ISSN 0798 1015

REVISTA ESPACIOS

HOME Revista ESPACIOS ÍNDICES / Index A LOS AUTORES / To the AUTORS

EDUCACIÓN • EDUCAÇÃO • EDUCATION Vol. 40 (Nº 15) Año 2019. Pág. 8

Big Data e inteligencia de negocios aplicados al estudio de mercado de estudios posgraduales de la Universidad Autónoma de Colombia

Big Data and business intelligence applied to the market study of postgraduate studies of the Universidad Autónoma de Colombia

CASTILLO, Rafael 1 y MORENO, Fernel 2

Recibido: 13/11/2018 • Aprobado: 24/04/2019 • Publicado 06/05/2019

Contenido

RESUMEN:

Este artículo presenta el trabajo realizado en la Universidad Autónoma de Colombia, en investigación de inteligencia de negocios y BigData, para determinar las necesidades de los estudiantes egresados de la universidad y así ofrecer programas de estudios de posgrados adecuados al momento actual o del futuro próximo. Se utilizó el método científico, se preseleccionaron las herramientas de apoyo y realizaron pruebas con datos recopilados de fuentes de información de redes sociales, encuestas y bases de datos estructuradas.

Palabras clave: BigData, Inteligencia de Negocios, Estudio de mercado

ABSTRACT:

This article presents the work carried out at the Autonomous University of Colombia, in both business intelligence and BigData research, to determine the needs of students who have graduated from the university and offer postgraduate programs suitable now or for the near future. The scientific method was used, support tools were pre-selected and tests were carried out with data collected from social network information sources, surveys and structured databases.

Keywords: BigData, Business Intelligence, Market study.

1. Introducción

La principal fuente de financiación económica de una Universidad es el ingreso de estudiantes y que cursen los semestres del programa elegido. Para que ello suceda, al igual que con cualquier producto, se deben presentar programas atractivos, de proyección, con suficiencia en calidad académica, precios asequibles y facilidades de pago que igualen o

superen la competencia. El estudiante al que se quiere llegar en primer lugar, para ofertar tales programas de estudio post-graduales es el estudiante egresado de pregrado, de la institución, pues él conoce y tiene referencia de la calidad, disponibilidad y facilidades que proporciona la Universidad. Sin embargo no se debe desconocer que cursar estos estudios en otra Universidad, le da oportunidad al estudiante, de conocer nuevos métodos de enseñanza-aprendizaje, otros docentes y en general un nuevo ambiente de estudio. El ideal es que, al graduarse, esté satisfecho con lo ofrecido y obtenido por o de la universidad, quiera volver en futuras ocasiones y que a su vez se convierta en un factor multiplicador para atraer nuevos estudiantes.

El proyecto a partir del cual se propone este artículo, tiene como objetivos determinar las preferencias de esos estudiantes, para dimensionar los presupuestos, hacer proyecciones de planta física, planta de personal y realizar mejoras en infraestructura física y tecnológica. Se usa el estudio de mercadeo para determinar las necesidades de los egresados en materia de capacitación, usando herramientas de búsqueda, empleando técnicas de minería de datos y Big Data para reemplazar el proceso actual y se utilizan encuestas enviadas por correo electrónico. Se pretende sistematizar y agilizar el proceso de seguimiento a esta población, para identificar y conocer sus experiencias académicas y medir su impacto en el mercado laboral y empresarial, donde ellos intervienen como imagen de la Universidad.

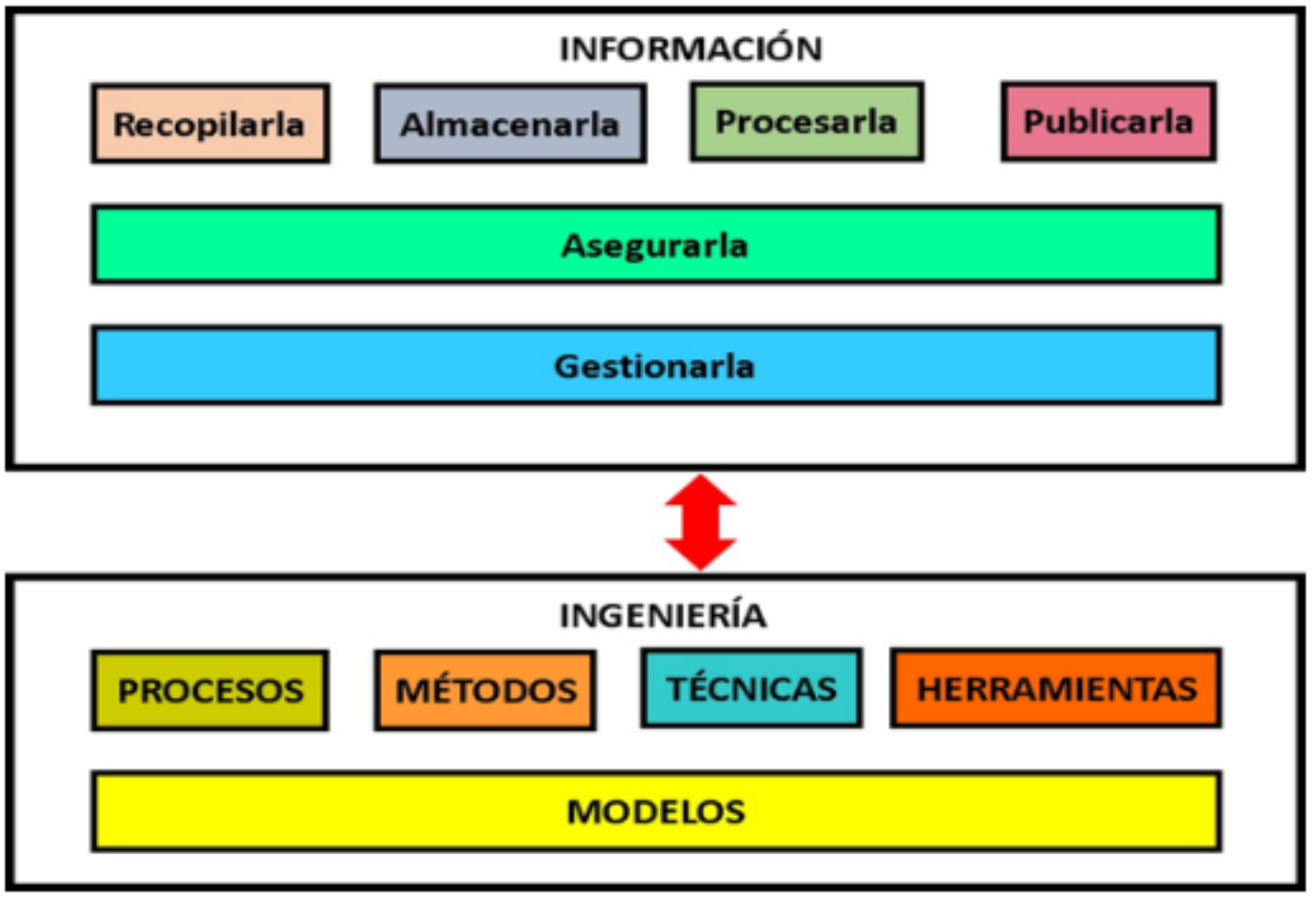
Partiendo de la absorción de datos, que incluye la extracción, transformación y carga (ETL por sus siglas en inglés), el proceso a seguir es el de procesamiento y descubrimiento de información, con los datos almacenados. Para datos numéricos y procedentes de fuentes estructuradas se usan los métodos tradicionales de Inteligencia de Negocios y para el procesamiento de datos no estructurados o semiestructurados se usan algunas de las técnicas de Big Data, tales como minería de opinión, análisis de sentimiento, modelamiento de tópicos, análisis de contenido automatizado o almacenamiento y procesamiento distribuido.

