BIG DATA for BUSINESS

2.3 Gobierno del dato

Conecta Empleo

Contenido desarrollado por
Synergic Partners

Índice del módulo

2.2. GOBIERNO DEL DATO

Data Governance

El Gobierno del Dato es una disciplina que permite garantizar el valor estratégico del dato en la plataforma a partir de la elaboración de un marco de Gobierno del Dato en el que se definan las capacidades organizativas, disciplinas tecnológicas y procesos y herramientas que garanticen el correcto despliegue del mismo.

Data Governance - Definiciones

"Orquestación formal de procesos, personas y tecnología para permitir que una organización convierta sus datos en un activo estratégico"

Fuente: *The MDM Institute*

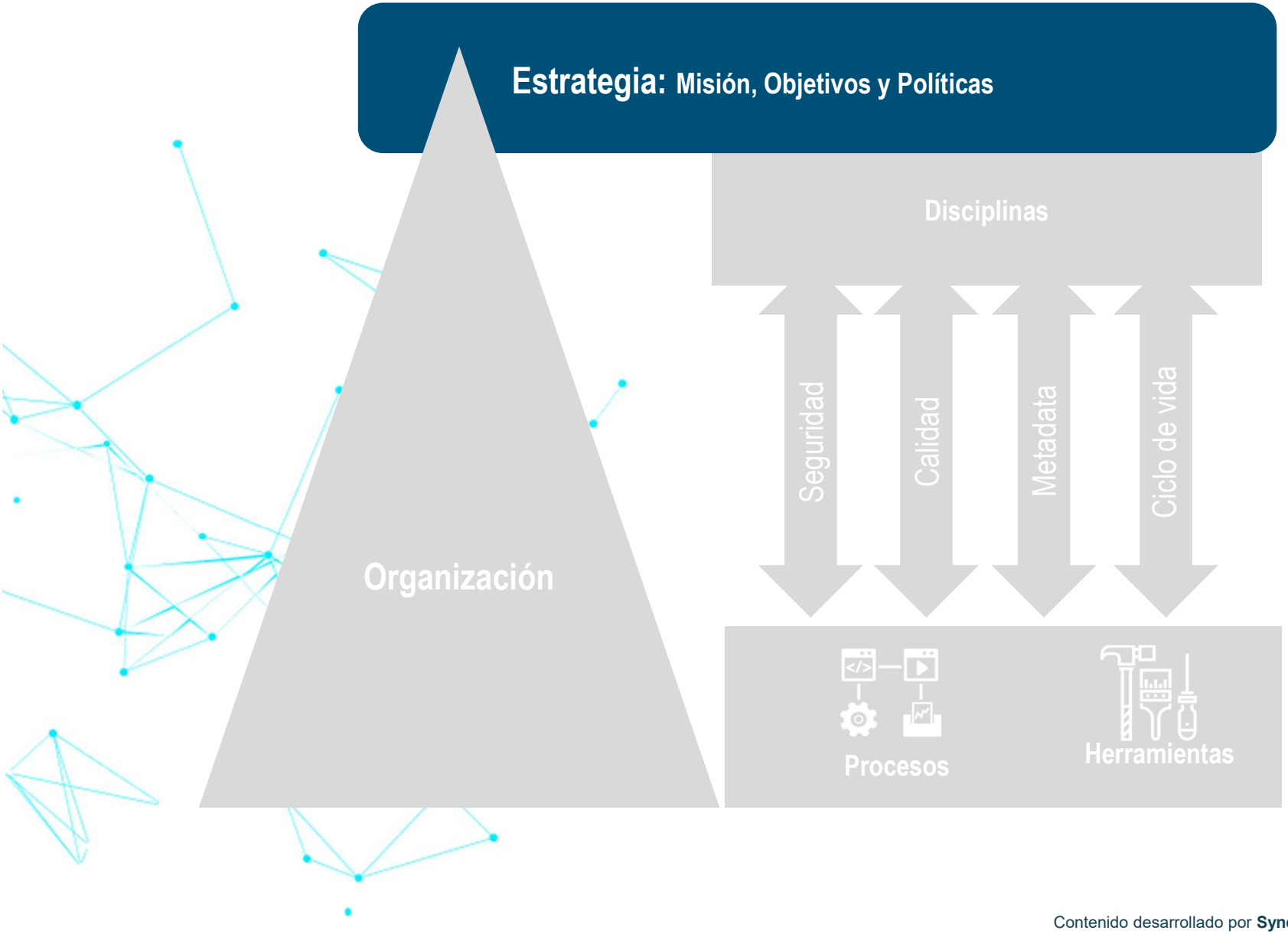
"Estructura Organizativa que crea y promueve Políticas y Procedimientos de Datos para su uso por negocio y por tecnología, a través de toda la Organización"

Fuente: *TDWI*

"Sistema que define las responsabilidades y deberes de cualquier proceso relacionado con los datos, en base a unas políticas existentes, las cuáles, describen quien puede hacer qué sobre qué datos y en qué circunstancias"

Fuente: *The Data Governance Institute*

Data Governance



Data Governance - Objetivos

Los propósitos que persigue el Gobierno de Datos (Data Governance) son:

- Asegurarse que los datos son siempre fiables y válidos en cada contexto empresarial.
- Que su calidad se mantiene a lo largo del tiempo.
- Que existen mecanismos de control sobre quién puede hacer qué con los datos en cada momento.

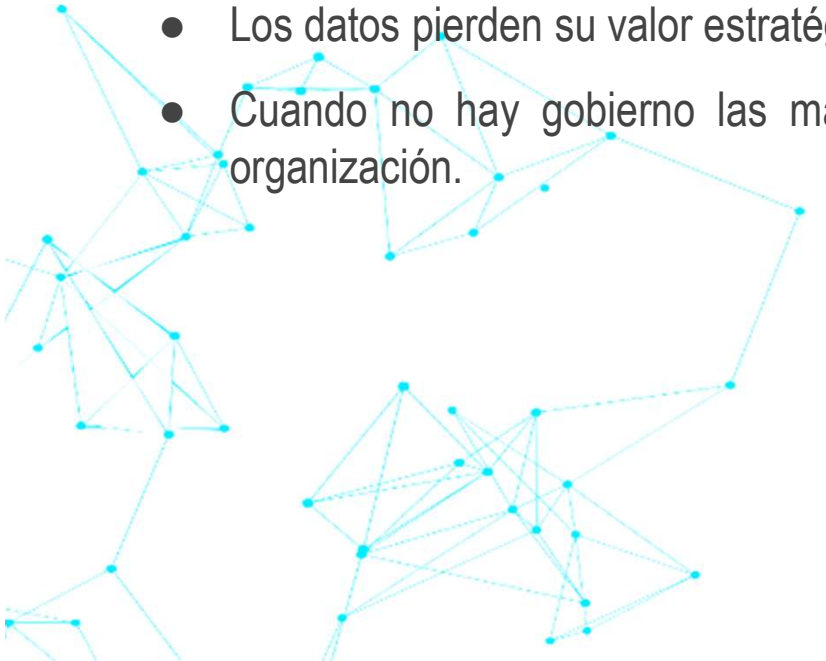
→ Todo ello con el objetivo de **apalancar los datos como un activo corporativo de gran valor empresarial.**

Para lograr estos objetivos es necesario establecer un conjunto de estándares, procesos y políticas que rijan el desarrollo y la utilización de los datos a nivel corporativo.

