CLUSTER PARA APRENDIZAJE Y PRÁCTICA DE *BIGDATA* Y SERVICIOS DE *LEARNING ANALYTICS*

Analia N. Herrera Cognetta, Nilda M. Pérez Otero, Francisco N. Colarich, Gustavo D. Castillo, Dalila J. Mamani, Mauro R. Patagua, Natalia E. Rodriguez, Roque E. Talavera y Diego M. Verrastro
Facultad de Ingeniería – Universidad Nacional de Jujuy – Argentina
{aniherrera012, nilperez, fcolarich, castilloguty, daly.jaquelin, mpatagua, nataliarod314, ema.tala015} @gmail.com, diegoxtr@hotmail.com

RESUMEN

Big Data es la nueva generación de almacenamiento, análisis de datos y de negocios. En el fenómeno del *Big Data* se considera, como fuentes de generación de datos a personas, *smartphones*, equipos que generan data de proyectos y experimentos, y, principalmente, a Internet. La tecnología *Big Data* permite recolectar, almacenar y preparar grandes volúmenes de datos, para analizar o visualizar la relación entre ellos, inclusive a partir de datos que se generan en tiempo real y provienen de redes sociales, sensores, dispositivos de diversa índole o fuentes de audio y video. En esta temática, el presente proyecto pretende, en una primera etapa, investigar tecnologías y arquitecturas para implementar un clúster en una plataforma web, donde estudiantes de la Facultad de Ingeniería de la UNJu, puedan aprender y practicar *Big Data*; en una segunda etapa se comenzará con una investigación en Analítica de Datos para evolucionar a *Learning Analytics*, utilizando algoritmos para medición, recopilación, análisis e informe de datos, sobre estudiantes universitarios y sus contextos, a fin de comprender y optimizar el aprendizaje y entornos en que se produce.

Palabras clave: Arquitectura escalable, *Big Data*, Plataforma web, Análisis predictivo de datos, Analítica de aprendizaje, *Learning Analytics*.

CONTEXTO

La línea de investigación aquí presentada se encuentra inserta en el proyecto **Cluster para aprendizaje y práctica de *Big Data* y servicios de *Learning Analytics***, ejecutado a partir de enero del corriente año por el grupo de la Facultad de Ingeniería de la Universidad Nacional de Jujuy. El proyecto, acreditado y financiado por la Secretaria de Ciencia y Técnica y Estudios Regionales de la Universidad Nacional de Jujuy, se encuentra bajo el Programa de Incentivos.

1. INTRODUCCIÓN

El *Big Data* es un concepto que hace referencia al tratamiento y análisis de enormes repositorios de datos, tan grandes, que resulta muy difícil y hasta imposible manejarlos con las herramientas de bases de datos y analíticas convencionales [1].

Esta tecnología se ocupa de las actividades relacionadas con los sistemas que manipulan grandes conjuntos de datos. Las dificultades vinculadas a la gestión de datos se centran en la recolección, almacenamiento, búsqueda, compartición, análisis y visualización de la información. El principal objetivo de manipular enormes cantidades de datos es, poder emplear dicha información en la creación de informes estadísticos y modelos predictivos, que pueden ser utilizados en muchas áreas del quehacer humano [2].

En el ámbito educativo, esta tecnología ha comenzado a brindar grandes beneficios. Por ejemplo, en la mejora de la gestión educativa, al desarrollo de nuevos métodos para la enseñanza y el aprendizaje, la creación de nuevas carreras y opciones profesionales, así como en la generación y almacenamiento de acervos digitales, que constituyen el producto de años de actividad académica, docente y de investigación.

Actualmente, se están gestando nuevos métodos sustentados en la tecnología para poder hacer el seguimiento de alumnos, mejorar su rendimiento, obtener datos objetivos de sus evaluaciones, predecir los riesgos académicos, o simplemente comprender el comportamiento de los grupos escolares. En este continuo cambio los profesores están más conscientes de la necesidad de actualizarse tecnológicamente, para ofrecer una educación más efectiva y adecuada a las necesidades de la población escolar actual. Es por ello que la analítica del *Big Data* se ha convertido en el recurso clave actual para entender y mejorar el proceso educativo [2].

Learning Analytics

Con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce el proceso educativo, no solo es importante conocer a los estudiantes, también es esencial analizar los datos sobre sus acciones y contextos de aprendizaje [3].

