The Modern Cloud Data Platform: Rise of the Lakehouse

Las organizaciones líderes entienden la importancia de hacer que los datos de alta calidad sean accesibles, utilizables y confiables. Una encuesta de McKinsey de 2019 encontró que las empresas con mayor crecimiento en ganancias en los tres años anteriores atribuyeron al menos el 20% de ese crecimiento directamente a sus iniciativas de datos.

¿Cómo lo lograron? Estas empresas de alto rendimiento implementan una estrategia de tres frentes, según McKinsey. En primer lugar, articulan estrategias de datos claras y a largo plazo. En segundo lugar, fomentan una cultura basada en los datos haciendo que los datos formen parte integrante de los puestos de trabajo de los empleados y educándolos en la gobernanza adecuada de los datos. Y tercero, implementan plataformas de datos modernas para soportar todas sus actividades de datos a escala.

Pero ¿qué es una "plataforma de datos moderna"? ¿Es un almacén de datos? ¿Un lago de datos? ¿Puede todo o parte de él estar en las instalaciones, o debe involucrar a la nube (o incluso múltiples nubes)? ¿Cuáles son los beneficios y desafíos de estos diversos enfoques? Y si hubiera una arquitectura de plataforma de datos ideal, ¿cómo se vería?

En 2020, O'Reilly Media, en colaboración con Databricks, per formó una encuesta global de más de tres mil profesionales de datos para determinar el estado de las arquitecturas modernas de plataformas de datos en la nube. Se pidió a los encuestados que evaluaran sus arquitecturas actuales de plataformas de datos, especialmente los desafíos que tenían con ellos, y cómo esos desafíos impactan en el éxito del negocio y del equipo. También se les pidió que recomendaran criterios que fueran importantes a tener en cuenta al evaluar nuevos tipos de arquitecturas de datos.

En este informe, primero hablaremos de las personas responsables de asegurar que las empresas están avanzando en sus viajes de datos, y cómo, como equipos en lugar de individuos, son fundamentales para transformar los procesos de datos en silos en procesos integrados que den a las empresas una comprensión más amplia de sus paisajes de datos. Luego hablaremos de las diversas arquitecturas de datos que se utilizan hoy en día, y sus fortalezas y debilidades rela tivas. Finalmente, presentaremos la idea de una nueva arquitectura de datos unificada que agrupe las ventajas y mitigue las desventajas de los modelos de datos tradicionales, y concluiremos con consejos sobre cinco pasos que puede dar para tener éxito en sus iniciativas de datos hoy.

Data Teams and Their Challenges

A medida que los datos se vuelven más importantes para las empresas, la necesidad de saber qué datos tienen y cómo monetizar se ha convertido en fundamental para el crecimiento exitoso. Esto es cierto para todas las empresas, independientemente de si son organizaciones tecnológicas tradicionales o no, y esta necesidad sólo va a ser más importante en los próximos meses y años a medida que se generen aún más datos, y su uso se convierte efectivamente en una necesidad competitiva.

La atención se centra cada vez más en los profesionales de datos que trabajan juntos como un equipo para aprovechar al máximo los datos de su organización. ¿Por qué? Porque no es suficiente que los profesionales de datos trabajen de forma autónoma. La sinergia entre las diferentes funciones es demasiado importante. Tienen que unirse como comunidad.

Tradicionalmente, las funciones clave de los datos existían por separado en cuatro compartimentos de profesionales de datos: científicos de datos, ingenieros de datos, arquitectos de datos, y analistas de datos. Aunque todos son parte de la misma organización, estos roles de datos han dependido históricamente de completamente diferentes conjuntos de herramientas y procesos, y cada uno trabajó con los datos en su propio silo. Este patrón conduce a una mayor complejidad organizativa y costo. También toma más tiempo y es más difícil obtener valor de datos cuando los profesionales de datos operan por separado.

Hoy en día, las principales empresas basadas en datos se están arquitectura que satisfaga las necesidades de todos estos profesionales de datos.

The Importance of the Cloud to the Data-Driven Company

Resulta que el mundo de los datos está tan enamorado de la nube como el resto del universo. Un 81% de los encuestados dijo que sus organizaciones habían adoptado servicios en la nube e infraestructura como arquitecturas de datos al menos en cierta medida. Solo 2 de cada 10 (19%) informaron no haber movido ninguna carga de trabajo de datos a la nube.

Entonces, ¿qué tipos de arquitecturas de plataforma de datos están utilizando las empresas de forma rentable, ya sea en la nube, en las instalaciones, o ambos? Se animó a los entes de respuesta a comprobar todo lo que se aplicaba. Como muestra la Figura 1, los almacenes de datos están ligeramente por delante (57%), seguidos de cerca por los lagos de datos (53%) y los llamados "sistemas especializados" (54%), que abarcan bases de datos especializadas como SAP ERP y Oracle Peo pleSoft en las instalaciones, y Salesforce y Workday en la nube.

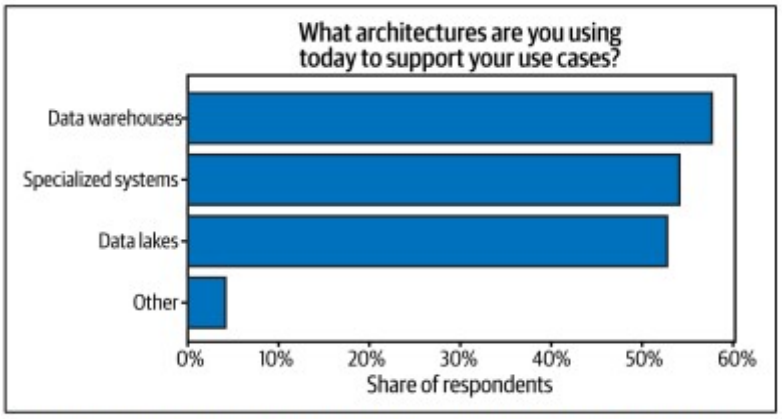

Figure 1. A close race between today's most popular data platform architectures

Figura 1. Una carrera cerrada entre las arquitecturas de plataforma de datos más populares de la actualidad

Uno de los hechos más importantes revelados por las respuestas a esta pregunta, por supuesto, fue que numerosas organizaciones están ejecutando múltiples arquitecturas de plataformas de datos. Como vemos en la Figura 2, los desafíos planteados por esto son frecuentes.

El principal de los desafíos de ejecutar múltiples arquitecturas de datos es la complejidad operativa, simplemente manteniendo la infraestructura estable y los repositorios de datos funcionando en tal entorno fue un problema para más del 70% de las empresas.

La calidad de los datos (planteada como problemas de compartimentos estancos, duplicación e incoherencia) y la gobernanza de los datos (incluidos la seguridad y la seguridad de los datos) también fueron identificados como retos clave por el 67% y el 66% de los encuestados, respectivamente. La calidad de los datos es fundamental porque las decisiones empresariales se

tomarán sobre la base de los datos almacenados en un depósito de datos. Las empresas necesitan tener absoluta confianza en que están trabajando desde un solo punto de la verdad. La gobernanza también es cada vez más importante, especialmente en lo que respecta a la privacidad de los datos, debido al creciente número de reglamentos que protegen los datos confidenciales del acceso de personas no autorizadas.

