CELLULE DATA
GRENOBLE ALPES

Démystifier le stockage

Violaine Louvet, GRICAD

Administratrice des Données, Algorithmes et Codes de l'UGA

Journée « Gestion des données de recherche en SHS »

23 novembre 2021

GRICAD
GRENOBLE ALPES RECHERCHE
INFRASTRUCTURE DE CALCUL INTENSIF ET DE DONNÉES

Stockage, de quoi parle-t-on ?

- Stockage = enregistrement d'une information sur un support physique

- Ce support physique peut avoir des caractéristiques très variées :

  - En fonction du matériel utilisé
    - Par exemple, quand vous achetez un portable vous pouvez choisir un disque SSD ou SATA.

  - En fonction de la technologie utilisée pour accéder à ce support physique
    - Par exemple, vous ne pouvez pas ouvrir sous Windows un disque dur qui a été formaté sous Linux car les technologies utilisées pour organiser le stockage ne sont pas les mêmes.

Stockage vs sauvegarde

- Sauvegarde : dupliquer des données pour les mettre en sécurité sur des supports de stockage différents
  - Recopie des données à l'identique
  - Sur des supports différents et localisés en général dans des endroits différents
    - Par exemple : je fais régulièrement des copies des fichiers de mon portable sur un disque dur externe
  - Objectif : pouvoir facilement récupérer des données en cas de perte ou de mauvaise manipulation
    - Permet de récupérer les données à la date de la dernière sauvegarde !

Stockage vs sauvegarde vs archivage

- Archivage = ensemble d'actions qui a pour but de garantir l'accessibilité sur le long terme d'informations (dossiers, documents, données) que l'on doit ou souhaite conserver pour des raisons juridiques, historiques ou culturelles. Il comprend à la fois des règles (procédures), des compétences et des infrastructures. (wikipedia)

- Dans notre cas : archivage à long terme de données numériques

- Donc ce n'est pas seulement du stockage sur une longue durée !

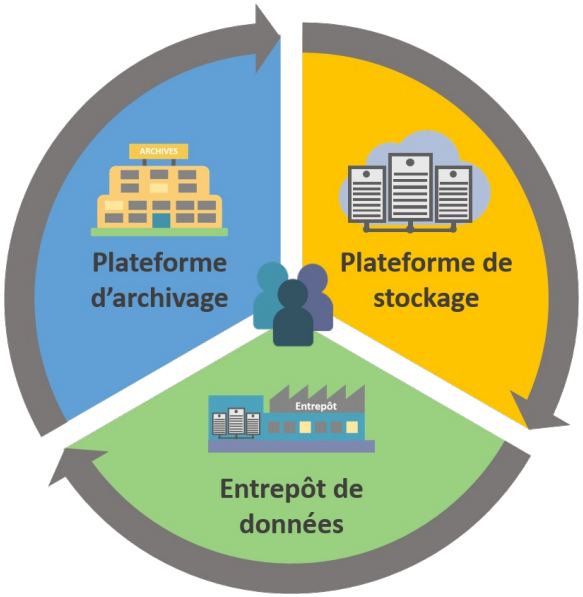
- Nécessité :
  - D'assurer la pérennisation des supports de stockage sur du long terme
  - D'assurer l'accès au contenu même quand les formats des données deviennent obsolètes
  - D'assurer l'intégrité des données

- Cadre juridique propre

- En France : un seul opérateur pour l'archivage des données de l'ESR : le CINES (Centre Informatique National de l'Enseignement Supérieur)

Plateforme de stockage, de diffusion, d'archivage, entrepôts de données

- Différentiés par les usages :
  - **Plateforme de stockage** : infrastructure proposant un stockage des données, avec des fonctionnalités de gestion des accès et éventuellement de sauvegarde intégrée
  - **Plateforme d'archivage** : infrastructure qui intègre dans son fonctionnement tout le processus nécessaire à l'archivage des données
  - **Entrepôt de données, plateforme de diffusion** : Réservoir de données de recherche, brutes ou dérivées, qui peuvent être retrouvées et réutilisées grâce à une description par des métadonnées. Un identifiant pérenne ou numéro d'accès unique est attribué à chaque jeu de données. Il peut être disciplinaire ou thématique, être institutionnel ou centralisé.



