Universidad&Ciencia
Pág. 115-123
Recibido: 16/06/2017 Aceptado: 16/09/2017
Publicado 30/10/2017

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

ANÁLISIS DE DATOS SEMIESTRUCTURADOS CON FORMATO JSON EN EL SISTEMA DE ARCHIVOS DISTRIBUIDOS HADOOP

ANALYSIS OF SEMI-STRUCTURED DATA WITH JSON FORMAT IN THE HADOOP DISTRIBUTED FILES SYSTEM

Autores: Jimmy Josué Peña Koo
Jorge Alfredo Colli Chi
José Ildefonso Espinosa Pacho
Institución: Instituto Tecnológico Superior del Sur del Estado de Yucatán
Correo electrónico: jimjpk@itsyucatan.edu.mx

RESUMEN

El presente trabajo de investigación demuestra el proceso de tratamiento para datos semi-estructurados en formato JSON, por medio de técnicas de Big Data. Para su desarrollo se empleó como herramienta principal la distribución de Linux Cloudera, este sistema operativo contiene un entorno de trabajo para gestionar información masiva y herramientas para el procesamiento de datos estructurados y semi-estructurados. Para la demostración, se trabajó con un caso de estudio que contiene la información de contaminación del estado de México, los indicadores analizados son: ozono, dióxido de azufre, dióxido de nitrógeno, monóxido de carbono y partículas suspendidas pm10.

Palabras clave: Big Data, Cloudera, Hadoop, JSON

ABSTRACT

This paper demonstrates the treatment process for semi-structured data in JSON format, using techniques of Big Data. For its development was used as the main tool Cloudera Linux distribution, this operating system contains a working environment to manage massive information and tools for processing structured and semi-structured data. For the demonstration, we worked with a case study containing information pollution Mexico City, the

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

indicators analyzed are: ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide and suspended particles pm10.

Keywords: Big Data, Cloudera, Hadoop, JSON.

INTRODUCCIÓN

Internet y los dispositivos móviles permiten que diariamente cada persona genere una enorme cantidad de información sobre intereses, valores y preferencias de consumo.

Globalmente el estudio sobre internet y las redes sociales determinó a enero de 2016, que de los más de 7,395 millones de habitantes del planeta, 3,419 millones tienen acceso a internet, 2,307 millones usan regularmente las redes sociales, 3,790 millones utilizan un teléfono móvil y 1,968 millones de personas acceden a las redes sociales a través de éstos (We are social, 2016).

La recolección, almacenamiento, manejo y análisis en tiempo real de estos volúmenes gigantescos de información originó Big Data.

Estos datos vienen de todas partes: sensores utilizados para recopilar información sobre el clima, mensajes a sitios de medios sociales, fotos, vídeos digitales, registros de transacciones de compra, señales de telefonía celular, entre otros.

Big Data trabaja con datos estructurados, semi-estructurados y no estructurados, tales como texto, datos de sensores, audio, vídeo y archivos. Del análisis de estos tipos de datos se obtienen nuevos conocimientos empleados para toma de decisiones.

Big Data trabaja con las herramientas de software comúnmente utilizadas en disciplinas de análisis avanzadas, tales como: análisis predictivo, minería de datos, análisis de texto y análisis estadístico. Pero los datos semi-estructurados y no estructurados no son siempre compatibles con los sistemas de datos tradicionales basados en bases de datos relacionales. Los sistemas manejadores de bases de datos relacionales no son capaces de manejar las demandas de procesamiento que plantea Big Data, que necesitan ser actualizadas con frecuencia, por ejemplo, datos en tiempo real para aplicaciones móviles. Por lo anterior expuesto, se propone demostrar el tratamiento de datos semi-estructurados por medio de la herramienta Hadoop, incorporada en Cloudera.

