

Extrait de Mines-Ponts 2016

Pour suivre la propagation des épidémies, de nombreuses données sont recueillies par les institutions internationales comme l'OMS. Par exemple, pour le paludisme, on dispose de deux tables :

- La table **palu** recense le nombre de nouveaux cas confirmés et le nombre de décès liés au paludisme; certaines lignes de cette table sont données en exemple (on précise que **iso** est un identifiant unique pour chaque pays):

nom	iso	annee	cas	deces
Bresil	BR	2009	309 316	85
Bresil	BR	2010	334 667	76
Kenya	KE	2010	898 531	26 017
Mali	ML	2011	307 035	2 128
Ouganda	UG	2010	1 581 160	8 431
...				

- la table **demographie** recense la population totale de chaque pays; certaines lignes de cette table sont données en exemple:

pays	periode	pop
BR	2009	193 020 000
BR	2010	194 946 000
KE	2010	40 909 000
ML	2011	14 417 000
UG	2010	33 987 000
...		

- Au vu des données présentées dans la table **palu**, parmi les attributs **nom**, **iso** et **annee**, quels attributs peuvent servir de clé primaire? Un couple d'attributs pourrait-il servir de clé primaire? (on considère qu'une clé primaire peut posséder plusieurs attributs). Si oui, en préciser un.
- Écrire une requête en langage SQL qui récupère depuis la table **palu** toutes les données de l'année 2010 qui correspondent à des pays où le nombre de décès dus au paludisme est supérieur ou égal à 1 000.

On appelle *taux d'incidence d'une épidémie* le rapport du nombre de nouveaux cas pendant une période donnée sur la taille de la population-cible pendant la même période. Il s'exprime généralement en «nombre de nouveaux cas pour 100 000 personnes par année ». Il s'agit d'un des critères les plus importants pour évaluer la fréquence et la vitesse d'apparition d'une épidémie.

- Écrire une requête en langage SQL qui détermine le taux d'incidence du paludisme en 2011 pour les différents pays de la table **palu**.
- Écrire une requête en langage SQL permettant de déterminer le nom du pays ayant eu le deuxième plus grand nombre de nouveaux cas de paludisme en 2010 (on pourra supposer qu'il n'y a pas de pays *ex æquo* pour les nombres de cas).

On considère la requête R qui s'écrit dans le langage de l'algèbre relationnelle:

$$R = \pi_{\text{nom,deces}}(\sigma_{\text{annee}=2010}(\text{palu}))$$

On suppose que le résultat de cette requête a été converti en une liste Python stockée dans la variable **deces2010** et constituée de couples (chaîne,entier).

- Quelle instruction peut-on écrire en Python pour trier la liste **deces2010** par ordre croissant du nombre de décès dus au paludisme en 2010?

Extrait de Centrale-Supélec 2018

On dispose d'une version plus générale de la fonction `simulation` pour laquelle toutes les particules ne sont plus nécessairement identiques. Cette fonction enregistre ses résultats dans une base de données dont la structure est donnée à la figure suivante :

SIMULATION		REBOND		PARTICULE	
SI_NUM	integer	SI_NUM	integer	PA_NUM	integer
SI_DEB	datetime	RE_NUM	integer	PA_NOM	varchar(100)
SI_DUR	float	PA_NUM	integer	PA_M	float
SI_DIM	integer	RE_T	float	PA_R	float
SI_N	integer	RE_DIR	integer		
SI_L	float	RE_VIT	float		
		RE_P	float		

1 Structure physique de la base de données des résultats de simulation

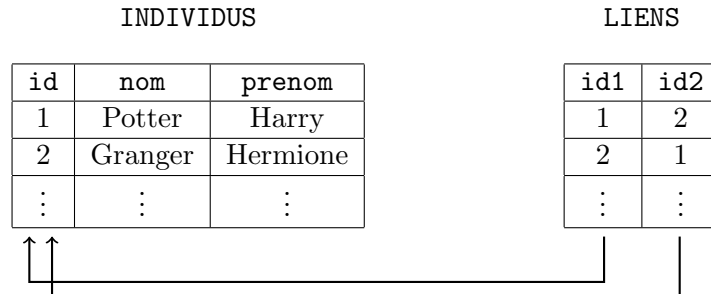
Cette base comporte les trois tables suivantes :