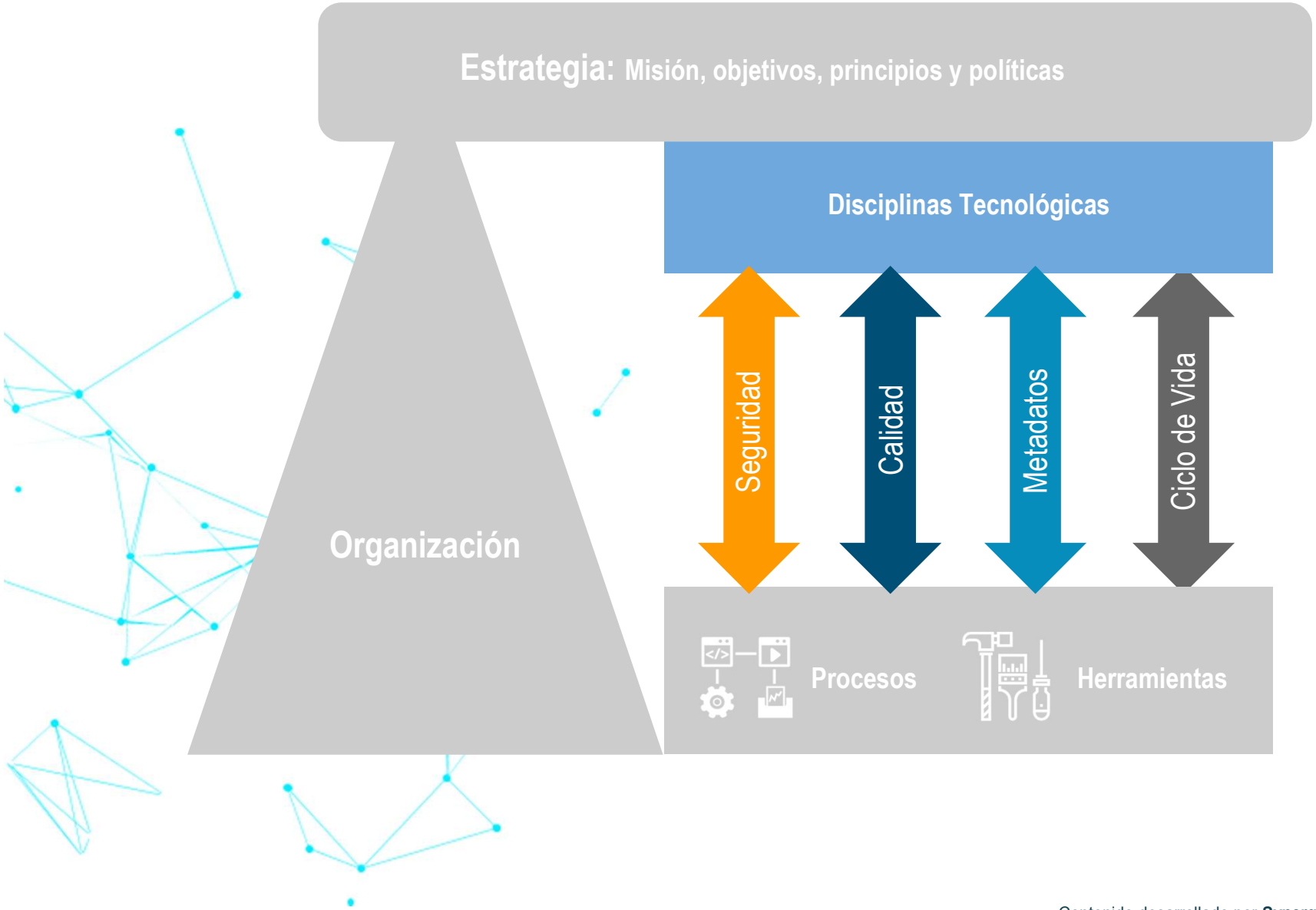
Consecuencias de no adoptar Big Data Governance

Son diversas en cuanto a naturaleza, gravedad y coste para la organización. Algunas de ellas son:

- La compañía se expone a sanciones económicas y/o pérdida de reputación.
- Se pierde productividad.
- Los datos pierden su valor estratégico para toma de decisiones.
- Cuando no hay gobierno las malas prácticas se vuelven costumbre dentro de la organización.



Data Governance



Data Governance - Disciplinas Tecnológicas

Seguridad

La **Seguridad** es uno de los **aspectos más importantes**. Como disciplina, la seguridad velará por garantizar los diferentes accesos autorizados a la plataforma y la información contenida en ella, impidiendo aquellos no consentidos.

Calidad

La **Calidad** está orientada a ofrecer información **fiable, correcta y de gran valor** para los consumidores. Para ello se deben establecer y cumplir un conjunto de procesos, estándares y buenas prácticas en cuanto al flujo y utilización de dicha información

GOBIERNO DEL DATO

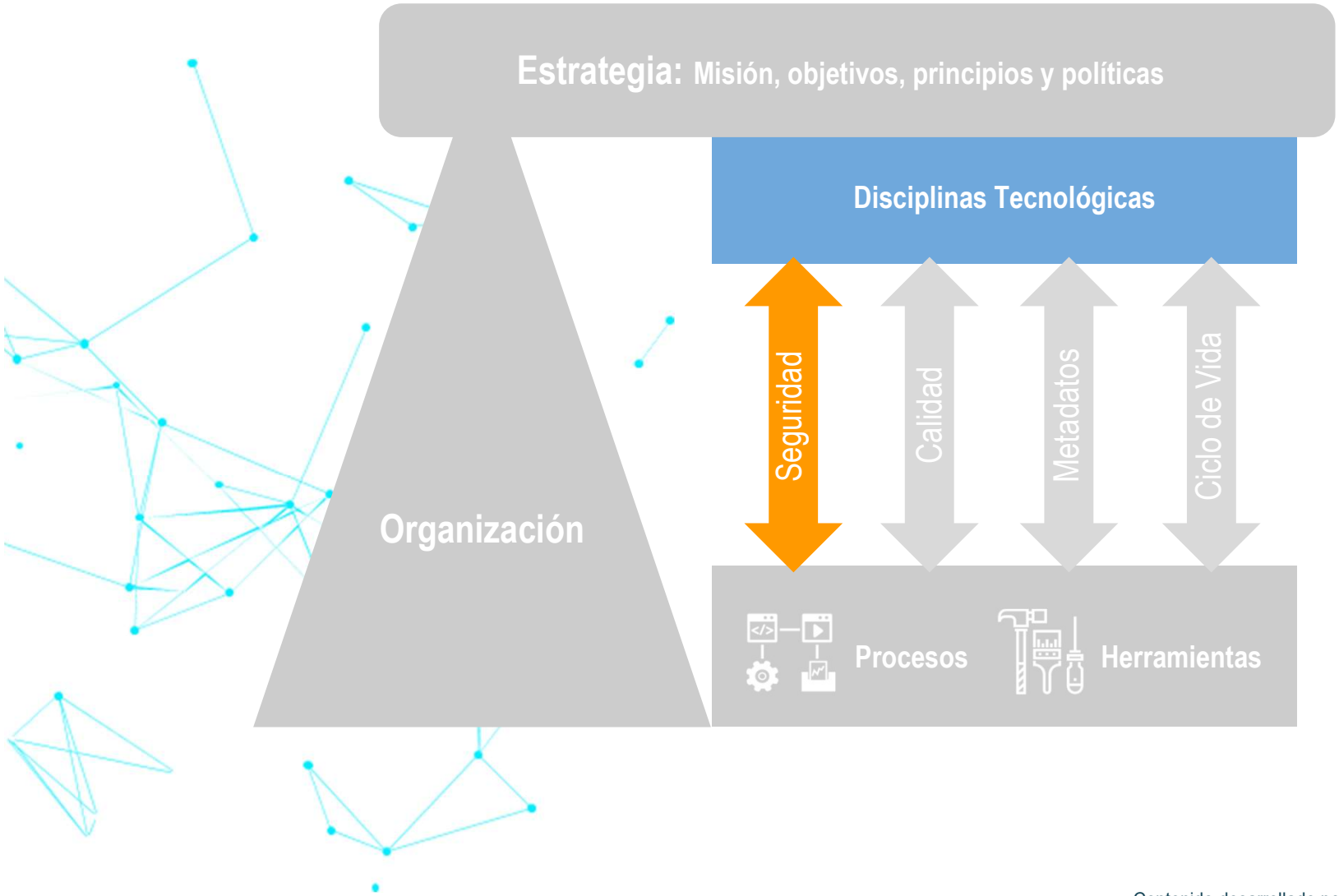
Ciclo de vida del dato

La gestión del **Ciclo de Vida Big Data** tiene como **objetivo optimizar el flujo de la información desde que se incorpora a la plataforma Big Data, hasta que se consume.** Todo ello con el objetivo de maximizar su utilidad para la compañía.

