Chapitre 04 :
Les données structurées et leur traitement

1 Qu'est-ce qu'une donnée

Les « données » sont à la base de l'informatique des données, car elle est base justement sur le traitement de ces données afin d'en extraire des informations utiles ou de les transformer, rassembler, d'en déduire un raisonnement ou une prédiction.

En informatique, tout est donné, depuis les 0 et les 1 qui décrivent l'état des transistors dans un circuit électronique, jusqu'à une vidéo, en passant par les photos, les adresses, un relevé de température ou l'âge d'une personne. Les données sont souvent rassemblées pour caractériser un objet comme l'adresse d'une personne (composée du numéro de rue, du nom de la rue, du code postal, de la ville et du pays par exemple).

Lorsque les données sont ainsi rassemblées pour décrire quelque chose avec plusieurs informations, on parle de données structurées.

La conservation des données est un enjeu qui existe depuis l'aube des civilisations, bien avant l'informatique, car on peut considérer que les textes de loi, les comptes et la mémoire des évènements historiques sont autant de données qu'il a fallu faire passer de génération en génération (tablettes d'argile, parchemins, livres manuscrits, imprimerie…).

Lors de l'avènement du traitement informatique des données, celles-ci ont d'abord été conservées sur des cartes perforées avec un système de lecture optique, avant de passer sur des supports magnétiques (bandes, disques durs), puis à des supports optique (CD, DVD, Bluray) avant de passer à des stockages dans des cellules mémoires (non volatiles) pour les systèmes actuellement utilisés dans les systèmes informatiques (cartes mémoires sd, …).

Ces stockages de données sont de plus en plus rapides pour la lecture et l'écriture et leurs capacités augmentent très vite à mesure que toutes les informations analogiques de notre vie sont numérisées pour un traitement informatique de plus en plus massif.

2 Formats de stockage des données

Même si, au final, toutes les données numérisées vont être conservées en binaire (0 et 1) qui correspond au format traité par les ordinateurs, pour qu'elles soient faciles à traiter, elles vont être organisées en fonction de leurs types.

Pour des données qui doivent pouvoir être triées, recoupées et traitées ultérieurement pour en faire des rapports, des analyses, des graphiques… on utilise principalement des tableaux ou des listes. Ceux-ci peuvent être inscrits dans des fichiers textes lisibles avec un indicateur de séparation entre champs de données. Le plus courant est le séparateur par virgule (comma en anglais) : '**comma separated values'** (**csv**). Ce format convient bien pour des petites collections de données. Pour de plus grosses quantités, on utilisera des ensembles de tables, reliées entre elles par des règles et constituants des « **bases de données** » (database).

Quand les données sont plus spécifiques, on utilise de nombreux autres formats de stockage identifiés par leurs extensions : PNG, JPEG, HEIF… pour des images ; MP3, WAV, M4A… pour des sons ; MP4, AVI, M4V… pour des vidéos…

À cette extension est associé une structure logique des données et un en-tête de fichier qui permettra à un programme d'avoir des détails sur les informations conservées dans le fichier.

