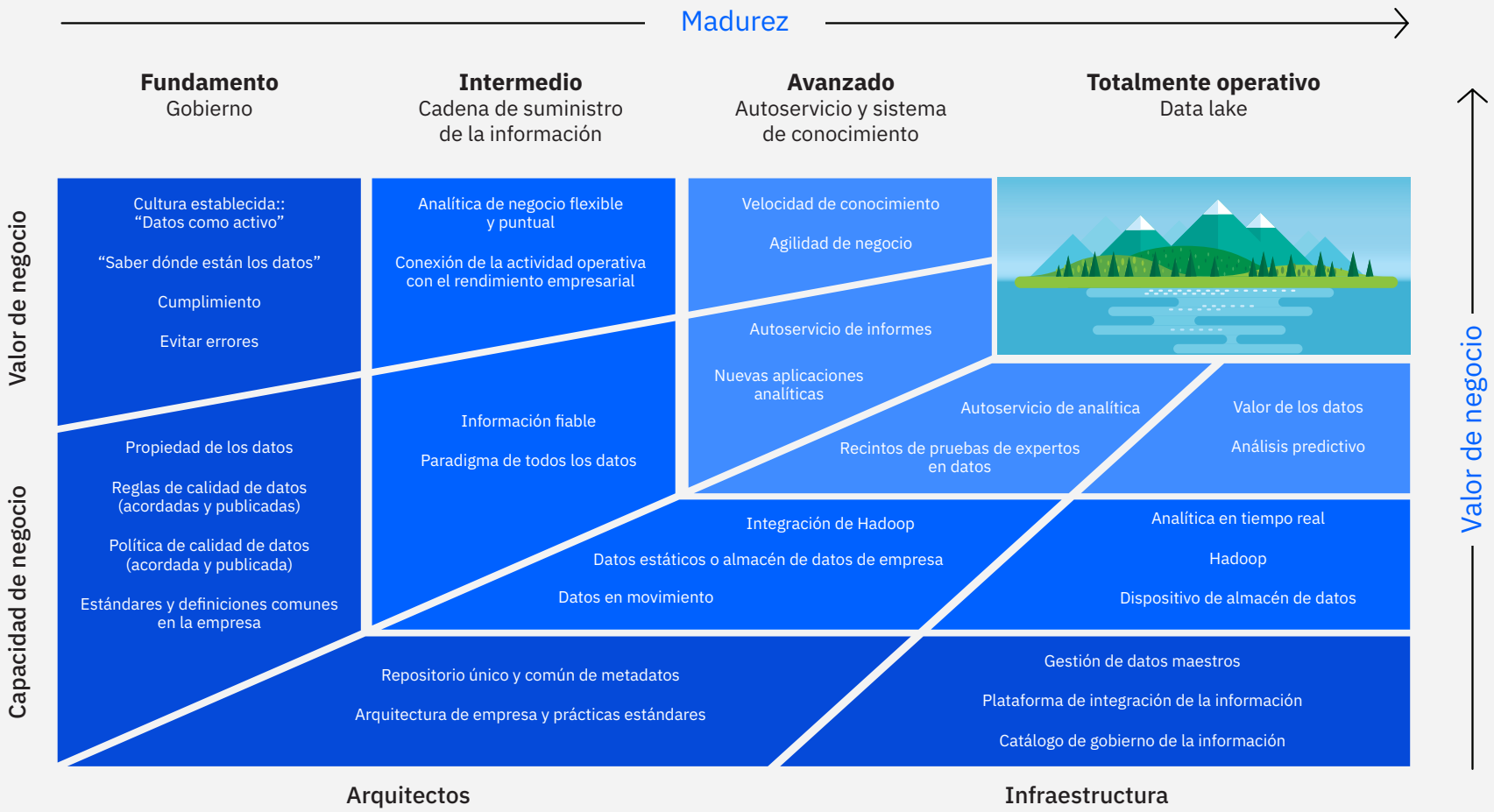
# Data lake gobernado para conocimiento empresarial