- la table **SIMULATION** donne les caractéristiques de chaque simulation effectuée. Elle contient les colonnes
 - SI_NUM numéro d'ordre de la simulation (clef primaire)
 - SI_DEB date et heure du lancement du programme de simulation
 - SI_DUR durée (en secondes) de la simulation (il ne s'agit pas du temps d'exécution du programme, mais du temps simulé)
 - SI_DIM nombre de dimensions de l'espace de simulation
 - SI_N nombre de particules pour cette simulation
 - SI_L (en mètres) taille du récipient utilisé pour la simulation
- la table **PARTICULE** des types de particules considérées. Elle contient les colonnes
 - PA_NUM numéro (entier) identifiant le type de particule (clef primaire)
 - PA_NOM nom de ce type de particule
 - PA_M masse de la particule (en grammes)
 - PA_R rayon (en mètres) de la particule
- la table **REBOND**, de clef primaire (SI_NUM, RE_NUM), liste les chocs des particules avec les parois du récipient. Elle contient les colonnes
 - SI_NUM numéro d'ordre de la simulation ayant généré ce rebond
 - RE_NUM numéro d'ordre du rebond au sein de cette simulation
 - PA_NUM numéro du type de particule concernée par ce rebond
 - RE_T temps de simulation (en secondes) auquel ce rebond est arrivé
 - RE_DIR paroi concernée : entier non nul de l'intervalle $[-SI_DIM, SI_DIM]$ donnant la direction de la normale à la paroi. Ainsi -2 désigne la paroi située en $y = 0$ alors que 1 désigne la paroi située en $x = L$
 - RE_VIT norme de la vitesse de la particule qui rebondit (en $m \cdot s^{-1}$)
 - RE_VP valeur absolue de la composante de la vitesse normale à la paroi (en $m \cdot s$)

1. Écrire une requête SQL qui donne le nombre de simulations effectuées pour chaque nombre de dimensions de l'espace de simulation.
2. Écrire une requête SQL qui donne, pour chaque simulation, le nombre de rebonds enregistrés et la vitesse moyenne des particules qui frappent une paroi.
3. Écrire une requête SQL qui, pour une simulation n donnée, calcule, pour chaque paroi, la variation de quantité de mouvement due aux chocs des particules sur cette paroi tout au long de la simulation. On se rappellera que lors du rebond d'une particule sur une paroi la composante de sa vitesse normale à la paroi est inversée, ce qui correspond à une variation de quantité de mouvement de $2m|v_{\perp}|$ où m désigne la masse de la particule et v_{\perp} la composante de sa vitesse normale à la paroi.

Extrait X-ÉNS 2016

Une représentation simplifiée, réduite à deux tables, de la base de données d'un réseau social est donnée dans la figure suivante :



2 Schéma de la base de données du réseau social

La table **INDIVIDUS** répertorie les individus et contient les colonnes :

- **id** (clé primaire), un entier identifiant chaque individu ;
- **nom**, une chaîne de caractères donnant le nom de famille de l'individu ;
- **prenom**, une chaîne de caractères donnant le prénom de l'individu.

La table **LIENS** répertorie les liens d'amitiés entre individus et contient les colonnes :

- **id1**, entier identifiant le premier individu du lien d'amitié ;
- **id2**, entier identifiant le second individu du lien d'amitié ;

On supposera par ailleurs que pour tout couple (x,y) dans la table **LIENS**, le couple (y,x) est également présent dans la table.

1. Écrire une requête **SQL** qui renvoie les identifiants des amis de l'individu d'identifiant x .
2. Écrire une requête **SQL** qui renvoie les noms et prénoms des amis de l'individu d'identifiant x .
3. Écrire une requête **SQL** qui renvoie les identifiants des individus qui sont amis avec au moins un ami de l'individu d'identifiant x .

X-ÉNS 2018**Introduction**

Une requête sur une base de données est décrite au moyen d'un langage **déclaratif**. Le langage SQL est le plus connu. Pour évaluer une requête, un système de gestion de base de données (SGBD) établit un plan d'exécution combinant les opérateurs de l'**algèbre relationnelle**. L'objectif de ce sujet est l'étude de ces opérateurs.

Nous étudierons en partie I l'implémentation en Python de ces opérateurs. Nous appliquerons ensuite en partie II ces résultats à des requêtes SQL. Nous verrons en partie III comment il est possible de tirer parti des propriétés des données pour améliorer les performances.

Les parties peuvent être traitées indépendamment. Néanmoins, chaque partie utilise des notations et des fonctions introduites dans les parties précédentes.

Notion de complexité

La complexité, ou le coût, d'un algorithme ou d'une fonction Python est le nombre d'opérations élémentaires nécessaires à son exécution dans le pire des cas. Lorsque cette complexité dépend d'un ensemble de paramètres (n, p, \dots) , on pourra donner cette estimation sous forme asymptotique. On rappelle qu'une application $c(n, p, \dots)$ est dans la classe $\mathcal{O}(f)$ s'il existe une constante $\alpha > 0$ telle que $|c(n, p, \dots)| < \alpha \times f(n, p, \dots)$ pour toutes les valeurs de n, p, \dots assez grandes. Nous préciserons plus loin le coût de chacune des opérations utilisées dans ce sujet.

Bases de données, tables, attributs, enregistrements

Nous détaillons ici la représentation des données dans le modèle relationnel. Une **base de données** est un ensemble de **tables**. Chaque table porte un nom et est associée à un vecteur d'**attributs** de longueur au moins 1. Le nombre d'attributs d'une table est appelé l'**arité** de la table. Le vecteur des attributs $\langle a_0, a_1, \dots, a_{k-1} \rangle$ d'une table T d'arité k est noté **attributs**(T) et la table est notée $T[a_0, a_1, \dots, a_{k-1}]$.