Par exemple, un fichier vidéo enregistré sur un téléphone portable contiendra les informations suivantes :

```
   0 0000001C 66747970 6D703432 00000001 69736F6D 6D703431 6D703432 ....ftypmp42....isommp41mp42
  28 00000001 6D646174 00000000 0199A702 000002A1 0605FFFF 9DDC45E9 ....mdat.....ôß....°..ÝÜEÈ
  56 BDE6D948 B7962CD8 20D923EE EF783236 34202D20 636F7265 20313532 ΩÊŸH∑ñ,ÿ Ÿ#ÓÔx264 - core 152
  84 202D2048 2E323634 2F4D5045 472D3420 41564320 636F6465 63202D20  - H.264/MPEG-4 AVC codec -
 112 436F7079 6C656674 20323030 332D3230 3137202D 20687474 703A2F2F Copyleft 2003-2017 - http://
 140 7777772E 76696465 6F6C616E 2E6F7267 2F783236 342E6874 6D6C202D www.videolan.org/x264.html -
 168 206F7074 696F6E73 3A206361 6261633D 31207265 663D3320 6465626C  options: cabac=1 ref=3 debl
 196 6F636B3D 313A303A 3020616E 616C7973 653D3078 333A3078 31313320 ock=1:0:0 analyse=0x3:0x113
 224 6D653D68 65782073 75626D65 3D372070 73793D31 20707379 5F72643D me=hex subme=7 psy=1 psy_rd=
 252 312E3030 3A302E30 30206D69 7865645F 7265663D 31206D65 5F72616E 1.00:0.00 mixed_ref=1 me_ran
 280 67653D31 36206368 726F6D61 5F6D653D 31207472 656C6C69 733D3120 ge=16 chroma_me=1 trellis=1
 308 38783864 63743D31 2063716D 3D302064 6561647A 6F6E653D 32312C31 8x8dct=1 cqm=0 deadzone=21,1
 336 31206661 73745F70 736B6970 3D312063 68726F6D 615F7170 5F6F6666 1 fast_pskip=1 chroma_qp_off
 364 7365743D 2D322074 68726561 64733D31 32206C6F 6F6B6168 6561645F set=-2 threads=12 lookahead_
 392 74687265 6164733D 3220736C 69636564 5F746872 65616473 3D30206E threads=2 sliced_threads=0 n
 420 723D3020 64656369 6D617465 3D312069 6E746572 6C616365 643D3020 r=0 decimate=1 interlaced=0
 448 626C7572 61795F63 6F6D7061 743D3020 636F6E73 74726169 6E65645F bluray_compat=0 constrained_
 476 696E7472 613D3020 62667261 6D65733D 3320625F 70797261 6D69643D intra=0 bframes=3 b_pyramid=
 504 3220625F 61646170 743D3120 625F6269 61733D30 20646972 6563743D 2 b_adapt=1 b_bias=0 direct=
 532 31207765 69676874 623D3120 6F70656E 5F676F70 3D302077 65696768 1 weightb=1 open_gop=0 weigh
 560 74703D32 206B6579 696E743D 32353020 6B657969 6E745F6D 696E3D32 tp=2 keyint=250 keyint_min=2
 588 33207363 656E6563 75743D34 3020696E 7472615F 72656672 6573683D 3 scenecut=40 intra_refresh=
 616 30207263 5F6C6F6F 6B616865 61643D34 30207263 3D637266 206D6274 0 rc_lookahead=40 rc=crf mbt
```

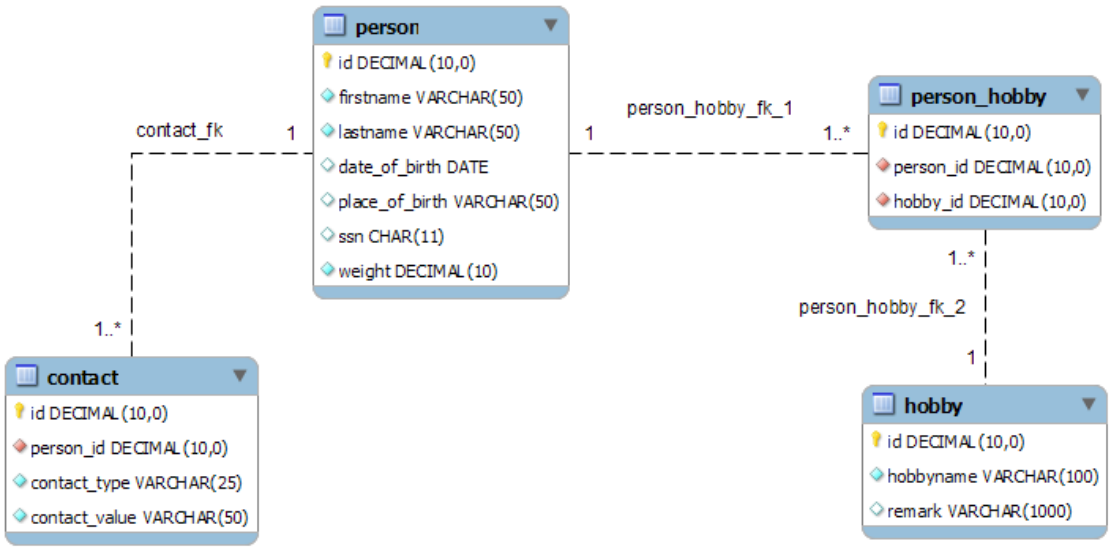
Les colonnes de gauche contiennent le codage du fichier en hexadécimal (comptage en base 16 très utilisée en informatique) et à droite sa traduction en ASCII (american standard code for information interchange) qui permet de lire ce contenu « en clair ». On constate que le début du fichier contient des informations sur le type de codage utilisé pour la vidéo (H.264/MPEG-A AVC Codec) suivie de nombreuses informations nécessaires au décodage de cette vidéo.

