SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCIÓN I.- TESIS

TRABAJO PROFESIONAL

"ANÁLISIS COMPARATIVO DE TÉCNICAS DE VISUALIZACIÓN PARA BIG DATA".

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:

I.S.C. Luis Eder Velázquez Peña

DIRECTOR DE TESIS:

Dra. Lisbeth Rodríguez Mazahua

CODIRECTOR DE TESIS:

Dr. Giner Alor Hernández

INSTITUTO TECNOLÓGICO DE ORIZABA

ORIZABA, VERACRUZ, MÉXICO. **ABRIL 2018**

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

FECHA: 18/04/2018
DEPENDENCIA: POSGRADO
ASUNTO: Autorización de Impresión
OPCIÓN: I

C. **LUIS EDER VELAZQUEZ PEÑA**
CANDIDATO A GRADO DE MAESTRO EN:
SISTEMAS COMPUTACIONALES

De acuerdo con el Reglamento de Titulación vigente de los Centros de Enseñanza Técnica Superior, dependiente de la Dirección General de Institutos Tecnológicos de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

"ANALISIS COMPARATIVO DE TECNICAS DE VISUALIZACION PARA BIG DATA".

Comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

A T E N T A M E N T E

RUBEN POSADA GOMEZ
JEFE DE LA DIV. DE ESTUDIOS DE POSGRADO

C.A. TITULACIÓN
SECRETARIA DE EDUCACIÓN PÚBLICA
INSTITUTO TECNOLÓGICO DE ORIZABA

ggc

Avenida Oriente 9 Núm. 852, Colonia Emiliano Zapata, C.P. 94320 Orizaba, Veracruz, México
Teléfonos: (272) 7 24 40 96 Fax. (272) 7 25 17 28
e- mail: Orizaba@itorizaba.edu.mx www.itorizaba.edu.mx

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Subsecretaría de Educación Superior
Tecnológico Nacional de México
Instituto Tecnológico de Orizaba

FECHA : 02/03/2018

ASUNTO: **Revisión de Trabajo Escrito**

C. RUBEN POSADA GOMEZ
JEFE DE LA DIVISION DE ESTUDIOS
DE POSGRADO E INVESTIGACION.
P R E S E N T E

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C. :

LUIS EDER VELAZQUEZ PEÑA

la cual lleva el título de:

"ANALISIS COMPARATIVO DE TECNICAS DE VISUALIZACION PARA BIG DATA".

Y concluyen que se acepta.

A T E N T A M E N T E

PRESIDENTE : **DRA. LISBETH RODRIGUEZ MAZAHUA** FIRMA

SECRETARIO : **DR. GINER ALOR HERNANDEZ** FIRMA

VOCAL : **MCE BEATRIZ ALEJANDRA OLIVARES** FIRMA

VOCAL SUP. : **M.C. SILVESTRE GUSTAVO SERGIO PELAEZ** FIRMA

EGRESADO(A) DE LA MAESTRIA EN **SISTEMAS COMPUTACIONALES**

OPCION: **I** **Tesis**

Avenida Oriente 9 Núm. 852, Colonia Emiliano Zapata, C.P. 94320 Orizaba, Veracruz, México
Teléfonos: (272) 7 24 40 96 Fax. (272) 7 25 17 28
e- mail: Orizaba@itorizaba.edu.mx www.itorizaba.edu.mx

Agradecimientos

Este trabajo representa un objetivo y un reto tanto personal como profesional que he alcanzado en mi vida, gracias a la colaboración de diferentes personas a las cuales les quiero dedicar las siguientes palabras.

Antes que nada debo a gradecer a Dios, por brindarme la sabiduría necesaria para la elaboración de este proyecto y permitir que me rodee de grandes personas que contribuyeron a su manera a alcanzar este objetivo.

Agradezco a mis padres por brindarme la vida y su inmenso apoyo. A mi madre Alejandra Peña por estar siempre ahí conmigo y ser una persona maravillosa en todo momento. A mi padre José Luis Velázquez por sus consejos y enseñanzas. A mi hermana Vannia Velázquez por los buenos y malos momentos que hemos vivido a lo largo de todo este tiempo.

También agradezco de manera especial a la Dra. Lisbeth por ser la guía en este camino, respetar siempre mi opinión y brindarme su apoyo en todo momento para la elaboración de esta investigación.

Agradezco al jurado revisor de este proyecto, al Dr. Giner, a la Maestra Beatriz y al Maestro Camarena por sus consejos y observaciones, para entregar de la mejor manera esta investigación de calidad.

De igual manera agradezco al Dr. José María, por darme la oportunidad de realizar mi estancia en la UC3M, brindarme su asesoría y consejos que fueron vítales para la investigación.

Quiero agradecer a todo el consejo de la Maestría en Sistemas Computacionales por todas las enseñanzas que han dejado en mí y siempre estar comprometidos en su trabajo.

También agradezco a mi compañero y amigo de toda la vida Samuel, por todas las experiencias que hemos vivido a lo largo de todo este tiempo, por su sincera amistad y su incondicional apoyo.

A mis compañeros de clase Rodolfo y Josefina, porque compartimos esta experiencia juntos y además descubrí en ellos a unas excelentes personas.

Finalmente, agradezco al Instituto Tecnológico de Orizaba y al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme las facilidades necesarias para la elaboración de este proyecto.

“La innovación distingue a los líderes de los seguidores”
Steve Jobs

Índice general

Índice de tablas

Índice de figuras

Resumen

Big Data se refiere a los grandes conjuntos de datos que en su tamaño sobrepasan la habilidad de las herramientas típicas de bases de datos para capturar, almacenar, administrar y analizar. Este término cobró fuerza en los últimos años debido a la creciente explosión en la generación y almacenamiento de información. *Big Data* representa una gran oportunidad de extraer información útil y el análisis visual proporciona las técnicas necesarias para el entendimiento de dichos datos. Dado lo anterior, el presente proyecto de tesis tiene como objetivo realizar un análisis comparativo de las diferentes técnicas de visualización para *Big Data* y así determinar cuáles son las mejores para los conjuntos de datos masivos, esto mediante un marco de evaluación propuesto que permitirá conocer las ventajas y desventajas de dichas técnicas.

Dado que el análisis visual de *Big Data* es un campo nuevo, la actual literatura no maneja algún marco referencial que ayude a determinar qué técnica de visualización aplicar a un conjunto de datos *Big Data*, por lo que el presente proyecto representa una gran oportunidad de proveer la información necesaria a los investigadores y beneficiarlos en la reducción del tiempo y esfuerzo que cada un aplica en el análisis de *Big Data*.

Se realizó una exhaustiva revisión del estado del arte para conocer qué técnicas de visualización son las que más se utilizan para el análisis de *Big Data*, así como los conjuntos de datos más utilizados. Se implementaron dichas técnicas y se evaluaron mediante el marco de evaluación propuesto (número de variables, número de tuplas, curva de aprendizaje, tiempo de construcción, tiempo de visualización, facilidad de interpretación, volumen, velocidad, variedad y veracidad) para determinar sus ventajas y desventajas. Todo esto mediante la instalación y uso de las herramientas necesarias para el manejo de *Big Data*, como lo es Hadoop y su sistema de archivos para el almacenamiento de grandes datos, Apache Pig como herramienta de análisis, la plataforma Cloudera que integra estas dos herramientas y el software estadístico R para la visualización de los datos.

Abstract.

Big Data refers to large data sets that in their size exceed the ability of typical database tools to capture, store, manage and analyze. This term has gained strength in recent years due to the growing explosion in the generation and storage of information. Big Data represents a great opportunity to extract useful information and the visual analysis provides the necessary techniques for the understanding of such data. Given the above, the present thesis project aims to perform a comparative analysis of the different visualization techniques for Big Data and thus determine which are the best for the massive data sets, this through a proposed evaluation framework that will allow to know the advantages and disadvantages of these techniques.

Given that the visual analysis of Big Data is a new field, the current literature does not handle any referential framework that helps to determine which visualization technique to apply to a Big Data dataset, so the present project represents a great opportunity to provide the necessary information to the investigators and benefit them in the reduction of the time and effort that each one applies in the Big Data analysis.

