**Initiez-vous à l’algèbre relationnelle avec le langage SQL**

**Relation** : tableau de données regroupant des objets de même nature

**Objet** : Représentation d’une ligne d’une relation

Dans algèbre relationnelle :

Relation = **table**

**Tuple** = ligne (ou aussi n-uplet, enregistrement ou vecteur)

**Attribut** = colonne

**Schéma** = ensemble des attributs d’une relation

**Domaine** = ensemble des valeurs que peuvent prendre les attributs

Une relation ne peut contenir deux tuples identiques : chaque ligne est unique identifié par une clé

**Clé**: groupe d’attribut minimum déterminant un tuple unique.

Ensemble des clés possibles d’une table = **clés candidates**

**Clé primaire** = clé choisie parmi clés candidates selon logique et intelligibilité

Si ce n’est pas possible, on crée une **clé artificielle**.

**Clé artificielle** : attribut ajouté à une relation ayant pour seule fonction d’identifier de manière unique les tuples de la relation.

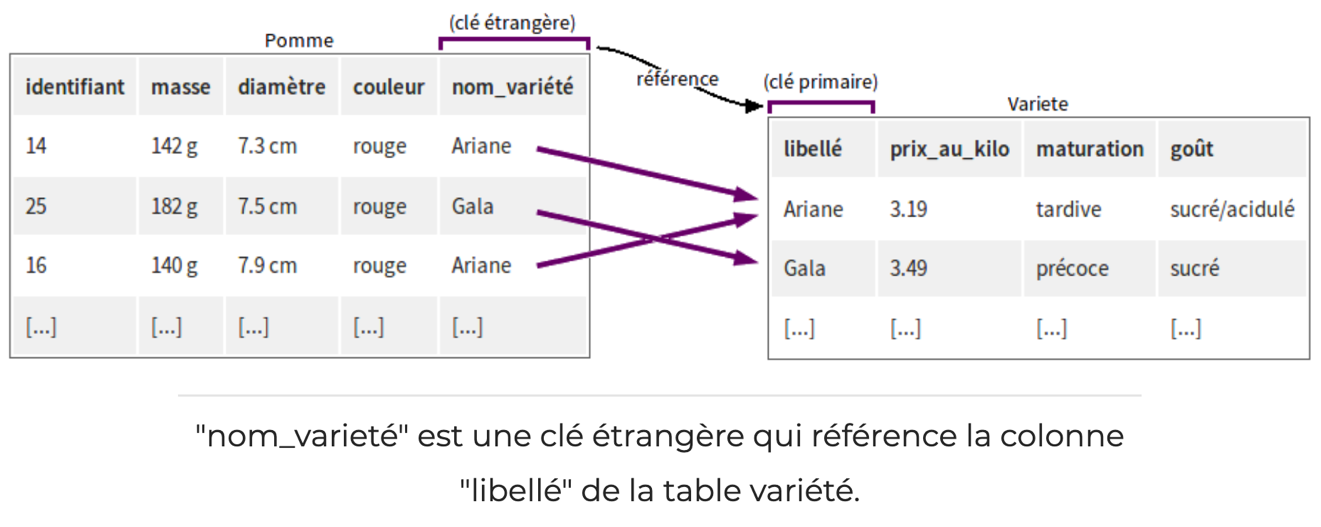
La clé primaire est très souvent une clé artificielle (identifiant).

La notation permettant d’identifier la clé primaire est [PK] après le ou les attributs choisis.

Sinon les attributs sont notés en gras.

Base de données très souvent composée de plusieurs tables.

**Clé étrangère** : attribut d’une table qui n’est pas sa clé primaire mais qui fait référence à la clé primaire d’une autre table. Peut être identifiée par [FK].



**Redondance de donnée** :

Principe à éviter et pour le faire, bien scinder la base en plusieurs relations dès que possible :

Dans une relation, si un attribut A dépend uniquement d’un groupe d’attributs G (et que ce groupe d’attributs n’est pas une clé candidate), alors il est possible de créer une nouvelle relation qui contiendra les attributs A et G.

Il faut cependant s’assurer que G soit minimal (cad que l’on ne puisse pas enlever d’attributs au groupe G sans casser la dépendance entre A et G).

G sera alors une clé candidate pour la nouvelle relation créée.

Si un attribut B dépend également uniquement de G, alors il faut aussi le déplacer dans la nouvelle relation créée.

