

AWS DataSync : L'accélérateur de migration de données

Introduction

Le besoin

Tu as des millions de fichiers (des téraoctets ou pétaoctets) en local et tu veux les **migrer massivement et rapidement** vers AWS, ou mettre en place une synchronisation régulière (ex: toutes les nuits).

Le contexte & Fonctionnement

DataSync est un outil de **transfert de données à grande vitesse**, optimisé pour saturer proprement ta bande passante réseau et aller 10 fois plus vite que des outils classiques (comme rsync ou un simple script de copie).

- **On-Prem vers AWS** : Tu as **besoin d'installer un Agent** (sous forme de VM) dans ton centre de données local pour lire les fichiers et les compresser/chiffrer avant l'envoi.
- **AWS vers AWS** : Si tu synchronises des données entre deux services AWS (ex: d'un compte AWS à un autre, ou d'une région à une autre), **aucun agent n'est requis**, AWS gère tout en interne.
- **Vers quels services AWS peut-il synchroniser ?**
 - Amazon S3
 - Amazon EFS
 - **AWS FSx** (toutes les versions : Windows, Lustre, ONTAP, OpenZFS).
- **Préservation des métadonnées** : C'est capital pour l'examen. DataSync préserve les permissions des fichiers, les dates de modification, les ownerships (UID/GID pour Linux, ACL pour Windows).

Exemple

? Le Scénario : La sauvegarde et migration d'un studio d'animation 3D

Tu es l'architecte cloud d'un studio de cinéma d'animation. Vos graphistes travaillent en local sur des serveurs de stockage ultra-rapides (des NAS de marque NetApp ou des serveurs de fichiers

Linux Windows). Chaque nuit, les artistes génèrent **20 To de nouvelles images et vidéos haute définition** (des millions de petits fichiers de rendu).

- **Le besoin** : Tu veux répliquer automatiquement ces 20 To de données chaque nuit vers AWS (dans un système **AWS FSx for Windows** ou **Amazon S3**) pour que l'équipe de production à l'autre bout du monde puisse travailler dessus le lendemain.
- **Le défi** : Un simple script de copie (`rsync` ou `robocopy`) mettrait trop de temps, saturerait la ligne internet, et perdrait les droits d'accès (permissions Windows ACL / Linux) des fichiers.

?? Comment ça se passe en pratique (Étape par Étape)

Voici comment tu mets en place **AWS DataSync** pour automatiser cela :

Étape 1 : Installer l'Agent en local

Dans ton centre de données, tu télécharges l'**Agent DataSync** fourni par AWS sous forme de machine virtuelle (VM) et tu l'installes sur tes serveurs (VMware ou Hyper-V).

- Cet agent va s'occuper de "scanner" ton stockage local très rapidement, de compresser les données et de les chiffrer avant l'envoi.

Étape 2 : Configurer la "Tâche" (Task) dans AWS

Dans la console AWS, tu crées une **Task DataSync** :

1. **Source** : Tu indiques le chemin de ton NAS local (ex: via le protocole NFS ou SMB).
2. **Destination** : Tu choisis ta cible AWS (ex: un système de fichiers **AWS FSx for NetApp ONTAP** ou un bucket **S3**).
3. **Fréquence** : Tu planifies l'exécution (ex: tous les soirs à 1h du matin).

Étape 3 : La synchronisation intelligente en action

À 1h du matin, la tâche démarre :

1. L'agent local analyse les fichiers. Contrairement à un outil classique, DataSync utilise un protocole réseau propriétaire d'AWS ultra-optimisé qui **parallélise le transfert**. Il s'adapte pour saturer au maximum (mais proprement) ta bande passante.
2. **Transfert incrémental** : DataSync compare la source et la destination. Si un fichier vidéo de 5 Go n'a pas changé, il ne le renvoie pas. Il n'envoie que les nouveautés ou les modifications.
3. **Préservation des métadonnées** : Si le fichier appartenait à l'utilisateur "Graphiste_Jean" avec des droits de lecture spécifiques en local, DataSync réapplique exactement les mêmes droits une fois le fichier arrivé dans AWS FSx.
4. **Vérification de l'intégrité** : À la fin du transfert, DataSync calcule une clé de vérification (checksum) pour chaque fichier pour s'assurer qu'aucun octet n'a été corrompu pendant le voyage sur internet.

? Ce qu'il faut retenir pour l'examen SAA-C03 (La différence clé)

On confond souvent **Storage Gateway** et **DataSync**. Voici le moyen infaillible de ne plus les mélanger :

- **AWS Storage Gateway est un PONT (Temps réel) :** Tes applications locales écrivent dedans *au fil de l'eau* tout au long de la journée. C'est une extension de ton stockage qui garde un cache local.
- **AWS DataSync est un CAMION DE DÉMÉNAGEMENT (Planifié / Batch) :** Il est là pour **déplacer ou synchroniser** de gros volumes de fichiers existants d'un point A à un point B, le plus vite possible, de manière planifiée (ex: une fois par jour, ou pour une migration définitive).

Revision #2

Created 2026-06-04 09:44:53 UTC by Victor

Updated 2026-06-04 10:08:22 UTC by Victor