An exhaustive review of the state of the art was carried out to know which visualization techniques are the most used for the Big Data analysis, as well as the most used data sets. These techniques were implemented and evaluated through the proposed evaluation framework (variables number, tuples number, learning curve, construction time, visualization time, interpretability, volume, velocity, variety and veracity) to determine their advantages and disadvantages. All this through the installation and use of the necessary tools for managing Big Data, such as Hadoop and its file system for the storage of large data, Apache Pig as an analysis tool, the Cloudera platform that integrates these two tools and the statistical software R for the visualization of the data.

Introducción

Actualmente el mundo está viviendo la denominada era de la información, donde grandes cantidades de datos se generan y recolectan diariamente tanto para su almacenamiento como su análisis. Todos estos grandes conjuntos de datos se engloban en el término de *Big Data*, el cual cobró fuerza en los últimos 20 años. *Big Data* representa una nueva oportunidad para descubrir información que sirva a las empresas e investigadores en las diferentes áreas que se desempeñan y así ayudar en la toma de decisiones.

Debido a que el análisis de *Big Data* es un campo relativamente nuevo, el análisis visual surgió como una herramienta para descubrir patrones y/o relaciones en los datos que antes eran difíciles de percibir u observar. Esto es consecuencia de que el análisis visual se ayuda de las capacidades de percepción humana, por lo que es más fácil encontrar información útil al momento de representar los datos de manera gráfica.

Al ser un campo nuevo el análisis visual, hoy en día no hay muchos reportes que ayuden a determinar cuáles son las mejores técnicas de visualización para los conjuntos de datos de *Big Data*, por lo que en este proyecto se realizó un análisis comparativo de dichas técnicas y, mediante un marco de evaluación, se determinaron sus ventajas y desventajas.

El presente documento se estructura en cinco capítulos para una mejor comprensión al lector. En el capítulo 1 se plasman los conceptos básicos necesarios para el entendimiento del proyecto de tesis, así como el planteamiento del problema, justificación, objetivo general y objetivos específicos que se alcanzaron en la presente investigación. El capítulo 2 muestra una revisión de los trabajos existentes relacionados con el tema de investigación y un análisis comparativo de la literatura mencionada, además de la propuesta de solución que se eligió. En el capítulo 3 se presenta el desarrollo de la metodología necesaria para el manejo de *Big Data* y el análisis de las técnicas de visualización y conjunto de datos más utilizados. El capítulo 4 expone el resultado de las visualizaciones para cada conjunto de datos y el análisis comparativo de las mismas. Finalmente, en el capítulo 5 se muestran las conclusiones y referencias de este trabajo.

Capítulo 1. Antecedentes

En este primer capítulo se presentan los conceptos más relevantes que se utilizaron a lo largo del proyecto de tesis y que representan la base para el desarrollo del mismo. Se da lugar a la problemática a resolver, los objetivos a alcanzar y la justificación de lo que se pretende realizar.

1.1. Marco teórico

En esta sección se presentan los conceptos relacionados con el tema de investigación.

1.1.1. Dato

De acuerdo al diccionario de la Real Academia Española [1], la palabra dato proviene del latín *datum* "lo que se da". Es un documento, una parte de información, testimonio o fundamento que permite llegar al conocimiento de algo o deducir las consecuencias legítimas de un hecho. Se refiere a información dispuesta de manera adecuada para su tratamiento mediante herramientas computacionales. De forma general, un dato es una representación simbólica de un atributo o variable cuantitativa o cualitativa. Aisladamente, un dato a veces no contiene información relevante, sólo cuando se examina un conjunto de datos a la luz de un enfoque, hipótesis o teoría se aprecia la información contenida en dichos datos.

1.1.2. Conjunto de datos

Es una colección de medidas tomadas de algún ambiente o proceso, también conocido como *"dataset"*. En el caso más simple es una colección de objetos, y para cada objeto se tiene un conjunto de medidas [2]. También se define como un grupo de información estructurada que responde a determinada cuestión de interés y que se representa como tabla o matriz, donde cada columna responde a una variable y cada fila representa un registro. En general, cada registro del conjunto de datos se caracteriza por los valores de sus atributos, que miden diferentes aspectos de la instancia [3].

1.1.3. *Big Data*

Big Data se refiere a los conjuntos de datos que en su tamaño sobrepasan la habilidad de las herramientas de bases de datos típicas para capturar, almacenar, administrar y analizar [4].
A lo largo de los últimos 20 años, los datos se incrementaron a gran escala en diversos campos. Bajo el explosivo incremento de los datos globales, el término de *Big Data* se usa principalmente para describir enormes conjuntos de datos. *Big Data* típicamente incluye masas de datos no estructurados que necesitan analizarse en tiempo real. También proporciona nuevas oportunidades para el descubrimiento de nuevos valores, ayudando a un entendimiento más profundo de los valores ocultos, y además para incursionar en nuevos desafíos [5]. *Big Data* viene de todas partes para influenciar en la vida, y por lo tanto, es demasiado grande, demasiado complejo y se mueve demasiado rápido. Por ejemplo, la publicación de fotografías y los comentarios que se escriben en Facebook™, subir y ver videos en YouTube™, mandar y recibir mensajes a través de teléfono inteligentes, enviar mensajes de voz a través de WhatsApp™, todo eso cuenta como *Big Data*.

1.1.3.1. Retos

De acuerdo con [6] hay una gran cantidad de retos cuando se manejan los problemas de *Big Data*, los cuales son:

- **Captura de datos.** *Big Data* se recolecta desde distintas fuentes: transacciones, medios sociales, sensores, experimentos, entre otros. Debido a la variedad de las fuentes de datos y el enorme volumen, es difícil recoger e integrar los datos con escalabilidad desde ubicaciones distribuidas.
- **Almacenamiento.** *Big Data* no solo requiere de enormes cantidades de almacenamiento, sino que también demanda una nueva gestión de datos en grandes sistemas distribuidos, esto debido a que los sistemas de bases de datos convencionales tienen dificultades para administrar *Big Data*.
- **Búsqueda.** Los datos sirven para realizar toma de decisiones de manera precisa, es por esto que es necesario que los datos estén disponibles de manera exacta, completa y

oportuna. La optimización de consultas se ha vuelto crucial a la hora de manejar grandes cantidades de datos para obtener respuestas eficientes.

- **Compartición.** La compartición de datos es ahora tan importante como la producción de los mismos. Los profesionales continuarán produciendo y consumiendo información de acuerdo a las necesidades de sus negocios, pero ahora se generan de una manera que se conecten y compartan con otros aspectos de la empresa. En otras disciplinas, como Medicina, Ecología, Biología, entre otras, se enfrentan a cuestiones relativas a la conservación y compartición de los datos.
- **Análisis.** El análisis oportuno y económico sobre *Big Data* es ahora un ingrediente clave para el éxito en diferentes negocios, disciplinas científicas y de ingeniería, y gobiernos. Es necesario optimizar los marcos de trabajo para el análisis de *Big Data*, así como los ambientes para un análisis profundo.
- **Visualización.** El trabajar con *Big Data* hace que la visualización de datos sea una tarea desafiante. El análisis visual es un campo emergente, en el cual enormes conjuntos de datos se presentan al usuario de una manera visualmente atractiva con la esperanza de que los usuarios serán capaces de descubrir relaciones interesantes.

1.1.3.2. Características

Big Data posee cinco características principales [6], [7] como se ilustran en la Figura 1.1 (aunque en ciertos trabajos presentan un modelo de 7Vs que define actualmente a *Big Data*, dicho modelo no se utiliza en esta investigación, ya que consideran características adicionales como: variabilidad, visualización, validez, viabilidad, entre otras [8]–[11]. Debido a que no existe un acuerdo en cuanto al modelo de las 7Vs, se prefirió usar el modelo de las 5Vs). Este modelo de las 5Vs contempla:

- **Volumen.** Describe cómo los conjuntos de datos de *Big Data* son extremadamente grandes y fácilmente alcanzan los terabytes, incluso yottabytes de información. Esta característica no solo conlleva a un problema de almacenamiento, sino también a un problema en el análisis.

