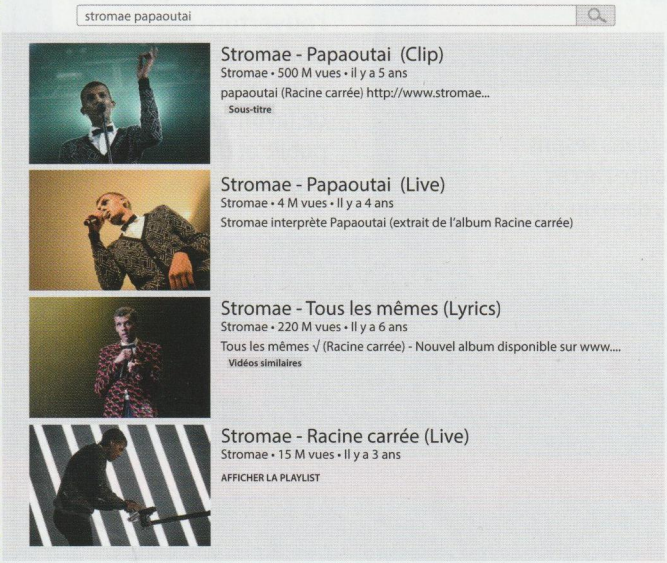
ACTIVITÉ

1 DES DONNÉES NON STRUCTURÉES AUX COLLECTIONS

Les applications que nous utilisons quotidiennement doivent manipuler de grandes quantités de données afin de répondre à nos requêtes (recherche d'une information sur Google, d'une vidéo sur YouTube, etc.). Structurer les données permet aux applications de les exploiter efficacement afin de répondre au mieux à nos attentes.

1 Le résultat d'une recherche dans YouTube



▲ **Une collection de propositions suite à une recherche lancée pour *Stromae Papaoutai* (simulation).**

Lorsqu'on lance la recherche *Stromae Papaoutai* sur YouTube, on obtient des propositions de clips pertinentes, c'est-à-dire en lien avec le clip demandé (d'abord, la vidéo précisément demandée, puis, par exemple, une version live, d'autres vidéos du chanteur, etc.).

Chaque clip proposé s'affiche avec des informations le décrivant (artiste, titre de la chanson, etc.). Pour arriver à ce résultat, il faut au préalable que des **données** aient été structurées.

VOCABULAIRE

Donnée

Valeur décrivant un objet, une personne, un événement digne d'intérêt pour celui qui choisit de la conserver. Par exemple, le numéro de téléphone d'un contact ou encore la chaîne de caractères *Jeanne* sont des données.

2 Comment ont été structurées les données ?

Artiste	Titre	Vues	Ancienneté
Stromae	Papaoutai	500 M	5 ans
Stromae	Papaoutai Live	4 M	4 ans
Stromae	Tous les mêmes	220 M	6 ans

◀ **Cette table est une représentation simple de la collection de propositions YouTube (doc. 1) : les objets en ligne, les descripteurs en colonne, et les données à l'intersection. Les données sont alors dites structurées.**

Pour faire des propositions pertinentes de vidéos (doc. 1), YouTube a sélectionné en amont des données parmi le très grand nombre disponible puis les a structurées selon des critères (nom de l'artiste, titre de la chanson, nombre de vues, etc.). La structuration de données peut se faire en créant une collection. Ci-dessus, la collection simplifiée d'une partie de la liste des propositions YouTube, sous la forme d'une table.

• Les valeurs dans les cases sont des données : *Stromae*, *Papaoutai* ou encore *500 M*. Les objets considérés ici sont des vidéos (ou clips) ; chacun est représenté par une ligne. Les descripteurs servent à décrire chaque objet : l'artiste, l'ancienneté de la mise en ligne, etc.

• Pour un objet donné (par exemple, la première ligne), à chaque descripteur (► *doc. 3*) correspond une valeur de l'objet, ce qui peut se formuler ainsi : *L'artiste Stromae a écrit une chanson dont le titre est Papaoutai, qui a été mise en ligne il y a 5 ans et a été vue 500 millions de fois.*

La collection des propositions du doc. 1 est modélisée par la collection des lignes de cette table.

