

Chapter 1. The Journey to Becoming Data-Driven

Piethein Strengholt

La Idea Principal: El Mundo Cambió y los Datos Ya no Pueden Vivir en un Sótano Centralizado

Imagina que los datos de una empresa están guardados en un único y gigantesco almacén en el sótano. Un solo equipo de “guardianes” se encarga de todo. Cada vez que alguien de un departamento (Ventas, Marketing, etc.) necesita algo, tiene que bajar, llenar un formulario, y esperar. Este proceso es lento, frustrante, y los guardianes del sótano no entienden realmente para qué necesita los datos el equipo de Marketing. Como resultado, la empresa es lenta, toma decisiones a ciegas y no puede competir.

Este capítulo explica que **este modelo centralizado ya no funciona**. El mundo digital es demasiado rápido, complejo y competitivo. Las empresas que sobrevivirán y prosperarán son las que tratan los datos como el corazón de su negocio, no como algo guardado en un sótano.

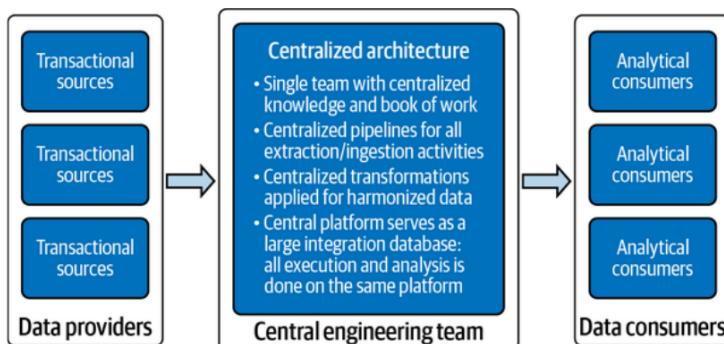


Figure 1-2. Centralized architectures are a bottleneck for many organizations: one team must wait for another team to finish its work

El viaje para convertirse en una “organización impulsada por los datos” consiste en **desmantelar ese sótano centralizado y darle a cada equipo (o “dominio”) la responsabilidad y las herramientas para gestionar sus propios datos**. Sin embargo, para evitar el caos, se necesita una estrategia clara que equilibre esta nueva autonomía con estándares y una visión común. En resumen, el futuro es la **descentralización**, pero

una descentralización organizada.

¿Por Qué Hay que Cambiar Ahora? Las 6 Tendencias que lo Están Rompiendo Todo

El autor identifica varias tendencias tecnológicas y de la industria que están obligando a las empresas a cambiar su forma de gestionar los datos.

1. **La Analítica está Fragmentando el Paisaje:** Cada vez hay más usos para los datos (Inteligencia Artificial, Machine Learning, reportes, etc.). Cada caso de uso necesita los datos de una forma específica. Esto provoca que se creen muchas copias y versiones de los datos por todas partes, un fenómeno llamado **proliferación de datos**. Es muy difícil saber cuál es el dato correcto y mantener el control.
2. **El Software se Entrega Más Rápido:** Con metodologías como **DevOps** y arquitecturas de **microservicios**, las aplicaciones ya no son grandes monstruos monolíticos, sino un conjunto de pequeños servicios que se actualizan constantemente. Esto significa que los datos también están más dispersos entre estos pequeños componentes, haciendo su gestión mucho más compleja.
3. **El Impacto Inmenso de la Nube:** Gracias a la nube y a redes más rápidas, ya no es necesario tener los datos en un solo lugar. Ahora es fácil mover terabytes de datos a donde se necesite la capacidad de cómputo (por ejemplo, a servicios de IA de terceros). Esto fragmenta aún más el panorama de datos.
4. **La Privacidad y la Seguridad son Prioridad Máxima:** Con regulaciones estrictas como el GDPR (en Europa) o CCPA (en California), las empresas están obligadas a saber exactamente qué datos personales tienen, de dónde vienen, cómo se usan y quién los ve. Escándalos como el de Cambridge Analytica y hackeos masivos han puesto esto en el centro de atención. Es casi imposible cumplir con estas reglas en un modelo centralizado y desordenado.
5. **Los Sistemas Operacionales y Analíticos Necesitan Integrarse:** Antes, el sistema que usabas para registrar una venta (operacional) era totalmente independiente del sistema que usabas para analizar las ventas del mes (analítico). Ahora, se necesita análisis en tiempo real para mejorar las operaciones diarias (ej: un modelo de IA que predice si un cliente va a abandonar la compra *mientras* está en la web). Esto exige una arquitectura que integre ambos mundos.
6. **Las Empresas Operan en Ecosistemas Colaborativos:** Las compañías ya no trabajan solas. Colaboran constantemente con socios, proveedores y clientes,

