Nota Técnica

LABORATORIO DATA FAE. UN IMPACTO EN LA EDUCACIÓN UNIVERSITARIA

DATA LABORATORY AT FACULTAD DE ADMINISTRACIÓN Y ECONOMÍA UTEM. ITS IMPACT IN HIGHER EDUCATION.

Amaru Fernández Durán*
Alejandro Lizama**
Diana López Avilés***
José Tobar Ríos****
Data FAE UTEM*****

1. INTRODUCCIÓN

Las ciencias de la información están avanzando a pasos agigantados, repercutiendo en todos los ámbitos de la sociedad, desde la forma en cómo se relaciona, comunica e interactúa el ser humano hasta en el desarrollo de los sectores económicos y político.

Estos avances y transformaciones tienen origen muchos años atrás, en la llamada Era Digital, que va ligada a las tecnologías de la información y comunicación, iniciando con la *revolución digital* a finales de la década de 1950, y tomando fuerza a finales de la década de 1970, con la adopción y la proliferación de las computadoras y almacenamiento de registros digitales. Cabe mencionar que los primeros atisbos de esta Era Digital fueron vislumbrados en 1936, año en que Alan Turing desarrolló la primera máquina *inteligente*, creando por primera vez un lenguaje sobre la base de reglas. En 1953 se desarrolló el primer transistor capaz de trabajar con corrientes eléctricas controladas, siendo este la base de la tecnología moderna. En 1965 Gordon E. Moore postuló que el tamaño del transistor disminuiría exponencialmente cada año, aseveración confirmada en el tiempo, y que provocó cambios en las máquinas computacionales, desde los servidores y granjas que manejan y procesan volúmenes incuantificables de información, hasta relojes que son capaces de realizar cálculos que tomarían años en efectuarse manualmente, dando origen a la *Era de la Información*.

El internet es la herramienta actual de entretenimiento, comunicación, trabajo y un sinnúmero de actividades, generando millones de datos por segundo, y la oportunidad única para la extracción y comprensión de esta. En este sentido, metodologías que se relacionan bien con data masiva y no estructurada toman un papel protagónico, como los *modelos de análisis supervisado* y *no supervisado* (Aprendizaje Profundo o *Deep Learning*), que se remontan a los años cincuenta y que ahora tienen cabida en el mundo de la estadística *moderno*.

* Ingeniero Informático con Mención en Cs. De la información.
**Técnico Administrador de Redes Computacionales.
***Magíster en Finanzas, Analista Banco Central de Chile, Docente Econometría Aplicada, Big Data, Finanzas Internacionales.
****Ingeniero Civil en Informática Gerente de Innovación en Wholemeaning. Docente Data Analytics.
*****Centro de investigación y análisis de datos de la Facultad de Administración y Economía UTEM.

Entender estos procesos y lograr ser parte de esta transformación digital es una tarea fundamental para estas nuevas generaciones y la nueva forma de educar; por tal motivo, se hace necesario crear una herramienta donde alumnos, investigadores y docentes puedan gestionar, procesar y utilizar datos de manera fácil y oportuna, siendo el Laboratorio DATA FAE el que cumple con estos requisitos.

2. CONTEXTO LABORATORIO DE DATOS

2.1. Repositorio de datos

En la actualidad, el crecimiento de la producción de información ha sido exponencial, principalmente estos se distribuyen en un 80% en no estructurado y en un 20% estructurado en todo el mundo (Marr, 2018). Estos atributos toman importancia en la existencia de los repositorios de data masiva, los cuales necesitan tecnologías para procesar esta información. Algunas de las principales fuentes de data masiva usadas son; datas históricas, sensores (*smart cities*, *internet of the things*), *social networks*, etc. Estas presentan dos dilemas importantes. Por un lado, está el almacenamiento y procesamiento de esta información y, por otro, entender los datos no estructurados. Desde el primer punto nacen las tecnologías de alta performance computacionales, relacionadas con arquitecturas capaces de manipular esta gran cantidad de datos, de las cuales la más usada es la de *procesamiento distribuido*, que utiliza softwares como: Hadoop, Spark y Apache, entre otros. Respecto del entendimiento de los datos no estructurados (segundo dilema), este permite que interactúen modelos estadísticos de *machine learning* o aprendizaje de máquinas, los cuales son capaces de realizar análisis de datos complejos, como textos, imágenes, audios, etc., acercando el comportamiento y racionalidad de los seres humanos a las máquinas (inteligencia artificial).

