
Universidad
Tecmilenio®

Universidad Tecmilenio

Ingeniería de datos masivos

Data modeling orientado a Big Data

Semana 6

Introducción

Kaggle es una comunidad de científicos de datos que pertenece a Google, en la cual se pueden encontrar, por ejemplo, bases de datos públicas e información y competencias sobre el aprendizaje automático (machine learning), así como publicaciones de múltiples organizaciones sobre diversas problemáticas, de ahí que diversos equipos intenten resolverlas para obtener reconocimiento, además de un premio económico sustancial por lograrlo.

Explicación

La **modelación de datos** para big data consiste en crear modelos para almacenar y analizar datos.

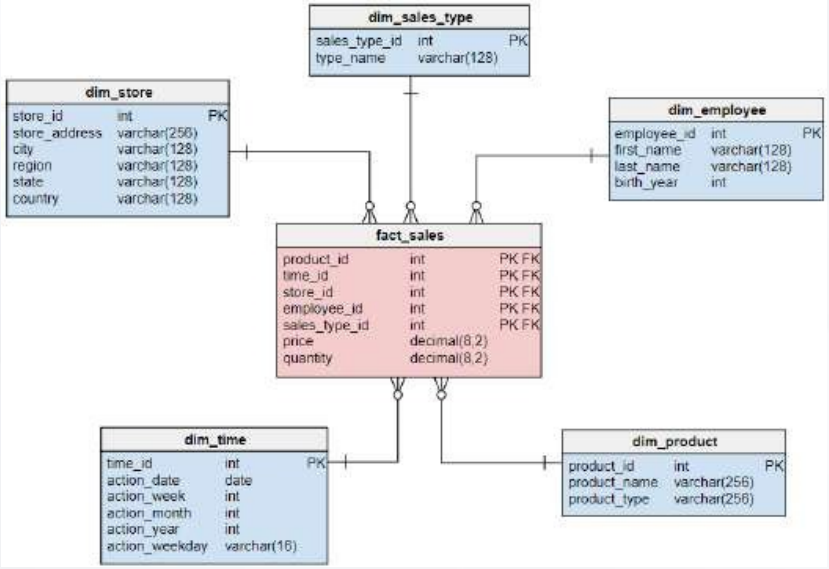
Desde una perspectiva científica, un **modelo** es una representación conceptual que describe de forma simplificada una realidad más compleja, facilitando entender el fenómeno y su análisis.

Los modelos para almacenar datos describen la estructura que se diseña para el almacenamiento de los mismos.

Estos modelos incluyen relaciones, restricciones y la forma adecuada para almacenar datos para poder manipularlos mejor y hacer análisis con ellos. Los datos describen personas, lugares, cosas de la vida real y los eventos entre ellos, por lo tanto, el modelo de datos representa la realidad que se va a analizar.

Explicación

El **modelo estrella** es el más utilizado, ya que consta de una tabla de hechos principal que describe los eventos ocurridos. Además, contiene tablas de dimensiones que explican los objetos de la realidad que participan en esos hechos.



Fuente: Ahmed, I. (2018). *Modeling Your Dimensional Data Warehouse: Sta. Schema vs. Snowflake Schema* . Recuperado de https://datawarehouseinfo.com/data-warehouse-star-schema-vs-snowflake-schema/

De acuerdo con IBM Knowlegde Center (s.f.), las tablas de hechos y dimensiones se definen de la siguiente forma:

Tabla de hechos o entidad de hecho

Guarda las medidas para realizar la evolución del negocio, por ejemplo, las ventas o el costo de las mercancías.

Total de dimensiones o entidad de dimensiones

Almacena información descriptiva sobre los valores numéricos presentes en una tabla de hechos.

Explicación

Un **modelo de análisis de datos** es el algoritmo o programa que se construye para representar el comportamiento que tienen los datos y se usa para hacer el análisis de los datos.

Se pueden diseñar para representar lo que ha acontecido en el **pasado**, o bien, se pueden diseñar para aprender de los datos y predecir el **futuro**.

