



Día, Fecha:	Lunes, 23/01/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Laboratorio Seminario de Sistemas 2

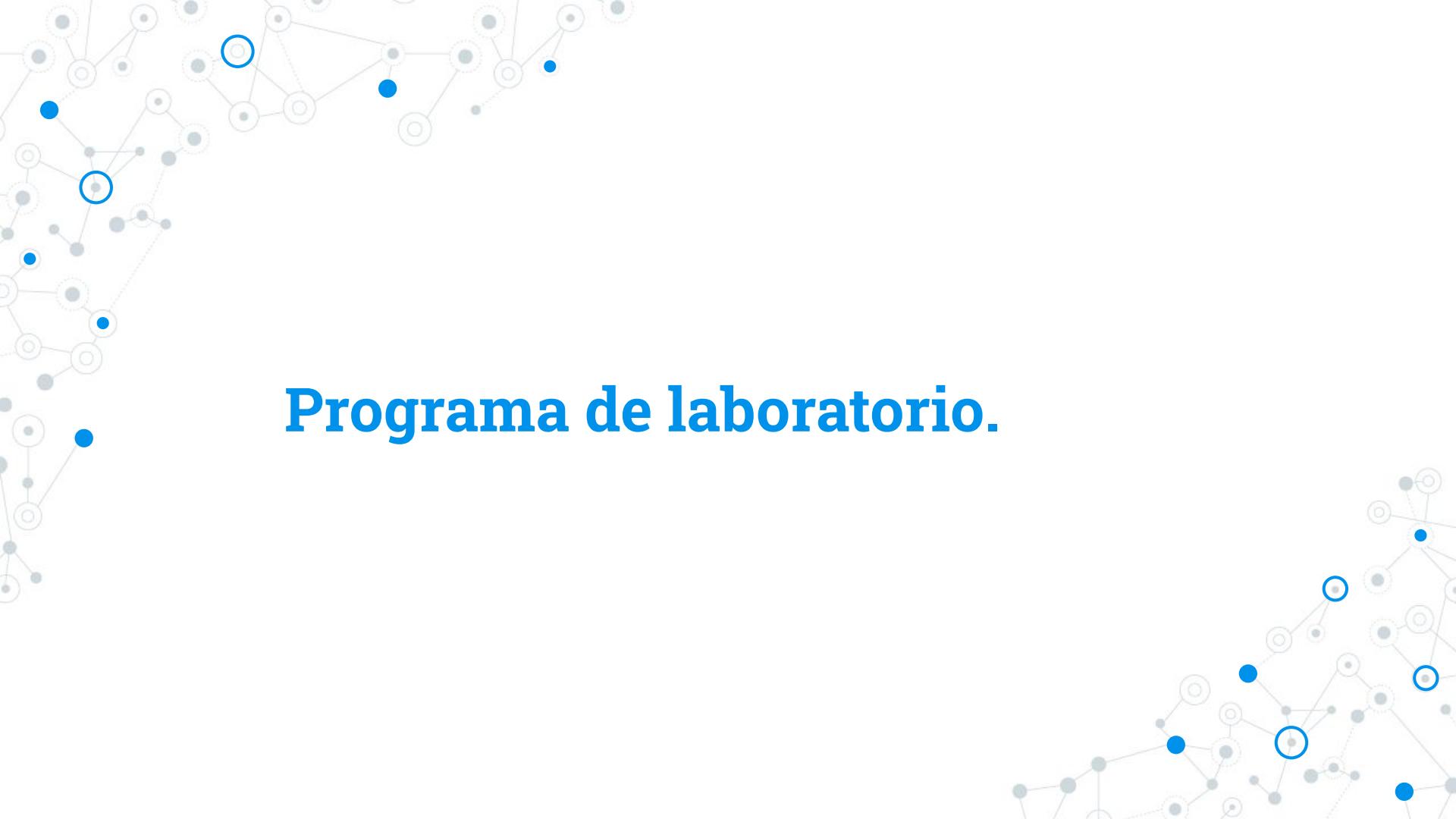
Sección “A”

Agenda

- Presentación.
- Detalle de programa y contenido.
- Información General.
- Clase 1.

Información Personal

- **Nombre de Tutor:**
Escarleth Andrea Velasco Campos
- **Correo electrónico:**
2301603280101@ingenieria.usac.edu.gt
velandreas@gmail.com



Programa de laboratorio.

Información General.

- **Horario:**

Lunes 17:20 – 19:00

- **Asunto de correos:**

[SS2]Duda

[SS2]Asunto De Correo

- **Formulario a llenar:**

Asistencia: <https://forms.gle/vmEBxs6ThhADau4N9>

Asignación: <https://forms.gle/e3BTJ4URyuZ9qdK97>

Grupo

WhatsApp:<https://chat.whatsapp.com/BXbSba2GyaDDMeBnaZN6vR>

Ponderación

Proyecto Fase 1	15 pts
Proyecto Fase 2	20 pts
Práctica 1	10 pts
Práctica 2	10 pts
Tareas (5)	10 pts
Hojas de Trabajo (5)	10 pts
Cortos (3)	15 pts
Examen Final	10 pts
Total	100 pts

Normas de trabajo

- Tareas, prácticas y proyecto se trabajarán de forma individual.
- Entregas tarde **no se calificarán**.
- Las copias detectadas tendrán 0 y su respectivo informe a la escuela.
- Uso del grupo de WhatsApp lo más **profesional** posible.

Aspectos Generales

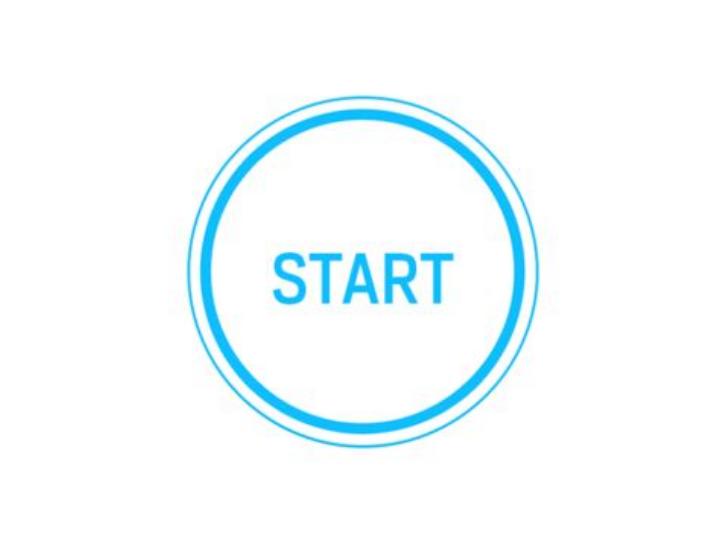
- Para entregas de tareas se tendrá el siguiente formato:
Tarea#_carnet.pdf
- Para hojas de trabajo se tendrá el siguiente formato:
HT#_carnet.pdf
- Los exámenes cortos se realizarán por UEDI.
- Domingos a las 9:00 PM

Contenido

- **Unidad 1:** Cubos Multidimensionales.
- **Unidad 2:** Solución de BI con herramientas Microsoft.
- **Unidad 3:** Procesamiento masivo paralelo y Hadoop.
- **Unidad 4:** Procesando Big Data con Apache Spark.

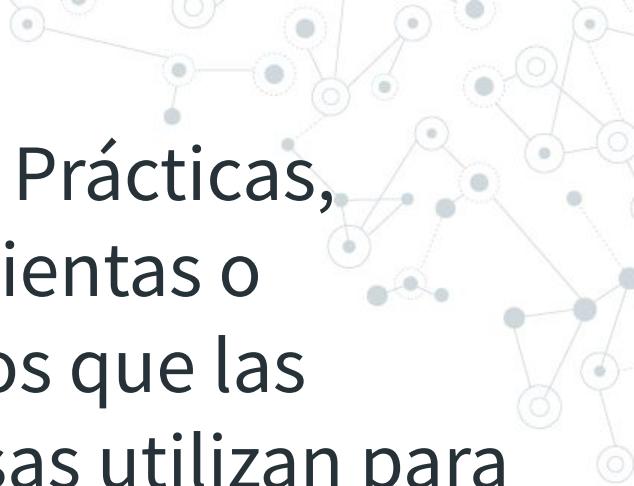


Comencemos...



START

¿Qué es Business Intelligence
o Inteligencia de Negocios?



Son las Prácticas,
herramientas o
métodos que las
empresas utilizan para
observar, analizar y
comprender todos los
datos según el modelo
de negocio.

¿Qué Beneficios Tiene?

- Hace más sencillo la generación de cambios en las estrategias.
- Controlar la información.
- Contar con información actualizada.
- Mejora acciones de marketing
- Incrementa las ventas.
- Aumenta la rentabilidad de una compañía al permitir el acceso a información de clientes, productos, competencia, mercado, etcétera.
- Contribuye en la segmentación de clientes, lo cual aumenta la adquisición y conversión.
- Provee de información a la empresa.
- Mejora la productividad con información.
- Disminuye gastos y controla los costos.
- Es más sencillo identificar tendencias.
- Se crea conocimiento



DATA SCIENCE LIFECYCLE

sudeep.co

01

BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02

DATA MINING

Gather and scrape the data necessary for the project.

03

DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04

DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05

FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

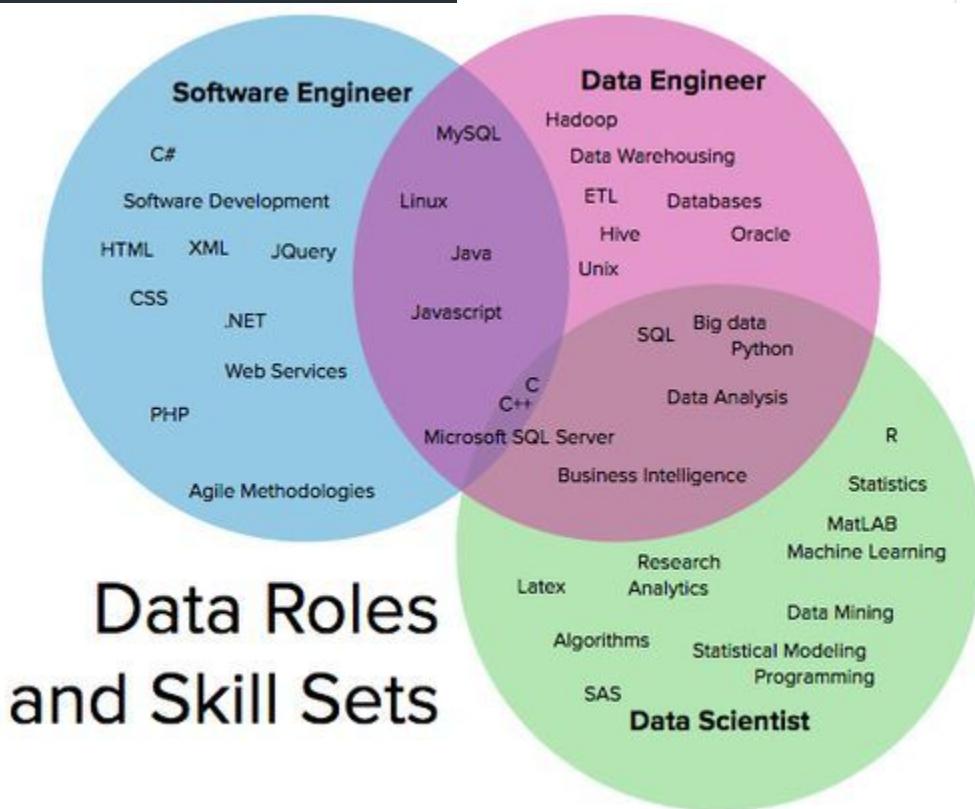
06

PREDICTIVE MODELING

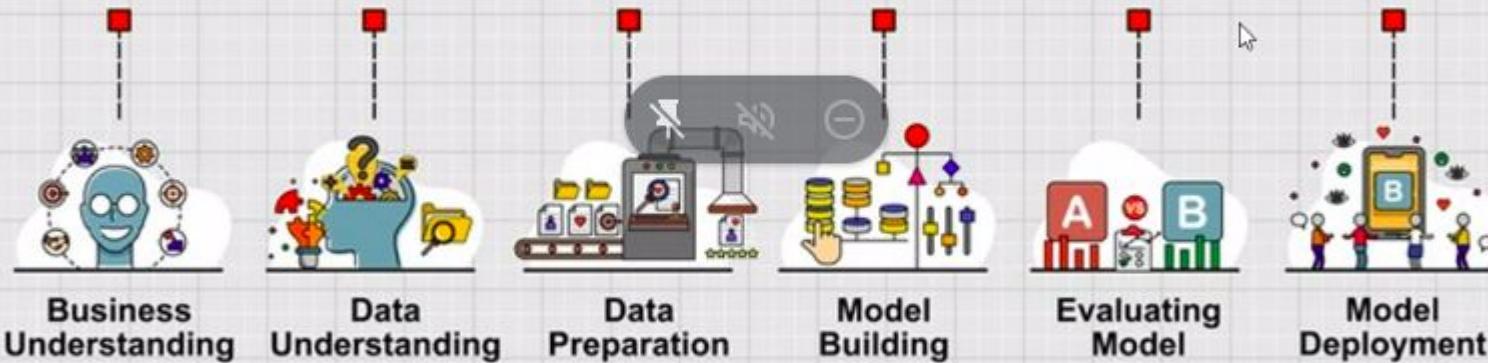
Train machine learning models, evaluate their performance, and use them to make predictions.

DATA VISUALIZATION

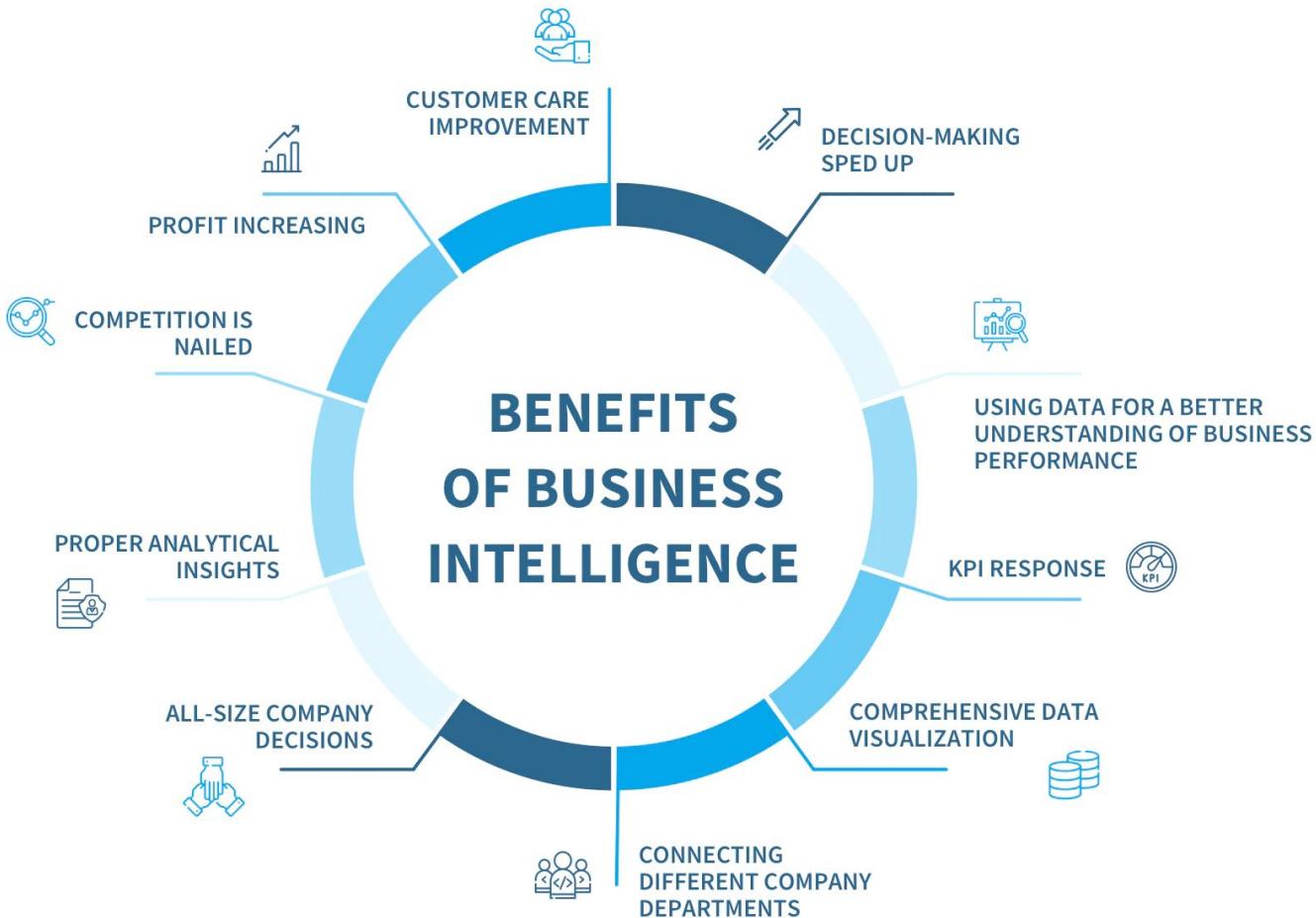
Communicate the findings with key stakeholders using plots and interactive visualizations.

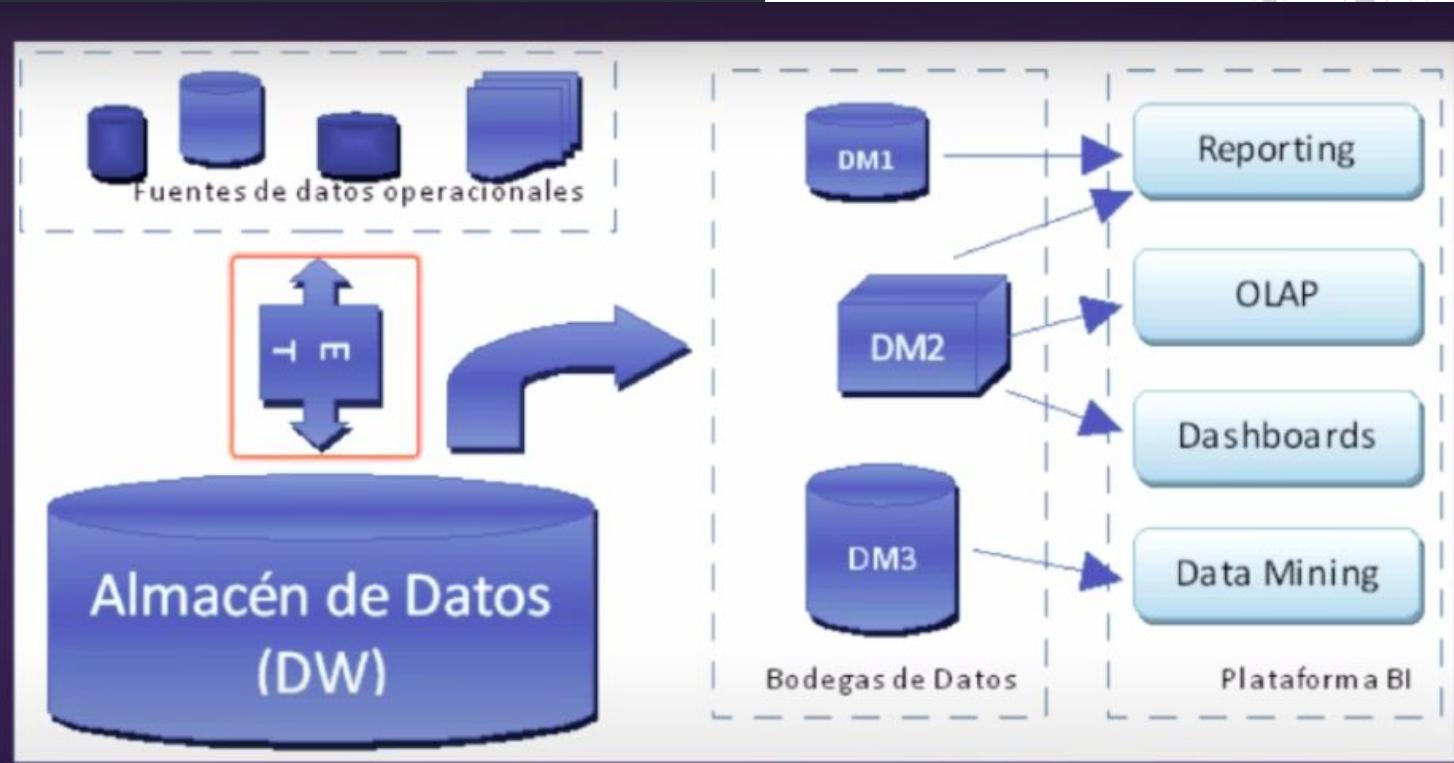


DATA SCIENCE PROCESS



shutterstock.com · 1833523507





Componentes BI

Herramientas de BI



QlikView



Power BI

Cubos Multidimensionales OLAP

- Estos cubos son estructuras multidimensionales las cuales nos permiten **analizar** bases de datos relacionales de gran volumen con gran **facilidad** y **rapidez** esto ya que reducen en gran parte el tiempo y los recursos para el análisis.