En este documento en primer lugar se describen a manera de contextualización la ingeniería de la Información, tomando como base el hecho de que el mundo es cada vez más digital; luego se describe BigData; algunas características de redes sociales como fuentes importantes de información no estructurada; la metodología usada para el desarrollo del proyecto; la descripción de los resultados obtenidos y las conclusiones.

1.1. Ingeniería de la información

La era digital, incluye la información generada desde cualquier parte, por cualquier medio y dispositivo en forma digital, potencializada por el internet de las cosas (*Internet of Things*, IoT), la cual permite que cualquier dispositivo que tenga un microchip y acceso a internet, envíe o reciba datos, que luego serán recibidos en algún medio de almacenamiento, para posteriormente ser procesado. La generación de información está disponible actualmente en forma estructurada (la que está en las bases de datos), de manera semiestructurada (a manera de atributo – valor) y No estructurada (Pérez, 2015, p. 4), como la que se tiene en los contenidos de las redes sociales. Los procesos aplicados a la información (casos de uso) son básicamente cuatro: Recopilación, almacenamiento, procesamiento, publicación y transmisión (Rajiv, 2014, p. 738), pero adicionalmente a manera transversal están los procesos que se encargan de dar la seguridad requerida y la gestión para el uso adecuado, como se muestra en la figura 1.

Fig. 1
Interrelación entre Ingeniería e Información



Fuente los autores

De otra parte, se tiene la ingeniería que tiene como fin fundamental crear modelos del mundo en su respectiva área de conocimiento y dar solución a problemas que se presentan entre el ser humano y la naturaleza, el ser humano y las máquinas y la interrelación de las máquinas entre sí. Por lo tanto, la Ingeniería de la Información tiene como principal objetivo aplicar procesos, técnicas, métodos y herramientas, a la información en cada una de los procesos mencionados en el párrafo anterior, como se muestra en la figura 1. Uno de los paradigmas que tiene que ver con las grandes cantidades de datos que se generan minuto a minuto (del orden de los *petabytes*) es el Big Data (Jiménez, 2016), tema que se detalla en la siguiente sección.

1.2. Big Data

Como factor de supervivencia de las organizaciones, en esta era de información digital, es el orientar los esfuerzos en gestionar proyectos que les permita cambiar sus estrategias de forma rápida, anticipándose a la competencia, ofreciendo productos o servicios novedosos que satisfagan las necesidades y demandas, tanto de los clientes actuales como de los clientes potenciales, requiriendo para ello tener un perfil detallado de los mismos. Para el caso de estudio de mercado, es necesario establecer con precisión y oportunidad qué productos/servicios se les debe ofrecer, qué opinan o dicen de la organización y el grado de satisfacción o fidelización que tienen, situación en la cual, la aplicación de la Ingeniería de la Información junto con Big Data, permitirá determinar este perfil y los servicios de formación posgradual que la Universidad podrá ofrecer como resultado de dicho estudio.

El término Big Data es relativamente nuevo, pues las primeras referencias datan del año 1997. (Mayer), (Bryson, 1999), (Marr, 2015) (Press, 2017). Hace referencia a los grandes volúmenes de datos que actualmente se generan día a día, y que dadas las condiciones de soporte tecnológico es posible recolectarlos, almacenarlos, procesarlos y publicarlos. Además, tiene la condición favorable que cada vez en mayor medida, el mundo del conocimiento humano está en medio digital.

Como contextualización dos definiciones de Big Data: una es la dada por IBM "el concepto de Big Data aplica para toda aquella información que no puede ser analizada o procesada utilizando procesos o herramientas tradicionales" (Barranco, 2012) y otra "El propósito del Big Data como metodología, es el de transformar una cantidad enorme de datos anodinos en una información útil para la toma de decisiones, convertir el éter de datos en oportunidades

de negocio contante y sonante” (Barranco, 2014).

La mayoría de facilidades existentes en determinada época, tan solo son posibles si se tiene el soporte tecnológico. El caso de Big Data no es ajeno. La existencia de hardware y software muy potentes (Rouda, 2015), de bajo costo o por acceso a facilidades tales como arriendo compartido (computación en la nube), hace realidad el hecho de hacer inteligencia de negocios con técnicas de analítica avanzadas, predictivas y prescriptivas. Luego de recopilar, analizar, procesar y publicar datos procedentes de fuentes tan diferentes como: datos existentes en bases de datos relacionales (datos estructurados); datos semiestructurados o no estructurados (Méndez, 2010) tomados de redes sociales, encuestas, páginas web, registros de navegación en la web, registros de llamadas de telefonía celular y de centros de llamadas, sensores y datos de geoposicionamiento entre muchas otras fuentes.

El **volumen** de los datos (del orden de Zetabytes), la **variedad** en la que se encuentran los datos al existir en múltiples formatos, la **velocidad** a la que se generan (Gigabytes por minuto), el **valo**r que tienen para la organización, la **veracidad** como grado de confiabilidad de que los datos son verdaderos, la **viabilidad** en el sentido de que los recursos necesarios y disponibles sean costo-efectivos y la forma como estos datos serán **visualizados**, son las características que dan forma al concepto de BigData, conocidos como las V´s de Big Data, resaltadas en negrilla. (IBM, 2016)

Paralelo al uso de la plataforma computacional, es necesaria la existencia de nuevos roles de Ingenieros de Sistemas o Ingenieros Informáticos multidisciplinares (Uniandes, 2016), pues ellos son quienes tendrán a cargo el correcto funcionamiento y gestión de los sistemas informáticos que se desarrollen o empleen. Adicional a los roles existentes en los departamentos informáticos, están los Científicos de datos (*data scientist*), Gerente de datos (*Chief Data Officer*, CDO), Ingenieros de datos, expertos en Análisis de datos, quienes entre muchas otras funciones tienen a su cargo el análisis descriptivo, predictivo y prescriptivo de los datos; recopilación, almacenamiento, aseguramiento (protección) y gestión de los datos y presentar los resultados para la toma de decisiones relacionadas con el tema para el cual se hace todo este proceso.

1.3. Redes sociales

Las redes sociales han aparecido para cambiar el modo de comunicarnos en el siglo XXI, se puede ver como una intromisión a nuestros hogares y han cambiado la manera de asociarnos y de comunicarnos. La participación de los estudiantes egresados y no egresados de la Universidad, en diferentes redes sociales, tales como las que se muestran en la figura 2, es una manera de obtener opiniones indirectas acerca de sus preferencias y planes futuros de estudio. Algunas de estas redes sociales, se describen brevemente a continuación.

Figura 2

Representación gráfica de logos de las redes sociales

Fuente los autores

Facebook, es la más popular. Creada en el año 2004, cuando Mark Zuckerberg de la Universidad de Hardvard, junto a Eduardo Saverin, Chris Hughes y Dustin Moskovitz esta red originalmente era un sitio para estudiantes donde podían intercambiar opiniones en forma rápida. Según cifras de Facebook, actualmente la red cuenta con 1.230 millones de usuarios diarios activos y de esos 1.150 millones se conectan a través de un dispositivo móvil. (Perú21, 2017).

En Facebook, existen dos tipos de cuentas: una para personas, totalmente gratuita y otra con costo para las empresas, en esta última se pueden crear páginas con publicidad para sus productos.

Esta red social es base para proyecto en ejecución ya que se adapta a la personalidad de los egresados con grado de educación universitario, quienes actúan de manera descomplicada, opinando libre y espontáneamente y en la que se forman grupos diversos con intereses comunes.

