Desarrollo de Capacidades Científico-Tecnológicas para la Gestión de Datos Masivos

Fernando Emmanuel Frati[1], Jose Texier[1], Daniel Robins[1,2], Fernanda Carmona[1], Alberto Riba[1], Javier Ruitti[1], Jonatan Alvarez[1,2], Cristian Rios[1], Lucas Loto[2]

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

Resumen

El incremento en la capacidad de almacenamiento y procesamiento de los equipos de cómputo, sumado a conexiones a Internet cada vez más veloces, permiten diseñar soluciones de datos masivos -antes restringidas a las ciencias tradicionales- para problemas de diferentes áreas de la sociedad. Actualmente, es posible poner a la par ciencia y sociedad como grandes generadores de datos masivos y, en consecuencia, es necesario reconocer la oportunidad estratégica de formar recursos humanos en ésta área de conocimiento. Sin embargo, al igual que toda tecnología emergente, el tema de gestión de grandes datos demora en llegar a la currícula de las carreras de grado. Mientras tanto, se genera una brecha entre lo que la ciencia/sociedad/industria requiere y lo que la instituciones educativas están en condiciones de ofrecer. La forma en que la comunidad académica lidia con esta brecha es fomentando la investigación y desarrollo en esos temas. Esta línea de I/D/I corresponde al diseño, desarrollo e implementación de proyectos que fortalecerán la capacidad científico-tecnológica necesaria para abordar problemas en el campo de la gestión y aprovechamiento de datos masivos (Big Data) que sean de interés para el desarrollo regional y nacional, sobre la base del recurso humano presente en la Universidad Nacional de Chilecito (UNdeC).

Palabras clave: *Big Data, Bases de datos analíticas-columnares, información no estructurada (NoSQL), visualización de grandes volúmenes de datos*

Contexto

El equipo de trabajo ha presentado un proyecto en esta línea en la convocato- ria "Proyectos de Investigación Científica y Tecnológica 2016, Plan Argentina Innovadora 2020" en la categoría Equipo de Reciente Formación (PICT-2016-4293). Además, durante el año en curso se adquiri- rá equipamiento tecnológico solicitado pa- ra dar soporte a esta línea con fondos del PROMINF.

Dos de los miembros dirigen proyectos vinculados a esta línea de trabajo, aprobados por la UNdeC en la convocatoria 2013-2014 del programa "Financiamiento para el Estímulo y Desarrollo de la Investiga-

ción Científica y Tecnológica". Es importante destacar que el estudiante involucrado en el proyecto ha sido beneficiado con una beca de Estímulo a las VocacionesCientíficas, convocatoria EVC 2016(CIN).

Introducción

La caracterización típica de un problema de Big Data es si cumple con un gran volumen, con variedad de fuentes (tanto estructuradas o no), requiere velocidad y frecuencia de las actualizaciones y con veracidad de la información [1]. A continuación, se presentan posibles orígenes de problemas de datos masivos presentes en la comunidad regional pero también a nivel nacional.

Actividades agroindustriales. El clima de la región se caracteriza por la extre- ma aridez, con grandes amplitudes térmi- cas, escasas lluvias anuales 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 va- lles áridos, lleva adelante una intensa actividad agrícola industrial. La mayor concentración de cultivos en el subsector de fruticultura superando las 20000 hectáreas lo tienen el olivo, la vid y el nogal, los cua- les se comercializan a nivel local, regional, nacional e internacional. El proceso de industrialización de algunos cultivos comoel de la vid se lleva adelante en 15 bodegas que se distribuyen en la ciudad y distritos del Departamento Chilecito, mientras que la fabricación de aceite de oliva se encuentra en pleno crecimiento. El sector mantiene un estrecho vínculo con la universidad, dispuesto a colaborar en trabajos de investigación y transferencia tecnológica. En este sentido, existe un gran potencial de análisis de grandes volúmenes de datos (control de plagas, monitorización del crecimiento de cultivos, gestión de datasets agrometeorológicos, etc.).

