Desarrollo de un Modelo de Procesos para Selección de Bases de Datos ante Problemáticas Determinadas

Giovanni Rottoli[1,2], Patricio Boffino[1], Martín Mihura[1],

Juan Zaffaroni[1] & Ma. Florencia Pollo-Cattaneo[1,2]

[1]*Grupo de Estudios en Metodologías de Ingeniería de Software (GEMIS), Facultad Regional Buenos Aires, Universidad Tecnológica Nacional, Argentina.*

[2]*Ingeniería en Sistemas de Información, Facultad Regional Concepción del Uruguay, Universidad Tecnológica Nacional, Argentina*

rottolig@frcu.utn.edu.ar, flo.pollo@gmail.com

Abstract

Distintas tecnologías de almacenamiento y gestión de datos, tales como las bases de datos NoSQL, se desarrollan constantemente para solucionar los nuevos problemas que ocasiona el fenómeno denominado Big Data. Sin embargo, ante esta gran variedad de opciones es menester contar con un método sistemático para evaluar y seleccionar aquellas tecnologías más adecuadas ante problemáticas particulares. Se presenta en consecuencia una primera aproximación a un modelo de procesos que se encuentra actualmente en desarrollo para la evaluación y selección de tecnologías de almacenamiento y gestión de datos. Se ilustra además la aplicación del modelo propuesto sobre un caso real a manera de prueba de concepto.

1. Introducción

Motivados por los constantes cambios que sufren las organizaciones debido a las variaciones del contexto en el cual se encuentran insertas, surgen nuevas soluciones y herramientas en materia de almacenamiento y gestión de datos [1,2].

Nuevos requerimientos como el poder escalar horizontalmente, poseer alta disponibilidad de datos, o esquemas de datos flexibles, entre tantos otros, surgieron como consecuencia del fenómeno denominado *Big Data*. Debido a esto, surgieron nuevas tecnologías de bases de datos, denominadas NoSQL, como alternativa a las tradicionales bases de datos relacionales que no pueden paliar este nuevo abanico de problemas [3-6].

Dentro de estas nuevas tecnologías de bases de datos se distinguen cuatro tipos principales, según la manera en la que representan los datos [4; 6-11]:

- Bases de datos Documentales: almacenan datos en forma de documentos, sin necesidad de poseer un esquema predefinido.
- Bases de datos Clave-Valor: modelan los datos en duplas <clave, valor>, donde la clave corresponde a un identificador único del dato, y el valor dato en sí mismo. Debido a su sencillez suelen ser muy eficientes.
- Bases de datos Columnares: almacenan un conjunto de valores – columnas – que son referenciados por una clave única, de forma similar a las bases de datos clave valor. Posibilitan agrupar los mismos en familias de columnas.
- Bases de datos Gráficas: representan los datos como nodos y relaciones entre ellos. Utilizan la teoría de grafos para realizar consultas de forma eficiente.

Ante determinados problemas, estas alternativas, incluyendo a las bases de datos relacionales, poseerán distinto desempeño, haciendo que una correcta evaluación y selección de los motores de bases de datos que serán utilizados sea indispensable, no existiendo

actualmente herramientas para que los analistas y diseñadores realicen esta tarea de forma sistemática en el marco de proyectos de sistemas de información [5, 12].

En el contexto detallado se plantea entonces la definición de una primera aproximación a un modelo de procesos que permita la evaluación y selección de tecnologías de bases de datos según el problema que se aborde. Para ello, el presente documento se estructura de la siguiente manera: en la sección 2 se presenta el problema en cuestión, se desarrolla la solución al mismo en la sección 3, se muestra un ejemplo de aplicación de la solución en la sección 4 y se presenta conclusiones y futuras líneas de trabajo en la sección 5.

2. Definición del Problema

Como se mencionó anteriormente, existe una amplia variedad de herramientas de almacenamiento y gestión de datos que han surgido en los últimos años y que poseen características particulares para hacer frente a problemas relacionados al actual fenómeno denominado Big Data.

Estas herramientas poseen características particulares que las diferencian entre sí y muchas veces sacrifican cuestiones propias de las bases de datos relacionales tradicionales (tales como el aseguramiento de la consistencia plena, esquemas rígidos, integridad referencial y de entidad, la utilización de transacciones complejas, el uso de índices secundarios, entre otros) [13, 14]. Por este motivo, la selección de una tecnología de almacenamiento y gestión de datos, impacta directamente en el desempeño de la misma frente a la problemática que se aborda [5,12].

