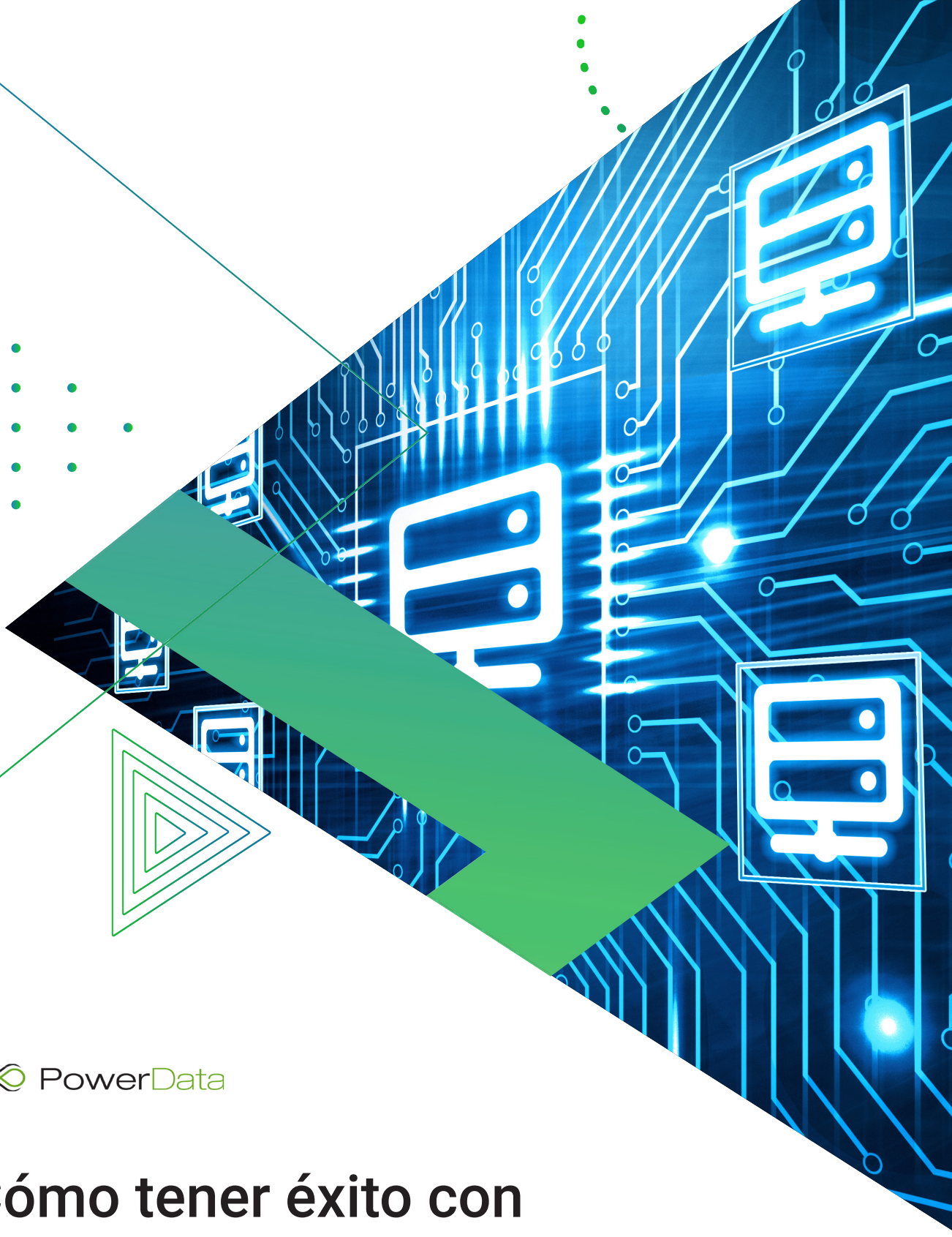


PowerData

Cómo tener éxito con un data lake en la nube

10 claves para superar los retos, obtener una ventaja competitiva y mejorar la experiencia de clientes.

Índice

Desde hace mucho tiempo las organizaciones se preocupan por capturar datos estructurados, semiestructurados y no estructurados; la novedad es que ahora lo hacen para impulsar iniciativas de inteligencia artificial y análisis avanzado para lograr un mayor impacto en el negocio, conocer con mayor profundidad la situación de la organización, y tomar mejores decisiones con mayor celeridad y precisión. Aquí entran en juego los data lakes, veamos cómo implementarlos con éxito.