Twitter, otra red social que permite expresar opiniones en 140 caracteres. Fue creada por Jack Dorsey en el año 2006, y se calcula que tiene más de 500 millones de usuarios, quienes envían más de 70 millones de tweets al día y recibe más de 800.000 peticiones de búsqueda (El tiempo, 2017). Está red social, se ha convertido en una red con grado importante en los diferentes niveles de la sociedad. Permite intercambiar videos, imágenes, gráficos y bloquear a usuarios indeseados por motivos de frecuentes ataques. También esta red sirve para dar a conocer su perfil profesional, encontrar familiares y amigos de los cuales hace mucho tiempo no se sabía algo de ellos.

A pesar de la apariencia sencilla de Twitter, cuenta con muchas herramientas con las cuales los usuarios pueden interactuar. Una de ellas es el Hashtag (etiquetas o metadatos) que permite al usuario crear tendencias, diferenciar, destacar y agrupar por palabras específicas. Como otras características, Twitter es (Zúñiga, 2017):

Hipertextual.

Intuitivo.

Asimétrico.

Multiplataforma.

Sincrónico.

Social.

Descentralizado.

Puede reenviarse (Retweet).

Dar Me gusta,

Crear listas.

2. Metodología

El método utilizado para el estudio es el inductivo que parte de los conocimientos particulares para encontrar las incidencias determinadas. Se aplica desde lo cualitativo y lo cuantitativo. Se establecen categorías de análisis y de observación de interés e intencionalidad.

Como paso inicial se establece la muestra por conveniencia, en virtud que no se cuenta con una línea base y con un universo determinado. El ámbito de intervención para la recopilación de información en lo interno se da desde la información que mantiene la oficina de Bienestar Universitario de la Universidad Autónoma de Colombia y en lo externo, se cuenta con la posibilidad que presenta este proyecto para hacer minería de datos aplicando el software requerido, que recolecta datos de las redes sociales. El componente técnico se aplica en la construcción de instrumentos tales como: encuestas dirigidas y entrevistas estructuradas. La encuesta dirigida se usó para el análisis cuantitativo y la entrevista estructurada, para el análisis cualitativo. Adicionalmente se trabajó con dos líneas de acción una participativa y otra inducida. Para el caso de datos extraídos de redes sociales se usó la metodología asalto de fuerza bruta, utilizando filtros relacionados con información pertinente, y relacionados con la Universidad Autónoma de Colombia. Posteriormente, luego de almacenar los datos, se procedió a utilizar la técnica de MapReduce y finalmente se utilizó herramientas y frameworks disponibles en el mercado para el análisis de datos.

3. Resultados

Tal como se comentó al inicio de este documento, los procesos que se realizan con la información son obtención de los datos, conocido como *data gathering* (Pajarillaga, 2012) o *data ingestion*; almacenamiento de los datos recopilados, que al igual que en el proceso anterior tiene en inglés el nombre de *Data Lake* (Maroto, 2016), por constituirse un sitio al cual fluyen múltiples y dispares fuentes de datos; el procesamiento de esta información para la obtención de resultados y finalmente la publicación de los resultados.

Las herramientas de las cuales se hizo uso, tienen versiones de prueba o son de uso libre son:

3.1. Para la recopilación de la información

Maltego (Paterva, 2017): herramienta que posibilita crear un grafo que tiene en sus nodos direcciones de correo electrónico, sitios web, nombres de personas, cuentas de redes sociales (twitter, Facebook, etc.) y documentos. Se hizo uso de esta herramienta por las facilidades que brinda, al conocer los enlaces que las páginas web de la Universidad tienen y a su vez, las cuentas de correo que se enlazan de los usuarios, facilitan la búsqueda tanto de personas como de sitios web que tienen mensajes o interrelaciones de dichos usuarios, para determinar qué relaciones pueden tener mayor impacto en una selección de fuentes a seguir.

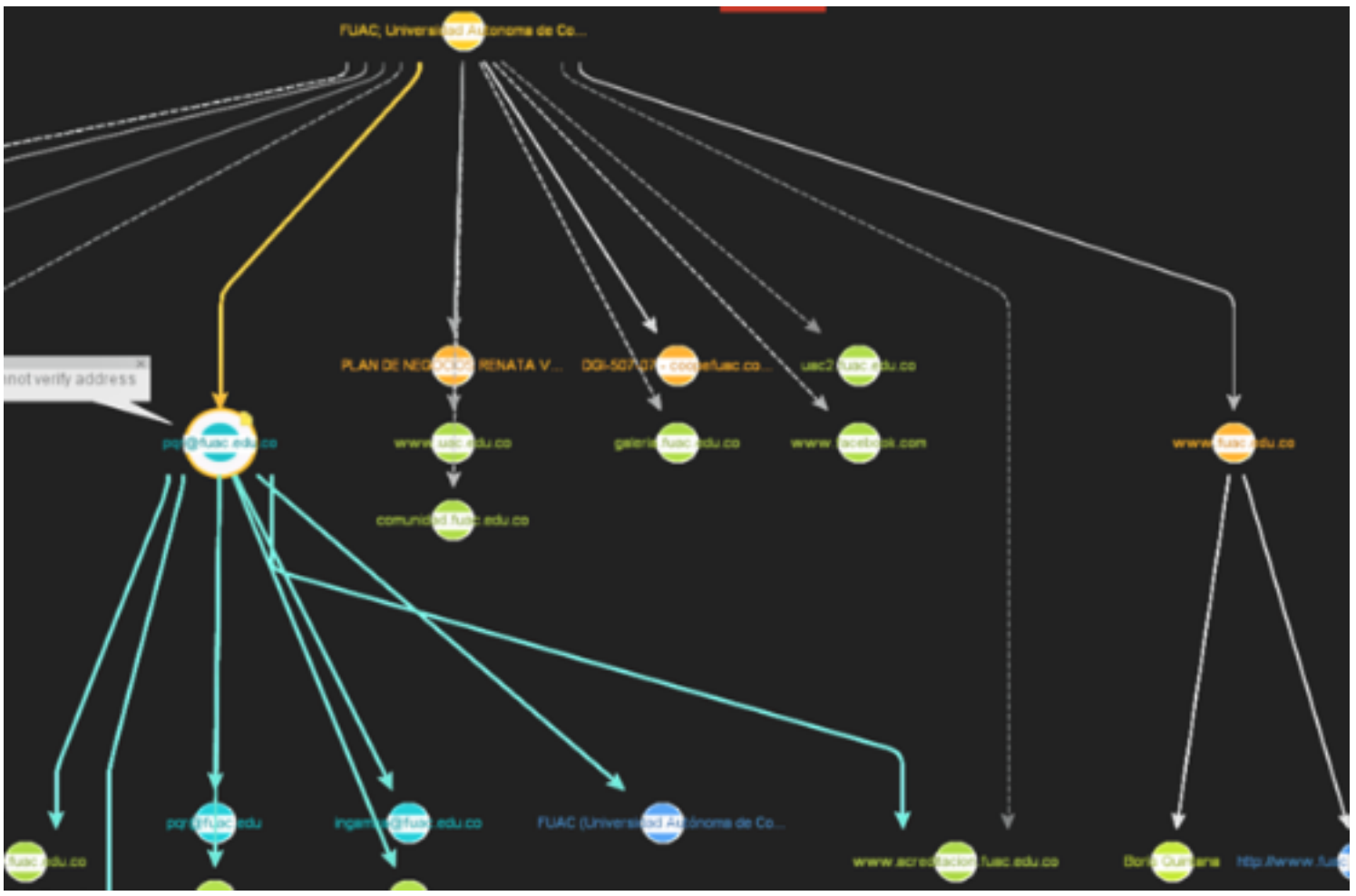
Aunque la herramienta tiene limitaciones en cuanto al uso a manera de prueba (versión *trial*), permitió establecer la potencialidad que tiene en el proyecto. A manera de exploración se tomó el sitio web de la Universidad Autónoma de Colombia (www.fuac.edu.co), el correo de egresado@fuac.edu.co y las cuentas creadas para el semillero de investigación SeInvEnTA. Con los resultados mostrados en la figura 3.