*Explore los principales componentes para ofrecer datos fiables de forma eficaz*

IBM **Cloud**

# Los data lakes gobernados añaden valor

Los data lakes son soluciones ideales para las organizaciones que priorizan los datos en su estrategia de operaciones. El uso compartido seguro de datos es un factor crucial cuando múltiples equipos deben acceder a datos empresariales. Para contribuir a gestionar dicha utilización, las organizaciones pueden depender de un data lake gobernado que aloje datos estructurados y no estructurados sin tratamiento previo, de forma fiable, segura y administrada. Para las organizaciones que generan valor a partir de sus datos, incluidos los datos sobre clientes, empleados, transacciones y otros activos, los data lakes gobernados crean oportunidades para identificar, comprender, compartir y actuar de forma fiable en dicha información.



# Arquitectura de un data lake gobernado

Las principales decisiones de diseño caracterizan la arquitectura de un data lake gobernado. Un embalse de datos está formado por tres componentes principales. Los cierres del data lake proporcionan plataformas que almacenan datos y ejecutan operaciones analíticas lo más cerca posible de los datos. Los servicios del data lake localizan, acceden, preparan, transforman, procesan y trasladan datos desde y hacia los repositorios del embalse de datos. Por último, la gestión de la información y el entramado de administración contribuyen a gobernar y gestionar los datos del data lake.

Las capacidades de gobierno validan y aumentan la calidad de los datos y están diseñadas para proteger los datos de su uso incorrecto. Esta medida garantiza la renovación, retención y finalmente la eliminación de los datos en los puntos adecuados de su ciclo de vida.

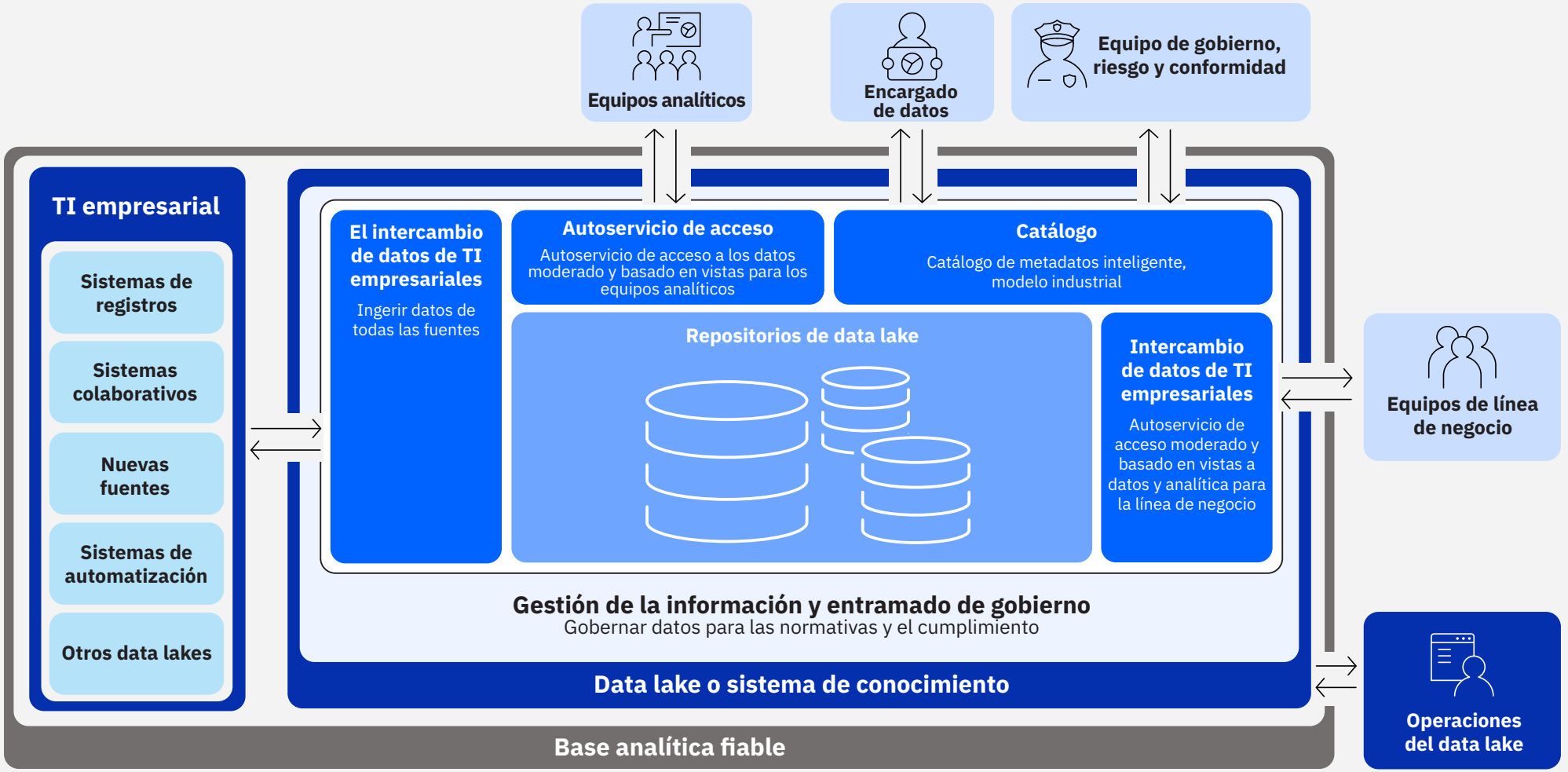
El gobierno, la organización de los datos y la capacidad para tener confianza en su calidad, es un aspecto importante de la gestión del data lake. Aunque los data lake se hayan diseñado para ofrecer un acceso flexible a los datos, es necesario disponer de un sistema de gobierno para que los datos sean seguros, estén protegidos y sigan siendo útiles. El data lake gobernado se puede ilustrar mediante sus capas, que son las siguientes:

