La Gestión de Datos en la era de la Ciencia de Datos

Autor
Martín Vladimir Alonso Sierra Galvis

Gestión de Datos, Maestría en Ciencia de Datos
Pontificia Universidad Javeriana Cali

Versión 1.0
Santiago de Cali, febrero de 2023

Tabla de contenido

Los datos

¿Qué es un dato?

Se conoce como dato a toda aquella medida recolectada a partir de una observación y que actúa como una fuente de información. Ahora bien, esta es la definición general, pero si la vemos desde la perspectiva tecnológica, que es a grandes rasgos la que nos va a interesar más, los datos son cantidades, caracteres o símbolos almacenados en medios magnéticos y transmitidos en forma de señales eléctricas por un computador. Existen muchos tipos de datos y muchas formas de representar esos datos, por ejemplo, texto, números o multimedia. En general, se pueden dividir los datos en dos categorías principales: datos cuantitativos y datos cualitativos. Los datos cualitativos se pueden expresar con números y pueden ser contados o comparados mientras que, los datos cualitativos, son generalmente datos en formato texto que pueden representar a su vez categorías.

Ejemplos de datos pueden ser el número de alumnos que matricularon determinada materia en el presente semestre, el número de copias de un libro que fueron vendidas en cierta librería durante el último mes, la temperatura que se registró en una ciudad en un día de la semana. la última localización registrada por un usuario en su dispositivo móvil e incluso un texto, un audio o un video compartido por un usuario en una red social.

¿Para qué sirven los datos?

Aunque el curso está orientado al área tecnológica, tenemos que tener muy presente que los datos han existido desde mucho antes que aparecieran las computadoras y los dispositivos electrónicos. Civilizaciones antiguas ya llevaban registros, por ejemplo, del número de comida almacenada en bodega, datos que les servían para, posiblemente, evitar hambrunas en tiempos de sequía. Grandes pensadores y científicos recolectaban datos de sus experimentos y realizaban análisis de los mismos que les permitían formular respuestas a muchas de sus teorías. Incluso los datos se han usado para contar historias. Hoy por hoy, los datos son considerados uno de los mayores bienes intangibles de las empresas y organizaciones. Su relevancia actual es tal que se le ha denominado como el oro o el petróleo del siglo

XXI. La razón: gracias al desarrollo tecnológico, la cantidad de datos que recolecta y almacena una empresa es inmensa. Tal cantidad le permite hacer un análisis, por ejemplo, del comportamiento o gustos de sus clientes y, en función de ello, crear campañas publicitarias que le permitan tanto retener clientes antiguos como captar nuevos. Así mismo la empresa puede visualizar tendencias de sus productos, saber cuáles productos pueden ser exitosos y cuáles no. Todo esto representa mejoras en procesos y, por ende, ganancias monetarias. Pero no solamente se beneficia el área comercial; también lo hacen otras áreas completamente diferentes, como el área de la investigación. Grandes cantidades de datos son utilizados para entrenar algoritmos de Machine Learning, una de las bases de la Inteligencia Artificial, tecnología que puede representar una gran ventaja para la raza humana. Esto porque los agentes inteligentes son capaces de realizar tareas y cálculos en tiempos mucho menores de los que puede llegar a alcanzar un humano. Y a mayor cantidad de datos, mayor es el perfeccionamiento de estos algoritmos. En fin, en nuestra era moderna, los datos son sumamente importantes para la mayor parte de los procesos que realizan los humanos diariamente. Pero, el mayor provecho no se obtiene a partir de los datos en bruto, sino de los datos procesados, de los datos convertidos en información. Y el reto está en seleccionar y procesar adecuadamente los datos correctos entre todo el gigantesco número de datos existentes. Es en este escenario en el que la Gestión de los Datos en la Ciencia de Datos se vuelve fundamental.

¿Son los datos equivalentes a la información?

Es muy común que las personas ajenas al tema de los datos e incluso profesionales del área confundan el término dato con el término información. Desde ya debemos tener muy claro que no significan lo mismo. Mientras los datos son medidas que se recolectan y almacenan en bruto a partir de observaciones del mundo real para luego ser transmitidos, la información es el resultado de tomar dichos datos y, a partir de un proceso de análisis, obtener un significado de ellos. Además de esto, para que los datos se conviertan en información, deben estar enmarcados en un contexto determinado. Pongamos como ejemplo el siguiente número.

156435765

Cuando vemos este número y tratamos de analizarlo, no nos dice nada. Es, simplemente, un número más que, para nosotros, significa cualquier cosa que nos podamos imaginar. Esto ocurre porque no hay un contexto específico en el que se encuentre enmarcado. Ahora si al número le añadimos el símbolo COP $ de la siguiente forma.

