Ús de Bases de Dades

PAC 3: Components d'emmagatzematge d'una base de dades.

Proposta de solució

Presentació

En aquesta Prova d'Avaluació Continuada s'exerciten els aspectes que convé tenir en compte en el disseny físic d'una base de dades. L'objectiu d'aquesta prova és comprovar el grau de comprensió del mòdul 3. Aquesta prova consta de 4 exercicis. Cal destacar que és necessari haver assimilat el contingut del mòdul 3 per a la correcta realització d'aquesta PAC.

Competències

En aquesta PAC es desenvolupen les següents competències del Grau en Multimèdia:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament, emmagatzematge i administració de dades.

Objectius

Els objectius concrets d'aquesta Prova d'Avaluació Continuada són:

- Distingir els components d'emmagatzematge i poder situar-los en una arquitectura que distribueix la funcionalitat de l'emmagatzematge de les dades gestionades pels SGBD en nivells diferents.
- Descriure els components del nivell físic d'aquesta arquitectura.
- Entendre el concepte de pàgina i conèixer-ne la funcionalitat i l'estructura.
- Comprendre la utilitat del nivell virtual.
- Conèixer la funcionalitat i l'estructura de l'espai virtual.
- Discernir els diferents tipus d'espais virtuals, entendre'n la utilitat i conèixer les característiques de cadascun.
- Formar-se una idea global de les diverses necessitats d'espai que sorgeixen en una base de dades.

Exercici 1 [25%]

Un lloc web on s'ofereixen diferents MOOC (*Massive Open Online Courses*) disposa d'una base de dades per tal de realitzar la gestió de les qualificacions dels alumnes que té inscrits en els diferents cursos. Algunes de les taules que en formen part són:

STUDENT(id_student, firstname, lastname,username, password, email)

COURSE(id_course, title, description, url)

ENROLLMENT (id_student, id_course, starts, finishes, global_mark)
{id_student} REFERENCES STUDENT(id_student)
{id_course} REFERENCES COURSE(id_course).
{starts} data d'inici del curs en que un alumne està matriculat
{finishes} data de finalització del curs en que un alumne està matriculat
{global_mark} amb valors de 'N' ,'D','C-','C+','B','A' que determinen la qualificació d'un determinat alumne en un determinat curs.

LESSON (id_course , id_lesson, title, description, uri_test, video)
{id_course} REFERENCES COURSE(id_course).
{title} conté el títol de la lliçó.
{description} conté una breu descripció de la lliçó.
{uri_test} conté la ubicació lògica, en un servidor, de l'script que permet formular un test basat en preguntes sobre la lliçó.
{vídeo} emmagatzema un vídeo que s'utilitza en la lliçó.

ASSESSMENT (id_student, id_course, id_lesson, date-time,mark)
{id_student, id_course} REFERENCES ENROLLMENT(id_student, id_course).
{id_course , id_lesson} REFERENCES LESSON(id_course, id_lesson).
{mark} recull el resultat de l'avaluació generada per la interacció de l'alumne amb el formulari del test corresponent a una lliçó.

Es considera que hi ha enregistrats de l'ordre de 100 cursos on cadascun d'ells consta de 10 lliçons i 10.000 alumnes. Cada alumne s'inscriu anualment a 3 cursos de mitjana. Es considera que de mitjana els alumnes realitzen el 80% de les lliçons de cada curs als que estan inscrits. En aquests moments es tenen enregistrades dades dels últims 5 anys.

Cal tenir en compte que cada mes es realitza el balanç de les qualificacions de les lliçons realitzades interactivament pels alumnes mitjançant la consulta corresponent. En aquesta consulta es vol conèixer el tant per cent d'alumnes que han superat cada lliçó, de la que tan sols es necessari conèixer l'identificador del curs i l'identificador de la lliçó.

Una segona consulta que es realitza de manera freqüent és la de conèixer el nom dels alumnes, la qualificació final i l'identificador de cada curs en el que s'ha inscrit.

Els vídeos emmagatzemats a la taula `LESSON` s'utilitzen en l'aplicació web per tal de realitzar el procés d'aprenentatge als alumnes inscrits en cada curs.

Nota: Per a la realització d'aquest exercici descarteu l'existència d'índexs.

Sense tenir en compte cap valoració sobre taules i/o atributs no especificats cal que identifiqueu els espais virtuals que caldria implementar per a optimitzar les consultes esmentades en aquest supòsit. Cal argumentar la resposta que realitzeu indicant les raons que determinen la vostra organització de les dades en diferents espais virtuals.