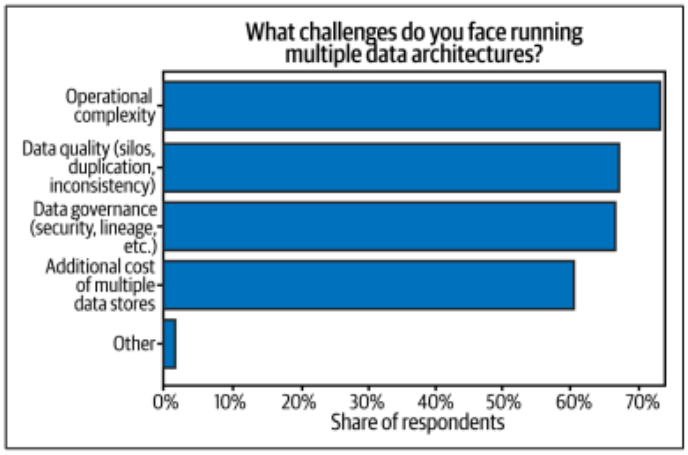

Figura 2. Desafíos de ejecutar múltiples arquitecturas de datos

Por último, un robusto 60% de los encuestados identificó como problema los costes adicionales de tener que soportar múltiples almacenes de datos.

Como resultado directo de estos desafíos, un 63% de las empresas están implementando o evaluando activamente nuevas arquitecturas de datos, como se muestra en la Figura 3. Solo 2 de cada 10 (19%) profesaron que estaban contentos con lo que tenían, lo que podría significar que estaban considerablemente por delante de la manada en la obtención de una mejor forma de plataforma de datos, o que de alguna manera están haciendo frente al nivel de dolor involucrado en el apoyo de múltiples arquitecturas.

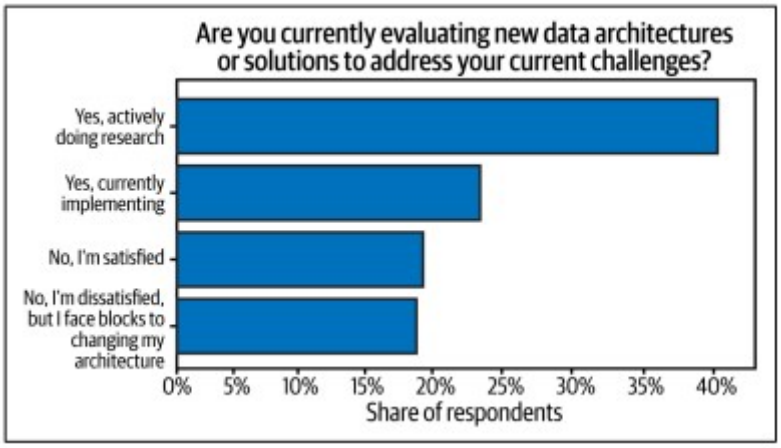

Figura 3. Muchos profesionales de datos están buscando nuevas opciones de arquitectura de forma de plataforma de datos

Para entender con mayor precisión qué empresas han instalado actualmente, y por qué podrían estar buscando soluciones más modernas, la siguiente sección definirá cada tipo de plataforma de datos y sus beneficios y desafíos.

Data Architecture Options

Hasta hace poco, las empresas han tenido tres opciones básicas de arquitecturas de plataformas de datos: almacenes de datos, lagos de datos y sistemas de datos especializados. Los tres se pueden desplegar ya sea en las instalaciones o en la nube. Vamos a examinar cada uno a su vez.

Data Warehouses

Un depósito de datos es un depósito central en el que los datos de una o más fuentes diferentes se integran y se utilizan para la presentación de informes y análisis empresariales. Se considera fundamental para la capacidad de una empresa de aprovechar sus datos para la inteligencia empresarial (BI).

Un atributo clave de un almacén de datos es que está altamente estructurado. Los datos almacenados en un almacén de datos se han preparado y "trans formado"-limpiado y deduplicado y formateado para cumplir con las normas establecidas. De hecho, los datos normalmente no se ponen en un almacén de datos hasta que los profesionales de datos están bastante seguros de que saben cómo se utilizará, y para qué fines. La mayoría de los almacenes de datos, ya sea en locales o en la nube, siguen las directrices y marcos definidos por Ralph Kimball y Bill Inmon a mediados de la década de 1980.

El almacenamiento de datos cambió fundamentalmente la forma en que las empresas obtenían datos y tomaban decisiones estratégicas. Antes de su aparición, los datos transaccionales y operativos estaban encerrados en diferentes silos, lo que dificultaba garantizar que hubiera coherencia en toda la empresa en la forma en que se definían los datos, para poner los datos directamente en las manos de los usuarios de negocios que lo necesitaban para hacer sus trabajos, y para obtener una imagen completa del negocio de una organización. Hoy en día, las casas de datos son muy populares, de hecho, como muestra la Figura 2, son actualmente la plataforma de datos líder.

Aunque los almacenes de datos tradicionales se encuentran en las instalaciones, los almacenes de datos basados en la nube están ganando rápidamente el favor, por razones de costo y escalabilidad, así como el hecho de que liberan a las organizaciones de tener que adquirir, desplegar y mantener la infraestructura necesaria para sostener los almacenes (más adelante).

Benefits of data warehouses

La principal directiva del almacén de datos es ayudar a la organización a tomar mejores decisiones comerciales. Al ayudar a los profesionales de los datos y a los consumidores de datos a alcanzar este objetivo, se obtienen otros beneficios. Almacenamiento de datos:

Proporcione inteligencia de negocios
: Poner datos de diferentes fuentes en un solo almacén de datos y hacerlo accesible para usuarios autorizados dentro de una organización significa que las empresas ya no tienen que depender de los instintos de los empleados y ejecutivos para hacer decisiones críticas. En su lugar, estas decisiones pueden apoyarse en datos.

Mejore el rendimiento de las consultas
: Las consultas constantes de los usuarios empresariales pueden llevar la infraestructura de análisis, como las martas de datos y las bases de datos heredadas hasta los límites. Un almacén de datos puede ser más eficiente en la gestión de consultas, aliviando la carga sobre el ecosistema en general.

Mejorar la calidad de los datos y la decisión
: Los datos se transforman antes de colocarlos en el almacén de datos. Esto significa que los datos de múltiples fuentes se ponen en un formato estandarizado, y los usuarios de toda la organización pueden ver y acceder a información consistente que les permite dirigir el negocio en una dirección común y consistente.

Democratizar los datos

> Más recientemente, gracias a los avances en las propias bases de datos, así como herramientas de análisis y visualización, Las organizaciones avanzadas basadas en datos están intentando democratizar completamente los datos en toda la organización permitiendo que más y más usuarios tengan acceso a sus almacenes de datos. Pero como verás, este beneficio está vinculado a uno de los mayores desafíos que los encuestados tuvieron con los almacenes de datos: la escalabilidad.

Challenges of data warehouses

Aunque los beneficios del uso de almacenes de datos son significativos, también existen desafíos. Un almacén de datos es, casi por definición, una base de datos supergrande. Diseñar y desplegar con éxito uno es una tarea enorme. Se requiere planificación, colaboración y coordinación tanto de personas como de recursos.

Cuando preguntamos a los profesionales de datos sobre los retos que enfrentaban con sus almacenes de datos, como muestra la Figura 4, el costo encabezó la lista (50%), seguido por operaciones complejas (47%) y escalabilidad (46%).