Laboratorio de Altura y Laboratorio de Alta Complejidad. La UNdeC cuenta con dos importantes laboratorios que ofrecen servicios a la comunidad y que representan un enorme potencial de trabajo con grandes volúmenes de datos y procesos complejos con requerimientos de tiempo real. El *Laboratorio de Altura* es el primero de América de esta clase. Se encuentra a 5200 metros sobre el nivel del mar, es de fácil acceso y posee excelentes condiciones atmosféricas. Este laboratorio permite realizar mediciones imposibles de hacer a nivel del mar en campos como medicina, biología, astronomía, física, etc. Por otro lado el *Laboratorio de Alta Complejidad* 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, entre otros. 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 al- ta resolución, lo que permitirá ampliar los servicios ofrecidos.

Servicios de información con valor agregado. Es posible generar información con valor agregado para la comunidad a partir de un gran volumen de datos que está disponible públicamente en Internet. Una expresión importante del fenómeno conocido como sociedades de la información y el conocimiento [2, 3] son las redes sociales Twitter, Facebook, Instagram, Snapchat, Whatsapp, entre otras [4, 5]. Los datos que circulan a través de ellas pueden ser transformardos en información de relevancia y utilidad sobre tendencias de consumo, pensamientos políticos, ideologías, preferencias y costumbres [6, 7, 8]. Por ejemplo, es posible predecir tendencias en elecciones a partir de un análisis de opinión de los

datos disponibles en estas redes [9, 10,11]. Otro enfoque en pleno crecimiento consiste en relacionar datos abiertos [12]difundidos por organizaciones públicas o privadas en búsqueda de ofrecer nuevos servicios a la comunidad. En Argentina varios organismos gubernamentales comenzaron a promover el uso de Datos Abiertos, con el objetivo de facilitar información a los ciudadanos para su consulta y libre uso [13, 14]. Sin embargo, también aparecen expresiones en el sector privado que utilizan es- tos recursos para ofrecer mejores servicios. Como ejemplo se puede mencionar el proyecto del diario La Nación Data [15] que utiliza datos abiertos para generar noticias que han tenido gran impacto en la comunidad, o emprendimientos como Properati Data [16], una inmobiliaria con un modelo de negocios basado en el cruce de información con datasets públicos en búsqueda de atraer la atención de clientes.

Líneas de Investigación, Desarrollo e Innovación

Bases de datos, Minería de datos, ba- ses de datos columnares; visualización de grandes volúmenes de datos; sistemas de procesamiento estadístico; Cloud Computing, Arquitecturas paralelas; Análisis social web; Simulación; Internet de lascosas; Agromática, Vitivinicultura, Genética; Repositorios institucionales y bibliotecas digitales, Análisis semántico de la información

Resultados y Objetivos

Desde el año 2008 se coopera con empresas del sector agrícola dedicadas especialmente al cultivo de olivo, lo que dio lugar a los primeros proyectos de investigación con financiamiento interno de la UNdeC para su desarrollo. Estos proyectos trabajaron sobremonitorización, planificación y automatización del riego de los cultivos, y su ejecución derivó en desarrollos de software a medida que continúan siendo utilizados. Las relaciones de cooperación entre las empresas del sector y los miembros del grupo representan una potencial fuente de problemas de datos masivos. Es de esperar que como parte de la ejecución de esta línea sea posible llevar esos desarrollos a gran escala.

Uno de los miembros está desarrollando su tesis de maestría en el tema “Mejora de la precisión posicional utilizando receptores GPS de bajo costo”. Es un tema de gran impacto para la región debido a queel único tipo de agricultura sustentable es la de precisión, y esta sólo es posible con la tecnología adecuada. Posee financiamiento propio y está indirectamenterelacionado con esta línea a través de los servicios que pueden ser ofrecidos al sector agrícola.

