Unidad N°1 Panorama General sobre Data Science

Sitio: Centro de E-Learning - UTN.BA
Curso: Curso de Data Science
Libro: Unidad N°1 Panorama General sobre Data Science

Imprimido por: Virginia Marich
Día: Wednesday, 27 de December de 2023, 09:44

Descripción

Tabla de contenidos

1. Introducción

¡Bienvenidos a la emocionante primera unidad del curso de Data Science! En la actualidad, el mundo de los datos es un universo en constante evolución. El volumen y la complejidad de la información generada en diferentes sectores hace que el papel del Data Scientist sea aún más crucial. Como futuros profesionales en este área, aprenderán a manejar grandes conjuntos de datos, aplicar técnicas avanzadas de análisis y extraer información valiosa para la toma de decisiones estratégicas en las organizaciones.

En esta primera unidad, nos sumergiremos en el mundo de los datos y descubriremos qué son los datos, qué tipos existen y cómo obtenerlos. También exploraremos el ciclo de vida de los datos y el flujo de trabajo que se sigue para obtener conclusiones enriquecedoras a través de ellos.

Además, conocerán los diferentes roles que existen en el mundo de los datos, explorando puestos en LinkedIn y las habilidades básicas que se necesitan para ejercer cada rol. Por supuesto, nuestro enfoque principal será el rol del Data Scientist, en el que les enseñaremos las habilidades y software que deben conocer para convertirse en un verdadero experto en la materia. También veremos ejemplos de actividades laborales diarias del rol.

Finalmente, les recordamos que en el mundo laboral, el trabajo en equipo es fundamental, y por eso, las instancias evaluativas de este curso están diseñadas para grupos de 3 a 6 estudiantes, tanto en las tareas como en el proyecto final. Como en un trabajo real, no pueden elegir con quién trabajar, por lo que aprenderán a trabajar con diferentes personas y desarrollarán habilidades importantes para el éxito en el mercado laboral actual. ¡Estamos emocionados de tenerlos aquí y esperamos que disfruten esta primera unidad del curso!

2. ¿Qué son los datos? ¿Qué tipos de datos hay? ¿Cuál es el flujo de trabajo que siguen los datos para obtener conclusiones enriquecedoras a través de ellos?

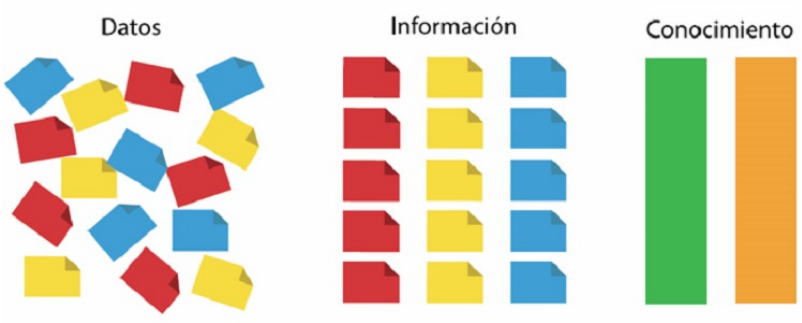
¿Qué son los datos?

El matemático británico Clive Humby afirmó en 2006 que "los datos son el nuevo petróleo", haciendo referencia a que los datos en el siglo XXI son como el petróleo en el siglo XVIII: un activo de gran valor y sin explotar.

Los datos pueden definirse como una representación simbólica de una característica o atributo que describe la realidad. Se dice que los datos son simplemente "piezas de información" y no dicen nada por sí solos, ya que necesitan un contexto para convertirse en información útil.

Por ejemplo, si te dijera: "35, 23, 21, 43", no habrías aprendido nada nuevo, ya que solo te habría dado datos. Sin embargo, si te digo que estos son los números de edad de cuatro alumnos de este curso, entonces ese conjunto de datos se convierte en información útil una vez que se le da contexto y por lo tanto el consumo de información produce conocimiento.