Cubos Multidimensionales OLAP

- Comúnmente utilizados para **reportería**, la data es categorizada por **dimensiones** que usualmente están precalculadas para incrementar drásticamente el desempeño de las consultas a comparación de una base de datos relacional.

Cubos Multidimensionales OLAP

- Uno de los lenguajes más utilizados para consulta y realización de tareas con cubos OLAP es MDX (MultiDimensional eXpressions)



Facilidad de uso:

- Cuando el cubo está construido cualquier usuario así sea con pocos o nulos conocimientos técnicos puede consultarla en cualquier momento.

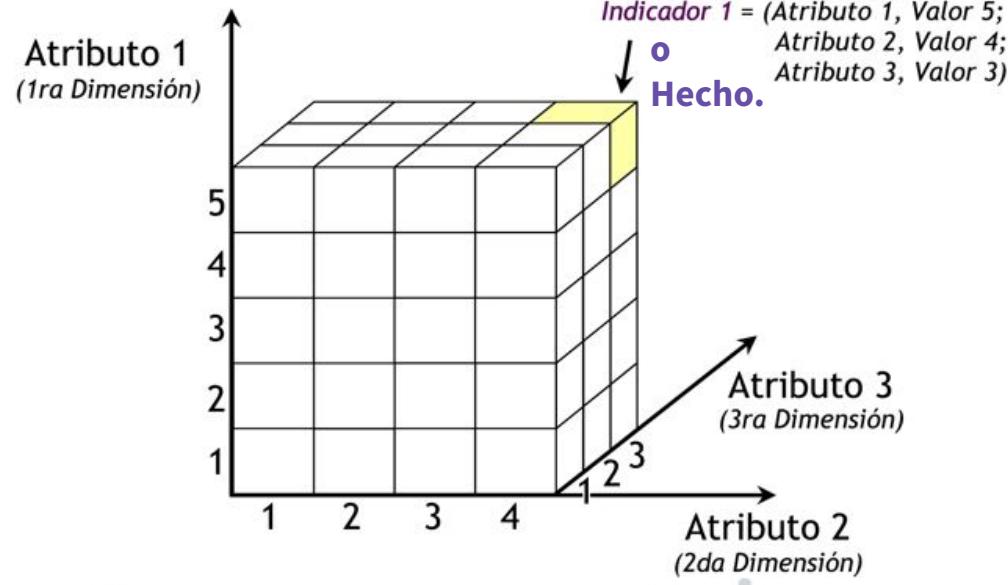
Rapidez:

- Si el cubo está bien construido este suele tener distintas agregaciones precalculadas, y esto hace que los tiempos de respuesta sean cortos.

Componentes de un Cubo OLAP

- Hechos o indicadores.
- Dimensiones
- Jerarquías

Cubo OLAP



Hechos o indicadores

- Son definiciones a partir de las cuales podremos obtener valores numéricos que ayudan para el análisis.
- Dependen de las Dimensiones y Jerarquías.

Dimensiones

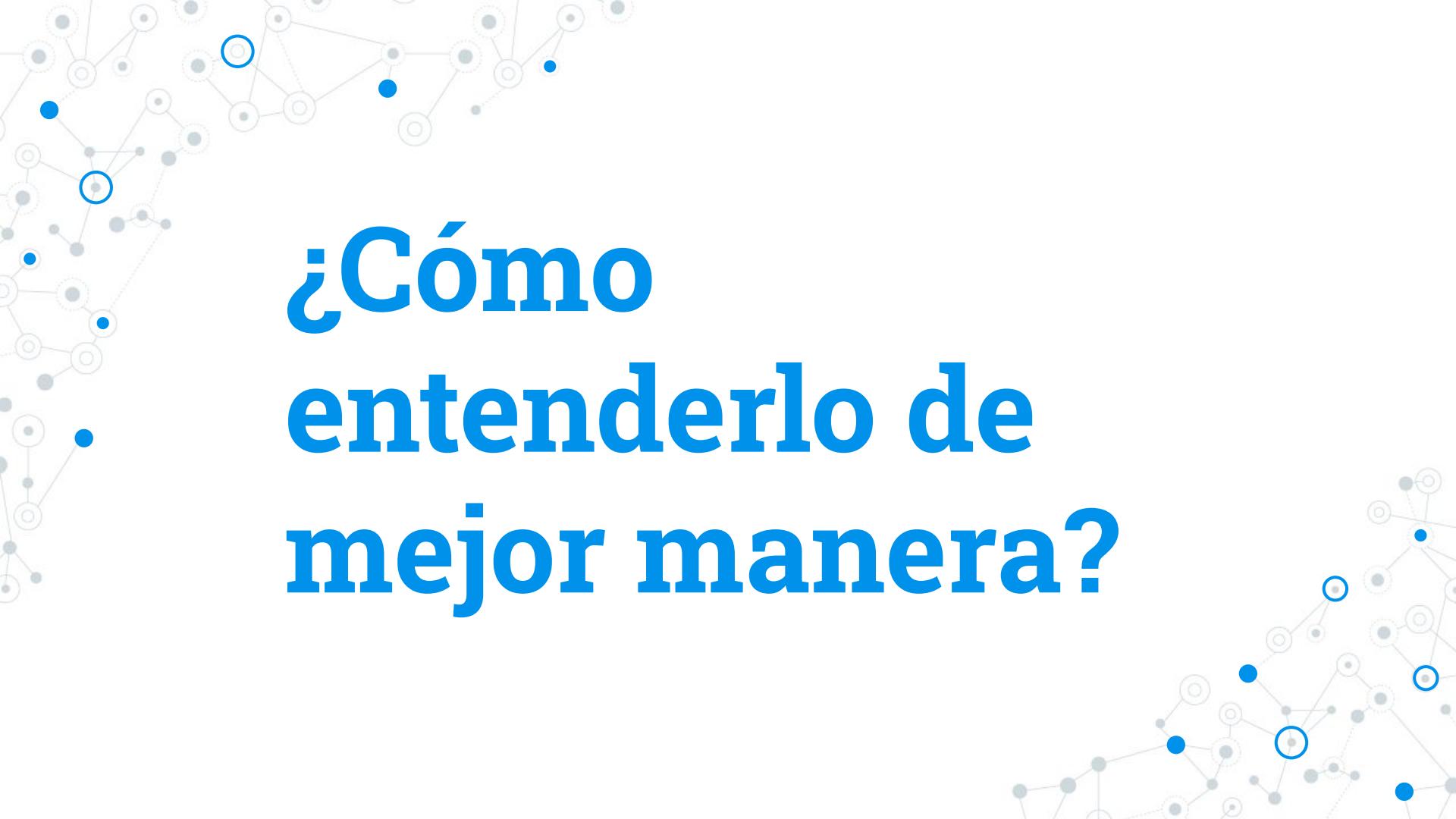
- Son también llamados atributos ya que describen los datos, son criterios que se utilizarán para analizar los indicadores dentro de un cubo multidimensional.

Jerarquías

- Es una relación lógica de tipo padre-hijo entre las dimensiones o atributos, al utilizar estas se pueden analizar datos desde el nivel más general hasta el más detallado.

Cubos Multidimensionales OLAP

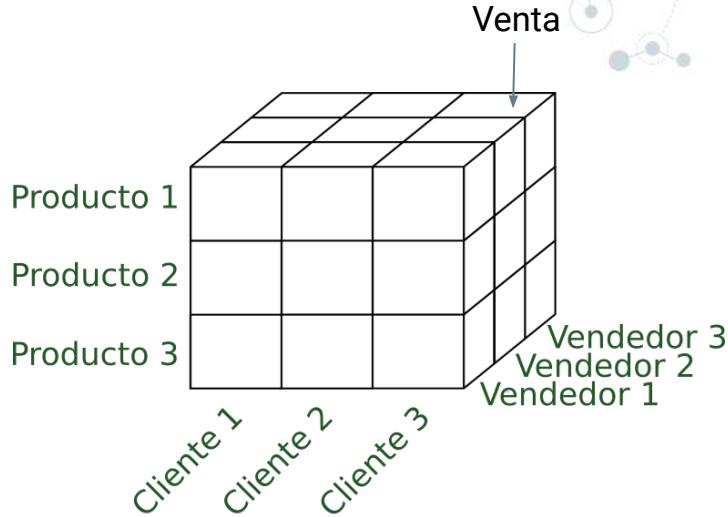
- Cabe destacar que no en todas las organizaciones es factible hacer uso de cubos OLAP, por lo que previamente es conveniente analizar las ventajas y desventajas que conlleva su implementación.



¿Cómo entenderlo de mejor manera?

Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?
- ¿Cuáles serían los hechos?



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?

R//

Dim1: Producto

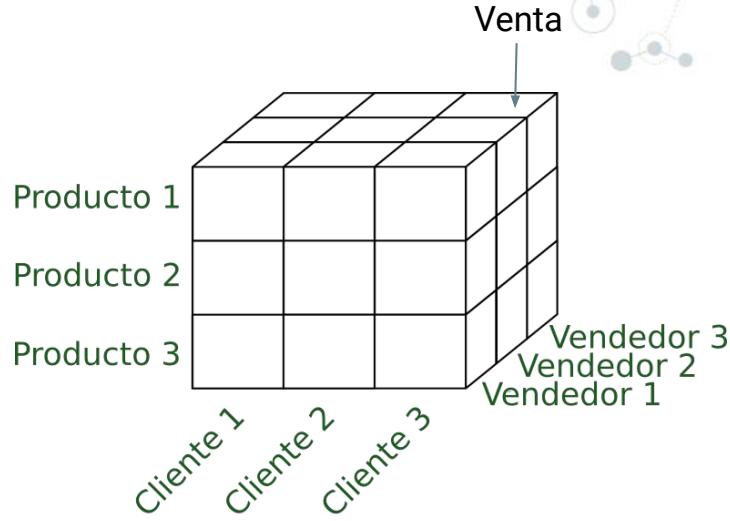
Dim2: Cliente

Dim3: Vendedor

- ¿Cuáles serían los hechos?

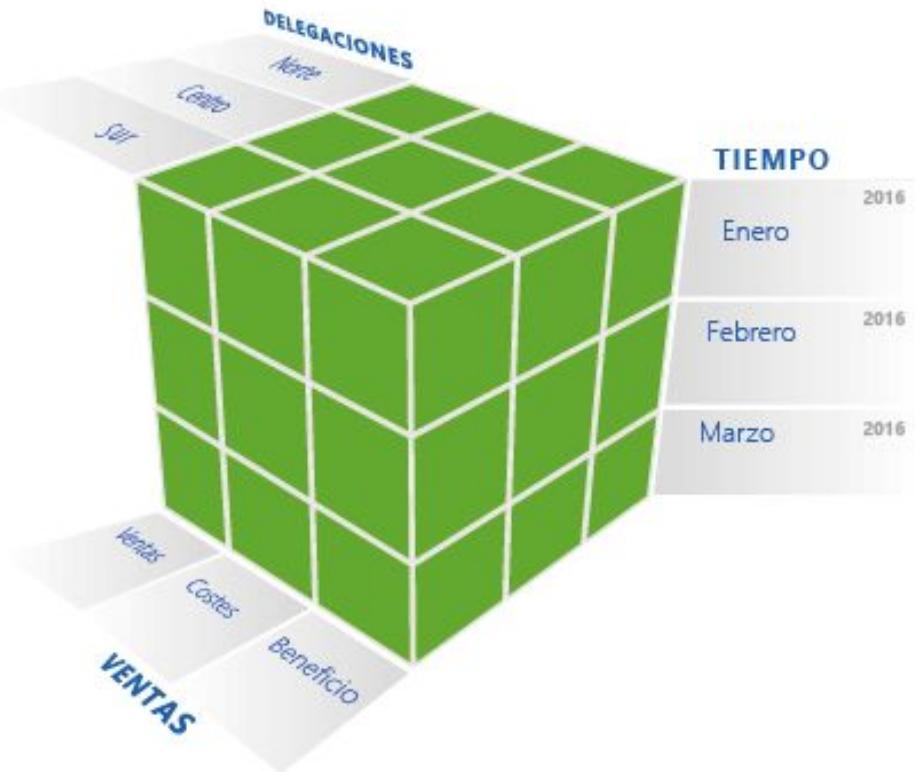
R//

Venta



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?
- ¿Cuáles serían los hechos?



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?

R//

Dim1: Delegación / Región

Dim2: Tiempo

- ¿Cuáles serían los hechos?

R//

Hecho1: Venta

Hecho2: Costes

Hecho3: Beneficios



Viéndolo desde otra perspectiva

- Supongamos que queremos conocer las **ventas** en la delegación Norte en el mes de Enero.

¿En donde se situarían esos datos usando las dimensiones y los hechos?



Viéndolo desde otra perspectiva

- Y por último queremos conocer las **ventas** y los **beneficios** en la delegación Sur y Centro en los meses de enero y febrero.

¿En donde se situarian esos datos usando las dimensiones y los hechos?



Viéndolo desde otra perspectiva

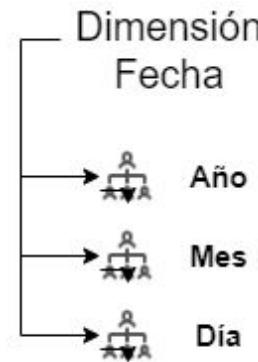
- Ahora queremos conocer los **beneficios** en la delegación Sur en los meses de enero y marzo..

¿En donde se situarian esos datos usando las dimensiones y los hechos?



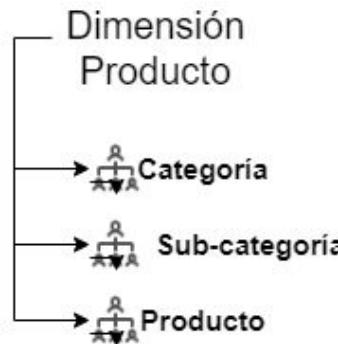
Ejemplo jerarquía:

- Supongamos que tendremos una dimensión **Fecha** con la siguiente jerarquía.



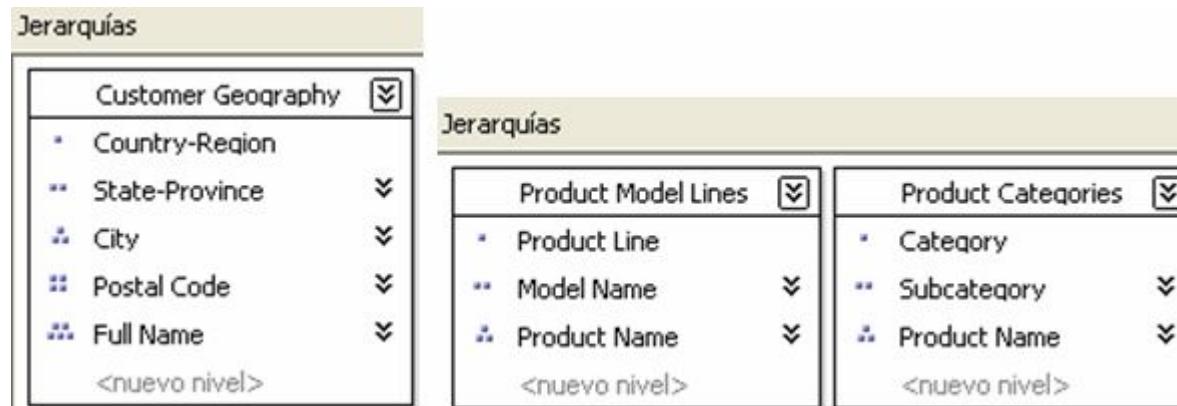
Ejemplo 2 jerarquía:

- Supongamos que tendremos una dimensión **Producto** con la siguiente jerarquía.



Ejemplo 3 jerarquía:

- Las jerarquías pueden incluir diferentes datos media vez estos tengan una relación.



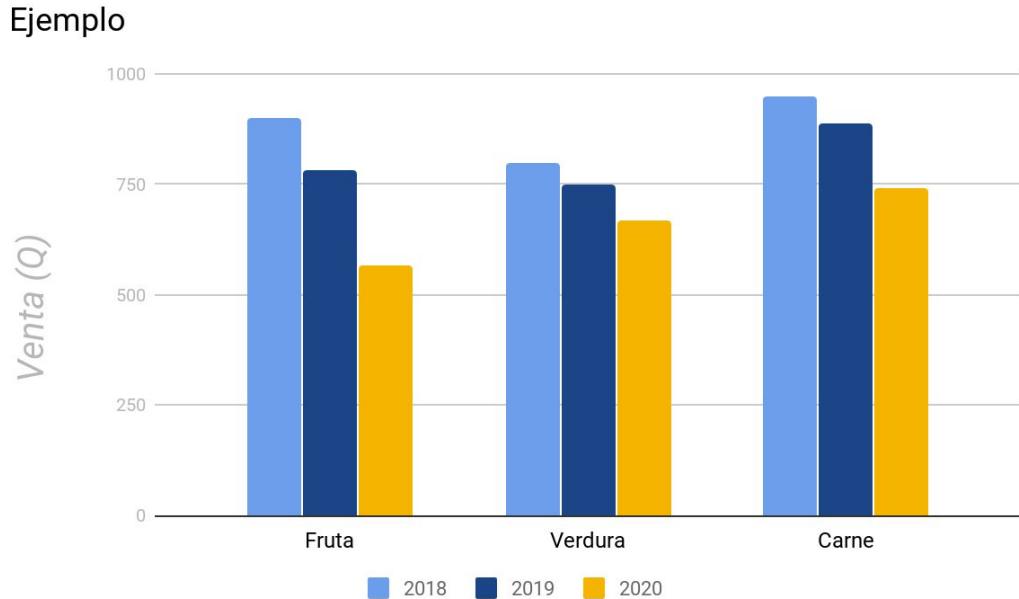


¿Cuál sería otro ejemplo de una jerarquía?



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?
- ¿Cuáles serían los hechos?



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?

R//

Dim1: Producto

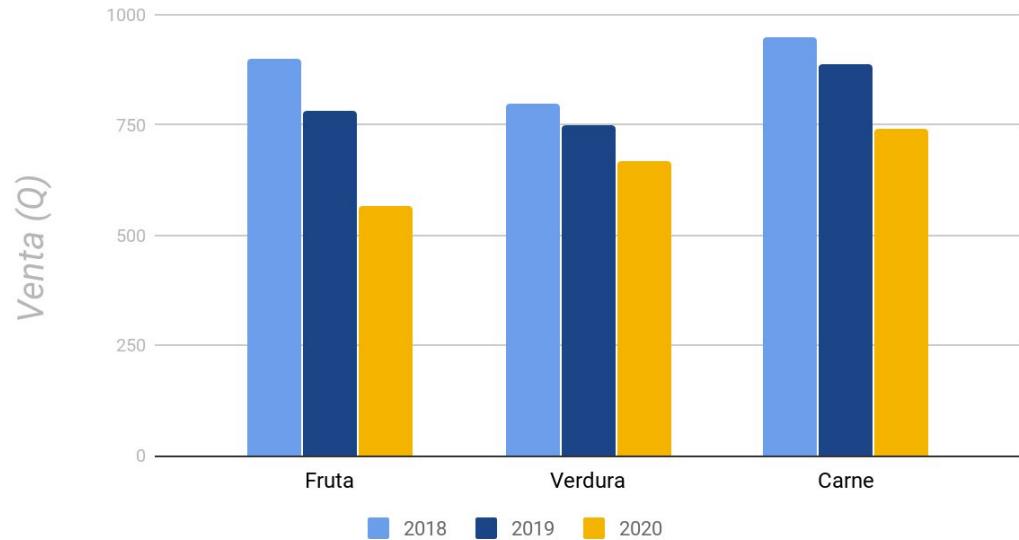
Dim2: Fecha / Año

- ¿Cuáles serían los hechos?

R//

Venta

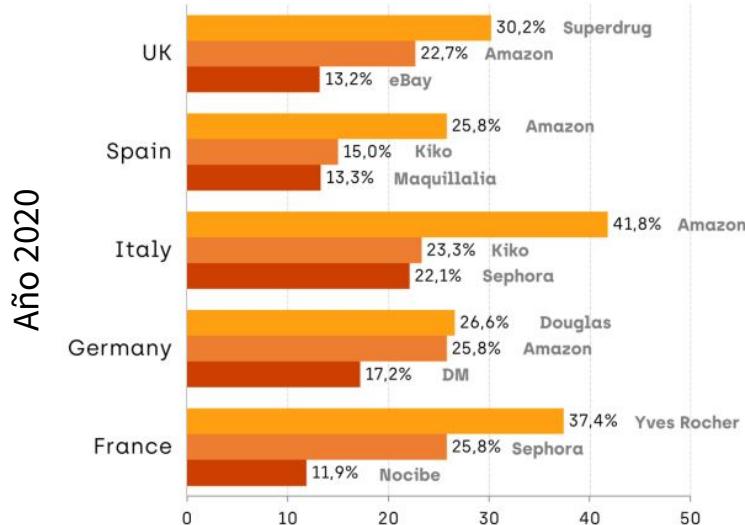
Ejemplo



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?
- ¿Cuáles serían los hechos?

