TEMA 75. TECNOLOGÍAS Y SISTEMAS DE EXPLOTACIÓN DE DATOS: DATA LAKE, DATA WAREHOUSE, LAKEHOUSE, DATA FABRIC, DATA MESH, TECNOLOGÍAS PARA LA PROTECCIÓN DE LA CONFIDENCIALIDAD (PET). ENTORNOS DE COMPARTICIÓN DE DATOS: ESPACIOS DE DATOS, ASPECTOS TECNOLÓGICOS Y ORGANIZATIVOS.

Actualizado a 16/05/2023

1. CONCEPTOS

- **Gestión de Datos**: Datos -> Información -> Conocimiento -> Sabiduría
- **Business Intelligence (BI)**: Conjunto metodologías, aplicaciones, prácticas y capacidades enfocadas a creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización. Un sistema de BI engloba diferentes tecnologías: DWH, análisis OLAP, cuadro de mando integral, dashboards, Minería de datos, integración de datos.
- **Data Warehouse (DW) o almacenes de datos**: base de datos corporativa caracterizada por la integración y depuración de información procedente de múltiples fuentes de datos tanto internas como externas a la organización. Es integrado, no volátil y variable en el tiempo y su fin es el de procesar la información para poder analizarla.
- **Data Warehousing (DWH) o almacenes de datos**: proceso que emplea un Data Warehouse.
- **Data Mart**: almacenes de datos especializados por áreas o temas. Es un subconjunto de los datos guardados en un DW, destinado a satisfacer las necesidades de un segmento de negocio en particular.
- **Strategic Marts** : Especializados en negocios estratégicos. Concepto similar a Data Mart.
- **Minería de datos**: se encarga de la extracción de patrones o de información implícita u oculta contenida en los datos.
- **Data Lake o lago de datos**: repositorio de almacenamiento centralizado que contiene gran cantidad de datos en bruto. Se diferencia de un data warehouse en que utiliza una arquitectura plana para almacenar los datos.
- **Data Lakehouse**: Un data lakehouse es una arquitectura híbrida de gestión de datos que combina las ventajas de flexibilidad y escalabilidad de un lago de datos con las estructuras y características de gestión de datos de un almacén de datos.
- **Data Fabric**: Un Data Fabric consiste en una arquitectura de servicios y funcionalidades que contribuye a procesar mejor los volúmenes de datos procedentes de multitud de fuentes. Es decir, tiene la capacidad de agruparlos bajo una misma nube o sistema de administración, sin importar de donde provienen esos datos.
- **Data Mesh**: Data Mesh es un enfoque sociotécnico descentralizado para compartir, acceder y gestionar datos analíticos en un entorno complejo y a gran escala dentro o entre las organizaciones. Consiste en cambiar los propietarios de los datos y cambiar el enfoque de cómo son consumidos y usados en la arquitectura de datos.

2. DATA LAKE

Un data lake o lago de datos es un repositorio centralizado que permite almacenar, compartir, gobernar y descubrir todos los datos estructurados y no estructurados de una organización a cualquier escala. Es el lugar en el que se vuelcan los datos en bruto.

Entre las características más importantes de los data lakes se encuentran su **flexibilidad, agilidad, trazabilidad y la escalabilidad**.

Los componentes de un Data Lake son:

- **Ingesta de datos**: Un sistema de capas de ingesta fácilmente escalable que extrae datos de fuentes diversas, incluidas páginas web, aplicaciones móviles, redes sociales, dispositivos IoT y sistemas de gestión de datos existentes.

- **Almacenamiento de datos**: Un sistema de almacenamiento de datos debe ser capaz de almacenar y tratar datos sin procesar.
- **Seguridad en los datos**: Deben ofrecer máxima seguridad, utilizando sistemas de autenticación y autorización, así como niveles de acceso basado en roles, protección de datos, etc.
- **Análisis de datos**: Una vez realizada la ingesta, los datos deben poder ser analizados de manera ágil y eficiente.
- **Gobierno de datos**: El proceso de ingesta, preparación, catalogación, integración y aceleración de consultas de datos debe simplificarse en su totalidad.