Maltego, permite que cualquier rama del grafo generado se puede ampliar, tal como se muestra en la figura 4. Otra de las facilidades que proporciona Maltego es la de exportar como tabla en Excel, los datos del grafo generado, que permiten posteriormente con otras

herramientas generar procesos estadísticos o de minería de datos y con ello continuar un proceso de análisis de datos, ya sea en *Power Bi Desktop*, o en el aplicativo de Excel Office, tal como se ve en la figura 4, que muestra de otra forma los datos del grafo de la figura 3.

Para la generación de este grafo, se tomó como base la dirección (URL) de la Universidad Autónoma de Colombia (www.fuac.edu.co), luego se expandió para encontrar tanto las direcciones entrantes como las direcciones salientes que tienen algún registro histórico relacionado con ella, y luego relaciones de sitios web a partir de uno de estos enlaces.

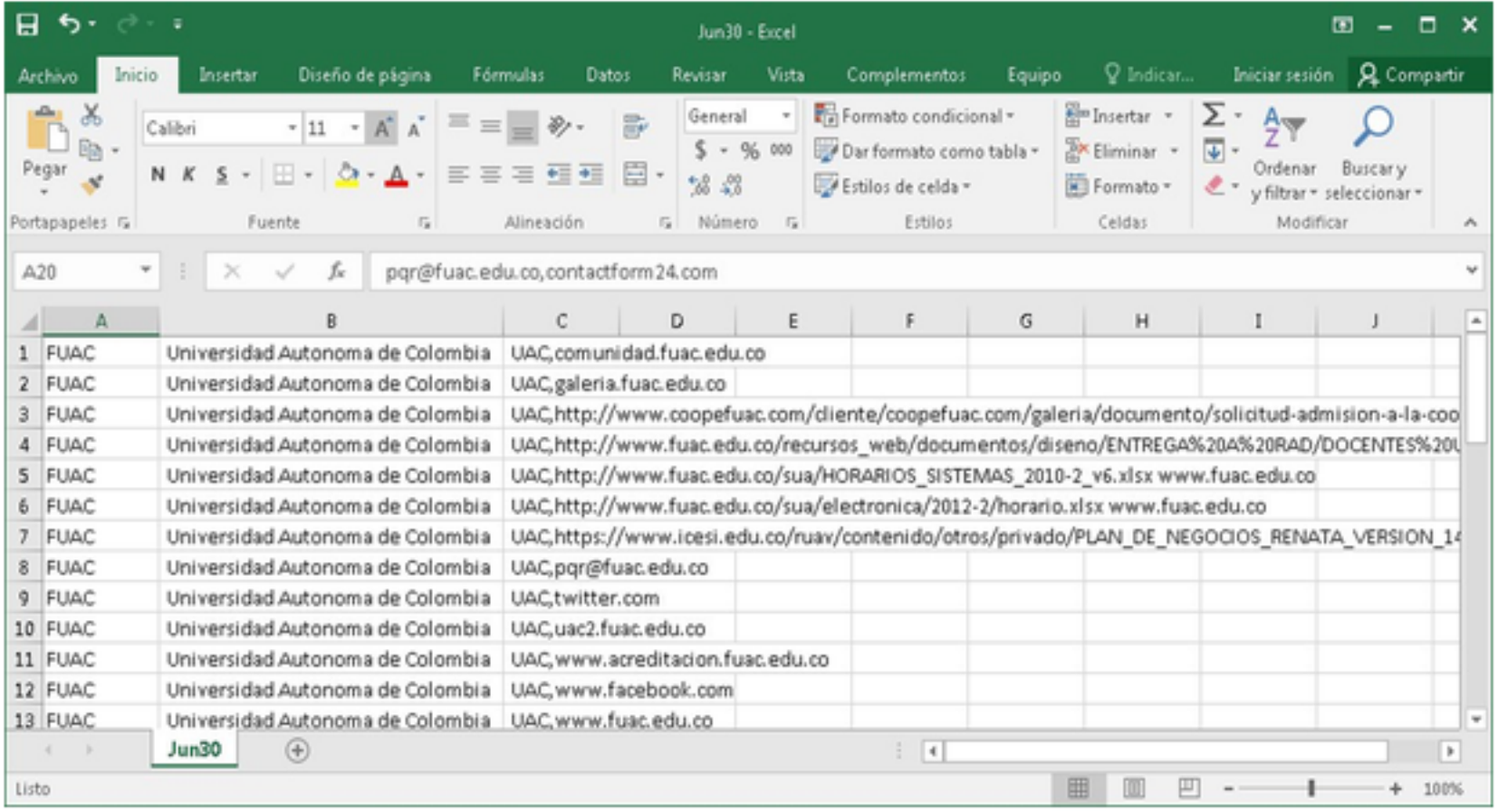
Fig. 3
Grafo generado con Maltego para una dirección de correo electrónico



Fuente los autores

Como limitante, dadas las restricciones que tiene Twitter, se tiene que la información obtenida tanto de correos institucionales, cuentas y mensajes de twitter, son pocos, imposibilitando que a partir de lo encontrado se pueda hacer algún análisis. Esta dificultad se soluciona al adquirir la versión licenciada de Maltego y pagar el canon respectivo que pide Twitter, por el uso de la información.

Fig. 4
Documento en Excel generado con datos proporcionados por Maltego.



Fuente los autores

Flume (Apache, 2017): que viene integrada a la plataforma Cloudera, que a modo *stream*, busca y recolecta la información de diferentes fuentes, tales como Twitter, Facebook y Linkedin, usando las API´s que proporcionan estas redes, con base en las llaves y claves para acceder a ellas. Con esta herramienta se hizo análisis de sentimiento con los mensajes generados en Twitter y relacionados con la Universidad Autónoma de Colombia; algunas de las cuentas de Twitter de funcionarios, estudiantes o docentes vinculados a la Universidad; algunas de las cuentas de Facebook con características similares a las descritas para Twitter, las cuales en su mayoría se obtuvieron con los datos generados por Maltego.