Metadatos

Los metadatos **proporcionan contexto a los datos y facilitan la comprensión de su significado.** Mantienen la **coherencia de los datos** utilizados, **optimizando los criterios de almacenamiento** y **facilitando la búsqueda de los datos.**

Data Governance - Disciplinas Tecnológicas



Data Governance - Data Security

El almacenamiento de grandes volúmenes de datos en infraestructuras **Big Data** las convierte en objetivos sensibles dentro de las empresas y pueden ser **mal utilizadas** deliberada o accidentalmente

Data Governance - Data Security
Autenticación

"**Autenticación** *[authentication]* es el proceso de verificar que un individuo, entidad o servidor es quien dice ser." (OWASP)

En este proceso se comparan las credenciales aportadas con las almacenadas en el sistema operativo o en un servidor de autenticación. Si coinciden, el proceso termina y se concede el acceso. Sistemas:

- LDAP
- Active Directory
- Kerberos

Nota: En la lógica del proceso la autenticación precede a la autorización, aunque a veces estos dos conceptos aparecen combinados.

Data Governance - Data Security
Autorización

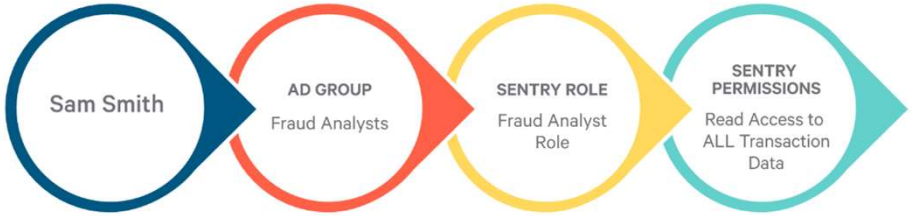
Control de acceso

"**Autorización** es el proceso por el cual el servidor determina si el cliente tiene permiso para usar determinado recurso o acceder a un archivo."

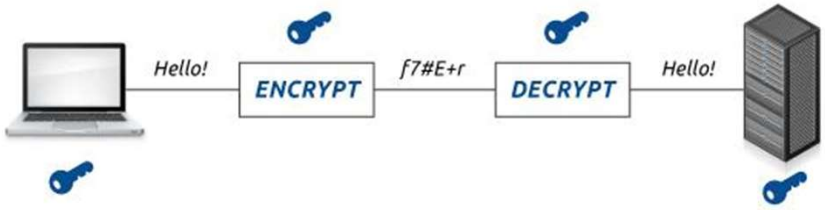
Una forma común de autorización son las **access control lists** (ACL), tablas que especifican qué derechos de uso tiene cada usuario o grupo de usuarios sobre un objeto.

File permissions in Linux

El sistema de permisos en los sistemas Linux, que definen tres derechos de uso (lectura, escritura y ejecución) y tres conjuntos (usuario, grupo y otros).



Data Governance - Data Security
Encriptación, anonimización y tokenización



Existen varios métodos para la protección de los datos que esencialmente buscan ofuscar la información para complicar su uso malintencionado:

- **Encriptación**: procedimiento que utiliza un algoritmo de cifrado para transformar un mensaje (por ejemplo, AES)
- **Hashing**: transformación irreversible, al contrario que la encriptación (por ejemplo SHA-256)
- **Tokenización**: transformación de un mensaje sin un procedimiento matemático que lo relacione con el contenido original

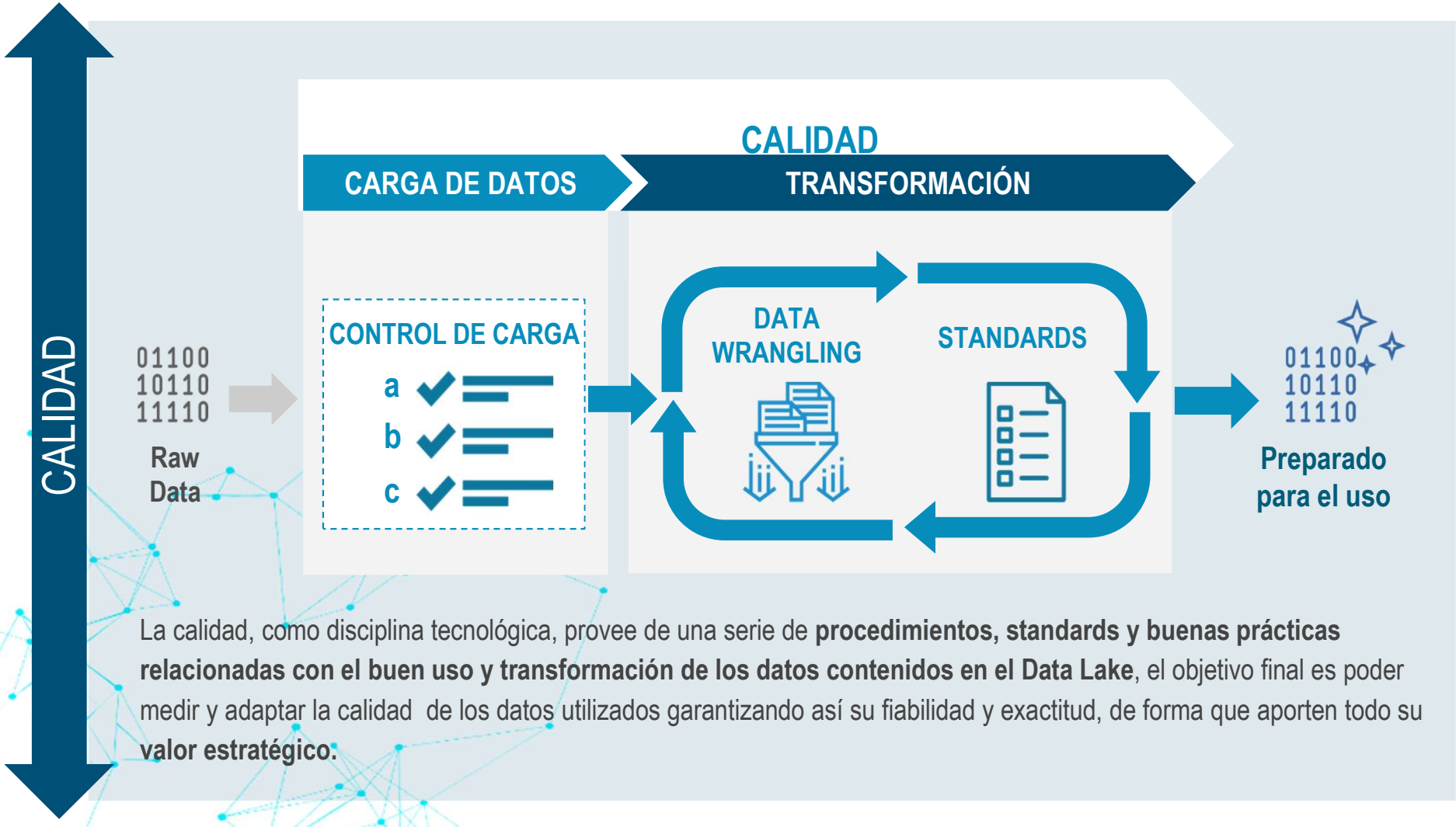
Data Governance - Disciplinas Tecnológicas

Data Governance - Data Quality

Data Quality (Calidad de datos) se refiere a los procesos, técnicas, algoritmos y operaciones encaminados a mejorar la calidad de los datos existentes en las bases de datos de empresas y organismos.