3 La recherche dans une collection YouTube

Lorsqu'on lance la recherche musée d'Orsay dans YouTube, voici ce qui apparaît : une partie de la collection résultat de cette recherche, que YouTube vient de sélectionner après une recherche dans sa table sous-jacente qui contient la totalité de ses vidéos. On y reconnaît, entre autres, les **descripteurs** suivants : *titre de la vidéo, nombre de vues, ancienneté.*

VOCABULAIRE

Descripteur
Il sert à décrire un objet. Par exemple, le nom, le prénom, l'adresse et le numéro de téléphone sont des descripteurs d'un contact dans un agenda.

4 Un algorithme de recherche dans une collection

L'algorithme suivant parcourt une **collection** d'élèves et affiche la sous-collection de ceux de plus d'un certain âge. Il manipule une table et ses lignes.

```
Variables      c est une table dont les colonnes sont prénom et âge
               résultat est une table dont les colonnes sont prénom et âge
               e est une ligne de table dont les colonnes sont prénom et âge
               a est un entier
Entrée         saisir a
Initialisation  c prend comme valeur la table de tous les élèves de la classe
               résultat prend comme valeur l'ensemble vide
Traitement     Pour chaque ligne e de c faire
               | Si e.âge>= a
               | | Alors ajouter e à résultat
               | Fin Si
               Fin Pour
Sortie         Pour chaque ligne e de résultat faire
               | afficher e.prénom, e.âge
               Fin Pour
```

VOCABULAIRE

Collection
Elle regroupe des objets partageant les mêmes descripteurs (par exemple, la collection des contacts d'un carnet d'adresses). La structure de table permet de présenter une collection.

QUESTIONS

1 **a.** **Doc. 1, 2** Pour chaque objet de la table (doc. 2), indiquez chacun de ses descripteurs, la donnée correspondante et à quel objet du doc. 1 il correspond.

b. Ajoutez dans la table (doc. 2) une quatrième ligne correspondant à la dernière proposition de YouTube du doc. 1 et précisez ses descripteurs et ses données.

c. Observez le résultat de la recherche YouTube : quelles autres données n'apparaissent pas dans la table ? À quels descripteurs ces données correspondent-elles ?

2 **Doc. 1, 2** En vous appuyant sur les documents et sur vos réponses à la question 1, proposez des données structurées sous la forme d'une collection (représentée par une table) pour modéliser toutes les informations de l'emploi du temps de votre classe, avec un objet pour chaque créneau de cours (matière, professeur, salle, etc.). Puis construisez à la main une table contenant uniquement les jours, les créneaux et les salles de cours de mathématiques.

3 **Doc. 3, 4** Proposez les grandes lignes d'un algorithme qui pourrait être utilisé par YouTube pour effectuer une recherche comme dans le doc. 3. Il prendra en entrée un nom d'artiste, et travaillera sur la collection de tous les clips de YouTube.

ACTIVITÉ

2 LES PRINCIPAUX FORMATS

Pour enregistrer des données structurées, il faut d'abord organiser ces données en collections, puis il faut choisir le format dans lequel les collections seront stockées. Les formats CSV et JSON permettent de stocker facilement des relations.

1 Les données dans une table

a. Un extrait de la table des gares SNCF

Intitulé gare	Code postal	Département	Région SNCF
Alet-les-bains	11580	Aude	Occitanie
Bar-sur-Aube	10200	Aube	Grand Est
Bourg-en-Bresse	01000	Ain	Auvergne-Rhône-Alpes
Bram	11150	Aude	Occitanie
Coat Guégan	22390	Côtes-d'Armor	Bretagne
Couffoulens-Leuc	11250	Aude	Occitanie
Pomas	11250	Aude	Occitanie
Pont-Melvez	22390	Côtes-d'Armor	Bretagne
Verzeille	11250	Aude	Occitanie

Identifiant

◀ Un extrait de la table des gares telle qu'on la visualise dans un tableur. La collection complète (avec ses métadonnées) disponible sur le site SNCF Open Data.

VOCABULAIRE

Donnée ouverte
Donnée librement accessible et utilisable.

b. Comment sont organisées les données dans une table ?

