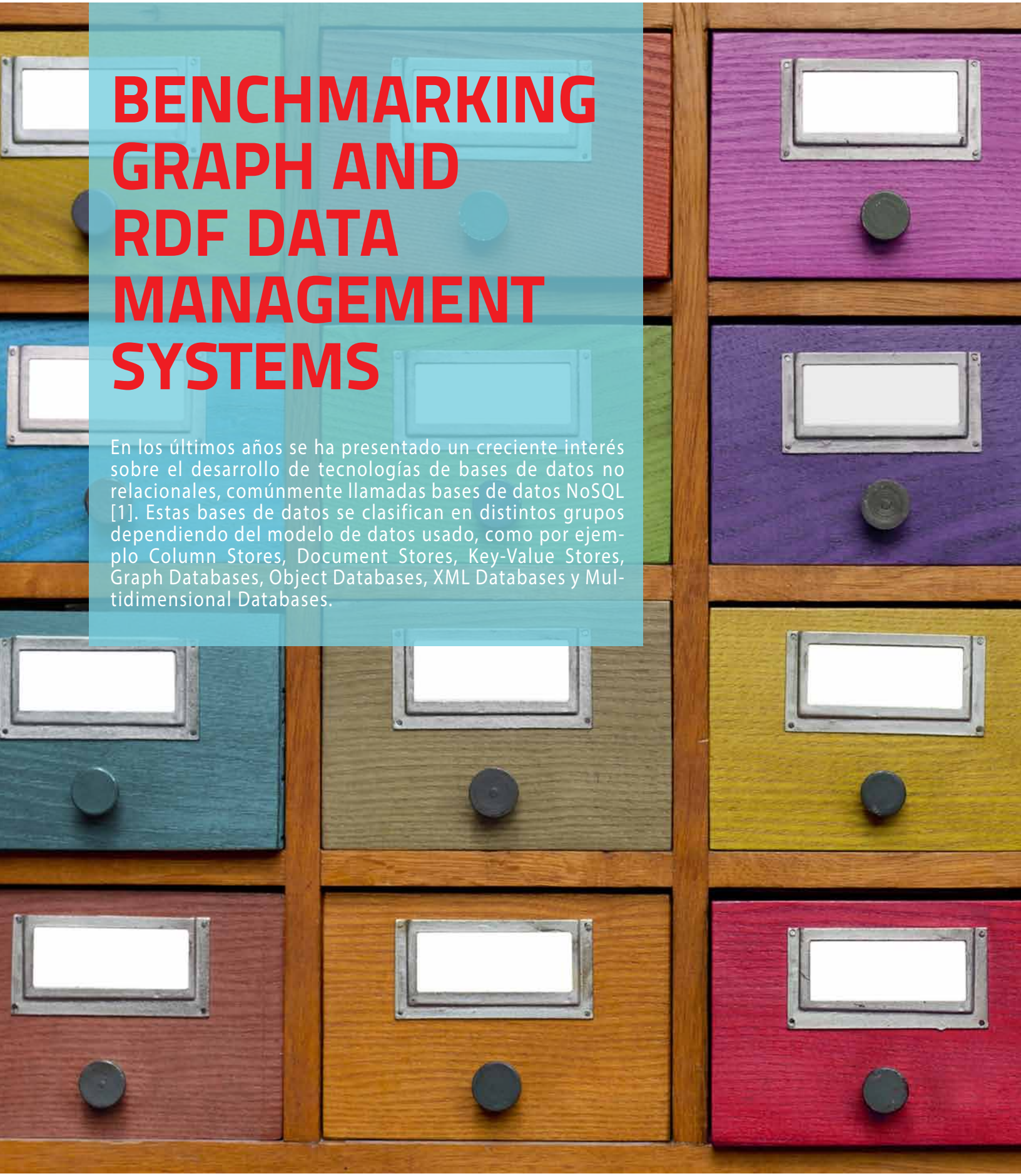

BENCHMARKING GRAPH AND RDF DATA MANAGEMENT SYSTEMS

En los últimos años se ha presentado un creciente interés sobre el desarrollo de tecnologías de bases de datos no relacionales, comúnmente llamadas bases de datos NoSQL [1]. Estas bases de datos se clasifican en distintos grupos dependiendo del modelo de datos usado, como por ejemplo Column Stores, Document Stores, Key-Value Stores, Graph Databases, Object Databases, XML Databases y Multidimensional Databases.