3.2. Para el almacenamiento

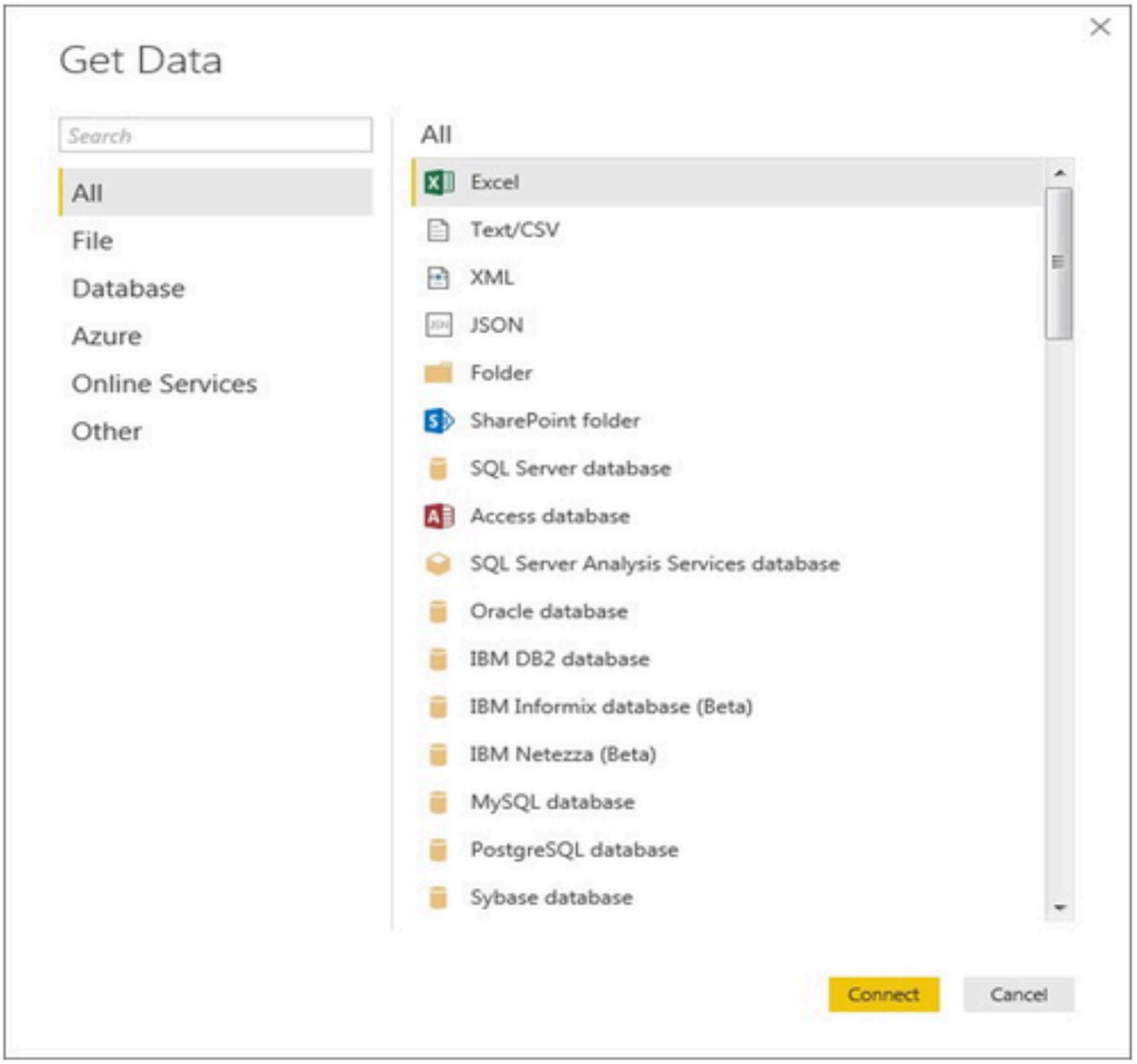
Hive (Hive, 2017), herramienta que se utiliza con HDFS (Hadoop, 2017) (*hadoop distributed file System*) para almacenar los datos como tablas (de datos), también sirve para realizar la visualización de resultados. Para el análisis de sentimiento de los mensajes se evaluó Solr (Hortonworks, 2017), pero se decidió por varias razones utilizar *Power BI de Microsoft*.

3.3. Para el procesamiento y despliegue

Se realizaron pruebas con las herramientas Power BI disponibles en Excel 2016 (Micrfosoft, 2015) *(power query, power pivot, power view y power map*) las cuales permiten hacer todos los procesos ya descritos, relacionados con la información (recopilación, etc.) con funcionalidades similares a las disponibles en Power Bi Desktop y Power BI service en la nube de Microsoft con Windows Azure, las cuales se pueden acceder en el sitio web de Microsoft (Microsoft, 2017).

Power BI Desktop de Microsoft, para usarla se instaló una versión de prueba por 60 días. La siguiente gráfica (Figura 5) muestra las opciones que da, para recopilar los datos, las cuales van desde datos en bases de datos, datos en la Nube de Azure, servicios en línea o archivos comunes en texto u hojas electrónicas.

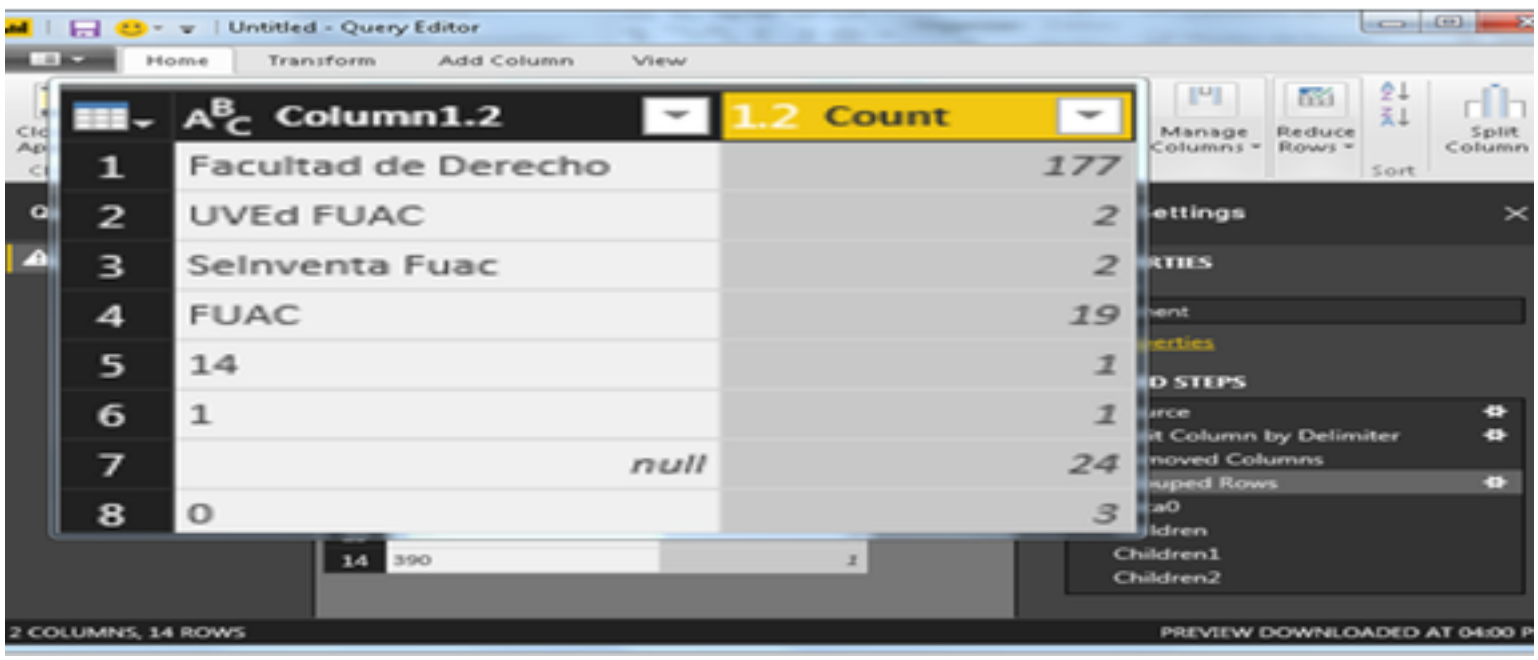
Fig. 5
Algunas de las fuentes de datos para incluir en *Power BI Desktop*



Fuente los autores

La ventana de trabajo proporcionada por Power BI, con la configuración básica de la herramienta Power Pivot, facilita la extracción de los datos, ponerlos a manera de tablas, realizar el proceso ETL, hacer el análisis de datos, visualizar por medio de muchos tipos de gráficas, generar reportes y publicar los resultados obtenidos.

Fig. 6
Pantalla de trabajo de *Power BI Desktop*



Fuente los autores

La figura 6 muestra el resultado de cargar los datos extraídos con el API de Twitter (200 twitts), y después de haber aplicado agrupamiento por origen del Twitt, indicando que la Facultad de derecho, según esa muestra, es la más activa.

