**Documentation pour la Reprise de Données dans une Table Hive**

Cette documentation décrit les étapes à suivre pour effectuer une reprise de données sur une table Hive. Les cas suivants sont couverts :

* Ajout d'une nouvelle colonne,
* Mise à jour d'une colonne existante,
* Suppression de données insérées dans une table après une certaine date.

Un processus de sauvegarde des tables dans une base temporaire avant la reprise est également inclus.

**Processus de Backup des Tables Avant la Reprise**

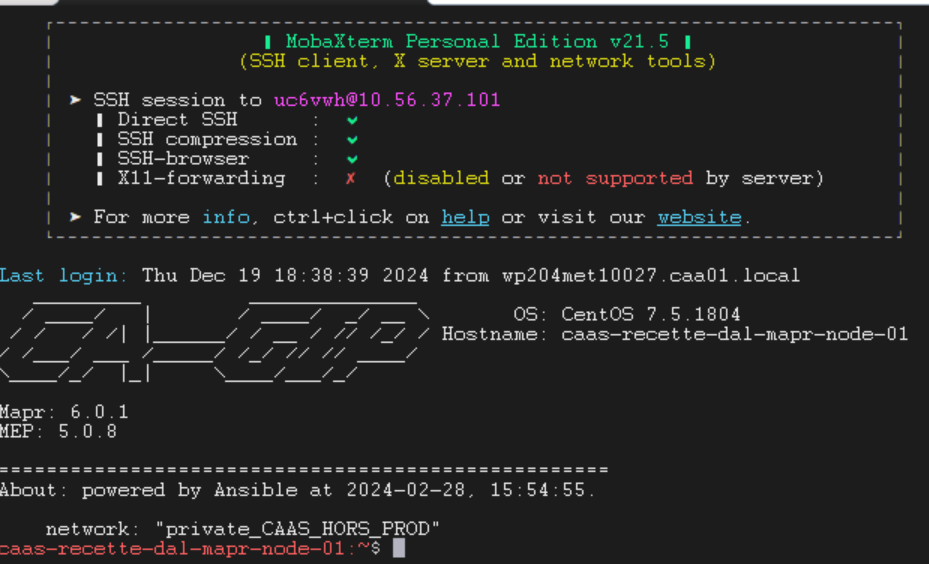
1. Créer une sauvegarde des tables concernées dans une base temporaire :

Avant de faire la reprise, il est fortement recommandé de faire un backup des tables à reprendre dans une base temporaire. Ces backups serviront lors de la reprise des données, mais également en cas de fausse manip.



Pour cela il faudra utiliser le script backup.sh ci-joint.

Pour faire la reprise des données sur l’environnement ***d5*** par exemple il :

* Se connecter au serveur edge de dev (10.56.37.101) via MobaExterm en utilisant son compte LDAP 
* Il faut utiliser le user applicatif: `ud8ddd1` pour se connecter :

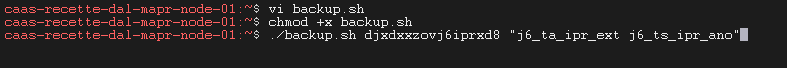
*sudo su – ud8ddd1*

* Une fois connecter, créer un fichier script « backup.sh » sur le serveur en utilisant la commande : *vi backup.sh* et y coller le contenu du script ***backup.sh*** joint à ce document.
* Rendre le fichier exécutable avec la commande : *chmod +x backup.sh*
* Pour exécuter le script il faut lui donner 2 arguments :

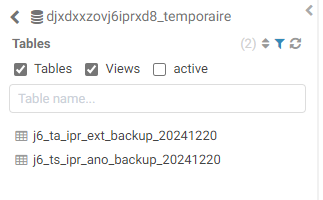
*# Usage: ./backup\_tables.sh database\_name "table1 table2 table3"*

* + L’argument 1 correspond au nom de la base de données dans laquelle se trouve les tables à reprendre
  + L’argument 2 correspond à liste des tables dont nous souhaitons faire un backup
  + A la fin de l’exécution, le script va créer une nouvelle base de données avec le nom de la base passé en argument suffixer de « \_temporaire » dans laquelle seront créés les backups des tables passées en avec le nom des tables suffixé de « **\_backup\_yyyyMMdd** ». yyyyMMdd correspond à la date du jour au format yyyyMMdd

Exemple :

