versión, versión anterior y actual, fechas auditorias

* + **Hay dimensiones que cambian rápidamente**: Ej. Puntaje de compras

**Estrategias**: Separar parte dinámica y estática, separo los volátiles, llevar la información a fact table.

* + **Dimensión** **Junk**: Atributos que no pertenecen a una dimensión.

**Estrategias**: Fact table, cada atributo como una dimensión aparte, sacarlos del diseño, Junk.

#### Problemas a tratar en proceso ETL

* Data Quality
* Detección de los cambios en sistemas fuentes.
* Normalización
* Procedimientos por contingencia

### Beneficios

* Poder crear un Master Data Managment (Repositorio central), estandarizado
* Posibilidad de tomar decisiones estratégicas basadas en análisis de datos
* Integrar sistemas (Cuando las organizaciones crecen se agregan más fuentes de datos)
* Tenes visión global de los datos consolidades en un data warehouse

## Cloud Computing

**Cloud** o nube, es el símbolo que representa la internet, **computing** representa conceptos de informática, lógica, coordinación y almacenamiento

La computación en la nube es un **paradigma que permite ofrecer servicios de computadoras a través de Internet** sin la necesidad de tenerlos localmente. Es de manera remota

Esto lo hace **escalable,** pudiendo tener **acceso remoto** con **pagos de uso** en los cuales el proveedor del servicio se encarga del **mantenimiento y actualizaciones gestionadas**. El objetivo es **compartir recursos entre consumidor, socio y vendedor en la cadena de valor de la nube**

**Tradicional App/Server**

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

**Virtual Server Model**

Interfaz de usuario gráfica

Descripción generada automáticamente

### Tipo de nubes

**Nube publica**

Es cuando el servicio en la nube es **ofrecido al publico** por ejemplo Google, Amazon o Microsoft. Ej.: Google Drive. Lo administran por **terceros**, los contenidos de distintos clientes pueden estar en **mismos servidores**

**Nubes privadas**

Es cuando una empresa **crea su propia nube que solo utilizan ellos mismos**, siendo exclusivo, se controla internamente. Ej.: Un banco que almacena sus datos

El proveedor es propietario del servidor, red y disco y **puede decidir que usuarios están autorizados a utilizar infraestructura**

**Nube Hibrida**

**Mezcla nube pública y privada,** algunas cosas quedan exclusivas para la empresa, otras para el público. Ej.: Empresa con datos confidenciales en la privada y en la publica la aplicación los datos menos sensibles

La infraestructura de nube se compone de una serie de nubes de todo tipo, pero las nubes tienen la capacidad de comunicarse a través de sus interfaces para permitir que los datos y / o aplicaciones pueden mover de una nube a otra

**Características principales**

* Autoservicio **bajo demanda**
* Permitir el acceso desde la **red** (pública, privada, híbrida, comunitaria)
* **Asignación de recursos en modo multiusuario**
* Capacidad de rápido crecimiento
* Servicio medido
* Elasticidad y escalabilidad
* Seguridad

**Amenazas**

* Seguridad y confidencialidad
* Rendimiento
* Los proveedores de tecnología

**Ventajas**

* Integración fácil y rápida
* Prestación de servicios a nivel mundial
* **Sin** **instalación** de **ningún tipo de hardware para la empresa**
* **Menor inversión para empezar a trabajar (ROI)**
* **Implementación más rápida, en cuestión de días u horas y con menos riesgos.**
* **Uso eficiente de energía ya que se consume sólo la necesaria.**

**Desventajas**

* Los datos "sensibles" del negocio no residen en las instalaciones de la empresa.
* **Dependencia a la disponibilidad de acceso a Internet**
* **Interdependencia a los proveedores de servicios**
* La información de la empresa debe recorrer diferentes nodos y cada uno de ellos son un foco de inseguridad
* Velocidad de respuesta, para los sistemas muy críticos puede no salta

### Arquitectura de la nube

**Modelos de servicio y despliegue:**

La adopción de servicios de conmutación en la nube está creciendo a gran velocidad y una de las razones se debe a su arquitectura que acentúa los beneficios de servicios compartidos en lugar de productos aislados

Gráfico, Gráfico de embudo

Descripción generada automáticamente

#### Tipos comunes de XaaS:

1. **SaaS (Software as a Service**): Aplicaciones de software accesibles a través de la nube, como Gmail, Microsoft 365, o Dropbox.
2. **PaaS (Platform as a Service):** Proporciona una plataforma para que los desarrolladores creen y desplieguen aplicaciones sin gestionar la infraestructura. Ejemplo: Google App Engine.
3. **IaaS (Infrastructure as a Service):** Ofrece infraestructura de computación como servidores virtuales, redes y almacenamiento. Ejemplo: Amazon Web Services (AWS).

