Travaux pratiques : Utilisation des faits et des tables de dimensions

Temps estimé nécessaire : 30 minutes

Objectif du laboratoire :

Ce laboratoire est conçu pour vous guider dans le processus de conception d'un entrepôt de données pour un fournisseur de services cloud. Il se concentre sur l'utilisation des données de facturation fournies dans un fichier CSV pour créer un schéma en étoile, y compris la conception de tables de faits et de dimensions. Ce schéma prendra en charge les requêtes complexes liées à la facturation, telles que la facturation moyenne par client, la facturation par pays, secteur et catégorie, ainsi que les tendances au fil du temps.

Avantages de l'apprentissage en laboratoire :

En effectuant ce laboratoire, vous acquerrez des compétences pratiques en matière d'organisation et d'analyse de grands ensembles de données à l'aide de techniques d'entreposage de données. Ces compétences sont essentielles pour prendre des décisions commerciales éclairées, optimiser la récupération des données et améliorer la compréhension des relations entre les données. Ces connaissances sont particulièrement utiles dans des scénarios réels, tels que l'analyse des données de facturation cloud, où elles peuvent conduire à une gestion des données plus efficace et à des analyses approfondies.

Objectifs

Dans ce laboratoire, vous allez :

- Étudiez le schéma du fichier csv donné
- Concevoir les tables de faits
- Concevoir les tables de dimensions
- Créer un schéma en étoile à l'aide des tables de faits et de dimensions

À propos de Skills Network Cloud IDE

L'IDE Cloud de Skills Network (basé sur Theia et Docker) fournit un environnement pour les travaux pratiques liés aux cours et aux projets. Theia est un IDE open source (environnement de développement intégré), qui peut être exécuté sur un ordinateur de bureau ou sur le cloud. Pour réaliser ce laboratoire, nous utiliserons l'IDE Cloud basé sur Theia exécuté dans un conteneur Docker.

Avis important concernant cet environnement de laboratoire

Veuillez noter que les sessions de cet environnement de laboratoire ne sont pas permanentes. Un nouvel environnement est créé pour vous à chaque fois que vous vous connectez à ce laboratoire. Toutes les données que vous avez pu enregistrer lors d'une session précédente seront perdues. Pour éviter de perdre vos données, prévoyez de terminer ces laboratoires en une seule session.

Exercice 1 : Étudiez le schéma du fichier csv donné

Dans ce laboratoire, nous allons concevoir un entrepôt de données pour un fournisseur de services cloud.

Le fournisseur de services cloud nous a fourni ses données de facturation dans le fichier csv `cloud-billing-dataset.csv`. Ce fichier contient les données de facturation de la dernière décennie.

Voici les détails par champ des données de facturation.

Nom du champ	Détails
identifiant client	Id du client
catégorie	Catégorie de client. Exemple : Particulier ou Entreprise
pays	Pays du client
industrie	À quel domaine/secteur appartient le client. Exemple : juridique, ingénierie
mois	Mois facturé, enregistré au format AAAA-MM. Exemple : 2009-01 fait référence au mois de janvier de l'année 2009
montant facturé	Montant facturé par les services cloud fournis pour ce mois en USD

Nous devons concevoir un entrepôt de données capable de prendre en charge les requêtes répertoriées ci-dessous :

- facturation moyenne par client
- facturation par pays
- Top 10 des clients
- Top 10 des pays
- facturation par secteur d'activité
- facturation par catégorie
- facturation par année
- facturation par mois
- facturation par trimestre
- facturation moyenne par secteur d'activité par mois
- facturation moyenne par secteur d'activité et par trimestre
- facturation moyenne par pays et par trimestre
- facturation moyenne par pays par secteur d'activité par trimestre

Voici cinq lignes choisies au hasard dans le fichier csv.

Exercice 2 : Concevoir les tables de faits

Le fait dans ces données est la facture qui est générée mensuellement.

Les champs `customerid`et `billedamount`sont les champs importants dans la table de faits.

Nous avons également besoin d'un moyen d'identifier les informations client supplémentaires, autres que l'identifiant et les informations de date. Nous avons donc besoin de champs qui font référence aux informations client et de date dans d'autres tables.

Le tableau des faits final du projet de loi ressemblerait à ceci :

Nom du champ	**Détails**
facture	Clé primaire - Identifiant unique pour chaque facture
identifiant client	Clé étrangère - ID du client
moisid	Clé étrangère - ID du mois. Nous pouvons résoudre les informations du mois facturé à l'aide de cette clé
montant facturé	Montant facturé par les services cloud fournis pour ce mois en USD

Exercice 3 : Concevoir les tables de dimensions

Notre fait (facture mensuelle) comporte deux dimensions.

1. Informations client
2. Informations sur les dates

Organisons tous les champs qui donnent des informations sur le client dans une table de dimension.

Nom du champ	**Détails**
identifiant client	Clé primaire - ID du client
catégorie	Catégorie de client. Exemple : Particulier ou Entreprise
pays	Pays du client
industrie	À quel domaine/secteur appartient le client. Exemple : juridique, ingénierie

Organisons ou dérivons tous les champs qui donnent des informations sur la date de la facture.

Nom du champ	**Détails**
moisid	Clé primaire - ID du mois
année	Année dérivée du champ mois des données d'origine. Exemple : 2010
mois	Numéro de mois dérivé du champ mois des données d'origine. Exemple : 1, 2, 3
nom du mois	Nom du mois dérivé du champ mois des données d'origine. Exemple : mars
quart	Numéro de trimestre dérivé du champ mois des données d'origine. Exemple : 1, 2, 3, 4
nom de quartier	Nom du trimestre dérivé du champ mois des données d'origine. Exemple : T1, T2, T3, T4

Exercice 4 : Créer un schéma en étoile à l'aide des tables de faits et de dimensions

Sur la base des deux exercices précédents, nous sommes maintenant arrivés à 3 tableaux, nous pouvons les nommer comme dans le tableau ci-dessous.