Los modelos de análisis de datos pueden agruparse de la siguiente manera:

Los modelos de análisis pueden ser **supervisados** cuando los datos tienen etiquetas, las cuales son categorías o respuestas incluidas en los datos que se usan para aprender y hacer predicciones. En este caso, los modelos son **predictivos**.

Los modelos de análisis pueden ser **no supervisados** cuando los datos no tienen categorías o respuestas asociadas. En este caso, los modelos son **descriptivos** y se analizan los patrones que existen en los datos.

Explicación

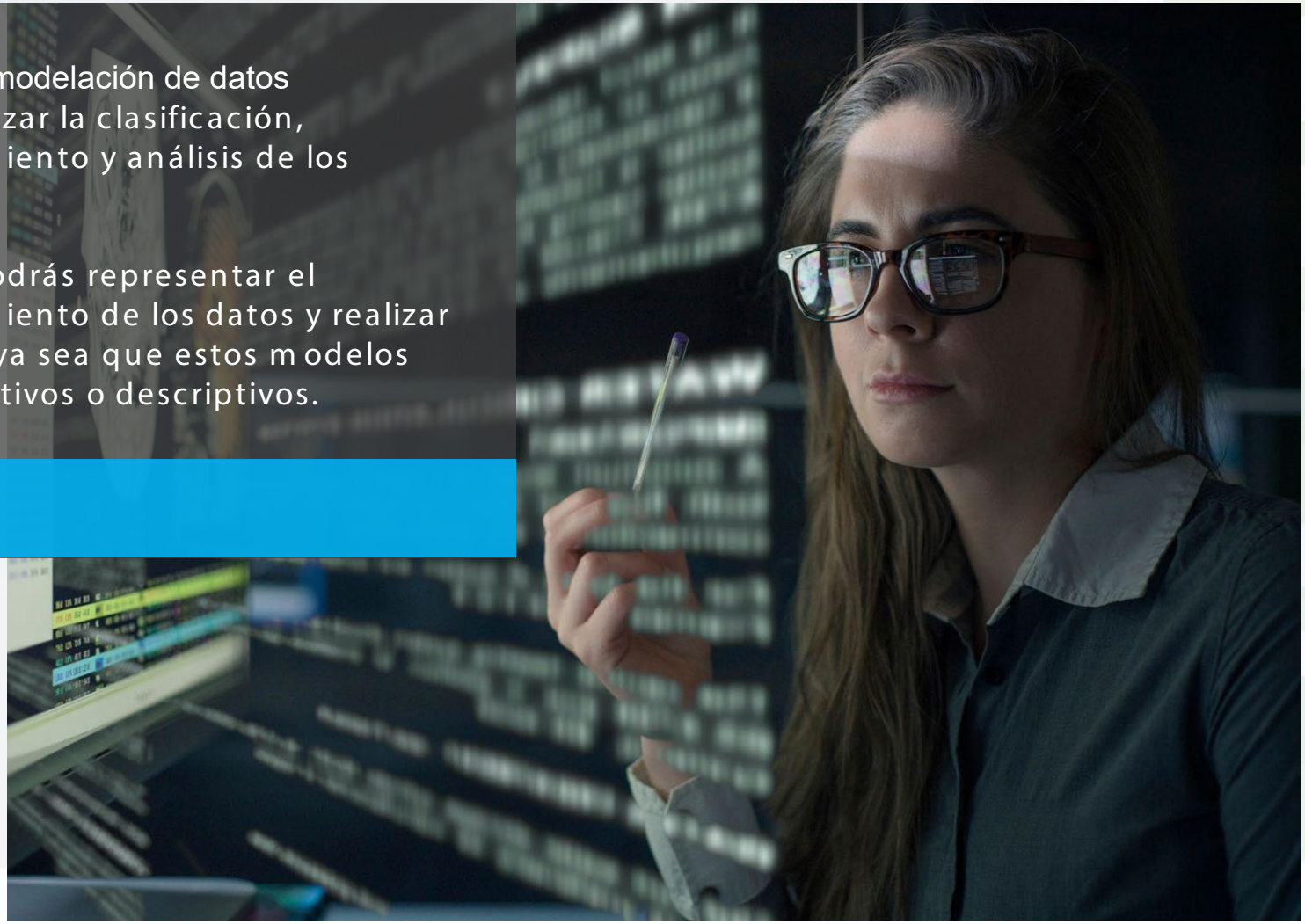
A partir de lo anterior, podemos señalar las siguientes definiciones sobre los diferentes tipos de modelos de análisis de datos:

Clasificación	Asignar etiquetas de clase a los datos; se le proporciona a un modelo clasificador un conjunto de ejemplos que ya están clasificados y a partir de estos aprende a asignar ejemplos no vistos.
Regresión	Para definir la relación entre una variable dependiente con una o más variables independientes.
Series de tiempo	Comprender los factores determinantes y la estructura detrás de una serie de tiempo y elegir un modelo para pronosticar y mejorar la toma de decisiones.
Agrupación	Consiste en dividir todos los datos en grupos (clústeres) en función de los patrones de los datos.

Cierre

Mediante la modelación de datos podrás realizar la clasificación, almacenamiento y análisis de los mismos.

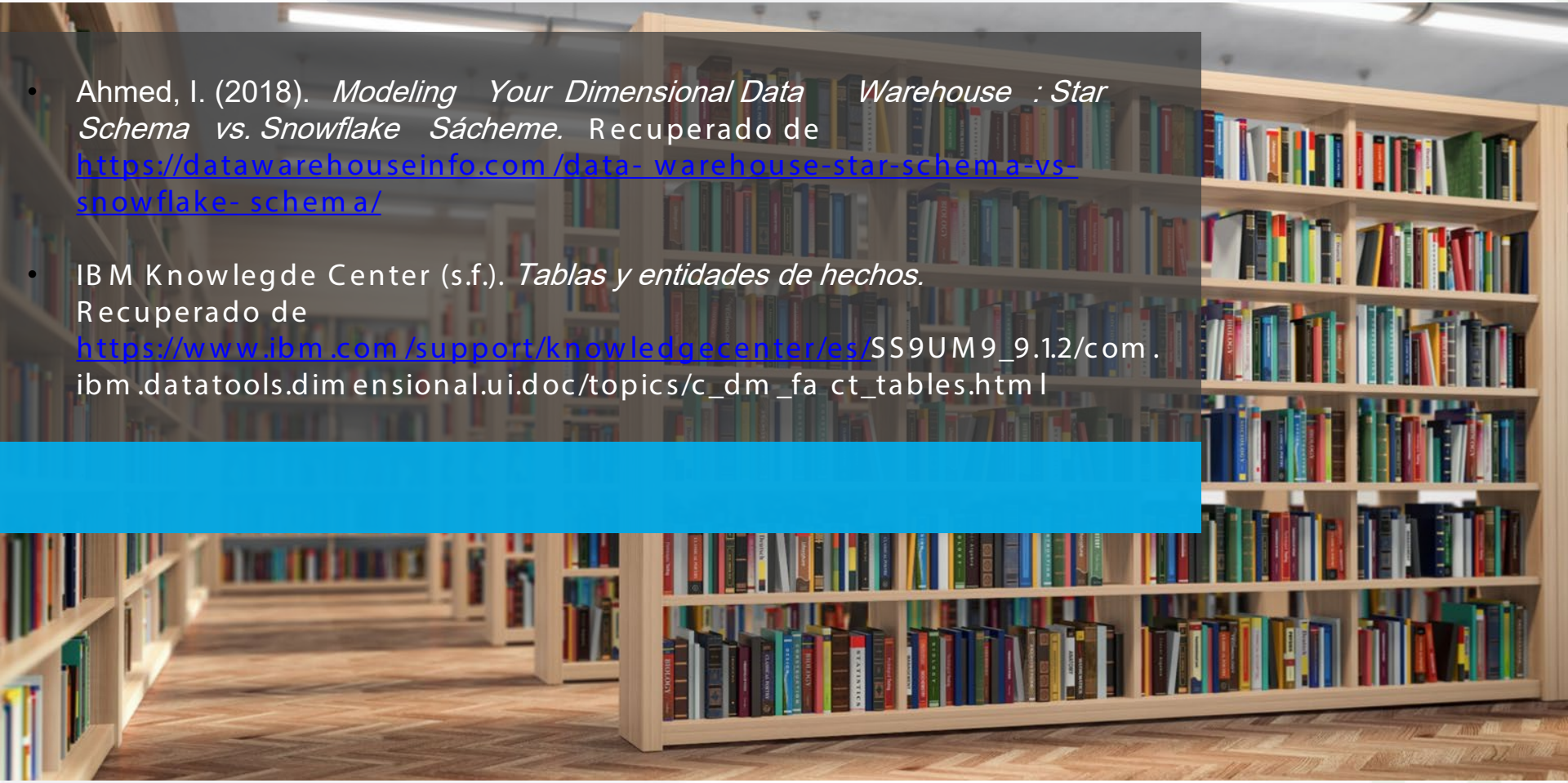
También podrás representar el comportamiento de los datos y realizar su análisis, ya sea que estos modelos sean predictivos o descriptivos.