Ante una amplia variedad de problemas y alternativas en almacenamiento y gestión de datos, tal como se plantea en [15], se considera de interés desarrollar un modelo de procesos que permita llevar a cabo de manera sistematizada un análisis y selección de las herramientas adecuadas para hacer frente a los problemas específicos que se presentan. De esta forma se podrá contar con mecanismos estandarizados en las etapas de diseño de sistemas de información que permitan considerar las distintas alternativas de bases de datos que surgen constantemente en la actualidad.

2.1. Metodología

Se utilizará un enfoque de investigación clásico con énfasis en la producción de tecnologías, mediante un método prototipado experimental, desarrollando una solución inicial y refinándolo de manera evolutiva por prueba de aplicación de la solución a casos de estudio de complejidad creciente [16, 17].

3. Solución Propuesta

A partir del problema identificado en la sección anterior, se propone la construcción de un modelo de proceso para la evaluación y selección de la tecnología de almacenamiento y gestión de datos adecuada en función a un problema de negocio determinado.

Se describe en este trabajo una aproximación inicial, constituida por cuatro fases, tal como se puede observar en la Figura 1, que serán expuestas individualmente en las subsecciones 3.1 de Análisis del problema, 3.2 de Mapeo de Características, 3.3 de Evaluación de Tecnologías y finalmente, 3.4 Selección de Alternativa.

3.1. Análisis del Problema

La primera fase del proceso consiste en el análisis y comprensión del dominio del problema.

Es necesario identificar en primer lugar el problema de negocio que requiere la utilización de tecnologías de gestión de datos, para lo cual se debe contar con personas con nivel de experticia a fin de realizar una adecuada captura de requerimientos, a través de herramientas como la descripción del escenario actual, glosarios del negocio, definición de requisitos estructurados (SRD), entrevistas, entre otros [18].

Posteriormente a la definición del problema de negocios, es necesario realizar un desglose del mismo en sub-problemas a fin de poder determinar cuáles son los requerimientos particulares asociados a dicho problema. Para ello, se sugiere utilizar herramientas como el árbol de descomposición funcional [19], el cual permite obtener una representación gráfica de dicho desglose.

3.2. Mapeo de Características

En la fase de mapeo se busca relacionar cada uno de

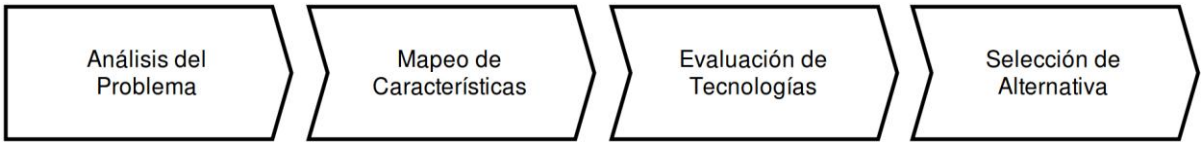

Figura 1.Fases del modelo de proceso propuesto

los requerimientos (producto de la fase de análisis) con características de herramientas de almacenamiento y gestión de bases de datos deseadas. Estas características pueden ser: poseer aseguramiento de la consistencia eventual, poseer esquema flexible o no tener esquema alguno, posibilidad de realizar escalado horizontal, replicación, distribución, entre otros [5]. Una vez más, un alto nivel de experiencia de aquellas personas que llevan a cabo estas actividades es primordial a fin de poder determinar correctamente cuáles son estas características, las cuales son extraídas principalmente de distintas fuentes bibliográficas. Para ello, se sugiere la utilización de una tabla de mapeo de dos columnas: la primera para los requerimientos y la segunda para las características asociadas a cada uno de ellos. Alternativamente, se puede utilizar grafos para una representación visual de estas relaciones.

Es importante tener en cuenta que un requerimiento puede estar relacionado con varias características y, a su vez, una características puede vincularse con varios requerimientos.

3.3. Evaluación de Tecnologías

Una vez que todas las características de almacenamiento y gestión de datos necesarias para resolver el problema de negocio hayan sido especificadas, se procede con la fase de evaluación de las distintas tecnologías disponibles. Para esto, es necesario determinar en qué grado las distintas tecnologías cumplen con las características por cada uno de los requerimientos.

