Plateformes de stockage

Eléments d'information à l'usage des communautés scientifiques
Mars 2020

Ce document a pour but d'informer les communautés scientifiques sur les différentes possibilités offertes à l'échelle locale et nationale pour stocker des données numériques. Ce n'est pas un catalogue exhaustif. Pour plus de renseignements, ne pas hésiter à contacter la cellule d'accompagnement data stewardship de Grenoble : uga-cellule-data@univ-grenoble-alpes.fr

Quelques définitions

Données chaudes et données froides

Les données froides sont des données qui ne sont que très peu utilisées et donc très peu accédées. C'est un peu l'équivalent d'un carton de documents dans un grenier. On peut potentiellement un jour en avoir besoin, mais on ne l'ouvre quasiment jamais. On s'attend ainsi à ce que les temps de récupération de données froides soient un peu plus longs.

Au contraire, les données chaudes sont des données actives, qui sont accédées souvent, voir de façon intensive. Ainsi, les espaces de stockage dit « scratch » correspondent à des données chaudes non sauvegardées, prêtes à être exploitées par es ressources de calcul ou de traitement.

Cette notion de température des données se distinguent de la notion de durée de conservation.

Stockage, sauvegarde et archivage

Il faut distinguer ces termes car il s'agit bien d'approches différentes.
On parle globalement de stockage c'est-à-dire le moyen / le support / l'endroit où les données sont conservées.
La sauvegarde concerne la réplication de données sur au moins deux supports distincts afin de pouvoir restaurer des données endommagées ou perdues.
La sauvegarde peut aussi s'apparenter à un stockage moyen terme de données liées à des projets terminés (donc froides) mais que l'on souhaite conserver pour réutilisation ultérieure éventuelle.
L'archivage est la conservation à long terme de données numériques de façon sécurisée et pérenne, c'est-à-dire que cela s'appuie sur des processus de vérification de l'intégrité des données (et des métadonnées associées), de relecture du contenu et des métadonnées associées, et de maintien des données sur un support exploitable.

Certaines données comme les données de santé impliquent des contraintes sur leur lieu de stockage : elles doivent être localisées soit à l'endroit où elles ont été produites (les CHU hébergent leurs données propres par exemple), ou sur une infrastructure labellisée « hébergement de données de santé ».

Entrepôt de données

Un entrepôt est un répertoire de données permettant le dépôt de jeux de données pour leur mise en visibilité et leur réutilisation par des tiers. Ce dépôt se distingue de la notion d'archivage qui inscrit la donnée dans un processus de pérennisation.
L'utilisation d'un entrepôt de données ne doit pas occulter la question de la durée de conservation de ces données.
Il existe de nombreux entrepôts de données, disciplinaires ou génériques (comme Zenodo par exemple).

Exploitation des données

La problématique du stockage est très souvent liée à ce qu'on veut faire des données, comment on veut les exploiter. Beaucoup d'approches sont possibles selon le type des données, leur taille, … En particulier, la catégorisation des données selon leur caractère structuré ou non structuré, c'est-à-dire des données organisées (dans une base de données, un tableau ou un fichier formaté selon une nomenclature précise) ou non organisées et complexes (textes, extraction de données du Web …) a un impact sur les outils et infrastructures à utiliser pour leur exploitation.

Quelques exemples de traitements pour des données structurées :

- Traitement ou analyse massive via l'utilisation d'infrastructures comme les grilles de calcul ou les supercalculateurs
- Utilisation d'algorithmes d'apprentissage automatique via des infrastructures à base de GPU
- Visualisation via des outils ou des infrastructures dédiés, ou consultation simple
- Utilisation d'outils spécifique via des infrastructures de cloud
- …

Quelques exemples de traitements pour des données non structurées :

- Selon la taille des données, outils de parcours, de fouille, de visualisation qui peuvent être utilisés en local ou sur des infrastructures de type cloud, ou consultation simple
- Utilisation d'algorithmes d'apprentissage automatique via des infrastructures à base de GPU
- Utilisation d'infrastructure « big data » à base d'outils comme la suite elasticsearch ou hadoop
- ...

Tous ces usages peuvent influer sur l'endroit où on va stocker les données.

Volumétrie, organisation des données

Les volumétries des données à stocker sont extrêmement différentes d'un projet à l'autre. On peut actuellement commencer à parler de grosses volumétries à partir de quelques dizaines de To. Ces tailles de données complexifient en particulier les transferts des données par le réseau.
Par exemple, pour un transfert entre une plateforme de stockage située dans un datacentre de l'UGA et un portable dans un laboratoire (sur le réseau UGA), le débit du réseau est de 10 GB/s, et donc le transfert de 10 To prendra 1000 secondes soit presque 17 mn (au mieux selon la charge du réseau).

