Big Data y su impacto sobre las comunidades

Fernando Emmanuel Frati[1], Jose Texier[1,2], Fernanda Carmona[1], Daniel Robins[1], Alberto Riba[1], Cristian Rios[1]

[1] Departamento de Ciencias Básicas y Tecnológicas, Universidad Nacional de Chilecito
9 de Julio 22, Chilecito, La Rioja, Argentina
{fefrati, jtexier, fbcarmona, drobins, ariba}@undec.edu.ar,
riosbourne555@gmail.com

[2] Universidad Nacional Experimental del Táchira
Venezuela

Resumen

La necesidad de administrar grandes volúmenes de información no estructurada (heterogénea) es un tema que en la actualidad está en crecimiento. El auge se debe a la enorme producción de información digital, tanto la información producida en Internet como la cuantiosa información producida por las instituciones (llámense, empresas u organismos estatales). La cuestión es que en su mayoría tal cantidad de información no es administrada correctamente lo cual conlleva pérdidas o complicaciones en el funcionamiento. El propósito de la creación de esta línea de investigación es analizar el contexto de oportunidades presentes en la región de Chilecito y ofrecer alternativas a las diferentes problemáticas presentes desde la perspectiva del Big Data. Para tal fin, esta línea de I/D/I corresponde al diseño, desarrollo e implementación de proyectos que fortalecerán la investigación y las diferentes labores relacionadas con las necesidades de la región sobre la base del recurso humano presente en la Universidad Nacional de Chilecito (UNdeC).

Palabras clave: *Big Data, Chilecito, procesamiento de información*

Contexto

Esta línea de I/D/I corresponde al diseño, desarrollo e implementación de proyectos que fortalecerán la investigación y las diferentes labores relacionadas con las necesidades de la región sobre la base del recurso humano presente en la Universidad Nacional de Chilecito (UNdeC).

En la actualidad, se están desarrollando ideas-proyectos a través de una tesina de grado sobre el tema de Big Data aplicados a la Vitivinicultura. Para la próxima convocatoria (2016) del programa "Financiamiento para el Estímulo y Desarrollo de la Investigación Científica y Tecnológica" de la Secretaría de Ciencia y Tecnología en la UNdeC (FiCyT - UNdeC) se presentará un proyecto sobre Big Data en los procesos de elaboración de vinos. También se está desarrollando un proyecto de análisis social mediante tweets en determinados temas, por ejemplo, elecciones presidenciales de la Argentina del 2015. Es importan-

te destacar que el estudiante involucrado en el proyecto ha sido beneficiado con una beca de Estímulo a las Vocaciones Científicas, convocatoria EVC 2015 (CIN).

Por otra parte, la UNdeC cuenta con la estructura tecnológica y de RRHH necesarios para ejecutar los proyectos que surjan de la línea de investigación.

Introducción

La Universidad Nacional de Chilecito se ubica en uno de los centros urbanos más importante de la región de los valles áridos del noroeste argentino. Históricamente la investigación y desarrollo de la UNdeC ha sido impulsada por las necesidades del sector agrícola. Sin embargo, las políticas de la universidad en términos de inversión y vinculación tecnológica llevadas a cabo en los últimos años ha derivado en la instalación de dos importantes laboratorios para la región que ofrecen un enorme potencial de trabajo.

Este trabajo explora las oportunidades de I+D+I en Big Data[1], entendiendo que estas actividades deben ser realizadas en su contexto, beneficiándose del trabajo interdisciplinario como un importante motor de desarrollo regional. Se denomina Big Data a aquellos problemas que requieren procesar grandes volúmenes de datos en poco tiempo [2, 3, 4]. Por ello, el procesamiento y posterior análisis debe ser realizado (preferiblemente) en tiempo real para poder mejorar la toma de decisiones con base en la información generada. Las características del Big Data se centran en las tres V:

- Volumen de grandes volúmenes de datos.
- Variedad de diversos tipos de fuentes de datos, ya sean estructurados o no estructurados.
- Velocidad de los datos, es decir, la frecuencia de las actualizaciones de estas grandes bases de datos.

