



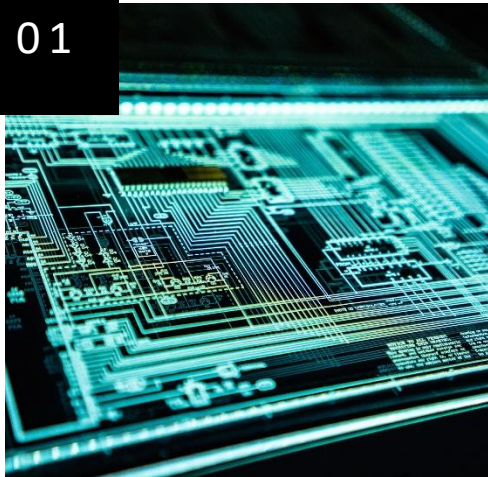
Módulo 2

Data Governance

Ciclo de vida y gobierno del dato dentro de
la organización

La IA en el mundo empresarial

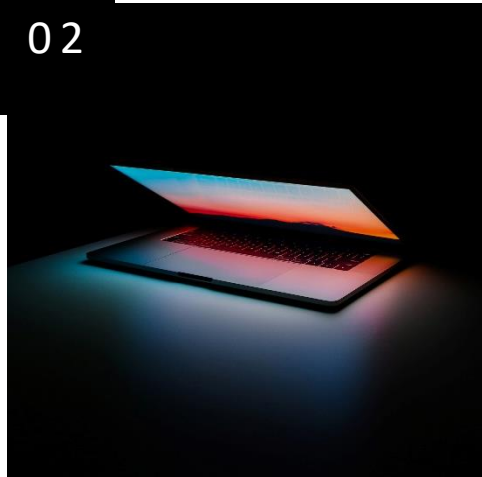
01



Introducción

La necesidad de gobierno del dato dentro de la organización

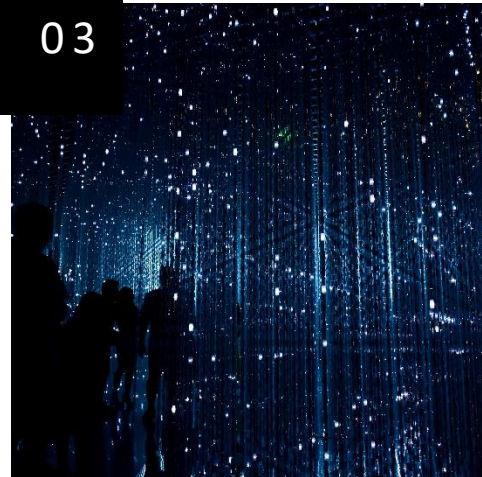
02



Data Governance I

Fundamentos básicos de Data Governance

03



Data Governance II

La calidad y disponibilidad de los datos, integración con los procesos de la organización

04



Data Governance III

Data Mesh y Data as a Product

Introducción

La necesidad de un correcto gobierno de los datos dentro de la organización

En la actualidad, las organizaciones manejan grandes volúmenes de datos, lo cual plantea diversos retos y oportunidades. En este contexto, exploraremos por qué el gobierno de datos es crucial para gestionar efectivamente esta información, apoyando la toma de decisiones y asegurando el cumplimiento normativo en un entorno digital complejo.



¿Por qué es tan importante?

1. Toma de Decisiones Informada:

- Un gobierno de datos efectivo asegura la calidad y accesibilidad de la información, posibilitando decisiones estratégicas fundamentadas y reduciendo riesgos asociados a datos inexactos o incompletos.

2. Cumplimiento Normativo:

- La gestión rigurosa de los datos ayuda a cumplir con regulaciones y leyes de protección de datos, mitigando posibles sanciones y fortaleciendo la confianza de los stakeholders.

3. Optimización Operativa:

- Facilita la operatividad eficiente, permitiendo que los procesos de negocio se ejecuten de manera más fluida y coherente mediante el uso de datos consistentes y bien gestionados.

4. Seguridad de la Información:

- Protege la información organizacional de accesos no autorizados y filtraciones, garantizando la confidencialidad, integridad y disponibilidad de los datos.

5. Innovación y Competitividad:

- Habilita la innovación al proporcionar una base de datos fiable y accesible para el desarrollo de nuevos productos/servicios y mejora la competitividad mediante insights basados en datos que impulsan estrategias orientadas al mercado.

¿Qué estilos de gobernanza de datos existen?

1. **Gobernanza de datos centralizada:** las organizaciones centrales son en última instancia responsables de las declaraciones de misión, las políticas, la elección de herramientas y más. Las acciones del día a día muchas veces se transfieren a las líneas de negocio (LOB).
2. **Gobernanza de datos federada:** esta permite que las unidades empresariales individuales o las iniciativas empresariales operen de la manera que mejor se adapte a sus necesidades. Con la gobernanza de datos federada, todavía hay un equipo centralizado más pequeño que dedica su trabajo a resolver los problemas que se repiten con más frecuencia, incluidas las herramientas de calidad de datos para toda la empresa, por ejemplo.
3. **Gobernanza de datos descentralizada o de autoservicio:** cada LOB hace lo que necesita para su proyecto específico. Cada proyecto utiliza cualquier herramienta o proceso de otros proyectos que se ajuste a sus necesidades. A medida que aumentan en popularidad temas como la malla de datos (en sí misma descentralizada), también lo hace la gobernanza de datos de autoservicio.

Tema 1

Ciclo de vida y gobierno del dato dentro
de la organización I: Fundamentos
básicos de data governance

Objetivos

01

Introducir la relevancia y la criticidad de disponer de un modelo de Data Governance robusto para la organización, así como los beneficios y oportunidades que proporciona

02

Presentar los principales elementos del modelo de data governance, haciendo foco sobre los componentes básicos, los conceptos clave, los roles y responsabilidades y el modelo de relación con el resto de roles dentro de la organización

03

Proporcionar contexto y ejemplos de implantación exitosa en diversas organizaciones.

ÍNDICE

01: Roles del modelo básico de data governance

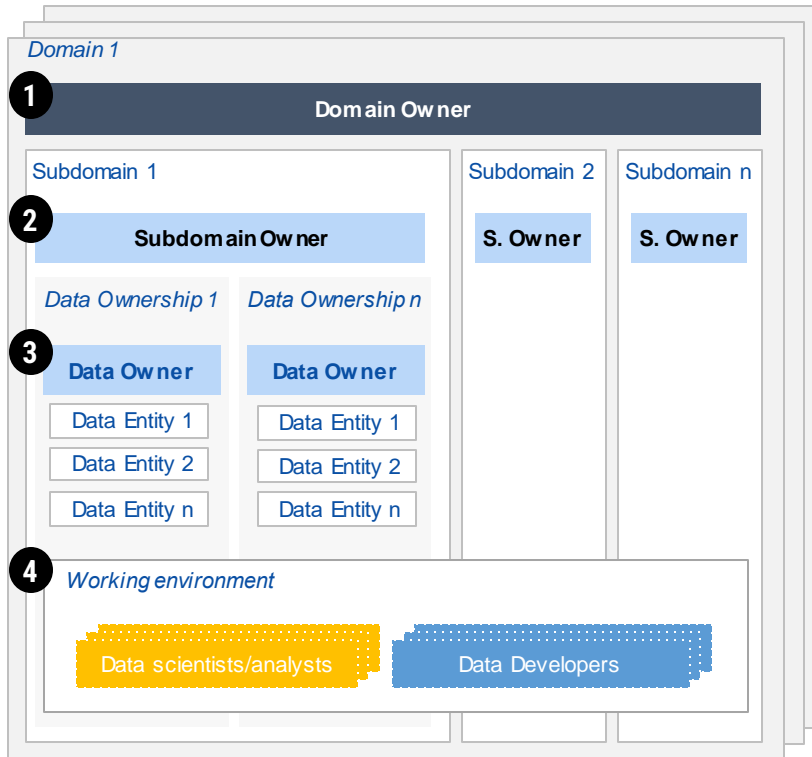
02: Conceptos clave

03: Órganos de gobierno y procesos

04: Cultura del dato

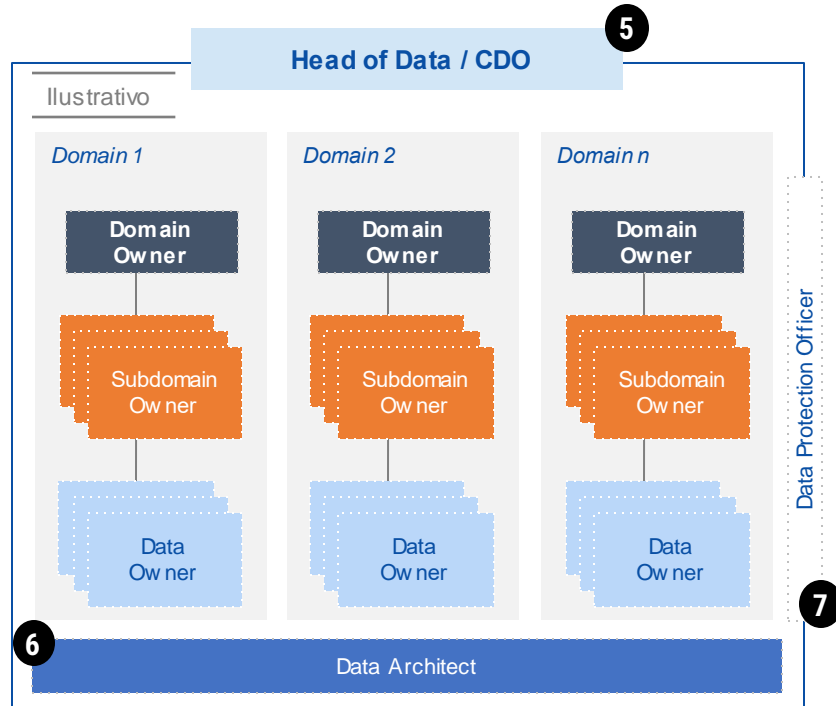
Roles básicos (1/2)

Ilustrativo



- 1 **Domain owners** son responsables del **gobierno de dominios** de datos (subdomains correlacionados)
- 2 **Subdomain owners** son responsables del **gobierno de los subdominios de datos** (data ownerships correlacionados)
- 3 **Data owners** son **responsables** de la definición, calidad, trazabilidad y disponibilidad de sus datos
- 4 Cuando los data owner necesitan analizar, descubrir o desarrollar soluciones, cuentan con **entornos de trabajo** en los que:
 - **Data scientists/ analysts:** colaboran para crear soluciones analíticas/modelos de datos y/o data insights
 - **Data developers:** responsables de aprovisionar datos, desarrollar soluciones e implantarlas en producción

Roles básicos (2/2)



- 5 El **Head of Data o Chief Data Officer** es responsable de fortalecer el **modelo de data governance** en todo el grupo. Igualmente, se encargará de:
 - **Definir la estrategia global de datos**
 - **Democratizar el acceso a los datos** en el Grupo
 - Promover y operar una **“data” community of practice** global
 - **Asegurar el cumplimiento** de las regulación existente
 - Crear y propagar una **“cultura de dato”** homogénea **en el Grupo**
- 6 **Data Architect** provee la **arquitectura e infraestructura de datos** necesaria para soportar las **necesidades del negocio**
- 7 **Data Protection Officer** define políticas y directrices para ayudar y aconsejar a los usuarios de datos a diseñar e implementar **controles adecuados**

El rol de un data owner

Principal Función

Responsable de la correcta **definición, calidad, trazabilidad y disponibilidad de sus datos**, así como del **cumplimiento de las políticas y procedimientos** de data governance

Responsabilidades

- 1 **Asegurar** que los **datos** bajo su responsabilidad **cumplen con el modelo de Data Governance** en términos de:
 - **Contenido:** siendo responsables de la **definición funcionalmente de sus datos** y dando el visto bueno de los metadatos técnicos definidos por los equipos de engineering
 - **Calidad:** siendo responsables de **definir las métricas funcionales de calidad**, asegurando que se operativizan en la herramienta corporativa y monitorizando sus resultados en cuadros de mando de calidad
 - **Lineage:** responsable de **documentar la trazabilidad funcional** del dato, **identificando las TDS y SoR de las que provienen sus datos y los usos que se dan de los mismos** (hasta lo que conozca). Adicionalmente se deben asegurar que en el futuro se implemente en una herramienta específica de trazabilidad (trazabilidad técnica).
- 2 **Identificar** los **procesos críticos y dependencias** de los datos que los componen en su ámbito de responsabilidad
- 3 **Definir un plan detallado para garantizar que se alcanzan los objetivos definidos**

Ejemplos

Todos los datos cuentan con una definición funcional en el diccionario de datos

Los datos poseen al menos un indicador de calidad definido, implementado en una herramienta y monitorizado

Se documenta la trazabilidad funcional para todos los datos: TDS, SoR, reportes en los que se usa, etc.