Bibliografía

- Ahmed, I. (2018). *Modeling Your Dimensional Data Warehouse : Star Schema vs. Snowflake Sácheme.* Recuperado de https://datawarehouseinfo.com/data- warehouse-star-schema-vs-snowflake- schema/

- IBM Knowlegde Center (s.f.). *Tablas y entidades de hechos.* Recuperado de https://www.ibm.com/support/knowledgecenter/es/SS9UM9_9.1.2/com.ibm.datatools.dimensional.ui.doc/topics/c_dm_fact_tables.html



Universidad Tecmilenio

Ingeniería de datos masivos

Categorías de grupos de datos para Big Data

Semana 6

Introducción

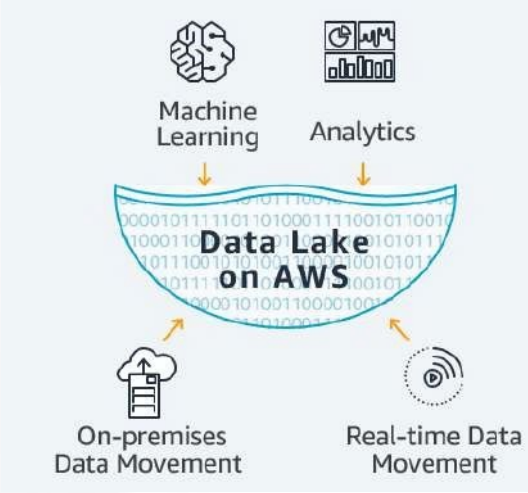
Los datos son un nuevo activo de las organizaciones, ya que con ellos se logran mejores decisiones, mejores productos y más clientes.

Asimismo, en una organización se recopilan, almacenan y analizan datos en los sistemas que soportan la operación de sus procesos para hacer analítica mediante el uso de las bases de datos.

Explicación

En la actualidad, la gran mayoría de las actividades que realizamos de forma cotidiana generan un rastro digital, el cual se utiliza como materia prima para mejorar los productos y servicios, es decir, la abundancia de los datos ha cambiado la naturaleza de la competencia entre las organizaciones.

Lo anterior se refleja a través de los **lagos de datos** (data lake) destinados al análisis, en donde diversas organizaciones ofrecen un conjunto de servicios para su administración, por ejemplo, **Amazon Web Services.**



Fuente: Amazon Web Services. (s.f.). *Data Lakes and Analytics* . Recuperado de https://aws.amazon.com/big-data/datalakes-and-analytics/?nc1=h_ls