Los aspectos que se tienen en cuenta para valorar la calidad de los datos son: **precisión, completitud, actualización, fiabilidad, integridad, etc.**

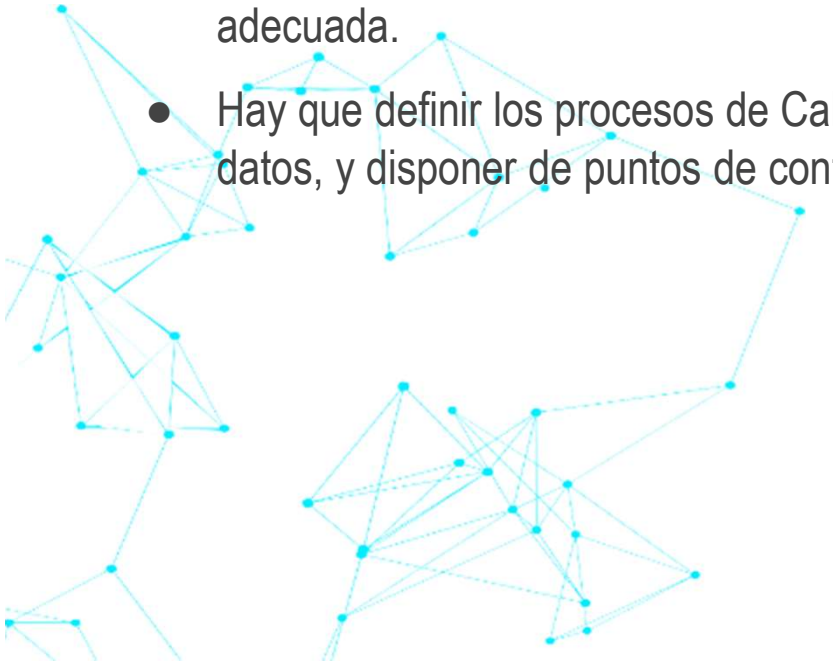
Data Governance - Data Quality



La calidad, como disciplina tecnológica, provee de una serie de **procedimientos, standards y buenas prácticas relacionadas con el buen uso y transformación de los datos contenidos en el Data Lake**, el objetivo final es poder medir y adaptar la calidad de los datos utilizados garantizando así su fiabilidad y exactitud, de forma que aporten todo su **valor estratégico.**

Data Governance - Data Quality

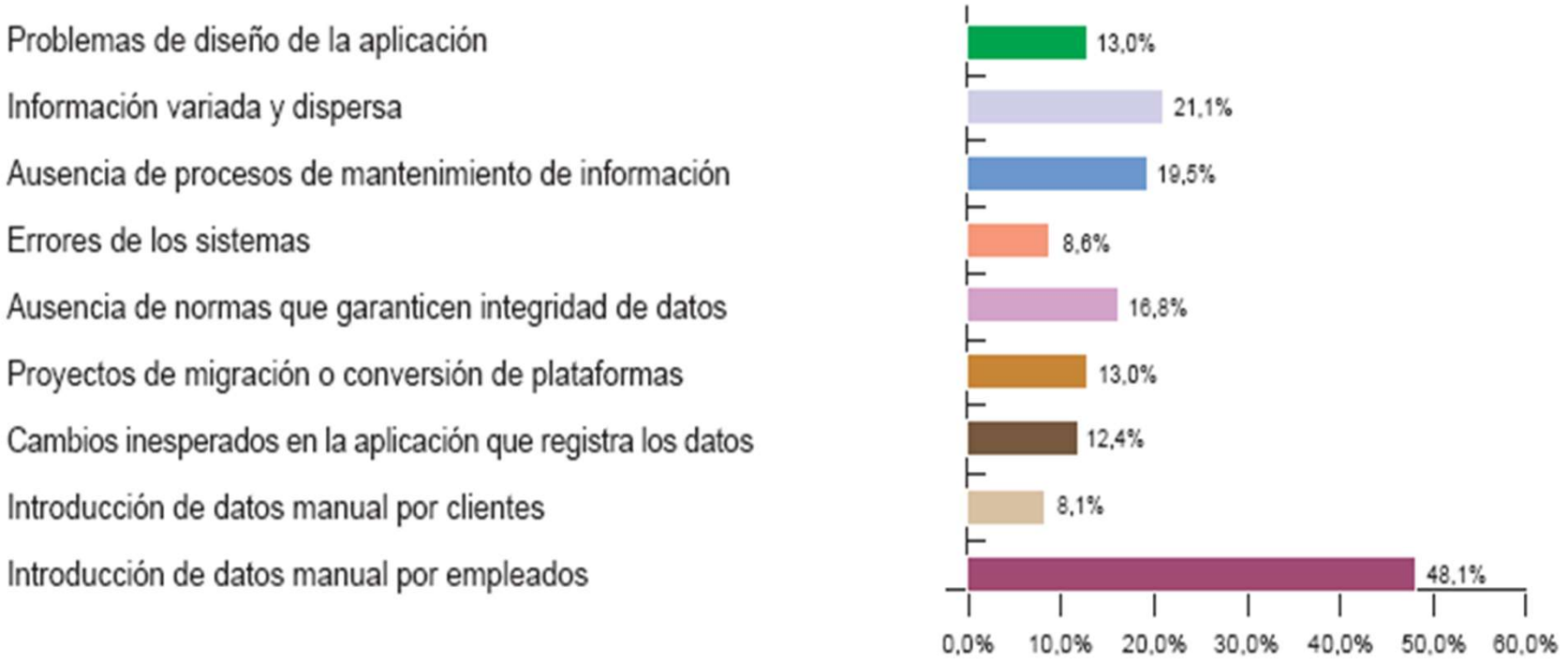
- El **aseguramiento de la calidad de los datos** es una parte integral de todo proyecto de datos.
- Se deben detectar los problemas de calidad durante todo el ciclo de vida para **identificar las prioridades y requisitos necesarios** para implantar la lógica empresarial más adecuada.
- Hay que definir los procesos de Calidad de Datos a lo largo de todo el ciclo de vida de los datos, y disponer de puntos de control en todos los puntos de contacto de la organización.



Data Governance - Data Quality

Principales causas de errores en la información

Hay diversos factores que inciden en la mala calidad de los datos:



Data Governance - Data Quality

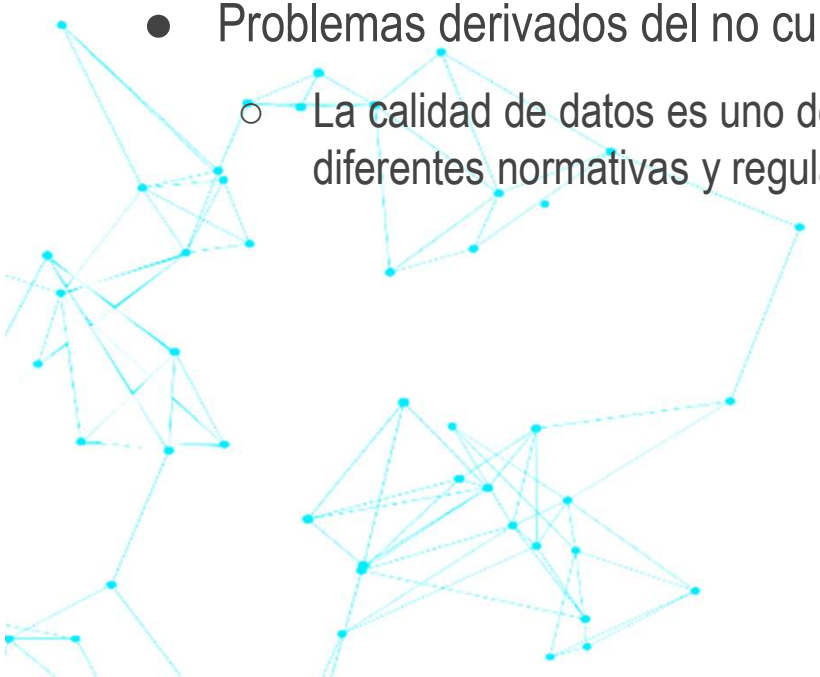
¿Cuáles son los síntomas de falta de calidad del dato?

