Capítulo 4
Analizar y hacer más fiables los datos

1. Introducción

El objetivo de este capítulo es repasar las herramientas y recursos que se pueden utilizar para comprender y analizar mejor los datos. Los datos pueden tener distintas facetas, entre ellas la de estar estrechamente ligados a un contexto, de modo que se pueden interpretar de una manera en un contexto determinado y de otra completamente distinta (incluso opuesta), en otro. Además, los datos tienen vida propia y se pueden variar o alterar con el tiempo o simplemente sufrir cambios durante su transporte o en su soporte de almacenamiento.

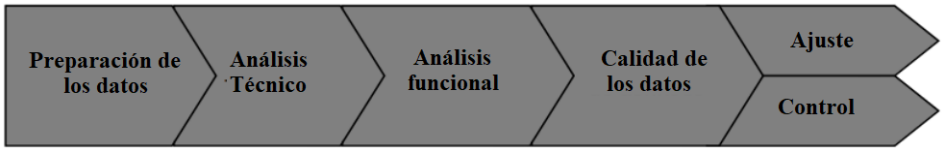
Por eso, antes de utilizar cualquier dato, es importante analizarlo para comprobar que se corresponde con nuestras expectativas en el momento de utilizarlo. Imaginemos que recuperamos conjuntos de datos no documentados ni explicados. En este caso, será esencial pasar por el proceso de análisis. Es una buena práctica asegurarse de que los datos que se van a utilizar son realmente conformes, y este es el objetivo de este capítulo.

Primero veremos cómo analizar nuestros datos desde una perspectiva técnica o estructural: es lo que se conoce como perfilado de datos. Este análisis se centra principalmente en los tipos, formatos y número de ocurrencias de los datos y no requiere ningún conocimiento particular de los mismos. El objetivo de esta fase es establecer una visión objetiva de los componentes estructurales de los datos y destacar las características cuantificables que se pueden extraer de su conjunto.

A continuación, veremos cómo analizar los datos desde un punto de vista funcional y, por tanto, más cualitativo. El aspecto cuantitativo también será posible, pero en este caso habrá que adaptarlo a un contexto empresarial. En esta sección, nos centraremos especialmente en cómo presentar nuestros datos para que sean más significativos.

Por último, como un médico que ha hecho un diagnóstico (gracias a las fases de análisis anteriores), estudiaremos las medidas correctoras (control, solución, etc.) que hay que aplicar para que nuestros datos sean más fiables. Esto es la calidad de los datos. En cierto modo, esta última y crucial etapa es el primer gran paso hacia el éxito de un enfoque de gestión de datos.

Estos son los pasos prácticos que se deben seguir en un proyecto de datos:



Fases de cualificación de los datos

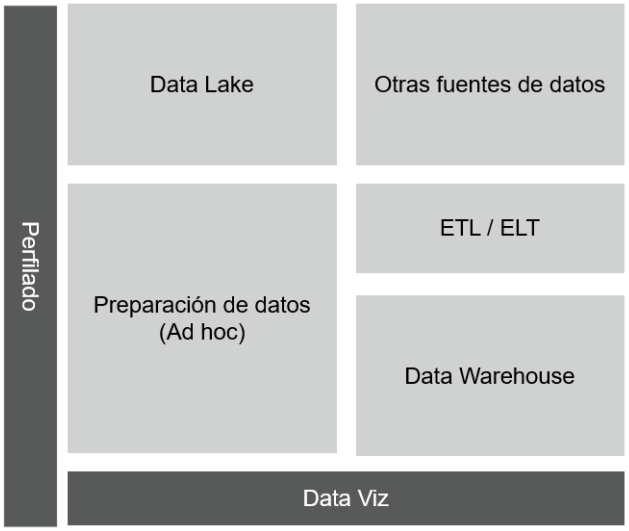
2. Preparar los datos

Por desgracia, los datos no siempre están listos para ser utilizados. Sin trabajarlos, rara vez es posible extraer los datos para obtener los resultados esperados (como la evolución de las ventas, el cálculo de un indicador de satisfacción, etc.). En la gran mayoría de los casos, habrá que reelaborar los datos en bruto para hacerlos utilizables para el análisis, la ingesta o incluso la modelización (Machine Learning).

