REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 1 • Nro. 2 • julio-septiembre 2020 • págs. 64-73

ARTÍCULO ORIGINAL

Geocodificación de direcciones postales cubanas en un entorno *Big Data*

Cuban Addresses Geocoding in a Big Data Environment

Ofir Alfonso Cantillo
ofir4991@gmail.com • https://orcid.org/0000-0003-24014565
Eduardo Sánchez Ansola
esanchezansola@gmail.com • https://orcid.org/0000-0001-5977-1633
UNIVERSIDAD TECNOLÓGICA DE LA HABANA "JOSÉ ANTONIO ECHEVERRÍA", CUJAE, CUBA

Recibido: 2019-12-19 • Aceptado: 2020-03-17

RESUMEN

Debido a la gran cantidad de datos digitales con que se cuenta, se ha convertido en una necesidad la creación de sistemas capaces de procesarlos. La información espacial no queda exenta de esto y debido a que gran parte de esta se encuentra en forma de direcciones postales los procesos de geocodificación han adquirido una gran importancia. La geocodificación es el proceso de convertir una dirección postal en coordenadas geográficas. Una de las aproximaciones en la actualidad para procesar una gran cantidad de datos son las tecnologías asociadas a *Big data.* Un problema se considera *Big data* cuando los datos a procesar se convierten en un inconveniente para los sistemas actuales. El objetivo del presente trabajo es diseñar un proceso de geocodificación para direcciones cubanas haciendo uso de un diseño *Big data.* Los resultados del presente trabajo demuestran que es factible el uso de tecnologías Big data para resolver el problema de la geocodificación pues se logra la disminución de los tiempos de respuesta con respecto a otros servicios existentes en la geocodificación de gran cantidad de direcciones postales.

Palabras clave: *Big data;* dirección postal; Geocodificación; *MapReduce.*

ABSTRACT

Due to the large amount of digital data available, the creation of systems capable of processing them has become a necessity. Spatial information is not exempt from this and because much of it is in the form of postal addresses, geocoding

processes have acquired great importance. Geocoding is the process of converting a postal address into geographic coordinates. One of the current approaches to process a large amount of data is the technologies associated with Big Data. Big Data is considered a problem when the data to be processed becomes inconvenient for current systems. The objective of this paper is to design a geocoding process for Cuban addresses using a Big Data design. The results of this article demonstrate that it is feasible to use Big Data technologies to solve the problem of geocoding because the reduction of response times is achieved compared to other services when geocoding a large number of postal addresses.

Keywords: *Big Data; postal address; address geocoding; MapReduce.*

INTRODUCCIÓN

El uso de la información geográfica en el día a día ha tomado un auge significativo en los últimos años. Servicios populares de Internet, como *Google Maps, Bing Maps* y *OpenStreetMap,* que tienen como base el uso de datos geográficos de diversas fuentes, han provisto a los usuarios de herramientas atractivas para la consulta y manipulación de información espacial.

Los datos geográficos son la clave para diferenciar un Sistema de Información Geográfica (SIG) de otro Sistema de información. Los SIG son un potente conjunto de herramientas para recolectar, almacenar, recuperar a voluntad, transformar y presentar datos espaciales procedentes del mundo real (Burrough, 1986). Una de las funcionalidades que ofrecen los SIG es la geocodificación de direcciones postales.

La geocodificación es un proceso fundamental que ayuda a organizaciones a refinar y enriquecer información existente relacionada con direcciones y localizaciones en una base de datos. Esta genera una latitud/longitud a partir de una dirección postal (Oracle, 2007). Este proceso puede ser llevado a cabo no solo para una dirección, sino para lotes de direcciones. Un lote puede estar compuesto por una gran cantidad de direcciones.

Son varias las entidades cubanas que han realizado sistemas donde intervienen procesos de geocodificación. Entre 2011 (Cruz, 2011) y 2013 (de Armas & Cruz, 2013) se presentó en el Complejo de Investigaciones Tecnológicas Integradas (CITI) un sistema de geocodificación nacional el cual fue evolucionando hasta convertirse en una funcionalidad incluida en *GeoServer* mediante servicios Web en el año 2017 (Girón & Sánchez, 2017).