- Desconfianza en el proceso de toma de decisiones
  - Imposibilidad de tomar decisiones de negocio sólidas debido a la falta de confianza en los datos
- Iniciativas de negocio que no alcanzan objetivos previstos
  - CRMs
  - Campañas de marketing
- No es posible conocer al cliente
  - Falta una visión unificada
  - La relación con el cliente se ve afectada: disminuye su satisfacción

Data Governance - Data Quality

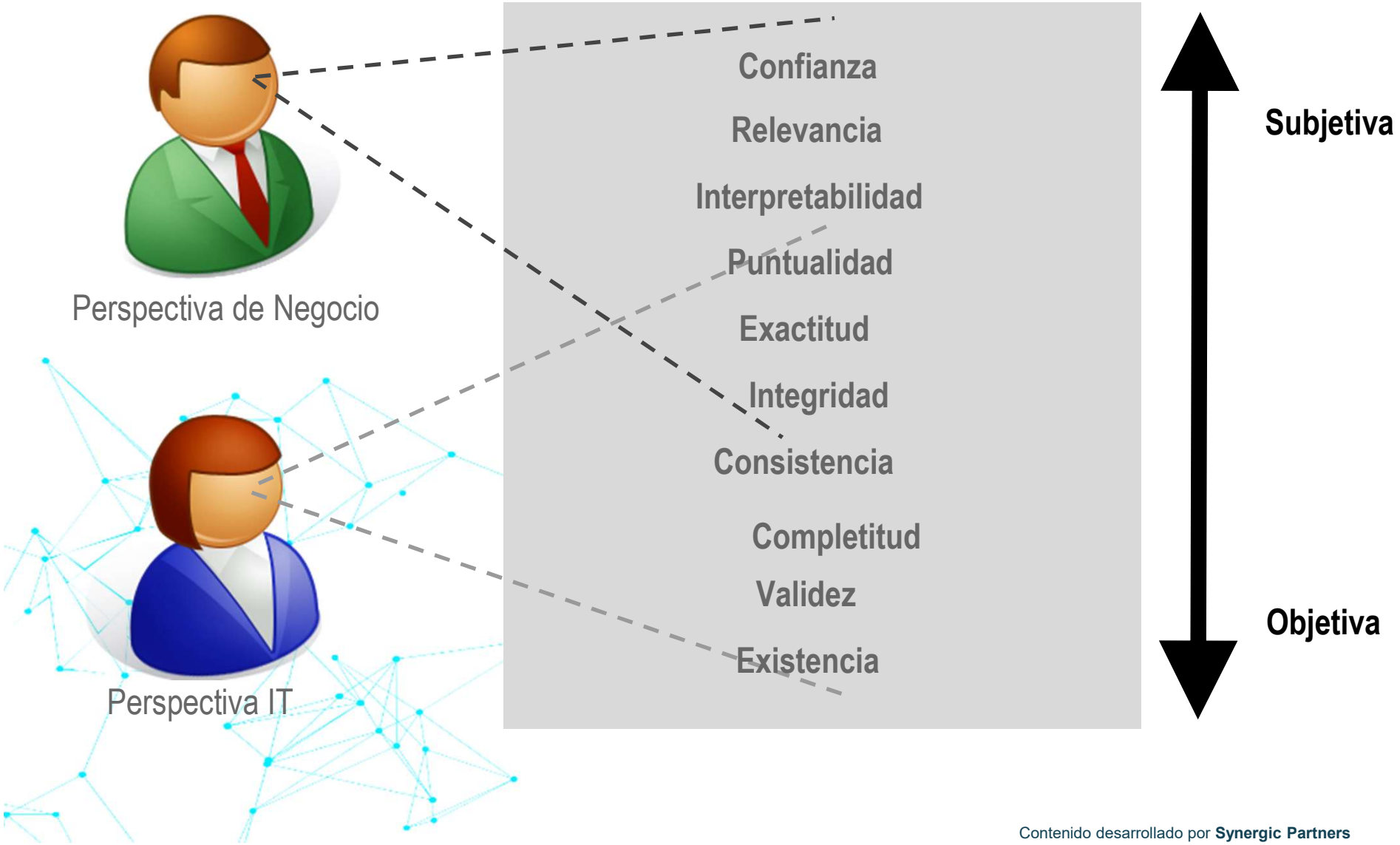
¿Cuáles son los síntomas de falta de calidad del dato?

- Aumenta la complejidad para detectar fraudes, sobrepagos, etc.
  - No se identifican duplicados, unidades familiares (households), relaciones entre empresas, etc.
- Problemas derivados del no cumplimiento normativo
  - La calidad de datos es uno de los pilares fundamentales de conformidad con las diferentes normativas y regulaciones (ej. LOPD)



Data Governance - Data Quality

Distintas visiones de lo que significa la Calidad de Datos



Data Governance - Data Quality
Etapas Proceso DQ II

A continuación se definen las diferentes etapas y técnicas de DQ:

Data Assessment - Profiling	Proceso de examinar los datos que existen en las fuentes de origen de una organización y recopilar estadísticas e información sobre los mismos.
Data Cleansing	Data Cleansing (depuración y limpieza de datos) es el proceso de detectar o descubrir y luego corregir datos corruptos, incoherentes o erróneos de un conjunto de datos.
Match and Consolidate	Detectar registros duplicados y consolidarlos en uno solo.
Data Auditing - Data Validation	Data Auditing es el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización.

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Todos los registros tienen todos los registros informados (414 with data) excepto **Longitud y Latitud** (2 no tienen)

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

1

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Código Local es la “Clave Principal” de la tabla: Tiene los 414 registros con dato y todos son distintos.

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

2

1

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Hay 12 registros que tienen al menos otro registro con el mismo **Nombre de Local** y 20 registros en que el **Nombre del Titular Explotación** iguales.

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

1

2

3

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Todos los locales pertenecen a un mismo tipo de local y provincia.

