Inserm

La science pour la santé
From science to health

Guide d'utilisation

Calcul HPC avec utilisation d'un jeu de données disponible sur un bucket S3

Inserm - Cloud self-service - Offre HPC

Objectif du document

Ce tutoriel vous présente comment exploiter des jeux de données externes hébergés sur un bucket S3 afin de les intégrer de manière sécurisée et performante dans des traitements réalisés sur l’infrastructure HPC.

Table des matières

1 Eléments contextuels

Rôle du bucket S3

- Stockage centralisé, durable et partagé des jeux de données
- Point d'entrée unique pour les données d'entrée et de sortie
- Découplage entre stockage et ressources de calcul
- Mutualisation et traçabilité des données entre utilisateurs et projets

Origine et nature des données (e.g.)

- Données brutes de séquençage (e.g. FASTQ, BAM, VCF)
- Séquences de référence et fichiers d'index (e.g. FASTA, index Bowtie2)
- Résultats intermédiaires et finaux
- Scripts, outils ou codes nécessaires aux traitements

Intégration des données dans le HPC

1. Configuration des accès sécurisés au bucket (AK/SK)
2. Synchronisation locale des fichiers via `rclone`
3. Copie sur le stockage local du cluster
4. Exécution des calculs sur les données locales pour garantir des performances optimales

La copie locale permet de limiter les latences réseau et d'assurer une efficacité maximale des traitements.

Contraintes réglementaires – Hébergement de Données de Santé (HDS)

- Données potentiellement sensibles ou à caractère personnel
- Hébergement sur infrastructure certifiée HDS
- Contrôle d'accès strict et authentification sécurisée
- Traçabilité des opérations
- Transferts chiffrés et absence d'exposition publique

Le recours à un bucket S3 HDS garantit la conformité réglementaire et la sécurité des données.

2 Configuration des accès au bucket S3 via ses AK/ SK

1) Cliquer sur le menu Clusters > HPC Shell Access pour accéder au home directory depuis la plateforme OOD HPC

2) Exécuter la commande ci-dessous pour lancer le script dans le dossier shared : /shared/inits3.sh.

```
$> /shared/inits3.sh s3_ecoli
```

Dans l'exemple ci-dessous, le profil se nomme **s3_ecoli**

Ce script permet de simplifier l'utilisation du bucket par la mise en place d'un profil stockant les informations **AK.SK** qui seront utilisées par l'outil `rclone`.

> Renseigner un nom de profil selon votre préférence pour configurer et utiliser plusieurs buckets au sein de son environnement :
>
> - 1 profil = 1 bucket
> - Attention à bien se souvenir du nom de profil choisi.

3) Saisir les Access Key (AK) et Secret Key (SK)

Le profil est créé, nous passons à l'étape suivante.

3 Duplication du bucket S3 sur l'environnement HPC

Afin de ne pas dégrader les performances de calcul en raison des échanges de données entre le HPC et le serveur S3, une **copie locale du bucket** doit être réalisée avec l'outil **rclone** (slide suivante).

Pour effectuer l'opération, 3 éléments sont nécessaires :

1️⃣ Le **nom du profil** configuré précédemment

2️⃣ L'**URL du bucket S3** source

3️⃣ Un **répertoire local** où seront stockés les fichiers téléchargés depuis le bucket

Dans notre exemple :

- nom du profil : s3_ecoli
- URL du bucket : https://s3.hds.inserm.fr:10444/xxxx-yyyyy-26-imp-37-hds
- répertoire local : /$HOME/bucket_s3_ecoli_rel606/

Étapes de duplication

Toujours depuis le terminal « cluster HPC Shell Access » :

Étape 1 : création du répertoire local

```
$> mkdir /$HOME/bucket_s3_ecoli/
```

Étape 2 : duplication du bucket dans le répertoire

Dans l'url du bucket remplacer l'entête : « https://s3.hds.inserm.fr:10444/ » par le nom du profil pour obtenir : « **s3_ecoli:**xxxx-yyyyy-26-imp-37-hds »

```
$> rclone sync -P \
s3_ecoli:xxxx-yyyyy-26-imp-37-hds/ \
/$HOME/bucket_s3_ecoli/
```

Étape 3 : vérification du contenu local

```
$> tree /$HOME/bucket_s3_ecoli/
```

Voici le résultat que nous obtenons :

```
[saucouturier@ood ~]$ mkdir bucket_s3_ecoli
[saucouturier@ood ~]$ rclone sync -P s3_ecoli:ioan-valerian-bulgaria-26-imp-37-hds/  bucket_s3_ecoli/
Transferred:        4.973 GiB / 4.973 GiB, 100%, 339.458 MiB/s, ETA 0s
Transferred:           20 / 20, 100%
Elapsed time:        30.3s
[saucouturier@ood ~]$ tree bucket_s3_ecoli/
bucket_s3_ecoli/
└── ecoli_rel606
    ├── data
    │   ├── SRR030257_1.fastq
    │   └── SRR030257_2.fastq
    ├── out
    │   ├── alignement-genomics-77_NC_012967.1_variants.txt
    │   ├── alignement-genomics-77_SRR030257.bam
    │   ├── alignement-genomics-77_SRR030257.raw.bcf
    │   ├── alignement-genomics-77_SRR030257.sam
    │   ├── alignement-genomics-77_SRR030257.sorted.bam
    │   ├── alignement-genomics-82_NC_012967.1_variants.txt
    │   ├── alignement-genomics-82_SRR030257.bam
    │   ├── alignement-genomics-82_SRR030257.raw.bcf
    │   ├── alignement-genomics-82_SRR030257.sam
    │   └── alignement-genomics-82_SRR030257.sorted.bam
    ├── ref
    │   ├── NC_012967.1.fasta
    │   └── NC_012967.1.fasta.fai
    └── tmp
        ├── NC_012967.1.1.bt2
        ├── NC_012967.1.2.bt2
        ├── NC_012967.1.3.bt2
        ├── NC_012967.1.4.bt2
        ├── NC_012967.1.rev.1.bt2
        └── NC_012967.1.rev.2.bt2
```