RENZOANGLES

Profesor Asistente, Departamento de Ciencias de la Computación, Universidad de Talca. Ingeniero de Sistemas, Universidad Católica de Santa María, Perú. Doctor en Ciencias mención Computación, Universidad de Chile. En 2013 realizó un postdoctorado en la VU University Amsterdam, participando en el proyecto "Linked Data Benchmark Council (LDBC)". Sus áreas de investigación son Bases de Datos de Grafos y Web Semántica, específicamente en Lenguajes de Consulta y Benchmarking de Bases de Datos para Grafos y RDF.

rangles@utalca.cl

BENCHMARKS PARA GRAPH/RDF DATABASES

La aparición del enfoque NoSQL se debió principalmente a las limitaciones que tienen las bases de datos tradicionales (aquellas basadas en el modelo relacional) para satisfacer los requisitos de gestión de datos en dominios de aplicación no tradicionales, los cuales requieren lidiar con grandes cantidades de datos de estructura compleja, como por ejemplo Big Data [2] o Linked Open Data [3]. En este sentido, las tecnologías NoSQL buscan mejorar la flexibilidad y el desempeño de los sistemas de gestión de datos basados en características como escalabilidad horizontal, independencia de esquema de datos, consistencia de datos parcial, replicación de datos y computación distribuida [4].

Las bases de datos orientadas a grafos (en inglés, *Graph Databases*) y las bases de datos RDF (en inglés, RDF *databases*, también llamadas *Triple Stores*) son dos enfoques NoSQL orientados a lidiar con datos no estructurados (heterogéneos, no relacionales) y altamente conectados. Las graph databases están diseñadas para almacenar datos con estructura de grafo y consultarlos a través de operaciones y/o lenguajes de consulta pensadas para explorar los grafos almacenados. Las RDF databases son bases de datos de grafo especialmente diseñadas para gestionar datos semiestructurados y metadatos creados en base al modelo de datos RDF, y permiten consultar dichos datos usando el lenguaje de patrones SPARQL, además de operadores especiales que permiten inferencia sobre los datos.

Actualmente existen diversas graph databases (Sparksee, Neo4j, AllegroGraph) y RDF databases (OpenLink Virtuoso, OWLIM, Sesame). Sin embargo, esta diversidad de sistemas genera las siguientes preguntas: ¿cuál es el desempeño de una graph/RDF database? ¿Cuál es la graph/RDF Database con mejor desempeño? Para poder responder a estas consultas, debemos evaluar y comparar los sistema de bases de datos usando herramientas denominadas benchmarks.

En el contexto general, un benchmark es una herramienta que permite comparar el desempeño de los sistemas. En el contexto de las bases de datos, un benchmark permite evaluar la capacidad de los sistemas de gestión de bases de datos, en particular su eficiencia para responder a las operaciones sobre los datos. De esta manera, los benchmarks muestran las fortalezas y debilidades de los sistemas.

Cabe destacar que los benchmarks no son pensados únicamente para evaluar los sistemas, más importante aún, estos buscan estimular el avance tecnológico a través de la identificación de posibles mejoras a nivel de desempeño y funcionalidad. En conclusión, los benchmarks ayudan a los usuarios finales de las bases de datos en la elección de productos de software competitivos y guían a la industria hacia el desarrollo de nuevas tecnologías.

La existencia de procesos estándar de benchmarking que definan la ejecución correcta de un benchmark sobre un sistema, son fundamentales para asegurar la confianza y aceptación de los benchmarks. Por ejemplo, el Transaction Processing Performance Council (TPC) [5] es un consorcio creado para supervisar el desarrollo y ejecución de benchmarks estándar para bases de datos relacionales; esto con la finalidad de asegurar resultados de benchmarking confiables para la industria y el mercado de usuarios finales.

En el contexto de las graph/RDF databases, no existen benchmarks estándar ni tampoco una autoridad independiente que se encargue de controlar los procesos de benchmarking. Sin bien existen algunos benchmarks provenientes del ámbito académico, estos no cumplen totalmente con las características deseadas en los benchmarks industriales, como por ejemplo alcance, relevancia, verificabilidad, en otras [6]. Además, los benchmarks académicos no modelan escenarios caracterizados por operaciones complejas sobre datos asimétricos y altamente correlacionados, como aquellos encontrados en casos de uso reales como Big Data o Linked Open Data [7].

THE LINKED DATA BENCHMARK COUNCIL

The Linked Data Benchmark Council (LDBC) [8] es un proyecto europeo que reúne una comunidad de académicos y expertos de la industria, cuyo objetivo común es el desarrollo de benchmarks estándar para la industria de graph/RDF databases.

El proyecto LDBC busca crear benchmarks siguiendo los principios de relevancia, simplicidad, confiabilidad y sostenibilidad. De manera especial, el LDBC busca el desarrollo de benchmarks que evalúen funcionalidades críticas de los sistemas, yendo más allá de los benchmarks creados en la academia. Con este fin, el LDBC entregará benchmarks de código abierto, desarrollados por grupos de trabajo integrados por expertos en arquitectura de bases de datos quienes conocen las funcionalidades críticas dentro de los motores de gestión de datos, y soportados por una comunidad de usuarios que entregan casos de uso y retroalimentación. Además, el LDBC espera incluir un mecanismo que asegure que los resultados de benchmarking sean revisados por una entidad independiente para verificar su conformidad.