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

1 2 3 4

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Hay 5 registros que tienen el domicilio duplicado

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

1 2 3 4 5

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

En el campo población se han escrito hasta 70 poblaciones distintas

Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

1 2 3 4 5 6

Data Governance - Data Quality

Ejemplo de Profiling con Herramienta DQ I

Habría que investigar si las longitudes y latitudes duplicadas son correctas. (10 y 6 duplicados)

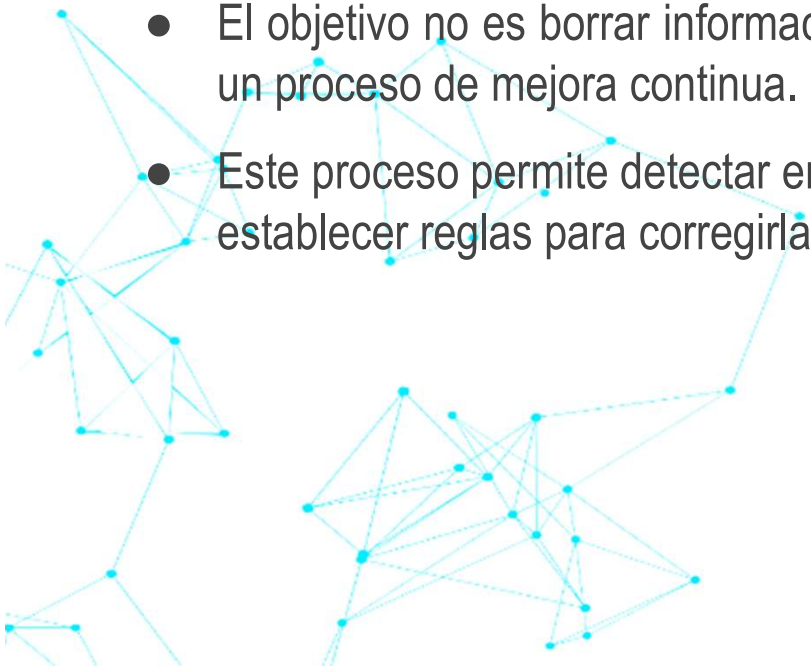
Input Field	Record Total	With Data	Without Data	Singleton	Duplicates	Distinct Values
Código Local	414	414	0	414	0	414
Nombre Local	414	414	0	402	12	408
Nombre Titular Explot.	414	414	0	394	20	404
Nombre Tipo Local	414	414	0	0	414	1
Nombre Provincia	414	414	0	0	414	1
Domicilio	414	414	0	409	5	411
Código postal	414	414	0	19	395	73
Población	414	414	0	22	392	70
Latitud	414	412	2	404	10	409
Longitud	414	412	2	408	6	411
Municipio	414	414	0	19	395	66
Población Total	414	414	0	19	395	66
Hombres	414	414	0	19	395	66
Mujeres	414	414	0	19	395	65

Data Governance - Data Quality

Etapas Proceso DQ - Data Cleansing I

Data Cleansing (Limpieza de datos) es el proceso de detectar o descubrir y luego corregir datos corruptos, incoherentes o erróneos de un conjunto de datos.

- Después del proceso la información será consistente con otros conjuntos similares de datos.
- El objetivo no es borrar información sino mejorar la calidad de los datos construyendo un proceso de mejora continua.
- Este proceso permite detectar entradas duplicadas, incompleta, erróneas, etc. y establecer reglas para corregirlas



Data Governance - Data Quality

Etapas Proceso DQ - Data Cleansing I

El proceso de Data Cleansing se centra básicamente en 3 acciones sobre los datos:

Parse (Separar)

Dividir un registro en partes según un patrón

Standardize (Estandarizar)

Transformar los datos a un formato común

Correct (Corregir)

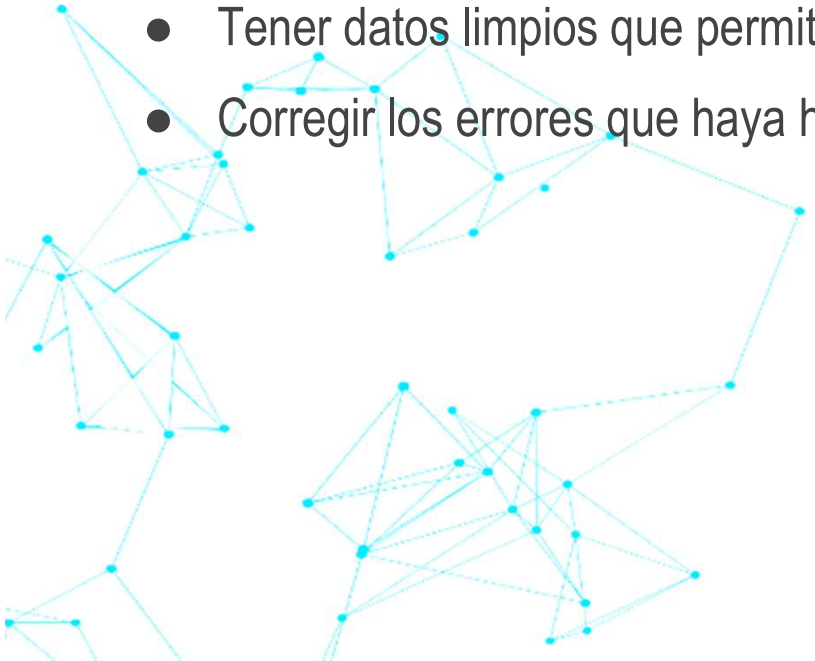
Eliminar errores sintácticos o semánticos de los datos

Data Governance - Data Quality

Etapas Proceso DQ - Data Cleansing II

Objetivos de aplicar Data Cleansing

- Tener la misma estructura de datos para cada campo.
- Tener datos limpios que permitan hacer análisis fiables.
- Corregir los errores que haya habido en la introducción datos.



Data Governance - Data Quality

Etapas Proceso DQ - Data Cleansing II

Ejemplos Data Cleansing:

A continuación se van a mostrar algunos ejemplos de las posibilidades que ofrecen las herramientas de Data Quality. Cada uno de estos procesos es automatizable para todos los registros.

La columna de la izquierda muestra el dato sin estandarizar y la de la derecha el dato después de haber sido estandarizado.

Población	Población estandarizada
MADRID	Madrid
MADRD	Madrid
Madri	Madrid
Madrid	Madrid
Madrid	Madrid
Mmadrid	Madrid
MAD	Madrid
Madrid Centro	Madrid

Unificar Nombres

Si no se unificaran, al hacer análisis se considerarían como poblaciones distintas.

Data Governance - Data Quality

Etapas Proceso DQ - Data Cleansing II

Ejemplos Data Cleansing:

A continuación se van a mostrar algunos ejemplos de las posibilidades que ofrecen las herramientas de Data Quality. Cada uno de estos procesos es automatizable para todos los registros.

La columna de la izquierda muestra el dato sin estandarizar y la de la derecha el dato después de haber sido estandarizado.

Nombre	Nombre estandarizado
Rodríguez, Juan	Juan Rodríguez
Pablo García	Pablo García
Álvaro Iniesta	Álvaro Iniesta
Alonso, Miquel	Miquel Alonso
Esteban, Francisco	Francisco Esteban
Font, Bernat	Bernat Font
Mónica Gónzalez	Mónica Gónzalez
Prats, Alfonso	Alfonso Prats

Ordenar según patrón

Detectar los registros en que el patrón es distinto y luego ordenarlo.

Data Governance - Data Quality

Etapas Proceso DQ - Match and Consolidate I

Detectar posibles duplicados

ID	Nombre	Fecha Nacimiento	Domicilio	Población	Código Postal
23	Pedro Gómez	10/11/1856	Avenida España 36	TERRASSA	8224
67	Sr. Pedro Gómez	10/11/1856	AV ESPAÑA	Terrassa	8224
84	PEDRO GOMEZ	10/11/1856	Avenida España 36	Terrassa	8224

Unificar registros en uno solo

ID	Nombre	Fecha Nacimiento	Domicilio	Población	Código Postal
23	Pedro Gómez	10/11/1856	Avenida España 36	Terrassa	8224

Data Governance - Data Quality

Etapas Proceso DQ - Match and Consolidate II

Matching - Reglas de Comparación:

Para determinar los registros que se consideran duplicados la técnica de *matching* puede utilizar **reglas estrictas** (que el dato sea exactamente igual) **o mediante el uso de *fuzzy logic*** (reglas de comparación difusa que internamente llevan algoritmos matemáticos).

Algunas de estas reglas que permiten comparar y encontrar registros duplicados son:

- Porcentaje del campo idéntico (>70%)

ID	Domicilio
34	Calle San Pedro 23
569	Calle San Pedro 23 ATICO

- Empiezan idénticamente n valores (>10)

ID	Nombre
344	PEDRO GARCÍA
445	PEDRO GARCÍA IGLESIAS

Data Governance - Data Quality

Etapas Proceso DQ - Match and Consolidate III

Matching - Reglas de Comparación (cont.)