4 Exemple d'exécution d'un calcul génomique

Notre script **sbatch** a été conçu pour prendre en paramètres :

- Le chemin du répertoire local qui contient les données de séquençage.
- Un NC : une séquence de référence issue de la base RefSeq du NCBI, au format FASTA.
- Un SRR : identifiant de run de séquençage dans la base de données SRA (Sequence Read Archive) du NCBI.

Avec notre jeu de données :

- Chemin du répertoire local : /$HOME/bucket_s3_ecoli/ecoli_rel606
- NC = NC_012967.1 (Escherichia coli B str. REL606)
- SRR = SRR030257

Nous exécutons notre script sur le HPC, sur 1 nœud CPU-16 cœurs (paramétré dans le script), le traitement de ce séquençage prend 2 minutes sur ce jeu de données.

Résultats de l'exécution

Résultats temporaires	Résultats finaux
/$HOME/bucket_s3_ecoli/ecoli_rel606/tmp	/$HOME/bucket_s3_ecoli/ecoli_rel606/out

Les fichiers sont indexés avec l'entête alignement-genomics et le numéro de job HPC.

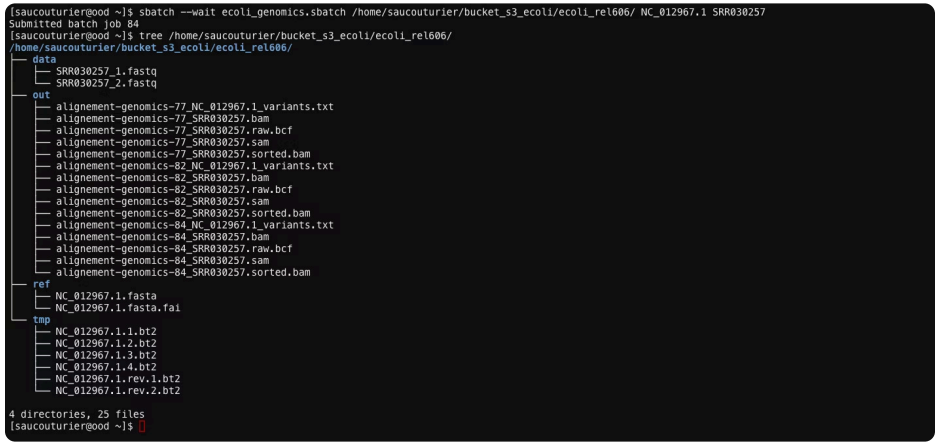
1) Lancer la commande ci-dessous pour exécuter le job, toujours depuis le terminal « cluster HPC Shell Access » :

```
$> sbatch --wait ecoli_genomics_s3.sbatch /$HOME/bucket_s3_ecoli/ecoli_rel606/ NC_012967.1 SRR030257
```

Le job avait le numéro 84.

Nous retrouvons les résultats de ce run avec ceux des run 77 et 82 précédant qui ont déjà été stockés dans le bucket s3.

Une fois le calcul fini, récupérer les données locales dans le bucket S3 pour exploiter les résultats en externe.



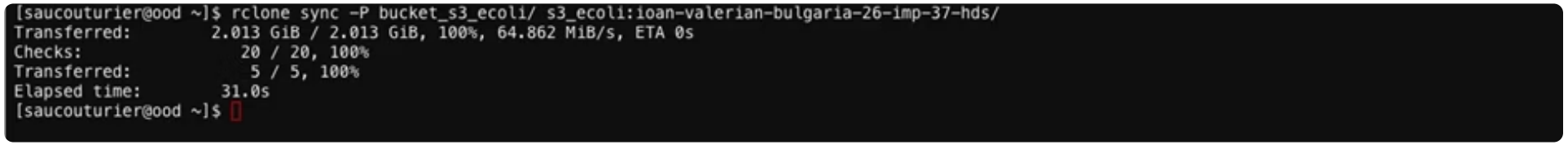
5 Recopie des données sur le bucket S3

Pour extraire des données locales vers votre bucket S3 depuis le terminal « cluster HPC Shell Access » :

1) Lancer la commande rclone en inversant source et destination par rapport à l'étape 2 du chapitre 3 :

```
$> rclone sync -P \
/$HOME/bucket_s3_ecoli/ \
s3_ecoli:xxxx-yyyyy-26-imp-37-hds/
```

Voici le résultat attendu :



Si vous avez des questions, contacter vos référents habituels sur le HPC.