Se procede entonces con la construcción de una tabla de dos entradas por cada requerimiento con las características asociadas al mismo en las filas y las tecnologías a evaluar en las columnas. La tabla será completada en cada celda con un resumen de cómo la tecnología en cuestión aborda la característica especificada. Por ejemplo, ante una característica como la implementación de agregados, se describiría de qué forma la tecnología evaluada implementa dichos agregados.

Tabla 1. Valores de grado de cumplimiento

Descripción	Valor Asociado
No cumple	20
Cumple parcialmente	40
Cumple	60
Cumple ampliamente	80

En función de esta tabla construida, se utiliza la Tabla 1 para otorgar un puntaje a cada intersección, generándose otra tabla de valores de grado de cumplimiento

Esta asignación puede realizarse en distintos niveles. Por ejemplo, es posible realizar un análisis a nivel de Paradigma, tales como el relacional, documental, entre otros, a nivel de base de datos, entre distintos motores, o bien a nivel de versiones de motores de bases de datos determinados. Por otro lado, si una característica se encuentra asociada a más de un requerimiento, es menester realizar la evaluación por cada uno de éstos, debido a la necesidad de evaluar, en la fase siguiente, como influye cada una de estas características en el requerimiento al que se encuentran asociados.

3.4. Selección de Alternativa

Una vez finalizada la fase anterior se prosigue con la selección de la alternativa que mejor se adapte a las necesidades manifestadas. Para ello, en primera instancia es necesario calificar los requerimientos obtenidos en la primera fase en función al grado de importancia de los mismos, según se considere, utilizando la Tabla 2.

Por cada requerimiento, luego, se multiplica el valor asociado al mismo con los valores de grado de cumplimiento asignados a cada característica en la fase anterior. Posteriormente se sumarán los resultados obtenidos por columna a fin de obtener valores de comportamiento de cada tecnología frente al requerimiento específico.

Tabla 2. Puntajes según importancia del requerimiento

Descripción	Puntaje
Requerimiento Opcional	20
Requerimiento Potencialmente Requerido	40
Requerimiento requerido a corto plazo	60
Requerimiento importante	80

En esta instancia es evidente el por qué es necesario evaluar una característica en múltiples oportunidades en caso de estar asociada a varios requerimientos: estos últimos podrían poseer un puntaje asociado diferente, por lo cual los valores de comportamiento resultarían distintos.

Luego de realizar este procedimiento por cada uno de los requerimientos, se suman los valores totales

obtenidos por cada tecnología, a fin de obtener un valor total general que representa cómo se comporta la tecnología en cuestión frente al problema original.

Un mayor valor total general obtenido sugiere la selección de dicha tecnología para ser implementada ante el problema presentado. Si algunas tecnologías evaluadas presentan valores totales generales iguales, o bien muy semejantes, queda a criterio del experto el realizar otra evaluación revisando los puntajes asignados en las etapas previas.

4. Caso de Ejemplo de Aplicación: Empresa de Seguridad de Mercaderías

En esta sección se presenta un caso de ejemplo de aplicación del proceso propuesto sobre una problemática real a modo de ilustración para ejemplificar la propuesta sin profundizar en la selección de la alternativa tecnológica para cada caso. Por otro lado, no se detalla información sobre la organización en la cual se enmarca el problema debido a cuestiones de confidencialidad.

El caso de estudio en cuestión corresponde a una empresa internacional de seguridad de mercaderías en tránsito que actualmente se encuentra operando con contenedores en distintas terminales portuarias de Argentina y, en custodia de mercaderías y control de seguridad de portones en depósitos de Aeropuertos de Argentina.

En la primera fase del proceso y mediante la utilización de encuestas sobre los *stakeholders*, se obtiene la necesidad de construir un sistema web para la contratación y seguimiento de los servicios por parte de los clientes en cualquier momento y en cualquier lugar del mundo. Es menester contar con datos replicados cerca de las ciudades principales donde se realizan las operaciones.