3.4. Resultados análisis de la encuesta

Este instrumento (encuesta), se tomó con base en el documento elaborado y puesto en la WEB, por la oficina de atención al egresado y puede ser adaptado con pequeños cambios a las características de todos los programas de la Universidad. Para ello se tomó como base, los estudiantes graduados de Ingeniería de sistemas desde el año 2010 al 2017, Por medio de un mensaje enviado al correo electrónico, que ellos suministraron al momento del trámite de su grado, se solicitó la participación en la encuesta. La encuesta en cuanto a las respuestas recibidas, es anónima, sin embargo, se tiene el registro de quienes la respondieron, lo cual está dentro de las condiciones legales de Habeas Data. La encuesta está basada en las preguntas que el área de atención al egresado tiene ya elaboradas, más otras que se consideraron particulares al programa de Ingeniería de sistemas. Tiene cinco secciones: Datos Personales, datos del programa, información laboral, aspectos académicos y aspectos institucionales, todas ellas orientadas a conocer la opinión y sentimiento que tienen respecto a la universidad en general, al programa en particular y conocer su situación actual y ubicación laboral y profesional.

Los datos de la muestra y resultados son los siguientes:

Total de invitaciones enviadas 284

Hombres: 205 (72.18%)

Mujeres: 79 (27.82%)

Total, de encuestas completadas totalmente 45

Parcialmente respondidas: 24

Total respuestas: 69 (24.3%)

Personas que no quieren que sus datos sean usados: 4 (por lo tanto, la mayor parte de los datos personales no están).

Para el cálculo del tamaño de la muestra se utilizó la siguiente fórmula:

$$n = \frac{Z^2 * s^2 * N}{e^2 * (N-1) + Z^2 * s^2}$$

Z = Nivel de confianza del 95% (correspondiente a 1.96)
s = 0.5 (desviación estándar asumida, cuando no se conoce)
N = 284 (Población de la encuesta)
e = 0.1 (error muestral)

Con base en la formula anterior el tamaño de la muestra válida es:

n = (3.841*0.25*284)/(0.1*283+4.091) =272.75/3.79 = 71.95

n = 72

3.4.1. Análisis resumido

Las respuestas a las preguntas de cada encuesta fueron trabajadas en una hoja de Excel, y con base en ella se obtuvieron los resultados mostrados a continuación.

El 71% de quienes respondieron, son hombres y 29% mujeres,

El 69% (45) de los estudiantes ingresaron entre los 17.5 y 24.5 años

El 55.6% (10) mujeres ingresaron con edad igual o mayor a 21.5 años

El 56.4% (26) hombres ingresaron con edad entre los 17.5 y 21.5 años

La mayor cantidad de estudiantes que ingresaron 42 (64.6%) son de estrato 3, seguidos por 19 (29.2%) de estudiantes de estrato 2.

Los valores por género varían muy poco con relación al total. Así por ejemplo el estrato 3 está conformado por un 61.1% de mujeres y 65.2% hombres. Tampoco el estrato socio-económico tiene influencia en alguno de los otros parámetros (tiempo para el grado o duración de los estudios), pero si hay influencia de la edad en el tiempo para graduarse,

pues 44 estudiantes de los 65, con edades comprendidas entre 17 y 21 años, duran menos de un año y medio para graduarse.

La duración de los estudios no se ve afectada por la edad de ingreso, los datos obtenidos muestran una distribución uniforme en este aspecto. Para está respuesta en particular, la duración promedio de los estudios, entre el rango de 4.5 y 6.5 años, está el 49%. Tiempo que puede considerarse normal, o equivalente 1 de cada dos estudiantes termina sus estudios en este intervalo.

La mayoría de los egresados (26 de 59), viven en la zona Occidental de la ciudad, tanto al norte como al sur, siendo Kenedy, Engativa y San Cristobal, las localidades donde hay mayor número de residentes. Una vez terminados los estudios, el tiempo promedio para graduarse en el término de un año es el 37%.

En cuanto al estado civil, tampoco afecta ni la duración de los estudios ni el tiempo para graduarse.

Después de haberse graduado, 6 estudiantes realizaron una maestría, uno de ellos tiene especialización y ha realizado cursos de actualización en la profesión. 15 estudiantes han realizado especializaciones y algunos de ellos han realizado diplomados y o cursos de actualización. Finalmente 4 han realizado diplomados. En total 35 estudiantes han realizado alguna capacitación posgradual.

En la encuesta hay una pregunta de respuesta abierta, "Por favor dar algunas sugerencias de acciones de mejora para la Universidad", el análisis realizado para las respuestas dadas, se hizo con herramientas para análisis de texto, tales "Calculadora de densidad de palabras", hecha en Excel, y detección de polaridad, con lenguaje de programación Python, basada en el algoritmo afinn.py, el cual se modificó, para obtener mejores resultados. Los resultados se muestran a continuación.

Número de opiniones: 71

Palabras diferentes: 490

Con base en esto, en primer lugar, se halló la frecuencia de aparición de cada palabra, luego la frecuencia de aparición de dos palabras (bigramas), la frecuencia de aparición de tres palabras (trigramas) y con más de cuatro palabras, ya el resultado solamente arrojaba una frase en la cual el resultado no fue significativo para el análisis.

Los resultados se muestran en la tabla 1, con las palabras relevantes, Mejorar 22 ocurrencias y estudiantes con 10 ocurrencias, dado que las demás son las que en análisis de texto se conocen como palabras "stopwords". No se muestran las de menor número de ocurrencias, ya que no son significativas para el análisis a realizar.

Tabla 1
Palabras con frecuencia de aparición mayor o igual a 10 veces.

Palabra	Nro. Ocurr.	Palabra	Nro. Ocurr.	Palabra	Nro. Ocurr.
de	84	a	30	no	13
la	47	en	29	se	11
y	43	con	28	estudiantes	10
que	40	Mejorar	22	una	10
el	39	para	21		
los	33	las	14		

En cuanto a la distribución por bigramas el resultado se muestra en la tabla 2, de los cuales los relevantes son: “desarrollo de” con 7 ocurrencias, “mejorar la” con 6 y con 4 ocurrencias, “actualización de”, “los docentes”, “mejorar el”, “mejorar las” y “mejorar los”.

Tabla 2
Agrupación de frecuencia de aparición por bigramas.

Palabra	Nro. Ocurr.	Palabra	Nro. Ocurr.	Palabra	Nro. Ocurr.	Palabra	Nro. Ocurr.
de los	9	los estudiantes	7	para que	5	los docentes	4
en el	9	de Sistemas	6	que no	5	Mejorar el	4
a la	8	la carrera	6	ya que	5	Mejorar las	4
con el	8	Mejorar la	6	a los	4	Mejorar los	4
de la	8	de software	5	actualización de	4	que lo	4
desarrollo de	7	la universidad	5	la interacción	4	que los	4

Y para comparar el resultado de agrupación por trigramas, con el fin de ver si es o no significativo, se muestra en la tabla 3, dando como relevantes “desarrollo de software” 5 ocurrencias y “mejorar la interacción” con 3 ocurrencias.

Tabla 3
Agrupación por frecuencia de aparición en trigramas.

Palabra	Nro. Ocurr.
desarrollo de software	5
de la carrera	4
a los estudiantes	3
con el objetivo	3
Mejorar la interacción	3
para que los	3

Dado que la palabra más repetida fue mejorar, relacionada con mejor, con actualizar y calidad, se buscaron las frases completas donde esta aparece; al igual se hizo con docente(s), profesor(es), maestro(s) dando como resultado:

Relacionadas con los docentes (profesores, maestros), se encontraron 12 opiniones negativas y 2 positivas

Con relación al contenido del programa, 11 opiniones orientadas a mejorar, actualizar o profundizar y 5 a enfatizar en el desarrollo de software.