C’est un cas favorable car on ne traverse pas de firewall et le nombre de routeurs traversés est le plus réduit possible, le temps sera beaucoup plus long pour le transfert sur un autre site.
Le temps de transfert sera aussi fortement impacté par le nombre de fichiers, transférer un volume de 1 To composé de 10000 fichiers sera beaucoup plus long que pour un seul fichier de 1 To.

Modalités d’accès

Différentes modalités d’accès aux données sont possibles en fonction de leur lieu de stockage. Le choix du stockage va donc imposer des contraintes sur la façon dont on va pouvoir utiliser les données.
Ces modalités d’accès se déclinent selon des protocoles techniques et des technologies de stockage très différents.
Il est donc important d’identifier les besoins de chaque projet avant tout choix technique :

- Comment veut-on accéder aux données ?
- Qui va accéder aux données ? Avec quels droits (lecture, modification) ?
- Que doit permettre cet accès aux données : traitement, analyse, manipulation, diffusion … ?

Du point de vue de l’utilisateur, les accès possibles sont :

- Accès par le web, l’utilisateur accède à ses données directement via son navigateur.
- Accès manuel par le réseau, l’utilisateur télécharge ses données sur sa machine soit via un client graphique, soit en utilisant la ligne de commande. Il peut aussi procéder à un montage d’un lecteur réseau
- Accès automatique par le réseau (le système se charge des connexions de façon transparente), l’utilisateur voit directement ses données depuis son ordinateur
- ...

Description des besoins pour les données, plan de gestion des données

Le choix d’une ou de plusieurs plateformes de stockage nécessite de répondre à un certain nombre de questions, que l’on retrouve dans un plan de gestion des données :

- Que dois-je stocker ?
  - Des données scientifiques, des codes, des documents ?
  - Des fichiers binaires ou ascii ?
  - Quelle volumétrie ?
- Qui va pouvoir accéder aux données ?
- D’où vais-je devoir accéder à ces données ?
- Quels vont être les flux / débits de données (continu & régulier, épisodique, ...) ?

- Pourquoi faire ?
  - Des traitements ou du calcul
  - Du partage
  - De la préservation
- Quand vais-je avoir besoin de ces données ?
  - Rapidement, de façon régulière
  - Dans quelques mois
  - Peut-être dans quelques années
- Quel financement va permettre d’assurer ce stockage ?

Différents opérateurs fournissent du service de stockage au niveau local ou national, chaque plateforme correspondant à un usage relativement précis. Les fiches suivantes synthétisent les caractéristiques de ces infrastructures.