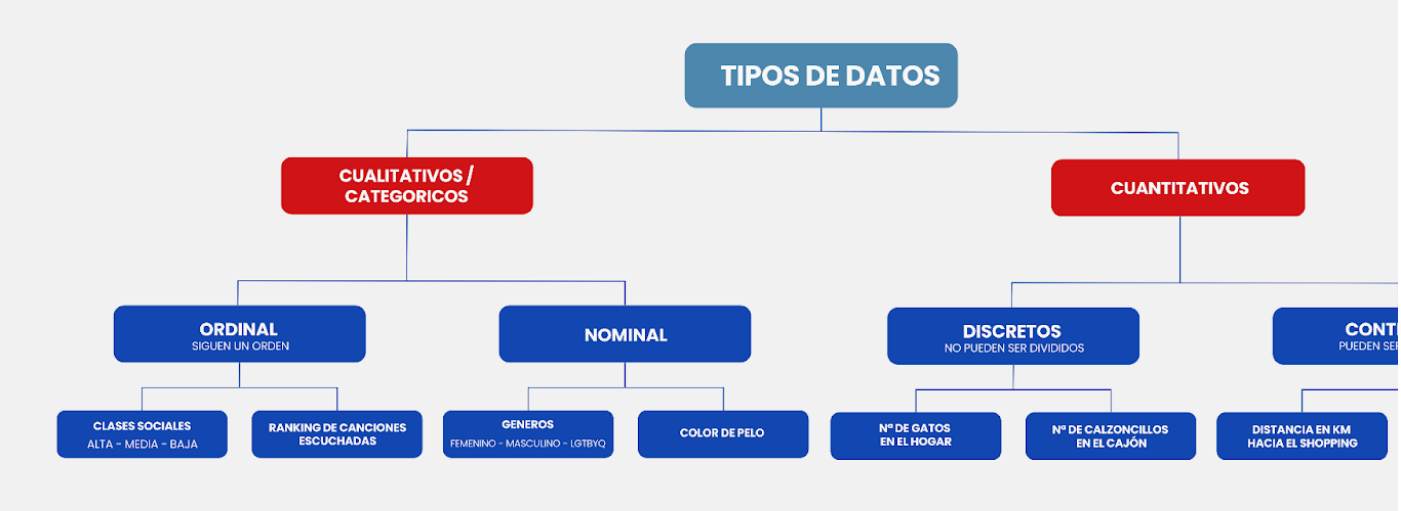
De manera similar, cuando somos pequeños y aprendemos palabras o letras, al principio sólo asimilamos "datos", pero cuando nos enseñan que "Sebastián" es el nombre de mi hermano, recién en ese momento adquirimos información útil. De manera similar, en el ámbito empresarial, si se nos proporciona una base de datos sin contexto, solo podríamos considerarla como datos, pero al proporcionar el contexto adecuado, podemos identificar esa base de datos como información útil y utilizarla para tomar decisiones informadas.



2.1. ¿Qué tipos de datos hay y cómo los obtengo?

Con los ejemplos dados, podemos clasificar coloquialmente los datos de dos maneras:

1. Datos cuantitativos: son numéricos y representan una cantidad medida, como un resultado de medición, recuento o cálculo matemático. Según las matemáticas pueden ser enteros, decimales, reales o imaginarios.
2. Datos cualitativos: son descriptivos y representan atributos o características, como nombres, títulos, letras o colores. Aunque también pueden ser números, como en el caso de definir un ID oposición de una persona en una base de datos como "001", lo que describe una característica y no una cantidad medida. Lo mismo ocurre si digo que "22" es el día de mi cumpleaños, que también describe una característica y no una cantidad medida.



La distinción entre los datos ordinales y nominales es simplemente una subclasificación. Los datos ordinales representan atributos que pueden ordenarse jerárquicamente, como las clases sociales o rankings, mientras que los datos nominales simplemente clasifican, como Femenino y Masculino.

En matemáticas, los datos cuantitativos tienen un fondo conceptual más complejo. Los datos discretos son números enteros que no pueden tomar valores fraccionarios. Por ejemplo, el número de visualizaciones de una publicación puede ser 2 o 3, pero no puede ser 2.5. Por otro lado, existen datos continuos que representan medidas que pueden tomar cualquier valor, como el peso del pan, que puede ser 1 kg y 300 gramos, y no solo 1 kg o 2 kg.

2.2. ¿Cómo se encuentran los datos en la realidad?

Ahora teniendo en cuenta estos conceptos, es importante reflexionar sobre nuestra percepción de los datos. Si les pido que cierren los ojos y piensen en datos, es probable que la mayoría visualice una tabla o un archivo de Excel. Sin embargo, en la vida real, especialmente en las dos fuentes principales de generación de datos como son las organizaciones y el internet, la mayoría de los datos no tienen este formato.

Si retomamos la idea de imaginar los datos como una tabla nos estamos refiriendo únicamente a un tipo de estructura de datos conocido como "datos estructurados".