Dans une table, chaque ligne représente un objet avec toutes ses informations (ici : une gare, son code postal, son département, sa région SNCF). Les collections comportent souvent au moins un descripteur qui caractérise de façon unique chaque objet. On appelle ce descripteur « **identifiant** » : ici, chaque ligne a un intitulé gare distinct ; c'est donc l'identifiant pour cette collection.

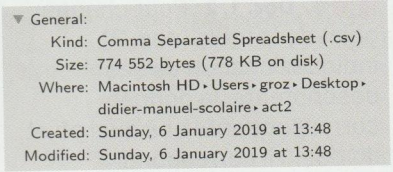
VOCABULAIRE

Identifiant
Il permet d'identifier un objet précis dans un ensemble d'objets.

2 La table des gares SNCF dans un fichier CSV

```
Intitulé gare,Code postal,Département,Région SNCF
Alet-les-Bains,11580,Aude,Occitanie
Bar-sur-Aube,10200,Aube,Grand Est
Bourg-en-Bresse,01000,Ain,Auvergne-Rhône-Alpes
Bram,11150,Aude,Occitanie
Coat Guégan,22390,Côtes-d'Armor,Bretagne
Couffoulens-Leuc,11250,Aude,Occitanie
Pomas,11250,Aude,Occitanie
Pont-Melvez,22390,Côtes-d'Armor,Bretagne
Verzeille,11250,Aude,Occitanie
```

Voici des **métadonnées** associées au fichier **CSV** ci-contre.



VOCABULAIRE

Format CSV
Dans les fichiers CSV (*Comma-Separated Values*), les valeurs sont séparées par des virgules. Certains fichiers CSV utilisent d'autres symboles : « ; » ou « | » ou encore des tabulations, par exemple.

VOCABULAIRE

Métadonnée
Littéralement, une « donnée sur une donnée ». Les métadonnées aident à manipuler et à interpréter les données qu'elles décrivent.

3 La table des gares SNCF dans un fichier JSON

```
[ {
    "Intitulé gare": "Alet-les-Bains",
    "Code postal": "11580",
    "Département": "Aude",
    "Région SNCF": "Occitanie" },
  {
    "Intitulé gare": "Bar-sur-Aube",
    "Code postal": "10200",
    "Département": "Aude",
    "Région SNCF": "Grand Est" },
  ... ]
```

VOCABULAIRE

Format JSON

Dans un fichier JSON, les descripteurs sont toujours entourés de guillemets. Le format JSON (*JavaScript Object Notation*) permet aussi de représenter des données plus complexes que des tables.
Le format JSON est souvent utilisé pour récupérer et échanger des données sur le Web.

4 Une répartition des données en deux collections

Intitulé gare	Code postal
Alet-les-bains	11580
Bar-sur-Aube	10200
Bourg-en-Bresse	01000
Bram	11150
Coat Guégan	22390
Couffoulens-Leuc	11250
Pomas	11250
Pont-Melvez	22390
Verzeille	11250

Code postal	Département	Région SNCF
01000	Ain	Auvergne-Rhône-Alpes
10200	Aube	Grand Est
11150	Aude	Occitanie
11250	Aude	Occitanie
11580	Aude	Occitanie
22390	Côtes-d'Armor	Bretagne

VOCABULAIRE

Base de données

Ensemble de collections reliées entre elles.

La redondance

En informatique, on essaie d'éviter la redondance, c'est-à-dire écrire deux fois la même information. Le département et la région SNCF étant les mêmes pour toutes les gares qui partagent le même code postal, il peut être judicieux de stocker dans une collection à part les informations qui ne dépendent que du code postal. La **base de données** ci-dessus contient les mêmes informations que celles du doc. 1 : pour retrouver le département de Pont-Melvez, il suffit de croiser les données des deux tables à travers le code postal.

