# Big Data en las organizaciones

## Big Data in organizations

***Luis Ernesto Gerena Salas***[1]


## *Resumen*

La información generada en las empresas viene en crecimiento exponencial, gracias a la variedad de plataformas tecnológicas que se implementan para mejorar, controlar o simplemente mantener registro de eventos en el día a día. Eventos que van desde las ventas de un producto, hasta la hora de ingreso de los empleados. En el primer caso se obtiene información del cliente, su edad, género, tarjeta de crédito, datos de contacto, entre otra información para después continuar enviándole información o para generar perfiles y comportamiento de clientes a los que se les puede ofrecer cualquier producto. Esta información correlacionada es la que utilizan hoy las empresas, con el fin de generar nuevos productos, crecer en ventas y reconocimiento de las marcas.


***Palabras clave: Big Data, NoSQL, Hadoop, IoT, sensores, seguridad.***


## *Abstract*

Nowadays Digital information generated by companies is growing exponentially, thanks to the variety of technological platforms that are being implemented to improve, control or simply to keep the record of events in the day-to-day. Those events can simply go from the sales of a product, to the clock-in time of employees. In the first case, customer's information is obtained such as their age, gender, credit card, and contact information, among others for the purpose of sending them subsequent information or simply to define profiles and check their behaviors so that any information can be send of any product. Nowadays most companies are using this information to generate new products, grow in sales and to have brand recognition in their own products.


***Keywords: Big Data, NoSQL, Hadoop, IoT, sensors, security***



1 Ingeniero de telecomunicaciones, Universidad Santo Tomás, especialista en Gerencia de Empresas de Telecomunicaciones, Universidad de los Andes. Docente Escuela Tecnológica Instituto Técnico Central, Investigador grupo K-Demy. Correo electrónico: legerenas@itc.edu.co

## 1. Introducción

El grupo de investigación K-Demy, ha buscado en los últimos proyectos, realizar investigaciones que permitan a la escuela el aprovechamiento de su datacenter m*óvil, así como de la infraestructura de red y los laboratorios con los que cuenta, al mismo tiempo*, generar proyectos que le permitan a la comunidad académica entender y conocer las nuevas tecnologías, infraestructuras y desarrollos que existen en organizaciones reconocidas del sector.

Big Data es la unión de elementos de infraestructura con el objetivo de generar valor a las personas. Para comenzar a explicar qu*é* es Big Data, se debe hablar en términos de capacidades, generalmente se define en términos de petabytes o exabytes. Según las redes sociales, en un día normal se pueden llegar a crear contenidos con un tamaño aproximado de 500 TB (Joyanes, 2013). Con esta cantidad de información, se comenzó a hablar de Big Data y la tecnología que conlleva para el manejo de esta información.

El crecimiento de esta tecnología ha permitido nuevas investigaciones realizadas por empresas, universidades y los profesionales que trabajan en el sector, quienes argumentan que en Big Data ya no es solo la capacidad la que se sobrepasa, sino los límites de los mismos datos. Es decir, se quiere que los datos se muestren cada vez más rápido (información en tiempo real), que pueda mostrar cualquier tipo de información a la que se requiera acceder (tráfico en las calles, temperatura en una nevera, hasta contenido en un televisor), generalmente dispositivos del Internet de las Cosas (IoT).

Según algunos autores, se hablaba al inicio del Big Data de 3 V: velocidad, volumen y variedad, que quieren decir: velocidad en el procesamiento de la información; volumen en la cantidad de datos que hoy en día se pueden llegar a almacenar; variedad se refiere a los diferentes formatos existentes y que requieren ser procesados, almacenados y posteriormente manipulados.

Durante el desarrollo del proyecto y con autores y documentos actualizados, se encontró una cuarta "V", el valor que la información (posterior a un análisis) puede entregar a cualquier entidad, organización o persona, con el fin de que esta genere nuevas formas de ver y cambiar el mundo.

