Conceptos básicos y modelos NoSQL
PEC1

Ejercicio 1 (30%)

A partir de la lectura de los apuntes (locuciones de los vídeos) de los temas I y II se pide responder de manera concisa (una página y media en total) a las siguientes preguntas:

1. ¿Qué significa que las bases de datos NoSQL orientadas hacia agregados favorecen los esquemas de crecimiento horizontal (o escalabilidad horizontal)?
2. Explica las ventajas del modelo de procesamiento map-reduce.
3. Explica brevemente qué es la persistencia políglota y cómo se refleja en un proyecto.
4. Explica las razones por las que un modelo relacional no es una buena opción cuando se deben procesar flujos de datos.
5. ¿Qué diferencias hay entre el modelo relacional y el modelo en grafo en cuanto a la representación de las relaciones?

Solución:

1. En general, las bases de datos NoSQL orientadas hacia agregados se ejecutan en ambientes distribuidos, conformados por clústeres de máquinas denominados nodos. De esta forma, se puede dimensionar la base de datos para mejorar su eficiencia o capacidad de almacenaje/procesamiento simplemente añadiendo más nodos a un clúster. Esta operación es sencilla, más flexible y no influye en el funcionamiento del resto de nodos. Por esta razón se dice que favorece el crecimiento horizontal también conocido como escalabilidad horizontal.
2. El modelo de procesamiento map-reduce presenta dos ventajas:
    a. Minimiza el tiempo de respuesta al ejecutarse parte de la consulta en paralelo.
    b. Minimiza el transporte de datos irrelevantes a través de la red (ya que los datos son consultados localmente en los nodos donde se almacenan, descartando los datos irrelevantes localmente).
3. La persistencia políglota consiste en el uso de diferentes tecnologías de almacenamiento para dar respuesta a las diferentes necesidades de almacenamiento que se producen en un proyecto. Así, dentro de un mismo proyecto se utilizará la tecnología, o conjunto de tecnologías, que mejor se

adapte a cada necesidad permitiendo la comunicación entre ellas y su acceso por una aplicación externa.

4. Los sistemas gestores de bases de datos relacionales utilizan una arquitectura de almacenamiento pull. Esto significa que los datos se deben almacenar en la base de datos antes de poder ser consultados. Por ello, esta arquitectura puede no ser adecuada en escenarios donde se trabaja con flujos de datos(se reciben datos de forma continua que deben ser procesados en tiempo real antes de ser almacenados). La forma más eficiente de trabajar suele ser mediante una arquitectura push, donde los datos son transmitidos desde el origen hasta el destino sin la necesidad de ser almacenados previamente en la base de datos. Normalmente, los datos se almacenan en memoria y, con el objetivo de realizar una escritura más rápida, se vuelcan conjuntamente en la base de datos por intervalos de tiempo o en tramos de datos. En algunos casos, no es necesario almacenar todo el flujo de datos, y en su lugar se almacena la información extraída de los datos y los cálculos de valores agregados. Por último, en los sistemas que trabajan con flujos de datos, la eficiencia y el rendimiento suelen primar sobre la consistencia, por lo que puede ser conveniente no cumplir con las propiedades ACID a cambio de más velocidad y disponibilidad.
5. En el modelo en grafo, las relaciones suelen estar explícitamente almacenadas en la base de datos. Es decir, cuando se almacena un nodo se indica los punteros a sus nodos relacionados. Así, para navegar de un nodo a sus nodos relacionados, se hace de forma rápida, siguiendo los punteros de los nodos relacionados. Sin embargo, en los modelos de datos relacionales,las relaciones entre datos están representadas de forma implícita (mediante claves foráneas). Si queremos obtener los datos relacionados, debemos “calcularlos” mediante una operación de combinación (en inglés join).

Ejercicio 2 (30%)

A partir de la lectura de los apuntes (locuciones de los vídeos) de los temas I y II indica si te parecen ciertas o falsas las siguientes afirmaciones.