- Fundamental, principalmente basado en el gobierno de datos
- Intermedio, que amplía los repositorios iniciales del data lake con nuevos tipos de datos y comportamientos de datos adicionales
- Avanzado, que permite el autoservicio de analítica

Cada capa contiene un valor específico para los diferentes grupos de la organización. Los arquitectos se benefician de una arquitectura de referencia publicada, soportada por un único repositorio común de metadatos. Los expertos en datos se benefician de un área controlada en la que pueden depositar recintos de pruebas en curso.

Las ventajas fundamentales de un data lake se derivan del gobierno. El gobierno impulsa una cultura de “primero los datos” en los de los usuarios comerciales se apropian de los datos y se ponen de acuerdo en las reglas y políticas. Las definiciones compartidas crean un conocimiento mutuo que contribuye a evitar confusiones entre los equipos. Con este terreno común se puede acceder a datos fiables y acelerar la obtención de información de las aplicaciones analíticas. El valor empresarial se desplaza desde la conciencia de la existencia de datos y su importancia hasta una analítica flexible en cualquier instante.

Un data lake modular y escalable consiste en varios elementos que fomentan el acceso de autoservicio en toda la organización.

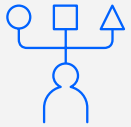


# Los cuatro tipos de consumidores de datos

Los usuarios que consumen datos del data lake varían de varias formas. Conocer la diferencia entre los enfoques que adoptan en los datos es un aspecto importante para el éxito del gobierno.