COP $2564357650

Entonces tenemos un primer contexto, muy pequeño eso sí, pero que nos está indicando que el número es una cifra monetaria en pesos colombianos. Pero, ¿una cifra mentaria relacionada con qué? Aquí el contexto sigue sin ser suficiente. Ahora bien, si además de esto nos dicen que dicha cifra corresponde al total de ingresos de una empresa durante un mes, entonces, en este punto ya comenzamos a tener información y no simplemente un dato. Para sacar el mayor provecho, el contexto debería ser más detallado: total de ingresos de Oracle en el mes de mayo del año 2022, por poner un ejemplo.

Los datos son significativos cuando se pueden reconocer, están completos y expresan una idea clara de forma que se pueda obtener información de ellos. Por esto, es importante, tal como se mencionó en la sección anterior, seleccionar y separar aquellos datos que pueden brindar un significado de aquellos que no tienen relevancia, esto con el objetivo de generar un mensaje congruente. Una vez que tenemos en nuestra mano la información, podemos generar conocimiento.

La Gestión de Datos

¿Qué es la Gestión de Datos?

A nivel general, la Gestión de Datos es un campo en el área de los datos que se encarga de la recopilación, almacenamiento y mantenimiento de los datos. Su principal objetivo es ayudar a crear estrategias para optimizar el uso de los datos con el fin de tomar decisiones que ayuden a maximizar beneficios. Por ejemplo, una gestión de datos sólida puede ayudar a una empresa a ajustar sus recursos y procesos de acuerdo a su entorno cambiante a través de una mejora en la toma de decisiones. La Gestión de Datos comprende:

1. Generar, almacenar y organizar los datos en la infraestructura de datos, por ejemplo, una base de datos.
2. Establecer que los datos tengan una alta disponibilidad para su uso.
3. Planificar la recuperación ante posibles desastres, sean naturales o causados por el ser humano.
4. Controlar el acceso a los datos.
5. Verificar y eliminar datos de acuerdo a normativas vigentes.

Un profesional encargado de la gestión de los datos debe asegurar la calidad de los mismos y para ello debe tener en cuenta elementos como el origen o fuente de los datos; el proceso y lugar de almacenamiento; la organización de los datos almacenados; la disponibilidad de los datos para ser utilizados; la seguridad de los datos, es decir, que los datos no sean manipulados ni eliminados por personas o agentes no autorizados. Así mismo, debe tener conocimiento sobre cómo preparar y limpiar datos con el propósito de corregir errores en los mismos; y sobre herramientas para realizar procesos de extracción, transformación y carga de datos que le permita la consolidación de fuentes adicionales de datos como, por ejemplo, almacenes de datos para la generación de informes. En resumen, el profesional debería ser capaz de administrar lo que se conoce como el **ciclo de vida** de los datos, término que hace referencia a todas las fases por las que pasan los datos, desde que se originan hasta que se eliminan.

Desafíos de la Gestión de Datos

Con el auge del Big Data, término que veremos más adelante en este curso, la Gestión de Datos se enfrenta a nuevos desafíos como:

1. El volumen de los datos es cada vez mayor y sus formatos son muy diferentes por lo que se dificulta el mantenimiento y reconocimiento de los mismos.
2. Relacionado con el punto anterior, al provenir de diferentes fuentes, los datos se tornan más complejos. Esto porque pueden estar en diferentes formatos, lo que dificulta su integración y consolidación en las estructuras de almacenamiento. Los datos que no se pueden consolidar se convierten en datos inservibles.
3. Al ser el volumen de datos tan grande, se debe tener presente la infraestructura en la que se almacenarán y la organización de los datos en ella. Todo esto teniendo en cuenta los diferentes formatos de las fuentes de datos, lo que muy seguramente requiere un proceso de transformación sólido para evitar inconsistencias.
4. La cantidad de datos puede hacer que su acceso se torne menos eficiente, esto debido también a su organización. Entre más estricto es el orden de los datos en la infraestructura de almacenamiento, más complejo es su acceso.
5. Al ser tantos los datos almacenados se puede incurrir fácilmente en problemas de seguridad y violaciones a las normativas de privacidad de datos lo que puede resultar en multas y sanciones.

Todos estos desafíos han hecho que Ingenieros de Software e Ingenieros de Datos desarrollen herramientas que permiten la gestión de grandes volúmenes de datos. Desde Sistemas Gestores de Bases de Datos, pasando por herramientas de limpieza y ETL, hasta herramientas para realizar procesamiento distribuido de datos como Hadoop, el objetivo de la Gestión de Datos siempre será el aseguramiento de la calidad de los mismos en todas las fases de su ciclo de vida de forma que sean una herramienta poderosa para la generación de información y la toma de decisiones.

