PowerData

Data Lake: Superando las limitaciones del Data Warehouse

El aliado del Big Data que te ayuda a cambiar las reglas

Índice

¿Qué papel juega el Data Lake en el mundo del Big Data?

Ya han pasado unos años desde la explosión Big Data y la mayoría de las organizaciones han madurado en su estrategia de grandes datos. De la euforia inicial se ha hecho la transición hacia un estado mucho más consciente, en el que se entiende el valor del dato y se asume que la importancia del acceso a la información de calidad es crítica.

Primero fueron las herramientas más amigables, que facilitaban las consultas, después un modelo de autoservicio usuario que empoderaba a diferentes perfiles y hoy, el Data Lake, termina de redondear el enfoque estratégico de la gestión de datos en la organización, al democratizar el acceso al conocimiento, modelándolo a la medida de las necesidades del usuario.

El Data Lake va más allá del Data Warehouse superando algunas de sus limitaciones. Puede que en tu negocio los almacenes de datos ya se hayan empezado a quedar pequeños, dándoos la pista necesaria para buscar una nueva fuente de información. Un repositorio donde no sea necesario tener que elegir, donde se puedan trabajar los activos informacionales que se emplearán mañana y en el que, mientras tanto, el rendimiento aplicado a consultas, procesamiento y análisis toma un impulso sin precedentes.

Entre los beneficios de usar Data Lake para el trabajo con Big Data se encuentran:

- Facilidad de búsqueda de nuevas maneras de innovar en la utilización de los datos de la organización, tanto para reporting como para analytics.

- Mayor agilidad gracias a su trabajo que se adapta a la necesidad del usuario respondiendo en el tempo adecuado que puede, perfectamente, ser en tiempo real si así es requerido

- Más capacidad y flexibilidad en un entorno en el que confluyen datos procedentes de la organización con información proveniente del exterior.

Data Lake es la próxima generación en materia de almacenamiento de datos, una evolución que el usuario ya viene precisando para poder hacer su trabajo. Se trata de una forma distinta de orientar la utilización de los datos organizacionales y el reporting y la única manera de poder responder a consultas sobre el propio negocio, haciéndolo sin barreras

¿Cómo integrar el Data Lake en tu arquitectura de datos (Data Lake y Data Warehouse)?

El Data Warehouse es el primer paso para poder tomar decisiones en una empresa, por eso no hay que prescindir de él. Sin embargo, el Data Lake permite ir más allá, al mejorar algunas de las ineficiencias del repositorio tradicional, como:

- Requiere de demasiado tiempo para analizar las fuentes de datos y las necesidades de negocio, dando como resultado un modelo muy estructurado de los datos que es el único que permite construir informes.

- Para estar listo para consumir determinados datos necesita que toda la información se someta a un proceso de modelado previo, a diferencia del Data Lake, donde no se pierde tiempo en esa transformación, sino que se invierte la mayor parte del tiempo en la ingesta de datos.

- Debido a su capacidad limitada, obliga a seleccionar muy bien el tipo de datos a almacenar. Además, hay que prestar atención a su formato y fuente de origen, puesto que no todos se pueden recoger en el Data Warehouse. Por el contrario, el Data Lake permite pensar en el futuro y nutrir la fuente de conocimiento de la empresa con datos que, aunque quizás hoy no se vayan a utilizar, es posible que se necesiten mañana.

- El Data Warehouse está estructurado del todo y el Data Lake no. Los Data Lakes soportan todos los tipos de datos (estructurados, semiestructurados y no estructurados). Archivos Swift, pdf, Excel, twitter, Facebook, gps, etc...

- Un almacén de datos está pensado para que el consumo de información se lleve a cabo de una forma determinada, casi prefijada, y que limita las posibilidades de explotar el valor de la información. Más allá del reporting o la analítica, con el lago de datos el modo en que cada usuario consume la información depende de él.

- El Data Lake puede recibir información del Data Warehouse para nutrirse con el feedback procedente de un análisis o una consulta, pero, al revés, el proceso no es posible si no existe una estructuración previa.