Une table $T[a_0, a_1, \dots, a_{k-1}]$ est constituée d'**enregistrements**. La **taille** d'une table est le nombre des enregistrements qu'elle contient. Dans ce sujet, nous considérerons qu'un enregistrement est un vecteur $\langle v_0, v_1, \dots, v_{k-1} \rangle$ de longueur l'arité k de la table. Chaque élément de ce vecteur est la valeur de cet enregistrement relativement à l'attribut correspondant de la table. La valeur v_i à l'indice i de l'enregistrement est ainsi la valeur associée à l'attribut a_i à l'indice i du vecteur d'attributs de cette table. On pourra donc identifier un attribut et son indice et parler de **la valeur d'un enregistrement associée à un indice**. La valeur d'un enregistrement e associée à un indice i est notée $e[i]$.

Nous considérons dans ce sujet que toutes les valeurs d'attributs sont des chaînes de caractères et que la comparaison entre deux valeurs d'un attribut a un coût unitaire quelles que soient ces valeurs.

Deux enregistrements représentés par des vecteurs contenant les mêmes valeurs aux mêmes indices sont égaux.

Une table peut contenir des enregistrements égaux. L'élimination des enregistrements égaux est une opération complexe qui est l'objet de l'opérateur SQL appelé DISTINCT, que nous étudierons plus loin.

Exemple (Tables et enregistrements). Considérons une agence de voyages qui vend des trajets et des chambres d'hôtel. La table

`Vehicule[IdVehicule, Type, Compagnie]`

contient les données relatives aux divers véhicules disponibles. Pour chaque enregistrement e représentant un véhicule dans la table `Vehicule`, la valeur de e associée à l'attribut `IdVehicule` est l'identifiant du véhicule ; la valeur de e associée à l'attribut `Type` est le type de véhicule ; la valeur associée à l'attribut `Compagnie` est le nom de la compagnie qui gère ce véhicule.

Cette table contient trois **enregistrements** qui décrivent des véhicules : un bus de la compagnie **IBUS**, un train de la compagnie **SNCF** et un avion de la compagnie **Hop** !.

`<98300, Bus, IBUS>`

`<1562, TGV, SNCF>`

`<30990, A320, Hop!>`

Considérons d'autre part la table

`Trajet[[IdTrajet, VilleD, VilleA, DateD, HeureD, IdVehicule]]`.

Cette table contient les trajets élémentaires possibles avec les valeurs des attributs associés : l'identifiant du trajet, la ville de départ, la ville d'arrivée, la date du départ de ce trajet, l'heure de départ du trajet, l'identifiant du véhicule utilisé pour le trajet. On rappelle que toutes ces valeurs sont des chaînes de caractères.

Cette table contient trois trajets possibles le **5 octobre 2016** pour aller de Lille à Rennes. Ils partent respectivement à **9h00**, à **10h00** et à **14h00**.

`<Trajet1, Lille, Rennes, 5 oct. 2016, 09h00, 30990>`

`<Trajet2, Lille, Rennes, 5 oct. 2016, 10h00, 98300>`

`<Trajet3, Lille, Rennes, 5 oct. 2016, 14h00, 1562>`

Représentation des tables et des enregistrements en Python

Dans ce sujet, nous représentons un enregistrement d'une table d'arité k par une **liste Python** de longueur k . L'élément d'indice i de cette liste représente la valeur de l'enregistrement pour l'attribut d'indice i de la table.

Nous représentons une table d'arité k par une liste d'enregistrements. Une table vide est représentée par une liste vide. Voici par exemple une représentation en Python de la table `Vehicule` d'arité 3:

```
>>> Vehicule
[['98300', 'Bus', 'IBUS'], ['1562', 'TGV', 'SNCF'], ['30990', 'A320', 'Hop !']]
```

Notez qu'une table peut être représentée par plusieurs listes différentes. Voici une autre représentation possible de cette table:

```
>>> Vehicule
[['1562', 'TGV', 'SNCF'], ['98300', 'Bus', 'IBUS'], ['30990', 'A320', 'Hop !']]
```

Rappel sur les listes Python

Nous rappelons brièvement les opérateurs sur les listes en Python.

Il est attendu que les candidats rédigent leurs réponses à l'aide de ces fonctions seulement. En particulier, l'opérateur d'égalité entre listes ne doit pas être utilisé.

Longueur. L'opération `len(L)` renvoie la longueur de la liste `L`. On considérera que cette opération a un coût unitaire.

Ajout. L'opération `L.append(x)` ajoute l'élément `x` à la fin de la liste `L`. On considérera que cette opération a un coût unitaire, indépendamment de la longueur de la liste et de la valeur de l'élément.

Extraction. L'opération `L.pop()` enlève le dernier élément de la liste `L` et renvoie cet élément. Une erreur est signalée si la liste est vide. On considérera que cette opération a un coût unitaire, indépendamment de la longueur de la liste et de la valeur de l'élément.