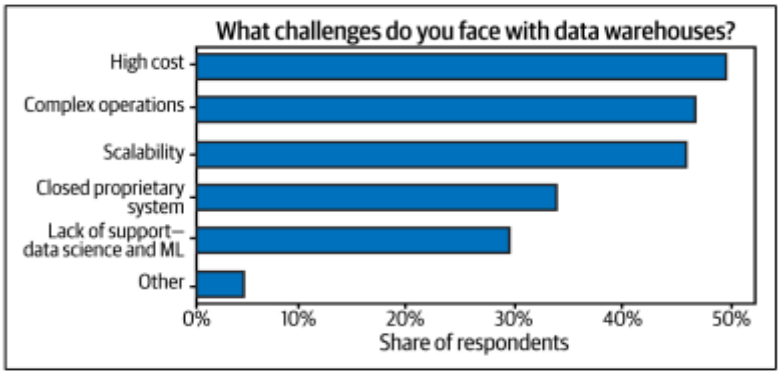

Figure 4. Challenges with data warehouses

Costo potencialmente alto. La mitad de los encuestados dijo que el alto costo de un almacén de datos era su mayor inconveniente. Esto puede ser un problema tanto con los almacenes de datos locales como en la nube.

En el caso de los almacenes de datos locales, los elevados derechos de licencia crean una base de costos costosa. Luego están los costos operacionales: toda la responsabilidad de adquirir, instalar y mantener la infraestructura de almacenamiento de datos recae en la organización. Y a medida que el almacén de datos crece -como inevitablemente lo hará- más personas y recursos tienen que estar comprometidos con la iniciativa. Para asegurar que siempre hay suficiente cómputo y almacenamiento, muchas organizaciones sobreprovisionan su infraestructura para permitir picos inesperados o picos en las cargas de trabajo; por ejemplo, los minoristas deben prepararse para este tipo de cosas durante la temporada de vacaciones. Esto puede significar que durante gran parte del año han infrautilizado la capacidad.

Por otro lado, la estructura de precios de los almacenes de datos en la nube, que es más comúnmente un modelo de alquiler, le permite pagar tanto o tan poco como necesite. Aunque esto elimina el riesgo de pagar por recursos no utilizados, todavía puede ser un esfuerzo costoso a medida que el almacén de datos crece.

Operaciones complejas. Casi la mitad (48%) de los encuestados dijeron que las operaciones de los almacenes de datos son demasiado complejas.

En el caso de los almacenes de datos locales, la tecnología de la información y, con frecuencia, el equipo de datos participan activamente en el despliegue, las mejoras, la implantación de la seguridad y las operaciones en curso. Esto no es trivial. Las plataformas de datos necesitan ser sintonizadas regularmente para garantizar un rendimiento constantemente alto a lo largo del tiempo, especialmente a medida que los volúmenes de datos escalan. De lo contrario, el almacén de datos puede volverse dolorosamente lento, ineficaz o incluso no funcional.

Las operaciones también pueden ser complejas con los almacenes de datos en la nube, tanto con respecto a los precios como al soporte de infraestructura. Al igual que otros servicios y soluciones en la nube, este mercado todavía está madurando. Diferentes ven dors utilizan diferentes modelos de costo. Algunos cobran una tarifa mensual fija, mientras que otros ofrecen esquemas de precios de pago por uso. Diferentes proveedores también abordan el soporte de infraestructura de manera diferente. Algunos productos de cloud data ware house requieren que usted proporcione y administre sus recursos de cloud, mientras que otros adoptan un enfoque sin servidores, en el que la carga de aprovisionamiento y administración de servidores cloud se abstrae completamente de la empresa. Luego están las tareas operacionales de hacer cumplir los acuerdos de nivel de servicio (SLA), integrar el almacén de datos con los procesos existentes, tanto basados en la nube como en los preparativos, y asegurar que las medidas de seguridad y recuperación ante desastres sean sólidas.

Por último, dado que muchas organizaciones primero depositan datos en un lago de datos, es necesario mantener numerosas canalizaciones de datos para trasladar los datos fuera del lago de datos a uno o más almacenes de datos. En los casos en que esos almacenes de datos hagan cambios, también se requieren tuberías de datos para mover los datos modificados de vuelta al lago de datos.

Escalabilidad. La escalabilidad es un problema muy real para los almacenes de datos locales, y el 47% de los encuestados dice que es una preocupación clave. El departamento de TI debe estar alerta para asegurarse de que hay suficientes recursos en todo momento, especialmente para hacer frente a los choques inesperados en el tráfico. La ampliación es una tarea que consume mucho tiempo y recursos, ya que normalmente implica la compra e instalación de nuevo hardware.

Para los almacenes de datos en la nube esto no es un problema, ya que las organizaciones pueden adquirir más cómputo o almacenamiento en cualquier momento que lo necesiten, incluso para el tráfico "bursty". Sin embargo, la escalabilidad sigue siendo un problema aquí, ya que es difícil mantener los cientos o miles de tuberías de datos necesarios para alimentar todos los diferentes informes que los grandes almacenes de datos deben servir. Esta preocupación se ve exacerbada por dos factores: la mayoría de los proveedores personalizados tienen múltiples proveedores de almacenamiento de datos en sus arquitecturas, y las arquitecturas de datos se dividen cada vez más entre múltiples proveedores de nube.

Sistemas propietarios cerrados. Un tercio (33%) de los profesionales de datos consideran que este problema es crítico. Desafortunadamente, muchos almacenes de datos locales no juegan bien con otros. El bloqueo es real, y puede ser una molestia costosa cuando se desea pasar a una solución de almacenamiento de datos diferente.

Incluso la nube no escapa a este desafío, ya que diferentes proveedores de nube tienen diferentes funciones y capacidades, y mover un almacén de datos de, por ejemplo, Google Cloud a Microsoft Azure no es un proceso sin problemas.

Cuando se les preguntó acerca de sus actitudes hacia el bloqueo causado por formatos de datos cerrados o software propietario, una gran mayoría (86%) de los profesionales de datos expresaron estar al menos "un poco" preocupados, con 58% siendo "preocupado" o "muy preocupado", como

se muestra en la Figura 5. Bloqueo de proveedores-en, por supuesto, ha sido un reto para las organizaciones desde los albores de la era digital, ya que los fabricantes de hardware y software tratan de dificultar a los clientes a salir de sí mismos después de comprometerse con una plataforma en particular. Este desafío continúa molestando a los profesionales de datos en la era de los datos.

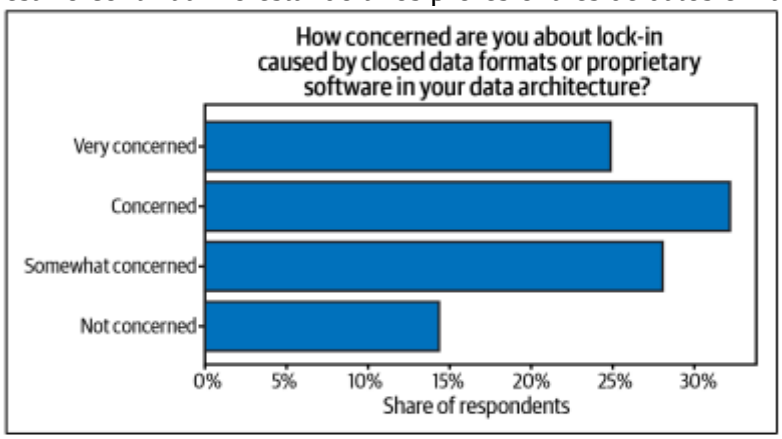

Figura 5. Abundan las preocupaciones sobre el bloqueo

Falta de apoyo para la ciencia de datos y el aprendizaje automático. El desafío aquí para el 29% de los profesionales de datos es que los almacenes de datos se basan en tecnología de 40 años que nunca fue diseñada para manejar nada más que datos estructurados. Audio, video, texto en lenguaje natural y otros tipos de datos no estructurados no encajan en los esquemas del almacén de datos. La creciente prevalencia de este tipo de datos como combustible para la ciencia de datos y el aprendizaje automático es lo que impulsó el aumento de los lagos de datos y la consiguiente complejidad de tratar de mantener tanto los lagos de datos como los almacenes de datos en una arquitectura de datos empresarial.