La redondance n’est pas un problème lorsque l’on manipule les données pour parfois concaténer des données de plusieurs tables et les questionner. Mais il faut **éviter redondance dans le stockage de données.**

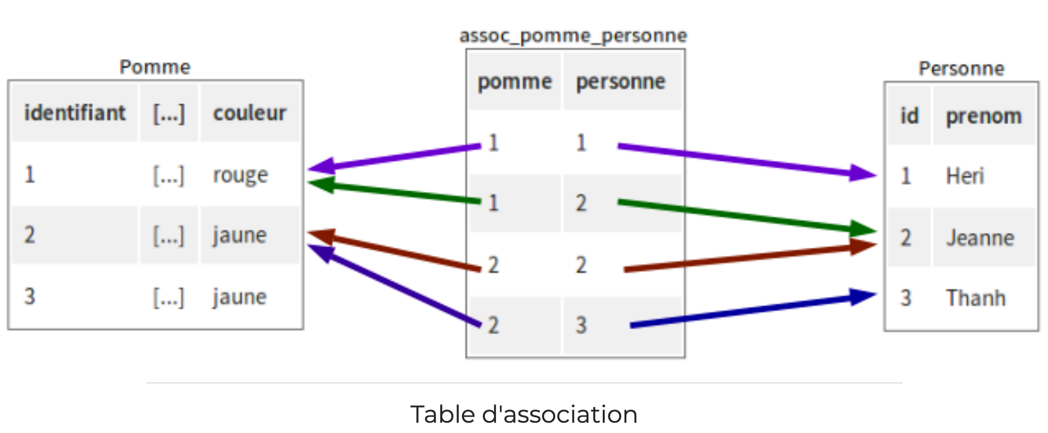
Le fait d’enlever les redondances est la **normalisation**.

**Table d’association** :

Plusieurs cardinalités du lien possibles entre deux tables :

* De 1 à plusieurs (une table *personne* mangeant une seule variété de pomme issue d’une table *variété*)
* De plusieurs à 1 (plusieurs variétés de pomme issue d’une table *variétés* mangées par une personne)
* De plusieurs à plusieurs (Problème)
* De 1 à 1 (Une seule table puisqu’une pomme par personne)

Dans le cadre d’une cardinalité plusieurs-plusieurs entre deux objets, on utilise une **table d’association**, composée d’au moins **deux clés étrangères référençant chacune l’un des 2 objets**.



