Universidad Tecnológica de El Salvador

Aplicación de herramientas *big data* al Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas de El Salvador

Investigadores:
Verónica Idalia Rosa
José Guillermo Rivera

Colección Investigaciones | 74

Universidad Tecnológica de El Salvador

Aplicación de herramientas *big data* al Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas de El Salvador

Investigadores:
Verónica Idalia Rosa
José Guillermo Rivera

Esta investigación fue subvencionada por la Universidad Tecnológica de El Salvador. Las solicitudes de información, separatas y otros documentos relativos a este estudio pueden hacerse a la siguiente dirección postal: Universidad Tecnológica de El Salvador, edificio *Dr. José Adolfo Araujo Romagoza*, Vicerrectoría de Investigación y Proyección Social, Dirección de Investigaciones, calle Arce y 19.ª avenida Sur, 1045, o a veronica.rosa@utec.edu.sv.

San Salvador, 2018

005.74
R788a Rosa, Verónica Idalia, 1968-
Aplicación de herramientas big data al Viceministerio de Vivienda
sv y Desarrollo Urbano del Ministerio de Obras Públicas de El Salvador /
Verónica Idalia Rosa, José Guillermo Rivera. -- 1ª ed. -- San
Salvador , El Salv. : Universidad Tecnológica (UTEC), 2018.
114 p. ; 23 cm. -- (Colección investigaciones ; v. 74)

ISBN 978-99961-48-97-2 (impreso)

1.Bases de datos. 2. Administración de bases de datos. 3.
Hadoop-Programa par computador. 4. Sistemas de almacenamiento y
recuperación de información. I. Rivera, José Guillermo, 1968-,
coaut. II. Título.

BINA/jmh

Aplicación de herramientas *big data* al Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas de El Salvador

Verónica Idalia Rosa • José Guillermo Rivera

Noel Castro
Revisión y corrección

Mauricio Gálvez
Diseño de carátula

Licda. Evelyn Reyes de Osorio
Diseño y diagramación

PRIMERA EDICIÓN
150 ejemplares
Abril, 2018

Impreso en El Salvador
Por Tecnoimpresos, S.A. de C.V.
19 Av. Norte, nº. 125, San Salvador, El Salvador
Tel.:(503) 2275-8861 • gcomercial@utec.edu.sv

ÍNDICE

Índice de ilustraciones

Índice de tablas

Ficha técnica	
Título de la investigación	Aplicación de herramientas *big data* al Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas
Equipo de investigación	Inga. Verónica Idalia Rosa Urrutia Ing. José Guillermo Rivera Pleitez
Línea de investigación	Desarrollo e innovación tecnológica
Área de conocimiento	Gestión y administración de bases de datos
Tipo de estudio	Descriptivo, experimental
Técnicas e instrumentos	No se utilizó ningún instrumento para la recolección de datos, debido a que no era necesario, más bien, lo que el Viceministerio de Vivienda y Desarrollo Urbano nos proporcion**ó** fueron los conjuntos de datos (*dataset*) para realizar las pruebas de la aplicación de las herramientas *big data.*
Muestra o participantes	Viceministerio de Vivienda y Desarrollo Urbano, proporcionando toda la información necesaria para la aplicación de las herramientas *big data*.
Fecha de realización	De febrero a noviembre de 2016
Alcance geográfico	San Salvador

Objetivos	**Objetivo general:** Aplicación de herramientas *big data* para el Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas. **Objetivos específicos:** • Desarrollar una exploración de campo en el Viceministerio de Vivienda y Desarrollo Urbano para conocer en detalle el problema existente. • Realizar el análisis del sistema de información para conocer la estructura de los datos almacenados. • Demostrar el uso de herramientas *big data,* para el análisis y visualización de la información.
Presupuesto	Sin presupuesto
Beneficiarios (Grupos de interés del estudio)	Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas

1. Introducción

Los macrodatos (*big data)* han sido muy usados en la informática y en las grandes empresas, ya que en estas se puede visualizar la gran cantidad de información que se maneja hoy en día. Es tanta la información que entra y sale que a la vez es un reto su manejo.

Big data es un término que hace referencia a una cantidad de datos tal que supera la capacidad del *software* habitual para ser capturados, gestionados y procesados en un tiempo razonable. El volumen de los datos masivos crece constantemente. En 2012 se estimaba su tamaño de entre una docena de terabytes hasta varios petabytes en un único conjunto de datos.

En el 2001 se realizó un informe de investigación en el que el analista Doug Laney del META Group [ahora Gartner] (Laney, 2016), definía "el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad".

Hoy en día, se continúa usando datos masivos y en mayor escala que hace 14 años, por lo tanto, para las empresas se hace necesario buscar herramientas que permitan dar soluciones a la demanda de grandes cantidades de datos para su procesamiento y análisis, tales son los casos de MapR, Cyttek Group, Cloudera y Hadoop, entre otros.

"*Big Data* es desde hace unos años el término de moda dentro del mundo de la informática. Dicho de otra manera, durante 2012 y parte de 2013 el 60 % de los artículos de opinión de tecnología avanzada hablan de *Big Data* como la nueva estrategia indispensable para las empresas de cualquier sector, declarando, poco menos, que aquéllos que no se sumen a este nuevo movimiento se quedarán 'obsoletos' en cuanto a la capacidad de reacción en sus decisiones, perdiendo competitividad y oportunidades de negocio contra su competencia."[1]

Debido a todo lo anterior, estamos ante una realidad que no se puede cambiar y en la que se debe ir en la misma dirección con los avances de la ciencia y la tecnología, por lo tanto, existe la necesidad de trabajar con una gran cantidad de datos, pero un mayor porcentaje de empresas no saben cómo hacerlo.

Esta investigación va a servir como referencia para dar a conocer el uso de herramientas de *big data* en El Salvador, específicamente a un sector del Gobierno.

1 Aitor Moreno, responsable de inteligencia artificial de Ibermática.

El Salvador, ubicado en Centroamérica, es un país muy pequeño en extensión territorial y población en comparación con otros países del mundo.

En cuanto a la tecnología, se trata de ir a la vanguardia sobre todo en ámbitos como el de las telecomunicaciones. El concepto de *big data* es algo novedoso, pero con mucho impulso para incursionar con él como herramienta indispensable en las telecomunicaciones, pues las empresas se preguntan cómo procesar y almacenar grandes volúmenes de datos y para luego analizarlos.

Es tanta la información que se genera a diario en la web mediante las redes sociales, los buscadores y el almacenamiento de datos en la nube, etc.; por lo que resulta abrumador. Solo el hecho de saber cómo se consigue captar y analizar dicha información es sorprendente.

También se sabe que las redes sociales hoy en día aportan mucha información relevante que los usuarios comparten libre y públicamente en la web. Para los que están inmersos en este medio, no es desconocido que a muchas personas les encanta publicar los lugares en los que están en un momento dado; las marcas que prefieren, ya sea de ropa, zapatos, accesorios, perfumes, comidas, restaurantes, etc.

Todo esto es aprovechado por las empresas para detectar tendencias en el mercado y para enfocar las acciones que se van a llevar a cabo, algo que ayuda a tomar mejores decisiones y a que los resultados sean mejores.

Por supuesto, las ventajas las obtendrán aquellas empresas que sepan cómo procesar y analizar esos datos; y es allí donde muchas se quedan estancadas al seguir haciendo los procedimientos cotidianos, por la ignorancia del uso de herramientas que facilitarían el procesado masivo de datos en poco tiempo.

Por otro lado, están los *dataset* públicos, que son archivos que se encuentran alojados en la nube de forma pública en distintos formatos; y es allí donde también surge el problema cuando los datos ya no son estructurados como comúnmente se ha acostumbrado a utilizarlos en las bases de datos relacionales tradicionales, pues estos se encuentran en formatos tales como JSON, CSV, DAT, ARFF, NCOL, etc. En estos casos se hace necesario el uso de herramientas que permitan almacenar y procesar ese tipo de ficheros.

De allí que, en el Viceministerio de Vivienda y Desarrollo Urbano, del Ministerio de Obras Públicas (MOP), está enfrentando serios problemas para el almacenamiento de grandes cantidades de información relacionada con la vivienda en El Salvador, ya que los recursos actuales

mediante bases de datos relacionales están sobrepasando los umbrales de almacenamiento por contener demasiada información; y porque la estructura SQL presenta grandes dificultades para administrarla. El MOP necesita encontrar una solución que le permita ser replicada en otros viceministerios con problemas similares, como el de Transporte.

Debido a la problemática existente en el Ministerio, tuvimos a bien tomarla en cuenta para poder ayudarles, y, en ese sentido, tener una relación Universidad-Gobierno para poder hacer uso de herramientas propias de *big data* y así hacer una propuesta que logre solucionar los problemas del procesamiento masivo de la información, del análisis de los resultados y de la visualización de los datos (ver anexo 1). Para ello se trabajó con *datasets* proporcionados por el Viceministerio, los cuales estaban en formato CSV (*Comma Separated Value*) y contenían una gran cantidad de datos sobre postulantes a vivienda de los 14 departamentos del país y sus 262 municipios, además de incluir a los extranjeros que también solicitan vivienda. Uno de los *dataset*, con 326,358 registros y 11 campos, tales como Id_Persona, P_Nombre, P_Nombre2, P_Apellido, P_Apellido2, P_Apellido3, P_sexo, P_Fecha_nacimiento, P_Id_Estado_civil, ID_Depto, ID_Municipio.

El otro *dataset*, con igual cantidad de registros y con 8 campos: Id_Persona, P_Apellido, P_Nombre, P_sexo, P_edad, P_Estado_civil, P_Depto, P_Municipio. Este último con datos filtrados y sin basura, es decir, sin datos nulos o erróneos.

Lo que se pretendía con esa información es que al hacer uso de herramientas *big data*, el procesamiento de los datos y su análisis respectivo para la toma de decisiones se hicieran en el menor tiempo posible para satisfacer la demanda de petición de vivienda de los habitantes postulantes.

2. Objeto de estudio

2.1 *Planteamiento del problema*

Debido a todo lo anterior, surge la necesidad de solucionar la problemática tomando en cuenta el conocimiento que se tiene de *big data*, en la que se hará uso de algunas herramientas para almacenar, procesar, analizar y visualizar datos de *dataset* proporcionados, los cuales contienen registros de postulantes a vivienda a nivel de todo El Salvador, por lo que esto podría ser de gran interés para instituciones

gubernamentales y privadas con problemas similares en el manejo de grandes volúmenes de datos.

2.2 *Justificación*

La justificación de la realización de dicho proyecto se debe a la gran notoriedad que está teniendo esta tendencia y porque es parte de las nuevas tecnologías.

Cualquier persona con o sin conocimientos tecnológicos se pregunta cómo se almacena toda la información que se genera en el mundo: redes sociales Facebook, Twitter, Smartcities, Instagram; o cómo Google es capaz de manejar todas las transacciones que se hacen a diario. Pero no solo se trata de eso, ya que *big data* alcanza todos los ámbitos: bolsa de valores, climatología, astronomía, *marketing*, etc., por lo que la cantidad de datos que se genera actualmente es abrumadora. Solo el hecho de llegar saber cómo se consigue captar y analizar dicha información parece una justificación bastante razonable para buscar herramientas que proporcionen soluciones atractivas.

Por otra parte, el almacenamiento de la información cada día se incrementa, por eso se ha decidido implementar nuevas tecnologías que cumplan con los requisitos de las grandes empresas, ya que almacenan cantidades enormes de información y requieren de mecanismos que les permitan realizar sus procesos de forma rápida y eficiente.

En la actualidad, la tecnología de los *big data* está tomando cada vez más realce dentro del mundo de los negocios y las estrategias. El conocimiento de esta tecnología puede ser aprovechado por cualquier empresa, con el fin de ofrecer de mejor forma sus productos y servicios.

La explotación de la tecnología de los *big data* permite que las empresas conozcan más de cerca a sus clientes, prestarles un mejor servicio, mejorar la calidad de sus productos, generar oportunidad para ingresar a nuevos mercados, completar su portafolio de clientes, entre otras tareas que generan beneficios al negocio.

Por lo tanto, la investigación sobre la tecnología de *big data* y el uso de herramientas que faciliten el procesamiento, análisis y visualización de los datos está basada en los siguientes indicadores: 1. Explorar los conocimientos que se tienen sobre *big data,* 2. Uso de las distintas herramientas para el procesamiento de los datos, 3. Conocimiento de herramientas de visualización de grandes cantidades de datos y 4. Conocer las preferencias y los elementos necesarios que se pueden utilizar para mejorar los procesos en las empresas.

2.3 *Objetivos*

2.3.1 *General*

- Aplicación de herramientas *big data* para el Viceministerio de Vivienda y Desarrollo Urbano del MOP.

2.3.2 *Específicos*

- Desarrollar una exploración de campo en el Viceministerio de Vivienda para conocer en detalle el problema existente.
- Realizar el análisis del sistema de información para conocer la estructura de los datos almacenados.
- Demostrar el uso de herramientas *big data* para el análisis y visualización de la información.

2.4 *Alcances*

- Uno de los alcances, como profesionales, es profundizar el conocimiento sobre el uso de nuevas herramientas tecnológicas para poderlas implementar en nuestras áreas de trabajo, ya que día a día los datos van aumentando y por lo tanto se necesitarán nuevas tecnologías que sean capaces de almacenar grandes cantidades de información.
- Hacer uso de las herramientas necesarias de *big data* que permitan procesar y analizar grandes volúmenes de datos para adquirir un mejor conocimiento y habilidad en la implementación de nuevas tecnologías, ya que, a medida que el tiempo avanza, los sistemas de almacenamiento van creciendo; y para ello debemos de estar preparados.

2.5 *Delimitación*

La investigación se basa principalmente en aplicar herramientas *big data* y la forma en cómo se almacena la información aumentante, por lo que se necesita que las personas que están a cargo conozcan los procesos que se deben de seguir para trabajar de una forma más efectiva.

Esta tecnología es usada en las grandes empresas que a diario generan enormes cantidades de información. Es de suma importancia que las empresas puedan invertir en la implementación de nuevos

equipos que cumplan con las características necesarias para que se genere una mayor seguridad en la transferencia de la información.

En esta investigación, se trabajará con *datasets* proporcionados por el Viceministerio de Vivienda y Desarrollo Urbano en San Salvador y se utilizarán las herramientas tecnológicas necesarias para el procesamiento de los datos, para su posterior análisis y presentación de la información.

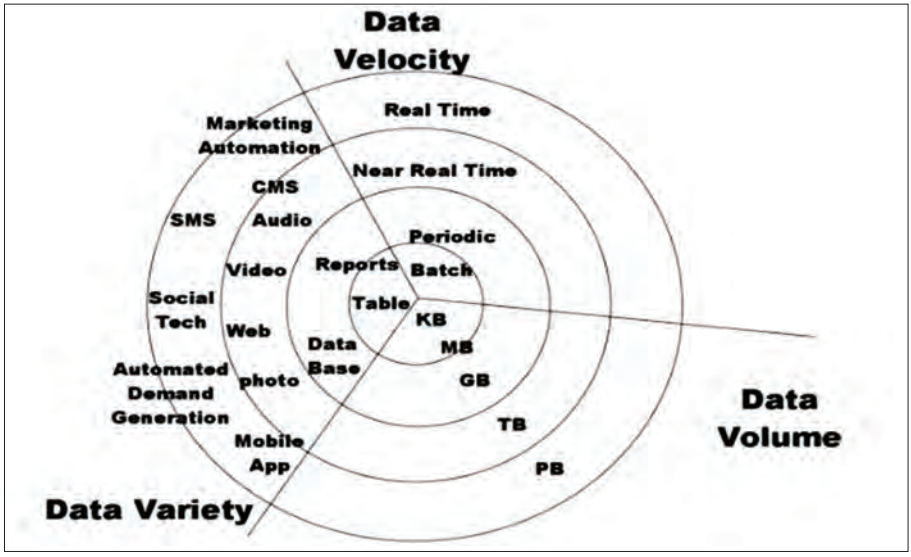
3. Marco teórico

3.1 *Qué es* big data

Big data es un término aplicado a conjuntos de datos que superan la capacidad del *software* habitual para ser capturados, gestionados y procesados en un tiempo razonable. Se considera un conjunto de datos que crecen rápidamente y que no pueden ser manipulados por las herramientas de gestión de bases de datos tradicionales (Aguilar, 2013).

En 2012, Gartner definió *big data* como "activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones". Esta definición hace mención a las tres famosas *V* de los *big data*: Volumen, Velocidad y Veracidad (figura 1), cuyos detalles se pueden consultar en el libro blanco de Fujitsu-Mitchell (Mitchell, I., 2012) y en Zicari (Zicari, 2014). Adicionalmente se han propuesto nuevas *V*. como Valor, Veracidad y Visualización; o incluso Volatilidad, Validez y Viabilidad (Jiménez, revista *Anales*, 2014).

Figura 1. Las tres primeras V de *big data*



Fuente: http://velvetchainsaw.com/2012/07/20/three-vs-of-big-data-as-applied-conferences/

El ser humano se ha visto en la necesidad de crear nuevas formas de comunicación y de almacenar la información de manera constante, siendo está de rápido crecimiento. Esta contribución a la acumulación masiva de datos se puede encontrar en diversas industrias; las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores y operaciones; de la misma manera sucede con el sector público.

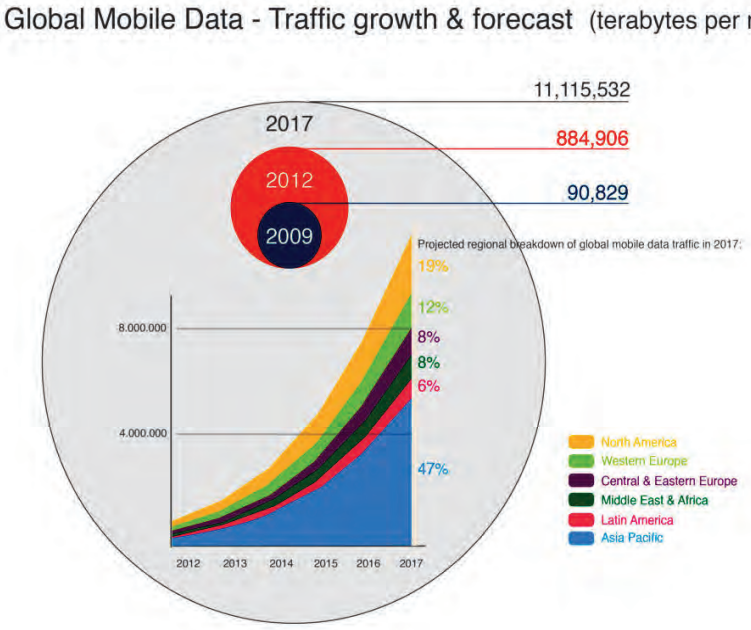
De acuerdo con un estudio realizado por Cisco, entre el 2011 y el 2016, la cantidad de tráfico de datos móviles crecería a una tasa anual de 78 %; y el número de dispositivos móviles conectados a internet excederá el número de habitantes del planeta. Las Naciones Unidas proyectan que la población mundial alcanzaría los 7.5 billones (7.5^{12}; el número seguido de doce ceros) para el 2016, de tal modo que habría cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 exabytes mensuales o 130 anuales.

Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVD anuales o 813 cuatrillones (813^{24}; el número seguido de veinticuatro ceros) de mensajes de texto. Pero no solamente son los seres humanos

quienes contribuyen a este crecimiento enorme de información, existe también la comunicación denominada *máquina a máquina* (M2M, *machine-to-machine*), cuyo valor en la creación de grandes cantidades de datos también es muy importante (Barranco Fragoso, 2014).

En la siguiente gráfica se muestra el crecimiento de datos móviles globales, la predicción y el crecimiento del tráfico (terabytes por mes), lo cual indica que cada año el crecimiento de los datos es cada vez mayor.

Figura 2. Datos móviles globales 2014. Predicción y crecimiento del tráfico



Fuente: http://www.scidev.net/filemanager/root/site_assets/global/data_spotlight/graph_global_data_fileminimizer_.jpg

En la siguiente infografía se observa una representación del uso de *big data*, sus mejores y peores aliados, así como los sectores que le sacan beneficio.

Figura 3. Infografía *big data*

Fuente: (N-economia, N economia, 2015)

3.2 *Tipos de datos*

a. **Datos estructurados (*Structured Data*).** Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.

b. **Datos no estructurados (*Unstructured Data*).** Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla, ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, *e-mails* o documentos de texto.

c. **Datos semiestructurados (*Semistructured Data*).** Datos que no se limitan a campos determinados, pero que contienen marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados que

describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Algunos ejemplos son HTML, XML, JSON y CSV. También se tiene la transformación que a continuación se explica.

Transformación

Una vez encontradas las fuentes de los datos necesarios, muy posiblemente dispongamos de un sinfín de tablas de origen sin estar relacionadas. El siguiente objetivo consiste en hacer que los datos se recojan en un mismo lugar y darles un formato.

Aquí entran en juego las plataformas ETL (Extract, Transform and Load). Su propósito es extraer los datos de las diferentes fuentes y sistemas, para después hacer transformaciones (conversiones de datos, limpieza de datos, cambios de formato…) y finalmente cargar los datos en la base de datos o Data Warehouse especificada. Un ejemplo de plataforma ETL es el Pentaho Data Integration, más concretamente su aplicación Spoon.

Almacenamiento NoSQL

El acrónimo NoSQL se refiere a Not Only SQL y son sistemas de almacenamiento que no cumplen con el esquema entidad-relación. Proveen un sistema de almacenamiento mucho más flexible y concurrente y permiten manipular grandes cantidades de información de manera mucho más rápida que las bases de datos relacionales.

Distinguimos los siguientes cuatro grandes grupos de bases de datos NoSQL.

Almacenamiento Clave-Valor (*Key-Value*). Los datos se almacenan de forma similar a los *maps* o diccionarios de datos, donde se accede al dato a partir de una clave única. Los valores (datos) son aislados e independientes entre ellos, y no son interpretados por el sistema. Pueden ser variables simples como enteros o caracteres, u objetos.

Por otro lado, este sistema de almacenamiento carece de una estructura de datos clara y establecida, por lo que no requiere un formateo de los datos muy estricto. Son útiles para operaciones simples basadas en las claves. Un ejemplo es el aumento de velocidad de carga de un sitio web que puede utilizar diferentes perfiles de usuario, teniendo mapeados los archivos que hay que incluir según el id de usuario y que han sido calculados con anterioridad.

Almacenamiento Documental. Las bases de datos documentales guardan un gran parecido con las bases de datos Clave-Valor, diferenciándose en el dato que guardan. Si en la anterior no requería una estructura de datos concreta, en este caso guardamos datos semiestructurados. Estos datos pasan a llamarse *documentos*, y pueden estar formateados en XML, JSON, Binary, JSON o en el que acepte la misma base de datos. Todos los documentos tienen una clave única con la que puede ser accedido e identificado explícitamente.

Almacenamiento en Grafo. Las bases de datos en grafo rompen con la idea de tablas y se basan en la teoría de grafos, donde se establece que la información son los nodos y las relaciones entre la información son las aristas, algo similar a lo que sucede en el modelo relacional. Su mayor uso se contempla en casos de relacionar grandes cantidades de datos que pueden ser muy variables.

Almacenamiento Orientado a Columnas. Por último, el almacenamiento *Column-Oriented* es parecido al Documental. Su modelo de datos es definido como "un mapa de datos multidimensional poco denso, distribuido y persistente". Se orienta a almacenar datos con tendencia a escalar horizontalmente, por lo que permite guardar diferentes atributos y objetos bajo una misma Clave.

A continuación, se presenta una tabla en donde se muestran las unidades de medida utilizadas en términos informáticos para los datos, y sobre todo para conocer hasta dónde hemos llegado con los grandes volúmenes de información y por ende utilizados en *big data*.

Tabla 1. Unidades de medida de los datos

Unidad	Tamaño	Significado
Bit (b)	1 o 0	Abreviatura de "dígito binario", usado por las computadoras después del código binario (1 o 0) para almacenar y procesar datos, incluyendo textos, números, imágenes, videos, etc.
Byte (B)	8 bits	Información suficiente para crear un número o una letra en inglés en código informático. Es la unidad básica de la informática.
Kilobyte (KB)	1.000 o 2^{10} bytes	Proviene de "mil" en griego. Una página de texto es de 2KB.
Megabyte (MB)	1.000KB o 2^{20} bytes	Proviene de "grande" en griego. Un archivo típico de una canción en formato MP3 es de aproximadamente 4MB.