QUESTIONS

1 *Limiter la redondance permet de réduire l'espace de stockage et de faciliter les mises à jour (moins de risques d'incohérences).*

a. **Doc. 1, 4** Quelle est la meilleure organisation entre ces deux documents lorsque beaucoup de gares ont le même code postal ? Et lorsqu'aucune paire de gares n'a le même code postal ? Justifiez vos réponses.

b. Sachant que toutes les gares du département ont la même région SNCF, comment pourrait-on mieux organiser les données ?

c. Selon vous, les deux requêtes ci-dessous seront-elles, a priori, plus efficaces sur le doc. 1 ou sur le doc. 4 ?
- Donner la liste des gares de la région Bretagne.
- Donner la liste des départements de la région Bretagne.

2 **Doc. 1, 2** Téléchargez la collection au format CSV sur le site des données ouvertes SNCF. Ouvrez le fichier dans un éditeur de texte, puis dans un tableur. Si nécessaire, précisez au tableur que le séparateur est ";". Pourquoi n'a-t-on pas utilisé le séparateur "," ?

3 **Doc. 1** Afficher le code source d'un courriel. Dans l'en-tête volumineux qui précède le texte du courriel, relevez des descripteurs. Faites le même exercice avec une page Web.

4 Consultez les propriétés d'un fichier PDF ou JPG et relevez des métadonnées.

ACTIVITÉ

3 OÙ SONT STOCKÉES LES DONNÉES ?

La quantité de données numériques est en constante augmentation. Par exemple, des sites comme Deezer ou Spotify stockent des pétaoctets de données. Comment assurer le stockage de ces données massives ? Quel en est l'impact sur l'environnement ?

1 L'ère du « Big Data »

Nos activités (écoute de musique en ligne, partage d'images, capteurs d'objets connectés, etc.) génèrent la circulation d'énormes quantités de données. Pour extraire les informations pertinentes rapidement, on peut procéder à l'indexation des données (ce que font les moteurs de recherche ► *p. 38*). On utilise aussi, de plus en plus, des algorithmes d'apprentissage pour analyser ces données (reconnaissance de personnes sur des images, programmation de véhicules autonomes, etc.). Cela demande de grandes capacités, à la fois en stockage et en calcul, et nos activités exigent que les données soient accessibles en permanence par Internet dans des délais très courts. Des data centers se sont développés partout dans le monde pour répondre à ces besoins.

Quelques chiffres clés

Dans le monde

9 milliards d'appareils
2 milliards de smartphones
1 milliard d'ordinateurs
5 à 7 milliards d'objets connectés

En 1 heure

8 à 10 milliards de mails échangés (hors spams)
180 millions de recherches Google

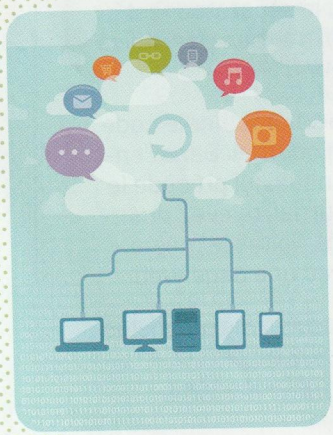
Total des données stockées dans le monde

Des dizaines de zettaoctets (10^{21})
Par exemple, le site Alibaba stocke 1 exaoctet (10^{18})

D'après ademe.fr, *La face cachée du numérique*, 2017.

2 Les supports de stockage

Les fichiers sont toujours stockés sur des supports magnétiques, par exemple un disque dur d'ordinateur, une clé USB, une carte de téléphone portable, etc., qui permettent un stockage local. Ces supports peuvent subir des altérations, même dans le cas d'une utilisation normale, d'où l'importance de faire des copies pour sauvegarder les données.



Plutôt que de stocker ou traiter les données localement, les applications font de plus en plus souvent appel au **cloud**, un système de stockage en ligne : des entreprises mettent à la disposition des clients des machines permettant de stocker des données ; les clients peuvent y avoir accès à tout moment et de n'importe quel endroit. Ces machines sont généralement regroupées dans des **data centers**.

SNT2037

Le cloud : ses usages, ses avantages et ses limites

VOCABULAIRE

Cloud
L'informatique en nuage (en anglais, *cloud computing*) consiste à exploiter à travers Internet des ressources informatiques (stockage, services) proposées par des entreprises sur des serveurs distants.

VOCABULAIRE

Data center
Un centre de données est un bâtiment qui regroupe un grand nombre de serveurs permettant de stocker de très grandes quantités de données.