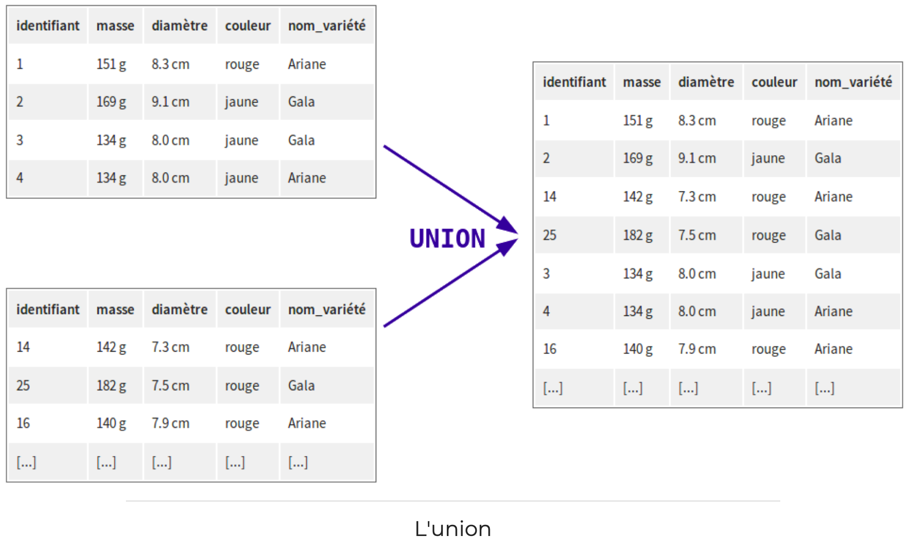
**Manipulation des relations**

Projection : opération pour afficher moins de colonnes

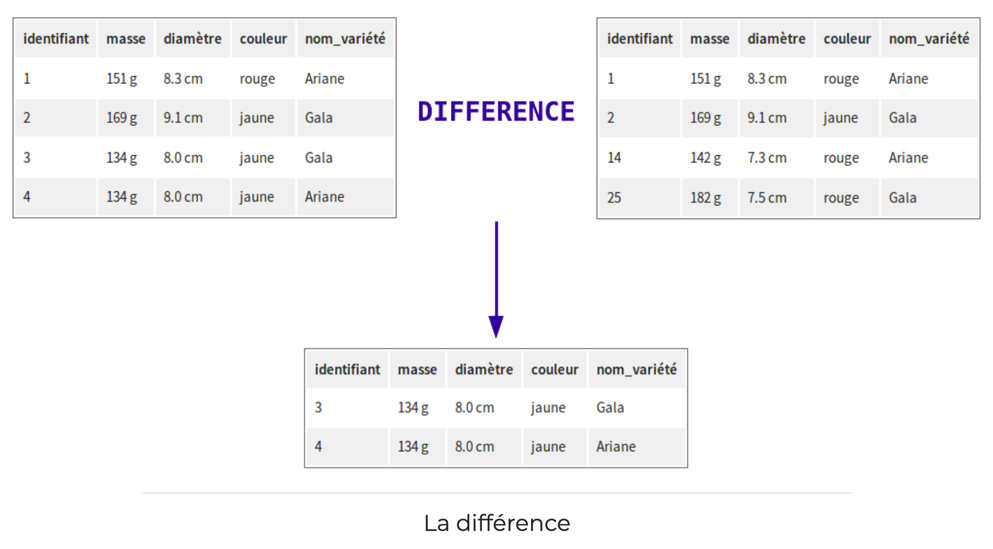
Restriction : opération pour afficher moins de lignes

Opérateurs ensemblistes : union, différence, intersection **aboutissant à + ou – de lignes**

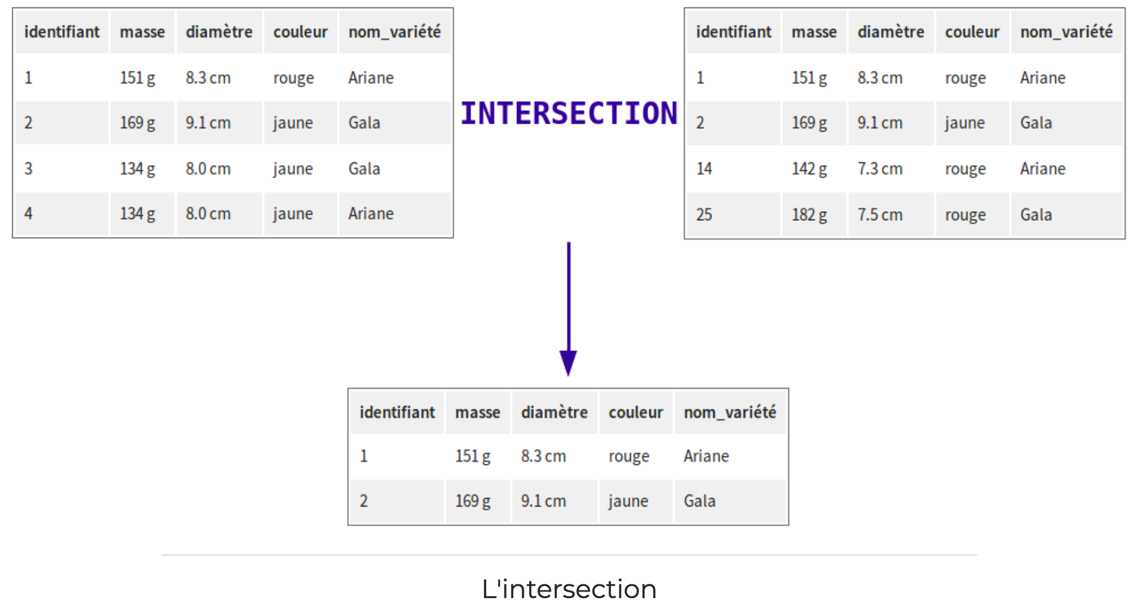
Union : réunion de deux tables avec mêmes attributs