compartiendo datos a través de APIs. Esto requiere una arquitectura de datos abierta, flexible y descentralizada.

Las Arquitecturas Antiguas y Por Qué Fallan (“El Dolor de la Centralización”)

Las empresas están lastradas por dos tipos de arquitecturas centralizadas que, según el autor, ya no sirven para el mundo actual.

1. El Enterprise Data Warehouse (EDW) - El Intento de la “Única Fuente de Verdad”

- **La idea:** Crear un único almacén de datos central con todos los datos de la empresa, perfectamente limpios, unificados y estructurados. El objetivo era tener una “única fuente de la verdad”.
- **El problema:**
 - **Cuello de botella:** Un equipo central se encarga de todo. Si 10 equipos piden datos, tienen que hacer cola. Esto es increíblemente lento y frena la agilidad del negocio.
 - **Rígido y Frágil (“Big Ball of Mud”):** El EDW se convierte en una “gran bola de barro”. Es un sistema tan complejo y con tantas dependencias entrelazadas que nadie se atreve a tocarlo por miedo a romper algo. Cualquier cambio es un proyecto de meses.
 - **Pérdida de Contexto:** El equipo central no tiene el conocimiento de negocio de cada departamento. Al intentar unificar los datos (ej: definir “cliente”), crean un modelo que al final no tiene sentido para nadie.
 - **Calidad de Datos y Propiedad:** Si los datos de origen son malos, el equipo central intenta arreglarlos con parches, pero no solucionan el problema en la fuente. Nadie se siente dueño de la calidad.

2. El Data Lake - El Intento de “Guardarlo Todo”

- **La idea:** Como estructurar los datos en un EDW es muy lento, mejor “arrojamos” todos los datos crudos (estructurados, no estructurados, de cualquier tipo) a un gran “lago de datos” y ya veremos cómo los usamos después.
- **El problema:**

- **Se Convierte en un Pantano (Data Swamp):** Sin un gobierno claro, el lago de datos se llena rápidamente de datos de mala calidad, duplicados, sin documentación e incomprensibles. Se convierte en un “pantano de datos” inútil.
- **Fontanería de Datos (Data Plumbing):** Los científicos e ingenieros de datos terminan pasando el 80% de su tiempo buscando, limpiando y conectando “tuberías” de datos, en lugar de generar valor.
- **Alto Acoplamiento a la Fuente:** Como los datos están crudos, si el sistema original cambia su estructura, todos los análisis que dependen de esos datos se rompen.

El problema fundamental de ambos enfoques es el **pensamiento centralizado**. Aleja a los expertos en datos de los expertos del negocio, frena la innovación y, simplemente, no puede escalar a la velocidad y complejidad del mundo actual.

Hacia una Gestión de Datos Moderna: DAMA y los Nuevos Enfoques

El autor introduce el marco **DAMA-DMBOK**, que es como la “biblia” de la gestión de datos. Este marco define 11 áreas funcionales clave (veáse la figura 1-1), con la **Gobernanza de Datos** en el centro. Es una guía útil para entender todas las piezas del rompecabezas (Calidad, Seguridad, Arquitectura, Metadatos, etc.).