Para cada una de las afirmaciones indica si es cierta o falsa, justificando la respuesta mediante lo que has leído en los materiales. En cada justificación deberá indicarse la cita de los apuntes, vídeo o libros en la que se sustenta.

No serán válidas las respuestas que no se justifiquen.
Se valorará la concisión (una página y media para las 5 afirmaciones como máximo).

Afirmación 1

En un modelo de agregación, la estructuración de agregados de un mismo tipo no puede variar en ningún caso. Es decir, todos los agregados del mismo tipo deben seguir la misma estructura.

Afirmación 2

El modelo en grafo es igual de fácil de escalar que los modelos agregados.

Afirmación 3

En un proyecto donde se debe priorizar la disponibilidad frente a la consistencia de los datos, son igual de recomendables una base de datos relacional que una base de datos NoSQL.

Afirmación 4

Los modelos de agregación son una buena elección en el caso de modelos conceptuales donde predominen asociaciones jerárquicas, asociaciones 1 a 1 y relaciones de tipo parte-todo.

Afirmación 5

El modelo documental es un caso particular del modelo clave-valor, pero más flexible que el clave-valor.

Solución:

Afirmación 1

La afirmación es **falsa**. Tal como aparece en la página 7 del documento B2_T3_2_ModelosAgregacionCaracteristicas, el modelo de agregados presenta un esquema flexible de forma que una de las propiedades que se cumplen es precisamente que la estructuración de agregados de un mismo tipo puede variar

Afirmación 2

La afirmación es **falsa**. Tal como aparece en la página 17 del documento B2_T4_1_ModelosEnGrafo, los modelos en grafo no son tan fácilmente escalables como los modelos de agregación puesto que los datos están altamente relacionados. Esto implica que distribuir los datos en diferentes ordenadores debe hacerse con mucho cuidado para no “romper” relaciones entre los datos. Por este hecho, la distribución de datos en estos modelos es compleja y requiere de información del dominio para realizarse de forma correcta.

Afirmación 3

La afirmación es **falsa**. Tal como aparece en la página 16 del documento B1_T2_PersistenciaPoliglota, las bases de datos relacionales priorizan la consistencia a la disponibilidad. En cambio, las tecnologías NoSQL suelen priorizar la

disponibilidad a la consistencia, por lo que escalan mejor horizontalmente que los sistemas gestores de bases de datos relacionales. Por lo tanto, los sistemas NoSQL pueden ser más aptos en aquellos proyectos donde la escalabilidad es un factor determinante pero a costa de sacrificar, en cierto grado, la consistencia de los datos.

Afirmación 4
La afirmación es **verdadera**. Tal como aparece en la página 10 del documento B2_T3_2_ModelosAgregacionCaracteristicas, cuando no existan interrelaciones (o asociaciones) complejas (de muchos a muchos) en el dominio de aplicación a representar los modelos de agregación son adecuados. En general, éste será el caso de modelos conceptuales donde predominan asociaciones jerárquicas (asociaciones con multiplicidad 1 a *), asociaciones 1 a 1 y relaciones de tipo parte-todo.

Afirmación 5
La afirmación es **falsa**. Tal como aparece en la página 6 del documento B2_T3_3_ModelosAgregacionTipos, el modelo documental se considera un caso particular del modelo clave-valor pero a diferencia de este último, en el modelo documental los agregados tienen una estructura interna. Esta estructura interna simplifica el desarrollo de aplicaciones, pero reduce la flexibilidad del modelo clave-valor.

Ejercicio 3 (20%)

La biblioteca de la UOC ha diseñado un sistema para conocer los hábitos de las personas que solicitan préstamos de sus libros. Para ello se quiere explotar la información que se almacena cada vez que un usuario solicita el préstamo de un recurso de la biblioteca.

Actualmente la información se encuentra almacenada en una base de datos relacional de tipo Oracle. La dirección de la biblioteca ha solicitado ayuda al equipo docente de la asignatura de Arquitectura de bases de datos no tradicionales, y de acuerdo al tipo de consultas que les gustaría realizar, el equipo docente ha recomendado realizar una migración de la información a una base de datos de tipo documental.