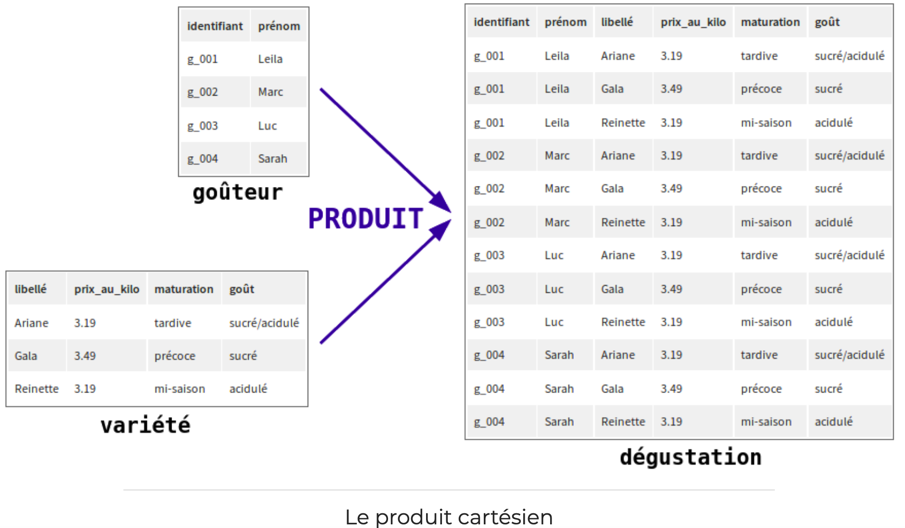
Différence : donne toutes les lignes d’une table non présentes sur la table qui lui est soustraite.



Intersection : donne toutes les lignes communes à deux tables. Équivalente à deux différences successives. R1 intersection R2 = R1 différence (R1 différence R2)



Produit cartésien :

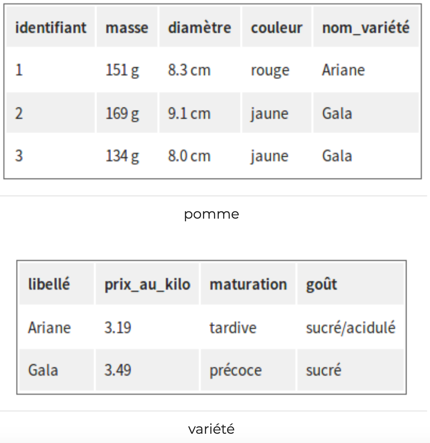


Division : très rarement utilisée



**Jointure** **interne**:

Créer une grande table à partir de deux tables selon une condition liées par une clé étrangère, **aboutissant à + ou – de colonnes**



**=**



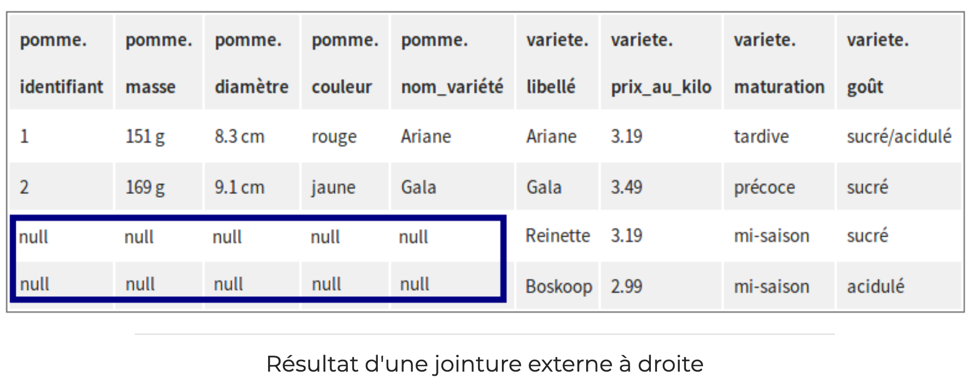
Si une clé primaire fait référence à plusieurs attributs alors la jointure peut être faite sur plusieurs colonnes.