Conceptualmente, los datos estructurados son aquellos que tienen una estructura y formato definido. Los datos estructurados son datos organizados en una estructura predefinida, con una disposición específica y con relaciones bien definidas entre los distintos elementos. Estos datos suelen almacenarse en bases de datos relacionales y se caracterizan por tener un formato consistente y fácilmente accesible. Los datos estructurados se organizan en filas y columnas, y se pueden categorizar según su tipo de dato (numérico, texto, fecha, etc.).

La forma más popular de base de datos hoy en día es la base de datos relacional.

Una base de datos relacional es aquella en la que los datos se organizan en una o más tablas. Cada tabla tiene un conjunto de atributos, que definen la naturaleza de los datos almacenados en la tabla como “columnas”, mientras que las “filas” son los registros. En el siguiente ejemplo, tenemos una tabla de información del estudiante, con cada fila representando a un estudiante y cada columna representando una pieza de información sobre el estudiante.

Características de los Estudiantes

ID ALUMNO	NOMBRE	APELLIDO	DNI	FECHA DE NACIMIENTO	ID CURSO INSCRIPTO
001	Valeria	Romano	37490011	22/02/1994	012
002	John	B	33976995	19/11/1988	011
003	Cristian	Natured	34596000	20/5/1990	012
004	Salomon	Paretto	44149240	2/1/2002	014

En una base de datos relacional, todas las tablas están relacionadas por uno o más campos, de modo que es posible conectar todas las tablas de la base de datos a través de los campos que tienen en común. Esta columna en común se conoce en la jerga como: columna clave o key column.

Tomando la tabla anterior como ejemplo, podemos decir que forma parte de una base de datos de los estudiantes de este curso. Además de esta tabla en la que ustedes se registran con sus datos para este curso en particular, la plataforma e-learning ofrece otros cursos.

ID CURSOS	TITULO	DURACIÓN
011	Data Science	2 meses
012	Taller de escritura académica	1 mes
013	Marketing Digital	6 meses
014	Data Analytics con R	2 meses
015	Coaching Ontológico	5 meses

Si a su vez, la plataforma registra a cada profesor con sus datos podemos tener lo siguiente:

ID Profesor	Nombre y Apellido	Profesión
111	Roberto Brignoli	Técnico Comercial
112	Ana Fucilli	Bióloga
113	Cesar Martinez	Licenciado en Matemáticas
114	Juan Mazzieri	Economista

Bien, ahora veamos: cada vez que un profesor es asignado a un curso, la plataforma le asigna un ID de curso. Sin embargo, un mismo profesor puede impartir más de un curso, por lo que se registra la siguiente tabla:

ID Profesor	ID Curso
112	012
114	011
112	015
113	014
111	013

Estas tablas están relacionadas entre sí mediante claves foráneas, tales como el número de identificación del estudiante, el número de curso y el número de identificación del profesor, entre otros. Y a su vez cada tabla tiene registros de identificación únicos conocidos como clave primaria, como en la tabla profesores el “id profesor”, en la tabla alumnos el “ID Alumno”, en la tabla cursos el “ID CURSO”, pasando a comportarse como foránea en la tabla Estudiantes.

Con una base de datos estructurada de esta manera, es posible realizar consultas y análisis de datos para entender mejor qué cursos son los más populares, qué estudiantes están teniendo más éxito, qué profesores tienen mejores resultados, entre otros aspectos.

En la jerga de la informática, es común utilizar los "diagramas de entidad-relación" para visualizar cómo se relacionan las diferentes tablas entre sí en una base de datos. Si hacemos el diagrama de nuestro ejemplo, tendremos lo siguiente:

Los formatos comúnmente utilizados en Datos Estructurados son archivos CSV y hojas de cálculo XLSX, los cuales pueden ser consultados y procesados por herramientas convencionales como Excel y SQL.

Como sucede con todo en la vida, las cosas se vuelven más complejas con el tiempo, y esto también se aplica a la estructura de los datos, especialmente en grandes organizaciones e internet. Por lo tanto, aparecen los llamados Datos Semiestructurados y No Estructurados.

Los datos semiestructurados son un tipo de datos que se encuentran comúnmente en el lenguaje de internet, como HTML o XML. Cuando se inspecciona una página web, se puede ver el lenguaje en el que se diseñó la página. Aunque sigue ciertos patrones, no se puede tratar de la misma forma que los datos estructurados convencionales, por lo que se los considera "semiestructurados".