Además, se ha acordado que los estudiantes de esta asignatura ayudarán a diseñar los tipos de documentos más eficientes para las siguientes dos consultas que desean realizar:

Consulta 1: Teniendo en cuenta la titulación en la que están matriculados los estudiantes, interesa conocer información acerca de los estudiantes que solicitaron un préstamo por cada mes y titulación. El objetivo es hacer un estudio de estacionalidad, por lo tanto, no consideraremos el año, sólo consideraremos los meses. Concretamente para cada mes y por cada titulación, se desea recuperar los datos (dni, nacionalidad, país donde vive) de los estudiantes de esa titulación, así como, por cada

estudiante, la lista de recursos que tomaron en préstamo. De cada recurso se quiere obtener el título, isbn, tipo de recurso y asignatura de la titulación para la cual está recomendado el recurso prestado (se supone que todo los préstamos que solicitan los estudiantes están relacionados con alguna asignatura de la titulación que están cursando).

Consulta 2: También interesa conocer información de los recursos prestados agrupados por las asignaturas y titulación en las que están recomendados. Concretamente para cada titulación y por cada asignatura (una asignatura queda especificada en el plan de estudios indicando su nombre y el curso en el que se imparte) que forman parte del plan de estudios de la misma, se desea recuperar los datos principales de cada recurso prestado: título, isbn, tipo de recurso y una lista con la información (dni, nacionalidad, país donde vive) de cada uno de los estudiantes que solicitaron el préstamo del recurso.

Cada consulta requiere un tipo concreto de documento. En este sentido, se pide indicar:

- Una representación gráfica del documento propuesto utilizando un diagrama de cajas anidadas como el que se explica en los apuntes sobre "Diseño de agregados" suministrado en el aula virtual. Como alternativa se puede presentar el documento en formato JSON pero se debe elegir entre una representación u otra.
- Una breve explicación de la estructura del documento y una justificación del porqué de su estructura.

Solución:

Consulta 1

De acuerdo a la descripción de la consulta, para poder realizarla eficientemente se propone un tipo de documento que tendrá como campo más externo el mes, y a continuación contendrá un campo titulación que contendrá una colección de subdocumentos por cada posible titulación. Cada subdocumento contendrá una lista con información de cada estudiante de esa titulación que solicitó un préstamo de un recurso de la biblioteca durante ese mes. La información de un estudiante será a su vez un documento que contendrá la información personal del estudiante (dni, nacionalidad y país donde vive) y una lista con información de cada recurso del que ha solicitado un préstamo. La información de cada recurso también será un documento con los campos de información: título, isbn, tipo de recurso y asignatura de la titulación para la cual está recomendado el recurso prestado.

Dicho agregado corresponde al 100% a la petición realizada y por tanto proporcionará toda la información relevante en una sola lectura de la base de datos.

En el siguiente diagrama se muestra el diseño planteado:

En formato JSON, un ejemplo de la consulta 1 quedaría de la siguiente manera:

```
{
  "Mes": "Enero",
  "Titulaciones": [
    {
      "Titulacion": "Grado en informática",
      "Estudiantes": [
        {
          "DNI": "12345345Z",
          "Nacionalidad": "Francesa",
          "Pais de residencia": "Portugal",
          "Recursos": [
            {
              "Titulo": "Lógica matemática",
              "ISBN": "978-84-269-0466-9",
              "Tipo de recurso": "Libro",
              "Asignatura donde se recomienda": "Lógica matemática"
            }]
        }]
    }]
}
```


Consulta 2

De acuerdo a la descripción de la consulta, para poder realizarla eficientemente se propone un tipo de documento que tendrá como campo más externo la titulación, y a continuación contendrá un campo asignaturas que contendrá una colección de subdocumentos por cada posible asignatura (nombre y curso de impartición). Cada subdocumento contendrá una lista con información de cada recurso asociado a la asignatura del que se solicitó un préstamo: título, isbn, tipo de recurso y una lista con información (dni, nacionalidad, país donde vive) de cada uno de los estudiantes que solicitaron el préstamo del recurso.