Documentar los principales procesos, identificando dependencias con otras áreas

Plan detallado para alcanzar las tareas definidas en cada dimensión de trabajo

El rol de un data owner (II)

Otras responsabilidades del data owner

- Revisar semanalmente los **dashboards del subdomain** y de los **componentes del data governance**
- Monitorizar nuevos **requerimientos de datos**, indentificando los **procesos críticos** para data ownership
- Asegurar que la asignación de data entities cumplen con las **leyes y regulaciones de políticas de datos internas y externas, estándares y protección de datos** - revisar e implementar las acciones necesarias para asegurar el cumplimiento
- Interactuar y apoyar al group/local data owners, proporcionando orientación y criterios para garantizar la homogeneidad de la definición y gestión de los datos: definir el mecanismo de gobernanza, los criterios, el escalamiento y flujo, etc.
- Colaborar con el Domain owner en la definición del plan anual del dominio y escalar al subdomain/domain owner las necesidades y requerimientos de datos
- Monitorizar el progreso de la implementación del plan– identificando incidencias, barreras, dependencias, etc.
- Administrar y resolver problemas (e.g. ajustes manuales, nuevos criterios requeridos, etc.) para garantizar el correcto gobierno de sus entidades– mantener un registro de problemas (inventario)
- Supervisar / estar a cargo de métricas de datos para reportar al subdomain/domain owner
- Fomentar la **cultura basada en datos** en toda la organización: ser proactivo y participar en capacitaciones

ÍNDICE

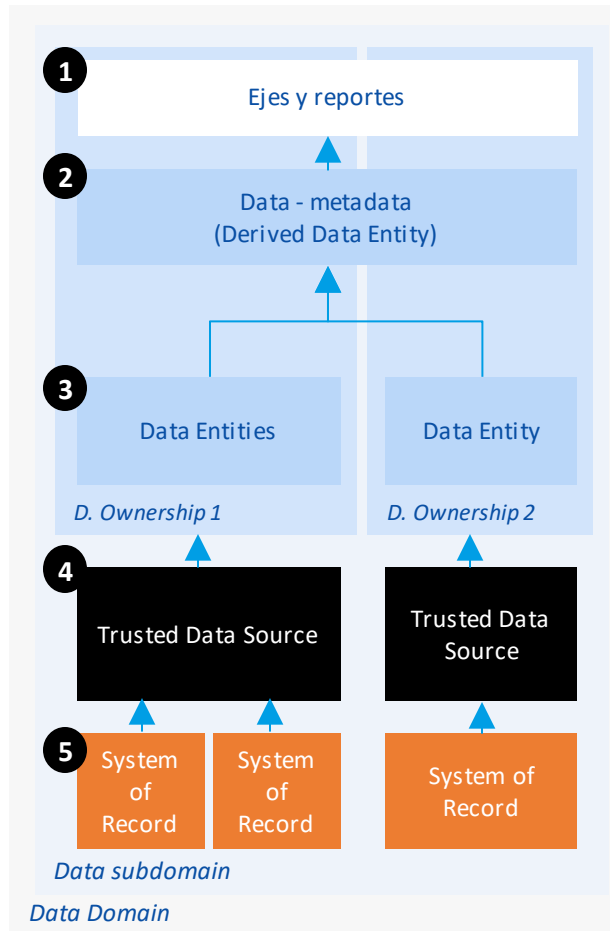
01: Roles del modelo básico de data governance

02: Conceptos clave

03: Órganos de gobierno y procesos

04: Cultura del dato

Conceptos básicos a considerar



- 1 Los datos pueden ser agregados o visualizados por diferentes ejes (temporal, geográfico, asset class,...) que generan diferentes vistas del mismo dato.
Estos datos se visualizan en informes/reportes con un diversas finalidades
- 2 Por dato entendemos cualquier **objeto, concepto o evento** para el que la organización **considera valioso almacenar información**. Existen **datos puros** (raw) y **derivados** (procesados)
- 3 Los datos se pueden **categorizar en diferentes tipologías**. Estas categorizaciones se denominan **entities**, y son el **mínimo nivel de granularidad** utilizado en éste modelo para **asignar responsabilidad** (ownership)
- 4 A su vez, **las data entities** se obtienen de fuentes de datos. Cuando estas fuentes **son consideradas como las únicas válidas y fiables** para obtener los datos se consideran “fiables” o **Trusted Data Sources (TDS)**
- 5 Estos TDS, se aprovisionan de los **system of record**, que suponen la fuente de aprovisionamiento fiable de información de forma **completa, exacta y actualizada**. **Importancia del reconocimiento e identificación de los shadow ITs.**

Por último, los dominios de datos se refieren a un **set de datos que puede abarcar un área completa o un set específico de datos** (evitando el solapamiento de información entre dominios)

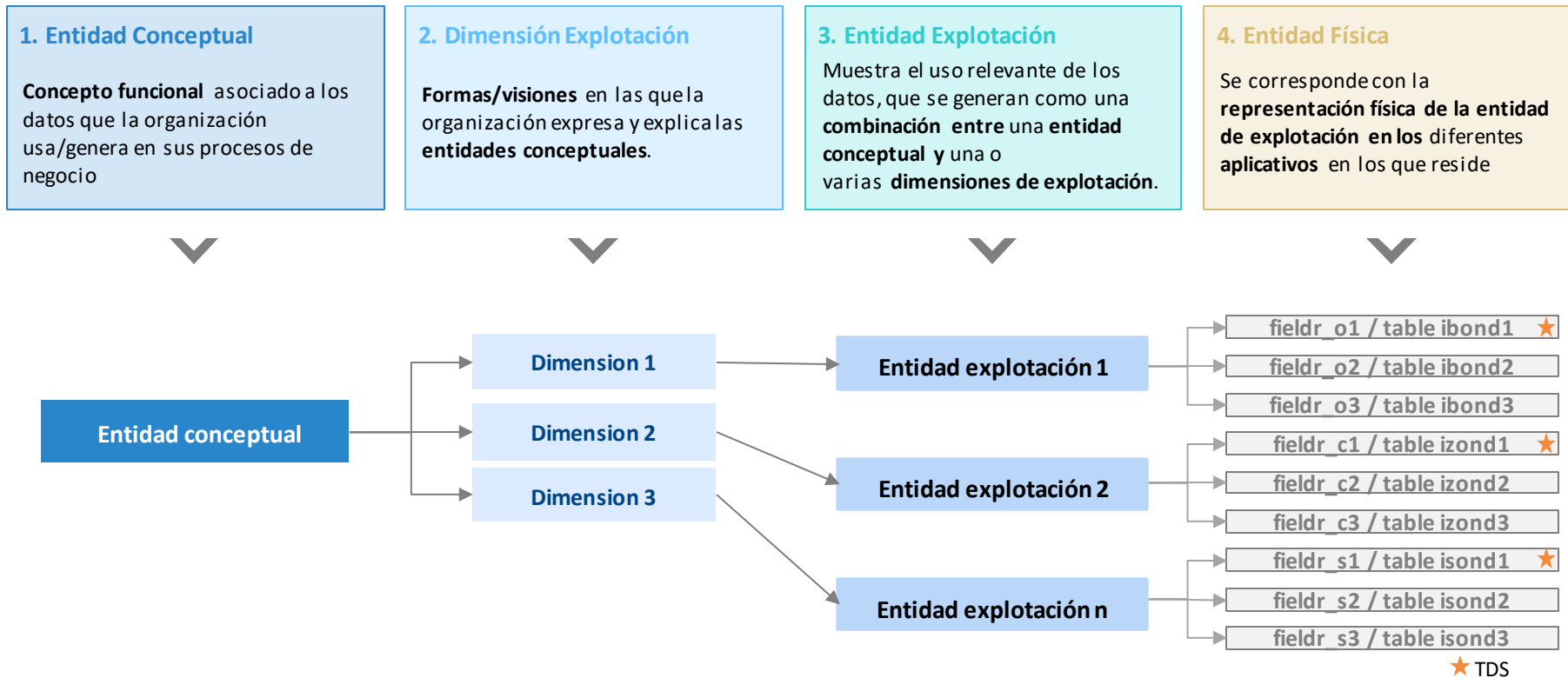
Principales componentes

1	Data Content	Asegura que todos los datos cuentan con una definición funcional y técnica en un diccionario de datos homogéneo (data dictionary)
2	Data Quality	Garantiza que el dato tiene la calidad requerida a través de reglas basadas en los 7 principios de calidad ¹ Permite a la organización certificar, auditar y verificar la calidad de los datos
3	Data Lineage	Definición del flujo del cálculo del dato identificando “pain points” (“hand-offs”, actividades manuales, re-procesos,...) Asegura la trazabilidad funcional de los datos así como el uso posterior de los mismos
4	Data Availability	Elaboración de un mapa de herramientas, infraestructuras y entornos de usuario utilizados por los data owner en la gestión y uso de sus datos, que permiten la disponibilización de los datos
5	Data Skills	Definición de mapa de roles/conocimiento que permitirá la creación de comunidades de práctica relacionadas con la gestión y tratamiento de los datos

¹ Principios de calidad: completitud, validez, integridad, consistencia, accesibilidad, disponibilidad y precisión

Data content

El **objetivo** de este componente es asegurar que todos sus **entidades conceptuales/dimensiones, entidades de explotación (usos) y entidades físicas asociadas**, cuentan con una adecuada definición en un diccionario de datos homogéneo



Data lineage

Persigue documentar información sobre el **origen, procesos de cálculo** y el **uso** de las **entidades** de explotación gestionadas dentro de cada ownership, a fin de poder identificar **potenciales mejoras** en la ejecución durante el proceso del ciclo de vida del dato.

Trazabilidad funcional

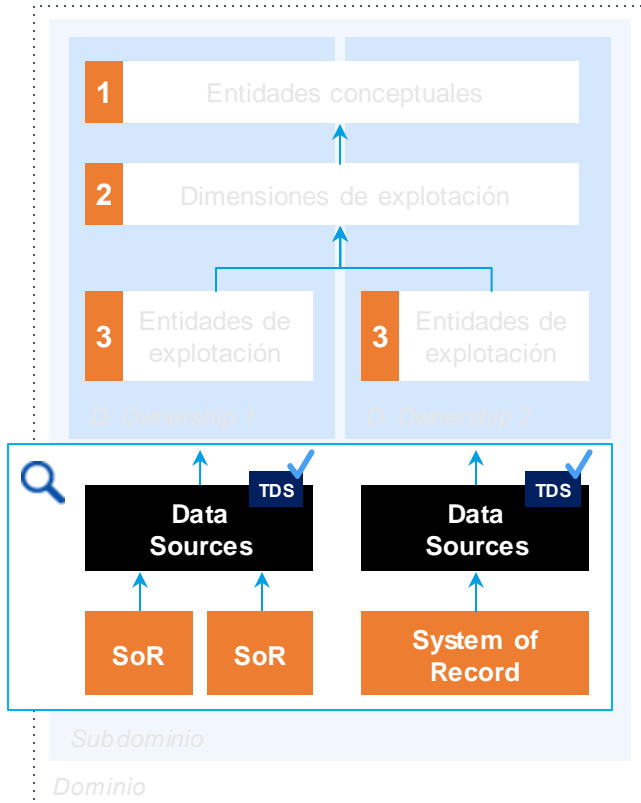
Definir la **trazabilidad funcional para cada dato** ('donde llegan los datos y hacia dónde van los datos'), identificando para cada proceso de transformación oportunidades para **simplificar el proceso, automatizar las interacciones manuales, eliminar controles innecesarios, minimizar las transferencias y reducir los riesgos operacionales**.

Área	Lineage nos ayuda a diagnosticar situaciones como:
Data Governance	<ul style="list-style-type: none"> • Desalineamiento entre las definiciones de data entities y requerimientos de negocio y IT • Aprovisionamiento informal de datos y 'fugas' de datos del proceso principal • Ownership y responsabilidades poco claras entre stakeholders • Incapacidad para rectificar problemas de datos a través de silos organizacionales
Data processes	<ul style="list-style-type: none"> • Captura de datos altamente manual que conduce a problemas de eficiencia y eficacia • Problemas que implican rechazos múltiples entre las personas en el flujo de datos • Gran cantidad de ajustes manuales a los datos en los sistemas
Data quality	<ul style="list-style-type: none"> • Controles y conciliaciones superpuestas o redundantes • Controles de calidad críticos llevados a cabo manualmente • Tipos de controles y frecuencia para asegurar la calidad
Systems and IT	<ul style="list-style-type: none"> • Roles ineficientes de ingeniería dentro del flujo de datos • Ineficiencia debido a la reutilización de sistemas, modelos de datos y aplicaciones • Mala gestión y supervisión de los controles de seguridad y los derechos de acceso
Business processes	<ul style="list-style-type: none"> • Personal propenso a errores o poco cualificado tocando datos • Dependencias de empleados clave y riesgo clave del hombre • Seguimiento disperso del estado del procesamiento de datos

Niveles de madurez

	1 Iniciado	2 Gestionado	3 Establecido	4 Medido	5 Predecible
Data Content	Business Glossary para algunos dominios de información (no completo)	BG de 'core data' completo con aprobación de workflows (automáticos o manuales)	Aprobación de workflows (automáticos o manuales) auditables: evidencias y reportes que cumplen con estándares regulatorios	Aprobación de workflows automáticos medibles: "Balanced Scorecards"	Aprobación de workflows automáticos predecibles
Data Quality	Reglas de Data Quality definidas; soporte manual a través de templates		Reglas de DQ Operativizadas y medibles en Data Quality Dashboards de forma automática (asegurando que se realizan los controles oportunos)		Problemas en DQ predecibles basados en patrones históricos y machine learning
Data Lineage	Data Maps estáticos para procesos de datos (lineage)	Trazabilidad en tiempo de diseño (end-to-end) operativizada en una herramienta específica	Trazabilidad en tiempo de diseño y ejecución; generación de evidencias que cumplen con el reporte regulatorio	Resultados de trazabilidad medibles mediante KPIs para distintos fines (P.ej: DQ, Cumplimiento, ...)	Problemas en reconciliación predecibles basados en patrones históricos y machine learning
Data Skills	Roles genéricos para el uso/tratamiento de datos		Establecimiento de roles y responsabilidades específicos para la gestión/tratamiento de datos. Definición de planes formativos para ellos.		Creación de nuevos roles y especializaciones relativas a la gestión de datos de forma dinámica (teniendo en cuenta las Community of Expertise /Practices)
Data Avail.	Censo estático de herramientas (indicando si la infra. es local o corporativa y si está disponible para toda la organización)		Herramientas para el tratamiento de datos con un alto nivel de automatización y con retroalimentación de información a la infraestructura corporativa		Herramientas e infraestructura que posibilitan la disponibilidad de los datos como servicio (DaaS)

Otros conceptos clave a considerar



¹Pueden alimentarse de una o varias fuentes de aprovisionamiento (SORs)

System of Record

Fuente o sistema de aprovisionamiento considerado como el origen de los datos, que alimenta a uno o varios aplicativos

