Evaluación del mantenimiento de la consistencia lógica en Cassandra

Pablo Suárez-Otero, María José Suárez-Cabal, Javier Tuya

[1]Departamento de Informática, Universidad de Oviedo, Gijón, España
{suarezgpablo, cabal, tuya}@uniovi.es

Resumen. En las bases de datos NoSQL como Cassandra es común que exista duplicidad de los datos entre las tablas, a diferencia de los modelos relacionales normalizados. Esto se debe a que las tablas se diseñan en base a consultas y a la ausencia de relaciones entre ellas. Por tanto, si los datos no se modifican convenientemente se pueden producir inconsistencias en la información almacenada. A su vez, es relativamente fácil que se introduzcan defectos que ocasionen inconsistencias en Cassandra, siendo éstos difíciles de detectar utilizando técnicas convencionales de pruebas dinámicas. Con el objetivo de ayudar al desarrollador a evitar la producción de inconsistencias, proponemos un nuevo método que, usando un modelo conceptual, es capaz de establecer los procesos necesarios para asegurar la calidad de los datos desde el punto de vista de su consistencia a través de pruebas estáticas. En este trabajo evaluamos este método ante un caso de estudio en el que insertamos tuplas en entidades y relaciones del modelo conceptual y extraemos qué es necesario para mantener la consistencia en el modelo lógico. Los resultados muestran cómo la desnormalización de los datos puede aumentar la complejidad del mantenimiento de la consistencia, no solo necesitando saber dónde se debe mantener la consistencia sino también cómo hay que mantenerla.

Palabras clave: Cassandra, consistencia lógica, pruebas estáticas, evaluación

1 Introducción

Cassandra [1] es una base de datos NoSQL en la que, a diferencia las bases de datos relacionales, no existen restricciones de integridad ni relaciones entre tablas. Debido en parte a esto y a que las tablas son creadas para consultar datos específicos [2], los datos pueden estar duplicados en diferentes tablas. La consistencia de estos datos debe ser mantenida en la aplicación cliente que trabaje con Cassandra, pudiendo dar lugar a la producción de posibles inconsistencias difíciles de detectar utilizando técnicas convencionales de pruebas dinámicas. Esta producción de inconsistencias se ve reflejado en la Fig. 1 con un ejemplo en el que se tienen dos entidades (“Autor” y “Libro”) y dos tablas Cassandra en las que consultar información de libros: “Libros” para obtener información de un libro localizándolo por su identificador y “Libros_por_autor” para consultar la información de libros por su autor. En el caso de existir una función destinada a insertar libros, si solo los insertase en una de las dos tablas, como se ve reflejado en la Fig. 1, se incurriría en la producción de inconsistencias.

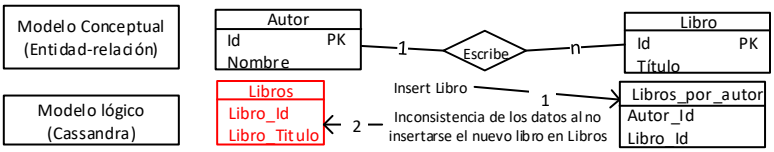

Fig. 1. Inserción en una sola tabla produciendo inconsistencia

Este trabajo es una continuación de [4], donde se explica nuestra propuesta, la cual se basa en un enfoque preventivo para evitar inconsistencias a través de pruebas estáticas. Este enfoque se divide a su vez en dos enfoques: Top-Down y Bottom-Up:

1. Top-down: Partiendo de un cambio en un dato correspondiente a una entidad o relación (modelo conceptual), se determinan qué datos en tablas y columnas Cassandra (modelo lógico) se deben modificar y cómo, a través de un análisis estático donde se consideran ambos modelos.
2. Bottom-up: Partiendo de un cambio en un dato en tablas y columnas Cassandra, se determinan qué otros datos en Cassandra se deben modificar y cómo. Para ello se determinará primero qué entidades y atributos se ven afectados por el cambio y, partiendo de ellos, se aplicaría el enfoque Top-Down.

La contribución de este artículo es la experimentación del enfoque Top-Down en un caso de estudio para comprobar la eficacia de este enfoque. Para ello ha sido necesario la generalización del enfoque Top-Down, profundizándose en los casos en los que para mantener la consistencia es necesario extraer información ya almacenada en la base de datos a través de consultas (operaciones SELECT).

El resto del artículo es como sigue: en la sección 2 se detallan los resultados de aplicar el enfoque Top-Down en un caso de estudio y se finaliza en la sección 3 con las conclusiones y las posibles líneas de trabajo futuro.

2 Experimentación de aplicación del enfoque Top-Down

El caso de estudio empleado es el descrito en [3], siendo su modelo conceptual el mostrado en la Fig. 2. El modelo lógico de este caso de estudio está formado por 9 tablas creadas para satisfacer 9 consultas. Se han evaluado 126 inserciones de tuplas en las entidades y relaciones de este modelo conceptual, variando la cantidad de información (número de atributos con valor asignado) en la tupla, con el objetivo de determinar las operaciones (INSERT, UPDATE o SELECT) necesarias para mantener la consistencia en el modelo lógico.

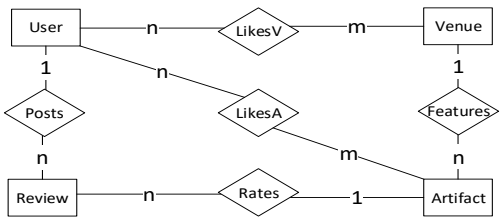

Fig. 2. Modelo conceptual utilizado para la experimentación

En el caso de las inserciones en relaciones, la tupla puede contener solo información de la relación (las claves primarias de las entidades relacionadas) o puede contener además los atributos de las entidades relacionadas. En la Tabla 1 se muestra el resumen de estas inserciones: número de inserciones de tuplas evaluadas (en entidades y relaciones), y total, media y máximo número de operaciones INSERT, UPDATE y SELECT necesarias para mantener la consistencia en la base de datos Cassandra.