- Cada vez que el Data Warehouse crece hay que hacer un análisis que permita averiguar dónde está la información, cómo funcionará el proceso y cómo se estructurarán los datos. Se trata de una resistencia al cambio implícita totalmente opuesta a la adaptabilidad que permite el lago de datos, que está diseñado para asimilarse al cambio de forma sencilla y rápida.

La llegada del Data Lake no desplaza al Data Warehouse, sino que sirve para ampliar las capacidades de la organización en lo que concierne a su gestión de datos. Así, el lago de datos puede utilizarse para grandes cantidades información no estructurada y el almacén de datos permite que el negocio logre una visión única de la verdad.

Uno y otro cuentan con ventajas e inconvenientes, como pueden ser la necesidad de implementar herramientas que faciliten la interacción con los datos, en el caso del Data Lake; o la falta de agilidad para fines analíticos que va indefectiblemente unida al Data Warehouse.

Pero es posible utilizar ambos. El Data Lake permitirá a los usuarios experimentar con diferentes modelos de datos y transformaciones, antes de configurar un nuevo esquema en un almacén de datos y servirá como área de ensayo, desde la cual suministrar datos al Data Warehouse. Y, a su vez, este almacén de datos limpios con valor conocido podrá nutrir y retroalimentar al lago corporativo que, cada vez concentrará mayor valor.

Obtener lo mejor de ambas soluciones es cuestión de visión y arquitectura. Es la manera de no dejar escapar ni una sola oportunidad

¿Cuál es el impacto del Data Lake en el negocio?

La misión del Data Lake es facilitar la ingesta de información, aunque su funcionamiento podría resumirse en tres fases:

- Entrada de datos
- Procesamiento y análisis de la información
- Retroalimentación

Sí, a diferencia de un almacén de datos, en el Data Lake las consultas no son el fin de cada proceso, sino que toda esa información que, gracias a distintos procesos analíticos se ha convertido en conocimiento, vuelve al lago de datos para darle el feedback necesario que lo hace más inteligente. De esta forma, la organización se sumerge gracias al Data Lake en un proceso de aprendizaje continuo muy fructífero.

El impacto de esta nueva forma de trabajar en el negocio es equiparable a una revolución interna.

Cada empresa tiene el poder de sacar todo el partido a sus datos, llegando a conocerse como nunca antes y exprimiendo sus posibilidades de formas muy distintas.

La construcción de un Data Lake implica la llegada de grandes oportunidades para organizaciones de diferentes sectores, como las siguientes:

Salud: disponer de un Data Lake permite hallar el mejor tratamiento para cada paciente.

Las enfermedades no afectan igual a todos los organismos. Pese a tener el mismo nombre y apellidos, una patología sigue un curso distinto en cada cuerpo y, para evaluar su impacto, las posibilidades de cura y las opciones más indicadas de tratamiento es necesario hacer un trabajo de investigación importante.

Sin la ayuda de la tecnología un médico necesitaría semanas para acceder y leer todos los historiales clínicos relevantes y la información sobre tratamientos que se requiriese. Pese a todo, su trabajo no estaría completo y le faltaría consistencia. Si contase con un Data Warehouse estaría limitando su labor a la información clínica procedente de su organización, privándose de la riqueza que podría aportarle la consulta de fuentes externas.

Un Data Lake permite al profesional de la salud absorber todo el conocimiento que se deriva de:

- Datos enviados desde las fuentes operativas o más orientadas a reporting.
- Datos históricos.
- Información proveniente del exterior de la clínica, como fuentes oficiales, estudios y datos de otras clínicas.

De esta forma, su criterio se forja tras analizar datos de muchos miles de casos, permitiéndole diseñar un tratamiento a la medida del paciente, totalmente personalizado y adecuado a sus necesidades particulares. Y esta completitud y visión no son las únicas ventajas, sino que también hay que valorar otro beneficio que permite el Data Lake, que tiene que ver con que este proceso puede estar impulsado y llevarse a cabo por una única persona, por ejemplo, el médico de cabecera del paciente.

