Modelado de Variedad de Activos de Dominio en Sistemas Big Data

Agustina Buccella, Alejandra Cechich, Juan Luzuriaga
Líam Osycka, Carolina Villegas, Marcos Cruz, Franco Corgatelli,
Rodolfo Martínez, Rafaela Mazalu, Marcelo Moyano
GIISCO Research Group
Departamento de Ingeniería de Sistemas
Universidad Nacional del Comahue
Neuquen, Argentina
agustina.buccella@fi.uncoma.edu.ar

1. Resumen

Un cambio importante con respecto a depósitos de datos tradicionales, es que en los Sistemas Big Data (SBDs) la naturaleza no estructurada de algunos datos puede provenir de diversas fuentes, entre ellas sensores, redes sociales, entorno y la misma empresa. La diversidad de esos datos puede analizarse abordando distintas características. Precisamente, la propiedad de los SBDs con respecto a diversidad de los datos se denomina *Variedad*.

La variedad en SBDs ha sido relacionada con diversas propiedades como interoperabilidad, seguridad, reusabilidad, etc. En este contexto, y respondiendo a la pregunta de investigación: *¿Cómo puede modelarse la variedad de la información de dominio de manera de incorporar reusabilidad en el desarrollo de SBDs?* , nuestro proyecto propone modelar variedad a modo de líneas de productos. A diferencia de otras propuestas, la nuestra toma como partida una estructura de actividades asociadas al desarrollo de SBDs, instanciada en artefactos software producidos durante esas actividades e incorpora el modelado de variedades de manera similar a líneas de productos.

Palabras Clave: Reusabilidad - Líneas de Producto de Software - Big Data

2. Contexto

La línea presentada se inserta en el contexto del *Proyecto UNComa: Modelado de Variedad en Sistemas Big Data*. Directora: Dra. Agustina Buccella, Co-directora; Dra. Alejandra Cechich, que se encuentra en etapa de evaluación (2022-2025).

3. Introducción

En [1], la *Variedad* se clasifica en una taxonomía que divide el análisis en cuatro casos de diversidad: estructural, de las fuentes, de contenido y de procesamiento. Por ejemplo, la diversidad estructural denota la variedad en formatos y tipos de datos, clasificándolos como estructurados, semi-estructurados y no estructurados; la diversidad de las fuentes se clasifica en tres grupos - datos generados por humanos, generados por máquinas o mediados por procesos; la diversidad de contenido aborda diferentes tipos de soporte (único medio, multimedio o gráfico); y la diversidad de procesamiento enfoca en las distintas necesidades de procesamiento algorítmico (batch, interactivo, streaming o gráfico).

La variedad en los datos también ha sido considerada desde el punto de vista de incorporación de semántica al proceso de modelado de arquitecturas en SBDs; por ejemplo, en

[2] se utiliza modelado de contextos para mejorar las aplicaciones Big Data en el ámbito de ciudades inteligentes y a efectos de analizar su sustentabilidad. En [11], en cambio, se sugiere el uso de técnicas de Deep Learning para extraer patrones complejos en los datos, así como para permitir indexación semántica.

Por otra parte, la variedad en SBDs ha sido relacionada con diversas propiedades como interoperabilidad, seguridad, reusabilidad, etc. Por ejemplo, en [9], se presentan seis arquitecturas de referencia propuestas actualmente en la literatura para SBDs y se analiza el cumplimiento de las cinco Vs (Volumen, Velocidad, Variedad, Variabilidad y Veracidad) como requerimientos tradicionales. En el análisis, la variedad se asocia a los siguientes requerimientos: a (R1) la heterogeneidad de los datos en formatos no estructurados (ej. texto y video), semi-estructurados (ej. basados en XML o JSON) y estructurados (ej. tablas relacionales); a (R2) el análisis eficiente de los datos lo cual requiere entender exactamente qué está almacenado y a (R3) resolver conflictos de interoperabilidad, lo que implica el manejo de metadatos apropiados. R2 puede abordarse por medio de métodos para integración de datos (ej. transformación basada en reglas), que permiten resolver conflictos en los esquemas de datos. Sin embargo, R3 requiere de una efectiva integración de los datos para resolver conflictos de interoperabilidad: (1) conflictos de dominio que se relacionan con diferentes interpretaciones del mismo dominio, incluyendo homónimos, sinónimos, acrónimos y restricciones de integridad; (2) conflictos de granularidad que se relacionan con diferentes unidades de medida y agregación de los datos; y (3) conflictos de completitud que se relacionan con diferentes piezas de datos que pertenecen a la misma entidad. De las comparaciones realizadas, se desprende que casi todas las arquitecturas de referencia propuestas satisfacen el requerimiento R1, pero sólo algunas abordan parcialmente R2 y ninguna al momento satisface R3, dejando un amplio margen para aportes en el área.

