

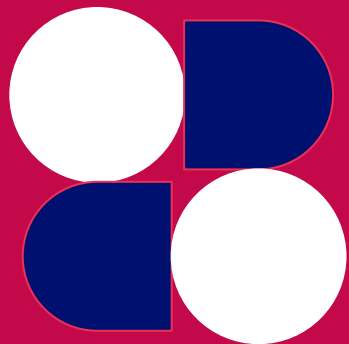


Programa PUCP DATA DRIVEN

Formación tecnológica

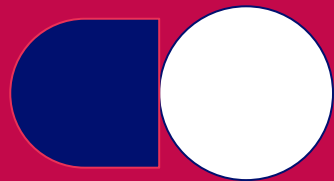
Dirección Académica de Planeamiento y Evaluación
Dirección de Gestión del Talento Humano





SESIÓN 10:

Conceptos para entender el modelado de datos dentro del proyecto de data lake y transformación digital de la PUCP





Al finalizar la sesión, lograrás:

- Comprender la importancia del proceso de transformación digital de la PUCP
- Comprender los conceptos clave en términos de datos para el entendimiento de la transformación digital de la PUCP



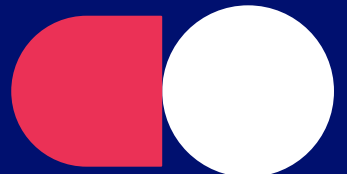
Contenido:



- 1 Data Lake
- 2 Data Warehouse
- 3 Data Flow
- 4 Datamart
- 5 Power BI Services
- 6 Data Mesh
- 7 Gobierno de Datos



1. Data Lakes





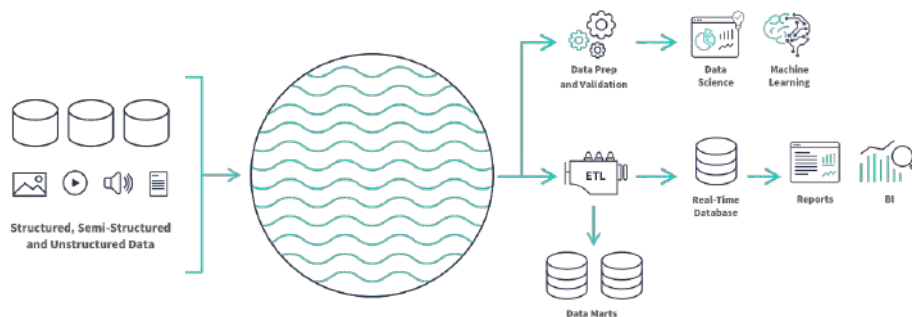
1. ¿Qué es un Data Lake?



Definición:

Un Data Lake es un repositorio de almacenamiento que contiene una gran cantidad de datos en su forma nativa y cruda.

Estos datos se almacenan tal como están, sin necesidad de preprocesamiento o estructuración.





1. Tipos de datos en un Data Lake



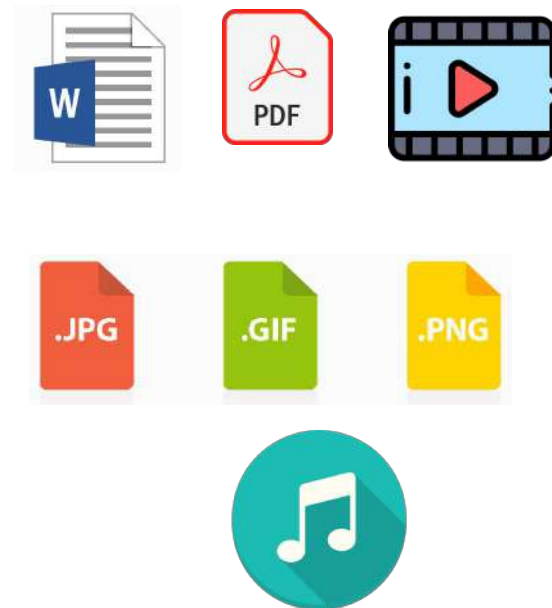
Estructurados:



Semiestructurados:



No estructurados:

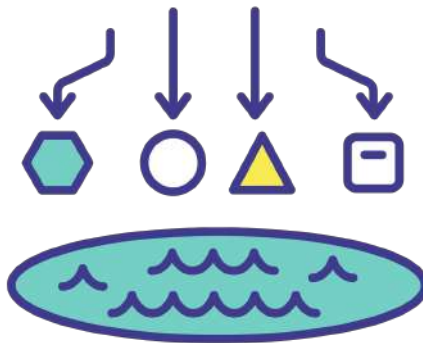




1. ¿Por qué los Data Lakes son importantes?



**Almacenamiento
Escalable y
Económico**



**Flexibilidad de
Datos**

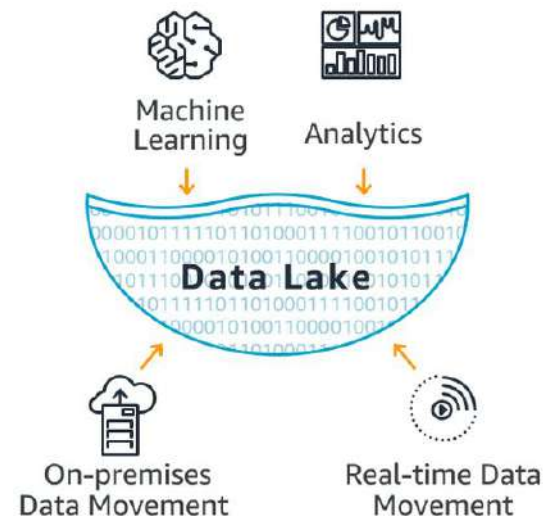


**Facilita el Análisis y la
Innovación**



1. Data Lakes - Características Claves

- **Escalabilidad:**
Capacidad para manejar y escalar con grandes volúmenes de datos.
- **Flexibilidad:**
Almacena todo tipo de datos en su formato nativo.
- **Agilidad:**
Permite un análisis rápido y flexible de datos.
- **Accesibilidad:**
Datos disponibles para ser consultados y analizados por diferentes herramientas y usuarios.