# Preguntas de examen

1. Depende un texto realizar

* DFM
* Modelo estrella y Snowflake

1. Definir conceptos de Hecho y Dimensión. ¿Cuál es su relación?

Hecho son datos cuantitativos o métricos, la actividad del negocio. Son numéricos y medibles en DW (Ventas, Inventario, tiempo de espera)

Dimensión es un dato cualitativo, describe contexto de hechos. Son en general datos textuales (Tiempo, cliente, producto, etc.)

Los hechos y las dimensiones están **relacionados** mediante claves, formando un modelo que permite a los usuarios analizar los hechos en diferentes contextos. La **tabla de hechos** contiene claves foráneas que se conectan con las claves primarias en las **tablas de dimensiones**.

1. Describir proceso ETL básico

Creación y mantenimiento de un **Data Warehouse**. Se refiere a los pasos necesarios para recopilar, procesar y almacenar datos de diversas fuentes en un sistema centralizado para análisis y toma de decisiones

Extracción, se sacan los datos

Transformación, se limpian los datos según las necesidades de DW

Cargan, se cargan en el DW

1. 5 diferencias entre OLTP y OLAP

* Propósito: OLTP se centra en la gestión de transacciones diarias; OLAP se orienta al análisis de datos históricos y toma de decisiones.
* Datos: OLTP maneja datos actuales y en tiempo real; OLAP utiliza datos históricos y consolidados.
* Consultas: OLTP realiza consultas rápidas y sencillas; OLAP ejecuta consultas complejas y multidimensionales.
* Normalización: OLTP tiene una estructura altamente normalizada para evitar redundancia; OLAP suele estar desnormalizado para mejorar la velocidad de consulta.
* Usuarios: OLTP es usado por empleados operativos para transacciones diarias; OLAP es usado por analistas y ejecutivos para análisis y reportes.

1. Defina de forma precisa que es un DataMart y diferenciar Kimball e Inmon

Un **Data Mart** es un subconjunto de un Data Warehouse que almacena datos específicos para un área o departamento (como ventas, finanzas o marketing), facilitando el acceso a la información relevante para usuarios finales de ese sector. Su propósito es optimizar la consulta y análisis de datos para necesidades específicas sin requerir acceso al Data Warehouse completo

* **Kimball**: Adopta un enfoque **Bottom-Up**; el Data Warehouse se construye integrando múltiples Data Marts creados primero para cada departamento. Los Data Marts son dimensionales y desnormalizados, facilitando consultas rápidas.
* **Inmon**: Emplea un enfoque **Top-Down**; primero se construye un Data Warehouse centralizado y normalizado, que luego alimenta Data Marts departamentales específicos. El Data Warehouse principal asegura la integridad y consistencia de los datos.

1. ¿Qué es latencia y la granularidad en DW?

La **latencia** se refiere a la **frecuencia o rapidez con la que los datos se actualizan o cargan** en el Data Warehouse. En otras palabras, es el tiempo que tarda un dato en estar disponible para consulta después de su creación o modificación en la fuente original

* Baja latencia: Casi a tiempo real se actualizan
* Alta latencia: Con menos frecuencia

La **granularidad** es el **nivel de detalle** de los datos almacenados en el Data Warehouse. Determina qué tan específicos o agregados están los datos

* **Alta granularidad**: Datos muy detallados (por ejemplo, transacciones individuales), que permiten análisis profundos, pero ocupan más espacio y son más complejos de manejar
* **Baja granularidad**: Datos más agregados (por ejemplo, ventas mensuales o anuales), que facilitan el análisis general y requieren menos espacio, pero limitan el detalle disponible