Gigabytes (GB)	1.000MB o 2^{30} bytes	Proviene de "gigante" en griego. Una película de dos horas de duración se puede comprimir en 1-2GB. Un archivo de texto de 1GB contiene mil millones de caracteres, o aproximadamente 290 copias de las obras completas de Shakespeare.
Terabyte (TB)	1.000GB o 2^{40} bytes	Proviene de "monstruo" en griego. Todos los libros catalogados en la Biblioteca del Congreso de los Estados Unidos equivalen a 15TB. Todos los *tweets* enviados antes de finalizar el 2013 llenarían aproximadamente 18.5TB de archivos de texto. La impresión de ese archivo (a un ritmo de 15 páginas, tamaño A4 por minuto) demoraría más de 1.200 años.
Petabyte (PB)	1.000TB o 2^{50} bytes	La NSA supuestamente analiza el 1,6 por ciento del tráfico mundial de Internet, o aproximadamente 30PB por día. Tocar ininterrumpidamente 30PB de música demoraría más de 60.000 años, que corresponde al tiempo transcurrido desde que el primer *Homo sapiens* salió de África.
Exabyte (EB)	1.000PB o 2^{60} bytes	1EB de datos corresponde a una capacidad de almacenamiento de 33.554.432 dispositivos de iPhone5 con una memoria de 32GB. Se prevé que, para 2018, el volumen total del tráfico mensual de datos móviles será aproximadamente la mitad de un EB. Si este volumen de información fuera almacenado en dispositivos iPhone5 de 32GB apilados uno encima de otro, su tamaño sería 283 veces la altura del *Empire State Building*.
Zettabyte (ZB)	1.000EB o 2^{70} bytes	Se estima que en 2013 la humanidad generó 4 a 5 ZB de datos, que excedieron en 46 billones la cantidad de información de las ediciones impresas de *The Economist*. Si esa cantidad de revistas fuera puesta hoja por hoja en el suelo, cubriría la superficie total de la Tierra.
Yottabyte (YB)	1.000ZB o 2^{80} bytes	El contenido del código genético humano puede ser almacenado en menos de 1,5GB, lo que implica que 1YB de almacenamiento contendría el genoma de más de 800 billones de personas, o de aproximadamente 100.000 veces la población total del mundo.

Fuente: Emmanuel Letouzé. Big data para el desarrollo: oportunidades y desafíos. (Pulso Global de la ONU, 29 de mayo de 2012). Robert Kirkpatrick. Señales de humo globales. (Pulso Global de la ONU, 21 de abril de 2011).

3.2.1 *Qué tipo de datos se deben explorar en* big data

Ya se ha mencionado que es tanta la información que se maneja hoy en día y existe mucha en la web, por lo tanto, se tiene que tener claro lo que se desea analizar y el problema que se quiera resolver.

Como bien se sabe, existe una gran variedad de datos y en distintos formatos, su clasificación se puede observar a continuación, aunque esta puede variar de acuerdo con los avances tecnológicos (*IBM developerWorks*, 2014).

Figura 4. Clasificación de los datos *big data*

Fuente: Sitio web de IBM.

1. **Web y Redes Sociales (*Web and Social Media*).** Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, blogs, etc.

2. **Máquina a Máquina (*Machine-to-Machine M2M*).** M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura,

presión, variables meteorológicas, variables químicas como la salinidad, etc.), los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3. **Grandes Transacciones de Datos (*Big Transaction Data*).** Incluye registros de facturación en telecomunicaciones, registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4. **Biometría (*Biometrics*).** Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5. **Generado por Humanos (*Human Generated*).** Las personas generamos diversas cantidades de datos, como la información que guarda un *call center* al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

3.3 *Componentes de una plataforma* big data

Las empresas a escala mundial han atacado esta problemática desde diferentes ángulos. Todas esas montañas de información generan un costo al no descubrir el valor asociado. Actualmente, quien tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto Hadoop.

Hadoop

Es utilizado en la actualidad por numerosas compañías para satisfacer sus necesidades de procesado de *big data*. Algunas de las grandes compañías que emplean Hadoop son Yahoo!, para realizar los cálculos requeridos por su motor de búsqueda, y Facebook, que presume de tener el clúster más grande de Hadoop con más de 100 petabytes de datos.

Hadoop está inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación *MapReduce*, el cual consiste

en dividir en dos tareas (*mapper-reducer*), para manipular los datos distribuidos, a nodos de un clúster, logrando un alto paralelismo en el procesamiento. Hadoop está compuesto de tres piezas: *Hadoop Distributed File System* (HDFS), *Hadoop MapReduce* y *Hadoop Common*.

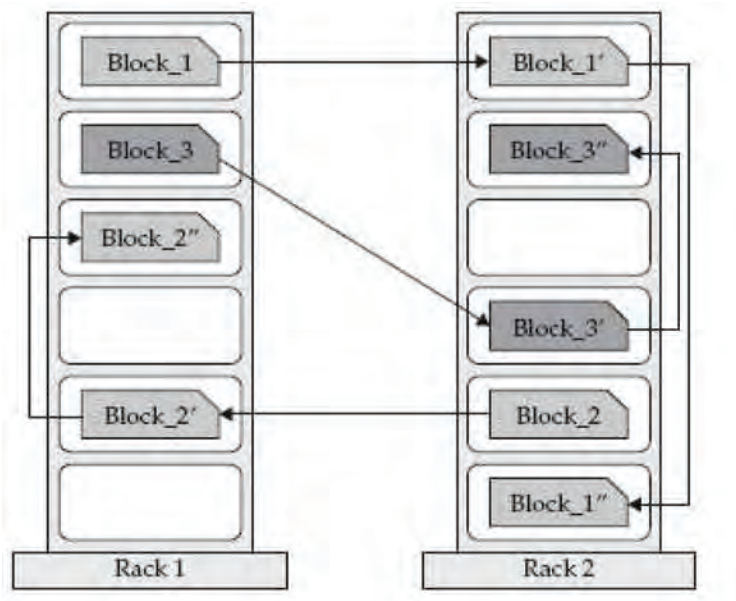
- **Hadoop Distributed File System (HDFS)**

  Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y son distribuidas a través del clúster; de esta manera, las funciones *map* y *reduce* pueden ser ejecutadas en pequeños subconjuntos; y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes de datos.

  HDFS es un sistema de ficheros que está especialmente diseñado para funcionar bien cuando se almacenan archivos grandes, que posteriormente se leerán de forma secuencial (Ghemawat, S. G., 2003).

  La siguiente figura ejemplifica cómo los bloques de datos son escritos hacia HDFS. Se observa que cada bloque es almacenado tres veces, y al menos un bloque se almacena en un diferente *rack* para lograr redundancia.

Figura 5. Ejemplo de HDFS



Fuente: Sitio web IBM.

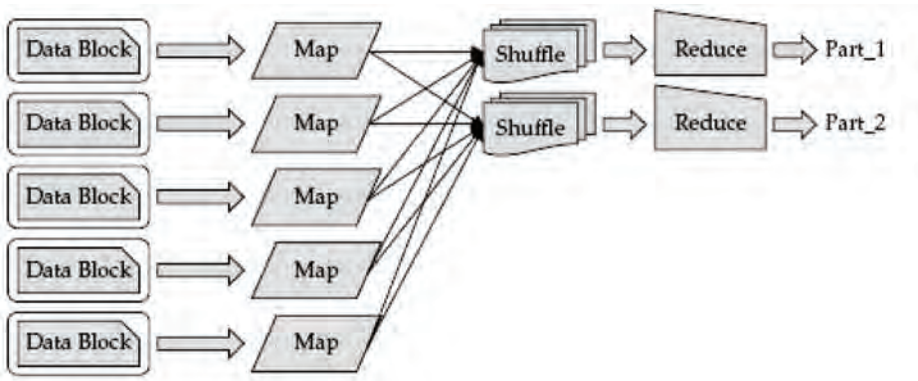
- **Hadoop MapReduce**

  El motor MapReduce es un sistema que gestiona los mecanismos para ejecutar tareas *MapReduce* de forma distribuida entre los diferentes nodos del clúster Hadoop. De nuevo, la forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador.

  Además, el ecosistema de Hadoop se compone de otros proyectos que, sin ser vitales para su funcionamiento, permiten realizar determinadas tareas de un modo más sencillo o más eficiente (Dean, *"MapReduce: Simplified Data Processing on Large Clusters"*, 2004).

  La siguiente figura ejemplifica un proceso sencillo de MapReduce.

Figura 6. Ejemplo de MapReduce



Fuente: Sitio web IBM.

- **Hadoop Common**

  Hadoop Common Components son un conjunto de librerías que soportan varios subproyectos de Hadoop.

3.3.1 *Principales distribuciones de Hadoop*

En principio, todos los componentes del proyecto Hadoop, así como los demás proyectos relacionados, se pueden descargar del sitio web de Apache, y también encontrar documentación para llevar a cabo su instalación.

No obstante, muchas compañías han decidido lanzar sus propias distribuciones de Hadoop. Las características de cada una de estas dis-

tribuciones dependen del fabricante, pero en general tienen las siguientes características:

- ✓ Ofrecen un ecosistema completo e interoperable. Como se comentó anteriormente, son muchos los proyectos que surgen en torno a Hadoop y al evolucionar cada uno de ellos de forma independiente. En ocasiones hay que ser cuidadoso de escoger una combinación de versiones que funcione correctamente. Las distribuciones de Hadoop ofrecen un ecosistema completo que ha sido testeado para garantizar que todos sus componentes funcionen correctamente.

- ✓ Ofrecen soporte (más allá del ofrecido por la comunidad de Hadoop), si bien este soporte no tiene por qué ser gratuito.

Además de las distribuciones (que pueden ser gratuitas o de pago), muchos fabricantes ponen a disposición de los usuarios lo que llaman un *sandbox*, en el que ofrecen una máquina virtual con Hadoop preinstalado (no apto para entornos en producción), que complementan con tutoriales u otros recursos útiles para los desarrolladores que no tengan experiencia.

a. Hortonworks

Una de las distribuciones más extendidas de Hadoop es Hortonworks Data Platform, que se presenta a sí misma como la distribución cien por ciento *opensource* de Hadoop y una de las que más ha contribuido al desarrollo de código del proyecto Hadoop. Hortonworks incorpora numerosos proyectos que se integran con Hadoop para aumentar el abanico de posibilidades que ofrecer a los desarrolladores y a los usuarios.

La siguiente figura muestra los diferentes componentes incluidos en Hortonworks, así como las versiones que se han escogido.

Una de las particularidades de Hortonworks es que la distribución de Hadoop que ofrece está disponible tanto para sistemas GNU/Linux como para Microsoft Windows, siendo la única distribución a día de hoy que soporta este último sistema operativo.

Además, Hortonworks ofrece una *sandbox* consistente en una máquina virtual con Hadoop preinstalado, complementado con el proyecto Hue (www.gethue.com), que ofrece una interfaz web sencilla y manejable para realizar algunas operaciones con Hadoop.

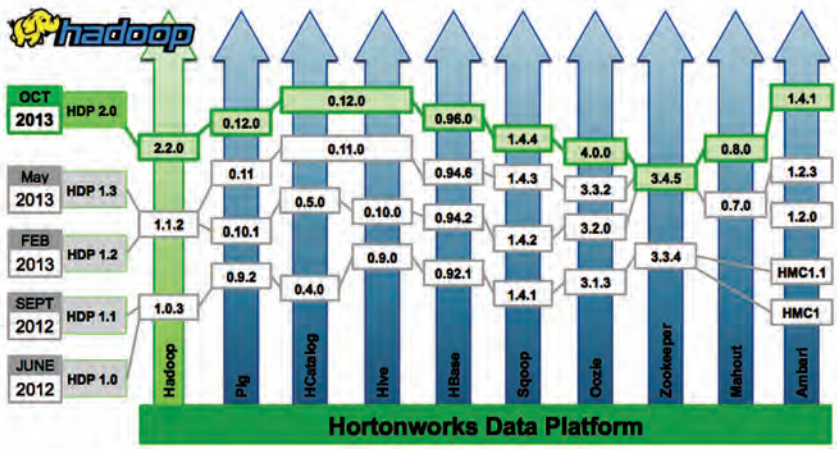
Además, Hortonworks ofrece la posibilidad de instalar y gestionar Hadoop a través de Ambari (ambari.apache.org), lo que simplifica sustancialmente la tarea de desplegar Hadoop en un clúster.

b. Cloudera

Cloudera (Cloudera, 2016) ofrece CDH (Cloudera Data Hub) como distribución de Hadoop, que también se presenta como una distribución cien por ciento *opensource* al igual que Hortonworks. Según indica la propia empresa, esta distribución cuenta con un volumen de descargas que supera a la suma de todas las demás distribuciones juntas.

Basándose en esta distribución, Cloudera ofrece diferentes soluciones, tales como Cloudera Express (que es gratuita de forma ilimitada), o diferentes paquetes de Cloudera Enterprise, que ofrece funcionalidades complementarias y soporte adicional. Además de estas soluciones, también pone a disposición de los usuarios Cloudera Live, que permite probar Hadoop *online* utilizando la interfaz Hue sin necesidad de instalarlo ni de descargar ninguna máquina virtual, por lo que resulta una alternativa interesante a Hortonworks Sandbox para usuarios que están empezando con Hadoop y no tienen recursos para desplegar una distribución. Finalmente, Cloudera es reconocido por ofrecer numerosos cursos de formación y también certificaciones en Hadoop, si bien estos cursos no son gratuitos.

Figura 7. Componentes de Hortonworks Data Platform



Fuente: http://hortonworks.com/

c. **MapR**

MapR (MapR Technologies, 2016) ofrece M3 como distribución básica de Hadoop, disponible de forma gratuita para su descarga. Al igual que las otras distribuciones discutidas anteriormente, combina Hadoop con otros proyectos de Apache que los complementan y que ofrecen funcionalidad extra, además de una consola propia para la gestión del clúster.

Por encima de M3, MapR ofrece las distribuciones de M5 y M7, que añaden más funcionalidad para entornos en producción (tales como protección de datos, alta disponibilidad, etc.) y soporte extendido.

Finalmente, MapR también ofrece la posibilidad de descargar un *sandbox* para comenzar con Hadoop, funcionando sobre una máquina virtual que el usuario puede ejecutar en su ordenador.

3.4 *Investigación y ejemplos de aplicación de* big data

En este apartado se hace referencia a investigaciones realizadas anteriormente, así como de la historia general sobre el auge de *big data* en la actualidad y a cómo beneficiarse con las herramientas que existen en esta área.

A escala de Centroamérica, y algunos países de América Latina, encontramos lo siguiente:

En México, según un artículo publicado en CentralAmericaData.com, ya no se toman decisiones sin antes analizar los *big data*, en el cual se menciona "el mercado de *Big Data* y analítica crecerá 11 % en el presente año. En el 2015 las ventas de este tipo de sistemas habían sumado $122.000 millones, y según las proyecciones de IDC (International Data Corporation), se espera que en los próximos cuatro años el ritmo se mantenga a casi 12 % (CentralAmericaData, 2016).

En la revista *TechTarget* se hizo una investigación en la que publicaron que "Big Data en América Latina avanza a pasos pequeños" (Pérez Arbesú, 2016). Encuestas recientes muestran que las organizaciones en la región no terminan de entender el concepto y que además deben afianzar la gestión de sus TIC (tecnologías de la información y la comunicación).

De acuerdo con la Encuesta de Prioridades de TI 2013 de *TechTarget*, los programas de *big data* y analítica están rezagados en América Latina. Casi una tercera parte (32 %) aún no tiene planes para incluir esta

tecnología en el negocio, mientras que un 23,9 % apenas está evaluando la opción de usar y gestionar grandes datos en su organización; sumadas, ambas cifras conjuntan más de la mitad de las respuestas de empresas que aún no utilizan *big data*, contra un 36,4 % de respuestas que indicaron que este año comenzarán su programa de grandes datos, y que lo expandirán o continuarán con el que tienen.

Tal vez este rezago se deba a un vago entendimiento de lo que *big data* realmente es. Suele interpretarse como un enorme conjunto de datos generados por grandes organizaciones, o procesos minuciosos como las transacciones financieras, pero el concepto en realidad va más allá. Manuel Cormille, *Business Information Manager* de Capgemini, explica que detrás de las aplicaciones que se usan diariamente (como el correo electrónico, por ejemplo) existe y se utiliza *big data.*

Cormille habló sobre los grandes volúmenes de datos durante la presentación de los resultados de un estudio que Capgemini realizó, en conjunto con *The Economist Intelligence Unit*, cuyo objetivo fue determinar el uso de *big data* en las organizaciones, así como dónde y cómo el uso de grandes datos está haciendo una diferencia.

El informe arrojó que un 26 % de la mejora en el desempeño actual de las compañías se ha logrado gracias a los *big data*. A su vez, un 41 % de los participantes en la investigación esperan que, tras la aplicación de esta tecnología en los próximos tres años, el negocio mejore su desempeño. Y más del 60 % de los encuestados a nivel global apuesta fuerte por *big data* como otra alternativa que posibilite un mejor desempeño de su organización.

La Estrella de Panamá publicó en el 2015 "el impacto de Big Data en las Pymes" (Quiros, 2015), en el cual se hace énfasis a que la vida *on line* se ha vuelto una forma de vida, por lo tanto, va en aumento la cantidad de datos que se encuentra en línea. Este fenómeno nos lleva a reflexionar sobre cómo la adopción de la tecnología impulsa este crecimiento de datos.

El autor del artículo hace énfasis en que, si bien el acceso a tanta información trae consigo beneficios para las empresas, también les genera retos como la discriminación adecuada de cuál es la información relevante, definir los costos de la infraestructura tecnológica, realizar reclutamiento de personal calificado para el correcto manejo de esta información, el análisis de datos y la elección de tecnología, de manera adecuada, para almacenar los datos obtenidos.

Es bien sabido que la tecnología puede ser un gran aliado para las compañías, siempre y cuando se elija acertadamente. Con la adopción

de herramientas que apoyen en el análisis de lo que llamamos *big data* se pueden tener resultados muy positivos, como mayor eficiencia en las operaciones del negocio, reducción de costos de TI, agilización en los tiempos de respuesta y aumento en la atracción y retención de clientes, entre otros.

En conclusión, el acceso a la información y su crecimiento pueden llegar a ser abrumadores, sin embargo, si se utilizan las herramientas adecuadas, estos datos se pueden volver en un fuerte aliado para las empresas, obteniendo resultados que impactarán directamente en las cifras de manera positiva.

De acuerdo con el sitio revistaitnow.com, en su publicación "Las vías para que los gobiernos lleguen a *Big Data*" (revistaitnow, 2016), el análisis de datos en el sector gobierno puede darse en Centroamérica debido a que hay países implementando diferentes soluciones. Pero, surge la pregunta ¿por qué no se está dando del todo?

Fabián Calderón responde: "¿El gobierno aplicando *Big Data*? Suena como una misión imposible de lograr por los altos costos, desorden en las áreas administrativas y datos que se van perdiendo con el tiempo sin control alguno. Pero la generación de una estrategia no está muy lejos en los gobiernos centroamericanos, existen acercamientos en la analítica de datos".

Panamá

El director de Gobierno Abierto de la Autoridad de Innovación Gubernamental (AIG), Carlos Díaz, comentó que las prioridades deberán estar enfocadas en la educación, salud, transporte, seguridad pública y medio ambiente.

Para Díaz, el uso de herramientas de analítica y *big data* en el gobierno permitirá contar con información más precisa y a tiempo al momento de establecer estrategias que conformen sus planes. Adicionalmente permitirá dar seguimiento casi en tiempo real si las acciones realizadas están llegando a los ciudadanos tal como fueron conceptualizadas.

El Salvador

Según el representante Duque Mártir Deras, del Tribunal Supremo Electoral la institución utiliza herramientas para la consulta, reporte y análisis que funcionan en torno a gráficos, para su fácil operación.

"Su principal fin es cumplir con las actividades de fiscalización designadas por ley; y depende de la orientación que definen los entes fiscalizadores, ya que cada uno determina sus procesos y recursos respectivos", precisó Mártir.

Costa Rica

Según Marcelo Jenkins, jerarca del Ministerio de Ciencia, Tecnología y Telecomunicaciones (Micitt), el gobierno y las instituciones públicas tienen iniciativas muy incipientes sobre el uso de grandes bases de datos, ya que no han llegado al punto de utilizar grandes dispositivos digitales masivamente utilizados y que generen información en la nube.

Una de las instituciones que está buscando cómo mejorar los procesos de la información es la Caja Costarricense del Seguro Social; y a pesar de que aún no cuenta con una iniciativa puntual del manejo de los *big data*, tiene un proyecto en búsqueda de la modernización del sistema.

Nicaragua

Norman René Trujillo Zapata, ingeniero en Computación y profesor de la Universidad Nacional de Ingeniería de Nicaragua, comentó que, aunque en el país no se esté implementando nada de *big data* ni análisis de datos, la educación sería la principal beneficiada en caso de concretarse alguna de estas soluciones.

Para él, en este tema se mejora en la gestión educativa, el desarrollo de nuevos métodos para la enseñanza y el aprendizaje, en la creación de nuevas carreras y opciones profesionales para los estudiantes, así como en la explotación y aprovechamiento de la producción científica generada por las instituciones educativas.

Guatemala

Andrés Guirola, gerente de ventas de Grupo Innovaciones Unidas en Guatemala, expresó que el gobierno debe invertir en la cobertura completa del recinto portuario, incluyendo accesos, prepuerto y almacenaje.

"La estrategia debería de ir todo sobre una sola plataforma, y que todos los que implementen estén sobre una sola, para así no tener un mix de sistemas", agregó.

El análisis debe romper barreras

La región tiene la capacidad necesaria para aplicar análisis de datos, pero hay diferentes retos que deben asumir los gobiernos centroamericanos para que funcionen. Neuman, representante panameño, propone como reto un cambio de actitud general en la parte administrativa de las TIC.

"Hay varios principios que deben ser parte de la 'personalidad' de los funcionarios no solo de TIC, sino en todos los niveles. El primer principio es el de que 'lo que no se mide no se puede corregir ni mejorar'. Si solo nos enfocamos en ejecución presupuestaria, perdemos de vista muchos otros indicadores que nos pueden permitir tomar mejores decisiones y gestionar más eficientemente", explicó el funcionario.

Por otro lado, Marcelo Jenkins, ministro del Micitt, precisó tres puntos claves en los que se debe apoyar el área de las TIC del gobierno, para hacerle frente a un análisis de punta en el sector gobierno.

– El primero es tener los dispositivos digitales para generar datos en el lugar y momento en el que se producen.

– El segundo es procesar esos datos que se almacenan en la nube u otro lugar, para que un encargado tenga que procesarlos estadísticamente para generar algún tipo de información.

– El tercero es la toma de la decisión, una vez obtenida la información. ¿Qué se hará con ella?

Para Jenkins, con estas tres barreras están las tres áreas siguientes en las que se podrían aplicar las TIC definitivamente:

– **Salud:** los productos *wearables* pueden ser la clave. "Todas las tecnologías de computación ambiental que generan datos sobre la salud humana, desde la presión arterial y el ritmo cardiaco hasta cuestiones como con el expediente médico digital de cada persona, generarían una gran cantidad de datos de la salud de las personas que podrían alimentar grandes bases de datos."

– **Transporte público:** "Si tuviéramos un buen sistema de pagos en autobuses, en el que se pague con un celular o tarjeta de algún tipo, los datos fluirían a la hora de tomar decisiones". Desde datos relacionados con ingreso en temas de tributación sobre lo que pagan los ciudadanos en autobuses o trenes hasta cómo se mueven las personas de un lado a otro, o cuál es su flujo, uno podría hacer un control de dónde se subió a dónde se bajó".

Por último, definió que hay que dar un giro a la orientación de la forma en que se atienden los problemas de analítica en la actualidad,

pasando de un ambiente más enfocado en análisis de gestión del desempeño de lo sucedido con los datos (pasado) a un enfoque más científico, tratando de encontrar "comportamiento oculto" en los datos (futuro), y con esto mejorar los procesos a diferentes niveles en las instituciones del país.

Todo lo anterior es lo que está sucediendo en Centroamérica, donde se puede observar que aún falta mucho por hacer en lo que respecta a este tema, sin embargo, se están haciendo algunas investigaciones; y en ciertas empresas se están utilizando herramientas *big data* para el análisis masivo de los datos.

Existen investigaciones sobre la aplicación de *big data*, las cuales se han dado en varios países del mundo; entre ellas están las siguientes:

- ✓ Investigación de *big data* en los entornos de defensa y seguridad (Carrillo Ruiz, y otros, 2013).

- ✓ Los *big data* pueden ayudar en el diagnóstico y tratamiento del cáncer (Bernardo, 2013).

- ✓ *Big data*, recurso para ciudades inteligentes (Souto, 2015); el nuevo recurso natural para las ciudades inteligentes, el cual aprovecha múltiples fuentes de datos; este analiza los datos a través de la tecnología analítica y permite a los líderes servir mejor a los ciudadanos y negocios en un mundo cambiante. Este recurso aprovecha algoritmos predictivos para resolver los problemas proactivamente.

- ✓ *Big data* ayuda al transporte inteligente en Dublín (Irlanda). Mediante el uso de herramientas de *big data*, se puede conocer el estado actual de toda la red de buses de un vistazo, y en forma detallada en áreas donde hay problemas para identificar la causa de la congestión antes de que se extienda por otras rutas. La población de Dublín en el 2013 era de 1.660.000. La información archivada ha servido para analizar *a posteriori* y entender lo que pasó¿ y así tomar medidas para optimizar el tráfico (Zikopoulos, 2015).

- ✓ *Big data* ayuda a cerveceras a delinear su *marketing* de acuerdo con el país. La brasilera Vortio (2013) analizó todas las conversaciones en las redes sociales sobre la palabra clave

cerveza en diferentes idiomas y países. Resultado: a diferentes culturas, diferentes comportamientos frente a la cerveza. En dos semanas de análisis, parece que los norteamericanos beben para relajarse; los italianos por causa de problemas de relación con su pareja; los alemanes para abastecer el tanque; los franceses para apreciar el gusto; los argentinos porque es saludable; los brasileros cuando salen de fiesta. Para esta investigación se analizaron y midieron percepciones, opiniones, etc.