Alfonso (Alfonso, Sánchez, & Guerra, 2018) presenta una mejora a este sistema donde se utilizó las ventajas de la computación paralela para disminuir el tiempo de geocodificación de direcciones cubanas. Los experimentos presentados comprobaron que se cumplió con el objetivo del sistema, pero aun así los autores consideraron que el sistema no sería capaz de procesar una gran cantidad de datos.

Estos métodos "tradicionales" para realizar el proceso de geocodificación pueden llegar a demorar tanto en un proceso de geocodificación por lotes que su utilización no sea factible en determinadas circunstancias. En (Xu, Flexner, & Carvalho, 2012) se realiza un estudio comparativo entre diferentes servicios geocodificadores disponibles en Internet (por ejemplo *Google Maps, Bing* y *Yahoo Place Finder)* donde se geocodificaron diferentes bases de datos de direcciones de los Estados Unidos de América, entre ellas una con un millón de instancias, y su geocodificación completa duró poco más de un año. El tiempo que duró la geocodificación fue, de acuerdo a los autores, muy superior a lo que necesitaban para cubrir las necesidades de geocodificación que tenían, por lo que tuvieron que desarrollar un sistema que pudiera procesar este volumen de datos. Como resultado del trabajo, los autores ofrecen un sistema de geocodificación de direcciones postales estadounidenses para un entorno de datos masivos o *Big data*.

El objetivo del presente trabajo es disminuir el tiempo del proceso de geocodificación de lotes de direcciones postales cubanas. El trabajo se encuentra enfocado a resolver las necesidades de geocodificación de un usuario con grandes cantidades de direcciones.

Se plantea la utilización de *Apache Spark* como la plataforma para el desarrollo de algoritmos *Big data*. Esta plataforma está diseñada para ser rápida debido a que extiende el popular modelo *MapReduce* para brindar soporte eficiente a un sinfín de operaciones computacionales sobre datos y tiene la capacidad de ejecutar cálculos en memoria, lo que lo hace más eficiente (Holden, Konwinski, Wendell, &Zaharia, 2015).

METODOLOGÍA

Estado actual de Big data

No existe una definición rigurosa acerca de lo que puede ser considerado como *Big data* o Datos Masivos (Kenneth & Mayer-Schonberger, 2013). Se refiere a la idea de que los datos generados y almacenados por una empresa o aplicación no pueden ser manejados por los sistemas tradicionales de procesamiento de datos (Baesens, 2014).

Según (O'Reilly, 2017) un problema se considera *Big data* cuando las características en que se presentan los datos los transforma a ellos mismos en una parte del problema a resolver.

La empresa *Google* fue una de las primeras en afrontar los problemas que traía consigo la masividad de los datos debido a que querían obtener información de la totalidad de las páginas disponibles en Internet para su posterior indexado y con ello dar soporte a la consulta de información en su conocido buscador. Para ello se basaron el principio "divide y vencerás" y crearon el paradigma *MapReduce* (Aboul-Ella, *et al.*, 2015).

MapReduce permite el procesamiento de datos en paralelo. Consta de dos etapas: la etapa *Map* (mapeo) es la que trabaja sobre partes distribuidas de un conjunto de datos en varios elementos de procesamiento que trabajen en paralelo. Los resultados de este proceso en cada nodo son clasificados y enviados a la siguiente fase nombrada *Reduce* (reducción). En esta segunda etapa se combinan los resultados de acuerdo a las necesidades del negocio y se le ofrece al usuario un resultado al problema inicial (Srinath&Gunarathne, 2013).

Geocodificación en Big Data

Las aplicaciones *Big data* donde intervienen procesos de geocodificación abundan en la literatura o el mercado. La Salud es una de las esferas donde más se han utilizado estas técnicas. En (Sen, *et al.,* 2012) se describe el proceso de desarrollo de un geocodificador de direcciones postales para un grupo de instituciones de salud en los Estados Unidos de América. Utilizan el sistema con el objetivo de localizar en el mapa las áreas más afectadas con determinada enfermedad para con ello tomar las medidas necesarias para su prevención en caso de que fuera posible. Experimentan con servicios geocodificadores comerciales de empresas como *Yahoo!* y *Google,* pero el tiempo que demoraría todo el proceso es muy superior a sus necesidades por lo que deciden desarrollar un sistema geocodificador para un entorno *Big data* utilizando *Apache Spark.* De acuerdo a los experimentos desarrollados, el tiempo de procesamiento para un millón de direcciones se redujo de unos cuantos meses (tiempo que demoraría utilizando las licencias públicas de *Google* o *Yahoo!)* a unos días utilizando el sistema *Big data* desarrollado.