En el contexto de esta línea, se está desarrollando un trabajo final de grado en el tema “Ambiente colaborativo para mejorar las prácticas de viticultores independientes en la zona de Pituil”. Se espera organizar y estructurar la captura y almacenamiento de la información que circula por la red para ser puesta al servicio de la comunidad científica y académica a través de datasets públicos. Se está trabajando con investigadores, profesionales y productores del sector, lo que sienta las bases para futuros trabajos en el área. Se prevé su finalización para julio de 2017.

Otro de los miembros está desarrollando su tesis de maestría con el tema “Métricas de calidad de los datos obtenidos en un Sistema de Integración de Datos”. El trabajo destaca el efecto negativo que pueden tener los distintos tipos de inconsistencias para la integración de datos y propone definir un conjunto de métricas basadas en distintos algoritmos de detección de inconsistencias para predecir la calidad de la información integrada. Estas métricas serán luego utilizadas en un framework para la publicación de datasets públicos. Se espera con es-

te trabajo avanzar en una infraestructura para dar soporte a otros proyectos vinculados a la administración pública para la publicación de datos abiertos (universidad, municpalidad, poder judicial, consejo deliberante, etc.). Está planificada la finalización de la tesis para marzo de 2018.

Se está trabajando con la empresa cordobesa Kunan S.A., sobre la idea de predecir comportamientos sociales a partir de análisis de sentimiento en redes sociales. Como resultado de esta colaboración se presentó un artículo sobre el Balotaje Argentina 2015 en las IV Jornadas de Cloud Computing & Big Data 2016. Para el trabajo se adquirieron los comentarios vertidos voluntariamente en la red social twitter por los usuarios referidos al balotaje presidencial con el agente Apache Flume de Hadoop, y se utilizó el motor de base de datos Vertica con su componente Pulse, y el software de visualización Tableau. Para el análisis de correlación de los datos se utilizó Stata.

En esta misma línea a finales de 2016 dos docentes en colaboración con Kunan S.A. utilizaron las técnicas de análisis de sentimiento en tweets para determinar el interés energético de sus usuarios a nivel mundial, catalogadas por tipo de energías. Este trabajo fue realizado para plantas generadoras de energía, y permitió conocer rápidamente la opinión de la población sobre el tema, determinar competidores y en función de ello proyectar inversiones en zonas con mayor oportunidad de aceptación. Igual que en el caso anterior, se emplearon Vertica, Tableau y Stata. Los resultados finales aún están pendientes de publicación.

Dos miembros del equipo ofrecieron un curso de Vertica durante las "VIII Jornadas de Informática y Comunicaciones 2016". A partir de este año, estos miembros comenzarán a dar una asignatura sobre Big Data en las carreras ofrecidas por la UNdeC.

Objetivos

Crear dentro del marco de la UNdeC la capacidad científico-tecnológica necesaria para abordar problemas en el campo de la gestión y aprovechamiento de datos masivos (Big Data) que sean de interés para el desarrollo regional y nacional.

- Definir los requerimientos de una plataforma de experimentación, desarrollo y producción de soluciones a problemas de datos masivos.
- Desarrollar una infraestructura acorde a los requerimientos anteriores.
- Estudiar técnicas y herramientas para la gestión y aprovechamiento de datos masivos.
- Difundir a nivel regional y nacional la potencialidad de trabajo del equipo.
- Explorar oportunidades de colaboración con la comunidad académica, productiva y en general que deriven en problemas de datos masivos.
- Canalizar esas oportunidades a través de trabajos de finalización de grado y tesis de postgrado.
- Promover el diseño y desarrollo de algoritmos paralelizados orientados a la optimización de problemas de cómputo con grandes volúmenes de datos.
- Consolidar un grupo de investigación multidisciplinario en la UNdeC.

