**Guide de correction d’anomalie dans le Data Warehouse**

Ce guide suppose une connaissance pratique de l'Agent SQL Server. Il donne une explication détaillée du chargement et la manière appropriée de résoudre les problèmes lorsqu'ils surviennent. Il est conseillé de conserver un fichier d'instructions EXEC de chaque proc qui exécute le chargement. Ce document doit être utilisé conjointement avec le Guide du développeur ETL.

**Flux de données**

Les données circulent dans l'entrepôt de données via l’ODS et l’ODS uniquement. L’ODS est la porte d'entrée du système et cela doit toujours rester vrai pour garantir la facilité de maintenance du système.

Les jobs d'agent qui gèrent le chargement de l'entrepôt sont nommés par ordre d'exécution. Ce schéma ne devrait pas non plus dériver loin de son état initial. Vous pouvez insérer des éléments supplémentaires, mais les relations ordinales doivent être définies dans la pierre. Par exemple, les tables de faits ne doivent jamais être traitées avant les tables de dimension.

Toutes les étapes du chargement d'entrepôt sont couplées de manière séquentielle, ce qui signifie que vous pouvez exécuter toutes les étapes indépendamment, à condition que cette étape précédente se soit exécutée à un moment donné. En d'autres termes, si une étape explose, vous n'avez pas besoin de revenir en arrière et d'exécuter l'ensemble du processus. Cela dit, une fois que vous vous êtes familiarisé avec le processus, le moyen le plus rapide de résoudre les problèmes consiste généralement à purger le staging, à extraire des données spécifiques et à tout réexécuter.

Les données circulent dans l'entrepôt de données de la manière suivante. Dans tous les cas, vous ne traitez que les données marquées comme nettoyées, non traitées et sans erreur.

1. Les données sont extraites des systèmes source dans une table de staging spécifique, puis dans ODS.
2. Les données sont nettoyées.
3. Les données des dimensions de production sont extraites dans le modèle commun dans ODS.
4. Les données dimensionnelles existantes sont comparées aux données dimensionnelles entrantes et des actions sont prises en fonction des règles entourant une dimension particulière et chargées dans Production.
5. Les champs dimensionnels utilisés pour identifier les enregistrements « de la table des faits » sont renvoyés vers ODS pour être mis en correspondance avec les faits.
6. Les enregistrements de faits sont mis en correspondance avec leurs dimensions et chargés dans le modèle commun.
7. Le modèle commun est mis en production.

**Les steps de correction d’anomalie**

Tout processus de correction est à la fois un travail de procédure et un travail de détective. Lire le guide des développeurs ETL et vous assurer que vous avez une bonne compréhension du système vous aidera à résoudre les problèmes plus rapidement que tout processus étape par étape. Cela étant dit, ce qui suit est une bonne heuristique pour trouver des problèmes dans le processus ETL de l'entrepôt.

1. Ouvrez l'étape qui est tombé ko et déterminez quelle était l'erreur.
2. S'il était lié à un proc, ouvrez la sous-étape et extrayez les instructions EXEC et exécutez-les à la main une par une. Le problème va apparaitre. (C'est là que votre fichier .sql d'instructions EXEC de chaque proc est utile.)
3. Déterminez et corrigez l'erreur. (C'est ici que vous sortez votre pipe Sherlock Holmes et que vous allez au travail. Il n'y a pas de méthode pour le faire.)
4. Relancez l'étape qui est tombé ko.

Le processus est conçu pour que tout problème avec les dimensions de chargement arrête le processus. Le chargement des dimensions est une condition suffisante nécessaire pour le chargement des faits. Les faits, en revanche, peuvent tomber ko et le reste des tables de faits sera autorisé à poursuivre le traitement.

Le traitement des dimensions doit être très rapide, de sorte que tout temps à long terme sur le traitement des dimensions doit être considéré comme une source de suspicion. Voici les principales causes des problèmes de chargement d'entrepôt.

1. Enregistrements en double dans le staging suite à une mauvaise logique dans l'extraction source. L'instruction SQL MERGE ne peut pas traiter les enregistrements en double à partir d'une seule source.
2. Troncature de données. Les tailles des types de données de chaîne sont déterminées en profilant les données existantes, puis en doublant la valeur de caractère maximale existante. Cela ne fonctionne pas toujours avec le temps. Vous devrez systématiquement déterminer quelle colonne est à l'origine du problème.
3. Extraire à nouveau les données source avant de traiter l'exécution précédente. Cela entraînera une erreur d'enregistrement en double.

