La gestión de datos

Introducción a la gestión de datos

Para que las organizaciones se centren en los datos con el fin de proporcionar valor a los clientes o tomar decisiones empresariales más informadas, necesitan recopilar una gran cantidad de datos de diferentes fuentes de datos, como flujos de clics, datos de registro, sistemas transaccionales y archivos planos, y almacenarlos en diferentes almacenes de datos dependiendo de las transformaciones que necesiten y su finalidad . Una vez que estos datos se almacenan en diferentes almacenes de datos, es necesario limpiarlos, transformarlos, organizarlos y unirlos a partir de diferentes fuentes de datos para proporcionar información más significativa a las aplicaciones posteriores, como los modelos de aprendizaje automático, para proporcionar recomendaciones de productos o buscar las condiciones del tráfico. También pueden ser utilizados por empresas o analistas de datos para extraer información empresarial significativa:

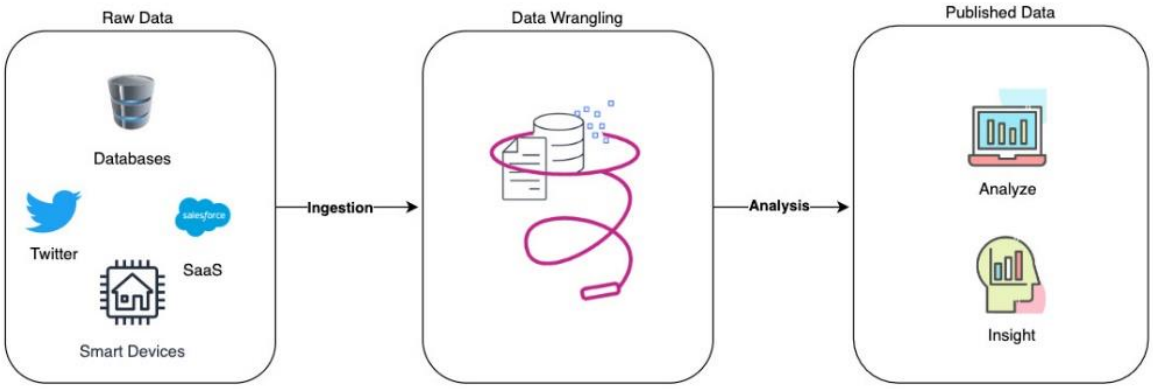

Figura 1: Canalización de datos

La regla 80-20 del análisis de datos

Cuando las organizaciones recopilan datos de distintas fuentes, al principio no sirven de mucho. Se calcula que los científicos de datos dedican aproximadamente el 80% de su tiempo a limpiar los datos. Esto significa que sólo el 20% de su tiempo se dedicará a analizar y crear ideas a partir del proceso de ciencia de datos:

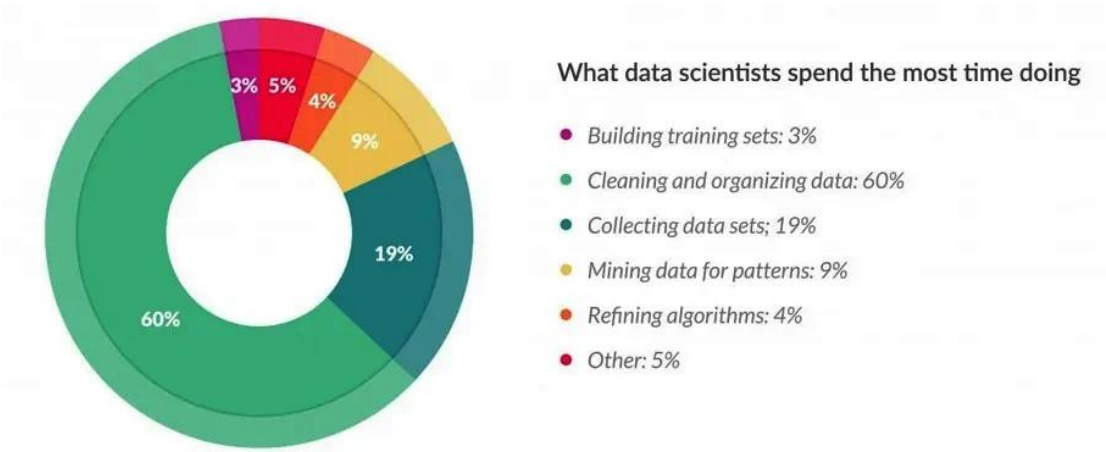

Figura 2: Distribución del trabajo de un científico de datos

Esto es lo que se conoce como *data wrangling* (gestión de los datos), vamos a aprender por qué es esencial, y los diversos beneficios que obtenemos de ella.