A partir del contexto descrito, esta propuesta analiza las potencialidades de la zona en conjunto con el RRHH presente en la UNdeC. Por ello, a continuación se proponen algunas ideas dirigidas a aplicar estos conceptos en la región:

Actividades agroindustriales: el clima de Chilecito se caracteriza por la extrema aridez, con grandes amplitudes térmicas, lluvias anuales medias de 200 milímetros, concentradas en época estival; fuerte insolación anual, frecuentes vientos desecantes y baja humedad atmosférica. Pese al marcado déficit hídrico típico de la región de los valles áridos, lleva adelante una intensa actividad agrícola industrial. La mayor concentración de cultivos en el subsector de fruticultura lo tienen el olivo con 11.000 hectáreas, la Vid con 6.500 hectáreas y el Nogal con 2.800 hectáreas, los cuales se comercializan a nivel local, regional, nacional e internacional. El proceso de industrialización de algunos cultivos como el de la vid se lleva adelante en 15 bodegas que se distribuyen en la ciudad y distritos del Departamento Chilecito. La fabricación de aceite de oliva está en pleno crecimiento, dando cuenta de un reducido número de productores que utilizan tecnología de avanzada para su elaboración.

Laboratorios de Altura y de Alta Complejidad: el laboratorio de altura es el primero de América de esta clase, y constituye un proyecto de excelencia en las actividades científicas-tecnológicas. Se encuentra a 5200 metros sobre el nivel del mar, en las Sierras del Famatina, de fácil acceso, excelentes condiciones atmosféricas y con la Universidad al pié de las sierras, lo que aporta el valor agregado de disponibilidad de recursos humanos formados

(y en formación). Este laboratorio permite hacer mediciones que no se pueden hacer al nivel del mar en campos como medicina, biología, astronomía, física, etc. Por otro lado el LAC presta los siguientes servicios a la comunidad: análisis de suelos, análisis de aguas para riego, análisis de aguas para consumo, análisis microbiológico de agua, análisis de efluentes, análisis en vías de implementación. Cuenta con una gran cantidad de instrumental de laboratorio y de campo, y actualmente está en proceso la adquisición de un secuenciador de ADN, un Secuenciador Genómico y un Microscopio Electrónico de Barrido de alta resolución, lo que permitirá ampliar los servicios ofrecidos. Ambos laboratorios cuentan con una importante cantidad de profesionales e investigadores de otras disciplinas e instrumental muy especializado. En consecuencia, existe una necesidad de trabajo con grandes volúmenes de datos y procesos complejos con requerimientos de tiempo real.

Oportunidades: se han desarrollado varias tesinas de grado (dirigidas por integrantes de este trabajo) afines a esta línea, involucrando tecnologías de microcontroladores y redes de sensores para resolver distintos problemas relacionados a la captura de información, automatización de procesos de control y predicción de variables asociadas a los procesos agrícolas. Sin embargo, estos proyectos se han desarrollado como sistemas a medida destinados solamente a la organización beneficiaria de cada proyecto. Por otro lado se están realizando reuniones de trabajo con las personas a cargo de ambos laboratorios en búsqueda de espacios de colaboración. Se propone llevar a gran escala estos proyectos, coordinando la participación de múltiples empresas de la región. Es de esperar que esto nos permita evaluar técnicas existentes e implementar desarrollos experimentales para clasificar, ordenar, jerarquizar y analizar información sobre grandes volúmenes de datos heterogéneos. Como objetivo principal nos proponemos implementar un repositorio de datos públicos a disposición de los investigadores de la región, aportando a cuestiones específicas sobre agrodatos de vid, olivo y nogales para Chilecito y zonas de influencia. Estos "Data Set" públicos nos habilitarán a pensar nuevos temas derivados de esta línea de trabajo.

La propuesta del proyecto permitirá establecer las capacidades necesarias con las que debería contar una base de datos de información masiva, tanto desde la perspectiva de almacenamiento y técnicas de indexación, como de distribución de las consultas, escalabilidad y rendimiento en ambientes heterogéneos, unido con el análisis de las oportunidades de aplicación del Big Data a la región.

Líneas de Investigación, Desarrollo e Innovación

- Bases de datos, Minería de datos
- Cloud Computing[11], Arquitecturas paralelas
- Análisis social web[12]
- Simulación
- Internet de las cosas
- Agromática, Vitivinicultura, Genética
- Repositorios institucionales y bibliotecas digitales, Análisis semántico de la información

Para llevar a cabo el proyecto propuesto, se plantean las siguientes actividades:

- Exploración de oportunidades de aplicación del Big Data en la región y relacionado con el recurso humano existente en la UNdeC.
- Obtención de datos. Por ejemplo, datos correspondientes a compras realizadas por personas, datos médicos de pacientes, estructuras de ADN, datos de redes sociales[5, 6].
- Modelado de los datos de forma relacional, documental, gráfica, clave-valor, y familia-columnas.
- Traslado de cada modelo a los diferentes motores de base de datos SQL y NoSQL [7, 8, 9].
- Análisis de los datos capturados, transformados y almacenados[10].
- Visualización de los datos.
- Pruebas mediante consultas complejas en cada uno de los motores para cada uno de los modelos desarrollados.
- Determinar si existen casos donde una estructura no influye en la eficiencia de las consultas bajo un determinado paradigma y si los resultados son comprensibles.
- Diseñar e implementar Data Set según las necesidades de la UNdeC.