- **Velocidad.** Describe qué tan rápido se producen los conjuntos de datos, también se refiere a que la recolección y análisis de datos se lleve a cabo rápida y oportunamente.
- **Variedad.** Indica cómo los conjuntos de datos de *Big Data* contienen tanto datos estructurados como no estructurados, tal como documentos, correos electrónicos, archivos de audio, imágenes, videos, datos móviles, archivos de registros de red, entre otros.
- **Valor.** Es importante entender que hoy en día los datos tienen un costo o valor, por ejemplo: económico, estadístico, entre otros.
- **Veracidad.** Es una necesidad validar la precisión de los datos mediante la eliminación de datos ruidosos a través de diversas metodologías, esto con la finalidad de asegurar la calidad de los datos.

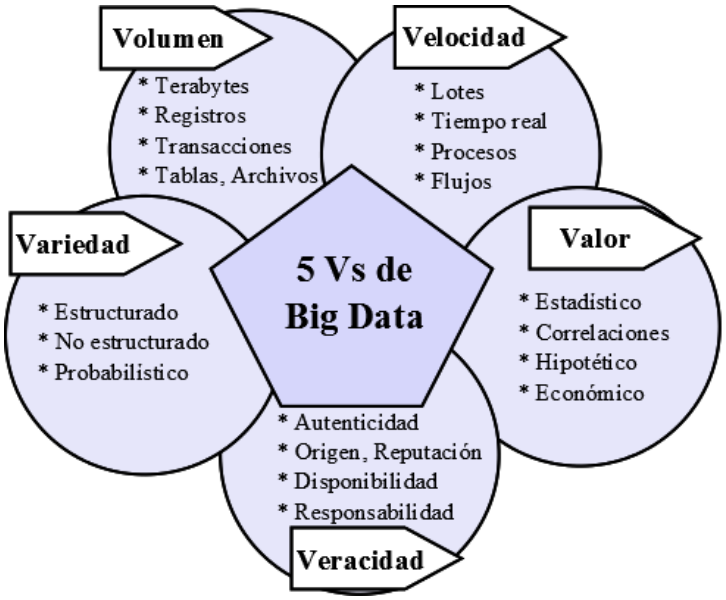


Figura 1.1 Modelo de las 5Vs de *Big Data*.

1.1.4. Análisis visual

El análisis visual para *Big Data* es el proceso de examinar grandes cantidades de datos de forma visual y demostrar una variedad de patrones ocultos, tendencias y correlaciones desconocidas [12]. Surgió como una técnica para descubrir conocimiento a partir de los datos. El análisis visual combina métodos de análisis de datos con la visualización interactiva para

permitir la exploración comprensiva de los datos. Además, al análisis visual toma ventaja de las capacidades de percepción y razonamiento humano para llevar a cabo un análisis exhaustivo de datos, tanto en el panorama general y niveles detallados [13].

1.1.5. Técnicas de visualización

Técnicas usadas para la creación de imágenes, diagramas o animaciones para comunicar, entender y mejorar los resultados del análisis *Big Data*. Presentar la información de tal manera que la gente la consuma de manera efectiva es un desafío fundamental que necesita cumplirse si el análisis de datos es para llevar a una acción concreta [4].

1.1.5.1. Orientadas a píxeles

Una manera simple de visualizar el valor de una dimensión es el uso de un píxel, donde el color del píxel refleja el valor de la dimensión. Para un conjunto de datos de *m* dimensiones, esta técnica crea *m* ventanas en la pantalla, una para cada dimensión. Los valores de la dimensión *m* de un registro se asignan a *m* píxeles en las posiciones correspondientes en la ventana [14].

1.1.5.2. Proyección geométrica

Un inconveniente de las técnicas de visualización orientadas a píxeles, es que no ayudan mucho en el entendimiento de la distribución de los datos en un espacio multidimensional. Las técnicas de proyección geométrica ayudan a encontrar interesantes proyecciones de conjuntos de datos multidimensionales. La tarea fundamental es tratar de resolver cómo visualizar un espacio de alta dimensión en una pantalla 2D. Los gráficos de dispersión muestran los puntos de datos en 2D usando coordenadas cartesianas. Una tercera dimensión se agrega usando diferentes colores o formas para representar diferentes puntos de datos [14].

1.1.5.3. Basadas en íconos

Utiliza pequeños íconos para representar valores de datos multidimensionales. Dos técnicas populares basadas en íconos son: *"Chernoff faces"* y *"stick figures"*.

"Chernoff faces" muestran datos multidimensionales hasta para 18 variables (o dimensiones) como una cara humana de dibujos animados. Esta técnica ayuda a revelar tendencias en los datos. Los componentes de la cara, como los ojos, oídos, boca y nariz, representan valores de las dimensiones por su forma, tamaño, ubicación y orientación.

La técnica de visualización *"stick figure"* mapea datos multidimensionales para figuras de palo de cinco piezas, donde cada figura tiene cuatro miembros y un cuerpo. Dos dimensiones se asignan a la pantalla (x e y) de los ejes y las dimensiones restantes se asignan al ángulo y/o la longitud de los miembros [14].

1.1.5.4. Basadas en jerarquías y grafos

Para grandes conjuntos de datos de altas dimensiones sería difícil visualizar todas las dimensiones al mismo tiempo. Por lo que estas técnicas reparten todas las dimensiones en subconjuntos, por ejemplo sub-espacios. Los sub-espacios se visualizan de una manera jerárquica. Como ejemplo de este tipo de técnicas se tiene al método *"Worlds-within-Worlds"* (Mundos dentro de Mundos), también conocido como "*n*-Vision". Otro ejemplo de métodos de visualización jerárquicos es *"tree-maps"* (mapas de árboles), que muestran datos jerárquicos como un conjunto de rectángulos anidados [14].

Finalmente se tiene a la técnica *"Heatmaps"* (mapas de calor), la cual es una representación visual de datos cuantitativos en dos ejes, el eje X refleja generalmente muestras individuales y el eje Y consta de grupos de parámetros medidos. El campo entre los ejes está compuesto por una colección de cajas contiguas codificadas por color para reflejar la cuantificación. Son una herramienta de visualización flexible para agrupar datos y explorar patrones [15].

1.1.5.5. Relaciones y datos complejos

Originalmente las técnicas de visualización eran principalmente para datos numéricos. Recientemente, más y más datos no numéricos, como texto y redes sociales, están disponibles para su análisis. Visualizar y analizar estos datos atrae mucho interés. Hay muchas técnicas de visualización nuevas dedicadas para este tipo de datos. Por ejemplo, muchas personas en la Web etiquetan varios objetos como imágenes, entradas de blogs y reseñas de productos. Una

"tag cloud" (nube de etiquetas) es una visualización de las estadísticas de las etiquetas generadas por el usuario [14].

Otra técnica es *"Clustergram"*, técnica de visualización usada para el análisis de agrupamiento mostrando cómo los miembros individuales de un conjunto de datos se asignan a los grupos mientras aumenta el número de grupos. La elección del número de grupos es un parámetro importante en el análisis de agrupamiento. Esta técnica permite al analista llegar a un mejor entendimiento de cómo los resultados de agrupamiento varían con diferentes números de grupos.

"History flow" (Flujo de la historia) es una técnica de visualización que traza la evolución de un documento que se editó por varios autores. El tiempo aparece en el eje horizontal, mientras que las contribuciones al texto están en el eje vertical; cada autor tiene un código de color diferente y la longitud vertical de una barra indica la cantidad de texto escrito por cada autor.

Una técnica más es la denominada *"Spatial information flow"* (Flujo de información espacial), la cual se encarga de visualizar los datos de forma "espacial" dentro de un mapa geográfico. Por ejemplo, teniendo un mapa mundial donde se representa la cantidad de datos IP (*Internet Protocol* o Protocolo de Internet) que fluyen entre Nueva York y el resto del mundo mediante un resplandor, se visualiza que entre mayor sea el brillo mayor será el flujo de datos. Esto permite determinar rápidamente qué ciudades están más estrechamente relacionadas con Nueva York en términos de su volumen de comunicaciones [4].

1.2. Planteamiento del problema

Actualmente se está viviendo la era de la información, donde grandes cantidades de datos se recolectan diariamente, ya sea para su análisis o simplemente para su almacenamiento. Estos grandes volúmenes de conjuntos de datos se engloban en el término de *Big Data*, es necesario analizar estos datos para extraer información útil que le sirva tanto a empresas como a investigadores en las diferentes áreas en las que se desempeñan. El análisis visual de *Big Data* es un campo nuevo donde surgen las técnicas de visualización, las cuales ayudan en este análisis de grandes conjuntos de datos para descubrir relaciones, patrones u otras características que no se observan fácilmente.