Ventajas de la gestión de datos

Usando la analogía del petróleo, cuando lo extraemos por primera vez, lo hacemos en forma de crudo, que no tiene mucha utilidad. Para que sea útil, tiene que pasar por una refinería, donde el crudo se introduce en una unidad de destilación. En este proceso de destilación, los líquidos y vapores se separan en componentes del petróleo llamados fracciones según sus puntos de ebullición. Las fracciones pesadas están en la parte inferior, mientras que las ligeras están en la parte superior, como se ve aquí:

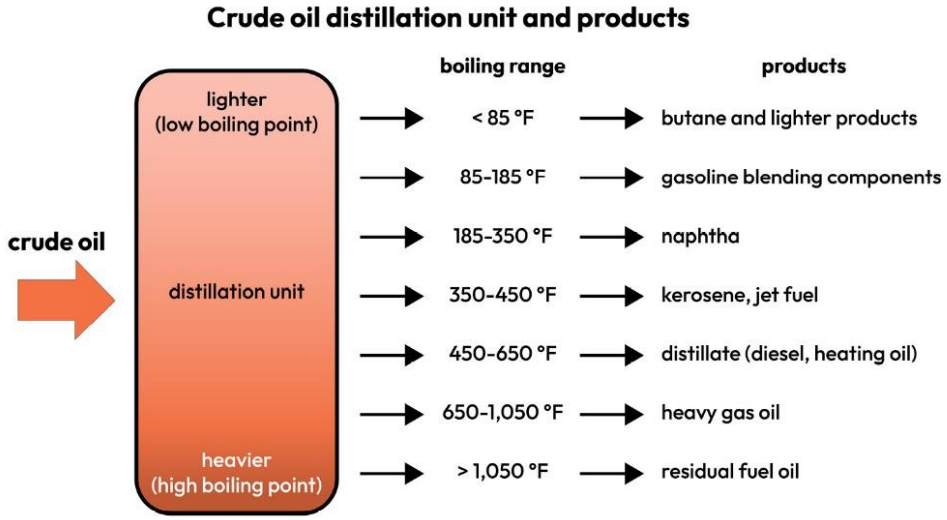

Figura 3: Procesamiento del petróleo crudo

La siguiente figura muestra la correlación entre el tratamiento del petróleo y el proceso de extracción de datos:

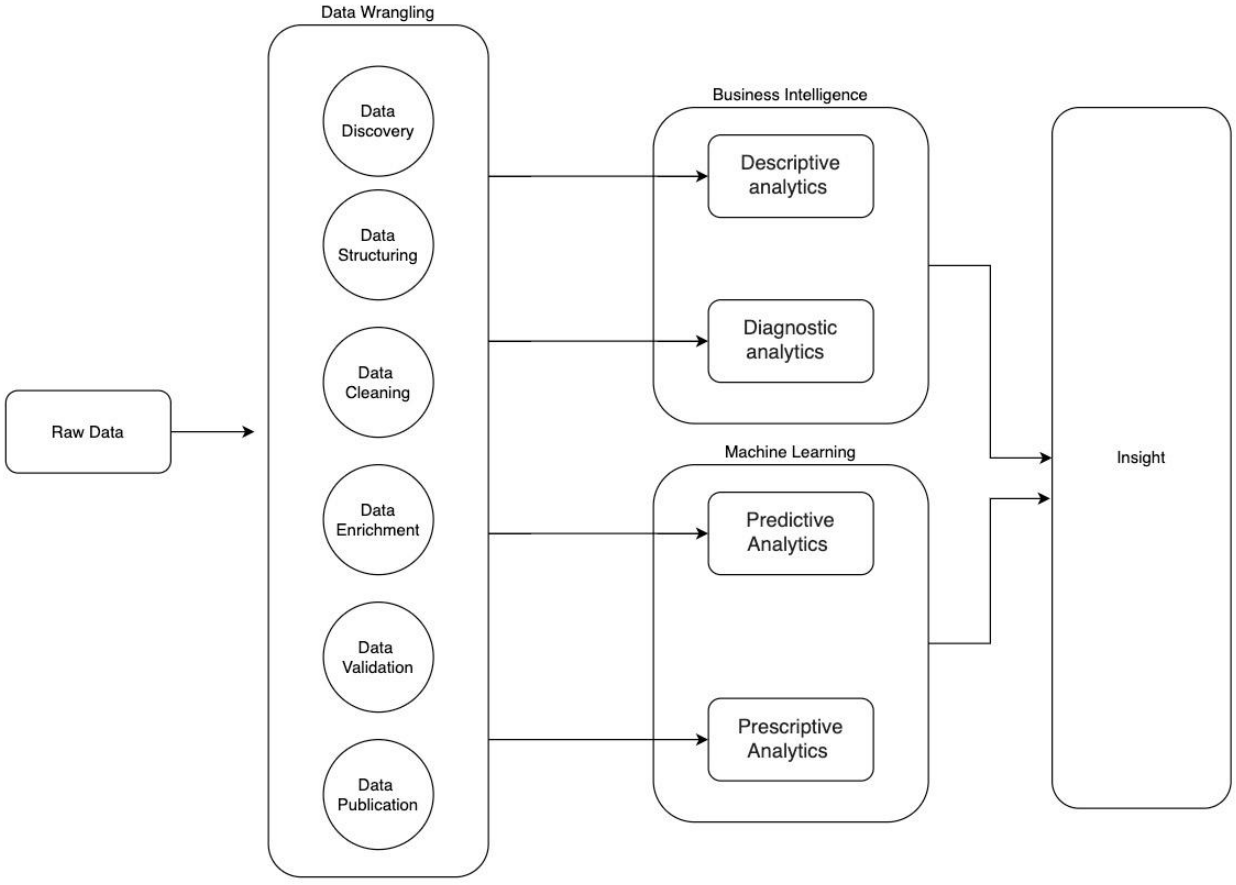

Figura 4: El proceso de gestión de datos

La gestión de datos ofrece muchas ventajas:

Mejora de la calidad de los datos: El reagrupamiento de datos ayuda a mejorar la calidad general de los datos. Implica identificar y tratar los valores que faltan, los valores atípicos, las incoherencias y los errores. Al abordar estos problemas, el reagrupamiento de datos garantiza que los datos utilizados para el análisis sean precisos y fiables, lo que permite obtener resultados más sólidos y fiables.

Ejemplo:

Una base de datos de ventas contiene precios negativos, fechas imposibles (como 32/13/2024) y celdas vacías en la columna *cantidad vendida*.

Durante la gestión de datos:

- Se eliminan o corrigen los precios negativos.
- Se corrigen las fechas erróneas.
- Se revisan o completan las cantidades faltantes

Mejora de la coherencia de los datos: A menudo, los datos proceden de diversas fuentes o tienen formatos distintos, lo que provoca incoherencias en las convenciones de nomenclatura, las unidades de medida o la estructura de los datos. La gestión de datos permite estandarizar y armonizar los datos, garantizando su coherencia en todo el conjunto de datos. La coherencia de los datos facilita la integración y la comparación de la información, lo que facilita un análisis y una interpretación eficaces.

Ejemplo:

En la columna *producto* aparecen valores como:

- "Laptop"
- "laptop"
- "Portátil"
- "PORTATIL"

Durante la gestión de datos:

- Se decide usar un único nombre, por ejemplo "Laptop".
- Se unifican mayúsculas/minúsculas y el idioma.