Formación de Recursos Humanos

El equipo de trabajo está formado por: Dos doctores especializados en repositorios institucionales, bibliotecas digitales, desarrollo de software, cómputo paralelo y tecnología grid. Tres estudiantes de maestría en informática en su etapa final, dos de los cuales trabajan en temas relacionados con esta línea. Un estudiante de grado, el cual

presentará su trabajo final en julio en un tema directamente relacionado a esta línea. Tres de los miembros están categorizados en el programa de incentivos. Además, el grupo mantiene vínculos de colaboración con los miembros del III-LIDI (UNLP) y con los miembros del grupo ARTECS de la Universidad Complutense de Madrid.

Referencias

[1] M. Tascón, "Introducción: Big Data. Pasado, presente y futuro," *Telos: Cuadernos de comunicación e innovación*, no. 95, pp. 47–50, 2013.

[2] I. Murua Anzola, M. L. Cacheiro Gonzá- lez, and D. J. Gallego Gil, "Las cibercomunidades de aprendizaje (cca) en la formación del profesorado," *RED.*, vol. XIII, no. 43, p. 29, 2014.

[3] M. Meirinhos and A. Osório, "Las comunidades virtuales de aprendizaje: el papel central de la colaboración," *Pixel-Bit. Revista de Medios y Educación*, no. 35, pp. 45–60, 2009.

[4] A. G. Sans, "Las Redes Sociales co- mo Herramientas para el Aprendizaje Colaborativo: Una Experiencia con Facebook," *Re-Presentaciones: Periodismo, Comunicación y Sociedad*, no. 5, pp. 48–63, 2009.

[5] R. V. Argüelles, "Las redes sociales y su aplicación en la educación," *Revista Digital Universitaria*, vol. 14, no. 4, pp. 1–14, 2013.

[6] M. Gil Mediavilla, V. Ausín Villaverde, and F. Lezcano Barbero, "Redes sociales educativas como introducción a los entornos personales de aprendizaje (PLE's)," *EduSer-Revista de educação*, vol. 4, no. 1, pp. 17–29, 2012.

[7] J. Lorca and L. Pujol, "Redes socia- les: descripción del fenómeno, situa- ción actual y perspectivas," *RevistaeSa- lud. com–Fesalud. Fundación para la eSalud*, vol. 4, no. 15, pp. 1–15, 2008.

[8] V. Miguel, M. Fernández, E. V. de Enseñanza Aprendizaje, and P. de Gestión, "Redes Sociales y Construcción del Conocimiento," *AB Martinez y N. Hernán- dez "Comunidades Virtuales de Aprendizaje". Caracas, Venezuela. Consejo de Desarrollo Científico y Humanístico, Universidad Central de Venezuela*, 2013.

[9] L. Deltell Escolar, "Predicción de tenden- cia política por Twitter: Elecciones Anda- luzas 2012," *Ambitos: Revista internacio- nal de comunicación*, no. 22, pp. 91–100, 2013.

[10] A. Ceron, L. Curini, and S. M. Iacus, "Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters– Evidence From the United States and Italy," *Social Science Computer Review*, vol. 33, no. 1, pp. 3–20, Feb. 2015.

[11] S. Unankard, X. Li, M. Sharaf, J. Zhong, and X. Li, *Predicting Elections from So- cial Networks Based on Sub-event De- tection and Sentiment Analysis*, ser. Lec- ture Notes in Computer Science, B. Be- natallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, Eds. Springer International Publishing, Oct. 2014, no. 8787.

[12] "El manual de Open Data." [Online]. Available: http://opendatahandbook.org/guide/es/

[13] S. Fumega, "Opening Cities: Open Da- ta in Buenos Aires, Montevideo and Sao Paulo," Exploring the Emerging Impacts of Open Data in Developing Countries (ODDC), Tech. Rep., Apr. 2014.

[14] "Datos Argentina." [Online]. Available: http://datos.gob.ar

[15] "LA NACION Data - LA NACION." [Online]. Available: http://www.lanacion.com.ar/data

[16] "Properati Data." [Online]. Available: http://www.properati.com.ar/data