

Quand utiliser la fusion et quand utiliser la jointure

Bienvenue dans cette vidéo consacrée à la comparaison entre les fusions de données et les jointures entre bases de données.

Vous pouvez télécharger le classeur d'exercices pour suivre les étapes dans votre propre logiciel Tableau.

Pour résumer, les jointures entre bases de données permettent de créer une source de données unique en intégrant deux sources de données au niveau des lignes. La fusion permet d'assembler les résultats des différentes requêtes de chaque source de données dans la vue. Pour plus d'informations concernant la jointure entre bases de données et la fusion de données, veuillez consulter les vidéos correspondantes.

Relations multiples

Pour commencer, nous allons configurer la fusion des sources de données Produits 2016 et Ventes 2016, utilisées lors de la vidéo sur les jointures entre bases de données. Analysons rapidement nos catégories de ventes par marché.

- Depuis la source de données **Produit**, faites glisser Catégorie vers Lignes.
- Passez ensuite à la source de données **Ventes**.
 - Cliquez sur l'icône Lier à côté d'ID produit pour créer la relation de fusion,
 - puis faites glisser Ventes vers Colonnes.
- Tout a l'air correct. La fusion nous donne un aperçu de nos ventes globales par catégorie.
- Mais que se passe-t-il si nous déplaçons Marché vers Couleur ? La majeure partie des barres est maintenant grise, ce qui renvoie à un astérisque. Pourquoi ?

Nous avons établi notre fusion à partir du champ ID produit, qui présente une relation multiple avec Marché. Les barres turquoise clair correspondent aux ID produit vendus uniquement sur le marché États-Unis/Canada, mais tout ce qui apparaît en gris est vendu sur plusieurs marchés, d'où la présence d'un astérisque. Le principe de la fusion, qui consiste à agréger séparément les résultats de chaque source de données et à les afficher ensemble dans la vue, ne nous permet pas de visualiser facilement les données à cause de cette relation multiple.

Dans un tel cas de figure, la jointure entre bases de données est préférable à la fusion. Cette source de données a été jointe au niveau des lignes en fonction de l'ID produit. Reproduisons la première vue.

- Catégorie vers Lignes
- Ventes vers Colonnes
- Marché vers Couleur

Nous avons à présent le comportement voulu : les catégories de ventes sont réparties par marché. Lorsque nous avons lié les données, nous avons simplement complété les informations de produit avec les informations de vente, sans modifier le nombre d'enregistrements de Ventes.

Jointures augmentant le nombre de lignes

La jointure entre bases de données est-elle toujours la meilleure solution ? Choisissons deux autres ensembles de données : Office City et Coffee Chain, que nous avons abordés dans la vidéo sur la fusion des données. Les deux enseignes appartiennent à la même société mère. Office City est présent aux États-Unis dans 50 États, et Coffee Chain dans 20 États. Supposons que nous voulions voir les ventes combinées des deux entreprises dans les 50 États.

Ici, nous avons une source de données liée sur la base d'un champ commun : « État ».

- Nous allons ajouter le champ État à partir de la connexion Office City, qui comprend les 50 États.
- Si nous effectuons un clic droit et que nous modifions le calcul « Ventes ajoutées », nous constatons que les ventes d'Office City et de Coffee Chain ont été ajoutées
 - avec une fonction « zéro-null ». La valeur nulle sera donc traitée comme un zéro pour les États où Coffee Chain n'est pas présent.
- Déplaçons Ventes ajoutées vers Colonnes.
- Waouh ! Ces entreprises doivent être vraiment en pleine expansion si elles enregistrent plusieurs millions de ventes dans certains États ! Mais c'est totalement improbable. Aussi, approfondissons notre recherche.

Ce tableau de bord affiche les ventes d'Office City et de Coffee Chain dans les quatre mêmes États. Nos ventes combinées devraient correspondre aux calculs au centre de l'écran : de simples additions de valeurs.

- Les ventes combinées fusionnées (en jaune) affichent le bon résultat, mais les valeurs de la source de données liée (en violet) sont bien trop élevées.

Si nous analysons le nombre d'enregistrements d'Office City et de Coffee Chain, ainsi que la source de données liée, nous pouvons voir que les valeurs sont démesurées.

Lorsque nous avons effectué la jointure en fonction du champ État, Tableau a créé une ligne pour chaque combinaison d'État avec une seule information : Produit et Date de Coffee Chain, et ID de ligne d'Office City. Ce produit vectoriel signifie donc que nous avons comptabilisé les valeurs des ventes plusieurs fois.

À l'inverse, nous obtenons les bons résultats à partir de la fusion : celle-ci renvoie la valeur de ventes d'une source de données pour un État et la valeur de ventes de l'autre source de données pour ce même État, nous permettant ainsi de les additionner. Il s'agit du comportement recherché dans cet exemple.

Comme vous pouvez le voir, notre analyse dépend fortement du champ à partir duquel nous effectuons la jointure ou la fusion. Il est très important de choisir le bon champ et de comprendre le principe de la jointure et de la fusion pour savoir laquelle utiliser et dans quels cas.

Avantages de chaque méthode

Il existe de nombreux avantages à effectuer une jointure entre bases de données pour obtenir une source de données intégrée.

- Puisque la jointure se fait au niveau des lignes, vous ne serez pas confronté aux limitations multiples de la fusion des données, et vous n'aurez donc plus aucun astérisque.
- À l'inverse d'une fusion des données, vous pourrez extraire, sauvegarder, publier et facilement partager avec d'autres utilisateurs la source de données unique. Comme n'importe quelle autre source de données.

Toutefois, si la granularité du champ lié est incorrecte, une jointure risque d'augmenter artificiellement la taille de votre ensemble de données et d'afficher des informations erronées.

La fusion de données envoie des requêtes distinctes à chaque source de données et les assemble au niveau de granularité souhaité dans la vue. Il peut s'agir de la configuration optimale pour l'analyse.

Conclusion

Merci d'avoir suivi cette vidéo sur la combinaison de sources de données par la fusion ou la jointure entre bases de données. Nous vous invitons à découvrir les autres vidéos de formation gratuite pour en apprendre davantage sur l'utilisation des produits Tableau.