Mayor exhaustividad de los datos: Los datos incompletos pueden plantear problemas durante el análisis y la modelización. Los métodos de tratamiento de datos permiten gestionar los datos que faltan aplicando técnicas como la imputación, en la que los valores que faltan se estiman o rellenan a partir de la información existente. Al tratar adecuadamente los datos que faltan, la gestión de datos ayuda a garantizar un conjunto de datos más completo, reduciendo los posibles sesgos y mejorando la precisión de los análisis.

Ejemplo:

En un conjunto de datos de clientes, algunos registros no tienen edad registrada.

Durante la gestión de datos:

- Se imputan las edades faltantes usando la edad promedio de clientes similares.
- O se marca explícitamente como "Edad no informada".

Facilita la integración de datos: Las organizaciones suelen tener datos dispersos en múltiples sistemas y fuentes, lo que hace que la integración sea una tarea compleja. La gestión de datos ayuda a fusionar e integrar datos de diversas fuentes, lo que permite a los analistas trabajar con un conjunto de datos unificado. Esta integración facilita una visión holística de los datos, lo que permite realizar análisis exhaustivos y obtener perspectivas que podrían no ser posibles cuando se trabaja con datos fragmentados.

Ejemplo:

- Un sistema tiene datos de ventas.
- Otro sistema tiene datos de clientes.

- Ambos usan un *ID de cliente*.

Durante la gestión de datos:

- Se unen ambas bases usando el ID común.
- Se resuelven diferencias de formato (por ejemplo, fechas).

Transformación de datos racionalizada: El tratamiento de datos proporciona las herramientas y técnicas para transformar los datos brutos en un formato adecuado para el análisis. Esta transformación incluye tareas como la normalización, la agregación, el filtrado y el reformateo de los datos. Al agilizar estos procesos, el *data wrangling* simplifica la fase de preparación de los datos, lo que ahorra tiempo y esfuerzo a los analistas y les permite centrarse más en el análisis real e interpretar los resultados.

Ejemplo:

Las ventas están registradas por transacción individual.

Durante la gestión de datos:

- Se agrupan las ventas por mes.
- Se calcula el total mensual de ingresos.
- Se convierten monedas a una sola unidad.

Permite una ingeniería de características eficaz: La ingeniería de características implica la creación de nuevas variables derivadas o la transformación de variables existentes para mejorar el rendimiento de los modelos de aprendizaje automático. El tratamiento de datos proporciona una base para la ingeniería de características al preparar los datos de forma que permitan transformaciones significativas. Al realizar tareas como el escalado, la codificación de variables categóricas o la creación de términos de interacción, el procesamiento de datos ayuda a derivar características informativas que mejoran el poder predictivo de los modelos.

Ejemplo:

A partir de los datos originales:

- Fecha de compra
- Precio
- Cantidad

Durante la gestión de datos:

- Se crea una nueva variable: *ingreso total = precio × cantidad*.
- Se crea otra variable: *día de la semana* a partir de la fecha.

Admite la exploración y visualización de datos: La gestión de datos a menudo implica el análisis exploratorio de datos (AED), en el que los analistas obtienen información y comprenden los patrones de los datos antes del modelado formal. Al limpiar y preparar los datos, la gestión de datos permite una exploración eficaz de los mismos, ayudando a los analistas a descubrir relaciones, identificar tendencias y visualizar los datos mediante tablas, gráficos u otras representaciones visuales. Estos pasos exploratorios son cruciales para formular hipótesis, tomar decisiones basadas en datos y comunicar los conocimientos de forma eficaz.

Ahora que ya conocemos las ventajas de la gestión de datos, vamos a entender los pasos que hay que dar en el proceso de gestión de datos.

Ejemplo:

Tras limpiar y preparar los datos:

- Se construye un gráfico de ventas mensuales.
- Se visualiza un diagrama de dispersión entre precio y cantidad vendida.

Gracias a la gestión de datos:

- Los gráficos no muestran errores.

- Los patrones son claros y comprensibles.

