Activos de datos mineros

Por Mark O’Connor, CPA, CMA

Los científicos de datos y otros susurradores de datos se pueden utilizar para entrenar datos para su uso por inteligencia artificial y análisis para tomar mejores decisiones comerciales.



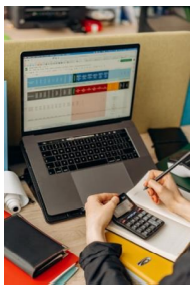
**Susurro de Datos**

Domar y aprovechar los activos de datos corporativos puede ser un desafío para los trabajadores de datos empresariales. Los científicos de datos y otros susurradores de datos se pueden utilizar para entrenar los datos para su uso por la inteligencia artificial y el análisis para tomar mejores decisiones comerciales. Los susurradores de datos pueden preparar los datos para ser utilizados para inteligencia artificial, análisis e impulsar mejores decisiones comerciales.

Cualquier líder empresarial con análisis de datos y recursos de TI (el equipo de Data Whisperers'**1**), ahora puede desbloquear, limpiar y recolectar datos comerciales para impulsar la generación de informes, la ejecución y el control de las operaciones. La ciencia de datos necesaria está cada vez más disponible a través de nuevas aplicaciones modernas, como Power BI, que pueden verse y funcionar de manera muy similar a una hoja de cálculo inteligente. Irónicamente, los activos de datos, a menudo un tesoro invaluable, no tienen un valor monetario que deba informarse en el balance. Los datos no son los únicos a este respecto. Otros activos digitales o capacidades comerciales comparten las mismas anomalías de reconocimiento del balance.

Los datos se almacenan en aplicaciones empresariales y en procesos empresariales asociados. Algunas aplicaciones ofrecen la opción de exportar datos, y algunas aplicaciones ya utilizan Business Intelligence (BI) para tener datos fácilmente disponibles. Para realizar extracciones periódicas, TI puede establecer descargas programadas de datos. Para crear más valor añadido y hacer que las extracciones de datos sean más ampliamente compartibles, se podrían crear catálogos, combinando los datos extraídos con otros datos y reutilizándolos para crear nuevos activos de datos. Estos activos de datos también se pueden utilizar de manera rentable para mejorar la base para la toma de decisiones de gestión.

Los datos extraídos se pueden organizar, almacenar y combinar en data warehouses (DW). Estos contenedores de datos se utilizan para actividades de informes de BI y, a veces, se convierten en lagos de datos sin procesar para análisis ad hoc. Los almacenes de datos se utilizan fundamentalmente para leer solo instantáneas de datos. El estándar es que no deben editarse directamente. Aunque el flujo de trabajo del proceso se puede establecer para realizar ediciones sistemáticas de las aplicaciones.



Los contenedores de datos DW están formando la base de los procesos de IA y demostrando ser altamente accesibles para la generación de informes y el uso en sistemas de autoservicio. Sin almacenes de datos efectivos, el valor de los datos corporativos puede enterrarse permanentemente en silos de aplicaciones corporativas. En las aplicaciones, estos datos no siempre están diseñados para una alta disponibilidad y un uso frecuente. Las aplicaciones de TI a menudo no están integradas con otras actividades, son difíciles de usar, no están disponibles para los usuarios comerciales y los susurradores de datos para una comprensión profunda. Los almacenes de datos están madurando y se están volviendo más completos y precisos. Por lo tanto, estos contenedores de datos se están convirtiendo en la única fuente de verdad para las consultas de las empresas y los clientes.

**¿Un active de valor incalculable o un pasivo?**

Las aplicaciones de TI que son anticuadas, con procesos en su mayoría manuales, automatización mínima o nula, baja funcionalidad y soporte deficiente del proveedor pueden convertirse fácilmente en una responsabilidad y una deuda técnica. Reunir datos y vincularlos por relaciones, por otro lado, agregará valor a los activos y evitará la deuda técnica. Sin embargo, este valor, buena o mala voluntad, no se divulgará de manera consistente y clara en el balance general bajo los principios contables actuales. Pero el valor de los datos se puede representar como parte de la capitalización de un activo del sistema de TI o, indirectamente, como un activo intangible en el balance. Esto también se puede hacer cuando una empresa ha sido vendida y si el nuevo precio de mercado excede el valor contable.

Los almacenes de datos están madurando y convirtiéndose

más completos y más precisos.