- ✓ *Big data* ayudó a Obama a ganar las elecciones el 2012 (Scherer, 2012). El equipo de dirección de campaña creó una mega base de datos con información de votantes y simpatizantes a partir de múltiples fuentes desde las elecciones de 2008. Analizaron la información a fin de identificar los gustos y preferencias de sus seguidores. Crearon diferentes aplicaciones para lograr lo siguiente:
  1. Mejores decisiones con mayor volumen de datos.
  2. Traducir datos en bruto para realizar análisis predictivo.
  3. Establecer las preferencias de los votantes.
  4. Solucionar problemas de volumen de información.

- ✓ *Big data* ayuda a las tiendas Macy's, de EE. UU., a incrementar sus ventas (De Juana, 2015). Hasta el 2010, Macy's seguía utilizando hojas de cálculo Excel para analizar grandes volúmenes de datos de clientes. Ahora, con *big data*, analiza decenas de millones de terabytes (1012) de información cada día, y ha pasado de 22 horas a 19 minutos para rehacer el precio de sus artículos. Han conseguido lo siguiente:
  1. Tomar las mismas decisiones en menos tiempo.
  2. Un incremento del 10 % en sus ventas.

- ✓ *Big data* ayuda a General Electric (GE) a mejorar sus productos (FondosFidelity, 2012). En el 2011, GE invirtió mil millones de dólares en un centro de investigación para mejorar sus diferentes productos. Allí analizan un gran volumen de datos procedentes de multitud de sensores y otros dispositivos digitales.

Las investigaciones citadas anteriormente solo son una pequeña muestra de las aplicaciones de *big data* en algunas áreas, pero en términos generales los principales sectores que usan esta tecnología son los siguientes:

- ***Marketing***, conocido también como *bussiness intelligence* (*inteligencia de negocios*) y es muy utilizado en las empresas de muchos países del mundo, debido a que mediante estrategias bien definidas logran crear y administrar conocimiento sobre el medio, a través de análisis de los datos existentes dentro de la empresa.
- **Redes sociales:** estas son las que más beneficios han obtenido con el auge de los *big data*, pues mediante técnicas de recolección de datos se puede llegar a saber las preferencias de las personas, lo cual es bien utilizado por las empresas para ser más competitivas.
- **Meteorología:** manejan datos de gran tamaño y bases de datos de los matemáticos desde hace mucho tiempo, por lo tanto, las herramientas *big data* han ayudado con ese problema.
- **Ingeniería:** esta área es muy amplia; y es por ello que se ve beneficiada con las tecnologías *big data*, como el transporte, sector energético, las telecomunicaciones, etc.
- **Medios de comunicación:** debido a que estos están inmersos en todos los aspectos en el manejo de la información masiva.

3.4.1 *Ventajas de utilizar herramientas de* big data *(Alten, 2014)*

El uso de las tecnologías *big data*, en la actualidad, ha venido a ayudar en gran manera en el manejo de grandes volúmenes de datos, eso se ha podido comprobar en las investigaciones que se han mencionado anteriormente.

El uso de estas tecnologías permite el tratamiento y análisis de grandes repositorios de datos que de otra manera no fuera posible si se quisiera lograr con las herramientas de bases de datos tradicionales, ya que estas se vuelven insuficientes en todos los aspectos.

Se ha mencionado que hoy en día existe una mayor cantidad de datos que se almacenan, los cuales proceden de páginas web, aplicaciones de imágenes y videos, redes sociales, dispositivos móviles, apps o sensores, por lo tanto, es necesario contar con herramientas potentes que permitan almacenar, procesar y analizar esos datos para fines diversos, los cuales pueden ser negocios, salud, comercialización de productos, etc.

3.4.2 *Desventajas de utilizar herramientas de* big data

Se mencionan mucho las ventajas que trae consigo el uso de las herramientas *big data*; y se podría pensar que no hay desventajas que mencionar, pero realmente sí las hay, sobre todo en países subdesarrollados en los que existe mucho desconocimiento en esta área, pero que tienen enormes deseos por incursionar en las nuevas tecnologías.

Las desventajas que se pueden mencionar, por la experiencia propia en El Salvador, son las siguientes:

- Falta de profesionales expertos.
- Resistencia al cambio por miedo al fracaso.
- Falta de inversiones destinadas a implementar soluciones *big data*.
- Dificultad de integración en los procesos internos de las empresas.
- Falta de interés por innovar y por capacitar al personal de las empresas.

Por estas desventajas, las empresas en El Salvador, y probablemente en otros países del mundo, no se deciden a incursionar en la tecnología *big data*, pues la disposición de profesionales en esta área es bien limitada o nula, aparte de que hay que invertir en equipo y capacitaciones, y, por otro lado, hay que cambiar todos los procesos internos para migrar a las nuevas tecnologías.

Muchas veces la resistencia al cambio y el miedo al fracaso truncan las posibilidades de mejorar y ser mucho más productivos profesionalmente. Pero las empresas que han asumido ese reto se han dado cuenta de los beneficios que han obtenido, debido a que pueden almacenar grandes volúmenes de datos y hacer los procesos en menos tiempo.

En El Salvador, muchas empresas trabajan con las bases de datos relacionales tradicionales, como SQL, Oracle, SyBase, MySQL, trabajando con datos que están estructurados. En el peor de los casos, utilizan Excel o Access para el almacenamiento de los datos, por lo que el tiempo de respuesta es mucho mayor y la cantidad de datos que se pueden almacenar no es en grandes volúmenes.

Se sabe que los datos en la nube carecen de estructura y están almacenados en formatos que no se pueden trabajar con las bases de datos relacionales. Es por ello que se necesita de herramientas que permitan procesar y analizar grandes volúmenes de datos en el menor tiempo posible.

A pesar de las desventajas que se mencionan, las empresas se van dando cuenta de que cada día se encuentran con mayor cantidad de datos,

los cuales provienen de fuentes diversas; y el problema es que muchas veces no son datos estructurados, y es allí donde surge la necesidad de hacer uso de otras herramientas que no sean las convencionales.

Las herramientas *big data* proporcionan la factibilidad de procesar grandes volúmenes de datos en solo segundos, lo cual permite elevar la competitividad y además da lugar a nuevos enfoques e ideas, y por ende se puede cambiar la forma fundamental en la que se gestiona la empresa.

3.5 *Teorías*

Big data es un sistema genérico que debe tratar una gran cantidad de datos al que le hace falta integrar muchas herramientas para que sea lo que dice su nombre, dependiendo de la cantidad de datos, su tipo, relación entre ellos, modelos y algoritmos que han de ejecutar.

En esencia, se trata de un conjunto de tecnologías y arquitecturas diseñadas para conseguir un mejor rendimiento de grandes volúmenes de información. Como ocurre con cualquier modelo de negocio, el factor clave para obtener beneficios de *big data* no depende de la capacidad tecnológica, sino de la capacidad humana para realizar la correcta interpretación de la información que permita obtener valor de su análisis.

3.5.1 *Potencial* big data

Big data no es solo una herramienta o una tecnología, sino un conductor de una disciplina de toma de decisiones mejorada basada en análisis predictivos, que marca el comienzo de una era de cambio cultural y mejora del rendimiento. La experiencia del usuario será clave no solo en la venta de servicios, sino también en los productos.

Con *big data* la venta de productos o servicios podrá diferenciarse haciendo que su consumo suponga una experiencia personalizada para los gustos y preferencias de cada cliente. *Big data* permitirá llevar a cabo la gestión de emociones a la hora de enriquecer el consumo de los productos y servicios.

Big data no es una actividad aislada. Para el éxito, se necesita más que nunca el conocimiento del negocio que permita hacer las preguntas correctas y establecer las correlaciones oportunas. Negocio y TIC deben ir de la mano desde el primer momento.

Sin duda alguna, uno de los retos de *big data* es incorporar a su capacidad analítica información de contexto que permita adaptar y comprender el resultado del análisis con base en las condiciones del

entorno. Para ello, el verdadero conocimiento será aquel que incorpore los atributos que contextualicen el análisis.

La contextualización del dato trata de responder e incorporar al análisis información relativa a cuándo se obtuvo la fuente de origen, cómo se obtuvo, de dónde procede, cuál es su naturaleza.

Existe una gran complejidad para realizar un análisis cuando el número de variables es muy alto, mucha de la información puede no ser útil o considerarse falsa. *Big data* puede derivar en que se encuentren correlaciones falsas o falsos positivos. Para intentar solventar esta problemática en el rigor del análisis de los datos, existen ciertas premisas que ayudan a evitar errores. Las más importantes son las siguientes:

- Es primordial comenzar con pequeños pilotos para ganar experiencia y conocimiento de las nuevas tecnologías.
- Es recomendable trabajar con expertos para evitar cometer grandes errores.
- Construir un modelo que permita conocer a futuro y corrija los errores permitiendo la optimización de los procesos de negocio.

Un ejemplo claro de esto lo podemos encontrar en Google. Respecto a las predicciones sobre la epidemia de gripe en América del Norte, Google hizo un estudio para conocer cómo es que se había propagado la epidemia. Dos años más tarde, el estudio ya no era válido debido a que los datos habían cambiado y el sistema no había sido realimentado de forma asistida.

Big data es mucho más que volumen de información, muchos tipos de variables, muchos tipos de observaciones, muchos resultados. Lo realmente importante es lo que se muestra a continuación:

- Cómo están extraídos los datos.
- De dónde provienen.
- Su fiabilidad.
- De todos los datos, cuáles son los que son relevantes para integrar en el sistema.
- La relevancia forma parte de la respuesta.
- Esto limita el alcance del sistema.
- Influye sobre la definición misma del sistema.

Por otro lado, para la meteorología, el tema ya está abordado completamente por la riqueza en cuanto a la gran cantidad de datos continuos con los que se alimenta el sistema. La evolución de orígenes

de datos es constantemente reevaluada con el fin de poder integrarlas de forma adecuada en el sistema.

3.6 *Contexto*

Como se ha venido hablando anteriormente, *big data* está siendo utilizado en muchas aplicaciones e incluso en nuestra vida cotidiana. La economía familiar, la educación, el entretenimiento, la política o incluso nuestra salud se ven implicados e influenciados.

Según el libro del ingeniero informático Joyanes Aguilar, *Big Data: Análisis de grandes volúmenes de datos en organizaciones* (Joyanes Aguilar, 2013), podemos darnos cuenta hasta qué punto revoluciona el *big data* nuestra vida cotidiana.

Hoy en día, en el ámbito económico, los grandes almacenes se benefician de estrategias basadas en la comparación analítica de datos para proporcionar ofertas a los clientes. En Estados Unidos, el conocido *Black Friday* usa tecnologías *big data* para ofrecer productos a diferentes sus clientes dependiendo de la hora del día.

Otro aspecto en el que encontramos la intervención de *big data* es en la educación, una de las áreas más importantes de nuestras vidas. En últimas investigaciones realizadas, se ha descubierto el uso de *data analytics* para métodos pedagógicos, o para itinerarios personalizados para cada alumno. Sin duda, es uno de los mayores beneficios que podemos encontrar, teniendo en cuenta la decadencia que dicha área sufre en los últimos años. Con esta nueva tecnología, se adapta personalmente cada alumno a sus propios intereses, su talento o aptitudes, mejorando el rendimiento escolar.

Otro de los aspectos más importantes en la vida cotidiana es la salud, en la cual *big data* se implica también de lleno. Gracias a las nuevas tecnologías, es posible monitorizar nuestra actividad física para obtener pautas personalizadas o consejos de rendimiento a largo plazo, así como diagnosticar enfermedades.

Encontramos, también, la influencia de *big data* en el entretenimiento. Interviene proporcionando a los usuarios recomendaciones personales, teniendo en cuenta los servicios que ya han consumido. Una de las tiendas virtuales más importantes en la actualidad, Amazon, ofrece ya el servicio de predecir cuáles serán nuestras propias transacciones, haciendo más fácil así la elección de nuestros productos. Esto lo logra haciendo uso de técnicas de inteligencia artificial, específicamente haciendo uso de sistemas de recomendación.

Se ha visto, por lo tanto, que *big data* se encuentra en la vida cotidiana en casi todos los aspectos. Facilita los servicios personalizados, predice elecciones y monitoriza actividades para un mayor control sobre cada persona. Sin duda, el *big data* puede llegar a cambiar la vida tal y como se concibe hoy.

Y por último, como dato curioso, y eso se pudo apreciar en el marco teórico, el presidente Barack Obama también incluyó la utilización de técnicas *big data* para su campaña electoral.

Es asombroso todo lo que se ha investigado y cómo *big data* está involucrado en muchos aspectos de nuestra vida. Por lo tanto, no nos podemos quedar atrás en la aplicación de esta tecnología. Es por ello que surge la necesidad de trabajar con el Viceministerio de Vivienda y Desarrollo Urbano para el uso de herramientas de *big data*, y lograr de esa manera ayudar a resolver el problema de trabajar con grandes volúmenes de datos. También, esto servirá para poder dar a conocer a otras empresas en El Salvador que puedan estar interesadas en incursionar en esta área, mejorando así sus procesos de almacenamiento y el análisis de los datos para llegar a ser más productivos.

La aplicación de las herramientas *big data* consiste en la elaboración de los pasos necesarios para la utilización de dichas herramientas, tales como Hadoop y Hive, para el almacenamiento, procesamiento y análisis de los datos más completos, o para facilitar algunas tareas y realizar consultas posteriores, utilización del programa R para el análisis estadístico de los datos y luego hacer uso de herramientas de visualización, ya sea Google Chart o D3.js, según sea necesario.

3.7 *Breve historia del Ministerio de Obras Públicas*

El inicio de carreteras, en la historia de El Salvador (Mined, 1994), data de 1528, fecha en la cual fue fundada por los españoles la Villa de San Salvador, en la cual tardaron quince días en trazar las calles, la plaza y la iglesia. En ese entonces, las calles de los diferentes poblados eran únicamente de tierra, y las principales, reforzadas de piedra, ya que el vehículo de transporte utilizado eran los carretones o caballos.

La modernización de la infraestructura de transporte, que comenzó con los ferrocarriles, también se pudo apreciar en las principales ciudades de San Salvador y Santa Ana. Las carretas y carruajes que llevaban a las personas de un punto de la ciudad a otro fueron reemplazados, primero, por tranvías de tracción animal y, después, por tranvías eléctricos. Ya en la década de 1920 fueron asfaltadas las principales calles de San

Salvador, y la mejoría de las calles obedecía también a otra consideración fundamental: la llegada del automóvil allá por 1915, y, pocos años más tarde, del camión y del autobús.

A partir de entonces el crecimiento de la infraestructura vial urbana e interurbana ha ido incrementándose aceleradamente, de acuerdo con la expansión de centros industriales, de producción, de servicios, así como de los habitacionales; prueba de ello es la ampliación de la "mancha urbana", en la ciudad de San Salvador, la cual siempre se ha considerado la principal fuente generadora de crecimiento económico del país. Esto genera una demanda de servicios, especialmente de comunicación y transporte, ya que sin ellos no se puede lograr la movilidad de productos para su comercialización, además de que influyen directamente en los costos de los artículos mediante los importes en concepto de producción.

En 1905 (Oficial, No. 159, 1905) es creada una oficina bajo el nombre de Cuerpo de Ingenieros Oficiales. A esta oficina le correspondía la Dirección General de Obras Públicas como dependencia directa del Ministerio de Fomento, con la salvedad de que los trabajos de caminos eran realizados por el Ministerio de Gobernación. A este le correspondía la inmediata inspección técnica en la ejecución de todas aquellas obras que sin ser nacionales se auxiliaban con fondos del tesoro público, asignándole funciones de ejecución y mantenimiento de las obras públicas, así como la construcción y mantenimiento de los edificios destinados al servicio público, y en general, todas las obras de ornato y mejora de las poblaciones de la República, entre otras.

En 1916 (Oficial, Decreto No. 198, 1916), el Poder Ejecutivo, considerando la necesidad urgente de poseer buenas vías de comunicación en relación con el tráfico de ese entonces, así como por las necesidades individuales, comerciales, industriales y agrícolas del país, y estimando que esto debe ser, por su gran importancia, objeto de dirección y estudio especial, totalmente separados del gran número de trabajos que tenía encomendado el Cuerpo de Ingenieros Oficiales y Dirección General de Obras Públicas, emitió el Decreto de creación de la Dirección General de Caminos, la cual funcionaría como una entidad técnica- consultiva, anexa al Ministerio de Gobernación y Fomento, la cual tendría a su cargo todo lo relacionado con las vías de comunicación de la República, puentes y obras que tengan relación con estas.

Fue hasta en 1917 (Oficial, Decreto No.278 y No. 279, 1917) que se emite un Decreto Legislativo de creación del Ministerio de Fomento y Obras Públicas, el cual posteriormente asumiría todas las funciones encomendadas a las anteriores oficinas de regulación vial.

En 1920, la Dirección General de Obras Públicas, dentro del ramo de Fomento, contaba con una Sección de Caminos, así como una Sección de Arquitectura, Saneamiento y Aguas y una Sección de Caminos, Puentes y Calzadas.

En 1936, la Dirección de Obras Públicas estaba integrada por el Departamento de Hidráulica y Mantenimiento del Servicio de Aguas y de la Pavimentación de la Capital y por el Departamento de Urbanización y Arquitectura.

En 1948, el Ministerio de Fomento y Obras Públicas contaba con la Dirección General de Carreteras.

En 1949, el ramo de Fomento y Obras Públicas estaba formado por:

- Secretaría de Estado
- Comisión Nacional de Electricidad
- Oficina de Cartografía y Geografía
- Bodega
- Dirección General de Obras Públicas
- Dirección General de Carreteras

En 1951, el ramo de Fomento y Obras Públicas estaba formado por:

- Secretaría de Estado
- Dirección de Bodegas, Talleres y Canteras
- Dirección de Caminos
- Dirección de Urbanismo y Arquitectura
- Dirección de Obras Hidráulicas, y
- Dirección de Cartografía

En 1952, la Dirección de Urbanización y Arquitectura cambia nombre a Dirección de Urbanismo y Arquitectura.

En 1954, la Dirección de Urbanismo y Arquitectura y la Dirección de Caminos se convierten en direcciones generales dentro del ramo de Obras Públicas. Todos estos cambios son producto de la necesidad de ordenar el crecimiento de las ciudades, tanto en su parte arquitectónica como en infraestructura, por lo cual se le encomiendan las funciones específicas de construir, mantener y rehabilitar la infraestructura urbana y vial del país. En esta última se incluyen las carreteras interurbanas, rurales y urbanas, las cuales se constituyen en uno de los pilares que sostienen la economía nacional.

Actualmente el MOP, dentro de su organización, cuenta con tres viceministerios: de Transporte, el cual se encarga de la reglamentación del tráfico, tanto rural como urbano, así como de los transportes aéreo,

terrestre y marítimo; de Vivienda y Desarrollo Urbano, que se encarga de todo lo relativo a las proyecciones de desarrollo urbano, planificación y ejecución de los diferentes programas, cuyo objetivo primordial es disminuir el déficit habitacional del país; y de Obras Públicas, que es el encargado de dirigir la planificación, construcción, rehabilitación, reconstrucción, ampliación y mantenimiento de la infraestructura vial del país.

3.7.1 *Viceministerio de Vivienda y Desarrollo Urbano*

Misión
Ser una organización moderna, innovadora, efectiva, transparente y con liderazgo institucional, rectora del desarrollo y ordenamiento territorial, la política de vivienda y el desarrollo de asentamientos humanos integrales en ambientes sostenibles.

Visión
Planificar, promover, normar, coordinar y facilitar el desarrollo y el ordenamiento territorial, de la política de vivienda y asentamientos humanos sostenibles que garanticen el progreso y bienestar de la población.

Valores MOPTVDU (Ministerio de Obras Públicas, Transporte, Vivienda y Desarrollo Urbano)

- Responsabilidad social y orientación al usuario
- Compromiso
- Eficiencia
- Integridad
- Lealtad
- Respeto
- Responsabilidad individual
- Trabajo en equipo

Adquiere la denominación actual de Ministerio de Obras Públicas mediante decreto Legislativo N.º 1059, publicado en el Diario Oficial

del 19 de junio de 1953, considerando que es conveniente armonizar las disposiciones administrativas con las del aspecto puramente fiscal.

El Reglamento anterior del Poder Ejecutivo, dado en 1958, fue sustituido por el actual Reglamento Interno del Órgano Ejecutivo, emitido mediante decreto ejecutivo Nº 24, publicado en el Diario Oficial 70 el 18 de abril de 1989.

Art. 43. - Compete al Ministerio de Obras Públicas, Transporte y de Desarrollo Urbano:

Área de Vivienda y Desarrollo

1. Formular y dirigir la Política Nacional de Vivienda y Desarrollo Urbano; así como elaborar los planes nacionales y las disposiciones de carácter general a que deban sujetarse las urbanizaciones, parcelaciones, asentamientos en general y construcciones en todo el territorio de la República.
2. Planificar, coordinar y aprobar las actividades de los sectores de Vivienda y Desarrollo Urbano en todo el territorio nacional.
3. Dirigir como órgano rector de las Políticas Nacionales de Vivienda y Desarrollo Urbano; determinando en su caso, las competencias y las actividades respectivas, de las entidades del Estado en su ejecución y orientando la participación del sector privado en dicha política.
4. Elaborar, planificar y velar por los planes de desarrollo urbano de aquellas localidades cuyos municipios no cuentan con sus propios planes de desarrollo local.
5. Planificar y coordinar el desarrollo integral de los asentamientos humanos en todo el territorio nacional.
6. Aprobar y verificar que los programas que desarrollen las instituciones oficiales autónomas que pertenecen al ramo, sean coherentes con la Política de Vivienda y Desarrollo Urbano emitida por el Ministerio, debiendo coordinar con las mismas todo lo relacionado con los asentamientos humanos dentro del territorio de la República y verificar que estos sean coherentes con los planes de desarrollo emitidos por las municipalidades competentes.
7. Adecuar y vigilar el cumplimiento de las Leyes y Reglamentos que en materia de urbanismo y construcción existieren.

Con las herramientas *big data* que se proponen, se dará a conocer lo que se puede hacer con cualquier conjunto de datos que se tenga, y, al final, el Viceministerio de Vivienda y Desarrollo Urbano decidirá si utiliza las nuevas tecnologías o continúa realizando los procedimientos como lo hace hasta el momento.

No hay estudios previos acerca de este tema en El Salvador, sin embargo, hay empresas que ya están inmersas en el uso de esta tecnología, tal es el caso de Dell en El Salvador, Telemóvil y algunos bancos. En otros países sí hay investigaciones relevantes acerca del uso de *big data* y de cómo ha ayudado a mejorar los procesos que se realizan en muchas áreas de la vida cotidiana. Anteriormente se hizo mención de esas investigaciones y aplicaciones que se han realizado sobre el uso de *big data*.

El objetivo principal que se pretende lograr con la investigación es solucionar el problema de los datos del Viceministerio de Vivienda y Desarrollo Urbano, proporcionando herramientas de *big data* para el manejo de grandes volúmenes de datos, para su respectivo almacenamiento, análisis y representación gráfica de la información generada.

Con toda la información que se tiene, se procesarán los datos haciendo uso de Hadoop; y luego, con la herramienta Hive se realizarán las consultas necesarias. También se utilizará el programa R para hacer análisis estadístico de los datos, y por último se hará uso de herramientas de visualización, tales como Google Chart y D3.js, para presentar visualmente los resultados obtenidos. A continuación, se encuentra la descripción de las herramientas *big data* a utilizarse en la investigación.

3.8 *Descripción de las herramientas que han de utilizarse en la investigación*

3.8.1 *Qué es Hadoop*

En la actualidad, Hadoop (White, *Hadoop: The definitive guide*, mayo 2012) es un proyecto de *software* libre, con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos. Los principales componentes que constituyen Hadoop son el sistema de archivos distribuido Hadoop Distributed File System (HDFS) y el motor MapReduce.

El HDFS está inspirado en el GFS, que permite distribuir los datos entre distintos nodos de un clúster (llamados *datanodes*), gestionando la distribución y la redundancia de forma transparente para el desarrollador que vaya a usar esos datos.

El motor MapReduce es un sistema que gestiona los mecanismos para ejecutar tareas MapReduce de forma distribuida entre los diferentes nodos del clúster Hadoop. De nuevo, la forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador, además, el ecosistema de Hadoop se compone de otros proyectos que, sin ser vitales para su funcionamiento, permiten realizar determinadas tareas de un modo más sencillo o más eficiente.

Existen varias distribuciones de Hadoop, tales como Hortonworks, Cloudera, MapR, pero para este trabajo de investigación se utilizará la primera. Anteriormente se habló de esta distribución de una forma más específica y de las ventajas que tiene para utilizarla.

Para llevar a cabo el despliegue de Hadoop, es necesario comenzar realizando los siguientes pasos:

1. **Decidir la arquitectura física del sistema**. Este probablemente es el paso más complicado, en un principio, pues requiere tener una visión global del uso que se le dará al sistema. Esta decisión incluye conocer el número de nodos del clúster, los servicios que ejecutará cada uno de ellos, la distribución física de los equipos, etc. Una de las principales ventajas que ofrece Hadoop consiste en que es relativamente sencillo adaptar la infraestructura física a las necesidades que surjan en un futuro (por ejemplo, añadiendo nuevos nodos).

2. **Decidir la distribución que se va a desplegar**, en caso de que se haga. Como se ha comentado anteriormente, la distribución que se utilizará será Hortonworks, aunque todas las demás ofrecen combinaciones de diversos proyectos del ecosistema Hadoop, cuya interoperabilidad ha sido testeada, además de funcionalidades extra y soporte.