2.2. Experiencia internacional y nacional

Según los datos proporcionados por la Red Mexicana de Repositorios Institucionales en México (Remeri), el 38% de las instituciones de educación superior cuenta con un repositorio (Rodríguez y Nava, 2013).

La Universidad Autónoma de Nuevo León puso en marcha el proyecto Repositorio Académico Digital de la UANL. Este sistema ofrece difusión de toda la producción intelectual generada en la universidad por medio de la iniciativa de *acceso abierto* (Open Access), teniendo como objetivo incrementar la visibilidad e impacto de las publicaciones, permitiendo estimular la innovación, facilitar el análisis cualitativo y apoyar las tareas de enseñanza-aprendizaje (Serna y Villanueva, 2014).

La Universidad Autónoma de Nuevo León tuvo como uno de sus objetivos principales subir posiciones en el ranking mundial de universidades, que realiza el Laboratorio de Cibermetría del CSIC de España (Webometrics), debido a que esta clasificación valora con especial importancia la adopción de políticas Open Access, que permiten a los investigadores ser citados frecuentemente.

En esta misma línea, existe una Red de Repositorios Latinoamericanos, creada en 2006, que brinda acceso abierto a más de un millón de documentos académicos. La capacidad de búsqueda de esta plataforma web permite relacionar más de 90 repositorios de universidades como la de São Paulo, Nacional Autónoma de México y Buenos Aires (Universidad de Chile, 2018). Por otro lado, en términos más cercanos a lo que hace DATA FAE, están entidades como IMFD, UC-DATA y DATA Observatory[1].

1. Entre otros repositorios que existen en el ámbito nacional e internacional.

IMFD
El Instituto Milenio Fundamentos de los Datos es un centro científico que desarrolla investigación de frontera y multidisciplinaria en torno a los problemas fundamentales en materia de datos.

UC-DATA
Es principal archivo de datos y estadísticas digitalizadas de ciencias sociales de UC Berkeley. El objetivo es apoyar las necesidades de datos de ciencias sociales de los investigadores de UC Berkeley, proporcionando acceso a una amplia gama de datos computarizados de ciencias sociales a profesores, personal y estudiantes de UC Berkeley.

DATA OBSERVATORY
DO es una organización público-privada sin fines de lucro destinada a potenciar al máximo el beneficio que obtenemos de los datos públicos, de valor global y únicos que se están generando en nuestro país.

Por otro lado, el Gobierno chileno también tiene proyectos de digitalización asociados tanto para el sector público como para las pymes, siendo el principal objetivo de estas últimas que incorporen a sus procesos las nuevas tecnologías. Datos.gob.cl es un sitio que contiene un buscador y catálogos con diversas categorías para encontrar conjuntos de información pública del gobierno de manera fácil, debido a que recopila información de los distintos sitios del gobierno. Por otro lado, Data Chile es una plataforma que integra, visualiza y distribuye datos públicos chilenos, con la idea de revelar brechas en servicios públicos e identificar oportunidades de diversificación industrial. Este tipo de plataformas muestra una evolución en Chile en materia tecnológica y manejo de datos en distintos formatos y estructuras, donde la información juega un rol fundamental en la toma de decisiones. Sin embargo, la información en muchas ocasiones se encuentra desfasada, en distintas frecuencias y formatos, lo que impide un rápido análisis estadístico. Por esta razón, es prioritaria la necesidad de oportunidad, frecuencia y longitud de estos datos, ya que son los desafíos que plantean las economías desarrolladas. De este modo, tecnologías computacionales que procesen y almacenen *big data*, permiten desarrollar el interés por comprender las distintas fuentes de información y tipologías de datos, logrando, a través de modelos de Machine Learning, predecir e implementar algoritmos que acercan a la máquina al comportamiento del ser humano (robótica, inteligencia artificial, etc.).