Lagos de datos

Una diferencia crítica entre lagos de datos y almacenes de datos es el tipo de datos y casos de uso que pueden manejar. Sin embargo, muchos piensan que los lagos de datos y los almacenes de datos son la misma cosa. Estos son los atributos que comparten:
• Son repositorios para almacenar datos.
• Pueden estar basados en la nube o en locales.
• Se despliegan cuando las organizaciones buscan democratizar los datos.

Pero ahí es donde las similitudes entre lagos de datos y almacenes de datos generalmente terminan. En la tabla 1 se resumen las diferencias.

Table 1. Main differences between data lakes and data warehouses

	Data lakes	Data warehouses
Types of data that can be stored and processed	Structured, semi-structured, and unstructured data	Structured and semi-structured data only
Purpose of data	Undefined purpose for data	Data defined for specific use cases
Types of users	Users are data scientists and data engineers	Users are nontechnical business users
Structure	Flexible and easy to change	Rigid and difficult to change

Entonces, ¿qué es un lago de datos? Es un sistema o repositorio de datos almacenados en su formato natural/ crudo, por lo general archivos u objetos blobs. Puede contener datos estructurados y no estructurados en su forma bruta, incluyendo datos estructurados de bases de datos relacionales o transaccionales (filas y col umns), datos semiestructurados (CSV, registros, XML, JSON), datos no estructurados (correos electrónicos, documentos, PDFs), y datos binarios (imágenes, audio, vídeo).

El objetivo de un lago de datos es poner todos estos datos a disposición de la transformación y la minería para generar informes, realizar visualizaciones y realizar análisis avanzados y aprendizaje automático con el fin de, en última instancia, obtener una ventaja comercial competitiva.

Beneficios de los lagos de datos

Los lagos de datos ofrecen algunos beneficios significativos sobre los almacenes de datos:

Los lagos de datos pueden ingerir y mantener todos los datos empresariales.

Gran parte del trabajo en la construcción de un almacén de datos gira en torno a la comprensión y la toma de decisiones sobre los datos. ¿De dónde viene? ¿Qué hay que hacer al respecto? ¿Cómo informará a los procesos institucionales? El objetivo es construir un modelo de datos altamente estructurado diseñado para la presentación de informes, con cierta capacidad para consultas ad hoc por usuarios más avanzados. Típicamente, si los datos no son necesarios para responder preguntas específicas o ser incluidos en un informe en particular, serán excluidos del almacén. Hay dos razones para esto: simplificar el modelo de datos y, específicamente para los almacenes de datos locales, para evitar llenar el costoso almacenamiento en disco. El almacenamiento es costoso porque en las casas de datos locales, se acopla con el cómputo. Aumentar el almacenamiento significa que usted tiene que comprar más cómputo, también, incluso si no lo necesita. Lo contrario también es cierto.

En contraste, un lago de datos retiene todos los datos de la empresa. Todos los datos generados o recogidos por la empresa pueden ser, y por lo general es, poner en el lago. ¿La razón? Es imposible predecir de antemano qué datos serán útiles para la ciencia de los datos exploratorios y el aprendizaje automático, o incluso para futuras necesidades de BI. Los lagos de datos le dan esa flexibilidad. Los datos también se conservan indefinidamente, para que las empresas puedan comprobar y volver a comprobar los datos antiguos según sea necesario.

Los lagos de datos pueden almacenar y procesar todos los tipos de datos.

Mientras que los almacenes de datos se ocupan de datos estructurados, como los datos de los sistemas transaccionales tradicionales, los lagos de datos pueden absorber datos estructurados y no estructurados. Esto incluye transmisión de datos como registros de servidores web, datos de sensores, actividad de redes sociales, texto e imágenes. Mientras que en el pasado este tipo de datos era difícil y costoso de almacenar y analizar, el lago de datos lo acepta todo.

Los lagos de datos hacen que todos los datos sean accesibles a todos los usuarios.

Hay tres tipos generales de usuarios de datos: usuarios empresariales, analistas de datos y científicos de datos. El almacén de datos está bien estructurado, es fácil de usar y está diseñado específicamente para responder a preguntas que el primer tipo de usuario -los usuarios empresariales- podría tener. Los analistas de datos son los constructores de los informes y paneles que los usuarios de negocios consumen. Por último, la tercera categoría de usuarios -científicos de datos, ingenieros de datos y otros profesionales de datos- generalmente eluden el almacén de datos por ser demasiado limitante. Están interesados en hacer análisis estadísticos profundos, a menudo utilizando herramientas de inteligencia artificial (IA). Los lagos de datos sirven a todos estos tipos de consumidores de datos por igual.

Los lagos de datos se pueden cambiar fácilmente.

Debido a que las organizaciones ponen tanto esfuerzo en el diseño de almacenes de datos y estructuras de datos para ser como ellos quieren, cambiarlos requiere asignar una gran cantidad de recursos de desarrollo -y tiempo- a la tarea. Por el contrario, el lago de datos almacena todos los datos en su forma bruta y lo hace accesible a cualquiera que lo necesite, para utilizarlo de la manera que desee. Y lo que es más importante, los lagos de datos tienen un marco de esquema sobre lectura (en lugar de la práctica de esquema sobre escritura de los almacenes de datos) y utilizan el proceso de extracción, carga de datos brutos y transformación según sea necesario (ELT) en lugar del proceso convencional de extracción, transformación y carga (ETL) proceso que siguen los almacenes de datos. Esto permite a los usuarios construir modelos y explorar los datos y esquemas que deseen. Cualquier resultado de experimentos de datos que no sean útiles puede ser simplemente desechado, sin cambios en las estructuras de datos y sin necesidad de involucrar a los desarrolladores de TI para obtener ayuda. Esto hace que los lagos de datos sean infinitamente más flexibles; no requieren cambios estructurales para responder a nuevas preguntas.

Los lagos de datos pueden proporcionar información procesable más rápidamente.

Precisamente porque los lagos de datos contienen todos los datos y tipos de datos, y porque permiten a los usuarios de todo tipo acceder a los datos antes de que hayan sido estructurados y transformados, los usuarios obtienen sus resultados más rápido que si tuvieran que esperar a que los profesionales de datos limpien y estandaricen los datos para ellos. Desafortunadamente, los lagos de datos también pueden convertirse en pantanos de datos, precisamente porque pueden convertirse en vertederos de datos que no se ajustan a ninguna norma (más sobre esto en la siguiente sección).

En resumen, los beneficios de los lagos de datos son sustanciales, especialmente cuando se trata de almacenes de datos más grandes y no estructurados para los que las empresas pueden tener numerosos fines finales aún no determinados.

Desafíos de los lagos de datos

Sin embargo, también hay muchos desafíos para la aplicación de los lagos de datos. No es de extrañar que la gobernanza sea la preocupación número uno, como se muestra en la Figura 6, con más del 60% de los encuestados diciendo que era un desafío. Y la falta de gobernanza resulta inevitablemente en datos desordenados y poco fiables, que fue el desafío número dos citado por más de la mitad (52%) de los profesionales de datos encuestados. Las operaciones complejas llegaron en un tercio cercano, con el 51% de los profesionales de los datos catalogando esto como una preocupación importante.