Dado el párrafo anterior, en este trabajo se analizaron y compararon las técnicas de visualización orientadas a píxeles, técnicas de proyección geométrica, técnicas basadas en íconos, técnicas basadas en jerarquías y grafos, así como visualización de relaciones y datos complejos. Se revisó la literatura para determinar los tipos de conjuntos de datos más comunes en los problemas de análisis *Big Data* y aplicar las técnicas de visualización descritas anteriormente, determinando un marco de evaluación para comparar las distintas técnicas, y con base en factores como velocidad, precisión, facilidad de interpretación, entre otros, decretar cuál es la mejor opción de visualización para los conjuntos de datos de *Big Data*.

1.3. Objetivo general y específico

En esta sección se plantea el objetivo general y los objetivos específicos a cubrir en el proyecto de tesis.

1.3.1. Objetivo general

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

1.3.2. Objetivos específicos

1. Estudiar los conceptos básicos de *Big Data* y análisis visual para conocer el alcance de la investigación.
2. Analizar el estado del arte para determinar cuáles son las técnicas actuales que se utilizan para el análisis visual de *Big Data.*
3. Seleccionar el conjunto de técnicas que se analizarán de cada tipo, así como los conjuntos de datos que se visualizarán.
4. Identificar las tecnologías a utilizar, esto es, herramienta de análisis, sistema gestor de bases de datos, entre otras.
5. Determinar un marco de evaluación que permita comparar cada técnica.
6. Implementar las técnicas para visualizar los conjuntos de datos seleccionados.
7. Validar las técnicas implementadas en un caso de estudio.

8. Comparar cada técnica con base en el marco de evaluación para conocer cuáles son las mejores.

1.4. Justificación

En los últimos años se popularizó el término de *Big Data*, esto por la creciente explosión en la generación y recaudación de datos por distintas empresas e investigadores. Es por esto que cada día se requieren más y mejores técnicas para el entendimiento de los datos, y así ayudar en la toma de decisiones. El análisis visual representa una gran ayuda para dicho entendimiento, ya que son extremadamente adecuadas para:

1. Inspeccionar datos a múltiples escalas en conjunción con el análisis estadístico.
2. Proveer una forma de mantener el contexto mostrando datos como un subconjunto de una mayor parte de los datos o mostrar variables correlacionadas, y
3. Ayudar a identificar patrones a través del tiempo en flujos de datos.

Dado que el análisis visual de *Big Data* es un campo nuevo, hoy en día existen pocos reportes técnicos o marcos referenciales [16]–[18] donde se analicen las diferentes técnicas de visualización para el análisis de *Big Data*. En los análisis existentes no consideran una revisión exhaustiva de la literatura por medio de la que se determinen los conjuntos de datos más utilizados, así como cuál es la mejor técnica para aplicar a dichos conjuntos de datos. Es por eso que es de vital importancia realizar una revisión más completa del estado del arte, un análisis comparativo y plasmar dicho análisis en un reporte técnico para su posterior difusión. Con dicho reporte se pretende beneficiar y ayudar a la comunidad de investigadores de *Big Data* en sus respectivos análisis, dado que conocerán las ventajas y desventajas de aplicar las técnicas de visualización para ciertos conjuntos de datos, reduciendo el tiempo y esfuerzo que cada uno de los investigadores aplica.

Capítulo 2. Estado de la práctica

En este capítulo se presenta el estado de la práctica, es decir, una recopilación y análisis de trabajos relacionados con el proyecto de tesis, esto con el fin de obtener información relevante y adquirir conocimiento acerca de los enfoques con los que se trabajó en el tema de tesis.

2.1. Trabajos relacionados

Con el desarrollo de nuevas tecnologías se incrementó el volumen de los datos en diversas áreas. Por otra parte, la información nunca estará limitada por dos o tres dimensiones y esto conduce a la generación de altas dimensiones de datos. Frente a las inherentes limitaciones de las habilidades cognitivas del humano para tratar con grandes volúmenes de datos, múltiples técnicas de visualización se utilizan ampliamente como herramientas efectivas para la abstracción de datos para su análisis y descubrimiento de conocimiento.

En [19] el principal objetivo fue la búsqueda de un método para proyectar los datos multi-dimensionales al espacio en 3-D para visualizar la información de grandes conjuntos de datos. Para esta visualización se requirió minimizar la distancia entre las clases. De acuerdo con los autores, si los datos tienen características de agrupamiento es razonable esperar que la visualización de los datos también tenga estas características. Por lo tanto, para [19] se usó el método de agrupamiento *k-means*.

El método de visualización en 3-D para *Big Data* utilizó *k-means* en las bases de datos originales para obtener buenos centros de agrupamiento. Después, se diseñó un mapa para proyectar los datos en el espacio original de los centros de agrupamiento. A los puntos de datos se le asignaron pesos específicos para estas tres direcciones de agrupamiento. Finalmente se obtuvieron los valores de las coordenadas en el espacio 3-D.

Para la verificación de los resultados de este enfoque, se realizó un experimento con las bases de datos extraídas del repositorio UCI (*University of California, Irvine*), incluyendo las bases de datos *iris*, *wine* y *pendigits*.

En [20] se desarrolló una herramienta llamada *Irregular Trend Finder* (ITF) para participar en el VAST (*Visual Analytics Science and Technology*) *Challenge* 2012. ITF es una herramienta

interactiva diseñada para el análisis de grandes cantidades de datos teniendo en mente el tiempo y una estructura jerárquica de los datos, por lo que un usuario ve primero la información general y luego, a petición, obtiene una información más detallada. En este desafío, los participantes recibieron un conjunto de datos de gran tamaño con el estatus de salud de cada máquina de una red. La misión fue encontrar "anomalías" dentro de la red usando alguna clase de herramienta de análisis.

Se diseñó ITF con un enfoque "primero resumir, luego enfocar y filtrar, y después los detalles a demanda" para que los usuarios encontraran las anomalías de esa forma. La cantidad de datos proporcionada para el desafío fue demasiado grande, por lo que era difícil mostrar todos esos datos al mismo tiempo. Si el usuario se daba cuenta de que había algunos tipos de anomalías, la herramienta proporcionaba una forma para ver la información detallada sobre las mismas.

ITF se desarrolló en *Processing* con una biblioteca JSON (*JavaScript Object Notation*). Tiene una estructura jerárquica y consiste en tres vistas. La primera vista, donde el usuario ve el porcentaje del estatus de todas las máquinas de la compañía; la segunda vista es por Región o *DataCenter*, y la tercera vista por rama o sede, indicando el estatus individual de cada máquina. El usuario ve la información más detallada con vistas profundas.

Con el uso de esta herramienta, se descubrió y confirmó que ver primero el resumen y después observar el detalle de la información es un método eficiente para buscar información particular en caso de grandes volúmenes de datos.

El uso generalizado y adopción de la Web basada en geo-mapas proporcionan un conjunto familiar de interacciones para explorar grandes espacios de geo-datos. Se requieren nuevas herramientas para la caracterización de *Big Data* a través del análisis exploratorio de datos (EDA, *Exploratory Data Analysis*) para sugerir hipótesis iniciales para pruebas. Derivado de estas técnicas en [21] se desarrolló un sistema basado en el análisis visual (TBVA, *Tile Based on Visual Analysis*) y se aplicó a conjuntos de datos de Twitter™ para demostrar la visualización interactiva.

Un ejemplo del conjunto de datos de Twitter™ estuvo compuesto por dos colecciones. La primera fue un conjunto de 300 millones de registros de geo-etiquetas. La segunda fue una

colección de un billón de *tweets* de más de 15 meses sin la consistencia del geo-etiquetado. Diagramas interactivos de todos los datos permitieron examinar todo el conjunto para hipótesis iniciales.

El enfoque de TBVA para EDA atrajo muchas interacciones paralelas en *browsers* basados en mapas geográficos, como lo es en *Google Maps*™ o *Bing Maps*™. Estos sistemas pre-procesan la imagen a múltiples resoluciones almacenándola en una pirámide de alimentación usando una cuadricula predefinida, que se presenta a demanda del cliente Web. El enfoque presentado en [21] abstrae el mapa geográfico a una superficie general de trazado, permitiendo a los analistas navegar desde una visión general hasta el nivel más bajo de detalle.