Una clave para el éxito en la construcción de un almacén de datos de Inteligencia de Negocios o el diseño de procesos de Inteligencia Artificial es comprender completamente los datos y las relaciones de datos, así como garantizar la calidad, integridad, integridad y organización de los datos. Para obtener una ventaja competitiva con BI o IA, la calidad de los datos también debe estar en un grado adecuado de granularidad.

**Con inteligencia**

La parte de "inteligencia" de BI e IA puede sugerir el uso de comparaciones significativas con otros participantes de la industria para facilitar y mejorar la comprensión de su propio negocio. Los datos de referencia externos se pueden combinar con datos corporativos para tener en cuenta los factores influyentes del entorno empresarial, como el historial o las predicciones de eventos meteorológicos, la demografía, las variaciones estacionales, la escasez o anomalías de suministro, la regulación, la salud pública,**2** y la política.

BI se utiliza para ilustrar tales comparaciones de negocios, para dashboardear indicadores clave, para monitorear controles y para brindar una mejor comprensión de los clientes, las partes interesadas y el mercado. Puede resaltar la eficacia de las operaciones comerciales y las nuevas estrategias. La Inteligencia Artificial puede utilizar datos de alta calidad para impulsar las operaciones y los procesos comerciales.

La forma en que se utilizan los datos para impulsar las decisiones influye en el grado de garantía de calidad que se requiere. Tener datos de IA meticulosos para este propósito es fundamental. Si bien las opiniones de auditoría de los estados financieros pueden proporcionar garantías de calidad para la presentación de informes sobre los estados financieros, es posible que no sean lo suficientemente granulares o precisas para tomar decisiones operativas.

Las herramientas de gestión de datos y las visualizaciones de BI pueden

ser útiles en la remediación, limpieza y establecimiento de una

base de datos de calidad para datos y metadatos.

**Precisión**

En el caso de ejecutar procesos de machine learning (ML), las reglas de guía o algoritmos**3** forman parte de un conjunto de procesos de automatización. El flujo de trabajo del algoritmo se utiliza para categorizar transacciones, hacer cálculos precisos y resolver problemas.

Para algunos entornos de datos, lograr una calidad de datos aceptable puede ser difícil sin la ayuda de científicos de datos altamente calificados, administradores de datos, analistas de TI y susurradores de datos que también tienen un profundo conocimiento de las operaciones comerciales. La ciencia de los datos puede ser complicada y laboriosa. Es importante preparar y organizar los datos para que sean "aptos para su uso". El conocimiento adquirido al explorar la ciencia de datos puede ayudar a marcar la diferencia.

Dependiendo de las estrategias de datos y los recursos disponibles, hacer análisis de datos y limpiar los datos al mismo tiempo puede ser el enfoque más efectivo para hacer un uso eficiente de los datos a largo plazo. Para al menos los activos de datos críticos y de mayor valor, la base, la base y las relaciones de los datos deben determinarse al seleccionar y agregar valor a esos activos de datos. Una vez que un activo de datos o un grupo de activos han sido descubiertos, limpiados y catalogados secuencialmente, pueden ser reportados a través de BI y reutilizados para impulsar procesos basados en IA.

Las herramientas de gestión de datos y las visualizaciones de BI pueden ser útiles para remediar, limpiar y establecer una base de datos de calidad de datos y metadatos. Los datos deben estar bien documentados para ser compartidos de manera efectiva. Los conceptos y categorías aplicables al área temática deben indicar las propiedades de los datos, las relaciones, los flujos de información y las ontologías de los datos.**4** Esta es el alma del negocio.

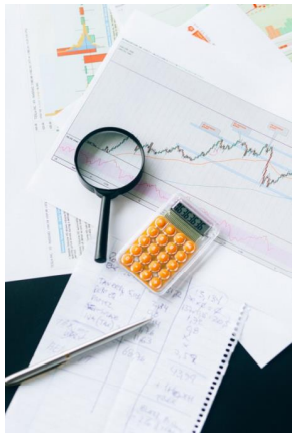
**Descubrimiento de Datos**

Salvaguardar el inventario de activos de datos de una empresa (su catálogo) abre la puerta a que las personas adecuadas descubran los datos de la empresa o la industria en el momento adecuado. Los catálogos pueden ser útiles para ayudar a mantener y asegurar los datos, así como para estandarizar los metadatos para facilitar el descubrimiento.**5** Muchos proyectos de IA fallan cuando esto no está correctamente implementado. A medida que las empresas se expanden y maduran y se dan cuenta del valor de sus propios activos de datos, están, o estarán, en condiciones de aprovechar esta información para BI e IA.