Telecomunicaciones: el Data Lake es su principal aliado a la hora de economizar recursos.

Cuando las empresas de telecomunicaciones tienen un Data Lake donde volcar los datos crudos están poniendo en manos de sus usuarios una herramienta muy potente. Cada individuo puede hallar diferentes formas de explotar la información contenida en este repositorio para generar valor para su organización.

Uno de los miles de modos de hacerlo es, por ejemplo, analizando el consumo de dispositivos. En vez de centrarse en estudiar la utilización de las antenas, como se viene haciendo, profundizar en la forma en que se consume aporta una perspectiva mucho más útil al permitir ganar conocimiento acerca de:

- Las necesidades diarias de las personas con respecto a las telecomunicaciones.
- El consumo que se lleva a cabo desde el móvil o la Tablet.
- El consumo que no se realiza desde dispositivos móviles sino desde el ordenador.
- Las actividades de comunicación de los usuarios.

Porque, realmente, no son necesarias todas las antenas a todas horas. Pero es complicado decidir qué antenas se pueden apagar y en qué franjas horarias hacerlo si no se conocen los hábitos de los clientes. Sólo el Data Lake posibilita entender que, a partir de una cierta hora la demanda crece mucho, y desde el momento en que el día va terminando, cuando llegan las horas frecuentemente destinadas al sueño, esa demanda cae, hasta situarse en un reducido 10% con respecto a la del día.

El coste de mantener encendidas esas antenas por la noche, esas 4, 5 o 6 horas es enorme y puede desglosarse en gastos asociados a:

- Mantenimiento.
- Servicios de bases de datos.
- Impacto ambiental.
- Consumo de energía.

Las empresas que cuentan con un Data Lake pueden acceder a los datos exactos que les permiten conocer puntualmente todo lo necesario sobre los hábitos de consumo para poder tomar decisiones respecto a apagar las antenas menos necesarias.

Además, el lago de datos consigue que las empresas sean más inteligentes, permitiéndoles aprender de sus errores y optimizar las soluciones. Los errores detectados al ingerir millones de registros pueden analizarse, permitiendo conocer esos patrones peligrosos que podrían convertirse en nuevos problemas en el futuro. Los resultados de este análisis pueden emplearse para lograr una mejora continua en la calidad de las llamadas. Los márgenes de servicio y la satisfacción del cliente se verían impactados de forma muy positiva.

De la misma forma, todos los datos que se conservan y que hacen posible entender cuándo los usuarios recurren al servicio de telecomunicaciones, en qué circunstancias, con qué objetivo y desde dónde, son la vía más rápida hacia la innovación. Esta valiosa fuente de conocimiento que no deja de crecer es la responsable de la minimización de los fracasos y el aumento de los éxitos, los lanzamientos que terminan convirtiéndose en best sellers.

Seguros: disponer de un Data Lake permite mejorar los resultados de las acciones de marketing y optimizar la detección del fraude.

Es frecuente contar con asegurados que son clientes de distintos tipos de producto dentro de la misma familia. El padre ha elegido la protección del hogar y el hijo la de su vehículo. Las acciones de marketing de las compañías aseguradoras orientadas al up selling y cross selling se benefician de un Data Lake, puesto que les garantiza esa visión 360 real.

La efectividad de una campaña de marketing tiene mucho que ver con la personalización y también con el momento. Un Data arehouse permite hacer una buena aproximación a partir de la que diseñar una promoción bastante certera, con un alto índice de éxito. Sin embargo, sus limitaciones dejan un elevado porcentaje de incertidumbre que podría echar por tierra todos los esfuerzos, la inversión e incluso hacer perder un cliente.

En lo que respecta al servicio al cliente hay que actuar con cautela y precisión. No caben los errores puesto que cometer un fallo podría afectar a:

- La imagen de la empresa.
- La confianza de los clientes y su lealtad.
- La percepción de transparencia y honestidad que se tiene de la marca.

