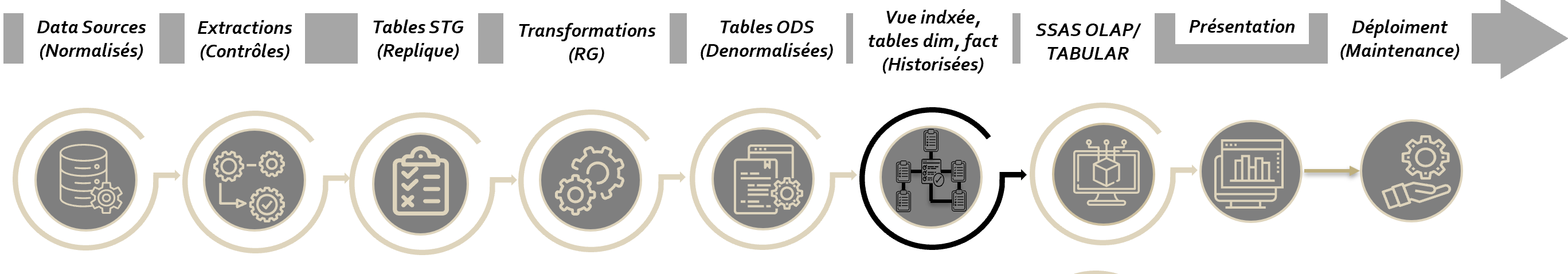
Guide des développeurs ETL

Aperçu :

L'environnement ETL existe dans un cadre bien défini. Ce cadre sera précisé dans ce document. L'environnement est conçu pour accepter des données provenant de nombreuses sources et les combiner en un seul flux de données qui se termine dans l'entrepôt de données. Les données se déplacent par étapes. Chaque étape sera décrite ci-dessous.



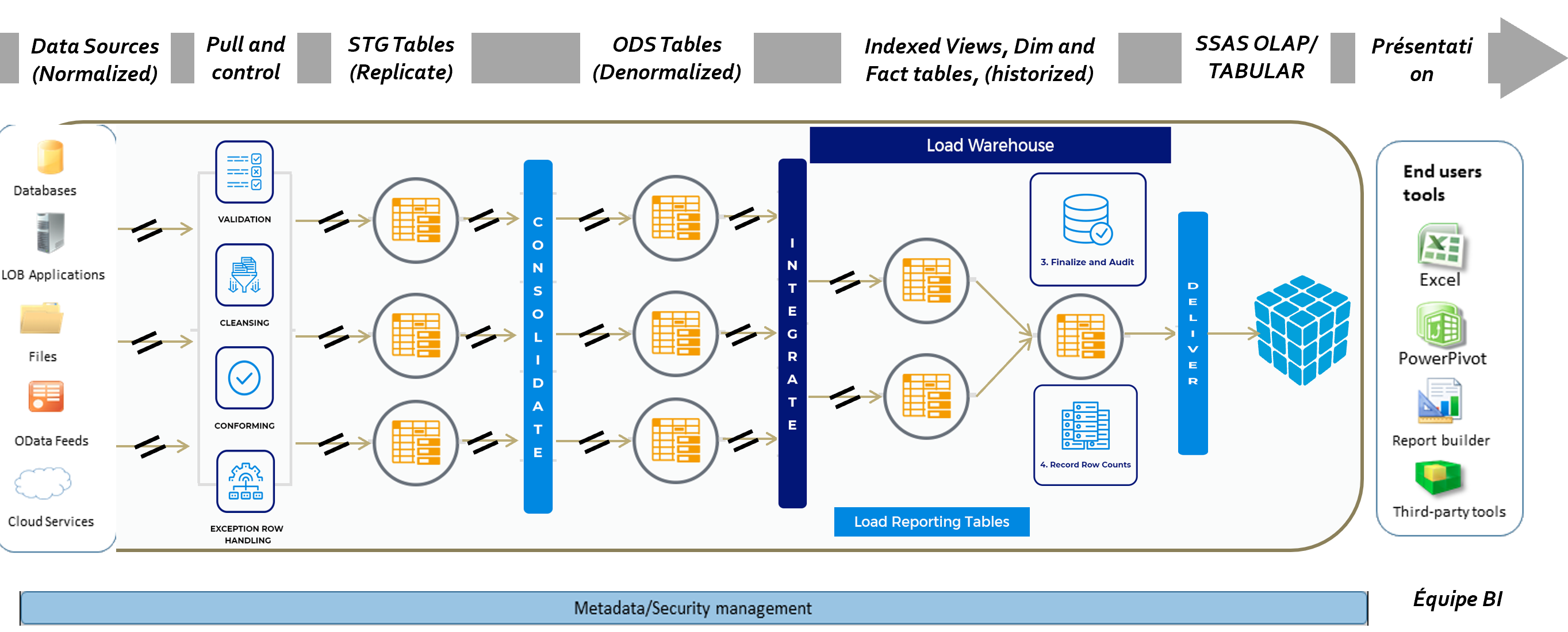


Figure 1. Processus de chargement du DWH généralisé.

