

Thème 2 : Les données structurées et leur traitement

L'évolution des capacités de stockage, de traitement et de diffusion des données fait qu'on assiste aujourd'hui à un phénomène de surabondance des données et au développement de nouveaux algorithmes capables de les exploiter. L'exploitation de données massives (Big Data) est en plein essor dans des domaines aussi variés que les sciences, la santé ou encore l'économie.

I. Qu'est-ce qu'une donnée

Les « données » sont à la base de l'informatique, car toute l'informatique est justement 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 (voir plus loin).

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 événements historiques sont autant de données qu'il a fallu faire passer de génération en génération (tablettes d'argile, parchemins, livres manuscrits, imprimerie...).

II. 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...

III. Métadonnées

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 :

IV. Données dans le nuage

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).

V. 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 CO₂ et une requête sur Internet correspond à 7g de CO₂ 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.