A continuación describiremos brevemente los elementos que conforman un benchmark, según la guía de diseño elaborada al interior de LDBC, y luego describiremos brevemente dos benchmarks que se encuentran en pleno proceso de desarrollo: el Social Network benchmark y el Semantic Publishing benchmark.

DISEÑO DE BENCHMARKS EN LDBC

Un benchmark está compuesto generalmente de tres elementos: un generador de datos (*data generator*), el cual permite crear datos en base a un esquema de datos definido; un generador de carga de trabajo (*workload generator*), el cual define el conjunto de operaciones (workload) que el sistema bajo evaluación (system under test, SUT) tendrá que procesar; y un conductor de pruebas (test driver), el cual es usado para ejecutar el workload sobre el sistema bajo evaluación, siguiendo reglas de ejecución precisas y midiendo el desempeño según métricas bien definidas.

En la **Figura 1**, se muestra la estructura del Social Network benchmark tomando en cuenta los tres componentes descritos anteriormente. Observe que cada componente contiene subelementos con características específicas; por ejemplo, los datos generados serán no-estructurados e incluirán correlaciones y distribuciones no uniformes. Nótese además, que la **Figura 1** incluye características alrededor de los componentes del benchmark (ej. simplicidad), las cuales dirigen su diseño y construcción.

El desarrollo de un benchmark implica un número significante de detalles que deben tomarse en cuenta, en particular durante la etapa de diseño. Entre estos detalles podemos mencionar:

- *El caso de uso debe ser real, claro, comprensible y relevante para los usuarios y la comunidad.*
- *El workload debe ser representativo de las operaciones encontradas en el caso de uso seleccionado.*

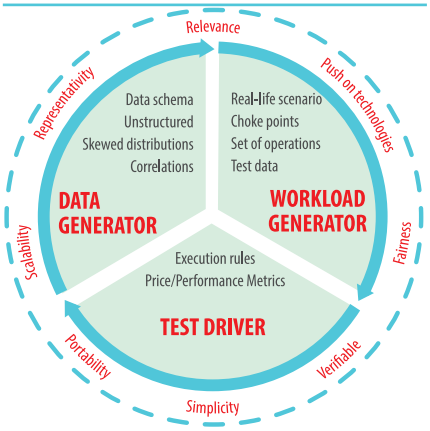

FIGURA 1.
ELEMENTOS Y CARACTERÍSTICAS DEL LDBC SOCIAL NETWORK BENCHMARK.

- *El workload debe ser diseñado cuidadosamente para asegurar que el benchmark incentive la innovación tecnológica. En este sentido, cada workload será definido en base a desafíos técnicos bien identificados denominados choke points. El objetivo de un diseño basado en choke points es asegurar que un workload presiona las funcionalidades técnicas más importantes de los sistemas actuales.*
- *Las reglas de ejecución y las métricas del benchmark deben ser definidas cuidadosamente para asegurar una evaluación confiable y una comparación justa de los sistemas.*

El desarrollo de benchmarks en el LDBC está a cargo de grupos de desarrollo denominados *task forces*. Una task force está formada por miembros del LDBC, incluyendo tanto usuarios finales como proveedores de tecnologías de base de datos. Se espera que los participantes de la industria sean personas expertas en los aspectos técnicos de los sistemas, de manera que faciliten la definición de choke points. En el caso de los usuarios finales, se espera recibir sus requisitos respecto a casos de uso relevantes así como la entrega de retroalimentación durante el desarrollo de un benchmark.

THE SOCIAL NETWORK BENCHMARK

El Social Network Benchmark (SNB) es un benchmark pensado para evaluar diversas funcionalidades de sistemas usados en la gestión de datos con estructura de grafo. El escenario de

este benchmark es una red social, y fue elegido con la siguientes metas en mente: deberá ser entendible para una gran audiencia, y esta audiencia deberá comprender la relevancia de gestionar dichos datos; deberá cubrir un conjunto de desafíos interesantes acorde con los alcances del benchmark; si bien los datos y el workload se crearán de manera sintética, estos deberán ser realistas en el sentido de reflejar características encontradas en la vida real.

El generador de datos del SNB está siendo diseñado para crear datos sintéticos con las siguientes características: el esquema de datos debe ser representativo de una red social; el método de generación debe considerar las propiedades existentes en redes sociales reales, incluyendo correlaciones entre los datos y distribuciones estadísticas; las herramientas de software generadas deben ser fáciles de usar, configurables y escalables.

