PATROCINADO POR

IBM®

# GEEK GUIDE

# Por qué los desarrolladores de aplicaciones prefieren OSDBMS de alta velocidad

LINUX™ JOURNAL

# Tabla de contenido



Ted Schmidt es un asesor especializado en soluciones de marketing y comercio electrónico para el sector manufacturero. Ted ha trabajado en la gestión de proyectos y productos desde antes de que empezara el movimiento "agile" en el 2001 y ha gestionado la entrega de proyectos y productos para fabricantes de bienes de consumo, dispositivos médicos, electrónica y telecomunicaciones durante más de 20 años. Cuando no está inmerso en el desarrollo de productos, Ted escribe novelas y dirige un pequeño taller de diseño gráfico en http://FloatingOrange.com. Ted ha sido ponente en conferencias de PMI y tiene un blog en http://FloatingOrangeDesign.Tumblr.com y en su página web en http://FloatingOrange.com.

## GEEK GUIDES:

Información de misión crítica para las personas más técnicas del mundo.

## Acerca del patrocinador

### IBM

IBM es una empresa global de asesoría y tecnología integrada con sede en Armonk, Nueva York. Con operaciones en más de 170 países, IBM atrae y retiene a algunas de las personas con más talento del mundo para ayudar a resolver problemas y ofrecer una ventaja competitiva a empresas, gobiernos y entidades sin ánimo de lucro.

La innovación es el centro de la estrategia de IBM. La compañía se ha reinventado a lo largo de diferentes eras tecnológicas y ciclos económicos, creando valor diferenciador para sus clientes. En la actualidad, ante el ritmo sin precedentes con el que cambia fundamentalmente el sector de las TI, IBM es mucho más que una empresa de “hardware, software y servicios”. IBM se está transformando en una empresa de soluciones cognitivas y plataformas cloud.

Las soluciones cognitivas basadas en el cloud son la clave de la transformación digital de los clientes. Esta transformación requiere avances en todos los niveles de la base TI de la empresa, desde los procesadores y diseño de sistemas hasta el almacenamiento, la red y la capa de integración. IBM Power Systems, fabricados con tecnologías abiertas y diseñados para aplicaciones de misión crítica, ofrecen la infraestructura diseñada para cargas de trabajo cognitivas.

# Por qué los desarrolladores de aplicaciones prefieren OSDBMS de alta velocidad

TED SCHMIDT

## Introducción

Como puede atestiguar cualquier desarrollador de aplicaciones sociales, móviles o IoT, los modelos de bases de datos relacionales existentes ya no resuelven todas nuestras necesidades. Los sistemas de gestión de bases de datos tradicionales no tienen nada de malo, simplemente no se han diseñado para dar respuesta a la variedad y volumen de datos que circulan por el mundo digital actual, sin mencionar la demanda de velocidad de proceso de todos estos datos con el fin de ofrecer las funciones útiles y orientadas a datos que todos nosotros esperamos encontrar cada vez más en prácticamente todas las cosas. Por fortuna, se ha creado un mundo totalmente nuevo de sistemas de gestión de bases

de datos de código abierto (OSDBMS) precisamente para gestionar la diversidad y complejidad de los datos actuales y para poder almacenar, analizar y actuar en ellos con la velocidad necesaria para que tengan un valor hasta ahora imposible.

En esta guía se examinarán algunos de los sistemas OSDBMS disponibles y las soluciones que ofrecen a problemas a los que nos enfrentamos en el desarrollo de aplicaciones innovadoras, para nuevos datos de fuentes tales como las redes sociales, móviles e IoT. Aunque el estudio de *todos* los actores OSDBMS va más allá del alcance de esta guía, se ofrece una mirada equilibrada de los actores más importantes en cada una de las principales categorías: SQL de código abierto, NoSQL (incluyendo almacenes de grafos, documentos y calve-valor) e incluso los productos en memoria acelerados por GPU disponibles en la actualidad.

Empezará analizando algunos de los principales retos que plantea el nuevo entorno de aplicaciones, como el Big Data que las nuevas aplicaciones crean y consumen. Ofreceré una visión general del paisaje OSDBMS, seguido de un examen con mayor profundidad de las principales ofertas de OSDBMS. Por último, concluirá con una mirada en la mejor plataforma tecnológica disponible para que estas DBMS modernas funcionen a su máximo nivel, resolviendo los retos planteados y rompiendo nuevos récords de velocidad, transferencia y escala para dar servicio a las aplicaciones más innovadoras actuales.

## Retos a los que se enfrentan los desarrolladores de aplicaciones innovadoras

Seguimos dando vueltas a los retos asociados al “Big Data” y su evolución. Aunque el Big Data pueda tener diferentes requisitos para las distintas organizaciones y sectores, cuando aquí hablemos de Big Data, nos referiremos específicamente a él en términos de los problemas planteados cuando


FIGURA 1. Las cuatro dimensiones del Big Data.

se diseñan y construyen aplicaciones innovadoras en un mundo que exige acceso instantáneo a grandes volúmenes de datos. Estoy considerando las fuentes externas de Big Data y los problemas asociados a la captura, almacenamiento, análisis y visualización de estos datos. Para un desarrollador de aplicaciones centrado en la creación de soluciones innovadoras basadas en datos, son de interés 4 dimensiones del Big Data: tamaño, velocidad, complejidad y fiabilidad. En esta guía me centro mayoritariamente en los aspectos de la velocidad y la complejidad, así como en el valor de las soluciones OSDBMS y NoSQL para resolverlos.

La dimensión de la velocidad se refiere a la rapidez con que se crean los datos, se almacenan, se procesan, se recuperan y, en última instancia, pueden utilizarse en el análisis. Las organizaciones exigen cada vez más que los datos puedan procesarse en tiempo real y enviarlos directamente a los procesos de toma de decisiones de una empresa. Piense en los sistemas de control de tráfico. Aunque la demanda

de velocidad siga aumentando, nuestra capacidad para seguir su ritmo se ve obstaculizada por un par de elementos.