**Jointure externe** :

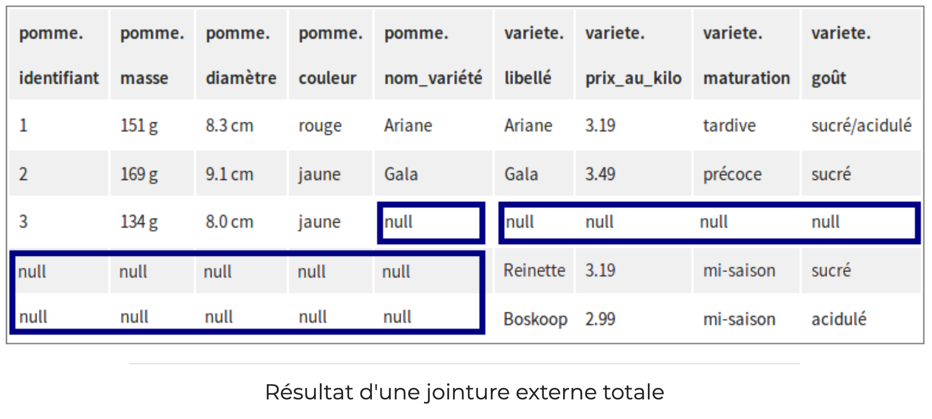
Une clé ne peut avoir de valeur nulle. Or si pour une clé étrangère, il manque une info pour une ligne, cela signifie qu’elle n’apparaîtra pas avec une jointure interne. On fait alors une jointure externe. Si la **valeur nulle est dans la table de gauche, on fait alors une jointure externe à gauche** pour conserver la ligne où l’info manque.



Si nous avons des variétés de pomme qui ne sont pas encore vendues alors nous avons plus d’infos à droite et il faut faire une **jointure externe à droite** :



Et si nous voulons garder toutes les infos et sur les pommes et sur les variétés, on fait une **jointure externe totale** :



La jointure est finalement le résultat d’un produit cartésien puis d’une restriction (on enlève toutes les lignes où la condition de la jointure n’est pas respectée).

**Jointure naturelle** :

Jointure faite grâce au même nom de l’attribut de la clé étrangère et de celui de la clé primaire de la table jointe. Ainsi, pas besoin de condition.

**Agrégats :**

Répond à une question qui donne un résultat par catégorie.

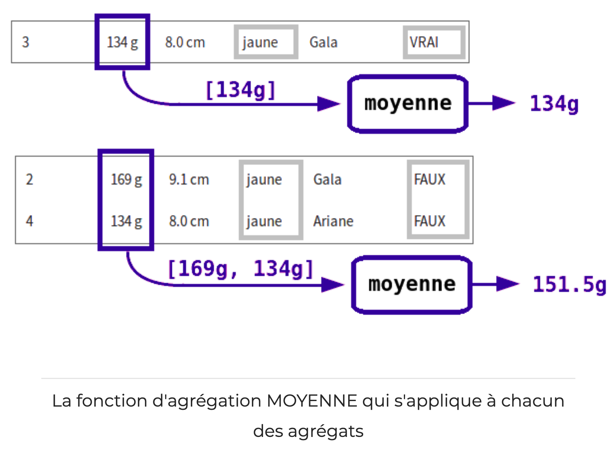
Ex : Quelle est la masse moyenne de chaque catégorie de pommes ?

Il y a donc deux composantes :

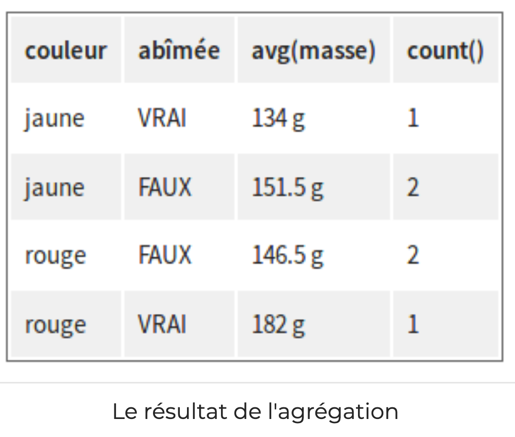
1. Un groupe d’attributs de partitionnement
2. Une (ou des) fonctions d’agrégation

Le partitionnement peut se faire par un attribut booléen.

La fonction d’agrégation prend en entrée les valeurs issues du partitionnement puis retourne une unique valeur pour chaque groupe.



Le résultat d’une agrégation donne une table avec autant de lignes que d’agrégats et ces lignes ne représentent pas les mêmes objets.



**L’importance des clés**

Il est impératif pour ces opérations de jointure de bien avoir choisi ses clés primaires et ses clés étrangères.

Une clé primaire ne doit pas avoir de doublons !

Il faut éviter toute redondance de données.

Méthode :

1. Comptez le nombre de lignes de votre table
2. Supprimez les doublons
3. Recomptez le nombre de lignes de votre table

Si 1 différent de 3, alors redondance de données supprimée à l’étape 2.

Pour supprimer doublons :

En SQL, utiliser SELECT DISTINCT

En python, méthode drop\_duplicates() de librairie Pandas