Resultados y Objetivos

Objetivos

- Generar un entorno de desarrollo sobre tecnología Big Data.
- Consolidar un grupo de investigación multidisciplinario en la UNdeC.
- Fomentar, incentivar y difundir las tareas de investigación.
- Mejorar la formación de recursos humanos con capacidades de investigación y desarrollo.
- Generar y establecer Data Sets públicos con información regional para ofrecerlos a los investigadores de la universidad.

Resultados esperados

- Formación del recurso humano.
- Identificación de las diferentes oportunidades de proyectos para la región.
- Diseño de un conjunto de estrategias para abordar las oportunidades en la región.
- Generación de un marco de trabajo que permite a la comunidad contar con una estructura para solucionar sus problemas, enfocados sobre el dominio de datos masivos.

Formación de Recursos Humanos

El equipo de trabajo está formado por docentes de las carreras Ingeniería en Sistemas y Licenciatura en Sistemas de la UNdeC (acreditadas por CONEAU), dos doctores especializados en repositorios institucionales, bibliotecas digitales, desarrollo de software, cómputo paralelo y tecnología grid. Otra docente está definiendo su tesis de Maestría en Informática. También participa un alumno avanzado de grado. En otras palabras, se cuenta con un recurso humano con habilidades y formación académica en las diversas áreas de la propuesta, asegurando la concreción de la línea. Adicionalmente, se destaca que dos están categorizados en el programa de incentivos.

Los integrantes son docentes de las asignaturas Programación I, Sistemas I, Arquitecturas Paralelas, Teoría de la Computación, Bases de Datos I y II, y Herramientas de Ingeniería de Software. Estas asignaturas contemplan la aprobación mediante la participación en proyectos de investigación, por lo que pueden surgir nuevos trabajos en esta línea.

Referencias

[1] D. López, "Análisis de las posibilidades de uso de Big Data en las organizaciones," *Universidad de Cantabria, Santander, España*, 2012.

[2] A. Rajaraman, J. D. Ullman, J. D. Ullman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press Cambridge, 2012, vol. 1.

[3] B. Dana and C. Kate, "'Critical Questions for Big Data'," *Information, Communication and Society*, vol. 15, p. 5, 2012.

[4] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, Sep. 2008.

[5] N. Antonopoulos and L. Gillam, *Cloud Computing: Principles, Systems and Applications*. Springer Science & Business Media, Jul. 2010.

[6] W. Hall, D. D. Roure, and N. Shadbolt, "The evolution of the Web and implications for eResearch," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1890, pp. 991–1001, Mar. 2009. [Online]. Available: http://rsta.royalsocietypublishing.org/content/367/1890/991

[7] E. Serrano and C. A. Iglesias, "Validating viral marketing strategies in Twitter via agent-based social simulation," *Expert Systems with Applications*, vol. 50, pp. 140–150, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417415008234

[8] B. Li, B. F. Liu, and N. X. Liu, "The Big Data how to Influence the Development of E-Government," in *Advanced Materials Research*, vol. 989. Trans Tech Publ, 2014, pp. 4717–4722. [Online]. Available: http://www.scientific.net/AMR.989-994.4717

[9] H. Ramírez and J. F. Herrera, "Un viaje a través de bases de datos espaciales," *NoSQL: Redes de ingeniería, Univ. Distrital Francisco J de Caldas, Bogotá, Colombia*, vol. 4, pp. 35–47, 2013.

[10] H. G. del Busto and O. Y. Enríquez, "Bases de datos NoSQL," *Revista Telem@ tica*, vol. 11, no. 3, pp. 21–33, 2013.

[11] C. Nance, T. Losser, R. Iype, and G. Harmon, "Nosql vs rdbms-why there is room for both," in *Proceedings of the Southern Association for Information Systems Conference*, 2013, pp. 111–116.

[12] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 647–657, Sep. 2010.