2.3. Laboratorio DATA FAE

En este contexto se define al Laboratorio DATA FAE como un centro de investigación y análisis de datos que contribuye al progreso científico, desarrollo económico y bienestar social.

Los ejes centrales en los cuales se basa son investigación, datos y vinculación (Figura 1).

Figura 1. Ejes de DATA FAE

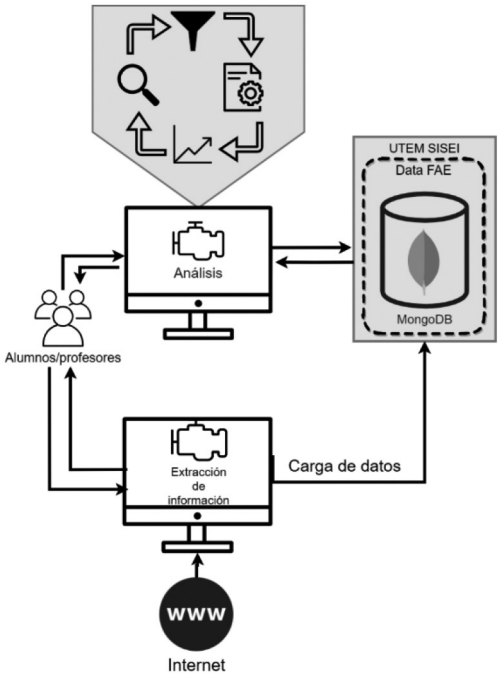
Fuente: elaboración propia.

2.4. Arquitectura tecnológica

El laboratorio DATA FAE cuenta con una infraestructura orientada al manejo de grandes volúmenes de datos no estructurados. El repositorio principal de datos, con una capacidad de almacenamiento inicial de 4 terabyte

de disco y 16 gigabytes de memoria[2], funciona sobre una base de datos no relacional llamada MongoDB. Esta base de datos tiene como principal característica que es orientada a documentos, por lo que no requiere seguir un esquema de almacenamiento de información, lo cual es clave para el trabajo con información no estructurada (Figura 2).

Figura 2. Arquitectura Laboratorio 2020



Fuente: elaboración propia.

El laboratorio se encuentra a disposición de la comunidad universitaria, junto con el acceso a bases de datos consolidadas, las cuales corresponden a las líneas de trabajo que se han ido trabajando: datos históricos del sector inmobiliario, dividendos, riesgo sistémico, entre algunas temáticas.

3. IMPACTO EN LA UNIVERSIDAD TECNOLÓGICA METROPOLITANA

3.1. Cursos asociados

Uno de los objetivos de este proyecto es integrar de manera activa al estudiantado, a través del traspaso de conocimientos de lenguajes de programación, manejo de bases de datos y análisis de los mismos. En este sentido, se incorporaron dentro de la malla curricular dos electivos que van en esta dirección, Big Data y Data Analytics, que intentan interconectar a los alumnos con estas nuevas herramientas, de manera que ellos logren explotar información que pueda ser incluida en el laboratorio para futuras investigaciones docentes, tesis y otros trabajos, creando una red de datos académica amplia, posibilitando la empleabilidad y profundización en áreas tecnológicas.

Las asignaturas se describen de la siguiente forma:

Big Data
Esta asignatura se basa en la extracción, procesamiento y análisis de información de fuentes no estructuradas, y se encuentra en la línea de la ciencia de los datos, rama fundamental para el desarrollo de algoritmos capaces de modelar el comportamiento de la información presentes en la cotidianidad del uso de redes, con el objetivo de construir estadísticas oficiales procedentes de fuentes informales.

Data Analytics
La asignatura de Data Analytics busca que los alumnos aprendan las distintas dimensiones del análisis de datos y lo apliquen de forma práctica a casos reales. Adicionalmente se enseñan distintas técnicas de aprendizaje de máquina y cómo estas se relacionan con el análisis de datos. El curso se apoya con el software RStudio Cloud[3] para todas las unidades.