1. Data Lakes - Ventajas y Desventajas

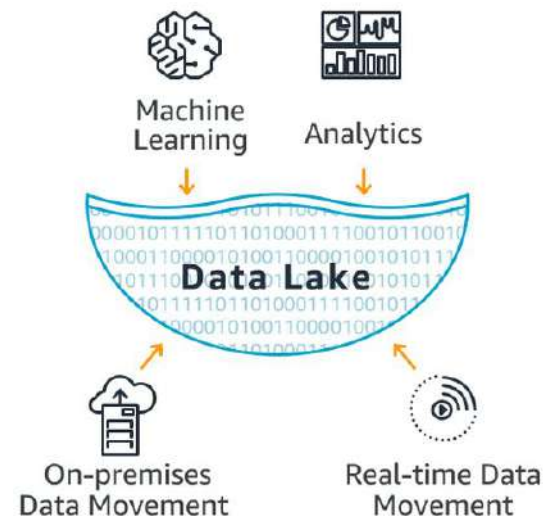


Ventajas:

- **Costos Bajos:** Almacenar grandes volúmenes de datos es más económico.
- **Flexibilidad:** Admite cualquier tipo de datos en su forma nativa.
- **Agilidad:** Permite análisis rápidos y flexibles.

Desventajas:

- **Calidad de Datos:** Sin una gestión adecuada, la calidad de los datos puede ser un problema.
- **Seguridad y Gobernanza:** Necesita estrategias robustas para asegurar y gobernar los datos.





1. Data Lakes - Maduréz

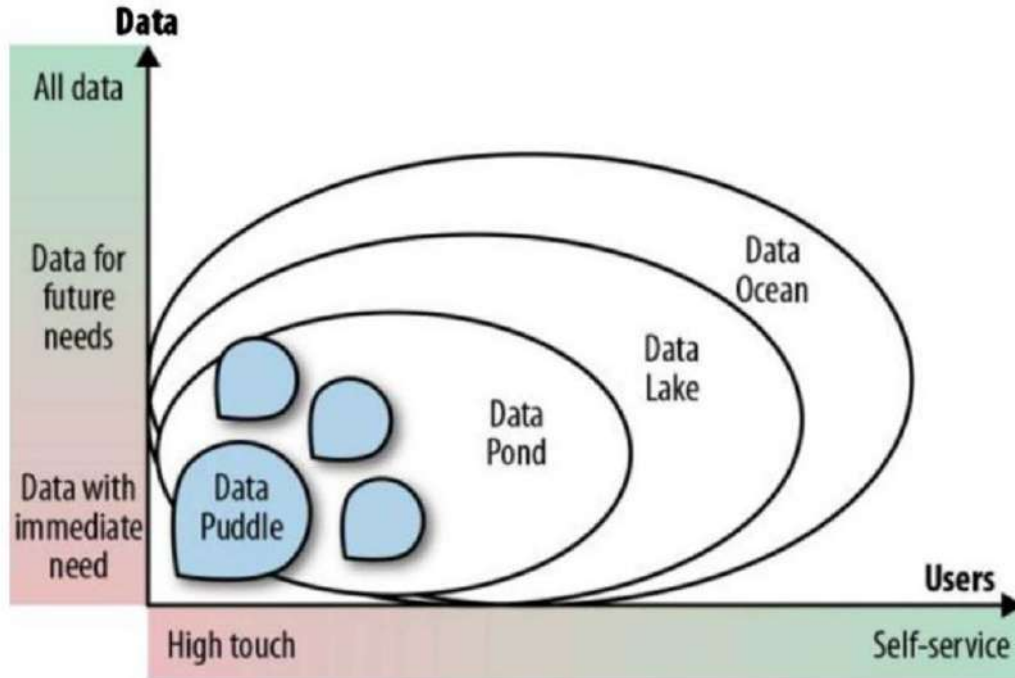


Figure 1-1. The four stages of maturity

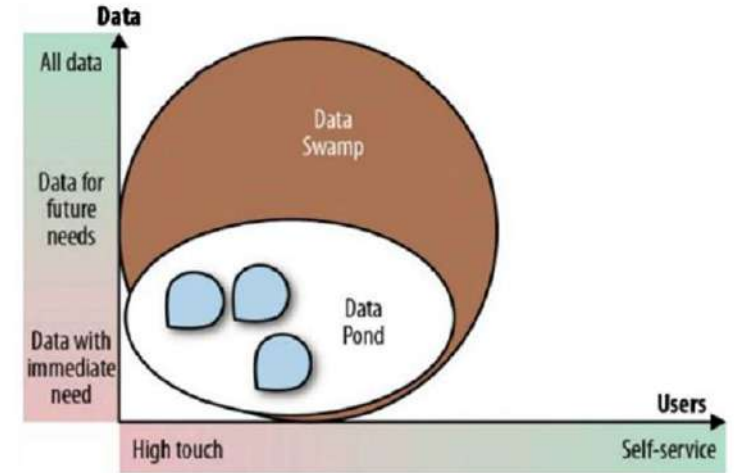
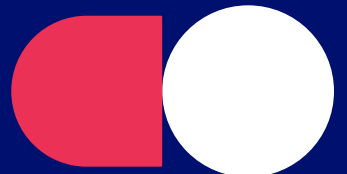


Figure 1-6. A data swamp



2. Data Warehouse





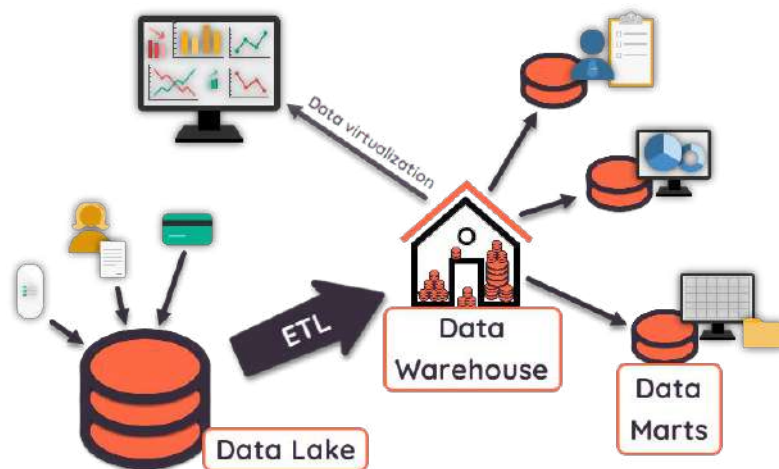
2. ¿Qué es un Data Warehouse?