El Data Lake pone en manos del corredor de seguros una herramienta muy potente, mediante la que, con gran agilidad, se puede poner en la piel del cliente y entender cuáles son sus verdaderas necesidades, las de hoy, las de ahora mismo. Desde esa posición es más sencillo elaborar una propuesta que no podrá rechazar, una oferta que, además, le demostrará que le conocen bien y buscan su mejor interés.

Por otra parte, las aseguradoras se encuentran bastante solas en su lucha contra el fraude, otra de las principales preocupaciones de sus responsables. Pero el Data Lake les ayuda a ser infalibles a la hora de identificar a los sujetos que no deberían incluir entre sus clientes. Basta con volcar en sus lagos de datos toda la información sobre cada asegurado, productos contratados, fechas de alta y todo sobre las pólizas para conocer rápidamente quiénes son las personas que tienden a litigar.

Partiendo de ciertos rasgos en común entre los sujetos que presentan perfiles de este tipo, resulta cada vez más fácil predecir quién generará más costes que otros clientes. De esta forma, se pueden prevenir esos gastos que ahogan la rentabilidad de un cliente y la diluyen entre costes legales, administrativos, etc...

Banca: mediante el Data Lake las tareas de búsqueda de nuevos productos se hacen más sencillas.

Además de la prevención del fraude, que pueden impulsar de forma similar a como se hace en la industria del seguro, con la ventaja de poder hacerlo en tiempo real para ganar en confiabilidad respecto de los usuarios; o de atraer nuevos clientes mediante técnicas de marketing que les aporten beneficios, el Data Lake impulsa el perfil innovador de las empresas del sector bancario.

Para diferenciarse de la competencia hace falta usar la creatividad y ser originales, pero, al mismo tiempo, los esfuerzos deben dirigirse en la dirección correcta. Allí es donde puede apuntarse con el conocimiento extraído del lago de datos, que hace posible:

- Analizar la cartera de productos.
- Relacionar los diferentes productos con calidad.
- Enriquecer los resultados, buscando duplicados, faltas de completitud y errores, pero también aplicando acciones de limpieza y perfilado de datos para corregirlos.
- Buscar nuevas soluciones que aporten valor a los clientes y se ajusten a sus necesidades actuales

El Data Lake permite aumentar la fidelidad del cliente, posicionarse mejor en el mercado, evitar gastos administrativos, costes legales y trabajar de forma más eficiente y segura.

Pero en esta industria, contar con el soporte de un lago de datos implica muchas más ventajas. Por ejemplo, hasta hace unos años, los sistemas de liquidación en su mayoría trabajaban de manera independiente y transferían datos a través de interfaces o informes para propósitos de control o ventas. Hoy en día, cada vez hay más requisitos que exigen análisis en tiempo real y llega un punto en que el Data Warehouse no puede responder a la demanda, le falta flexibilidad, y ello puede perjudicar a la entidad en términos de cumplimiento. Las iniciativas regulatorias y los requisitos para información precisa y en detalle van en aumento, exigiendo la disponibilidad de información de liquidación en tiempo real en las aplicaciones de ventas.

La cuestión es que normalmente no es posible un acceso directo a los sistemas de liquidación, bien por cuestiones arquitectónicas, bien por motivos de estabilidad y rendimiento de los sistemas, bien por obstáculos técnicos relacionados con la horizontalidad y la descentralización. La forma de superarlos todos pasa por integrar ese almacén de datos que ya no da más de sí con un Data Lake, que hace posible alinearse con las necesidades actuales del negocio ofreciendo una respuesta sólida y puntual.

Retail: disponer de un Data Lake les permite ganar en rentabilidad.