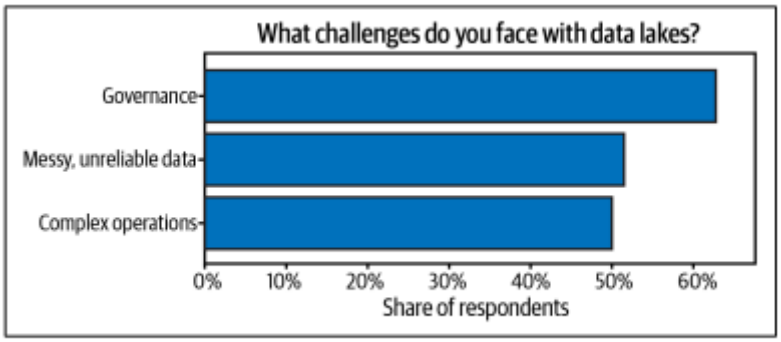

Figure 6. Challenges with data lakes

Vamos a profundizar un poco más en estos temas principales:

Gobernanza.

Precisamente debido al enorme volumen de datos que se encuentran en un lago de datos, con usuarios de todo tipo que se sumergen, preguntan, consumen o informan sobre los datos a voluntad, es un desafío considerable para garantizar que los datos siguen siendo seguros y privados. Cumplir con la creciente lista de reglas de privacidad -de la industria, los estados, el gobierno federal e incluso los organismos internacionales- es importante para evitar multas fuertes y vergüenza pública. Es esencial que un profesional de datos de alto nivel (tal vez incluso el director de datos) establezca las políticas adecuadas para los datos a lo largo de todo su ciclo de vida.Los datos deben permanecer seguros y privados en su estado bruto, y permanecer de esa manera incluso cuando los científicos de datos lo exploran, los analistas lo curan y los usuarios de negocios lo analizan.

Datos sucios y poco fiables.

A medida que los lagos de datos continúan acumulando más y más datos en diferentes estructuras y formatos, mantenerlos consistentes y limpios puede ser una tarea formidable. La arquitectura del lago de datos está más dispersa y tiene menos limitaciones en el formato y la escala de los datos almacenados que un almacén de datos. Lagos de datos también toman tiempo para reflejar escribe a sus clientes. Esto resulta en consultas que muestran datos inconsistentes hasta que todos los nodos del lago se vuelven consistentes. Además, los lagos de datos carecen de mecanismos para alertar a los usuarios cuando las escrituras fallan. Puede tomar semanas o meses descubrir que subconjuntos de datos están dañados o incompletos.

Operaciones complejas.

Los lagos de datos locales tienen los mismos problemas operacionales que los almacenes de datos internos. El rendimiento y la seguridad son lo más importante, y la TI debe mantener los lagos de datos y funcional, funcionando a un rendimiento óptimo en todo momento, para que sean exitosos. Si un lago de datos obtiene la reputación de ser lento o no responder, el equipo no lo utilizará. Una forma diferente de pensar en esto es simplemente invertir el problema de escalabilidad del almacén de datos: con los lagos de datos no necesariamente tiene un montón de tuberías para mantener, pero es necesario hacer una gran cantidad de ingeniería para escalar los tiempos de respuesta con un mayor uso.

Construir, hacer la transición o mantener un lago de datos en la nube también puede ser un reto operativo, especialmente cuando una organización tiene datos tanto en las instalaciones como en la nube para gestionar. Además, hoy en día las soluciones multi-cloud son cada vez más comunes, por tres razones principales. En primer lugar, las empresas a menudo necesitan diversificar su infraestructura para cumplir con la normativa o mitigar el riesgo. En segundo lugar, la toma de decisiones independiente en las grandes empresas con frecuencia da lugar a diferentes departamentos que invierten en diferentes proveedores de nube. Y tercero, la actividad de fusión y adquisición (M&A) obliga al adquirente a absorber las inversiones tecnológicas de la adquirida.

En resumen, el uso de lagos de datos que son diferentes de los que se encuentran en los almacenes de datos plantea importantes desafíos. A pesar de ello, las empresas se desplazan cada vez más a los lagos de datos debido a su flexibilidad y capacidad para facilitar el acceso a todos los datos.

Sistemas especializados

El tercer y último tipo de arquitectura de plataforma de datos cae en una categoría llamada "sistemas especializados." Estas son aplicaciones que tienden a ser grandes repositorios de datos para tipos específicos de datos. Por ejemplo, Salesforce es un gran repositorio de datos que muchas empresas utilizan para gestionar los datos de gestión de las relaciones con los clientes (CRM). El día de trabajo es otro ejemplo, que alberga datos que están particularmente relacionados con los recursos humanos.

El principal beneficio de los sistemas especializados es que los datos están estrechamente controlados y organizados de acuerdo con las especificaciones del proveedor de la aplicación. Por lo general, hay formas bien establecidas de consultar el sistema y obtener informes sobre temas comunes. Los desafíos surgen cuando desea integrar datos de uno de estos sistemas especializados, como combinar datos de facturación en un almacén de datos local con datos de clientes que probablemente se encuentren en la nube de Salesforce.

Las empresas que poseen arquitecturas de plataformas de datos de sistemas especiales son las únicas que ponen las operaciones complejas primero cuando se les pregunta sobre los desafíos (ver Figura 7).

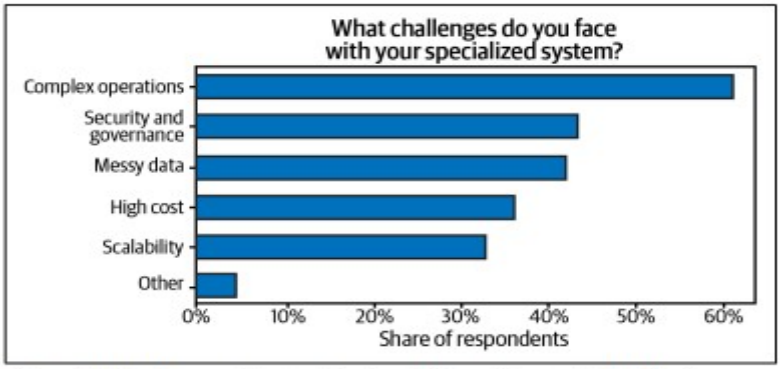

Figure 7. Challenges data professionals face with specialized systems

El principal desafío se reduce a una palabra: la integración. Aunque las API están disponibles para hacer integraciones comunes bastante sencillas, es complicado eliminar muchos de los silos que pueden surgir en estos sistemas especiales. El mapeo de datos, el mastering, la deduplicación y -quizás lo más importante- la migración de datos desde formatos propietarios pueden ser grandes dolores de cabeza.

El impacto empresarial de los datos.

Los datos marcan la diferencia. Casi tres cuartas partes (67%) de los profesionales de datos encuestados dijeron que la mejora en la toma de decisiones empresariales impulsa sus casos de uso de datos. Más de la mitad también indicó que las experiencias de los clientes (60%) y los procesos empresariales clave (56%) eran factores determinantes. De hecho, estos fueron los tres principales conductores independientemente del caso de uso (ver Figura 8).

