Análisis comparativo de técnicas de visualización para Big Data.

Maestría en Sistemas Computacionales — Instituto Tecnológico de Orizaba
ISC. Luis Eder Velázquez Peña — levp8915@hotmail.com
Dra. Lisbeth Rodríguez Mazahua — lrodriguez@itorizaba.edu.mx

Objetivo: Realizar un análisis comparativo entre las diferentes técnicas de visualización para determinar cuáles son las más adecuadas para realizar un análisis de Big Data.

Introducción

El mundo está viviendo la denominada era de la información, donde grandes cantidades de datos se generan diariamente ya sea para su análisis o simplemente para su almacenamiento. Estos nuevos datos se engloban en el término de Big Data, el cual cobró fuerza en los últimos 20 años.

El análisis visual representa una gran oportunidad para descubrir patrones o tendencias en los datos, pero se vuelve una tarea desafiante si se realiza en Big Data, es por esto que el objetivo principal de esta investigación es proporcionar la suficiente información sobre las diferentes técnicas de visualización que se utilizan en Big Data y así realizar un análisis comparativo de las mismas para determinar sus ventajas y desventajas, esto mediante un marco de evaluación propuesto.

Metodología

Se realizó una exhaustiva revisión de trabajos relacionados con el análisis visual en Big Data, con el fin de determinar las técnicas de visualización y conjuntos de datos más utilizados.

Las técnicas implementadas en este análisis, determinadas después de la revisión de la literatura, fueron las gráficas de líneas, de barras, de dispersión, mapas de calor (*Heatmap*), mapas de árbol (*Treemap*), clustergram, coordenadas paralelas y *Spatial Information Flow*.

Datos como exportaciones e importaciones de países, flujos de clics, calificaciones de productos y calificaciones de películas, fueron utilizados para estas técnicas de visualización.

El marco de evaluación contempla puntos como número de variables, número de tuplas, curva de aprendizaje, tiempo de construcción, tiempo de visualización, facilidad de interpretación, volumen, velocidad, variedad y veracidad.

Todo esto se hizo mediante el uso de herramientas necesarias para el manejo de Big Data, como lo es Hadoop y su sistema de archivos [1], Apache Pig como herramienta de análisis [2], la plataforma Cloudera que integra estas dos herramientas [3] y R para la visualización de los datos [4].

Resultados

Usando las herramientas antes descritas, se implementaron cada una de las técnicas de visualización sobre los conjuntos de datos, por ejemplo, en la Figura 1 se aprecia la implementación de *Spatial Information Flow* sobre datos de exportaciones de los principales países en los últimos 50 años. Después de todas las implementaciones, se realizó un cuadro comparativo de acuerdo al marco de evaluación, de esta forma, el investigador evaluará los resultados y determinará qué técnicas son las más apropiadas para visualizar Big Data.

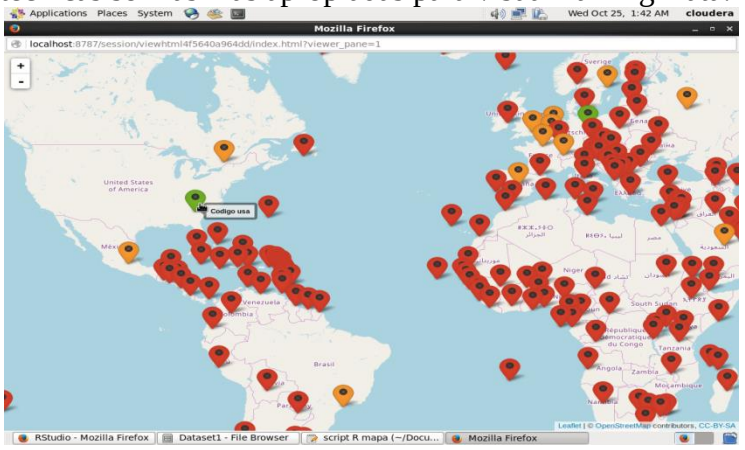

Figura 1 Gráfica de *Spatial Information Flow* implementada.

Conclusiones

Dada la poca información que existe respecto a la comparativa de técnicas de visualización en Big Data, esta investigación proporciona la información necesaria para determinar las ventajas y desventajas de aplicar dichas técnicas de visualización en conjuntos de datos Big Data.

Referencias

[1] "Welcome to Apache™ Hadoop®!", 20-jul-2016. [En línea]. Disponible en: http://hadoop.apache.org/. [Consultado: 20-jul-2016].
[2] "Welcome to Apache Pig!" [En línea]. Disponible en: https://pig.apache.org/. [Consultado: 24-nov-2017].
[3] "Download QuickStarts for CDH 5.12 | Cloudera". [En línea]. Disponible en: https://www.cloudera.com/downloads/quickstart_vms/5-12.html. [Consultado: 24-nov-2017].
[4] "R: The R Project for Statistical Computing", 16-oct-2016. [En línea]. Disponible en: https://www.r-project.org/. [Consultado: 16-oct-2016].