Tratamiento de Grandes Volúmenes de Datos en Ciudades Inteligentes
Una Propuesta de Big Data con NoSQL

Sonia Formia, Luis Vivas, Mauro Cambarieri, Nicolás García Martínez, Horacio Muñoz Abatte, Marcelo Petroff
Laboratorio de Informática Aplicada (LIA)
Licenciatura en Sistemas, UNRN – Sede Atlántica
{ sformia, lvivas, mcambarieri, ngarciam, hmunoz, mpetroff}@unrn.edu.ar

Resumen. El despliegue de sensores, la participación de los ciudadanos en las redes sociales y la generación de contenidos, entre otras razones, han producido un aumento de la información disponible en las ciudades, dando origen a un fenómeno conocido como BIG DATA. La reducción de los costos de almacenamiento y la implementación de tecnologías que antes no existían, permite almacenar cantidades muy grandes de información, con gran variedad de formatos, en tiempos cada vez mas cortos. El volumen y la diversidad involucrados han generado el surgimiento de nuevas tecnologías de bases de datos, las que comúnmente se resumen bajo el término NoSQL. Este, engloba hoy en día muchas bases de datos diferentes que ofrecen esquemas mas flexibles que el tradicional entidad-relación y trabajan mejor para manipular grandes cantidades de datos de manera eficiente para el procesamiento analítico, aunque no siempre proveen un lenguaje de consulta declarativo [1], por lo que se requiere de mayor experticia para manipular los datos. Se hace necesario definir los alcances de aplicabilidad del movimiento NoSQL al procesamiento analítico de información desde los sistemas de Inteligencia de Negocios públicos, (*Business Intelligence, BI*) tradicionales hasta el Big Data.

Palabras clave: Big Data Analytics. Bases de Datos NoSQL, Data Science. Business Intelligence. Smart Citie.

1 Introducción

Se suele relacionar Big Data con la necesidad que tienen las organizaciones, tanto públicas como privadas, de aprovechar la información disponible para tomar mejores decisiones, mejorar sus tiempos de respuesta, conocer a sus ciudadanos o hacer más eficiente el gasto público. Introduciendo mejoras en las administraciones públicas, con el objetivo de generar Estados más eficientes y mejorar la calidad de vida de las personas. También Big Data se relaciona con la necesidad de generar ciudades más eficientes, en donde se pueda prever cuál es el movimiento del tráfico, cuánta energía se va a consumir en un determinado momento, cuáles son los fenómenos meteorológicos con mayor porcentaje de ocurrencia, cuáles son las necesidades de los

ciudadanos y hasta cómo es el rendimiento de los alumnos de las instituciones públicas [2].

Si bien no es algo nuevo, su alcance ha ido variando con el tiempo, es por esto que se hace necesario definir de manera consistente lo que hoy se denomina Big Data, clasificar las diferentes tecnologías que ofrece el mercado para luego centrarse en el estudio e implementación de las que ofrezcan mejores soluciones a las necesidades de almacenamiento y análisis de datos, en particular para los desarrollos para ciudades inteligentes.

Big Data es el nombre que se le da a conjuntos de información que crecen de una manera tan exponencial que resulta prohibitivo almacenarlos y/o procesarlos con métodos o técnicas tradicionales del mundo de base de datos relacionales. La cantidad de información disponible exige recurrir a nuevas herramientas y procesos para recopilar datos (tanto estructurados como no estructurados) y para almacenar, manipular, administrar, integrar y analizar dichos datos.

Cuando se piensa en bases de datos relacionales acuden a la mente los mismos nombres, algunos comerciales y otros de software libre. Aunque cada una tiene sus peculiaridades, no es difícil elegir entre un sistema y otro. La decisión de cuál elegir, se basará en sus características y precio.

Al hablar de bases de datos NoSQL (*not only SQL*, no sólo SQL) la elección se complica. Al día de hoy existen muchos sistemas de bases de datos NoSQL, elegir uno de ellos puede ser muy difícil, ya que ninguno ha obtenido todavía la fama, estabilidad y estandarización que sí han conseguido las bases de datos relacionales. Por otro lado, la taxonomía de bases de datos no relacionales no es clara, dependiendo de los autores se pueden encontrar bases de datos analíticas (*Analytic Data Stores*) dentro de las clasificaciones de NoSQL o fuera de ellas, como un tipo diferente de base de datos [3]. Si bien puede pensarse que solo son etiquetas, es importante determinar al menos funcionalidades básicas, aunque no se las pueda catalogar taxativamente.

