Resumen #4

Bases de Datos II

Luis Diego Delgado Muñoz

Prof. Nereo Campos

Fecha de Entrega: 21/09/2022

Schema-Agnostic Indexing with Azure DocumentDB

Azure DocumentDB es el servicio de base de datos distribuida de múltiples usuarios de Microsoft para administrar documentos JSON a escala de Internet. Excepcionalmente, DocumentDB proporciona consultas consistentes en tiempo real frente a tasas muy altas de actualizaciones de documentos. Como servicio multiusuario, DocumentDB está diseñado para operar dentro de presupuestos de recursos extremadamente frugales mientras proporciona un rendimiento predecible y un sólido aislamiento de recursos para sus usuarios.

Capacidades

DocumentDB se basa en el modelo de datos JSON y el lenguaje JavaScript directamente dentro de su motor de base de datos. Este enfoque habilita las siguientes capacidades de DocumentDB:

- El lenguaje de consulta admite consultas ricas relacionales y jerárquicas. Tiene sus raíces en el sistema de tipos de JavaScript, la evaluación de expresiones y el modelo de invocación de funciones.
- El motor de la base de datos está optimizado para atender consultas consistentes frente a escrituras de documentos de alto volumen sostenido. De forma predeterminada, el motor de la base de datos indexa automáticamente todos los documentos sin requerir un esquema o índices secundarios de los desarrolladores.
- Ejecución transaccional de la lógica de la aplicación proporcionada a través de procedimientos almacenados y disparadores, creada completamente en JavaScript y ejecutada directamente dentro del motor de base de datos de DocumentDB.
- Como un sistema de base de datos geodistribuido, DocumentDB ofrece niveles de coherencia bien definidos y ajustables para que los desarrolladores elijan y las garantías de rendimiento correspondientes.

El motor de la base de datos DocumentDB, a su vez, consta de componentes que incluyen la máquina de estado replicada (RSM) para la coordinación, el tiempo de ejecución del lenguaje JavaScript, el procesador de consultas y los subsistemas de almacenamiento e indexación responsables del almacenamiento transaccional y la indexación de documentos. Para brindar durabilidad y alta disponibilidad, el motor de la base de datos de DocumentDB conserva los datos en los SSD locales y los replica entre las instancias del motor de la base de datos dentro del conjunto de réplicas, respectivamente.

Indexación Agnóstica de Esquema

Con el objetivo de eliminar el desajuste de impedancia entre la base de datos y los modelos de programación de aplicaciones, DocumentDB aprovecha la simplicidad de JSON y su falta de especificación de esquema. No hace suposiciones sobre los documentos y permite que los documentos dentro de una colección de DocumentDB varíen en el esquema, además de los valores específicos de la instancia. A diferencia de otras bases de datos de documentos, el motor de base de datos de DocumentDB funciona directamente en el nivel de la gramática JSON, siendo independiente del concepto de un esquema de documento y desdibujando el límite entre la estructura y los valores de instancia de los documentos. Esto, a su vez, le permite indexar documentos automáticamente sin necesidad de esquemas o índices secundarios.

La técnica que ayuda a borrar el límite entre el esquema de los documentos JSON y sus valores de instancia es representar los documentos como árboles. La representación de documentos JSON como árboles a su vez normaliza tanto la estructura como los valores de instancia en los documentos en un concepto unificador de una estructura de ruta codificada dinámicamente.

Organización del índice Lógico

La representación lógica del índice se puede ver como un conjunto ordenado de tuplas clave-valor, cada una de las cuales se denomina entrada de índice. La clave consta de un término que representa la información de ruta codificada del nodo en el árbol de índice y un PES (selector de entrada de publicación) que ayuda a dividir las publicaciones horizontalmente. El valor consiste en una lista de publicaciones que representan colectivamente los identificadores de documentos codificados (o fragmentos de documentos).

Una lista de publicaciones captura los identificadores de todos los documentos que contienen el término dado. El tamaño de la lista de publicaciones es una función de la frecuencia del documento: la cantidad de documentos en la colección que contiene un término determinado, así como el patrón de ocurrencia de los identificadores de documentos en la lista de publicaciones.

Organización del índice Físico

La indexación coherente en DocumentDB proporciona nuevos resultados de consulta frente a la ingestión sostenida de documentos. Esto plantea un desafío en un entorno de múltiples inquilinos con presupuestos frugales para memoria, CPU e IOPS. El mantenimiento eficiente del índice del documento sin ningún conocimiento previo de los esquemas del documento depende de la elección de la estructura de datos más "eficiente en escritura" para administrar las entradas del índice. Además de los requisitos anteriores para la actualización del índice, la estructura de datos del índice debe ser capaz de atender consultas de búsqueda de puntos, rangos y comodines de manera eficiente, todo lo cual es clave para el lenguaje de consulta de DocumentDB.