Chapitre 4
Analyser et fiabiliser les données

1. Introduction

L'objectif de ce chapitre est de passer en revue les outils et moyens à mettre en œuvre afin de mieux comprendre et analyser les données. En effet, une donnée peut avoir différentes facettes dont celle d'être intimement liée à un contexte. Une donnée peut avoir une interprétation dans un contexte donné et une autre totalement différente – voire opposée – dans une autre perspective. Par ailleurs, une donnée a une vie et peut varier, s'altérer dans le temps ou tout simplement subir des changements lors de son transport ou dans son support de stockage.

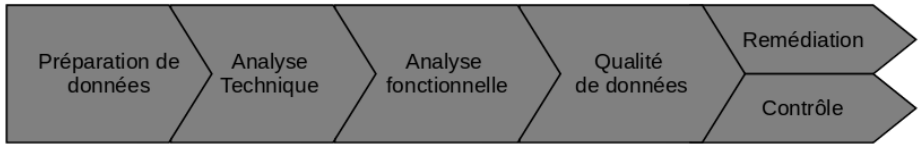
Avant de pouvoir utiliser une donnée, il est donc important de l'analyser afin de vérifier qu'elle correspond à nos attentes au moment de son utilisation. Imaginez que l'on récupère des jeux de données non documentés et non expliqués. Dans ce cas précis, il sera indispensable de passer par la case analyse. Une bonne pratique est de s'assurer que les données que l'on va utiliser sont bel et bien conformes, et c'est tout l'objet de ce chapitre.

Nous verrons tout d'abord comment analyser nos données sous un prisme technique ou structurel : c'est ce que l'on appelle le profilage de données (ou le Data Profiling). Cette analyse se focalise principalement sur les types, formats, nombre d'occurrences de la donnée et ne demande aucune connaissance particulière sur la donnée. Cette phase a pour objectif de poser un état des lieux factuel sur les composantes structurelles de la donnée et va faire ressortir les caractéristiques quantifiables que l'on peut extraire du jeu de données.

Ensuite, nous verrons comment analyser d'un point de vue fonctionnel et donc plus qualitatif les données. L'aspect quantitatif sera aussi possible mais dans ce cas il devra être adapté à un contexte métier. Nous nous focaliserons particulièrement dans cette partie sur la manière de présenter nos données pour les rendre plus parlantes.

Pour terminer, et tel un médecin ayant posé un diagnostic (grâce aux phases d'analyses précédentes), nous aborderons les moyens curatifs (contrôle, remédiation, etc.) qu'il faut mettre en œuvre afin de fiabiliser nos données. C'est la qualité de données. Cette dernière étape primordiale préfigure en quelque sorte un premier grand pas vers une démarche de gouvernance de données réussie.

Voici de manière très pragmatique les différentes étapes à suivre dans un projet de données :



Étapes de qualification de la donnée

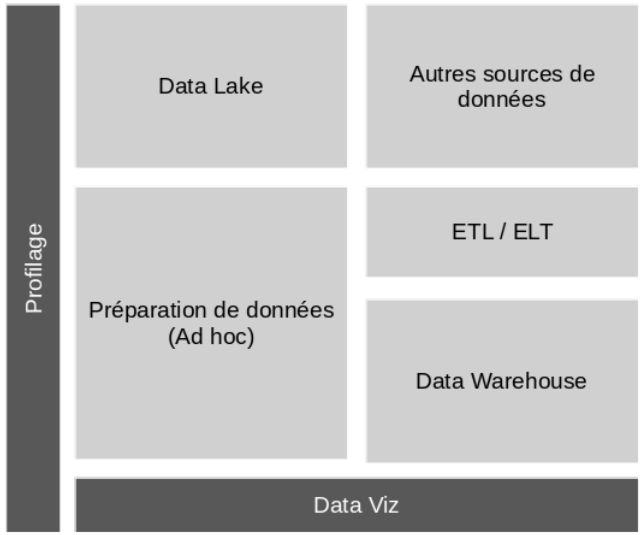
2. La préparation de données

Malheureusement les données ne sont pas toujours prêtes à l'emploi. Telles quelles, il est rarement possible d'en extraire afin de produire les résultats attendus (comme l'évolution d'un chiffre d'affaires, le calcul d'un indicateur de satisfaction, etc.). Dans l'immense majorité des cas, il va falloir retravailler les données brutes afin de les rendre exploitables dans la perspective d'une analyse, d'une alimentation ou même d'une modélisation (Machine Learning).