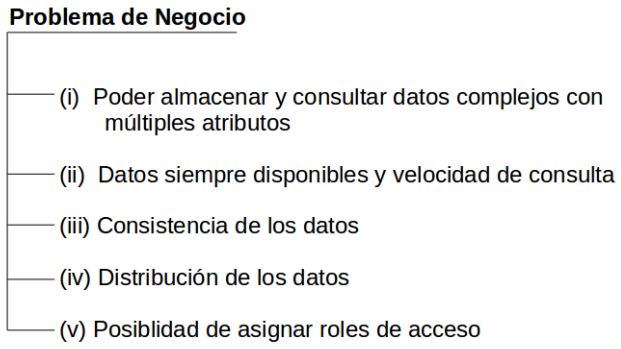

Figura 2. Árbol de descomposición funcional para el primer caso de ejemplo

Tabla 3. Mapeo de requerimientos del primer caso de ejemplo con características de tecnologías de gestión de datos

Requerimiento	Características
Almacenar y consultar datos complejos con múltiples atributos	- Datos con múltiples atributos. - Utilización de índices secundarios. - Uso de agregados
Datos siempre disponibles y velocidad de consulta	- Alta disponibilidad. - Distribución de carga
Consistencia en los datos	- Aseguramiento de la consistencia. - Aseguramiento de la integridad.
Distribución de los datos	- Replicación. - Alta disponibilidad
Posibilidad de asignar roles de acceso	- Seguridad

La problemática identificada fue dividida en sub-problemas utilizando el árbol de descomposición funcional [19] tal como se ve en la Figura 2, resultando los requerimientos que se muestran en las hojas del mismo: (i) Poder almacenar y consultar datos complejos con múltiples atributos; (ii) Datos siempre disponibles y velocidad de consulta; (iii) Consistencia en los datos; (iv) Distribución en los datos; (v) Posibilidad de asignar roles de acceso.

Posteriormente, en la fase de mapeo de características, se procede con la realización de la tabla de mapeo sugerida, tal como se puede observar en la Tabla 3.

Tabla 4. Tabla de descripción del comportamiento de características sobre distintas tecnologías

Característica	Oracle 12c	MongoDB 3.2.x	Cassandra 3.05
Múltiples atributos	- Lenguaje SQL. - Acceso/búsqueda multi-atributo por clave única o rango.	- Lenguaje JavaScript. - Acceso/búsquedas multi-atributo, por clave única o rango.	- Lenguaje CQL. - Acceso y búsquedas por ROW KEY o por rango en un 2do nivel.
Utilización de índices secundarios	- Índices únicos, duplicados, simples, compuestos.	- Índices únicos, duplicados, simples, compuestos.	- Índices sobre 2do nivel. - Índices basados en hashing.
Utilización de agregados	- Agregación mediante GROUP BY.	- Frameworks de agregación. - Map-Reduce nativo/externo con Hadoop.	- Map-Reduce externo con Hadoop

Tabla 5. Tabla de grados de cumplimiento de características sobre distintas tecnologías

Característica	Oracle 12c	MongoDB 3.2.x	Cassandra 3.05
Múltiples atributos	80	80	20
Utilización de índices secundarios	80	80	40
Utilización de agregados	80	60	20

Tabla 6. Tabla de gardos de cumplimiento de características sobre distintas tecnologías afectados por el puntaje asignado al requerimiento

Característica	Oracle 12c	MongoDB 3.2.x	Cassandra 3.05
Múltiples atributos	640	640	640
Utilización de índices secundarios	640	640	320
Utilización de agregados	640	480	160
Total	1920	1760	1120

Se ha seleccionado el requerimiento (i) del caso analizado y las características asociadas al mismo según la Tabla 3 y se realizó la evaluación con las tecnologías de base de datos relacional Oracle 12c, base de datos documental MongoDB3.2.x, y la base de datos columnar Cassandra 3.05, por ser las bases de datos más utilizadas en el mercado y con mejor ranking para su tipo al mes julio de 2016 [20].

Los resultados obtenidos, esto es, la tabla de descripción y la tabla de valores de grado de cumplimiento de cada característica se pueden observar en la Tabla 4 y la tabla 5.

Por último, en la fase de selección de la alternativa se determina el puntaje 4, requerimiento importante, a asignar al requerimiento evaluado y se realizan las operaciones correspondientes sobre la tabla de grado de cumplimiento de las características construida anteriormente, a fin de evaluar el comportamiento de las tecnologías elegidas ante el requerimiento de forma particular. Los resultados obtenidos se pueden observar en la Tabla 6.

Posteriormente, este proceso debe ser realizado por cada uno de los requerimientos y sumar los valores obtenidos a fin de poder seleccionar una alternativa.

Al evaluar todos los requerimientos mediante este proceso la base de datos NoSQL Documental MongoDB3.2.x fue seleccionada para brindar solución a la problemática propuesta, a pesar de que en el requerimiento evaluado como ejemplo la alternativa relacional muestra mejor desempeño.