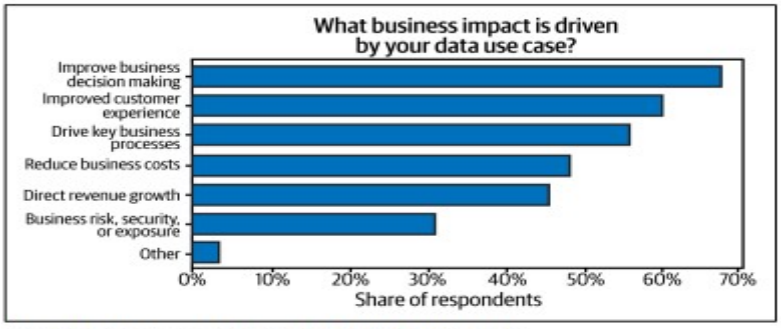

Figure 8. The business impact of data initiatives

Estos datos sugieren que la mayoría de las organizaciones favorecen una visión a largo plazo de sus inversiones en datos en lugar de buscar solo recompensas a corto plazo como reducciones en los costos de negocio (49%) o aumentos en el crecimiento de ingresos (45%). Esto refleja la opinión generalmente aceptada en el mercado de que la capacidad de hacer uso eficaz y eficiente de los datos va a determinar qué organizaciones tienen éxito en el futuro previsible.

Impacto de la gestión de arquitecturas de datos complejas.

A pesar de que la mayoría de las empresas buscan el valor comercial de sus actividades de datos, parece que la complejidad de las plataformas de datos, infraestructuras y arquitecturas impiden a los equipos de datos alcanzar plenamente sus objetivos (véase la Figura 9).

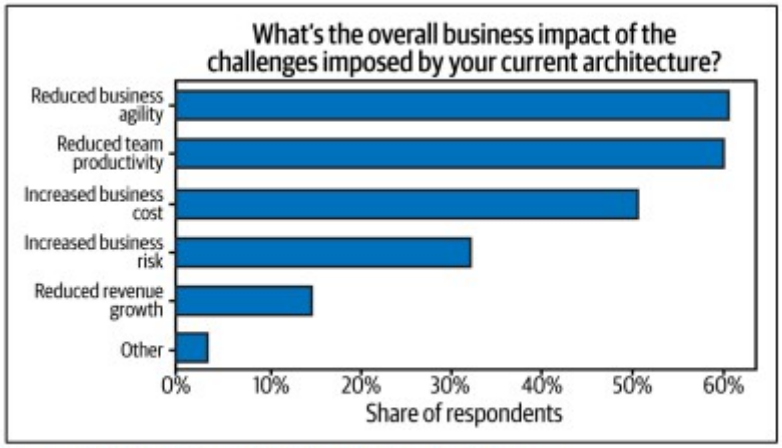

Figure 9. Overall impact of architectural challenges

Más del 60% de los encuestados dijo que tener que gestionar una infraestructura de datos compleja reducía la agilidad empresarial y la productividad del equipo. Sólo un poco más de la mitad (51%) dijo que el alto costo de las arquitecturas de datos complejas afectó a su negocio. Esto de nuevo refuerza la conclusión de que la reducción de costes no es necesariamente el principal motor para simplificar una arquitectura de datos. Los controladores más suaves como la agilidad y la productividad del negocio parecen importar más cuando se trata de obtener valor de los datos.

Prever un mejor entorno de datos.

No es de extrañar, entonces, que el 89% de los encuestados sintieran que simplificar la arquitectura de datos al mismo tiempo que apoyar todos los casos de uso necesarios sería valioso, con la pluralidad (47%) diciendo que sería "muy valioso", como se muestra en la Figura 10.

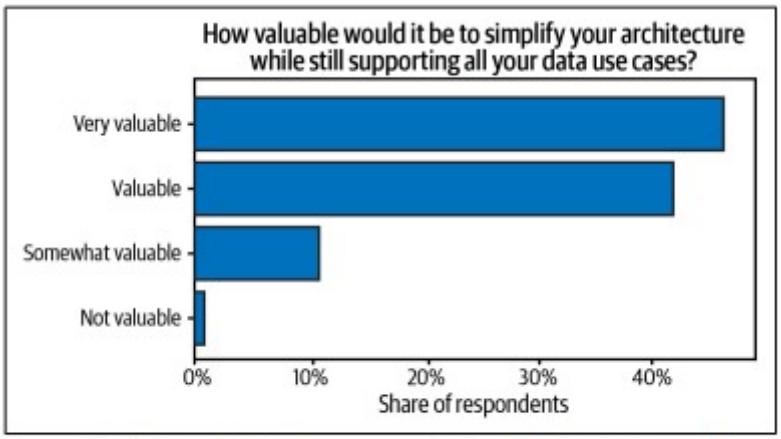

Figure 10. Data professionals are eager to simplify their data architecture platforms

Y cuando se les preguntó qué cualidades eran importantes al concebir una nueva arquitectura de datos, como se muestra en la Figura 11, una gran mayoría de los encuestados dijo que las siguientes características eran "importantes" o "muy importantes" para un entorno de datos ideal: datos centralizados, datos abiertos, características empresariales, arquitectura nativa de la nube, una relación precio/rendimiento eficiente y una que admite todos los casos de uso.

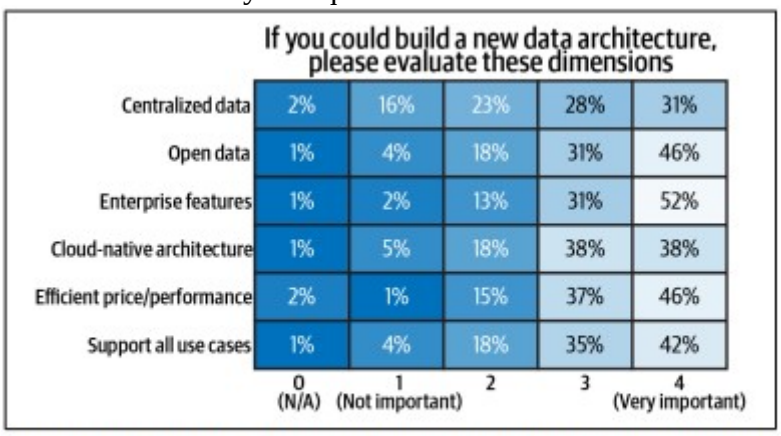

Figure 11. Importance of proposed dimensions for a new type of data architecture

Echemos un vistazo más de cerca a estas cualidades "imprescindibles":

Datos centralizados.

El hecho de tener datos centralizados es un aspecto importante o "muy importante" de ser impulsados por datos para aproximadamente 6 de cada 10 profesionales de datos (59%). Centralizar sus datos le da una única fuente de verdad. Se asegura de que sus equipos de datos y de hecho todo el mundo en la organización está utilizando los mismos datos, lo que conduce a informes más alineados entre los departamentos y, en última instancia, una mejor toma de decisiones.

Tener un almacén de datos centralizado es fundamental para ayudar a las empresas a determinar qué procesos de negocio funcionan de manera eficiente y cuáles no. También es esencial para tomar decisiones inteligentes basadas en datos que aseguren la relevancia del negocio. Es el mejor enfoque para garantizar que los datos siguen siendo un activo y no un limitador del éxito empresarial.

Datos abiertos.

Incluso más profesionales de datos (77%) quieren una plataforma que sup ports formatos de almacenamiento que sean abiertos y estandarizados, como Apache Parquet. Las plataformas que ofrecen formatos abiertos y estandarizados también deben proporcionar API, de modo que cualquier tipo de herramienta o motor, incluyendo bibliotecas de aprendizaje automático y Python o R, pueda acceder a los datos directamente de una manera eficiente. En general, los profesionales de datos creen cada vez más que el futuro de los datos y la inteligencia artificial depende de plataformas abiertas y agnósticas para la nube.

Características de Enterprise.