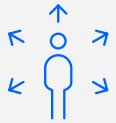
**Equipos analíticos**

- Expertos en datos que gestionan datos y construyen modelos
- Desarrolladores analíticos que convierten modelos en aplicaciones
- Desarrolladores de aplicaciones que incorporan aplicaciones analíticas en sistemas operativos



**Encargado de datos**

- Optimiza la calidad de los datos y preparan trabajos de ETL
- Cataloga datos y lleva a cabo la gestión de los metadatos
- Busca el equilibrio entre la privacidad y la protección de datos



**Equipo de gobierno, riesgo y conformidad**

- Especialistas en gobierno de datos que crean políticas de gobierno de datos y seguridad
- Protegen los datos para asegurarse de que se aplican los controles de privacidad en todos los procesos
- Compilan requisitos de retención, archivado y eliminación, así como se aseguran de que los datos son conformes a las políticas y normativas

**Equipos de línea de negocio**

- Ejecutivos de línea de negocio (LOB) como CMOs, CFOs o CHROs
- Directores ejecutivos de datos que emergen como propietarios comerciales de los datos
- Ejecutivos de LOB que implementan sistemas para resultados de negocio o conocimientos factibles específicos

# Componentes de un data lake gobernado

Un data lake gobernado es una arquitectura de referencia independiente de una tecnología específica, que incluye procesos de gobierno y gestión. No es Hadoop o un almacén de datos de empresa que se pueda comprar o sustituir. Un data lake gobernado es una solución local o basada en nube para organizaciones que desean colocar los datos en el centro de sus operaciones. Los componentes de un data lake gobernado incluyen los siguientes elementos:

**El intercambio de datos de TI empresariales** puede extraer, analizar, refinar, transformar e intercambiar datos entre data lakes y sistemas de TI empresariales, así como trasladarlos desde charcos de datos hasta data lakes. Limpia los datos y monitoriza la calidad de los datos de forma continua.

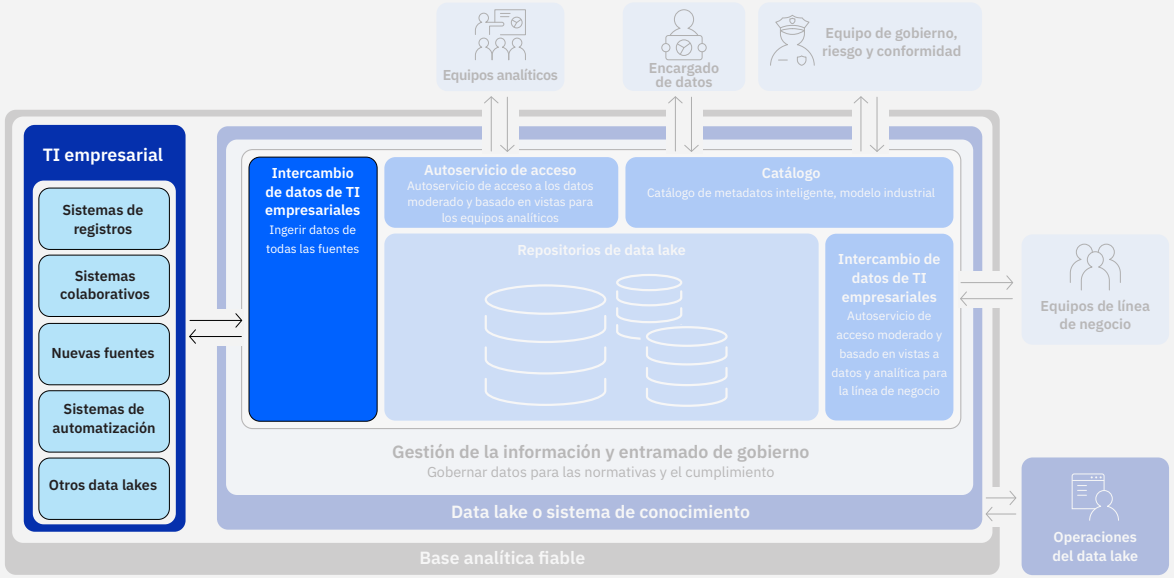
**Servicios de catalogación** que describen los datos del data lake: su significado, cómo se clasifican y los requisitos de gobierno resultantes que colocan en los datos.