Toutes ces informations sont nécessaires pour que le fichier puisse être ouvert et exploité par d'autres ordinateurs. On parle alors **d'interopérabilité.**

3 Données structurées et traitement

3.1 Description des données structurées : descripteurs et valeurs

On parle de données structurées quand un ensemble de données donne des clefs d'accès simples aux données qu'il contient. C'est typiquement le cas d'une base de données qui contient des tableaux et des clefs d'indexation permettant d'identifier rapidement chaque ensemble de données (comme un numéro d'article ou un numéro de client), mais aussi d'un fichier csv qui contient des identificateurs de colonne permettant un tri rapide.



Exemple de la structure d'une base de données d'association (source Wikipédia).

Dans l'image ci-dessus, on voit que chaque information sur une personne de la base est identifiée par un **descripteur** qui décrit ce qu'elle doit contenir (firstname – prénom ; lastname – nom ; date-of-birth : date de naissance….) et chacun de ces champs est défini par un **type** précis (varchar(50) : 50 caractères libres ; Date…). Pour chaque utilisateur il y aura une ligne dans le tableau « person » et cette ligne contiendra les valeurs saisies lors de l'enregistrement d'un nouveau membre de l'association.

3.2 Récupérer des données structurées

La protection des données personnelles fait que de nombreuses informations précises ne sont heureusement pas accessibles librement sur Internet. Il existe toutefois des sites d'information ouverts regroupant des bases de données à usage publiques : les Open Data.

Le site https://data.education.gouv.fr permet par exemple d'accéder à de très nombreuses informations générales sur le fonctionnement de l'éducation nationale en France (budget, élèves, réussites scolaires…)

3.3 Recherches, tri et calculs dans des tables de données

En choisissant un ensemble de données il est possible d'y effectuer de **recherches** spécifiques, de mettre en place un **filtre** (par année par exemple), puis de cliquer sur les colonnes du tableau pour effectuer un **tri** (croissant ou décroissant).

Il est également possible de récupérer les données au format csv afin de les utiliser pour effectuer des **calculs** ou des **analyses** graphiques en créant des représentations à partir des données.

4 Métadonnées (Big Data)

Lorsque les données stockées ne sont pas du texte et donc pas facilement lisibles par un ordinateur pour effectuer une recherche ou tri (par exemple), on y associe souvent des métadonnées qui vont permettre de faciliter la classification et les opérations sur les données.
Ces métadonnées sont généralement stockées dans l'en-tête du fichier et parfois accessible en étudiant les propriétés du fichier (selon le système d'exploitation).
Par exemple pour un fichier audio musical sur Mac :

On voit ici que ce fichier a été acheté sur le site musical Qobuz et de nombreuses autres informations utiles pour faire une rechercher : titre du morceau, type de codage (Audio Apple MPEG-4), fréquence d'échantillonnage (44,1 kHz), artiste, label…

Ces métadonnées sont de simples informations textuelles qu'il est possible de modifier facilement avec un éditeur dédié. C'est même parfois nécessaire lorsqu'on partage une photo ou une vidéo afin d'éviter de diffuser en même temps votre nom, l'endroit où elle a été prise, la marque de votre téléphone ou d'autres informations confidentielles (plus d'informations à ce propos dans les prochains chapitres).

5 Données dans le nuage

5.1 Le stockage dans le nuage

Le « nuage » (ou « cloud ») est basé sur l'image d'Internet vu comme une couverture mondiale (nuageuse) accessible de partout, tout le temps et avec de nombreux types d'appareils différents. Les serveurs qui y sont connectés sont accessibles en permanence.