En los orígenes del Big Data, el equipo Proyecto Manhattan en el año 1946 consigue utilizar computadoras para analizar y predecir el comportamiento que podría causar una reacción

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

nuclear en cadena (Azcárraga, 2016). En el año 1995, surge un nuevo enfoque cuando se lanzan los sitios web de eBay y Amazon, que suponen el comienzo de la carrera para la personalización de las diversas experiencias en línea para cada uno de los usuarios.

De acuerdo con Camargo (2015), Big Data se refiere a cantidades masivas de datos que se acumulan con el tiempo, que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos.

Una clasificación de los datos empleados por los sistemas de explotación de la información Big Data, es datos semi-estructurados. Consiste en información procesada y con formato definido pero con estructura variable, como las bases de datos basadas en columnas y los archivos con información en un lenguaje de marcado (Russom, 2011).

Para este trabajo se propone demostrar el proceso de tratamiento de datos semi-estructurados con Hadoop, por medio del entorno proporcionado por la distribución de Linux Cloudera.

Un trabajo relacionado publicado por Vasilenko y Kurapati (2014), trata del procesamiento eficiente de documentos XML en Hadoop Map Reduce, este trabajo describe un enfoque genérico para el manejo de XML basado en la arquitectura Apache Hive.

Otra investigación relacionada es la propuesta de Urmila Pol (2014), en ella demuestra el uso del administrador Cloudera para el proceso de formación de un clúster Hadoop de cuatro nodos con VirtualBox.

MATERIALES MÉTODOS

Con base en la clasificación de Hernández (2010) esta investigación cuantitativa tiene un diseño no experimental transeccional descriptivo, porque se realizó sin manipular deliberadamente las variables, se analizó el tratamiento de datos semi-estructurados con formato JSON en el Sistema de Archivos Distribuidos Hadoop (HDFS, por sus siglas en inglés Hadoop Distributed File System), tomado como caso de estudio los niveles de contaminación de la ciudad de México resultantes de mediciones en un único momento en el tiempo, correspondiendo al mes de mayo del año 2015, censado a partir de técnicas de Internet de las Cosas.

El diseño transeccional descriptivo indaga la incidencia de las modalidades, categorías o niveles de una o más variables en una población, son estudios puramente descriptivos.

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

El trabajo estudia la incorporación de archivos de formato JSON para su tratamiento con HDFS, por medio de las categorías de contaminantes registrados entre las cuales figuran: ozono, dióxido de azufre, dióxido de nitrógeno, monóxido de carbono y partículas suspendidas pm10, en las zonas noroeste, noreste, centro, suroeste y sureste de la Ciudad de México, para luego proporcionar su descripción.

Para el desarrollo del proyecto se empleó la metodología ICAV propuesta por la empresa Big Data SAC, empresa de consultoría especializada en Big Data en el Perú (Big Data SAC, 2013). La metodología es basada en cuatro pasos: primero, identificar las necesidades del área usuaria; segundo, consolidar la información que se requiere para el análisis de los datos; tercero, análisis predictivo con la información consolidada; y cuarto, visualizar los resultados para su distribución a los usuarios finales.

RESULTADO Y DISCUSIÓN

Para el análisis de la información se trabajó a partir del censo de los contaminantes que se encuentran en el aire de la ciudad de México, recopilado en mayo del año 2015, almacenado con formato JSON, el cual es definido por Álvarez (2014) como un formato estándar abierto que utiliza texto legible para transmitir objetos de datos que consisten en pares atributo-valor.

Este trabajo está basado en el entorno que presenta la distribución de Linux Cloudera, la cual ofrece soluciones para almacenamiento, gestión y acceso a soluciones Big Data basadas en Hadoop.

La recopilación de la información fue cargada en Apache Hadoop, un framework de software que soporta aplicaciones distribuidas bajo licencia libre que permite a las aplicaciones trabajar con miles de nodos y petabytes de datos (The Apache Software Foundation, 2016). Para esta etapa se utilizó la herramienta Hue, una interfaz Web para el análisis de los datos con Apache Hadoop tal como se puede ver en la figura 1.