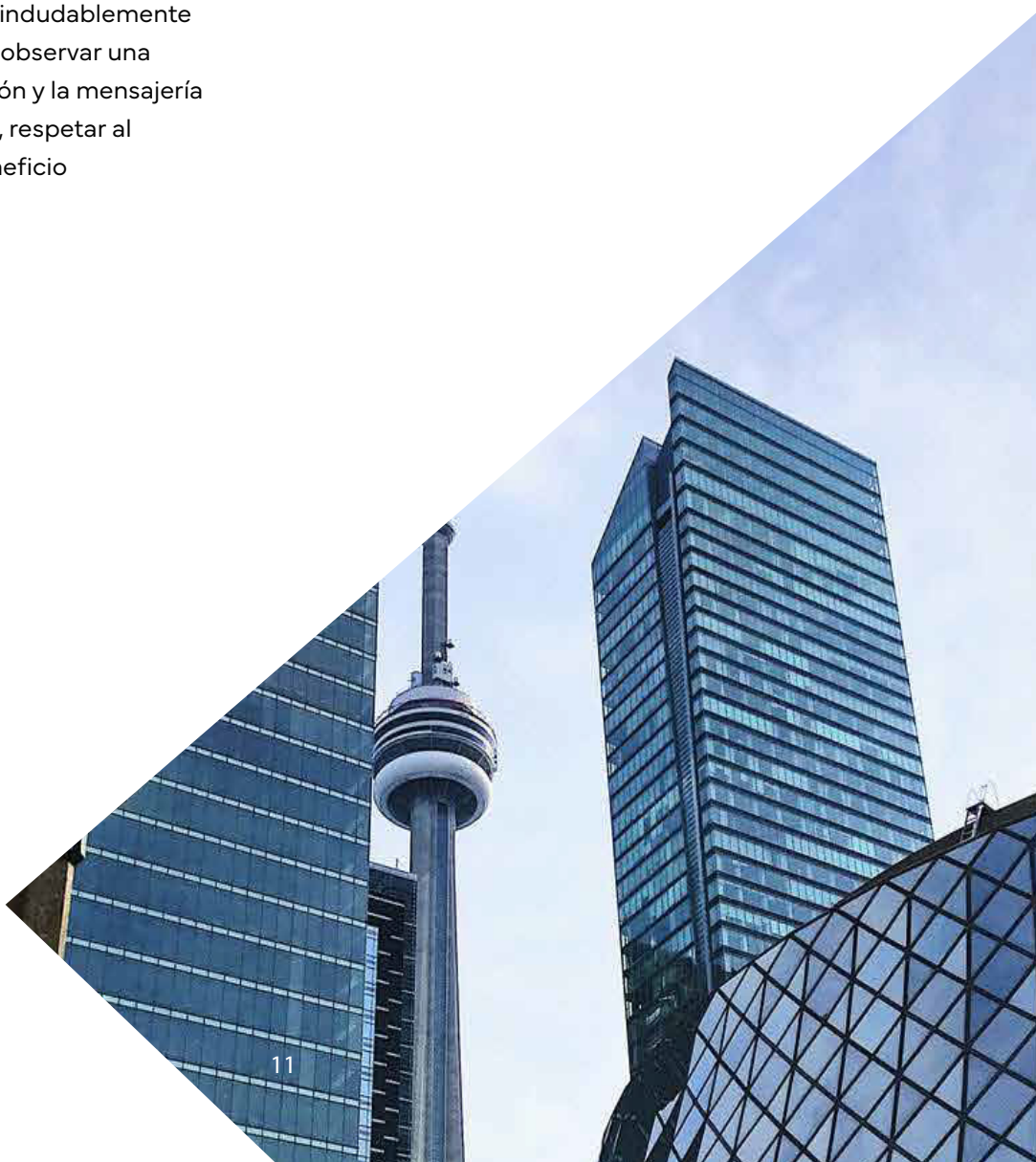
No se trata solo de tener más clientes y mejores, sino que también hace falta saber ahorrar. Pero en una industria como la minorista, la calidad o la puntualidad no deben nunca verse afectadas, por lo que es esencial conocer bien a los proveedores.