La figure 1 représente le processus de chargement.

PRESENTER / DECRIRE LA FIGURE 1

Ordre des processus ETL

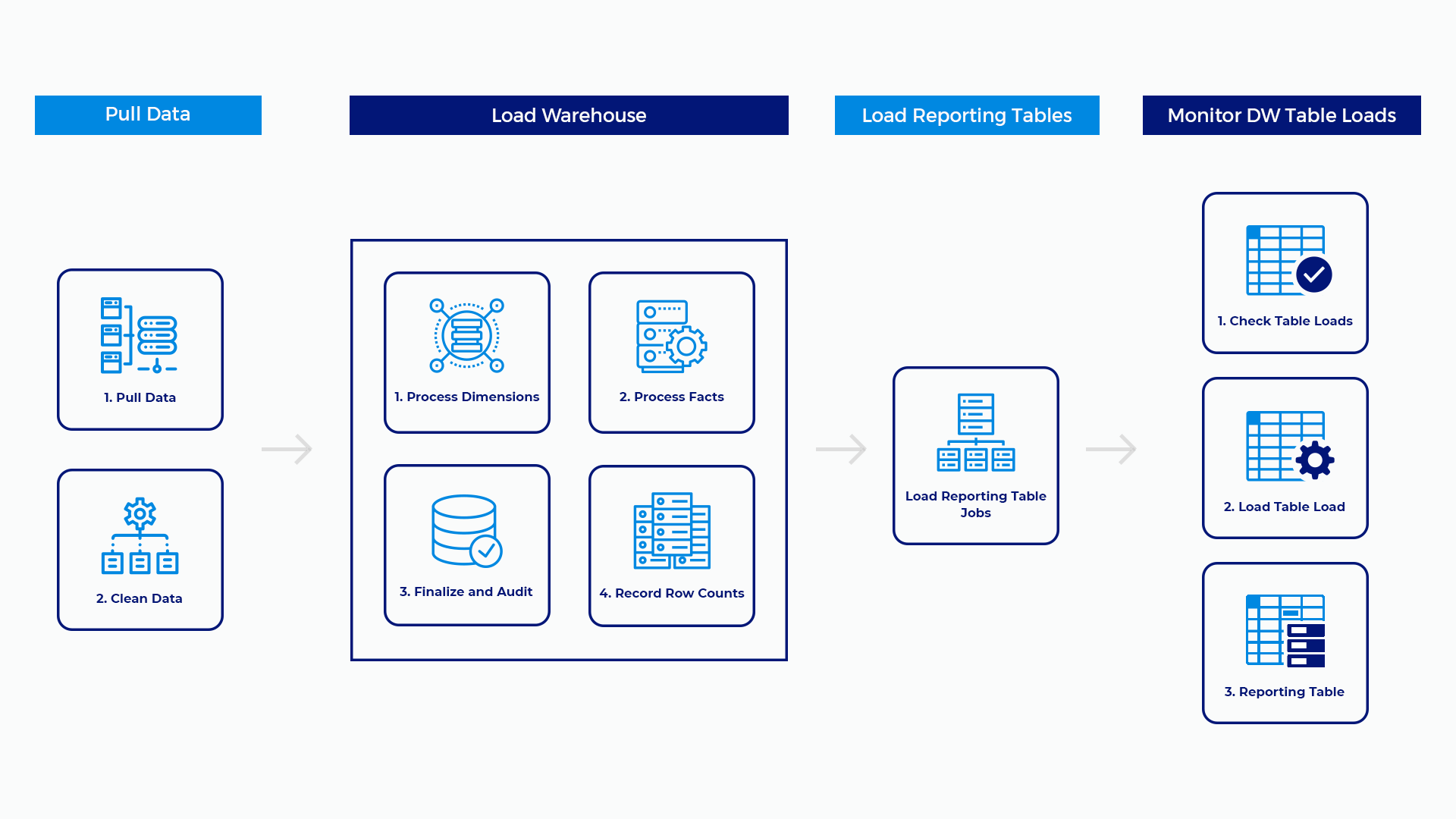


Figure 2. Séquence de travaux ETL généralisée

**Extraction de données**

En général, tous les processus qui acquièrent des données de systèmes externes extraient et nettoient les données au cours de la même étape.