Tendencia de preferencia en **compras** de cosméticos en tiendas online.



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?

R//

Dim1: Fecha / Año

Dim2: País

Dim3: Tienda

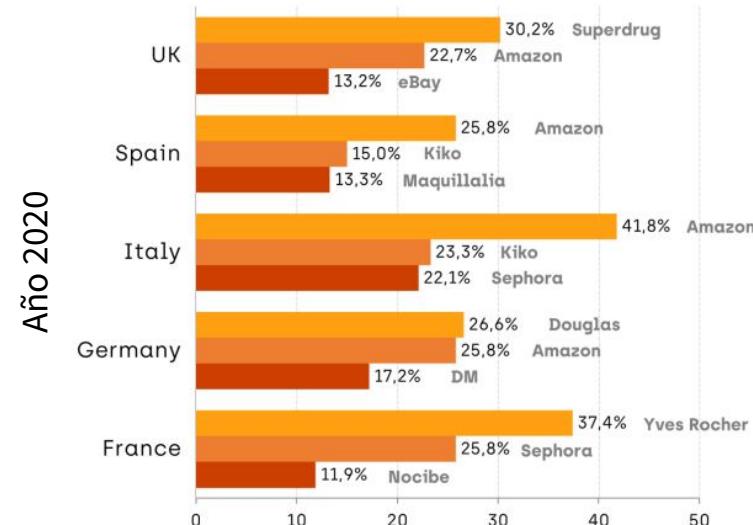
Dim4: Producto

- ¿Cuáles serían los hechos?

R//

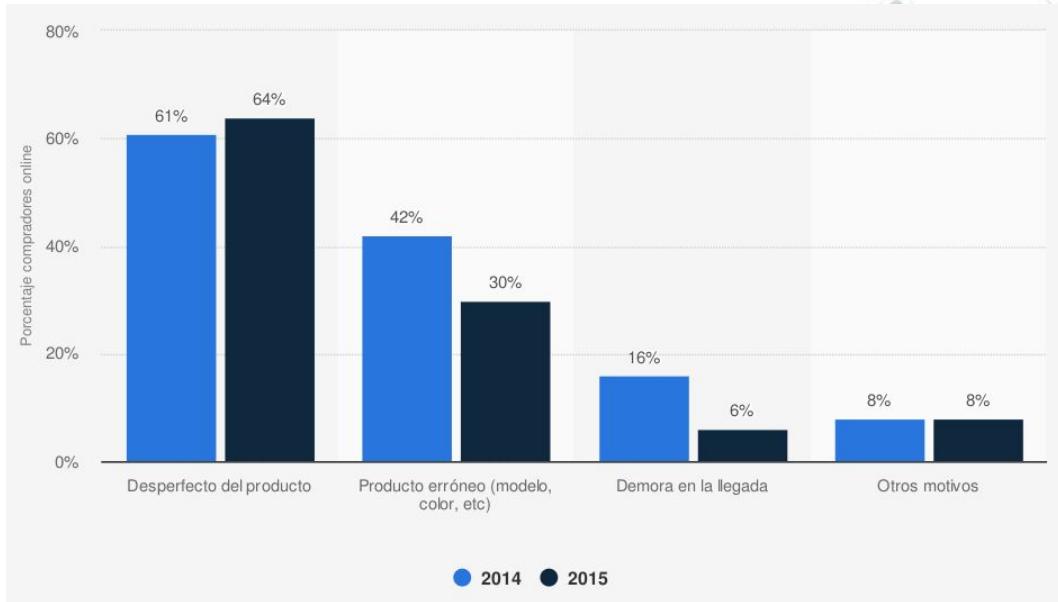
Compra

Tendencia de preferencia en **compras** de cosméticos en tiendas online.



Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?
- ¿Cuáles serían los hechos?



Razones por las cuales hubieron devoluciones en la empresa GuateUsac.

Ejemplo:

- ¿En el siguiente ejemplo cuáles serían las dimensiones?

R//

Dim1: Fecha / Año

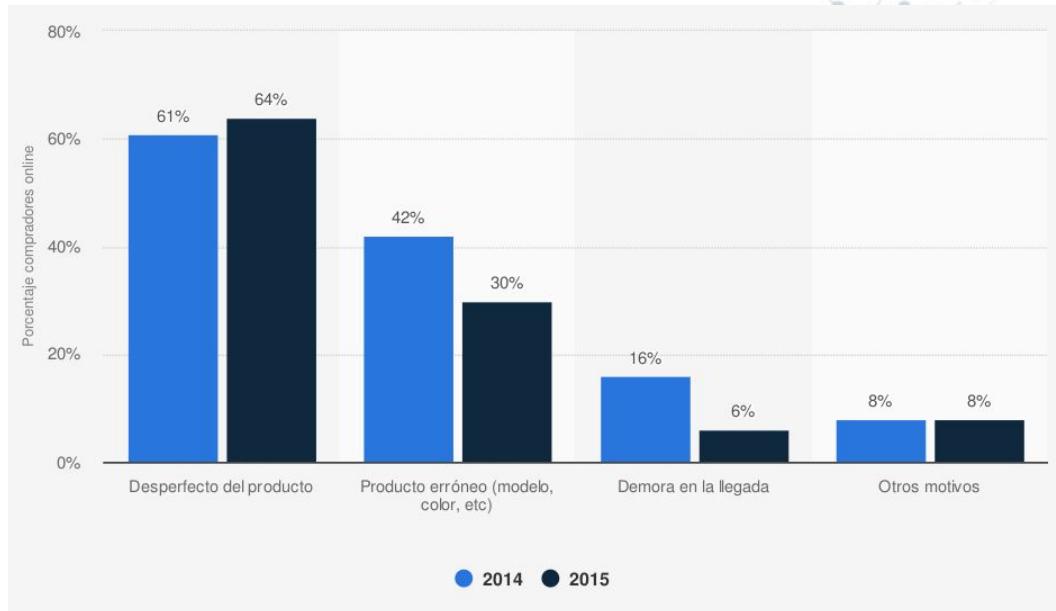
Dim2: Producto

Dim3: Razón / Motivo

- ¿Cuáles serían los hechos?

R//

Devolución



Razones por las cuales hubieron devoluciones en la empresa GuateUsac.

Uso de los cubos multidimensionales OLAP.

Los cubos OLAP permiten el análisis rápido de los datos, gracias a su particular estructura multidimensional, de acuerdo con las múltiples dimensiones con las que se cuenta. Su uso puede extenderse a diferentes áreas de negocio como:

- Ventas
- Contabilidad
- Marketing
- Logística

Herramientas a utilizar en el laboratorio

- Visual Studio (cualquier versión).
- SQL Server 2012.
- Microsoft SQL Server Data Tools - Business Intelligence(para la versión de VS que tengan). **Nota:** este debe ser el mismo idioma que su Visual Studio.
- Otras que más adelante se detallarán.

¿Dudas o Preguntas?





**GRACIAS POR SU
ATENCIÓN**

A LOS QUE PUSIERON, CLARO



Día, Fecha:	Lunes, 30/01/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Agenda 30/01/2023

- Avisos
- Clase 2.
 - Tablas de dimensión y hecho.
 - Datawarehouse
 - Proceso de ETL
- Tarea 1.
- Hoja de Trabajo

Avisos generales

- Notas
- Asignación DTT
- Fechas

Clase 2



Tablas de Dimensión

- Las tablas de dimensiones contienen atributos que describen las entidades de negocio.
- Una tabla de dimensión almacena información descriptiva sobre los valores almacenados en la tabla de hecho
- Cada tabla posee un identificador único(**llave subrogada**) que lo une a la tabla de hechos.

Llave subrogada

- Es un identificador único que se asigna a cada registro de una tabla de dimensión.
- Esta clave, generalmente, no tiene ningún sentido específico de negocio.
- Son siempre de **tipo numérico**. Preferiblemente, **un entero autoincremental**.

Llave subrogada

Beneficios:

- Facilita el particionamiento eficiente de los datos físicos.
- Crear una separación de modelos multidimensionales para facilitar el control de cambios.
- Mejorar el rendimiento de operaciones.

Tablas de Dimensión

CLIENTES	
id_Cliente	NombreCliente

PRODUCTOS	
id_Producto	
Rubro	
Tipo	
NombreProducto	

FECHAS	
id_Fecha	
Año	
Trimestre	
Mes	
Día	

Tabla de Hechos

- Estas tablas contienen los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones.
- Son los indicadores del negocio, por ejemplo, ventas , pedidos, reclamos, compras, devoluciones, etc. Cada hecho representa una transacción o evento.
- Cada registro de esta tabla posee una clave primaria que está compuesta por las claves primarias(**llaves subrogadas**) de las tablas de dimensiones relacionadas a este.
- Es importante resaltar que la tabla de hechos idealmente debe almacenar solo valores numéricos.

Tabla de Hechos - Llaves Subrogadas

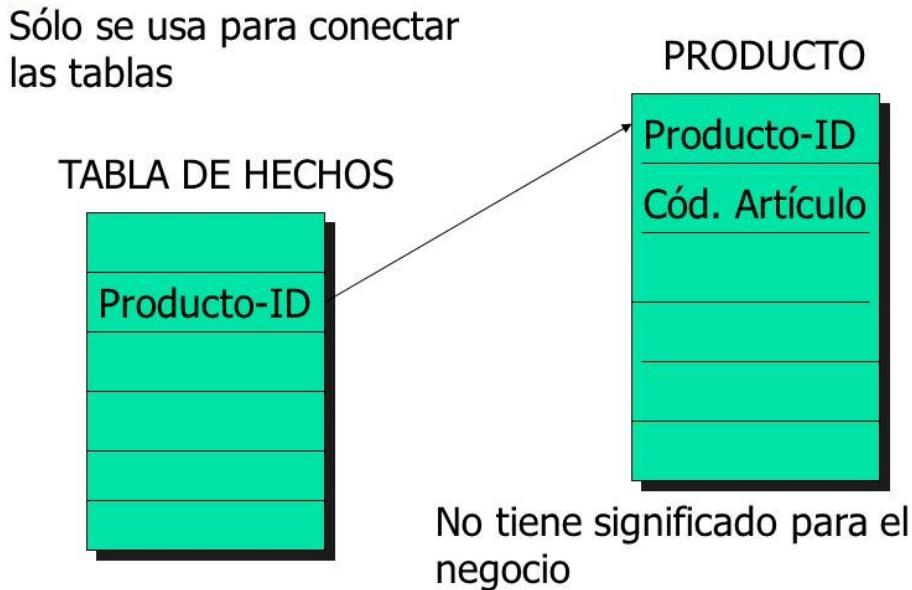


Tabla de Hechos

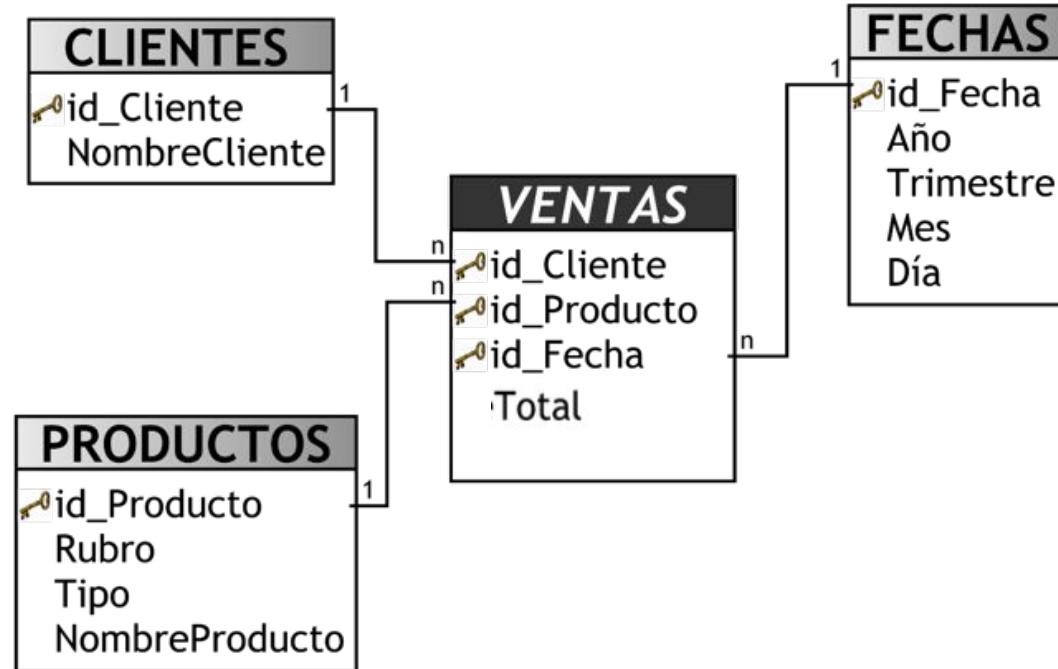
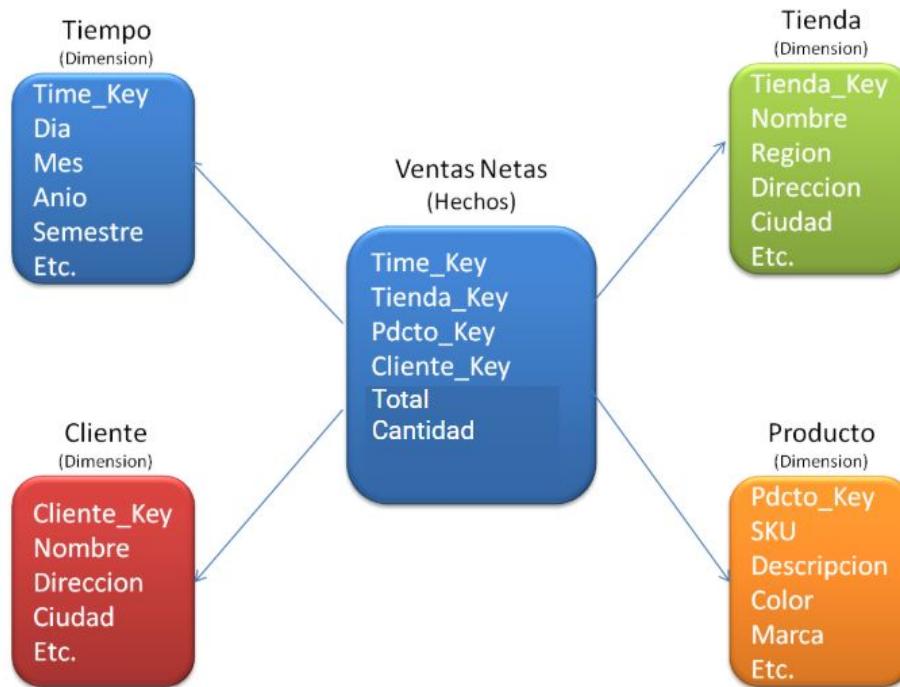


Tabla de Hechos



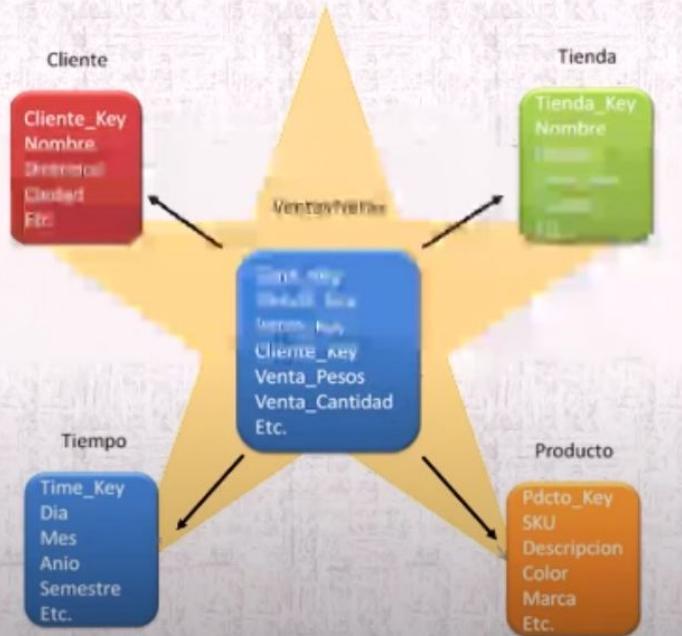
Modelo Dimensional

- Disciplina de modelado de datos alternativa al modelo E-R.
- Con un formato simétrico que permite:
 - Facilidad de entendimiento
 - Eficiencia en consultas
 - Resistencia al cambio

2 tipos de Tablas:

- Hechos
- Dimensiones

- Esquema estrella -> Un diagrama por cada proceso de negocio



Datawarehouse

- Un datawarehouse es una base de datos corporativa o un almacén de datos que tiene como característica la integración y depuración de todos los datos que recogen los diversos sistemas de una empresa.

Datawarehouse

- Cuando se habla de querer implementar una solución fiable de BI(Business Intelligence) el **primer paso** es la creación de un Datawarehouse.

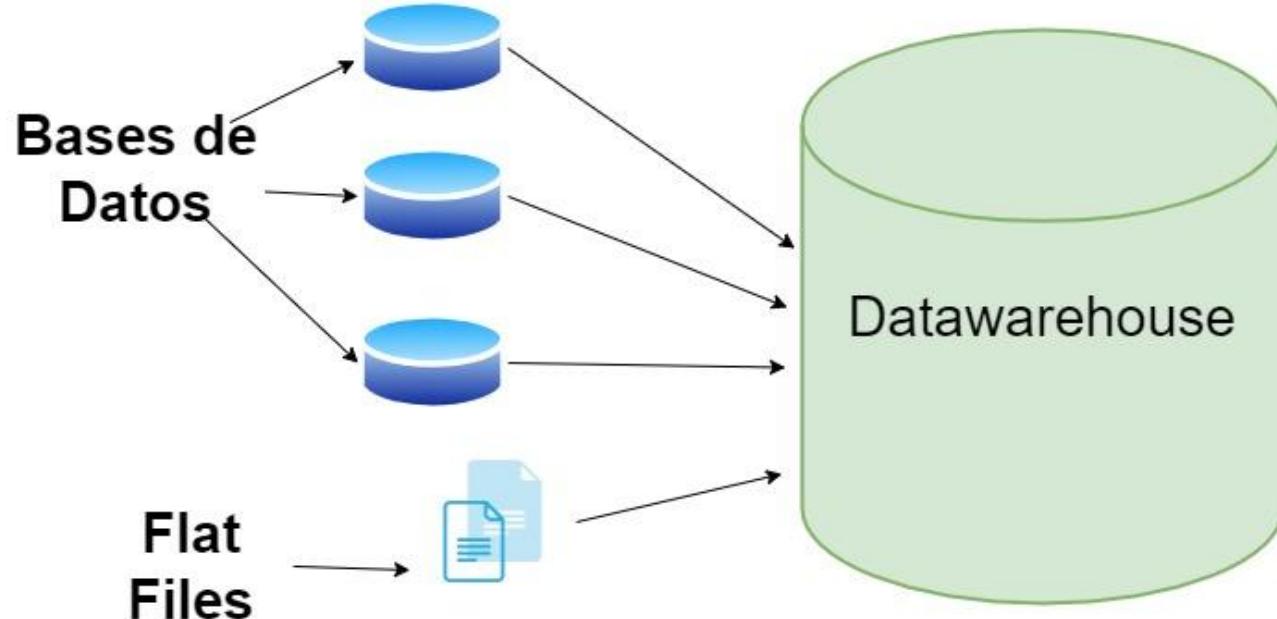
Datawarehouse

- La **función principal** de un datawarehouse es la de contener los datos necesarios o útiles para una organización o empresa y así poder utilizarlos en un futuro para extraer información ventajosa para la compañía y sus clientes.