Definición:

Un Data Warehouse (Almacén de Datos) es un sistema de almacenamiento centralizado que recopila, integra y gestiona grandes cantidades de datos estructurados procedentes de múltiples fuentes.

Estos datos se organizan y estructuran para facilitar el análisis y la generación de informes.





2. ¿Por qué los Data Warehouse son importantes?



**Consolidación
de Datos**



Análisis Eficiente



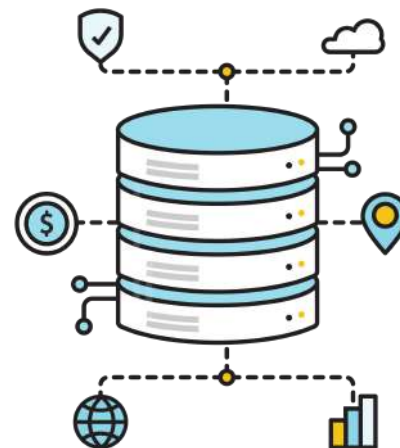
Histórico de Datos



2. Data Warehouse - Características Claves



- **Integración:**
Consolida datos de diversas fuentes en un único repositorio.
- **Consistencia:**
Proporciona datos precisos y coherentes para el análisis.
- **Optimización de Consultas:**
Diseñado para realizar consultas y generar informes de manera eficiente.
- **Historial de Datos:**
Almacena datos históricos para análisis de tendencias y patrones.





2. Data Warehouse - Ventajas y Desventajas



Ventajas:

- **Análisis Rápido:** Optimizado para consultas y generación de informes.
- **Datos Integrados:** Proporciona una visión completa y coherente del negocio.
- **Seguridad:** Mejores prácticas de seguridad y gobernanza de datos.

Desventajas:

- **Costos Elevados:** Puede ser costoso implementar y mantener.
- **Rigidez:** Menos flexible para cambios rápidos en la estructura de datos.



DATA WAREHOUSE



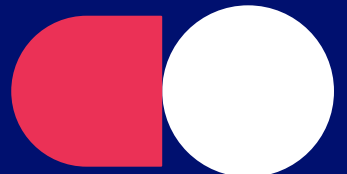
2. Data Warehouse Vs. Data Lake



Característica	Data Warehouse	Data Lake
Datos	Procesados, estructurados	Crudos, en su forma nativa
Estructura	Esquemático, con estructuras definidas	Flexible, sin esquema definido
Costo	Más costoso	Más económico
Usabilidad	Ideal para BI y reportes operativos	Ideal para Data Scientists y Analistas
Velocidad	Alta para consultas y reportes	Alta para ingesta de datos



3. Flujos de datos - Data flows



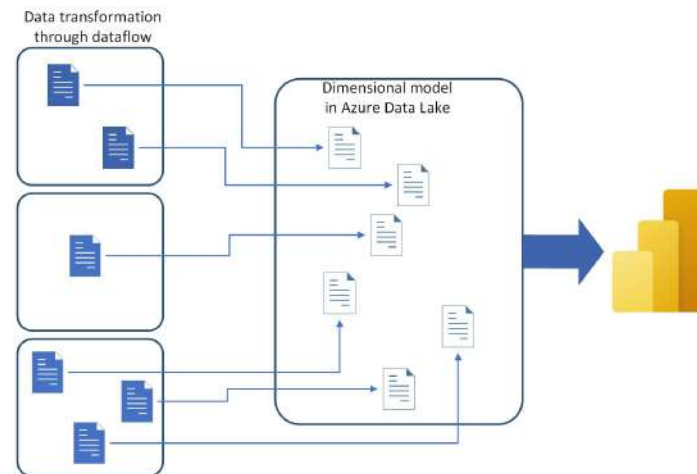


3. ¿Qué es Data flow?



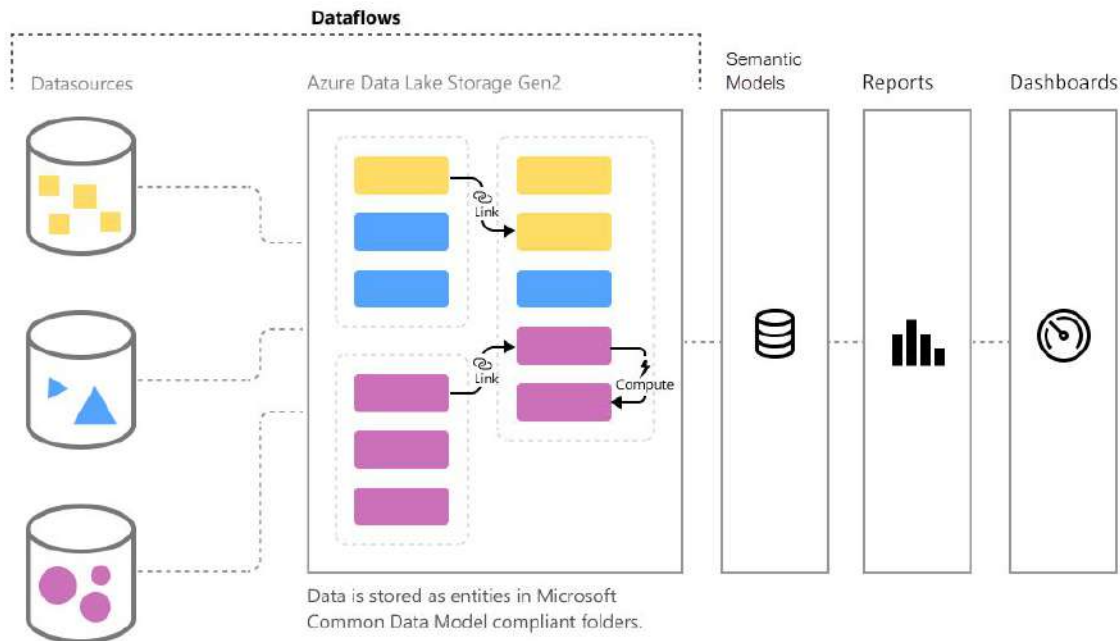
Los Dataflows permiten a los usuarios crear, gestionar y reutilizar conjuntos de datos a través de un enfoque más estructurado y repetible.