Con relación a las instalaciones de laboratorios o salas de computo 6 relacionadas a actualizar o mejorar.

con relación a la interacción con los estudiantes y egresados, 4 opiniones orientadas a

mejorar.

De la universidad en cuanto a Sede, organización y sistema académico, 5 opiniones.

En cuanto a la detección de polaridad el algoritmo usado (afinn-165), tiene en cuenta la valoración de cada palabra en un rango entre -5 y 5. El resultado obtenido muestra 11 opiniones con valoración negativa; 14 con valoración neutra y 46 con valoración positiva.

Dado el resultado, se encuentra que es necesario optimizar el algoritmo para que incluya bigramas o trigramas que permitan tratar combinaciones de palabras tales como "No muy bueno" , "es poco agradable", "no es malo", etc, las cuales por separado dan un resultado de polaridad positivo o neutro, pero que en contexto, su resultado es diferente.

3.5. Datos obtenidos de twitter y de Facebook.

Como complemento a obtener el perfil del egresado y con base en los resultados de trabajar con la herramienta "maltego", tomando como base la cuenta el twitter de la dirección del programa de Ingeniería de sistemas y los seguidores de la misma o a quienes se sigue, se hizo el análisis de mensajes publicados en Twitter, desde el 17 al 22 de mayo, con los siguientes resultados:

Cantidad de mensajes recopilados 200, de los mensajes recopilados tan solo uno hace referencia a la universidad, los demás a la situación política de ese momento, Ministerio TIC, donde los hashTag del momento fueron: YoamoInternet, Teletrabajo y WiFiGratis con más de 100 retwitts.

Indicando que no había algo que publicar, comentar o decir de la Universidad en ese momento.

Para Facebook, el seguimiento que se hizo fue manual, ya que el algoritmo para realizarlo no se pudo perfeccionar y por ampliación del *habeas data*, los dirigentes de Facebook, restringieron el acceso de los datos y por otra parte no se pudo enlazar los vínculos dinámicos, lo cual para efectos de los resultado se hizo manualmente con un barrido de los usuarios (estudiantes y egresados Fuac), contactos de las cuentas de la Universidad Autónoma de Colombia, evidenciando los resultados (entre el 26 de junio y 7 de julio de 2018) mostrados en la tabla 4.

Tabla 4
Resultados de seguimiento a publicaciones en Facebook

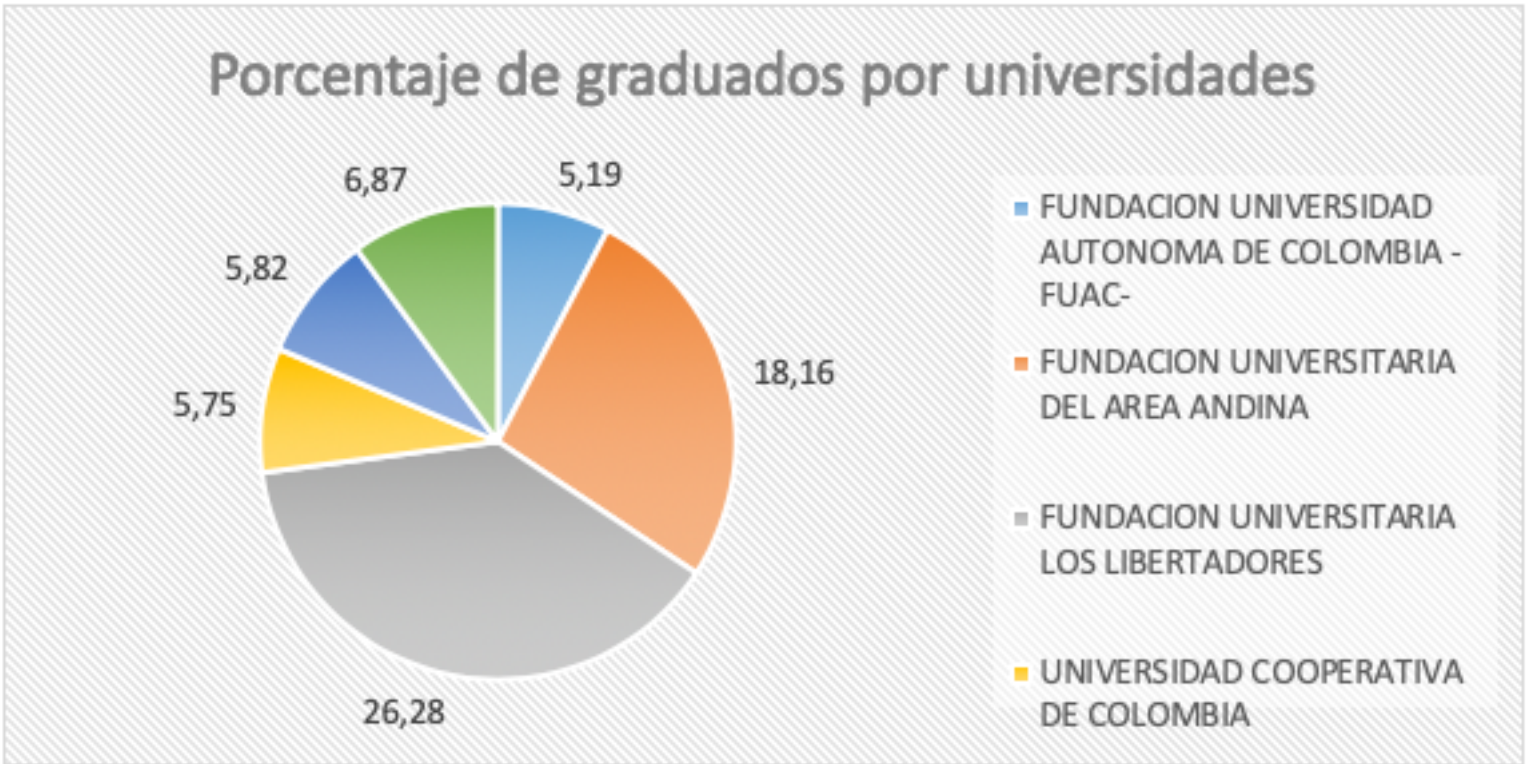
	ME GUSTA	**Me encanta**	**Me divierte**	**Me asombra**	**Me entristece**	**Compar-tido**	**Comen-tarios**
Actividad 1	33	1					
Actividad 2	218	44	8			72	6
Actividad 3	25			12	25		14
Actividad 4	248	36	7				
Actividad 5						2	
Actividad 6	312	56		6		47	
Actividad 7	7						
Actividad 8	9						

Actividad 9	46	10					4
Actividad 10	315	20	7			55	25
Actividad 11							10
Actividad 12	136						21

3.6. Segmentación del mercado de programas de posgrado.

Con respecto al objetivo del proyecto, "Criterios para segmentar el mercado de programas de posgrado de la Fuac", en los programas, que ésta ofrece y la participación de los demás centros de educación superior, se tomó la página oficial del Ministerio de Educación Nacional, y en los servicios que ofrece, está un documento con más de 378.000 registros de los programas de los diferentes centros de educación superior, donde figura el número de egresados, por programas académicos. De esta información se seleccionaron aproximadamente 4.000 registros de aquellos posgrados de todas las universidades que pudieran ser competencia, para nuestras ofertas posgraduales, aplicando tablas dinámicas que ofrece la herramienta Microsoft Excel. La hoja electrónica, resumen de la tabla inicial, puede ser utilizada por cualquier programa de la universidad, en busca de sus posibles competidores, ya que la herramienta, con la tabla dinámica creada, así lo permite. Como resultado, la Universidad Autónoma de Colombia tiene participación del 5.19% con un total de 2916 graduados, entre el año 2001 y el 2017.