2. Se está ampliando la capacidad del servidor, esperando que en 2021 cuente con 96 gigabytes.

3. Lenguaje de programación para computar estadística y gráficos a través de un navegador (Véase: https://rstudio.cloud/).

3.2. Web scraping y R como herramientas de extracción y análisis

Desde esta perspectiva se han desarrollado múltiples trabajos mezclando estas herramientas, algunos de estos son : Ecoosfera[4] y TheWeatherChannel[5]

Ecoosfera
Ecoosfera es una página web que nació en un Laboratorio de Conciencia Digital. Su objetivo es expandir el conocimiento de información respecto de la sustentabilidad, creatividad y nuevas formas de entender la realidad. Para ello presenta noticias sobre Arte, Evolución, Guía Gaia, Medio ambiente, Noticias, SCI-Innovación, Wellness, Columna especial.

El trabajo de investigación se basó en determinar cuál de estas últimas categorías es la de mayor interés para el público Ecoosfera, mediante la extracción de compartidos en redes sociales (Facebook, Twitter y Pinterest), a través del conocimiento y uso de HTML[6] y R-Studio[7], con la librería *rvest*[8].

TheWeatherChannel
Esta página se encarga de subir información meteorológica, tanto de los días anteriores como la predicción climática para los próximos días, de distintos puntos geográficos. Dentro de la información que maneja, almacena información de 12 meses atrás y los próximos 14 días, por lo que se obtuvo información de distintos años, generando comparaciones del comportamiento climático y los cambios que estos puedan vivir en sus temporadas en los distintos años. Este análisis fue hecho con R-Studio y manejo de HTML.

4. ALCANCES

DATA FAE pretende trabajar con centros privados de datos y otras universidades del ámbito local y externo, generando recursos educativos, redes de contacto y empleabilidad en los alumnos, permitiendo el desarrollo científico y el traspaso de conocimientos de docentes a estudiantes y de estudiantes a docentes.

4. Véase Ecoosfera en https://github.com/ariamp21/Ecoosfera (Aranda, 2019).

5. Véase TheWeatherChannel en https://github.com/borisff/TheWeatherChannel (Fernández, 2019).

6. Hypertext Markup Language (HTML) es el lenguaje de marcado estándar para documentos diseñados para mostrarse en un navegador web (Berners-Lee y Connolly, 1995).

7. RStudio es un entorno de desarrollo integrado para R, un lenguaje de programación para computar estadística y gráficos (Véase: https://rstudio.com/).

8. Contenedores alrededor de los paquetes *xml2* y *httr* para facilitar la descarga, y luego la manipulación HTML y XML (Wickham, 2019).

REFERENCIAS BIBLIOGRÁFICAS

Berners-Lee, T. y Connolly, D. (1995). Hypertext Markup Language - 2.0. Recuperado de: https://tools.ietf.org/pdf/rfc1866.pdf

Marr, B. (2018). *Data strategy: Cómo beneficiarse de un mundo de big data, analytics e internet de las cosas: como beneficiarse de un mundo de big data, analytics e internet de las cosas.* Bogotá: Ecoe Ediciones

REMERI (2020). Red Mexicana de Repositorios Institucionales. Recuperado de: http://www.remeri.org.mx/portal/index.html

Rodríguez, T. y Nava, L. (2013). Diagnóstico de la situación de los repositorios institucionales en las IES mexicanas. Recuperado de: https://docplayer.es/17945900-Diagnostico-nacional-de-repositorios-institucionales-en-las-ies-mexicanas.html

Serna, N. y Villanueva, C. (2014). Implementación del acceso abierto al conocimiento y repositorio institucional UANL. Cuarta Conferencia de Directores de Tecnología de Información, TICAL2014.

Universidad de Chile (2018). Red de repositorios Latinoamericanos brinda acceso a más de un millón de documentos académicos de acceso abierto. Recuperado de: https://www.uchile.cl/noticias/144031/red-de-repositorios-mas-de-un-millon-de-documentos-de-acceso-abierto

Wickham, H. (2019). Easily Harvest (Scrape) Web Pages. Recuperado de: https://cran.r-project.org/web/packages/rvest/rvest.pdf