La analítica del aprendizaje o *Learning Analytics* (LA) es un término relativamente nuevo que ha ido creciendo en los últimos años por cuatro razones principales [4]:

1. **Aumento sustancial de datos**: la cantidad de datos disponibles en los contextos educativos, se obtienen a través de dispositivos digitales y tecnologías en línea, que permiten la captura y posterior análisis de la información.
2. **Formatos de datos mejorados**: en la actualidad hay formatos estandarizados para el registro de datos educativos.
3. **Avances en la computación**: la analítica es impulsada también por los avances de la computación, que permiten el análisis de grandes cantidades de datos.**Incremento de herramientas más sofisticadas para el análisis**: como es el caso de *Google Analytics* que desarrolló *MapReduce*, base importante para el proyecto *Hadoop*.

La LA puede ser de varios tipos [5]:

- **Descriptiva**: destinada a la agregación de datos y generación de información y visualizaciones relacionadas con los eventos y la interacción de los estudiantes con los cursos.
- **Predictiva**: destinada al desarrollo de modelos estadísticos y aprendizaje automático para, en función de la información disponible (cruda o tras la analítica descriptiva), obtener una visión lo más precisa posible de las potenciales situaciones a tratar en el futuro. Entre sus aplicaciones, se destaca la detección de estudiantes en riesgo de abandono y la predicción de resultados.
- **Prescriptiva**: nutrida a partir de la información proporcionada por los anteriores procesos de análisis y destinada a dar respuesta a sus resultados a través de funcionalidades como orientar a los estudiantes que están en riesgo de abandono para evitar que eso suceda.

Aspectos técnicos para el uso de *Big Data*

Una de las principales limitaciones para la adopción de la tecnología *Big Data* en cualquier institución son los recursos económicos, de infraestructura y de capital humano experimentado, que se requieren para su instrumentación. Sin embargo, están comenzando a aparecer alternativas relativamente económicas para poder realizar analítica con *Big Data,* y herramientas especializadas diseñadas para facilitar diferentes aspectos de esta tecnología. Por ejemplo, *Hadoop* es un *framework* de código abierto, diseñado para el almacenamiento y procesamiento a gran escala de conjuntos de datos en un gran número de máquinas, que permite la creación de aplicaciones para procesar grandes volúmenes de información, haciendo uso de un modelo sencillo de programación; es escalable y ofrece un buen nivel de tolerancia a fallos [6]

En este sentido, *Apache Hadoop* permite resolver varios aspectos importantes como [2]:

- Establecer el punto de contacto con la ciencia de los datos (data science) que provee
- Las técnicas necesarias para manipular y tratar la información desde un punto de vista estadístico/matemático.

- Hacer posible el procesamiento distribuido de grandes conjuntos de datos en arreglos de computadoras (clusters), utilizando modelos sencillos de programación.
- Escalar la infraestructura de cómputo, es decir, utilizar desde unos pocos servidores (nodos) y sistemas de almacenamiento de información, hasta miles de equipos, todos ellos ofreciendo una calidad idéntica de servicio.
- Permitir a las aplicaciones desarrolladas trabajar con miles de nodos y petabytes de datos, tantos como necesidades de análisis de información se requieran.
- Disponer de facilidades para realizar consultas avanzadas sobre los conjuntos de datos.
- Usar extensiones que facilitan el trabajo, manipulación y seguimiento de toda la información que maneja.
- Permite ejecutar procesos en paralelo en todo momento.

Se trata, en definitiva, de un *framework* de software de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos. Éste se orienta hacia la computación distribuida, en la cual la escalabilidad y la fiabilidad son los dos atributos más importantes. En otras palabras, *Hadoop* es el complemento perfecto de *Big Data*.

2. LINEAS DE INVESTIGACION Y DESARROLLO

El proyecto presenta dos etapas bien diferenciadas: en una primera etapa (aprendizaje) se investigará tecnologías software y arquitecturas hardware para el desarrollo y la implementación de un clúster, en una plataforma web que servirá de laboratorio a estudiantes de la Facultad de Ingeniería de la UNJu, para la práctica y aprendizaje de *Big Data*. La segunda etapa, capacitación, investigación y desarrollo en analítica de datos, se iniciará con investigación en Analítica de Datos para evolucionar a *Learning Analytics*, y analítica masiva de datos personalizados, utilizando algoritmos con herramientas orientadas por teorías del aprendizaje, técnicas pedagógicas y algoritmos para medición, recopilación, análisis e informe de datos sobre estudiantes universitarios y sus contextos, a fin de comprender y optimizar el aprendizaje y entornos en que se produce.