La creación de estas cantidades de información no solo se obtienen de los contenidos de las redes sociales, existen equipos que en todo momento crean los datos que requieren ser almacenados, dispositivos como sensores del clima que pueden obtener la temperatura, la humedad; puertas de acceso que obtienen el registro de qui*é*n accede o sale, entregando la hora exacta; cámaras de videovigilancia, que obtienen los videos de los circuitos de televisión en una empresa, toda esta información puede ser guardada en sistemas de almacenamiento creados para este fin (Pérez, 2015).

## 2. Metodología

**Fase 1:** implantar en la Escuela Tecnológica Instituto Técnico Central – ETITC, dispositivos capaces de obtener por s*í* solos y en determinados momentos, información sobre la temperatura, humedad, tráfico en la red de datos de la ETITC o y el momento exacto de estas mediciones. Los sensores realizan las mediciones 2 veces por minuto durante 1 mes, lo que permite recolectar más de 250.000 datos, que generalmente no es posibles manejarlos en una simple hoja de cálculo.

**Fase 2:** b*úsqueda de herramientas existentes que permitan almacenar toda la cantidad de información obtenida de los sensores (con la menor cantidad de fallas), procesarla y permitir el análisis de los datos. Determinar el mejor software existente, ya sea con licenciamiento o herramientas open source* que no solo permitan su instalación, sino mejoras y adecuaciones de acuerdo con los recursos que se tienen para su implementación en el datacenter de la Institución o en los equipos dispuestos para ello.

**Fase 3:** ejecución y desarrollo de pruebas con el objetivo de determinar de acuerdo a cada una de las características, la mejor herramienta a ser utilizada en la ETITC, su implementación y el costo por su uso.

## 3. Avances de resultados

### 3.1. Sensores

Uno de los objetivos iniciales fue la obtención de datos y el tipo de información que puede llegar a ser eficiente en algún momento, tanto para la Escuela como para una empresa o la sociedad en sí, por lo que desde el inicio, el grupo de investigación determinó que la información que se puede llegar a obtener (también gracias a la tecnología con los sensores) son los datos del clima (temperatura, humedad, luminosidad del sol), así mismo, llegar a obtener datos del tráfico que viaja por la red, con el fin de llegar a identificar, de ser posible, qué tráfico se observa en la red de datos de la ETITC.

Para la obtención de datos, se realizaron pruebas con algunos dispositivos Arduino ethernet, los cuales permiten la conexión de sensores y la conexión a la red de datos y así mantener una conexión con la red de datos de la ETITC y realizar las capturas del tráfico.

Debido a la cantidad de datos obtenidos y por la frecuencia de tomas, se decidió cambiar a un equipo con un mayor procesamiento y capacidad, como lo es el Raspberry Pi, con las mismas funcionalidades y mayores capacidades tanto en el almacenamiento como en el procesamiento y reenvío de los datos a un servidor. Este último es uno de los puntos más importantes en la implementación de sensores, debido a la cantidad de datos que se pueden obtener en un momento determinado en el dispositivo y al tiempo utilizar recursos para el envío de información hacia los servidoresdonde van a quedar almacenados para su posterior manejo. Se realizó de esta manera porque en los Arduinos no es posible realizar procesos de análisis sin afectar el procesamiento, velocidad y capacidad, necesarias para las tareas de obtención de datos.

### 3.2. Almacenamiento:

Al tener los datos de los sensores, estos deben ser enviados a un sistema capaz de almacenarlos de forma ordenada, segura y que estén disponibles en el momento que se necesiten. Además, los datos deben ser enviados desde los sensores de la manera más eficiente hacia los servidores, sin perder información y sin que esta sea cambiada en su transporte hacia los servidores de almacenamiento. Otro de los objetivos de este tipo de sistemas es cumplir con los tres pilares de seguridad de la información, donde la información debe estar disponible en todo momento, ser confidencial (solo la deben ver los usuarios permitidos) e íntegra (es decir, no debe ser alterada por personas no autorizadas).

Estos servidores deben tener la capacidad de guardar cualquier tipo de información, en el formato que exista, deben poder almacenar (en el caso de la ETITC) datos de horas, grados, URL de páginas web, imágenes, audio, video y mantener estos datos accesibles.