El esquema de datos del SNB modela una red social con perfiles de usuario enriquecidos con intereses, etiquetas (tags), mensajes (posts) y comentarios. Adicionalmente, los datos generados exhiben correlaciones reales entre valores (ej., los nombres de las personas son creados de acuerdo con su nacionalidad), correlaciones de estructura (ej., dos personas amigas en su mayoría viven en lugares cercanos geográficamente), y distribuciones estadísticas (ej., la relación de amistad sigue una distribución de ley de potencias). El generador de datos está implementado para ejecutarse en Hadoop, lo cual permite una generación rápida y escalable de archivos de datos de gran tamaño.

Con el objetivo de cubrir los requisitos más relevantes de las aplicaciones que gestionan datos sobre redes sociales, el SNB entrega tres workloads distintos (de cierto modo el SNB es tres *benchmarks* en uno): un *interactive workload*, orientado a evaluar consultas relativamente simples y operaciones de actualización concurrentes; un *business intelligence workload*, compuesto de consultas complejas que simulan un análisis en línea del comportamiento de los usuarios, esto con el propósito de realizar marketing; y un *graph analytics workload*, pensado para evaluar la funcionalidad y escalabilidad de los sistemas para el análisis de grafos a través de operaciones complejas, las cuales usualmente no pueden ser expresadas usando un lenguaje de consulta.

Adicionalmente, cada workload incluirá una o más métricas para medir el desempeño de los sistemas, como por ejemplo el tiempo de respuesta o el throughput (métrica que mide el número de operaciones por unidad de tiempo, por ejemplo, transacciones por minuto).

THE SEMANTIC PUBLISHING BENCHMARK

El Semantic Publishing Benchmark (SPB) está diseñado para simular la gestión y consumo de metadatos RDF sobre contenido multimedia. El escenario específico se basa en una organización de noticias, la cual mantiene descripciones RDF de su catálogo de noticias además de los trabajos creativos. El SPB simula un workload donde un gran número de agentes consultan el catálogo de artículos noticiosos y al mismo tiempo se tienen operaciones de edición y descripción del contenido multimedia.

Para la generación de datos, el SPB emplea una ontología que define numerosas propiedades para el contenido, por ejemplo fecha de creación, resumen, descripción, entre otros. Además, una ontología de etiquetas es usada para clasificar los trabajos creativos en diversas categorías como deportes, geografía o información política.

El SPB incluye dos workloads que demandan alto desempeño para la ejecución de consultas (que pueden calcularse de manera paralela), así como para operaciones de actualización continuas y concurrentes. El *Editorial* Workload simula la creación, actualización y borrado de metadatos sobre trabajos creativos. Este workload se basa en que las compañías de medios usan procesos manuales y semiautomáticos para gestionar descripciones de trabajos y clasificarlas de acuerdo a categorías definidas en ciertas ontologías, además de incluir referencias a otras fuentes de datos. El *Aggregation workload* simula la agregación dinámica de contenido para su inmediato consumo (ej., a través de un sitio web). La acción de publicar contenido es considerada dinámica ya que el contenido no se selecciona ni arregla manualmente, en lugar de esto se usan plantillas que entregan formato al contenido, el cual es seleccionado cuando el usuario lo accede. En este workload se usan consultas SPARQL para encontrar contenido relevante.

De manera preliminar, el SPB considera el throughput como métrica para medir las operaciones de actualización y consulta que ejecutan los agentes editores y consumidores de contenido durante una cantidad definida de tiempo.

NOTAS FINALES

El software y los documentos asociados al SNB y al SPB pueden ser descargados desde GitHub [9] y la información sobre su desarrollo se encuentra disponible en el Wiki [10] administrado por las Task Force. Se invita al lector a unirse a la comunidad del LDBC para colaborar e influenciar en el desarrollo de benchmarks para graph/RDF databases. ■

REFERENCIAS

[1] NoSQL Databases. http://nosql-database.org
[2] C. Snijders, U. Matzat and U. Reips. Big Data: Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science, Vol. 7, Nº 3, 2012.
[3] T. Heath and C. Bizer. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 1, Nº 1, 2011.
[4] M. Stonebraker. SQL Databases V. NoSQL Databases. Communications of the ACM, Vol. 53, Nº 4, 2010.
[5] Transaction Processing Performance Council (TPC). http://www.tpc.org/
[6] K. Huppler: "The Art of Building a Good Benchmark", in: TPCTC, 2009.
[7] S. Duan, A. Kementsietsidis, K. Srinivas and O. Udrea. Apples and Oranges: A Comparison of RDF Benchmarks and Real RDF Datasets. Proc. of the International Conference on Management of Data (SIGMOD), 2011.
[8] Linked Data Benchmark Council (LDBC). http://www.ldbc.eu
[9] LDBC Software and Documentation. https://github.com/ldbc/
[10] LDBC Technical User Community Wiki. http://138.232.65.142:8090/display/TUC/