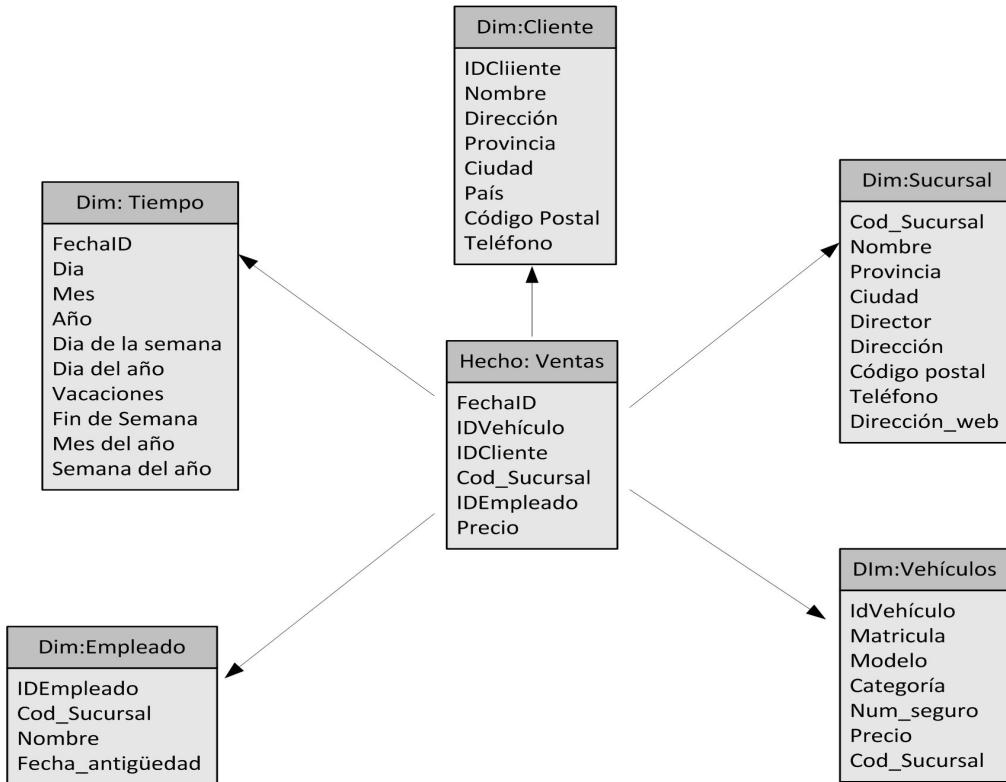
Datawarehouse

- A diferencia de una base de datos que es un mero almacén para el ingreso de datos, un datawarehouse se encuentra especialmente estructurado para favorecer la comprensión y el análisis de los datos

Representación gráfica de un Datawarehouse.



Representación gráfica de un Datawarehouse.



¿Cuándo me interesa implementar un Datawarehouse?

- Si necesito integrar muchas diferentes fuentes de datos casi en tiempo real.
- Si tengo gran cantidad de datos históricos a tratar o debo mantener registros históricos, incluso si los sistemas de transacción de origen no lo hacen.
- Si necesito limpiar o mejorar la calidad de los datos para analizar.
- Si tengo riesgo de que los usuarios puedan provocar errores o pérdidas de datos durante sus consultas.

¿Cómo deben de almacenarse los datos en un Datawarehouse?

- De forma segura.
- De forma fiable.
- Fácil de recuperar
- Fácil de administrar

Ventajas y Desventajas del Datawarehouse

VENTAJAS

- Proporciona una comunicación fiable entre los departamentos.
- El acceso a la información es más rápido.
- Permite conocer en cualquier momento los buenos y malos resultados de la empresa.
- Inteligencia histórica.

DESVENTAJAS

- Requiere mucho mantenimiento transformación y limpieza.
- El coste es alto.
- El diseño es complejo.

Aplicaciones

Predicción de Mercado

- Predecir el flujo de un mercado con información histórica y detección de patrones a través del tiempo.

Análisis de Comportamiento

- Estudiar y clasificar el comportamiento de los clientes y negocios de acuerdo a parámetros específicos.

Modelado de Costos y Presupuestos

- Utilizando funciones de agregación y agrupamientos, se pueden analizar los costos de operación para hacer mejoras en el negocio.

Beneficios

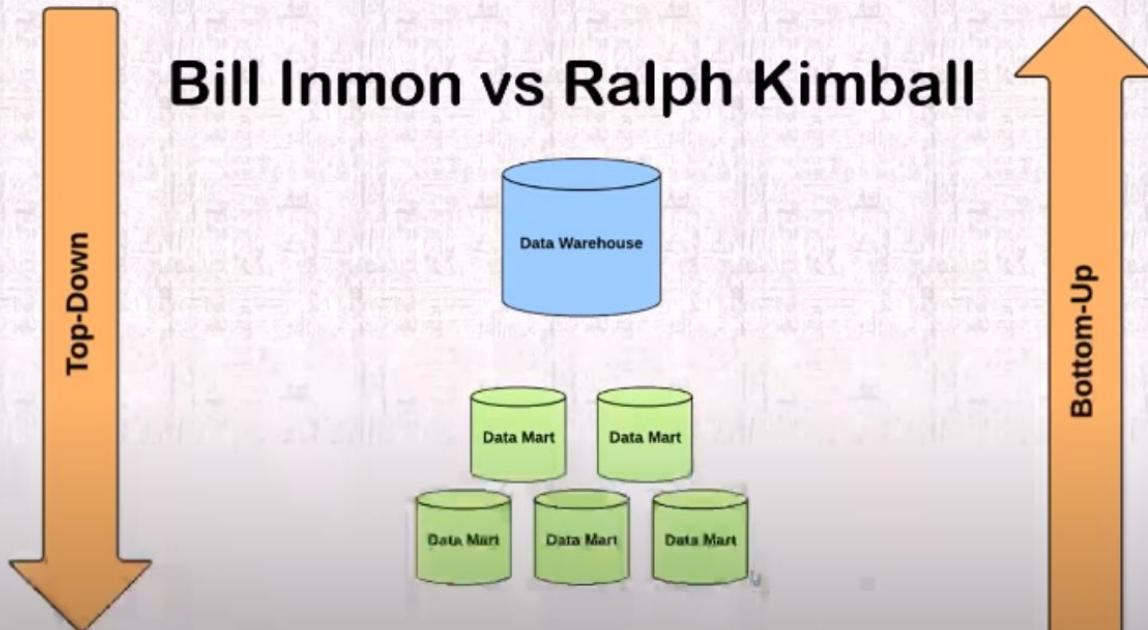
Calidad y
Consistencia
de Información

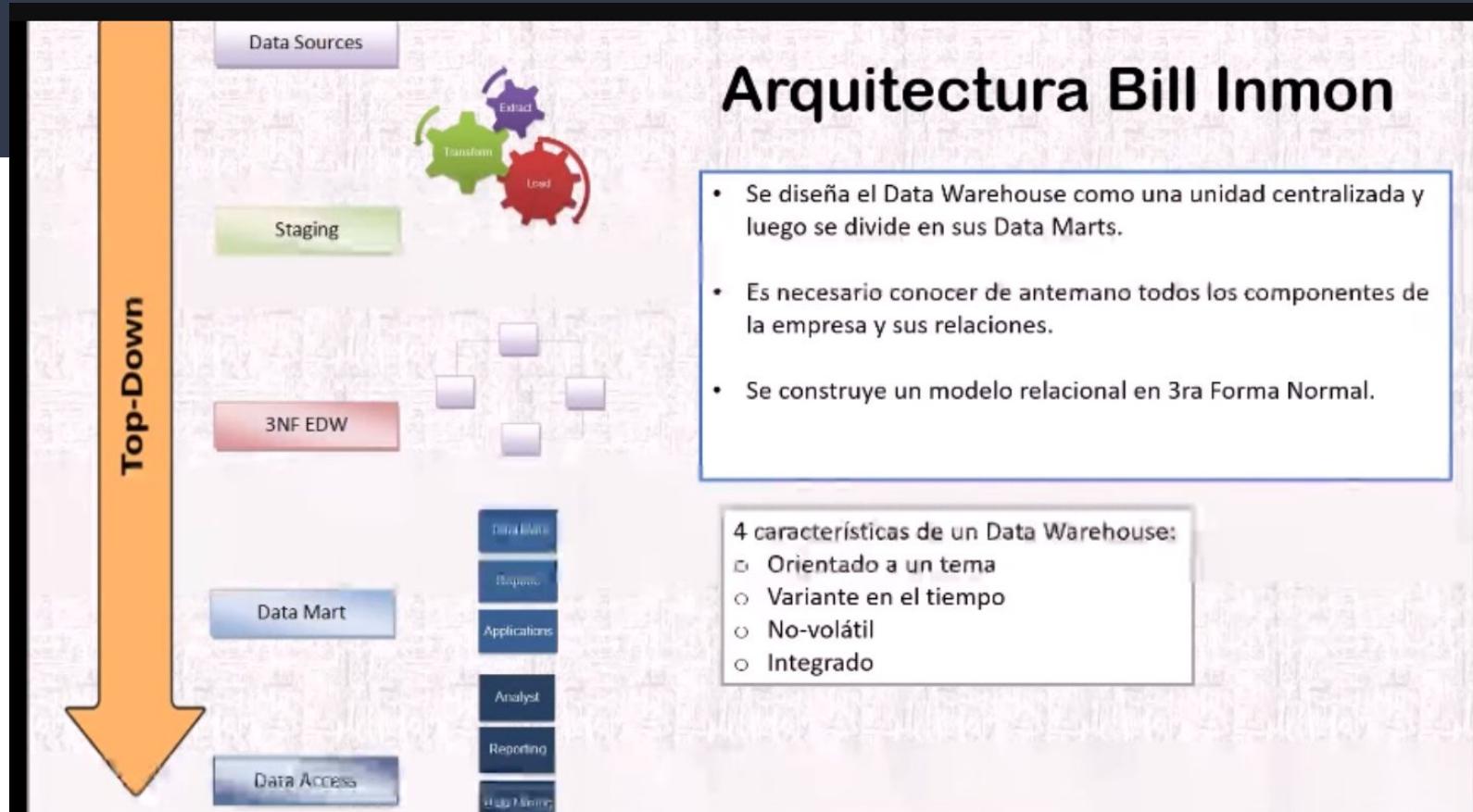
Rapidez de
Respuesta

Visualización
Intuitiva

Arquitecturas

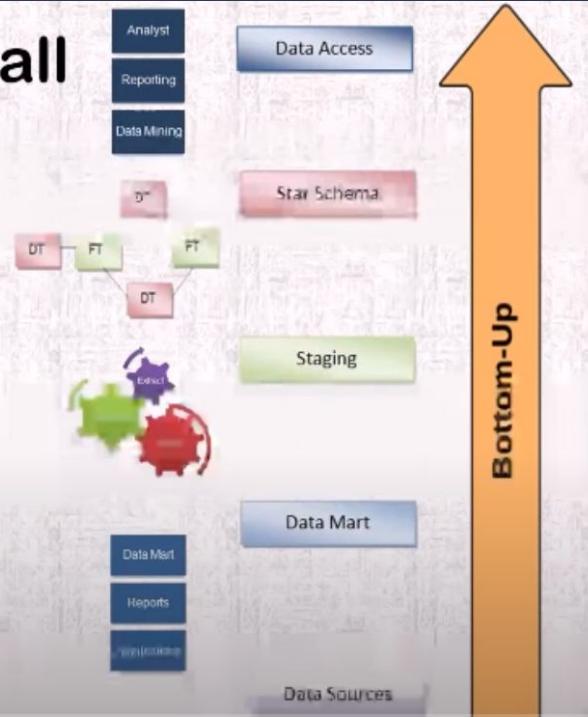
Bill Inmon vs Ralph Kimball





Arquitectura Ralph Kimball

- Diseño modular: Los Data Marts se crean primero, por cada departamento o unidad de negocio.
- Se construye un modelo dimensional de los datos.
- El Data Warehouse se completa por medio de un “bus” que integra los distintos Data Marts como si fueran uno solo.



Bill Inmon (Top – Down)	Ralph Kimball (Bottom – Up)
Mayor tiempo en diseño previo, se necesita conocer toda la estructura de la empresa y sus procesos.	Modular, cada departamento crea su propio Data Mart de forma independiente.
Modelo relacional: 3ra Forma Normal. Parecido a Bases transaccionales. Poco eficiente para análisis.	Modelo dimensional: Esquema estrella, diseñado para análisis de datos y facilidad de lectura.
Datos actualizados de forma continua e integrada.	Datos se actualizan independientes en cada Data Mart, de forma asíncrona.
Proceso de carga y transformación de datos unificado.	Cada Data Mart se encarga de su carga y transformación de datos.
Mantenimiento y escalabilidad complejos.	Mantenimiento depende de cada departamento y se puede escalar agregando Data Marts.

PROCESO DE ETL

¿Qué significa ETL?

E	T	L
Extract	Transform	Load
Extracción	Transformación	Carga

¿Qué es el proceso de ETL?

- Es un proceso mediante el cual nos permite mover datos desde múltiples fuentes (Excel, bases de datos, archivos, Internet) para integrarlos en un lugar, que se sugiere este sea un Datawarehouse.

Extract - Extracción



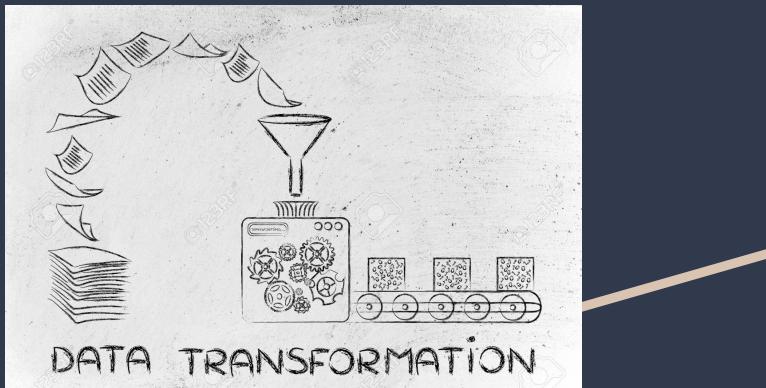
- Es la primera fase del proceso de ETL, en esta se obtiene la “materia prima” en este caso la data desde las distintas fuentes que se proporcionen.
- Esta data es con la que se trabajara en las siguientes dos fases.

Extract - Extracción



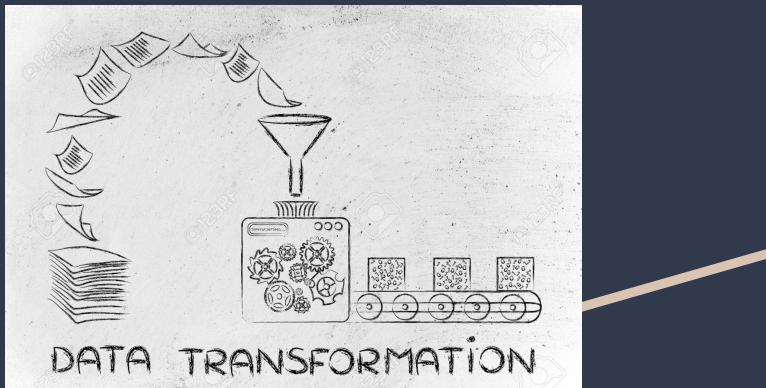
- El volumen de datos extraídos, así como el intervalo de tiempo entre extracciones, depende de las necesidades y requisitos del negocio.

Transform - Transformación



- Es la fase más crítica ya que es la que lleva más trabajo para realizar ya que la data que se trae de la **Fase 1** necesita ser limpiada, mapeada y transformada.
- Esta fase es clave ya que agrega valor y cambia los datos para que tengan sentido y puedan ser utilizados para generar informes.

Transform - Transformación



- Cuando se realiza la transformación se debe mantener la integridad de los datos al realizar operaciones como:
 - Validacion
 - Calculos
 - Filtrado
 - Remocion de duplicados.

Load - Carga



- En esta **fase 3** se llega al objetivo final que es la carga de datos en la base de datos del **Datawarehouse**.
- En caso de fallas, se deben contar con mecanismos de recuperación.

Load - Carga



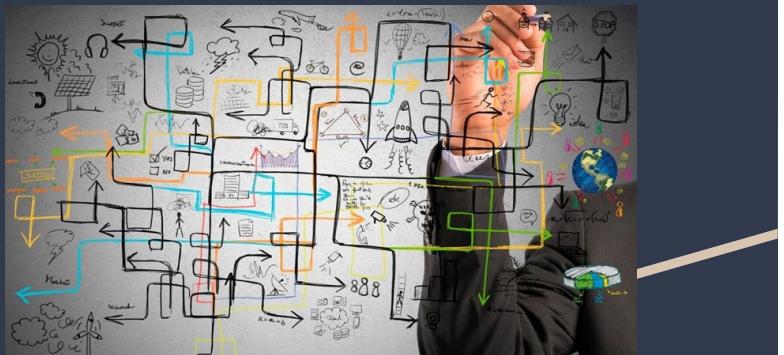
- La carga de datos debe ser de forma consistente en el Datawarehouse de destino.

Características del proceso ETL

Cada empresa llega a tener diferentes datos y necesidades distintas, pero hay características comunes en todo proceso de ETL:

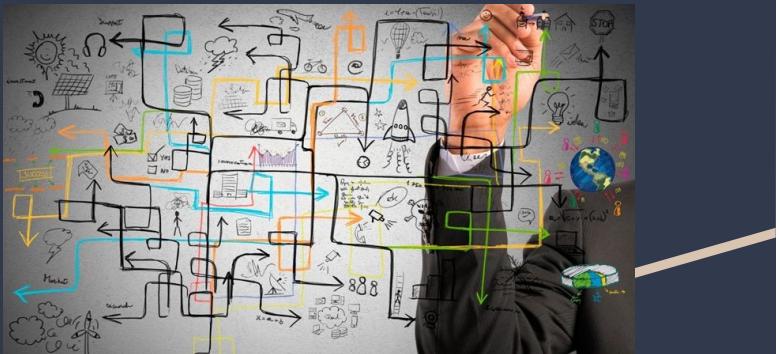
- Complejidad
- Continuidad
- Criticidad

Complejidad



- Las empresas pueden contar con grandes cantidades de datos almacenados por años y generadas por distintos departamentos, repartidos en distintas fuentes como:
 - Bases de Datos.
 - Archivos de texto
 - Flat files
 - Excel
 - CVS

Complejidad



- Extraer, realizar el tratamiento y consolidar toda esa información es una tarea **bastante compleja**.

Continuidad



- Para poder contar con análisis precisos, vamos a necesitar mantener el Datawarehouse constantemente actualizado ya que pueden agregarse nuevas fuentes o nuevos datos a las fuentes.

Continuidad



- Por esto, es importante que el proceso de ETL se realice cada cierto tiempo en intervalos regulares, para detectar dichos cambios, extraer los nuevos datos, transformarlos y cargarlos al Datawarehouse.

Criticidad



- Generalmente los datos que se poseen en las empresas no vienen por defecto en una forma en la cual se puedan usar para la resolución de problemas del negocio.

Criticidad



- Sin los procesos de ETL, las empresas pueden llegar a encontrarse con una cantidad de datos muy grande que **no se puede** llegar a utilizar.

Ventajas y Desventajas del proceso de ETL.

VENTAJAS

- Permite extraer y consolidar datos de múltiples fuentes.
- Permite adaptar e integrar nuevas fuentes de datos.
- Facilita el análisis y el reporte de datos de forma sencilla.

DESVENTAJAS

- Alto coste inicial.
- Se requiere un nivel avanzado de conocimientos para las herramientas.
- El mantenimiento tiene que ser constante.

Utilidades del proceso ETL

- Mover datos de una o múltiples fuentes.
- Formatear datos y realizar limpieza cuando esto sea necesario.
- Una vez alojados en el destino(Datawarehouse) se pueden analizar los datos según las necesidades de la empresa.

Desafíos del proceso ETL

- Procesamiento de datos en tiempo real.
- Aumentar la velocidad del procesamiento de datos.
- Integración de nuevas fuentes de datos.

Procesamiento de datos en tiempo real.



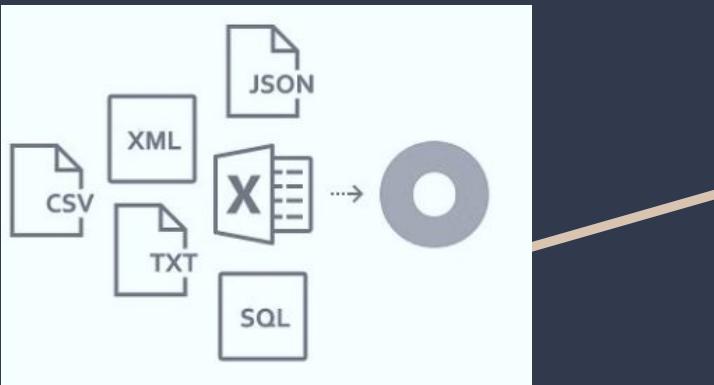
- Cada dia se necesita mayor velocidad para la toma de decisiones, el proceso de ETL tiene que adecuarse para poder operar lo más cercano posible al tiempo real.

Aumentar la velocidad del procesamiento de datos.



- El aumento de cantidad de datos como de complejidad en los datos, puede llegar a dificultar la tarea de transformación.

Integración de nuevas fuentes de datos.