Dicho agregado corresponde al 100% a la petición realizada y por tanto proporcionará toda la información relevante en una sola lectura de la base de datos.

En el siguiente diagrama se muestra el diseño planteado:

En formato JSON, un ejemplo de la consulta 2 quedaría de la siguiente manera:

```
{ "Titulación": "Grado en informática",
  "Asignaturas": [
    {
      "Nombre": "Iniciación a la programación",
      "Curso": "Primero",
      "Recursos": [
        {
          "Titulo": "Lógica matemática",
          "ISBN": "978-959-13-0499-5",
          "Tipo de recurso": "Libro",
          "Estudiantes": [
            {
              "Dni": "12345345Z",
              "Nacionalidad": "Francesa",
              "Pais de residencia": "Portugal"
            }]
        }]
    }]
}
```

Ejercicio 4 (20%)

Las limitaciones que presenta el modelo relacional fue una de las razones que hicieron que surgieran las bases de datos NoSQL. Sin embargo, las limitaciones de los sistemas NoSQL también motivaron otras bases de datos denominadas NewSQL, que son sistemas que adoptan el modelo relacional para ofrecer algunas de sus ventajas junto con algunas de las mejoras que proporcionan las bases de datos NoSQL. Para saber más de las bases de datos NewSQL se propone leer los apartados 1, 2 y 3 del artículo titulado “What’s Really New with NewSQL?” y los apartados 1, 3 y 4 del artículo titulado “NewSQL Through the Looking Glass”.

Una vez leídos los artículos de referencia, se propone buscar un caso de aplicación de una base de datos NoSQL y otro caso de aplicación de una base de datos NewSQL. A continuación, contesta a las siguientes preguntas por cada caso de aplicación (1 página como máximo para cada caso):

1. Indica el enlace al caso analizado.
2. Describe el problema de persistencia de datos que se ha resuelto.
3. Justifica las razones por las que es recomendable la base de datos que se ha utilizado como solución.
4. Justifica las razones por las que no sería recomendable utilizar otro tipo de bases de datos diferentes a la utilizada en la solución.
5. Indica las referencias extra utilizadas para desarrollar el ejercicio.

No se puede reproducir de forma textual ninguna frase de las referencias utilizadas. A continuación, se muestran algunos ejemplos. **Estos ejemplos no pueden utilizarse para resolver el ejercicio (si se usan alguno de estos ejemplos la pregunta estará anulada), deben buscarse otros diferentes**:

- Uso de un modelo en grafo para analizar los papeles de Panamá.
- Uso de un modelo documental para ofrecer un servicio de contenidos personalizado y en tiempo real.
- Uso de TIDB para análisis de datos en tiempo real.

Criterios de valoración

Los apartados 1 y 2 tienen un peso del 30% cada uno, y los apartados 3 y 4 tienen un peso del 20% cada uno. Se valorará, para cada apartado, la validez de la solución y la claridad de la argumentación. Cualquier solución no justificada se considerará incompleta.

Formato y fecha de entrega

Tenéis que enviar la PEC al buzón de Entrega y registro de EC disponible en el aula (apartado Evaluación). El formato del archivo que contiene vuestra solución puede ser .pdf,. odt, .doc y .docx. Para otras opciones, por favor, contactar previamente con vuestro profesor colaborador. El nombre del fichero debe contener el código de la asignatura, vuestro apellido y vuestro nombre, así como el número de actividad (PEC1). Por ejemplo nombreapellido1_nosql_pec1.docx. La fecha límite para entregar la PEC1 es el **17 de octubre.**

Propiedad intelectual

Al presentar una práctica o PEC que haga uso de recursos ajenos, se tiene que presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL etc.). El estudiante tendrá que asegurarse que la licencia que sea no impide específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por el copyright.

Será necesario, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales, y su código fuente, si así corresponde.