Explicación

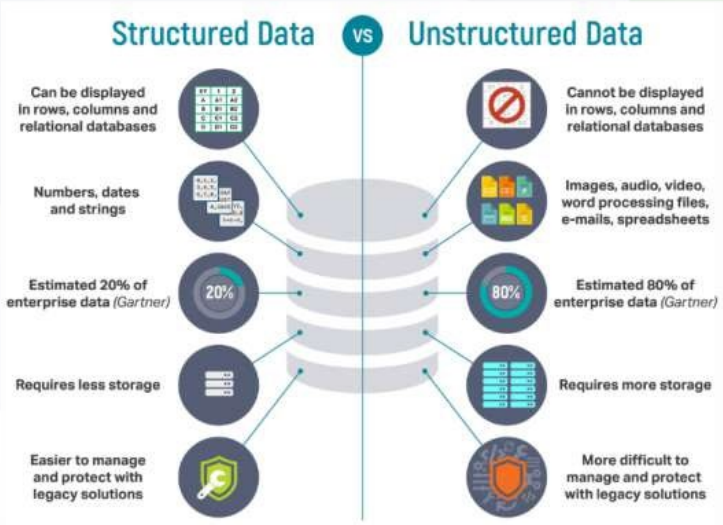
Los tipos de datos que se recopilan, almacenan y analizan en una organización son archivos de:

- Texto
- Video
- Audio
- Sensores
- Redes sociales
- Json, XML

Los **datos** son generados en diferentes áreas de operación como ventas, servicio, gestión de pedidos, fabricación, compras, facturación, cuentas por cobrar y cuentas por pagar.
Las transacciones o eventos que se registran incluyen sus descripciones, por ejemplo, para una venta sería la hora, fecha, lugar, precio, descuento, producto y cliente.

Hoy en día los datos no estructurados representan alrededor del 80% de los datos existentes en las organizaciones, mientras que el resto son datos estructurados.

Los datos tienen una **estructura** y una forma particular en la que se organizan y se almacenan en una computadora para que se puedan acceder y modificar eficientemente.



Fuente: Chaing, C. (2018). *In the Machine Learning Era, Unstructured Data Management is More Important Than Ever*. Recuperado de https://www.igneous.io/blog/structured-data-vs-unstructured-data

Explicación

Se le llama **dato** a cada uno de los valores que toma una variable y que además, tiene un tipo.

Son tres datos (56, 48 y 90), por tanto, la **variable** es el peso de la persona y el **tipo** es cuantitativo.

Dentro de una población se pueden recoger datos de diferentes variables con los siguientes **tipos** (Patel, 2018):

Las variables *cuantitativas* son las que se expresan con números o cantidades.

Discreta	Cuando solo acepta valores aislados, es decir, no existe ninguna cantidad intermedia.
Continua	Cuando puede tomar cualquier valor entre un intervalo.

Las variables *cualitativas* expresan una cualidad o categoría.

Nominal	Cuando no tiene un criterio de orden, por ejemplo, una profesión que puede ser abogado, médico, etc.
Ordinal	Cuando puede tomar cualquier valor entre un intervalo, por ejemplo, las categorías: primero, segundo, etc.
Binaria	Cuando solo se tienen dos categorías, por ejemplo, hombre o mujer.

Explicación

Los **datos** son el elemento básico que se procesa para construir información, lo cual sirve para generar conocimiento para tomar decisiones y actuar. Lo anterior se puede desglosar de la siguiente forma (Rao, 2018):

Datos	Información	Conocimiento
Valores puntuales cualitativos o cuantitativos de hechos y objetos de nuestra realidad.	Datos contextualizados, categorizados, calculados y condensados. Asimismo, son datos procesados y organizados que tienen un significado.	Información a la que se integra la experiencia o que es evaluada en múltiples escenarios para generar nuevas conclusiones o tomar acciones.

Con datos se puede hacer analítica descriptiva para generar información y con esta, hacer analítica predictiva y así tomar decisiones. Además, dichas predicciones se evaluarán en diferentes escenarios por la analítica prescriptiva para recomendar o automatizar.

Cierre

En un mundo globalizado y digital, los datos son un recurso clave en las organizaciones, ya que permiten oportunidades de innovación materializadas a través de productos y servicios con un alto valor agregado para los consumidores.