Fiche 1 : Plateforme SUMMER https://summer.univ-grenoble-alpes.fr/accueil	**Fiche 2 : Plateforme MANTIS** https://gricad-doc.univ-grenoble-alpes.fr/hpc/data_management/mantis/
Idéal pour : Stocker, sauvegarder des données, quelque soit leur volume et vouloir y accéder via le réseau depuis un système bien identifié (serveur de laboratoire, machine de calcul …). Pour des données chaudes ou froides. **Peu adapté pour :** Accéder à des données hors UGA, partager des données avec des utilisateurs hors UGA. **S'adresse à :** A tous les utilisateurs ayant une problématique de stockage centralisé, membre de la communauté UGA en priorité, et bénéficiant d'un support informatique pour la mise en place. **Coût :** Tarification annuelle complète accessible sur le site. Pour une volumétrie inférieure à 5 To : 40 € To/an Pour une volumétrie supérieure à 5 To, avec accès performant : 80 € To/an **Politique d'accès :** Utilisateur UGA dispensé des coûts d'infrastructure. Coût d'infrastructure pour les extérieurs à l'UGA : 5400 € HT/an. Pas de limite de volumétrie. **Type d'accès :** Accès via le réseau depuis un système identifié. Protocoles ISCSI, NFS et CIFS. **Exploité par :** Le Comité Technique Stockage constitué de membre de la DGDSI UGA, de laboratoires, de GRICAD et de composantes d'enseignement, coordonné par la DGDSI UGA. Localisé à Grenoble. **Eléments de pérennité :** La plateforme est soutenue par l'établissement UGA et en partie financée par les contributions des utilisateurs.	**Idéal pour :** Stocker des données dans le cloud, quelque soit leur volume et vouloir y accéder depuis les infrastructures de traitement de données et de calcul. Accessible depuis certaines machines de calcul de centres nationaux (IDRIS). Pour des données chaudes ou froides. **Peu adapté pour :** Sécuriser des données car la plateforme n'assure pas de sauvegarde. **S'adresse à :** A tous les utilisateurs ayant une problématique de stockage centralisé, en particulier pour l'usage de la grille de calcul, et soit issu de la communauté scientifique grenobloise, soit collaborateur extérieur d'un membre de cette communauté. **Coût :** Pas de coût associé à l'utilisation de la plateforme. En cas d'usage important (au delà de quelques dizaines de To), l'utilisateur peut être amené à contribuer financièrement à la plateforme mutualisée via l'achat ou la participation à l'achat d'un nœud par exemple (coût actuel des nœuds environ 10 k€). **Politique d'accès :** Plateforme ouverte à tous les utilisateurs autorisés. Le quota de base est de 500 Go, révisable sur demande en cas de besoins plus importants. **Type d'accès :** Accès via un client IRODS, depuis les infrastructures autorisées, en particulier les machines de calcul de GRICAD et de l'IDRIS (ADA, Jean Zay). L'utilisation des données nécessitent leur copie sur un système local. **Exploité par :** L'UMS GRICAD. Localisé à Grenoble. **Eléments de pérennité :** La plateforme est un des services de stockage de l'UMS GRICAD, qui a pour tutelle le CNRS, l'UGA, GINP et INRIA. La plateforme est financée par les apports des projets scientifiques, des laboratoires et des tutelles de l'UMS.

Fiche 3 : Plateforme BETTIK https://gricad-doc.univ-grenoble-alpes.fr/hpc/data_management/bettik/	**Fiche 4 : Plateforme NextCloud UGA** https://cloud.univ-grenoble-alpes.fr/
Idéal pour : Stocker des données potentiellement massives pendant leur traitement sur les machines de calcul. Pour des données chaudes. Espace « scratch ». **Peu adapté pour :** Stocker à moyen terme, sauvegarder des données. **S'adresse à :** A tous les utilisateurs utilisant les machines de calcul Luke et Dahu. **Coût :** Pas de coût associé à l'utilisation de la plateforme. En cas d'usage important, l'utilisateur peut être amené à contribuer financièrement à la plateforme mutualisée via l'achat ou la participation à l'achat d'un nœud par exemple (coût actuel des nœuds environ 10 k€). **Politique d'accès :** Plateforme ouverte à tous les utilisateurs autorisés. Le quota de base est de 500 Go, révisable sur demande en cas de besoins plus importants. **Type d'accès :** Accès direct depuis les machines de calcul Luke et Dahu. **Exploité par :** L'UMS GRICAD et ses partenaires. Localisé à Grenoble. **Eléments de pérennité :** La plateforme est le service de stockage performant de l'UMS GRICAD, qui a pour tutelle le CNRS, l'UGA, GINP et INRIA. La plateforme est financée par les apports des projets scientifiques, des laboratoires et des tutelles de l'UMS.	**Idéal pour :** Partager, stocker et éditer de manière collaborative des fichiers et documents de taille raisonnable (quelques Go). Pour des données chaudes ou froides. **Peu adapté à :** Stockage de grand volume de données. **S'adresse à :** Utilisateurs UGA. **Coût :** Gratuit. **Politique d'accès :** Accès ouvert à tous les membres de l'UGA avec une limite de volumétrie de 50 Go. **Type d'accès :** Accès Web, ou par un client de synchronisation Nextcloud à travers un VPN en cas de connexion extérieure au réseau UGA. **Exploité par :** La DGDSI UGA. Localisé à Grenoble. **Eléments de pérennité :** Plateforme financée par l'établissement UGA.