**El gobierno** ayuda a gobernar los datos del data lake y aplicar políticas adecuadas, seguridad, calidad de datos y privacidad en los datos almacenados en el lago.

**El acceso de autoservicio** consiste en tres conjuntos de servicios que proporcionan acceso on demand al data lake. El acceso de autoservicio de los usuarios analíticos permite acceder a los datos en bruto tal como se han almacenado. Para los equipos de LOB, el servicio proporciona datos normalizados en estructuras de datos simplificadas. Para los equipos de gobierno, riesgo y conformidad, el servicio proporciona datos gobernados a efectos de auditorías.

# Ingestión de datos de varias fuentes

La ingestión es el proceso de extracción, transformación, procesamiento de calidad e intercambio de datos entre el data lake, los sistemas de TI empresariales y otros data lakes existentes. Muchos de los datos del data lake proceden de los sistemas de TI de la organización. Estos tipos de datos pueden ser estructurados, semiestructurados o no estructurados. Las fuentes de datos pueden ser sistemas que operan en la empresa, un registro de sitio web u otras fuentes que monitorizan actividades. IBM ofrece la escalabilidad en el volumen de datos y la riqueza de transformación y réplica.






**Cuando se hace y se hace bien**

- Los datos fluyen en el data lake sin interrupción
- Se analizan datos transformados, estandarizados y enriquecidos
- Disminuyen los costes de almacenamiento incluso si aumentan los volúmenes de datos
- Uso de un recinto de pruebas para análisis exploratorio

**Cuando no se hace o se hace incorrectamente**

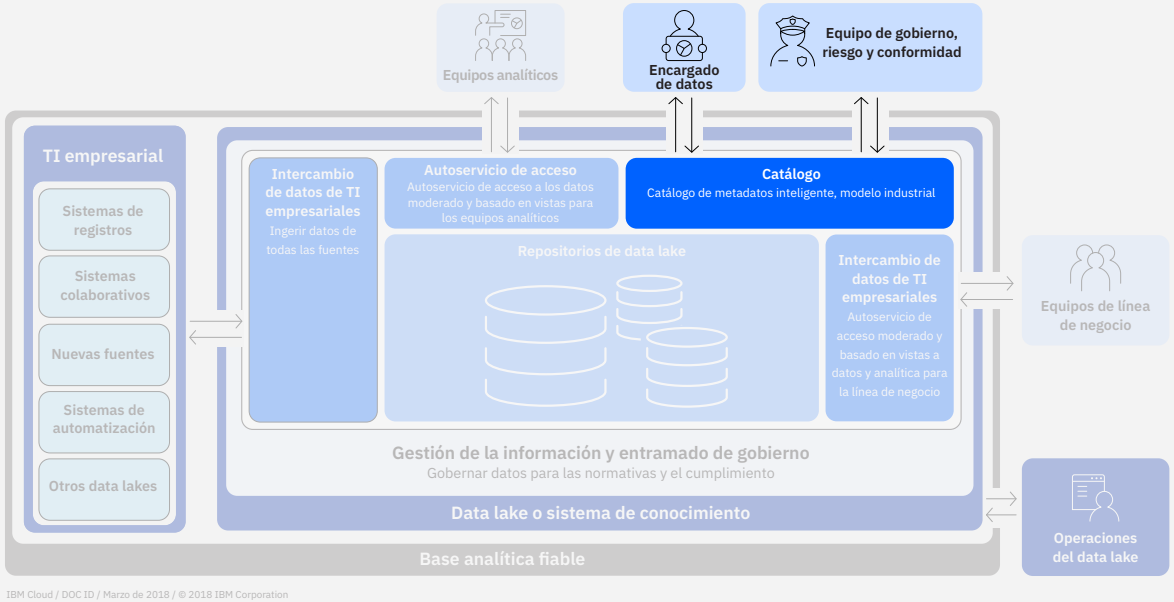
- Dificultades para mantener la frescura de los datos cuando aumenta el volumen
- Incapacidad para utilizar activos de información no estructurada
- Costes más altos del almacenamiento
- Limpieza complicada de los datos, lo que se traduce en costes más altos del proceso de datos