Accès. L'opération `L[i]` renvoie l'élément d'indice i de la liste `L` de longueur n . Cette opération ne peut être utilisée dans ce cadre qu'avec un indice compris entre 0 et $n - 1$. On considérera que cette opération a un coût unitaire, indépendamment de la longueur de la liste et de la valeur de l'indice.

Concaténation. L'opération `L1 + L2` renvoie la concaténation des deux listes `L1` et `L2`. Les listes ne sont pas modifiées. On considérera que cette opération a un coût unitaire.

Itération. Il est possible de parcourir une liste `L` par la commande d'itération

```
for x in L: ...
```

Ce parcours respecte l'ordre des éléments apparaissant dans la liste. On considérera que le coût d'un parcours est la somme des coûts des opérations effectuées, sans surcoût additionnel.

I. Implémentation des opérateurs de l'algèbre relationnelle en Python

Dans toute la suite, on supposera que les arguments des fonctions Python à rédiger sont bien formés : toutes les listes représentant les enregistrements d'une table ont la même longueur, qui est l'arité de cette table, les entiers représentant des indices d'attributs appartiennent bien à l'intervalle attendu, etc.

Sélection avec test d'égalité à une constante

L'opérateur $\sigma_{Constante}$ prend en argument une table T d'enregistrements, un attribut de cette table identifié à son indice i dans le vecteur `attributs(T)` et une valeur c . Il renvoie une table T' associée aux mêmes attributs que T . Elle est constituée des enregistrements de T tels que la valeur de l'attribut d'indice i est égale à la valeur c . Cette table peut être vide.

Exemple. $\sigma_{Constante}(\text{Trajet}, 1, \text{Lille})$ renvoie une table T' avec les mêmes attributs que la table `Trajet`. Elle contient tous les voyages dont la ville de départ est Lille. Dans notre exemple, c'est le cas de tous les voyages. La table T' contient donc les mêmes enregistrements que la table `Trajet`.

Q.I.1. Implémentez la fonction

```
SelectionConstante(table, indice, constante)
```

qui prend en arguments une table `table` représentée par une liste d'enregistrements, un entier `indice` associé à un attribut de cette table `table` et une valeur `constante`. Elle renvoie une nouvelle liste représentant la table $\sigma_{\text{Constante}}(\text{table}, \text{indice}, \text{constante})$.

Q.I.2. Donnez la complexité de votre implémentation de la fonction `SelectionConstante` par rapport à la taille de la table `table`. Justifiez votre réponse en vous appuyant sur la structure du programme.

Sélection avec test d'égalité entre deux attributs

L'opérateur σ_{Egalite} prend en argument une table T d'enregistrements et deux attributs de T identifiés par leurs indices respectifs i et j dans `attributs(T)`. Notez qu'il est possible que $i = j$. Il renvoie une table T' associée aux mêmes attributs que T . Elle est constituée des enregistrements de T tels que la valeur pour l'attribut d'indice i est égale à la valeur pour l'attribut d'indice j . Cette table peut être vide.

Exemple. $\sigma_{\text{Egalite}}(\text{Trajet}, 1, 2)$ renvoie une table avec les mêmes attributs que la table `Trajet`. Elle contient tous les voyages dont la ville de départ est la même que la ville d'arrivée. Le résultat est une table vide.

Q.I.3. Implémentez la fonction

`SelectionEgalite(table, indice1, indice2)`

qui prend en argument une table `table` d'enregistrements et deux attributs identifiés par leurs indices `indice1` et `indice2`. Elle renvoie la table $\sigma_{\text{Egalite}}(\text{table}, \text{indice1}, \text{indice2})$ sous la forme d'une nouvelle liste.

Projection sur des indices

L'opérateur Π prend en argument une table d'enregistrements $T \llbracket a_1, \dots, a_{k-1} \rrbracket$ d'arité k et un vecteur $L = \langle l_0, \dots, l_{k'-1} \rangle$ tel que $0 < k' \leq k$ d'indices identifiant des attributs $\langle a_{l_0}, \dots, a_{l_{k'-1}} \rangle$ de la table T .

On se restreint au cas où la liste L est ordonnée dans le sens croissant, sans répétition. On suppose que les valeurs du vecteur L sont bien comprises entre 0 et $k - 1$.

L'opérateur Π renvoie la table T' d'arité k' associée au vecteur d'attributs $\langle a_{l_0}, \dots, a_{l_{k'-1}} \rangle$. Les enregistrements de T' sont obtenus à partir des enregistrements de T en conservant uniquement les valeurs de ces enregistrements pour les attributs de T' . Deux enregistrements distincts et différents de T peuvent ainsi créer deux enregistrements égaux dans T' .

Exemple. $\Pi(\text{Trajet}, \langle 1, 2 \rangle)$ renvoie une table associée aux attributs $\langle \text{VilleD}, \text{VilleA} \rangle$. Ici, on obtient trois enregistrement égaux à $\langle \text{Lille}, \text{Rennes} \rangle$. La table n'en est pas moins constituée de trois enregistrements; sa taille est 3.

