Extracción de datos

Retomando el concepto de ETL (Extract: extracción de la información, Transform: preparación de la información y Load: carga para continuar el proceso), hablaremos de la extracción, la primera etapa dentro de la preparación de la información.

En esta etapa no interesa comprender la información, sino obtenerla.

Debemos ser cuidadosos porque hay una variedad enorme de fuentes para analizar (digitales y no digitales).

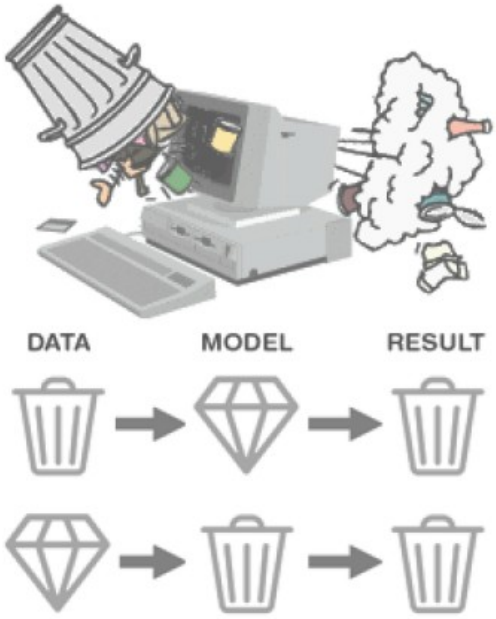
Algunas de las fuentes no están relacionadas con otras (silos de información). Por ejemplo, puede ocurrir que los datos de dos sectores de la misma empresa no estén vinculados, que estén en diferentes formatos, e incluso puede que ambos sectores estén utilizando software diferentes para sus procesos.

Tendremos que ser capaces de unificar los datos de interés en un programa en el que podamos realizar el análisis requerido.

Los tipos de archivos más comunes que encontramos urante una extracción son xslx, txt y csv.

Limpieza de datos

Es el segundo paso de ETL. Después de la extracción, la información no siempre llega de una forma estandarizada, con todos los datos que necesitamos. Incluso puede llegar con errores de ortografía, de estructura, campos en blanco, campos inexistentes o datos erróneos. Debemos intentar que toda la información sea entendible para los siguientes pasos.



GIGO: Garbage in, garbage out

Si ingresamos datos basura a nuestro proceso, la salida también será información basura. Por esto es importante que el proceso de limpieza sea eficiente y de calidad al momento de la extracción.

Este será nuestro primer acercamiento en el que empezaremos a entender qué datos tenemos, cómo lo podemos utilizar y cuáles son las primeras preguntas que puedo formular.

Podemos utilizar cualquier software para la limpieza de datos, lenguajes de programación y hasta planillas de cálculo. Esto nos permitirá filtrar información, cambiar valores, excluir datos, etc.

Algunos ejemplos de las tareas más comunes para comenzar a limpiar nuestros datos:

- Eliminar espacios en blanco
- Eliminar celdas y filas en blanco
- Convertir números almacenados como texto a números
- Eliminar celdas o registros duplicados
- Resaltar errores
- Cambiar texto a minúculas/mayúsculas o primera letra mayúscula para que los datos sean consistentes
- Distribuir texto en columnas
- Chequeo ortográfico
- Eliminar formatos
- Utilizar el Buscar/Reemplazar

Actividad

Ejemplo de Limpieza de datos en planillas de cálculo. En la primera pestaña se ven los datos tal cual nos los entregaron, y en la segunda el resultado final. Es posible hacer una copia del documento para que puedas experimentar.
No te preocupes si no sabes cómo proceder. En las próximas clases veremos funciones en planillas de cálculo.
Planilla: *Actividad BI ETL – Limpieza* (busca el archivo para descargar en el aula virtual)

Exploración de datos

Es la manera en que vamos a tratar de comprender los datos que estamos analizando, entender qué historias nos quieren contar y cuáles son los descubrimientos que podemos hacer mientras exploramos estos datos.