Proposta de solució:

Primerament analitzem la cardinalitat de cadascuna de les taules a partir de les dades aportades:

STUDENT: 10.000 registres
COURSE: 100 registres
LESSON: 100 cursos* 10 lliçons/curs = 1000 registres
ENROLLMENT: 10.000 alumnes * 3 inscripcions/alumne-any * 5 anys = 150.000 registres
ASSESSMENT: 10.000 alumnes * 8 test/curs (de mitjana) * 3 cursos/alumne-any * 5 anys = 1.200.000 registres

A continuació analitzem les taules implicades en cadascuna de les consultes:

```
SELECT id_course, id_lesson, count(*) AS pass_number
FROM assessment
WHERE mark >= 5 AND date-time BETWEEN <start_date> AND <finish_date>
GROUP BY id_course, id_lesson
```

```
SELECT firstname, lastname, id_course, global_mark
FROM student s, enrollment e
WHERE s.id_student = e.id_student
```

La taula `ASSESSMENT` conté un elevat nombre de files i és l'única taula involucrada en la primera consulta a optimitzar, que s'executa mensualment. Per tant, resulta convenient associar la taula a un espai virtual fragmentat *EV_assessment*, de manera que les diferents files de la taula s'emmagatzemin en diferents fragments. El criteri a emprar per decidir a

quin fragment aniran a parar les diferents files pot ser un criteri temporal: per exemple, mensual o anual

A la segona consulta s'accedeix simultàniament al contingut de les taules `STUDENT` i `ENROLLMENT`. En conseqüència, per optimitzar la consulta, interessarà que aquestes taules es troben tan a prop físicament com sigui possible, minimitzant el temps d'accés global i augmentant el rendiment de la consulta. Per a això, s'han d'associar les dues taules a un mateix espai virtual d'agrupació *EV_student_enroll.*

Per al camp `vídeo`, de la taula `LESSON`, on s'emmagatzemen els vídeos de les lliçons, utilitzarem un espai virtual d'objectes grans *EV_video*.

Ens queda doncs determinar com associar les taules `COURSE` i `LESSON` a diferents espais virtuals. Una primera consideració podria basar-se en el fet que `LESSON` està molt lligada a la taula `COURSE` (degut a que conceptualment la entitat `LESSON` és una entitat dèbil de `COURSE`) En aquest cas seria convenient emprar un espai virtual d'agrupació si consideréssim que hi ha consultes que s'efectuen molt freqüentment. No obstant, l'enunciat no esmenta res al respecte, per tant, també podem considerar un espai virtual de taula per a cadascuna d'elles.

Exercici 2 [25%]

Cerca informació i descriu d'una manera concisa en què es basa la tècnica que utilitza el SGBD PostgreSQL per emmagatzemar objectes grans anomenada TOAST. (límit de la resposta 1 pàgina).

Proposta de solució:

L'SGBD PostgreSQL emmagatzema les files corresponents a una taula en pàgines de grandària fixa (generalment 8 kB) i no permet que una fila s'estengui sobre diverses pàgines. Òbviament cal una estratègia d'emmagatzematge alternativa per a aquelles files en què la seva mida excedeixi la mida de la pàgina.

Per superar aquesta limitació, PostgreSQL implementa una tècnica anomenada TOAST (*The Oversized-Attribute Storage Technique*) aplicable únicament als tipus de dades que suporten una representació de longitud variable `varlena` (*variable length array*), als quals anomenarem *TOASTables*. Quan una taula disposa de columnes amb tipus de dades *TOASTables*, l'SGBD associa una taula TOAST identificada mitjançant un OID (*identificador d'objecte*) que actua com una extensió de la taula principal.

Quan la mida d'una fila és superior a 2 kB (o al valor fixat en el paràmetre `TOAST_TUPLE_THRESHOLD`) els valors dels camps *TOASTables* són comprimits i/o dividits en múltiples files que s'emmagatzemen *fora de línia* a la taula TOAST associada. A la taula principal, en lloc d'emmagatzemar-la dada, s'emmagatzema un punter cap a les files de la taula TOAST que alberguen els diferents fragments en què la dada s'ha dividit. Aquest procés es realitza automàticament per part de l'SGBD, de manera transparent per a l'usuari, fins a aconseguir que la longitud de la fila resulti inferior al valor prefixat en el paràmetre `TOAST_TUPLE_TARGET` (normalment 2 kB).

Les dades emmagatzemades fora de línia es divideixen en fragments. La fragmentació es realitza després de la compressió (si aquesta és aplicable). La longitud màxima de cada fragment està prefixada per l'SGBD en el paràmetre `TOAST_MAX_CHUNK_SIZE`. Aquests fragments són inserits com files en la taula TOAST associada, la qual està composta de 3 columnes: `chunk_id`, `chunk_seq` i `chunk_data` que corresponen respectivament a un identificador per al valor emmagatzemat, un nombre seqüencial per a cada fragment i el fragment real de la dada emmagatzemada .

El punter que substitueix a la dada en la taula principal inclou l'identificador OID de la taula TOAST en què està emmagatzemat el valor, i l'identificador `chunk_id` del valor específic.