Evidentemente, la latencia (la diferencia entre cuando se recogen los datos y cuando están disponibles para su uso) va a afectar a la velocidad. Además de la latencia tenemos al acecho el fantasma de la ley de Moore. La Ley de Moore afirma que la potencia de proceso informático global se duplica cada dos años. Aunque esta ley se haya cumplido durante un tiempo, el sentido común nos dice que debe existir un límite físico en lo que pueden proporcionarnos los chips de sílice. Es física y afecta directamente a la capacidad para que las aplicaciones sigan obteniendo las ganancias de velocidad que necesitan para procesar en tiempo real crecientes volúmenes de datos cada vez más complejos.

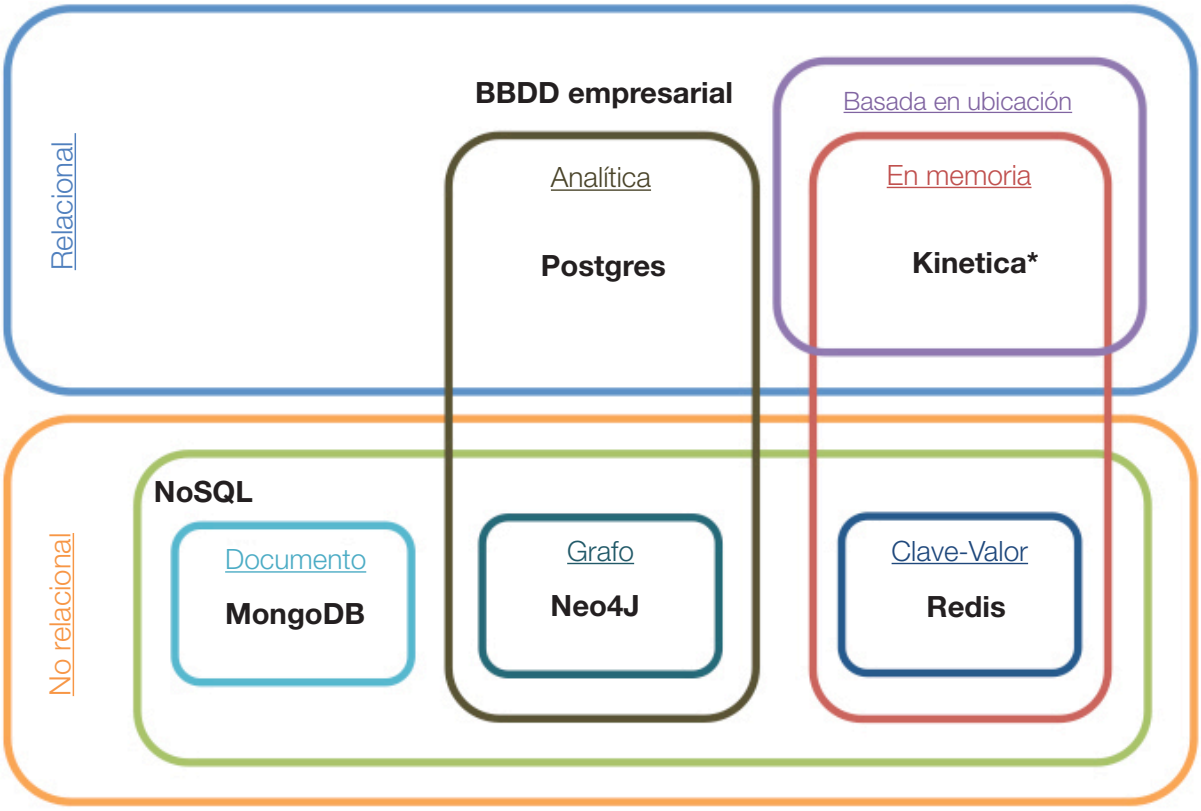
Lo que esto significa es que, como desarrolladores, tenemos más problemas ahora ya que simplemente no podemos confiar en la Ley de Moore como respuesta a la demanda de más velocidad. Tenemos que ser más innovadores y esto significa que debemos mirar las ventajas de las soluciones de código abierto en lugar de seguir confiando en las bases de datos tradicionales, para poder gestionar los enormes volúmenes de datos con los que estamos trabajando actualmente y de los cuales nos esforzamos en ofrecer funciones, características y conocimiento tan rápidamente como sea posible.

## Qué aporta el código abierto

Existen decenas de soluciones DBMS de código abierto. Aquí se analizan 5 que ofrecen ventajas concretas en el desarrollo de aplicaciones innovadoras: MongoDB, Redis, Neo4j, PostgreSQL y Kinetica. (Precisión: aunque Kinetica no sea de código abierto, forma parte de un ecosistema de tecnología abierta mediante su pertenencia y participación en la OpenPOWER Foundation. En este sentido, y debido

a sus ventajas concretas en el proceso en tiempo real de grandes streamings de conjuntos de datos, lo incluso en este análisis. La OpenPOWER Foundation, de la que hablará más tarde en esta guía, es una organización abierta de miembros tecnológicos que básicamente busca dar respuesta a los límites de la Ley de Moore, permitiendo a las empresas miembro personalizar CPUs de POWER de formas nuevas e innovadoras para exprimir hasta el último gramo de velocidad y potencia posible. Recomiendo encarecidamente que obtengan más información sobre la OpenPOWER Foundation en https://openpowerfoundation.org, pero será suficiente con decir que es aquí donde se produce la innovación.)