3 Les data centers

Les data centers abritent des milliers de serveurs informatiques destinés à stocker et traiter les données via un réseau interne ou un accès à Internet. Pour proposer des services cloud à leurs clients, les géants du cloud (Microsoft, Google, Amazon, Apple, OVH, Orange, etc.) mettent à leur disposition des dizaines de data centers gigantesques : il en existe plus de 400 dans le monde. Ces centres sont conçus pour garantir une haute disponibilité – le taux de disponibilité atteint 99,6 %, voire 99,995 % – et optimiser les coûts d'exploitation. Ils consomment énormément d'énergie (près de 30 % des coûts d'exploitation). Toutefois, de gros efforts sont réalisés pour optimiser la consommation : par exemple, en s'implantant dans des régions froides pour éviter de climatiser. Dans les 27 centres d'OVH, l'énergie « gaspillée » (refroidissement, transformateurs et câbles électriques) ne représente que 9 % de l'énergie utilisée par les serveurs.

SNT2036

Le plus grand centre de données d'Europe

4 L'impact du numérique sur l'environnement

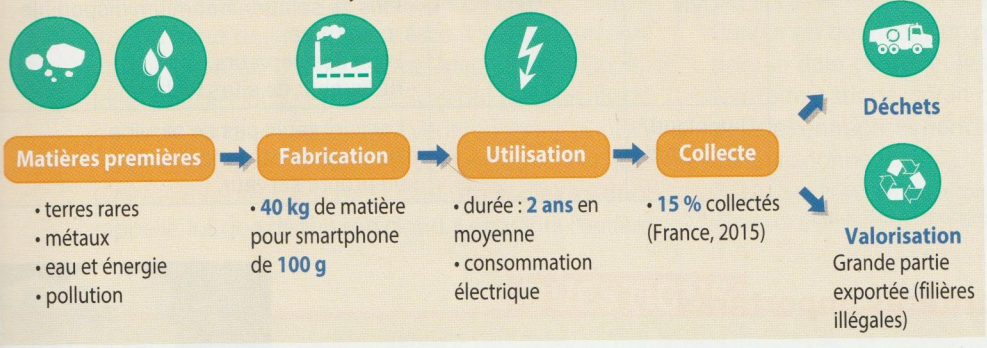
Les activités numériques permettent des économies d'énergie mais elles engloutissent près de 10 % de la consommation électrique mondiale. En 2015, les data centers ont consommé 416 TWH ; cela représente presque la consommation d'un pays comme la France. Par exemple, l'envoi d'un courriel consomme autant qu'une ampoule allumée une demi-heure, les visionnages en streaming du clip *Gangnam style* ont consommé l'équivalent de la production annuelle d'une centrale. Cette consommation qui explose est concentrée dans les pays les plus riches : un Américain possède en moyenne dix périphériques, alors qu'un Indien n'en possède qu'un. 60 % de la population mondiale est exclue du numérique. Par ailleurs, la fabrication des objets connectés pollue et consomme beaucoup de ressources limitées (métaux rares).

ZOOM SUR...

La pollution invisible

L'économie numérique est de plus en plus énergivore et, donc, source d'émission de CO_2. La partie concernée par Internet serait équivalente au 3e pays pollueur. Internet pollue 1,5 fois plus que le transport aérien.

Cycle de vie d'un smartphone



Tendances :

- Cycles de vie courts et augmentation des quantités.
- Baisse de la valeur marchande car moins de matériel facile à recycler (trop grande variété de petits composants).
- Développement de filières de recyclage et de réemploi.

QUESTIONS

1 **Doc. 1, 2, 3** En quoi consiste le cloud ? Quels sont les avantages et les inconvénients du stockage dans le cloud ?

2 **Doc. 3** Combien de minutes un data center peut-il être indisponible ? Comment une telle disponibilité est-elle assurée ?

3 **Doc. 4** Il ne suffira pas d'alimenter les data centers en électricité solaire/éolienne pour régler les problèmes environnementaux posés par le numérique. Pourquoi ? Donnez deux raisons.

