

SNT synthèse données structurées

Version pdf

Repères historiques

- 1930 : utilisation des cartes perforées, premier support de stockage de données ;
- 1956 : invention du disque dur permettant de stocker de plus grandes quantités de données, avec un accès de plus en plus rapide ;
- 1970 : invention du modèle relationnel (E. L. Codd) pour la structuration et l'indexation des bases de données ;
- 1979 : création du premier tableur, VisiCalc ;
- 2009 : Open Government Initiative du président Obama ;
- 2016 : Règlement Général sur la Protection des Données dans l'Union Européenne.

Données et informations



"Point de cours 1"



"Données"

Une **donnée** est une valeur décrivant un objet, une personne, un événement digne d'intérêt pour celui qui choisit de la conserver.



"Descripteurs"

Si on veut gérer une collection de chansons, on a besoin de décrire chaque objet chanson par les mêmes catégories de données qu'on appelle des **descripteurs**. La valeur d'un descripteur est une donnée et doit appartenir au domaine qui caractérise le descripteur, on peut ainsi comparer les données de deux chansons (leur année par exemple).

☰ "Données structurées"

Une collection regroupe des objets partageant les mêmes descripteurs (par exemple, une collection de chansons). La structure de **table** permet de présenter une collection : les objets en ligne, les descripteurs en colonne et les données à l'intersection. Les données sont alors dites **structurées**.

Descripteurs →

Une valeur du descripteur «DURÉE» →

Un objet →

ID	DURÉE	TITRE	STYLE	DATE	INTERPRÈTE
MORCEAU#1	04:34	ROUX COOL	REGGAE	1993	BOB
MORCEAU#2	04:34	LOVE BIRD	JAZZ	1956	ELSA
MORCEAU#3	03:25	RE VOLE LOVE	RAP	2017	MC
MORCEAU#4	08:45	BLUE PIGEON	ROCK	1987	PIXEES
MORCEAU#5	02:23	FLYING GRASS	JAZZ	1979	LITTLE PINK DOTS

Source : cours de [Cédric Gouygou](#)

☰ "Information"

Des données structurées peuvent être interprétées pour en extraire de l'information.
L'informatique est la science du traitement automatique de l'information.

✎ "Point de cours 2"

☰ "Format de fichier"

Pour assurer leur durabilité et leur partage, les données sont conservées sous différents **formats de fichiers**.

☰ "Format CSV"

Le format de fichier **CSV (Comma-Separated Values)** est un format de fichier texte utilisé pour enregistrer des données structurées en table :

- chaque titre de colonne correspond à un **descripteur**
- chaque ligne contient l'enregistrement des données d'un objet : les valeurs pour chaque descripteur sont séparées par un symbole (virgule, point virgule ou tabulation selon le type de format CSV).

Un exemple de table avec deux descripteurs, trois enregistrements et le symbole ',' comme séparateur :

```
Nom, Prénom, Âge
Yamal, Lamine, 17
Mbappé, Kylian, 25
Foden, Phil, 24
```

"Format ouvert"

Un format de données est considéré comme **ouvert** (ou libre) si son mode d'organisation est rendu public par son créateur et qu'aucune restriction légale (telles que le droit d'auteur, les brevets ou le copyright) n'empêche son utilisation libre.

"Interopérabilité"

Les formats ouverts sont souvent conçus pour favoriser **l'interopérabilité** : un document sauvegardé dans un format ouvert ne dépend pas du logiciel utilisé pour le créer, le modifier ou le lire. Ainsi, l'utilisateur peut choisir librement le logiciel qu'il souhaite utiliser.

"Métadonnées"

Les **métadonnées** d'un fichier sont des informations descriptives sur les données contenues dans ce fichier, telles que le titre, l'auteur, la date de création ou le format. Elles facilitent la recherche, la gestion et l'utilisation des données.

Machines



"Point de cours 3"

≡ "Bandes magnétiques (1950)"

Les **bandes magnétiques** utilisent une bande de matériau magnétisable pour enregistrer les données de manière séquentielle, c'est-à-dire que pour atteindre une donnée il faut parcourir toutes celles stockées précédemment.

≡ "Disquettes (années 1970)"

Les **disquettes** sont des supports de stockage magnétique portables avec des capacités limitées, utilisées pour le transfert et le stockage de données sur les premiers ordinateurs personnels.