Son una colección de tablas, que son creadas y gestionadas en el servicio Power BI. Estas tablas son alimentadas por diferentes fuentes de datos y pueden ser transformadas mediante Power Query, que es la misma herramienta utilizada para la manipulación de datos en Power BI Desktop





3. ¿Qué es Data flow?

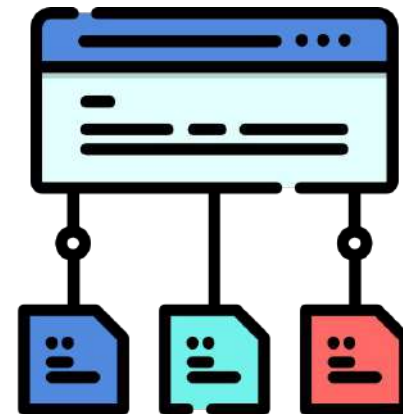




3. Beneficios - Data flows



- **Centralización:**
Cualquier cambio realizado en un Dataflow se refleja automáticamente en todos los informes y dashboards que utilizan ese Dataflow, asegurando consistencia en los datos.
- **Reutilización:**
Permiten reutilizar los conjuntos de datos transformados en múltiples informes y dashboards sin necesidad de repetir las mismas transformaciones en Power BI.
- **Colaboración mejorada:**
Acceso compartido y actualizaciones sincronizadas.
- **Menos Dependencia del Escritorio:**
Los procesos de ETL (Extracción, Transformación, y Carga) se manejan en la nube, reduciendo la carga en el sistema local y permitiendo que los usuarios continúen trabajando en sus dispositivos sin interrupciones.

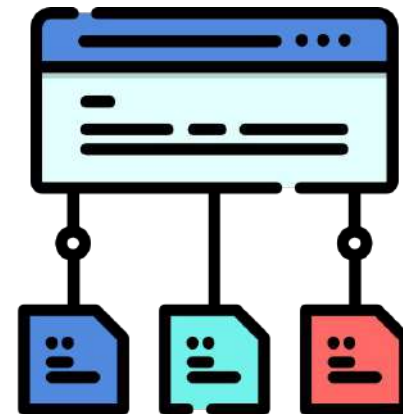




3. Beneficios - Data flows



- **Escalabilidad y rendimiento:**
Los Dataflows aprovechan la capacidad de procesamiento en la nube de Power BI.
- **Mejora del rendimiento:**
Al preprocesar y almacenar los datos en Dataflows, los informes y dashboards pueden cargar más rápido, ya que no necesitan repetir las transformaciones cada vez que se acceden.
- **Gobernanza y seguridad:**
Los Dataflows permiten una mejor gestión del control de acceso y la seguridad de los datos. Los administradores pueden definir quién tiene acceso a qué datos, mejorando la gobernanza de los datos en la organización.





3. ¿Qué es Modelo Semántico en Power BI?



Un modelo semántico es una representación estructurada de datos que define la manera en que los datos se organizan, relacionan y se interpretan. Está diseñado para proporcionar una comprensión clara y coherente de los datos mediante la aplicación de significados y contextos específicos a los datos.

¿Qué es el modelo semántico?



Semántico: los datos tienen un significado para una persona u organización

ProductCode
VS003462

Cantidad
2

Significado: se refiere a un producto específico con características, es decir, motores de nitrógeno

Significado: refleja el desempeño del negocio



Modelo: Los objetos y los cálculos representan un proceso del mundo real

Modelo de venta: Representa el proceso de ventas con datos. Ej: tablas, relaciones y DAX medidas



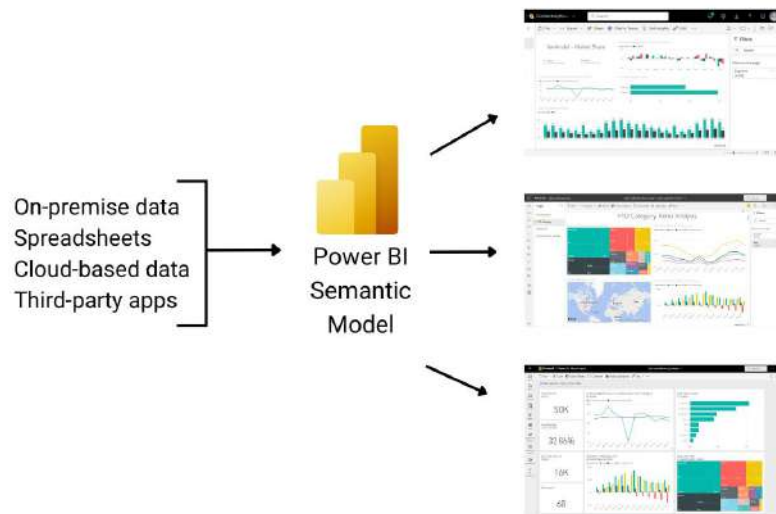
```
SUMX (
    'Sales',
    'Sales'[Quantity]
    * RELATED ( 'Product'[Unit Price] )
)
```



3. ¿Qué es Modelo Semántico en Power BI?