4 **Doc. 4** Comment chaque individu peut-il contribuer à limiter les effets négatifs de ses activités numériques sur l'environnement ?

COURS

LES DONNÉES STRUCTURÉES ET LEUR TRAITEMENT

Les données sont au cœur de toute activité numérique. Leur structuration est essentielle pour pouvoir produire de l'information.

1 Comment structurer des données ?

Qu'elles soient issues du Web, de réseaux sociaux, de capteurs, etc., les **données** sont au cœur de notre quotidien : toute activité numérique consiste essentiellement à acquérir, transformer ou échanger des données.

Pour transformer une donnée en information utile, il faut connaître son **contexte**, l'objet qu'elle décrit. C'est le rôle des **descripteurs** et des **métadonnées**.

Une **collection** est un ensemble d'objets partageant les mêmes descripteurs. On peut représenter une collection sous forme de **table**. Lorsque les données sont ainsi organisées en collections, on parle de **données structurées**. Une **base de données** est un ensemble de collections reliées entre elles.

Descripteurs
Donnée
Collection

Nom	Téléphone	Adresse électronique
Marie	06...	marie.dupond@orange.fr
Erwan	04...	erwan.dupond@orange.fr
Sarah	05...	sarah.dupond@gmail.fr

Donnée

Représentation dans un système informatique d'une information décrivant un sujet particulier (âge d'une personne, photographie d'une voiture, etc.).

Métadonnée

Donnée qui sert à décrire une autre donnée (une « donnée à propos d'une donnée »). La date d'enregistrement d'un fichier ou encore le nom de son auteur sont des métadonnées.

Données structurées

Données organisées en collections, c'est-à-dire en ensemble d'objets partageant les mêmes descripteurs.

2 Comment stocker et traiter des données structurées ?

Pour stocker une donnée de façon persistante, l'ordinateur l'enregistre dans un **fichier**. On associe à tout fichier des **métadonnées**, qui décrivent les données. Le choix d'un **format** pour le fichier dépend à la fois du type de donnée à stocker (par exemple, JPEG ou PNG pour des images, CSV ou JSON pour des tables) et de ce que l'on veut en faire. Par exemple, certains appareils photographiques enregistrent comme métadonnées la date et la géolocalisation d'une photographie ; pour du code informatique, les métadonnées peuvent comporter l'auteur et la licence.

Des programmes spécialisés ont été développés pour gérer les **bases de données** et extraire des informations à partir des données. Il s'agit de stocker de façon **fiable** des données, de les filtrer rapidement, de réaliser des prédictions à partir de ces données ou encore de les visualiser.

3 Le cloud : exploiter les données à distance

Cloud

L'informatique en nuage (*cloud computing*) consiste à exploiter, à travers Internet, des ressources informatiques (stockage, services) hébergées sur des serveurs distants.

De plus en plus d'applications font appel au **cloud** pour stocker, traiter, partager ou rendre les données accessibles. Il permet des économies d'échelle grâce au partage des ressources (matériel, logiciels et maintenance) entre les utilisateurs.

Les **data centers** hébergent les machines permettant d'offrir les services cloud. Ils sont sécurisés mais cela ne dispense pas d'être vigilant sur les données sensibles. En France, la **commission nationale de l'Informatique et des Libertés** veille au respect des droits des citoyens en matière de stockage de **données personnelles**.

4 Exploitation des données massives : quel impact ?

Donnée personnelle

Information qui permet d'identifier, directement ou indirectement, une personne physique, par exemple son nom, son adresse, son numéro de téléphone, l'adresse IP de son ordinateur, son numéro d'immatriculation, etc.

L'**explosion du volume de données** disponibles et le développement de **techniques Big Data** pour traiter ces données permettent de proposer ou d'améliorer des services dans de nombreux domaines : **santé, science, économie**, etc. Par exemple, les suggestions d'itinéraires sur Google Maps reposent sur l'exploitation de données créées par les utilisateurs.

Ces grands volumes de données, notamment celles des utilisateurs de services Web, deviennent un énorme **enjeu économique**. Le **Règlement général sur la protection des données** encadre le stockage de **données personnelles** pour que les usagers puissent contrôler au mieux les données qu'ils fournissent en échange de services.