Data Source

Cualquier soporte en el que el usuario mantiene, modifica, transforma y/o calcula datos, considerando desde una herramienta ofimática hasta aplicativos corporativos¹

DS Fundacional

Soportes que almacenan los datos recibidos desde las fuentes de aprovisionamiento sin que sufran ninguna alteración (estáticos por naturaleza)

DS Derivada

Soportes que almacenan datos que ya han sido sometidos a procesos de transformación, dando lugar a datos derivados

Trusted Data Source (TDS)

Fuente de datos universalmente conocida y adoptada como la única válida y fiable para obtener un dato concreto (*"golden source"*)

Glosario de términos (I)

	Término	Descripción
01	Dato	Consideramos como dato cualquier concepto del cual la organización requiere almacenar información para su negocio. Los datos pueden ser clasificados en diferentes tipologías: master data, reference data, transactional data, historical data, metadata, etc Adicionalmente, existen dos tipos de datos dependiendo del nivel de enriquecimiento que tengan: 1) datos brutos (o datos no procesados) y datos derivados (o datos procesados)
02	Data entity	Concepto abstracto de alto nivel con significado por sí mismo que proporciona contexto a la actividad de negocio (p.ej: cliente, producto). Por cada data entity es común que la organización almacene varios datos finales (p.ej: nombre del cliente, dirección del cliente, etc).
03	Trusted Data Source (TDS)	Fuentes de datos universalmente conocidas y adoptadas como las únicas válidas y fiables para obtener un dato concreto (golden source). Existen dos tipologías de TDS: •TDS fundacional: almacenan datos brutos sin modificar (estáticos por naturaleza); •TDS derivada: almacenan datos procesados o derivados en los dominios
04	System of Record (SoR)	Fuente o sistema de aprovisionamiento considerado como el origen de los datos. Cada TDS tiene uno o varios SoR.
05	Ejes (o atributos)	Información que se combina con los datos (o métricas) para facilitar su agregación o explotación en un informe. P.ej: visualizar un dato por el eje país, o por el eje mes.
06	Tipología de dato (atributo o indicador)	Dependiendo del dato, se puede catalogar como un atributo (eje de agregación, p.ej: epígrafe o centro) o un indicador (métrica, p.ej: dudosos)
07	Reporte	Informe o documento compuesto por varios datos/métricas para transmitir información con un determinado objetivo. P.ej: informe de sostenibilidad por región (para realizar el seguimiento del área y ayudar en la toma de decisiones)

Glosario de términos (II)

	Término	Descripción
08	Fase de aprovisionamiento de datos	Fase en la que se recogen las principales fuentes de datos (SoR) que ceden información a un área en concreto. P.ej: el área de compras recibe información de precios de los distintos proveedores en el ERP corporativo procedentes de los distintos delegados comerciales de aprovisionamiento.
09	Fase agregación / cálculo	Fase en la que se registran todos los sistemas/bbdd (TDS) donde el área consolida y transforma la información para la generación de nuevos datos P.ej: el área de compras aplica los márgenes o mark-ups establecidos por la compañía y publica los precios de referencia para contratación por parte de cada uno de los departamentos de la empresa
10	Fase explotación de datos	Fase en la que los datos finales son usados y reportados en informes. P.ej: el área de compras recopila la información de todas las contrataciones a terceros y el valor total de los contratos
11	Data Domain	Set de datos que puede abarcar un área completa o un conjunto de datos específicos. Debe estar alineado con los procesos de negocio más importantes del banco y al mayor nivel de gestión. Las compañías definen los data domains para establecer responsabilidades sobre los datos

ÍNDICE

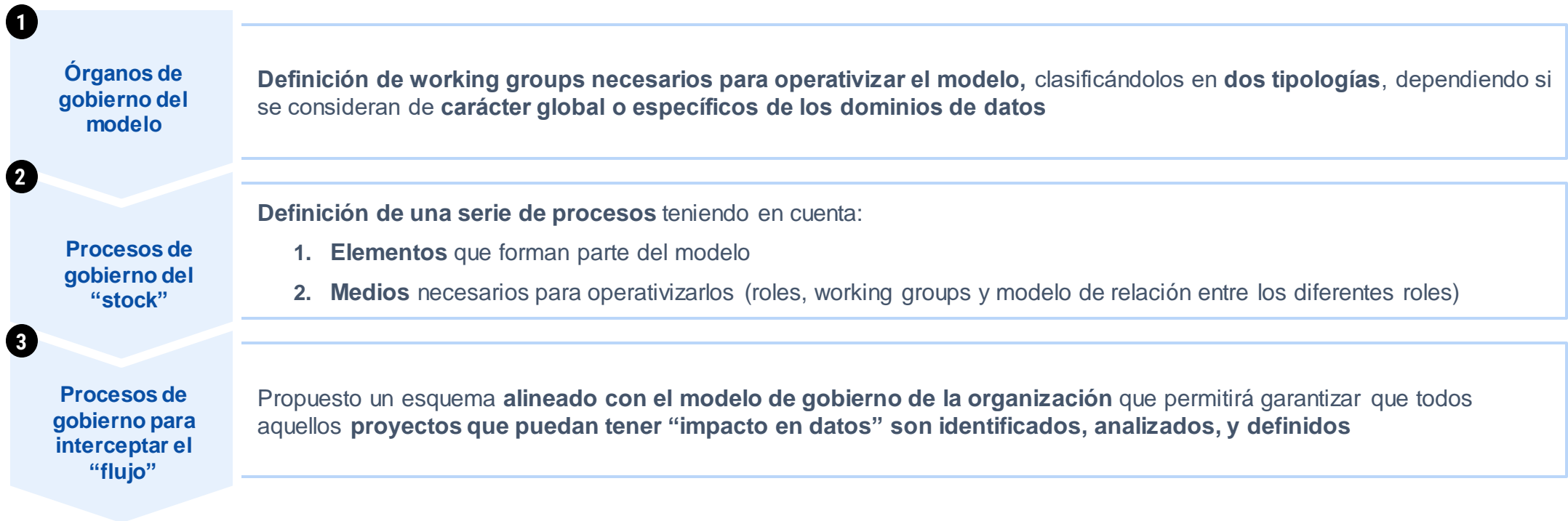
01: Roles del modelo básico de data governance

02: Conceptos clave

03: Órganos de gobierno y procesos

04: Cultura del dato

Necesidad de órganos de gobierno y procesos



Hay una infinidad de maneras en la que éstos órganos de gobierno pueden ser compuestos y configurados, lo realmente importante es la permeabilización en todos los estratos de la organización, desde la dirección general y hasta el usuario final, teniendo en cuenta a todos los owners de cada pieza de información (y asumiendo desde el principio que en muchos casos es imposible definir un propietario único para todos los datos)

ÍNDICE

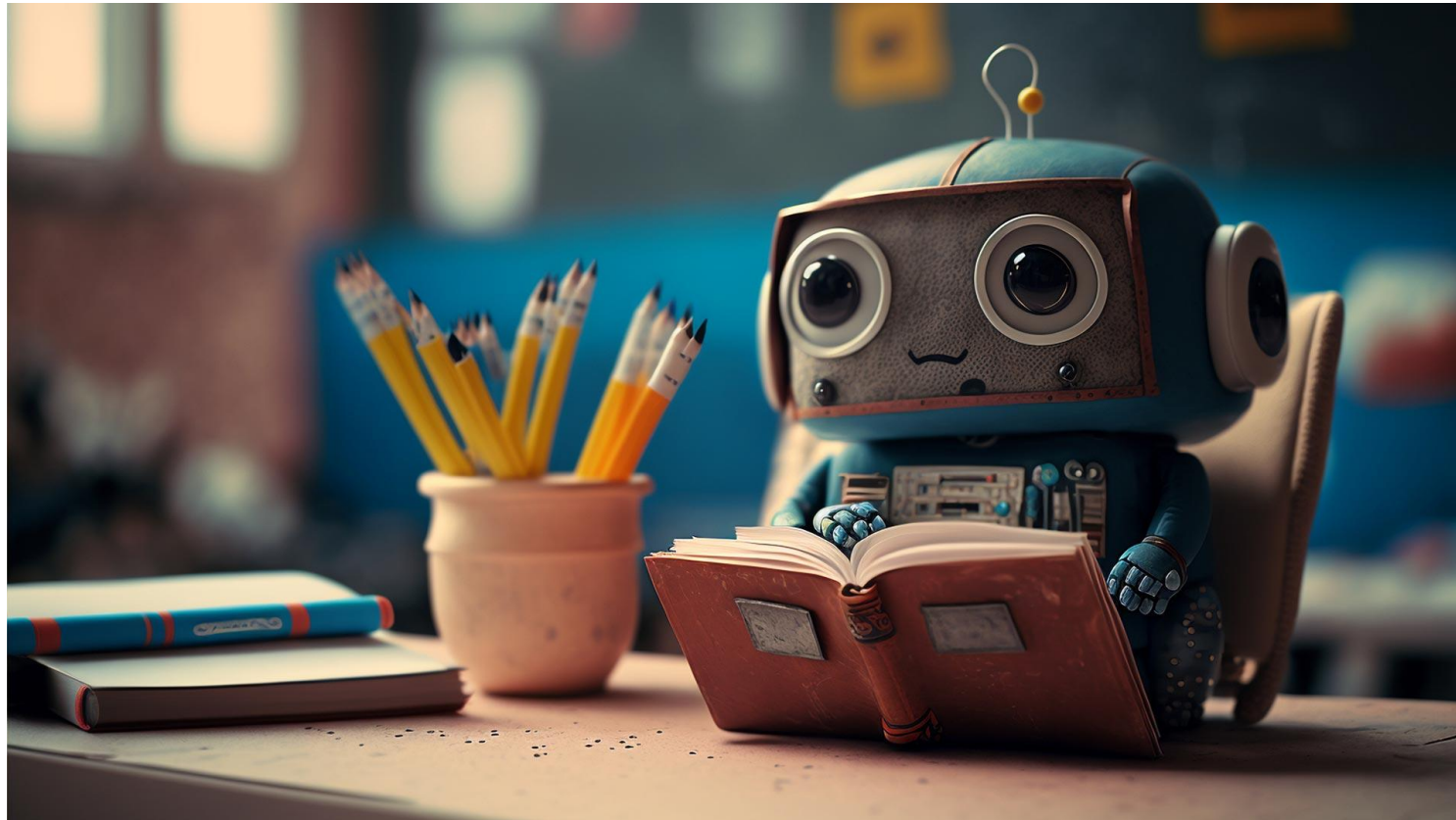
01: Roles del modelo básico de data governance

02: Conceptos clave

03: Órganos de gobierno y procesos

04: Cultura del dato

La cultura del dato



La cultura del dato

Una cultura de datos son las creencias y los comportamientos colectivos de las personas que valoran, aprovechan y promueven el uso de datos para mejorar la toma de decisiones. Como resultado, los datos se integran en las operaciones, la mentalidad y la identidad de una organización. Una cultura de datos permite a todos acceder a la información que necesitan para realmente basarse en los datos y superar los desafíos empresariales más complejos.

- **Adoptando comportamientos basados en datos:** Alineando los datos y el análisis con los resultados empresariales
- **Aprovechando los datos estratégicos:** Priorizando los datos en la toma de decisiones y en los procesos empresariales
- **Fomentando la colaboración y la participación en la comunidad:** Trabajando en equipo para lograr una misión compartida de liderar con datos.



Para tener en cuenta

1. Tratar el dato como un activo:

- Dar al dato una identidad más allá de las aplicaciones de ingeniería que lo procesan, para mejorar su re-usabilidad, seguridad y futuro desarrollo (aplicando disciplinas de gestión de proyectos)

2. Codificar regulaciones y controles, automatizar los procesos:

- Ofrecer un camino guiado hacia el "cómo sí" en lugar de caer al "no" por defecto. Utilizar perfilados de grano fino y formalizar requerimientos semánticos (para qué, para quién), en los procesos, en los sistemas y en las personas.

3. Cerrar el gap entre operaciones y analítica:

- Evitar la optimización temprana (ser más agile y permitir una mayor abstracción y exploración) de las representaciones analíticas de los datos, reemplazar ETLs frágiles y manuales con contratos sólidos basados en formatos y layouts de datos extensibles (Flexibilidad by design)

4. Evitar una excesiva centralización:

- Pero hasta cierto punto, hay que evaluar dónde es necesario y dónde es perjudicial, favoreciendo modelos de data mesh allí donde los datos sean volátiles, cambiantes, inciertos o en función de aspectos regulatorios y de control

La gobernanza de datos para la IA

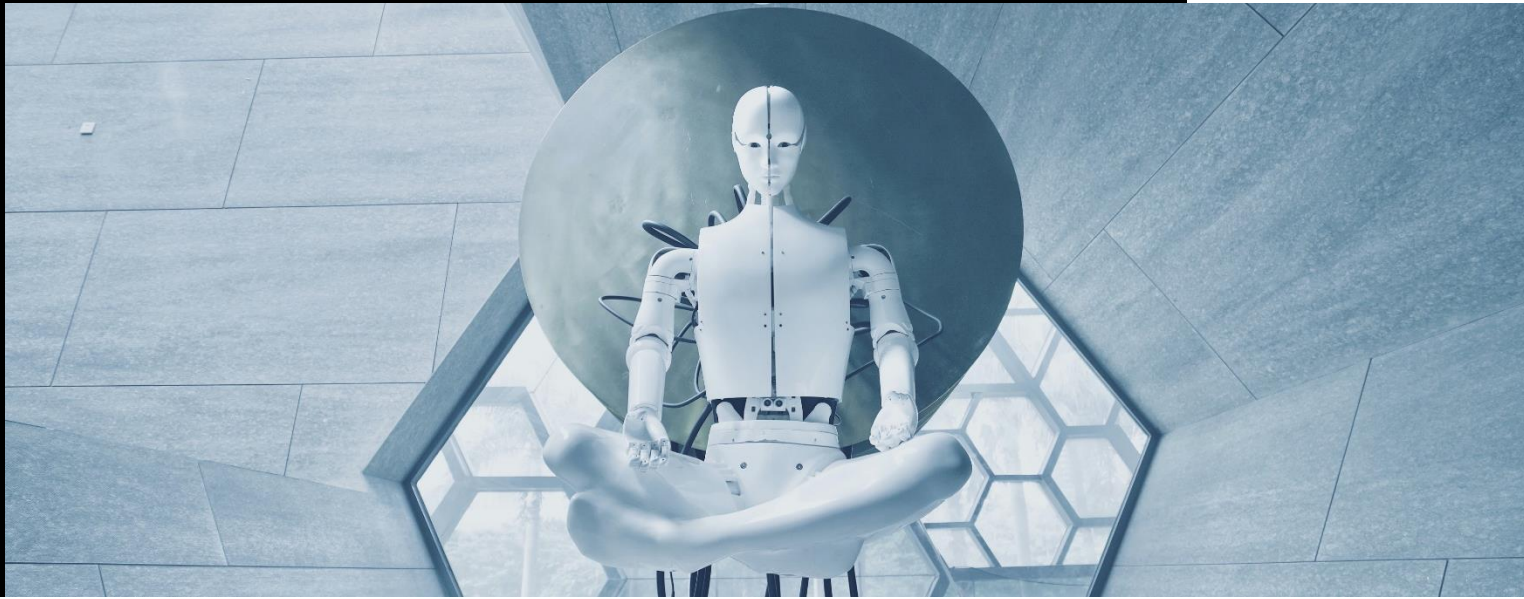
La gobernanza de la IA o del ML aplica muchas de las mismas prácticas de gobernanza de datos.