Un modelo semántico en Power BI puede considerarse una capa lógica que contiene las transformaciones, cálculos y relaciones entre las fuentes de datos necesarias para crear informes y cuadros de mando. Un modelo semántico sirve como única fuente de verdad para los informes de toda una organización.



3. Ejemplos de transformaciones de datos

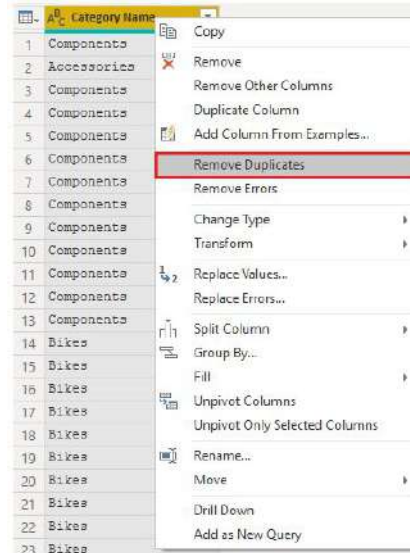


Transform Add Column View Tools Help

Transpose Data Type: Text Replace Values Unpivot Columns
Reverse Rows Detect Data Type Fill Move
Count Rows Rename Pivot Column Convert to List
Split Column Format Merge Columns Extract Parse Statistics

Any Column Text Column

	ID	Subcategory Name	Attribute
1	1	Mountain Bikes	January
2	1	Mountain Bikes	February
3	1	Mountain Bikes	March
4	1	Mountain Bikes	April
5	1	Mountain Bikes	May
6	1	Mountain Bikes	June
7	1	Mountain Bikes	July
8	1	Mountain Bikes	August
9	1	Mountain Bikes	September
10	1	Mountain Bikes	October
11	1	Mountain Bikes	November
12	1	Mountain Bikes	Dezember
13	2	Road Bikes	January
14	2	Road Bikes	February



Category Name
1 Components
2 Accessories
3 Bikes
4 Clothing



3. Ejemplos de transformaciones de datos



Home Transform Add Column View Tools Help

Transpose Data Type: Text

- Decimal Number
- Fixed decimal number
- Whole Number
- Percentage
- Date/Time
- Date
- Time
- Date/Time/Timezone
- Duration
- Text
- True/False
- Binary

Replace Values Unpivot Columns

Fill Move

Pivot Column Convert to List

Any Column

Split Column Format Text Column

Use First Row as Headers

Table

	ABC ID	Category Name	ABC Month	ABC Value
1	1	kes	January	780000
2	1	kes	February	790000
3	1	kes	March	800000
4	1	kes	April	810000
5	1	kes	May	820000
6	1	kes	June	830000
7	1	kes	July	840000
8	1	kes	August	850000



3. ¿Cómo crear un Modelo Semántico en Power BI?



5 PASOS PARA CREAR UN MODELO SEMÁNTICO



Conéctate a tus fuentes de datos



Transforma y limpia los datos



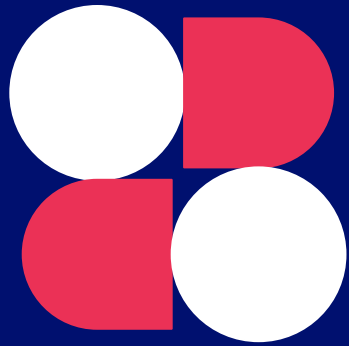
Define la relación entre las bases de datos



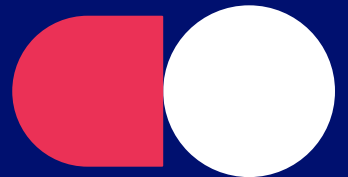
Crea las medidas y cálculos



Publica en el servidor de Power BI



4. Almacenes de datos - Datamarts



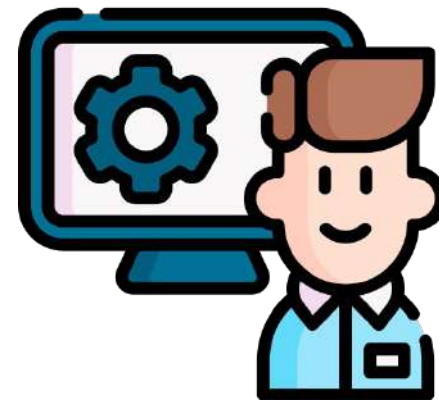


4. ¿Qué es Datamarts?



Los Datamarts son soluciones de análisis de autoservicio, lo que permite a los usuarios almacenar y explorar datos que se cargan en una base de datos totalmente administrada.

Es un subconjunto de un Data Warehouse que se enfoca en un área específica.



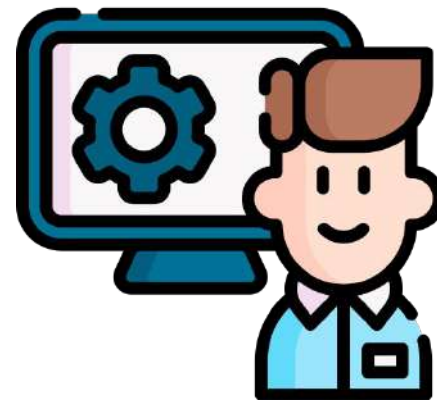


4. ¿Qué es Datamarts?