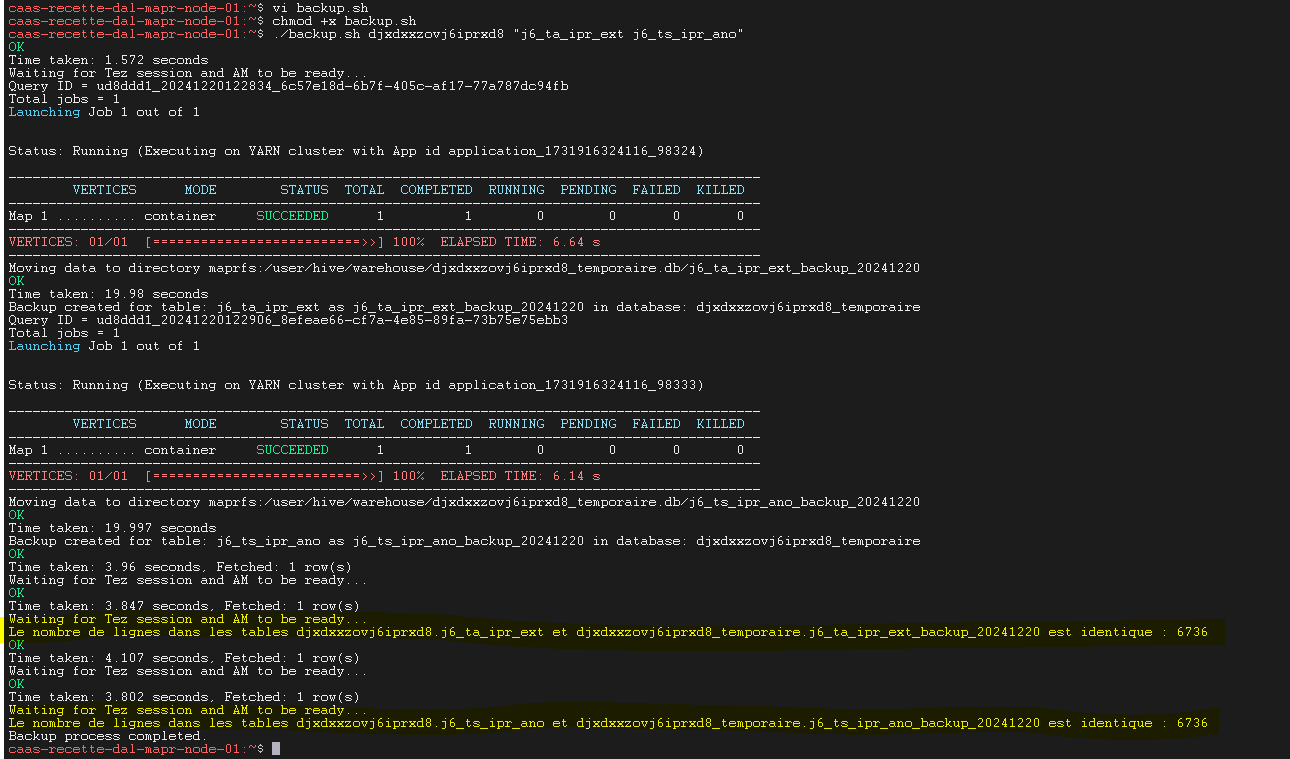


L’exécution du script avec les arguments ci-dessus donnera le résultat suivant :



1. Vérifier que les tables de backup contiennent toutes les données nécessaires :

A la fin du script, une comparaison de volume est faite entre la table et son backup pour s’assurer que les données ont été bien backupées.



Il faudra ensuite se connecter à Hue pour afficher un échantillon de données pour les backup en se servant de la requête ci-dessous :

*SELECT \* FROM <temp\_db>.<nom\_table\_backup> limit 100;*

**Cas 1 : Ajout d'une Nouvelle Colonne**

En cas de rajout d’une nouvelle colonne à une table existante, pour ne pas perdre les données existantes il faut suivre les **Étapes :**

1. Créer d’abord un backup de la table (se référer à **Processus de Backup des Tables Avant la Reprise**)
2. Une fois que la table est backupée, il faut alors réinstaller la table (en s’assurant que la nouvelle colonne est dans le catalogue de la version de l’usage déployée). En lançant le script de clean install. Exemple de lanement ci-dessous pour un usage et une source
   * Source

*/mapr/caas-recette-cluster/dd/d8d/apps/socle/current/management/data/data-catalog/bin/sfdb-db-util-caas.sh -cmd clean-install -objs rsrg.dd -tables dd\_tr\_ipr\_par\_dcl -ctype SOURCE*

* + Usage

*/mapr/caas-recette-cluster/dd/d8d/apps/socle/current/management/data/data-catalog/bin/sfdb-db-util-caas.sh -cmd clean-install -objs iprx.j6 -tables j6\_ts\_ipr\_dqa\_rej -ctype SOURCE*

/!\ Ces lancements sont à adapter selon l’usage.