**Définitions des colonnes d'audit**

Chaque table possède un ensemble normalisé de colonnes d'audit utilisées pour la provenance des données. Les définitions de ces colonnes et leurs utilisations sont ci-dessous.

Tables de staging

ETLKey - il s'agit d'un identifiant unique qui fait plus qu'identifier des enregistrements uniques dans le staging. Il est utilisé dans le processus pour faire correspondre les enregistrements de faits avec leur dimension. Pendant le dépannage, vous pouvez utiliser ETLKey dans la table de faits du modèle commun, pour suivre les faits dans leur dossier d'origine.

UniqueDims - Une valeur qui représente l'intersection de toutes les dimensions qui composent un fait. La valeur est créée à l'aide de l'algorithme SHA. Cette valeur est créée par une colonne calculée dans le modèle et la production communs et renvoyée à la table des étapes. Cette colonne peut être utilisée pour identifier les enregistrements en double. UniqueDims est utilisé pour déterminer ce qui a été intégré en production.

UniqueRows - Rarement utilisé, UniqueRows est également créé par une colonne calculée de la même manière que UniqueDims, sauf qu'il utilise l'algorithme de hachage MD5. UniqueRows est souvent utilisé dans le cas de tables de faits qui doivent être mises à jour là où elles représentent un hachage des faits actifs. Il est également utilisé dans le cas où la table de staging est une relique des données de dimension ET est également une dimension indésirable(junk dimension) auquel cas, là encore, il s'agit d'un hachage des valeurs de cet enregistrement.

SourceSystem - Une colonne contenant le nom du système d'où proviennent les données. Dans le cas où les données sont développées à partir de plusieurs sources, la valeur sera «multi-source».

Cleansed - Une valeur binaire qui indique si l'enregistrement a été nettoyé ou non. Utilisé dans toutes les tables, il est particulièrement utile lorsque le nettoyage implique des instructions CASE où le traitement à plusieurs reprise peut modifier la valeur.

ErrorRecord - Valeur binaire utilisée pour indiquer un problème avec l'enregistrement. Il s'agit d'un fichier multi-usage qui peut être utilisé pour toute erreur définie par l'utilisateur.

Processed - Valeur binaire qui indique si l'enregistrement a été traité(déplacé en production) ou non.

RunDate - Horodatage de la date à laquelle les données ont été extraites. Utile pour déterminer s'il y a eu plusieurs exécutions.

Pour qu'un enregistrement soit traité, Cleansed, ErrorRecord et Processed doivent contenir respectivement les valeurs 1, 0, 0.

Tables de production (schema dim et fact)

CreatedOn - Date de création de l'enregistrement. Cette valeur doit TOUJOURS être remplie.

CreatedBy - Qui a créé l'enregistrement. Cette valeur doit TOUJOURS être remplie.

UpdatedOn - Date de mise à jour de l'enregistrement. Cette valeur doit TOUJOURS être remplie. Lorsque vous effectuez des mises à jour manuelles avec INSERT ou UPDATE, assurez-vous que votre instruction SQL remplit cette valeur ou vous risquez que des choses vraiment terribles se produisent.

UpdatedBy - Qui a mis à jour l'enregistrement. Cette valeur doit TOUJOURS être remplie. Lorsque vous effectuez des mises à jour manuelles avec INSERT ou UPDATE, assurez-vous que votre instruction SQL remplit cette valeur ou vous risques que des choses vraiment terribles se produisent.

SourceSystem - Identique au staging.

SourceSystemKey - La clé d'identification unique du système source. Il s'agit généralement d'une clé naturelle mais peut être une clé primaire dans certains cas d'utilisation. Utilisé dans les tables qui ne sont pas des dimensions indésirables.

EffectiveFrom - Utilisé dans les tables de dimensions pour déterminer quand l'enregistrement a commencé à être effectif.

EffectiveTo - Utilisé dans les tables de dimension pour déterminer quand l'enregistrement a expiré.

IsMostRecentRecord - Indicateur binaire pour déterminer quel enregistrement est la valeur actuelle de la dimension.

RowHash - Utilisé dans les tables de dimensions indésirables. Il s'agit d'un champ calculé utilisant l'algorithme de hachage MD5 qui représente les valeurs de cette ligne qui est généralement une combinaison unique de colonnes. Peut également être utilisé dans des dimensions de Type II à évolution lente où vous ne souhaitez mettre à jour que certaines colonnes.

UniqueDims - Identique au staging.