Comprendre les dimensions à évolution lente (DAL)

Les **dimensions à évolution lente (SCD)** sont des méthodes utilisées dans les entrepôts de données pour gérer les changements dans les données dimensionnelles au fil du temps, tout en maintenant l'intégrité des données historiques et des rapports analytiques.

Les Données en cours d'utilisation sont les méthodes utilisées pour surveiller les changements dans les attributs des dimensions, gérer les mises à jour, aider les entreprises à préserver les données historiques et assurer l'exactitude des rapports. Dans l'Entrepôt de données, un problème courant est de gérer les changements dans les données dimensionnelles au fil du temps. C'est là que nous utilisons le concept de "Slowly Changing Dimensions" (SCD). Cette lecture fournira une brève explication sur les types de SCD et discutera également de leurs avantages, de leur utilisation et de leurs considérations dans la conception d'un entrepôt de données.

Différents types de SCD :

Il existe quatre principaux types de SCD :

- **Type 0 : Conserver la valeur d'origine**
- **Type de données 1 : Écraser les données existantes**
- **Type 2 : Préservation des données historiques**
- **Type 3 : ajout d'un nouvel attribut**

Cependant, dans la plupart des implémentations avancées, les types ci-dessous sont également utilisés, qui combinent ou étendent certains des types de base.

- **Type 4 : Tableau historique**
- **Type 6 : Approche hybride**

Type 0 : Conserver la valeur d'origine

Ce type peut être utilisé pour une dimension statique, ce qui signifie qu'une fois qu'une valeur est insérée, elle reste statique. Aucune modification ne sera apportée aux données de la dimension dans le Type 0. De même, les données historiques ne sont pas mises à jour. Cette approche est utile pour les données qui doivent rester constantes dans le temps. Il s'agit par exemple de codes de produits ou de numéros de compte. Le principal avantage du type 0 est qu'il est simple à mettre en œuvre et plus efficace pour les dimensions qui changent rarement.

Type de données 1 : Écraser les données existantes

Le Type de données 1, c'est-à-dire l'écrasement des données existantes, applique les changements à la dimension directement en écrasant les données existantes. Cette méthode ne conserve pas d'enregistrement des modifications historiques. Par

conséquent, si la valeur d'un attribut est mise à jour, l'ancienne valeur est perdue. Par exemple, lorsque seul l'état actuel des données est important, comme la correction de fautes d'orthographe ou la mise à jour de toute information de contact.

Avantages :

- Facile à mettre en œuvre.
- Permet d'économiser de l'espace de stockage.

Inconvénients :

- Aucune donnée historique n'est conservée.
- Peut conduire à des rapports historiques inexacts.

Exemple : Si un client change d'adresse, la nouvelle adresse écrase l'ancienne

Type de données 2 : Conserver les données historiques (versionnement des lignes)

Le Type de données 2, c'est-à-dire la conservation des données historiques, vous permet de suivre les changements en ajoutant de nouvelles lignes dans la table de dimension à chaque fois qu'il y a une mise à jour. Chacune des lignes comportera la version actuelle et la version historique des données et les dates de début et de fin (ou) les drapeaux sont utilisés pour indiquer si la ligne est la version actuelle. Par exemple, lorsqu'il est essentiel de conserver un historique complet des modifications, comme le suivi des changements d'adresse des clients à des fins de conformité légale/d'audit.

Avantages :

- Dans le Type de données 2, l'intégralité des données historiques est conservée.
- Il est plus facile d'extraire des données en cours d'utilisation, telles qu'elles existaient à un moment donné.

Inconvénients :

- Le type 2 augmente la taille de la table de dimension.
- Cela nécessite principalement une gestion minutieuse des champs de version.

Exemple : Une nouvelle ligne sera créée avec la nouvelle adresse lorsqu'un client met à jour son adresse, tandis que l'ancienne ligne sera marquée comme historique.

Type 3 : Ajout d'un nouvel attribut (suivi d'un historique limité)

Le type 3, qui consiste à ajouter un nouvel attribut, permet de suivre les modifications historiques en ajoutant de nouvelles colonnes à la table de dimension. Chacune de ces colonnes représente une version différente de l'attribut. Cette méthode est utile lorsque seule une quantité limitée de données historiques doit être stockée, comme les valeurs précédentes et actuelles. Par exemple, lorsque vous devez suivre un petit nombre de changements et qu'il est seulement nécessaire de comparer l'état précédent et l'état actuel.