Cuando se pretende que los datos se compartan ampliamente, como se ve con algunos gobiernos y asociaciones de la industria, el catálogo o los metadatos,**6** también debe tener muestras disponibles de los registros de datos. Los conjuntos de datos pueden utilizarse para estudios de mercado y fuentes de referencia de información, por ejemplo, en el comercio de productos básicos de recursos entre países.

Las muestras de datos y los datos en sí se pueden distribuir a grupos cerrados o abiertos. La creación de controles de acceso sobre un catálogo de datos es esencial para proteger la información y garantizar la privacidad. No todos deberían ver los datos o cualquier parte de los datos, especialmente cuando se trata, por ejemplo, de intereses nacionales secretos, competencia en el mercado o privacidad.

Para algunos conjuntos de datos, partes de los datos podrían ser redactadas, mediante el cifrado específico de campo implantado**7** para prevenir y bloquear la exposición de violaciones de privacidad o secreto. Sin embargo, el conjunto de datos masivo o restante aún puede estar disponible siempre que no contenga datos referenciales que permitan la recreación de la información secreta o privada.



El Grupo Gartner tiene revisiones de software y recomendaciones para evaluaciones continuas de datos de calidad.**8** Las organizaciones que deseen pasar a una gestión de la calidad de datos más compleja y la preparación para el descubrimiento de datos pueden estar interesadas en:

o SAP Data Services.

o IBM InfoSphere utilizando IBM Watson.

o Talend de Informática.

o Productos basados en la nube como Atlan, que pueden ayudar a proporcionar herramientas que ayudan a garantizar la calidad.

**Los equipos pueden producir inteligencia como un susurrador de datos**

Las herramientas de análisis de datos sofisticadas, pero menos complicadas están evolucionando y cambiando el panorama y el uso de los datos. El descubrimiento de datos, la IA y la BI no deben ser el ámbito exclusivo de los recursos técnicos o los científicos de datos. Las herramientas de gestión de datos y los conjuntos de habilidades especializadas en arquitectos de negocios pueden ser útiles para documentar y describir explícitamente los flujos y ontologías previstos. Las herramientas poderosas pueden ayudar a moderar los riesgos de no usar recursos dedicados a científicos de datos.

Estamos aprendiendo rápidamente y a veces dolorosamente cómo

distinguir y discernir la calidad de los datos en la web

Microsoft Power BI, por ejemplo, puede ofrecer la integración de vistas de datos y empresas internas y externas. Mediante el uso de la metáfora de la vista similar a una hoja de cálculo, las herramientas pueden ayudar a proporcionar una vista inicial bajo el capó de los activos de datos de la empresa existentes, así como para los datos externos publicados. La capacidad de hoja de cálculo de la herramienta puede producir un prototipo visualmente de alta calidad y puede emplear iteraciones ágiles que forman la base de informes periódicos, como el producido por Tableau o IBM Cognos.

**Herramientas de descubrimiento de catálogos y gestión de metadatos**

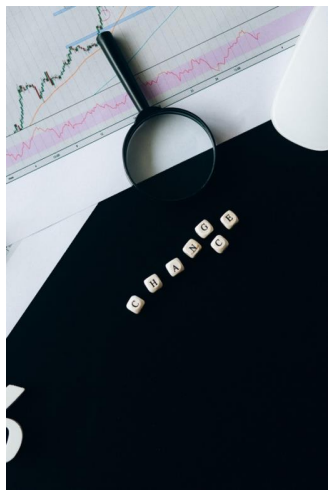
Las prioridades de datos de una organización a menudo se refuerzan fuertemente si se instala un director de Datos (CDO) en la C-Suite. La protección de los activos de datos se está convirtiendo cada vez más en un punto focal en los niveles superiores. Cada vez más, las empresas con un énfasis en la estrategia de datos están inventariando sus activos de datos en catálogos y creando capacidades de datos para convertir los datos en su ventaja. Los datos están cada vez más disponibles para el personal y las partes interesadas externas a través de la web. Las partes interesadas autorizadas pueden obtener una vista previa de los datos para determinar que son "aptos para su uso" y se pueden usar para crear iteraciones y análisis comerciales. Ser aprobado para el catálogo de datos de una empresa implica la seguridad de que los conjuntos de datos son de alta calidad, actuales y listos para ser utilizados. La calidad mejora cuando los comentarios de los usuarios y los comentarios sobre la calidad de un activo digital se adjuntan a los conjuntos de datos en el descubrimiento. Si bien la catalogación de datos sobre una base empresarial es nueva y está creciendo, ha evolucionado rápidamente a través de productos de aplicaciones que pueden ayudar a proporcionar herramientas para garantizar la calidad y la gestión de metadatos.