Una cosa es cierta: aunque de algún modo esta fase es la cara oculta del iceberg, no deja de ser una etapa importante y, sobre todo, puede llevar mucho tiempo si no se lleva a cabo con las competencias y los recursos adecuados. Por ejemplo, se dice que los analistas de datos y otros científicos de datos dedican más del 80% de su tiempo a prepararlos. Esto les deja solo un 20% de su tiempo para trabajar con ellos.

Por lo tanto, la preparación de los datos consiste en recogerlos y transformarlos para poder trabajar con ellos. Son las dos primeras fases de nuestro famoso ETL (o ELT). Y con razón, el enfoque es muy similar, aunque la finalidad suele ser muy distinta. En un enfoque de tipo ETL, los datos se transportan en última instancia a una (o varias) fuentes de datos de destino, con vistas a un amplio análisis. Con la preparación de datos, el objetivo es prepararlos para un fin específico.

La tendencia que estamos observando es que los ETL siguen totalmente centrados en casos de uso de tipo transporte de datos (alimentar Data Warehouse, Data Lake, migración de datos, etc.) y que las herramientas de preparación de datos (que también tienen las mismas funcionalidades de transformación), están ofreciendo estas funcionalidades directamente a los usuarios de negocio (incluidos los científicos de datos). De este modo, estos consumidores de datos pueden trabajar directamente sobre ellos, o incluso recuperarlos directamente del Data Lake para filtrarlos, transformarlos, etc. En cierto modo, ahora pueden preparar los datos como mejor les parezca.



Análisis e Iniciativas sobre los datos

La idea es sencilla: permitir el acceso libre (potencialmente seguro) a todos los datos, pero también ofrecer las herramientas adecuadas a los usuarios que los conocen. Esta es una buena forma de empezar a analizarlos, aunque más adelante veremos que la fase de preparación se puede ampliar a fines más amplios, como la modelización de Machine Learning.

Desde un punto de vista práctico, y de forma similar al ETL, la preparación de datos consta de cinco fases principales:

– **Importación o adquisición de datos**: esta etapa requiere la conectividad con los distintos sistemas fuente.
– **Descubrimiento**: un análisis previo suele comenzar con una llamada a las funciones de perfilado de datos.
– **Depuración de datos**: a menudo es necesaria una depuración inicial de los datos. Por ejemplo, formatos que no coinciden (las fechas suelen ser un problema), o categorías que presentan problemas de coherencia, etc. Por tanto, es necesario alinear los datos para poder analizarlos posteriormente.

– **Enriquecimiento**: a este nivel, puede ser interesante añadir datos auxiliares (es decir, datos que no proceden de la fuente de datos). Si tenemos datos de localización, ¿por qué no añadir datos demográficos?
– **Publicación**: se trata de poner los datos preparados a disposición de la herramienta que los va a utilizar. Si el objetivo es analizar esos datos, quizá haya que exportarlos en un formato determinado o, quizá, haya que estudiar la solución de preparación de datos para ver si puede ponerlos a disposición de forma nativa.

3. Análisis descriptivo

En primer lugar, hay que señalar que el análisis de datos requiere que estos estén disponibles en formato tabular (filas y columnas). Por tanto, este capítulo trata de los datos estructurados. Actualmente, todas las soluciones (o casi todas) trabajan con datos estructurados de esta forma. Una vez disponibles los datos en este formato, se trata de analizarlos desde un punto de vista técnico, describirlos e identificar, por qué no, un primer nivel de excepciones (por ejemplo, la detección de valores atípicos).

Esta etapa tiene varios nombres: se denomina análisis descriptivo o perfilado de datos (Data Profiling).

Este análisis debe describir la muestra de datos (o todo su conjunto), perfilando cada columna para descubrir información importante sobre atributos, como frecuencia y distribución de los valores de los datos, formatos, patrones y nulos, mínimos y máximos. A continuación, se leen todos los datos para proporcionar un análisis y un inventario exhaustivos.