En SBDs, la reusabilidad ha sido abordada también desde diversos ángulos. Por ejemplo, en [13] se discuten conceptos de reusabilidad en el contexto de analítica de datos distinguiendo entre uso y reuso del dato. Más específicamente, pero en el mismo sentido, en [8] se profundizan aspectos de privacidad en el contexto de reusabilidad de datos. Allí se propone una taxonomía en reuso de datos que pueda ser útil para determinar en qué medida ese reuso debe ser permitido y bajo qué condiciones para preservar privacidad. Otras propuestas, abordan aspectos de reuso en términos de aumentar la colaboración en el desarrollo de SBDs mediante el uso de nuevas tecnologías (ej. computación en la nube). Por ejemplo, en [15] se propone un enfoque de gestión de SBDs mediante capacidades de almacenamiento y procesamiento en una nube pública y se ejemplifica su uso. Adicionalmente, las distintas plataformas de soporte para desarrollo de SBDs también se abordan desde un punto de vista de reuso; por ej. en [3] se analiza la mejora en la eficiencia de herramientas como Apache Hadoop[1] y Spark[2] debido al reuso de artefactos entre diversos proyectos. Para ello, se analizan aspectos comunes y se provee de un flujo de trabajo implementado de manera escalable y extensible.

En un sentido similar, incorporando la detección de aspectos comunes y variables a modo de familia de sistemas, en [10] la arquitectura de referencia se ve acotada por medio de casos de uso (ej. visualización y análisis de información geoespacial estratégica, análisis inteligente de señales, etc.). De esos casos, se identifican requerimientos relevantes al SBD, incluyendo categorías, como tipos de datos (ej. texto no estructurado, geoespacial, audio), transformaciones en los datos (ej. clustering, correlación), visualizaciones (ej. imágenes, redes), etc. Luego, la arquitectura se organiza como una colección de módulos que descomponen la solución en elementos realizando funciones o capacidades para un conjunto de aspectos (requerimientos externos, módulos reusables, roles y participantes, paquetes de datos comerciales y open source). Finalmente, se descompone el SBD en 13 módulos agrupados en las categorías: (1) Proveedor de aplicaciones de Big Data (incluyendo la ló-

[1] https://hadoop.apache.org/
[2] https://spark.apache.org/

gica de negocios del sistema), (2) Proveedor del framework de Big Data (incluyendo plataformas, almacenamiento, etc.), y (3) Módulos transversales (abordando distintos aspectos relevantes al desarrollo de SBDs).

Considerando estas propuestas, en [6] hemos definido una primera aproximación de los elementos que componen nuestra arquitectura de referencia para SBDs basada en reuso. En ella, los aspectos de negocios (dominio), aplicación (software y análisis) y tecnológicos se abordan en niveles separados; siendo transversales aspectos como el uso/reuso de estándares, taxonomías y conocimiento.

Uno de los componentes principales de esta arquitectura agrupa los denominados *activos de dominio*, constituidos por artefactos de software que son creados para el dominio en el que se esta trabajando. Así, además de incluir a los participantes del desarrollo del SBD, involucra los requerimientos del proyecto y del dominio, restricciones, modelos y casos de uso. Es importante resaltar que estos activos deben generarse a partir de *taxonomías de dominio y estándares* y de *activos basados en conocimiento*. De esta forma, se deben crear artefactos enfocados en que puedan ser reusados en el mismo dominio e incluso en otros dominios relacionados (artefactos para reuso), y/o que puedan desarrollarse en base a otros artefactos ya creados (artefactos con reuso).

Para identificar variedad en activos de dominio, proponemos dos enfoques: (1) identificación desde los datos (bottom-up) y/o identificación desde los requerimientos (top-down). Ambos casos no son excluyentes, ya que se puede iniciar una búsqueda a modo exploratorio desde los datos, a la vez que se establecen algunos requerimientos de búsqueda.

En la Figura 1 mostramos la visión global del enfoque bottom-up de nuestra propuesta [12], es decir, la identificación de variedad a partir de los datos (Paso (1) "¿Qué dicen los datos?"). Se procede luego a realizar el estudio exploratorio (Paso (2)), que brindará información de correlaciones posibles en los datos y variaciones detectadas en su análisis.

En principio, al centrarnos en SBDs, el primer elemento a considerar es el proceso de desarrollo (Paso (3)), donde las etapas básicas pueden resumirse en: (1) Adquisición de Datos, que consiste en extraer los datos desde las fuentes, agregando un proceso de carga y filtrado para que los datos sean adecuados a su posterior procesamiento; (2) Preparación de Datos, que consiste en estructurar el formato de los datos, realizar la limpieza de los mismos y eventualmente, también su integración; y (3) Análisis de Datos, que contiene las funcionalidades que permiten derivar conocimiento a partir de los datos, enfocando en análisis descriptivo, predictivo y/o prescriptivo.