1. Réalimenter la table en utilisant les données de la table de backup et en appliquant la règle de gestion sur la nouvelle colonne. Il faudra via Hive utiliser la commande **INSERT INTO TABLE.**

Ci-dessous un exemple de requête pour la reprise de données pour ajout de la colonne ia\_etp\_app avec la valeur « **13** »

* Installation de la nouvelle structure de la table hive

*/mapr/caas-recette-cluster/dd/d8d/apps/socle/current/management/data/data-catalog/bin/sfdb-db-util-caas.sh -cmd clean-install -objs rsrg.dd -tables dd\_tr\_eai\_etp -ctype SOURCE.*

* Via Hive en ligne de commande, la requête ci-dessous permet d’alimenter la table.

*insert into table ddddxxzbpddrsrgd8.dd\_tr\_eai\_etp select id\_\_ ,ct\_dcl ,cd\_etp\_pcs ,ll\_etp\_pcs ,13 as ia\_etp\_app ,dt\_maj\_par ,fl\_act ,ing\_sender ,ing\_tms\_sender ,ing\_sender\_id ,ing\_channel ,ing\_tms\_ingestion ,ing\_correlation\_id ,pck\_tms\_data ,pck\_state ,pck\_completude ,pck\_structure ,pck\_tms\_packing ,pck\_packing\_id from ddddxxzbpddrsrgd8\_temporaire.dd\_tr\_eai\_etp\_backup\_20241220;*

1. Une fois la reprise terminée il faut faire un count pour s’assurer qu’il n’y a pas de perter de données. Dans le cas de notre exemple il faut exécuter les deux requêtes les deux requêtes puis comparer les résultats.

*SELECT count(\*) FROM ddddxxzbpddrsrgd8.dd\_tr\_eai\_etp;*

*SELECT count(\*) FROM ddddxxzbpddrsrgd8\_temporaire.dd\_tr\_eai\_etp\_backup\_20241220;*

**Cas 2 : Mise à Jour d'une Colonne Existante**

L’étape de mise à jour d’une colonne existante est assez similaire à celle de l’ajout d’une nouvelle colonne. A la différence que lors de la mise à jour d’une colonne existante nous n’avons pas besoin de faire un *clean install*

**Étapes :**

1. Créer d’abord un backup de la table (se référer à **Processus de Backup des Tables Avant la Reprise**)
2. Il faut vider toute la table si la table n’est pas partitionnée, vider la partition concernée par la reprise si la table est partitionnée.

truncate table <nom\_base>.<nom\_table> ;

ou

alter table <nom\_base>.<nom\_table> drop partition(parti1=XXX, parti2=XXX, ….) ;

1. Alimenter la table en utilisant les données de la table de backup et en appliquant la règle de gestion sur la nouvelle colonne. Il faudra via Hive utiliser la commande **INSERT INTO TABLE.**

Ci-dessous un exemple de requête pour la reprise de données pour ajout de la colonne ia\_etp\_app avec la valeur « **13** »

* Installation de la nouvelle structure de la table hive

*/mapr/caas-recette-cluster/dd/d8d/apps/socle/current/management/data/data-catalog/bin/sfdb-db-util-caas.sh -cmd clean-install -objs rsrg.dd -tables dd\_tr\_eai\_etp -ctype SOURCE.*

* Via Hive en ligne de commande, la requête ci-dessous permet d’alimenter la table.

*insert into table ddddxxzbpddrsrgd8.dd\_tr\_eai\_etp select id\_\_ ,ct\_dcl ,cd\_etp\_pcs ,ll\_etp\_pcs ,13 as ia\_etp\_app ,dt\_maj\_par ,fl\_act ,ing\_sender ,ing\_tms\_sender ,ing\_sender\_id ,ing\_channel ,ing\_tms\_ingestion ,ing\_correlation\_id ,pck\_tms\_data ,pck\_state ,pck\_completude ,pck\_structure ,pck\_tms\_packing ,pck\_packing\_id from ddddxxzbpddrsrgd8\_temporaire.dd\_tr\_eai\_etp\_backup\_20241220;*

1. Une fois la reprise terminée il faut faire un count pour s’assurer qu’il n’y a pas de perte de données. Dans le cas de notre exemple il faut exécuter les deux requêtes les deux requêtes puis comparer les résultats.