Q.I.4. Considérons une table d'enregistrements $T \llbracket a_1, \dots, a_{k-1} \rrbracket$ d'arité k . Implémentez la fonction

`ProjectionEnregistrement(enregistrement, listeIndices)`

qui prend en argument un enregistrement de cette table et une liste d'indices $\langle l_0, \dots, l_{k'-1} \rangle$ telle que $0 < k' \leq k$ identifiant des attributs de cette table. Elle renvoie une nouvelle table représentant l'enregistrement d'attributs $\langle a_{l_0}, \dots, a_{l_{k'-1}} \rangle$.

On se restreint au cas où la liste `listeIndices` est ordonnée dans le sens croissant, sans répétition. On supposera également que tous les indices l_i de cette liste sont compris entre 0 et $k - 1$.

Q.I.5. Implémentez la fonction

`Projection(table, listeIndices)`

qui prend en arguments une table d'enregistrements d'arité k et une liste l'indices $\langle l_0, \dots, l_{k'-1} \rangle$ telle que $0 < k' \leq k$ d'indices identifiant des attributs de cette table. Cette fonction renvoie une nouvelle liste représentant $\Pi(\text{table}, \text{listeIndices})$.

Produit cartésien

L'opérateur \times prend en argument deux tables T_1 et T_2 d'enregistrements. La table T_1 , d'arité k_1 , est constituée de n_1 enregistrements. La table T_2 , d'arité k_2 , est constituée de n_2 enregistrements. La table T' résultante est d'arité $k_1 + k_2$. Son vecteur d'attributs `attribut(T')` est la concaténation des vecteurs d'attributs de T_1 et de T_2 .

La table T' est constituée de $n_1 \times n_2$ enregistrements. Ces enregistrements sont créés par la concaténation de chaque enregistrement de T_1 avec chaque enregistrement de T_2 . Les n_1 premiers attributs sont ceux de T_1 dans l'ordre de T_1 , les n_2 suivants sont ceux de T_2 , dans l'ordre de T_2 . L'ordre des enregistrements ainsi synthétisés dans T' est arbitraire.

Exemple. $\times(\text{Vehicule}, \text{Trajet})$ renvoie une table T' . Les enregistrements de T' sont formés par la concaténation deux à deux des enregistrements de la table `Vehicule` et de ceux de la table `Trajet`.

```
<98300, Bus, IBUS, Trajet1, Lille, Rennes, 5 oct. 2016, 09h00, 30990>
<98300, Bus, IBUS, Trajet2, Lille, Rennes, 5 oct. 2016, 10h00, 98300>
<98300, Bus, IBUS, Trajet3, Lille, Rennes, 5 oct. 2016, 14h00, 1562>
<1562, TGV, SNCF, Trajet1, Lille, Rennes, 5 oct. 2016, 09h00, 30990>
<1562, TGV, SNCF, Trajet2, Lille, Rennes, 5 oct. 2016, 10h00, 98300>
<1562, TGV, SNCF, Trajet3, Lille, Rennes, 5 oct. 2016, 14h00, 1562>
<30990, A320, Hop!, Trajet1, Lille, Rennes, 5 oct. 2016, 09h00, 30990>
<30990, A320, Hop!, Trajet2, Lille, Rennes, 5 oct. 2016, 10h00, 98300>
<30990, A320, Hop!, Trajet3, Lille, Rennes, 5 oct. 2016, 14h00, 1562>
```

Q.I.6. Implémentez la fonction

`ProduitCartesien(table1, table2)`

qui prend en argument deux tables `table1` et `table2` d'enregistrements. Elle renvoie une nouvelle liste représentant $\times(\text{table1}, \text{table2})$.

Jointure

L'opérateur \bowtie prend en argument deux tables, T_1 , d'arité k_1 et de taille n_1 , et T_2 , d'arité k_2 et de taille n_2 . Il prend aussi en argument un attribut de T_1 , identifié par son indice i_1 tel que $0 \leq i_1 < k_1$ dans le vecteur `attributs(T_1)`, noté A_1 , et un attribut de T_2 , identifié par son indice i_2 tel que $0 \leq i_2 < k_2$ dans le vecteur `attributs(T_2)`, noté A_2 . Posons $A_2 = \langle a_0, \dots, a_{i_2}, \dots, a_{k_2-1} \rangle$.

La table T' résultante est d'arité $k_1 + k_2 - 1$. Son vecteur d'attributs `attribut(T')` est la concaténation du vecteur A_1 et du vecteur $A'_2 = \langle a_0, \dots, a_{i_2-1}, a_{i_2+1}, \dots, a_{k_2-1} \rangle$, obtenu en effaçant la coordonnée i_2 de A_2 .

La table T' est constitué d'au plus $n_1 \times n_2$ enregistrements. Les enregistrements de T' sont créés par concaténation des enregistrements e_1 de T_1 et e_2 de T_2 tels que $e_1[i_1] = e_2[i_2]$, en supprimant la valeur d'indice $k_1 + i_2$ pour éviter la répétition avec celle d'indice i_1 . L'enregistrement résultant de cette opération est appelé *jointure* des deux enregistrements e_1 et e_2 . Notez qu'il est possible que plusieurs couples (e_1, e_2) produisent des jointures égales dans T' .