Universidad&Ciencia
Pág. 115-123
Recibido: 16/06/2017 Aceptado: 16/09/2017
Publicado 30/10/2017

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Figura 1. Entorno Web Hue para el análisis de datos.

Posterior a la carga de datos se empleó la herramienta Hive, un controlador de bases de datos para Cloudera que permite a los usuarios de la empresa acceder a información de Hadoop a través de Inteligencia de Negocios (BI, por sus siglas en inglés Business Intelligence) desde aplicaciones con soporte ODBC. La herramienta Hive permitió realizar consultas de tipo SQL a la información previamente cargada, obteniendo un mayor procesamiento e interpretación de los datos, ya que cada consulta visualiza la tabla de resultados y permite graficar los resultados generados por la consulta. Esta interfaz permite generar informes de los datos previamente obtenidos.

Como ejemplo del caso de estudio analizado, se presenta información por día de los índices más altos de monóxido de carbono en la zona centro del estado de México (figura 2), obtenido de la consulta SQL “select dia, max(centro_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia”, cuya gráfica resultante generada por la misma interfaz se puede observar en la figura 3.

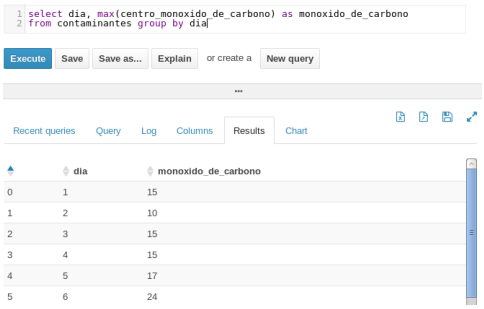


Figura 2. Consulta de niveles de contaminación de monóxido de carbono en zona centro.

Universidad&Ciencia
Pág. 115-123
Recibido: 16/06/2017 Aceptado: 16/09/2017
Publicado 30/10/2017

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

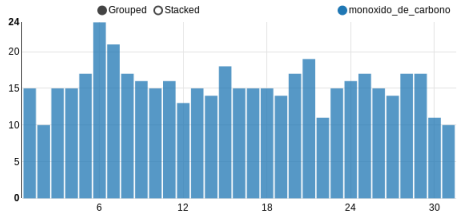
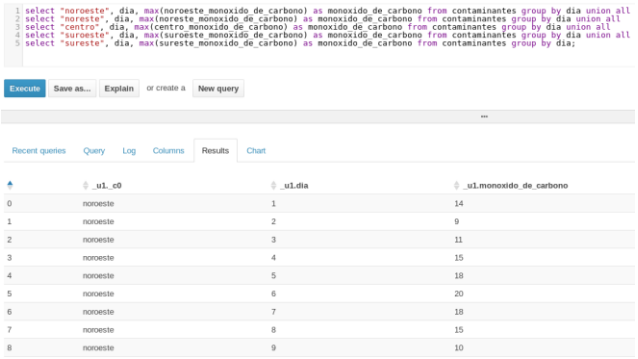


Figura 3. Gráfica de niveles de contaminación de monóxido de carbono en zona centro.

De los resultados obtenidos anteriormente se puede observar en la figura 3 que el incremento más alto de monóxido de carbono en este periodo, fue después de la conmemoración de un día festivo, Batalla de Puebla.

Otra información de interés con respecto a la variable monóxido de carbono, fue realizar un comparativo por zonas en el mismo periodo de tiempo presentado en la figura 4. Resultado de este análisis, se observa en la gráfica de la figura 5 que en las zonas sureste y suroeste se emiten los menores niveles de contaminación de monóxido de carbono en el estado de México, con respecto a las zonas centro, noreste y noroeste.



```
select "noroeste", dia, max(noroeste_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia union all
select "noreste", dia, max(noreste_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia union all
select "centro", dia, max(centro_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia union all
select "suroeste", dia, max(suroeste_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia union all
select "sureste", dia, max(sureste_monoxido_de_carbono) as monoxido_de_carbono from contaminantes group by dia;
```

Execute | Save as... | Explain | or create a | New query

Recent queries | Query | Log | Columns | Results | Chart

	_u1._c0	_u1.dia	_u1.monoxido_de_carbono
0	noroeste	1	14
1	noroeste	2	9
2	noroeste	3	11
3	noroeste	4	15
4	noroeste	5	18
5	noroeste	6	20
6	noroeste	7	18
7	noroeste	8	15
8	noroeste	9	10

Figura 4. Consulta comparativa de monóxido de carbono por zonas.