≡ "Disques durs (années 1980)"

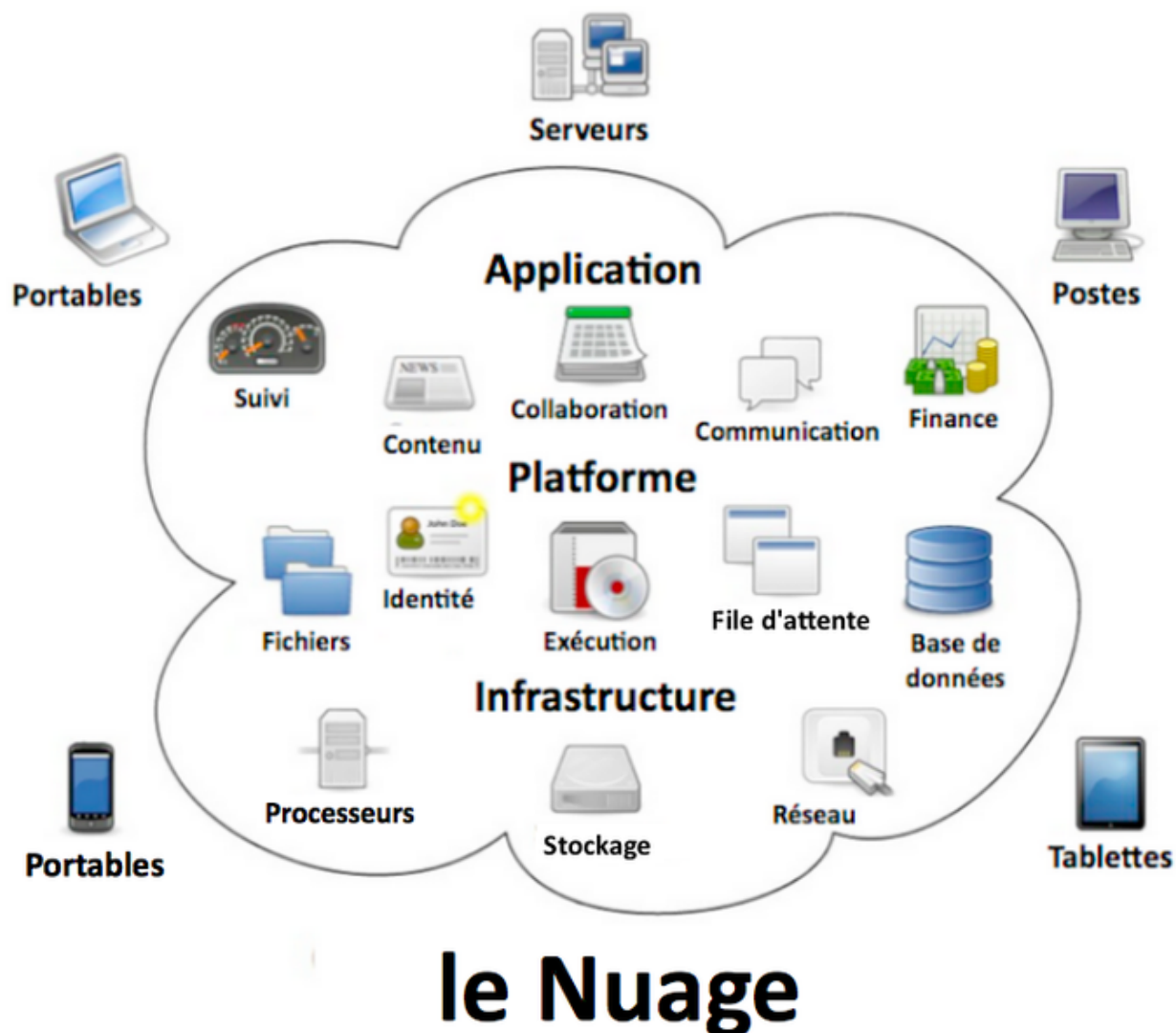
Inventé en 1956, le **disque dur** utilise des plateaux rigides recouverts de matériau magnétique pour stocker les données. Offrant de grandes capacités et des temps d'accès rapides, les disques durs accompagnent la démocratisation de l'informatique dans les années 1980 et deviennent le principal support de stockage interne pour les ordinateurs.

≡ "SSD (années 2000)"

Les **mémoires SSD** utilisent une technologie de mémoire flash à base de transistors, pour stocker les données. Elles offrent des vitesses de lecture/écriture supérieures à celles des disques magnétiques et une faible consommation d'énergie.

≡ "Stockage en ligne (années 2010)"

Le **stockage en ligne**, ou "dans le nuage", permet de stocker des données sur des serveurs distants accessibles via Internet, offrant une capacité évolutive et un accès global, facilitant la collaboration et les sauvegardes automatisées.



Source : Sam Johnston, Wikimedia Commons

Algorithmes et programmes



"Point de cours 4"



"Recherche dans une table"

Une table de données structurées peut faire l'objet de différentes opérations :

- rechercher une information précise dans la collection
- trier la collection selon un ou plusieurs descripteurs
- filtrer la collection selon un ou plusieurs tests sur les valeurs des descripteurs
- effectuer des calculs
- mettre en forme les informations produites pour une visualisation par les utilisateurs.

