UN ESTADO ACTUAL SOBRE CIENCIA DE DATOS Y BIG DATA

Claudio Carrizo* [(1)]; Fernando Cardona[(1)]; Raúl Navarro Peláez[(1)]; Sofía Racca[(1)]; Facundo Barrera[(1)]
Pablo Vacca[(2)]

[(1)] Facultad Regional San Francisco – Universidad Tecnológica Nacional - Av. de la Universidad 501 San Francisco – Pcia. De Córdoba - Tel. 03564-421147
[(2)] Facultad Regional Córdoba – Universidad Tecnológica Nacional - Maestro M. Lopez esq. Cruz Roja, Córdoba Capital - Pcia. De Córdoba - Tel. 0351-5986000

INTRODUCCIÓN

Actualmente existe un creciente interés en las organizaciones por extraer información y producir conocimiento a partir de la cantidad masiva de datos creados diariamente (Fayyad et al., 2017). La disponibilidad de Big Data permite a las organizaciones de todas las industrias aprovechar el análisis de datos, con el fin de extraer conocimiento procesable que puede utilizarse para la toma de decisiones y predicciones comerciales sólidas (Molina-Solana, M. et al., 2017). Al utilizar Big Data, el análisis empresarial abre el potencial predictivo del análisis de datos para mejorar la gestión estratégica, la eficiencia operativa y el rendimiento financiero (Newman, R. et al., 2017). Pero no es solo la masividad lo que hace que todos estos datos nuevos sean interesantes o plantee desafíos (Van der Aalst, W.M., 2016), son datos en sí, y su comportamiento en tiempo real (Rupp, G.M. et al., 2017), los convierten en componentes básicos en la búsqueda de conocimiento.

Una característica importante de los datos es la alta diversidad con la que cuentan. Estos pueden ser desde los tradicionales: numérico, categórico o binario hasta más complejos como los son: texto (correos electrónicos, tweets, artículos científicos, comentarios), registros (datos a nivel de usuario, datos de eventos con marcas de tiempo, logs), datos de ubicación geográfica, red, sensores o imágenes. Por tanto, los principales desafíos científicos que dan paso al surgimiento la ciencia de los datos, están dados por la necesidad de analizar datos diversos, incompletos y desordenados; conjuntos de datos muy grandes, que cambian en el tiempo (Kormos, M. et al., 2017), y la necesidad de encontrar hallazgos que impulsen decisiones sobre operaciones y productos en las organizaciones.

Hoy, la ciencia de datos está entrando en una nueva era, donde la tecnología de la información ahora es capaz de soportar negocios basados en datos, en tiempo real (Norbert, D., Andreas et al., 2017) con el fin de facilitar decisiones informadas basadas en evidencia científica confiable para proporcionar herramientas a los responsables de las políticas y las decisiones (Dalkir, K. et al., 2017). Según (Naivy Pujol Méndez et al., 2018), la ciencia de datos es el campo interdisciplinario con bases en informática, estadística y matemáticas, que se ocupa de la teoría, la práctica y la comunicación de los resultados con el fin de extraer conocimiento relevante de los datos. En la Figura 1 se ilustra la relación de las diferentes disciplinas (informática/ciencias de la computación, matemáticas y estadística, y dominio específico del negocio) con la ciencia de datos.

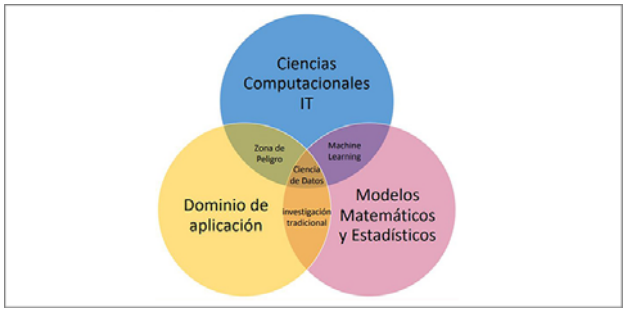

Figura 1: Disciplinas Ciencia de Datos (Ayankoya, K. et al., 2014).

Por lo tanto, la Ciencia de Datos es un nuevo campo que implica nuevos especialistas, con habilidades muy variadas. A las personas que se dedican a la Ciencia de Datos se las conoce como "Científico de Datos", los cuales son expertos y tienen la capacidad de poder extraer un valor significativo de los datos y también administrar todo el ciclo de vida de los datos. Deben poseer habilidades en las disciplinas relacionadas con la ciencia de datos (Schutt, R. et al., 2013); ser capaces de estudiar las diversas fuentes de información disponibles en una organización; extraer datos a partir de diversos formatos (Blei, D.M. et al., 2017); depurarlos, analizarlos, idear y desarrollar algoritmos; realizar inferencias, preparar y comunicar los resultados de dichos análisis y transmitir conclusiones que ayude a tomar mejores decisiones (Cao, L., 2017). En la industria, el papel del científico de datos se está convirtiendo rápidamente en una carrera muy solicitada y solicitada. Un número creciente de empresas, como Google, Facebook, IBM, PayPal y Amazon, también están buscando científicos de datos para unirse a sus equipos de ciencia de datos y ayudarlos a mantener una ventaja innovadora en la era de los grandes datos (Yangyong Zhu and Yun Xiong, 2015). Big Data (en español, grandes datos) es un conjunto de datos a gran escala que no puede ser procesado y analizado por técnicas tradicionales y métodos dentro de un tiempo aceptable (Yang Zhao-hong

et al., 2015). A las características que más identifican a Big Data se las conoce como el modelo de la 3Vs (Volumen, Velocidad, Variedad) (B.Gerhardt, K. Griffin and R. Klemann, 2012), aunque en los últimos años han surgido otros modelos que incluyen más características, como el modelo de las 5Vs (Cheng Xue-qi et al.), 7Vs (Harshawardhan S. et al., 2014) y 10Vs (George Firican).