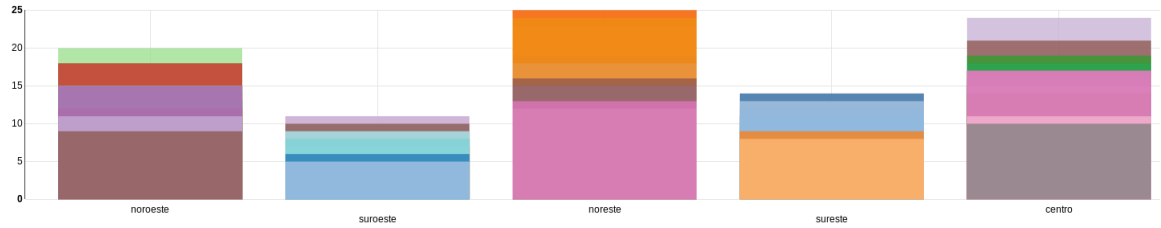


Figura 5. Gráfica comparativa de monóxido de carbono por zonas.

Universidad&Ciencia
Pág. 115-123
Recibido: 16/06/2017 Aceptado: 16/09/2017
Publicado 30/10/2017

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

El comportamiento de la gráfica presentada en la figura 6, demuestra que los fines de semana son las ocasiones cuando se presenta una disminución favorable del nivel de contaminación de monóxido de carbono.

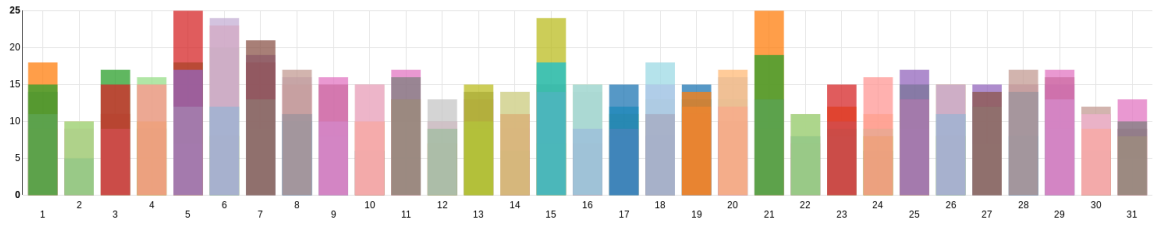


Figura 6. Gráfica de niveles de monóxido de carbono.

Adicional a la observación del monóxido de carbono, se analizaron los niveles de contaminación ocasionadas por el ozono, dióxido de azufre, dióxido de nitrógeno y pm10, cuyas estadísticas se presentan en la figura 7. Como apoyo a la interpretación de esta información obtenida se realizó la gráfica de la figura 8, para observar cuáles de estos indicadores afectaron en mayor medida al estado de México durante el tiempo de censado, resultando una afectación en menor medida del dióxido de azufre.