Para la elaboración de la propuesta metodológica se empleará una arquitectura basada en dos máquinas virtualizadas (de las cuales una hará de máster y la otra de esclavo), esto por la carencia de recursos físicos para llevar a cabo el despliegue.

En cuanto a la distribución, se sugiere instalar Hortonworks, puesto que dispone de un instalador y un configurador del clúster basado en Apache Ambari, lo que simplifica enormemente el proceso.

Por qué Hadoop. En la actualidad, la cantidad de datos que se generan a cada segundo es inmensa. En 2012, la International Business Machines Corporation (IBM) publicó una infografía (IBM, 2014) en la que, basándose en fuentes tales como estudios de IDC (International Data Corporation) o EMC (**EMC** Corporation, empresa estadounidense de software y hardware), resumía el estado actual en lo referente a la cantidad de datos que inundaba la web.

Algunas de las cifras más significativas son las siguientes:

- Se mandan 294.000 millones de *e-mails* diariamente.
- Se suben 100 terabytes de datos a Facebook diariamente.
- Se generan 5 exabytes (millones de terabytes) cada dos días.
- Existen en el universo digital 2.7 zettabytes (miles de millones de terabytes) de datos.

Esta inmensidad de datos se puede explicar fundamentalmente en base a los siguientes tres orígenes:

- La interacción entre humanos que emplean un sistema informático que registra información mientras se produce dicha interacción, son los casos de los *e-mails*, de los foros de internet y de las redes sociales, en los que los datos los generamos los humanos y son almacenados o procesados por máquinas.
- La interacción entre un humano y una máquina. Este caso se da cuando navegamos por internet y los servidores web generan logs con información sobre el proceso de navegación, o cuando compramos en una plataforma de comercio electrónico o empleamos la banca *online* y un sistema registra nuestras transacciones.
- La interacción entre máquinas (M2M), en la que son varias máquinas las que intercambian información entre ellas y la registran. Algunos ejemplos son los sistemas de monitorización, en los que un sistema de sensores proporciona la información que recibe a otras máquinas para que realicen algún procesado sobre ella.

Probablemente, más importante que la cantidad de datos que se genera en la actualidad es entender que este ritmo crecerá en el futuro,

puesto que cada vez son más las personas que tienen acceso a internet y la variedad de dispositivos que se conectan a la red. En esta misma infografía, IBM revela que en 2020 se generarán 35 zettabytes de datos anualmente.

Este ritmo de generación de datos introduce numerosos desafíos en lo que concierne al modo en que se almacenan y, especialmente, en la forma en la que deben ser procesados. Resulta evidente que los sistemas tradicionales son incapaces de manejar esta información de una forma eficiente, puesto que no están preparados para soportar la explosión de datos de los últimos años.

Las principales compañías que vinieron notando esta necesidad de sistemas para el almacenamiento y procesado más eficiente de grandes cantidades de datos fueron los buscadores de internet.

La labor de un buscador de internet es, en teoría, relativamente sencilla. Lo primero que debe hacer es rastrear la web, siguiendo los hipervínculos de cada página para ir construyendo un grafo (una estructura de datos que relaciona diferentes nodos, en este caso páginas web entre sí por medio de enlaces). A continuación, debe elaborar un índice invertido en el que se pueda localizar fácilmente una web dados unos términos de búsqueda.

Además, los buscadores suelen emplear una función de *ranking* mediante la que asignan a cada página web un peso en función de su relevancia, que se puede calcular con base en la cantidad de páginas que enlazan con la página asignada y, al mismo tiempo, a la relevancia de cada una de las páginas.

La principal dificultad que encuentran los buscadores es que la cantidad de páginas web disponibles es inmensa, lo que dificulta llevar a cabo todo el proceso de indización en un tiempo razonable; como para mantener los resultados de búsqueda actualizados.

En el año 2003 (Ghemawat, Gobioff, & Leung, 2003), Google publica su famoso artículo en el que describe Google File System, un sistema de ficheros escalable y distribuido que pretende subsanar la dificultad de tener que almacenar grandes cantidades de datos de forma confiable y proporcionando un alto rendimiento para aplicaciones que realizan un uso intensivo de estos datos.

Al año siguiente, Google publica otro artículo (Dean, *"MapReduce: Simplified Data Processing on Large"*, 2004) en el que describe el paradigma de programación MapReduce, cuya finalidad es llevar a cabo un procesado distribuido de grandes cantidades de datos de forma eficiente.

MapReduce

MapReduce (White, Hadoop: The Definitive Guide, 2012) es un desarrollo que responde a la necesidad de Google de procesar grandes cantidades de datos de manera eficiente, de forma paralela. Además, es un paso intuitivo tras el desarrollo de GFS. Puesto que ahora hay grandes cantidades de datos almacenadas de forma distribuida entre varios equipos, resulta oportuno realizar un procesamiento también distribuido de estos datos. La idea detrás de MapReduce es sencilla: una aplicación MapReduce cuenta con una rutina *map*() y otra rutina *reduce*(), que son las que dan nombre a este modelo de programación.

La rutina *map*() recibe una tupla clave-valor (<*k, v*>) y devuelve un conjunto de tuplas clave-valor (<*km, vm*>).

Posteriormente, todas las claves devueltas por las rutinas *map*() ejecutadas se ordenan y se agrupan por clave, resultando un conjunto de tuplas que contienen una clave y una lista de valores. Por ejemplo, si las rutinas *map*() habían devuelto las tuplas <kmi, vmi,1>, <kmi, vmi,2>, …, <kmi, vmi,n>, tras esta fase todas estas tuplas se agruparán en una tupla <kmi, [vmi,1, vmi,2, …, vmi,n]>.

Por último, la rutina *reduce*() recibe como entradas estas tuplas agrupadas y devuelven, para cada una de ellas, un conjunto de valores (<*vr*>).

Probablemente trabajar bajo este paradigma de programación, para muchos, no sea fácil de comprender y aplicarlo, pero por esa razón se trabajará con una herramienta que viene incluida en Hadoop, la cual es mucho más fácil, sobre todo para aquellos que en alguna ocasión han trabajado con base de datos relacionales como SQL. Esta herramienta es conocida como Hive, la cual incorpora el procesado MapReduce.

Hive

Apache Hive (Capriolo, Wampler, & Rulberglan, 2012) es un proyecto que forma parte del ecosistema Hadoop y, por ello, viene incluido en muchas distribuciones de Hadoop, y también en la distribución Hortonworks.

El propósito de Hive es, en cierto modo, emular un sistema de bases de datos relacional encima de Hadoop.

Así, el usuario podrá crear tablas e insertar datos (o crearlas a partir de ficheros existentes en HDFS), para posteriormente consultarlas empleando un lenguaje de modelado y de consulta muy similar a SQL.

Es importante entender que esta lógica funciona bien cuando trabajamos con datos que son estructurados, puesto que el concepto de *tablas* en el modelo relacional estructura los datos en columnas (campos) y en filas (registros).

Hive es una herramienta adecuada para usuarios que estén familiarizados con las bases de datos relacionales. Permite crear tablas y hacer consultas sobre ellas empleando un lenguaje similar a SQL, si bien estas consultas se traducirán automáticamente a rutinas MapReduce.

3.8.2 *Qué es R*

Se puede definir R (Foundation, s/f) desde las siguientes dos perspectivas:

- R es un entorno de *software*.
- R es un lenguaje de programación.

Fundamentalmente R puede ser definido como un entorno *software* para el análisis matemático y estadístico de datos, en cierto sentido similar a herramientas tales como Microsoft Excel. Mediante el entorno de R, vamos a ser capaces de manipular datos (por ejemplo, cargarlos desde ficheros, editarlos, volverlos a almacenar...), analizarlos y presentar los resultados gráficamente para facilitar su interpretación.

El entorno *software* viene acompañado de un lenguaje de programación que pone a nuestra disposición las funcionalidades típicas de un lenguaje de propósito general (manejo de variables, tipos y estructuras de datos, operadores, mecanismos de control del flujo de ejecución, funciones, etc.) combinadas con librerías y herramientas específicas para facilitar el análisis de datos. Utilizando este lenguaje es relativamente sencillo implementar nuestras propias funciones y *scripts* para automatizar el procesamiento de ciertos datos.

En la práctica, estas dos perspectivas están muy relacionadas. Así, por ejemplo, para interactuar con el entorno de R se utilizarán expresiones escritas en el lenguaje R.

Por qué utilizar R

Actualmente existe una amplia gama de herramientas que pudiéramos pensar en utilizar a la hora de llevar a cabo análisis de datos (por ejemplo, Microsoft Excel, S-PLUS, una versión comercial del lenguaje S, SAS, SPSS de IBM, etc.). Así pues, una de las cuestiones que las empresas pudieran

plantear en esta metodología es por qué elegir R como herramienta de análisis de datos.

Algunas de las razones que se podrían emplear a la hora de justificar la decisión incluyen lo siguiente:

- R es *software* de código libre con licencia GNU GPL (General Public License).
- Mientras que las principales herramientas de análisis tienen un costo (algunas con precios bastante elevados), R es completamente gratuito.
- Existen versiones para los sistemas operativos más comunes: Windows, Mac OS X y Linux.
- Posee una comunidad de usuarios amplia y muy activa, por lo que va a resultar relativamente sencillo encontrar documentación o ayuda en foros si resulta necesario.
- El entorno es fácilmente extensible, mediante el desarrollo de paquetes. Debido a esto, evoluciona rápidamente: nuevos algoritmos y técnicas de análisis se incorporan con regularidad.
- R y sus extensiones nos ofrecen una gran variedad de herramientas de análisis y visualización de datos. Actualmente existen más de 5.000 paquetes disponibles para ser instalados en el entorno.

3.8.3 *Herramientas de visualización*

Google Chart (Charts, 2012) es una aplicación de Google para realizar estadísticas web, de fácil uso para desarrolladores de *software* web, usado en muchos campos, como Google Analytics; se puede usar con diferentes formatos, Json, JavaScript y *plugins* que se pueden integrar con varios lenguajes de programación.

Esta herramienta permite realizar gráficos atractivos, y existe una gran variedad de galerías disponibles en el sitio de Google para utilizarlas y adaptarlas a las necesidades de análisis de cada persona.

D3.js (D3js.org, 2015), o simplemente D3, de documentos basados en datos. Es una biblioteca JavaScript para producir visualizaciones de datos dinámicos e interactivos en los navegadores web. Hace uso ampliamente de SVG, HTML5 y estándares CSS. En contraste con muchas otras bibliotecas, D3.js permite un gran control sobre el resultado visual final. Para poder hacer uso de esta herramienta, es necesario conocer de JavaScript, por consiguiente, hay que aprender ese lenguaje de programación.

4. Metodología de la investigación

4.1 *Metodología*

La investigación se realizó con el objetivo de aplicar herramientas *big data* en el procesamiento, análisis y visualización de los datos del Viceministerio de Vivienda y Desarrollo Urbano del MOP. Se hizo un estudio descriptivo experimental usando conjuntos de datos de postulantes a vivienda en todo el territorio nacional, los cuales fueron proporcionados por dicho Viceministerio.

Con esos datos se trabajó para aplicar las herramientas *big data* y poder demostrar la efectividad en el procesamiento y análisis de datos masivos.

4.2 *Participantes*

Viceministerio de Vivienda y Desarrollo Urbano, proporcionando toda la información necesaria para la aplicación de las herramientas *big data*.

4.3 *Instrumento para la recolección de datos*

No se utilizó ningún instrumento para la recolección de datos, debido a que no era necesario, más bien, lo que el Viceministerio de Vivienda y Desarrollo Urbano nos proporcionó fue los conjuntos de datos (*dataset*) para realizar las pruebas de la aplicación de las herramientas *big data*.

4.4 *Procedimiento*

Lo primero que se realizó fue revisar las bases de datos para verificar cómo están almacenados y su estructura. Posteriormente se revisaron los procesos que se llevan a cabo y cuáles son las consultas necesarias para la generación de reportes.

Por último, se solicitaron los *dataset* para poder hacer las pruebas y demostraciones con las herramientas *big data* propuestas.

5. Discusión de resultados

Después de haber explorado los datos del Viceministerio Vivienda y Desarrollo Urbano y revisar los *dataset* proporcionados, en términos generales la aplicación de las herramientas *big data* consistió en lo siguiente:

1. Almacenar y procesar los *dataset*, los cuales contenían información sobre postulantes a vivienda en todo el territorio salvadoreño, usando Hadoop.
2. Con la herramienta Hive, que viene en la distribución Hortonworks de Hadoop, se hicieron las consultas necesarias, ya que esta herramienta es similar a las instrucciones que se utilizan en SQL, por lo que, para los que están acostumbrados a trabajar con base de datos relacionales, les será fácil entender la lógica de cómo trabaja Hive. En nuestro país, SQL es el *software* más utilizado para bases de datos; esa es la razón por la que se seleccionó esta herramienta.
3. Para el análisis de datos estadístico, se utilizó el programa RStudio. Las razones por las que se seleccionó este programa fueron explicadas en el marco teórico. Este programa nos devolvió información sobre datos importantes de los postulantes a vivienda que están almacenados y, a la vez, permitió que realizáramos conclusiones con base en los resultados.
4. Después de haber hecho un análisis estadístico y las consultas pertinentes de los datos, se procedió a realizar los gráficos necesarios para una mejor comprensión de los resultados y, con base en ello, sacar conclusiones y poder tomar decisiones. Las herramientas que se pueden utilizar son Google Chart y D3.js

6. Propuesta: "aplicación de herramientas *BIG DATA* al viceministerio de vivienda y desarrollo urbano del ministerio de obras públicas"

6.1 *Desarrollo y metodologías utilizadas*

En este apartado se tratará de explicar los procesos fundamentales que será necesario implementar para lograr el objetivo planteado. Estos deben ser de forma ordenada, iniciando desde la elección del *dataset* y luego con el uso de las herramientas necesarias para el almacenamiento, procesamiento, análisis y visualización de los datos.

Para este caso, como se ha mencionado anteriormente, se hizo uso de dos *dataset*, los cuales contienen registros de los postulantes a vivienda en todo el país. Estos ficheros han sido manipulados de tal manera que pueden ser almacenados con Hadoop; y luego, con la herramienta Hive, hacer las consultas necesarias; posteriormente, utilizar el programa RStudio, con el que se harán análisis estadísticos de los datos más relevantes de los ficheros utilizados.

Con la información obtenida, se procederá a elaborar los gráficos, usando cualquiera de las herramientas de visualización mencionadas anteriormente, los cuales reflejarán lo más importante de los datos y así obtener conclusiones que servirán en la toma de decisiones.

Ficheros utilizados

Los ficheros están en formato CSV, los cuales han sido proporcionados por el Viceministerio de Vivienda y Desarrollo Urbano. Anteriormente se mencionó la cantidad de registros que tienen y sus campos.

6.1.2 *Almacenando y procesando datos con Hadoop*

Como se dijo antes, Hadoop es un proyecto de *software* libre, con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos. Los principales componentes que constituyen Hadoop son el sistema de archivos HDFS y el motor MapReduce.

Por lo tanto, el uso de Hadoop se puede hacer de dos formas: primero, instalando CentOS, que es la opción recomendada para desplegar posteriormente Hortonworks. Para realizar las consultas respectivas de los *dataset*, en este proyecto, se hizo uso de la herramienta

Hive. Y segundo, utilizando la versión más compacta de Hadoop, que es Sandbox, esta requiere de menos recursos de *hardware*. Por lo tanto, para esta investigación se utilizó la segunda opción.

Con la herramienta Hive, se realizaron consultas que son importantes para el Viceministerio de Vivienda y Desarrollo Urbano, tales como:

- Listado de postulantes a vivienda por género.
- Listado de postulantes a vivienda de un municipio determinado.
- Listado de postulantes a vivienda por rango de edades.
- Listado de postulantes a vivienda por estado civil, etc.

6.1.3 *Analizando datos con R*

Dentro de los objetivos se encuentra la demostración de herramientas *big data*. Una de ellas es el programa R para el análisis estadístico de los datos. Por lo tanto, se importó el *dataset* al programa R y se comenzó a realizar los análisis estadísticos respectivos.

Debido a que el Viceministerio de Vivienda y Desarrollo Urbano tiene conjuntos de datos sobre postulantes a viviendas, le interesa conocer lo siguiente:

- Cantidad de postulantes por género
- Cantidad de postulantes por municipio
- Cantidad de postulantes por estado civil
- Cantidad de postulantes por edades
- Cantidad de postulantes por género y edad
- Municipios más demandados en vivienda
- Gráficos representativos de barras y de pastel

6.1.4 *Visualizando datos*

Cada día se generan billones de datos, e incluso de forma individual estamos inundados de libros y correos electrónicos, fotografías, videos, música, y, aparte de ello, se almacena más información y documentos en la nube, como, por ejemplo, Google Drive, o mediante cualquier disco virtual, lo cual hace que exista demasiada información y de la que es imposible conocer el poder que puede tener.

Asimismo, si se añade el concepto de *big data*, "grandes conjuntos de datos (*datasets*)", se podrá comprender que la principal dificultad no es la captura y almacenamiento de los datos, sino más bien el análisis y su posterior visualización estática o interactiva.

Los datos por sí solos no tienen sentido, a menos que se conviertan en información comprensible y útil, para luego acceder al conocimiento.

Se sabe que los datos son la materia prima de la información, y esta, la del conocimiento; y, por supuesto, al tener conocimiento se pueden tomar decisiones acertadas.

Para las empresas, el tiempo en procesar todos los datos generados es valioso. Pero se sabe que esto no se puede lograr sin las herramientas adecuadas, que permitan transformar la multitud de datos. Una vez se haya logrado transformar los datos en información, es necesario presentarla de manera atractiva y que sea fácil de comprender, por lo tanto, la magia de la visualización de información radica en la captura y síntesis previa de la información mediante el uso de diversas técnicas visuales, como diagramas, gráficas, esquemas, nubes de palabras, conexiones, grafos, para que pueda ser transformada y con ello facilitar su compresión.

Es por ello que, después de haber hecho uso de herramientas para el almacenamiento, procesamiento y análisis de los datos, es necesario hacer uso de herramientas de visualización que permitan comprender los resultados, pero en un formato gráfico, debido a que para la mente humana es mucho más fácil entender imágenes que únicamente texto o números.

De esto se encarga la visualización de datos, que no es otra cosa que el diseño de la comprensión de manera atractiva. Los datos deben ser comprendidos de manera efectiva. El objetivo de toda buena visualización debe ser centrar la atención del interesado en la información que es realmente relevante e importante.

Una vez se disponga de ese nuevo conocimiento, se facilitará hacer análisis y sacar conclusiones que serán importantes para la toma de decisiones.

Figura 8. Diagrama de estructura de los datos hacia la sabiduría

Fuente: Hey, J.: The Data, information, knowledge, Wisdom Chaim: The Metaphorical Link.

Lo que se pretende con el diagrama anterior es pasar de los datos a la sabiduría, porque de eso dependerá tomar buenas decisiones.

Las herramientas de visualización que se pueden utilizar son Google chart, Jqplot o D3.js. Cada una de ellas tiene sus propias características, tal y como se mencionó en apartados anteriores, por lo que, dependiendo del uso que quiera dárseles en la representación gráfica de los datos, así será la selección de cualquiera de ellas; o se pueden utilizar las tres si se desea.

Para la implementación de la metodología propuesta, es necesario tomar en cuenta ciertos requisitos técnicos, los cuales se detallan a continuación.

6.2 *Identificación de requisitos*

Los requisitos técnicos y las tecnologías por utilizar son los siguientes:

Tabla 2. Requisitos T*écnicos para usar herramientas big data*

Herramienta tecnológica	Requisito de *hardware*	Requisito de *software*
Hadoop	✓ 5.ª generación del procesador Intel® Core™ i3, i5 o i7 ✓ Memoria de 16 GB expandible a 32 GB ✓ Disco duro de 1 a 2 TB	✓ Windows y Linux ✓ Linux de 64 bits ✓ Oracle VM VirtualBox ✓ CentOS 6.5 o más reciente ✓ Distribución Hortonworks o Sandbox
Programa R	No requiere de características especiales.	✓ Windows, Mac o Linux ✓ R o RStudio
Google chart, D3.js	No requiere de características especiales.	✓ Cualquier sistema operativo ✓ Cualquier editor de texto: bloc de notas, *sublime text* o *notepad* ++ ✓ Navegador web ✓ Servidor web

Fuente: elaboración propia

Hadoop es la única herramienta que requiere de muchos recursos de *hardware*, sobre todo de memoria RAM y de disco duro. Es por ello que lo más recomendable es que esté instalado en un servidor, pero debido a las limitantes encontradas y no disponer de suficientes recursos en el equipo utilizado se optó por crear máquinas virtuales para simular un máster y un esclavo, y hacer el despliegue en un entorno similar a uno de producción.

Los demás programas que han de utilizarse son gratuitos y con lecturas de manuales en la web se facilita.

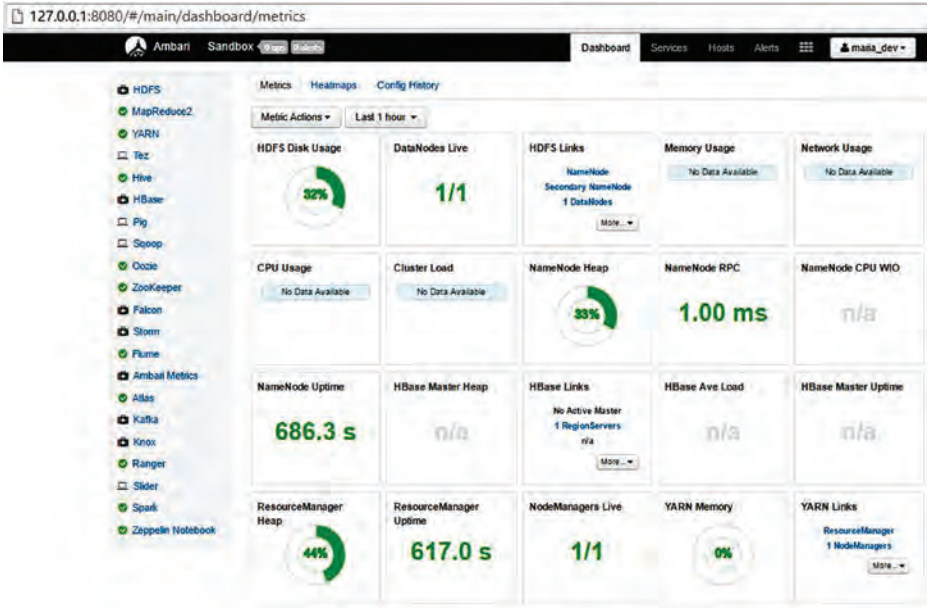
6.3 *Utilización de Hadoop*

1. Como se va a utilizar Hadoop para el procesamiento de los datos y posteriormente realizar las consultas necesarias con Hive, será necesario desplegarlo. Para ello es recomendable tomar en cuenta lo que se mencionó en el apartado 3.8.1 con respecto a la decisión de la arquitectura física y la distribución.

2. Como se ha mencionado en la "Identificación de los requisitos", Hadoop requiere de muchos recursos de *hardware*, por lo que se recomienda instalarlo en un servidor.

3. Una vez teniendo instalado Hadoop, verificamos que el despliegue este correctamente como se muestra en la figura 9, en la que se observa que todos los servicios están activos. Para esta investigación se utilizó una versión más compacta de Hadoop, que es Sandbox. Hortonworks Sandbox (Hortonworks, 2011-2016) es un entorno personal portátil de Apache Hadoop®, que viene con docenas de tutoriales interactivos de los ecosistemas de Hadoop y de las novedades más interesantes de la última distribución HDP.

Figura 9. Despliegue de Hadoop

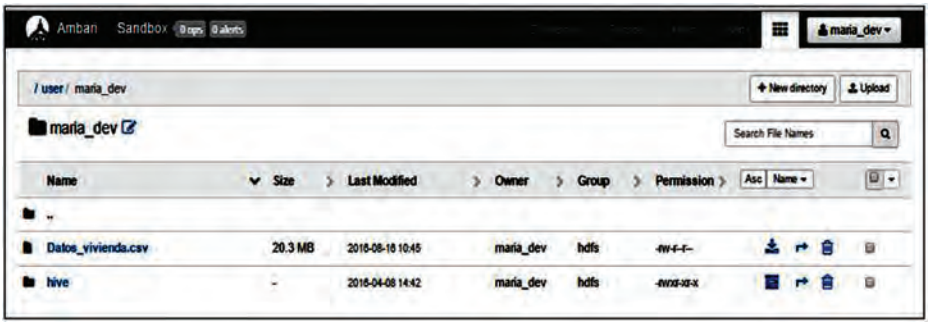


Fuente: elaboración propia

6.3.1 *Trabajando con Hadoop*

Una vez se haya hecho el despliegue correctamente de Hadoop, se procede a cargar el fichero para que esté en formato HDFS y que pueda ser utilizado posteriormente con la herramienta Hive, para las consultas necesarias.