¿Qué es XML? - Explicación del lenguaje de marcado extensible (XML) - AWS (amazon.com)

2.3. Datos no estructurados

Se refiere a datos que no tienen una estructura u organización definida, pueden tomar muchas formas, como texto libre, imágenes, videos, archivos de audio, mensajes de redes sociales, correos electrónicos, documentos PDF, entre otros. Estos datos no siguen un esquema predefinido y, por lo tanto, no se pueden organizar fácilmente en tablas o columnas.

A pesar de que los datos no estructurados presentan un desafío para la organización y el análisis, se han vuelto cada vez más importantes en el mundo actual. Muchas empresas y organizaciones trabajan con grandes cantidades de datos no estructurados, lo que ha llevado al desarrollo de nuevas herramientas y tecnologías para su procesamiento y análisis, como el aprendizaje automático, el procesamiento de lenguaje natural y la inteligencia artificial.

Existen empresas que buscan detectar acciones o movimientos peligrosos de los clientes frente a una cámara, para poder predecir posibles robos, ataques o agresiones. Los videos que detectan movimientos son los "datos". Para guardar estos datos en un formato analizable, se requiere un tipo de esfuerzo y herramientas informáticas diferentes a las utilizadas para datos estructurados o semiestructurados. Adicionalmente, como datos no estructurados, se pueden incluir visitas a páginas (reviews), patrones de comportamiento de clics en páginas de interés, comentarios y likes, entre otros.

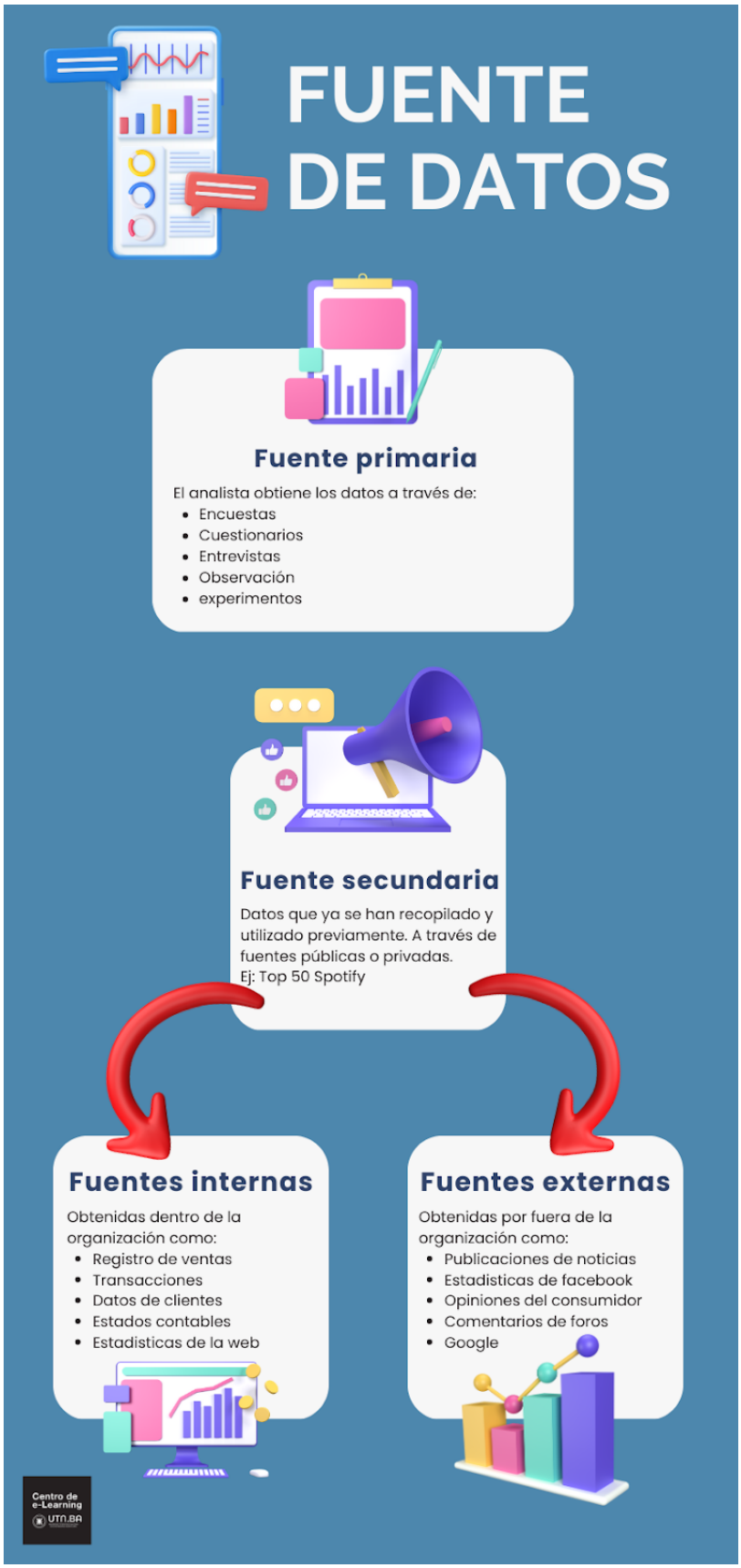
2.4. ¿Cómo se obtienen los datos?