Este tipo de servicios son los utilizados en la actualidad por las redes sociales que existen en internet, donde se mantiene cualquier tipo de información y donde cada usuario es el dueño de ella; igualmente, tiene acceso en cualquier momento para revisarla y/o compartirla con otros usuarios.

En el caso de las organizaciones, las redes sociales permiten esto y otras opciones adicionales, como analizar quiénes ven las publicaciones, los que siguen a la compañía, los que compran, sus comportamientos, gustos y así generar mejor publicidad, encontrar nuevos productos para sus clientes y nuevas formas de generarles valor.

Este tipo de sistemas con sensores y servidores obteniendo, analizando y entregando información es usado en cualquier sector, uno de los más empleados son las aplicaciones, con la finalidad de evitar el tráfico en las ciudades.

Existen aplicaciones que gracias al sensor GPS de los teléfonos inteligentes son capaces de capturar

la posición en cualquier parte de la tierra, medir la velocidad y enviar toda esta información a un servidor donde se procesa, almacena, analiza y finalmente se entrega a otros teléfonos inteligentes una ruta más rápida para llegar desde dos puntos diferentes en las ciudades.

Los datos entregados por este tipo de aplicaciones es lo que hace de Big Data la tecnología que puede ser implementada por cualquier organización con el fin de generar valor a sus clientes nuevos o existentes.

La topología de herramientas capaces de almacenar gran cantidad de datos se implementa uniendo en red varios equipos servidores (en vez de uno solo), donde la capacidad de procesamiento y almacenamiento se agrega como si fuera un solo equipo. Esto quiere decir que, en vez de tener 10 servidores con amplios recursos que no se utilizan, se tiene un splo equipo con una capacidad 10 veces mayor y donde la información se almacena de manera repetida y secuencial.

Esto permite mantener la integridad y disponibilidad de la información en caso de falla de cualquiera de los equipos físicos, donde los demás son capaces de continuar realizando las funciones (actuando siempre como un sistema).

Al conectar los equipos en paralelo, se comparte también la cantidad de información que cada disco tiene que almacenar y que debe entregar en un momento dado. La figura 1 muestra el tiempo (en segundos) que se toma acceder a 1 TeraByte de datos, que resulta de los siguientes datos: hoy en día un equipo normal tiene 1 TB en disco y una velocidad de acceso de 5 Gbps, es decir, leer todo el disco, se puede tomar aproximadamente 200 segundos. Al tener 10 equipos en paralelo, ese mismo tamaño (1 TB) se puede almacenar de a 100 GB en cada equipo y el tiempo de acceder a la información puede bajar hasta 20 segundos. Cada día crecen los datacenter, la cantidad de dispositivos conectados en paralelo y la velocidad en los discos, lo que hace que cada vez sea más rápido (Miller, 2015).