Pasar a un OSDBMS ofrece ventajas tangibles y, a medida que las herramientas de gestión siguen madurando a nivel de empresa, las limitaciones anteriores están desapareciendo. El menor coste asociado a la libertad del software no propietario, e incluso del hardware no propietario, es la ventaja más evidente del código abierto en general. Pero, como verá cuando cubra los casos de uso específicos que mejor se adaptan a cada una de estas soluciones OSDBMS, la velocidad, la flexibilidad y la toma de decisiones inteligente son las más importantes. Esto no quiere decir que nos olvidemos de la ventaja que el código abierto aporta a la velocidad de la evolución y la innovación. Por su naturaleza, las soluciones propietarias evolucionan más lentamente que la demanda del entorno en constante evolución. No obstante, las soluciones de código abierto se benefician de la aportación de una amplia comunidad de muchas voces centradas en la resolución de problemas reales. En este sentido, las soluciones de código abierto evolucionan e innovan más rápidamente de lo que cualquier solución propietaria puede hacer.




* Aunque Kinetica no sea de código abierto, forma parte del ecosistema Open Tech.

FIGURA 2. Paisaje de los OSDBMS.

Consideremos las dos principales categorías del paisaje OSDBMS: no relacional (que incluye a MongoDB, Redis y Neo4j) y relacional (que incluye a EDB Postgres y Kinetica).

En el lado no relacional se encuentra Redis, una base de datos de clave-valor NoSQL particularmente útil en el sector de los juegos (entre mucho otros), en el que predominan las operaciones de datos de alta velocidad, simples pero elegantes. A continuación, Neo4j es una base de datos de grafos NoSQL especialmente eficaz en el almacenamiento de relaciones entre puntos de datos. Por último, MongoDB es una base de datos de documentos NoSQL, ideal como

base de datos de propósito general pero particularmente útil por su falta de esquema, lo que significa que se puede almacenar todo tipo de datos distintos de una forma muy flexible.

En la categoría relacional, Kinetica no es de código abierto pero merece su reconocimiento por la pura velocidad que pone encima de la mesa. Kinetica es una base de datos relacional que también se basa en ubicaciones, y una base de datos en memoria acelerada por GPU. La clave de su gran velocidad en el procesamiento de grandes streamings de datos reside en la última parte de la frase anterior: acelerada por GPU (más detalles a continuación). Por último, pero no menos importante, la otra base de datos relacional que analizaremos, EDB Postgres Advanced Server, es realmente el paquete de PostgreSQL para empresas creado por EDB, un producto del desarrollo de código abierto por parte de una comunidad en la que participa activamente y da soporte religiosamente. También es magnífica para la analítica.

En la vertiente no relacional del paisaje se encuentra otra base de datos en memoria: Redis. Redis es una base de datos de clave-valor NoSQL particularmente útil en el sector de los juegos. Neo4j es otra base de datos NoSQL que, como base de datos de grafos, es especialmente eficaz en el almacenamiento de relaciones entre puntos de datos. Por último, MongoDB es también una base de datos NoSQL, pero es una base de datos de documentos, lo que significa que, aunque sea una buena base de datos de propósito general, su ventaja real proviene de su falta de esquemas. Sin un esquema es posible almacenar todo tipo de datos distintos, como un aumento de temperatura o rotaciones por segundo.

Tenga presente que, aunque las ventajas de cada OSDBMS, muchas veces solapadas, en última instancia dependan de su caso de uso específico, todas estas bases de datos comparten una ventaja similar: formar parte de un modelo de desarrollo

de tecnología abierta. No sólo eso, sino que, en la infraestructura correcta, cada OSDBMS proporciona el añadido desesperadamente necesario de un avance en la velocidad para la innovación real en el desarrollo de aplicaciones que hagan un gran uso de los datos.

## DBMS de almacén de documentos MongoDB

MongoDB es una base de datos de código abierto con un modelo de datos orientado a documento. Almacena datos en formato JSON Binario (BSON), que amplía JSON para incluir otros tipos de datos, tales como int, long, fecha, coma flotante, etc. Estos documentos BSON permiten modelar de una forma mucho más rápida y fácil los datos de la aplicación en la base de datos, puesto que los documentos BSON se alinean con la estructura de objetos del lenguaje de programación. Cada documento contiene varios campos y cada campo contiene un valor de un tipo de datos específico, como por ejemplo subdocumentos o matrices. Los documentos con estructuras similares se agrupan en colecciones. En un RDBMS, las colecciones serían tablas, los documentos serían filas y los campos serían columnas.

Un modelo de datos orientado a documento no tiene esquemas. Por ello, a diferencia de un RDBMS, que almacena valores NULL en los campos vacíos, en MongoDB, si no hay datos, tampoco existe el campo para almacenarlos. Esto significa que no debe preocuparse por los cambios realizados en un esquema existente mientras lo esté desarrollando, lo cual le permite ser más ágil en la adaptación de los requisitos de negocio en constante evolución. Esto abre la puerta a la innovación, ya que es mucho más fácil la evolución de las aplicaciones.

El modelo orientado a documento de MongoDB también disminuye la necesidad de crear uniones, ya que no se descomponen documentos normalizados en tablas más

MongoDB es especialmente destacado cuando es necesario desplegar rápidamente aplicaciones web basadas en JavaScript, que hagan uso de un gran número de contadores en tiempo real o almacenen muchas imágenes.

pequeñas. Con un menor número de uniones se logra una gran mejora en escalabilidad y velocidad. La ventaja de MongoDB, al contrario que las demás bases de datos NoSQL, es que se puede seguir utilizando uniones si se desea combinar datos de varias colecciones.

MongoDB también ofrece capacidades de sharding automático y soporte para aplicaciones geoespaciales, lo que lo hace ideal para el tipo de aplicaciones que estoy analizando aquí. En comparación con el particionamiento de los RDBMS, que se ve complicado por las varias tablas y uniones, el sharding de MongoDB se realiza con la partición del espacio de claves, ya que la clave es el ID de documento y el documento es el valor en los almacenes de documentos clave-valor. En realidad, todas las bases de datos NoSQL tienen algún tipo de sharding o particionamiento que disminuye la latencia y aumenta la escalabilidad.