Figura 10. Carga del fichero para que este en formato HDFS



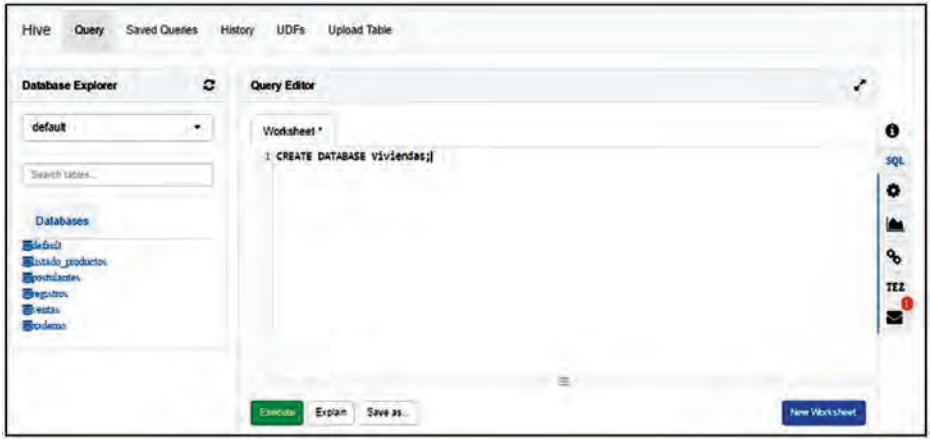
Fuente: elaboración propia.

6.3.2 *Uso de Hive*

1. Se procede a utilizar Hive, la cual es una de las herramientas de Hadoop. Anteriormente se mencionaron las razones del porqué se seleccionó.

2. En la opción Hive View de la ventana de Sandbox, y en el Query Editor, procedemos a crear la base de datos con el comando CREATE DATABASE.

Figura 11. Creación de la base de datos en Hive

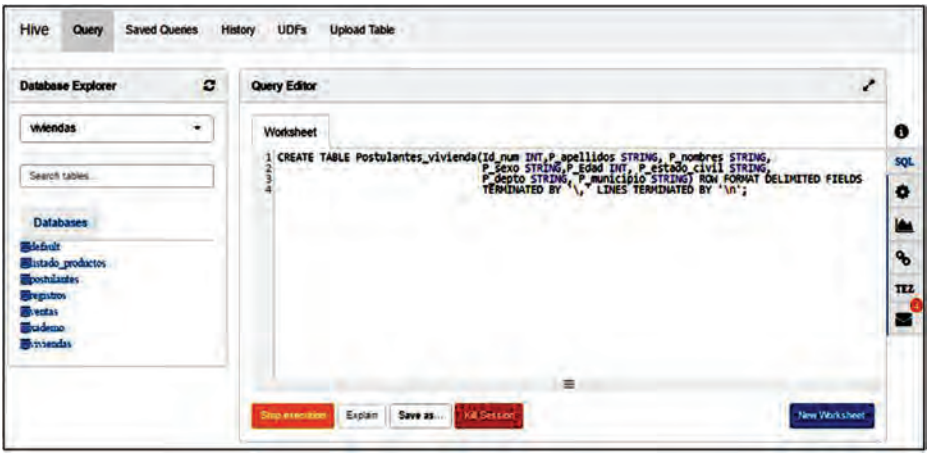


Fuente: elaboración propia.

3. Posteriormente, se crea la tabla con el comando CREATE TABLE. La tabla debe llevar todos los campos que contiene el fichero, con su respectivo tipo de datos.

A continuación, se presenta la imagen donde se aprecia la creación de la tabla.

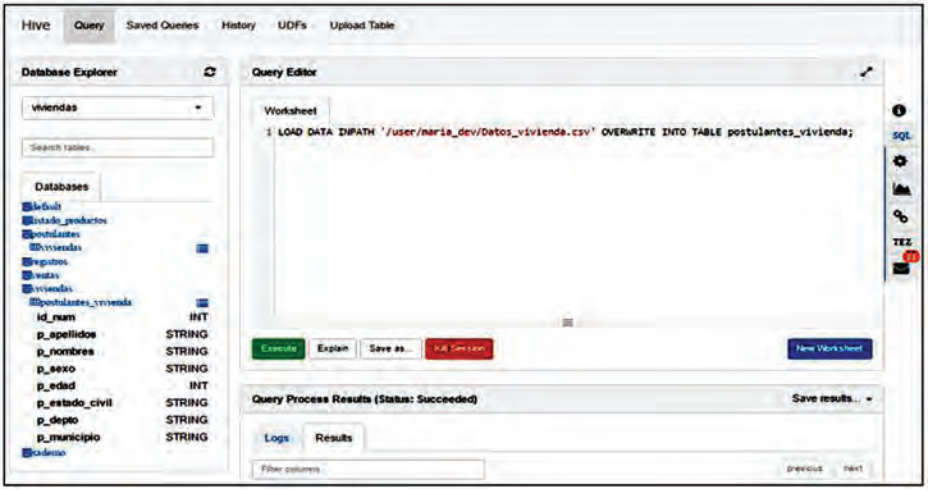
Figura 12. Creando la tabla dentro de la base de datos



Fuente: elaboración propia.

4. Procedemos a cargar el archivo dentro de la tabla creada utilizando el comando LOAD DATA INPATH.

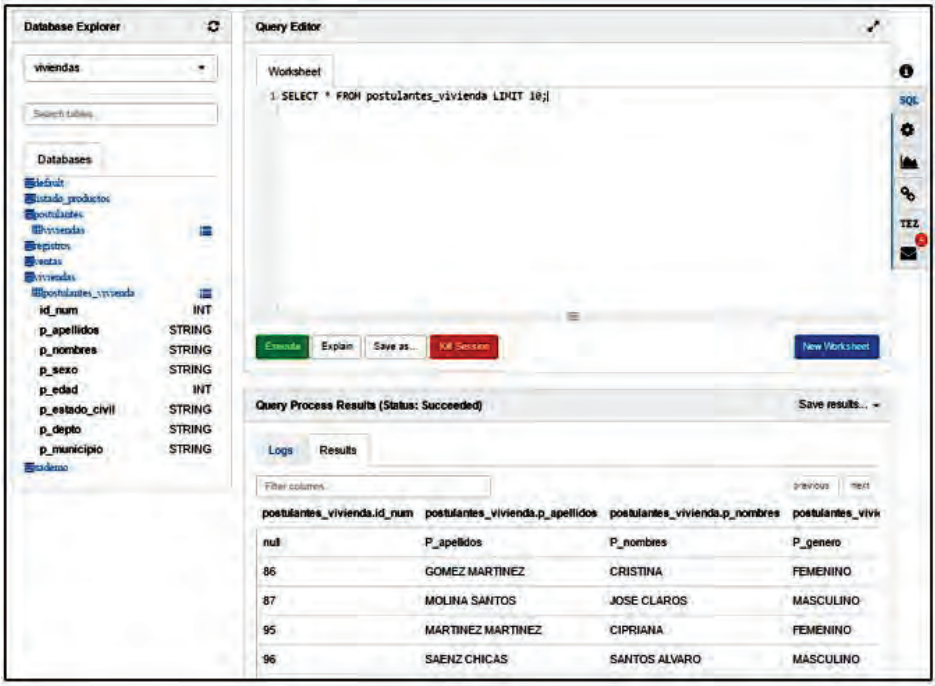
Figura 13. Cargando el archivo en la tabla dentro de la base de datos



Fuente: elaboración propia.

5. Se comprueba que los datos se hayan cargado en la tabla, utilizando el comando SELECT y que muestre al menos diez registros.

Figura 14. Mostrando diez registros de la tabla



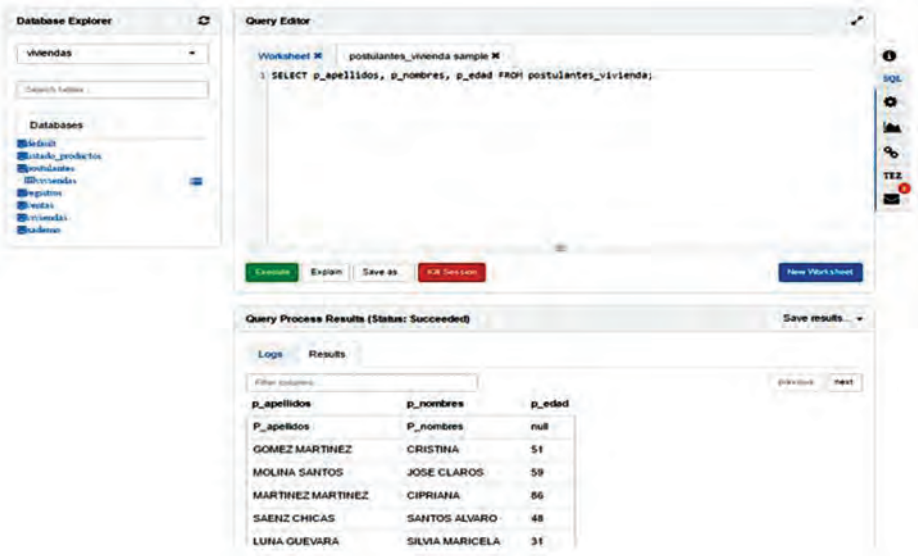
Fuente: elaboración propia.

6. Realizando consultas en Hive dentro de la tabla “Postulantes vivienda”.

   ✓ **Ver los registros por apellidos, nombres y edad.**

SELECT p_apellidos, p_nombres, p_edad FROM postulantes_vivienda.

Figura 15. Listado de los apellidos, nombres y edad de los postulantes a vivienda



Fuente: elaboración propia.

✓ **Mostrar solo los registros del sexo masculino**

Figura 16. Listado de postulantes del género masculino

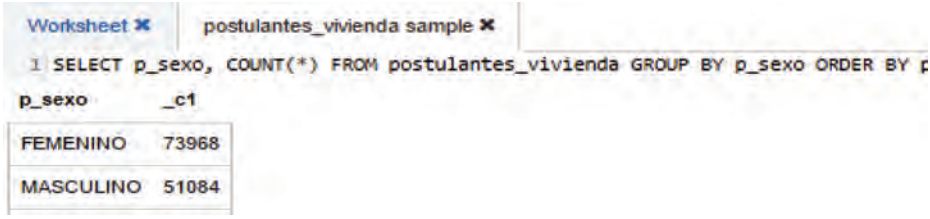
Worksheet ✖ postulantes_vivienda sample ✖

```
1 SELECT * FROM postulantes_vivienda WHERE p_sexo = "MASCULINO";
```

postulantes_vivienda.id_num	postulantes_vivienda.p_apellidos	postulantes_vivienda.p_nombres	postulantes_vivi
87	MOLINA SANTOS	JOSE CLAROS	MASCULINO
96	SAENZ CHICAS	SANTOS ALVARO	MASCULINO
100	MARTINEZ GARCIA	SANTOS DOMINGO	MASCULINO
104	GARCIA BONILLA	JOSE DONAI	MASCULINO
106	MARTINEZ MARTINEZ	SANTOS ROBERTO	MASCULINO
110	DIAZ ARGUETA	TEOFILO	MASCULINO
119	PEREZ PEREZ	SANTOS GREGORIO	MASCULINO
121	GONZALEZ	SANTOS LAZARO	MASCULINO
125	MARTINEZ PEREZ	SANTOS	MASCULINO
129	LUNA ORTIZ	JUAN EVANGELISTA	MASCULINO
131	CLAROS	SANTOS TRANSITO	MASCULINO
132	MARTINEZ MARTINEZ	SANTOS BRUNO	MASCULINO
134	MARTINEZ	MARIA GONZALO	MASCULINO
136	ARGUETA MARTINEZ	JUAN BAUTISTA	MASCULINO
139	BLANCO ARRIAZA	JOSE CATALINO	MASCULINO
143	MARTINEZ PEREZ	ISIDRO	MASCULINO

Fuente: elaboración propia.

✓ **Contar el número de registros masculinos y femeninos**



Worksheet ✖ postulantes_vivienda sample ✖

```
1 SELECT p_sexo, COUNT(*) FROM postulantes_vivienda GROUP BY p_sexo ORDER BY p
```

p_sexo	_c1
FEMENINO	73968
MASCULINO	51084

✓ **Mostrar las personas con estado civil casado(a)**

Figura 17. Listado de postulantes a vivienda cuyo estado civil es casado(a)

Worksheet ✖ postulantes_vivienda sample ✖

```
1 SELECT p_nombres, p_apellidos, p_estado_civil FROM postulantes_vivienda WHERE p_estado_civil="CASADO (A)";
```

p_nombres	p_apellidos	p_estado_civil
SILVIA MARICELA	LUNA GUEVARA	CASADO (A)
MARIA JUANA	MARTINEZ MARTINEZ	CASADO (A)
JOSE DONAI	GARCIA BONILLA	CASADO (A)
SANTOS ROBERTO	MARTINEZ MARTINEZ	CASADO (A)
VICENTA	RAMOS DE DIAZ	CASADO (A)
TEOFILO	DIAZ ARGUETA	CASADO (A)
MARIA VICTOR	MARTINEZ DE PEREZ	CASADO (A)
SANTOS GREGORIO	PEREZ PEREZ	CASADO (A)
SANTOS LAZARO	GONZALEZ	CASADO (A)
SANTOS	MARTINEZ PEREZ	CASADO (A)
JUAN EVANGELISTA	LUNA ORTIZ	CASADO (A)
MARIA EUGENIA	PEREZ DE MARTINEZ	CASADO (A)
ISIDRO	MARTINEZ PEREZ	CASADO (A)
MARIA MARCOS	ORTIZ CACERES	CASADO (A)
SANTOS PABLO	PEREZ HERNANDEZ	CASADO (A)

Fuente: elaboración propia.

Estas son una muestra de las consultas que se pueden realizar con el fichero, pero se pueden hacer más y de cualquier tipo, todo dependerá de las necesidades de cada usuario. Los comandos que se usan son bastante similares y, hasta cierto punto, iguales a los utilizados en SQL; y lo más importante es que los resultados los muestra en cuestión de segundos.

6.4 *Utilización de R*

R es el programa que se va a utilizar para el análisis estadístico de los datos, por lo tanto, hay que seguir algunos pasos para la instalación y luego se procede a trabajar en él.

6.4.1 *Instalación de R*

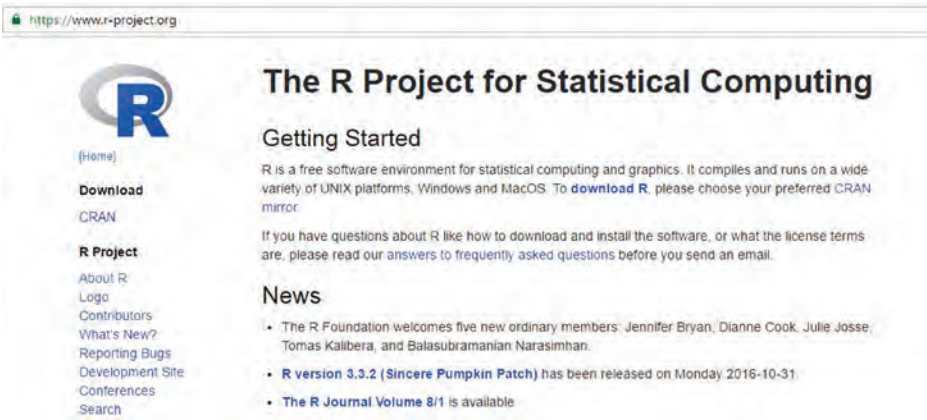
En primer lugar, hay que descargar el instalador específico de la plataforma utilizada en el equipo. Para este caso es Windows, pero hay versiones para Mac OS o Linux.

Los pasos que se deben seguir son los siguientes:

1. Descargar el instalador desde https://www.r-project.org/.

   Aparecerá la siguiente pantalla:

Figura 18. Instalación de R

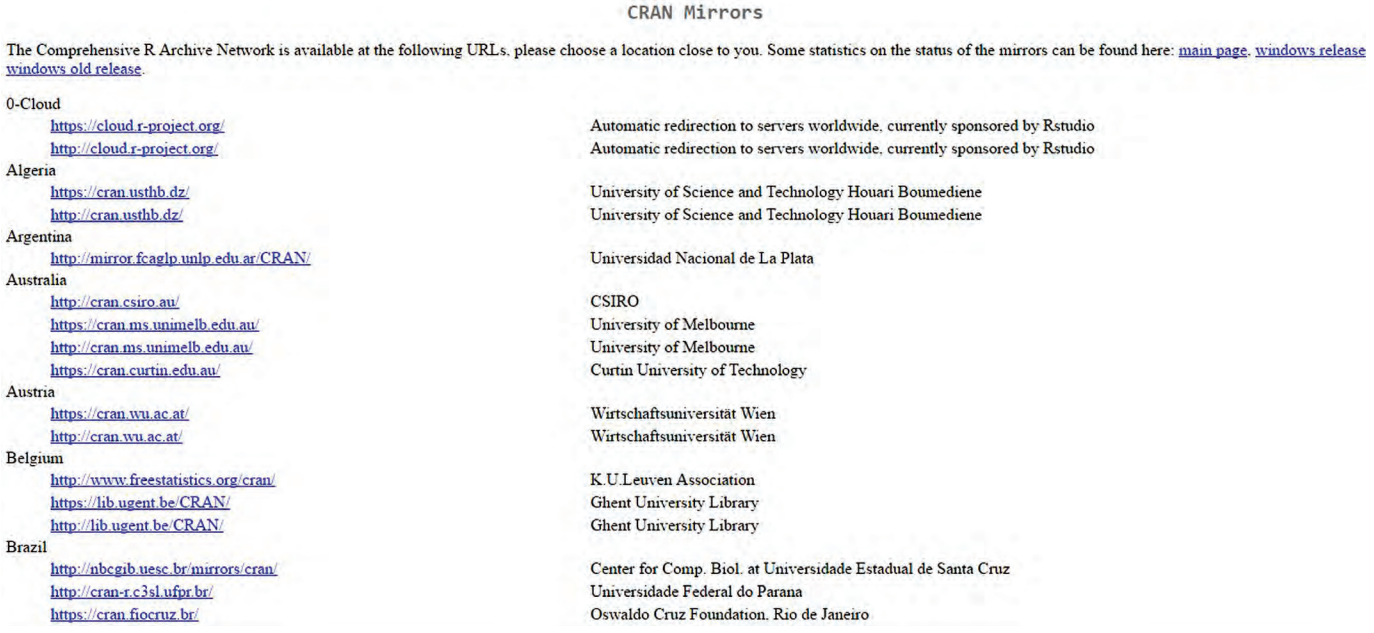


Fuente: elaboración propia.

2. Si se observa al lado izquierdo de la pantalla, aparece un enlace con el texto CRAN, acrónimo de *Comprehensive R Archive Network*, una red de servidores web y FTP distribuidos por todo el mundo que actúan como réplicas (*mirrors*) para la distribución del *software* y la documentación de R. Al dar un clic en ese enlace, observará que aparece una página web conteniendo la lista de servidores de la red CRAN desde los que podrá descargar el *software*.

   La lista está organizada geográficamente para que los usuarios escojan el servidor que se encuentre más próximo a su ubicación, porque es probable que también se encuentre más cerca en internet y, por tanto, los tiempos de descarga sean menores.

Figura 19. Listado de red de servidores CRAN



CRAN Mirrors

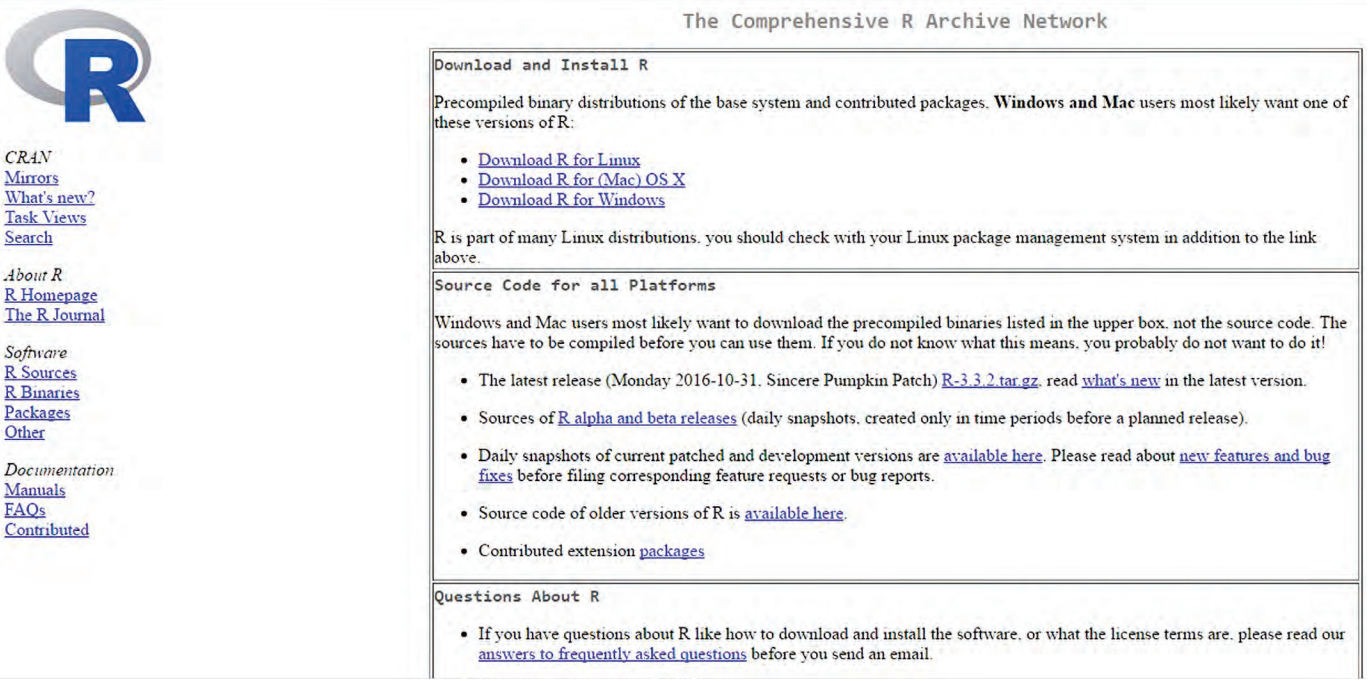
The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release windows old release.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
http://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
http://cran.csiro.au/	CSIRO
https://cran.ms.unimelb.edu.au/	University of Melbourne
http://cran.ms.unimelb.edu.au/	University of Melbourne
https://cran.curtin.edu.au/	Curtin University of Technology
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
http://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association
https://lib.ugent.be/CRAN/	Ghent University Library
http://lib.ugent.be/CRAN/	Ghent University Library
Brazil	
http://nbcgib.uesc.br/mirrors/cran/	Center for Comp. Biol. at Universidade Estadual de Santa Cruz
http://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
https://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro

Fuente: elaboración propia.

Para este caso se selecciona el servidor de El Salvador.

Figura 20. Pantalla para seleccionar el sistema operativo



Fuente: elaboración propia.

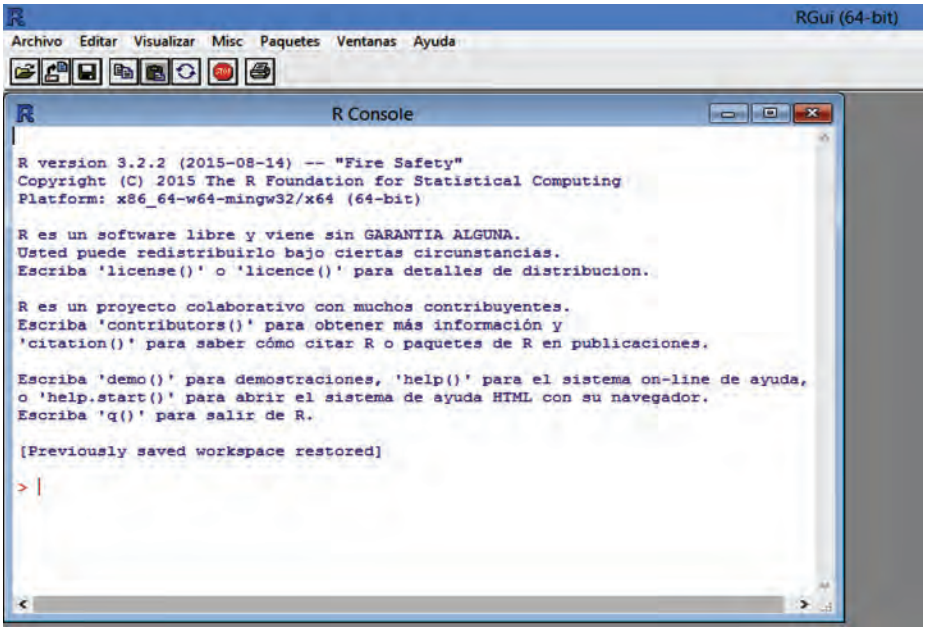
3. Una vez se haya seleccionado el servidor de descarga, verá que aparece una nueva página en cuya parte superior aparecen enlaces específicos para cada sistema operativo. Verá también que desde esta página se tiene acceso al código fuente de la herramienta, en caso de que se quiera realizar la instalación a partir de él.

4. Se selecciona el enlace *Download R para Windows* y lo llevará a otra página en la que le pedirá que seleccione qué es exactamente lo que se quiere descargar, si el sistema base, los paquetes de extensión (*contributed packages*) u otras herramientas. Para este caso, se selecciona el sistema base, por lo que lo llevará a la página donde se encuentra el instalador y procederá a descargarlo y a esperar para poder ejecutarlo.

5. Una vez instalado el programa, aparecerá el acceso directo en el escritorio siempre y cuando lo haya especificado al momento de la instalación.