Pasos de la gestión de datos

Al igual que el petróleo crudo, los datos sin procesar tienen que pasar por varias etapas de procesamiento para adquirir sentido. En esta sección, vamos a aprender el proceso de seis pasos que implica la gestión de datos:

1. Descubrimiento de datos
2. Estructuración de datos
3. Limpieza de datos
4. Enriquecimiento de datos
5. Validación de datos
6. Publicación de datos

Antes de empezar, es importante entender que estas actividades pueden o no tener que seguirse secuencialmente, o en algunos casos, puede saltarse cualquiera de estos pasos.
Además, hay que tener en cuenta que estos pasos son iterativos y difieren según el usuario, como analistas de datos, científicos de datos e ingenieros de datos.
Por ejemplo, el descubrimiento de datos para los ingenieros de datos puede diferir de lo que significa para un analista o un científico de datos:

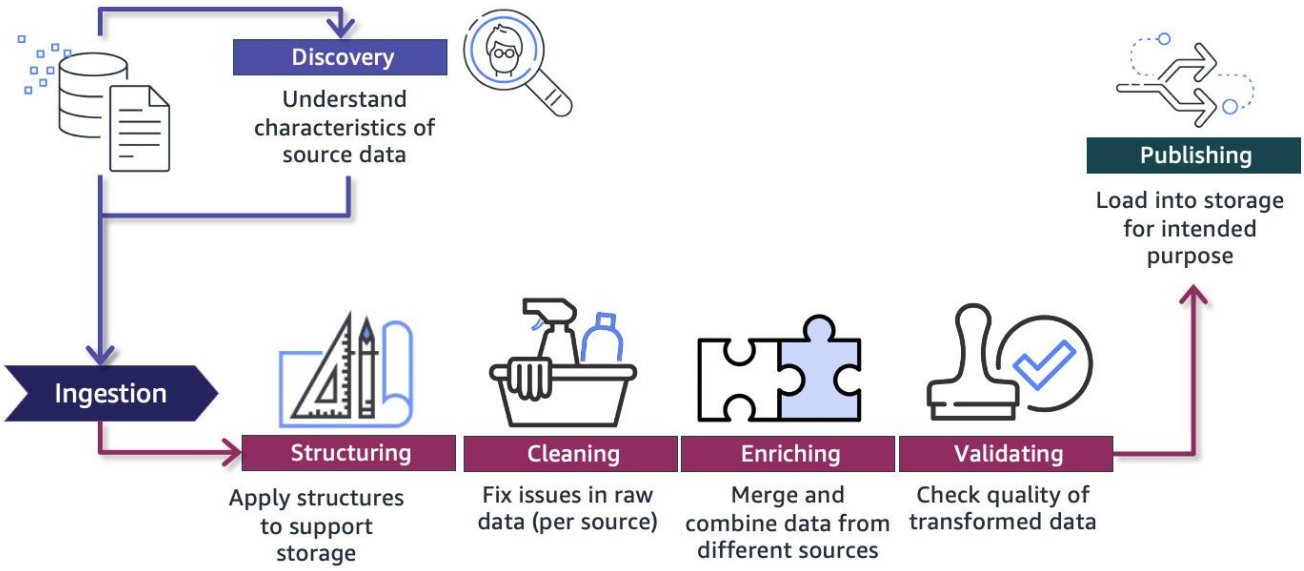

Figura 5: Etapas del proceso de análisis de datos

Empecemos a conocer estos pasos en detalle.

Descubrimiento de datos

El primer paso del proceso de búsqueda de datos es su descubrimiento. Es uno de los pasos más importantes del proceso. En el descubrimiento de datos, nos familiarizamos con el tipo de datos que tenemos como datos brutos, qué caso de uso queremos resolver con esos datos, qué tipo de relaciones existen entre los datos brutos, qué formato tendrán los datos, como *CSV* o *Parquet*, qué tipo de herramientas hay disponibles para almacenar, transformar y consultar estos datos, y cómo queremos organizar estos datos, por ejemplo, por estructura de carpetas, tamaño de archivo, particiones, etc. para facilitar el acceso.
Vamos a entenderlo con un **ejemplo**:

En este ejemplo, trataremos de entender cómo varía el descubrimiento de datos en función de la persona que ha de tratar con ellos. Supongamos que tenemos dos colegas, James y Jean. James es ingeniero de datos y Jean analista de datos, y ambos trabajan en una empresa de venta de coches.
Jean tiene que analizar las cifras de ventas de coches en un área determinado. Se ha puesto en contacto con James y le ha pedido los datos de la tabla de ventas del sistema de producción.
Este es el proceso de descubrimiento de datos para Jane (una ***analista de datos***):