Con el uso del auto-sharding y documentos BSON, MongoDB proporciona una base de datos flexible y de alta velocidad que permite adoptar un enfoque más ágil y con mayor capacidad de respuesta ante los requisitos de negocio en constante evolución. MongoDB es especialmente destacado cuando es necesario desplegar rápidamente aplicaciones web basadas en JavaScript, que hagan uso de un gran número de contadores en tiempo real o almacenen muchas imágenes. Es increíblemente rápida cuando se utiliza para satisfacer las necesidades de consulta y creación de informes en tiempo real de IoT.

Además, con su soporte geoespacial, es ideal para su uso en aplicaciones en las cuales saber dónde se encuentra el usuario o mostrar al usuario hacia dónde debe ir, es importante.

Los desarrolladores también apreciarán la formación bajo demanda que ofrece MongoDB. MongoDB proporciona soporte de desarrollo que se basa en proyecto, no en servidor, lo que es de gran ayuda tanto para desarrolladores como directores de operaciones.

## Servidor avanzado EDB Postgres

En realidad, EDB Postgres es una versión ampliada de la base de datos relacional PostgreSQL de código abierto, distribuida por Enterprise DB. Aunque EDB Postgres acostumbra a ir una versión por detrás de PostgreSQL, es una parte activa e integral de la comunidad PostgreSQL.

La velocidad y la escalabilidad son las principales ventajas que aporta PostgreSQL y EDB las complementa con un amplio conjunto de herramientas para la empresa. Tener acceso al ecosistema de la comunidad PostgreSQL es una clara ventaja, pero también surgen otras ventajas de esta base de datos escalable y de alto rendimiento.

PostgreSQL admite varios tipos de datos, incluidos los tipos de datos definidos por el usuario (como XML), colecciones de tablas y Varrays. Admite datos de texto, indexación y búsqueda, y permite ejecutar operaciones de lectura/escritura sin bloqueo mediante el empleo de un control de simultaneidad de varias versiones. Por último, los procedimientos almacenados se pueden escribir en lenguajes tales como C/C++, Java, JavaScript, Python, Perl y Ruby, lo que proporciona total libertad y flexibilidad.

EDB Postgres también ofrece funciones de seguridad empresarial, incluyendo los perfiles ampliados de contraseña. Mediante el plugin PostGIS de código abierto, es posible

implementar EDB Postgres como base de datos espacial de fondo para sistemas de información geográfica (GIS). También incluye una herramienta GUI para la creación y depuración de activadores y procedimientos almacenados.

Adicionalmente, EDB Postgres proporciona mejoras de optimización de escalado vertical y escalabilidad ampliada para los subsistemas de bloqueo, lo que aumenta el rendimiento. Hace posible la integración entre diferentes bases de datos (con soporte para bases de datos relacionales, de documentos y de clave-valor), lo que le permite combinar datos estructurados, no estructurados y transaccionales. También puede utilizarse para aplicaciones de sólo lectura en las que la alta velocidad es crítica. Permite que los DBAs puedan definir las prioridades en el consumo tanto de E/S como de CPU selectivamente entre los distintos procesos y es compatible con Oracle.

EDB ofrece una subscripción de desarrollador de Postgres que proporciona acceso directo a conocimientos técnicos de Postgres, vídeos técnicos, un gran volumen de información y una comunidad muy sólida. EDB también proporciona un conjunto de herramientas de nivel empresarial para aliviar las cargas habituales asociadas a la migración, integración y gestión.

También cabe destacar que EDB Postgres se vende como DBMS de suscripción, lo cual incluye todas las actualizaciones, mantenimiento y soporte, además del software.

Resumiendo, EDB Postgres Advanced Server es una gran solución de base de datos relacional moderna, porque es segura, escalable, flexible y rápida, especialmente cuando se ejecuta en una arquitectura de servidor optimizada. Ofrece todo lo que se espera de un RDBMS para la empresa, sin los precios de licencias asociadas a otros proveedores y con la innovación que todos apreciamos del código abierto.

## Base de datos de grafos Neo4j

Aunque al principio pueda parecerlo, las bases de datos de grafos no tienen realmente nada de mágico. Un grafo se compone de dos elementos básicos: nodos y relaciones. Cada nodo representa un fragmento de datos: un objeto, una entidad. Cada relación representa el modo en que dos nodos se relacionan entre sí. Los sitios web de redes sociales en las que los usuarios se siguen unos a otros, como Facebook o Tumblr, son ejemplos clásicos de esta idea. Los usuarios son los nodos y el "seguimiento" es la relación entre los nodos.