Este trabajo pone énfasis en el estudio de las bases de datos NoSQL desde el punto de vista del análisis de datos. Se plantea la hipótesis de la utilización de tecnologías NoSQL desde los comienzos de un proyecto de BI, independientemente de la cantidad de datos y de la naturaleza (estructurada o no estructurada) de los mismos. En este sentido se espera encontrar motivos para fomentar el uso de bases de datos NoSQL en aplicaciones de data warehouse (DW) y data mining tradicionales, es decir, sobre datos estructurados y sobre volúmenes de información que hoy en día no se consideran Big Data.

2 Contexto y Objetivos

El presente trabajo se encuentra enmarcado en el proyecto: Estudio y evaluación de tecnologías de la información y la comunicación para el desarrollo de ciudades inteligentes en Río Negro, del Laboratorio de Informática Aplicada (*LIA*) de la Universidad Nacional de Río Negro (*UNRN*).

El objetivo general de la investigación es estudiar, identificar y probar tecnologías disponibles para Big Data y determinar su aplicabilidad a la información producida por los desarrollos para las ciudades inteligentes.

El objetivo particular de este tramo es encontrar tecnologías alternativas, dentro del movimiento NoSQL que puedan ser utilizadas en pruebas de concepto para tareas de BI en entornos estructurados y para volúmenes limitados de datos, de manera de proveer al equipo de trabajo con los conocimientos necesarios para encarar un proyecto de Big Data utilizando las mismas herramientas. En el proceso se espera comparar las implementaciones tradicionales sobre bases de datos relacionales con las de bases de datos NoSQL en el contexto de problemas ya conocidos.

3 Big Data

Big data se aplica a la captura, gestión y procesamiento de conjuntos de datos que superan las capacidades del software habitual. Los tamaños de Big data están cambiando constantemente. Es importante porque es un concepto que engloba y se relaciona con otras tendencias como cloud computing, movilidad, Internet de las cosas, ciberseguridad, analytics, etc. [4].

La adopción de Big Data parece ser un hecho que tarde o temprano deben realizar las organizaciones [5]. Es de esperar que las oportunidades que esto brinde compensen la demanda económica y de especialistas requeridos para el análisis de estos grandes volúmenes de datos. Aparece, entonces, el perfil del científico de datos, que podría definirse como la evolución del analista de datos o de negocios en el contexto de Big Data. El científico de datos es una persona con habilidades diversas: ciencias de la computación, analítica, matemáticas, generación de modelos y estadística, además de buen comunicador.

Esas nuevas necesidades han llevado a nuevos requerimientos de análisis de la información, lo que anteriormente se conocía como BI actualmente requiere de un nuevo modelo de análisis, Big Analytics, que trabaje con los datos al nivel más bajo de granularidad disponible con modelos más ágiles que los actuales de BI, que permitan hacer análisis de manera escalable, en tiempo real y que integren datos no estructurados con facilidad. Las “tres V”: variedad, volumen y velocidad, son los temas clave.

4 Bases de Datos NoSQL

El movimiento NoSQL se inició hace algunos años impulsando el uso de nuevos sistemas de bases de datos que no se basan en el modelo entidad-relación y no responden solamente a consultas en lenguaje estándar SQL, y que vendrían a instalar una alternativa a ciertas limitaciones en flexibilidad, procesamiento y escalabilidad que enfrentan los sistemas tradicionales para abordar el camino a Big Data. Estas bases de datos no requieren esquemas de tablas fijas y no soportan operaciones de *join*. Están optimizadas para operaciones principalmente de lectura escalables sin tanto énfasis en la consistencia.