5. Conclusión

Se presenta en este trabajo un primer abordaje a la creación de un modelo de procesos para seleccionar tecnologías de almacenamiento y gestión de datos adecuadas para una problemática en particular, definiendo las fases del mismo y las tareas generales a realizar para evaluar las distintas alternativas, e ilustrándolo posteriormente mediante su aplicación sobre un caso de estudio real.

El presente trabajo se encuentra en pleno desarrollo, por lo cual se continúa trabajando en el mismo a fin de lograr una estructura más madura, definiendo y automatizando tareas y herramientas en cada etapa, para posteriormente realizar la validación de dicho proceso mediante su aplicación en diversos escenarios.

6. Referencias

[1] Moya J. (2008). Management Democrático. Cataluña: PreMya Consultores.

[2] Lopez, D. (2012). Análisis de las posibilidades de uso de Big Data en las organizaciones. Universidad de Cantabria, Santander, España.

[3] Abramova,V.,Bernardino,J., & Furtado,P. (2014). Experimental evaluation of NoSQL databases. International Journal of Database Management Systems, 6(3), 1.

[4] Nayak, A., Poriya, A., & Poojary, D. (2013). Type of NOSQL databases and its comparison with relational databases. International Journal of Applied Information Systems, 5(4), 16-19.

[5] Sadalage P. & Fowler M. (2013).“NoSQL Distilled, A Brief Guide to the Emerging World of Polyglote Persistence”, Addison-Wesley, Boston, USA, 1st. Edition.

[6] Strauch, C., Sites, U. L. S., & Kriha, W. (2011). NoSQL databases. Lecture Notes, Stuttgart Media University.

[7] Arora, R.,& Aggarwal, R. R. (2013). Modeling and Querying Data in MongoDB.International Journal of Scientific and Engineering Research, 4(7)

[8] Bugiotti, F., & Cabibbo, L. (2013). A Comparison of Data Models and APIs of NoSQL Datastores. Dipartamento di Ingegneria della Università di Roma.

[9] Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. 2011 International Conference on Cloud and Service Computing. 336-341

[10] Hewitt, E. (2011). Cassandra: The Definitive Guide. Sebastopol, CA: O’Reilly Media, Inc. p. 24.

[11] Moniruzzaman, A. B. M., & Hossain, S. A. (2013). NoSQL database: New era of databases for big data analytics classification, characteristics and comparison. arXiv preprint arXiv:1307.0191.

[12] Pollo-Cattaneo, M. F., Nocera, M. L., & Rottoli, G. D. (2014). Rendimiento de tecnologías NoSQL sobre cantidades masivas de datos. Cuaderno Activa, (6), 11-17. ISSN: 2027–8101.

[13] Copeland, R. (2013). MongoDB Applied Design Patterns. Sebastopol, CA: O’Reilly Media, Inc. p. 25.

[14] MongoDB. (2015, August). Top 5 Considerations When Evaluating NoSQL Databases. Disponible en: https://www.mongodb.com/collateral/top-5-considerations-when-evaluating-nosql-databases. Consultado el 24 de agosto de 2016.

[15] Róttoli, G.D., Zaffaroni, J., López Nocera, M., Pollo Cattaneo, M. F. (2016, Abril). Metodología para evaluación de impacto de migración entre versiones de bases de datos NoSQL. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). p. 258-262. ISBN: 978-950-698-377-2

[16] Riveros, H. & Rosas, L. (1985) El Método Científico Aplicado a las Ciencias Experimentales. México: Editorial Trillas. ISBN 96-8243-893-4.

[17] Sabato J. & Mackenzie M. (1982) La Producción de Tecnología: Autónoma o Transnacional. Instituto Latinoamericano de Estudios Transnacionales - Technology & Engineering. ISBN 9789684293489.

[18] Ochoa, A., & Martínez, R. G. (2006). Uso de Técnicas de Educción para el Entendimiento del Negocio (Doctoral dissertation, Tesis de Tesis de Magister en Ingeniería del Software. Escuela de Postgrado. ITBA).

[19] Gomez A., Juristo N., Montes C., & Pazos J. (1997). Ingeniería Del Conocimiento. Colección de Informática. Ed. Centro de Estudios Ramón Arces. Madrid. ISBN: 84-8004-269-9

[20] SolidIT (Julio, 2016). DB-Engines Ranking. DB-Engines-Knowledge Base of Relational and NoSQL Database Management Systems. Disponible en: http://db-engines.com/en/ranking. Consultado el 24 de agosto de 2016.