Las herramientas exploratorias de TBVA también difieren de otros mapas geográficos mediante el aplazamiento del procesamiento de imagen hasta la hora de la solicitud. TBVA generó los mosaicos "*tiles"* en tres escenarios: el escenario de agregación, el escenario de resumen y el escenario de procesamiento. Este enfoque permitió una modificación interactiva y control de la representación visual. También se creó un mapa de calor (*Heatmap*) para la localización de la densidad del *tweet*.

Usando un billón de puntos de datos de Twitter™, se demostró que TBVA permite el análisis exploratorio de datos para realizarse en grandes conjuntos de datos de tamaño prácticamente ilimitado.

El crecimiento en los volúmenes de datos afecta a las grandes organizaciones de hoy, donde comúnmente usan herramientas de software para capturar, administrar y procesar los datos, pero que no manejan grandes volúmenes de datos de manera efectiva. El desafío principal para esas organizaciones es analizar grandes cantidades de datos y extraer información útil o conocimiento para futuras acciones en un corto tiempo. Este tipo de demandas propicia el mercado para innovadoras herramientas de control para *Big Data*, tales como el análisis visual.

En [12] se propuso el análisis visual de *Big Data* usando el software estadístico R. El método propuesto se compone de tres pasos. En el primer paso se extraen los datos de un sitio Web destino. En el segundo paso se analizan los datos obtenidos de acuerdo al tipo y se almacenan en una base de datos. En el tercero se realiza un análisis visual de los datos almacenados en la

base de datos usando el software estadístico R. Los datos recolectados pertenecen al sitio Web de *Calcutta Communication*.

R es una suite integrada de servicios de software para la manipulación de datos, cálculo y representación gráfica. La herramienta R cuenta con facilidades gráficas para el análisis y visualización de los datos, ya sea directamente en la computadora o en papel. En años recientes, el software estadístico R emerge como una de las herramientas más usadas por los científicos de los datos. Grandes organizaciones como Facebook™, Google™ y LinkedIn™ lo adoptaron.

Se diseñaron varios escenarios para el análisis visual de los datos, incluyendo visualización de datos simple, agrupamiento, minería de texto y redes sociales. El resultado de estos métodos provee información útil para las organizaciones y obtiene ventajas competitivas sobre las organizaciones rivales.

En [22] se presentó una nueva y eficiente forma de analizar grandes conjuntos de datos usados en la investigación de las ciencias sociales. Se proporcionó y demostró una forma de lidiar con este tipo de conjuntos de datos sin la necesidad de realizar instalaciones de computación distribuida de alto rendimiento. Se usó un conjunto de datos censado de Internet y con la ayuda de herramientas libres disponibles y bibliotecas de programación, se visualizó la actividad global de IP en una dimensión espacio y tiempo. También se observó una considerable reducción en el tamaño de almacenamiento del conjunto de datos junto con un tiempo de procesamiento más rápido.

Los científicos sociales se basan en los datos recogidos por los experimentos y encuestas o los datos suministrados por las oficinas de estadística. Recientes avances tecnológicos han permitido a los profesionales utilizar conjuntos de datos nuevos y muy grandes, con frecuencia revelando, por primera vez, la información sobre las decisiones y preferencias individuales. En el estudio [22] el objetivo fue utilizar un censo de Internet como caso de estudio de la utilización de fuentes de datos alternativas para las ciencias sociales. El censo de Internet contiene direcciones IP regulares de todo el espacio IPv4 en 2012. Desde una perspectiva de ciencia social, cada variación en la actividad *on-line* y *off-line* es realmente una revelación de las preferencias individuales para el acceso a Internet.

Para la solución del problema de análisis de grandes datos, se desarrolló un programa Java con software libre disponible para filtrar y agregar el conjunto de datos, ejecutable sobre una computadora multi-núcleo. El código fuente está disponible sobre demanda. Se demostró el método para producir patrones reveladores de la actividad de Internet de alrededor de las principales ciudades de Australia en la dimensión tiempo y espacio.

Se encontró que el método entrega exitosamente una reducción significativa en los requerimientos de espacio de almacenamiento (alrededor de 18 veces), dando lugar a una velocidad similar en el tiempo de procesamiento. Por otra parte, el formato resultante se convirtió fácilmente a los tipos de datos de herramientas comunes usados por los científicos sociales, como son Qgis, R, Stata o Matlab.

Describiendo el método, se empezó por la conversión de 9 TB de archivos de texto a un formato binario compacto y luego se transformó el conjunto resultante en un formato legible para los programas estadísticos estándar en las ciencias sociales. Existen dos problemas de procesamiento de datos. Primero, los datos se convierten usando *HDF5* (*Hierarchical Data Format* 5) para tener un formato reducido, rápido y accesible, y segundo, la información necesaria se agrega y extrae. El sitio Web de *HDF5* lista varias implementaciones para el soporte de lenguajes de programación como C, Python y Java. Finalmente en [22] se usó *JHDF5*, el cual es un envolvente de Java y por su uso fácil y reducción de la barrera de entrada.

El conjunto de datos de salida se procesó exitosamente con herramientas estándar de visualización de datos como es R y Qgis, respectivamente.

Debido al incremento sin precedentes de la tecnología computacional de alto rendimiento, las simulaciones están evolucionando hacia una mayor fidelidad numérica y complejidad. Por lo tanto, las técnicas para analizar los datos eficientemente, particularmente las técnicas visuales interactivas, no siguen el ritmo del crecimiento. En [23] se describió y demostró un sistema de análisis visual, llamado *Exploratory Data analysis ENvironment* (EDEN), con aplicación específica para el análisis de los conjuntos de datos de la simulación del sistema complejo terrestre. EDEN representa el tipo de herramienta interactiva de análisis visual que es

necesario para transformar los datos en conocimiento, mejorando así la comprensión crítica de los procesos del sistema terrestre.

Este nuevo sistema, EDEN, está disponible libremente y facilita el descubrimiento de conocimiento interactivo y la generación de hipótesis para el análisis exploratorio más productivo de los datos de simulación climática.

A través de varias evaluaciones prácticas de EDEN en estudios climáticos del mundo real, se corroboró la idea de que usando un marco de trabajo de análisis visual interactivo se obtiene un proceso más eficiente para el análisis climático en comparación con las herramientas convencionales.

El foco de atención de [23] fueron los conjuntos de datos de CLM4 (*Community Land Model*). Las simulaciones globales de CLM4 contienen 360 variables de salida, la mayoría son de dos dimensiones, con algunas en tres dimensiones. Las simulaciones consisten en archivos de salida mensuales que son típicamente alrededor de 415 MB cada uno. Para una simulación de 100 años, se produjeron 1200 archivos, totalizando alrededor de 500 GB.

El éxito general de EDEN y su amplia adopción en la comunidad climática es mayormente atribuido al hecho de que expertos del dominio se integraron estrechamente en el proceso de desarrollo y sus subsecuentes iteraciones de diseño.

El rastreo y el almacenamiento del comportamiento de navegación de los usuarios en la Web, hasta con clics de ratón individuales, crean enormes *logs* de sesiones Web. Mientras tales datos de sesiones Web contienen información de gran valor acerca de los comportamientos de los usuarios, el incremento en el tamaño de los datos da lugar a grandes desafíos para el análisis y visualización de los datos. Con el objetivo del análisis visual “Primero analizar, mostrar lo importante, enfocar, filtrar y analizar más a fondo, detalles a demanda”, en [13] se desarrolló un sistema de análisis visual de dos niveles, *TrailExplorer2*, para descubrir conocimiento de grandes *logs* de datos.

La comprensión integral de la experiencia del usuario es crucial para la mejora de servicios en un sitio Web. Por lo general, un usuario navega en el sitio Web dejando un rastro de qué páginas visitó y cuánto tiempo permaneció en cada página. Este tipo de secuencia de eventos, almacenados como sesiones Web, refleja en gran medida el sentir de los usuarios acerca del

sitio Web. Analizando estas secuencias de eventos se avanza en el entendimiento de la experiencia de usuario.