Nom de la table	**Taper**	**Détails**
Facturation	Fait	Ce tableau contient le montant de la facturation et les clés étrangères des données client et mensuelles
Client Dim	Dimension	Ce tableau contient toutes les informations relatives au client
Mois sombre	Dimension	Ce tableau contient toutes les informations relatives au mois de facturation

Lorsque nous organisons les tables ci-dessus dans le style de schéma en étoile, nous obtenons une structure de table qui ressemble à celle de l'image ci-dessous.

L'image montre les tables de faits et de dimensions ainsi que les relations entre elles.

Exercice 5 : Créer le schéma sur l'entrepôt de données

Étape 1 : Démarrez le serveur postgresql.

Démarrez le serveur PostgreSQL en cliquant sur la commande ci-dessous :

Ouvrir et démarrer PostgreSQL dans l'IDE

Étape 2 : Créez la base de données sur l'entrepôt de données.

En utilisant la commande createdb du serveur PostgreSQL, nous pouvons créer directement la base de données depuis le terminal.

Tout d'abord, exécutez la commande ci-dessous pour définir votre mot de passe PostgreSQL pour l'authentification. Remplacez <votre_mot_de_passe> par votre mot de passe PostgreSQL réel, puis exécutez la commande :

```
export PGPASSWORD=<your_password>
```

Maintenant, exécutez la commande ci-dessous pour créer une base de données nommée billingDW.

```
createdb -h postgres -U postgres -p 5432 billingDW
```

Dans la commande ci-dessus

- -h indique que le serveur de base de données est accessible en utilisant le nom d'hôte « postgres »
- -U mentionne que nous utilisons le nom d'utilisateur postgres pour nous connecter à la base de données
- -p indique que le serveur de base de données fonctionne sur le port numéro 5432

Vous devriez voir un résultat comme celui-ci.

Étape 3 : Téléchargez le fichier de schéma .sql.

Les commandes pour créer le schéma sont disponibles dans le fichier ci-dessous.

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0260EN-SkillsNetwork/labs/Working%20with%20Facts%20and%20Dimension%20Tables/star-schema.sql

Téléchargez le fichier en exécutant la commande ci-dessous.

```
wget https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0260EN-SkillsNetwork/labs/Working%20with%20Facts%20and
```

Étape 4 : Créer le schéma

Exécutez la commande ci-dessous pour créer le schéma dans la `billingDW`base de données ci-dessous.

```
psql  -h postgres -U postgres -p 5432 billingDW < star-schema.sql
```

Vous devriez voir un résultat similaire à celui ci-dessous.

Exercices pratiques

Dans cet exercice pratique, vous analyserez le fichier csv ci-dessous, qui contient des données sur les ventes quotidiennes dans différents magasins d'un détaillant de mode international.

1. Problème:

   Concevez le schéma de la table de dimension DimStore.

▼ Cliquez ici pour un indice

Assurez-vous que ce tableau contient le pays et la ville du magasin.

▼ Cliquez ici pour la solution

Nom du champ	**Détails**
identifiant de magasin	Clé primaire - Identifiant unique pour chaque magasin
ville	Ville où se trouve le magasin.
pays	Pays où se trouve le magasin.

2. Problème:

   Concevez le schéma de la table de dimension DimDate.

▼ Cliquez ici pour un indice

Ici, le client a besoin de rapports avec une précision journalière. Assurez-vous d'inclure également le jour, le jour de la semaine et le nom du jour de la semaine dans ce tableau.

▼ Cliquez ici pour la solution

Nom du champ	**Détails**
identifiant de la date	Clé primaire - ID de la date
jour	Jour dérivé du champ de date des données d'origine. Exemple : 13, 19
jour de la semaine	Jour de la semaine dérivé du champ de date des données d'origine. Exemple : 1, 2, 3, 4, 5, 6, 7. 1 pour dimanche, 7 pour samedi

Nom du champ	Détails
nom du jour de la semaine	Nom du jour de la semaine dérivé du champ de date des données d'origine. Exemple : dimanche, lundi
année	Année dérivée du champ de date des données d'origine. Exemple : 2010
mois	Numéro du mois dérivé du champ de date des données d'origine. Exemple : 1, 2, 3
nom du mois	Nom du mois dérivé du champ de date des données d'origine. Exemple : mars
quart	Numéro de trimestre dérivé du champ de date des données d'origine. Exemple : 1, 2, 3, 4
nom de quartier	Nom du trimestre dérivé du champ de date des données d'origine. Exemple : T1, T2, T3, T4

3. Problème:

   Concevez le schéma de la table de faits FactSales.

▼ Cliquez ici pour un indice

Assurez-vous que le champ totalsales est capturé et qu'il existe un moyen de faire référence au magasin et à la date. Ajoutez également un rowid pour identifier de manière unique chaque ligne.

▼ Cliquez ici pour la solution

Nom du champ	Détails
Rowid	Clé primaire - Identifiant unique pour chaque ligne
identifiant de magasin	Clé étrangère - ID du magasin
identifiant de la date	Clé étrangère - ID de la date
ventes totales	Ventes totales

Félicitations ! Vous avez terminé avec succès ce laboratoire.

Auteurs

Ramesh Sannareddy

Autres contributeurs

Rav Ahuja