El mercado denominado data lake estará creciendo a razón de un 20% anual entre 2020 y 2027, de la mano de un proceso incontenible de generación de información al ritmo de la transformación digital.

Fuente: Grand View Research

Esos grandes volúmenes de datos ponen en jaque a las necesidades de almacenamiento y ahí aparece el paradigma de los lagos de datos, que remiten a grandes repositorios donde la información puede ser almacenada tanto si tiene estructura como si no la tiene (unstructured data). En cualquier caso, son datos que pueden ser utilizados por analistas o expertos en ciencia de datos especialmente, pero también por el conjunto de la organización y su ecosistema.

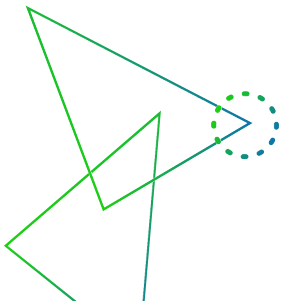
Lago de datos en la nube

La idea del lago se refiere a superar el aislamiento de islas (o silos) de información; también, a la posibilidad de moverse con flexibilidad y velocidad, a reunir información desde diferentes fuentes (o afluentes en términos fluviales) y, al mismo tiempo, al aprovechamiento del valor que se puede extraer de los datos con la mayor celeridad posible. En tanto evolución de paradigmas preexistentes, data lakes tiene un costo de implementación menor al conocido data warehouse.

Antiguamente, aunque hablemos de sólo unos años atrás, ese almacenamiento masivo se realizaba en servidores on premise alojados en el propio edificio corporativo o data center. Actualmente, las plataformas de nube pública ofrecen muchas alternativas para la construcción de data lakes. Por supuesto, siempre está la opción de rentar servidores privados virtuales (VPS) en cualquier proveedor de hosting para armar una nube privada que pueda contener también las funciones de almacenamiento de información.

No obstante, para organizaciones de mayor porte están disponibles las soluciones de las grandes plataformas de nube pública como Google, Amazon, Microsoft y otras.

Una de las claves aquí es que en estos lagos de información, pueden contenerse datos en su formato nativo, sin necesidad de forzar su amoldamiento a estructuras predeterminadas.



60%

De las organizaciones han obtenido una ventaja competitiva gracias a las iniciativas de data lake y casi la mitad ha logrado mejorar la experiencia de sus clientes.