La masividad de datos no solo se encuentra en las direcciones de entradas en un proceso de geocodificación. En (Xiang, Kardes, Wang, & Ang, 2014) se presenta un sistema de extracción de componentes de una dirección para datos masivos. El sistema que desarrollaron los autores es capaz de identificar cada elemento de la dirección con un 95.6% de precisión debido a que utilizaron 100 millones de direcciones en el proceso de entrenamiento de su sistema probabilístico para el reconocimiento de componentes dentro de una dirección postal.

Según (Teerayut, *et al.,* 2014), la geocodificación inversa también ha sido objeto de estudio debido a la incapacidad de muchos sistemas de afrontar la masividad de datos que se necesita procesar en el día de hoy. Los autores del artículo crearon un marco de trabajo para la geocodificación inversa de millones de puntos en el espacio con el objetivo de relacionar personas con actividades en una zona geográfica determinada.

Los *software*s comerciales encargados de geocodificar millones de direcciones también están ocupando un importante lugar en el mercado. Este es el caso de la herramienta *PitneyBowesSpectrum™ Geocodingfor Big data* desarrollado por la compañía *PitneyBowes Inc.* con el objetivo de ofrecer a sus clientes un sistema de geocodificación directa e inversa capaz de procesar un fichero que contenga millones elementos. Utilizan A*pache Spark* y *Hive* como marco de trabajo para el procesamiento en un clúster de computadoras y utilizan el sistema de ficheros distribuidos de *Hadoop* para el almacenamiento de los datos necesarios en el proceso de geocodificación de las direcciones postales. La herramienta no se encuentra distribuida de forma libre y solamente permite la geocodificación de direcciones en Estados Unidos de América (Pitney, 2019).

Propuesta de solución

La geocodificación de direcciones postales cuenta con cuatro etapas divididas en dos fases fundamentales. En una primera fase se extraen los elementos necesarios para el proceso de geocodificación a partir de una dirección postal escrita en formato libre y en la segunda fase se procesa cada componente de la dirección para encontrar su posición geográfica. El diseño

para el algoritmo *Geocoding_Big data* cuenta con dos etapas del patrón *MapReduce* que corresponden a cada fase definida en un proceso de geocodificación. El diagrama de actividades (figura 1) muestra el diseño del algoritmo.