TrailExplorer2 contempló dos niveles. Un nivel es un motor de búsqueda de datos de gran escala basado en Hadoop para extraer datos de interés; el otro nivel es una interfaz visual interactiva para apoyar la exploración visual.

Cuando *TrailExplorer2* inicia, se le pide al usuario que seleccione un conjunto de eventos de su interés. El sistema entonces busca todas las sesiones que contienen cualquiera de esos eventos. Para el manejo de enormes conjuntos de datos se escogió el modelo MapReduce.

La interfaz visual de *TrailExplorer2* contiene cuatro componentes, esto es, la vista principal, el panel de información a detalle, la leyenda y el gráfico de distribución del tiempo transcurrido. Las sesiones de usuario se visualizaron en la vista principal. Además de la vista de detalle de las sesiones individuales, las estadísticas agregadas se mostraron en el panel de información a detalle y la tabla de distribución. La vista de leyenda mostró los diferentes colores utilizados para representar todas las posibles ocurrencias de eventos.

La interacción es la clave para un ambiente de análisis visual. Este sistema manejó varias interacciones de usuario. Primero, los analistas destacaron los eventos de interés para investigar sus características. Los eventos destacados son de color más brillante. Segundo, usando el árbol de prefijo por defecto, las sesiones de salida estuvieron alineadas sobre sus primeros eventos. Otra interacción fue la modificación del tamaño del depósito de tiempo, lo que permite a los analistas visualizar e investigar los datos con diferentes niveles de granularidad.

TrailExplorer2 [13] se desarrolló por el hecho de que los analistas tienen dificultades para explorar la gran cantidad de datos de las sesiones Web para obtener ideas de los patrones de comportamiento de los usuarios. Para demostrar la efectividad del sistema se ocuparon los datos de las sesiones del mundo real de eBay.

Los conceptos de Almacén de Datos, Cómputo en la Nube y *Big Data* se propusieron durante la era de la sobreexplotación de datos. En [24] se introdujo un marco de trabajo para lograr una mejor visualización para *Big Data*. Este marco de trabajo redujo el costo de construcción de almacenes de *Big Data* mediante la división de datos en subconjuntos de datos y

visualizarlos respectivamente. Basándose en la herramienta de visualización D3.js y dirigida por el principio *"Whole-Parts"*, los datos actuales se presentaron a los usuarios desde diferentes dimensiones por diferentes gráficos estadísticos.

El proceso de análisis consume recursos, los cuales son limitados en un almacén de datos. Si estos cálculos se dividen y analizan respectivamente en el servidor fuente de los datos, entonces ayudaría a reducir la sobrecarga de doble conteo de datos de gran tamaño mediante el uso de una interfaz de acceso de datos. Es por esto que Fan Boa y Jia Chen [24] presentaron un marco de trabajo visual para *Big Data*, que utilizó la herramienta Web de presentación visual D3.js para lograr gráficos estadísticos. Basados en la idea "Primero darle un tamaño adecuado, seleccionar un resumen adecuadamente, y luego mostrar los detalles necesarios", los gráficos de visualización interactiva se mostraron desde una dimensión de análisis diferente. En [24] los gráficos de visualización interactiva se demostraron por las dimensiones categoría de libro, tiempo, entre otras.

D3.js es una biblioteca Java*Script* para manipulación de documentos basados en datos. Se logró mostrar la visualización de carga de datos, enlace de datos, elementos de transformación analítica y elementos excesivos.

Las pruebas se realizaron sobre los datos de ventas de libros, mediante la selección de diferentes dimensiones estadísticas, se utilizó un gráfico circular que muestra el perfil general, o un gráfico de barras que muestra las ventas específicas debajo de cada dimensión.

Este marco de trabajo para la visualización de datos, minería de datos y métodos estadísticos no fue exhaustivo y no satisface todas las necesidades de los usuarios. Sin embargo, basado en la idea de [24], el usuario es capaz de construir un marco de trabajo de acuerdo a su demanda y luego desarrollarlo con las restricciones visuales, escogiendo diferentes dimensiones de los datos.

El análisis visual de conjuntos de datos de referencia geográfica con un gran número de atributos es desafiante, debido al hecho de que las características de los atributos dependen altamente de las ubicaciones en las que se centran, y la escala y hora en que se miden. Se requieren métodos visuales interactivos y especializados para ayudar a los analistas en el entendimiento de las características de los atributos cuando estos múltiples aspectos se

consideran concurrentemente. Es por esto que en [25] se desarrolló *"attribute signatures"*, gráficos diseñados interactivamente que muestran la variabilidad geográfica de las estadísticas de atributos, a través de los cuales se explora visualmente el grado de dependencia entre los atributos y la geografía. Este método permitió la variación de múltiples resúmenes estadísticos de varios atributos para considerarse concurrente y geográficamente.

En algunos casos, los datos tienen un fuerte componente geográfico el cual domina la variación. Por ejemplo, el entendimiento de las características de la población es de gran importancia para los gobiernos y agencias involucradas en proveer servicios y diseño de políticas.

El diseño de mecanismos de apoyo a la exploración de la variación geográfica en múltiples atributos simultáneos es difícil, puesto que las distribuciones geográficas tienden a ser heterogéneas y a menudo están estrechamente relacionadas e influenciadas por las características topográficas. El mecanismo visual e interactivo propuesto está diseñado para ayudar al análisis de los datos geográficos para abordar estos desafíos.

Se utilizó un único conjunto de datos para demostrar el método desarrollado para el análisis de datos geográficos multivariados. El conjunto de datos consistió en registros tomados del Censo de Población del Reino Unido en 2001 y 2011 para las 181,000 *"Output Areas"* (OA) de Inglaterra y Gales. Cada OA tuvo 41 atributos asociados a ella.

Las diferentes perspectivas sobre la variación geográfica proporcionan una estructura que se ocupó para el diseño y desarrollo de los métodos de análisis. Se empezó con un marco de trabajo que establece la estructura del espacio de análisis. Después se discutieron métodos visuales interactivos para hacer frente a las diversas partes de este espacio.

El objetivo fue desarrollar técnicas para ayudar a entender cómo los atributos múltiples varían en el espacio como un medio para obtener el conocimiento de los fenómenos representados por los datos geográficos. El marco de trabajo, las técnicas y la herramienta que se presentaron facilitan esta actividad a través de un conjunto estructurado de perspectivas de análisis y visualizaciones asociadas y calculadas.

En [26] se presentó *ConnectomeExplorer*, una aplicación para la exploración interactiva y el análisis visual guiado por consultas de grandes conjuntos de datos de microscopía electrónica

(EM, *Electron Microscopy*) volumétrica en la investigación *conectómica*. Este sistema incorporó un álgebra de consulta basado en conocimiento que apoya la especificación interactiva de las consultas evaluadas dinámicamente, lo cual permitió a los neurólogos plantear y responder a las preguntas de dominio específico de una manera intuitiva. Se evaluó la aplicación sobre un conjunto de datos de más o menos un terabyte de datos de EM y 750 GB de datos de segmentación, que contiene alrededor de 4,000 estructuras segmentadas y 1,000 sinapsis.

Los neurólogos esperan obtener una comprensión de cómo el cerebro se desarrolla y funciona, y cómo se desarrollan y tratan algunas patologías. Para apoyar este objetivo, se desarrollaron métodos de alto rendimiento para obtener imágenes neuronales, para permitir a los científicos adquirir datos de imágenes a velocidad y resolución sin precedentes.

Los neurólogos actualmente tienen enormes colecciones de EM volumétrica de alta resolución y sus segmentaciones, pero carecen de medios eficaces para el análisis de ellos o responder directamente preguntas de alto nivel de dominio específico. La mayoría de las herramientas actuales solo proporcionan visualizaciones 2D de subconjuntos de datos EM o las correspondientes segmentaciones.

ConnectomeExplorer es un sistema integrado para el análisis interactivo y visualización 3D de datos de neurociencia a gran escala. Posee un álgebra de consultas que proporciona a los científicos en la investigación *conectómica* una forma intuitiva de especificar consultas de alto nivel de manera dinámica que se evalúan en el conjunto de datos completo. También cuenta con una interfaz de usuario intuitiva que permite especificar combinaciones de consultas en diferentes dominios, y explorar y analizar simultáneamente los resultados en múltiples vistas vinculadas. Este sistema de visualización está basado en un marco de trabajo de representación flexible que es escalable a datos petaescala y que soporta múltiples volúmenes y vistas vinculadas.