Existen numerosas discusiones sobre las ventajas y desventajas de las bases de datos NoSQL comparadas con las bases de datos relacionales [3][6][7]. Entre las desventajas se pueden mencionar:

- No existe un líder en el mercado NoSQL.
- No existen standard (como SQL para los RDBMS).
- Muy alta especialización, lo que dificulta elegir una única base de datos para todas las necesidades.
- La escalabilidad (que es promocionada como una ventaja) no es sencilla de conseguir en la práctica.
- Dificultades para proveer consistencia y disponibilidad absolutas.
- Cambio en los modelos de datos: EAV- modelo Entidad-Atributo-Valor vs. Modelo Entidad-Relación.

Entre las ventajas están las siguientes:

- Flexibilidad en el cumplimiento de ACID (atomicidad - ***A****tomicity*, consistencia - ***C****onsistency*, aislamiento - ***I****solation* y durabilidad- ***D****urability*).
- Rendimiento mayor, incluso de hasta varios órdenes de magnitud [6].
- Escalabilidad horizontal (empleo de hardware más económico).
- Mejor y mayor manejo de información no estructurada.
- Evita el mapeo objeto-relacional [3].
- Solución a escenarios que no cubren las bases de datos relacionales, en cuanto a volumen de datos y a su procesamiento en tiempo real.
- Los modelos de datos centralizados no pueden ser distribuidos fácilmente (DDB). Las bases de datos NoSQL se construyen desde el inicio con la escalabilidad en mente.

5 Tipos de Bases de Datos NoSQL

Existen varias aproximaciones diferentes para clasificar las bases de datos NoSQL [8], una de ellas, basada en la evolución de la arquitectura de las mismas, arroja los siguientes tipos [7] (aunque pueden encontrarse productos con características de más de un tipo):

- **Clave – valor**: Son las más sencillas de entender, guardan tuplas que contienen una clave y su valor. Este modelo favorece la escalabilidad sobre la consistencia, y omite/limita las funcionalidades analíticas y de consultas complejas ad-hoc. Es conceptualmente similar a una tabla hash y es muy adecuado para almacenar datos no estructurados dada su flexibilidad y velocidad. Las bases de datos NoSQL clave-valor más populares son: Riak, Redis, Amazon DynamoDB, Voldemort, Membase, Dynamite, Tokio Cabinet, Cloudant y Cassandra (si bien esta última también tiene propiedades de bases de datos orientadas a columnas).
- **Orientadas a columnas**: Este tipo de bases de datos están pensadas para realizar consultas y agregaciones sobre grandes cantidades de datos. Funcionan de forma parecida a las bases de datos relacionales, pero almacenando columnas de datos en lugar de registros. Muchas de estas bases de datos están inspiradas en la tecnología BigTable de Google, que consiste en

un sistema de almacenamiento distribuido para manipulación de datos estructurados, diseñado para escalar a grandes tamaños (petabytes). Es un mapa ordenado multidimensional, persistente, distribuido y disperso, indexado por una clave fila, una clave columna y tiempo. Cada valor del mapa es una colección (*array*) ininterrumpida de bytes [5]. Algunos ejemplos son: Google BigTable, Apache HBase (proyecto Hadoop), Hypertable, Infobright y Cassandra (híbrido clave-valor). Riak también se encuentra en algunas clasificaciones dentro de este tipo [3].

- **Orientadas a documentos**: Son aquellas que gestionan datos semi-estructurados, como documentos. Estos datos son almacenados en algún formato estándar como puede ser XML, JSON o BSON. Son las bases de datos NoSQL más versátiles, se pueden utilizar en gran cantidad de proyectos. Guardan la información como un listado de documentos desestructurados. Al acceder a un documento se puede ingresar en un número no especificado de campos con sus respectivos valores. Son ejemplos de este tipo de bases de datos MongoDB y CouchDB.
- **En grafo**: Basadas en la teoría de grafos, utilizan nodos y aristas para representar los datos almacenados. Son muy útiles para guardar información en modelos con muchas relaciones, como redes y conexiones sociales. Permiten contar con un modelo de negocio más complejo, con flexibilidad en las relaciones entre entidades. Por definición, una base de datos orientada a grafos es cualquier sistema de almacenamiento que provea libre indexado por adyacencia, esto significa que cada elemento contiene un puntero directo a su elemento adyacente y no requiere búsqueda por índices [8]. Algunas bases de datos en grafo son: Neo4j, HyperGraphDB, AllegroGraph y VertexDB.