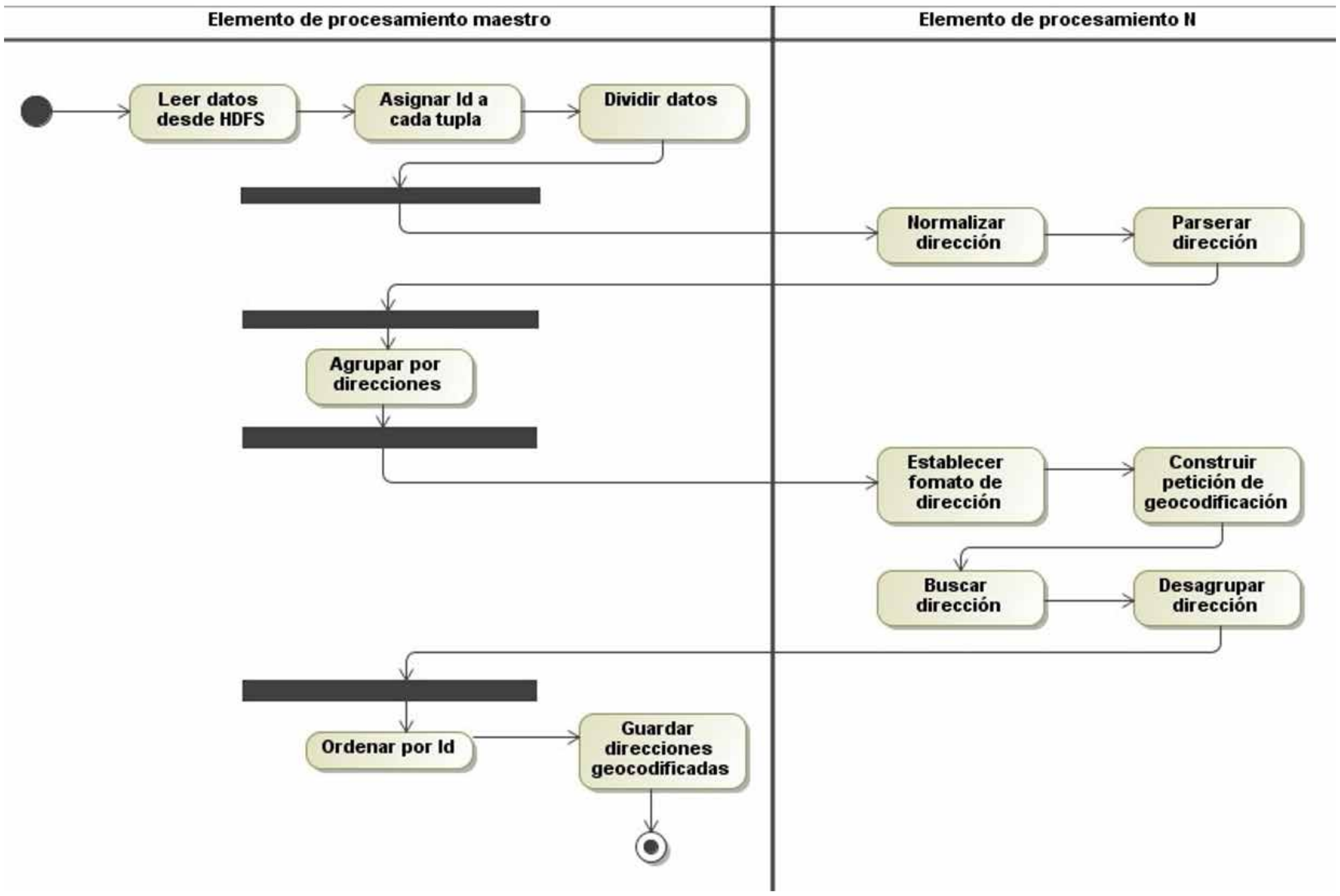

Figura 1. Diagrama de actividades del algoritmo *Geocoding_Big Data.*

Para la primera fase se ejecutan los siguientes procedimientos:

1. Asignar un identificador a cada dirección.
2. Dividir los datos de entrada (direcciones postales escritas en texto libre) de la forma más homogénea posible.
3. Normalizar cada una de las direcciones.
4. Para cada dirección en cada partición encontrar sus componentes.
5. Agrupar las direcciones por sus componentes.

Se propone realizar los procedimientos 3 y 4 en una fase *Map* donde se obtendrían los elementos necesarios para el proceso de geocodificación. Seguidamente agrupar, en una función *Reduce,* las direcciones comunes con el objetivo de geocodificar únicamente una dirección por cada grupo y evitar un procesamiento de datos innecesarios. De esta forma también se identificarían las direcciones que no pudieron ser procesadas por contener errores significativos en su escritura. Estas direcciones no son enviadas a la segunda fase con lo que se ahorra tiempo en el proceso transferencia de datos. Esta segunda fase contiene los procesos siguientes:

1. Establecer el tipo de dirección a geocodificar.
2. Construir petición de geocodificación.

3. Encontrar dirección con la mayor precisión posible.
4. Desagrupar direcciones.
5. Ordenar direcciones por el identificador.

Para la segunda fase se propone realizar los tres primeros pasos en una fase *Map* con el objetivo de encontrar la posición geográfica de cada dirección que contiene la partición o, en caso de no ser encontrada, notificar que la dirección no pudo ser geocodificada. Se complementa esta etapa con una función *Reduce* encargada de asignarle a cada dirección en un grupo la respuesta correspondiente y organizarlas para ofrecerle la respuesta al usuario en el mismo orden en que fue recibida la petición.

La figura 2 muestra la vista de la arquitectura del algoritmo *Geocoding BigData* separado por capas.