Algunos ejemplos de herramientas de descubrimiento de catálogos y gestión de metadatos disponibles públicamente**9** incluyen:

o Atlan, que actúa como un centro virtual privado para activos de datos que van desde tablas y paneles hasta modelos y código.

o Data World para el descubrimiento de datos, el gobierno ágil de datos y la información empresarial procesable.

o Google y las redes sociales se han convertido en lugares favoritos (o catálogos) para descubrir hechos (y teorías de conspiración) sobre personas, lugares, alimentos, medicamentos, técnicas y cosas que hacer. Estamos aprendiendo rápidamente y a veces dolorosamente cómo distinguir y discernir la calidad de los datos en la web.



Algunos gobiernos jurisdiccionales ofrecen datos como "Datos Abiertos". Algunos datos digitales departamentales, incluidos los registros, las métricas de servicio, los datos del censo nacional, los registros de aeronaves y las presentaciones financieras del mercado de valores de las empresas, son de uso gratuito. Algunos de estos datos son útiles en marketing y desarrollo de productos. Además, los datos obtenidos de los informes anuales y las presentaciones de las empresas públicas están siendo recopilados por los agregadores de datos, que luego los vuelven a empaquetar para la venta.

Algunos datos disponibles externamente (como XBRL, una estructura estandarizada XML), vienen con metadatos totalmente incrustados. Está disponible para su transmisión directa en modelos comerciales o de comercio patentados que pueden usarse para inducir órdenes de compra o venta automatizadas o semiautomatizadas impulsadas por IA y ML. En esta misma categoría de informes financieros y sostenibilidad, los países y continentes ofrecen servicios de portal. Servicios como EDGAR de la Comisión de Bolsa y Valores de los Estados Unidos, EDINET de Japón y el PAAS de la Autoridad Europea de Valores y Mercados tienen puntos únicos de acceso para datos digitales en varios formatos y para taxonomías como XBRL y JSON.

De hecho, es un noble llamado a ser reconocido como un susurrador de datos para una organización.

**1** An analogy in this is to succeed in gaining control of data and adding value. The term is derived from how a practitioner can understand the psychology and soul to mold or gain control and appear to succeed in achieving dominance.

**2** Nextstrain is an open-source project to harness the scientific and public health potential of pathogen genome data. It provides a continually updated view of publicly available data alongside powerful analytic and visualization tools. (<https://nextstrain.org/>).

**3** For example, a robotic corrective algorithmic rule to solve a data element quality issue that might cause significant inaccuracy or incompleteness. The automated rule, sometimes called a “Bot,” could retrieve a new quantity from some specified location and replace the data element with something that is more appropriate.

**4** The documentation or metadata should contain an information architecture framework, sometimes called an ontology. Ontology is a complex word and may be too complicated given that business has a long history in philosophy, in which it refers to the subject of existence. An ontology for data can be defined as an explicit methodical description or the narrative of a conceptualization.

**5** Data scientists, business managers or data workers need to review their datasets to ensure that they are “fit for use” and to learn about relationships between datasets, ontologies, data taxonomies, associated processes and other metadata.

**6** Metadata is information about the data, but not the data itself. Metadata is as important as the data. Metadata includes descriptions of data, features, attributes, facets, properties, fields and columns. It can classify and include information regarding legal ownership, the origins of Data/pedigree/lineage and provenance, and required permissions.

**7** Encryption of selected fields is not a common technology yet.

**8** Gartner Magic Quadrant for Data Quality Solutions, Published September 29, 2021. ID G00733919.

**9** Guido De Simoni, Alan Dayley, Mark Beyer, “Market Guide for Active Metadata Management” (Gartner: July 2021).