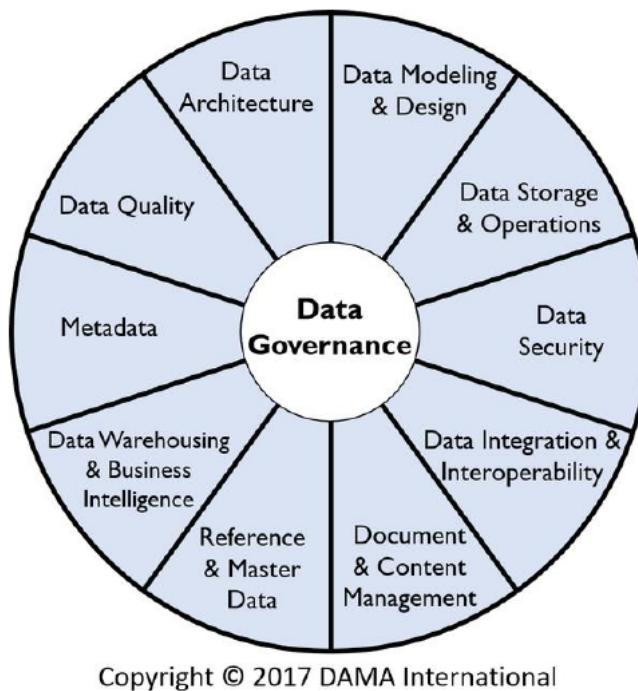


Figure 1-1. The 11 functional areas of data management

Sin embargo, el autor critica que DAMA-DMBOK se ha quedado anticuado en algunos aspectos, especialmente en su visión sobre la integración de datos, que sigue siendo muy centralista y no aborda bien las arquitecturas descentralizadas modernas.

Por eso, han surgido nuevos enfoques que sí lo hacen:

- **Data Mesh:** Es una visión **socio-técnica** (enfocada en las personas y la organización). Propone que los datos sean propiedad de los equipos de negocio que los generan (“dominios”). Estos equipos tratan sus datos como productos y los ofrecen al resto de la empresa a través de una plataforma de autoservicio.
- **Data Fabric:** Es una visión más **tecnológica**. Propone usar metadatos e inteligencia artificial para crear una capa “inteligente” que automatiza y simplifica el acceso, la integración y el gobierno de los datos, sin importar dónde estén físicamente.

El autor aclara que no son técnicas rígidas y que pueden (y deben) coexistir y complementar las inversiones ya existentes.

La Solución: Definir una Estrategia de Datos

No se puede pasar a un modelo impulsado por los datos sin un plan. Una **estrategia de datos** es el mapa que guía la transformación de la cultura, la gente, los procesos y la tecnología. El capítulo termina con una guía detallada de 16 puntos sobre cómo crear esta estrategia. Aquí están los pasos más importantes resumidos:

1. **Empezar por el Negocio, no por la Tecnología:** La estrategia debe responder a: “¿Cómo nos ayudarán los datos a alcanzar nuestros objetivos de negocio?”.
2. **Equilibrar “Defensa” y “Ataque”:**
 - **Defensa:** Proteger los datos, cumplir con regulaciones, asegurar la calidad (control total).
 - **Ataque:** Usar los datos para innovar, crear nuevos productos y ser más competitivos (flexibilidad).
3. **Pensar en las Personas y la Cultura:** ¿Tenemos el talento necesario? ¿Están los equipos dispuestos a compartir datos y asumir responsabilidad? La cultura es lo más difícil de cambiar.
4. **Crear un Plan con Hitos Medibles:** Definir objetivos claros y KPIs (indicadores) para saber si la estrategia está funcionando y generando valor.
5. **Comunicar la Estrategia a Toda la Organización:** Todo el mundo debe entender por qué se está haciendo este cambio y cuál es su papel en él.

Conclusión (Wrapping Up)

- El cambio hacia una organización impulsada por los datos es inevitable y fundamental para sobrevivir.
- Las viejas arquitecturas centralizadas (EDW, Data Lakes) han fracasado porque son cuellos de botella que no pueden escalar.
- **La descentralización es el futuro inevitable de los datos,** pero necesita ser mitigada con estándares y una alineación central para no caer en el caos.
- El punto de partida de toda transformación es una **estrategia de datos clara y dirigida desde el negocio**, no desde TI. Sin esta guía, cualquier esfuerzo de implementación, por bueno que sea, está destinado a fracasar a largo plazo.