La competitividad en este sector obliga a sus responsables a apoyarse en la tecnología para hacer análisis y, con el Data Lake pueden escudriñar a todos y cada uno de sus suppliers para cada uno de los cientos y hasta miles de productos que venden. Al mismo tiempo, el lago de datos les facilita el proceso de analizar a los clientes. Gracias a toda la información que contiene procedente de datos transaccionales, históricos de compras, datos de redes sociales, de blogs, de foros de opinión, de encuestas, de estudios de mercado, y mucho más pueden saber:

- El nivel de satisfacción de los consumidores.
- Cómo aumentar su lealtad.
- Cuándo van a realizar la próxima compra.
- Qué mecanismos les ayudarían a incentivar las compras en días diferentes de la semana.

Todo este conocimiento es fuente de ventaja competitiva. Porque, desde el momento en que se tiene la capacidad para cotejar datos procedentes de redes sociales con datos internos de los clientes, se incrementan las capacidades de segmentación y las habilidades analíticas del negocio.

El Data Lake permite lograr el nivel más elevado de personalización, el que conduce a una experiencia del todo única, indudablemente satisfactoria y sólo hay que observar una precaución: reducir la fricción y la mensajería irrelevante para los clientes, respetar al consumidor y buscarsu beneficio



Automoción: el Data Lake les permite mejorar sus resultados de Advocacy Marketing.

En el sector del automóvil ya llevan tiempo usando técnicas de inbound marketing para atraer y retener prospectos. Les funcionan. Hoy día, buscan perfeccionar aún más sus resultados aplicando el método de Advocacy Marketing para explotar el valor de sus influencers.

Lo que se dice en redes sociales cuenta, y mucho (no hay que olvidarse de las consecuencias del escándalo VolksWagen, del que aún se vierten comentarios en el social media); por eso, las principales compañías del sector de la automoción quieren trabajar esa conexión que les une a su público a través de su eslabón más cercano a esa audiencia que desean captar, los influencers.

El Data Lake les facilita la tarea de segmentar a los influencers para convencerles a través de ofertas. Ése es el primer paso y, una vez completado con éxito, tras estos perfiles más destacados, llegará el gran público.

Además, contar con un Data Lake facilita a las empresas del sector el alcanzar nuevos segmentos de clientes de forma directa, a través de la innovación. La conducción autónoma no es la única posibilidad, sino que es una de las tendencias que llegan, captando la atención de un tipo de consumidor diferente. Se trata de novedades como las que tienen que ver con la automatización de la gestión del vehículo, su integración con aplicaciones domóticas en el hogar o las nuevas posibilidades de entretenimiento que facilitan opciones avanzadas de conectividad integradas con el cuadro de mando del automóvil

Pasos para el diseño de un Data Lake

El diseño de un Data Lake debe hacerse a conciencia. Hay que tener en cuenta que, del acierto en su configuración dependerán los resultados que se obtengan del análisis y procesamiento que por medio de la información contenida en este lago se lleve a cabo.

En el proceso de construcción de una arquitectura de este tipo hace falta seguir los siguientes pasos:

Preparar la transición de un schema-on-write a un schema-on-read.

No es lo mismo el planteamiento a la hora de diseñar un data warehouse que la preparación que precede a la configuración de un data lake.

Por eso, hay que tener claro qué tipo de datos se trabajan, cuáles se necesitarán en el futuro, cómo se explota Big Data en la organización, hasta qué punto llega la necesidad de asegurar una visión única, si la escalabilidad resultaría beneficiosa para el negocio, si puede ser preciso el acceso simultáneo a aplicaciones y datos, si se está avanzando hacia el empoderamiento usuario a partir de su autonomía en el consumo de información o si realmente es necesario llevar a cabo análisis de datos en tiempo real.

Garantizar que se cubren las necesidades en materia de infraestructura.