"Recherche dans une base de données"

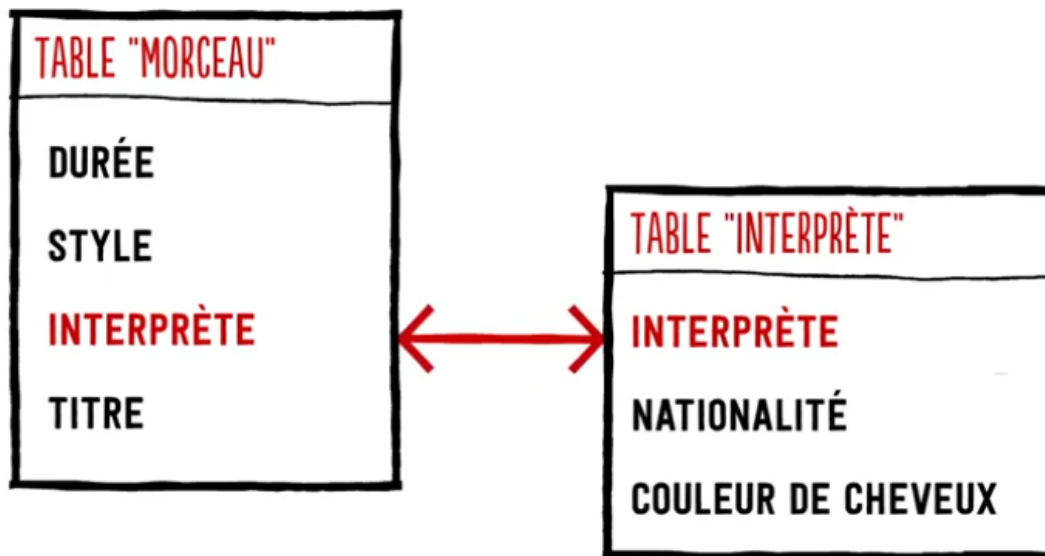
Une **base de données** est une collection de plusieurs tables de données structurées qu'on peut rapprocher sur les valeurs de descripteurs communs pour générer de nouvelles informations.

Les *Systèmes de Gestion de Bases de Données (SGBD)* sont très utilisés pour gérer les données structurées : abonnés d'une plateforme, stocks et commandes, systèmes de réservation ...

Les recherches dans une base de données sont effectuées à l'aide de **requêtes**.

"Exemple de la gestion d'une base de données musicale"

Pour gérer une base de données de morceaux de musique, il est préférable d'utiliser plusieurs tables plutôt qu'une seule. Séparer les données des morceaux de celles des interprètes permet de ne pas ressaisir le nom de l'interprète pour chaque morceau. En éliminant les redondances, on réduit les risques d'erreurs de saisie ou de modification.



Impact sur les pratiques humaines



"Point de cours 5"



"Big Data"

L'évolution des capacités de stockage, de traitement et de diffusion conduit à une explosion du volume de données disponibles. On parle de **Big Data**.

Ces données massives peuvent être exploitées grâce à l'augmentation de la puissance de calcul et aux progrès des systèmes d'Intelligence Artificielle, dont l'entraînement nécessite également de gros volumes de données.



"Open Data"

Certaines de ces données, en particulier celles produites par les services publics, sont ouvertes (**Open Data**) et considérées comme des biens communs. Deux beaux exemples

de réutilisation de ces données ouvertes : [Covidtracker](#) outil de visualisation de la progression de l'épidémie très populaire en 2020-2021 et [Suptracker](#) plateforme de visualisation de données sur l'orientation postbac.

☰ "RGPD"

L'exploitation de données personnelles par les grandes plateformes (GAFAM) nécessite un cadre juridique transfrontalier qui commence à se mettre en place comme le [Règlement Général sur la Protection des Données \(RGPD\)](#) dans la communauté européenne.

L'exploitation de données personnelles par des régimes autoritaires peut constituer une menace pour les libertés individuelles.

☰ "Impact environnemental"

Les **centres informatiques (data center)** qui hébergent l'infrastructure de stockage de l'internet mondial, sont responsables de 7 à 15 % des impacts du numérique. D'après un rapport commandé par le gouvernement à l'association [Green IT](#) en 2020, l'empreinte environnementale du numérique mondial est environ trois fois celle de la France.

%	Energie	GES	Eau	Elec.	Ress.
Utilisateurs	60	63	83	44	75
Réseaux	23	22	9	32	16
Centres informatiques	17	15	7	24	8

Illustration : répartition des impacts en 2019

Source : Greenit