# Catalogación

La catalogación ayuda a etiquetar los datos del data lake y crear un inventario de activos de información. Las interfaces de catalogación proporcionan a los usuarios del data lake información acerca de los datos en su clasificación, procedencia y modo de gobierno. Las ofertas de IBM para la catalogación de data lakes poseen las siguientes características:

- Capacidad para capturar activos de información no estructurada en el catálogo
- Integración de ecosistema abierto con prácticamente cualquier activo de información
  - Un catálogo de empresa para prácticamente todos los activos de información de la organización
  - Habilitadores de términos comerciales y datos específicos del sector
  - Capacidad de puntuación y etiquetación social como parte de los metadatos

Los datos llevados al conducto de gobierno deben ser conocidos, para que los datos técnicos tengan sentido desde el punto de vista comercial. Por ejemplo, un número de 9 cifras puede ser un número de la seguridad social o un número de ID de empleado, o ambos. El paso de clasificación y asignación de término comercial añade significado comercial a los datos técnicos. La automatización es el principal atributo para que este proceso pueda ampliarse hasta satisfacer el volumen y variedad de datos del lago. A continuación, los flujos de trabajo de conservación, la evaluación de la calidad y los controles de datos garantizan que se puedan trasladar los datos a su catalogación, lo que permite ponerlos a disposición de toda la empresa.




**Cuando se hace y se hace bien**

- Mejora el plazo de obtención de resultados y se tiene más tiempo para analizar los datos
- Captura conocimiento contextual del activo y aumenta la utilidad de los datos
- Seguimiento de la procedencia de los datos y mayor confianza en los datos
- Activos de información de mercado para un consumo más amplio
- Asistencia en el cumplimiento normativo de los datos

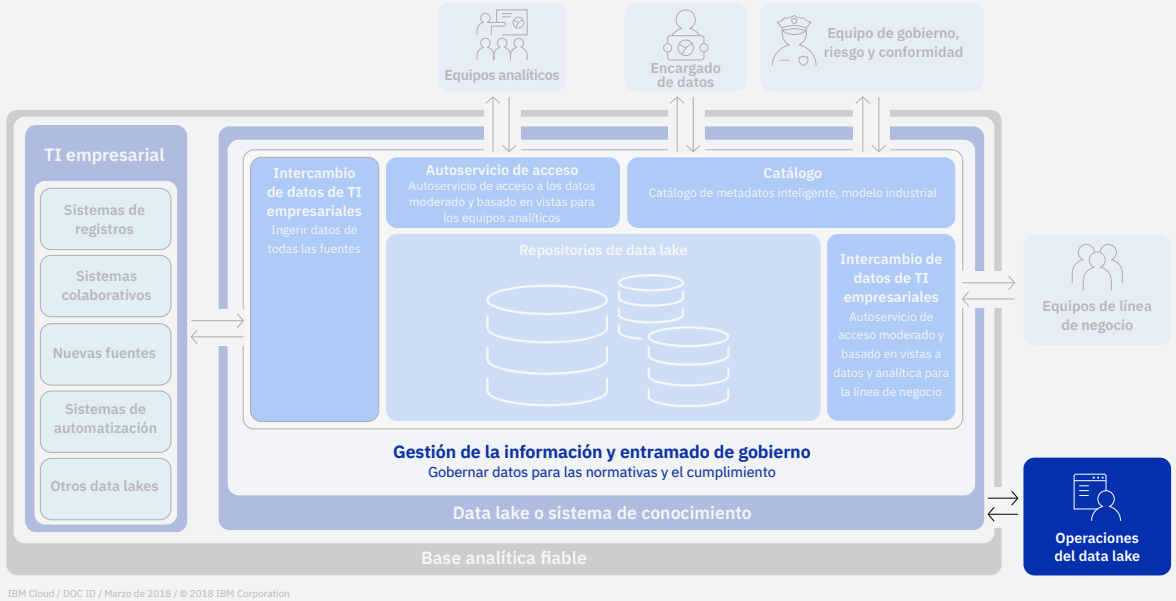
**Cuando no se hace o se hace incorrectamente**