Exemple. \bowtie (`Vehicule`, `Trajet`, 0, 5) renvoie les enregistrements qui décrivent les voyages de chaque véhicule suivi des informations le concernant:

```
<98300, Bus, IBUS, Trajet2, Lille, Rennes, 5 oct. 2016, 10h00>
<1562, TGV, SNCF, Trajet3, Lille, Rennes, 5 oct. 2016, 14h00>
<30990, A320, Hop !, Trajet1, Lille, Rennes, 5 oct. 2016, 09h00>
```

Q.I.7. Implémentez la fonction

```
Jointure(table1, table2, indice1, indice2)
```

qui prend en argument deux tables `table1` et `table2` et deux entiers représentant respectivement la position d'un attribut de `table1` et celle d'un attribut de `table2`. Elle renvoie une nouvelle liste représentant la table \bowtie (`table1`, `table2`, `indice1`, `indice2`).

On pourra commencer par implémenter une fonction qui prend en arguments deux enregistrements e_1 et e_2 et deux indices i_1 et i_2 tels que $e_1[i_1] = e_2[i_2]$ et qui renvoie leur jointure au sens ci-dessus.

Q.I.8. Donnez la complexité de `Jointure(table1, table2, indice1, indice2)` telle que vous l'avez implémentée. Justifiez votre réponse en vous appuyant sur la structure du programme.

Distinct

Nous ajoutons aux opérateurs précédents un nouvel opérateur `Distinct` qui n'appartient pas à l'algèbre relationnelle classique. Cet opérateur permet de supprimer les répétitions d'enregistrements égaux dans une table T . Il renvoie une table T' associée aux mêmes attributs que T . Cette table contient exactement un représentant pour chaque classe d'enregistrements égaux de T .

Exemple. `Distinct(Π (Trajet, $\langle 1, 2 \rangle$))` renvoie une table avec un unique représentant de chaque couple possible de villes de départ et d'arrivée. Cette table ne contient qu'un seul enregistrement : `<Lille, Rennes>`.

Q.I.9. Implémentez la fonction

SupprimerDoublons(table)

qui prend en argument une table `table` et renvoie une nouvelle liste représentant la table `Distinct(table)`.

On rappelle que l'opérateur Python d'égalité entre listes ne doit pas être utilisé dans ce sujet. Il est seulement possible de tester l'égalité de deux valeurs associées à un même attribut.

Q.I.10. Donnez la complexité de votre implémentation de `SupprimerDoublons(table)`. Justifiez votre réponse en vous appuyant sur la structure du programme.

II. Implémentation de requêtes SQL en Python

Les données de notre agence de voyage sont enregistrées par les tables suivantes:

- **Vehicule(IdVehicule, Type, Compagnie)** enregistre les véhicules disponibles par l'identifiant du véhicule, son type et sa compagnie.
- **Trajet(IdTrajet, VilleD, VilleA, IdVehicule)** enregistre les trajets élémentaires possibles par l'identifiant du trajet, la ville de départ, la ville d'arrivée ainsi que le véhicule utilisé.
- **Ticket(IdTicket, IdTrajet, Place, Date, Heure, Prix)** enregistre les tickets disponibles par l'identifiant du ticket, celui du trajet auquel ce ticket donne accès, le numéro de la place, la date, l'horaire, le prix.
- **Hotel(IdHotel, Classe, Ville)** enregistre les hôtels connus par l'identifiant de l'hôtel, sa classe, sa ville.
- **Chambre(IdReservation, IdHotel, Date, Prix)** enregistre les chambres d'hôtel qui sont disponibles par l'identifiant de réservation à utiliser, l'identifiant de l'hôtel où se trouve la chambre, la date et le prix.

L'objectif de cette partie est d'étudier l'implémentation de requêtes SQL en combinant les fonctions de l'algèbre relationnelle présentées dans la partie I. Tout commentaire expliquant et justifiant la traduction sera apprécié.

Par convention, la liste Python représentant une table aura le même nom que cette table. On représentera un attribut avec sa position. Par exemple, l'attribut `IdTrajet` de la table `Trajet` est représenté par l'entier 0.

Dans chaque cas, le résultat de la requête sera affecté à une variable nommée `resultat`. Par exemple, la requête SQL

```
SELECT Vehicule.Compagnie FROM Vehicule
```

pourra être implémentée par `resultat = Projection(Vehicule, [2])`, en supposant que la variable Python `Vehicule` représente la table `Vehicule`. Dans des cas plus complexes, on pourra simplifier l'expression en utilisant des variables auxiliaires pour stocker la valeur de certaines sous-expressions, comme dans l'exemple suivant:

```
r1 = Vehicule
resultat = Projection(r1, [2])
```

Il est attendu que les candidats rédigent leurs réponses en combinant uniquement les fonctions de l'algèbre relationnelle présentées dans la partie I, à l'exclusion de toute autre fonction ou structure de contrôle Python.