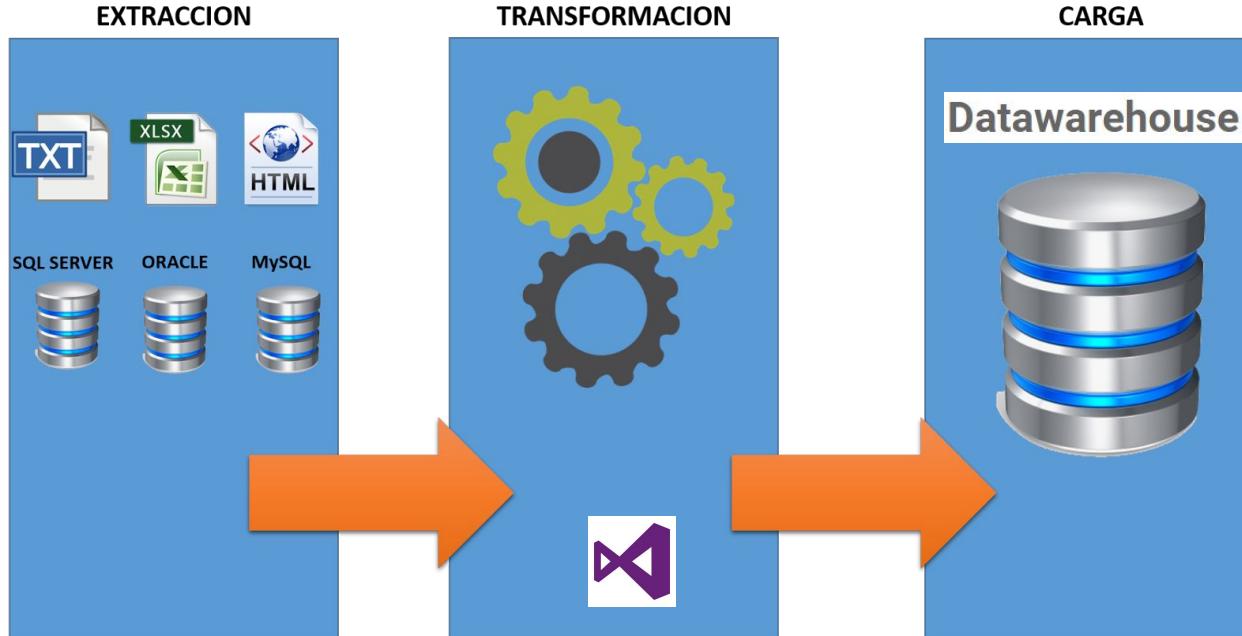
- El proceso de ETL necesita evolucionar para soportar nuevas fuentes de datos en cualquier momento.

Herramientas de ETL

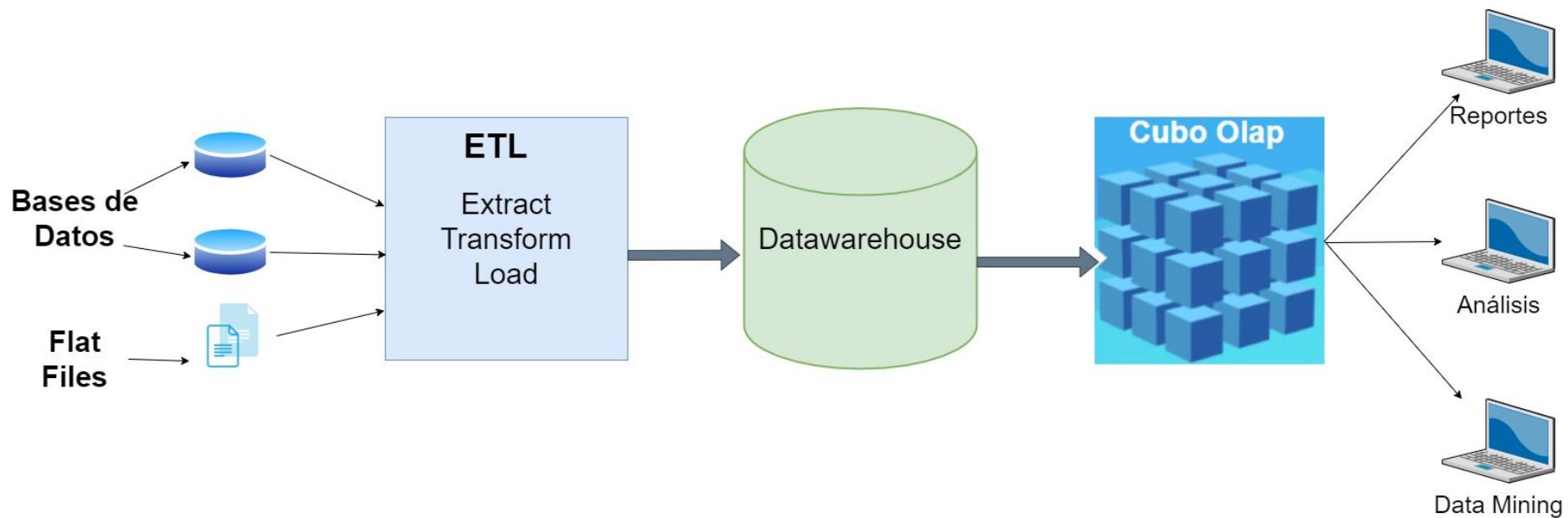
Hoy en día las principales herramientas para realizar ETL son:

- IBM InfoSphere DataStage
- Oracle Data Integrator.
- ***Microsoft SSIS.***
- Informatica PowerCenter.
- Pentaho Data Integration. (Open Source)

Representación gráfica del proceso de ETL



¿Qué hemos visto hasta ahora?



¿Dudas o Preguntas?



Tarea 1.

Ingresar a <https://www.kaggle.com/datasets>

Escoger un DataSet con más de 1000 datos y hacer la
limpieza de los datos o transformación



Nombre de la actividad:	Hoja de Trabajo 1
Cantidad de participantes:	60
Doy fe que esta actividad está planificada en dtt (Sí/No):	Sí

Hora de inicio:	18:40
Hora de fin:	19:00
Duración (min):	20 min

Participantes: llenar las siguientes cajas de texto (tomar información del chat del meet)

Luis Fernando Culajay Sandoval 201903838	Luis Diego de Leon Sanchez 201800987	William Alejandro Borrayo Alarcón 201909103
Estanley Rafael Cobar García 201700319	Ronald Geovany Ordoñez Xiloj 201314564	Katerine Adalinda Santos Ramírez 201908321
Angel Oswaldo Arteaga Garcia 201901816	Edwin Antonio Lopez Ordóñez 200313430	Yimmi Daniel Ruano Pernillo 201503470
Pedro Rolando Ordoñez Carrillo 201701187	Jaime Ismael Bellosio García 201325557	José Andrés Morales Calderón 201602754
Joel Estuardo Rodríguez Santos 201115018	Adrian Samuel Molina Cabrera 201903850	Denis Francisco Vasquez Flores 201212808
Gerson Aaron Quinia Folgar 201904157	Diego Manuel Morales Rabanales 201503958	Adrián Byron Ernesto Alvarado Alfaro 201700308
Daniel Eduardo López Alvarez 201700390	Alexandro Provenzale Pérez 201904012	José Francisco Santos Salazar 201643762
	Keila Avril vilchez Suarez	Ana Lucia Morales Gonzalez



Día, Fecha:	Lunes, 06/02/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Clase 3

Agenda 06/01

- Anuncios
- Datamart
- Resolución hoja de Trabajo 1
- Datawarehouse vs Datamart
- Modelos de datos
 - ◆ Estrella.
 - ◆ Copo de nieve.
 - ◆ Constelación.
- Practica 1
- Tarea 2
- Ejemplo Práctico

Anuncios

- Fechas de entrega
- Notas | Hoja de Trabajo 1, Tarea 1

Hoja de Trabajo 1.

Identifique las dimensiones y los hechos de las siguientes gráficas.

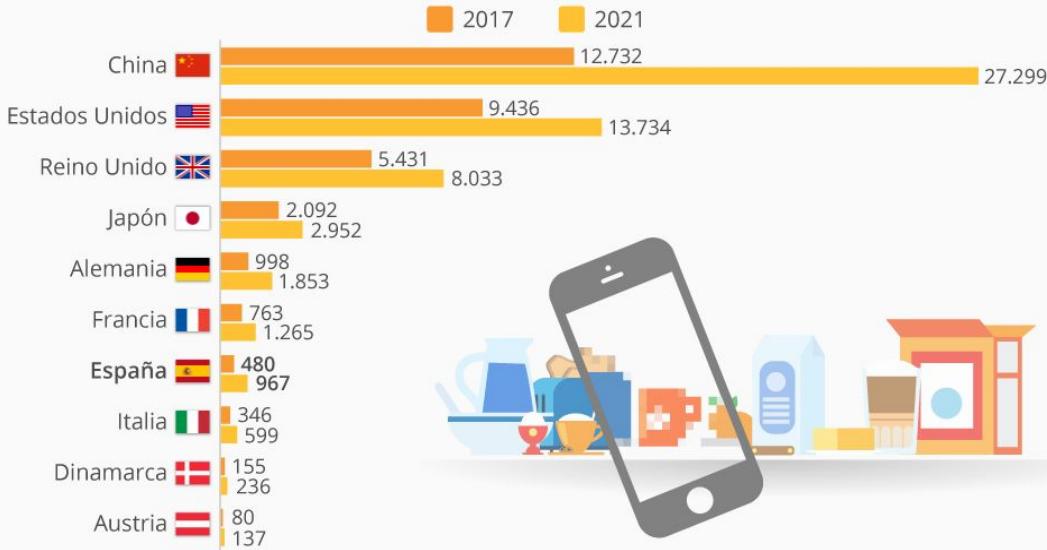
Dimensiones:

- País
- Producto
- Fecha | Año

Hechos:

- Compra
 - Gasto (\$)

Facturación de compras de alimentos, bebidas y productos de uso personal (en \$).



Hoja de Trabajo 1.

Identifique las dimensiones y los hechos de las siguientes gráficas.

Dimensiones:

- Tienda
- Producto
- Tiempo (años)

Hechos:

- Venta
 - Dinero

Ventas 2020



Hoja de Trabajo 1.

Identifique las dimensiones y los hechos de las siguientes gráficas.

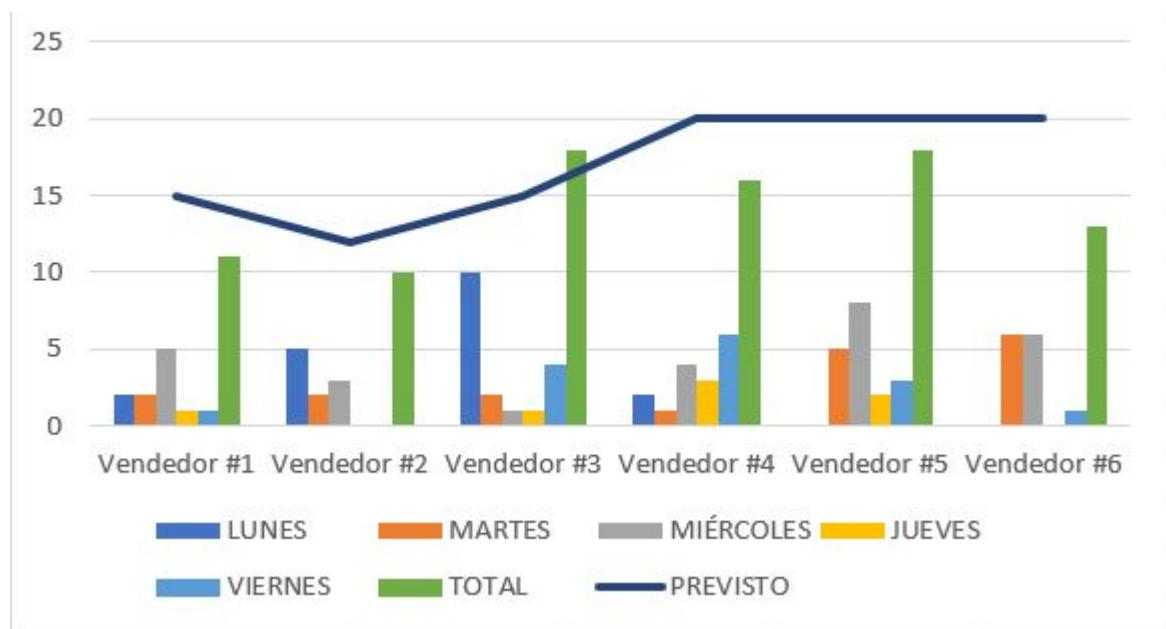
Reporte de ventas semanal por vendedor.

Dimensiones:

- Vendedor
- Fecha -> Dia | Semana

Hechos:

- Ventas total por dia
- Total Previsto



Datamart

- Un datamart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica.

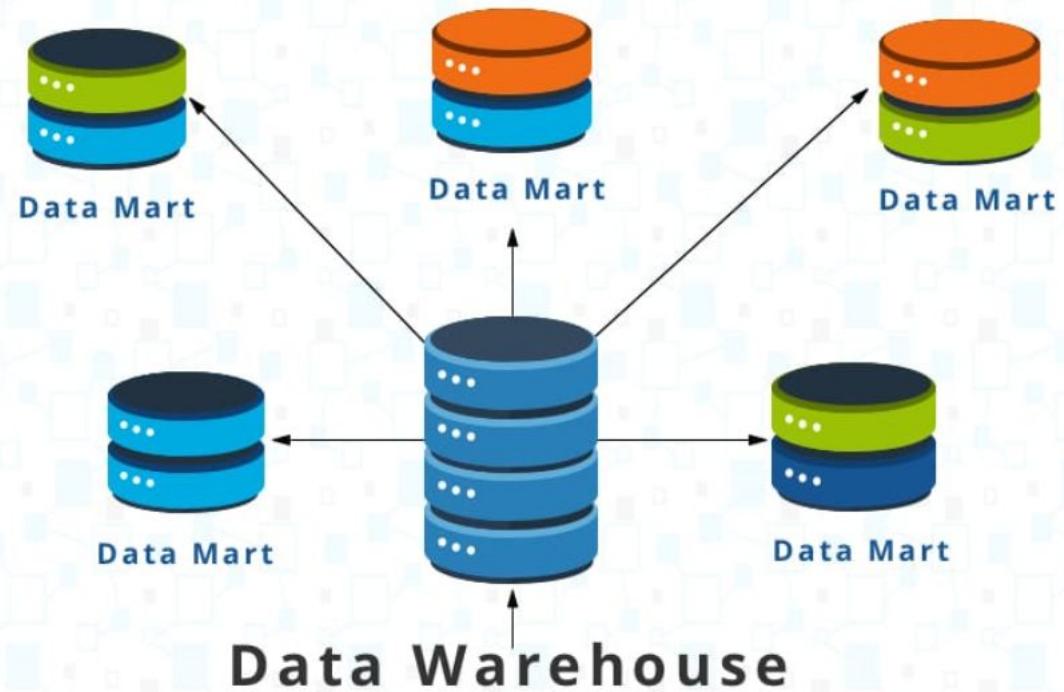
Datamart

- Se caracteriza por disponer la estructura óptima de datos para analizar la información a detalle desde todas las perspectivas que afecten a los procesos de dicho departamento.

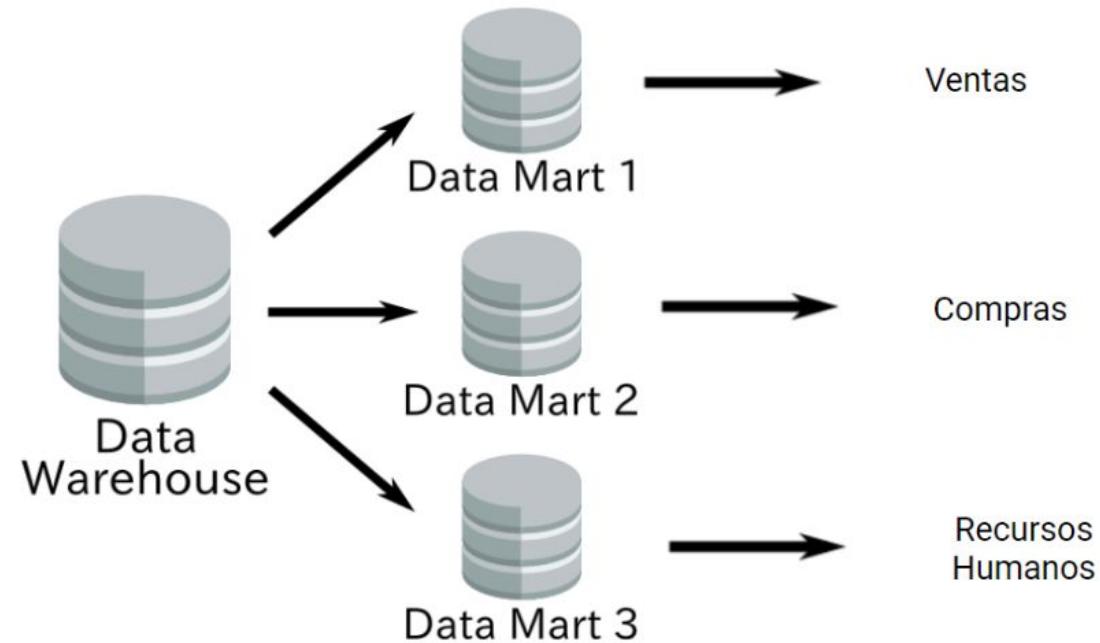
Datamart

- Los datamarts son subconjuntos de los datos del Datawarehouse con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

Representación gráfica de un Datamart.



Representación gráfica de un Datamart.



Datawarehouse y Datamart

- Si nos referimos a un Datawarehouse estamos hablando que este contiene **todos** los datos de una organización.
- Mientras que el Datamart solamente obtiene un **subconjunto** de los datos de una organización, lo que hace centrar lo en un área específica dentro de la organización.

Datawarehouse y Datamart

- Un problema que surge es cuando el **datawarehouse** llega a crecer y a tornarse muy complejo. Debido a esto el rendimiento de las consultas decae y el modelo deja de ser óptimo.
- En estos casos la solución es la creación de **datamarts** especializados por áreas como Ventas, Compras, etc.

Ventajas y Desventajas del Datamart

VENTAJAS

- Consultas más rápidas debido al poco volumen de datos a recorrer.
- Fácil acceso a los datos que se utilizan con frecuencia.
- Su costo de construcción es relativamente menor a la de un datawarehouse.

DESVENTAJAS

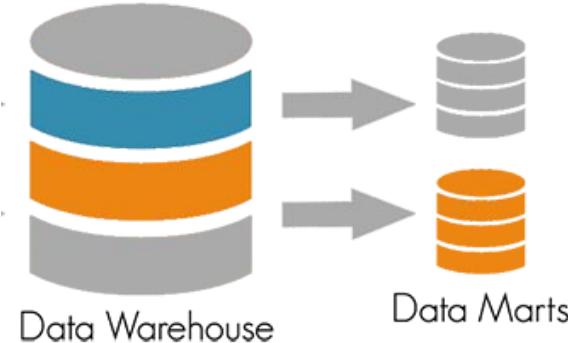
- No maneja grandes volúmenes de información.
- No considera otras fuentes de datos de la empresa.

Tipos de Datamart

- **Dependiente.**
- **Independiente.**
- **Híbridos.**

Tipos de Datamart

- **Dependiente:** se crea a partir de un datawarehouse existente.



Tipos de Datamart

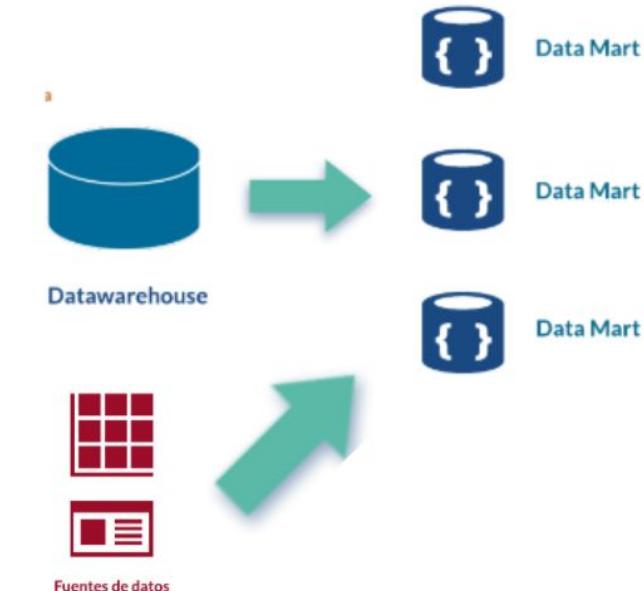
- **Independiente:** es un sistema autónomo que es creado sin utilizar ningun datawarehouse.

Este es conveniente para empresas pequeñas.



Tipos de Datamart

- **Híbridos:** en este tipo combina datos de un datawarehouse con otros sistemas de fuentes de datos.



Datawarehouse vs Datamart

	Datawarehouse	Datamart
Alcance	Almacena información relacionada con todo el sistema.	Se limita a almacenar información de un área de la organización en específico.
Tamaño e integración	Almacena grandes cantidades de datos provenientes de muchas fuentes de datos, por lo que suele ser más grande.	Se concentra en resúmenes de datos totalizados por lo que suele ser más pequeña.

Datawarehouse vs Datamart

	Datawarehouse	Datamart
Creación	La creación es más complicada ya que debe contemplar todos los datos del sistema.	La creación es más simple ya que tiene menos relaciones y están enfocados a sólo un tema.
Costo de manejo	Más costoso, porque requiere más recursos físicos para manejar grandes cantidades de datos.	Es menos costoso ya que requiere menos recurso físicos para manejar los datos requeridos.
Objetivo	Optimizar la obtención de datos, integrando y optimizando los datos fuente.	Es diseñado para entregar de manera óptima la información para el soporte de decisiones de negocio.

Modelos de datos.