- Riesgo de perder tiempo buscando y etiquetando datos
- Pérdida de conocimiento tribal al localizar los datos, pero no poder encontrar compañeros que comprendan los datos
- Desconocimiento de quién ha tenido acceso a los datos
- Incumplimiento de requisitos de cumplimiento y gobierno

# Gobernar y gestionar los datos

La integración de la información y el entramado de gobierno permiten al sistema realizar un seguimiento eficaz del data lake, lo que permite comprender la información entrante y aplicar automáticamente políticas de gobierno. La infraestructura de gobierno contribuye a documentar las políticas de gobierno y promulgar reglas que contribuyan a definir cómo debe estructurarse, almacenarse, transformarse y trasladarse la información.

Los requisitos del gobierno de la información se documentan en el catálogo en forma de políticas, reglas y clasificaciones. Los principales diferenciadores de IBM son que los activos no estructurados pueden formar parte del data lake y que se mantienen el volumen, variedad y velocidad de los niveles de datos.






**Cuando se hace y se hace bien**

- Seguir el ritmo del volumen de nuevos datos sin dejar de gobernarlo
- Cumplir con los requisitos normativos mediante herramientas de cumplimiento específicas del sector
- Acelerar la adopción de datos maestros
- Aumentar la exactitud de los conocimientos con datos de alta calidad
- Responder rápidamente a las auditorías de cumplimiento
- Aumentar la capacidad para proteger los datos

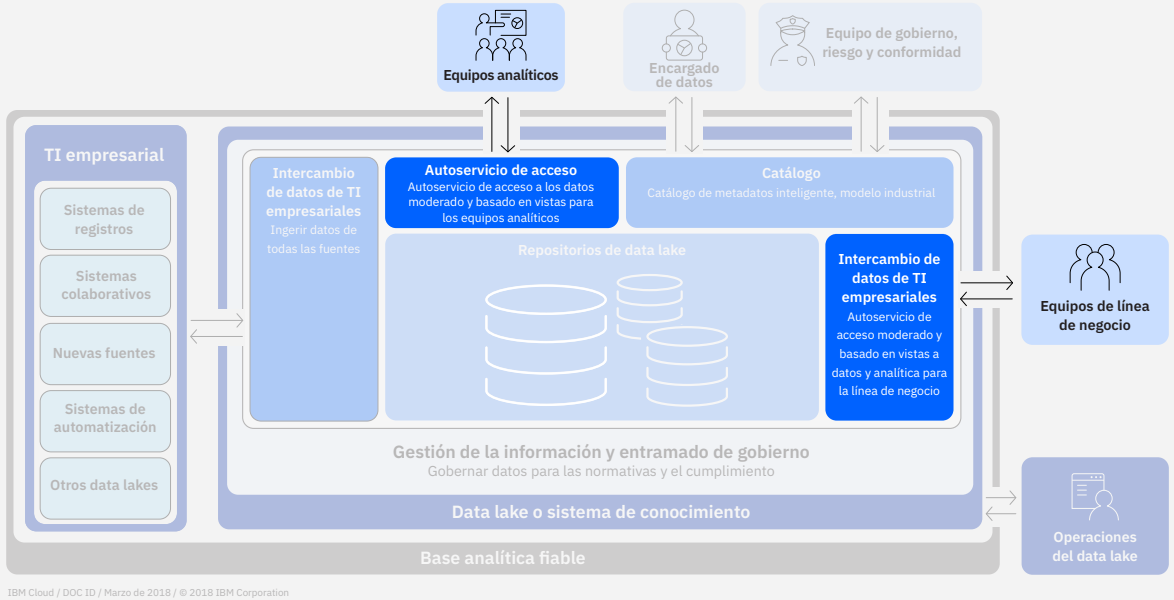
**Cuando no se hace o se hace incorrectamente**

