Bloque 3.

Tema 3: Data wrangling

Universidad Tecmilenio

¿Qué es data *wrangling*?

También conocido como data *munging*, consiste en el proceso de limpieza, reestructuración y mejoramiento de los datos disponibles en bruto para transformarlos en un formato idóneo, permitiendo estimular el proceso para la toma de decisiones estratégicas.

El proceso de organizar y limpiar los datos previo al análisis ha comprobado ser de gran utilidad para las empresas, ya que permite analizar grandes cantidades de datos rápidamente.

En el data wrangling existen seis pasos generales (Acadgild, 2018):

Con dicho proceso se busca el descubrimiento de errores o de información incompleta, caduca, repetitiva, etc., con el fin de limpiar y eliminar duplicaciones para realizar su orden, estructuración y enriquecimiento.

Importancia

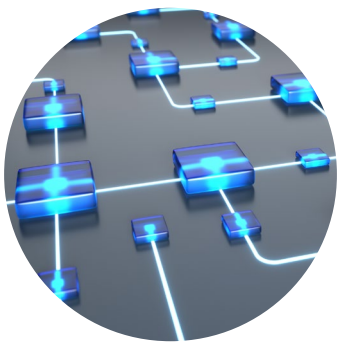
El mundo actual genera diariamente una gran cantidad de datos, por ejemplo, al momento de ingresar a las redes sociales, cuando miramos una película o alguna serie en plataformas de *streaming*, etc. Sin embargo, si estos datos están desactualizados tienen muy poco valor.

Para valorar la trascendencia de data wrangling, se consideran las siguientes razones (Imarticus, 2018):

- Credibilidad de los datos.
- Generación de confianza entre las partes interesadas.
- Apoyo para el aprendizaje automático (machine learning).

Por lo tanto, es posible afirmar que data wrangling ayuda a las empresas a tomar mejores decisiones y a ser más eficientes, puesto que contribuye a detectar errores, ahorrar tiempo, estructurar y homogenizar la información para su análisis. Asimismo, brinda información confiable y gestiona recursos e infraestructura.

Diferencia con data *mining*

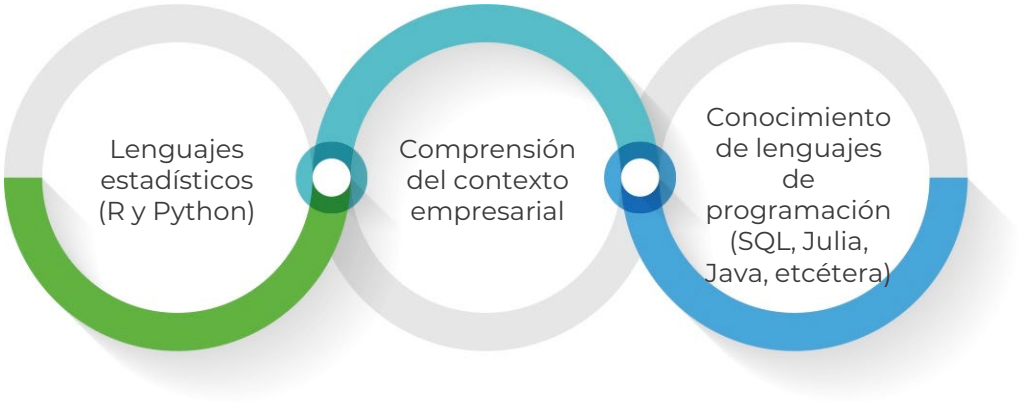


Data mining, también llamada como minería de datos, consiste en descubrir algunos patrones específicos escondidos dentro un conjunto vasto de datos, dando como resultado un patrón significativo; mientras que data wrangling es un superconjunto de minería de datos que implica varios procesos (transformación, limpieza, integración, etc.) y que genera como resultado una idea significativa.

Entonces, se puede considerar que data mining se encuentra dentro del proceso de data wrangling.

Habilidades requeridas

Una persona experta en data wrangling busca dar solución a todos los problemas relacionados con los datos, ya que estos se encuentran en todas partes (normalmente de forma cruda). Asimismo, un experto en datos necesita habilidades para integrar la información existente de varias fuentes, por lo que requiere conocimientos en los siguientes aspectos (Choudhury, 2019):



Retos

El desafío más grande es el tiempo que requiere, debido a la naturaleza estructural y detallada de los conjuntos de datos analíticos que se producen. Por esa razón, los principales desafíos son los siguientes (Pearlman, 2019):

- Dar sentido al resultado final.
- El acceso a los datos.
- Datos limpios.
- Integración manual de los datos.
- Ingeniería especifica.

No obstante, se tiene que considerar que la empresa debe tener un objetivo particular para que el proceso de data wrangling sea mucho más fácil y rápido, ya que al querer ordenar una gran cantidad de datos sin saber lo que se busca, podría ocasionar un gran desperdicio de recursos.

Referencias

- Acadgild. (2018). *What is Data Wrangling? 6 Steps in Data Wrangling*. Recuperado de https://acadgild.com/blog/6-steps-in-data-wrangling
- Choudhury, A. (2019). *The importance of data munging for data preparation in analytics*. Recuperado de https://analyticsindiamag.com/the-importance-of-data-munging-for-data-preparation-in-analytics/
- Imarticus (2018). *What is Data Wrangling and Why is it Important?*. Recuperado de https://imarticus.org/what-is-data-wrangling-and-why-is-it-important-data-analytics-blog/
- Pearlman, S. (2019). *Data wrangling: Speeding up Data preparation*. Recuperado de https://www.talend.com/resources/data-wrangling/

Para expandir tu conocimiento, te recomendamos los siguientes recursos adicionales:

Los siguientes enlaces son externos a la Universidad Tecmilenio, al acceder a estos considera que debes apegarte a sus términos y condiciones.

Podcasts

Para saber más sobre **data wrangling**, te recomendamos escuchar los siguientes podcasts:

- Roberts, B. (27 de noviembre de 2018). *BDB Podcast EP 53 – Data Wrangling Made Simple with Trifacta* [Audio podcast]. Recuperado de https://bigdatabeard.com/bdb-podcast-ep-53-data-wrangling-made-simple-with-trifacta/
- Kennedy, M. (21 de diciembre de 2016*). #90 Data Wrangling with Python* [Audio podcast]. Recuperado de https://bit.ly/36vDpmc
- The Data Science Happy Warriors Podcast. (27 de octubre de 2018*). Episode 2: Data Wrangling: Why you gotta do what you gotta do* [Audio podcast]. Recuperado de https://bit.ly/2ubC4Cv

MOOC

Para saber más sobre **data wrangling**, te invitamos a cursar los siguientes MOOC:

- Harvard. (s.f.). *Data Science: Wrangling*. Recuperado de https://www.edx.org/es/course/data-science-wrangling
- UC DAVIS. (s.f.). *Data Wrangling, Analysis and AB Testing with SQL*. Recuperado de https://es.coursera.org/learn/data-wrangling-analysis-abtesting
- Mongo DB. (s.f.). *Data Wrangling with MongoDB*. Recuperado de https://www.udacity.com/course/data-wrangling-with-mongodb--ud032