Tipos de modelos

- **Modelo Estrella o Star Scheme.**
- **Modelo Copo de Nieve o Snowflake Scheme.**
- **Modelo Constelación(Copo de Estrellas) o Starflake Scheme.**

Modelo Estrella o Star Schema.

Modelo Estrella

- Es el más sencillo en su estructura, consta con:
 - Una tabla central de Hechos.
 - Varias tablas de dimensión.
- Lo característico de este modelo es que la única tabla que tiene relación con otras tablas es la de hecho.

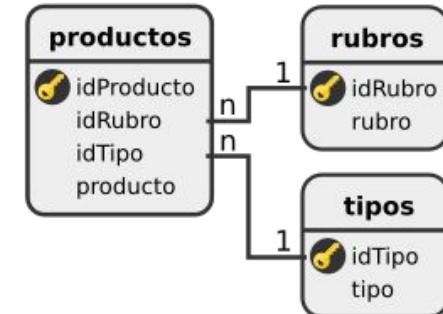
Modelo Estrella

- Las tablas de dimensión sólo están relacionadas con la tabla de hechos.
- Las tablas de dimensión se encuentran desnormalizadas.

Desnormalizado



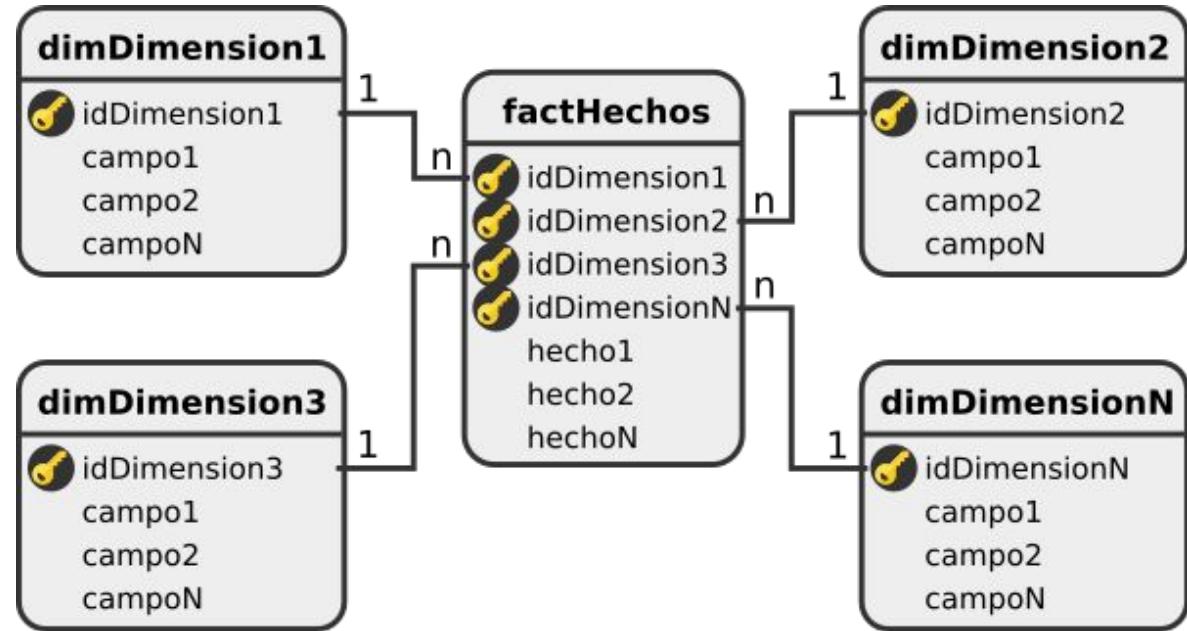
Normalizado



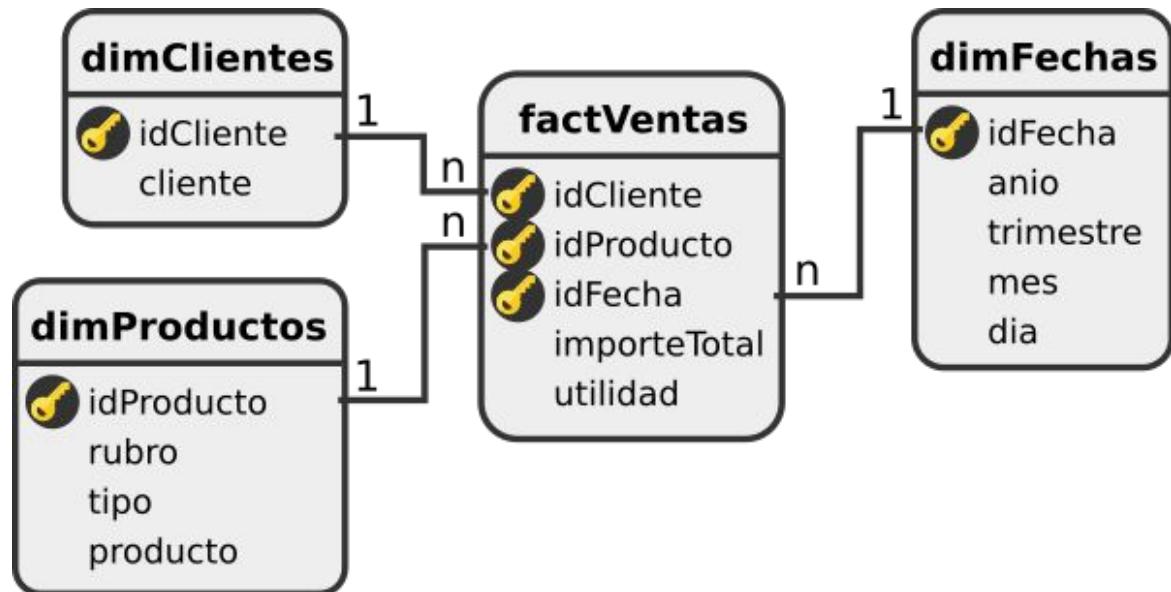
Modelo Estrella

- El esquema estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios.
- Este modelo es soportado por casi todas las herramientas de consulta y análisis.

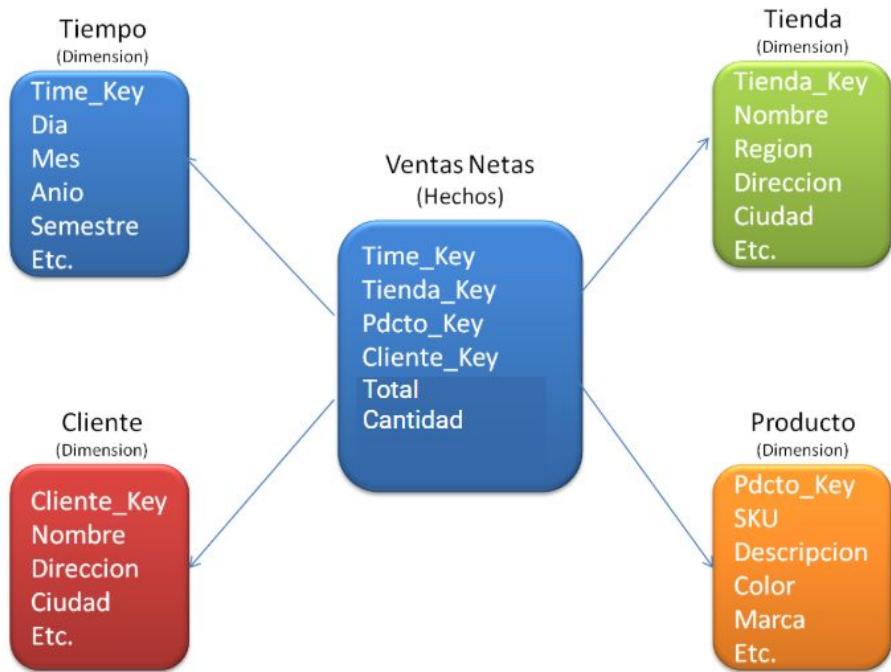
Representación gráfica del modelo Estrella



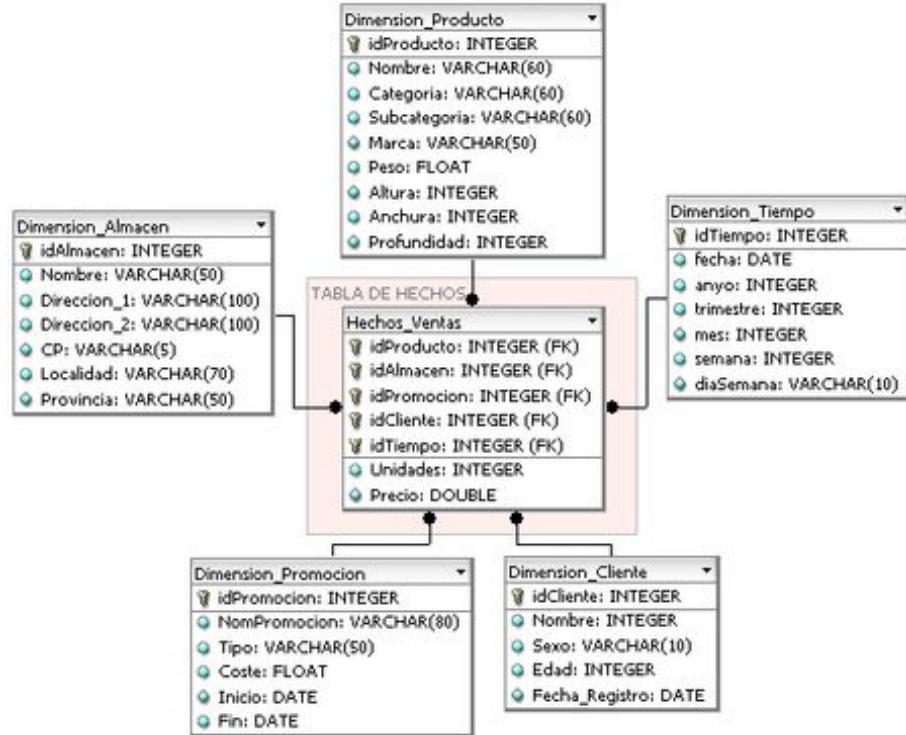
Representación gráfica del modelo Estrella



Representación gráfica del modelo Estrella



Representación gráfica del modelo Estrella



Modelo Copo de Nieve o Snowflake Scheme

Modelo Copo de Nieve

- Es una variación o desviación de un modelo estrella.
- En este modelo la tabla de hechos deja de ser la única relacionada con otras tablas ya que existen otras tablas que se relacionan con las dimensiones.
- Puede implementarse luego de haber desarrollado un Modelo Estrella.

Modelo Copo de Nieve

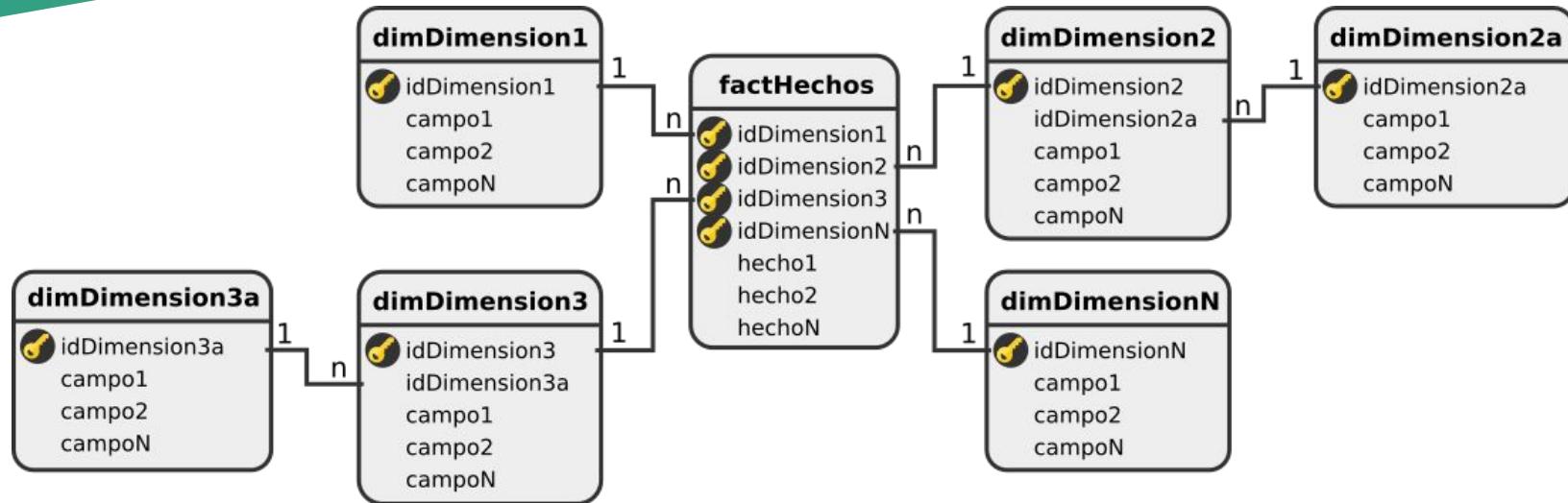
- Existen tablas que no tienen relación directa con la tabla de hechos.
- Este modelo fue creado para facilitar el mantenimiento de las dimensiones.
- La extracción de datos es más difícil y vuelve la tarea de mantener el modelo un poco más compleja.

Modelo Copo de Nieve

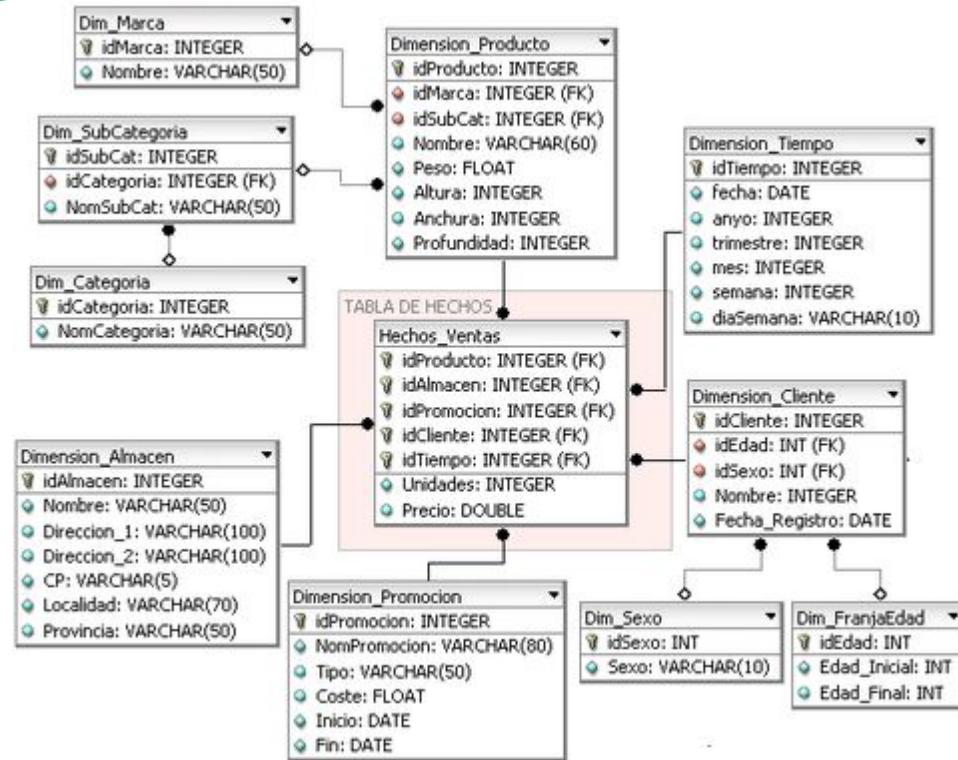
- Su finalidad es normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos.
- Este modelo puede poseer tablas de dimensiones organizadas en jerarquía



Representación gráfica del modelo Copo de Nieve



Representación gráfica del modelo Copo de nieve



Modelo Constelación (Copo de Estrellas) o Starflake Scheme

Modelo Constelación

- Está compuesto por una serie de Esquemas en Estrella.
- Posee lo siguiente:
 - Una tabla de Hechos **principal**.
 - Una o más tabla de Hechos **Auxiliares**, dichas tablas están relacionadas con sus respectivas tablas de Dimensiones.

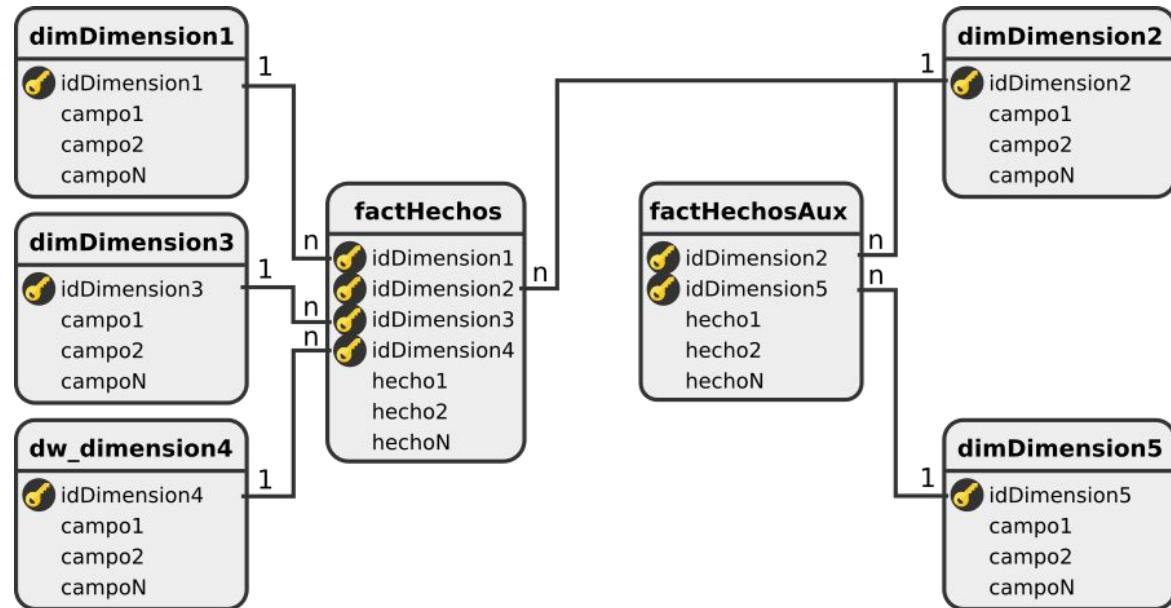
Modelo Constelación

- Las tablas de Hechos Auxiliares pueden vincularse con solo algunas de las tablas de Dimensiones asignadas a la tabla de Hechos Principal, y también pueden hacerlo con nuevas tablas de Dimensiones que se necesiten.

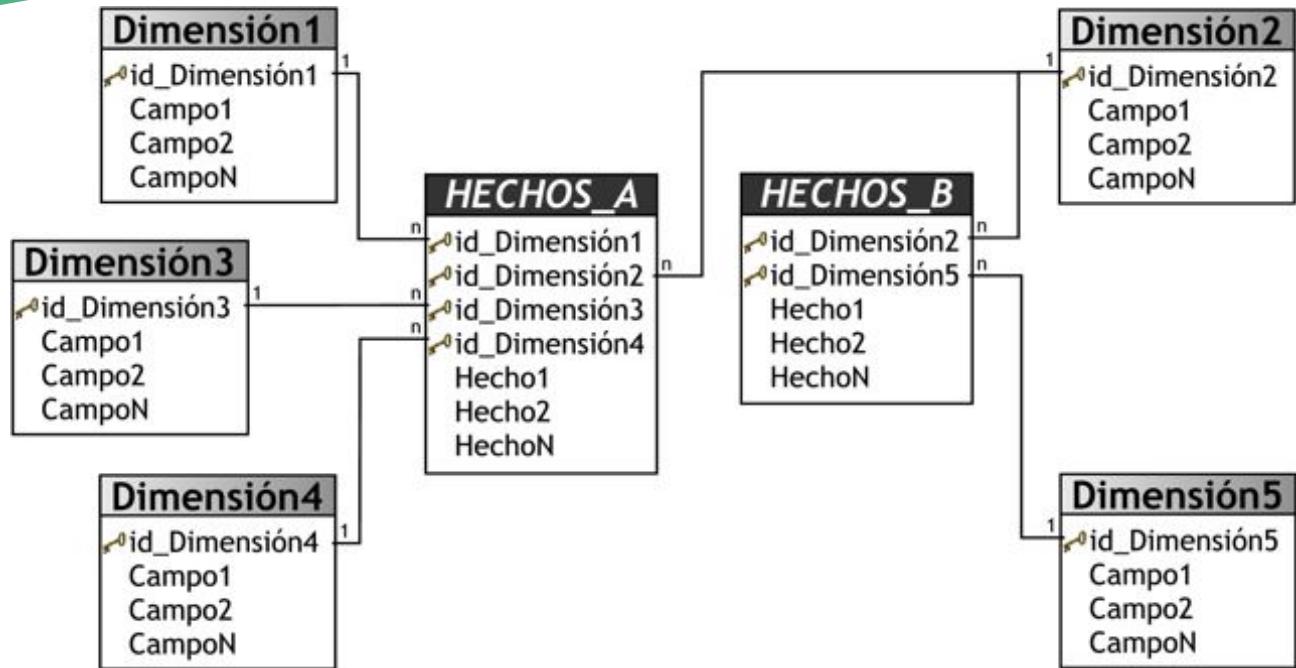
Modelo Constelación

- No es necesario pero se puede dar el caso que las diferentes tablas de Hechos comparten las mismas tablas de Dimensiones.
- Su capacidad analitica es mayor debido a que permite tener más de una tabla de hechos.
- Contribuye a reutilizar tablas de Dimensiones, ya que una misma tabla de Dimensión puede utilizarse para varias tablas de Hechos.