Ces serveurs ont d'abord servi à héberger des sites internet, des mails et d'autres types de fichiers précis. Avec l'augmentation importante des débits internet dans le monde, des sociétés ont commencé à proposer d'utiliser des serveurs afin de proposer un stockage des données « dans le nuage », hors de chez vous et avec un accès à Internet.

De nombreux prestataires proposent maintenant ces services, comme Apple (iCloud) et Microsoft (OneDrive) qui les ont intégrés dans leurs systèmes d'exploitation, ou des sociétés tierces comme DropBox qui offrent ce service moyennant un abonnement annuel.

Il est généralement nécessaire d'installer un logiciel sur l'ordinateur et c'est lui qui va se charger de gérer l'accès aux données. Le tarif de l'abonnement va ensuite dépendre de la quantité de données que l'on souhaite stocker.

Ces services sont de plus en plus utilisés en entreprise et ils permettent aussi de déporter les fichiers volumineux de certains serveurs web vers des serveurs ayant de grosses capacités de stockage et des grands débits internet. Des sociétés comme Akamai ou Amazon proposent ainsi des stockages rapides et sécurisées pour de très gros clients comme Microsoft, Apple ou Google (Youtube).

5.2 Partage et synchronisation de données

Lorsque les données sont stockées « dans le nuage », elles sont généralement aussi dupliquées sur votre espace de stockage local (si celui est assez volumineux et si vous le souhaitez). Tout nouveau fichier que vous enregistrez dans l'espace synchronisé du service auquel vous êtes abonné sera immédiatement dupliqué sur le serveur distant (si vous avez un accès à Internet !).

Cela assure que vos données ne soient pas perdues si vous vous faites voler (ou si vous cassez) votre ordinateur portable ou votre téléphone. Le prestataire que vous avez payé s'engage également à effectuer des sauvegardes régulières de vos données afin qu'il n'y ait pas de risque qu'elles soient perdues en cas de panne du serveur. Ces sauvegardes sont parfois multiples et vous permettent d'accéder à d'anciennes versions de vos fichiers en cas de modifications malencontreuses (cela dépend du « cloud » choisi).

Des options vous permettent également de rendre publics les fichiers que vous souhaitez partager et d'obtenir un simple lien que vous pourrez transmettre à vos correspondants. C'est très pratique pour envoyer de gros fichiers qui sont déjà sur votre espace de stockage dans le nuage : vous ne transmettez que l'adresse du fichier et celui-ci est accessible autant de temps que vous le voulez sans que vous ayez besoin de laisser votre ordinateur allumé.

5.3 Enjeux énergétiques et climatiques

Qui dit stockage sur des serveurs allumés en permanence, dit aussi grosse consommation d'énergie. C'est l'un des problèmes les plus alarmants de notre usage du numérique qui consomme plus de 10% de la production d'énergie dans le monde.

30% de cette énergie part dans le fonctionnement des serveurs et 40% dans le fonctionnement d'internet en lui-même (relais, routeurs, convertisseurs fibre-optique…). Selon Françoise Berthoud (informaticienne au Gricad), l'envoi d'un mail de 1Mo correspond à l'émission de 20g de CO_2 et une requête sur Internet correspond à 7g de CO_2 émis. Autant dire que les émissions deviennent vite énormes lorsqu'on utilise un enregistrement dans le cloud.

Il faut en effet alimenter les serveurs en énergie et surtout les refroidir, car la grande concentration d'appareils électriques dans les centres serveurs produit beaucoup de chaleur qui est nuisible aux ordinateurs (qui font des erreurs de calcul quand il fait trop chaud et risquent même de fondre si aucun refroidissement n'est prévu). Il est alors nécessaire d'installer des systèmes de climatisation qui consomment encore plus d'énergie.

Il existe des solutions pour réduire cette consommation en utilisant la chaleur générée par les centres serveurs pour réchauffer des habitations ou l'eau des piscines, mais cela nous interroge aussi sur notre usage souvent futile des réseaux (sociaux par exemple) alors que le réchauffement de la planète devient de plus en plus préoccupant et que l'effort de tous est nécessaire si on veut éviter des catastrophes climatiques de plus en plus dramatiques.