Otros autores [10] excluyen de la categoría NoSQL las bases de datos orientadas a columnas y generan una clasificación diferente que incluye tecnologías como bases de datos en memoria [5] y Hadoop, como se describe a continuación:

- **NoSQL**: incluye las documentales, grafo y clave-valor.
- **Analíticas**: diseñadas para ser usadas como motores de Data Warehouse. Incluyen columnares, MPP (*massively parallel processing* – procesamiento masivamente paralelo) y en memoria (In-Memory Analytics)
- **Hadoop**: (Almacenamiento HDFS + procesamiento Map/Reduce). Ideales para grandes volúmenes de datos que no cuadran en bases de datos transaccionales o NoSQL (emails, tweets, imágenes, logs, etc.). Arquitectura masivamente paralela, que incluye un sistema de archivos y un esquema de procesamiento distribuido.

5.1 Teorema CAP.

Otra forma de clasificar las bases de datos NoSQL deriva del teorema CAP para sistemas distribuidos, enunciado por Eric Brewer [9]. El teorema establece que es imposible para un sistema distribuido garantizar simultáneamente:

- Consistencia *(**C**onsistency*)
- Disponibilidad (***A**vailability*)
- Tolerancia a fallos (***P**artition Tolerance*)

Solo dos de las tres son posibles. Brewer señala que así como las propiedades ACID de los sistemas de bases de datos relacionales proveen consistencia, las propiedades BASE proveen disponibilidad y son las siguientes:

- Basicamente disponible (***B**asically **A**vailable*)
- Estado flexible (***S**oft-state*)
- Eventualmente consistente (***E**ventual consistency*)

Mientras ACID es pesimista y fuerza la consistencia al final de cada operación, BASE es optimista y acepta consistencia flexible en la base de datos para asegurar disponibilidad. Conociendo estas restricciones, se sugiere utilizar como criterio de selección los requerimientos que se consideren más críticos para el negocio. En caso de optar por las propiedades BASE, se deberá elegir la base de datos que provea el par de propiedades que son requerimiento para el problema.

6 Elección de un tipo de base de datos NoSQL para BI.

Como se puede apreciar en los apartados anteriores, el movimiento NoSQL y las herramientas para Big Data se han convertido en un gran universo de opciones. Varios autores, algunos citados en este trabajo, han intentado la tarea de clasificarlas y organizarlas. De la revisión bibliográfica realizada surge claramente que no se ha conseguido aún el consenso, aún así las clasificaciones descriptas colaboran en fragmentar el universo de opciones para permitir encarar el estudio de los productos que, a primera vista, pueden resultar útiles para el objetivo propuesto.

Teniendo en cuenta la amplitud del problema planteado y la necesidad de enfocar el análisis en productos específicos que provean factibilidad de implementación, se realiza una primera aproximación a la selección.

En primer lugar se hace necesario considerar el entorno de aplicación donde se realizarán las pruebas que permitan comparar el desempeño de la base de datos seleccionada versus un ambiente analítico ya existente. Para ello se toma la decisión de definir la prueba sobre un DW operativo existente en el LIA: el Sistema de Información Gerencial SIU-Wichi. Este DW maneja datos financieros, administrativos, académicos y de personal de la UNRN. SIU-Wichi utiliza la base de datos relacional PostgreSQL y la plataforma analítica Pentaho BI Server.

Una vez decidido el entorno que se utilizará para las pruebas, analizando las diferentes clasificaciones de bases de datos para Big Data, la elección se inclina hacia las bases de datos analíticas, justamente por el tipo de aplicación, y entre ellas, las orientadas a columnas. Estas bases de datos son de las más conflictivas en cuanto a su lugar en las clasificaciones, dado que pueden o no ser incluidas dentro de las bases de datos NoSQL, pero también son las que conceptualmente más se acercan al paradigma relacional, y más se asemejan al modelo multidimensional utilizado en el diseño de DW (ver Fig.1).

