Le 09/05/2024

Examen
Bases de Données Avancées

Durée : 1h30

Exercice 01 (11.5 pts)

Considérons une base de données **BDBio** hébergeant des informations biologiques relatives aux gènes. Cette base de données stocke des données détaillées pour chaque gène. L'ensemble des gènes de la base constitue *un seul* génome.

- Chaque gène peut générer des transcripts (transcrits). La base de données conserve l'identité de chaque transcrit, ainsi que ses positions de début (Start) et de fin (End) sur le gène. Un gène possède d'autres gènes orthologues (partageant des fonctions communes) et peut être lui-même orthologue à d'autres gènes.
- Pour chaque transcrit, on répertorie les protéines qui en dérivent. Une protéine possède aussi une identité et des positions de début et de fin.

Ci-après deux exemples de données relatives à la base de données **BDBio.**

```
{
    "Id_gene": "gene1",
    "Start": 100,
    "End": 200,
    "Transcripts": [
     {
      "Id_transcrit": "transcript1",
      "TrStart": 110,
      "TrEnd": 150,
      "Proteins": [
       {
        "Id_proteine": "protein1",
        "PrStart": 120,
        "PrEnd": 140
      },
      {
       "Id_proteine": "protein2",
       "PrStart": 130,
       "PrEnd": 145
     }
   ]
  }
],

Orthologues": ["gene_other1", "gene_other2"]
```

```
{
    "Id_gene": "gene2",
    "Start": 300,
    "End": 400,
    "Transcripts": [
     {
      "Id_transcrit": "transcript2",
      "TrStart": 310,
      "TrEnd": 350,
      "Proteins": [
       {
        "Id_proteine": "protein3",
        "PrStart": 320,
        "PrEnd": 340
       }
      ]
     }
    ],
    "Orthologues": ["gene_other3", "gene_other4"]
}
```

Exercice 02 (08.5 pts)

Une association œuvrant pour la promotion des *sports* de différentes catégories a effectué un sondage ouvert sur son site Internet afin de collecter les informations concernant la pratique des sports. La base de données se présente comme suit ; les clés primaires sont soulignées, les clés étrangères sont désignées par « * »

Sport (CodS, nomS, descriptionS, catégorieS, typeS)
Citoyen (CodC, nomC, prénomC, genreC, âgeC, adresseC, CodS*, numClub*)
Club (numCl, nomCl, responsableCl, adresseCl, CodS*)
Sponsor (numSponsor, nomSp, statutSp, secteurSp)
Sponsor-Club (numSponsor *, numCl*)

Spécifications

Le type de sport peut être collectif ou individuel.
*La catégorie du sport peut être : un sport **de ballon**, un sport **aquatique**, un sport de **combat** ou un sport d'**athlétisme.***
*Un **citoyen** doit choisir un seul sport à pratiquer (ou déjà pratiqué) de manière libre ou adhérent à un club.*
Un sport peut avoir plusieurs clubs. Un club est spécialisé dans un seul sport et peut avoir plusieurs sponsors. Un sponsor peut sponsoriser plusieurs clubs même dans différents sports.

A cause du volume important de données, l'association décide de répartir ses données sur **quatre** *(04) sites* distants (Site **A**, Site **B**, Site **C** et Site **D**) afin d'alléger le stockage et les traitements.

1. Proposer une fragmentation de la BD en justifiant vos choix. Préciser à chaque fois, le type de fragmentation utilisé et donner les expressions algébriques correspondantes.
2. Dessiner le schéma d'allocation de tous les fragments.
3. On considère la requête **R** suivante : *Trouver tous les sponsors (numSponsor) qui sponsorisent les sports collectifs dans les catégories « aquatique » ou de « combat » pratiqués par les jeunes âgés de moins de 30 ans et adhérents à des clubs.*
   a. Donner le script SQL associé à la requête **R**.
   b. Dessiner l'arbre algébrique global, l'arbre canonique et l'arbre simplifié de la requête **R**.
   c. Sachant que la fragmentation est transparente à l'utilisateur et que la requête **R** est émise à partir du **Site A**, donner les différentes *stratégies* d'exécution de cette requête. En déduire la meilleure en justifiant votre réponse.
   d. Donner les formules d'évaluation du *coût* d'exécution de la requête (selon la meilleure stratégie).

Bon courage !

Partie I – NoSQL

1. Rappeler le théorème **CAP**. Quelles sont les propriétés de CAP supportées par les bases de données orientées document ?
2. Quel est le patron de conception utilisé pour générer la base de données **BDBio**?
3. On considère le SGBD MongoDB pour implémenter la BD **BDBio**

   3.a/ Ecrire la requête permettant d'obtenir la *longueur totale du génome* décrit dans la base (somme des longueurs de tous les gènes).

   3.b/ Ecrire la requête permettant d'obtenir le nombre total de protéines pour chaque gène.

Partie II – Relationnel-Objet

4. Proposer le diagramme de classes UML relatif à la base de données **BDBio**
5. On souhaite implémenter le diagramme obtenu dans un SGBD *relationnel-objet* en utilisant le langage de définition de données SQL3 :

   5.a/ Définir *tous les types* nécessaires. Prendre en compte *toutes les associations* qui existent.

   5.b/ Donner les scripts de création des tables nécessaires à cette base de données.

   5.c/ Donner la signature et le corps de la *méthode* NbProteins qui permet de récupérer le nombre de protéines produites par un transcript.
6. On suppose que la base de données est déjà remplie. Ecrire la requête SQL3 permettant de supprimer tous les transcripts n'ayant pas de protéines.
7. Ecrire en SQL3 la requête suivante : Trouver les gènes qui ont des transcripts dérivant des protéines et afficher les positions de ces protéines.

NB : En MongoDB on a les opérateurs arithmétiques : **$add, $subtract, $multiply et $divide.**

- { $add: [<expression1>, <expression2>, ...] }
- { $subtract: [<expression1>, <expression2>] }
- { $multiply: [<expression1>, <expression2>, ...] }
- { $divide: [<expression1>, <expression2>] }