Une chose est certaine, si cette phase est en quelque sorte la face cachée de l'iceberg, ce n'en est pas moins une étape importante et surtout qui peut s'avérer très chronophage si elle n'est pas effectuée avec les bonnes compétences et ressources. On dit par exemple que les analystes de données et autres data scientistes passent plus de 80 % de leur temps à préparer leurs données. Il ne leur reste donc plus que 20 % de leur temps pour réaliser leur travail sur les données.

La préparation de données consiste donc à collecter et transformer les données afin que l'on puisse travailler dessus. On reconnaît bien là les deux premières phases de notre fameux ETL (voire ELT). Et pour cause la démarche est très similaire. À ceci près que la finalité est souvent très différente. Dans une démarche de type ETL, les données sont au final transportées vers une (ou plusieurs) source de données cibles, ce dans une perspective large d'analyse. Dans le cadre de la préparation de données, l'objectif est plutôt de préparer les données dans un objectif précis.

La tendance que l'on constate est que les ETL restent totalement attachés à des cas d'utilisation de type transport de données (alimentation de Data Warehouse, Data Lake, migration de données, etc.) et que des outils de préparation de données – récupérant par ailleurs les mêmes fonctionnalités de transformations – proposent directement ces fonctionnalités à des utilisateurs métier (jusqu'aux Data Scientists). Ainsi ces consommateurs de données peuvent travailler directement sur leurs données, voire même directement récupérer les données au sein de Data Lake afin de les filtrer, transformer, etc. En quelque sorte, il leur est dorénavant possible de les préparer comme bon leur semble.



Analyses et Initiatives sur la donnée

L'idée est simple : permettre un libre accès (potentiellement sécurisé) à toutes les données, mais aussi proposer les bons outils aux utilisateurs qui connaissent les données. Voilà qui permettra de bien commencer l'analyse de données, mais nous verrons plus tard que l'étape de préparation de données peut être étendue à de plus vastes desseins comme la modélisation en Machine Learning.

D'un point de vue concret et de manière similaire aux ETL, on trouve cinq grandes phases à la préparation de données :

– **Importation ou acquisition des données** : cette étape nécessite une connectivité aux divers systèmes sources.
– **Découverte** : une préanalyse commence bien souvent par un appel aux fonctionnalités de profilages de données.
– **Nettoyage des données** : un premier nettoyage de données est couramment nécessaire. Par exemple, des formats qui ne correspondent pas (typiquement les dates sont souvent un écueil) mais aussi des catégories ayant des problèmes de cohérence, etc. Il faut donc aligner les données pour pouvoir les analyser ultérieurement.

– **Enrichissement** : à ce niveau, il peut être intéressant d'ajouter des données annexes (c'est-à-dire des données qui ne proviennent pas de la source de données). Si on a des données de localisation, pourquoi ne pas y ajouter des données démographiques ?
– **Publication** : c'est la mise à disposition de vos données préparées pour l'outil qui va être chargé de l'utiliser. Si l'objectif est d'analyser ces données, peut-être doit-on les exporter dans un format particulier, ou alors peut-être faut-il regarder du côté de la solution de préparation de données si elle peut les mettre nativement à disposition ?

3. Analyse descriptive

Notons avant toute chose que l'analyse de données nécessite que les données soient mises à disposition sous un format tabulaire (lignes et colonnes). Ce chapitre concerne donc les données structurées. À l'heure actuelle, toutes les solutions (ou presque) fonctionnent avec des données structurées de la sorte. Une fois mises à disposition sous ce format, l'idée est d'analyser d'un point de vue technique les données fournies, de les décrire et d'identifier, pourquoi pas, un premier niveau d'exceptions (par exemple la détection d'outliers).

Cette étape a plusieurs noms : on l'appelle analyse descriptive ou profilage de données (Data Profiling).

Cette analyse doit décrire l'échantillon de données – ou tout le jeu de données – en effectuant le profil de chaque colonne pour y découvrir les informations importantes sur les attributs, notamment la fréquence et la distribution des valeurs de données, les formats, les patterns et les valeurs nulles, minimums et maximums. Toutes les données sont donc lues pour fournir une analyse et un état des lieux exhaustifs.