Avantages :

- Le type 3 est facile à mettre en œuvre.
- Il nécessite beaucoup moins d'espace que le type 2.

Inconvénients :

- Le type 3 ne peut suivre qu'une quantité limitée d'historique.
- Il ne permet pas de conserver un historique complet des modifications.

Exemple : Stocker l'adresse actuelle et l'adresse précédente d'un client dans des colonnes distinctes sur la même ligne.

Type 4 : Tableau historique (suivi des données historiques dans un tableau distinct)

Dans le Type 4, les données historiques sont stockées dans une table distincte des données de dimension courante. Dans ce cas, la table de dimension principale ne contient que les données actuelles, tandis qu'une table historique distincte stocke toutes les versions antérieures des données. Par exemple, lorsque vous souhaitez séparer les données actuelles des données historiques afin d'améliorer les performances et de simplifier la conception.

Avantages :

- Le type 4 conserve un enregistrement historique complet.
- Il sépare généralement les données actuelles des données historiques.

Inconvénients :

- Le type 4 est plus complexe à mettre en œuvre.
- Il nécessite principalement un stockage supplémentaire pour les tables historiques.

Exemple : Une table de clients actuels ne comportant que les dernières informations mises à jour, tandis qu'une table de clients historiques associée conserve les enregistrements plus anciens.

Type 6 : approche hybride

Le type 6 est une approche hybride qui combine des aspects des **types 1**, **2** et **3**. Elle conserve l'historique complet comme le type 2, dispose d'un indicateur d'actualité comme le type 1 et suit les versions antérieures comme le type 3. Cette méthode permet d'accéder aux données actuelles, de les comparer aux versions précédentes et de conserver un enregistrement historique complet. Par exemple, si vous avez besoin d'une solution flexible pour suivre à la fois les versions actuelles et historiques des données, et si vous avez également besoin de comparer les valeurs antérieures.

Pour :

- Le type 6 combine les avantages d'autres types comme le type 1, le type 2 et le type 3.
- Il permet de suivre l'historique complet en conservant l'état actuel.

Inconvénients :

- Le type 6 est plus complexe à gérer.
- Il nécessite principalement plus d'espace de stockage.

Exemple : Lorsqu'un client change d'adresse, la table de dimension comporte un champ d'adresse actuelle (type 1), de nouvelles lignes dans la table pour suivre les changements historiques complets (type 2) et un champ d'adresse précédente (type 3).
Éléments clés à prendre en compte pour la mise en œuvre de la DSC :

1. **Exigences de l'entreprise :** Avant de choisir un type de DSC, évaluez les besoins de l'entreprise. Avez-vous besoin de suivre les modifications historiques ? Dans l'affirmative, quelle quantité d'historique devez-vous conserver ?
2. **Le versionnage :** Comme indiqué précédemment, le type 2 nécessite souvent une date de début, une date de fin et un indicateur actuel pour gérer les différentes versions d'une même ligne de dimension. Veillez à manipuler ces champs avec soin afin d'éviter les erreurs dans le contrôle des versions.
3. **Stockage et performances :** Le suivi des données historiques peut augmenter la taille des tables de dimension. Tenez toujours compte de l'impact sur les performances des requêtes qui accèdent aux tables de dimension.
4. **Processus d'extraction, de transformation et de chargement (ETL) :** Le processus ETL doit être conçu correctement pour s'adapter au type de DSC utilisé. Par exemple, l'ETL de type 1 ne fait que mettre à jour les lignes existantes, tandis que l'ETL de type 2 doit détecter les changements et insérer de nouvelles lignes.

Conclusion :
Les dimensions à évolution lente (SCD) constituent toujours un moyen très efficace de gérer les changements dans les données de dimension au fil du temps. Les entreprises peuvent garantir l'exactitude des rapports, conserver les données historiques et optimiser les performances de leurs entrepôts de données en sélectionnant soigneusement le type de SCD approprié en fonction de leurs besoins. Qu'il s'agisse d'un simple écrasement (Type 1), d'un suivi historique complet (Type 2) ou d'une solution hybride (Type 6), la bonne Stratégie de données vous aidera à réussir votre gestion de données à long terme.