Representación gráfica del modelo Constelación



Representación gráfica del modelo Constelación



Modelo Estrella

vs

Modelo Copo de Nieve

	Estrella	Copo de Nieve
Tablas	Contiene una tabla de hechos rodeada de varias dimensiones.	Contiene una tabla de hechos rodeada de varias dimensiones, que a su vez pueden estar rodeadas de varias dimensiones.
Redundancia	Alta redundancia.	Poca redundancia
Facilidad de uso	Fácil uso.	Difícil de entender, uso mas complicado.

Modelo Estrella

vs

Modelo Copo de Nieve

	Estrella	Copo de Nieve
Joins	Con un solo join es posible relacionar la tabla de hechos y la de dimensiones.	Requiere múltiples joins para hacer los análisis.
Rendimiento de queries	Ejecuciones más rápidas.	Ejecuciones más complejas, debido a cruces.

Modelo Estrella

vs

Modelo Copo de Nieve

	Estrella	Copo de Nieve
Estructura	Descentralizada.	Normalizada.
Diseño de bases de datos	Simple.	Complejo.
Cuando usarlo	Cuando las tablas de dimensión tiene pocas filas.	Cuando las tablas de dimensión tienen un tamaño bastante elevado.

En resumen...



Características - Modelo Estrella

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

Características - Modelo Copo de Nieve

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Es muy útil en tablas de dimensiones de muchas tuplas.
- Las tablas de dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.

Características - Modelo Copo de Nieve

- Si se poseen múltiples tablas de dimensiones, cada una de ellas con varias jerarquías, se creará un número de tablas bastante considerable, que pueden llegar al punto de ser inmanejables.
- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido.

Características - Modelo Constelación

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.

¿Dudas o Preguntas?



TAREA 2

- Instalar SQL Server 2014 o Superior
- Adjuntar un Screenshot con la vista de SQL Server ya instalado en su computadora con Su nombre y carne en Pantalla

Entrega 12/2/2023 23:59

PRACTICA 1



Día, Fecha:	Lunes, 20/02/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Clase 4





Agenda - 20/02/2023

- Avisos
- Componentes SSIS
 - Apoyo
 - Extracción
 - Carga
 - Transformación
- Ejemplo práctico de carga.





Avisos

- Fechas de entrega
- Corto 1
- Tarea 3





Componentes de SSIS para:

- Apoyo
- Extracción
- Transformación
- Carga





Apoyo





Apoyo

Tarea Ejecutar SQL



Tarea Flujo de datos



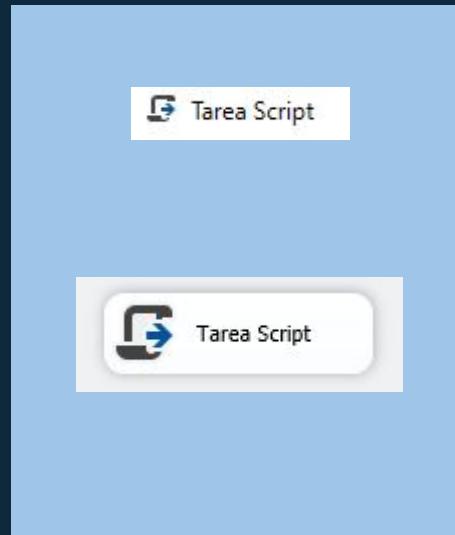
Contenedor de bucles Foreach



Contenedor de bucles
Foreach



Apoyo



Extracción





Extracción

 Origen de ADO NET

 Origen de ADO NET 

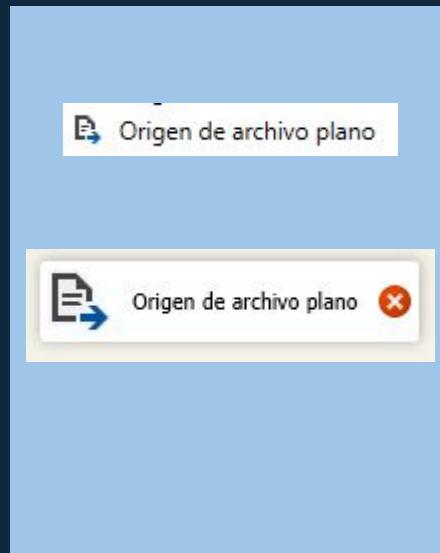
 Origen de OLE DB

 Origen de OLE DB 

 Origen de Excel

 Origen de Excel 

Extracción

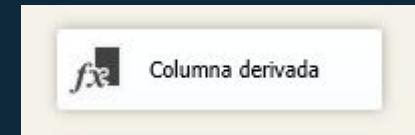


Transformación





Transformación



- Columna derivada:

Este crea una nueva columna dependiendo de la expresión que se evalúe, puede ser desde una concatenación, una suma, hasta una expresión condicional, apoyándonos con las opciones que nos dan en la parte superior derecha.

The screenshot shows the 'Derived Column Transformation Editor' window. On the left, there's a tree view with 'Variables y parámetros' and 'Columnas'. On the right, a large list of functions and operators is displayed under categories like 'Funciones matemáticas', 'Funciones de cadena', etc. Below the tree view, a table lists a single column named 'NuevaColumna' with the expression '<agregar como colum...>' and the value 'col1+col2'. A warning message at the bottom states: 'El componente no tiene ninguna columna de entrada disponible y no se puede configurar correctamente. Para corregir el problema, conecte una salida con columnas disponibles de otro componente a la entrada de este componente.' At the bottom right are buttons for 'Aceptar', 'Cancelar', and 'Ayuda'.

Nombre de columna d...	Columna derivada	Expresión	Tipo de datos
NuevaColumna	<agregar como colum...	col1+col2	



Transformación



- Conversión de datos:

Este componente se basa en realizar conversiones y de ser correctas las mismas crea una nueva columna con la conversión realizada.

1_0 Editor de transformación Conversión de datos

Configure las propiedades utilizadas para convertir el tipo de datos de una columna de entrada a otro tipo. Configure la longitud, la precisión, la escala y la página de códigos de la columna en función del tipo de datos al que se convertirá la columna.

Columna de entrada	Alias de salida	Tipo de datos	Longitud	Precisión	Escala	Página c
Der_Carne	CARNE	entero de cuatro bytes con signo [DT_I4]				
PossibleNota	NOTA	Booleano [DT_BOOL] cadena [DT_STR] cadena Unicode [DT_WSTR] decimal [DT_DECIMAL] entero de cuatro bytes con signo [DT_I4] entero de cuatro bytes sin signo [DT_I2] entero de dos bytes con signo [DT_I2] entero de dos bytes sin signo [DT_UI2]				

Configurar la salida de errores... Aceptar Cancelar Ayuda



Transformación

- **División condicional:**

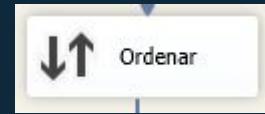
Este componente tendrá dos salidas una que cumpla con la condición que nosotros configuremos y otra en la que no se cumpla. Normalmente se usa para la verificación de datos.

The screenshot shows the 'Editor de transformación División condicional' (Conditional Split Transformation Editor) window. At the top, there's a title bar with the name of the transformation and standard window controls. Below the title bar, a message instructs the user to specify conditions for directing input rows to specific outputs. A tree view on the left lists available functions: Variables y parámetros, Columns, Funciones matemáticas, Funciones de cadena, Funciones de fecha y hora, and Funciones NULL. Under Funciones NULL, several functions are listed, with 'ISNULL(<expression>)' highlighted. A tooltip for this function states: 'Devuelve un valor nulo de un tipo de datos solicitado.' (Returns a null value of the requested data type). On the right side of the editor, there's a table for defining output paths based on conditions. The table has columns for Order, Output Name, and Condition. One row is visible: Order 1, Output Name 'CarneNull', and Condition 'ISNULL(Der_Carne)'. Below the table, there's a field for 'Nombre de salida predeterminado:' (Default output name) set to 'CarneNotNull'. At the bottom of the editor are buttons for 'Configurar la salida de errores...' (Configure error output...), 'Aceptar' (Accept), 'Cancelar' (Cancel), and 'Ayuda' (Help).

Orden	Nombre de salida	Condición
1	CarneNull	ISNULL(Der_Carne)



Transformación



- Ordenar:

Este componente ordenara el conjunto de datos que tengamos de la forma que le indiquemos, en el podremos quitar los elementos repetidos eligiendo las columnas que se quieren verificar.

Editor de transformación Ordenar

Especifique las columnas que se ordenarán y establezca el tipo y el criterio de ordenación. Las columnas no seleccionadas se copiarán sin ninguna modificación.

Columnas de entrada disponibles		
	Nombre	Paso a través
<input type="checkbox"/>	Carne	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Nombre	<input checked="" type="checkbox"/>
<input type="checkbox"/>	LlevaLab	<input checked="" type="checkbox"/>
<input type="checkbox"/>	PossibleNota	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Der_Carne	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	CARNE	<input type="checkbox"/>
<input type="checkbox"/>	NOTA	<input checked="" type="checkbox"/>

Columna de entrada	Alias de salida	Tipo de orden	Criterio de or...	Marcadores de comparación
CARNE	CARNE	ascendente	1	

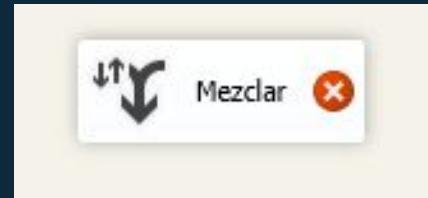
Quitar filas con valores de ordenación duplicados

Aceptar Cancelar Ayuda

Transformación

- **Mezclar:**

Este componente nos permitirá combinar dos orígenes de datos, convirtiendo estos en un solo flujo de datos.

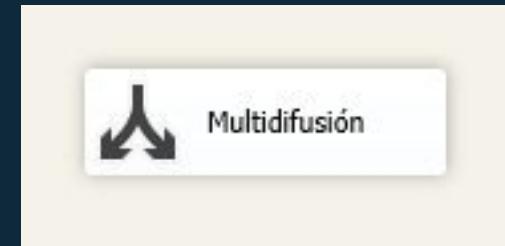




Transformación

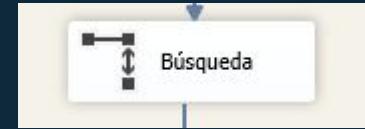
- **Multidifusión:**

Este componente tiene varias salidas a partir de un conjunto de datos distribuye a estas N salidas creando “copias” de estos datos.





Transformación



- Búsqueda:

Componente que nos ayudará a buscar registros desde una tabla para verificar si existen o hacer otro tipo de validaciones con ellos.

* Se recomienda que se use con orígenes y destinos OLE DB.

Editor de transformación Búsqueda

Esta transformación permite ejecutar combinaciones de igualdad simples entre la entrada y un conjunto de datos de referencia.

General
Conexión
Columnas
Opciones avanzadas
Salida de error

Columnas de entrada...

Nombre
Came
Nombre
LlevaLab
PossibleNota
cod_encargado
nombre_encargado
nombre_alumno
CARNE
COD_ENCARGA...

Columnas de búsqueda disponibl...

<input checked="" type="checkbox"/> Nombre	Ind...
<input checked="" type="checkbox"/> came	
<input type="checkbox"/> Nombre	
<input type="checkbox"/> LlevaLab	
<input type="checkbox"/> PossibleNota	

Columna de búsqueda Operación de búsqueda Alias de salida

carne	<agregar como columna nueva>	carne
-------	------------------------------	-------

Aceptar Cancelar Ayuda



Carga



Carga

 Destino de ADO NET

 Destino de ADO NET 

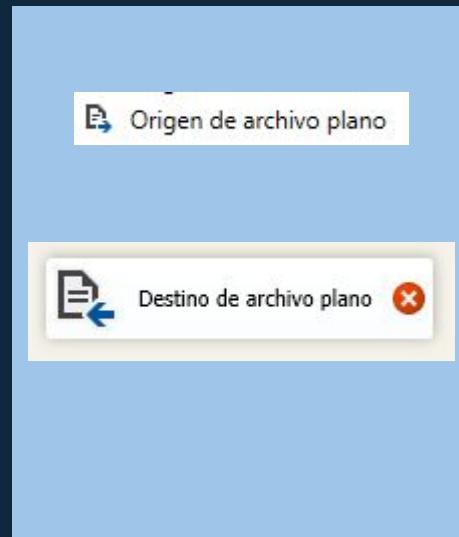
 Destino de OLE DB

 Destino de OLE DB 

 Destino de Excel

 Destino de Excel 

Carga





¿Dudas o Preguntas?





Tarea 3

Instalar las herramientas SSDT de Visual Studio y adjuntar los Screenshots de las instalaciones, la fecha y hora de la computadora deben ser visibles.

Nota:

- ◊ Deben de tener Visual Studio Instalado
 - ◊ Deben de ser las mismas versiones tanto de visual Studio como de la herramienta SSDT
 - ◊ Los 2 deben de estar en el mismo Idioma
 - ◊ Revisar Guía de apoyo proporcionada
- 



Día, Fecha:	Lunes, 27/02/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

clase 5

“

Agenda 27/02:

- Avisos
- Business Intelligence
 - Principales usos
 - Herramientas para BI
- Continuación de ejemplo práctico
- Lectura de la Fase 1 del Proyecto
- Hoja de Trabajo 2

“

**Business
Intelligence**

Business Intelligence

- Se trata de un concepto extremadamente importante para la gestión de proyectos, así como para la productividad de una empresa en general.

Business Intelligence

- Es un conjunto de metodologías y recursos que permiten realizar dos procesos:
 - Transformar datos en información.
 - Transformar información en conocimiento.

Business Intelligence

- A través de los datos que se han ido recopilando en el sistema, las empresas obtienen información valiosa para poder tomar, posteriormente, las decisiones necesarias para mejorar el futuro de la empresa.
- Es algo así como transformar datos primarios en conocimiento.

Principales usos

- Gestión de proyectos.
- Creación de presupuestos.
- Asignación de recursos.

Gestión de proyectos

- Cuestiones relacionadas con la planificación temporal, la organización de recursos o la motivación de los empleados que participarán en un proyecto.

Creación de presupuestos

- A la hora de comenzar un proyecto, la elaboración de presupuestos puede beneficiarse de la aplicación de Business Intelligence siempre y cuando exista una persona experta en el tema.

Asignación de recursos

- Si una persona experta dispone de información acerca de los recursos disponibles para un proyecto, podrá crear un conocimiento que permita asignarlos de la forma más eficaz posible.

Soluciones BI

- Garantiza la mejor forma de explotar y aportar valor a los datos de una compañía de forma que no sólo se trate de tener la información, sino de tenerla en el momento adecuado, en el lugar exacto y en el formato o dispositivo necesario.

Soluciones BI

- Previsión de resultados (forecasting).
- Cubos multidimensionales(OLAP)
- Almacenes de datos especiales (datawarehouse o datamarts)

Casos de BI en la vida real

- Toyota Motor Corporation.
- Cadenas de alimentación.

¿Qué debe tener una solución BI?

- Descubrimiento integral de datos.
- Recibe la historia completa.
- Datos confiables.
- Demanda y eficiencia.
- Flexibilidad y control.
- Rendimiento.
- Extensibilidad(capacidad de crecer).

Herramientas para BI



Tableau



QlikView



Microsoft Power BI



Power BI



Pentaho

Pentaho Report Designer - <Untitled Report>

File Edit View Format Data Window Help

<Untitled Report>

Total Quantity per Year

Revenue Distribution

Quantity Distribution

2003 Comedy 6,401.93 840

2003 Drama 5,172.43 645

2003 Horror 2,063.13 284

2003 SciFi 3,128.56 417

Structure Data

Master Report

- Page Header
- Report Header
 - chart
- Group: year4
 - chart
 - chart
- Details Body
 - Detail Header
 - Details
 - number-field: quantity
 - number-field: revenue
 - string-field: dvd_release_gen
 - string-field: year4
- No Data
- Details Footer

Style Attributes

Name	Value	Expr
common	legacy-chart	
type	field	
value		
name		
if-null		
image-map	query-meta...	
data-format		
style-format		
enable-style-bold		
enable-style-italics		
enable-style-underline		
enable-style-strike-through		
enable-style-font-family		
enable-style-font-size		

1 / 5 2 MB

The screenshot shows the Pentaho Report Designer interface. On the left, there's a toolbar with various icons for file operations, selection, and preview. The main workspace contains a dashboard with three charts: a bar chart titled 'Total Quantity per Year' showing data for Comedy, Drama, Horror, and SciFi genres across years 2003-2007; a pie chart titled 'Revenue Distribution' for the same genres; and another pie chart titled 'Quantity Distribution'. Below these charts is a table with four rows of data. To the right of the workspace is the 'Structure' panel, which displays the hierarchical report structure with nodes like 'Master Report', 'Report Header', 'Group: year4', and 'Details Body'. The 'Data' panel shows the detailed fields used in the charts. At the bottom right is the 'Style' and 'Attributes' panel, where styles for common elements like charts and tables can be defined. The overall interface is dark-themed.

Reporting Services

Report Project1 - Microsoft Visual Studio

FILE EDIT VIEW PROJECT BUILD DEBUG TEAM SQL FORMAT REPORT TOOLS TEST ANALYZE WINDOW HELP

Quick Launch (Ctrl+Q) X

Report Data Server Explorer Toolbox

Graph Sample.rdl [Design] * Matrix Sample.rdl [Design]

Design Preview

Chart Title

Day Type 1
Day Type 2
Day Type 3
Day Type 4
Day Type 5
Day Type 6

Order Day F
Order Day E
Order Day D
Order Day C
Order Day B
Order Day A

Chart Title

0 20 40 60 80

Row Groups Column Groups

Chart Data

- Σ Values
 - UniqueCustomer ($\text{Sum}(\text{UniqueCustomer})$)
- Category Groups
 - OrderDay
- Series Groups

Solution Explorer Team Explorer Class View properties

Ready

<http://www.misjournal.com>

¿Dudas o Preguntas?



“

Ejemplo Práctico

Hoja de Trabajo 2

Realizar Manual De Proceso ETL con un proyecto de Integration Services igual al ejemplo visto en clase, debe ingresar al menos 10 datos y el primer nombre del alumno debe ser el suyo, debe de tener al menos 2 datos nulos en el nombre, y debe de tener al menos 2 archivos de entrada.

Fecha de Entrega 2/3/2023



Nombre de la actividad:	Hoja de Trabajo 2
Cantidad de participantes:	49
Doy fe que esta actividad está planificada en dtt (Sí/No):	Sí

Hora de inicio:	18:50
Hora de fin:	19:00
Duración (min):	10 min

Participantes: llenar las siguientes cajas de texto (tomar información del chat del meet)

ronald geovany ordoñez	xiloj	201314564	Brandon Mauricio Noj Romero	201801028	Cesar Leonel Chamale Sican	201700634
Jorge Isaac Xicol			Estanley Rafael Cóbar García	201700319	Pedro Rolando Ordoñez Carrillo	201701187
Vicente		201807316	Edson Armando Guix Manuel	201701029	Virginia Sarai Gutierrez Depaz	201504443
Katerine Adalinda			Luis Fernando Culajay Sandoval	201903838	Marco Antonio Xocop Roquel	201122934
Santos Ram[irez		201908321	Carlos Antonio Velasquez Castellanos	201403654	Ricardo Antonio Alvarado Ramirez	201603157
ana lucia morales			W Guay Sen Rafael Herrador Reyes	200714200	Jose Augusto Martinez Villegas	201907131
gonzalez		201902207	José Francisco Santos Salazar	201643762	Ariel Rubelce Macario Coronado	201905837
Frederick Jonathan			Ramon Osvaldo Patzan Caballeros		Pablo Fernando Cabrera Pineda	
Faugier Pinto		201602842				
Horacio Ciraiz Orellana		201513758				
William Alejandro						
Borrayo Alarcón		201909103				
Hector Josue Orozco						
Salazar		201314296				



Día, Fecha:	Lunes, 06/03/2023
hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Clase 6

Agenda:

- Avisos
- Modelo Fase 1.
- SSIS, SSAS y SSRS.
- Ejemplo Práctico.
- Hoja de Trabajo 3

Modelo Constelación (Copo de Estrellas) o Starflake Scheme

Modelo Constelación

- Está compuesto por una serie de Esquemas en Estrella.
- Posee lo siguiente:
 - Una tabla de Hechos **principal**.
 - Una o más tabla de Hechos **Auxiliares**, dichas tablas están relacionadas con sus respectivas tablas de Dimensiones.