Bibliografía

- Amazon Web Services . (s.f.). *Data Lakes and Analytics* . Recuperado de https://aws.amazon.com/big- data/datalakes-and-analytics/?nc1=h_ls
- Bhageshpur, K. (2019). *Data Is The New Oil -- And That's A Good Thing* . Recuperado de https://www.forbes.com/sites/forbestechcouncil/
- 2019/11/15/data-is-the-new-oil- and-thats-a-good-thing/#3fa4165a7304
- Chaing, C. (2018). *In the Machine Learning Era, Unstructured Data Management is More Important Than Ever.* Recuperado de https://www.igneous.io/blog/structured- data-vs-unstructured-data
- Patel, A. (2018). *Chapter -2 Data and It's Different Types.* Recuperado de https://medium.com/ml-research- lab/chapter-2-data-and-its-different-types-3dfebcbb4dbe
- Rao, V. (2018). *From data to knowledge* . Recuperado de https://www.ibm.com/developerworks/library/ba-data-becomes-knowledge- 1/index.html

Universidad Tecmilenio

Ingeniería de datos masivos

Tratamiento de la información y los datos

Semana 6

Introducción

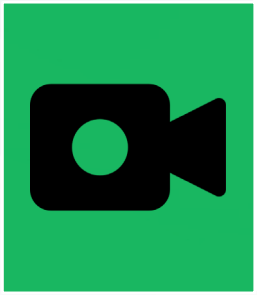
Los datos se encuentran en muchos lugares, tanto internos como externos a la empresa, por tanto, lograr datos limpios y completos es un reto. Asimismo, el tiempo que dedican los profesionales del sector para lograrlo ronda entre el 50% y 80%, ya que los datos generalmente no están listos para lograr hallazgos o hacer análisis, lo cual es muy importante, ya que son los objetivos de proyectos de analítica.

Explicación

La preparación de datos se convierte en un problema aún mayor cuando se considera big data, ya que provienen de una gran cantidad de fuentes.

Por lo tanto, una vez que se identifica la pregunta que se quiere responder se deben seleccionar los datos, los cuales estarán en diferentes formatos. Asimismo, estos incluirán una variedad de tipos de datos para lo cual se necesitarán de herramientas para analizarlos.

Para conocer más sobre la importancia de la preparación de los datos, revisa el siguiente video:



Microsoft Developer . (2018, 14 de noviembre). *Why Should I Care About Data Preparation ?* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=xrfbFKHpJ40

Explicación

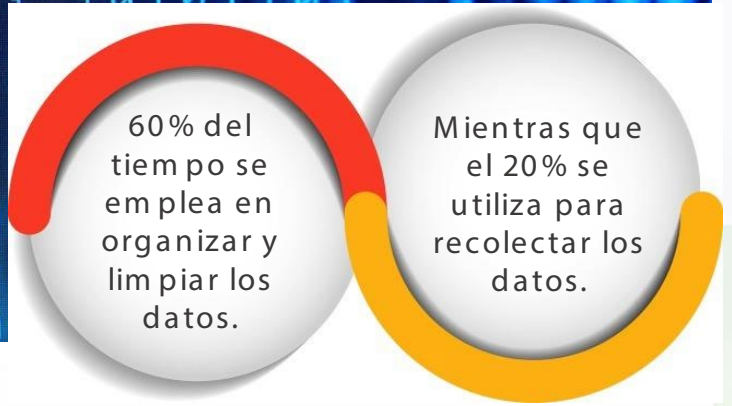
De acuerdo con la metodología CRISP-DM (Saluja, 2018), después de las etapas del entendimiento del negocio y de los datos, lo siguiente es prepararlos para iniciar con la etapa de la creación de un modelo que sirva para su análisis.

Cuando el objetivo es analizar con visualizaciones se tiene una etapa de preparación o tratamiento de datos para poder construirlas. Aunque se dispone de datos, el reto es que tengan la calidad necesaria para poder hacer análisis con ellos.

80%

Es el tiempo que representa la preparación y el manejo de los datos para el análisis en la actualidad.

De lo anterior se desprende lo siguiente:



Explicación

Después de identificar las fuentes de los datos, el objetivo es dejarlos en el formato requerido para la siguiente etapa . Por lo tanto, se requieren los siguientes pasos para lograrlo (Pearlman, 2020). Posteriormente se realiza la fusión o la integración de datos en una sola tabla .

Adquisición, ingesta, recopilación o captura de los datos.

Resolver dónde, cuándo y cómo se capturan los datos, así como registrarlo como **metadata** , los orígenes de los datos pueden ser de diferentes fuentes. Además de resolver cómo los vamos a guardar.