Un enfoque simple puede ser comenzar con algunas zonas genéricas:

- **Raw / Bronze**. Esta capa es un depósito que almacena datos en su estado original, sin filtrar ni limpiar.
- **Cleansed / Silver**. La siguiente capa se puede considerar como una zona de filtración que elimina las impurezas, pero también puede implicar un enriquecimiento.
- **Curated / Gold**. Esta es la capa de consumo, que está optimizada para análisis en lugar de ingesta. Puede almacenar datos en data marts desnormalizados o esquemas en estrella.
- **Laboratory**. Esta es la capa donde ocurre la exploración y la experimentación.

Un Data Lake se puede implementar de dos maneras. **Data Lake Local (On Premise)** o **Data Lake Cloud**.

Un Data Lake tiene muchas ventajas. Algunas de las destacables son:

- **Centralización.** El Data Lake permite centralizar todos los datos en un mismo lugar, vengan de la fuente que vengan.
- **Persistencia**. Puede que la fuente original del dato esté obsoleta o se haya desactivado, sin embargo, su contenido puede que siga siendo valioso para el análisis.
- **Flexibilidad**. Todo dato que llegue al Data Lake puede ser normalizado y enriquecido. Además los datos se preparan en función de la necesidad del momento. Esto permite reducir considerablemente los costes y los tiempos.
- **Disponibilidad**. Se puede acceder a la información y enriquecerla desde cualquier punto del planeta, por cualquier usuario autorizado por el Data Lake. Esto ayuda a la organización a recopilar más fácilmente los datos necesarios para la toma de decisiones.

El data lake adopta una estructura denominada "**Schema on Read**", la estructura no está predeterminada antes de que se almacenen los datos. Este tipo de estructura les permite adaptase a los cambios de uso y circunstancias.

Los silos de datos o Data Silos ocurren cuando no existe un lugar o un sistema centralizado en el que almacenar todos los datos de la organización.

Un Data Lake inteligente permite analizar eficazmente el grado de protección de la información que se guarda en los diferentes silos. Con la nueva normativa europea **GDPR (General Data Protection Regulation)**, esta seguridad en la privacidad de los datos se ve asegurada.

Otro de los escenarios a evitar son los denominados **Data Swamps**. Estos surgen cuando una organización recopila y almacena grandes cantidades de datos sin un plan o procesos efectivos de gestión y clasificación.

Algunas de las tecnologías para implementar un Data lake son:

- **Azure**
  - Almacenamiento: ADLS (Azure Data Lake Storage). Generation2 con Azure Blob Storage.
  - Clústers: Hadoop, Spark o Kafka
  - Ingesta de datos: Azure Data Factory
- **AWS**
  - Almacenamiento: S3 (Simple Storage Service)
  - Redshift: Servicio de DataWarehouse
  - Ingesta de datos: AWS Glue

3. DATA WAREHOUSE

Enmarcado dentro de los sistemas:

- **DSS (Decision Suport System)**: sistemas de soporte a la decisión.
- **EIS (Executive Information System**): especialización de DSS para la dirección.
- **MIS (Management Information Systems**): centrado en niveles operativo, táctico y estratégico.

Bill Inmon (TOP DOWN): Los almacenes de datos se caracterizan por ser "OVNI": **orientados a temas**, **variantes en el tiempo**, **no volátiles** e **integrados** (solución integral procedente de varias fuentes)

Kimball (BOTTOM UP): Define el almacén de datos como "una copia de las transacciones de datos específicamente estructuradas para la consulta y análisis).

3.1. APROVISIONAMIENTO DE UN DATAWAREHOUSE

Los DW han de recopilar la información de diferentes fuentes, para ello hacen uso de **herramientas ETL** (extract, transform and load), que se encargan de la extracción de la información, su transformación y posterior carga en el almacén de datos.

CDI (customer data integration) Unifica datos de varias bbdd dispersas, por ejemplo clientes en varias bbdd de la organización.

En todo este proceso son importantes los **METADATOS**: datos que describen a datos y que son empleados por el DW para simplificar y posibilitar la obtención de información. Se puede aprovisionar el DW desde dos posibles enfoques:

- **Bottom Up**: Se aprovisionan almacenes temáticos (DATAMARTS) y del conjunto de ellos se crea el DW.
- **Top Down:** Se aprovisiona el DW y, si se requiere especialización temática, se aprovisionan a partir del DW los DATAMARTS. Este enfoque es poco práctico.

Ejemplos de herramientas ETL: Pentaho Data Integration (Kettle), Scriptella, Ab Initio o AWS Glue (solución en la nube).