6. Se procederá a ejecutar el programa y se abrirá una ventana que le ofrece una consola (R console), la cual funciona de manera similar a la línea de comandos del sistema operativo MS-DOS, en la que se digitan los comandos (denominados *expresiones en R*) y se obtendrá mediante ella los resultados (incluyendo mensajes de error si algo va mal).

Figura 21. Interfaz de R (consola)

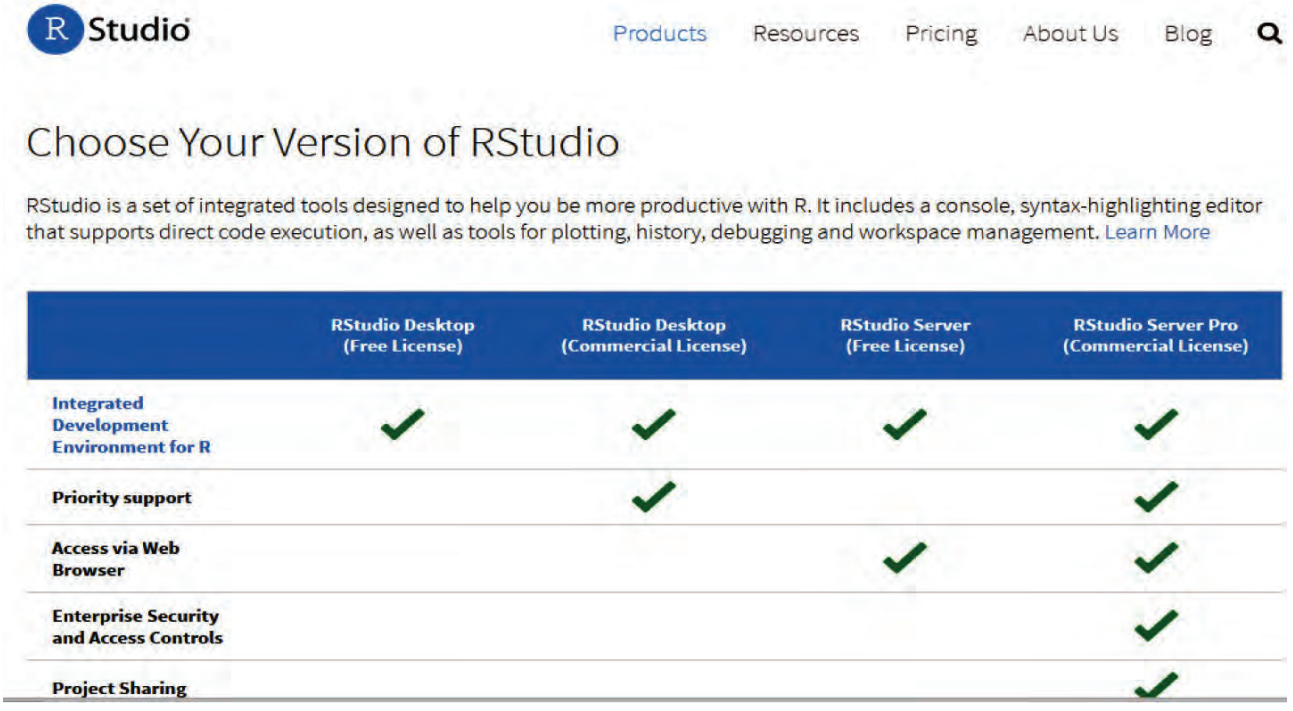


Fuente: elaboración propia.

La otra opción es descargar RStudio, que es una versión con ambiente gráfico, de https://www.rstudio.com/products/rstudio/download/. Ahí aparecen varias opciones, por lo que se deberá seleccionar la que mejor convenga de acuerdo con las necesidades y recursos disponibles de la empresa.

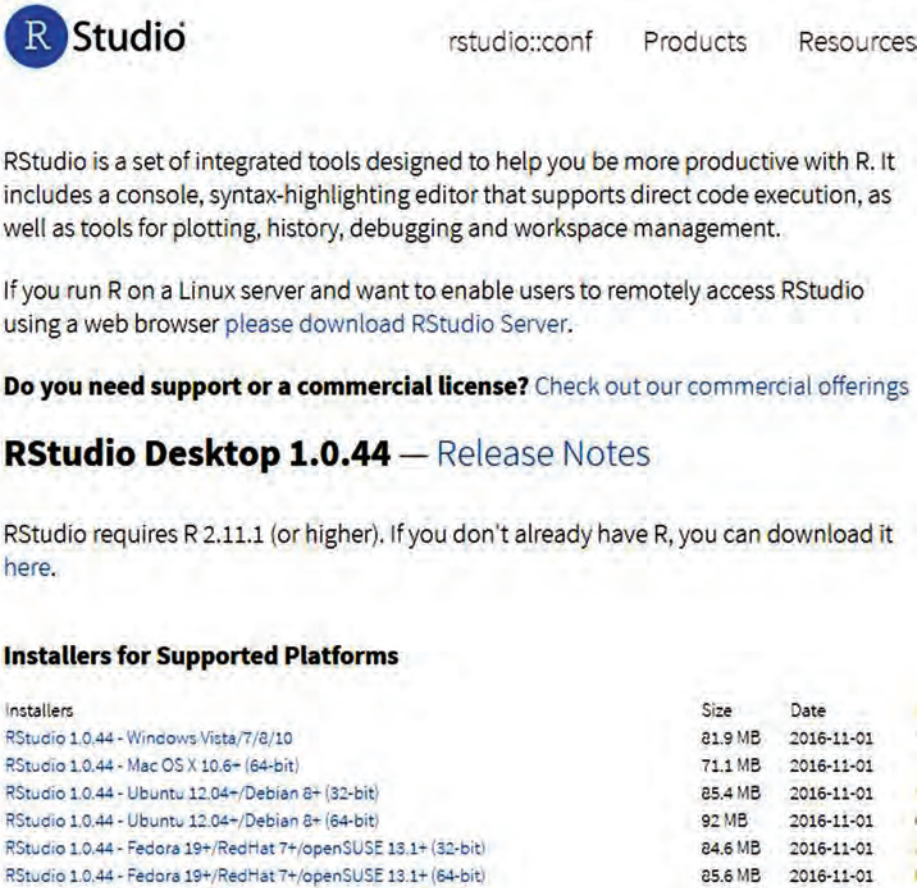
Para nuestra investigación, se seleccionó la versión para *desktop* con licencia libre. Posteriormente se debe seleccionar el sistema operativo que se tenga instalado en el equipo, donde se utilizará RStudio.

Figura 22. Descarga de RStudio



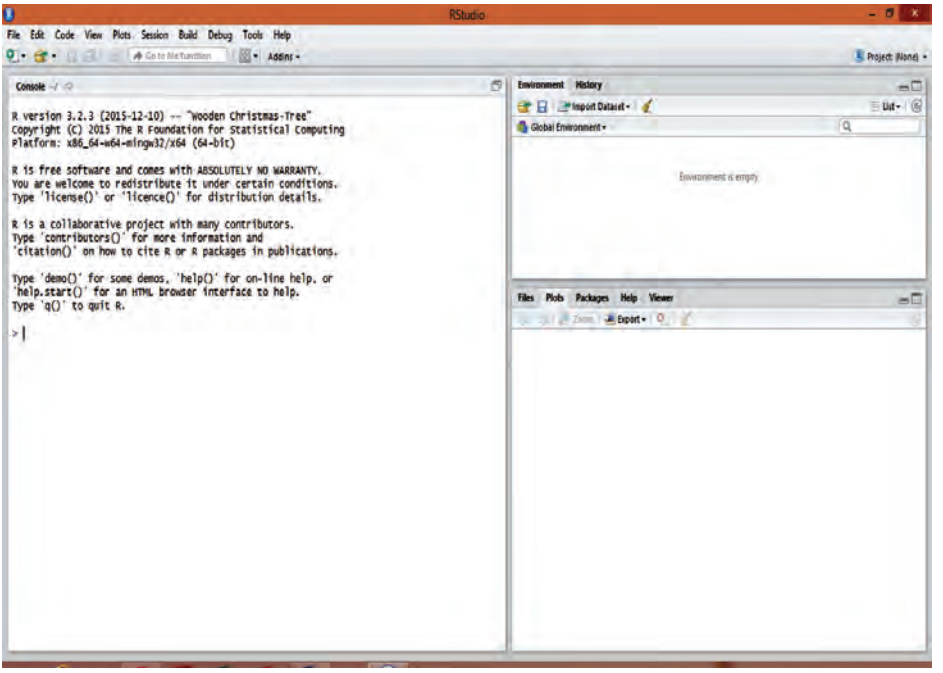
Fuente: https://www.rstudio.com/products/rstudio/download/

Figura 23. Selección del sistema operativo del equipo



Fuente: https://www.rstudio.com/products/rstudio/download/

Figura 24. Interfaz de RStudio



Fuente: elaboración propia.

6.4.2 *Utilización de R para el análisis estadístico de los datos*

Como ya se tiene instalado R, se procede a realizar los análisis estadísticos necesarios con el *dataset* seleccionado. Para este caso se hará uso de un *dataset* que contiene el registro de postulantes a viviendas en el Viceministerio de Vivienda y Desarrollo Urbano.

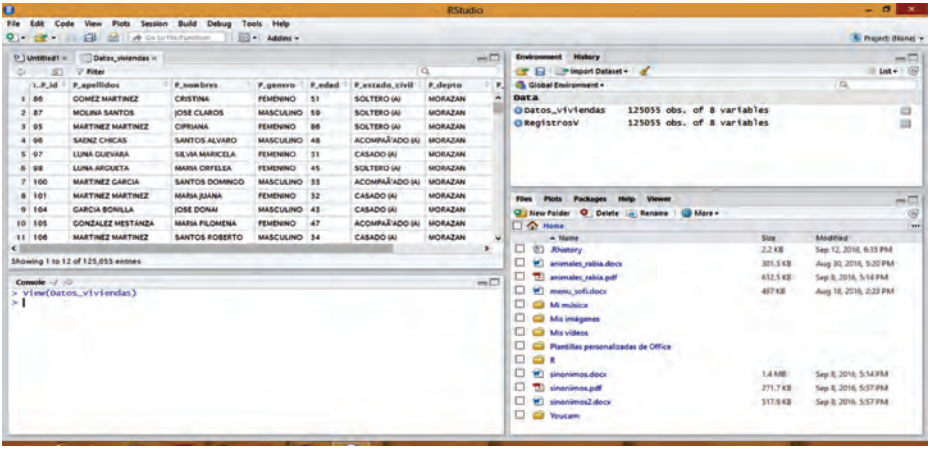
1. Primeramente, se debe importar el *dataset* a la consola de R con el siguiente comando:

```
RegistrosV<-read.csv("C:/Datos_viviendas.csv",sep=",",header=T)
```

En este caso lo que se ha hecho es cargar el *dataset* en una variable a la que se ha nombrado *RegistrosV*. En R el operador de asignación es <- y no el signo =, como es acostumbrado en algunos lenguajes de programación.

2. Al digitar el comando *View(RegistrosV)*, mostrará los datos del *dataset* de la siguiente manera:

Figura 25. Mostrando los datos del *dataset*



Fuente: elaboración propia.

3. Si se necesita saber cuántas filas y columnas tiene el fichero, se hace con el siguiente comando:

```
> dim(RegistrosV)
[1] 125052      8
```

Se puede observar que el fichero tiene 125.052 filas y 8 columnas.

4. Como el *dataset* almacena registros de postulantes a viviendas a escala de todo el país, se quiere saber la cantidad de postulantes por género, estado civil, municipio, etc. Para ello se utilizará la función *Summary*, la cual sirve para producir de las diversas funciones de ajuste del modelo.

Cantidad de postulantes por género

```
> summary(RegistrosV$P_genero)
 FEMENINO MASCULINO
    73968     51084
```

Se puede observar que hay 73.968 del género femenino y 51.084 del masculino.

Cantidades de postulantes por estado civil

En los resultados, se logra apreciar que los que más solicitan vivienda son las personas casadas y acompañadas, sin embargo, hay un buen número de solteros que solicitan una vivienda.

```
> summary(RegistrosV$P_estado_civil)
ACOMPA¥ADO (A)      CASADO (A) DIVORCIADO (A)   SEPARADO (A)     SOLTERO (A)
        34611           50284             862           6120           28968
    VIUDO (A)
         4207
```

Cantidades de postulantes por departamento

```
> summary(RegistrosV$P_depto)
  AHUACHAPAN      CABA¥AS CHALATENANGO     CUSCATLAN     EXTRANJERO  LA LIBERTAD
       15514         3051         6174          3459           124          7322
     LA PAZ     LA UNION      MORAZAN SAN MIGUEL SAN SALVADOR  SAN VICENTE
        7281         2411         7538         11054         14334          5924
  SANTA ANA    SONSONATE     USULUTAN
       14017        17460         9389
```

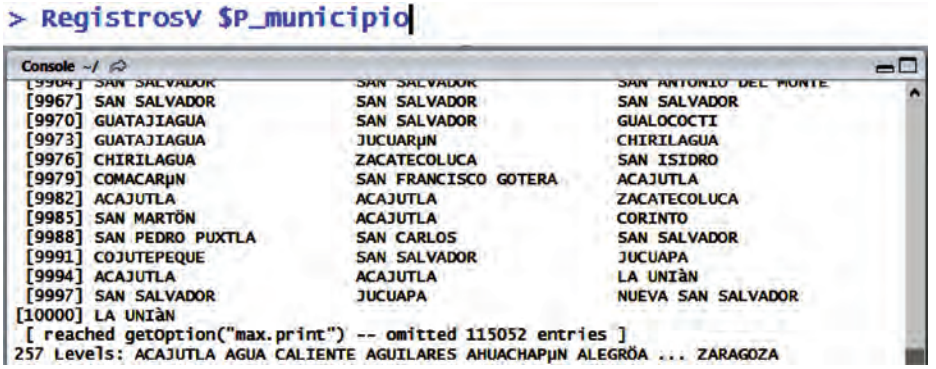
Se necesita conocer la cantidad de municipios y el listado de ellos.

Cantidades de municipios distintos

```
> length(levels(RegistrosV$P_municipio))
[1] 257
```

Listado de municipios solicitados

Figura 26. Listado de municipios con postulantes a vivienda



Fuente: elaboración propia.

6. Generando un *data frame* para mostrar el listado de apellidos, género, estado civil y la edad:

```
> apellidos<-RegistrosV$P_apellidos
> genero<-RegistrosV$P_genero
> estado<-RegistrosV$P_estado_civil
> edad<-RegistrosV$P_edad
> datos<-data.frame(apellidos,genero,estado,edad)
```

Visualizamos el *data frame* con solo escribir el nombre de la variable creada:

Figura 27. Listado de postulantes por apellido, género, estado civil y edad

```
> datos
Console ~/
2487                    ROSALES DE AYALA  FEMENINO DIVORCIADO (A)   57
2488                      RAMOS NAJARRO MASCULINO      VIUDO (A)   36
2489                            AREVALO  FEMENINO     CASADO (A)   50
2490                    TOBAR RODRIGUEZ MASCULINO     CASADO (A)   85
2491                    DELGADO GALEAS   FEMENINO    SOLTERO (A)   73
2492                      JACOBO ROQUE MASCULINO     CASADO (A)   42
2493                    MARTINEZ PUQUIR  FEMENINO ACOMPAÑADO (A)   31
2494                       JACOBO ROSA MASCULINO     CASADO (A)   34
2495                              ALAS   FEMENINO     CASADO (A)   57
2496                INTERIANO HERNANDEZ MASCULINO     CASADO (A)   60
2497                      PEREZ MATA    FEMENINO     CASADO (A)   61
2498                            ZEPEDA  MASCULINO ACOMPAÑADO (A)   52
2499                       VALENSUELA   FEMENINO    SOLTERO (A)   60
2500                    HERRERA CATANEDA MASCULINO    CASADO (A)   45
 [ reached getOption("max.print") -- omitted 122552 rows ]
```

Fuente: elaboración propia.

7. Generando un *data frame* para que nos muestre los apellidos de las personas del género femenino y que sean menores o iguales a 50 años.

Figura 28. Listado de postulantes del género femenino menores o iguales a 50 años

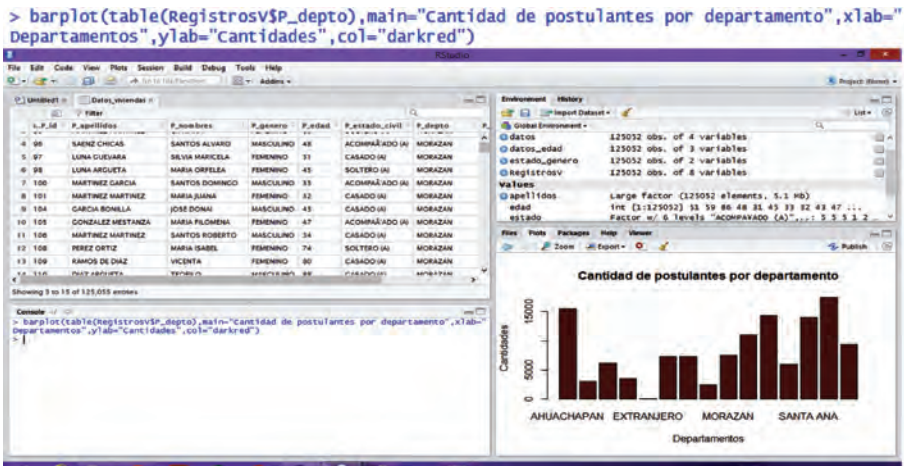
```
> apellidos<-RegistrosV$P_apellidos
> genero<-RegistrosV$P_genero
> edad<-RegistrosV$P_edad
> datos_edad[(genero=="FEMENINO")&(edad <= 50),]
Console ~/
8670                    SALAZAR CARDOZA  FEMENINO   48
8672                  AGUILAR MARTINEZ   FEMENINO   31
8673                        DIAZ LOZA    FEMENINO   31
8675                  ROMERO DE LOPEZ    FEMENINO   33
8676                 HENRIQUEZ GARCIA    FEMENINO   36
8677                  ALARCON MORENO     FEMENINO   48
8679                  MARQUEZ CARDOVA    FEMENINO   27
8681                      RICO ABREGO    FEMENINO   31
8682                  DURAN VAQUERANO    FEMENINO   45
8686                    CRESPO GALEAS    FEMENINO   35
8688                        BENITEZ      FEMENINO   49
8691                   FLORES VASQUEZ    FEMENINO   47
8692                         FLORES      FEMENINO   38
8695                        SANCHEZ      FEMENINO   24
 [ reached getOption("max.print") -- omitted 46637 rows ]
```

Fuente: elaboración propia.

En RStudio también se pueden crear gráficos, quizás no los mejores, que resultan bastante útiles en un momento dado.

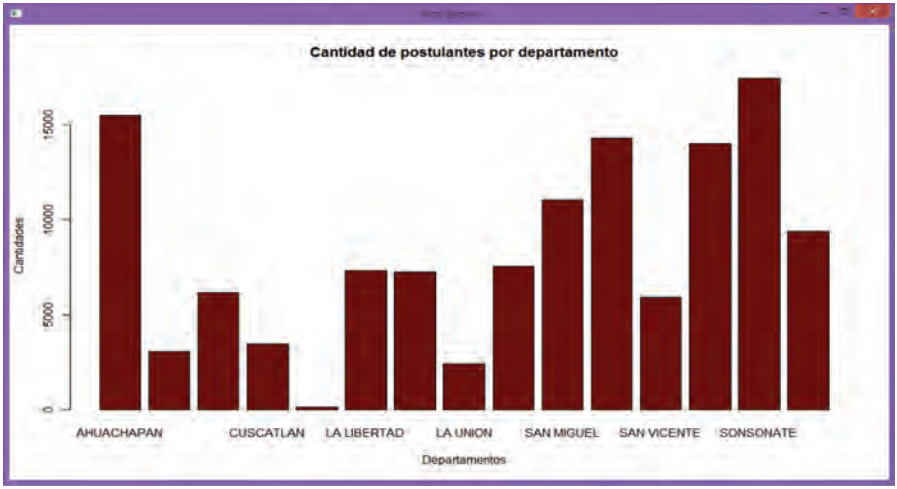
A continuación, aparece un gráfico de barras que muestra las cantidades de postulantes por departamento.

Figura 29. Gráfico de barras que muestra la cantidad de postulantes por departamento



Fuente: elaboración propia.

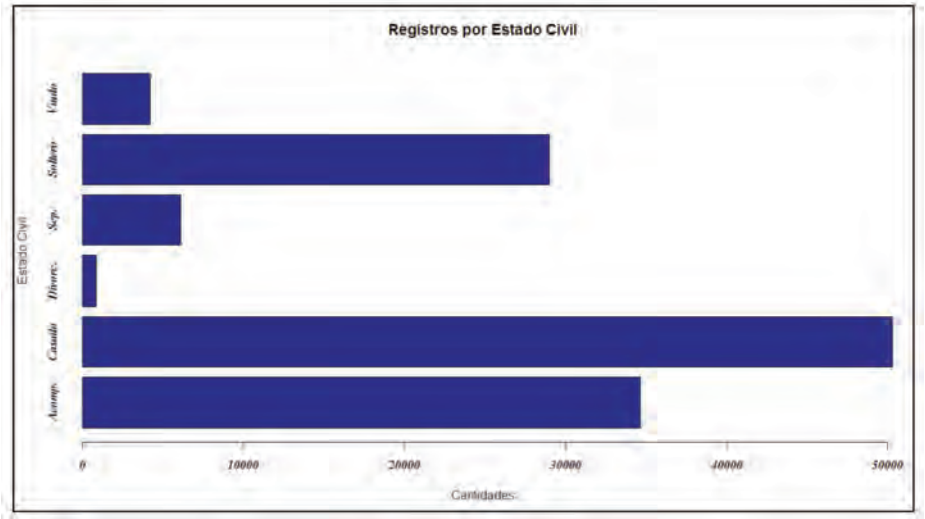
Figura 30. Gráfico de barras por departamento de forma amplia



Fuente: elaboración propia.

8. Gráfico de barras horizontales que muestra los registros por Estado Civil.

Figura 31. Gráfico de barras horizontales que muestra los registros por estado civil



Fuente: elaboración propia.

6.5 *Visualización de la información*

La visualización de la información es muy importante, debido a que mediante los gráficos se pueden entender mejor los resultados obtenidos al hacer uso de cualquier herramienta de procesamiento y análisis estadístico de los datos.

Anteriormente se usó Hive en Hadoop para consultas al estilo SQL, pero también se hicieron análisis estadísticos con RStudio y algunos gráficos, pero nos daremos cuenta que se pueden crear gráficos más vistosos y dinámicos que se pueden utilizar posteriormente en la creación de reportes en una página web. Partiendo de ello, se elaboran gráficos utilizando las herramientas Google Chart y D3.js.

En esta investigación se utilizan dos herramientas para demostrar lo que se puede hacer con cada una de ellas y cómo muestran los datos. Al final, es opcional con cual trabajar; y dependerá de la habilidad que se tenga en la programación con Java Script, HTML y CSS, solo que debe tomarse en cuenta las ventajas de cada una de ellas y de las necesidades que se tenga en un momento dado en la representación de los datos.

Los *dataset* que se han utilizado para esta investigación han sido revisados y depurados con Google Spreadsheets, debido a que contenían columnas con campos vacíos, los cuales no eran útiles en la representación de los datos, por lo tanto, esas filas han sido eliminadas para que la información sea lo más real posible.

Los códigos utilizados para la elaboración de los gráficos se han basado en galerías existentes en la web, debido a que hay ejemplos desarrollados y que pueden ser adecuados a las necesidades de cada usuario. Por ejemplo, para Google Chart existe una diversidad de galerías que pueden utilizarse, estas se encuentran en la siguiente URL: https://google-developers.appspot.com/chart/interactive/docs/gallery

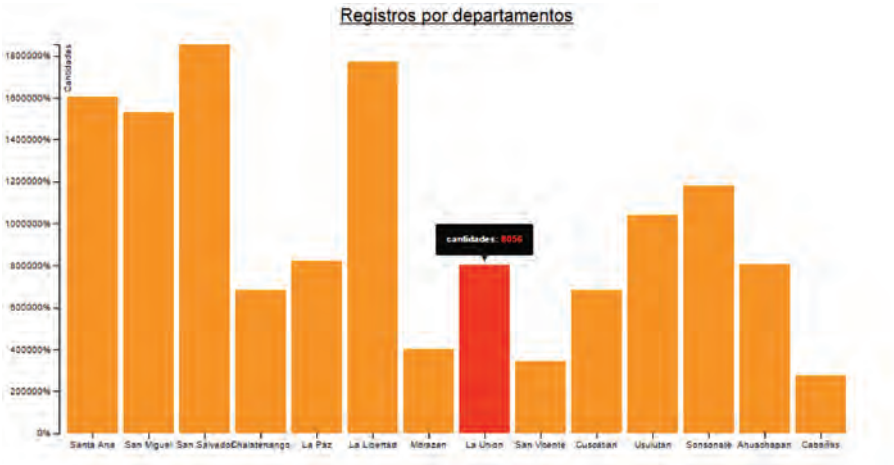
Para hacer uso de D3.js, al igual que en Google Chart, es necesario tener conocimiento de JavaScript. Sin embargo, también existen gallerías y ejemplos que pueden ayudar a crear los gráficos de forma fácil y adaptar los datos como se considere necesario. Estas galerías pueden encontrarse en https://github.com/mbostock/d3/wiki/Gallery.