Addicionalment el punter també emmagatzema la mida de la dada sense comprimir i la mida de la dada emmagatzemada.
Quan es vol accedir a una dada de la taula principal , que es troba emmagatzemat fora de línia , l'SGBD detecta automàticament que el valor s'emmagatzema a la taula TOAST , recupera els fragments a través dels identificadors (`OID` i `chunk_id`) proporcionats pel corresponent punter i els adjunta per tornar el resultat , descomprimint els fragments en cas necessari.
Hi ha 4 tipus de estratègies per emmagatzemar les dades.

PLAIN: es per a tipus de dades que no son del tipus TOAST.
EXTENDED: primer fa la compressió i després l'emmagatzemament fora de línia. S'utilitza per a la majoria del tipus TOAST.
EXTERNAL: permet fer l'emmagatzemament fora de línia però no la compressió. S'aconsegueix un accés més ràpid a les dades, però també ocupen més espai.

MAIN: permet la compressió però no l'emmagatzemament en diferents fragments.

Exercici 3 [25%]

Cerca informació i fes un resum sobre quin fitxer i quines variables de configuració intervenen en el funcionament del dietari, log file, a PostgreSQL. Es vol saber: com es gestiona la ubicació de l'arxiu log, quins esdeveniments podem enregistrar en aquest arxiu i quina informació ens permet enregistrar sobre aquests esdeveniments. (límit de la resposta 2 pàgines)

Proposta de solució:

La font principal on s'inclou la informació per tal de respondre adequadament a aquesta qüestió correspon a la secció **19.8. Error Reporting and Logging** de la documentació de PostgreSQL *:*

https://www.postgresql.org/docs/10/runtime-config-logging.html

Exercici 4 [25%]

Argumenta si les següents afirmacions són correctes o falses:

a. Si augmentem o disminuim la longitud d'una fila llavors el VAF es pot veure modificat en funció de l'augment o disminució.

CERT. Al VAF s'hi emmagatzema l'adreçament de les files dins de la pàgina, per tant la seva mida depèn del número de files que hi ha. La mida del VAF no depèn de la llargada de les files. No obstant, un canvi de longitud d'una fila ja existent dins d'una pàgina, com a conseqüència de fer servir la sentència UPDATE, pot afectar de diferents maneres la reorganització de l'espai de la pàgina. Com a conseqüència el contingut dels diferents elements del VAF sofreix modificacions, ja que el desplaçament de les files dins de la seva pàgina provoca que canviï les adreces que indiquen el començament de la fila dins la pàgina. És a dir, es modifica el contingut del vector. Hi ha una excepció, que seria que la fila que canviï de mida sigui l'última, en aquest cas el VAF no es veuria modificat.

b. El RID (*Row Identifier* o simplement *rowid*) permet la localització d'una fila dins de l'espai virtual.

CERT. Per la localització de la fila a l'espai virtual, el SGBD utilitza un element anomenat identificador de fila o RID, que té dues parts diferenciades: el número de la pàgina que conté la fila i el número d'element del VAF (vector d'adreces de fila) d'aquesta pàgina que conté l'adreça que apunta a la fila.

c. Les taules temporals no es tenen en compte en el càlcul de l'espai necessari per a una BD.

FALS. Com qualsevol taula cal tenir-la en compte. Aquestes taules també s'emmagatzemen segons l'arquitectura de components i fan servir l'espai del disc assignant-lo a un EV o més.

d. L'extensió és un concepte del nivell virtual que defineix com portar diverses pàgines físiques de dades, no consecutives, de memòria externa a memòria interna.

FALS. L'extensió és un conjunt de pàgines físiques, en principi consecutives, que el SO adquireix a petició de l'SGBD quan aquest detecta que necessita més espai per emmagatzemar dades

Recursos

Els següents recursos són d'utilitat per la realització de la PAC:

Bàsics

- Mòdul didàctic 3. Components d'emmagatzematge d'una base de dades.

Criteris de valoració

La ponderació dels exercicis és la següent:

- Exercici 1: 25%
- Exercici 2: 25%
- Exercici 3: 25%
- Exercici 4: 25%

Aquesta PAC s'ha de fer de manera estrictament individual. Qualsevol indici de còpia serà penalitzat amb un suspens (D) per a totes les parts implicades i la possible avaluació negativa de l'assignatura en la seva totalitat.

Format i data de lliurament

1. El format del fitxer ha de ser PDF.
2. El nom del fitxer ha de tenir el format següent: "nomUsuariUOC_PAC3.pdf", per exemple "jperezbr_PAC3.pdf".
3. El nom del alumne ha d'aparèixer a la portada i en cada pàgina del document.

Data límit de lliurament : 9 d'abril de 2019 a les 24:00 hores.

La data de lliurament d'aquesta PAC ha de ser estrictament respectada, i no s'acceptarà cap lliurament després de la data establerta. Si es considera per alguna raó justificada que no es va a poder complir amb aquesta data, l'estudiant s'haurà de posar en contacte amb el seu consultor de l'assignatura amb suficient anterioritat per poder buscar conjuntament una solució al respecte. Si s'acorda el lliurament amb posterioritat, la nota màxima d'aquesta PAC serà un aprovat (C+).