Para obtener datos existen diversas formas, conceptualmente pueden clasificarse en dos grupos:

- Fuentes primarias: son aquellas que proporcionan información directamente desde su origen. Los datos son recopilados por el investigador de primera mano, sin pasar por intermediarios. Ejemplos son las encuestas, censos, entrevistas, experimentos y observaciones directas de sensores, telemetría, etc.
- Fuentes secundarias: son aquellas que utilizan información que ya ha sido recopilada y publicada por otros. Pueden ser internas o externas de la organización que realizan la investigación. Ejemplos son investigaciones de mercado, bases de datos en línea, artículos de revistas científicas, etc.

Ejemplos dentro de una organización:

- Un empleado contratado por un local gastronómico de varias franquicias, desea relevar la valoración de los comensales, para esto propone en cada franquicia, realizar entrevistas y cuestionarios para evaluar el desempeño. ¿Ante qué fuente creen que nos encontramos? ¿ El profesional obtiene de primera mano los resultados? ¿o previamente alguien lo hizo por él?
- Una empresa de conexión de talentos profesionales con empresas con avidez de cubrir puestos, analiza las preferencias de búsqueda de los perfiles profesionales y de las empresas para evaluar si pueden hacer un buen match para proceder a una entrevista. Para esto se decide verificar las valoraciones de las empresas que buscan atraer talento a través del ranking de “Great Place to Work”. Esta fuente de datos de la que estamos hablando: ¿Creen que es una fuente primaria o previamente realizó esta recopilación alguien más?


FUENTE DE DATOS
Fuente primaria
El analista obtiene los datos a través de:
Encuestas
Cuestionarios
Entrevistas
Observación
experimentos
Fuente secundaria
Datos que ya se han recopilado y utilizado previamente. A través de fuentes públicas o privadas.
Ej: Top 50 Spotify
Fuentes internas
Obtenidas dentro de la organización como:
Registro de ventas
Transacciones
Datos de clientes
Estados contables
Estadisticas de la web
Fuentes externas
Obtenidas por fuera de la organización como:
Publicaciones de noticias
Estadisticas de facebook
Opiniones del consumidor
Comentarios de foros
Google
Centro de e-Learning
UTN.BA

2.5. Almacenamiento de los datos

Un punto importante a tener en cuenta es dónde analizan las empresas o grupos de investigación sus datos, esto es a donde se almacenan y cuales son sus fuentes de datos. Anteriormente, muchas empresas utilizaban y aún utilizan archivos de Excel compartidos, pero ¿consideran ustedes que es fiable almacenar datos importantes en Excel? ¿Creen que es inalterable?

La respuesta en su mayoría es no. Si bien es fácil manejar los datos en Excel y compartirlos, tiene sus limitaciones cuando la cantidad de datos es grande y cuando el nivel de usuarios lleva a que pueda no ser confiable. Es fácil de manipular y editar sin tener registro de los cambios.

Es por esto, que, el paradigma actual de las organizaciones busca la manera de "centralizar" todos los datos en un mismo lugar. Centralizar los datos se refiere a la práctica de almacenar todos los datos de una organización en un solo lugar, en lugar de tenerlos dispersos en diferentes ubicaciones y sistemas. Sin embargo, centralizar los datos también puede tener algunos desafíos, como el costo y la complejidad de implementar un sistema de almacenamiento centralizado y la necesidad de garantizar que los datos se mantengan actualizados y precisos.

Para entender los tipos de almacenamiento, primero necesitamos entender el concepto de servidores:

Los servidores son dispositivos que se utilizan para almacenar, procesar y administrar datos y recursos en una red. Existen dos tipos principales de servidores: los servidores físicos y los servidores en la nube.