Todos los gráficos elaborados son dinámicos, por lo que, para visualizar las animaciones, es necesario usar un servidor web, por ejemplo, Xampp. También se puede usar *plunker* y ejecutarlos desde http://plnkr.co/edit/?p=catalogue.

En algunos gráficos se han usado archivos CSV, es decir, se han importado los datos en el código. Este archivo CSV ha sido elaborado con los datos más importantes que se obtuvieron en los análisis realizados con las otras herramientas. Los códigos utilizados podrán verse en los anexos de este documento. A continuación, se presentan las visualizaciones generadas.

Ejemplo 1

Figura 32. Gráfico en D3 que muestra los registros de postulantes a vivienda por departamento



Fuente: elaboración propia.

El gráfico anterior ha sido elaborado con D3.js y muestra de forma dinámica los registros de postulantes por departamento, incluyendo los extranjeros que han solicitado viviendas en El Salvador.

En http://embed.plnkr.co/qtD1UnZieFAkU8hhfOhd se puede ver el gráfico animado, y el código, en el Anexo 2.

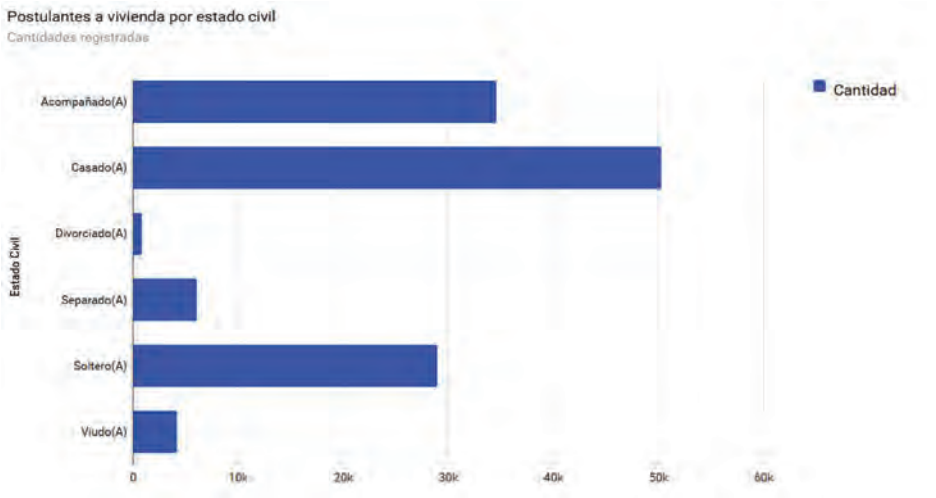
Ejemplo 2

En el siguiente gráfico se muestran los registros de postulantes por estado civil. El cual ha sido elaborado en Google Chart. Si se observa y

se compara con el generado en RStudio, es muy diferente debido a que con esta herramienta el gráfico es dinámico y mucho más presentable.

En https://embed.plnkr.co/5rMPZstLHXqM6bVWPuIN/ se puede ver el gráfico animado, y el código, en el Anexo 3.

Figura 33. Gráfica en Google Chart que muestra el registro de postulantes por estado civil

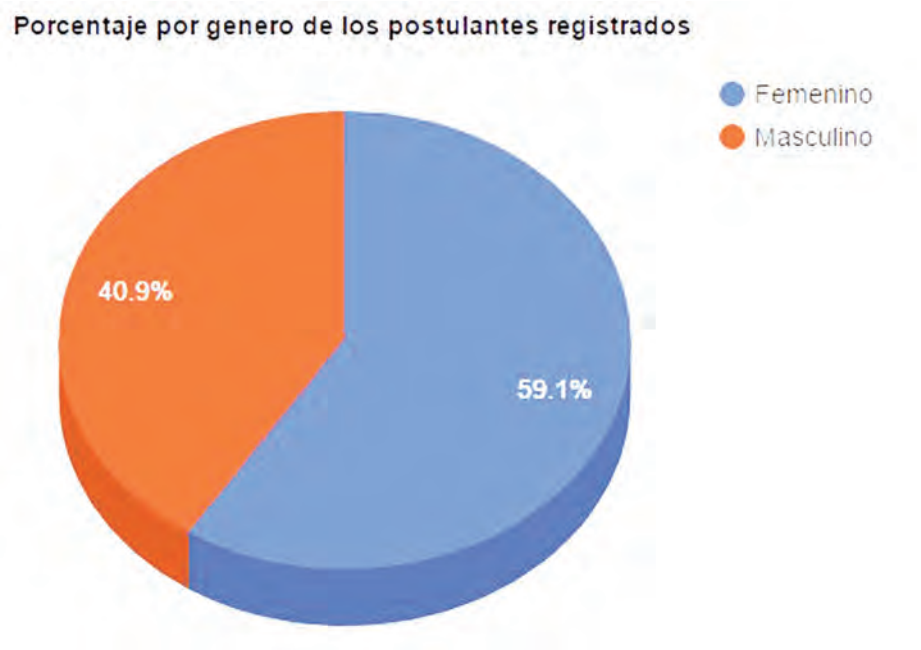


Fuente: elaboración propia.

Ejemplo 3

El siguiente gráfico ha sido elaborado en Google Chart; es un *pie chart* (gráfico de pastel) en 3D, el cual muestra los porcentajes de acuerdo con el género de los postulantes a vivienda registrados.

Figura 34. Gráfica en Google Chart que muestra los porcentajes por género de los postulantes a vivienda



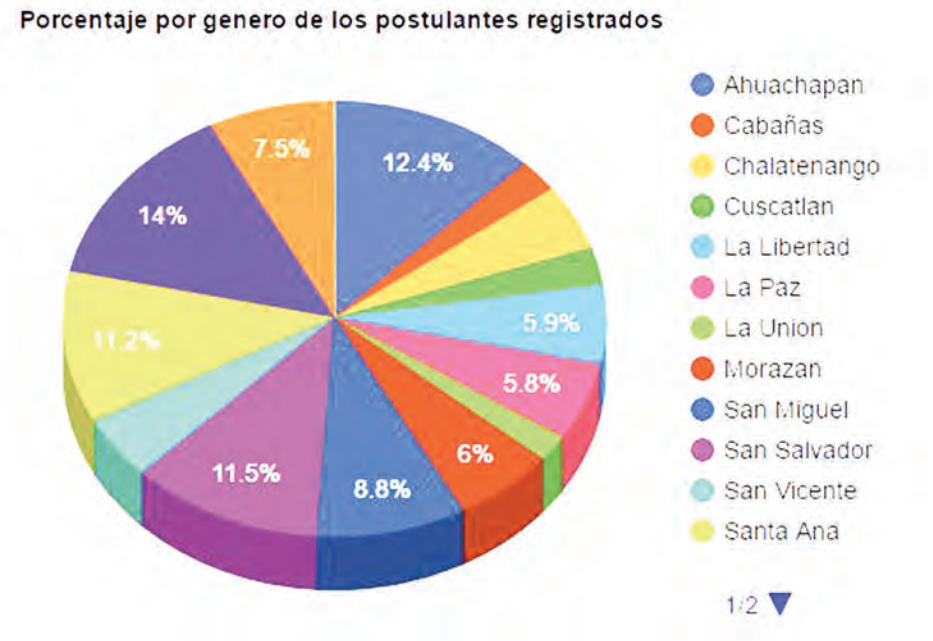
Fuente: elaboración propia.

En en https://embed.plnkr.co/h3zmfM8gnvb2Ro6FOjo5/
se puede ver el gráfico animado, y el código, en el anexo 4.

Ejemplo 4

El siguiente gráfico es un *pie chart* en 3D elaborado en Google Chart. Al compararlo con el gráfico generado en RStudio, se puede ver la diferencia, aparte de que es dinámico, la presentación de los datos lo hace más ordenado.

Figura 35. Gráfica en Google Chart que muestra los porcentajes por departamento de los postulantes registrados



Fuente: elaboración propia.

En https://embed.plnkr.co/10d4sDhNZs70Nw26IIKH/ se puede ver el gráfico animado, y el código, en el anexo 5.

7. Conclusiones

- El uso masivo de información genera que el procesamiento y análisis de los datos se realice de manera lenta, lo cual retrasa la toma de decisiones, sobre todo en una institución de gobierno que tiene una gran demanda en la adquisición de viviendas, por lo tanto, es necesario utilizar otras herramientas que permitan trabajar de manera óptima los datos y así agilizar los trámites.

- Debido a la problemática existente en el Viceministerio de Vivienda y Desarrollo Urbano con respecto al manejo de los datos, se procedió a implementar herramientas *big data* para el procesamiento, análisis y visualización de los datos.

- Al hacer uso de Hadoop, el Viceministerio pudo almacenar grandes volúmenes de información y los resultados de su procesamiento fue en menor tiempo; se realizaron consultas puntuales de los habitantes con la herramienta Hive, las cuales son determinantes para la toma de decisiones pertinentes.

- Con RStudio se realizaron análisis estadísticos, tales como la cantidad de postulantes por género, por departamento y municipio, porcentajes de personas por estado civil y por edades, etc. Además de algunos gráficos representativos de los resultados.

- Se demostró que se pueden crear gráficos dinámicos para poder utilizarlos en la presentación de reportes en una página web, lo cual permitirá tomar decisiones a corto plazo.

8. Recomendaciones

- Debido al manejo de grandes volúmenes de información, en el Viceministerio de Vivienda y Desarrollo Urbano, es recomendable que se vaya pensando en cambiar la forma de trabajar y no quedarse con las bases de datos relacionales, porque dentro de poco tiempo serán insuficientes para el procesamiento de los datos.

- La propuesta para el Viceministerio es que adquieran las herramientas necesarias de *big data,* que se les aplicaron a los dataset proporcionados, y que comprueben su eficiencia al usarlas para dar respuestas inmediatas, no importando la cantidad de datos que se tengan; y que no es necesario que los datos sean estructurados.

- Adquirir equipo necesario y licencias de *software* (cuando no sean gratuitas), para la implementación de las herramientas *big data*.

- Buscar capacitaciones, para el personal de Informática, en el uso de herramientas *big data,* para que puedan darle continuidad a las aplicaciones que se les demostraron con esta investigación.

- Desarrollar procesos de interoperabilidad para aprovechar los recursos de datos actuales y generar información predictiva para la toma de decisiones estratégicas.

9. Referencias

Aguilar, L. J. (2013). *Big Data, Analisis de los grandes volumenes de datos.* Mexico: Alfaomega.

Alten (Julio de 2014). *Alten.es*. Obtenido de http://www.alten.es/wp-content/uploads/2014/07/CONCLUSIONES-PROYECTO-BIG-DATA.pdf

Barranco Fragoso, R. (19 de marzo de 2014). *Evaluando Software.com*. Obtenido de http://www.evaluandosoftware.com/nota-3684-Que-es-el-Big-Data.html

Bernardo, A. (8 de mayo de 2013). *Think Big*. Obtenido de http://blogthinkbig.com/big-data-cancer/

Capriolo, E.; Wampler, D., & Rulberglan, J. (2012). Programming Hive. En E. Capriolo; D. Wampler, & J. Rulberglan. Estados Unidos: O'Reilly.

Carrillo Ruiz, J. A.; Marco de Lucas, J. E.; Dueñas López, J. C.; Cases Vega, F., Fernández, J. C.; Pereda Laredo, L. F., & González Muñoz de Morales, G. (Marzo de 2013). *http://www.ieee.es/*. Obtenido de http://www.ieee.es/Galerias/fichero/docs_investig/DIEEEINV03-2013_Big_Data_Entornos_DefensaSeguridad_CarrilloRuiz.pdf

CentralAmericaData (5 de octubre de 2016). *CentralAmericaData.com*. Obtenido de http://www.centralamericadata.com/es/article/home/Ya_no_se_toman_decisiones_sin_antes_analizar_la_Big_Data

Charts, G. (3 de abril de 2012). *Google Charts*. Obtenido de https://developers.google.com/chart

Cloudera, i. (2016). *Cloudera*. Obtenido de http://www.cloudera.com/

D3js.org (2015). *Data-Driven Documents*. Obtenido de https://d3js.org/

De Juana, R. (24 de febrero de 2015). *MuyComputerPro*. Obtenido de http://www.muycomputerpro.com/2015/02/24/ejemplos-reales-uso-inteligente-big-data

Dean, J. a. (2004). *MapReduce: Simplified Data Processing on Large*. Obtenido de http://static.googleusercontent.com/media/research.google.com/es//archive/mapred

Dean, J. a. (2004). *MapReduce: Simplified Data Processing on Large Clusters.*

FondosFidelity (2012). Obtenido de https://www.fondosfidelity.es/static/pdfs/informes-fondos/Fidelity_ArgInvSXXI_BigData_Sept12_ES.pdf

Foundation, T. R. (s/f). *R-project.org*. Obtenido de https://www.r-project.org/about.html

Ghemawat, S. G. (2003). *The Google File System.*

Ghemawat, S.; Gobioff, H., & Leung, S.T. (Octubre de 2003). *The Google File System. In Proceedings of the 19th ACM Symposium on Operating Systems Principles.* Obtenido de http://static.googleusercontent.com/media/research.google.com/es//archive/gfs-sosp2003.pdf

Hortonworks (2011-2016). *Hortonworks*. Obtenido de http://hortonworks.com/products/sandbox/

IBM (2014). *IBM Big Data & Analytics Hub*. Obtenido de http://www.ibmbigdatahub.com/infographic/flood-big-data

IBM developerWorks (13 de noviembre de 2014). Obtenido de https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Jiménez, C. M. (Noviembre-Diciembre de 2014). Revista *Anales*. Obtenido de http://www.revista-anales.es/web/n_29/pdf/10-16.pdf

Joyanes Aguilar, L. (2013). *Big Data: Análisis de grandes volúmenes de datos en las organizaciones.* Mexico: AlfaOmega.

Laney, D. (2016). *Gartner*. Obtenido de http://www.gartner.com/analyst/40872/Douglas-Laney

MapR Technologies, i. (2016). *MapR*. Obtenido de http://www.mapr.com/

Mined, (1994). Historia de El Salvador. San Salvador, El Salvador.

Mitchell I, L. M. (2012). *The White Book of Big Data. The definitive guide to the revolution in business analytics.* Obtenido de Fujitsu.com. http://www.fujitsu.com/global/Images/WhiteBookofBigData.pdf

N-economia (Noviembre de 2015). *N economia*. Obtenido de http://n-economia.com/notasalerta/trasformacion-digital-big-data-infografia/

Oficial, D. (8 de julio de 1905). No. 159. *Diario Oficial.* Tomo 59.

Oficial, D. (31 de agosto de 1916). Decreto No. 198. *Diario Oficial*. Tomo 31.

Oficial, D. (18 de diciembre de 1917). Decreto No.278 y No. 279. *Diario Oficial.* Tomo No. 143.

Pérez Arbesú, L. (2016). *TechTarget*. Obtenido de http://searchdatacenter.techtarget.com/es/cronica/Big-Data-en-America-Latina-avanza-a-pasos-pequenos

Quiros, R. (14 de julio de 2015). *La Estrella de Panamá*. Obtenido de http://laestrella.com.pa/economia/impacto-data-pymes/23878908

revistaitnow (Octubre de 2016). *revistaitnow.com*. Obtenido de https://revistaitnow.com/las-vias-los-gobiernos-lleguen-al-big-data/

Scherer, M. (9 de noviembre de 2012). *DataPrix*. Obtenido de http://www.dataprix.com/noticias-it/tendencias-tecnologicas/big-data/big-data-ayudaron-obama-ganar-las-elecciones

Souto, S. (4 de mayo de 2015). *Hechos de Hoy*. Obtenido de http://www.hechosdehoy.com/big-data-pilar-fundamental-para-el-desarrollo-de-las-ciudades-inteligentes-43285.htm

White, T. (2012). Hadoop: The Definitive Guide. En T. White. Estados Unidos: O'Reilly.

White, T. (Mayo 2012). Hadoop: The definitive guide. En T. White, *Hadoop: The definitive guide* (págs. 12-13). Estados Unidos: O'Reilly Media, Inc.

Zicari, R. (2014). *Big Data: Challenges and Opportunities.* Costa Rica: CRC Press.

Zikopoulos, P. (2015). *BeSmart*. Obtenido de http://www.besmart.company/blog/big-data-transforma-ciudades/

10. Anexos

Anexo 1

Carta del Viceministerio de Vivienda y Desarrollo Urbano

EL SALVADOR
UNIDOS CRECEMOS TODOS

Página 1 de 1

MOPTVDU - GII - 2016

San Salvador, 31 de Mayo de 2016

Señores
Universidad Tecnológica de El Salvador
Presentes.-

Estimados Señores:

Es un gusto saludarles y desearles éxitos en sus labores cotidianas en su aporte académico al crecimiento del país.

Es para el Ministerio de Obras Públicas, Transporte y de Vivienda y Desarrollo Urbano, muy importante lograr establecer la relación Universidad - Gobierno por medio de la investigación y nos sentimos agradecidos por tomarnos en cuenta para desarrollar su estudio de investigación experimental en el Viceministerio de Vivienda y Desarrollo Urbano, aplicado a las nuevas herramientas de Big Data, esperando que los resultados de dicha investigación sean de beneficio para ambas instituciones.

Les deseamos éxitos en su investigación y en su labor académica.

Cordialmente,

MINISTERIO DE OBRAS PÚBLICAS GERENCIA DE INFORMATICA INSTITUCIONAL

Ing. Carlos Alberto Trigueros Ayala
Gerente de Informática Institucional – Ad Honorem

MOP

MINISTERIO DE OBRAS PÚBLICAS, TRANSPORTE Y DE VIVIENDA Y DESARROLLO URBANO
ALAMEDA MANUEL ENRIQUE ARAUJO, KM. 5 ½ CARRETERA A SANTA TECLA CONTIGUO A MERCADO DE ARTESANIAS

Anexo 2

Código en D3.js para la elaboración del gráfico de barras, que muestra los registros de postulantes por departamento

```
<!DOCTYPE html>
<meta charset="utf-8">
<style>

body {
  font: 10px sans-serif;
}

.axis path,
.axis line {
  fill: none;
  stroke: #000;
  shape-rendering: crispEdges;
}

.bar {
  fill: orange;
}

.bar:hover {
  fill: orangered ;
}

.x.axis path {
  display: none;
}

.d3-tip {
  line-height: 1;
  font-weight: bold;
  padding: 12px;
  background: rgba(0, 0, 0, 0.8);
  color: #fff;
  border-radius: 2px;
}
```

```
36
37  /* Creates a small triangle extender for the tooltip */
38  .d3-tip:after {
39    box-sizing: border-box;
40    display: inline;
41    font-size: 10px;
42    width: 100%;
43    line-height: 1;
44    color: rgba(0, 0, 0, 0.8);
45    content: "\25BC";
46    position: absolute;
47    text-align: center;
48  }
49
50  /* Style northward tooltips differently */
51  .d3-tip.n:after {
52    margin: -1px 0 0 0;
53    top: 100%;
54    left: 0;
55  }
56  </style>
57  <body>
58  <script src="http://d3js.org/d3.v3.min.js"></script>
59  <script src="http://labratrevenge.com/d3-tip/javascripts/d3.tip.v0.6.3.js"></script>
60  <script>
61
62  var margin = {top: 40, right: 20, bottom: 30, left: 60},
63      width = 960 - margin.left - margin.right,
64      height = 500 - margin.top - margin.bottom;
65
66  var formatPercent = d3.format(".0");
67
68  var x = d3.scale.ordinal()
69      .rangeRoundBands([0, width], .1);
70
71  var y = d3.scale.linear()
72      .range([height, 0]);
73
74  var xAxis = d3.svg.axis()
75      .scale(x)
76      .orient("bottom");
77
78  var yAxis = d3.svg.axis()
79      .scale(y)
80      .orient("left")
81      .tickFormat(formatPercent);
82
83  var tip = d3.tip()
84    .attr('class', 'd3-tip')
85    .offset([-10, 0])
86    .html(function(d) {
87      return "<strong>Postulantes:</strong> <span style='color:red'>" + d.cantidades + "</span>";
88    })
89
90  var svg = d3.select("body").append("svg")
91      .attr("width", width + margin.left + margin.right)
92      .attr("height", height + margin.top + margin.bottom)
93    .append("g")
94      .attr("transform", "translate(" + margin.left + "," + margin.top + ")");
95
96  svg.call(tip);
97
98  d3.csv("data.csv", type, function(error, data) {
99    x.domain(data.map(function(d) { return d.departamento; }));
100   y.domain([0, d3.max(data, function(d) { return d.cantidades; })]);
101
102   svg.append("g")
103       .attr("class", "x axis")
104       .attr("transform", "translate(0," + height + ")")
105       .call(xAxis);
```

```
106
107     svg.append("g")
108         .attr("class", "y axis")
109         .call(yAxis)
110       .append("text")
111         .attr("transform", "rotate(-90)")
112         .attr("y", 6)
113         .attr("dy", ".71em")
114         .style("text-anchor", "end");
115
116
117     svg.append("text")
118         .attr("class","title")
119         .attr("x",(width/2))
120         .attr("y", -25)
121         .attr("text-anchor","middle")
122         .style("font-size","20px")
123         .style("text-decoration","underline")
124         .text("Registros por departamentos");
125
126     svg.selectAll(".bar")
127         .data(data)
128       .enter().append("rect")
129         .attr("class", "bar")
130         .attr("x", function(d) { return x(d.departamento); })
131         .attr("width", x.rangeBand())
132         .attr("y", function(d) { return y(d.cantidades); })
133         .attr("height", function(d) { return height - y(d.cantidades); })
134         .on('mouseover', tip.show)
135         .on('mouseout', tip.hide)
136
137 });
138
139 function type(d) {
140     d.cantidades = +d.cantidades;
141     return d;
142 }
143
144 </script>
```

Archivo csv utilizado (data.csv)

data.csv

```
1   departamento,cantidades
2   Ahuachapan,15514
3   Cabañas,3051
4   Chalatenango,6174
5   Cuscatlan,3459
6   Extranjero,124
7   La Libertad,7322
8   La Paz,7281
9   La Union,2411
10  Morazan,7538
11  San Miguel,11054
12  San Salvador,14334
13  San Vicente,5924
14  Santa Ana,14017
15  Sonsonate,17460
16  Usulutan,9389
```

Anexo 3

Código en Google Chart para la elaboración del gráfico de barras horizontales, que muestra el registro de postulantes por estado civil

```
<html>
  <head>
    <script type="text/javascript" src="https://www.google.com/jsapi"></script>
    <script type="text/javascript">
      google.load("visualization", "1.1", {packages:["bar"]});
      google.setOnLoadCallback(drawChart);
      function drawChart() {
        var data = google.visualization.arrayToDataTable([
       ['Estado Civil', 'Cantidad', { role: 'style' } ],
       ['Acompañado(A)', 34611 , 'color: gray'],
       ['Casado(A)', 50284, 'color: #76A7FA'],
       ['Divorciado(A) ',862, 'opacity: 0.2'],
       ['Separado(A)', 6120, 'stroke-color: #703593; stroke-width: 4; fill-color: #C5A5CF'],
       ['Soltero(A) ', 28968, 'stroke-color: #871B47; stroke-opacity: 0.6; stroke-width: 8; fill-color: #BC5679; fill-opacity: 0.2'],
       ['Viudo(A)', 4207, 'color: gray'],
       ]);

       var options = {
         chart: {
           title: 'Postulantes a vivienda por estado civil',
           subtitle: 'Cantidades registradas',
         },
         bars: 'horizontal' // Required for Material Bar Charts.
       };

       var chart = new google.charts.Bar(document.getElementById('barchart_material'));

       chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="barchart_material" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

Anexo 4

Gráfico en Google Chart para la elaboración del gráfico *pie chart* 3D, que muestra los porcentajes de los postulantes por género

```
<html>
  <head>
    <script type="text/javascript" src="https://www.google.com/jsapi"></script>
    <script type="text/javascript">
      google.load("visualization", "1", {packages:["corechart"]});
      google.setOnLoadCallback(drawChart);
      function drawChart() {
        var data = google.visualization.arrayToDataTable([
          ['Genero', 'Cantidades'],
          ['Femenino',  73968],
          ['Masculino',  51084]
        ]);

        var options = {
          title: 'Porcentaje por genero de los postulantes registrados',
          is3D: true,
        };

        var chart = new google.visualization.PieChart(document.getElementById('piechart_3d'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="piechart_3d" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