Pero las herramientas y las soluciones a menudo pueden llevarle mucho más lejos en términos de análisis. Veamos en detalle los diferentes análisis que se pueden realizar sobre un conjunto de datos, sin tener ningún conocimiento funcional del mismo. Para hacerlo, podríamos utilizar herramientas No Code como Informatica, Talend o SAS DataFlux que, con un simple clic, pueden obtener este tipo de resultados. Pero aquí vamos a utilizar la librería Python Pandas Profiling, para que cualquiera pueda probar este tipo de análisis en su propio ordenador.

Para instalar la librería, basta con ejecutar el comando pip:

```
$ pip install pandas_profiling
```

Todo lo que tiene que hacer es abrir un dataframe con la librería Pandas y lanzar el profiling:

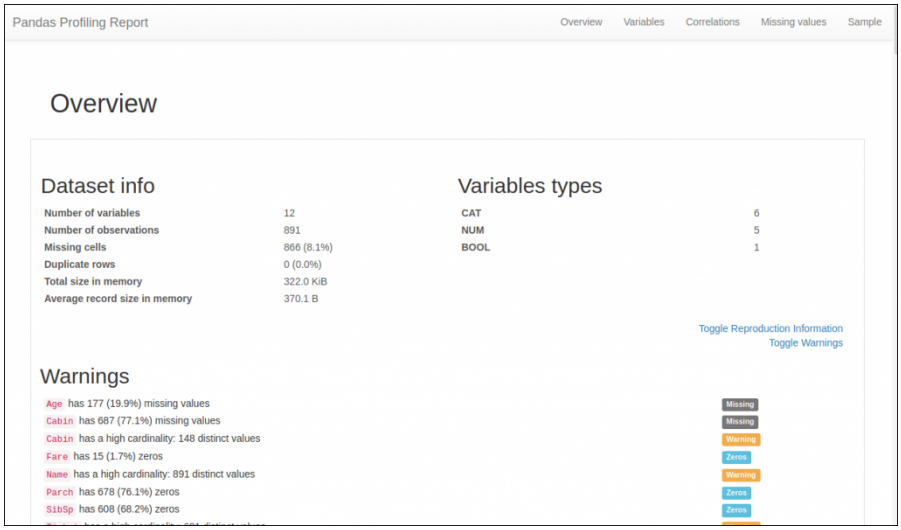
```
from pandas_profiling import ProfileReport
import pandas as pd
train = pd.read_csv('../datasources/titanic/tren.csv')
prof = ProfileReport(tren)
prof.to_file(output_file='informe.html')
```

A continuación, la librería proporciona un archivo HTML con los resultados de los análisis del conjunto de datos.

3.1 Análisis básicos

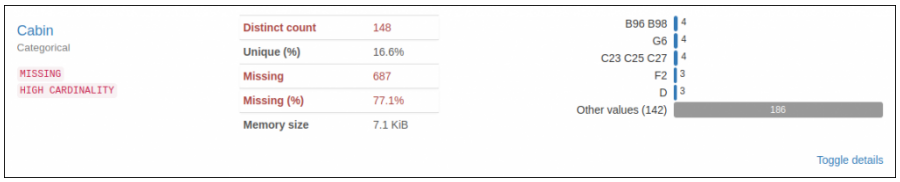
La primera pantalla proporcionada por la librería ofrece una visión general del conjunto de datos, que incluye:

- el número de filas y columnas del conjunto de datos,
- el número de datos que faltan,
- el número de líneas duplicadas,
- información sobre el espacio de memoria ocupado por estos datos,
- información sobre los tipos de datos que se encuentran en las diferentes columnas (encontramos la noción de variable categórica CAT, numérica o booleana/binaria).



Resultado proporcionado por Pandas Data Profiling

Lo que es aún más interesante es que en la parte inferior encontrará algunas estadísticas sobre los datos que faltan en cada columna (la columna Cabina, por ejemplo, tiene un 77,1% de datos que faltan). El análisis no se detiene ahí, por supuesto, y le permite profundizar en el análisis por columnas.



Detalle del análisis de columnas proporcionado por Pandas Data Profiling