3. RESULTADOS OBTENIDOS/ESPERADOS

Es una realidad hoy la necesidad de que los estudiantes de las carreras vinculadas con las TICs, adquieran los conocimientos y capacidades sobre las últimas tecnologías, en la Institución Universitaria donde cursan. El contar con una herramienta que pudiera reforzar el conocimiento y apoyar el desarrollo de las habilidades prácticas en *Big Data*, será de gran utilidad tanto para estudiantes como para docentes.

Este proyecto ambiciona obtener un cluster que sirva como laboratorio para estudiantes de la Facultad de Ingeniería de la UNJu, y que a través de un proceso de madurez sea posible implementar estudios de analítica de datos para contribuir con la formación a través de *Learning Analytics*.

Por ello, el objetivo del proyecto consiste en lograr, a través de la investigación de nuevas tecnologías, diseñar, desarrollar e implementar un *cluster* para estudio y práctica de *Big Data*, para estudiantes de grado, alcanzando un nivel de madurez suficiente para el estudio y aplicación de *Learning Analytics*, siendo los objetivos particulares:

- Investigar y evaluar las herramientas del ecosistema Hadoop, para diseño y desarrollo.
- Implementar el entorno de trabajo que cumpla con los requerimientos del diseño.
- Diseñar y realizar las pruebas de usabilidad del clúster.
- Capacitación del equipo en Analítica de Datos.
- Brindar servicios de *Learning Analytics* para la toma de decisión.

Al tratarse de un proyecto que recién inicia (enero de 2020 a diciembre de 2021), no se cuenta aún con resultados obtenidos.

4. FORMACION DE RECURSOS HUMANOS

El grupo de investigación está formado por dos docentes, dos egresados y cinco estudiantes de las carreras de Licenciatura en Sistemas e Ingeniería Informática, todos de la Facultad de Ingeniería de la Universidad Nacional de Jujuy. Los integrantes del grupo fueron elegidos según sus especialidades y perfiles, un egresado tiene un Mg en Innovación Educativa por Competencias, quien aportará su conocimiento en esta temática para Learning Analytics, otro egresado trabaja actualmente en una empresa de desarrollo de software y cada estudiante se destaca por su compromiso y cumplimiento con respecto a la planificación propuesta, y en la profundidad y nivel académico demostrado en la investigación preliminar. Si bien las docentes cuentan con trayectoria en investigación, para los egresados y los estudiantes, es su primera experiencia, lo que representa un desafío y una oportunidad, desafío para las docentes de dirigir y organizar las actividades y el aprendizaje en investigación, para que resulte una vivencia enriquecedora, y que a la vez incentive a los jóvenes a seguir esta línea de trabajo, Una oportunidad para ellos de adquirir hábitos y metodologías para la investigación aplicada, que puede llegar a convertirse en su quehacer cotidiano, o la expertiz lograda, colabore en su desempeño profesional. Las carreras que estos estudiantes eligieron, tienen como característica fundamental, la evolución continua de métodos, herramientas, equipos y desarrollos, sobre los que deben tener conocimiento y manejo para mantenerse activos en el mercado laboral.

5. REFERENCIAS

[1] DANS, Enrique, "Big Data: una pequeña introducción", 2011. En línea: < https://www.enriquedans.com/2011/10/big-data-una-pequena-introduccion.html >, consultado en marzo de 2020.

[2] SALAZAR ARGONZA, Javier, "Big Data en la educación", Revista Digital Universitaria, 1 de enero de 2016, Vol. 17, Núm. 1, ISSN: 1607-6079 Disponible en Internet: <http://www.revista.unam.mx/vol.17/num1/art06/index.html>

[3] LANG, Charles, *et al.* (ed.). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research, 2017.

[4] SIEMENS, George; BAKER, Ryan SJ d. Learning analytics and educational data mining: towards communication and collaboration. En *Proceedings of the 2nd international conference on learning analytics and knowledge*. 2012. p. 252-254.

[5] ROLDÁN, Xavier Alamán, *et al*. GHIA: Modelado de Estudiantes, Analítica de Aprendizaje, Atención a la Diversidad y e-Learning. *IE Comunicaciones: Revista Iberoamericana de Informática Educativa*, 2019, n° 30, p. 78-89.

[6] ZENTENO, José Antonio Carrillo. Big Data-Analítica del aprendizaje y minería de datos aplicados en la Universidad. *Pro Sciences*, 2018, vol. 2, no 8, p. 39-54.