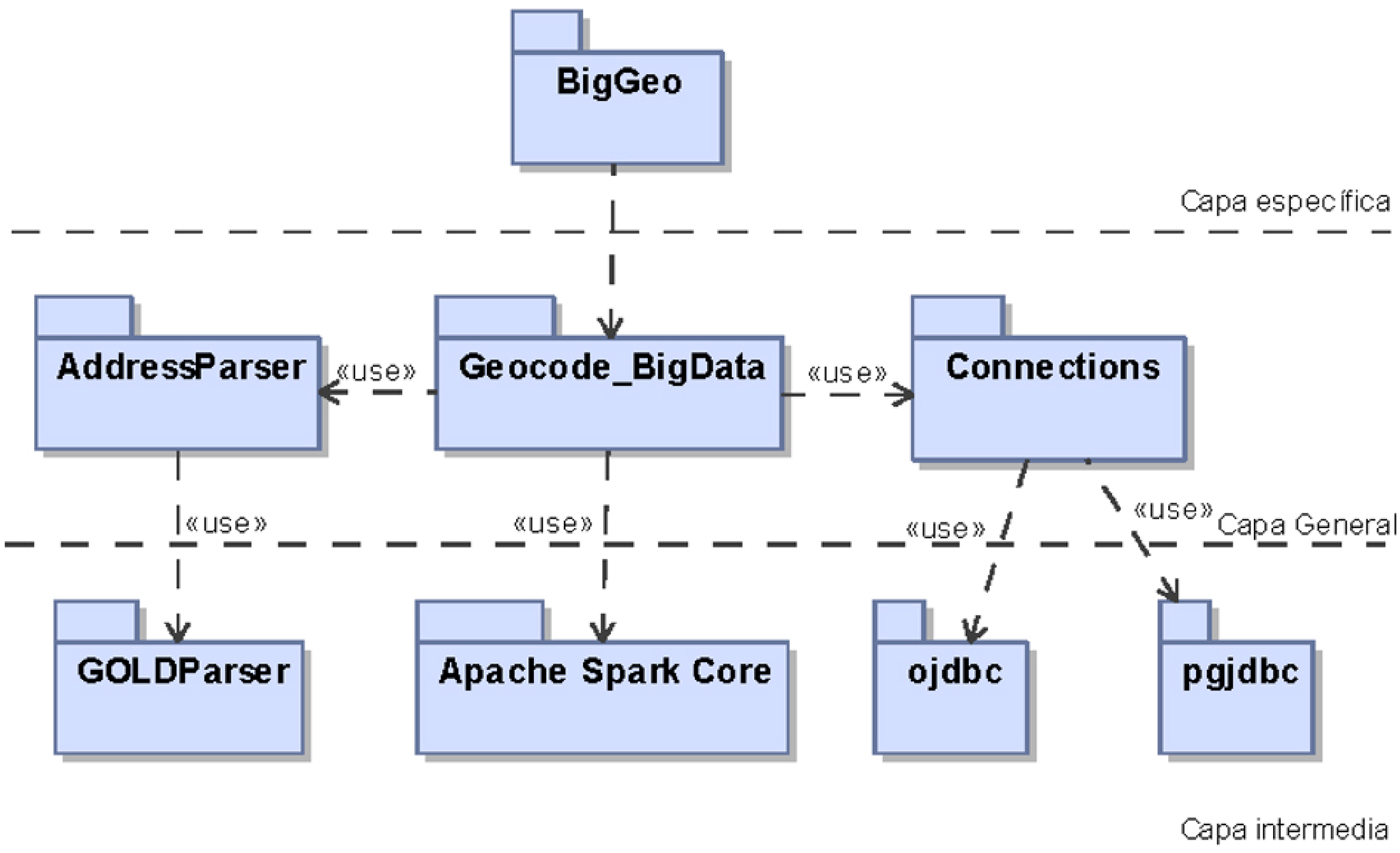

Figura 2. Vista de la arquitectura desde un punto de vista de reutilización.

El paquete *"geocode_BigData"* contiene todas las clases y objetos necesarios para el proceso de geocodificación y es el contenedor de la lógica fundamental del algoritmo. El paquete "BigGeo" es el sistema que utiliza el algoritmo de geocodificación de direcciones postales cubanas en un entorno *Big data*. La aplicación se desarrolló fundamentalmente en lenguaje PHP con el marco de trabajo *Yii2*. Se utilizaron versiones específicas de la plataforma *Apache Spark* y del sistema de ficheros distribuidos de *Hadoop* (HDFS) las cuales cuentan con interfaces REST para la comunicación del clúster de computadoras con la aplicación.