- Terminan idénticamente n valores:

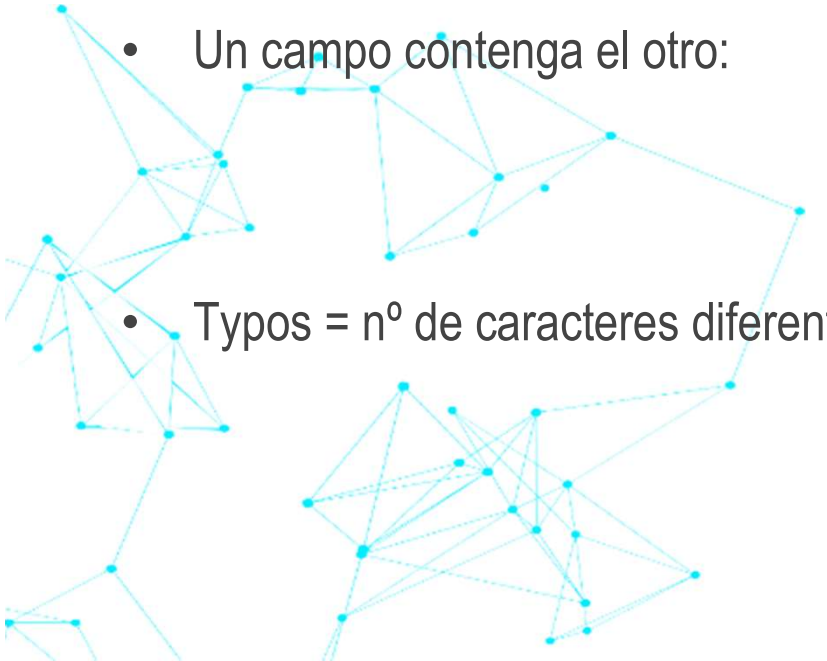
ID	Dispositivo
435	2007-GALAXY II
23	2009- GALAXY II
234	2014- GALAXY II

- Un campo contenga el otro:

ID	Nombre y apellidos
546	Sra. Mª Carmen Esteban Cruz
1234	Carmen Esteban

- Typos = nº de caracteres diferentes (<3)

ID	Población
34	GETAFE
285	GGETAFE



Data Governance - Data Quality

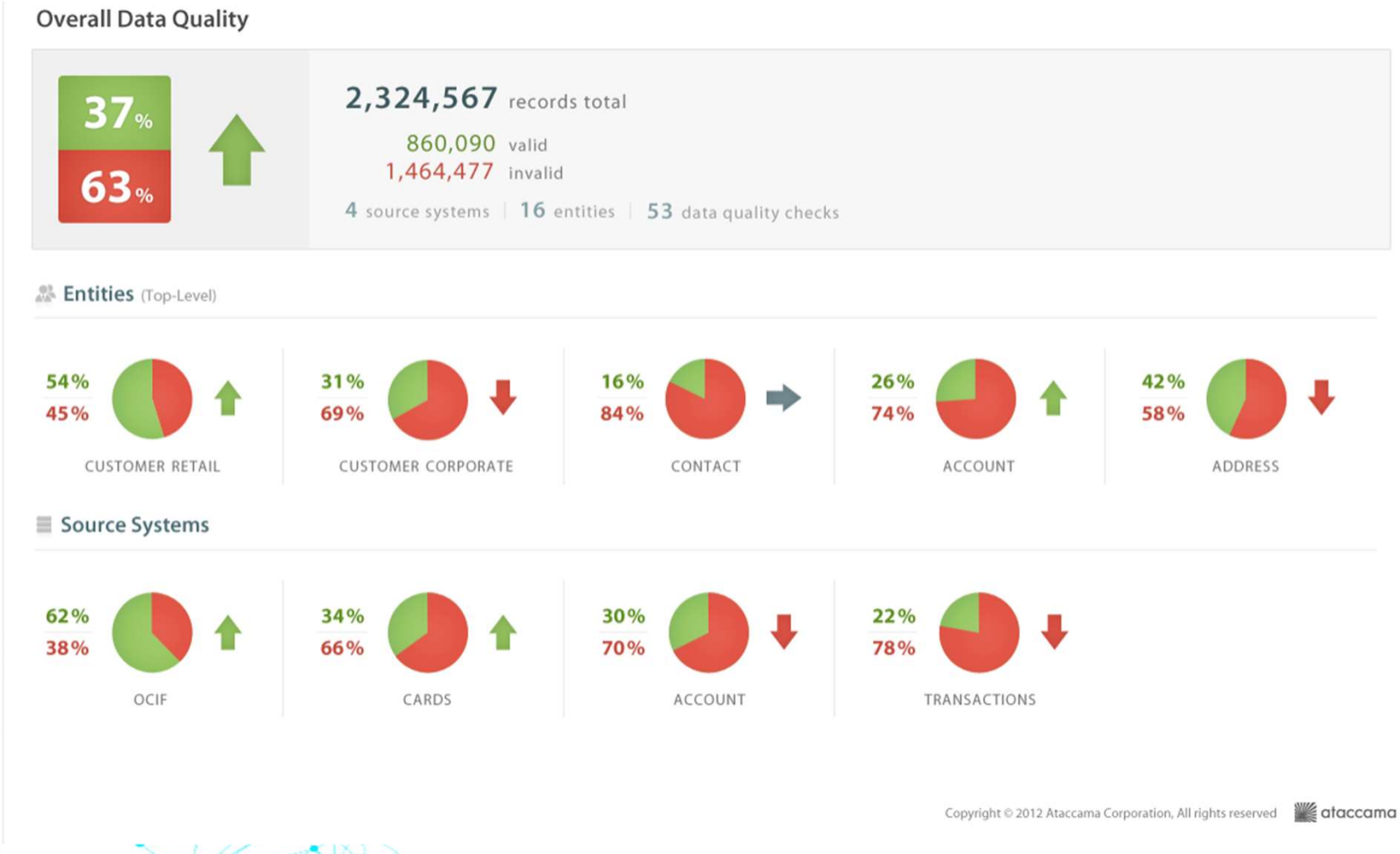
Etapas Proceso DQ - Data Auditing I

Data Auditing es el proceso de gestionar cómo los **datos** se **ajustan** a los **propósitos** definidos por la organización.

Se establecen **políticas** para gestionar los criterios de datos para la organización. No es suficiente con actuar sino que se debe **vigilar**.

Data Governance - Data Quality

Etapas Proceso DQ - Data Auditing II



Data Governance - Disciplinas Tecnológicas

DATA ABOUT DATA

"Datos que describen a otros datos"

Metadata de un canción

- Título: Beat it
- Duración: 4:18
- Artista: Michael Jackson
- Álbum: Thriller
- Género: Pop
- Mi rating: 5 stars
- Nº de reproducciones: 32
- Última reproducción: 2015/10/26 2:34 PM

Data Governance - Gestión de Metadatos

Metadata

Tipos de Usuario

- Usuarios analíticos y de negocios
- Data Owners y Data Stewards
- Equipo de Ingeniería de datos

Tipos de Metadatos

- Metadatos de Negocio
- Metadatos técnicos y funcionales
- Metadatos Operativos

Linaje del dato

El linaje incluye información sobre:

- El origen de los datos
- Las transformaciones sufridas por los datos

Etiquetado de datos

Las etiquetas describen el contenido de los datos y facilitan su localización en el Data Lake

Data Governance - Gestión de Metadatos

TIPO DE METADATOS

Metadatos de Negocio: Informan acerca del contexto del dato dentro de la compañía: tipo de dato (contrato, número de cuenta, fecha de alta, etc.), definición, quién es el *dueño* y su responsable (Data Owner).

