

Sobre los datos

Rodrigo Barrera

Universidad de Valparaíso

Introducción

Los datos son el alma de cualquier empresa. Uno de los requisitos más importantes para que cualquier empresa alcance un alto rendimiento es desarrollar una base sólida de datos de alta calidad y, a continuación, gestionar esos datos como un activo corporativo. Sólo si los datos de una empresa son precisos y fiables, podrá aprovecharlos para actividades de valor añadido, análisis sólidos y previsiones. Este factor permite a la empresa adaptarse a los cambios del ecosistema, obtener más información sobre los clientes, optimizar los costes operativos, planificar con mayor precisión y garantizar una mejor integración de sus diversas entidades empresariales.

Introducción

Dicho esto, se entiende que gestionar los datos como un activo corporativo supone un cambio de cultura para la mayoría de las empresas. Significa tratar los datos como un activo en una planta de fabricación. Requiere una atención implacable a su calidad, capacidad para gestionar y medir el rendimiento con métricas, y aplicar medidas de seguridad de datos herméticas. Significa gestionar los datos con procesos sólidos, gobernarlos con una organización adecuada y adoptar las herramientas tecnológicas apropiadas para tener éxito en la monetización de los datos. Sin embargo, una base sólida de datos permite a una organización comprender mejor las preferencias y el comportamiento de los clientes, gestionar mejor las relaciones con los proveedores y controlar mejor su cartera de productos y servicios, aunque también plantea una serie de retos.

Introducción

Es muy importante asegurarse de que los datos no sólo se capturan y almacenan, sino que se tratan e interpretan eficazmente. Las organizaciones deben adoptar tecnologías y diseños de arquitectura de datos adecuados para optimizar los costes de almacenamiento. Una empresa también tiene que garantizar el cumplimiento legal y normativo de sus activos de datos. Los mismos datos que proporcionan a una organización una ventaja competitiva al ofrecer información valiosa sobre los clientes, también pueden plantear graves problemas desde el punto de vista de la continuidad de la actividad, si se pone en peligro la confidencialidad de los datos de los clientes y si se vulnera la información confidencial o de identificación personal de los clientes.

Introducción

Por lo tanto, es muy importante para una organización encontrar el equilibrio adecuado entre la acumulación y el almacenamiento de datos como activo corporativo, y la aplicación de estrictos principios de archivo, gobernanza y seguridad para garantizar la seguridad y confidencialidad de la información sensible. Una vez establecido el inmenso valor de la captura, conservación y consumo de datos, la siguiente sección explora los diferentes tipos de datos con los que trata una organización moderna.

Tipos de datos

Datos estructurados

El concepto de datos estructurados es bastante fácil de entender. Los datos estructurados incluyen todo tipo de datos que puedan almacenarse en una tabla de base de datos con filas y columnas. Tienen claves relacionales que pueden referirse entre sí. Este tipo de datos constituye la forma más sencilla de gestionar la información. Además, los datos estructurados representan sólo entre el 5 y el 10% de toda la información que manejamos hoy en día. Los datos estructurados representan el formato más sencillo para capturar, almacenar, organizar y analizar datos, bien organizados para facilitar el acceso. En virtud de esta construcción básica, la gestión de datos estructurados es a menudo un enfoque conveniente para tratar grandes volúmenes de información.

Datos estructurados

En el ecosistema de datos, los datos estructurados se encuentran principalmente en bases de datos relacionales. Además, las hojas de cálculo también son fuentes habituales de datos estructurados. Independientemente de si los datos estructurados proceden de una base de datos relacional compleja o de una hoja de cálculo de Microsoft Excel, uno de los requisitos previos de la gestión de datos estructurados es la creación de un modelo de datos bien definido, que a su vez impulsa las decisiones sobre cómo se capturarán, conservarán y consumirán los datos. Por ejemplo, una de las decisiones básicas se refiere al formato de los datos, ya sea numérico, alfabético o una combinación de ambos.

Datos estructurados

La siguiente tabla resume las características clave de los datos estructurados y sus principales fuentes y ejemplos. Lista de comprobación de datos estructurados Características de los datos estructurados:

- Muy organizados
- Bien definidos
- Fácilmente accesibles
- Fáciles de analizar

Datos no estructurados

Los datos no estructurados representan alrededor del 80% de los datos que se utilizan hoy en día. Incluyen principalmente texto y contenido multimedia. Algunos ejemplos de datos no estructurados son los documentos de tratamiento de textos, mensajes de correo electrónico, vídeos, fotos, archivos de audio, presentaciones, páginas web, etc. Los datos no estructurados son omnipresentes. La mayoría de las personas y organizaciones generan a lo largo de su vida grandes volúmenes de datos no estructurados. Los datos no estructurados son producidos por una máquina o por un ser humano.

Datos no estructurados

El siguiente paso es elegir la herramienta de análisis adecuada para el trabajo, y hay varias opciones para empezar. Por ejemplo, si hay que analizar una única fuente de datos, como las publicaciones en redes sociales para segmentar a los clientes y diseñar campañas de marketing eficaces, la empresa puede elegir herramientas de análisis de redes sociales o de análisis de sentimientos. Sin embargo, si una organización quiere extraer información de datos textuales, debe utilizar herramientas de análisis de texto. Independientemente de la herramienta utilizada para el análisis de datos no estructurados, los resultados deben estar claramente tabulados y ser fáciles de visualizar. Además, en este ecosistema moderno, los informes deben ser accesibles a través de ordenadores y dispositivos móviles en clientes basados en navegadores.