Les **data centers**, qui gèrent de gros volumes de données, consomment beaucoup d'**énergie** et de **ressources limitées ou polluantes**, ce qui constitue un grand défi géopolitique, économique et sanitaire.

LES CAPACITÉS DU CHAPITRE

- **1.** Définir une donnée personnelle. → **Controverse**
- **2.** Identifier les principaux formats et représentations de données. → **Activité 2**
- **3.** Identifier les différents descripteurs d'un objet. → **Activités 1, 2**
- **4.** Distinguer la valeur d'une donnée de son descripteur. → **Activités 1, 2**
- **5.** Utiliser un site de données ouvertes, pour sélectionner et récupérer des données. → **Activité 2**
- **6.** Réaliser des opérations de recherche, filtre, tri ou calcul sur une ou plusieurs tables. → **Activités 1 et 2, Fab Lab, Page du Codeur**
- **7.** Retrouver les métadonnées d'un fichier personnel. → **Activité 2**
- **8.** Utiliser un support de stockage dans le nuage. → **Activité 4**
- **9.** Partager des fichiers, paramétrer des modes de synchronisation. → **Activité 4**
- **10.** Identifier les principales causes de la consommation énergétique des centres de données ainsi que leur ordre de grandeur. → **Activité 3**

FAB LAB

SNT2039

Installation et présentation de SQLite

OBJECTIF

CONSTRUIRE ET UTILISER UNE BASE DE DONNÉES grâce au logiciel libre SQLite. Nous allons créer une table, y insérer des lignes et les modifier pour modéliser des événements de la vie réelle. Ces actions apparemment anodines sont effectuées quotidiennement par toutes les applications de bases de données du monde entier.

1 Élaborer le cahier des charges

On considère une agence de voyages qui stocke les billets de train ou d'avion de ses clients, qui sont des lycéens. Pour simplifier, on utilisera une seule table très simple et on supposera que tous les clients ont des prénoms différents, chacun n'ayant qu'un billet.

Un billet a les descripteurs suivants : Prénom, Destination, Transport, Jour, Prix, Statut.

Pour chaque billet dans la table, la donnée pour le descripteur Transport est "train" ou "avion", et celle pour Statut est "payé" ou "effectué" (signifiant que le voyage a été effectué). Les traitements de cette application sont l'insertion dans la table d'un nouveau billet avec le statut "payé", la modification du statut "payé" en "effectué", le filtrage en fonction d'un descripteur.

2 Utiliser le langage SQL

Structured Query Language (« Langage de requêtes structuré », SQL) est le langage de bases de données le plus répandu au monde depuis des décennies. Nous allons avoir besoin des **ordres SQL** pour créer une table, ajouter une ligne et modifier une ligne, et faire une recherche dans la table.

On lance SQLite dans une fenêtre console en tapant la commande : `*sqlite3*`.

Après un message, on obtient le prompt : `*sqlite>*` . Tous les ordres ci-dessous sont tapés dans cette fenêtre.

VOCABULAIRE

Ordre SQL

Il est similaire à une commande Linux ou Windows, mais il opère sur la base de données, et non sur les dossiers, fichiers ou processus de l'ordinateur lui-même.

3 Créer et utiliser sa table

a. Se connecter à SQLite.

b. Créer la table de l'agence de voyages puis en afficher le contenu avant même d'y insérer une ligne. Notre agence de voyages peut ouvrir !

c. Un(e) client(e) arrive et demande un billet : insérer son billet dans la table. Afficher toute la table sans filtrage.

Deuxième client(e) et deuxième billet avec la même destination que le premier.

Afficher tout.

Afficher le résultat d'un filtre sur le nom d'un client ; combien de lignes obtient-on ?

Procéder de la même manière sur la destination ; combien de lignes obtient-on ?

Ajouter un troisième billet avec une destination différente.

d. L'un(e) des client(e)s fait son voyage : modifier la ligne correspondant à son billet.

Afficher de nouveau toute la table sans filtrage.