Uno de los grandes desafíos que afrontan la Ciencia de Datos y el Big Data, es la falta de personas que tengan experiencia y habilidades en el manejo de plataformas de Big Data y el análisis de datos.

MÉTODOS

Para la elaboración del estado actual, se llevaron a cabo 3 actividades:

Actividad 1: Desarrollo de una Revisión Sistemática de la Literatura (RSL) sobre las temáticas Ciencia de Datos y Big Data: se llevó a cabo en primera instancia una revisión sistemática de la literatura con el objetivo de poder relevar artículos de los últimos 5 años, que estén relacionados directamente con la Ciencia de Datos y Big Data, y que provengan de fuentes calificadas. Del total de artículos relevados, se hizo una selección a través de la relevancia del artículo según su Título, Resumen, Introducción y Conclusiones. Finalmente, de los artículos seleccionados, se hizo una lectura en profundidad acerca del desarrollo de cada artículo.

Actividad 2: Elaboración del Estado del Arte de Ciencia de Datos y Big Data: se llevó a cabo con el resultado obtenido en la Actividad 1.

Actividad 3: Elaboración del estado actual de Ciencia de Datos y Big Data: Con el resultado de la Actividad 2, se llevó a cabo un estado actual acerca de las 2 temáticas propuestas.

RESULTADOS

De las actividades definidas, hasta el momento se han podido alcanzar los siguientes resultados:

- Una Revisión Sistemática de la Literatura, en donde finalmente se seleccionaron 30 artículos para leer en profundidad y que están relacionados directamente con las temáticas abordadas.
- Se realizó la construcción de un repositorio bibliográfico con 30 artículos relevantes con una actualidad de hasta 5 años.
- Se elaboró un documento que contiene estado del arte de Ciencia de Datos y Big Data.
- Se elaboró un documento que contiene el estado actual de las temáticas propuestas.

CONCLUSIONES

Este trabajo consistió en presentar el resultado de un estado actual realizado sobre la Ciencia de Datos y el Big Data, el cuál fue producto de una Revisión Sistemática de la Literatura y un estado del arte, desarrollado en el marco del PID-UNT 4567.

REFERENCIAS

Ayankoya, K., Calitz, A., Greyling, J., 2014. Intrinsic Relations Between Data Science, Big Data, Business Analytics andDatafication, in: Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology, SAICSIT '14. ACM, New York, NY, USA, p. 192:192–192:198.

Blei, D.M., Smyth, P., 2017. Science and data science. Proc. Natl. Acad. Sci. 114, 8689–8692.

B.Gerhardt, K. Griffin and R. Klemann, "Unlocking Value in the Fragmented World of Big Data Analytics", Cisco Internet Business Solutions Group, June 2012.

Cao, L., 2017. Data Science: A Comprehensive Overview. ACM Comput Surv 50, 43:1–43:42.

Cheng Xue-qi, Jin Xiao-long, Wang Yuanzhuo, et al.A Literature Review on Big Data System and Analysis Technology[J].Journal of Software, 2014(9):1889-1908.

Dalkir, K., Beaulieu, M., 2017. Knowledge Management in Theory and Practice. MIT Press.

Fayyad, U.M., Simoudis, E., Srivastava, A., 2017. Foreword to the Applied Data Science:Invited Talks Track at KDD-2017, in:Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,KDD 17.ACM,New York, NY, USA, pp. 7–8.

George Firican. The 10 Vs of Big Data. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx.

Kormos, M., Collura, M., Takács, G., Calabrese, P., 2017. Real-time confinement following a quantum quench to a nonintegrable model. Nat. Phys. 13, 246.

Harshawardhan S. Bhosale, Devender P. Gadekar,"A Review paper on Big Data and Hadoop", International Journal of Scientific and Research Publication, Vol 4, 2014.

Molina-Solana, M., Ros, M., Ruiz, M.D., Gómez-Romero, J., Martin-Bautista, M.J., 2017. Data science for building energy management: A review. Renew. Sustain. Energy Rev. 70, 598–609.

Naivy Pujol Méndez, Joelsy Porven Rubier, 2018. Ciencia de datos: una revisión del estado del arte. Universidad de las Ciencias Informáticas (UCI). Habana, Cuba

Newman, R., Chang, V., Walters, R.J., Wills, G.B., 2016. Model and experimental development for Business Data Science. Int. J. Inf. Manag. 36, 607–617. https://doi.org/10.1016/j.ijinfomgt.2016.04.004.

Norbert, D., Andreas, G., Armin, K., Manuel, M., Andrea, H., 2017. Solutions for Cyber-Physical Systems Ubiquity. IGI Global.

Rupp, G.M., Opitz, A.K., Nenning, A., Limbeck, A., Fleig, J., 2017. Real-time impedance monitoring of oxygen reduction during surface modification of thin film cathodes. Nat. Mater. 16, 640. https://doi.org/10.1038/nmat4879.

Schutt, R., O'Neil, C., 2013. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc.

Van der Aalst, W.M., 2016. Process mining: data science in action. Springer.

Yang Zhao-hong, Wang Hui-yu, Zhao Bin, Han Zhi-he, Lu Wan-lin. A Literature Review on the Key Technologies of Processing Big Data. 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis.

Yangyong Zhu and Yun Xiong, 2015. Towards Data Science. Data Science Journal, 14: 8, pp. 1–7, DOI.