En la figura 3 se muestra el diagrama de despliegue del sistema encargado de ejecutar el algoritmo *Geocoding_BigData* en un clúster de computadoras.

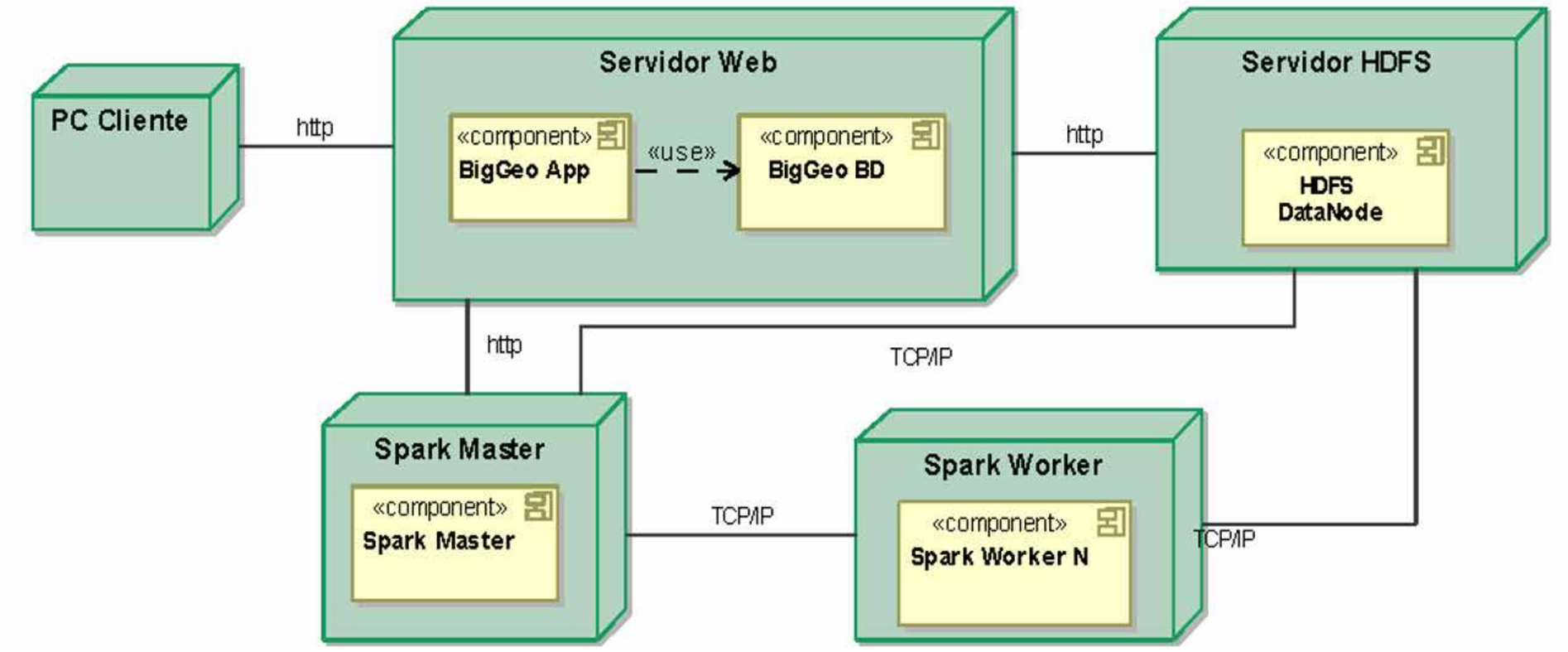

Figura 3. Diagrama de despliegue de la solución.

RESULTADOS Y DISCUSIÓN

El algoritmo *Geocoding_BigData* fue probado en un clúster heterogéneo con la configuración mostrada en la tabla 1.

Tabla 1. Características del ambiente de prueba.