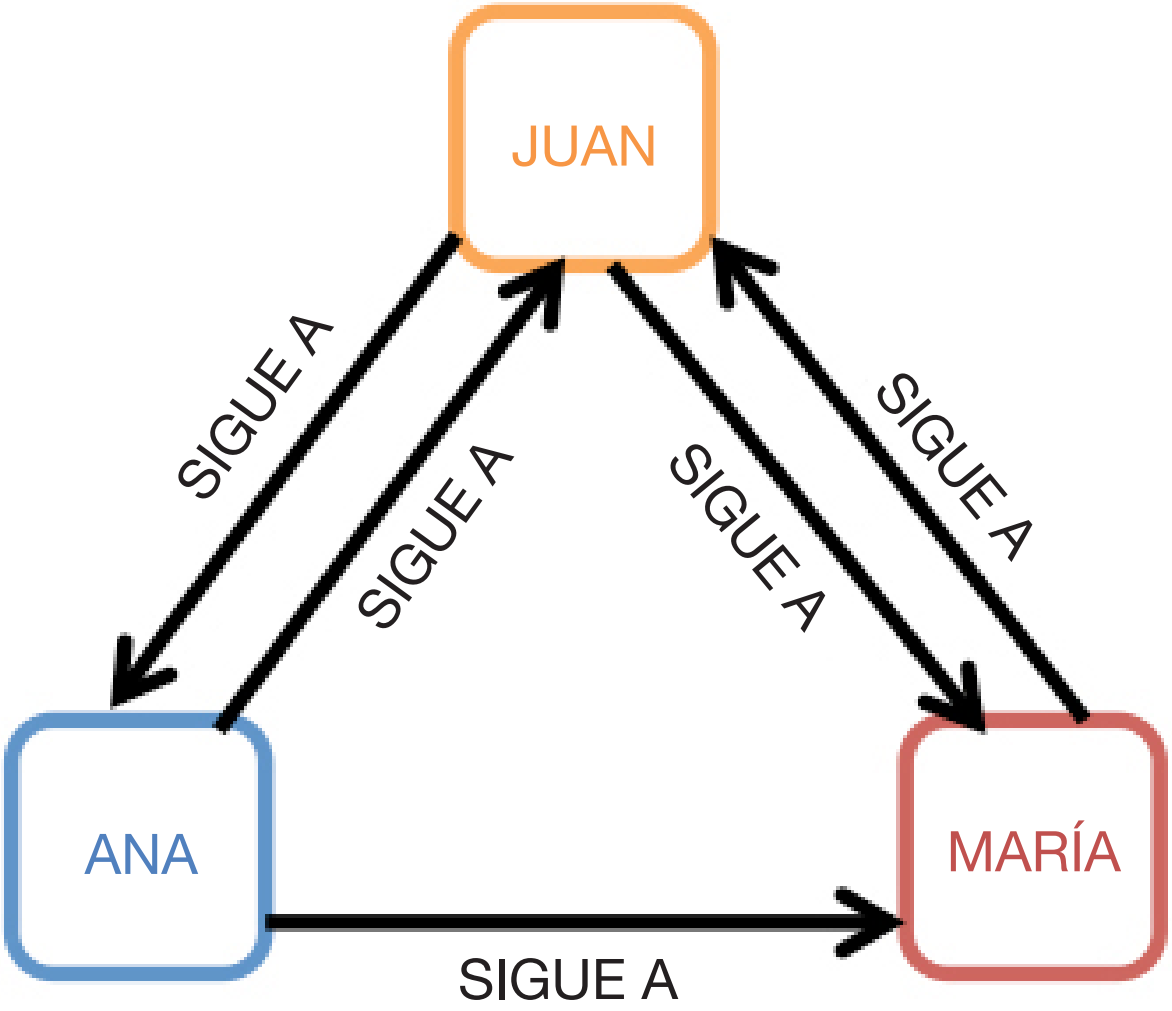



FIGURA 3. Nodos y relaciones.

En una base de datos de grafos, las relaciones tienen la máxima prioridad, lo que significa que los modelos de datos son más sencillos y más expresivos y no es necesario preocuparse de cosas tales como claves foráneas. También significa que nunca se puede tener una relación sin dos nodos y no se puede suprimir un nodo sin también suprimir sus relaciones.

Las bases de datos de grafos tienen dos características importantes a tener en cuenta para comprender las ventajas que estas bases de datos aportan al desarrollo de aplicaciones. La primera es la diferencia entre almacenamiento de grafos nativo y no nativo. Las bases de datos de grafos nativas como Neo4j se han diseñado específicamente para almacenar y gestionar grafos, al contrario que la adaptación de bases de datos relacionales u orientadas a objeto para que admitan grafos. Las bases de datos no nativas de grafos utilizan bases de datos relacionales u orientadas a objeto para el almacenamiento de los datos. Por ello, cuando el volumen de los datos y la complejidad de las consultas aumentan, una base de datos no nativa de grafos acaba sufriendo una latencia mucho mayor.

La segunda característica beneficiosa de Neo4j es su motor de procesamiento de grafos. El el procesamiento nativo de grafos, los nodos conectados apuntan directamente al otro. Esto tiene el nombre de adyacencia libre de índices y es el método más eficiente para procesar datos en una base de datos de grafos. El motor de proceso de grafos nativo de Neo4j proporciona un rendimiento constante en tiempo real ya que evita las costosas búsquedas de índice que las bases de datos no nativas deben llevar a cabo.

Estas características son muy buenas para las aplicaciones de gestión de acceso e identidad, en las que es necesario realizar un seguimiento a alta velocidad de los usuarios y sus autorizaciones, o motores de recomendaciones en tiempo real que utilizan muchas aplicaciones de personalización y productos

de comercio electrónico. Como se ha indicado anteriormente, estas características son indispensables cuando se debe realizar un análisis en tiempo real de los datos de aplicaciones sociales.

Neo4j es particularmente ágil gracias a su modelo de datos adaptable: puede responder a las necesidades de negocio que surjan con cambios en el modelo de datos sin preocuparse del posible impacto en la funcionalidad existente. Neo4j también se ha creado para ser rápido. Debido a las relaciones explícitas entre los nodos, Neo4j evita la consiguiente lentitud inevitable cuando crece el conjunto de datos.

Neo4j también proporciona un excelente soporte en línea para el desarrollador, que incluye un gran volumen de documentos, base de conocimientos, acceso a entornos de pruebas o "sandbox" y, al igual que los demás OSDBMS descritos, un sólido sistema de soporte de la comunidad.

## Sistemas de bases de datos en memoria

Los sistemas de bases de datos en memoria (IMDBS), tales como Redis y Kinetica, almacenan datos en la memoria principal, al contrario que los sistemas de bases de datos tradicionales, diseñados para almacenar los datos en soportes persistentes. Aunque técnicamente es posible poner una base de datos tradicional en RAM, sufriría la sobrecarga de un sistema diseñado para el almacenamiento en disco. Precisamente porque un IMDBS almacena los datos en memoria, evitando así la sobrecarga de operaciones de E/S y almacenamiento en caché, es incrementalmente más rápido que un DBMS tradicional. Los IMDBS también tienen unos requisitos de memoria y CPU mucho más bajos, debido a este diseño tan sencillo.