Áreas de datos

- **Staging Area**: Área temporal donde se almacenan los datos de origen para las cargas.
- **Operational Data Store (ODS)**: Área de soporte a los sistemas operacionales. Proporciona el último valor de los datos almacenados; no mantiene históricos.

3.2. EXPLOTACIÓN DE UN DATAWAREHOUSE

Explotación de la información:

- **Query & Reporting**. Herramientas para la elaboración de informes tanto en detalle como
- sobre información agregada, a partir de los DW y DM.
- **Cuadro de mando analítico (CM)**. Dashboards. Orientado a la obtención de indicadores **KPI (Key Performance Indicator)** para la dirección.
- **Cuadro de mando integral o estratégico**. **CMI**. Orientado a área estratégica de una organización. El cuándo de mando integral permite alinear los objetivos de las diferentes áreas o unidades con la estrategia de empresa. Combina KPI con **KGI (Key Global Indicator).**

Estructuras multidimensionales: los DW, a diferencia de los OLTP, se basan en el uso de estructuras (Cubos) multidimensionales que permiten la manipulación y visualización de los datos de manera más eficiente. Se componen de: **tablas de hechos, tablas de dimensiones y métricas**. La tabla de hechos es la tabla principal del modelo que contiene los "campos claves" que se unen a las tablas de dimensión.

Los tres esquemas más habituales de las BBDD multidimensionales son: **En estrella**, **copo de nieve** y **constelación**.

Las operaciones habituales que se pueden hacer en este modelo son:

- **DRILL-DOWN o ROLL-DOWN**: Desagregar (añade +nivel de detalle). Añade un atributo de una dimensión existente.
- **DRILL- UP o ROLL-UP**: Agregar (reduce el nivel de detalle). Sube un nivel en la jerarquía, quita un atributo de una dimensión que sigue existiendo.
- **DRILL-ACROSS**: Como drill-down, pero agrega un atributo de una nueva dimensión a la consulta como nuevo criterio de análisis.
- **ROLL-ACROSS**: Como drill-up, quita un atributo de la consulta y con ello una dimensión.
- **PIVOT-ROTATE**: Cambiar orden de visualización de los atributos e indicadores, para analizar la información desde una diferente perspectiva.
- **DRILL-THROUGH**: Expandir para apreciar los datos en su MÁXIMO nivel detalle.
- **PAGE**: Presenta el cubo dividido en secciones (páginas), según los valores de un atributo (útil para cuando una query devuelve muchos registros).
- **SLICE**: Selecciona un elemento de una dimensión y proyecta sobre el resto de dimensiones. El resultado es un nuevo sub-cubo.
- **DICE**: Selecciona dos o más dimensiones de un cubo, cuyo resultado es un nuevo sub-cubo.

Para explotar la información de esta manera se emplean sistemas OLAP (Online Analytical Processing).

Inconsistencias que pueden aparecer en los OLAP:

- **Data Sparsity**. Hay huecos en los datos, no tiene por qué haber datos en todas las celdas.
- **Data Explosion**. Al haber muchas dimensiones la información crece exponencialmente.

3.3. ARQUITECTURA OLAP

Ámbitos de aplicación: DataWarehouse, Sistemas de toma de decisiones (DSS), Sistemas de información ejecutiva (EIS), así como en el campo de la llamada Inteligencia de negocios (o Business Intelligence).

Objetivo: agilizar la consulta de grandes cantidades de datos. Es lo más rápido a la hora de ejecutar sentencias SQL de tipo SELECT (consultas analíticas complejas e iterativas) en contraposición con OLTP (INSERT, UPDATE Y DELETE).

Tipos de sistemas OLAP:

- **MOLAP**: la información se almacena directamente de forma **multidimensional** (cubos). Es la forma clásica.
- **ROLAP**: Relational OLAP. A nivel físico la información se almacena de forma **relacional,** pero para su explotación se construyen cubos dinámicamente.
- **HOLAP**: Hybrid OLAP. Mezcla de los dos anteriores.
- **DOLAP**: Desktop OLAP. Orientado a equipos de escritorio. BD relacional almacena los datos y se hacen las consultas y el análisis de datos desde una copia en local.
- **SOLAP:** Spatial OLAP
- **RTOLAP:** Real Time OLAP
- **WOLAP**: Web OLAP
- **In-memory OLAP**: La estructura dimensional se genera sólo a nivel de memoria.