- Jane tiene que identificar los datos que necesita para generar el informe de ventas (por ejemplo, datos de transacciones de ventas, datos de detalles de vehículos, datos de clientes, etc.).
- Jane tiene que encontrar dónde residen los datos de ventas (una base de datos, un archivo compartido, un CRM, etc.).
- Jane tiene que identificar cuántos datos necesita (de los últimos 12 meses, del último mes, etc.).
- Jane tiene que identificar qué tipo de herramienta va a utilizar (Amazon QuickSight, Power BI, etc.).
- Jane tiene que identificar el formato en el que necesita los datos para que funcionen con las herramientas que tiene.
- Jane tiene que identificar dónde quiere almacenar estos datos: en un lago de datos (Amazon S3), en su escritorio, en un archivo compartido, en un entorno *sandbox*, etc.

Este es el proceso de descubrimiento de datos para James (un ***ingeniero de datos***):

- ¿Qué sistema ha solicitado los datos? Por ejemplo, Amazon RDS, Salesforce CRM, ubicación SFTP de producción, etc.
- ¿Cómo se extraerán los datos? Por ejemplo, utilizando servicios como Amazon DMS o AWS Glue o escribiendo un script.
- ¿Cómo será el calendario? ¿Diaria, semanal o mensual?
- ¿Cómo será el formato del archivo? Por ejemplo, CSV, Parquet, ORC, etc.
- ¿Cómo se almacenarán los datos en el almacén proporcionado?

Estructuración de los datos

Para dar soporte a los casos de uso empresarial actuales y futuros con el fin de servir mejor a sus clientes, la organización debe recopilar cantidades de datos sin precedentes de diferentes fuentes de datos y en diferentes variedades. En la arquitectura de datos moderna, la mayoría de las veces, los datos se almacenan en **lagos de datos**, ya que un lago de datos permite almacenar todo tipo de archivos de datos, ya sean datos estructurados, datos no estructurados, imágenes, audio, vídeo o cualquier otra cosa, y tendrán diferentes formas y tamaños en su forma bruta. Cuando los datos están en bruto, carecen de una estructura definitiva, que es necesaria para almacenarlos en bases de datos o almacenes de datos, o para utilizarlos para construir modelos analíticos o de aprendizaje automático. En este punto, no están optimizados en cuanto a coste y rendimiento.
Además, cuando se trabaja con datos en flujo, como flujos de clics y análisis de registros, no todos los campos de datos (columnas) se utilizan en los análisis.
En esta fase de la gestión de datos, intentamos optimizar el conjunto de datos brutos para obtener ventajas en cuanto a costes y rendimiento, particionando y convirtiendo los tipos de archivo (por ejemplo, *CSV* en *Parquet*).
Una vez más, pensemos en nuestros amigos James y Jean.

- Para Jean, la analista de datos, la estructuración de datos significa que busca realizar consultas directas o almacenar datos en un almacén de memoria de una herramienta de BI, en el caso de Amazon QuickSight denominada capa SPICE, que proporciona un acceso más rápido a los datos.

- Para James, el ingeniero de datos, cuando extrae datos de un sistema de producción y busca almacenarlos en un lago de datos como Amazon S3, debe considerar cómo será el formato del archivo. Puede dividirlo por regiones geográficas, como municipio, provincia, región o estado, o por fecha, por ejemplo, año=YYYY, mes=MM y día=DD.

Limpieza de datos

El siguiente paso del proceso de manipulación de datos es la limpieza de estos. Los dos pasos anteriores nos dan una idea del aspecto de los datos y de cómo están almacenados. En la etapa de limpieza de datos, empezamos a trabajar con los datos en bruto para darles sentido y poder definir futuros casos de uso.