Metadatos Técnicos: Dan información acerca del contenido y origen de los datos. Esta categoría abarca el nombre de la tabla, su origen (o linaje), campos en la tabla y tipo de datos (fecha, numérico, etc.).

Metadatos Operacionales: Incluye información acerca del uso de los datos, por ejemplo, cuando fue la última vez que se actualizó, cuántas veces ha sido accedido, cuándo fue la última vez y por quién.

Data Governance - Disciplinas Tecnológicas

Data Governance - Gestión de Metadatos

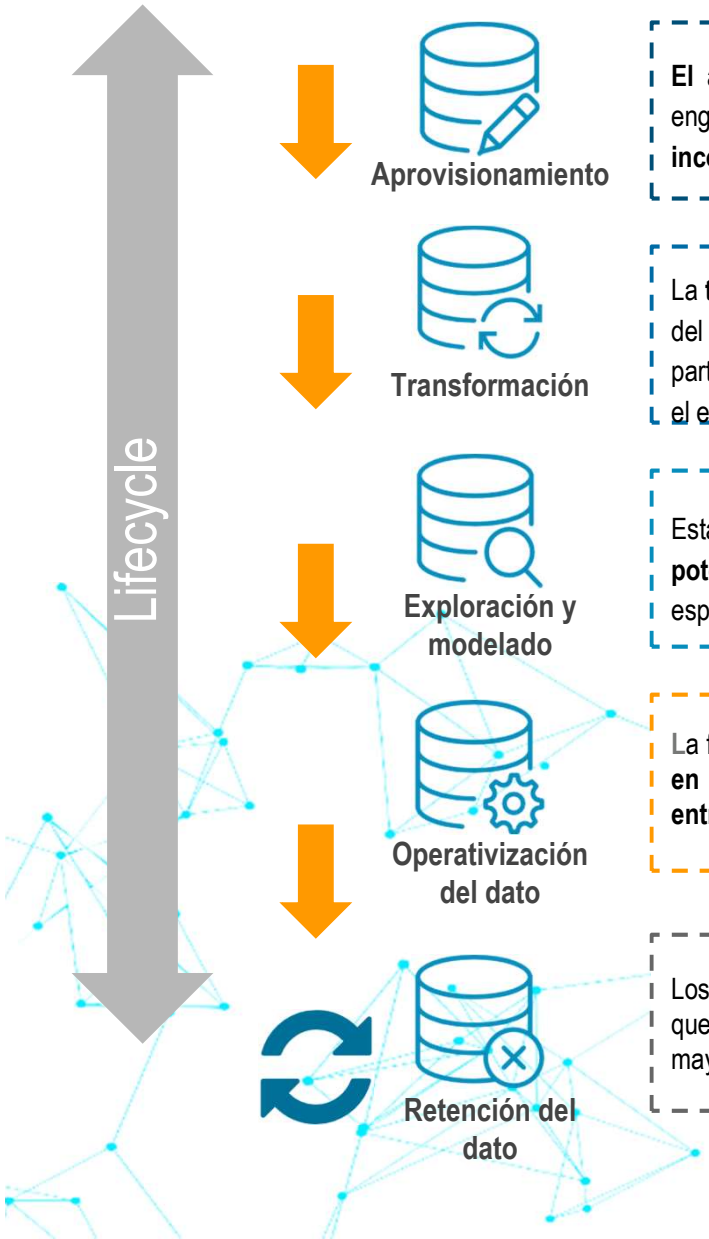
Linaje de Datos

Traza del **CICLO DE VIDA** del dato a través de los sistemas por los que circula y las operaciones que son realizadas sobre él:

- ¿De dónde procede el dato?
- ¿Qué transformaciones ha sufrido?
- ¿Cuál es su destino?

El linaje dibuja la **trazabilidad** completa del **dato** e incluye información acerca del **origen** del dato, así como las **transformaciones** que ha ido sufriendo durante su procesamiento y uso. Es fundamental conocer la trazabilidad de cada uno de los datos que se usan para tomar decisiones estratégicas, en caso contrario, no se puede garantizar que el dato es correcto y, por tanto, confiable.

Data Governance - Ciclo de vida del dato



El aprovisionamiento hace referencia a la inclusión de datos en el Data Lake. Esta fase engloba las **acciones** que se llevan a cabo **desde que se identifica la necesidad de incorporar determinados datos al Data Lake**, **hasta que finalmente estos son incluidos.**

La **transformación** hace referencia a todas **las modificaciones** o adecuaciones sobre los datos del Data Lake desde una perspectiva de **calidad y seguridad.** Las transformaciones se harán a partir de los datos ingestados en el Data Lake o generados por modelos/procesos ejecutados en el entorno Big Data.

Esta fase hace referencia a todas las **actividades que se realizan de estudio sobre la potencialidad de los datos**, así como el análisis de cómo extraerles el valor mediante funciones específicas o modelos que lo habiliten.

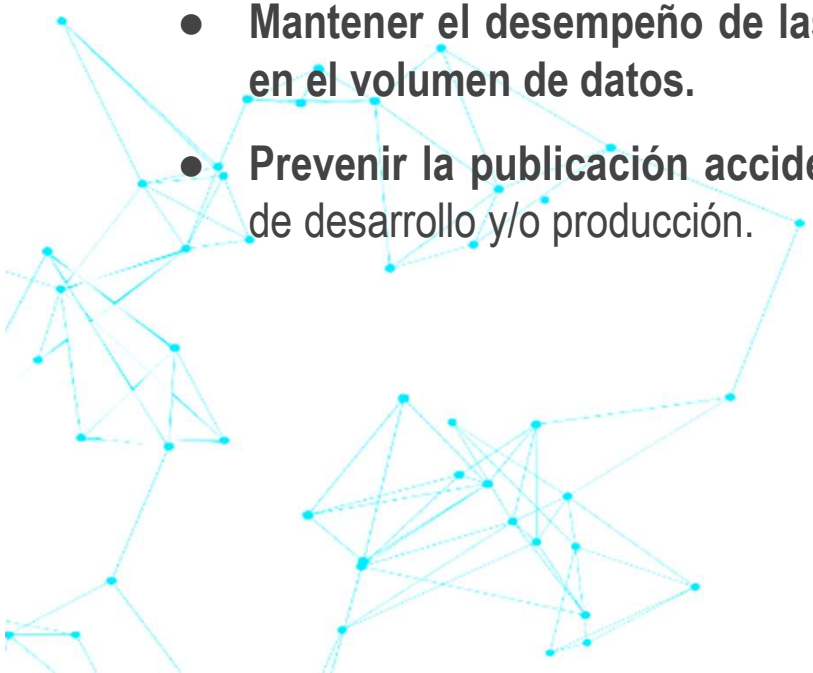
La fase de operativización engloba las **acciones** que hay que acometer cuando se decide **poner en producción modelos o procesos de mejora de eficiencia en relación a sus datos de entrada y de salida**.

Los datos han de **ser usados** mientras sigan siendo **productivos** y se desechan cuando han quedado **sin valor**. El reto esta fase es identificar qué datos ya no resultan útiles, ya que, la mayoría de los datos **pierden parte de su utilidad** con el paso del tiempo

Data Governance - Ciclo de vida del dato
Importancia

La instauración de **esta disciplina** de Gobierno del Dato para plataformas Big Data **es imprescindible a la hora de** enfrentar los siguientes retos:

- **Desarrollar, validar y productivizar aplicaciones analíticas** en el menor tiempo posible.
- **Mantener el desempeño de las aplicaciones** informáticas **a pesar del incremento en el volumen de datos.**
- **Prevenir la publicación accidental o intencional de datos sensibles** en entornos de desarrollo y/o producción.



Data Governance - Organización