Cómo analizar datos no estructurados

Teradata Aster es un ejemplo ideal. Una vez más, también se puede optar por una infraestructura de almacenamiento propia. Podría ser una arquitectura altamente distribuida utilizando Hadoop. También se puede ejecutar una herramienta de análisis de software como SPSS o R en las fuentes de datos. Independientemente de la pila tecnológica elegida o del método de despliegue elegido, hay algunas consideraciones clave de diseño para tratar los datos no estructurados. Si se prevé que el almacenamiento de datos se realice in situ, tendrá que ser escalable para volúmenes de datos masivos y alto rendimiento.

Cómo analizar datos no estructurados

Si el objetivo clave es la disponibilidad de resultados en tiempo real, entonces se debe diseñar para alta disponibilidad. Si el objetivo empresarial es el análisis significativo de tendencias históricas, entonces la durabilidad de los datos se convierte en el criterio clave de diseño. El análisis de datos para extraer información útil siempre ha sido un objetivo empresarial clave para cualquier empresa.

En el ecosistema moderno, un porcentaje significativo de esa información se encuentra en forma de datos no estructurados, ya sea en la web o en aplicaciones que se ejecutan in situ. Aquellas empresas que puedan aprovechar este activo aumentarán su eficacia, acelerando la toma de decisiones empresariales que serán de mayor calidad. Con la reducción de los costes de software y hardware, estos beneficios ya no se limitan a las grandes empresas. Los beneficios están disponibles incluso hoy en día para las pequeñas y grandes empresas que se toman en serio la recolección de inteligencia empresarial a partir de sus propios datos.

	Datos estructurados	Datos no estructurados
Características Clave	<ul style="list-style-type: none"> - Modelos de datos definidos - Mayormente textuales - Fácilmente buscables 	<ul style="list-style-type: none"> - Sin modelo de datos definido - Pueden ser de formatos variados como texto, imágenes, sonido, video - Difícil de buscar
Reside en	<ul style="list-style-type: none"> - Bases de datos relacionales - Almacenes de datos 	<ul style="list-style-type: none"> - Bases de datos NoSQL - Almacenes de datos - Aplicaciones en premisa/en línea - Lagos de datos
Generado por	<ul style="list-style-type: none"> - Humanos - Máquinas 	<ul style="list-style-type: none"> - Humanos - Máquinas
Ejemplos	<ul style="list-style-type: none"> - Nombres - Números telefónicos - Números de seguridad social - Números de tarjetas de crédito - Direcciones - Fechas - Nombres y números de productos 	<ul style="list-style-type: none"> - Documentos de texto - Mensajes de correo electrónico - Archivos de audio - Archivos de video - Imágenes - Imaginería de vigilancia

Datos semi-estructurados

Es importante entender que la mayoría de los datos son híbridos hasta cierto punto. Proporciona la capacidad de añadir etiquetas, palabras clave y metadatos a tipos de datos que, de otro modo, se considerarían datos no estructurados. Por lo tanto, no entran directamente en ninguno de los dos tipos de datos antes mencionados, es decir, datos estructurados y datos no estructurados. Estos datos entran en otra categoría que se denomina datos semiestructurados, principalmente debido al hecho de que poseen una mezcla de características tanto de los datos estructurados como de los no estructurados. Los datos semiestructurados son información que no cabe en una base de datos relacional, pero que lleva asociados algunos metadatos que le dan cierta apariencia de estructura y facilitan su análisis.

Datos semi-estructurados

Con cierto tratamiento, algunos tipos de datos semiestructurados pueden almacenarse en bases de datos relacionales, aunque esto puede resultar difícil para otros tipos de datos semiestructurados. La adición de elementos descriptivos y metadatos a imágenes, comunicaciones por correo electrónico y archivos de tratamiento de textos son algunos ejemplos de datos semiestructurados. Los lenguajes de marcado como XML se utilizan a menudo para gestionar datos semiestructurados. Así, los ejemplos de datos semiestructurados suelen ser documentos XML y JSON. Algunos ejemplos de datos semiestructurados en tiempo real son los siguientes.

Datos semi-estructurados

Texto (XML, correo electrónico o mensajes de intercambio electrónico de datos (EDI)): Carece de estructura formal, pero contiene etiquetas que le dan una apariencia de estructura. La mayoría de los contenidos de las redes sociales son un tema candente de análisis hoy en día. Facebook, Twitter y otros ofrecen acceso a los datos a través de una interfaz de programación de aplicaciones (API) y los datos están disponibles en un formato semiestructurado.

Datos semi-estructurados

Registros de servidores web y patrones de búsqueda: Los registros del servidor web se utilizan para capturar la navegación de un usuario por un sitio web. Todo tipo de actividades, como la búsqueda, el consumo de contenidos o las compras, se registran detalladamente en los registros del servidor web. Proporcionan información valiosa sobre los intereses, las preferencias y el comportamiento de navegación de una persona. Estos datos también están disponibles en formato semiestructurado.

Datos semi-estructurados

JSON (JavaScript Object Notation): JSON es un formato de intercambio de datos semiestructurados. Aunque el nombre JavaScript está implícito, los demás lenguajes de programación tipo C también lo reconocen. La estructura JSON consiste en pares nombre/valor (objetos, tablas hash, etc.) o en una lista ordenada de valores (matrices, secuencias, listas). La estructura es intercambiable entre lenguajes y, por tanto, JSON es eficazmente adecuado para la transmisión de datos entre aplicaciones.