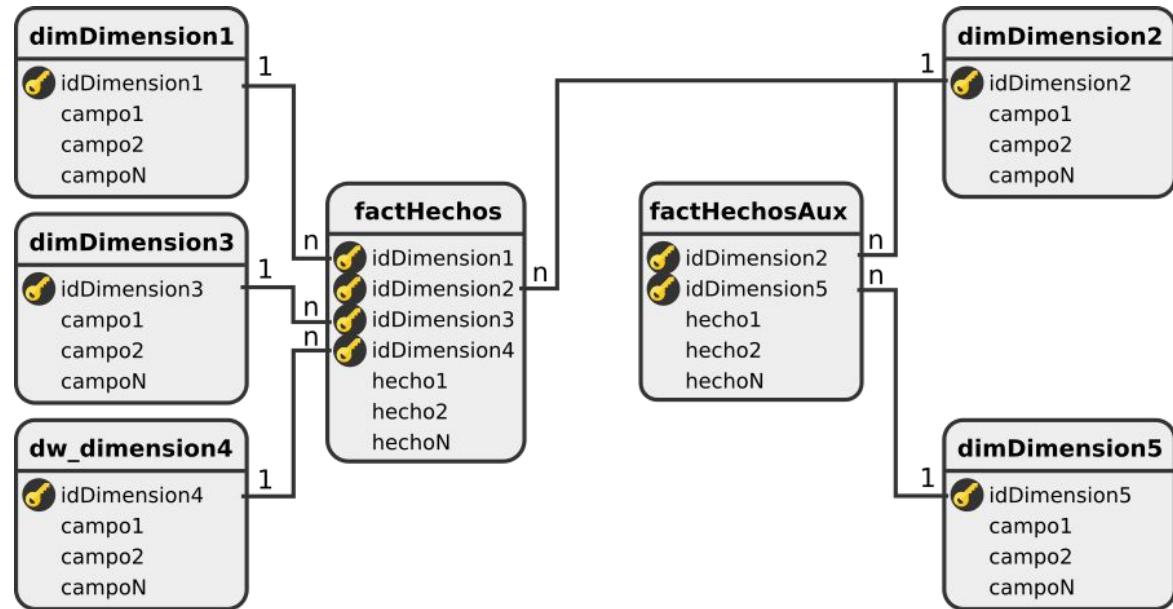
Modelo Constelación

- Las tablas de Hechos Auxiliares pueden vincularse con solo algunas de las tablas de Dimensiones asignadas a la tabla de Hechos Principal, y también pueden hacerlo con nuevas tablas de Dimensiones que se necesiten.

Modelo Constelación

- No es necesario pero se puede dar el caso que las diferentes tablas de Hechos comparten las mismas tablas de Dimensiones.
- Su capacidad analitica es mayor debido a que permite tener más de una tabla de hechos.
- Contribuye a reutilizar tablas de Dimensiones, ya que una misma tabla de Dimensión puede utilizarse para varias tablas de Hechos.

Representación gráfica del modelo Constelación



SSIS

- Integration Services es un componente que permite la migración de datos, permite mover los datos de un origen a un destino sin modificar los datos de origen.
- Tareas de flujos de datos:
 - División condicional
 - Ordenar
 - Agregación
 - Columna derivada



SSAS

- Analysis Services es un producto que cubre el área de Business Intelligence, permite realizar modelos:
 - Multidimensionales
 - Tabulares
 - El modelo multidimensional es el más utilizado y permite realizar modelos complejos y minería de datos
- 



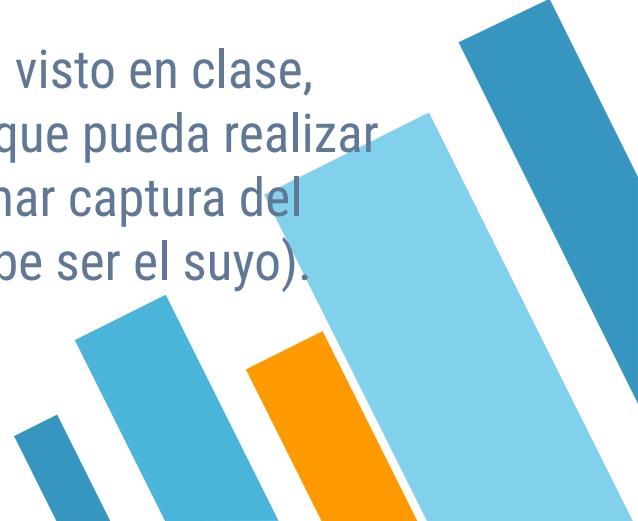
SSRS

- Reporting Services es un servicio que incluyen un conjunto de herramientas que permite el diseño, la administración y publicación de informes y reportes.
 - Como paso siguiente sería el entregarlos a los usuarios adecuados de diferentes maneras, ya sea que los vean en el navegador web, en dispositivo móvil o como un correo electrónico en sitio web.
- 



Hoja de Trabajo 3

Realizar una Investigación Sobre 5 Herramientas BI con sus Características y diferentes usos.



Realizar la Unión de Un Archivo plano diferente al visto en clase, (Se proporcionara el ejemplo visto en clase para que pueda realizar el ejercicio) con otras clases y otros alumnos, tomar captura del resultado.(El primer nombre del archivo plano debe ser el suyo).



Nombre de la actividad:	Hoja de Trabajo 3
Cantidad de participantes:	56
Doy fe que esta actividad está planificada en dtt (Sí/No):	Sí

Hora de inicio:	18:35
Hora de fin:	19:00
Duración (min):	

Participantes: llenar las siguientes cajas de texto (tomar información del chat del meet)

Uzzi Libni Aaron Pineda Solorzano 3006560840101@ingenieria.usac.edu.gt	201403541	Julio Enrique Wu Chiu	201906180	Marco Antonio Xocop Roquel	201122934
Keila Avril vilchez Suarez	201905837	Mike Leonel Molina García	201212535	Pablo Fernando Cabrera Pineda	201901698
Ariel Macario	201701187	Adrian Samuel Molina Cabrera	201903850	Maynor Octavio Piló Tuy	201531166
Pedro Rolando Ordoñez	201904157	Luis Diego de Leon Sanchez	201800987	Ronald Rodrigo Marín Salas	201902425
Gerson Aaron Quinia Folgar	201602952	José Daniel Alvarado Fajardo	201904061	Jeffry Emanuel Mendez Diaz	201901557
Widvin Josué Quiñónez Díaz	201643762	Ramon Osvaldo Patzan Caballeros	201216022		
José Francisco Santos Salazar	201314059	Jorge Isaac Xicol Vicente	201807316		
			201325557		



Día, Fecha:	Lunes, 20/03/2023
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Escarleth Andrea Velasco Campos

Clase 8

Agenda - 20/03

- **Clase 7:**
 - Notas
 - Corto
 - Big Data
 - Hadoop.
- Ejemplo práctico Hadoop.

Hoja de Trabajo 4

BIG DATA

- Se entiende como Big Data a las cantidades de datos a gran escala que llegan a sobrepasar las capacidades del software convencional para ser capturadas y procesadas en un tiempo razonable.

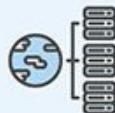
BIG DATA

- Es un término común bajo el que se agrupan toda clase de técnicas de tratamiento de grandes volúmenes de datos, fuera de los análisis y herramientas clásicas.
- Este concepto engloba muchas ideas y aproximaciones, pero todas con un objetivo común: extraer información de valor de los datos, de forma que pueda ser de ayuda para las decisiones y procesos de negocio.

BIG DATA

- ¿Sabían que un avión es capaz de generar aproximadamente 1 terabyte de datos durante todo su recorrido?
- ¿Y cuántos vuelos hay en un día?
- Esto da varios petabytes(1000TB) de información a diario.
- En Twitter, por ejemplo, sólo en un día se generan 9 Terabytes de datos.

Las “cinco V” de Big Data



VOLUME



VARIETY



VELOCITY



VERACITY



VALUE

Las “cinco V” de Big Data

- **Volumen:**
El incremento exponencial de los datos fruto de las nuevas tecnologías y la facilidad de generar datos digitales.
- **Variedad:** La variedad hace referencia a los diversos tipos de datos disponibles.
- **Velocidad:** La velocidad es el ritmo al que se reciben los datos y (posiblemente) al que se aplica alguna acción.

Las “cinco V” de Big Data

- **Veracidad:** Los datos deben de ser confiables. Eliminar los datos tomados de manera correcta y detectar patrones reales es todo un reto del Big Data.
- **Valor:** Es la transformación de los datos en información, convirtiéndose en conocimiento para la toma de decisiones. Los datos y su análisis tienen que generar un beneficio para las empresas.

Herramientas de Big Data



Herramientas de Big Data

- Como el Big Data es algo que no deja de crecer, las herramientas que se usan para gestionarlo evolucionan con él y se perfeccionan permanentemente.
- Se emplean herramientas como **Hadoop**, Pig, Hive, Cassandra, **Apache Spark**, Kafka, etc.,

“ *Tipos de datos
en Big Data*

Tipos de datos en Big Data

- **Datos estructurados (Structured Data):** tienen perfectamente definido la longitud, el formato y el tamaño de sus datos. Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales.
- **Datos no estructurados (Unstructured Data):** Los datos no estructurados se caracterizan por no tener un formato específico. Se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, ficheros multimedia de imagen, audio o video.
- **Datos semi-estructurados (Semistructured Data):** Son una mezcla de los dos anteriores no presenta una estructura perfectamente definida como los datos estructurados pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones, como por ejemplo HTML, XML o JSON.

Tipos de datos en Big Data

Datos estructurados (Structured Data)

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902

Datos no estructurados (Unstructured Data)

CAPÍTULO PRIMERO

Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocin flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda. El resto della concluían sayo de velarite, calzas de velludo para las fiestas con sus pantuflos de lo mismo, los días de entre semana se honraba con su vellori de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rociń como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años, era de complección recia, seco de carnes, enjuto de rostro; gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los autores que deste caso escriben), aunque por conjectura verosímiles se deja entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.

Datos semi-estructurados (Semistructured Data)

```
{  
  "marcadores": [  
    {  
      "latitude": 40.416875,  
      "longitude": -3.703308,  
    },  
    {  
      "latitude": 40.417438,  
      "longitude": -3.693363,  
      "description": "Paseo del Prado"  
    },  
    {  
      "latitude": 40.407015,  
      "longitude": -3.691163,  
      "city": "Madrid",  
      "description": "Estación de Atocha"  
    }  
  ]  
}
```

Hadoop

Hadoop

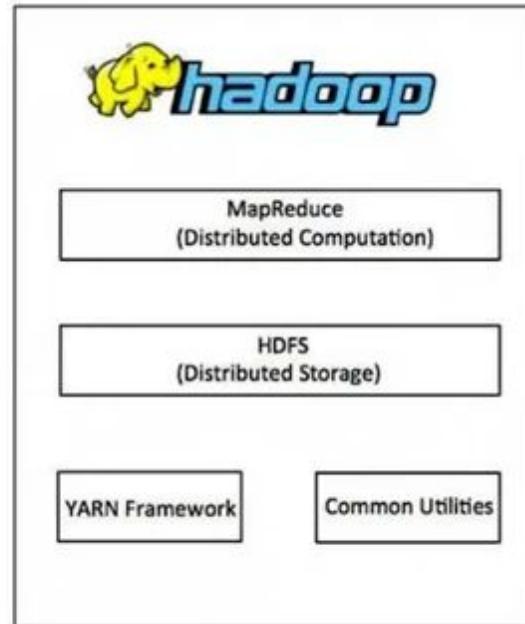
- Hadoop es un framework que puede almacenar grandes cantidades de datos y permitir consultas sobre dichos datos, que se presentan con un bajo tiempo de respuesta.
- Esto lo consigue mediante la ejecución distribuida de código en múltiples nodos.

Hadoop

- **Hadoop** es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos.
- **Hadoop** es utilizado en Big Data para ofrecer capacidades de analizar, descubrir y definir patrones de comportamiento mediante el procesamiento de las grandes cantidades de datos.

Arquitectura de Hadoop

Contiene 4 nodos principales que son:



Arquitectura de Hadoop

- **Common Utilities:** lo forma el hardware y las librerías que son necesarias para ejecutar Hadoop.
- **YARN:** es el gestor de recursos de Hadoop.
- **HDFS:** este es el sistema de archivo distribuido en todo Hadoop.
- **Los procesos MapReduce** son un sistema o una manera de implementar el software que nos permitan parallelizar los datos

HDFS

- **HDFS** (Hadoop Distributed File System) es el componente principal del ecosistema Hadoop.
- Es posible almacenar data sets masivos con tipos de datos estructurados, semi-estructurados y no estructurados como texto, imágenes, vídeo, etc.
- Es un sistema distribuido basado en **Java** que permite obtener una visión de los recursos como una sola unidad.

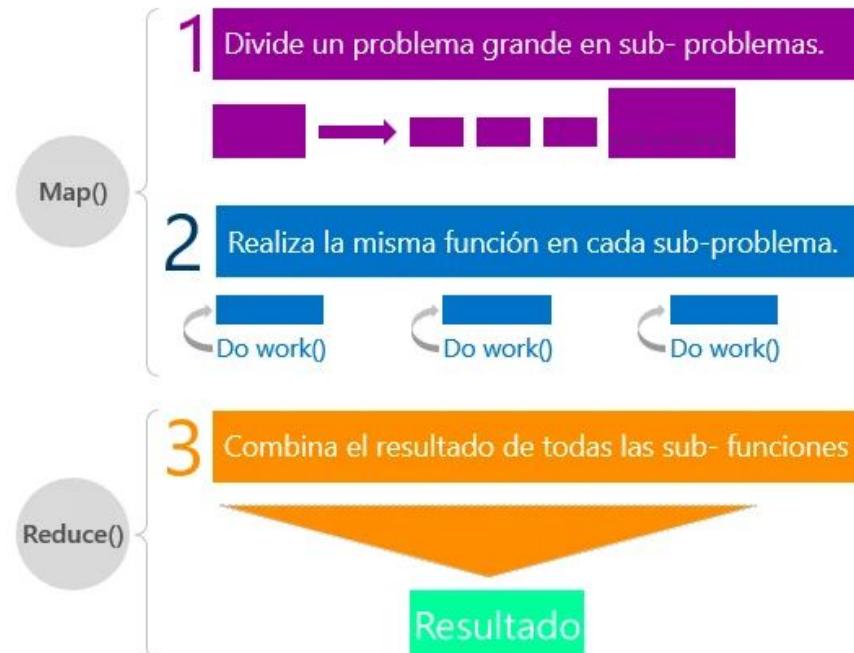
HDFS

- **HDFS** se encarga de almacenar los datos en varios nodos manteniendo sus metadatos.
- Distribuir los datos en **varios nodos** de almacenamiento aumenta la velocidad de procesamiento, el paralelismo en las operaciones y permite la replicación de los datos.

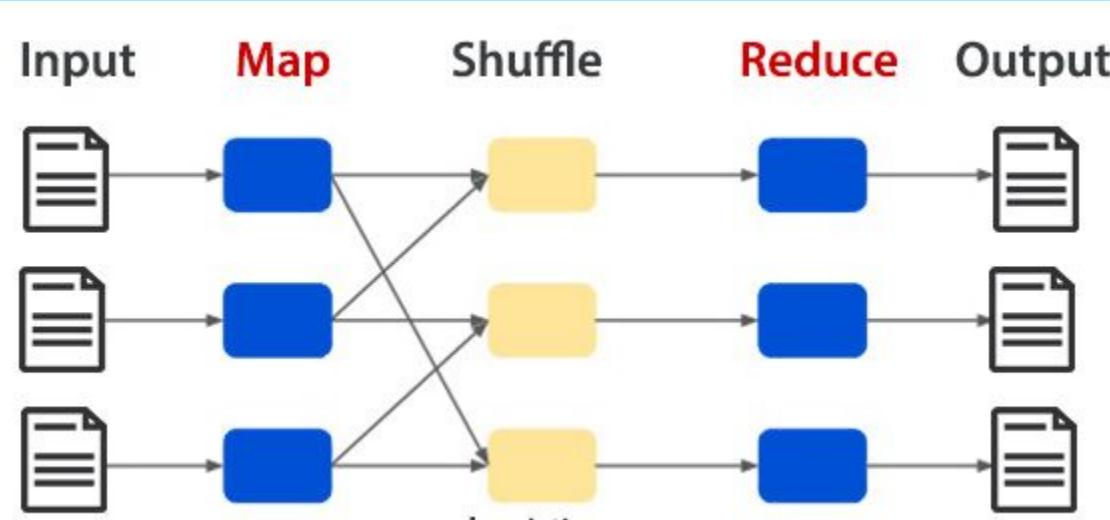
MapReduce

- **Map** – “Divide y vencerás” - divide la tarea de entrada en subtareas y las ejecuta entre distintos nodos.
- **Reduce** – “Combina y reduce la cardinalidad” - la función *Reduce* recoge las respuestas a las sub-tareas en cada subnodo y las combina y agrupa para obtener la respuesta final.

MapReduce

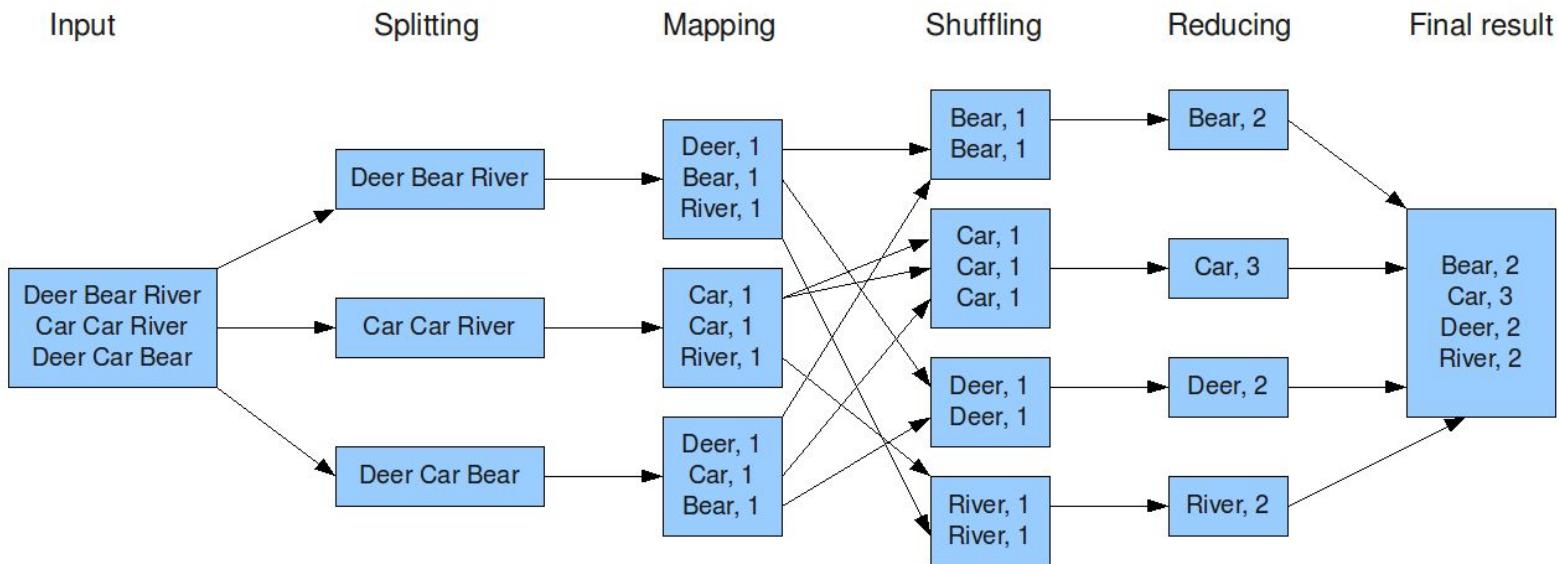


MapReduce



MapReduce

The overall MapReduce word count process



Ventajas de Hadoop

- Aísla a los desarrolladores de todas las dificultades presentes en la programación paralela.
- Permite distribuir la información en múltiples nodos.
- Tiene mecanismos para la monitorización de los datos.
- Dispone de módulos de control para la monitorización de los datos.

Desventajas de Hadoop

- Pocas personas tienen el nivel técnico para implementarlo.
- Actualizaciones constantes, aproximadamente cada semana.
- Su integración no es fácil lo que complica un poco su conectividad.

¿Cómo lo vamos a utilizar nosotros?



Hoja de Trabajo 4

Realizar la Instalación de Hadoop y enviar captura de la instalación.



Nombre de la actividad:

Hoja de Trabajo 4

Cantidad de participantes:

Doy fe que esta actividad está planificada en dtt (Sí/No):

Sí

Hora de inicio: 18:40

Hora de fin: 19:00

Duración (min): 10 min

Participantes: Llenar las siguientes cajas de texto (tomar información del chat del meet)