**Entrepôt de chargement**

Les processus de chargement du DWH réels sont une combinaison de tables de chargement et de processus d'audit qui sont utilisés pour vérifier et s'assurer que tout est bien chargé. Lorsque vous créez des processus ETL, vous devez vous assurer d'ajouter du code qui prend en compte ces processus. Toutes les dimensions sont traitées en premier; puis toutes les tables de faits sont traitées. Le chargement de la table de faits communiquent de nouveau avec le stg pour prendre note de ce qui a été exactement chargé dans l'entrepôt. Le processus de finalisation et d'audit vérifie les enregistrements intermédiaires et signale les écarts. Chaque chargement de table de faits que vous créez aura besoin de code ajouté à la procédure stockée usp\_MarkRecordsAsProcessed. Le nombre de lignes d'enregistrement est entièrement automatisé et documente le nombre d'enregistrements chargés dans chaque table de faits et le compare aux chargements historiques pour déterminer s'il y a une anomalie dans le processus de chargement.

**Table de reporting des chargements**

Ce travail est destiné aux processus qui chargent des tables de rapports dénormalisées résidant uniquement dans la base de données de rapports. Les table de reporting qui vivent dans ODS sont réservés aux professionnels des données et sont chargés par différents processus.

**Surveiller les chargements de table DW**

Vérifier les chargements des tables rendra compte de la part des chargements non effectués. De fréquent à pas du tout sur une période de trois jours sont géré par des triggers. Dans la dernière étape, les données volumétriques sont transférées dans un tableau de rapport dénormalisé pour une digestion facile.

**Nettoyage des données**

Comme pour tout processus ETL, certaines des données qui entrent dans le système doivent être nettoyées et normalisées avant d'être chargées. Le nettoyage des dates est une tâche courante et est décrit ci-dessous. Utilisez la fonction udf\_CleanDates () pour nettoyer correctement les dates. Dans le dossier SQL Scripts / Framework Objects, ouvrez le fichier CREATE FUNCTION udf\_CleanDates.sql et affectez les variables de date vide et d'erreur selon votre système. La fonction de udf\_Cleandates est décrite ci-dessous.

1. Définissez des dates vides sur une date arbitraire. Nous recommandons 99991230.
2. Définissez les dates d'erreur à une date arbitraire. Nous recommandons 99991231.
3. Convertir les dates restantes au format AAAAMMJJ

**Chargement des Dimensions**

La figure trois illustre le processus généralisé de chargement des dimensions de l'entrepôt. La plupart des processus de chargement de dimension suivent ce modèle de conception. Le processus général est expliqué comme suit: Chargez la table de modèle commune à partir de l'entrepôt. Fusionnez les données existantes avec de nouvelles données. Déterminez quels enregistrements sont nouveaux et chargez-les dans l'entrepôt. Cela peut également être un processus de fusion si vous mettez à jour d'anciens enregistrements. Tronquez la table de modèle commune. Chargez la table de modèle commune avec les champs clés nécessaires pour remplir les tables de faits.

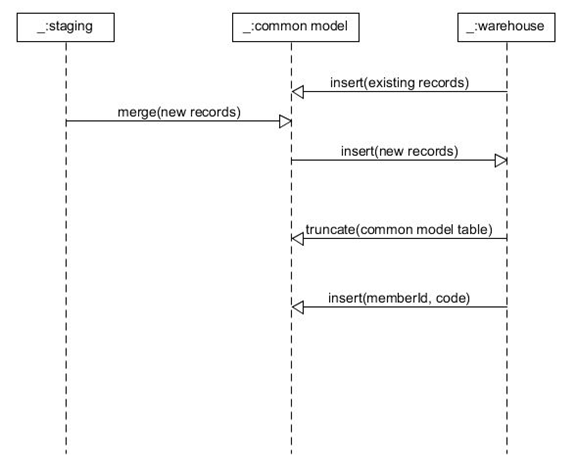


Figure 3. Diagramme de séquence de chargement des dimensions.

Commandes de chargement du DWH

1. Tu ne mettras pas de codes autonomes dans la base de données. Importez toujours les valeurs de code avec leurs explications en anglais si possible.
2. Tu ne placeras pas de données dans des colonnes nommées de manière générique comme Value1 et Value2. Chaque colonne doit avoir un nom clair et compréhensible de telle sorte que la signification des valeurs dans cette colonne soit entièrement sans ambiguïté, même sans avoir à consulter le dictionnaire de données.
3. Tu ne dois pas utiliser 1 ou 0 pour représenter les valeurs booléennes. Utilisez uniquement «Oui» et «Non».
4. Tu ne placeras pas de données de filtre textuelles dans des tables de faits.
5. Tu ne placeras pas de dates dans des tables de faits sans une connexion appropriée à une dimension de date.
6. Tu ne permettras pas si possible des drapeaux et des acronymes dans la base de données. Les systèmes sont généralement remplis de toutes sortes de drapeaux et d'autres acronymes qui nécessitent une expertise pour déchiffrer. Habituellement, un frontal décode ces drapeaux pour un utilisateur. Ces valeurs n'appartiennent pas à l'entrepôt de données et doivent être traduites dans le cadre du processus de nettoyage.