Tabla 1. Resumen resultados ante inserciones en entidades y relaciones

		Operación INSERT			Operación UPDATE			Operación SELECT		
Entidad/Relación	Nº Inserciones de tuplas	Total	Media	Máximo	Total	Media	Máximo	Total	Media	Máximo
Artifact	4	12	3	3	8	2	2	18	4,5	9
Review	4	4	1	1	8	2	2	10	2,5	4
User	4	0	0	0	0	0	0	0	0	0
Venue	4	0	0	0	0	0	0	0	0	0
Features	20	60	3	3	40	2	2	90	4.5	9
LikeA	25	135	5.4	6	40	1.6	2	215	8.6	16
LikeV	25	25	1	1	0	0	0	39	1.56	3
Posts	20	20	1	1	40	2	2	50	2.5	4
Rates	20	68	3.4	4	40	2	2	110	5.5	13
Total	126	324	2.57	6	176	1.4	2	532	4.2	16

De estos datos se puede extraer que, en general, la desnormalización de los datos obliga a determinar varias operaciones para alcanzar la consistencia de los datos, alcanzándose un total de 324 operaciones INSERT, 176 operaciones UPDATE y 532 operaciones SELECT en el conjunto de las 126 inserciones evaluadas. La ausencia de operaciones al insertar tuplas de las entidades Venue y User se debe a la ausencia de tablas con solo la información de Venue o User. Esto es debido a que no se consulta la información de un Venue o un User de forma aislada, sino que siempre se consulta junto con más información. La información de Venue y User se inserta en la base de datos cuando se insertan tuplas de las relaciones LikesA o LikesV.

También se ha detectado una relación inversamente proporcional entre el número de operaciones SELECT y el número de atributos con valor asignado en la tupla. Las inserciones en entidades pueden tener hasta 3 atributos con valor asignado mientras que en el caso de las relaciones puede haber hasta 6 atributos con valor asignado (la combinación de los 3 atributos de cada entidad de la relación). Esta relación se muestra en la Fig.3 donde se puede observar como la media de operaciones SELECT de todas las inserciones decrece según aumenta el número de atributos con valor asignado.

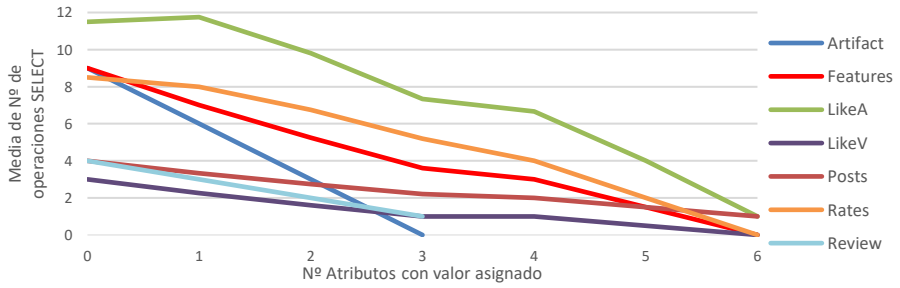

Fig. 3. Relación entre número de SELECTs y atributos con valor en la tupla

3 Conclusiones y trabajo futuro

En Cassandra, es común que los datos puedan estar desnormalizados en varias tablas, siendo necesario mantener la consistencia de estos datos en cada tabla donde estén almacenados. Esto se ve reflejado en los resultados de la experimentación realizada, donde, para mantener la consistencia en 9 tablas se llegaron a necesitar para algunas inserciones hasta 6 operaciones INSERT y 16 operaciones SELECT. Al aumentar el número de consultas sobre la misma información la complejidad para mantener la consistencia lógica irá en aumento, siendo más fácil la introducción de defectos. Para ayudar al desarrollador a la prevención de estos defectos en el mantenimiento de la consistencia lógica hemos desarrollado un método que ha sido implementado en una herramienta. Este método proporciona las operaciones necesarias para mantener la consistencia, ayudando por tanto al aumento de la calidad de los datos desde el punto de vista de la consistencia.

Como trabajo futuro se profundizará en el enfoque bottom-up y se abordará la creación de modelos conceptuales desde un modelo lógico con el objetivo de que los resultados de esta investigación puedan ser empleados en las aplicaciones diseñadas sin haber usado un modelo conceptual.

Agradecimientos

Este trabajo ha sido realizado bajo los proyectos de investigación TIN2013-46928-C3-1-R y TIN2016-76956-C3-1-R/-2-R, financiados por el Ministerio de Economía y Competitividad, y fondos FEDER. También ha sido realizado bajo el proyecto GRUPIN14-007, financiado por el Principado de Asturias y fondos FEDER.

Referencias

1. Cassandra (último acceso: Marzo 2018), http://cassandra.apache.org
2. Data Modeling Concepts (último acceso: Marzo, 2018), http://docs.datastax.com/en/cql/3.3/cql/ddl/dataModelingApproach.html.
3. Chebotko,A., Kashlev, A., Lu, S.: A Big Data Modeling Methodology for Apache Cassandra. Proc. IEEE Int. Congress on Big Data, pp. 238–245 (2015)
4. Suárez-Otero, P., Gutierrez, J., de la Riva, C., & Tuya, J. Mantenimiento de la Consistencia Lógica en Cassandra. JISBD 2017.