El soporte del Data Lake que le permitirá funcionar al rendimiento deseado depende de las directrices que se planteen a la hora de perfeccionar su diseño en materia de gobierno de datos, gestión de datos y datos maestros, gestión de metadatos, seguridad, modos de ingesta de datos, fuentes de datos, aplicaciones BI y accesos.

Crear una estructura básica.

Es la etapa inicial en la necesaria integración y será determinante para aprender a adquirir y transformar datos a escala. Es un paso importante para descubrir cómo hacer que Hadoop funcione para la organización.

Trabajar el potencial de análisis

Mejorar la capacidad de análisis de datos y su interpretación dependerá de la elección correcta de herramientas que permitan combinar y fusionar gradualmente el Data Warehouse y el Data Lake.

Sentar las bases para una buena integración con el Data Warehouse.

La colaboración sostenible entre ambos repositorios es crucial para la inteligencia de negocio.
Hay que garantizar una sinergia consistente que aproveche las fortalezas de ambas arquitecturas en base a la creación de un conjunto de datos que permita el intercambio de conocimiento en todas las direcciones.

Integrar el Data Lake con las capacidades existentes de gestión de datos para optimizar el rendimiento en todas las funciones

Beneficios para el negocio de contar con un logo de datos

¿Crees que el Data Lake es un sistema de almacenamiento de datos? ¿Te parece que sus funciones se reducen al procesamiento y análisis? ¿Lo confundes con Hadoop?

Las ventajas del Data Lake superan a las de contar simplemente con un Data Warehouse y eso lo demuestran los principales puntos de diferencia entre ambos:

Datos

Un almacén de datos sólo almacena datos que han sido estructurados, mientras que un lago de datos no hace diferencias. Lo almacena todo.

Agilidad

Un almacén de datos es un repositorio altamente estructurado y, aunque a nivel de arquitectura no es complicado obrar ambios, lo cierto es que hacerlo implica un gran consumo de tiempo. Por el contrario, como el Data Lake carece de estructura, está aportando a los desarrolladores y a los científicos de datos la capacidad de configurar y reconfigurar fácilmente sus modelos, consultas y aplicaciones a la medida de sus necesidades.

Tratamiento

El Data Warehouse se basa en un schema-on-write, que sólo funciona cuando los datos se modelan de forma previa a su carga.

El Schema-on-read del lago de datos permite cargar los datos sin procesar y ocuparse del modelado en función de la demanda y coincidiendo con su uso. Un importante ahorro en tiempo y recursos.

Usuarios

Pese a que el Data Lake está orientado a todos los usuarios, quienes mayor valor podrán extraer de él son los perfiles técnicos más especializados. A diferencia de otros sistemas de almacenamiento esto supone una importante diferencia, que debe tenerse en cuenta de manera especial a efectos de garantizar la seguridad y diferenciar los niveles de acceso sin que este tipo de medidas afecten al rendimiento.

Intuyes que el Data Lake es mucho más que una tendencia. Pero siguiendo corazonadas no se llega a ninguna parte, no en los negocios. Por eso, necesitas conocer cuáles son esas ventajas que conlleva la implementación de un lago de datos en tu empresa:

- **Capacidad de acceder a los datos a alta velocidad:** además de que parece no tener límites en cuanto a volumen de datos contenidos, el Data Lake permite usarlos en condiciones de gran velocidad, aportado agilidad a la completitud, al integrar información procedente de fuentes diversas y en formatos distintos.

- Escalabilidad: dentro del lago de datos se puede almacenar una cantidad creciente de información sin que su potencial de crecimiento suponga un inconveniente, sino todo lo contrario.

- Transformación a la medida de las necesidades: si en el Data Warehouse los datos se modelan como cubos en el momento de la ingesta o entrada de datos, en el Data Lake se invierte este orden. Este repositorio ofrece una flexibilidad inigualable, al permitir que el modelado pueda retrasarse hasta el momento de consumir.

- Conversión de las fuentes de datos: no existen restricciones ni en lo que respecta a las fuentes de origen ni en lo relacionado con el tipo de datos, ya que se admiten registros, multimedia, datos procedentes de chats e emails, XML, datos de sensores, datos de redes sociales, datos binarios y datos demográficos, entre otros.