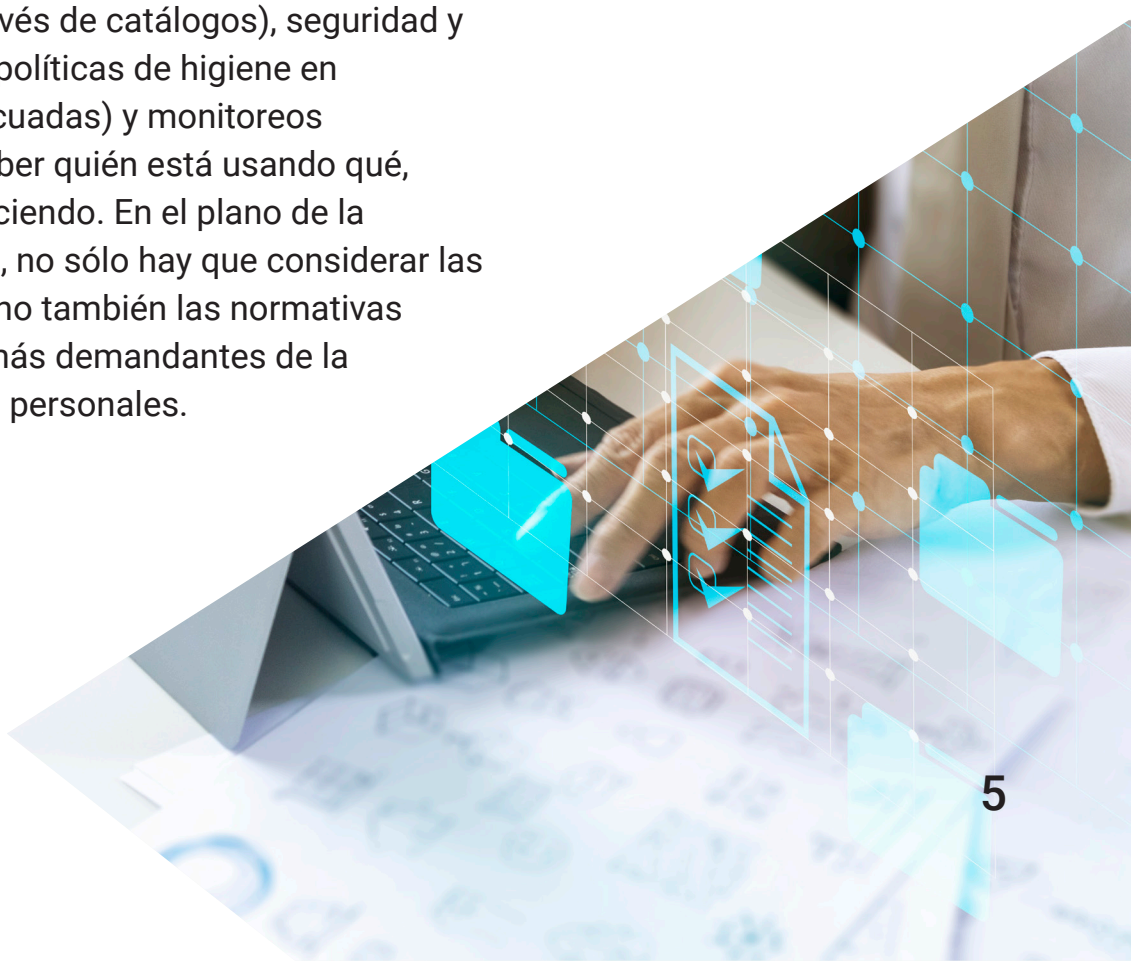
Fuente: Ventana Research

Sortear desafíos al construir data lakes

Para lograr una exitosa implementación de esta estrategia de gestión de datos, se pueden tener en cuenta las siguientes consideraciones:

01.

Contar con la plataforma para almacenar prácticamente toda la información que se genera en el proceso de negocios, en cualquier formato, puede generar un cierto caos. El **gobierno de datos** (data governance) es clave en términos de ordenamiento (a través de catálogos), seguridad y privacidad (con las políticas de higiene en ciberseguridad adecuadas) y monitoreos inteligentes para saber quién está usando qué, y cuándo lo está haciendo. En el plano de la protección de datos, no sólo hay que considerar las políticas internas sino también las normativas públicas, cada día más demandantes de la protección de datos personales.



02.

Un adecuado mecanismo de **agregación de datos** permite mantener la agilidad aún cuando estén sumando información las aplicaciones web y móviles, los sistemas transaccionales, las redes sociales y el correo electrónico, como así también un creciente enjambre de dispositivos de diverso tipo conectados bajo el paradigma IOT (Internet of Things). Esto requiere de planificación, una arquitectura acorde y la visión integradora desde la ciencia de datos.

03.

Aunque parece una realidad del pasado, aún hoy muchos sistemas en el Cloud almacenan información en silos o bases de datos desconectadas entre sí, lo que obliga al personal a malgastar tiempo en encontrar lo necesario. La perspectiva data lake permite colocar todo en un entorno centralizado, a lo que se suman herramientas de **visión integrada** que pueden nutrirse de diferentes fuentes para un mejor acceso a la información.

04.

Prestar atención a la **calidad de los datos** es otro de los aspectos clave para garantizar el aprovechamiento del valor que reporta la disponibilidad de datos, como así también para la toma de las mejores decisiones. Existen soluciones para monitorear y reconocer cuándo los datos están dañados, no tienen la precisión requerida o están desactualizados. En este punto es necesario establecer políticas y reglas de negocio que, incluso, puedan potenciarse con machine learning e inteligencia artificial para mantener limpio el lago de información.

05.

Una vez que se establecen los protocolos para el flujo de datos, es necesario manejar técnicas de **integración** que puedan prevenir las consecuencias de cualquier alteración o desvío, fruto de cambios inesperados o imprevistos en el ingreso de información (data ingestion) o en su formato. Aquí también es necesaria la inteligencia artificial para realizar un análisis y adaptación de lo necesario on the road. Automatizar todo lo posible esta tarea es fundamental.