Los Datamarts proporciona una experiencia sencilla y opcionalmente sin código para:

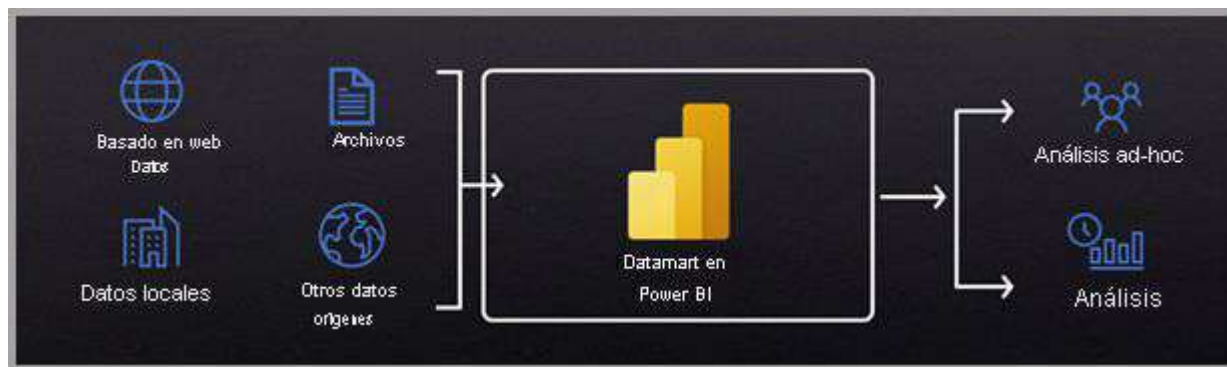
- Ingerir datos de distintos orígenes de datos
- Extraer la transformación y cargar (ETL) los datos mediante Power Query
- Cargarlos en una base de datos de Azure SQL totalmente administrada y no requiere ajuste ni optimización.





4. ¿Qué es Datamarts?

Datamarts genera automáticamente un conjunto de datos o un modelo semántico, que se puede usar para crear informes y paneles de Power BI.

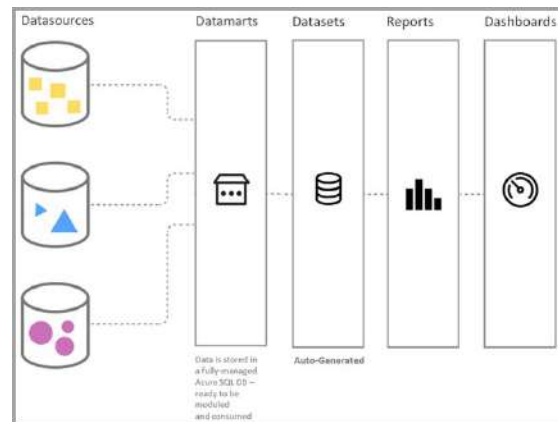




4. Datamarts - Características clave



- 100 % basado en web, no se requiere ningún otro software
- Una experiencia sin código que da como resultado un datamart totalmente administrado
- Ajuste del rendimiento automatizado
- Editor de consultas SQL y visual integrados para el análisis ad hoc
- Compatibilidad con SQL y otras herramientas de cliente populares
- Integración nativa con Power BI, Microsoft Office y otras ofertas de Análisis de Microsoft
- Se incluye con capacidades de Power BI Premium y Premium por usuario



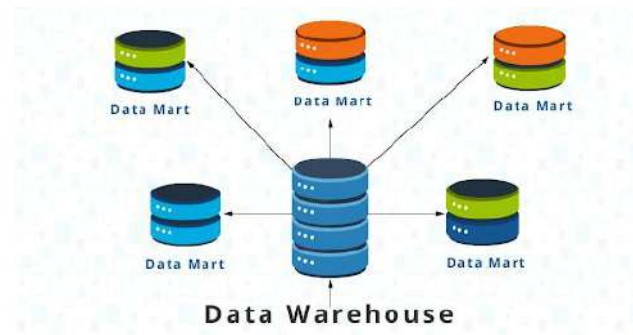


4. Datamarts - Ventajas



Ventajas:

- Los usuarios de autoservicio pueden realizar fácilmente análisis de bases de datos relacionales, sin necesidad de un administrador de bases de datos.
- Datamarts proporciona ingesta, preparación y exploración de datos de un extremo a otro con SQL, incluidas las experiencias sin código.
- Habilitación de la creación de modelos semánticos e informes dentro de una experiencia global





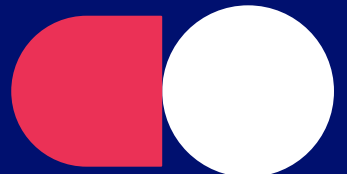
4. Data Flows Vs. Datamarts



Característica	Data Flows	Datamarts
Función Principal	Extracción, Transformación y Carga reutilizables (ETL)	Base de datos totalmente administrada para almacenamiento y exploración de datos
Exploración de datos	No se pueden explorar ni consultar sin un modelo semántico	Permiten explorar y consultar datos en una base de datos relacional
Formato de datos	Power BI o CDM (Common Data Model) si se utiliza un lago de datos propio	Almacenados en una base de datos Azure SQL
Reutilización	Lógica ETL reutilizable	Generación automática de un modelo semántico para cada datamart
Casos de Uso	Reutilización de lógica de ETL, preparación de datos compartibles para Power BI	Almacenamiento, exploración, análisis ad hoc, creación de informes en la web



5. Power BI Services

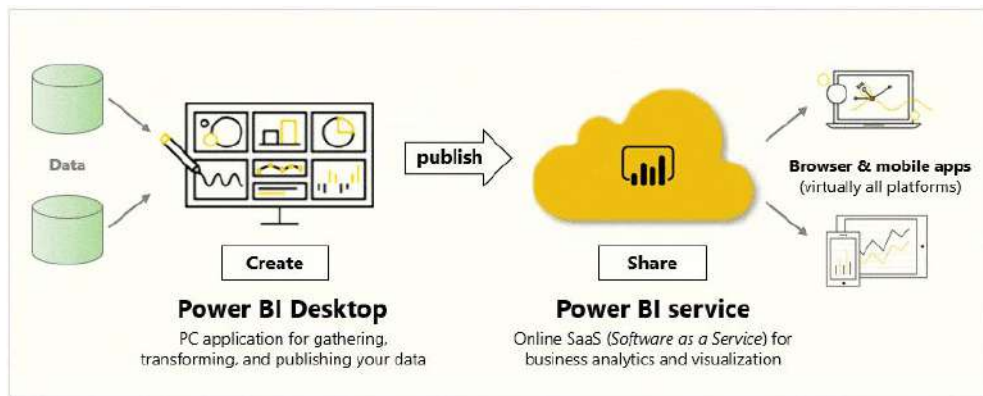




5. ¿Qué es Power BI Services?