- Incapacidad para gestionar el aumento del volumen de datos de fuentes estructuradas y no estructuradas
- Pérdida de tiempo buscando datos, lo que puede afectar a la preparación para las auditorías
- Pérdida de oportunidades para cumplir los requisitos de cumplimiento y gobierno

# Autoservicio o informes

El acceso de autoservicio contribuye a encontrar información relevante en los datos, mediante interfaces de búsqueda sencillas. Proporciona datos fiables y de alta calidad a compiladores autosuficientes que pueden utilizar los datos para construir modelos analíticos en sus iniciativas de ciencia de datos. También permite a los usuarios no técnicos transformar los datos antes de construir y desplegar modelos.

El acceso directo a los datos ayuda a los compiladores de TI en sus acciones de preparación y transformación de datos. Este acceso ayuda a los equipos de gobierno y cumplimiento normativo a conservar los datos con el fin de prepararlos para las auditorías. También ayuda a los consumidores de soluciones a crear informes personalizados para sus requisitos de negocio y tener acceso a datos preparados para fines comerciales, para que puedan tomar decisiones rápidas y generar conocimientos comerciales significativos a partir de los datos.






**Cuando se hace y se hace bien**

- Facultar a los usuarios de datos para que puedan acceder a datos contextuales
- Ayudar a los consumidores de datos a confiar en los datos mediante conocimiento tribal, etiquetaje social y puntuación cualitativa de los activos de información
- Ver cómo los datos se convierten en un activo de la organización accesible a todos los consumidores de datos
- Lograr una amortización más rápida
- Acelerar la innovación
- Facilitar una exploración y analítica de datos ágiles e interactivas

**Cuando no se hace o se hace incorrectamente**

- Dedicar más tiempo a encontrar y preparar datos que a analizarlos
- Incapacidad para encontrar o acceder a activos no estructurados
- Decisiones más lentas debido a la falta de acceso a datos fiables
- Innovación obstaculizada por la experiencia

# Por qué IBM

Según un estudio llevado a cabo por Radiant Advisors, el 72 % de líderes identificaron el gobierno y la seguridad como los principales retos, aunque máximos factores de éxito para sus organizaciones. El primer paso consiste en reconocer el gobierno y la arquitectura de la información como prioritarios. Esto abrirá la conversación en la organización para definir claramente lo que los usuarios de datos necesitan obtener de los datos. En un mundo en el que la entrada de datos erróneos equivale a la salida de datos erróneos, cada uno de los usuarios de datos forma parte de la conversación.

El despliegue de una plataforma única a nivel de toda la empresa para la integración de datos, el proceso de calidad de los datos y el gobierno de datos es esencial para tener éxito en las iniciativas de analítica. De este modo se podrá ingerir datos, asegurarse de que son de alta calidad y gobernarlos para inyectarlos en los procesos analíticos. Afrontar los retos con un método gobernado en el data lake permite crear una base desde la que suministrar datos fiables que servirán para muchos usos.

Ningún otro proveedor puede igualar la amplitud y profundidad de la plataforma IBM Unified Governance and Integration. Tanto si es la escalabilidad para gestionar grandes volúmenes de datos, aceleradores específicos del sector, capacidades para hacer que los datos estructurados, semiestructurados y no estructurados sean utilizables, como liderar conexperiencia en aprendizaje automático e inteligencia artificial, IBM proporciona una completa solución para construir un data lake gobernado y fiable.

Si desea obtener más información, visite **ibm.com**/governed-data-lake.

**IBM España, S.A.**
Tel.: +34-91-397-6611
Santa Hortensia, 26-28
28002 Madrid
Spain

La página de inicio de IBM se encuentra en:
**ibm.com**



03018403-ESES-01