La question des volumes de données

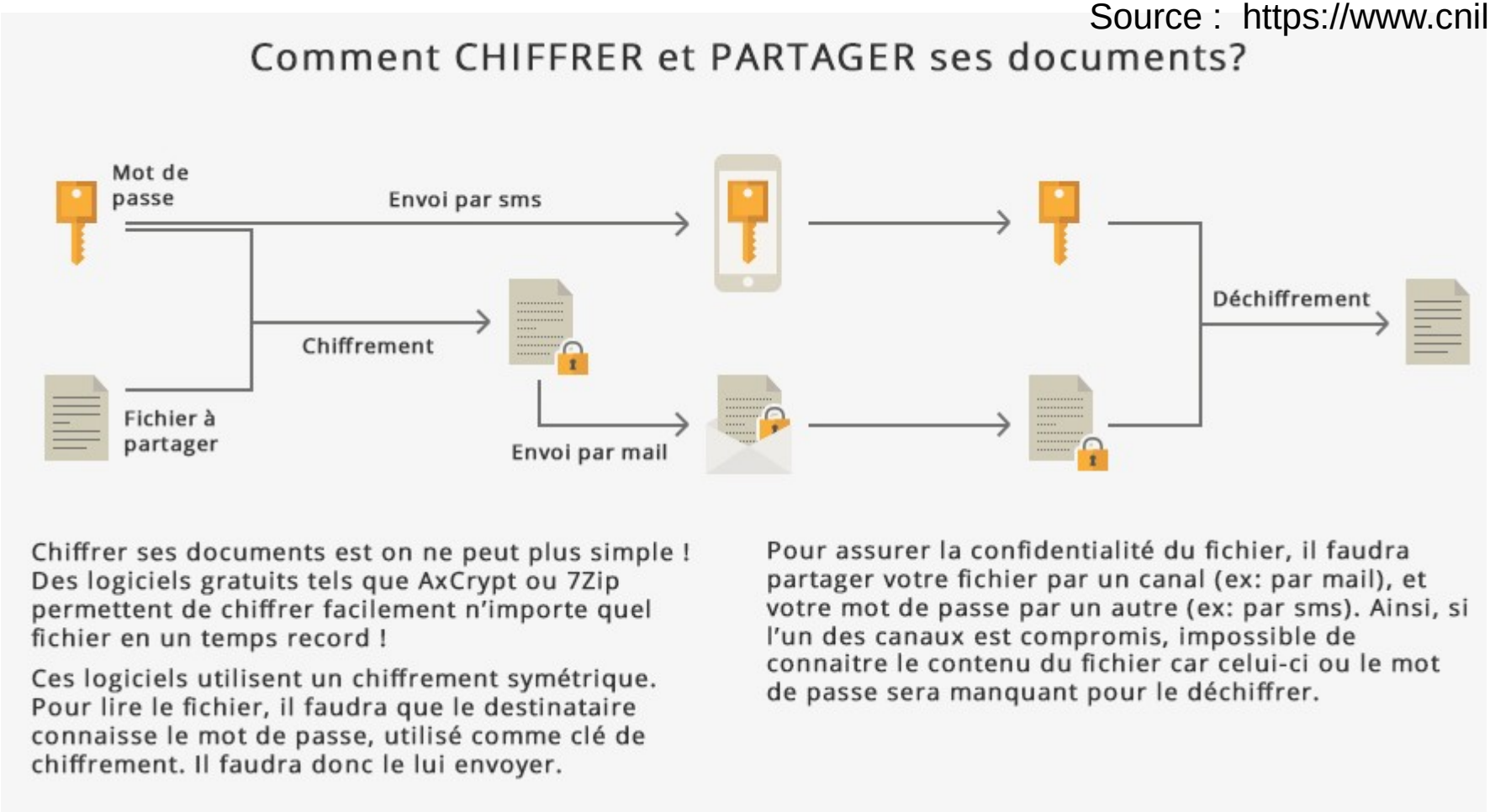
- 1 Po = 1 000 To ; 1 To = 1 000 Go ; 1 Go = 1 000 Mo
  - Capacité disque dur externe « moyen » : 2 To
  - Plateforme de stockage de site : quelques Po
  - Fichier texte < fichier bureautique ~ fichier audio ~ fichier image < fichier vidéo
- On peut commencer à parler de gros volumes à partir de quelques To à quelques dizaines de To : en fait à partir du moment où le temps de transfert ou le temps de chargement devient prohibitif
- Conséquences d'une volume important :
  - Problématique du transfert des données
  - Problématique du chargement en mémoire pour traitement / analyse / calcul
  - Problématique de la capacité de stockage nécessaire et de son coût
  - Problématique de la politique de sauvegarde ...