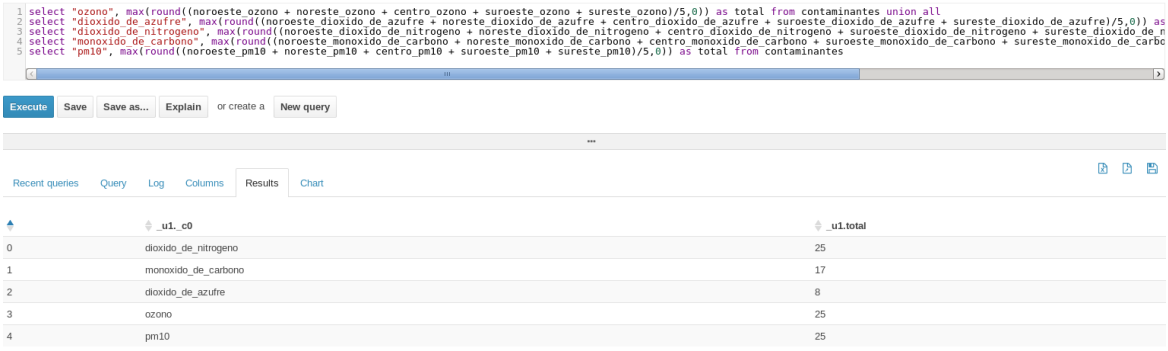


Figura 7. Consulta de niveles de contaminación.

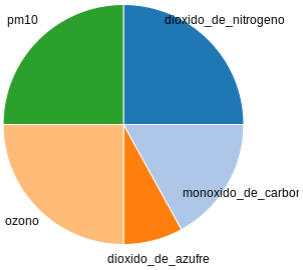


Figura 8. Gráfica de niveles de contaminación.

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

CONCLUSIONES

El tratamiento de datos semi-estructurados con formato JSON en el Sistema de Archivos Distribuidos Hadoop, presenta ventajas para su análisis al poder trabajarlos como datos estructurados. Siguiendo este tratamiento es posible realizar consultas tipo SQL, generando estadísticas y gráficas para describir la información de manera ágil. Con la aplicación del diseño transeccional descriptivo al caso de estudio de niveles de contaminación de la ciudad de México, se describieron las categorías de contaminación acorde a las zonas, empleando las herramientas de software libre proporcionadas por la distribución de Linux Cloudera basada en CentOS. Se propone como trabajos futuros el análisis de información de datos semi-estructurados recopilados por medio de sensores en la región frutícola del Estado de Yucatán, para su divulgación y aplicar mejoras en el tratamiento de los procesos. También se propone aplicar la metodología para datos no estructurados, con un enfoque en la generación de conocimiento que permita la aplicación en beneficio de la comunidad tecnológica y la sociedad del sur del estado de Yucatán.

BIBLIOGRAFÍA CONSULTADA

ÁLVAREZ, S.; GÉRTRUDIX, M. Y RAJAS, M.: «La construcción colaborativa de bancos de datos abiertos como instrumento de empoderamiento ciudadano», *Revista Latina de Comunicación Social*, Vol. 69, pp.661-683, 2014.

AZCÁRRAGA, J.: «Colaboraciones en Física», *Vida Científica*, Vol. 9, pp.1-12, 2016.

BIG DATA SAC.: *Análisis de Grandes Volúmenes de Información*. Disponible en http://www.bigdata.pe/. Visitado el 2 de octubre de 2016.

CAMARGO, J.; CAMARGO, F. Y JOYANES, L.: «Knowing the Big Data», *Revista Facultad de Ingeniería*, Vol. 24, pp.63-77, 2014.

HERNÁNDEZ, R.; FERNÁNDEZ, C. Y BAPTISTA, P.: *Metodología de la Investigación,* México: Mc Graw Hill, 2010.

POL, U.: «Big Data and Hadoop Technology Solutions with Cloudera Manager*», International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 4, pp.1028-1034, 2014.

RUSSOM, P.: *Big Data Analytics*, USA: TDWI (The Data Warehousing Institute), 2011.

Vol.6, No. Especial de CIVITEC,
octubre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

THE APACHE SOFTWARE FOUNDATION: *Apache Hadoop*. Disponible en http://hadoop.apache.org/. Visitado el 4 de octubre de 2016.

VASILENKO, D. Y& KURAPATI, M.: «Efficient Processing of XML Documents in Hadoop Map Reduce», *International Journal on Computer Science and Engineering*, Vol. 6, pp.329-333, 2014.