Los profesionales de datos también quieren una plataforma con características empresariales, con el 83% de los encuestados diciendo que se trata de una necesidad importante o "muy importante". Los sistemas de nivel empresarial requieren herramientas de seguridad y control de acceso. Las apuestas de la tabla de clase empresarial incluyen capacidades de gobernanza de datos como auditoría, retención y linaje. También se necesitan herramientas que permitan el descubrimiento de datos, como catálogos de datos y métricas de uso de datos.

Arquitectura nativa de la nube.

Más de tres cuartas partes (76%) de los encuestados pensaban que una arquitectura nativa de la nube era importante o "muy importante", porque es necesario aprovechar al máximo las aplicaciones nativas de la nube y los almacenes de datos nativos de la nube y lagos de datos. Las arquitecturas nativas de la nube ayudan a las organizaciones a ofrecer las soluciones digitales ágiles, automatizadas, escalables y altamente disponibles que son fundamentales para una organización impulsada por datos. Además, a medida que las empresas se despliegan cada vez más en más de un proveedor de nube, esto significa adoptar una tecnología que puede proporcionar una experiencia coherente independientemente de dónde se almacenen los datos.

Precio/rendimiento eficiente.

Según la encuesta, tener una relación precio/rendimiento eficiente es importante o "muy importante" para el 83% de los profesionales de datos. Esto invariablemente significa pasar a una arquitectura de nube, que permite la transformación de los requisitos de gastos de capital pesados (CAPEX) para una plataforma de datos en las instalaciones en gastos operativos (OPEX) que están alineados con lo que está pasando en el negocio. El crecimiento exponencial de los datos no justifica el crecimiento exponencial de los costes de la infraestructura de datos. Y cuando no tienes visibilidad de quién está haciendo qué con qué datos, resulta en costos descontrolados, incluyendo costos de infraestructura, costos de datos y costos laborales. Las empresas necesitan una plataforma que evite este exceso de gastos.

Soporte para todos los casos de uso.

Finalmente, más de tres cuartas partes de los encuestados (77%) dijeron que tener una plataforma que apoye todos los casos de uso es importante o "muy importante." Esto significa que la plataforma necesita apoyar la ingeniería de datos, análisis de datos, ciencia de datos y casos de uso de aprendizaje automático que implican datos estructurados y no estructurados, así como el apoyo de datos por lotes y streaming. En resumen, cualquier cosa que su organización quiera hacer con los datos, el formulario de plataforma de datos debería ser capaz de manejarlo.

Abordar los límites de la corriente.

Arquitecturas de datos En los últimos años, el soporte para un nuevo tipo de arquetipo de gestión de datos ha crecido. Este informe abarcó los depósitos de datos anteriores y cómo han evolucionado desde sus inicios a finales de los años 80 para adaptarse a las necesidades cambiantes de las empresas en materia de apoyo a la toma de decisiones e inteligencia empresarial. También se discutió cómo, aunque los almacenes de datos han sido excelentes para los datos estructurados, las empresas de hoy en día tienen enormes cantidades de datos semiestructurados y no estructurados que quieren utilizar. Es por eso que, a principios de la década de 2010, las organizaciones comenzaron a construir grandes depósitos de datos para datos brutos que soportan tanto datos estructurados como no estructurados.

Pero aquí también ha habido limitaciones. Los lagos de datos pueden almacenar cantidades masivas de datos, pero no pueden soportar transacciones, son débiles en la gobernanza de datos, y su falta de consistencia y aislamiento hacen muy difícil mezclar apéndices y lecturas y realizar trabajos tanto por lotes como por streaming.

Por estas y otras razones, los lagos de datos no han cumplido su promesa. Pero el deseo de las empresas por un sistema de datos flexible y de alto rendimiento sigue siendo fuerte. Como la encuesta ha demostrado, las empresas buscan sistemas que puedan manejar una amplia gama de casos de uso diversos, que abarcan análisis SQL, monitoreo en tiempo real, ciencia de datos, inteligencia artificial y aprendizaje automático.

Con respecto a este último, muchas innovaciones recientes de la IA se centran en el procesamiento de datos no estructurados como texto, imágenes y vídeo. Los almacenes de datos no pueden almacenar este tipo de datos, y los lagos de datos no son óptimos para otros casos de uso, por las razones indicadas anteriormente. Así que en muchos casos, las empresas implementan y administran múltiples sistemas: tal vez un lago de datos, varios almacenes de datos y otros sistemas especializados como streaming, series temporales, gráficos o bases de datos de imágenes.

Pero, de nuevo, como la encuesta mostró, ejecutar múltiples sistemas introduce complejidad, con todas sus dificultades concomitantes. ¿Qué solución hay para el desafío?

Recientemente, una nueva arquitectura llamada lakehouse ha surgido como una alternativa a las arquitecturas heredadas del pasado.

¿Qué es un Lakehouse ?

Una casa de lago combina los mejores elementos de lagos de datos y almacenes de datos para construir algo nuevo. Los lakehouses tienen estructuras de datos y características de gestión de datos similares a los almacenes de datos, pero utilizan el almacenamiento flexible y de bajo costo de los lagos de datos. En otras palabras, son como serían los almacenes de datos si fueran diseñados hoy, en un mundo donde el almacenamiento barato y altamente confiable, en forma de almacenes de objetos, está disponible (ver Figura 12).

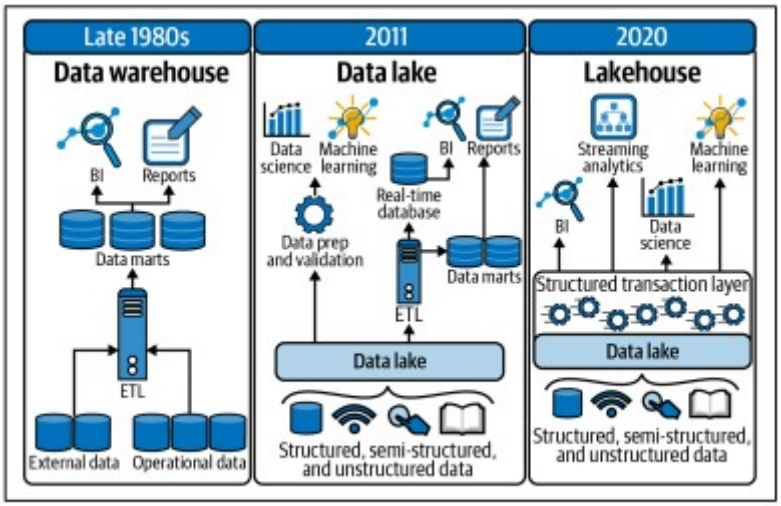

Figure 12. Each type of data repository has strengths and challenges

Una casa lakehe tiene las siguientes características clave:

Soporte de transacciones.
En una casa de lago, múltiples tuberías de datos con frecuencia leen y escriben datos simultáneamente, generalmente usando SQL. El apoyo a la atomicidad, consistencia, aislamiento y durabilidad (transacciones ACID) asegura la consistencia cuando esto sucede.

Esquema de aplicación y gobernanza.
La casa del lago debe apoyar la aplicación del esquema y la evolución, incluyendo paradigmas de esquemas de almacenamiento de datos como los esquemas estrella/ copo de nieve. El sistema debería poder garantizar la integridad de los datos y disponer de mecanismos sólidos de gobernanza y auditoría.

Soporte de BI.
Con una casa de datos, puede utilizar sus herramientas de BI directamente en su lago de datos. Esto mejora la frescura de los datos, reduce la latencia y reduce el gasto de tener que colocar y apoyar copias de los datos tanto en un lago de datos como en un almacén de datos.