1. Responder V o F, justificar F
   1. Un proceso ETL tiene el beneficio de integrar sistemas. VERD - permite extraer datos de múltiples fuentes, transformarlos y cargarlos en un Data Warehouse
   2. Un proceso ETL es vital para crear un repositorio central estandarizado de la organización. VERD - ETL es fundamental para consolidar datos de diversas fuentes en un Data Warehouse
   3. Un proceso ETL puede ser usado para migrar bases de datos legacy. VERD - ETL puede extraer datos de sistemas legacy, transformarlos y cargarlos en nuevas bases de datos o en un Data Warehouse, facilitando así la migración de datos de sistemas antiguos a sistemas modernos.
   4. Un proceso ETL no permite la integración de datos de diferentes sucursales de una empresa. FALSO- Permite la integración de datos de diferentes sucursales
2. Responda V (verdadero) o F (Falso), justifique cada una de las repuestas: Hablando de Cloud Computing, entre las ventajas podemos citar:
   1. EI Cloud Computing es eficiente ya que consume sólo la energía necesaria VERDADERO
   2. EI Cloud Computing es económica ya que ahorra la instalación de hardware a la empresa VERD - es económica, ya que elimina la necesidad de que las empresas inviertan en hardware costoso.
   3. EI Cloud Computing genera una dependencia a la disponibilidad de acceso a internet. VERD - Dado que los servicios de nube están basados en Internet, cualquier interrupción en la conectividad afectará el acceso a esos servicios, lo que podría generar problemas operativos si no hay una buena conexión o si se produce una falla en la red
   4. EI Cloud Computing nos da una menor inversión para empezar a trabajar (ROI). VERD - permite a las empresas evitar grandes gastos en hardware, software y personal para mantenimiento. En lugar de hacer una gran inversión inicial, las empresas pagan solo por lo que usan, lo que mejora el retorno de inversión (ROI) al reducir gastos fijos

## Para Practicar DFM

Una cadena de ventas minoristas posee varias aplicaciones de software independientes que trabajan en distintas bases de datos, como consecuencia recopilar toda la información necesaria para tornar decisiones se ha vuelto un trabajo muy engorroso que llegar a demorar días. Dada esta situación el director le ha encomendado a usted la tarea de implementar un nuevo sistema que le permita tener información de forma rápida y efectiva para la toma de decisiones. Inicialmente el director Ie planteo que necesita conocer el monto y la cantidad de ventas realizadas por zona, en un determinado tiempo y por los productos diferenciando en categorías que ellos necesarios dependiendo de la época del año.

La empresa tiene un aplicativo que registra todas las ventas realizadas en los diferentes puntos de venta que registra la información en una base de datos

1. Hacer DFM
2. Resolver diseño en modelo estrella y Snowflake

DFM

Hay que ver Hechos/Métrica (Cuantitativo) y Dimensiones/Atributo (Cualitativo)

Una dimensión que no debe faltar es la del tiempo (Dia, mes, año)

El ministerio del interior que se encarga de exponer los datos y resultados de las ultimas elecciones necesita diseñar un DFM que permita exponer la cantidad de votos efectuados por cada ciudadano. Lo novedoso es que los datos de entrada corresponden a la digitalización de cada uno de los conteos por urna ubicada en una mesa de un colegio que esta distribuido en cada municipio de la provincia del país. También se reciben de otras BD los votos electorales realizados por ciudadanos extranjeros en el exterior.

Se desea tener un diseño de exploración que permita armar un mapa del país con los votos por cada candidato. En la elección se elige Presidente, Gobernador e Intendente como así también legisladores provinciales y nacionales.

Dado que también se va a efectuar el conteo definitivo con todos los telegramas recibidos se desea también tener el detalle de día y hora en que cada telegrama fue digitalizado

La nueva base DW se va a cargar con datos de distintas elecciones: Aparte 2023 (Paso, Generales y Ballotage) también se intentará unificar con BD de elecciones legislativas y presidenciales pasadas: 2021, 2019, 2017, etc.

1. Hacer DFM
2. Resolver diseño en modelo estrella y Snowflake

*Es un ejercicio no verificado, puede tener errores:*

Si se necesita, se puede categorizar así:

* Municipio (PK\_MUN, desc, *FK\_PROV*)
* Provincia (PK\_PROV, desc, *FK\_PAIS*)
* País (PK\_PAIS, desc)
* Fecha (PK\_FEC, dia, mes, año, horario)
* **Telegrama (cant\_votos, *FK\_FEC, FK\_UR, FK\_CAND, FK\_PROV, FK\_MUN*)**
* Colegio (PK\_COL, nombre, dirección)
* Urna (PK\_UR, nro\_mesa)
* Voto (*FK\_DNI, FK\_UR*)
* Candidato (PK\_CAND, nombre, categoría)
* Ciudadano (PK\_DNI, nombre, dirección, *FK\_PAIS, FK\_PROV, FK\_MUN*)

Diagrama

Descripción generada automáticamente