Las aplicaciones que necesitan un acceso muy rápido a los datos, o a su manipulación, son los principales candidatos para un IMDBS. Los sistemas embebidos en tiempo real, aplicaciones de mercados financieros, aplicaciones de comercio electrónico y redes sociales son excelentes candidatos para un IMDBS, ya

No es poco habitual que un IMDBS crezca por encima del rango del terabyte y conservar todas las ventajas de rendimiento en comparación con las soluciones de DBMS tradicionales.

que pueden obtener ventajas reales de su velocidad. Un IMDBS no solamente es rápido, sino que también se escala muy bien. No es poco habitual que un IMDBS crezca por encima del rango del terabyte y conservar todas las ventajas de rendimiento en comparación con las soluciones de DBMS tradicionales.

**Redis** Una base de datos de clave-valor, o almacén, es una base de datos diseñada para almacenar, recuperar y gestionar matrices asociativas. Una matriz asociativa es un modelo de datos sencillo en el que a cada clave se le asocia un único valor en una colección, una relación conocida como par de clave-valor. Una serie arbitraria, por ejemplo, un nombre de archivo, hash o URI, representa la clave de cada uno de estos pares de clave-valor. El valor, que se almacena como blob, puede ser cualquier tipo de datos, como por ejemplo una imagen o un documento. Puesto que el valor se almacena como blob, no requiere ninguna definición inicial de esquema o modelo de datos. Esto también elimina la necesidad de indexar los datos para aumentar el rendimiento. No es posible filtrar o controlar lo que se devuelve de una petición en función del valor, ya que el valor es opaco.

Los almacenes de clave-valor utilizan mandatos get, put y delete en lugar de un lenguaje de consulta, lo que significa que el camino para recuperar datos es una petición directa al objeto en memoria. La relación entre los datos no se calcula, por lo que no existe sobrecarga de optimización. No es necesario preocuparse del almacenamiento de los índices, la velocidad

de red o el equilibrio en un sistema distribuido. Debido a su sencillez, el almacén de clave-valor es muy rápido y flexible, sencillo de utilizar, muy escalable y portátil. Redis es un almacén de estructura de datos de clave-valor, en memoria y de código abierto (licencia BSD) que se utiliza como base de datos, caché e intermediario de mensajes.

Una de las ventajas de utilizar Redis proviene del uso de primitivas comunes de Redis, tales como LPUSH, LTRIM y LREM. El uso de estas primitivas permite realizar tareas que son difíciles y lentas con los almacenes de datos tradicionales para llevarlas a cabo de una forma mucho más fácil. Por ejemplo, en una aplicación web, se pueden eliminar los artículos suprimidos de la caché por medio de LREM o se puede utilizar LPUSH para insertar un ID de contenido en la cabecera de la lista almacenado en una clave para mostrar la lista de elementos más recientes en una página de inicio, y se puede utilizar LTRIM para limitar el número de elementos de la lista. Con estas primitivas sencillas, Redis facilita mucho el trabajo de un desarrollador.

Debido a su sencillez, velocidad y baja latencia, Redis también es una gran solución para el desarrollo de aplicaciones de comercio electrónico, en las que interesa almacenar de forma eficiente preferencias y perfiles de usuarios, por ejemplo, para recomendar productos en función de lo que los usuarios estén visualizando o presentar anuncios y cupones en tiempo real, individualizados según los hábitos de compra de un cliente. Puesto que todos los datos están en memoria, se eliminan los retrasos para encontrar dichos datos, lo que se traduce en un rendimiento extremadamente veloz. Por medio de Redis como caché delante de otra base de datos, por ejemplo, también se logra un gran aumento en la velocidad.

Personalmente, también aprecio el soporte en línea disponible para Redis. Incluye una lista completa de mandatos formando parte de una guía de programación detallada, junto con varios tutoriales, guías administrativas y otros

recursos para el desarrollador. Para obtener un conocimiento completo y detallado de las grandes ventajas que ofrece Redis, visite https://redis.io.

**Aceleración GPU con Kinetica** Repetimos, Kinetica no es una base de datos de código abierto, pero forma parte del ecosistema de hardware y software abiertos de la OpenPOWER Foundation. Kinetica es una base de datos en memoria y distribuida, acelerada mediante unidades de proceso gráfico (GPU). Un GPU es simplemente un circuito diseñado para acelerar la creación de imágenes para su visualización, alterando y manipulando rápidamente la memoria. Donde una CPU tiene muchos núcleos y mucha memoria en caché, una GPU tiene miles de núcleos, lo que se traduce en aumentos de velocidad de más de 100 veces que con una CPU, en algunos casos. Puesto que son ideales para tomar grandes volúmenes de datos y realizar la misma operación una y otra vez, las GPUs se pensaron originalmente para la representación de juegos 3D. Más recientemente, las GPUs se han utilizado para acelerar cargas de trabajo de computación en funciones tales como el modelado financiero, investigación, exploración energética e inteligencia artificial. De hecho y debido a la arquitectura de procesamiento paralelo que genera velocidades de proceso hasta 100 veces más rápidas, Kinetica es ideal para analizar grandes streamings de datos. Esto lo hace perfecto para el análisis predictivo que definen las cargas de trabajo de IA.

Kinetica utiliza la potencia de proceso de la GPU para gestionar grandes conjuntos de datos, particularmente el streaming de datos, en un tiempo mucho menor y con una ocupación de hardware mucho más pequeña que las bases de datos tradicionales. Esto es especialmente útil en aplicaciones IoT y visualización geoespacial. Incluye herramientas de visualización capaces de representar volúmenes muy grandes de datos y no existe la necesidad de preparar el esquema

antes de poder analizar los datos. Kinetica es una gran herramienta complementaria para sistemas transaccionales, data warehouses y data lakes, y puesto que es totalmente compatible con SQL, es fácil de consultar. También admite REST, JSON, Java, JavaScript, C++ y Python, entre otros, lo que la hace muy intuitiva para el desarrollador. Esto implica que puede explorar grandes conjuntos de datos con mayor rapidez que antes sin tener que aprender un nuevo lenguaje de consulta o programación o construir nuevos modelos de datos.