Fiche 5 : Plateforme Gitlab https://gricad-gitlab.univ-grenoble-alpes.fr/	**Fiche 6 : Plateforme Alfresco** https://espaces-collaboratifs.grenet.fr/share/page/
Idéal pour : Le stockage de fichiers et de documents de volumétrie moyenne pour le travail collaboratif. Pour des données chaudes ou froides. **Peu adapté à :** Stockage de grosse volumétrie et de fichiers binaires. **S'adresse à :** Tous les membres de la communauté scientifique grenobloise et leurs collaborateurs extérieurs. **Coût :** Accès libre. **Politique d'accès :** Ouvert à tous. Seuls les membres de la communauté scientifique grenobloise (avec un compte sur le référentiel de l'UGA) peuvent créer des projets. La volumétrie autorisée est de quelques Go à quelques dizaines de Go. **Type d'accès :** Accès web ou avec un client git. **Exploité par :** L'UMS GRICAD et ses partenaires. Localisé à Grenoble. **Eléments de pérennité :** La plateforme est un service de l'UMS GRICAD, qui a pour tutelle le CNRS, l'UGA, GINP et INRIA. La plateforme est financée par l'UMS.	**Idéal pour :** Partager, stocker et éditer de manière collaborative des fichiers et documents de taille raisonnable (quelques Go), en particulier des documents issus des suites bureautiques. Pour des données chaudes ou froides. **Peu adapté à :** Stockage de grand volume de données. **S'adresse à :** Tous. Aux extérieurs à l'UGA sur invitation. **Coût :** Accès libre. **Politique d'accès :** Ouvert à toute la communauté UGA et sur invitation aux collaborateurs extérieurs. **Type d'accès :** Accès web. **Exploité par :** DSIM (ex-SIMSU). Localisé à Grenoble. **Eléments de pérennité :** Plateforme financée par la DSIM.

Fiche 7 : Stockage de masse du CC-IN2P3 https://doc.cc.in2p3.fr/fr/Data-storage/mass-storage.html	**Fiche 8 : Plateforme d'archivage du CINES** https://www.cines.fr/archivage/
Idéal pour : Stocker des données pouvant aller jusqu'à des volumétries très importantes. Stockage de masse sur bande magnétique. Pour des données froides. **Peu adapté à :** Stockage de données hors communauté IN2P3 **S'adresse à :** Très majoritairement aux communautés IN2P3, voir https://doc.cc.in2p3.fr/fr/Getting-started/access.html **Coût :** Coût à voir en fonction de chaque projet **Politique d'accès :** Accès après accord pour les projets autorisés **Type d'accès :** Accès via un client spécifique. **Exploité par :** CC-IN2P3. Localisé à Lyon. **Eléments de pérennité :** Infrastructure du CNRS	**Idéal pour :** Archiver à long terme des données. Pour des données froides. **Peu adapté à :** Stockage de données court et moyen terme, stockage de données ne nécessitant pas d'archivage **S'adresse à :** Tous **Coût :** Coût à prévoir en fonction de chaque projet **Politique d'accès :** Archivage accessible aux projets autorisés **Type d'accès :** Dépôt des archives avec l'aide des équipes du CINES **Exploité par :** CINES. Localisé à Montpellier. **Eléments de pérennité :** Le CINES est l'opérateur pour l'archivage des données et documents numériques produits par la communauté Enseignement Supérieur et Recherche française

Fiche 9 : Plateformes d'Huma-Num https://www.huma-num.fr/services-et-outils/stocker	**Fiche 10 : Plateforme FG-IRODS de France Grille** http://www.france-grilles.fr/catalogue-de-services/fg-irods/
Idéal pour : Stocker, échanger, partager des données de petite, moyenne ou grande taille. Pour des données chaudes ou froides. **Peu adapté à :** L'archivage des données **S'adresse à :** Communautés SHS **Coût :** Accès libre **Politique d'accès :** Limité aux communautés SHS **Type d'accès :** Différents type de stockage permettant différents type d'accès **Exploité par :** Infrastructure de Recherche Huma-Num. Localisé à Lyon. **Eléments de pérennité :** Huma-Num est une TGIR (Très Grande Infrastructure de Recherche)	**Idéal pour :** Stocker des données sur un espace de stockage hautement disponible et personnalisable. Pour des données chaudes ou froides. **Peu adapté à :** L'archivage des données, les très grosses volumétries **S'adresse à :** Tous après autorisation **Coût :** Accès libre, participation à l'infrastructure si volumétrie importante **Politique d'accès :** Accès aux membres de l'organisation virtuelle France Grille. Accès à cette VO sur demande **Type d'accès :** Technologie IRODS, accès depuis un client IRODS autorisé, via le web ou depuis les infrastructures de grille et cloud de France Grilles **Exploité par :** Infrastructure et service mutualisés par des partenaires de France Grilles. Localisé majoritairement à Lyon. **Eléments de pérennité :** France Grille est une infrastructure de Recherche structurée en GIS soutenue par différents organismes de recherche (MESRI, CNRS, CEA, CPU, INRA, INRIA, INSERM, RENATER).