Ejemplos de BBDD que permiten almacenar cubos MOLAP/HOLAP son: Hbase, Oracle OLAP o SQL Server.

Analysis Services.Una dimensión obligatoria en el procesamiento OLAP es el **TIEMPO**.

4. DATA LAKEHOUSE

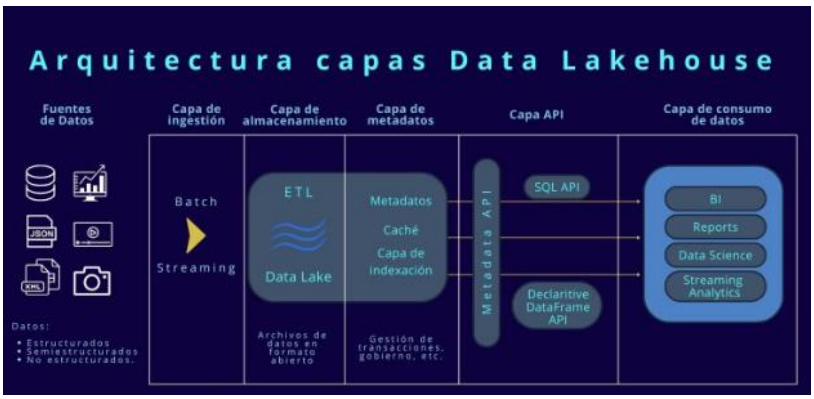
Data Lakehouse integra y unifica un Data Warehouse y un Data Lake para combinar lo mejor de ambos mundos y construir de manera flexible y elástica un ecosistema que respalde sin problemas la inteligencia empresarial y los informes, la ciencia de los datos y la ingeniería de datos, el aprendizaje automático e IA.

Los Data lakehouse suelen comenzar como data lakes que contienen todo tipo de datos. Luego, los datos se convierten al formato **Delta Lake** (una capa de almacenamiento de código abierto que brinda confiabilidad a los lagos de datos).

Entre algunas de las características comunes de un Data Lakehouse son:

- **Tipos de datos extendidos (Extended data)**: Data Lakehouse tienen acceso a una gama más amplia de datos que los data warehouse.
- **Transmisión de datos (Data streaming**): Permite realizar informes en tiempo real al admitir análisis de transmisión.
- **Esquemas**: A diferencia de los data lakes, Data Lakehouse aplica esquemas a los datos, lo que ayuda en la estandarización de grandes volúmenes de datos.
- **Soporte de Business Intelligence y Data Science**: Los profesionales de BI y análisis pueden compartir el mismo repositorio de datos.
- **Compatibilidad con transacciones**: Manejar transacciones simultáneas de escritura y lectura.
- **Apertura**: Data Lakehouse admite formatos de almacenamiento abiertos (p. ej., Parquet). De esta forma, los profesionales de datos pueden usar R y Python para acceder a ellos fácilmente.
- **Desacoplamiento de procesamiento/almacenamiento**: un Data Lakehouse reduce los costes de almacenamiento mediante el uso de clústeres que se ejecutan en hardware económico.

Un Data Lakehouse puede tener hasta cinco capas:



5. DATA FABRIC

Data Fabric es una **combinación** de **arquitectura de datos y soluciones de software** dedicadas que centralizan, conectan, gestionan y gobiernan datos entre diferentes sistemas y aplicaciones. Esto permite a las empresas acceder y usar datos en tiempo real, creando una única fuente confiable y automatizando los procesos de gestión de datos.

Algunos de los componentes clave de la arquitectura de Data Fabric incluyen:

- **Conectores de datos**: Conectores de datos como puentes que conectan los diferentes sistemas donde se almacenan datos a una ubicación central.
- **Gestión de datos**: Implica asegurarse de que los datos estén organizados, sean seguros y de alta calidad. Aquí se incluyen actividades como **integración de datos** (unir datos de diferentes fuentes); **gobernanza de datos** (establecer reglas sobre cómo se deben usar y gestionar); y **seguridad de datos** (proteger los datos confidenciales contra el acceso no autorizado).
- **Modelado de datos y capa semántica**: el modelado de datos ayuda a dar sentido a los datos creando un lenguaje común para los datos en diferentes sistemas.
- **Procesamiento y analíticas de datos**: Se procesan y analizan los datos para obtener información estratégica. Aquí entran en juego tareas como **almacenamiento de datos** (almacenar grandes cantidades de datos); **streaming de datos** (procesar continuamente los datos a medida que se generan); y **visualización de datos**.
- **Automatización de la gestión de datos**: El análisis de datos se puede usar para fundamentar la automatización en varias áreas del negocio, pero como término arquitectónico, la automatización ayuda a garantizar que los datos se gestionen eficiente y consistentemente.