En la etapa de limpieza de datos, tratamos de dar sentido a los datos haciendo lo siguiente:

- Eliminación de columnas no deseadas, valores duplicados y relleno de columnas con valores nulos para mejorar la preparación de los datos.
- Validación de los datos para identificar los valores que faltan en columnas obligatorias como nombre, apellidos, número de seguro social, número de teléfono, etc.
- Validación o corrección del tipo de datos para optimizar el almacenamiento y el rendimiento.
- Identificación y corrección de valores atípicos
- Eliminación de datos basura o valores no deseados, como caracteres especiales

Tanto James como Jane pueden realizar tareas de limpieza de datos similares; sin embargo, su escala puede variar. En el caso de James, estas tareas deben realizarse para todo el conjunto de datos. Para Jane, puede que sólo tengan que realizarlas en el área geográfica de su interés, y la granularidad también podría variar.

Enriquecimiento de datos

Hasta el paso de limpieza de datos, trabajábamos principalmente con fuentes de datos individuales y les dábamos sentido para su uso futuro. Sin embargo, en el mundo real, la mayoría de las veces, los datos están fragmentados y almacenados en múltiples almacenes de datos dispares, y para dar soporte a casos de uso como la creación de soluciones de personalización o recomendación, necesitamos unir los datos de diferentes almacenes de datos.

Por lo tanto, en el paso de enriquecimiento de datos, construimos el proceso que mejorará los datos sin procesar con datos relevantes obtenidos de diferentes fuentes.

Validación de datos

Hay un término muy interesante en informática que se llama ***garbage in*, *garbage out* (GIGO)**. GIGO es el concepto de que los datos de entrada defectuosos (basura) producen resultados defectuosos.

En otras palabras, la calidad de la salida viene determinada por la calidad de la entrada. Por lo tanto, si introducimos datos defectuosos, obtendremos resultados inexactos.

En la fase de validación de datos, abordamos este problema realizando diversas comprobaciones de la calidad de los datos:

1) Validación comercial de la exactitud de los datos
2) Validación de la seguridad de los datos
3) Validación de la coherencia de los resultados en todo el conjunto de datos.
4) Validación de la calidad de los datos mediante comprobaciones de calidad de los datos como las siguientes
    i) Número de registros

ii) Valores duplicados
iii) Valores que faltan
iv) Valores atípicos
v) Valores distintos
vi) Valores únicos
vii) Correlación

Existe un gran solapamiento entre la limpieza y la validación de datos, y sí, hay muchas similitudes entre estos dos procesos. Sin embargo, la validación de datos se realiza en el conjunto de datos resultante, mientras que la limpieza de datos se realiza principalmente en el conjunto de datos sin procesar.

Publicación de datos

Una vez completados todos los pasos del tratamiento de datos, éstos están listos para ser utilizados en el análisis con el fin de resolver problemas empresariales.

Por tanto, el último paso consiste en publicar los datos para el usuario final con el acceso y los permisos necesarios.

En este paso, nos concentramos principalmente en cómo se exponen los datos al usuario final y dónde se almacenan los datos finales, es decir, en una base de datos relacional, un almacén de datos, zonas limpias o de usuario en un lago de datos, o a través del Protocolo Seguro de Transferencia de Archivos (SFTP), etc.

La elección del almacenamiento de datos depende de la herramienta a través de la cual el usuario final desea acceder a los datos. Por ejemplo, si el usuario final desea acceder a los datos a través de herramientas de BI como Amazon QuickSight, Power BI, Informatica, etc., un almacén de datos relacional será la opción ideal. Si es un científico de datos quien accede a los datos, lo ideal sería almacenarlos en un almacén de objetos.

Prácticas recomendadas para la gestión de datos

Hay muchas formas y herramientas disponibles para realizar la gestión de datos, dependiendo de cómo se realice y quién la realice. Por ejemplo, si está trabajando en casos de uso en tiempo real, como la recomendación de productos o la detección de fraudes, la elección de la herramienta y el proceso para llevar a cabo la gestión de datos serán muy diferentes en comparación con la creación de un panel de inteligencia empresarial (BI) para mostrar las cifras de ventas.

Independientemente del tipo de casos de uso que busque resolver, se pueden aplicar algunas prácticas recomendadas estándar en cada caso que le ayudarán a facilitar su trabajo como gestor de datos.