La calidad y la integración de los datos deben proporcionar los datos necesarios para el entrenamiento de modelos y la implementación en producción (los almacenes de características son un aspecto importante de esto).

La inteligencia artificial (IA) responsable presta especial atención al uso de datos confidenciales para crear modelos.

Las capacidades adicionales de gobernanza del ML incluyen permitir que las personas participen en la creación, la implementación y la supervisión de los modelos; documentar el entrenamiento, el control de versiones, los casos de uso admitidos y guiar el uso ético del modelo; y supervisar el modelo en producción para verificar su precisión, desviación, sobreajuste y ajuste insuficiente.

La IA generativa requiere capacidades de gobernanza de datos adicionales, como la calidad y la integridad de los datos, para respaldar la adaptación de los modelos básicos para el entrenamiento y la inferencia, la gobernanza de la toxicidad y el sesgo de la IA generativa y las operaciones del modelo de base (FM): las FMOps.



Contar con un buen modelo de gobierno del dato aporta una serie de beneficios a la organización, desde la mejora del control y la asignación de responsabilidades, hasta el fomento de la calidad y el incremento de la productividad, pasando por una mayor accesibilidad y democratización de acceso a la información, una mejor toma de decisiones y un mayor compliance con los requerimientos regulatorios.

Conclusiones

Tema 2

Ciclo de vida y gobierno del dato dentro de la organización II: La calidad y disponibilidad de los datos. Integración con los procesos dentro de la organización

Objetivos

01

Entender la crucial importancia de disponer de datos de calidad para la toma de decisiones y su incorporación a modelos analíticos y de negocio

02

Mejorar la fiabilidad y la confiabilidad en los datos dentro de la organización

03

Mejorar la toma de decisiones y los datos en los que las decisiones se sustentan, con base en los procesos de aseguramiento de la calidad y la retroalimentación positiva de la que esos mismos procesos se benefician

Los problemas de la calidad

Sector Retail:

- **Best Buy:** Durante el Black Friday, Best Buy tuvo que cancelar miles de pedidos en línea debido a errores en los datos de inventario que mostraban productos en stock cuando en realidad no lo estaban.

Sector Transporte:

- **Aerolínea Qantas:** Un avión A380 sufrió un fallo en un motor debido a un error en la base de datos de mantenimiento que no reflejaba adecuadamente las horas de operación del motor.
- **Toyota:** La compañía tuvo que retirar del mercado millones de vehículos debido a problemas con los aceleradores. Parte del problema radicaba en la gestión y calidad de datos relacionados con los componentes de los vehículos y las quejas de los clientes.

Sector Energético:

- **Pacific Gas and Electric:** Un gasoducto explotó en San Bruno, California, causando la muerte de 8 personas. Se encontró que la compañía tenía datos incorrectos sobre la tubería, lo que contribuyó al desastre.

Sector Público:

- **Oficina del Censo de EE.UU:** La Oficina del Censo gastó aproximadamente 3 mil millones de dólares adicionales en el censo de 2010 debido a problemas con los datos y sistemas de información.
- **UK Home Office:** Más de 1.500 delincuentes extranjeros no fueron considerados para deportación debido a errores en la base de datos del Ministerio del Interior.

Sector Salud:

- **Hospital de la Universidad de California:** Un error en el software de prescripción llevó a que una paciente recibiera 38 píldoras de un medicamento en lugar de la dosis prescrita de una sola píldora. La paciente falleció debido a la sobredosis.
- **Veterans Health Administration:** Miles de pacientes fueron expuestos a posibles infecciones en clínicas de colonoscopia debido a un error de datos que resultó en equipos no esterilizados adecuadamente.

Sector Financiero:

- **JP Morgan Chase:** La entidad tuvo pérdidas multimillonarias debido a errores en los modelos de datos de riesgo que subestimaron la exposición al riesgo en sus carteras.

Introducción

■ El modelo de gobierno aspira a:

- Añadir expertise en la definición de reglas de calidad
- Medir los resultados de la implantación de dichas reglas
- Gestionar las alertas, de forma que trabajemos en la mejora continua de los datos en el Grupo

■ La aplicación de más y mejores reglas de calidad sobre los datos **aportará una serie de beneficios:**

Beneficios

- Comprobar los **resultados de las reglas de calidad definidas**, detectar **posibles incidencias**, y desencadenar **acciones de gestión** sobre las mismas
- Permitir **mejorar la fiabilidad de los datos**, consiguiendo que los datos que se usan estén certificados, auditados y verificados
- **Mejorar la toma de decisiones**, garantizando que los datos que las sustentan son precisos

Conceptos clave

Término	Descripción
Principio de calidad	Categoría sobre la cual es necesario definir reglas para mejorar la calidad de los datos. Como parte del modelo vamos a definir 6 principios divididos en básicos y expertos
Regla de calidad	Control definido y operativizado que contribuye a que el dato cumpla con la calidad requerida (i.e., que el valor del dato es el correcto)
Parámetros de la regla de calidad	Conjunto de atributos que forman parte de la definición de la regla de calidad
Umbrales de tolerancia	Conjunto de valores que sirven para determinar si el resultado de una regla de calidad es o no satisfactorio
Alerta	Incumplimiento de los umbrales definidos para una regla y sus parámetros. Conllevan la necesidad de valorar si una acción concreta de mejora por parte del Data Owner y/o los equipos técnicos correspondientes es o no necesaria

Tener un plan y un proceso es fundamental



- 1** **Definición funcional** de las reglas de calidad y determinación de sus parámetros y umbrales de tolerancia
Definición funcional de los outputs necesarios para seguimiento por parte del data owner/ Dominio
- 2** **Valoración** de las reglas priorizadas por parte del Dominio, determinación de **dónde** deben ser implantadas y **recolección** de los **resultados** de su aplicación
- 3** **Análisis y seguimiento** de los resultados de la aplicación de las reglas de calidad definidas para la **gestión y mejora continua de las alertas presentadas**
- 4** **Monitorización y análisis** de la calidad del dato en función de las reglas definidas para determinar la **necesidad de ajustes** a los parámetros y/o umbrales de las reglas de calidad existentes y/o definición de **nuevas reglas** de calidad adicionales

ÍNDICE

01: Los principios de calidad

02: Las reglas de calidad

03: La gestión, el aseguramiento y el
mantenimiento de la calidad en los datos

04: La retroalimentación de los procesos de
calidad

Los principios de calidad

Ejemplo			
1	Compleitud	Los datos necesarios están presentes e informados	La actividad del grupo empresarial está informado (sin valores nulos).
2	Validez	Los datos tienen unos valores válidos para su uso: formato, longitud, valores permitidos, etc	El porcentaje de reparto de la garantía está entre 0% y 100%.
3	Integridad	Existe una correcta relación entre los datos, siendo coherentes entre sí y cumpliendo las reglas de negocio relativas a los mismos	El país de residencia de un cliente se corresponde con su código postal.
4	Consistencia	Los datos son consistentes dentro de un repositorio de información, y en el resto de repositorios/sistemas involucrados en el proceso	Todos los clientes deben tener asociado al menos un código de contrato existente en la maestra de contratos
5	Accesibilidad	El dato debe estar disponible con un refresco de información adecuado al uso que se le dé al dato	El fichero TVFI_MES.csv contiene 210 datos para cada una de las series históricas
6	Disponibilidad	El dato debe estar disponible cuando se necesita y de forma actualizada	El fichero RDA_OPERACIONES.txt debe recibirse en la fecha comprometida (diario 7.50h)
7	Precisión	Los datos deben reflejar la realidad exacta	Todas las pérdidas RO que se encuentran en las cuentas tipo 2 deben estar reportadas.

ÍNDICE

01: Los principios de calidad

02: Las reglas de calidad

03: La gestión, el aseguramiento y el
mantenimiento de la calidad en los datos

04: La retroalimentación de los procesos de
calidad

Las reglas de calidad

Las reglas se clasifican entre básicas y expertas en función de la necesidad de conocimiento experto por parte del data owner para definir las

Clasificación

Características

Principios asociados

1 REGLAS BÁSICAS

- Son **necesarias**, pues proporcionan el mínimo nivel de control requerido, pero **no suficientes**
- Se pueden **estandarizar más fácilmente**, aunque requieren del **Data Owner** para la **definición** de sus parámetros y umbrales

1. Disponibilidad
2. Completitud
3. Validez

2 REGLAS EXPERTAS

- Son necesarias para incorporar la **lógica funcional y la definición intrínseca** de los datos
- Se necesita el **expertise funcional del Data Owner** para ser definidas adecuadamente

4. Consistencia
5. Integridad
6. Precisión

Las reglas deben evolucionar y ser la base de la mejora continua

Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas básicas:**

Principio	Regla	Descripción	Ejemplo
1- Disponibilidad	1.1-Comprobación que el fichero se ha recibido en la fecha y hora acordadas	Comprobar si el fichero se ha recibido en la fecha y hora establecida	<ul style="list-style-type: none"> El fichero XXX (Saldos por Contrato) de Canarias se recibe en D + 10 en la máquina de finanzas
	1.2-Actualización del valor del dato en la fecha requerida	Comprobar que el valor del dato de fecha de datos está actualizado a la fecha requerida	<ul style="list-style-type: none"> La fecha de datos del fichero mensual corresponde al cierre de mes en cuestión
2- Completitud	2.1-Completitud de registros	Comprobar que los valores de los datos están informados en el fichero correspondiente	<ul style="list-style-type: none"> El fichero no se ha recibido vacío
	2.2-Completitud de perímetro requerido	Comprobar que el número de registros es el apropiado con la especificación realizada en el fichero correspondiente	<ul style="list-style-type: none"> El fichero de XXX de Clientes tiene la totalidad de clientes de Colombia susceptibles de ser enviados al Motor El número de registros que nos ha enviado el país es el mismo que hemos cargado en global

Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas básicas:**

Principio	Regla	Descripción	Ejemplo
3- Validez	3.1-Comprobación del valor de dato nulo (NULL o vacío)	Comprobar que no existen valores nulos (NULL o vacío) en el fichero correspondiente	<ul style="list-style-type: none"> El 100% de los valores del campo no es NULL
	3.2-Comprobación del formato del campo	Comprobar que el formato del campo es el adecuado para la entidad conceptual/explotación respectiva	<ul style="list-style-type: none"> El valor del dato cumple con el formato requerido: alfanumérico y con una longitud exacta de 9 posiciones El valor del dato es un %
	3.3-Comprobación de valores no permitidos	Comprobar que el valor del dato no se encuentra entre los valores no permitidos	<ul style="list-style-type: none"> El campo fecha no puede informarse con 99/99/9999
	3.4-Comprobación del valor del dato dentro del rango esperado	Comprobar que el dato se encuentra dentro de los valores esperados para el mismo	<ul style="list-style-type: none"> El porcentaje de reparto de la garantía está entre 0% y 100%
	3.5-Comprobación del valor del dato en catálogo	Comprobar que el valor del dato es uno de los definidos en el catálogo correspondiente	<ul style="list-style-type: none"> El valor del campo Segmento Global es uno de los existentes en el Catálogo Segmento Global

Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas expertas:**

Principio	Regla	Descripción	Ejemplo
4- Consistencia	4.1-Transferencia de datos origen - destino	Comparar el valor de un mismo dato en fichero origen y fichero destino	<ul style="list-style-type: none"> La exposición total debe ser idéntica entre los pasos de proceso X e Y La información que se transfiere entre un repositorio y otro no sufre transformación en caso de no ser requerida
	4.2-Duplicidad de registros para un mismo dato	Comprobar que no hay datos duplicados si el tipo de dato es en un fichero determinado	<ul style="list-style-type: none"> El código de cliente es único para cada cliente y no existen duplicados No pueden existir dos registros con la misma clave en la tabla
	4.3-Conciliación de un mismo dato entre diferentes tablas/repositorios en el mismo momento de tiempo	Comparar el valor de un mismo dato en diferentes tablas o repositorios en los que se encuentre disponible	<ul style="list-style-type: none"> Todos los clientes que tienen rating existen en el repositorio de clientes La conciliación de un saldo desde inventarios a un epígrafe contable El sumatorio de los saldos por área de negocio es igual al saldo total del Banco Integridad referencial entre bases de datos

Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas expertas:**

Principio	Regla	Descripción	Ejemplo
5- Integridad	5.1-Réplica de cálculos	Comprobar que el valor de un dato es igual a la fórmula establecida	<ul style="list-style-type: none"> El riesgo en Watchlist se corresponde con el sumatorio del crédito a la clientela (sin dudosos), riesgos contingentes (sin dudosos), riesgo de contrapartida y riesgo emisor de cliente con marca de Watchlist
	5.2-Comparativa entre diferentes valores de datos relacionados entre sí	Comprobar que el valor del dato cumple una regla establecida respecto al valor de otro dato (>, >, =...)	<ul style="list-style-type: none"> La fecha de vencimiento del contrato es posterior a la fecha de apertura El país de residencia de un cliente es correcto en relación a su código postal Si el cliente pertenece al perímetro alto consumo energético, el sector AA debe corresponder con el sector

Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas expertas:**

Principio	Regla	Descripción	Ejemplo
6- Precision	6.1-Concentración de valores del mismo dato en un catálogo	Comprobar la adecuada distribución de valores de un catálogo (porcentaje de registros en un valor respecto al total)	<ul style="list-style-type: none"> La concentración de préstamos en el segmento PYMES no puede exceder el 90%
	6.2-Comprobación de la sensibilidad del valor de un dato a un umbral	Comprobar que el valor del dato no sobrepasa un umbral determinado	<ul style="list-style-type: none"> El ratio de morosidad no puede superar el 10%
	6.3-Comparativa de la evolución de diferentes valores de datos relacionados entre sí en el mismo momento de tiempo	Comprobar que el valor de un dato comparado con otro tiene una evolución razonable	<ul style="list-style-type: none"> Si el riesgo crediticio se mantiene y la tasa de mora aumenta, los dudosos también aumentan
	6.4-Comparativa del valor del mismo dato entre períodos	Comprobar que la tendencia (evolución) del valor de un mismo dato en el tiempo es aceptable	<ul style="list-style-type: none"> El número de clientes activos de un mes respecto al anterior no puede ser $>< 5\%$
	6.5-Correlación del valor de un dato en las últimas n observaciones	Comprobar que la correlación de los valores de un dato en sus últimas n observaciones no es inferior a un nivel determinado. Nivel y número de observaciones a definir por data owner	<ul style="list-style-type: none"> Variación de medidas estadísticas: La media (mediana, moda) del valor de las últimas n observaciones no varía más de respecto al período anterior La varianza de las últimas n observaciones no es superior a x
	6.6Comparativa entre valores de datos distintos entre periodos	Comprobar que la tendencia (evolución) del valor de distintos datos en el tiempo es aceptable	<ul style="list-style-type: none"> Si la variación de activos dudosos sube un %, la evolución de la provisión debe ser igual o mayor a ese %

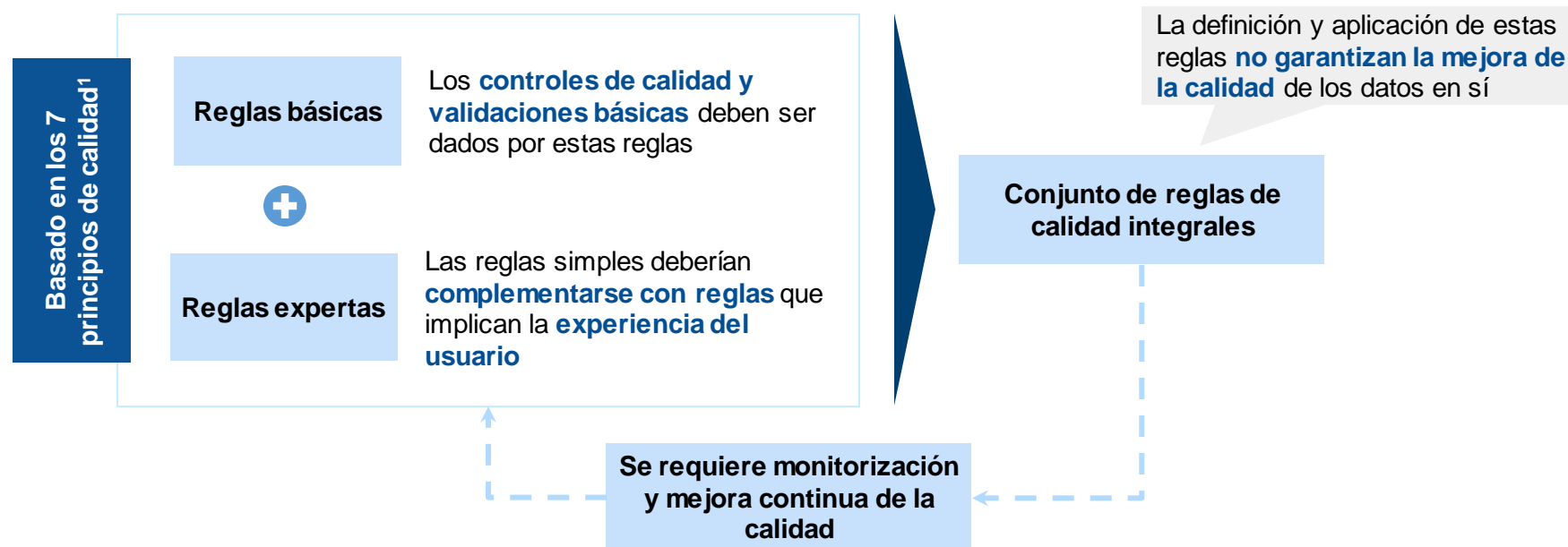
Las reglas de calidad

Se muestra a continuación un ejemplo de un marco de reglas de calidad (no exhaustivo) y ejemplos asociados: **Reglas expertas:**

Principio	Regla	Descripción	Ejemplo
6- Precision	6.7-Coincidencia con base de datos externa	Comprobar que el valor del dato coincide con el valor publicado en una base de datos externa en un momento determinado del tiempo	<ul style="list-style-type: none"> La calificación externa del emisor corresponde a la publicada por la Agencia de Calificación correspondiente El volumen de ventas de una empresa se corresponde con lo registrado en el Registro Mercantil
	6.8-Coincidencia con soporte físico	Comprobar que el valor del dato coincide con el valor en un soporte físico concreto en un momento determinado del tiempo	<ul style="list-style-type: none"> El número de identificación del cliente corresponde a lo que se refleja en su documento físico El índice de referencia de un producto, coincide con lo recogido en el expediente físico
	6.9-Tendencia de volumetría	El volumen de registros recibido/generado entre periodos no supera el umbral	<ul style="list-style-type: none"> La variación del número de registros recibido de venta de inmuebles entre este mes y el anterior no debe superar el 1%

Controles y monitorización

Para asegurar que la calidad del dato se mantiene en el tiempo, debe asegurarse un mecanismo de mejora continua



Para garantizar la calidad de los datos, se deben seguir ambas reglas (simples y complejas), así como procesos de mejora continua de la calidad

Ejemplos adicionales

Reglas Básicas

1 Completitud

- La actividad del grupo empresarial está informado (sin valores nulos)

2 Validez

- El dato cumple con el formato requerido: sin caracteres extraños y con una longitud exacta de 9 posiciones
- El porcentaje de reparto de la garantía está entre 0% y 100%
- El sector del cliente se corresponde con uno de los valores recogidos en el catálogo vigente, sin valores por defecto
- El saldo dispuesto es negativo (signo contable) o igual a cero

7 Disponibilidad

- La información del ámbito doméstica se recibe en la fecha comprometida.
- El valor de tasación está accesible y actualizado (última fecha de actualización inferior a un año)

Regla Experta

3 Integridad

- El país de residencia de un cliente se corresponde con su código postal.
- La fecha de vencimiento del contrato es posterior a la fecha de apertura
- Si el cliente pertenece al perímetro de Specialized Lending, el sector AA debe corresponder con el sector de project

4 Consistencia

- El código de cliente es único para cada cliente y no existen duplicados
- Todos los contratos tienen asociado un cliente
- El saldo de crédito almacenado en el informacional está conciliado con la contabilidad
- Si el riesgo crediticio se mantiene y la tasa de mora aumenta, los dudosos también aumentan
- El saldo dudoso cedido en el aprovisionamiento y presentado en el informe es consistente

5 Accesibilidad

- El volumen de registros de clientes del informacional incluye toda la población de intervinientes de la entidad
- El número de titulares cedidos en los ficheros de pre-carga coincide con los registrados en el informacional

6 Precisión

- El número del documento de identidad del cliente se corresponde con lo reflejado en su documento físico
- La calificación externa del emisor se corresponde con la publicada en la Agencia de Calificación correspondiente

ÍNDICE

01: Los principios de calidad

02: Las reglas de calidad

**03: La gestión, el aseguramiento y el
mantenimiento de la calidad en los datos**

04: La retroalimentación de los procesos de
calidad

Monitorización de la calidad

Existen dos conceptos clave para realizar una definición adecuada de la regla que permita conocer si el resultado de la misma es o no satisfactorio

Parámetros definidos en la regla de calidad

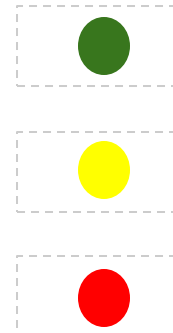
- Se trata de los atributos que **forman parte de la definición de la regla** que serán incluidos en la Template de Calidad
- Estos parámetros **sirven al equipo de TI para construir, parametrizar e implantar** la regla de Calidad sobre el fichero, registro o dato concreto

Umbrales de tolerancia definidos en la regla de calidad

- Se trata de los valores **de tolerancia** al incumplimiento en una determinada regla para un grupo de datos
- Estos umbrales **definen tramos de tolerancia** y típicamente requieren que se defina tanto un **valor objetivo como un valor mínimo**

La implantación de la regla y sus parámetros dará como resultado un % de cumplimiento que se informará en el **Quality Output Template** o portal de calidad correspondiente

% de cumplimiento de la regla (número de registros OK sobre el total de registros)



El % de cumplimiento de la regla se compara con los umbrales de tolerancia establecidos dando como resultado un indicador verde o amarillo o rojo

Ejemplos de tipos de regla, parámetros y umbrales (1)

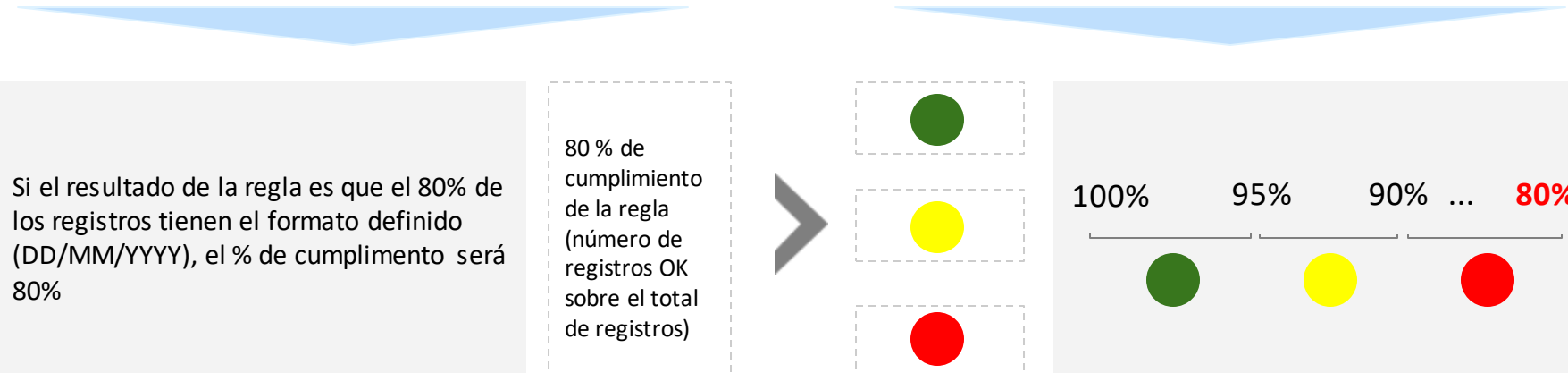
Ejemplo I. Tipo de Regla: Formato del campo; el formato del campo fecha de cancelación debe ser DD/MM/YYYY

Parámetros definidos en la regla de calidad

- El formato del campo **Fecha de Cancelación** debe ser DD/MM/YYYY
- Este **formato será utilizado por TI** de cara al análisis técnico, construcción e implantación de la regla.
- Si el dato cumple con este formato, el resultado será OK

Umbrales de tolerancia definidos en la regla

- **Objetivo:** 95%. **Mínimo:** 90%.
- Si el nº de registros OK es < 90% del total, el resultado es insatisfactorio (indicador rojo). Si está entre 90% y 95%, el resultado es mejorable (indicador amarillo). Por encima del 95% el resultado es satisfactorio (ind. verde)



Ejemplos de tipos de regla, parámetros y umbrales (2)

Ejemplo II. Tipo de Regla: Comparativa del valor del mismo dato entre períodos. La variación del nº de clientes activos respecto al mes anterior no puede ser >10% (en valor absoluto).