Figura 7
Participación porcentual de las universidades, en posgrados que tambien son ofrecidos por la Universidad Autónoma de Colombia



En la figura 7, se muestra la participación porcentual de cada universidad frente al universo total de egresados en el período 2001, 2017, con porcentajes iguales o superiores a los de la Universidad Autónoma de Colombia. Como se aprecia la Fundación Universitaria los Libertadores tiene una participación del 26.28% del total, seguida por la Fundación Universitaria del Área Andina con el 18.16%. Para otras Universidades su relevancia en el período analizado es la Universidad Libre con un porcentaje del 6,87%, la Universidad Externado de Colombia con una participación de 5.82% de un total de 3266 egresados, la Universidad Cooperativa de Colombia con un porcentaje de 5.75% de un total de 3227 egresados y la Universidad Autónoma de Colombia con un porcentaje de 5.19% de un total de 2916 egresados.

Las Universidades que tuvieron menor participación fue la Universidad Manuela Beltrán y La

Fundación Universitaria Konrad Lorenz con participación del 0.02 %.

Con esta participación estas universidades, no se infiere que gradúe pocos estudiantes, sino que los gradúa en otros posgrados que no son afines a los ofrecidos por la Universidad Autónoma de Colombia.

Se puede evidenciar que las universidades de gran reconocimiento no aparecen en este análisis debido a que las especialidades, maestrías y los doctorados y posdoctorados, no son las categorías afines a las que ofrece la Universidad Autónoma de Colombia.

4. Conclusiones

Aunque los datos existen en gran cantidad, para utilizar estos datos es necesario pedir la autorización de quienes estén relacionados con ellos, (Ley de Habeas Data).

Las facilidades que proporcionan las herramientas actuales, utilizadas con grandes volúmenes de datos, permiten desarrollos de aplicaciones y efectuar análisis de datos, en pocos días, contrario a lo que sucedía hace apenas pocos años, en los que elaborar una bodega de datos, hacer inteligencia de negocios y ponerlos al alcance de los usuarios finales, estaba del orden de años.

La integración de diferentes fuentes de datos, con diferentes tipos de datos, también es una facilidad existente, que permite ampliar la cobertura de cualquier proyecto relacionado con investigación de mercados, abarcando áreas que antes eran exclusivas de proyectos especializados y resultó de gran utilidad para realizar el proyecto objeto de este artículo.

De las herramientas evaluadas, las que están incorporadas en Cloudera (flume por ejemplo), son útiles, pero por ser mediante comandos en la consola, impiden el rápido desarrollo de una aplicación. mientras que usar las que tiene Microsoft, son intuitivas, de fácil aprendizaje, manejo por interfaz gráfica, pero limitadas para volúmenes de datos que superen unas cuantas decenas de gigabytes.

El uso de datos abiertos, disponibles por entidades estatales, también fue de gran ayuda al proyecto, ya que con ellos fue posible realizar el estudio de ofertas de posgrados en Colombia, por medio de tablas dinámicas y elaborar diferentes análisis con esos datos, resultando ser de mucha utilidad, para la toma de decisiones.

En el análisis de detección de polaridad o análisis de sentimiento en los mensajes de Twitter y opiniones dadas en las respuestas a una de las preguntas de la encuesta, escritos en español, los resultados son apenas iniciales, ya que los diccionarios y librerías con catálogos que clasifican las palabras, solamente están disponibles para usarlas adquiriendo herramientas con costos que no están cubiertos en el proyecto, condición que implica crear dichos diccionarios, librerías y corpus lingüísticos.

El proyecto permite crear las bases necesarias para ampliar el conocimiento de los estudiantes de la universidad, ampliando la cobertura a áreas como la académica, bienestar universitario, entre otras, lo que en argot comercial se conoce como “vista 360 del cliente”.

Referencias bibliográficas

Apache. (2017). Recuperado de https://flume.apache.org/

Barranco, R. (2012). ¿Qué es Big Data? Todos formamos parte de ese gran crecimiento de datos. IBM developer Works.

Barranco, R. (2014). ¿Qué es Big Data?. IBM developersworks. Recuperado de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Bryson, S; Kenwright, D. (1999) Visually exploring gigabyte data sets in real time. Communications of the ACM, Vol. 42 No. 8, pp. 82-90

El tiempo. (2017). Twitter llega a los 500 millones de usuarios, según Twopcharts. Recuperado de www.eltiempo.com/archivo/documento/cms-11200102.

Hadoop. (2017). Recuperado de https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Hive. (2017). Recuperado de https://hive.apache.org/

Hortonworks. (2017). Recuperado de https://es.hortonworks.com/apache/solr/

IBM. (2016). 4 Vs IBM Big Data & Analytics Hub. Recuperado de http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Jiménez, C. (2016) 1er foro de Ingeniería de la Información. Universidad de los Andes.

Maroto, C. (2016). A Data Lake Architecture With Hadoop and Open Source Search Engines. Recuperado de https://dzone.com/articles/a-data-lake-architecture-with-hadoop-and-open-sour

Marr, B. (2015). A Brief History of Big Data Everyone Should Read. Recuperado de https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr?trk=mp-author-card

Mayer-Schönberger, V. ; Cukier, K. Big data, la revolución de los datos masivos. Recuperado de http://www.eldiario.es/turing/Big-data_0_161334397.html

Méndez, M. (2010). Integración de Datos Estructurados, Semiestructurados y No Estructurados. Recuperado de https://privmario.wordpress.com/2010/07/28/integracion-de-datos-estructurados-semiestructurados-y-no-estructurados/

Microsoft. (2015). The poweer BI team. Publish to Power BI from Excel 16. Recuperado de https://powerbi.microsoft.com/en-us/blog/publish-to-power-bi-from-excel-16/

Microsoft. (2017). Microsoft Power Bi. Cree sorprendentes informes y visualizaciones con Power BI Desktop. Recuperado de https://powerbi.microsoft.com/es-es/desktop/

Pajarillaga, F. (2012). DATA GATHERING. Recuperado de https://es.slideshare.net/mschie/data-gathering

Paterva Home. (2017). Recuperado de https://www.paterva.com/web7/buy/maltego-clients/maltego-ce.php

Pérez, M. (2015). BIG DATA Técnicas, herramientas y aplicaciones. Editorial Alfa Omega.

Perú21. (2017). Facebook cumple 10 años con 1230 millones de usuarios. Recuperado de www.peru21.pe/opinion/facebook-cumple-10-anos-1230-millones-uusarios-2168536

Press, G. (2017). A Very Short History Of Big Data. Recuperado de http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#24cd66a655da.

Rajiv, R. (2014). Streaming Big Data Processing in Datacenter Clouds. IEEE-Clouds Computing, 1, 738 -739.

Revista PYM. (2017). Linkedin su historia y su caso de éxito. Recuperado de http://www.revistapym.com.co/destacados/linkedin-su-historia-su-caso-éxito

Rouda, N. (2015). The Surprising Economics of Engineered Systems for Big Data (with Oracle. Recuperado de http://www.oracle.com/us/technologies/big-data/eng-systems-for-big-data-esg-wp-2852701.pdf

Uniandes. (2016). ¿En qué roles se puede desempeñar un ingeniero de sistemas y Computación?. Recuperado de https://sistemas.uniandes.edu.co/es/isis-aspirantes/roles

Zúñiga, A. (2017). 10 Características comunicativas de Twitter. Recuperado de https://andrewzuniganajar.wordpress.com/2013/09/20/10-caracteristicas-comunicativas-de-twiiter/

---

1. Ingeniería de Sistemas. Universidad Autónoma de Colombia. Ingeniero de Sistemas. Rafael.castillo@fuac.edu.co
2. Ingeniería de Sistemas. Universidad Autónoma de Colombia. Ingeniero de Sistemas. Fernel.moreno@fuac.edu.co

---