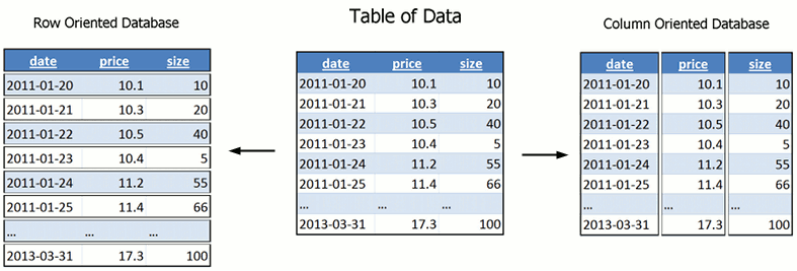

Fig. 1. Base de datos orientada a columna vs. orientada a fila [11].

El almacenamiento en filas sirve cuando todas las columnas son necesarias, ideal para un mundo transaccional donde usualmente se necesita todo el detalle de una entidad. En cambio, el almacenamiento en columnas sirve cuando solo se requieren algunas columnas para el análisis, cuando lo que se busca es información consolidada (sumas, cantidades, promedios), ideal para un mundo analítico donde la información se concentra en la métrica de distintas entidades.

Dentro de las bases de datos orientadas a columnas, se opta por iniciar las pruebas sobre Infobright. Se enumeran aquí las razones de tal decisión [10]:

- ✓ Es un motor de base de datos analítico orientado a columnas de alta performance que entrega rápidos tiempos de respuesta a consultas ad-hoc sobre Big Data con mínimo esfuerzo de administración y mantenimiento.
- ✓ La empresa provee un modelo de negocios "Try & Buy" basado en una versión Enterprise (Infobright Enterprise Edition, IEE) y una versión Open Source (Infobright Community Edition, ICE).
- ✓ Es socio tecnológico de varias empresas de BI, incluyendo Pentaho (plataforma de SIU-Wichi).
- ✓ Su base de clientes incluye empresas como Yahoo!, Xerox, Bwin, etc.
- ✓ A diferencia de otras bases analíticas, la mejor performance de Infobright está basada en modelos matemáticos, no en hardware.
- ✓ Se basa en la arquitectura MySQL (base de datos relacional muy utilizada actualmente). Por estar implantada sobre MySQL acepta SQL como lenguaje de consulta.

Dentro de las características distintivas de Infobright que se tomaron en cuenta para la decisión, surge como dato interesante que es una base de datos clasificada como NoSQL y sí acepta SQL. De los apartados anteriores se puede comprender que esto no es una contradicción, sino una corroboración que el movimiento NoSQL no significa "No SQL", sino "Not Only SQL" (no solamente SQL), y que si se está pensando en iniciar el camino hacia bases de datos que acerquen el BI tradicional al inminente Big Analytics, es quizá una buena opción hacer los cambios de manera gradual.

7 Arquitectura de Infobright.

Las principales componentes de la arquitectura de Infobright son [12]:

- **Data packs - data compression** (paquetes de datos y compresión de datos). Los datos se almacenan en data packs de medida fija que incluyen una cantidad determinada de valores de cada columna. Cada data pack se comprime individualmente utilizando el algoritmo de compresión optimal para esos datos, lo que resulta en tasas de compresión típicas de 10:1 hasta 40:1.
- **Knowledge Grid** (cuadrícula de conocimiento). Es una estructura en memoria que crea y almacena automáticamente información acerca de los datos al momento de carga y cuando se ejecutan consultas. Incluye información de la tabla completa y de cada *data pack* individual (*Knowledge nodes*).
- **Granular Computing Engine** (motor de computación granular). Procesa las consultas usando el *Knowledge Grid*, reduciendo o eliminando la cantidad de datos que deben descomprimirse para responder la consulta.
- **High-Speed Loading** (carga de alta velocidad). Varias opciones de módulos de carga de datos que hacen disponible los datos generados por máquinas en tiempos cercanos al tiempo real. Productos adicionales incluyen conectores para Hadoop que simplifican la extracción de datos desde HDFS.
- **DomainExpert** (Experto en el Dominio). Extensión a la inteligencia del *Knowledge Grid* que agrega información acerca de un dominio en particular (por ej. Web, servicios financieros, etc.). Contribuye a optimizar datos generados por máquina.
- **Rough Query** (consulta aproximada) Puede acelerar las consultas analíticas en un factor 20:1 sobre grandes volúmenes de datos. Permite que el usuario vaya reduciendo los resultados de manera iterativa antes de ejecutar la consulta completa.
- **Built on MySQL** Infobright está construido dentro de la arquitectura de MySQL. Esta integración le permite vincularse fácilmente con cualquier herramienta de ETL (*Extraction Transformation and Loading* – extracción, transformación y carga) y BI compatible con MySQL.