Parámetros definidos en la regla de calidad

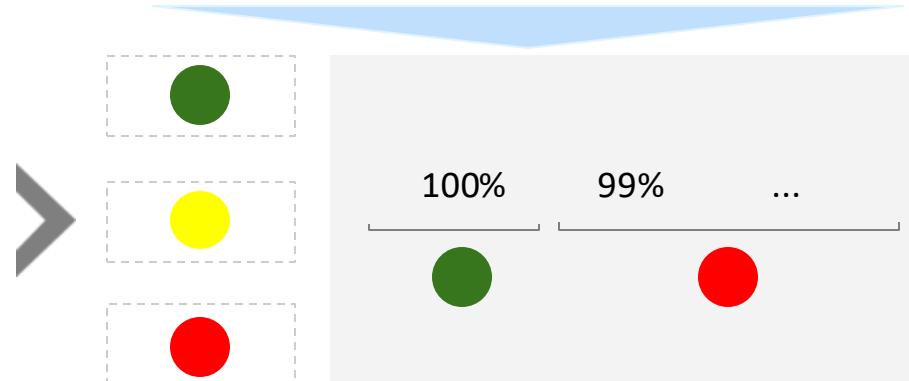
- Max. value: 10%, Min value: -10%
- Estos **valores** serán utilizados por engineering de cara al análisis técnico, construcción e implantación de la regla.
- Si la variación del valor oscila entre -10% y 10%, el resultado será OK

Si el resultado de la implantación de la regla arroja que la variación del nº de clientes ha sido de un 2%, el % de cumplimiento es el 100% (1 registro ok sobre 1 registro total)

100 % de cumplimiento de la regla (número de registros OK sobre el total de registros)

Umbrales de tolerancia definidos en la regla

- **Objetivo:** 100%. **Mínimo:** 100%.
- Al ser una regla aplicable a sólo un dato, el umbral de tolerancia es 100% (se cumple o no se cumple la regla definida).



ÍNDICE

01: Los principios de calidad

02: Las reglas de calidad

**03: La gestión, el aseguramiento y el
mantenimiento de la calidad en los datos**

04: La retroalimentación de los procesos de
calidad

Modelo de implementación de reglas de calidad

Temas a tener en cuenta en la priorización de la implementación de las reglas de calidad:

- Alineamiento con proyectos relevantes del dominio
 - Pain Points referidos a la calidad de datos
 - Prioridad de los datos según Dominio
 - Criticidad de las reglas de calidad
 - Número de reglas implantadas por DS
- Número de DO beneficiados por cada DS

Gestión de las reglas de calidad

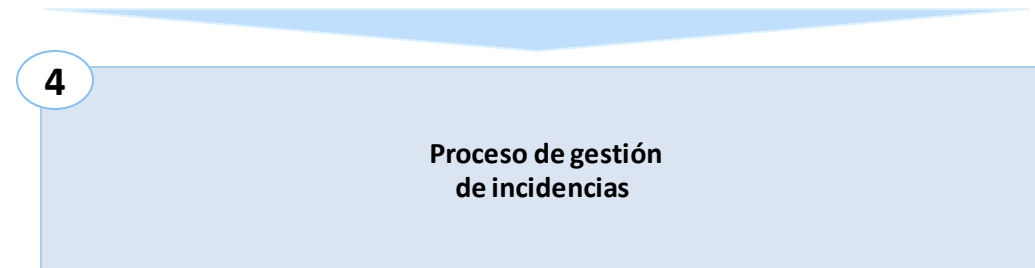
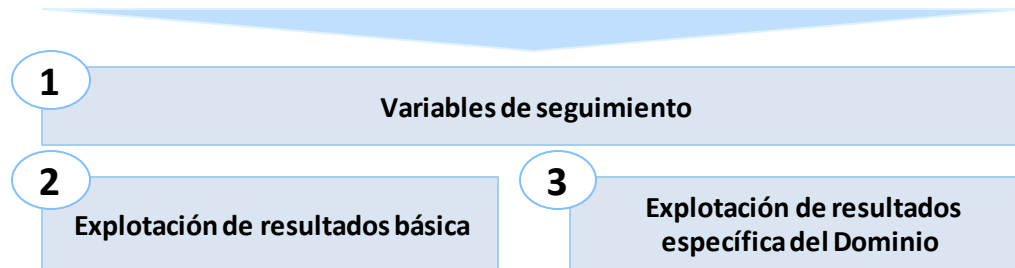
Medición y seguimiento del resultado de las reglas de calidad

- Medir **los resultados de las reglas de calidad de manera homogénea** (Quality Output Template) y complementar con soluciones adicionales del Dominio para gestión interna (explotación específica)
- Complementar con **datos de Content y Quality**



Gestión de incidencias

- Establecer un **circuito que permita** establecer planes de acción para **solventar las incidencias** que se identifiquen



Gestión de las reglas de calidad

- 1 Valorar bajo criterio experto de usuario, la evolución del resultado de cumplimiento de la regla.
- 2 Analizar detalladamente aquellos registros que no han cumplido la regla definida
- 3 Obtener conclusiones y/o justificaciones que expliquen el incumplimiento de la regla definida
- 4 Definir plan de remediación: análisis de incidencias y/o redefinición de reglas.
- 5 Informar al los responsables de gobierno acerca de los planes de acción definidos y planificados sobre las métricas más relevantes

Gestión de las reglas de calidad

1. Valoración de resultados globales

A partir de los dashboards de seguimiento se valora, bajo criterio experto del dominio, el resultado de cumplimiento de la regla

Ejemplo:

- Regla definida: “la capacidad básica a 30 días debe estar entre +/-1% respecto a su importe de referencia”
- Porcentaje de cumplimiento= 80% (amarillo)
- Criterio experto usuario: este mes se han producido muchas salidas en cash de las cuentas corrientes y los depósitos de clientes por temor a una nueva crisis financiera.
- Valoración: Intuimos que el resultado puede ser un tema puntual, es decir, la regla está bien definida pero existe una justificación económica y social que puede explicar el resultado obtenido

2. Análisis detallado incumplimiento de la regla

Para poder analizar por qué hay registros que no cumplen la regla definida, será necesario que el data owner defina los ficheros de apoyo que considere necesario para realizar este análisis.

Puntos relevantes a tener en cuenta:

- Mínimo nivel de información al que se identifica un registro
- Volumen de población sobre el que se ha implementado la regla
- Atributos que enriquecen el análisis de registros que no cumplen la regla
- Lugar donde se pueden poner a disposición los resultados de los registros evaluados,
- Frecuencia de generación del fichero de detalle
- Profundidad de almacenamiento de resultados requerida

Gestión de las reglas de calidad

3. Conclusiones/justificaciones

Tras el análisis realizado, las conclusiones que podemos obtener son las siguientes:

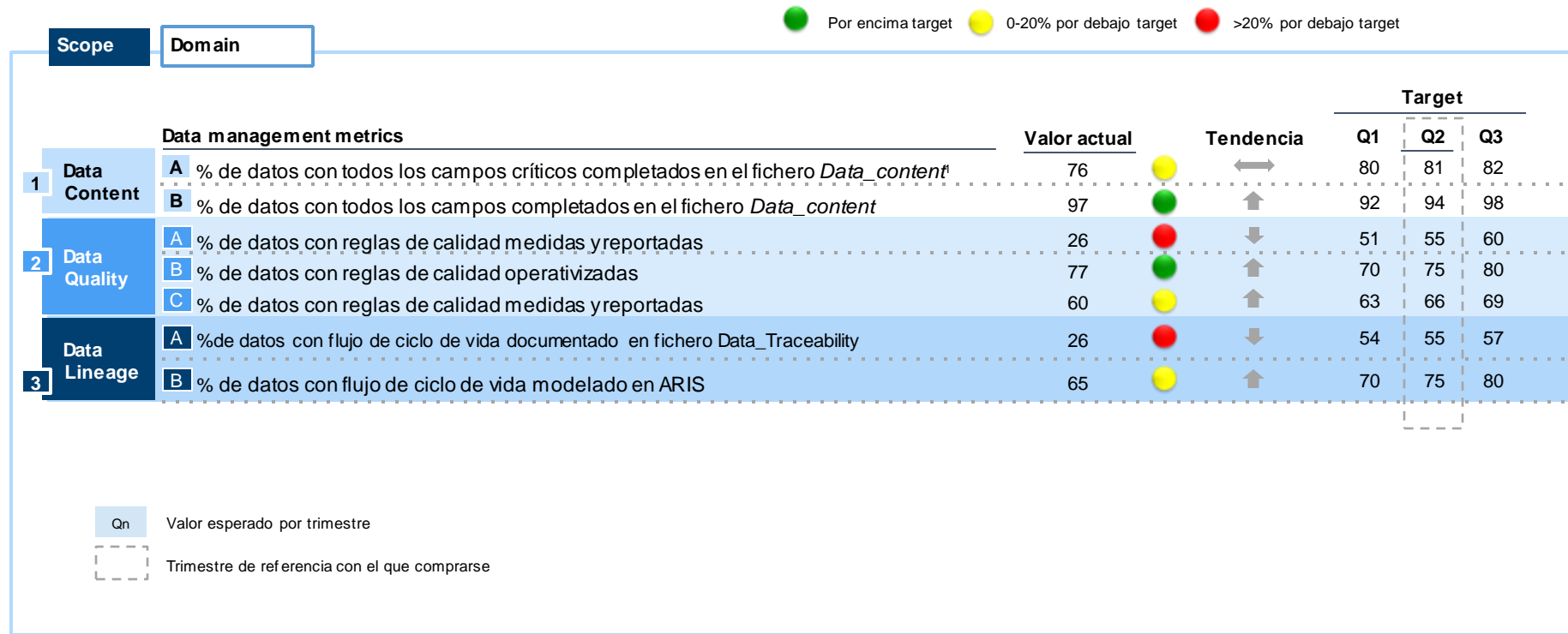
Tipo Conclusión	Descripción	Ejemplos
Justificación funcional	Razones funcionales bajo criterio experto del usuario de particularidades o comportamientos de los datos en ciertos escenarios	Este mes se ha provisionado un cliente relevante por entrar en suspensión de pagos que produce un incremento superior al esperado al contrastar este saldo con el del mes anterior.
Incidencia en datos	Identificado el error en alguno de los data sources implicados dentro del proceso del ciclo de vida del dato.	Para la regla <i>“siempre que el tipo de interés sea variable, el índice de referencia debe venir informado”</i> . Si para la aplicación de préstamos no viene informado, se trata de una incidencia en el operacional
Reglas mal definidas	Identificado que la regla de calidad definida no es tan exhaustiva como se esperaba o que se debe de ajustar a las especificaciones del dato que estamos evaluando	Para la regla <i>“Si un contrato está en mora, la fecha de entrada en inversión irregular debe venir informada”</i> . En esta definición no se ha tenido en cuenta que para contratos que nacen ya en default (compra de activos malos por fondos buitres), el contrato puede estar en mora pero nunca ha estado en situación de impago a 30 días.

4. Plan de remediación

Tipo Conclusión	Plan de remediación
Justificación funcional	Al tratarse de un tema puntual y justificado, no se llevaría a cabo ningún plan de remediación
Incidencia en datos	Comunicar la incidencia detectada al system owner responsable del data source para evaluar la solución
Reglas mal definidas	Redefinición de reglas por parte del data owner. Estas reglas se consideran descartadas temporalmente

El conocimiento experto del usuario sobre sus datos es necesario pero no suficiente para poder concluir sobre los resultados de calidad obtenidos

Ejemplo de cuadro de mando



ÍNDICE

01: Los principios de calidad

02: Las reglas de calidad

03: La gestión, el aseguramiento y el
mantenimiento de la calidad en los datos

**04: La retroalimentación de los procesos de
calidad**

Retroalimentación de los procesos de calidad

La retroalimentación en los procesos de calidad se refiere a la información que se recopila después de una evaluación de datos para identificar áreas de mejora y fortalezas en la gestión de la calidad de los datos.

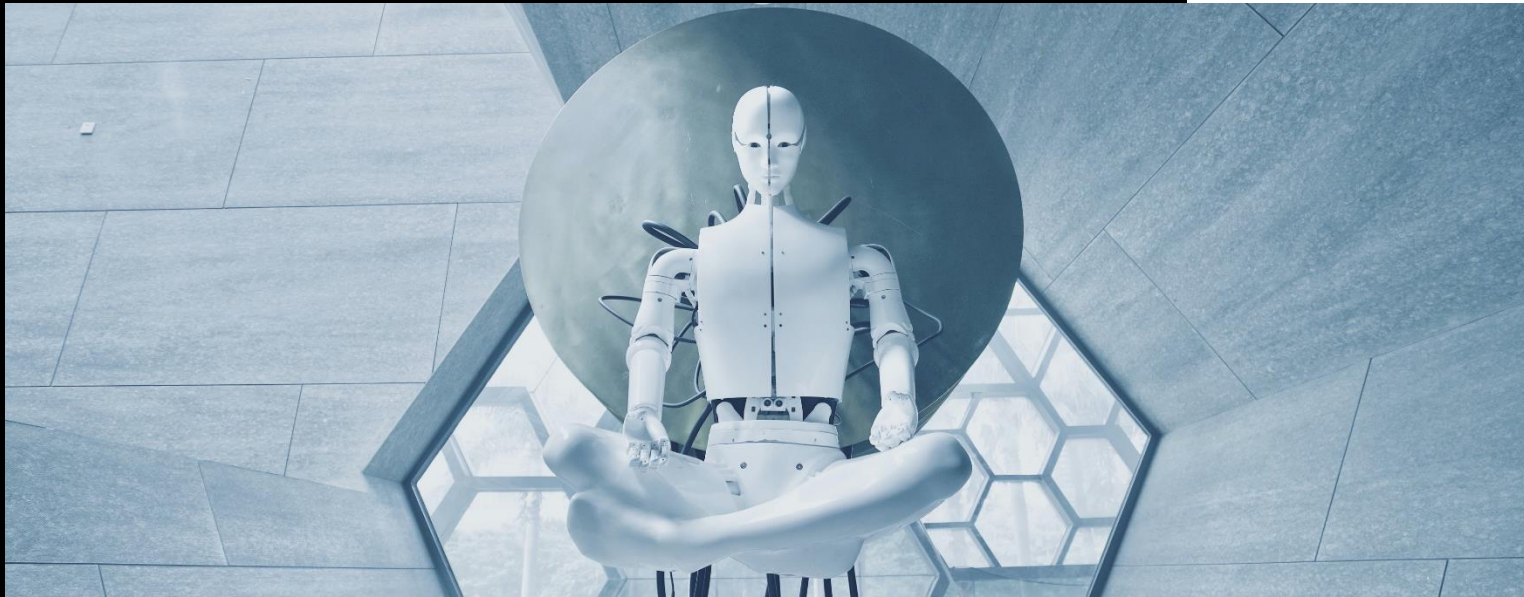
- **Identificación de Problemas:** Permite detectar errores y áreas de mejora en tiempo real.
- **Optimización de Procesos:** Al entender las debilidades, se pueden ajustar y mejorar los procesos.
- **Mejora Continua:** La retroalimentación es esencial para un ciclo de mejora continua en la gobernanza de datos.