Identificar el caso de uso empresarial

Es recomendable que decidas qué servicio o herramienta quieres utilizar para la gestión de datos antes de escribir una sola línea de código. Es muy importante identificar el caso de uso empresarial, ya que esto sentará las bases para los procesos de gestión de datos y facilitará la tarea de identificar los servicios que desea utilizar. Por ejemplo, si tienes un caso de uso empresarial como el análisis de datos de RR.HH. para organizaciones pequeñas, en el que sólo necesita concatenar algunas columnas, eliminar algunas columnas, eliminar duplicados, eliminar valores NULL, etc. de un pequeño conjunto de datos que contiene 10.000 registros, y sólo unos pocos usuarios van a acceder a los datos procesados, entonces no necesita invertir una tonelada de dinero para encontrar una herramienta de procesamiento de datos sofisticada disponible en el mercado; simplemente puedes utilizar hojas de Excel para tu trabajo.

Sin embargo, cuando se trata de un caso de uso empresarial, como el procesamiento de datos de siniestros que recibe de diferentes socios, en el que necesitamos trabajar con archivos semiestructurados como JSON, o conjuntos de datos no estructurados como archivos XML para extraer sólo algunos datos de los archivos, como el ID del siniestro y la información del cliente, y deseamos realizar procesos complejos de gestión de datos, como uniones, búsqueda de patrones mediante expresiones regulares, etc., ya deberemos escribir programas o suscribirnos a cualquier herramienta de nivel empresarial para nuestro trabajo.

Identificación de la fuente de datos y obtención de los datos adecuados

Tras identificar el caso de uso empresarial, es importante identificar qué fuentes de datos se necesitan para resolverlo. Identificar esta fuente nos ayudará a elegir qué tipo de servicios son necesarios para traer los datos, la frecuencia y el almacenamiento final. Por ejemplo, si se quiere crear una solución de detección de fraudes con tarjetas de crédito, es necesario aportar datos de transacciones con tarjetas de crédito en tiempo real; incluso la limpieza y el procesamiento de los datos deben hacerse en tiempo real. La inferencia de aprendizaje automático también debe ejecutarse sobre datos en tiempo real.

Del mismo modo, si está creando un cuadro de mando de ventas, es posible que tengamos que introducir datos de un sistema CRM como Salesforce o un almacén de datos transaccionales como Oracle, Microsoft SQL Server, etc.

Tras identificar las fuentes de datos adecuadas, es importante incorporar los datos correctos de estas fuentes de datos, ya que ayudarán a resolver los casos de uso empresarial y facilitarán el proceso de gestión de datos.

Identificación del público

Un aspecto importante de la gestión de datos es la identificación del público. Conocer a nuestro público nos ayudará a identificar qué tipo de datos desean consumir. Por ejemplo, los equipos de *marketing* pueden tener diferentes requisitos para la gestión de datos en comparación con los equipos de ciencia de datos o los ejecutivos de negocios.

Esto también nos dará una idea de dónde publicar los datos: por ejemplo, un equipo de científicos de datos puede necesitar datos en un almacén de objetos como Amazon S3, los analistas empresariales pueden necesitar datos en archivos planos como CSV, los desarrolladores de BI pueden necesitar datos en un almacén de datos transaccionales y los usuarios empresariales pueden necesitar datos en aplicaciones.

AWS Glue DataBrew

Lanzada en 2020, AWS Glue DataBrew es una herramienta visual de preparación de datos que facilita la limpieza y normalización de datos para que pueda prepararlos para el análisis y el aprendizaje automático. La interfaz de usuario visual que proporciona este servicio permite a los analistas de datos sin experiencia en codificación o *scripting* llevar a cabo todos los aspectos de la manipulación de datos. Viene con un amplio conjunto de acciones comunes de transformación de datos predefinidas que pueden simplificar estas actividades de manipulación de datos. Al igual que cualquier software como servicio (SaaS) (https://en.wikipedia.org/wiki/Software_as_a_service), los clientes pueden empezar a utilizar la interfaz de usuario web sin necesidad de aprovisionar ningún servidor y sólo tienen que pagar por los recursos que utilicen.