06.

Nunca es suficiente lo que se haga en términos de **protección de datos.** Tanto los errores humanos como los ciberataques ponen continuamente en jaque la consistencia y privacidad de la información. Las políticas de Confianza Cero (Zero trust), backups continuos, mecanismos de data lost protection y recuperación ante desastres son algunas de las estrategias que se deben integrar en los data lakes, sin olvidar la permanente inversión en capacitación al personal.

07.

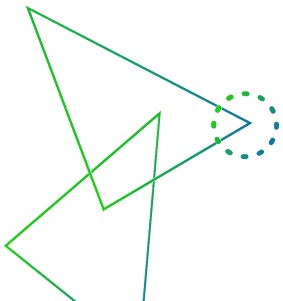
La cuestión del **costo** no es un asunto menor cuando se trata de manejar muchos procesos o grandes volúmenes de información. Combinar estrategias muticloud con modelos de pago por uso o consumo de recursos suele ser el antídoto para los potenciales sobrecostos que genere el ingreso al mundo del data lake. La llamada gestión autónoma de datos (AutoDM) utiliza metadatos, automatización e IA para estandarizar procesos y acelerar la entrega de datos al ritmo del negocio, manteniendo acotados los costos.

08.

DataOps y **DataSecOps** son estrategias clave para resolver la complejidad que pueda surgir del manejo de tanta y tan variada información en un lago de datos. La falta de personal específicamente entrenado en estas cuestiones de la ciencia de datos lleva a implementar este tipo de soluciones para mantener la agilidad operativa.

09.

Trabajar con **metadatos** es otra forma de sortear el desafío de moverse en un lago repleto de información diversa y dispersa. Una organización data first necesita contar con herramientas que gestionen los metadatos desde una perspectiva de gobernanza global.



10.

Las soluciones de **Machine Learning Ops (MLOps)** son la clave para poner en marcha soluciones de inteligencia artificial y automatización que materializan las ventajas del data lake.

Con la utilización de data lakes, las organizaciones logran reducir sus costos de TI hasta en un 70% para lo que es almacenamiento y gestión de datos.

Fuente: Aberdeen Strategy & Research

Cómo construir una solución de datos en la nube

Para armar un data lake en la nube que sea flexible y brinde la mayor productividad, es necesario contar con una solución integral de gestión de datos empresariales que integre todos los componentes necesarios, tales como:

- detección y comprensión de datos, compilación en catálogos;
- acceso e integración;
- conexión y automatización mediante APIs y aplicaciones;
- limpieza y confianza que garanticen la calidad de los datos;
- gestión y relación mediante MDM (master data management) y aplicaciones 360;
- gobierno de datos y protección;
- uso compartido a través de un marketplace de datos.

Todo esto conforma lo que se conoce como una plataforma inteligente de gestión de datos en la nube, utilizable por desarrolladores de ETL, ingenieros de datos, integradores, especialistas en datos, analistas y usuarios corporativos enfocados en el negocio propiamente dicho.

Algunas funcionalidades que resultan útiles en el proceso de armado de estos lagos de información incluyen la incorporación y replicación de datos por lotes, la transmisión en flujo y la captura de cambios en tiempo real, incluso desde grandes fuentes de datos como data warehouses o mainframes. En esa circulación de datos sin costura son útiles las técnicas conocidas como de extracción, carga y transformación (ELT) o bien extracción, transformación y carga (ETL). La correcta integración es una premisa fundamental de los data lakes.

En resumen, en una era en la que la transformación digital obliga a manejar grandes volúmenes de datos que pueden ser el sustento de las ventajas competitivas que contribuyan de manera crítica al negocio, son múltiples y variadas las estrategias, soluciones y herramientas que se deben articular. En ese sentido, el apoyo profesional experto resulta un complemento esencial para lograr una ecuación costo-beneficio acorde a la realidad de cada organización.

PowerData

PowerData, es una compañía multinacional de origen español con gran presencia regional, está enfocada en todo lo relacionado con la Gestión y Gobierno de Datos, tiene una trayectoria de más de 20 años impulsado una cultura Data-Driven en las empresas de la mano de sus aliados tecnológicos.

Te invitamos a explorar los proyectos donde aportamos valor con la gestión de datos. powerdata.es