

Una propuesta a la industria de bases de datos
No tres sino un: GQL

Un lenguaje estándar descriptivo de consulta de grafos de propiedad
Un complemento a SQL

La siguiente propuesta fue enviada originalmente por Neo4j a proveedores en el segmento de manejo de datos de grafos con propiedades, investigadores y expertos en industria activa en

- Linked Database Benchmarking Council (LDBC) Graph Query Language Task Force,
- Grupo de trabajo INCITS Ad Hoc for Property Graph Extensions (sub grupo de INCITS DM32.2, el cuerpo responsable de proponer cambios al estándar ISO SQL en los EEUU)

Equipo técnico de Neo4j ha contribuido activamente en ambos grupos mencionados.

—

GCORE es un lenguaje de investigación designado por colaboradores de LDBC, y descrito en un documento aceptado por el segmento industrial de SGMOD'18.

PGQL 1.1 es un lenguaje usado por el producto de Oracle PGX, el cual también ha sido implementado en contexto de investigación.

Cypher es un lenguaje originalmente creado por Neo4j y actualmente mantenido por el proyecto openCypher, el cual ha sido implementado en Neo4j Server y Neo4j Morpheus for Apache Spark (beta), Redis Graph, Agens Graph (sobre PostgreSQL), SAP HANA Graph, Memgraph; en el proyecto open-source Cypher for Apache Spark y Cypher for Gremlin, y en proyectos de investigación como InGraph y Cytosm (Cypher a SQL). Su historia y trayectoria es descrita en un documento aceptado por el segmento industrial de SGMOD'18.

Estos tres lenguajes comparten raíces, sintaxis y semántica común, y sus autores comparten múltiples objetivos técnicos para la evolución del lenguaje de consulta de bases de datos de grafos.

—

Viernes, 11 de Mayo 2018

Estimados colegas [de industria e investigación],

Hemos pensando acerca de varios esfuerzos recientes en lenguajes de consulta de bases de datos de grafos (SQL, LDBC, Cypher, PGQL).

He mantenido comunicación con Emil Eifrem, nuestro CEO, y Philip Rathle, nuestro líder de Producto, quienes son ejecutivos que promueven nuestro trabajo en esta área, en respecto a la decisión inminente de fijar un alcance de trabajo limitado para SQL:2020 Property Graph Querying.

Está claro que SQL:2020 va a soportar la definición de grafos en DDL, su mapeo a tablas existentes y la habilidad de ejecutar consultas de tipo MATCH y así extraer resultados tabulares. Al parecer por lo menos Oracle, IBM, SAP y Neo4j concuerdan que este es un objetivo realista para SQL, juzgando por discusiones y correspondencia reciente. Nosotros sentimos que llegar a ese punto en SQL sería un resultado muy valioso.

Al mismo tiempo, también es claro que Cypher, PGQL de Oracle y el lenguaje de investigación GCORE están entrelazando un espacio funcional más amplio, con frecuentes superposiciones en el contenido y perspectiva.

LDBC GCORE aboga (en esencia) algo que parece ser Cypher solo de lectura, adicionado con consultas de composición (cierres sobre grafos), lo que implica múltiples grafos nombrados / construidos, al igual que consultas de ruta regular.

PGQL es altamente similar sintáctica y semánticamente a Cypher solo de lectura, contiene múltiples fuentes grafos y consultas de ruta regular. Pero aún no cuenta con soporte de creación de grafos, aunque estamos seguros que sus creadores estarían interesados en seguir esa dirección

Cypher se ha expandido, con las características implementadas en Cypher para Apache Spark, para cubrir múltiples fuentes para entrada y salida de grafos, y grafos nombrados (cierres sobre grafos). Esta es la dirección que queremos seguir y finalizar, incluyendo grafos y vistas actualizables. Pero Cypher no cuenta aún con el nivel de consulta de ruta regular que PGQL cuenta, siendo una brecha que nos gustaría cubrir.

Además, muchos de los interesados, incluyendo especialistas en SQL, parecen estar interesados a cierto grado en evolucionar el pensamiento más allá en temas como morfismo controlable y consultas de ruta regular más sofisticadas.

Hace poco más de un año en la reunión Walldorf de LDBC, Jan Michels de Oracle planteó la posibilidad de dos iniciativas paralelas en ISO: un lenguaje nativo de consulta de grafos con propiedades, y extensiones para SQL. Para cuando el actual Ad Hoc fue creado in abril 2018 esto se redujo únicamente a extensiones para SQL.