- Reutilización y eficiencia: una vez que los datos son recogidos, limpiados y almacenados en un almace15 namiento SQL estructurado del Data Lake, es posible reutilizar las secuencias de comandos existentes. Todo ello con la flexibilidad de ejecutar consultas SQL paralelas de forma masiva, al tiempo que se integran con diferentes aplicaciones y bibliotecas avanzadas de algoritmos. De esta forma, se necesita mucho menos tiempo para cada consulta y se consumen menos recursos que los que habría que emplear para llevar a cabo el procesamiento de SQL fuera del Data Lake.

- Analítica avanzada más coherente: el Data Lake puede utilizar la disponibilidad de grandes cantidades de datos coherentes junto con algoritmos de aprendizaje para reconocer los elementos de interés que impulsarán el análisis de datos en tiempo real. Ésta es una de las facetas que más le diferencian de un Data Warehouse.

- Ahorro: El funcionamiento del Data Lake se basa en Hadoop, por lo que se beneficia de un coste de almacenamiento bajo, sobre todo si se compara con el Data Warehouse. Las razones de esta ventaja en costes tienen que ver con dos motivos. Uno es el hecho de que Hadoop es un software de código abierto, por lo que el licenciamiento y el soporte de la comunidad es gratuito y el otro está relacionado con la configuración de Hadoop, que está diseñado para ser instalado en hardware de bajo costo.

Cosas que debes de saber antes de lanzarte al Data Lake

Conocer los beneficios para el negocio del Data Lake hace que falte tiempo para dejar atrás métodos menos innovadores y apostar por este repositorio del futuro, que ya es una realidad. Las empresas líderes ya trabajan con el lago de datos pero, en el proceso de dejar atrás la dependencia al Data Warehouse, hay que tener en cuenta una serie de cuestiones.

De hecho, no es recomendable lanzarse al lago de datos sin más, sino que antes hay que:

Tener en cuenta las herramientas que ya se tienen implementadas en la organización. Sobre todo, en lo que respecta a las soluciones y fuentes de datos. Entender que, para poder explotar el Data Lake al máximo de su capacidad hace falta nutrirlo de datos. Sin esa ingesta previa, el lago estaría vacío y no podría ser de gran utilidad. Si se quieren analizar correos electrónicos, es posible hacerlo, pero antes hay que cargarlos. El Data Lake debe poder encontrar al usuario de negocio y sus datos.

Tomarse en serio la seguridad. Se trata de un desafío importante puesto que este tipo de repositorio contiene una cantidad y variedad de datos mucho mayor que cualquier otro de que disponga la organización. Atender al cumplimiento y controlar los accesos y autorizaciones es el punto de partida. Para que los resultados sean fiables, debe prestarse atención a la calidad de datos, una máxima que también se aplicaba al almacenamiento en el Data Warehouse y que en el lago de datos aún cobra mayor sentido. No hay que confundir datos crudos con datos libres de errores.

Es importante el marco de metadatos y su gestión, puesto que resultan indispensables para el acceso dinámico y adecuado a los datos brutos. Además de una descripción técnica y relacionada con el contenido del Data Lake, los metadatos deben disponer de información sobre si los datos brutos pueden fusionarse y transformarse.

El gobierno de los datos deberá graduarse a la medida de las necesidades de los diferentes niveles del Data Lake. A diferencia de los Data Warehouses, donde las políticas se definían y aplicaban de forma uniforme, gracias a la estructuración de su marco, en el lago de datos hay que modular el gobierno en función del propósito.

Por último, no hay que olvidarse de definir con qué objetivo la industria quiere construir el Data Lake. Las necesidades en cuanto a los tipos de datos, los tipos de análisis o la frecuencia nunca serán las mismas, como tampoco serán iguales los aspectos a tener en cuenta respecto a la privacidad de los datos o las cuestiones relativas a seguridad.

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