El motor (*granular computing engine*) resuelve las consultas consultando el *knowledge grid* y los *knowledge nodes*. La consulta puede ser resuelta directamente o identificar solamente los *data packs* requeridos, minimizando la descompresión.

9 Conclusiones y Futura Línea de Trabajo

Este trabajo presenta un análisis de las herramientas existentes para Big Data, en particular las orientadas a bases de datos NoSQL. Se revisan las diferentes clasificaciones de las bases de datos alternativas a las relacionales, poniendo el foco en aquellas que se adapten a las tareas de BI y que puedan ser utilizadas en todo el camino de una organización hacia el análisis de sus datos, desde los DW tradicionales a las aplicaciones de Big Analytics.

Se selecciona una categoría de bases de datos NoSQL, las orientadas a columnas, y se analizan sus características confrontándolas con las que se consideran necesarias para el objetivo propuesto.

En base a la categoría seleccionada, a la definición de un entorno de prueba factible y a las herramientas de BI existentes en el ámbito de la investigación, se elige

un producto específico de bases de datos orientadas a columnas: Infobright Community Edition, se estudia su arquitectura y funcionamiento y se lo implementa en el ambiente de trabajo, para la realización de pruebas de adaptabilidad y performance.

El trabajo deja planteado un camino para la inserción del BI tradicional en las nuevas herramientas de bases de datos para Big Data, que podrá ser utilizado en el marco del proyecto de Ciudades Inteligentes en Río Negro para dar soporte al almacenamiento y análisis de información sobre el volumen de datos que éste genere a futuro. En dicho camino se podrá ahondar en el uso de otros tipos y productos de base de datos NoSQL para Big Data, incluso aquellos que seguramente seguirán surgiendo en el mercado.

Referencias

1. Martín, A, Chavez, S., Rodriguez, N., Valenzuela, A., Murazzo, M. Bases de datos NoSQL en Cloud Computing. UNSJ. XV Workshop de Investigadores en Ciencias de la Computación. 2013
2. Bollatti, Rodolfo. Big Data en la educación. XV Workshop de Investigadores en Ciencias de la Computación. 2013
3. Antiñanco, Matías Javier. Bases de Datos NoSQL: Escalabilidad y alta disponibilidad a través de patrones de diseño. Diciembre, 2013.
4. Eaton, Chris. Deutsch, Tom. Deroos, Dirk. Lapis, George. Zikopoulos, Paul. Understanding Big Data. Analytics for Enterprise Class Hadoop and Streaming Data. Mc Graw Hill
5. Joyanes Aguilar, Luis. Big Data: análisis de grandes volúmenes de datos en organizaciones. Alfaomega. Julio 2013.
6. Pruebas de rendimiento bases de datos columnares vs bases de datos orientadas a filas. http://www.stratebi.es/todobi/abr12/DBColumn_OpenSource.pdf
7. Strauch, Christof. NoSQL Databases. Universidad Stuttgart Media. 2010.
8. Katsov, Ilya NoSQL Data Modeling Techniques http://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/ (2012).
9. Brewer, Eric. Toward Robust Distributed Systems. http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf (2000).
10. Pentaho Big Data Architecture. http://es.slideshare.net/datalytics/big-data-architecture-con-pentaho
11. Timestored http://www.timestored.com/kdb-guides/kdb-database-intro
12. Infobright Analytic Database. Architecture Overview. Whitepaper. http://www.infobright.org/
13. Infobright Data Loading Guide Revision 2.1 – November 11, 2010 http://www.infobright.org/downloads/ice/Infobright_Data_Loading_Guide.pdf
14. Infobright Best Practices May 2012

http://es.scribd.com/doc/133237610/Infobright-Best-Practices