Hemos puesto mucho trabajo en el proyecto de extensiones de grafos para SQL, y creemos que seguir este camino sería muy positivo.

Sin embargo, también creemos que hay oportunidades paralelas para llegar a una narrativa estándar (propiedad), un Graph Query Language (GQL), que cumpla la promesa de GCORE, la cual complementarían las limitadas capacidades de consulta de grafos de SQL.

Si recordamos las críticas de SAP a Cypher que Marcus Paradies hizo en la reunión de implementadores de openCypher en las oficinas de SAP en febrero del año pasado (ver imagen), entonces podemos ver que desde entonces PGQL, GCORE y openCypher han seguido empujando hacia un objetivo mayor, el cual SAP aboga en su presentación:

IMAGE****

Hoy podríamos resumir el resultado lógico de esta dirección como: La unión de PGQL y Cypher con las propuestas de GCORE. No tres, sino un, Graph Query Language nativo.

Nosotros sentimos que aún no estamos actuando colectivamente para perseguir esa lógica común. Existe riesgo real de que los últimos quince a dieciocho meses de trabajo intensivo y convergente en varios grupos terminen siendo una oportunidad desperdiciada. No es claro que alguien busque este resultado. Por nuestra parte, queremos evitar que esto suceda – y expresar firmemente nuestra disposición en participar en un proceso abierto y unido que pueda prevenir un resultado negativo.

Te estamos escribiendo a ti y a todos los participantes, a medida que avanzamos hacia la primera reunión de ISO WG3 para discutir seriamente el área de consulta de grafos, para pedir tu apoyo en la siguiente iniciativa:

- 1) Buscar un lugar de encuentro estándar formalmente constituido, capaz de formar un estándar internacional, o para ser referido por un estándar internacional, con trato equitativo a todos los participantes y apropiada protección de la Propiedad Intelectual para compartir de forma segura ideas y trabajo. ISO WG3 (que es en principio el responsable de los Lenguajes de Bases de Datos en lo plural y no necesariamente solo SQL) es una posibilidad con la que estaríamos contentos. Hay otras posibilidades, y nosotros en Neo4j no somos partidarios al respecto.
- 2) Tomar como base PGQL, Cypher y GCORE sin poner ataduras en el producto más allá de reconocer que estos lenguajes a) Intersectan ampliamente en sintaxis, semántica y aparente plan de trabajo, y b) PGQL y Cypher son bien establecidos e implementados en (múltiples) productos. Si existen otros lenguajes que deben de ser considerados, nosotros ciertamente apoyaríamos su inclusión como base.
- 3) Buscar unificar estas bases en un lenguaje – llamémoslo GQL como una propuesta – buscando ser adoptado por la industria, incluyendo a los nuevos servicios de nube como Amazon Neptune y Azure CosmosDB, como un estándar, para complementar SQL.
- 4) Buscar definir la relación de GQL (soportable, interoperable, referido) a SQL, en colaboración con WG3.

Por nuestra parte, nosotros nos comprometemos a entregar, sin restricciones de Propiedad Intelectual todos los artefactos de openCypher, software o textual, que hemos generado; a buscar acuerdo y soportar a otros colaboradores de openCypher; y a proveer un grupo significativo de personal técnico de nivel Sr. de Neo4j con conocimiento en este ámbito por lo menos por el periodo hasta junio 2020 para trabajar conjuntamente con otros participantes en el estándar propuesto de GQL, asumiendo que una masa crítica de proveedores industriales en este espacio se inifiquen para buscar esta iniciativa. (Para nosotros, desde luego, sería bienvenida la participación de expertos investigadores.)

Estamos considerando hacer esta propuesta en las reuniones de ISO WG3/ INCITS DM32.2/Ad Hoc la semana entrante en Toronto, y queremos darles a conocer nuestro pensar por adelantado, al igual que también queremos pedir su seria consideración para apoyar esta iniciativa, la cual creemos es de gran valor para toda la industria y todos los usuarios.

Adjunto un diagrama que (a pesar de mi mala caligrafía y pobreza en trazado) puede ayudar a ilustrar la esencia de esta propuesta.

Con un cordial salud,

Alaistar

Signature