**Conventions**

Lors du développement de processus ETL, il y a un certain nombre de conventions à respecter. Ces conventions sont largement basées sur le bon sens, mais elles changent en fonction de ce que vous faites et de l'endroit où vous le faites.

**Dois-je utiliser SSIS?**

Neuf fois sur dix, la réponse à cette question sera non. En règle générale, n'utilisez SSIS que pour développer des processus ETL que si vous devez déplacer un fichier physique réel ou si vous avez une raison de faire quelque chose que T-SQL ne peut pas faire comme interagir avec le système de fichiers ou le système d'exploitation d'une manière ou d'une autre. La plupart des processus ETL déplacent les données d'une base de données vers une autre même si un fichier plat a été créé dans un processus intermédiaire. Dans le cas où vous devez traverser des serveurs, utilisez un serveur lié.

**Conventions pour SSIS**

Il existe un package de modèle standard. Utilisez-le comme point de départ pour tout ETL qui nécessite SSIS comme indiqué dans la section ci-dessus. La plupart des normes sont déjà intégrées à ce package. Tandis que vous continuez à développer votre processus, gardez à l'esprit les points suivants:

1. Essayez de réduire à zéro le nombre de valeurs codées en dur dans votre package. Placez toutes les configurations dans la table de configuration.
2. Essayez de nommer des variables en suivant les exemples de conventions du package.
3. Le package a actuellement deux paramètres pour déplacer des fichiers. Assurez-vous de les utiliser et de les connecter à l'environnement global.
4. Ne placez pas d'instructions SQL non arbitraires dans vos packages. Si votre code dépasse trois lignes, transformez-le en procédure stockée et appelez-le à la place.
5. La tâche de script doit être développée en utilisant C #.

**Conventions pour T-SQL**

Chaque procédure stockée doit avoir un bloc de documentation. Le bloc doit inclure le nom du développeur, la date de création du script, une description de ce que fait le script et un journal des modifications apportées au script. Tous les scripts T-SQL doivent être conformes aux conventions suivantes:

1. Tous les scripts doivent être justifiés à gauche.
2. Tous les mots de réserve SQL doivent être mis en majuscule.

**Conventions pour C #**

Essayez de vous conformer aux normes Microsoft, mais ce n'est pas une règle stricte car il ne devrait pas y avoir de prépondérance de développement C # dans l'environnement.

**Conventions pour les ODS**

Les conventions de dénomination dans ODS ne semblent suivre aucun modèle. En effet, une règle principale pour nommer les objets dans cette base de données est qu'ils doivent être conformes au système source. Les conventions pour les ODS sont les suivantes:

1. AUCUNE table dans ODS ne doit utiliser le schéma dbo. Chaque table a besoin d'un schéma qui identifie CLAIREMENT la fonction de cette table. S'il s'agit d'une table intermédiaire, le schéma doit identifier le système source des données. Quelques exemples:
   1. cm = modèle commun
   2. boa = Bank of America
   3. ss = Smart Soft
2. Les conventions de dénomination des tables de staging POUR LES COLONNES doivent être conformes aux conventions de dénomination du système source. En fait, les noms des colonnes doivent être identiques au système source.
3. Les conventions de dénomination des tables intermédiaires pour les NOMS DE TABLE doivent suivre le modèle de mise en majuscule à chaque premier mot.
4. Les tables de modèle communes ou ODS doivent être presque identiques aux tables du DWH qu'elles imitent, mais dépourvues de tous les index, clés et la plupart des contraintes.
5. Chaque table de scène doit avoir les colonnes suivantes, que vous les utilisiez ou non:
   1. Dim unique. Cette colonne est utilisée pour aider à déterminer quels enregistrements ont réellement été transférés à l'entrepôt.
   2. Enregistrement d'erreur. Valeur binaire pour indiquer qu'il y a un problème avec cet enregistrement.
   3. Traité. Enregistrement binaire utilisé pour identifier les enregistrements réellement déplacés vers l'entrepôt.
   4. Date d'exécution. Horodatage de la fin du chargement.

**Conventions pour le reporting**

La base de données de rapports ne suit aucune convention de dénomination de base de données standard. La base de données de rapports est principalement constituée de vues exposées aux utilisateurs. En tant que tels, les objets sont nommés de manière à être facilement lisibles par les utilisateurs.

**A retenir**

Ce document tente de résoudre les principaux problèmes liés au développement d'ETL par rapport à l'entrepôt de données. Bien sûr, nous ne pouvons pas résoudre tous les problèmes possibles que vous pourriez rencontrer. Heureusement, il existe un référentiel d'exemples et de modèles pour les travaux futurs. Si vous voulez savoir comment faire quelque chose, recherchez les processus existants qui font quelque chose de similaire.