Métodos de Retroalimentación:

- **Auditorías Internas:** Revisión periódica de los datos y procesos internos.
- **Encuestas:** Obtener feedback directo de los usuarios y stakeholders sobre la calidad de los datos.
- **Análisis de Metadatos:** Revisar la calidad y consistencia de los metadatos para asegurar que los datos son comprensibles y utilizables.

Acciones Post-Retroalimentación:

- **Corrección de Errores:** Actuar sobre los errores identificados de inmediato.
- **Revisión de Procedimientos:** Ajustar y actualizar procedimientos basados en el feedback.
- **Formación y Capacitación:** Si se identifican errores recurrentes, puede ser necesario capacitar al equipo en prácticas de calidad de datos.



- La calidad de los datos es crucial para tomar decisiones informadas y estratégicas en la organización.
- Asegurar datos precisos y confiables minimiza riesgos y asegura el cumplimiento con regulaciones de datos.
- Datos de alta calidad potencian la innovación y optimizan el valor obtenido de las iniciativas relacionadas con los datos.

Conclusiones

Tema 3

Ciclo de vida y gobierno del dato
dentro de la organización III: Data
Mesh y Data as a Product

Objetivos

01

Más allá de los modelos tradicionales de aseguramiento de la calidad

02

Calidad delegada, calidad como proceso y calidad integrada

03

Data Mesh: Calidad embebida y Datos como producto

ÍNDICE

01: Más allá del data governance

02: Data Mesh: Los datos como producto

Principales retos en la adopción de los modelos de gobierno y calidad

- **Una misma talla no le vale a todos los datos.** El origen del dato (transaccional, insight, informacional, etc), los atributos del dato (confidencialidad, complejidad de elaboración, etc) y su uso (un solo dominio, compartido, etc) determinan las necesidades de gobierno y control (mayores, menores y más o menos centralizadas).
- **Data governance & Data Quality reactivo no funciona.** Ingesta de datos agresiva sin aterrizar los procedimientos de gobierno desde el principio (“shift left” de los esfuerzos de gobierno). El resultado a medio plazo es la “deuda de gobierno” que se traduce en la inversión de grandes esfuerzos para cubrir esa deuda.
- **Mantener viva la llama del compromiso de todas las partes.** Las áreas de negocio son los principales “consumidores” de datos y por tanto son el corazón de un programa de gobierno de datos. El compromiso por su parte es absolutamente esencial.
- **Dimensionamiento de Equipos de Gobierno.** Sin un buen enfoque de mantener los cimientos sólidos y unos niveles altos de automatización altos, el gobierno se desplaza hacia el consumo – lleno de manualidades - y eso multiplica la necesidad de recursos necesarios.
- **Iniciativas aisladas de Gobierno.** Heterogeneidad en los planteamientos y ejecución por parte de las áreas de gobierno funcionales y técnicas. Reglas claras que los diferentes dominios ejecutan a escala.
- **Segmentación de áreas Operacionales – Ingestas - Analíticas.** Desvinculación de los procesos y equipos de los sistemas fuente / operacionales, los equipos encargados de las ingestas / aprovisionamiento del dato de los equipos más centrados en el consumo (analíticos, etc).
- **Falta de claridad en los Roles y Responsabilidades.** Una de las razones por las que muchas iniciativas de Gobierno de Datos fracasan es porque las personas involucradas no tienen claras sus funciones y responsabilidades (¿Quién define qué? ¿Quién hace qué?).

Principales retos en la adopción de los modelos de gobierno y calidad

Situación de contexto:

- Cambio constante en el ecosistema de datos
- Proliferación de nuevas fuentes de datos
- Evolución de los casos de uso
- Continua evolución de las plataformas de datos

ÍNDICE

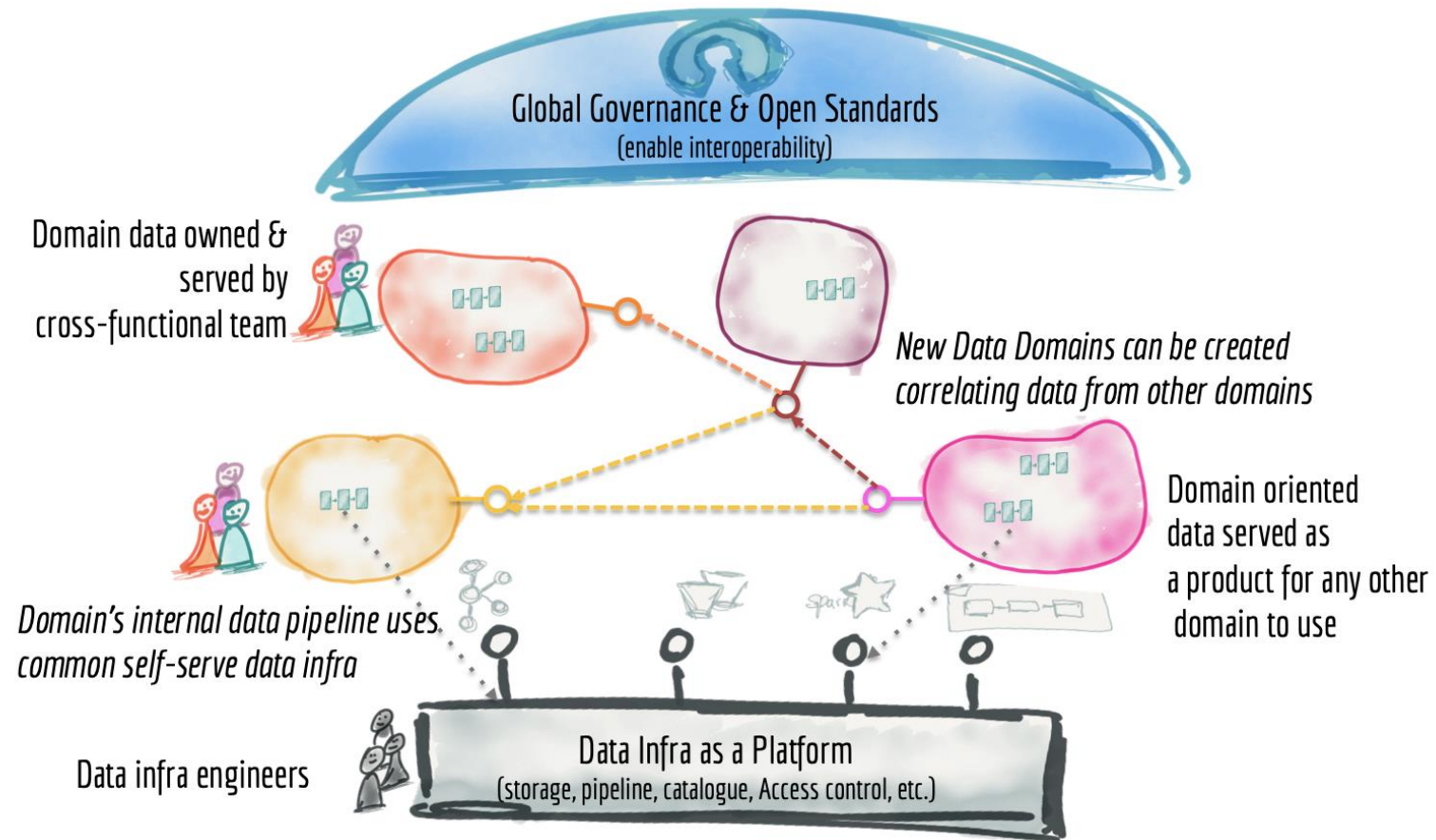
01: Más allá del data governance

02: Data Mesh: Los datos como producto

Data Mesh overview

1. **Data Ownership descentralizado orientado al dominio.** Para que el ecosistema de creación y consumo de datos sea escalable a medida que aumenta el número de fuentes de datos, el número de casos de uso y la diversidad de modelos de acceso a los datos, basta con aumentar los nodos autónomos en el Data Mesh
2. **Dato como producto.** Para que los usuarios de datos puedan descubrir, comprender y utilizar de forma segura los datos, éstos deben ser publicados bajo los estándares definidos por los equipos conocedores de los datos responsables de los dominios
3. **Gobierno federado.** Para que los usuarios de los datos puedan obtener valor de la agregación y composición de productos de datos independientes, Data Mesh impulsa el comportamiento como ecosistema que sigue las normas de interoperabilidad globales respetando la autonomía de los dominios locales

Data Mesh overview



Características básicas

Propiedad de datos por dominio

Los dominios de datos existentes pueden conservarse, ya que con frecuencia representan el estado actual de manera fiel. A su vez pueden construirse nuevos dominios que representen nuevos conocimientos para la empresa.

Datos como producto

Productos de datos en los que los datos se publican en lugar de extraerse, y la propiedad persiste.

Datos gobernados siempre

Gobierno Federado y global con autonomía local de dominio, normas y reglas globales

Datos disponibles siempre, autoservicio

Paradigma de infraestructura de autoservicio y puede implementarse en una plataforma de datos de autoservicio y herramientas

Equilibrio entre lo **centralizado** y lo **descentralizado**. Data Mesh propone lo mejor de ambos mundos

Owner del dominio

Poner en valor el conocimiento de los equipos que gestionan el BAU de las áreas que generan los datos, con **control total en la evolución de los datos, conocimiento del dato con precisión y visibilidad en el uso del dato**

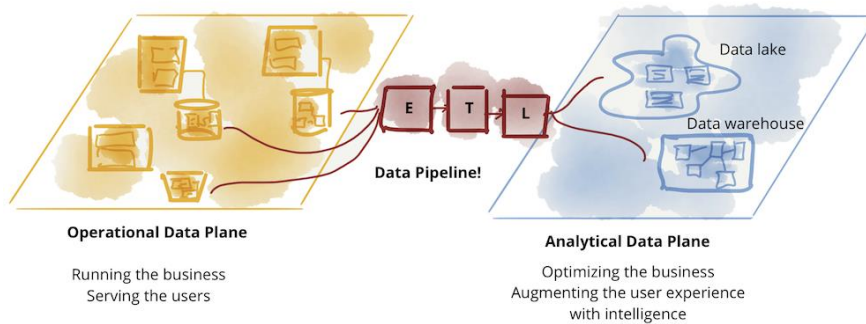
Organización central

Vista única federada, con un inventario completo (incluyendo los datos grises que no pertenecen a ningún dominio concreto), un gobierno programático y unas metodologías habilitadas, y priorizando los esfuerzos basados en el consumo y las necesidades reales.

Data Mesh no es una herramienta o un proceso, es una estrategia.

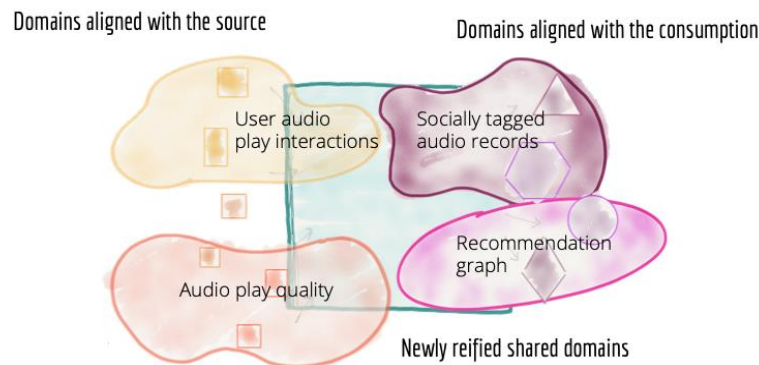
Data Ownership por dominio

Un approach de Data Mesh promueve la distribución de la responsabilidad y toma de decisiones a los equipos cercanos a los procesos que generan datos, con dominios específicos de información vinculados al contenido del dato



Dominios de datos alineados a los Data Sources: Representan los hechos y la realidad de las áreas de negocio

- Conocimiento de los equipos Operacionales que generan el dato
- Los datos de los dominios de datos son estructurales y cambian con poca frecuencia
- A nivel empresa, no hay mapeo 1:1 entre dominio y Sistema de origen de datos
- Existen muchas fuentes de datos que son agregaciones de datos puestos al servicio de la organización.