Anexo 5

Código en Google Chart, para la elaboración del gráfico *pie chart* 3D, que muestra los porcentajes de los postulantes por departamento

```
<html>
  <head>
    <script type="text/javascript" src="https://www.google.com/jsapi"></script>
    <script type="text/javascript">
      google.load("visualization", "1", {packages:["corechart"]});
      google.setOnLoadCallback(drawChart);
      function drawChart() {
        var data = google.visualization.arrayToDataTable([
          ['Departamento', 'Cantidades'],
          ['Ahuachapan',15514],
          ['Cabañas',3051],
          ['Chalatenango',6174],
          ['Cuscatlan',3459],
          ['Extranjero',124],
          ['La Libertad',7322],
          ['La Paz',7281],
          ['La Union',2411],
          ['Morazan',7538],
          ['San Miguel',11054],
          ['San Salvador',14334],
          ['San Vicente',5924],
          ['Santa Ana',14017],
          ['Sonsonate',17460],
          ['Usulutan',9389],
          ]);

        var options = {
          title: 'Porcentaje por genero de los postulantes registrados',
          is3D: true,
        };

        var chart = new google.visualization.PieChart(document.getElementById('piechart_3d'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="piechart_3d" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

Breve hoja de vida de los investigadores

Verónica Idalia Rosa. Docente investigadora Utec. Ingeniera en Sistemas y Computación. Candidata a doctora en Informática de la Universidad de Alicante, España. Máster en Visual Analytics y Big Data de la Universidad de La Rioja, España. Maestría en Docencia Universitaria de la Universidad Tecnológica de El Salvador.

José Guillermo Rivera. Docente investigador Utec. Ingeniero en Sistemas y Computación, con Maestría en Administración de Recursos Humanos y licenciado en Docencia Universitaria de la Universidad Evangélica de El Salvador. Tiene escalafón docente otorgado por la Maestría en Docencia Universitaria, Mined 2009.

Colección Investigaciones 2003-2018

Publicación	Nombre	ISBN
2003	Historia de la Economía de la Provincia del Salvador desde el siglo XVI hasta nuestros días. Primer Tomo Siglo XVI Jorge Barraza Ibarra	99923-21-12-1 (v 1) 99923-21-11-3 (Edición completa)
Diciembre 2003	Recopilaciones Investigativas. Tomos I, II y III	SIN ISBN
2004	Historia de la Economía de la Provincia del Salvador desde el siglo XVI hasta nuestros días. Segundo Tomo Siglos XVII y XVIII Jorge Barraza Ibarra	99923-21-14-8 (v 2) 99923-21-11-3 (Edición completa)
2004	Historia de la Economía de la Provincia del Salvador desde el siglo XVI hasta nuestros días. Tercer Tomo Siglo XIX Jorge Barraza Ibarra	99923-21-15-6 (v 3) 99923-21-11-3 (Edición completa)
2005	Historia de la Economía de la Provincia del Salvador desde el siglo XVI hasta nuestros días. Cuarto Tomo Siglo XIX Jorge Barraza Ibarra	99923-21-31-8 99923-21-11-3 (Edición completa)
2006	Historia de la Economía de la Provincia del Salvador desde el siglo XVI hasta nuestros días. Quinto Tomo Siglo XX Jorge Barraza Ibarra	99923-21-39-3 (v 5) 99923-21-11-3 (Edición completa)
2009	Recopilación Investigativa. Tomo I	978-99923-21-50-8 (v1)
2009	Recopilación Investigativa. Tomo II	978-99923-21-51-5 (v2)
2009	Recopilación Investigativa. Tomo III	978-99923-21-52-2 (v3)
Enero 2010	Casa Blanca Chalchuapa, El Salvador. Excavación en la trinchera 4N. Nobuyuki Ito	978-99923-21-58-4
Marzo 2010	Recopilación Investigativa 2009. Tomo 1	978-99922-21-59-1 (v.1)
Marzo 2010	Recopilación Investigativa 2009. Tomo 2	978-99922-21-60-7 (v.2)
Marzo 2010	Recopilación Investigativa 2009. Tomo 3	978-99922-21-61-7 (v.3)
Octubre 2010	Obstáculos para una investigación social orientada al desarrollo. Colección Investigaciones José Padrón Guillen	978-99923-21-62-1
Febrero 2011	Estructura familia y conducta antisocial de los estudiantes en Educación Media. Colección Investigaciones n.º 2 Luis Fernando Orantes Salazar	

Febrero 2011	Prevalencia de alteraciones afectivas: depresión y ansiedad en la población salvadoreña. Colección Investigaciones n.º 3 José Ricardo Gutiérrez Ana Sandra Aguilar de Mendoza	
Marzo 2011	Violación de derechos ante la discriminación de género. Enfoque social. Colección Investigaciones n.º 4 Elsa Ramos	
Marzo 2011	Recopilación Investigativa 2010. Tomo I	978-99923-21-65-2 (v1)
Marzo 2011	Recopilación Investigativa 2010. Tomo II	978-99923-21-65-2 (v2)
Marzo 2011	Recopilación Investigativa 2010. Tomo III	978-99923-21-67-6 (v3)
Abril 2011	Diseño de un modelo de vivienda bioclimática y sostenible. Fase I. Colección Investigaciones n.º 5 Ana Cristina Vidal Vidales Luis Ernesto Rico Herrera Guillermo Vásquez Cromeyer	
Noviembre 2011	Importancia de los indicadores y la medición del quehacer científico. Colección Investigaciones n.º 6 Noris López de Castaneda	978-99923-21-71-3
Noviembre 2011	Memoria Sexta Semana del Migrante	978-99923-21-70-6
Mayo 2012	Recopilación Investigativa 2011. Tomo I	978-99923-21-75-1 (tomo 1)
Mayo 2012	Recopilación Investigativa 2011. Tomo II	978-99923-21-76-8 (tomo 2)
Mayo 2012	Recopilación Investigativa 2011. Tomo III	978-99923-21-77-5 (tomo 3)
Abril 2012	La violencia social delincuencial asociada a la salud mental en los salvadoreños Colección Investigaciones n.º 7 Ricardo Gutiérrez Quintanilla	978-99923-21-72-0
Octubre 2012	Programa psicopreventivo de educación para la vida efectividad en adolescentes Utec-PGR. Colección Investigaciones Ana Sandra Aguilar de Mendoza Milton Alexander Portillo	978-99923-21-80-6

Noviembre 2012	Causas de la participación del clero salvadoreño en el movimiento emancipador del 5 de noviembre de 1811 en El Salvador y la postura de las autoridades eclesiales del Vaticano ante dicha participación. Claudia Rivera Navarrete	978-99923-21-88-1
Noviembre 2012	Estudio Histórico proceso de independencia: 1811-1823. José Melgar Brizuela	978-99923-21-87-4
Noviembre 2012	El Salvador insurgente 1811-1821 Centroamérica. César A. Ramírez A.	978-99923-21-86-7
Enero 2012	Situación de la educación superior en El Salvador. Colección Investigaciones n.º 1 Carlos Reynaldo López Nuila	
Febrero 2012	Estado de adaptación integral del estudiante de educación media de El Salvador. Colección Investigaciones n.º 8 Luis Fernando Orantes	
Marzo 2012	Aproximación etnográfica al culto popular del Hermano Macario en Izalco, Sonsonate, El Salvador. Colección Investigaciones n.º 9 José Heriberto Erquicia Cruz	978-99923-21-73-7
Mayo 2012	La televisión como generadora de pautas de conducta en los jóvenes salvadoreños. Colección Investigaciones n.º 10 Edith Ruth Vaquerano de Portillo Domingo Orlando Alfaro Alfaro	
Mayo 2012	Violencia en las franjas infantiles de la televisión salvadoreña y canales infantiles de cable. Colección Investigaciones n.º 11 Camila Calles Minero Morena Azucena Mayorga Tania Pineda	
Junio 2012	Obrajes de añil coloniales de los departamentos de San Vicente y La Paz, El Salvador. Colección Investigaciones n.º 14 José Heriberto Erquicia Cruz	

Junio 2012	San Benito de Palermo: elementos afrodescendientes en la religiosidad popular en El Salvador. Colección Investigaciones n.º 16 José Heriberto Erquicia Cruz Martha Marielba Herrera Reina	978-99923-21-80-5
Julio 2012	Formación ciudadana en jóvenes y su impacto en el proceso democrático de El Salvador. Colección Investigaciones n.º 17 Saúl Campos	
Julio 2012	Factores que influyen en los estudiantes y que contribuyeron a determinar los resultados de la PAES 2011. Colección Investigaciones n.º 12 Saúl Campos Blanca Ruth Orantes	978-99923-21-79-9
Agosto 2012	Turismo como estrategia de desarrollo local. Caso San Esteban Catarina. Colección Investigaciones n.º 18 Carolina Elizabeth Cerna Larissa Guadalupe Martín José Manuel Bonilla Alvarado	
Agosto 2012	Conformación de clúster de turismo como prueba piloto en el municipio de Nahuizalco. Colección Investigaciones n.º 19 Blanca Ruth Gálvez García Rosa Patricia Vásquez de Alfaro Juan Carlos Cerna Aguiñada Óscar Armando Melgar.	
Septiembre 2012	Mujer y remesas: administración de las remesas. Colección Investigaciones n.º 15 Elsa Ramos	978-99923-21-81-2
Octubre 2012	Responsabilidad legal en el manejo y disposición de desechos sólidos en hospitales de El Salvador. Colección Investigaciones n.º 13 Carolina Lucero Morán	978-99923-21-78-2
Febrero 2013	Estrategias pedagógicas implementadas para estudiantes de Educación Media y el Acoso Escolar (*bullying*). Colección Investigaciones n.º 25 Ana Sandra Aguilar de Mendoza	978-99923-21-92-8

Marzo 2013	Representatividad y pueblo en las revueltas de principios del siglo XIX en las colonias hispanoamericanas. Héctor Raúl Grenni Montiel	978-99961-21-91-1
Marzo 2013	Estrategias pedagógicas implementadas para estudiantes de educación media. Colección Investigaciones n.º 21 Ana Sandra Aguilar de Mendoza	978-99923-21-92-8
Abril 2013	Construcción, diseño y validez de instrumentos de medición de factores psicosociales de violencia juvenil. Colección Investigaciones José Ricardo Gutiérrez Quintanilla	978-99923-21-95-9
Mayo 2013	Participación política y ciudadana de la mujer en El Salvador. Colección Investigaciones n.º 20 Saúl Campos Morán	978-99923-21-94-2
Mayo 2013	Género y gestión del agua en la mancomunidad La Montañona, Chalatenango, El Salvador. Normando S. Javaloyes Laura Navarro Mantas Ileana Gómez	978-99923-21-99-7
Junio 2013	Libro Utec 2012 Estado del medio ambiente y perspectivas de sostenibilidad. Colección Investigaciones José Ricardo Calles Hernández	978-99961-48-00-2
Julio 2013	Guía básica para la exportación de la flor de loroco desde El Salvador hacia España, a través de las escuelas de hostelería del país vasco. Álvaro Fernández Pérez	978-99961-48-03-3
Agosto 2013	Proyecto Migraciones Nahua-pipiles del Postclásico en la cordillera del Bálsamo. Colección Investigaciones n.º 24 Marlon V. Escamilla William R. Fowler	978-99961-48-07-1
Agosto 2013	Transnacionalización de la sociedad salvadoreña, producto de las migraciones. Colección Investigaciones n.º 25 Elsa Ramos	978-99961-48-08-8

Septiembre 2013	La regulación jurídico penal de la trata de personas especial referencia a El Salvador y España. Colección Investigaciones Hazel Jasmin Bolaños Vásquez	978-99961-48-10-1
Septiembre 2013	Estrategias de implantación de clúster de turismo en Nahuizalco. Colección Investigaciones n.º 22 Blanca Ruth Gálvez Rivas Rosa Patricia Vásquez de Alfaro Óscar Armando Melgar Nájera	978-99961-48-05-7
Septiembre 2013	Fomento del emprendedurismo a través de la capacitación y asesoría empresarial como apoyo al fortalecimiento del sector de la Mipyme del municipio de Nahuizalco en el departamento de Sonsonate. Diagnóstico de gestión Colección Investigaciones n.º 23 Vilma Elena Flores de Ávila	978-99961-48-06-4
Septiembre 2013	Imaginarios y discursos de la herencia afrodescendiente en San Alejo, La Unión, El Salvador. Colección Investigaciones n.º 26 José Heriberto Erquicia Cruz Martha Marielba Herrera Reina Wolfgang Effenberger López	978-9961-48-09-5
Septiembre 2013	Memoria Séptima Semana del Migrante	978-99961-48-11-8
Septiembre 2013	Inventario de las capacidades turísticas del municipio de Chiltiupán, departamento de La Libertad. Colección Investigaciones n.º 33 Lissette Cristalina Canales de Ramírez Carlos Jonatan Chávez Marco Antonio Aguilar Flores	978-99961-48-17-0
Septiembre 2013	Condiciones culturales de los estudiantes de educación media para el aprendizaje del idioma Inglés. Colección Investigaciones n.º 35 Saúl Campos Morán Paola María Navarrete Julio Aníbal Blanco	978-99961-48-22-4

Septiembre 2013	Recopilación Investigativa 2012. Tomo I	978-99923-21-97-3
Septiembre 2013	Recopilación Investigativa 2012. Tomo II	978-99923-21-98-0
Noviembre 2013	Infancia y adolescencia como noticia en El Salvador. Camila Calles Minero	978-99961-48-12-5
Diciembre 2013	Metodología para la recuperación de espacios públicos. Ana Cristina Vidal Vidales Julio César Martínez Rivera	978-99961-48-4-9
Marzo 2014	Participación científica de las mujeres en El Salvador. Primera aproximación. Camila Calles Minero	978-99961-48-15-6
Abril 2014	Mejores prácticas en preparación de alimentos en la micro y pequeña empresa. Colección Investigaciones n.º 29 José Remberto Miranda Mejía	978-99961-48-20-0
Abril 2014	Historias, patrimonios e identidades en el municipio de Huizúcar, La Libertad, El Salvador. Colección Investigaciones n.º 31 José Heriberto Erquicia Martha Marielba Herrera Reina Ariana Ninel Pleitez Quiñonez	978-99961-48-18-7
Abril 2014	Evaluación de factores psicosociales de riesgo y de protección de violencia juvenil en El Salvador. Colección Investigaciones n.º 30 José Ricardo Gutiérrez	978-99961-48-19-4
Abril 2014	Condiciones socioeconómicas y académicas de preparación para la de los estudiantes de educación media. Colección Investigaciones n.º 32 Saúl Campos Paola María Navarrete	978-99961-48-21-7
Mayo 2014	Delitos relacionados con la pornografía de personas menores de 18 años: especial referencia a las tecnologías de la información y la comunicación con medios masivos. Colección Investigaciones n.º 34 Hazel Jasmín Bolaños Miguel Angel Boldova Carlos Fuentes Iglesias	978-99961-48-16-3

Junio 2014	Guía de buenas prácticas en preparación de alimentos en la micro y pequeña empresa	
Julio 2014	Perfil actual de la persona migrante en El Salvador. Utec-US COMMITTE	978-99961-48-25-5
Septiembre 2014	Técnicas de estudio. Recopilación basada en la experiencia docente. Flavio Castillo	978-99961-48-29-3
Septiembre 2014	Valoración económica del recurso hídrico como un servicio ambiental de las zonas de recarga del río Acelhuate. Colección Investigaciones n.º 36 José Ricardo Calles	978-99961-48-28-6
Septiembre 2014	Migración forzada y violencia criminal una aproximación teórica practica en el contexto actual. Colección Investigaciones n.º 37 Elsa Ramos	978-99961-48-27-9
Septiembre 2014	La prevención del maltrato en la escuela. Experiencia de un programa entre alumnos de educación media. Colección Investigaciones n.º 38 Ana Sandra Aguilar de Mendoza	978-99961-48-26-2
Septiembre 2014	Percepción del derecho a la alimentación en El Salvador. Perspectiva desde la biotecnología. Colección Investigaciones n.º 39 Licda. Carolina Lucero	978-99961-48-32-3
Diciembre 2014	El domo el Guegüecho y la evolución volcánica. San Pedro Perulapán (Departamento de Cuscatlán), El Salvador. Primer Informe. Colección Investigaciones n.º 41 Walter Hernández Guillermo E. Alvarado Brian Jicha Luis Mixco	978-99961-48-34-7
Enero 2015	Publicidad y violencia de género en El Salvador. Colección Investigaciones n.º 40 Camila Calles Minero Francisca Guerrero Morena L. Azucena Hazel Bolaños	978-99961-48-35-4

Marzo 2015	Imaginario colectivo, movimientos juveniles y cultura ciudadana juvenil en El Salvador. Colección Investigaciones n.º 42 Saúl Campos Morán Paola María Navarrete Carlos Felipe Osegueda	978-99961-48-37-8
Mayo 2015	Estudio de buenas prácticas en clínica de psicología. Caso Utec. Colección Investigaciones n.º 44 Edgardo Chacón Andrade Sandra Beatriz de Hasbún Claudia Membreño Chacón	978-99961-48-40-8
Junio 2015	Modelo de reactivación y desarrollo para cascos urbanos. Colección Investigaciones n.º 48 Coralia Rosalía Muñoz Márquez	978-99961-48-41-5
Junio 2015	Niñas, niños, adolescentes y mujeres en la ruta del migrante. Colección Investigaciones n.º 54 Elsa Ramos	978-99961-48-46-0
Julio 2015	Historia, patrimonio e identidades en el Municipio de Comasagua, La Libertad, El Salvador. Colección Investigaciones n.º 49 José Heriberto Erquicia Cruz Martha Marielba Herrera Reina	978-99961-48-42-2
Agosto 2015	Evaluación del sistema integrado de escuela inclusiva de tiempo pleno implementado por el Ministerio de Educación de El Salvador. (Estudio de las comunidades educativas del municipio de Zaragoza del departamento de La Libertad). Colección Investigaciones n.º 43 Mercedes Carolina Pinto Benítez Julio Aníbal Blanco Escobar Guillermo Alberto Cortez Arévalo Wilfredo Alfonso Marroquín Jiménez Luis Horaldo Romero Martínez	978-99961-48-43-9
Agosto 2015	Aplicación de una función dosis-respuesta para determinar los costos sociales de la contaminación hídrica en la microcuenca del Río Las Cañas, San Salvador, El Salvador. Colección Investigaciones n.º 45 José Ricardo Calles Hernández	978-99961-48-45-3

Octubre 2015	El derecho humano al agua en El Salvador y su impacto en el sistema hídrico. Colección Investigaciones n.º 50 Sandra Elizabeth Majano Carolina Lucero Morán Dagoberto Arévalo Herrera	978-99961-48-49-1
Octubre 2015	Análisis del tratamiento actual de las lámparas fluorescentes, nivel de contaminantes y disposición final. Colección Investigaciones n.º 53 José Remberto Miranda Mejía Samuel Martínez Gómez John Figerald Kenedy Hernández Miranda	978-99961-48-48-4
Noviembre 2015	El contexto familiar asociado al comportamiento agresivo en adolescentes de San Salvador. Colección Investigaciones n.º 52 José Ricardo Gutiérrez Quintanilla Delmi García Díaz María Elisabet Campos Tomasino	978-99961-48-52-1
Noviembre 2015	Práctica de prevención del abuso sexual a través del funcionamiento familiar. Colección Investigaciones n.º 55 Ana Sandra Aguilar de Mendoza María Elena Peña Jeé Manuel Andreu Ivett Idayary Camacho	978-99961-48-53-8
Diciembre 2015	Problemas educativos en escuelas de Cojutepeque contados por los profesores y profesoras. Escuela de Antropología. Julio Martínez	
Febrero 2016	Desplazamiento interno forzado y su relación con la migración internacional. Colección Investigaciones n.º 56 Elsa Ramos	978-99961-48-56-9
Marzo 2016	Monografía Cultural y socioeconómica del cantón Los Planes de Renderos. Colección Investigaciones n.º 57 Saúl Campos Paola Navarrete Carlos Osegueda Julio Blanco Melissa Campos	978-99961-48-60-6

Abril 2016	Modelo de vivienda urbana sostenible. Colección Investigaciones n.º 58 Coralia Rosalía Muñoz Márquez	978-99961-48-61-3
Mayo 2016	Recopilación de Investigaciones en Tecnología 2016: Colección Investigaciones n.º 59 Internet de las cosas: Diseño e implementación de prototipo electrónico para el monitoreo vía internet de sistemas de generación fotovoltaico. Omar Otoniel Flores Cortez German Antonio Rosa Implementación de un entorno de aprendizaje virtual integrando herramientas de *E-learning* y CMS. Marvin Elenilson Hernández Carlos Aguirre *Big data*, análisis de datos en la nube. José Guillermo Rivera Verónica Idalia Rosa Urrutia	978-99961-48-62-0
Julio 2016	Aplicación de buenas prácticas de negocio (pequeña y mediana empresa de los municipios de San Salvador, Santa Tecla y Soyapango en El Salvador.) Colección Investigaciones n.º 46 Vilma de Ávila	978-99961-48-44-6
Julio 2016	Afectaciones psicológicas en estudiantes de instituciones educativas públicas ubicadas en zonas pandilleriles. Colección Investigaciones n.º 60 Edgardo R. Chacón Manuel A. Olivar Robert David MacQuaid Marlon E. Lobos Rivera	978-99961-48-67-5
Octubre 2016	Los efectos cognitivos y emocionales presentes en los niños y las niñas que sufren violencia intrafamiliar. Colección Investigaciones n.º 61 Ana Sandra Aguilar Mendoza	978-99961-48-69-9
Noviembre 2016	Historia, patrimonio e identidad en el municipio Puerto de La Libertad, El Salvador. Colección Investigaciones n.º 62 José Heriberto Erquicia Cruz Paola María Navarrete Gálvez	978-99961-48-70-5

Febrero 2017	El comportamiento agresivo al conducir asociado a factores psicosociales en los conductores salvadoreños. Colección Investigaciones n.º 63 José Ricardo Gutiérrez Quintanilla Óscar Williams Martínez Marlon Elías Lobos Rivera	978-99961-48-72-9
Marzo 2017	Relaciones interétnicas: afrodescencientes en Centroamérica. Colección Investigaciones n.º 64 José Heriberto Erquicia Rina Cáceres	978-99961-48-73-6
Abril 2017	Diagnóstico de contaminiación atmosférica por emisiones diésel en la zona metropolitana de San Salvador y Santa Tecla. Cuantificación de contaminantes y calidad de combustibles. Colección Investigaciones n.º 65 José Remberto Miranda Mejía Samuel Martínez Gómez Yonh Figerald Kenedy Hernández Miranda René Leonel Figueroa Noé Aguirre	978-99961-48-75-0
Mayo 2017	Causas y condiciones del incremento de la migración de mujeres salvadoreñas. Colección Investigaciones n.º 66 Elsa Ramos	978-99961-48-76-7
Junio 2017	Etnografía del volcán de San Salvador. Colección Investigaciones n.º 67 Saúl Campos Morán Paola María Navarrete Carlos Felipe Osegueda	978-99961-48-77-4
Agosto 2017	Modelo de e-Turismo cultural aplicando tecnología *m-Learning*, georreferencia, visitas virtuales y realidad aumentada para dispositivos móviles. Colección Investigaciones n.º 68 Elvis Moisés Martínez Pérez Melissa Regina Campos Solórzano Claudia Ivette Rodríguez de Castro Ronny Adalberto Cortez Reyes Rosa Vania Chicas Molina Jaime Giovanni Turcios Dubón	978-99961-48-80-4

Octubre 2017	Influencia de la tradición oral, la cocina que practican los pueblos indígenas y las variantes dialectales en la conservación y difusión de la lengua náhuat pipil. Colección Investigaciones n.º 69 Morena Guadalupe Magaña de Hernández Jesús Marcos Soriano Aguilar Clelia Alcira Orellana Mercedes Carolina Pinto Julio Aníbal Blanco José Ángel García Tejada	978-99961-48-84-2
Noviembre 2017	Propuesta de políticas públicas frente al perfil demográfico de El Salvador Carolina Lucero Morán Guiomar Bay Saúl Campos Morán Lucía del Carmen Zelaya de Soto	978-99961-48-87-3
Noviembre 2017	El estado de las competencias de desarrollo de la mujer en la zona de La Libertad Ana Sandra Aguilar de Mendoza	978-99961-48-88-0
Diciembre 2017	Conocimiento financiero y económico entre estudiantes universitarios: un estudio comparativo entre El Salvador y Puerto Rico Modesta Fidelina Corado Roberto Filánder Rivas Ronald Hernández Maldonado	978-99961-48-89-7
Enero 2018	Situación actual del manejo de las aguas ordinarias en lotificaciones y parcelaciones habitacionales de la zona rural de El Salvador. Un análisis de cumplimiento técnico y legal aproximado Alma Carolina Sánchez Fuentes María Teresa Castellanos Araujo Ricardo Calles Hernández Erick Abraham Castillo Flores	978-99961-48-91-0

Este libro se terminó de imprimir
en el mes de abril de 2018
en los talleres de Tecnoimpresos, S.A. de C.V.
19ª. Av. Norte N.º 125,
ciudad de San Salvador, El Salvador, C.A.

Big data, en El Salvador, es un sistema tecnológico novedoso, lo que hace necesario incursionar en esta tendencia. Por esta razón, el objetivo de la presente investigación fue aplicar herramientas *big data* en el ámbito gubernamental para almacenar, procesar y analizar sus grandes cantidades de datos, con el fin de lograr conclusiones que ayudarían en la toma de decisiones en menor tiempo. Para esta investigación se hizo uso de *dataset* con información sobre postulantes a vivienda en el Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas. Los datos fueron almacenados y procesados mediante herramientas *big data,* tales como Hadoop y Hive para análisis estadístico, finalizando con la creación de visualizaciones en Google chart y D3. La investigacion se llevó a cabo durante el período de febrero a noviembre de 2016.

La Colección Investigaciones tiene el objetivo de evidenciar el trabajo científico de la Universidad Tecnológica de El Salvador ante la comunidad científica nacional e internacional, y la sociedad.

No hay enseñanza sin investigación ni investigación sin enseñanza
Pablo Freire

ISBN 978-99961-48-97-2

Vicerrectoría de Investigación y Proyección Social
Calle Arce y 19ª avenida Sur n.° 1045, edificio *Dr. José Adolfo Araujo Romagoza,*
San Salvador, El Salvador, (503) 2275 1013 / 2275 1011