Q.II.1. Proposez une implémentation pour la requête suivante.

```
SELECT *
  FROM Trajet
 WHERE Trajet.VilleD = Rennes
```

Q.II.2. Proposez une implémentation pour la requête suivante.

```
SELECT *
  FROM Trajet, Vehicule
```

Q.II.3. Proposez une implémentation pour la requête suivante.

```
SELECT *
  FROM Trajet, Vehicule
 WHERE Trajet.IdVehicule = Vehicule.IdVehicule
```

Q.II.4. Proposez une implémentation pour la requête suivante.

```
SELECT Classe, Ville, Date, Prix
  FROM Hotel JOIN Chambre
    ON Hotel.IdHotel = Chambre.IdHotel
```

Q.II.5. Proposez une implémentation pour la requête suivante.

```
SELECT Hotel.IdHotel
  FROM Hotel, Trajet, Ticket
 WHERE Hotel.Ville = Trajet.VilleA
    AND Trajet.IdTrajet = Ticket.IdTrajet
    AND Ticket.Prix = '50'
```

Q.II.6. Proposez une implémentation pour la requête suivante.

```
SELECT *
  FROM Chambre
 WHERE Chambre.Prix = '100'
    AND Chambre.IdHotel IN
      (SELECT Hotel.IdHotel
        FROM Hotel, Trajet, Ticket
       WHERE Hotel.Ville = Trajet.VilleA
          AND Trajet.IdTrajet = Ticket.IdTrajet
          AND Ticket.Prix = '50')
```

III. Amélioration des performances

Il est possible dans certains cas d'améliorer l'implémentation d'une requête en tenant compte de propriétés particulières de la représentation des données ou en utilisant des structures de données supplémentaires. Dans cette partie, nous allons montrer que l'on peut améliorer les performances en triant les données avant de les traiter ou en utilisant des tables associatives (dictionnaires) auxiliaires.

Tables triées par rapport à un indice

Une table d'arité k est représentée par une liste d'enregistrements, eux-mêmes représentés par des listes à k éléments. Supposons tout d'abord avoir à disposition une fonction

`TrieTableIndice(table, indice)`

qui trie par ordre croissant suivant l'ordre lexicographique les enregistrements de la liste `table` d'arité k par rapport à la valeur de l'attribut d'indice `indice` dans le vecteur des attributs de cette table. On suppose que la valeur `indice` est strictement inférieure à k .

Par exemple, la liste `Trajet` ci-dessous est triée par rapport à l'attribut d'indice 1 pour l'ordre lexicographique $<$ sur les chaînes de caractères de Python.

```
>>> Trajet
[['30990', 'A320', 'Hop!'], ['98300', 'Bus', 'IBUS'], ['1562', 'TGV', 'SNCF']]
```

Q.III.1. Implémentez la fonction `VerifieTrie(table, indice)` qui renvoie `True` si la table est triée pour l'indice `indice` et `False` sinon.

Q.III.2. Considérez la fonction de la question I.1

`SelectionConstante(table, indice, constante).`

Hypothèse. On suppose que la liste représentant la table `table` est triée selon l'indice `indice`.

Proposez une implémentation de cette fonction qui utilise cette hypothèse pour en améliorer les performances. Elle sera nommée comme suit :

`SelectionConstanteTrie(table, indice, constante).`

Q.III.3. Considérez la fonction de la question I.7

`Jointure(table1, table2, indice1, indice2).`

Hypothèse. On suppose dans cette question que les enregistrements de la table `table1` ont des valeurs deux à deux distinctes pour l'attribut d'indice `indice1` et qu'il en va de même pour les enregistrements de la table `table2` avec l'indice `indice2`. On suppose de plus que la liste représentant la table `table1` est triée selon l'indice `indice1` et que celle représentant la table `table2` est triée selon l'indice `indice2`.

Proposez une implémentation de cette fonction qui utilise cette hypothèse pour améliorer les performances. Elle sera nommée comme suit :

`JointureTrie(table1, table2, indice1, indice2).`

Q.III.4. Donnez la complexité de votre implémentation en vous appuyant sur la structure du programme. Donnez des exemples pour lesquels cette nouvelle approche est plus performante. Y a-t-il des cas où elle n'est pas plus performante ?

Utilisation d'un dictionnaire (index)

Dictionnaires Python

Un dictionnaire est une structure de données de Python qui permet d'associer à une clé c une valeur v . On parle aussi de *table d'association*. Dans notre cas, les clés sont des chaînes de caractères et les valeurs sont des listes d'entiers.

Nous donnons ici quelques opérations permettant de manipuler les dictionnaires en Python.

Il est attendu que les candidats rédigent leurs réponses en utilisant exclusivement ces opérations pour manipuler les dictionnaires

Création d'un dictionnaire. L'opération `dico = {}` crée le dictionnaire `dico` et l'initialise à vide. Cette opération a un coût unitaire.