*SELECT count(\*) FROM ddddxxzbpddrsrgd8.dd\_tr\_eai\_etp;*

*SELECT count(\*) FROM ddddxxzbpddrsrgd8\_temporaire.dd\_tr\_eai\_etp\_backup\_20241220;*

**Cas 3 : Suppression de Données Insérées Après une Certaine Date**

Très souvent nous avons besoin de vider, nettoyer une ou plusieurs tables d’un usage suite à des bugs ou des ingestions de fichiers incorrects et relancer les traitements à partir de l’étape en défaut

Pour se faire il faut suivre les étapes ci-dessous :

1. Back up des tables

Il faut d’abord faire un backup de la ou les table(s) qu’on souhaite nettoyer et la table de suivi en utilisant le script *backup.sh*.

Supposons vouloir faire une reprise de données de la table j6\_ta\_ipr\_ext. Il faut donc faire un backup de la table de suivi et de la table EXT.

*./backup\_tables.sh djxdxxzovj6iprxd8 "j6\_ta\_ipr\_ext j6\_ts\_ipr"*

1. Nettoyage de la table de Suivi

Il faudra ensuite nettoyer la table de suivi (supprimer toutes les lignes qui viennent après l’étape qu’on souhaite reprendre)

Si on souhaite par exemple faire la reprise à partir de l’étape **« 4 »**, il faut relever la date d’exécution de cette étape dans la table de suivi.

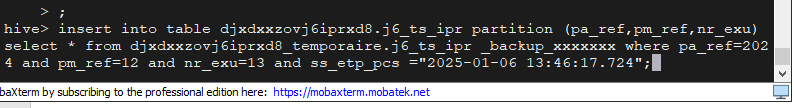


ss\_etp\_pcs = 2025-01-06 13:46:17.724

Faire un drop de la partition concerner :

*alter table djxdxxzovj6iprxd8.j6\_ts\_ipr drop partition (pa\_ref=2024,pm\_ref=12,nr\_exu=13);*

Alimenter maintenant la table de suivi par les données de la table de backup en faisant un filtre sur les données en ne prenant que les lignes qui ont une date avec la date d’exécution de l’étape « 4 » précédemment relevé.

*insert into table djxdxxzovj6iprxd8.j6\_ts\_ipr partition (pa\_ref,pm\_ref,nr\_exu) select \* from djxdxxzovj6iprxd8\_temporaire.j6\_ts\_ipr* *\_backup\_xxxxxxx where pa\_ref=2024 and pm\_ref=12 and nr\_exu=13 and ss\_etp\_pcs ="2025-01-06 13:46:17.724";*

1. Nettoyage des tables à reprendre

Une fois le nettoyage de la table de suivi terminé, il faut supprimer dans les tables à reprendre toutes les lignes qui ont été ingérés après la date de l’étape à laquelle on souhaite reprendre. Ces lignes seront identifiées par le champs prc\_tms\_data.

Si nous continuons notre exemple sur la table EXT

Il faudra faire le command hive en command line  :

* + Un drop de la partition concernée de la table EXT

*alter table djxdxxzovj6iprxd8.j6\_ta\_ipr\_ext drop partition(pa\_ref=2024,pm\_ref=12,no\_exu=13);*

* + Un insert des données de la table backup dans la partition supprimée de la table EXT en prenant le soin de bien récupérer que les données qui ont une date prc\_tms\_data < à la date de l’étape dans la table de suivi.

*insert into TABLE djxdxxzovj6iprxd8.j6\_ta\_ipr\_ext partition (pa\_ref,pm\_ref,no\_exu) select \* from djxdxxzovj6iprxd8\_temporaire.j6\_ta\_ipr\_ext\_backup\_yyyyMMdd where pa\_ref=2024 and pm\_ref=12 and no\_exu=13 and prc\_tms\_data < "2025-01-06 13:46:17.724";*

|  |  |
| --- | --- |
|  | S’il y’a eu modification ou génération de données pour une des étapes qui vient après l’étape « 4 » Il faudra faire ce nettoyage de données sur la table alimentée par l’étape. |