Luego, los resultados validados con los expertos (Paso (4)) podrán utilizarse en decisiones referidas al problema de dominio (Paso (5)), e incluso retroalimentar un nuevo ciclo exploratorio.

Entonces, para identificar variedad en los activos de dominio, el enfoque bottom-up de la propuesta parte de la definición de un problema dependiente del dominio e intenta detectar características variantes dentro de cada una de las etapas del proceso de análisis de datos (Paso (3)). En el ejemplo de la Figura 1, sólo se enfoca en identificar la diversidad de contexto (o dominio), manteniendo constantes las fuentes, contenido y procesamiento.

Al igual que en el desarrollo en líneas de producto software (LPSs), nuestra propuesta modela la variedad en activos de dominio a través de dos fases: la Ingeniería de Dominio y la Ingeniería de Aplicación. La primera es responsable de identificar y definir las funcionalidades y aspectos comunes y variables que forman parte de la plataforma que comparten todos los productos de la línea de productos software (todos los SBDs); mientras que la segunda está compuesta por las actividades que permiten realizar la derivación de productos particulares, es decir, que permiten instanciar en cada caso [14].

En trabajos previos, hemos presentado una propuesta de diseño de LPSs dirigida por funcionalidades, donde cada funcionalidad se documenta a través de una hoja de datos funcional (datasheet), representando el conjunto de servicios comunes y variantes [4, 5]. Para el caso de reusabilidad en SBDs, la Figura 1 muestra la hoja de datos funcional definida para reusar modelos de análisis de datos (activos

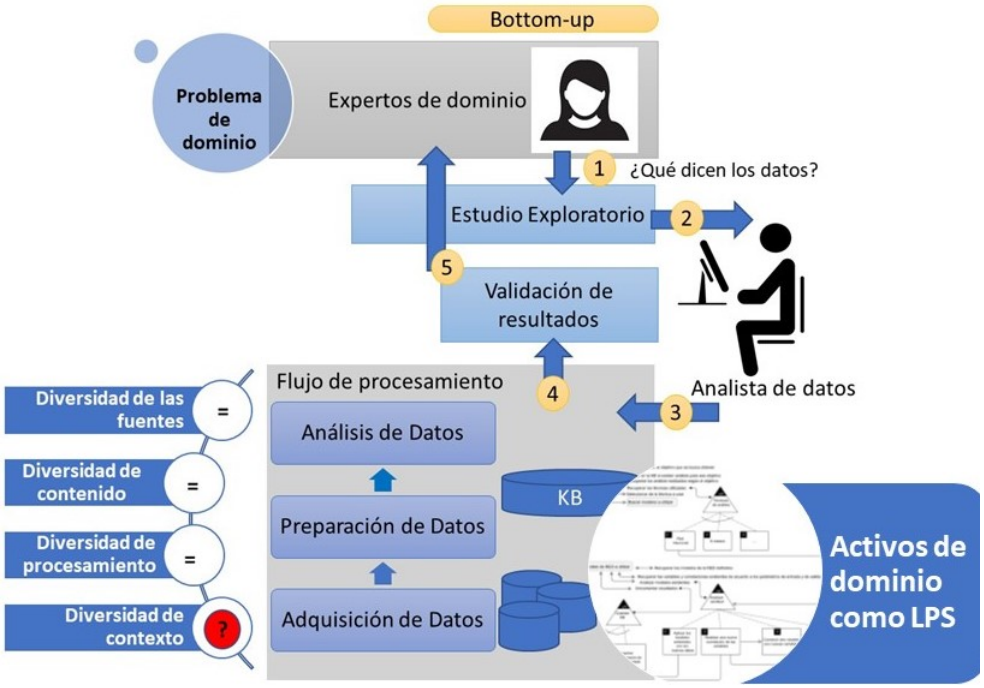

Figura 1. Vision global del enfoque bottom-up

de dominio a nivel de plataforma de LPS) y detectar variedades de contexto. Detalles de este modelo pueden verse en [12].

4. Líneas de Investigación y Desarrollo

En proyectos previos, hemos realizado amplios avances en lo que respecta al área de LPSs definiendo y refinando una metodología de desarrollo a nivel de subdominos. Dentro de la metodología, hemos presentado sus bases y diseñado artefactos que se utilizan en el análisis de dominios y en el análisis organizacional de una LPS [4] y tienen la particularidad de favorecer el reuso basado en una taxonomía de servicios. Es precisamente esta ventaja la que nos permitió luego realizar extensiones hacia otros subdominios. De esta forma hemos podido así avanzar en el desarrollo de múltiples LPSs basadas en la jerarquía de dominios definida. Sin embargo, para dicha extensión hemos tenido que formalizar varios aspectos respecto a los artefactos de software creados. En el caso de la taxonomía, se utilizaron los estándares reconocidos del subdomino, y se creó un proceso de desarrollo que se aplicó en la creación de las dos primeras LPSs [5, 7].