Hasta este momento hemos analizado varias ofertas de bases de datos del paisaje OSDBMS. En la categoría no relacional se encuentran MongoDB, con su modelo de datos orientado a documento; Neo4j, con su modelo de grafos; y Redis, que ofrece un método de clave-valor. En el campo relacional se encuentran EDB Postgres, una excelente base de datos para la analítica, y una opción en memoria, Kinetica. La ventaja que tienen en común todos estos sistemas de gestión de bases de datos es la velocidad, que es crítica cuando hablamos del desarrollo de aplicaciones analíticas de Big Data.

Ahora veamos, por último, una plataforma ideal para alojar estos OSDBMS y las aplicaciones construidas sobre ellos: servidores OpenPOWER LC diseñados desde cero para Big Data por IBM y sus socios de la OpenPOWER.

## Por qué la implementación de un OSDBMS en sistemas OpenPOWER de IBM es el enfoque adecuado

Para utilizar las capacidades de los OSDBMS con el fin de crear soluciones realmente innovadoras no solamente se necesita velocidad de proceso, sino también colaboración. Al principio de este libro electrónico he mencionado que la OpenPOWER Foundation, un consorcio de más de 250 miembros que incluye a algunos de los nombres más importantes en tecnología: IBM, Google, NVIDIA, Mellanox Technologies, Tyan, Xilinx y

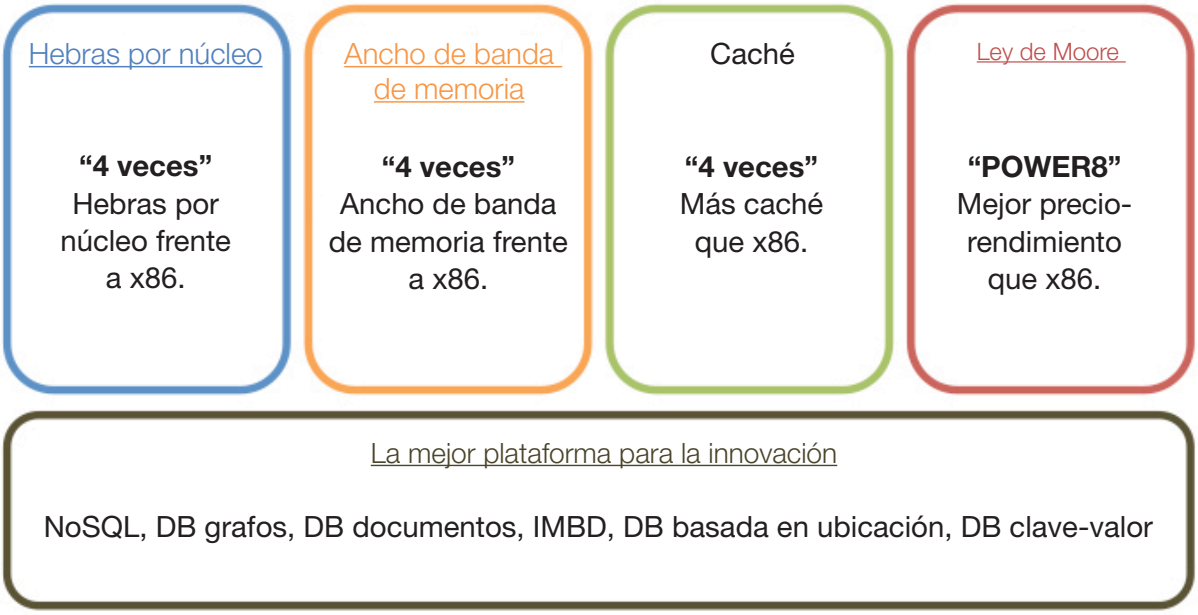



Figura 4. IBM POWER8 se ha creado para OSDBMS.

Canonical. La OpenPOWER Foundation ha estado colaborando desde hace varios años en diseños de sistemas basados en la arquitectura de procesador POWER de IBM. Encontrará una de las últimas manifestaciones comercialmente disponibles de dicha colaboración en los servidores OpenPOWER LC de IBM.

Los servidores OpenPOWER LC de IBM con tecnología de procesador POWER8 se han diseñado para cargas de trabajo de Big Data, incluyendo las soluciones OSDBMS que hemos analizado aquí. IBM POWER8 ofrece 4 veces más de caché de procesador, ancho de banda de memoria y multihebra que las plataformas comunes pueden proporcionar.

IBM POWER8 ejecuta Linux estándar de la industria de Red Hat, SUSE y Canonical. Esto hace que el paso de aplicaciones Linux x86 a Power sea más atractivo y sencillo que nunca. Linux en Power proporciona la plataforma innovadora que los desarrolladores realmente necesitan para aprovechar la potencia y escala de los OSDBMS para aplicaciones con un gran uso de datos.

El diseño del POWER8 combina potencia de cálculo, ancho de banda de memoria rendimiento de E/S para generar la velocidad necesaria para cargas de trabajo Big Data y analíticas. POWER8 se ha diseñado para ofrecer 4 veces más de hebras por núcleo que la infraestructura tradicional, 4 veces más de ancho de banda de memoria y mayor capacidad de memoria que la infraestructura tradicional, con sistemas scale-out que pueden ofrecer hasta 2 TB en un servidor de 2 zócalos y toda la gama hasta los 16 TB en servidores scale-up para la empresa. POWER8 también proporciona 4 veces más de memoria caché por procesador con una menor latencia, lo que le permite procesar más datos con más rapidez.

**MongoDB** Para MongoDB es ideal, ya que se obtiene una plataforma que ofrece una visión integrada y en tiempo real de todos los datos. Según IBM, MongoDB en POWER8 ofrece un rendimiento por servidor un 40% mejor que Intel Xeon. Esta es una excelente solución para la expansión de servidores en el centro de datos y, si también se toma en consideración los costes de despliegue, MongoDB en POWER8 ofrece el doble de rendimiento por euro comparado con los sistemas basados en x86; una gran noticia si se busca ahorrar dinero para una futura innovación.