Ajout d'une association. L'opération `dico[c] = liste` ajoute au dictionnaire une association entre la clé c et la liste `liste`. Si une association existait déjà pour la clé dans le dictionnaire, celle-ci est remplacée, donc perdue. Cette opération a un coût unitaire, indépendamment de l'état du dictionnaire et des valeurs de la clé et de la liste.

Extraction d'une valeur associée à une clé. L'opération `dico[c]` renvoie la liste associée à la clé c dans le dictionnaire `dico`. Cette opération n'est autorisée que si la clé c est effectivement associée à une valeur dans le dictionnaire `dico`. Si aucune association n'existe pour la clé, une erreur se produit. Cette opération a un coût unitaire, indépendamment de l'état du dictionnaire et de la valeur de la clé.

Test de présence. L'opération `c in dico` renvoie `True` si la clé c est associée à une valeur dans le dictionnaire `dico` et `False` sinon. Cette opération a un coût unitaire, indépendamment de l'état du dictionnaire et de la valeur de la clé.

On peut itérer sur les clés présentes dans un dictionnaire `dico` par la commande `for c in dico: ...`

Voici un exemple d'utilisation:

```
>>> dico = {}
>>> dico['aaa'] = [1]
>>> dico['bbb'] = [2]
>>> dico['ccc'] = [3]
>>> dico['aaa'].append(4)
>>> dico['ccc'] = [5]
>>> for c in dico: print(c, ' -> ', dico[c])
aaa -> [1,4]
bbb -> [2]
ccc -> [5]
>>> dico["ddd"]
KeyError: 'ddd'
```

Utilisation de dictionnaires pour indexer les bases de données

Une autre idée pour explorer les tables efficacement est d'utiliser des dictionnaires.

Considérons une table T d'arité k et de taille n représentée par une liste Python d'enregistrements $[e_0, \dots, e_{n-1}]$. Soit L cette liste. Considérons un indice i d'attribut de T tel que $0 \leq i < k$.

On peut associer à T et i un dictionnaire Python $\text{Dico}_{T,i}$ de la manière suivante. Les clés de ce dictionnaire sont les valeurs possibles v qui apparaissent pour l'attribut d'indice i dans la table T . L'image associée à une clé v est la liste des positions dans L des enregistrements e tels que $e[i] = v$.

Si l'attribut d'indice i de T ne prend la valeur v pour aucun enregistrement, cette clé n'est pas enregistrée dans le dictionnaire. L'image d'une clé est donc une liste non vide.

Exemple (Dictionnaire associé à une table). Considérons la table

`Vehicule` `[[IdVehicule, Type, Compagnie]]`

avec les enregistrements suivants:

```
<98300, Bus, IBUS>
<1562, TGV, SNCF>
<30990, A320, Hop !>
<1789, TGV, SNCF>.
```

Soit `dico` le dictionnaire `DicoVehicule,1` associé à l'attribut `Type` de position 1. Nous avons

```
>>> for c in dico: print(c, ' → ', dico[c])
Bus → [0]
A320 → [2]
TGV → [1, 3]
>>> dico['Ariane6']
KeyError: 'Ariane6'
```

Application à la selection

Q.III.5. Implémentez la fonction `CreerDictionnaire(table, indice)` qui prend en argument une table `table` et un indice `indice` d'attribut de table. Elle renvoie un dictionnaire de la table `table` pour l'attribut d'indice `indice`.

Q.III.6. Considérer la fonction de la question **I.1**

`SelectionConstante(table, indice, constante).`

Implémentez la fonction `SelectionConstanteDictionnaire(table, indice, constante, dico)` qui a la même fonctionnalité que `SelectionConstante`, mais qui prend en plus en argument un dictionnaire `dico` de la table `table` pour l'indice `indice`.

Q.III.7. Comparez la complexité de la fonction `SelectionConstanteDictionnaire` avec celle de la fonction `SelectionConstante`. Donnez des exemples pour lesquels cette nouvelle approche est plus performante. Y-a-t-il des cas où elle n'est pas plus performante ?

Application à la jointure

Q.III.8. Considérez la fonction de la question **I.7**

`Jointure(table1, table2, indice1, indice2).`

Implémentez la fonction

`JointureDictionnaire(table1, table2, indice1, indice2, dico2)`

qui a la même fonctionnalité que `Jointure`, mais qui prend en plus en argument un dictionnaire `dico2` de la table `table2` par rapport à l'indice `indice2`.

- Q.III.9.** Donnez la complexité de votre implémentation par rapport aux tailles et arités respectives des tables `table1` et `table2` ainsi que par rapport à la longueur maximale d'une liste renvoyée par le dictionnaire `dico2`, qui sera notée k_2 . Justifiez votre réponse en vous appuyant sur la structure du programme.
- Q.III.10.** L'opérateur de jointure prend en argument deux tables qui jouent des rôles analogues. Il serait donc possible d'utiliser un dictionnaire pour `table1` au lieu d'un dictionnaire pour `table2`. Comment pourrait-on choisir la table à indexer pour obtenir les meilleures performances ?