Almacenamiento desacoplado del cómputo.
Debido a que el almacenamiento y el cómputo utilizan grupos separados, las casas de lago son capaces de escalar a muchos más usuarios concurrentes y tamaños de datos más grandes.

Apertura.
Los formatos de almacenamiento que usan los lakehouses (como Apache Parquet, Delta Lake y Apache HUDI) son abiertos y estandarizados, y proporcionan API para que una variedad de herramientas y motores, incluyendo las bibliotecas de aprendizaje automático y Python/R, puedan acceder eficientemente a los datos directamente.

Soporte para diversos tipos de datos, desde datos no estructurados hasta estructurados.
Una casa de lago se puede utilizar para almacenar, refinar, analizar y acceder a los tipos de datos necesarios para muchas aplicaciones de datos nuevas que requieren datos semiestructurados y no estructurados, incluyendo imágenes, vídeo, audio y texto.

Soporte para diversas cargas de trabajo.
Una casa lakehouse admite todos los casos de uso y cargas de trabajo, incluyendo ciencia de datos, aprendizaje automático, y SQL y análisis. Puede que se necesiten múltiples herramientas para soportar todas estas cargas de trabajo, pero todas dependen del mismo repositorio de datos.

Streaming de extremo a extremo.
Los informes en tiempo real son la norma en muchas empresas. Lakehouses soporta streaming, lo que elimina la necesidad de sistemas separados dedicados a servir aplicaciones de datos en tiempo real.

Todas estas características clave se suman para hacer los datos lakehouses considerablemente más atractivo que los lagos de datos o almacenes de datos por sí solos.

Conclusión: Sumérgete en un Lake-house

Las empresas más exitosas en las próximas décadas serán- debajo de las trampas de superficie de industrias específicas-empresas de datos.

Para permitir la transformación masiva de datos que se discute en este informe, es necesario reunir a todos sus usuarios y todos sus datos, y dar a sus usuarios las herramientas y la infraestructura que necesitan para extraer información de los datos. Usted necesita una única plataforma de datos empresarial construida sobre estándares abiertos que se escala en todos los departamentos y equipos.

Las empresas que se inician en estos desafíos pueden adoptar un enfoque gradual o adoptar soluciones locales y trasladarlas a la nube. Pero sin un enfoque holístico, nativo de la nube, usted se está preparando para reemplazar una arquitectura anticuada con otra que no entregará los bienes a largo plazo.

Los siguientes cinco pasos pueden asegurar que usted está progresando hacia un sistema que puede soportar la prueba del tiempo.

Paso 1: Reúne todos tus datos.

Las empresas durante décadas han dependido de los almacenes de datos para agregar datos de negocios estructurados y tomar decisiones mediante la creación de paneles de BI utilizando herramientas de visualización. Cuando los lagos de datos debutaron a principios de la década de 2010, finalmente hicieron que la ciencia de datos, la inteligencia artificial y el aprendizaje automático fueran factibles para las empresas. Hoy en día, el modelo de casa de lago combina la fiabilidad de los almacenes de datos con la escalabilidad de los lagos de datos utilizando un formato abierto como Delta Lake o Apache HUDI. Independientemente de sus opciones de arquitectura específicas, debe elegir una plataforma que pueda almacenar todos sus datos estructurados y no estructurados por igual en formatos abiertos adecuados tanto para las cargas de trabajo de análisis de datos como de ciencia de datos, lo que le permite mantener un control a largo plazo sobre sus datos.

Paso 2: Habilitar a todos los usuarios para acceder de forma segura a los datos que necesitan para hacer su trabajo.

Asegúrese de que cada miembro de su equipo de datos, a través de varias funciones y unidades de negocio -ingenieros de datos, científicos de datos, arquitectos de datos y analistas de datos- tenga acceso a todos los datos que necesita, y ninguno de los datos que no está autorizado a acceder. Esto significa cumplir con regulaciones como GDPR, CCPA, HIPAA y PCI, dependiendo de su industria y ubicación geográfica.

También es crítico que todos sus datos permanezcan juntos, en un lugar cen tralizado. Si usted está fragmentando los datos copiando en un nuevo sistema para un subconjunto de usuarios-por ejemplo, en un almacén de datos para un determinado conjunto de sus usuarios de BI-usted sufrirá de "deriva de datos," que conduce a problemas en el paso 3. También significa que usted tendrá deriva verdad, donde alguna información en su organización será rancia o de calidad deficiente, lo que conduce a (en el mejor de los casos) la desconfianza organizacional de los datos, o, más probablemente, malas decisiones que conducen a malos resultados empresariales.

Paso 3: Administre su plataforma de datos como usted administra su negocio.

Cuando usted a bordo de un nuevo empleado, usted los configura para el éxito. Obtienen el ordenador adecuado, acceso a los sistemas adecuados, etc. Su plataforma de datos debe ser la misma, debe estar configurada para tener éxito.

Dado que todos sus datos están en un solo lugar, cada empleado verá una faceta diferente de esos datos, de acuerdo con el papel y las responsabilidades definidas de cada uno. Este tipo de acceso a los datos debe ser automatizado en función de sus procesos de incorporación, y debe ser transparente para que pueda ser fácilmente auditado.

Paso 4: Aproveche la seguridad nativa de la nube.

A medida que la nube se ha convertido en la ubicación de facto para el procesamiento masivo de datos y el aprendizaje automático, la tradicional "zona desmilitarizada" (DMZ) y la seguridad perimetral de la seguridad "in situ" se están reemplazando por redes de confianza cero y definidas por software. En consecuencia, las empresas deben garantizar que sus plataformas de procesamiento de datos están diseñadas para la nube y que aprovechan los mejores controles nativos de la nube de su clase.

Además, dado que cada usuario accede a los datos que necesita con sus propias credenciales en

línea, las capacidades de auditoría en la nube y telemetría le proporcionan un registro de acceso y modificación de datos a través de herramientas nativas de la nube. Esto hace posible el paso 3.

Paso 5: Automatizar para la escala.

Ya sea que esté desplegando su plataforma de datos en cientos de unidades de negocio o muchos miles de clientes, este proceso debe ser completamente automatizado. Esto significa que su plataforma de datos debe desplegarse con cero intervención humana.

Además, para cada espacio de trabajo (entorno para cada unidad de negocio), el acceso a datos, los modelos de aprendizaje automático y otras plantillas también deben configurarse automáticamente.

La potencia de esta escala exige controles potentes. Con el poder de cómputo de millones de máquinas al alcance de su mano, es fácil ejecutar facturas de nubes masivas. Para implementar en los departamentos de toda la empresa, las políticas de gasto correctas y los cobros deben diseñarse para asegurar que la energía se está implementando como el negocio espera. Las API pueden automatizar todo, desde el aprovisionamiento de usuarios y espacios de trabajo en equipo hasta la ejecución de tuberías de producción, el control de costos y la medición de resultados empresariales. Una plataforma totalmente automatizada es necesaria para impulsar su empresa.

Sobre el autor

Alice LaPlante es una escritora, editora y profesora de escritura galardonada, tanto de ficción como de no ficción. Becaria de Wallace Stegner y profesora de Jones en la Universidad de Stanford, Alice enseñó escritura creativa en Stanford y en el programa de San Francisco State MFA durante más de 20 años. Alicia, una de las escritoras más vendidas del New York Times, ha publicado cuatro novelas y cinco libros de no ficción, así como libros superventas editados para muchos otros escritores de ficción y no ficción. Regularmente consulta con firmas de Silicon Valley como Google, Salesforce, HP y Cisco sobre sus estrategias de marketing de contenidos. Alice vive con su familia en Palo Alto, California, y Mallorca, España.