Dominios alineados con el consumo: Promueven la estrategia del dato a través de casos de uso

- Los conjuntos de datos del dominio del consumidor y los equipos que los poseen, tienen como objetivo satisfacer un grupo de casos de uso estrechamente relacionados
- Estructuralmente pasan por más cambios
- Transforman los eventos del dominio fuente para agregar vistas y estructuras que se ajusten a un modelo de acceso particular

Las organizaciones han adoptado la propiedad y la responsabilidad sobre los datos por dominio al implementar capacidades operativas, pero no lo han hecho para la gestión de los datos

Data as a Product

Uno de los retos de las arquitecturas de datos analíticos existentes es la **elevada fricción y el coste de descubrir, comprender, confiar y en última instancia, utilizar datos de calidad.**

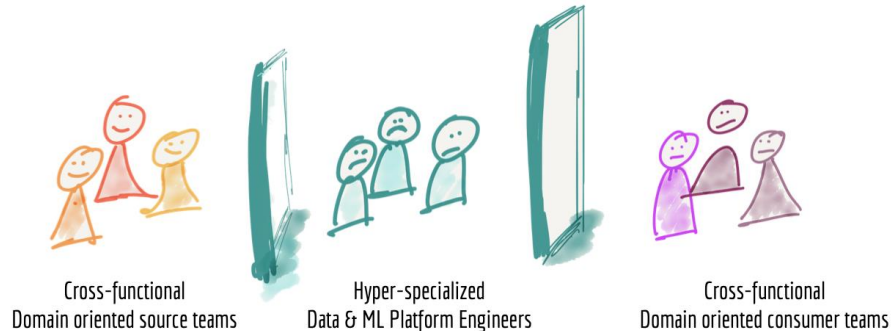
Modelo de responsabilidad invertido: la responsabilidad de la **calidad** de los datos se desplaza hacia arriba, lo más cerca posible de la **fuentes de los datos:**

- **Descubrible:** Un catálogo de datos disponibles con su metainformación, como sus propietarios, fuente de origen, linaje, conjuntos de datos de muestra, etc. Cada producto de datos del dominio debe registrarse en este catálogo de datos centralizado para facilitar su descubrimiento.
- **Direccionable:** Los dominios pueden almacenar y servir sus conjuntos de datos en diferentes formatos.
- **Confiable y veraz:** Dato con gobierno implementado según políticas definidas, calidad medible y lineage que represente con exactitud la descripción de su diccionario.
- **Auto-Explicativo:** La creación de conjuntos de datos como productos con un mínimo de fricción para que D.Eng y D.Sc los utilicen requiere una semántica y una sintaxis de los datos bien descritas, idealmente acompañadas de conjuntos de datos de muestra como ejemplos. Los esquemas de datos son un punto de partida para proporcionar activos de datos de autoservicio.
- **Inter-Operable:** La clave para una correlación eficaz de datos entre dominios es seguir ciertas normas y reglas de normalización. Estas estandarizaciones deben pertenecer a un gobierno global, para permitir la interoperabilidad entre conjuntos de datos de dominios múltiples.
- **Seguro:** Debe marcar las guidelines de seguridad, en términos de clasificación, acceso, etiquetado y privacidad

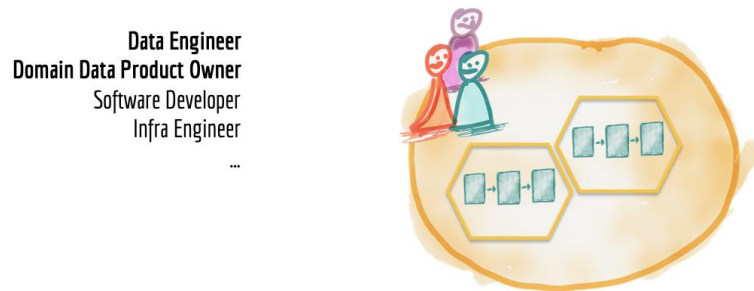
El principio de los datos como producto está pensado para resolver el problema de la calidad de los datos y de los antiguos silos de datos. Para ello, los datos analíticos proporcionados por los dominios deben ser **tratados como un producto**, y los dominios que hacen uso de esos datos deben ser tratados como **clientes**

Data Product Owner

Data Product Owner como responsable de **construir, mantener y servir los productos de datos del dominio.**



De equipos de origen desconectados, consumidores frustrados y un equipo de plataforma de datos sobrecargado.



A un equipo integrado con conocimientos, skills y autonomía para publicar datos al resto de la organización bajo unas premisas globales de aplicación local por dominio..

Equipos en el modelo tradicional

- Grupo de ingenieros de datos hiperespecializados y aislados de las unidades operativas, sin conocimiento del negocio y del dominio
- Desconocimiento de dónde se originan los datos ni dónde se utilizan, sin embargo toman decisiones y ejecutan acciones
- No existe ningún incentivo en proporcionar datos veraces y correctos

Células Data Product Owners en el modelo Data Mesh

- **Data Product Owner:** Define e implementa medidas objetivas para publicar un dato como producto responsable del grado de satisfacción de los usuarios que consumen el dato medido bajo NPS: **Data Governance y Data Quality. Documentación y diccionariado. Linaje funcional y técnico. Entendimiento del dato publicado**
- **Data Engineers:** Ejecutar las acciones dentro del ecosistema de datos de la organización: Aplicar **Calidad del dato a nivel Plataforma** e interacción con los equipos de Ingeniería, Arquitectura y Seguridad para asegurar los KPIs sobre el dato publicado. **Documentar, alta y actualización en el catálogo y aplicación del gobierno (F+T).** Colaborar con los equipos de ingeniería en los **procesos de ingesta y remediación del dato**
- **SW Developer & Infra Engineer:** Aportar la experiencia en la **operación de los sistemas como fuente de datos** para el desarrollo de “productos de datos”

Gobierno descentralizado

La implementación de Data Mesh requiere adoptar un **modelo de gobierno descentralizado**, que promueva la autonomía de los dominios y su interoperabilidad a través de unos estándares globales, descentralizando la función de gobierno a través de **células multidisciplinares**

Modelo de gobierno

- Esquema de toma de decisiones guiado mediante la federación de los llamados Domain Data Product owners, con autonomía y poder para la toma de decisiones a nivel de dominio local, mientras crean y se adhieren al conjunto de reglas globales – reglas aplicadas a todos los productos y sus interfaces – para asegurar un ecosistema saludable e interoperable.
- El **dominio es autónomo** a la hora de aplicar el gobierno, siguiendo las guías y reglas de un gobierno central.

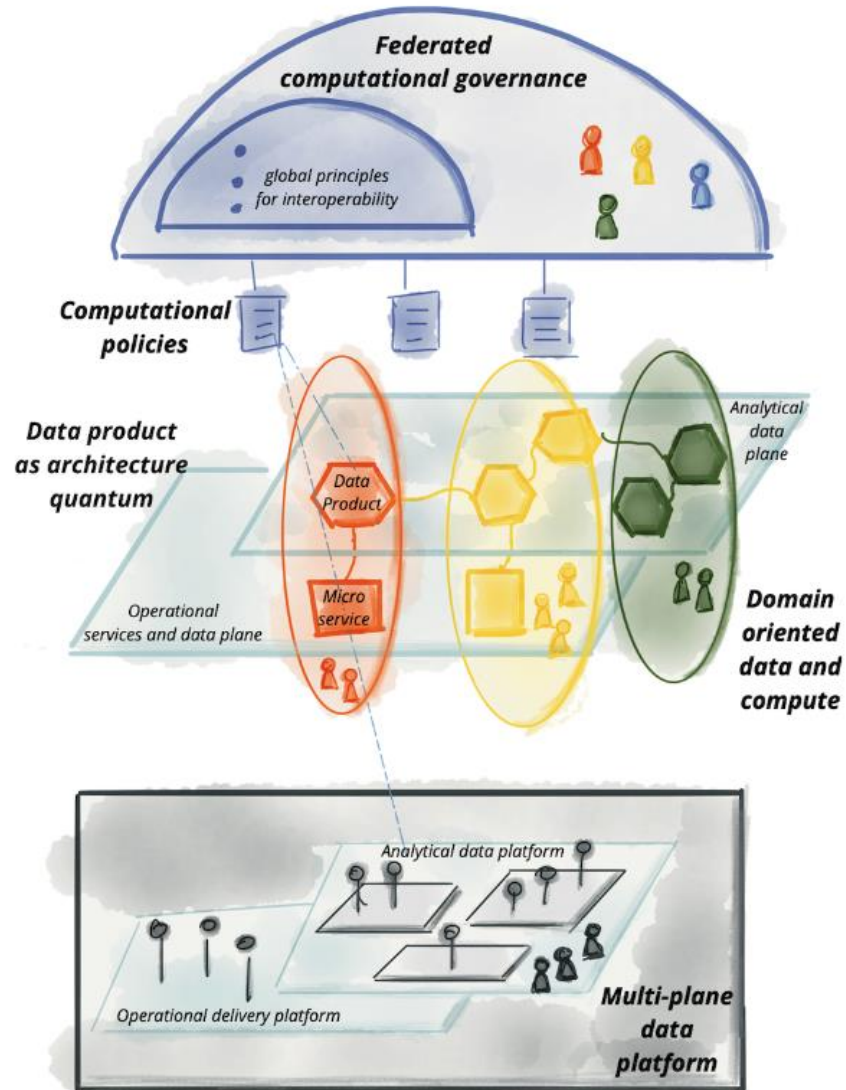
Balance

- Qué decisiones se necesitan tomar en cada dominio de forma particular y que decisiones deben ser tomadas globalmente para todos los dominios.
- Las decisiones globales tienen un propósito, crear interoperabilidad y el llamado **efecto de red compuesto**, a través del descubrimiento y composición de productos de datos que se pueden enlazar entre sí.

Estructura. Para que el modelo de gobierno federado funcione, se necesita una estructura organizativa de apoyo, un modelo de incentivos y una arquitectura desplegada:

- Para llegar a decisiones y normas globales de interoperabilidad.
- Asegurando que en todo momento se respeta la autonomía local de los dominios.
- Facilitando la correcta implementación de políticas globales de una manera eficiente, segura e interoperable.

Gobierno descentralizado



Domain-oriented
decentralized data
ownership and
architecture

So that the ecosystem creating and consuming data can scale out as the number of sources of data, number of use cases, and diversity of access models to the data increases; simply increase the autonomous nodes on the mesh.

Data as a product

So that data users can easily discover, understand and securely use high quality data with a delightful experience; data that is distributed across many domains.

Self-serve data
infrastructure as a
platform

So that the domain teams can create and consume data products autonomously using the platform abstractions, hiding the complexity of building, executing and maintaining secure and interoperable data products.

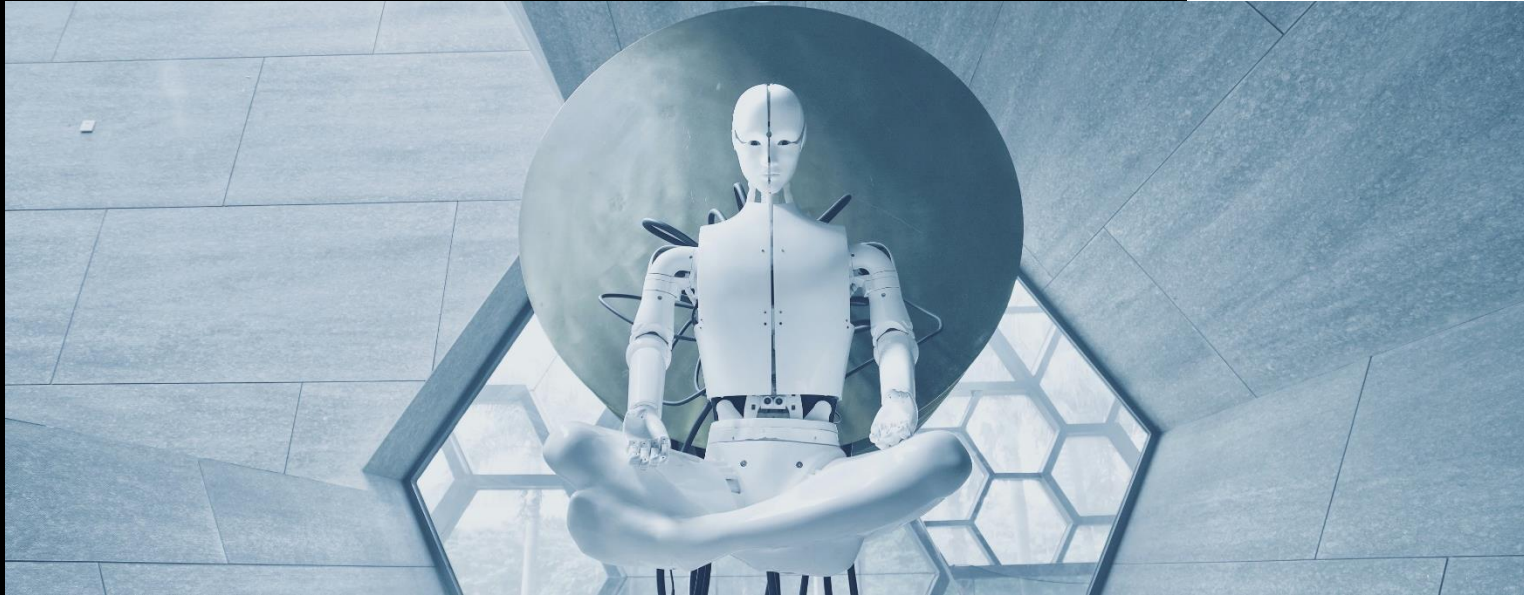
Federated
computational
governance

So that data users can get value from aggregation and correlation of independent data products - the mesh is behaving as an ecosystem following global interoperability standards; standards that are baked computationally into the platform.

Gobierno descentralizado

Pre data mesh governance aspect	Data mesh governance aspect
Centralized team	Federated team
Responsible for data quality	Responsible for defining how to model what constitutes quality
Responsible for data security	Responsible for defining aspects of data security i.e. data sensitivity levels for the platform to build in and monitor automatically
Responsible for complying with regulation	Responsible for defining the regulation requirements for the platform to build in and monitor automatically
Centralized custodianship of data	Federated custodianship of data by domains
Responsible for global canonical data modeling	Responsible for modeling <u>polysems</u> - data elements that cross the boundaries of multiple domains
Team is independent from domains	Team is made of domains representatives

Pre data mesh governance aspect	Data mesh governance aspect
Aiming for a well defined static structure of data	Aiming for enabling effective mesh operation embracing a continuously changing and a dynamic topology of the mesh
Centralized technology used by monolithic lake/warehouse	Self-serve platform technologies used by each domain
Measure success based on number or volume of governed data (tables)	Measure success based on the network effect - the connections representing the consumption of data on the mesh
Manual process with human intervention	Automated processes implemented by the platform
Prevent error	Detect error and recover through platform's automated processing



- You build data, you own it. Gobierno desde el origen
- Células multidominio con responsabilidad completa end to end
 - Capacitar al negocio para responsabilizarse de la calidad y confiabilidad del dato

Conclusiones

Demos



[Lake Formation](#)

[Data Zone](#)



[Data Catalog](#)



Collibra

[Data Quality &](#)

[Observability](#)

[Catalog & Lineage](#)



[Dataplex](#)

¡Gracias!

"La estadística es el arte de nunca tener que decir que estás equivocado"

- Vannebar Bush