5. Resultados Obtenidos/Esperados

El objetivo principal de la línea de investigación es *Desarrollar técnicas y herramientas que mejoren los procesos y técnicas aplicadas a la explotación de grandes volúmenes de datos, favoreciendo el desarrollo de ambientes inteligentes que permitan reusabilidad.*

Al momento, hemos planteado una arquitectura de referencia [6] y un modelo de procesos [12] para el modelado y gestión de la variedad en SBDs. Los resultados publicados son preliminares a modo de prueba de conceptos. Actualmente, estamos trabajando en colaboración con el Instituto de Tecnología Agropecuaria (INTA)-Alto Valle para la aplicación del proceso de modelado en el análisis de la napa freática, en función de la variedad de fuentes acuíferas de diversas zonas geográficas (variedad contextual).

6. Formación de Recursos Humanos

El proyecto reúne aproximadamente a 13 investigadores, entre los que se cuentan docentes y alumnos de UNComa, y colaboradoras expertas del dominio de aplicación, espe-

cíficamente pertenecientes al INTA. A su vez, el proyecto cuenta actualmente con dos doctores y un magister. Varios de los docentes-investigadores de GIISCo-UNComa han terminado o se encuentran próximos a terminar carreras de postgrado. Además, varios de los integrantes se encuentran finalizando sus tesis de grado. Por último, este año seguiremos con la supervisión del trabajo de 2 becarios EVC-CIN.

Referencias

[1] Jemal Abawajy. Comprehensive analysis of big data variety landscape. *International Journal of Parallel, Emergent and Distributed Systems*, 30(1):5–14, 2015.

[2] S. Bibri and J. Krogstie. The core enabling technologies of big data analytics and context-aware computing for smart sustainable cities: a review and synthesis. *Journal of Big Data*, 4(38), 2017.

[3] Reuben Borrison, Benjamin Klöpper, Moncef Chioua, Marcel Dix, and Barbara Sprick. Reusable big data system for industrial data mining - a case study on anomaly detection in chemical plants. In *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, pages 611–622. Springer International Publishing, 2018.

[4] A. Buccella, A. Cechich, M. Arias, M. Pol'la, S. Doldan, and E. Morsan. Towards systematic software reuse of gis: Insights from a case study. *Computers & Geosciences*, 54(0):9 – 20, 2013.

[5] A. Buccella, A. Cechich, M. Pol'la, M. Arias, S. Doldan, and E. Morsan. Marine ecology service reuse through taxonomy-oriented SPL development. *Computers & Geosciences*, 73(0):108 – 121, 2014.

[6] A. Buccella, J. Luzuriaga, A. Cechich, L. Osycka, F. Paterno, M. Pol'la, M. Cruz, R. Martinez, R. Mazalu, and M. Moyano. Reusabilidad en el contexto de desarrollo de sistemas para big data. In *Actas del XXIII Workshop de Investigadores en Ciencias de la Computación, Chilecito, La Rioja*, pages 525–529, 2021.

[7] Agustina Buccella, Alejandra Cechich, Juan Porfiri, and Domenica Diniz Dos Santos. Taxonomy-oriented domain analysis of gis: A case study for paleontological software systems. *ISPRS International Journal of Geo-Information*, 8(6), 2019.

[8] Bart Custers and Helena Uršič. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. *International Data Privacy Law*, 6(1):4–15, 2016.

[9] Ali Davoudian and Mengchi Liu. Big data systems: A software engineering perspective. *ACM Computing Surveys*, 53(5), 2020.

[10] John Klein, Ross Buglak, David Blockhow, Troy Wuttke, and Brenton Cooper. A reference architecture for big data systems in the national security domain. In *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*. ACM/IEEE, 2016.

[11] M. Najafabadi, F. Villanustre, T. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 2017.

[12] L. Osycka, A. Buccella, and N. A. Cechich. Identificación de variedad contextual en modelado de sistemas big data. In *Memorias del XXVII Congreso Argentino de Ciencias de la Computación (CACIC)*, pages 367–376. Red de Universidades con Carreras en Informática, 2021.

[13] I.V. Pasquetto, B.M. Randles, and C.L. Borgman. On the reuse of scientific data. *Data Science Journal*, 16(8), 201720.

[14] Klaus Pohl, Günter Böckle, and Frank J. van der Linden. *Software Product Line Engineering: Foundations, Principles and Techniques*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[15] Zhiwu Xie, Yinlin Chen, Julie Speer, Tyler Walters, Pablo A. Tarazaga, and Mary Kasarda. Towards use and reuse driven big data management. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, page 65–74. Association for Computing Machinery, 2015.