- 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.

## 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.

## Equipo investigador

**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 Maestria 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.



## Otras líneas de investigación Utec

- Turismo
- Democracia y gobernabilidad
- Comunicación para el desarrollo
- Vivienda y desarrollo urbano
- Desarrollo e innovación tecnológica

## Investigación en breve

Es una colección de fascículos que resumen los resultados de las investigaciones realizadas por la Vicerrectoría de Investigación y Proyección Social.

*No hay enseñanza sin investigación ni investigación sin enseñanza*
Pablo Freire

Vicerrectoría de Investigación y Proyección Social

Calle Arce y 19ª avenida Sur n.º 1045, edificio *Dr. José Adolfo Araujo Romagoza*.
San Salvador, El Salvador, (503) 2275-1013 / 2275-1011

www.utec.edu.sv | Centro de llamadas: 2275-8888 Maestrías: 2275-2700

HAGAMOS LA DIFERENCIA

Universidad Tecnológica de El Salvador

n.º 13
ABRIL 2018
INVESTIGACIÓN EN BREVE

Universidad Tecnológica de El Salvador

Vicerrectoría de Investigación y Proyección Social

# 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

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.

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

Shutterstock

## Introducción

*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.

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

La explotación de la tecnología del *big data* permite que las empresas conozcan más de cerca a sus clientes, prestarles un mejor servicio, mejorar la calidad de sus productos, generar oportunidades para ingresar a nuevos mercados, completar sus portafolios de clientes, entre otras tareas que generen beneficios en sus negocios.

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 están basados 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.

## Objetivos de la investigación

**General**

- Aplicación de herramientas *big data* para el Viceministerio de Vivienda y Desarrollo Urbano del Ministerio de Obras Públicas (MOP).

**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.

## 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.

## Participantes

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

## 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*.

## 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.

## Resultados

Esta investigación propuso solucionar el problema en el manejo 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.

Se procesaron datos usando Hadoop. Luego, con la herramienta Hive, se realizaron las consultas necesarias para su aplicación, así como se utilizó el programa R para hacer análisis estadístico de los datos, y, por último, se usaron herramientas de visualización, tales como Google Chart y D3.js, para presentar visualmente los resultados obtenidos.

A continuación se presentan las herramientas *big data* utilizadas en la investigación.

**Hadoop.** 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.

**MapReduce.** 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 Google File System (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.

**Hive.** Apache Hive 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.

**R.** Se puede definir R desde dos perspectivas: como un entorno de *software* y como 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. A través del 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.

## Herramientas de visualización

**Google Chart** 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**, 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.

Después de haber explorado los datos del Viceministerio Vivienda y Desarrollo Urbano y de 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.

## 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.