Power BI Services es una plataforma de análisis de datos basada en la nube que permite a las organizaciones compartir, colaborar y consumir información empresarial a través de informes, paneles y aplicaciones. Facilita la creación y distribución de contenido analítico, mejorando la toma de decisiones basada en datos.





5. Funcionalidades Principales de Power BI Services



- **Paneles Interactivos:** Visualización en tiempo real de métricas clave.
- **Informes:** Herramientas para crear informes detallados y personalizados.
- **Aplicaciones:** Paquetes de contenido que combinan paneles e informes.
- **Espacios de Trabajo:** Áreas colaborativas para equipos y departamentos.
- **Conexión a Datos:** Acceso a diversas fuentes de datos tanto en la nube como locales.





5. Ventajas de Power BI Services



- **Acceso en Tiempo Real:** Datos actualizados al instante.
- **Colaboración:** Facilita el trabajo en equipo con espacios de trabajo y comentarios.
- **Escalabilidad:** Adecuado para pequeñas y grandes empresas.
- **Seguridad:** Protege datos con controles de acceso y encriptación.
- **Automatización:** Programación de actualizaciones y notificaciones automáticas.





5. Power BI Services - Componentes Clave

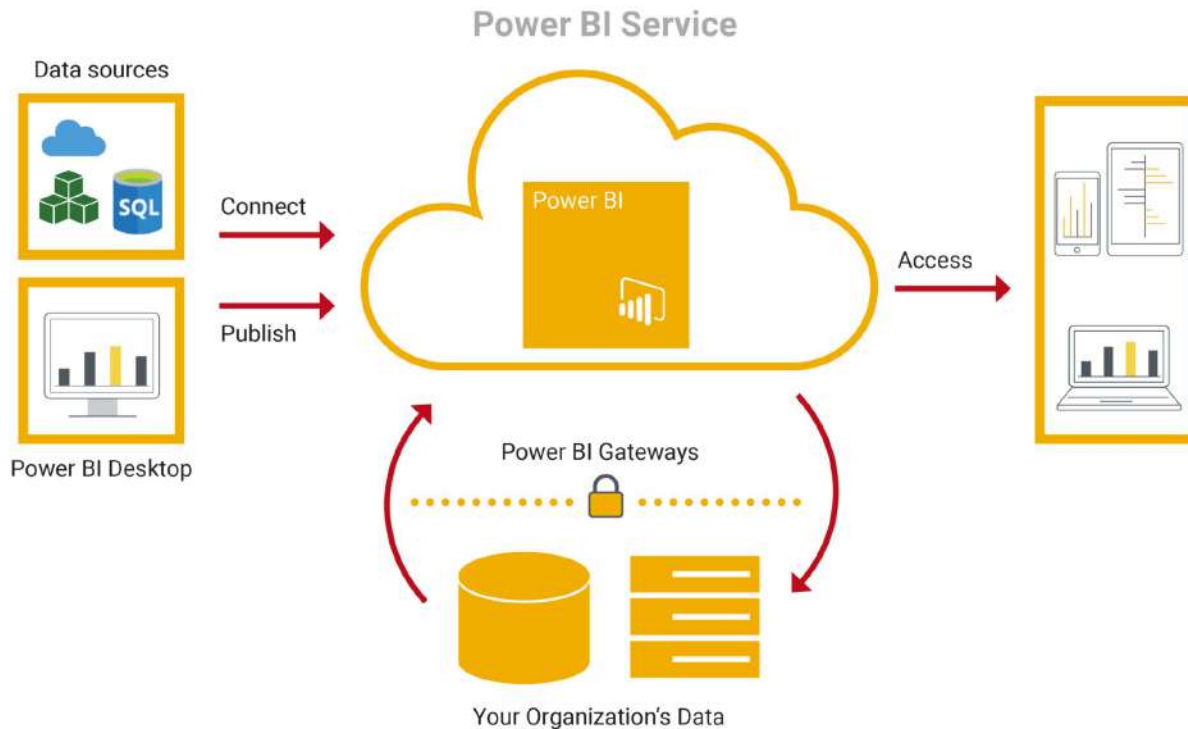


- **Informes y Paneles:** Herramientas de visualización y análisis.
- **Workspaces:** Espacios de trabajo colaborativos para equipos.
- **Apps:** Paquetes de contenido reutilizables.
- **Datasets:** Conjuntos de datos conectados a múltiples fuentes.
- **Dataflows:** Preparación y transformación de datos.



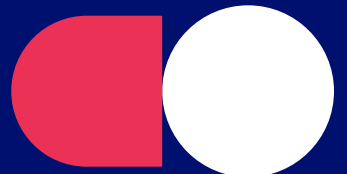


5. Power BI Services - Arquitectura





6. Data Mesh





6. ¿Qué es Data Mesh?



Data Mesh es una arquitectura de datos descentralizada que trata los datos como un producto y otorga la responsabilidad de los datos a los equipos que mejor los entienden.

A diferencia de las arquitecturas tradicionales centralizadas, Data Mesh permite una mejor escalabilidad y agilidad en el manejo de grandes volúmenes de datos.