Componentes arquitectónicos clave de Data Fabric son:

- Fuentes de datos para la ingesta
- Análisis y gráficos de conocimiento para el procesamiento
- Algoritmos avanzados para la generación de insights
- API y SDK para conectividad con interfaces de entrega
- Capa de consumo de datos
- Capa de transporte de datos
- El entorno de alojamiento

6. DATA MESH

Data Mesh no es sólo **tecnología**, también es una **cultura**. Data Mesh o malla de datos es un método sociotécnico para construir una **arquitectura** de datos **descentralizada** mediante el aprovechamiento de un diseño de autoservicio orientado al dominio (en una perspectiva del desarrollo de software)

Los cuatro principios de Data Mesh son:

- **Propiedad impulsada/orientada por el dominio**: transferir la propiedad de los datos a las manos de los dominios (ejemplo gerencias, departamentos, etc.), siendo los dueños y teniendo la responsabilidad de asegurar su calidad y seguridad.
- **Datos como producto**: El objetivo es que los dominios sean capaces de generar sus productos de datos, mantenerlos, validar y mejorar su calidad.
- **Infraestructura de autoservicio**: se refiere a que las tareas complejas implicadas en la generación de recursos y espacios de desarrollo, sean simples y en modalidad de autoservicio para que los usuarios de los dominios de negocios puedan escalar sin depender de especialistas.
- **Gobernanza Federada**: se refiere a los métodos que nos permiten obtener un equilibrio entre las políticas y las acciones realizadas por los dominios de negocio, con la finalidad de no arriesgar la privacidad, el incumplimiento de políticas y la escalabilidad de los dominios de negocio.

Además una arquitectura de este tipo debe ofrecer características **como la interoperabilidad, seguridad, gobierno, auto descriptiva, transversal en equipos, cambio de paradigma**.

7. PROTECCIÓN DE LA CONFIDENCIALIDAD

Confidencialidad es la propiedad de la información, por la que se garantiza que está accesible únicamente a personal autorizado a acceder a dicha información.

Organización Internacional de Estandarización (ISO) en la norma **ISO/IEC 27002** como "garantizar que la información es accesible sólo para aquellos autorizados a tener acceso".

Reglamento General de Protección de Datos, **RGPD (UE) 2016/679**, art. **5.1.f**, " Los datos personales serán tratados de tal manera que se garantice una seguridad adecuada de los datos personales, incluida la protección contra el tratamiento no autorizado o ilícito y contra su pérdida, destrucción o daño accidental, mediante la aplicación de medidas técnicas u organizativas apropiadas («integridad y confidencialidad»)."

De la misma forma, según el artículo 5 de la **Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, LOPDGDD**, "Los responsables y encargados del tratamiento de datos así como todas las personas que intervengan en cualquier fase de este estarán sujetas al deber de confidencialidad al que se refiere el artículo 5.1.f) del Reglamento (UE) 2016/679."

Privacy Enhancing Technologies (PET). La Agencia Ciberseguridad de la Unión Europea (**ENISA**) se ha referido a las PET **como sistemas que engloban procesos, métodos o conocimientos técnicos para lograr una funcionalidad específica de protección de la intimidad o de los datos de las personas físicas**.

Los PET se pueden dividir en cuatro categorías:

Tipos de PET	Tecnologías clave	Aplicaciones actuales y potenciales*	Retos y limitaciones
Herramientas de ofuscación de datos	Anonimización / seudonimización	Almacenamiento seguro	- Garantizar que la información no se filtre (riesgo de reidentificación)
	Datos sintéticos	Aprendizaje automático para preservar la privacidad	- Sesgo amplificado en particular para datos sintéticos
	Privacidad diferencial	Ampliación de las oportunidades de	- Habilidades y competencias insuficientes
	Pruebas de conocimiento cero	investigación Verificación de la información sin necesidad de divulgación (por ejemplo, verificación de la edad)	- Las aplicaciones aún están en sus primeras etapas.
Herramientas de procesamiento de datos encriptados	Cifrado homomórfico	Cómputo de datos cifrados dentro de la misma organización Cómputo de datos privados que son demasiado confidenciales para divulgarlos Seguimiento/descubrimiento de contactos	- Desafíos de limpieza de datos
	Cálculo de múltiples partes (incluida la intersección del conjunto de privados)		- Garantizar que la información no se filtre.
			- Mayores costos de cómputo
	Entornos de ejecución de confianza	Informática utilizando modelos que deben permanecer privados	- Mayores costos de cómputo - Desafíos de seguridad digital
Analítica federada y distribuida	Aprendizaje federado	Aprendizaje automático que preserva la privacidad	- Se necesita conectividad confiable
	Analítica distribuida		- La información sobre los modelos de datos debe estar disponible para el procesador de datos
Herramientas de responsabilidad de datos	sistemas contables	Establecer y hacer cumplir reglas sobre cuándo se puede acceder a los datos Seguimiento inmutable del acceso a los datos por parte de los controladores de datos	- Casos de uso limitados y carecen de aplicaciones independientes - Complejidad de configuración - Riesgos de cumplimiento de privacidad y protección de datos cuando se utilizan tecnologías de contabilidad distribuida
	Umbral de uso compartido de secretos		
	Almacenes de datos personales / Personal Gestión de la información	Proporcionar a los interesados el control sobre sus propios datos	- Desafíos de seguridad digital

8. ESPACIOS DE DATOS

Un **espacio de datos** es un ecosistema donde materializar la compartición voluntaria de los datos de sus participantes dentro de un entorno de soberanía, confianza y seguridad, establecido mediante mecanismos integrados de gobernanza, organizativos, normativos y técnicos.

La Unión Europea establece los **fundamentos** que deben seguir los espacios de datos, estos son:

- Descentralización
- Apertura
- Transparencia
- Soberanía
- Interoperabilidad

Para una correcta y segura compartición de espacios de datos se definen 5 pasos a seguir:

8.1. PRINCIPIOS

La estrategia de la Unión Europea establece como 4 los principios básicos de un espacio de datos:

- **Soberanía de datos**. Es la capacidad de una persona o de una entidad para la libre determinación en cuanto a sus datos.
- **Igualdad de condiciones**. Eliminación de barreras por situaciones de monopolio. Cuando existe una igualdad, se compite en la calidad del servicio, y no en la cantidad de datos que controlan.
- **Infraestructura descentralizada**. Los datos no deben estar almacenados en sistemas centralizados monolíticos si no en infraestructuras descentralizadas interoperables con un nivel de acuerdo funcional, técnico, operacional, legal y económico.
- **Gobernanza público-privada**. Definición de acuerdos y gobernanza entre personas, empresas públicas y privadas y servicios de datos.

IDSA (International Data Spaces Association) coalición que actualmente integran 133 empresas internacionales, sin ánimo de lucro, para trabajar en el concepto de espacio de datos y en los principios que debe seguir su diseño para obtener valor de los datos a través de la compartición, en base a mecanismos seguros, transparentes y con equidad que garanticen la soberanía y confianza.

8.2. ROLES

Los siguientes roles son base para el funcionamiento de un espacio de datos.

8.3. COMPONENTES

Componentes para el acceso al espacio de datos:

- **Componentes para el acceso - Conector.** Uno de los principales elementos de espacios de datos es el conector, por el cual los participantes acceden al espacio de datos y a los propios datos.
- **Componentes para la intermediación.** Permiten los servicios de intermediación antes mencionados. De todos ellos, el más fundamental es el catálogo de recursos.
- **Componentes para la gestión de identidad y el intercambio seguro de datos**. Estos componentes permiten garantizar la identidad y la seguridad de las transacciones.
- **Componentes para la gestión del espacio de datos**. Se trata de herramientas que permiten que el espacio de datos opere con normalidad, facilitando las operaciones diarias y la gestión.

8.4. REFERENCIAS

Europa:

- Estrategia Europea de datos
- Ley Europea de datos
- Ley de Gobernanza Europea de datos
- Espacios de datos
- Espacio Europeo de Datos Sanitarios

España:

- España Digital 2026
- Estrategia Nacional de Inteligencia Artificial ENIA
- Hub español Gaia-X
- Datos.gob.es
- Protección espacio de datos
- Oficina del Dato