Gestión de Datos en Ciencia de Datos

La definición que hemos visto hasta el momento de Gestión de Datos es la general, misma que está enfocada a profesionales como Administradores de

Bases de Datos o Ingenieros de Datos. Estos roles cuentan con lo que se denomina **gobernanza de datos**, son dueños de los datos y pueden ejecutar procesos de tipo CRUD (o **C**reate, **R**ead, **U**pdate y **D**elete), es decir, pueden crear datos, leer datos, actualizar datos y eliminar datos directamente de la infraestructura donde se encuentran almacenados. Además de esto, son responsables del diseño e implementación de dicha infraestructura. Como profesionales en Ciencia de Datos, nuestro rol es un poco distinto. Nosotros **no somos los dueños** sino, de cierta forma, consumidores de los datos. No nos interesa ni diseñar ni tampoco implementar infraestructuras "originales" de almacenamiento como una base de datos. Lo que nos interesa es acceder a los datos para realizar ciertos procesos como transformarlos, limpiarlos y analizarlos para, finalmente, convertirlos en insumo de diversos algoritmos y herramientas que permitan generar conocimiento para responder a diversos problemas. Es por esta razón que, generalmente, no trabajamos solos sino acompañados por otros profesionales como Ingenieros de Datos, quienes serán los encargados de proveernos la información necesaria relacionada a los datos con los que queremos trabajar. Pero entonces ¿qué tiene que ver la Gestión de Datos en la Ciencia de Datos? La respuesta radica en que, aunque nuestra labor principal tiene que ver más con el acceso a los datos y su procesamiento, lo cierto es que en la consecución de dicha labor tendremos que realizar algunas tareas propias de la Gestión de Datos. Por ende es importante tener conocimiento sobre algunos conceptos como:

1. Lenguajes de consulta de datos. Nuestra tarea más importante será la consulta de datos, es decir, la obtención de datos desde una fuente de datos o infraestructura de almacenamiento. Normalmente, dicha consulta se realiza con un lenguaje. El ejemplo más conocido es el lenguaje SQL.
2. Fuentes de datos. Nuestro rol no está enfocado a la generación de datos pero sí debemos conocer las diferentes fuentes u orígenes de datos. Posiblemente nos tocará consultar información de fuentes tan diferentes como una base de datos relacional, una base de datos NoSQL o un dataset en formato CSV.
3. Modelos lógicos de datos. Aunque estos modelos se utilizan en las etapas de diseño y creación de infraestructuras de datos, como científicos de datos deberíamos ser capaces de entenderlos, pues serán una herramienta muy útil cuando queramos acceder a los datos:

Nos ayudarán a identificar cómo están organizados y distribuidos los datos en determinada infraestructura de almacenamiento.

4. Estrategias para realizar preparación y limpieza de datos. Los datos almacenados no están exentos de errores. Como científicos de datos debemos ser capaces de identificar los problemas y solventarlos para que los datos que utilicemos nos brinden información verídica y confiable.
5. Estrategias para realizar procesos de extracción, transformación y carga de datos (ETL). Esto es fundamental para poder realizar análisis pues, normalmente, las infraestructuras de almacenamiento no están pensadas para analizar datos. Aquí también es importante conocer sobre modelos lógicos pues, aunque no estaremos encargados de implementar la infraestructura "original" de almacenamiento, si deberemos implementar la infraestructura de almacenamiento para analítica como lo son los almacenes de datos.

Para formarnos como científicos de datos, todos estos conceptos relacionados a la Gestión de Datos se complementarán con otros más avanzados como Análisis de Datos o Machine Learning, temas que se verán a profundidad en los siguientes semestres de la Maestría.

Referencias

1. Australian Bureau of Statistics. *Data*. [En línea]: disponible en https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/data. [Accedido en 2023].
2. Innovación Aprendizaje. *Video: los datos y la información*. Octubre de 2016. [En línea]: disponible en https://www.youtube.com/watch?v=IEx7KVfOqfM. [Accedido en 2023].
3. noupe. *What Is Data Management?* Febrero de 2020. [En línea]: disponible en https://www.noupe.com/inspiration/tutorials-inspiration/what-is-data-management.html. [Accedido en 2023].
4. Oracle. *¿Qué es la gestión de datos?* [En línea]: disponible en https://www.oracle.com/co/database/what-is-data-management/. [Accedido en 2023].
5. Red Hat. *¿Qué es la gestión de los datos?* Mayo de 2022. [En línea]: disponible en https://www.redhat.com/es/topics/data-services/what-is-data-management. [Accedido en 2023].
6. Shen, S. *What is Data? And why we need data management, data literacy and data analytics.* Noviembre de 2020. [En línea]: disponible en https://towardsdatascience.com/what-is-data-ade94b37204a. [Accedido en 2023].
7. U of G Library. *What is Data?* Abril de 2019. [En línea]: disponible en https://www.youtube.com/watch?v=pg12U1BAnoA. [Accedido en 2023].
8. University of Houston Libraries. *What is Data?* Septiembre de 2021. [En línea]: disponible en https://www.youtube.com/watch?v=WnP6jDvupiY. [Accedido en 2023].