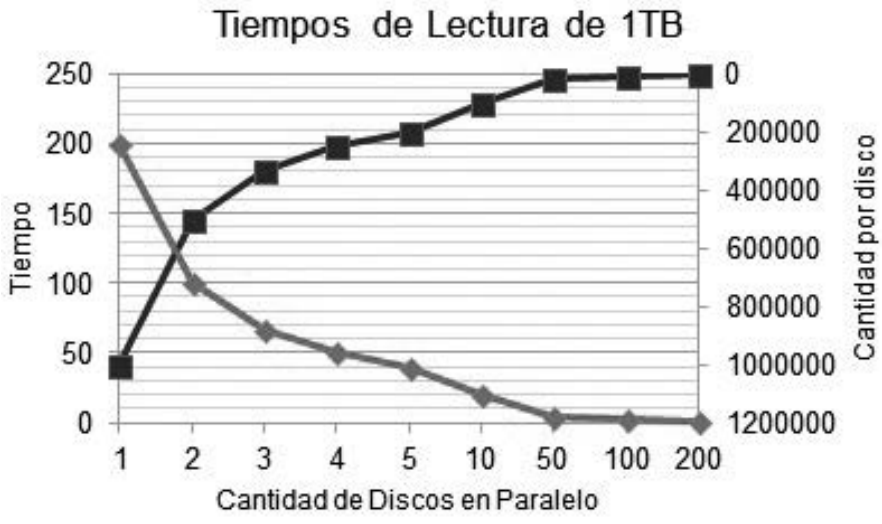


**Figura 1.** Tiempo de acceso a información con infraestructura en paralelo.
Fuente: autores

Después de las pruebas realizadas en el laboratorio, la herramienta que permite trabajar de esta manera es Hadoop, una infraestructura digital que está creada en *open source* con la licencia de Apache, que comenzó en empresas de internet como Yahoo y Google y quienes han invertido tiempo y

recursos financieros para desarrollar día a día esta tecnología (Marz y Warren, 2015). Con esta tecnología, otras empresas reconocidas han realizado desarrollos propios que permiten mejoras en el software y adecuaciones para cada tipo de empresa que lo implemente. Dentro de estas empresas está Oracle, IBM, Microsoft y SAS, (esta última ha ganado gran valor por sus innovaciones en Big Data en los últimos años), y cada una permite algunas características, aplicaciones y formas diferentes de manejar la información (White, 2015).

### 3.3. Bases de datos*:*

Una de las características más importantes de este tipo de sistemas, son las bases de datos no estructuradas (NoSQL), donde básicamente no se utiliza el lenguaje de consulta estándar (SQL), utilizado en las empresas y en bases de datos relacionales. A pesar de no ser un estándar aún, las grandes empresas de tecnología, como buscadores, redes sociales, entre otras, las utilizan con el fin de tener mayor velocidad al hacer búsquedas de datos históricos por cada usuario, ya que no tiene varias tablas para encontrar la información (como sucede con las BD SQL), sino que todo se encuentra gracias a una clave y un valor buscado en solo un índice (Uckelmann, 2011).

## *4. Conclusiones*

La información día a día crece de forma exponencial, se generan datos en la mayoría de las cosas que las personas hacen, las empresas buscas encontrar comportamientos y rutinas en el clima, las personas y dentro de poco hasta los objetos (vehículos, electrodomésticos) y así llegar a ofrecer nuevos e innovadores productos, tanto tangibles como intangibles, para esto, se necesita acceder a la información se necesita acceder en tiempo real, sin fallas y sin riesgos de haber sido alterada.

Si al hacer pruebas de seguridad informática de este tipo de infraestructura, es posible determinar que los riesgos pueden ser minimizados, las empresas, universidades y cualquier organización comenzar*án* a implementar esta tecnología para almacenar datos de gran tamaño y de acceso permanente.

Las bases de datos NoSQL permiten mayor velocidad en las búsquedas de información, son utilizadas por las empresas más grandes de tecnología de la actualidad, por lo que no solo deben ser investigadas, sino que se debe buscar la forma de mejorar este tipo de tecnología para que sea implementada en todas las organizaciones, si se quiere mejorar el acceso a la información por parte de los usuarios internos y externos.

Con el fin de aprovechar los recursos de los dispositivos como RAM y almacenamiento, se debe considerar en las organizaciones implementar este tipo de sistemas capaces de generar más capacidad, robustez y fiabilidad en los datacenters. Como se evidenció en la figura 1, entre más dispositivos sean configurados en la red en paralelo, será posible llegar a tiempos menores para acceder y recuperar información guardada.

## *5. Referencias bibliográficas*


Joyanes, L. (2013). *Análisis de grandes volúmenes de datos en organizaciones*.

Marz, N., y Warren, J. (2015). *Big Data. Principles and best practices of scalable real-time data systems.*

*Miller, M. (2015). The Internet of Things. How Smart TVs, Smart Cars, Smart Homes, and Smart Cities Are Changing the World.*

Pérez, M. (2015). *Big Data. Técnicas, herramientas y aplicaciones.*

*Uckelmann, D., Harrison, M., y Michahelles, F.(2011). Architecting the Internet of Things.*

*White, T. (2015). Hadoop. The Definitive Guide. Storage and analysis at Internet scale.*