LA PAGE DU CODEUR

SNT2085
Les bases de Python

→ Voir l'introduction à Python, p. 172

La librairie Pandas de Python permet de manipuler efficacement des collections. Elle est à la fois simple à utiliser et assez rapide pour traiter de grosses collections. L'objet qui stocke une table est appelé « DataFrame » dans Pandas.
Nous utiliserons aussi la librairie Matplotlib de Python, qui permet de tracer des graphiques (diagrammes à barres, nuages de points, etc.).
Télécharger le fichier CSV dans lequel sont enregistrées les productions électriques mensuelles pour chaque type d'énergie.

SNT2040
Fichier à télécharger

1 Observer les premières lignes

Pour visualiser le fichier, on peut soit l'ouvrir directement, soit utiliser la commande unix :
`head nomfichier.csv`.

a. Ouvrir le fichier. Les premières lignes sont reproduites ci-dessous (notre programme a ajouté le numéro des lignes).

```
1 Électricité, production mensuelle en France, en GWh (01/1981-11/2012),,,,,,;
2 Periode,Production totale nette d'électricité (en GWh),Production nette
  d'électricité primaire; y c. pompages  (en GWh),Production nette
  d'électricité nucléaire (en GWh),Production nette d'électricité hydraulique
  (en GWh),Production d'électricité éolienne (en GWh),Production nette
  d'électricité thermique classique (en GWh)
3 1981-01,26411,15125,8874,6251,0,11286;
4 1981-02,24108,13360,8189,5171,0,10748;
```

b. Que contient la première ligne du fichier ?
c. À quoi correspondent les éléments de la ligne 2 ?
d. Que contiennent les lignes 3 à 5 ?

2 Lire un programme

Que font les lignes de programme suivantes ?

```
import pandas as pd
from pandas import DataFrame, read_csv
import matplotlib
```

3 Charger des données

Pour charger les données dans une table à partir du fichier, nous utiliserons read_csv de Pandas, dont la syntaxe est :
nom_table = read_csv(chemin_du_fichier.csv, sep=',', ...)
`sep = ','` indique que le séparateur est une virgule.
`skiprows = x` indique qu'on souhaite sauter les x premières lignes.
`names = liste_de_noms` indique qu'on va donner nous-même le nom des descripteurs. On aurait pu les récupérer dans la 2e ligne du fichier, mais ces noms très longs seraient peu pratiques.

Charger les données dans une table nommée « elec » en renommant les descripteurs (mois, electr.totale, primaire, nucléaire, hydraulique, éolien et thermique).
Voici quelques éléments de syntaxe pour manipuler les DataFrames.

- **Afficher un DataFrame :** `print(elec.to_string())`
- **Sélectionner les 2 premières lignes du DataFrame «elec» :** `elec.head(2)`
- **N'afficher que ces 2 colonnes :** `elec[['nucleaire', 'hydraulique']]`
- **N'afficher que les lignes pour lesquelles l'éolien n'est pas nul :** `elec[elec['eolien'] != 0]`
- **N'afficher que les lignes pour lesquelles la somme des deux types d'électricité dépasse 20 000 GWh :** `elec[elec['nucleaire']+elec['hydraulique'] >20000]`
- **Afficher le DataFrame trié par production hydraulique croissante :** `elec.sort_values('hydraulique')`
- **Afficher le maximum sur chaque colonne du DataFrame :** `elec.max()`

Remarque : il est intéressant de noter que les quatre premières instructions ci-dessus prennent en entrée un objet de type DataFrame et renvoient un nouvel objet DataFrame. On peut donc stocker dans une variable le résultat et appliquer de nouvelles opérations au DataFrame ainsi obtenu, ou directement composer les opérations : `elec[elec['eolien'] != 0][['nucleaire', 'hydraulique']]`

4 Énergie primaire

On souhaite vérifier si l'énergie primaire est bien la somme de toutes les autres énergies, sauf la thermique (nucléaire, hydraulique, éolienne).

a. Écrire un programme qui affiche les lignes pour lesquelles ce n'est pas le cas.
b. Interpréter le résultat.

5 Tracer une figure

On exécute les lignes suivantes.

```
matplotlib.use('Agg')
import matplotlib.pyplot as plt
plot=elec.plot(kind='line')
plt.savefig('resultat.pdf')
```

Que représente la figure obtenue ?