CPU	RAM	HDD	SO	Nodos
Intel (R) Core (TM) i5-4200U @ 1.60HGz 2.30 GHz	8GB	1 TB	Windows 10	1
Intel (R) Core (TM) i7 @ 2.65 HGz 2.67 GHz	6 GB	1 TB	Windows 8.1	2
Intel (R) Core (TM) i5-5200U @ 2.20HGz 2.20 GHz	4 GB	1 TB	Windows 10	1
Intel (R) Core (TM) Duo CPU E730 @ 2.66 GHz 2.67 GHz	4 GB	500GB	Windows 7	1

El nodo máster del clúster fue configurado con 4 núcleos y 8 GB de RAM. Los nodos esclavos suman un total de 30 núcleos y 20 GB de memoria RAM.

Un lote de 10 mil direcciones postales cubanas fue geocodificado utilizando el algoritmo *Geocoding_BigData* y el sistema geocodificador implementado utilizando el servidor de mapas *GeoServer.* El tiempo de procesamiento de cada sistema se documentó en la tabla 2.

Con el objetivo de encontrar cuál sistema mejora el tiempo de geocodificación y establecer comparaciones entre ellas de acuerdo a los resultados obtenidos, se utilizarán pruebas estadísticas.

Tabla 2.

Tiempo de ejecución de cada variante del Sistema.

Petición	Geoserver (min)	Geocoding_Big Data (min)
1	113,78374	18,1409
2	127,7128	16,0125
3	107,37154	16,1899
4	108,37512	17,088
5	125,82375	15,3034
6	129,94275	16,8933
7	115,75412	26,9182
8	108,74985	17,3355
9	119,37914	15,1072
10	106,16754	15,0839

Dentro de las pruebas de hipótesis se encuentra la prueba de Kruskal-Wallis, la cual es utilizada para determinar si las medianas de dos o más muestras difieren. Esta prueba será utilizada para comprobar si existen diferencias entre los tiempos de ejecución de cada una de las variantes del algoritmo. Para determinar si existen diferencias entre los tiempos de ejecución de las variantes implementadas se definen como hipótesis las siguientes:

- H0: Los sistemas son iguales en cuanto a tiempo de respuesta.
- H1: Los sistemas son diferentes en cuanto a su tiempo de respuesta.

Donde H0 es la hipótesis nula y H1 es la hipótesis alternativa. Se utiliza un α=0,05 (nivel de significación). En la figura 4 se muestra el resultado de realizar la prueba usando el *software Minitab 16.* Para ello fueron utilizados los tiempos de ejecución de la tabla 2.
Al realizar la prueba se obtuvo un valor *p*=0,000 como resultado de un redondeo. Como este valor es menor que el α se rechaza la hipótesis nula y se acepta la hipótesis alternativa, por lo que se puede concluir que los sistemas difieren en cuanto al tiempo de respuesta.

Una vez comprobado que los tiempos de respuesta de los sistemas difieren entre sí se pasa a comprobar que el algoritmo *Geocoding_BigData* disminuye el tiempo de geocodificación de direcciones postales. La demostración se realizará apoyándose de la prueba de hipótesis Mann-Whitney. Para ello se define un α = 0.05 y se definen como hipótesis las siguientes:

- H0: Los sistemas son iguales en cuanto a tiempo de respuesta.
- H1: El sistema geocodificador en *GeoServer* presenta tiempos de ejecución mayores que el algoritmo *Geocoding_BigData*.

El resultado de la prueba se muestra en la figura 5.

Al realizar la prueba se obtuvo un valor *p*=0,0001, que es menor que el α = 0.05 por lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa, por lo que se puede concluir que algoritmo *Geocoding_BigData* reduce el tiempo de geocodificación de direcciones postales cubanas con respecto a la variante implementada en *GeoServer.*

Prueba de Kruskal-Wallis: Tiempo vs. Sistema

Prueba de Kruskal-Wallis en Tiempo

Sistema	N	Mediana	Clasificación del promedio	Z
Geocoding_BigData	10	16.54	5.5	-3.78
GeoServer	10	114.77	15.5	3.78
General	20		10.5	

H = 14.29 GL = 1 P = 0.000

Figura 4. Resultado de la ejecución de la prueba de Kruskal-Wallis.

Prueba de Mann-Whitney e IC: GeoServer; Geocoding_BigData

	N	Mediana
GeoServer	10	114.77
Geocoding_BigData	10	16.54

La estimación del punto para ETA1-ETA2 es 98.10
95.5 El porcentaje IC para ETA1-ETA2 es (91.29;108.93)
W = 155.0
Prueba de ETA1 = ETA2 vs. ETA1 > ETA2 es significativa en 0.0001

Figura 5. Resultado de la ejecución de la prueba Mann-Whitney.