Se probó el sistema sobre una máquina con CPU quadcore a 3.2 GHz con 12 GB de memoria RAM y una GPU NVIDIA GTX con 2 GB de RAM.

Debido a la complejidad y “oscuridad” de los conjuntos de datos de gran escala (*Big Data*), aumentó el interés en el análisis visual (VA) gracias a la capacidad humana para obtener

conocimiento rápidamente a través de los procesos de análisis y decisiones visuales. Desafortunadamente, la mayoría de las herramientas o técnicas VA no son apropiadas para *Big Data*. Para mejorar la escalabilidad visual, los métodos computacionales proveen información compacta y significativa de los datos. Tales métodos incluyen reducción de dimensión, agrupamiento y métodos que explotan el aprendizaje automático y la minería de datos. En [27] se sugirieron maneras para personalizarlos, para aprovechar los aspectos específicos para ambientes VA.

Muchos sistemas VA escogen un método computacional específico, lo tratan como una caja negra, y se centran en el análisis de salida. La mayoría de estos métodos implican computación fuerte. A medida de que se vuelven más avanzados y capaces, tienden a requerir mayor intensidad computacional.

Se sugirió cómo personalizar los métodos computacionales abordando la discrepancia de precisión y convergencia. Una de las maneras más fáciles de reducir la precisión y por lo tanto acelerar el cálculo es cambiar la precisión doble a precisión simple. Extendiendo esta idea para VA aceleraría muchos métodos computacionales, mientras se mantenga la calidad de sus resultados visuales cuando la estabilidad numérica no sea una preocupación crítica.

Un algoritmo llegará a una solución viable para VA mucho antes que su iteración convergente. Sin embargo, la forma de determinar dichos criterios de convergencia todavía no están claros. En lugar de ello, se propuso una visualización interactiva a nivel de iteración, que tiene por objeto la visualización intermedia de los resultados en varias iteraciones y dejar a los usuarios interactuar con estos resultados en tiempo real. Debido a que este enfoque visualiza dinámicamente los resultados intermedios, los usuarios obtienen rápidamente una visión general de ellos.

Para lograr una estrecha integración entre los métodos computacionales y VA, es necesario que los investigadores de cada lado se preocupen más por el otro lado. Por lo tanto, es imprescindible que los investigadores quienes diseñan los métodos computacionales se den cuenta que hacer un algoritmo más interactivo e interpretable en casos prácticos de análisis de datos es tan importante como hacer frente a las preocupaciones prácticas tales como el tamaño máximo aplicable de los datos, tiempo computacional y requisitos de memoria. Por otro lado, los investigadores quienes aplican los métodos computacionales para VA necesitan entender

los detalles del algoritmo tanto como sea posible y adaptarlos para hacer que se mezclen bien en tiempo real.

La reconstrucción del *conectoma* humano es uno de los principales esfuerzos científicos más importantes del siglo XXI. Al descifrar los circuitos neuronales del cerebro humano, que comprenden billones de neuronas y sus interconexiones (sinapsis), los investigadores *conectómicos* esperan mejorar su comprensión de la función cerebral, así como patologías como la enfermedad de Alzheimer y el autismo. Sin embargo, el *conectoma* mamífero es inmensamente complejo, y la enorme cantidad de datos imagen que se adquieren, almacenan y procesan presenta un gran desafío para los neurólogos.

La comprensión de los datos requiere visualizaciones interactivas 3D de los volúmenes escaneados, corrección visual de la segmentación y navegación 3D dentro del volumen. Nuevos paradigmas y sistemas de visualización son, por lo tanto, necesarios para facilitar la exploración interactiva y análisis de flujos de datos de microscopía a gran escala.

En [28] se desarrolló un sistema punto-a-punto escalable y diseñado para la exploración interactiva 3D y la navegación de datos EM segmentados de alta resolución. El sistema ofreció un marco de trabajo de procesamiento de volúmenes y visualización flexible que maneja volúmenes de datos petaescala y que trata con datos incompletos.

Este sistema tiene dos partes principales: un conducto basado en datos para manejar la adquisición de imágenes, almacenamiento de datos y la generación de *mipmap* 2D; y el flujo de visualización que apoya la construcción y visualización del bloque 3D.

El entorno del sistema empleó una arquitectura de red cliente-servidor. El archivo de visualización está en un sistema de archivos compartidos, permitiendo a múltiples usuarios acceder a los datos. Todas las comunicaciones de red están basadas en conexiones TCP (*Transmission Control Protocol*) y usan compresión de imagen para reducir el ancho de banda de la red.

El objetivo del flujo que siguen los datos fue procesar los datos nuevos tan pronto como el EM los haya escaneado y estén disponibles para su visualización.

Se probó el sistema sobre tres máquinas con 12 CPU dual-core a 3 GHz, cada una con 48 GB de memoria RAM y una Nvidia Quadro 6000 GPU con 4 GB de RAM GPU. Se Implementó

el sistema en C++, el *ray caster* en GLSL (*OpenGL Shading Language*), y el procesador de muestras en CUDA (*Compute Unified Device Architecture*) y *OpenMP*.

La enorme cantidad de datos de registro de usuarios recogidos por los proveedores de motores de búsqueda crean nuevas oportunidades para entender la lealtad del usuario y comportamiento de la deserción a una escala sin precedentes. Sin embargo, esto también plantea un gran desafío para analizar el comportamiento y recoger una visión de los datos grandes y complejos. En [29] se introdujo *LoyalTracker*, un sistema de análisis visual para rastrear la lealtad del usuario y cambiar el comportamiento hacia múltiples motores de búsqueda a partir de la gran cantidad de datos de registro de usuarios.

Los motores de búsqueda se han convertido en una necesidad en la vida diaria. Como proveedores de servicios en línea, las compañías de motores de búsqueda recolectan datos de registro a gran escala de sus usuarios quienes autorizan proporcionar sus datos de búsqueda, ofreciendo una fuente de datos más rica para un análisis profundo del comportamiento de los usuarios. Esta capacidad atrae considerable atención de las diferentes áreas de investigación, como la minería de datos y visualización.

La mayoría de los sistemas tienen como objetivo visualizar el tráfico Web y/o rutas de navegación a través del sitio Web, pero son incapaces de rastrear la retención de los usuarios rigurosamente o conducir a un análisis sistemático de los patrones de deserción. Un sistema capaz de rastrear la lealtad de los consumidores es críticamente importante para las empresas para evitar la pérdida de clientes.

LoyalTracker permitió a los analistas definir múltiples categorías de lealtad como leales incondicionales y conmutadores, siguiendo la práctica general en el mercado. El sistema tiene tres vistas vinculadas: una vista de flujo, una vista del mapa de densidad, y una vista de nube de palabras. La vista de flujo usó una nueva técnica de visualización interactiva, la cual visualiza el dinamismo en cada categoría de lealtad predefinida y su agregación a través del tiempo con claridad. En la vista del mapa de densidad y nube de palabras, se mostró más información detallada a demanda, lo que ayuda a los analistas en la detección rápida del comportamiento de los usuarios, determinando las razones subyacentes, y de manera intuitiva transmitir los resultados a un público más amplio.

Se implementó el sistema usando Java. Después del pre-procesamiento de los datos, el rendimiento interactivo se alcanzó con una PC con procesador Intel Core i7-2600 y memoria de 8 GB. A fin de evaluar el sistema *LoyalTracker*, se llevaron a cabo entrevistas con cinco expertos de dominio.

Como caso de estudio para el sistema, se extrajeron los 100,000 clientes más activos en el mercado de Estados Unidos durante la primera semana de Julio 2012 de los grandes registros de búsquedas. Después se adquirieron los datos sobre el comportamiento de búsqueda de estos usuarios a partir de Julio 2012 hasta Diciembre 2012. El sistema permitió a los expertos del dominio clasificar libremente a los usuarios en diferentes categorías de acuerdo a los niveles de lealtad de los mismos.