Température des données

- **Données froides** : très peu utilisées et donc très peu accédées.
  - Un carton de photos anciennes dans un grenier : on sait qu'il est là mais on ne va pas l'ouvrir tous les jours. Par contre, on a envie de pouvoir sortir une photo lors d'une occasion particulière
  - En général, on peut se permettre des temps d'accès un peu long sur ces données
  - Un disque dur sur lequel on a stocké des données d'un vieux projet
- **Données chaudes** : actives, accédées souvent voir de façon très intensive
  - Par exemple, des données issues d'une expérimentation en cours et que l'on doit analyser
  - Dans ce cas, on souhaite que les traitements soient rapides et donc que les temps d'accès aux données soient performants
  - Il existe des espaces de stockage spécifiques pour les traitements intensifs de données permettant de faire des calculs efficacement

Chiffrement des données

Source : https://www.cnil.fr/



Le chiffrement est une méthode qui consiste à protéger ses documents en les rendant illisibles par toute personne n'ayant pas accès à une clé dite de déchiffrement.

Attention : le chiffrement est irréversible, si vous perdez la clé de déchiffrement ou le mot de passe qui la déverrouille vous ne pourrez plus accéder à vos données.

Le coût du stockage

- Coût financier
  - Le stockage a un coût qu'il ne faut pas négliger
  - Par sa nature, contrairement au calcul par exemple, le stockage n'est pas vraiment mutualisable
  - On peut optimiser en partageant une plateforme et les coûts humains associés
  - Plus un stockage est performant (on accède rapidement aux données), plus il est cher
  - Exemple de coût :
    - disque dur externe 2 To : 35 à 70 € environ
    - 1 To sur la plateforme Summer UGA : de 25 € à 80 € annuel en fonction du type de stockage
    - 1 To sur la plateforme Microsoft Azur : de 50 à 180 € environ par mois en fonction du type de stockage
- Coût environnemental
  - Le stockage, et encore plus l'archivage, a un coût environnemental non négligeable
  - Exemple : pour la plateforme Bettik à l'UGA (stockage performant pour le calcul) : 1 Go.an émet 12 g CO2e – en 2019 : 1.4 Po.an → 17 t CO2e
- Coût humain
  - Il ne faut pas négliger l'importance de disposer de ressources humaines compétentes pour administrer les plateformes de stockage, quelle que soit leur taille

Les questions à se poser pour le choix du stockage

Importance du Plan de Gestion de Données où ces points sont abordés

- Que dois-je stocker ?
  - Des données scientifiques, des codes, des documents ?
  - Des fichiers binaires ou ascii ?
- Quelle est / sera la volumétrie des données ?
- Quel niveau de sécurisation / sauvegarde est nécessaire (les données sont-elles facilement reproductibles ?) ?
- Que vais-je faire avec ces données ?
  - Des traitements ou du calcul
  - De l'analyse, de la fouille
  - Du partage
  - De la préservation …
- Les données ont-elles un caractère sensible ou confidentiel ?
- Qui devra avoir accès aux données ?
- De quelle manière les données devront-elles être accédées ?
- Quels vont être les flux / débits de données (continu & régulier, épisodique, …) ?
- Quand vais-je avoir besoin de ces données ?
  - Rapidement, de façon régulière
  - Dans quelques mois
  - Peut-être dans quelques années …
- Quel est le financement prévu pour la gestion et le stockage des données ?