También nos ayuda a hacernos preguntas que ni sabíamos que existían. Al momento de ver patrones y tendencias, podemos empezar a hacer este ejercicio de reflexión: ¿Qué es

lo que produce este patrón? ¿Por qué hay tantas visitas en determinados páginas de nuestra web y por qué tan pocas en el resto?

Además propone cambios de hipótesis. Generalmente, al inicio de nuestro proceso tenemos algunas preguntas de las cuales queremos tener respuestas. Es muy común que nuestras preguntas cambien después de la exploración, cuando descubrimos que nuestros datos no nos van a permitir responder las preguntas iniciales Sin embargo, es posible que nos puedan brindar información más interesante que ni siquiera teníamos en nuestro radar.

Las herramientas más utilizadas para explorar datos son los lenguajes de programación como Python y R, las planillas de cálculo, y softwares más específicos como Looker Studio, Tableau y Power BI.

Actividad

Revisa los dos ejemplos de exploración de datos, ambos realizados a partir de la misma fuente

Google Sheets : https://docs.google.com/spreadsheets/d/1XHXjDyk5q5bJZyHXSbJxpM-52ue_EY0oyosBJRrhbws/edit?usp=sharing

Looker Studio: https://datastudio.google.com/reporting/451a5266-2e4f-4b1d-a1a9-42d1297f8b21

Durante este curso, aprenderemos a utilizar estas herramientas.

Highlights

Son los descubrimientos que tenemos en nuestros datos y que nos ayudan a tomar decisiones. Nos permiten saber qué está ocurriendo en nuestra información.

Aunque se pueden buscar infinidad de cosas en la información, debemos tratar de empezar a buscar según dos tendencias: empezar por lo más buscado (Top Data) y terminar por lo menos buscado (Bottom Data). ¿Cuál es el producto o servicio que más se vende? ¿Cuál es el que nos deja más margen de contribución? ¿Cuál es el menos vendido? ¿Cuál es el que menor margen de contribución nos deja?

Los highlights también nos permiten encontrar patrones, sobre todo por temporalidad, por ejemplo ¿Cómo se ven afectadas las compras según el mes, la estación o festividades?

Otro tipos de patrones pueden encontrarse en las correlaciones entre dos valores, como utilidades vs. ventas.

Es una tarea en la se invierte bastante tiempo porque estamos esforzándonos por “hacer hablar” a nuestros datos.

Reporting

Es clave tener en cuenta el mensaje, los descubrimientos, los highlights encontrados, cuáles son los que pueden interesarle a la audiencia y buscar la mejor manera de transmitirlo.

También es preciso pensar en nuestra audiencia. No se debe hacer la misma presentación para diferentes audiencias, porque tienen diferentes intereses, inquietudes y capacidad de comprensión de téminos técnicos.

Hay diferentes fomas de reporting: estáticos (impresos, pdf,Power Point), o dinámicos (Power BI, Tableau, Looker Studio).

Las buenas prácticas nos dan las pautas para una comunicación eficaz, sobre las tendencias y formas que tenemos de procesar la información a la hora comprender datos. Por ejemplo: está comprobado que es más fácil de comprender y sintetizar una imagen que una planilla repleta de datos.

No debemos olvidar el contexto al presentar un reporte. No todos los asistentes están al tanto de la situación, y deberíamos asegurarnos de darles una pauta de entendimiento previo, para darle un marco a nuestro mensaje.

La información debe ser concentrada en los **dashboards**, para un primer filtrado de las visualizaciones y quedarnos con lo más importante que pudimos encontrar. Es el soporte visual de nuestra presentación.

La manera en que nosotros vamos a poder relatar nuestros hallazgos se denomina **storytelling**, y trata de contar la historia sobre nuestros descubrimientos de forma que mantenga la atención de la audiencia.

No debemos olvidar la descripción de los patrones encontrados, de los descubrimientos y nuestras sugerencias para la toma de decisiones. Como profesionales del BI se aprecia nuestra opinión sobre cómo debe utilizarse esta información.