**Servidor avanzado EDB Postgres** El Servidor avanzado de EDB Postgres también se ejecuta en Linux little endian en POWER8, lo cual elimina los problemas de portabilidad. La ejecución del servidor avanzado de EDB Postgres en servidores OpenPOWER LC de IBM proporciona multihebra de alto rendimiento, más memoria caché, mayor ancho de banda para los datos y una relación precio/rendimiento aproximadamente dos veces mejor que en los sistemas basados en x86. Los benchmarks de IBM demuestran que los servidores OpenPOWER LC logran un rendimiento por núcleo un 60% mejor que Intel Xeon. Repetimos, esta es la oportunidad perfecta para que su empresa despliegue más

Mediante esta solución, que funciona con cualquier cliente Redis sin cambios en la API estándar de Redis, un único servidor POWER8 con aceleración CAPI-Flash puede procesar más de 200K operaciones/segundo con una latencia inferior al milisegundo: esto es rapidez.

cargas de trabajo en menos y gaste menos en despliegues de infraestructura para las aplicaciones de Big Data, dejando más recursos para la innovación.

**Redis** Redis Labs, que también es miembro de la OpenPOWER Foundation, e IBM Power Systems colaboran estrechamente para ofrecer una solución Redis optimizada para POWER8 y su interfaz de procesador acelerador coherente (CAPI) como parte del soporte de Redis para el IBM Data Engine for NoSQL, que ejecuta Redis en IBM 840 Flash System, la tarjeta IBM CAPI-Flash y Redis Labs Enterprise Cluster (RLEC) para software Flash como sustituto de la RAM. Mediante esta solución, que funciona con cualquier cliente Redis sin cambios en la API estándar de Redis, un único servidor POWER8 con aceleración CAPI-Flash puede procesar más de 200K operaciones/segundo con una latencia inferior al milisegundo: esto es rapidez. También puede almacenar el 90% de un conjunto de datos de varios terabytes en Flash y solamente el 10% en RAM. Cuando se compara con una solución Redis pura basada en RAM, esto ayuda a disminuir los costes de despliegue en más de un 70%. Las pruebas de Redis han demostrado que el rendimiento de los servidores IBM OpenPOWER LC es un 67% mejor que el de x86.

**Neo4j** Neo4j en servidores OpenPOWER LC de IBM proporciona la plataforma de base de datos de grafos más escalable del mundo, capaz de almacenar y procesar grafos increíblemente grandes. Uno de los principales retos del procesamiento de grafos a escala es el modo en que se gestiona el tamaño del conjunto de datos sin comprometer las capacidades en tiempo real. Con los 56 TB de memoria ampliada disponible en el servidor LC por medio de CAPI y Flash tal como se ha descrito anteriormente, se aumenta el tamaño de las consultas en tiempo real, puesto que el tamaño de los grafos que se pueden almacenar en memoria también ha aumentado. Pero la línea de servidores LC con POWER8 también está equilibrada. Cada núcleo puede gestionar 8 hebras de hardware al mismo tiempo, con un total de 96 hebras simultáneas en un chip de 12 núcleos. Los controladores de memoria en chip permiten tener un gran ancho de banda en la memoria y la E/S de sistema. Con la aceleración CAPI activada en un servidor IBM OpenPOWER LC, esta combinación ofrece 2 veces el rendimiento de Intel Xeon.

**Kinetica** El logro del mejor rendimiento de Kinetica depende realmente de la rapidez con que se muevan los datos entre la CPU y la GPU, ya que Kinetica se ha diseñado para utilizar la memoria del sistema. La nueva tecnología NVIDIA NVLink e IBM POWER8 proporcionan el enfoque más avanzado y asequible para ofrecer analítica de alto rendimiento con Kinetica. Las interconexiones, tales como NVIDIA NVLink, abren un camino más amplio entre la CPU y la GPU, que permite que Kinetica aproveche totalmente la memoria del sistema. El procesamiento GPU ya no está limitado a la velocidad con que se pueden mover los datos a través del subsistema de E/S, lo cual abre la posibilidad de que Kinetica acceda a conjuntos de datos más grandes.

IBM ha demostrado que Kinetica puede tener un rendimiento ejecutándose en un servidor IBM OpenPOWER LC con NVLink hasta 2,5 veces superior al de un sistema x86 similar.

## Conclusión

Existe ahí fuera un mundo amplio y complejo que genera datos grandes y complejos. Las bases de datos, enfoques de desarrollo y plataformas de procesamiento tradicionales no van a proporcionar el rendimiento necesario para satisfacer la demanda actual de capacidades de datos y analítica en tiempo real.

El OSDBMS ofrece varias ventajas a los desarrolladores innovadores: rendimiento y flexibilidad para el diluvio actual de datos diversos y, quizás más importante aún, las ventajas de los modelos de desarrollo abiertos, incluido el acceso a una vibrante comunidad que se adapta al cambio y a los problemas del mundo real de una forma rápida y eficaz. No solamente encontramos capacidades aceleradas y de respuesta en los OSDBMS, sino que las herramientas de nivel empresarial y la liberación de las restricciones de las soluciones propietarias hacen que los OSDBMS se puedan utilizar más y sean más accesibles.

A medida que la demanda ejercida en los desarrolladores de aplicaciones siga evolucionando, las mismas aplicaciones deben seguir innovándose para dar respuesta a dichas demandas. El abrir abierto de los OSDBMS proporciona el entorno necesario para que las ideas y la innovación fluyan hasta soluciones reales y que se puedan utilizar. Mediante una colaboración estrecha, IBM y los líderes del espacio de los OSDBMS proporcionan plataformas líderes del mercado para sacar el máximo partido de estas nuevas tecnologías de base de datos.

Descubra cómo puede aprovechar las ventajas que ofrece el desarrollo de sus aplicaciones con un OSDBMS en POWER siguiendo este enlace: https://www-01.ibm.com/marketing/iwm/dre/signup?source=mrs-form-12148&S_PKG=ov53321.■