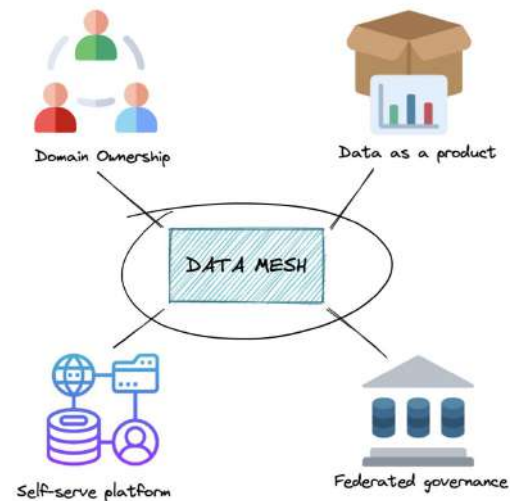




6. Principios fundamentales - Data Mesh



- **Dominio Propietario de Datos:**
Los equipos responsables de los dominios empresariales son también responsables de los datos.
- **Datos como Producto:**
Los datos deben ser tratados como productos con su propio ciclo de vida.
- **Infraestructura de Autoservicio:**
Proveer herramientas y plataformas que permitan a los equipos gestionar sus propios datos.
- **Gobernanza Federada:**
Establecer políticas y estándares comunes que sean implementados de manera descentralizada.





6. Componentes Clave - Data Mesh



- **Dominios de Datos:**
Divisiones lógicas de la organización que poseen y gestionan sus propios datos.
- **Plataforma de Autoservicio:**
Herramientas y servicios que permiten a los equipos gestionar, procesar y consumir datos.
- **Gobernanza Federada:**
Políticas y estándares comunes aplicados de manera descentralizada.
- **Productización de Datos:**
Tratamiento de los datos como productos completos con documentación, APIs y métricas de calidad.





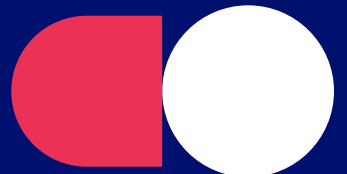
6. Data Mesh Vs. Arquitecturas Tradicionales



Característica	Data Mesh	Arquitecturas Tradicionales
Estructura	Descentralizada	Centralizada
Responsabilidad	Equipos de dominio	Equipos centrales de datos
Escalabilidad	Alta	Limitada
Gobernanza	Federada	Centralizada
Mantenimiento	Distribuido entre equipos	Equipo central



7. Gobierno de Datos





7. ¿Qué es Gobierno de Datos?

El gobierno de datos es el conjunto de políticas, procesos y controles que aseguran la calidad, integridad, seguridad y disponibilidad de los datos en una organización. Su objetivo es gestionar los datos como un activo valioso, garantizando su uso efectivo y responsable.





7. Importancia del Gobierno de Datos



- **Calidad de Datos:** Asegura que los datos sean precisos, completos y consistentes.
- **Cumplimiento Normativo:** Garantiza que los datos cumplan con regulaciones y normativas.
- **Seguridad de Datos:** Protege los datos contra accesos no autorizados y brechas de seguridad.
- **Toma de Decisiones:** Mejora la toma de decisiones basada en datos fiables y precisos.
- **Eficiencia Operativa:** Reduce la redundancia y los errores en los procesos de datos.





7. Componentes del Gobierno de Datos



- **Políticas y Normativas:** Directrices que gobiernan la gestión y el uso de los datos.
- **Roles y Responsabilidades:** Definición clara de quién es responsable de qué datos.
- **Calidad de Datos:** Procesos y herramientas para asegurar la precisión y consistencia.
- **Seguridad de Datos:** Medidas para proteger los datos contra accesos no autorizados.
- **Monitoreo y Auditoría:** Seguimiento continuo y revisión del cumplimiento de las políticas de datos.
- **Educación y Capacitación:** Formación para los empleados sobre buenas prácticas de manejo de datos.





7. Roles del Gobierno de Datos



Roles clave en el Gobierno y Gestión de Datos



Data Owner (DO)



Data Stewards (DS)



Chief Data Officer (CDO)



Data Protection
Officer (DPO)



Chief Information
Officer (CIO)



Oficina de Gobierno
del Dato



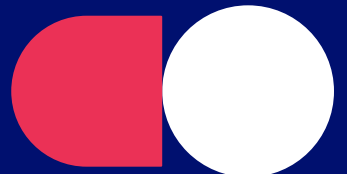
7. Roles del Gobierno de Datos



- **Chief Data Officer (CDO):** Responsable de la estrategia de datos a nivel organizacional. Asegura que los datos sean gestionados como un activo estratégico.
- **Data Governance Council:** Grupo de líderes y expertos que establecen políticas y estándares. Supervisa la implementación del gobierno de datos.
- **Data Stewards:** Garantiza la calidad y uso adecuado de los datos. Trabaja en estrecha colaboración con los propietarios de datos y usuarios.
- **Data Owners:** Propietarios de los datos dentro de un dominio específico. Responsable de la gestión y protección de sus datos.
- **Data Custodians:** Técnicos que administran el almacenamiento y la infraestructura de datos. Asegura la seguridad y disponibilidad de los datos.
- **Data Users:** Empleados que acceden y utilizan los datos para sus actividades diarias. Deben seguir las políticas y prácticas establecidas.



8. Actividad grupal



slideswith.com/PREPASOS10

Leer

Analizar

Medir

Describir





- ¿Qué es un Data Lake?

R // Un repositorio centralizado que almacena datos en su forma cruda y original

- ¿Qué es un Data Warehouse?

R// Un sistema utilizado para reportes y análisis de datos estructurados y organizados

- ¿Qué es un Dataflow en Power BI?

R// Un proceso para preparar y transformar datos antes de usarlos en Power BI



- ¿Qué es un Datamart?

R// Una subdivisión de un Data Warehouse enfocada en un área específica

- ¿Qué es Power BI Services?

R// Una plataforma en la nube para compartir y colaborar en informes y dashboards de Power BI

- ¿Qué es Data Mesh?

R// Un enfoque descentralizado para la gestión y acceso a datos



Gracias