Mais les outils ou solutions permettent souvent d'aller bien plus loin en matière d'analyse. Voyons en détail les différentes analyses qu'il est possible de réaliser sur un jeu de données sans pour autant en avoir une connaissance fonctionnelle. Pour ce faire nous pourrions utiliser des outils No Code tels qu'Informatica, Talend ou SAS DataFlux qui, avec un simple clic, permettent d'obtenir ce type de résultat. Mais nous allons plutôt utiliser ici la librairie Python Pandas Profiling, afin que chacun puisse se frotter à ce type d'analyse sur son ordinateur.

Pour installer la librairie, il suffit de lancer la commande pip :

```
$ pip install pandas_profiling
```

Ensuite, il suffit d'ouvrir un dataframe avec la librairie Pandas et de lancer le profiling :

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

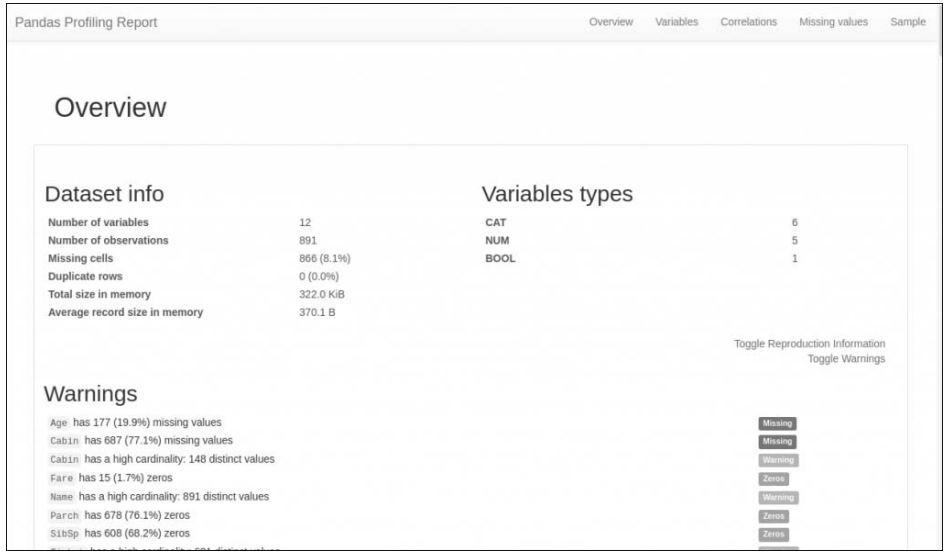
La librairie fournit alors un fichier HTML avec le résultat des analyses sur le jeu de données.

3.1 Analyses basiques

Le premier écran fourni par la librairie donne un aperçu global du jeu de données.

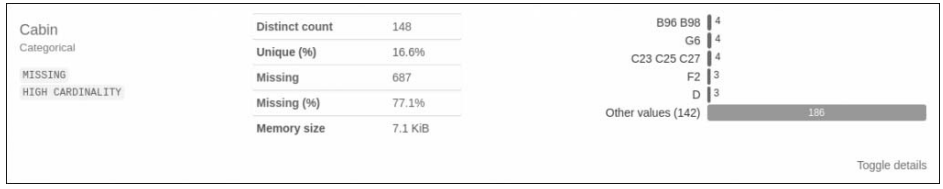
On y trouve :

– le nombre de lignes et de colonnes du jeu de données ;
– le nombre de données manquantes (non renseignées) ;
– les nombres de lignes en double ;
– des informations sur l'espace mémoire occupé par ces données ;
– des informations sur les types de données constatés sur différentes colonnes (on retrouve la notion de variable Catégorielle CAT, numérique ou booléenne/binaire).



Résultat fourni par Pandas Data Profiling

Beaucoup plus intéressant, on trouve dans la partie du bas quelques statistiques sur les données manquantes pour chaque colonne (la colonne Cabin par exemple 77,1 % de données manquantes). L'analyse ne s'arrête pas là, bien sûr, et permet d'aller en profondeur sur l'analyse par colonne.



Détail d'analyse d'une colonne fourni par Pandas Data Profiling