CONCLUSIONES

Las necesidades de geocodificación de direcciones postales han crecido tanto en la actualidad que se ha convertido en una necesidad la creación de sistemas *Big data* capaces de procesar ese flujo de datos. El algoritmo *Geocoding_BigData* fue implementado utilizando el paradigma *Ma-*

pReduce sobre la plataforma *Apache Spark.* Su diseño permitió la reducción del tiempo de geocodificación de direcciones postales cubanas en comparación con otros servicios existentes. No obstante, se hace necesario la creación de un sistema capaz de procesar grandes cantidades de direcciones postales provenientes de un número significativamente alto de usuarios. Asimismo, con el fin de validar completamente la solución propuesta es recomendable desplegar la misma en un entorno real en la nube o al menos con una configuración completamente homogénea.

REFERENCIAS

Aboul-Ella Hassanien, Ahmad Taher Azar, Vaclav Snasel, Janusz Kacprzyk, & Jemal H. Abawajy. (2015). *Big data* in Complex Systems: Challenges and Opportunities: Springer International Publishing.

Alfonso Cantillo, O., Sánchez Ansola, E., & Guerra Denis, L. (2018). Geocodificación de direcciones postales cubanas utilizando computación paralela. Paper presented at the IV Congreso internacional de ingeniería informática y sistema de información, La Habana, Cuba.

Baesens, B. (2014). Analytics in a *Big data* World: The Essential Guide to Data Science and its Applications: Wiley India.

Burrough, P. A. (1986). Principles of geographical information systems for land resources assessment (Vol. 12). New York: Clarendon Press.

Cruz Gutiérrez, A. A. (2011). Servicio de Geocodificación. Implementación en una Infraestructura de Datos Espaciales. Paper presented at the VII Congreso Internacional GEOMÁTICA 2011, La Habana.

de Armas García, C. J., & Cruz Gutiérrez, A. A. (2013). Deployment of a National Geocoding Service: Cuban Experience. URISA Journal, 25.

Girón Lima, L., & Sánchez Ansola, E. (2017). Servicios Basados en la Localización para GeoServer. Paper presented at the VIII Convención Agrimensura, La Habana.

Holden K., Konwinski,A., Wendell P., & Zaharia M., (2015). Learning Spark, lightning-fast data analysis. (First ed.): O'Reilly Media.

Kenneth Cukie, & Viktor Mayer-Schonberger. (2013). *Big data*. A Revolution That Will Transform How We Live, Work, and Think. United Kingdom: Eamon Dolan/Houghton Mifflin Harcourt.

O'Reilly Media, I. (2017). *Big data* Now. United States of America: O'Reilly Media, Inc.

Oracle. (2007). Oracle Spatial 11g: Administración Avanzada de Datos Espaciales para Aplicaciones Empresariales. United States of America.

Pitney Bowes, I. (2019). Spectrum™ Geocoding for *Big data* User Guide. In P. B. S. Inc. (Ed.), (3.2 ed.). United States of America.

XuSen, FlexnerSoren, & CarvalhoVitor. (2012, Septembre 18-12, 2012). Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with *Big data*. Paper presented at the GIScience in the *Big data* Age (GIScience 2012), Colombus, OH, USA.

Srinath Perera, & GunarathneThilina. (2013). Hadoop *MapReduce* Cookbook. Birmingham: Packt Publishing.

Teerayut Horanont, Jiranuwat Prapakornpilai, Santi Phithakkitnukoon, Apichon Witayangkurn, & Ryosuke Shibasaki. (2014). Space Profile-Based Reverse Geocoding Service Using Cloud Platform. Paper presented at the 2014 IEEE International Conference on Services Computing, Anchorage, AK, USA.

Xiang Li, Hakan Kardes, Xin Wang, & Ang Sun. (2014). HMM-based Address Parsing with Massive Synthetic Training Data Generation. Paper presented at the Proceedings of the 4th International Workshop on Location and the Web, Shanghai, China.

Copyright © 2020 Alfonso-Cantillo, O., Sánchez-Ansola, E.

Este obra está bajo una licencia de Creative Commons Reconocimiento 4.0 Internacional.