Formato y limpieza de los datos.

Resolver que cada dato tenga el formato adecuado.
Al hacer limpieza con los datos: hay que analizar si los datos son correctos, si hay faltantes, si son atípicos o nulos.

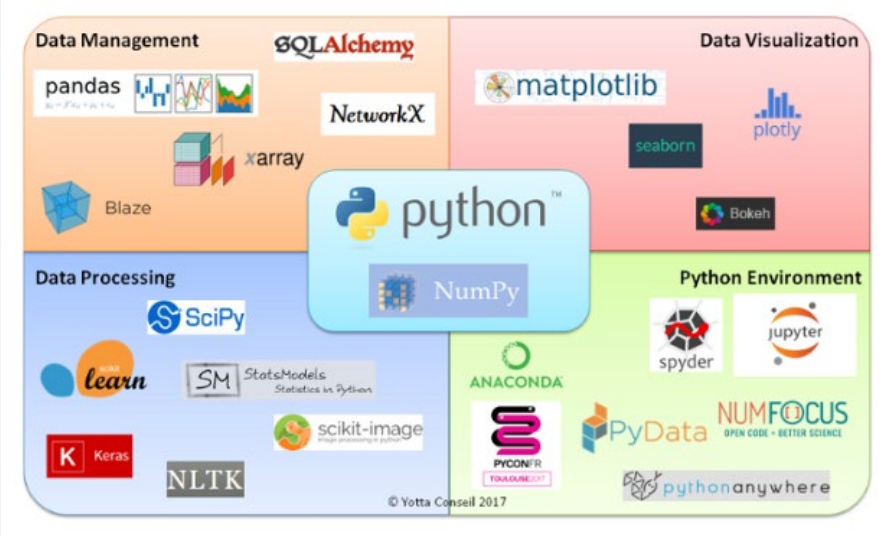
Creación de datos para el modelo.

Construir una tabla para el modelo a partir de los datos procesados para tener el formato requerido y estar limpios.

Explicación

Para hacer el tratamiento de los datos existen diferentes herramientas que de forma interactiva te permiten manipular los datos, por ejemplo, **Alteryx** o **Tableau Prep** .

Asimismo, este proceso también se puede realizar mediante lenguajes de programación como **Python** , utilizando librerías que nos permiten manejar tablas de datos (Pandas) y hacer visualizaciones (Matplotlib).



Fuente: Atrebas. (2019). *Self-studying Python and Machine Learning: 2018 in review.* Recuperado de https://atrebas.github.io/post/2019-01-15-2018-learning/

Para conocer un ejemplo de preparación de datos, mira el siguiente video:

Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 5. Preparación de los datos* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=nUHp8fmrFCM

Cierre

Entre los beneficios más sobresalientes de la preparación de datos podemos mencionar los siguientes (Pearlman, 2020):

- ✓ Solución rápida de errores en los datos antes de procesarlos.
- ✓ Producción de datos de alta calidad.
- ✓ Permite tomar decisiones empresariales más precisas y eficientes.

Asimismo, para realizar la preparación de datos existen diversas de herramientas disponibles como Tableau, Power BI, Qlik, Altair Monarch, Microstratregy, entre otras.

Bibliografía

- Atrebas . (2019). *Self-studying Python and Machine Learning : 2018 in review .* Recuperado de https://atrebas.github.io/post/2019-01-15- 2018-learning/
- Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 5. Preparación de los datos* [Archivo de video]. Recuperado de https://www.youtube.com /watch?v=nU Hp8fmrFCM&feature=youtu.be
- Microsoft Developer. (2018, 14 de noviembre). *Why Should I Care About Data Preparation ?* [Archivo de video]. Recuperado de https://www.youtube.com /watch?v=xrfb FKHpJ40
- Saluja, C. (2018). *Data Preparation — A crucial step in Data Mining .* Recuperado de https://medium .com /@ chhavi.saluja1401/data-preparation-a-crucial-step-in- data-mining-dba35772f281
- Pearlman, S. (2020). *What is Data Preparation ?* Recuperado de https://www.talend.com /resources/what-is-data-preparation/