Los servidores físicos son dispositivos físicos/hardware que se encuentran en una ubicación física específica, generalmente dentro de una organización o empresa. Estos servidores son administrados y mantenidos por el personal de TI de la organización y se utilizan para almacenar y procesar datos, aplicaciones y otros recursos. Los servidores físicos se pueden configurar para satisfacer las necesidades específicas de la organización y pueden ser más seguros y confiables que los servidores en la nube, pero también son más costosos de adquirir y mantener.

Por otro lado, los servidores en la nube son servidores que se ejecutan en la nube y se pueden acceder a través de Internet. Estos servidores son administrados y mantenidos por un proveedor de servicios en la nube, como Amazon Web Services o Microsoft Azure. Los servidores en la nube son escalables y flexibles, lo que significa que la capacidad de almacenamiento y procesamiento se puede aumentar o disminuir según las necesidades de la organización. Además, los servidores en la nube son generalmente más rentables que los servidores físicos, ya que no es necesario realizar una inversión inicial significativa en hardware. Sin embargo, la seguridad y privacidad de los datos pueden ser preocupantes, ya que los datos se almacenan en servidores externos que son gestionados por terceros.

En base a los servidores disponibles en la organización se construye el sistema de almacenamiento centralizado, esto es dependiendo qué tengo disponible, cuánta memoria, y cómo se va a acceder a esos datos.

Van a escuchar mucho sobre los términos 'arquitectura' e 'infraestructura' en lo referente al almacenamiento de datos en las organizaciones. Si bien el data scientist no se encarga de la construcción de estos sistemas, es importante tener nociones sobre su origen y cómo funcionan, ya que es como construir una casa propia. Normalmente, uno no construye su propia casa, pero quiere conocer el proceso de construcción.

Lo que van a escuchar muy seguido en el mundo de la ciencia de datos es el término "Data Warehouse". Se trata de una infraestructura o base de datos que almacena de forma centralizada todos los datos que requiere la organización, esto es tiene una gran capacidad de almacenamiento, y permite a los analistas acceder a datos históricos de distintas fuentes. Es una infraestructura importante para el análisis de datos, y aunque es posible profundizar en detalles más técnicos, este concepto es suficiente para tener una primera noción sobre el tema.

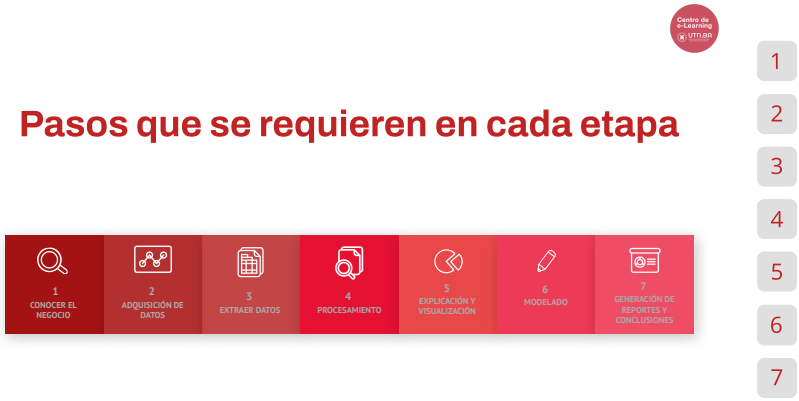
Para ampliar un poco más sobre el significado de servidores, data warehouse, arquitectura y almacenamiento en la nube, recomiendo la siguiente bibliografía "The Definitive Guide to Cloud Computing" por Dan Sullivan (Editorial Apress, 2015).

2.6. Ciclo de vida de los datos: ¿Cuál es el flujo de trabajo que siguen los datos para obtener conclusiones enriquecedoras a través de ellos?

Bien, una vez asimilado el concepto de los datos, los tipos de datos que existen, cómo se obtienen y cómo se almacenan normalmente en las organizaciones, veamos cuál es el ciclo de vida de los datos hasta que se obtiene valor o se toman decisiones clave a partir de ellos.

Enfocándonos en la temática del Data Science, si bien el Data Scientist no es el responsable de ejecutar cada etapa del ciclo de vida de los datos, sí es responsable de tener un conocimiento global del camino de los datos dentro de la organización.

Si bien no existe una receta oficial sobre cómo es el ciclo de vida de los datos paso a paso, con la bibliografía utilizada en este curso y el conocimiento práctico en organizaciones, se puede definir el siguiente ciclo:



genially Education