Las técnicas de visualización de datos se usan ampliamente para analizar y explorar datos para permitir a los usuarios seleccionar y reducir interactivamente la amplitud del alcance de la vista para un mejor entendimiento de los datos. Con el surgimiento de *Big Data*, tienen que desarrollarse nuevos métodos analíticos para la visualización para manejar conjuntos de datos muy grandes y complejos en un corto periodo de tiempo.

En [7] se clasificaron los atributos de *Big Data* dentro de las dimensiones 5Ws basado en el comportamiento de los datos. Cada instancia de dato contiene estas dimensiones 5Ws, que se aplican para múltiples conjuntos de datos a través de diferentes formas de datos.

El modelo 5Ws contiene, ¿Por qué ocurrieron los datos? (*Why*), ¿De dónde vienen los datos? (*Where*), ¿Cuándo ocurrieron los datos? (*When*), ¿Quién recibió los datos? (*Who*), y ¿Cómo se transfirieron los datos? (*How*).

El modelo de coordenadas paralelas 5Ws se probó y evaluó usando el conjunto de datos ISCX2012 (*Information Security Centre of Excellence*). El conjunto de datos contiene 130,288 instancias y 20 dimensiones que indican que el total de tráfico es 130,288, incluyendo 37,375 ataques, 36 fuentes IP, 23,653 puertos de origen, 1,656 IP de destino y 22 puertos de destino.

Antes de clasificar los datos dentro de las dimensiones 5Ws, el conjunto de datos tenía 130,288 nodos de datos dentro de 20 dimensiones. Después de aplicar el patrón 5Ws, los datos se redujeron a 4,178 nodos en 5Ws dimensiones sin pérdida de información. La agrupación de datos se redujo más del 95% usando el patrón 5Ws lo que mejoró significativamente.

La visión es el sentido dominante de los seres humanos, que les permite reconocer patrones recurrentes, rápidamente se atiende a lo inesperado, y visualmente se resuelven problemas espaciales complejos y abstractos.

En [30] se mencionó que gran parte de las investigaciones científicas de datos de hoy se crean, almacenan y analizan digitalmente. Los científicos están observando fenómenos con nuevos tipos de instrumentos digitales, sensores y vehículos de exploración robótica que recolectan datos cada vez a mayor resolución. Pero a pesar de muchos avances recientes, la escala, heterogeneidad y complejidad de los conjuntos de datos científicos modernos siguen planteando grandes desafíos para los diseñadores de visualización.

Los científicos están construyendo nuevos tipos de pantallas de computadoras con lentes digitales de alta resolución que se usan efectivamente para visualizar grandes cantidades de datos. Muchas disciplinas científicas están adoptando rápidamente las lentes digitales modernas como *Cyber-Commons* y CAVE2 (*Cave Automatic Virtual Environment*), ambas del Laboratorio de Visualización Electrónica *(Electronic Visualization Laboratory)* de la Universidad de Illinois en Chicago. Estas lentes proveen alta resolución, superficies de visualización estereoscópica, creando espacios de realidad híbrida (HR) que eliminan la barrera entre los ambientes virtuales y paneles de visualización.

Un reto principal en la visualización de *Big Data* es evitar la confusión visual. El gran número de puntos de datos a menudo hacen visualizaciones demasiado densas para ser útiles. Las visualizaciones desordenadas incrementan el flujo de trabajo cognitivo de los usuarios, haciendo difícil la lectura de variables, la comparación de elementos o el reconocimiento de las tendencias en los datos. Una visualización en 2D exacerba la confusión al forzar a los diseñadores a empaquetar todos los elementos en un único plano 2D. Una visualización 3D permite a los diseñadores la libertad de capas de información en 3D para reducir la confusión y mejorar la comprensión y el rendimiento potencialmente. Los ambientes HR mejoraron la resolución estereoscópica que es crucial para dejar a los espectadores utilizar percepción profunda cuando se razona sobre grandes conjuntos de datos.

Una de las principales contribuciones de los ambientes HR fue transformar los tradicionales muros de visualización en sistemas altamente inmersos, incorporando rastreo, dispositivos de

entrada de seis grados de libertad *(six-degree-of-freedom)*, y en algunos casos, un campo de visión panorámica.

Los ambientes HR tienen un gran potencial para ayudar a hacer frente a algunos desafíos de visualización compleja que involucran grandes cantidades de información heterogénea.

2.2. Análisis comparativo

La Tabla 2.1 muestra el análisis comparativo de los trabajos relacionados con el proyecto de tesis.

Artículo	Problema/Objetivo	Tecnología	Conjunto de datos	Técnica de visualización	Resultado	Estado
[19]	Desarrollo de nuevas tecnologías incrementan el volumen de datos, por lo que se vuelve difícil el análisis de la información.	Método *k-means*.	*Iris.* *Wine.* *Pendigits.*	Visualización 3D.	Algoritmo K3DV, *k-means* basado en visualización 3D.	Finalizado.
[20]	Participar en el VAST *Challenge* 2012 con una herramienta de visualización original.	*Processing.* JSON.	Estatus de las máquinas en una red.	Técnica basada en jerarquía.	ITF – *Irregular Trend Finder*.	Finalizado. (Mejoras a futuro)
[21]	Adopción de la Web basada en geo mapas requieren de nuevas herramientas para su análisis.	No se menciona.	*Tweets* con geo-etiquetas.	Mapas de calor.	TBVA – *Tile Based Visual Analytics*.	Finalizado.
[12]	Crecimiento de los volúmenes de datos en las organizaciones, incapaces de realizar un análisis efectivo.	R.	Datos de "*ranking"* de descargas provenientes de Calcutta Comm.	Diferentes gráficas.	Método de análisis visual usando R.	Finalizado.
[22]	Nueva forma de reducir conjuntos de datos para su análisis visual.	Java. HDF5.	Actividad de direcciones IP.	No se menciona.	Programa Java para reducir el conjunto de datos.	Finalizado.
[23]	Simulaciones cada vez más fiables y complejas requieren de mejores herramientas para análisis visual.	Java.	Datos climáticos de simulaciones.	No se menciona.	EDEN - *Exploratory Data analysis ENvironment*.	Finalizado.

Artículo	Problema/Objetivo	Tecnología	Conjunto de datos	Técnica de visualización	Resultado	Estado
[13]	Conocer el comportamiento de los usuarios a través de los "*logs"* de sesiones Web.	Hadoop. MapReduce.	Sesiones de eBay.	Técnica basada en jerarquía.	*TrailExplorer2*.	Finalizado. (Mejoras a futuro).
[24]	Realizar el análisis visual en las fuentes de datos antes de que lleguen al almacén de datos.	HTML. Java*Script*. C#. JSON.	Venta de libros.	Técnicas de proyección geométrica. Orientadas a pixeles.	Marco de trabajo de análisis visual usando la herramienta Web D3.js	Finalizado.
[25]	Dificultades en el análisis de grandes conjuntos de datos de referencia geográfica.	No se menciona.	Censo de población del Reino Unido.	Técnica basada en jerarquía y grafos.	*Attribute Signatures*.	Finalizado.
[26]	Se requieren nuevas herramientas para el análisis interactivo de datos neurológicos.	No se menciona.	Datos de EM.	Visualización 2D y 3D. Diferentes gráficas.	*ConnectomeExplorer*.	Finalizado. (Mejoras a futuro)
[27]	La mayoría de las herramientas o técnicas de análisis visual no son apropiadas para *Big Data*. Se requiere conocer los métodos computacionales a ocupar.	Método *k-means*.	Servicio de noticias.	Gráficos de dispersión.	Algoritmo iterativo de análisis.	Finalizado.
[28]	Se requiere de análisis complejos de las imágenes cerebrales para un mejor entendimiento de enfermedades.	C++. Protocolo TCP. GLSL. CUDA. OpenMP.	Corteza cerebral de ratón.	Visualización 2D y 3D.	Sistema para visualización y análisis de imágenes cerebrales.	Finalizado.
[29]	Existe un gran desafío por analizar el comportamiento y lealtad de los usuarios en los diferentes motores de búsqueda.	Java.	Clientes de Estados Unidos.	Mapas de densidad.	*LoyalTracker*.	Finalizado. (Mejoras a futuro)

Artículo	Problema/Objetivo	Tecnología	Conjunto de datos	Técnica de visualización	Resultado	Estado
[7]	Para un mejor tratamiento de *Big Data* se requiere reducir la amplitud de los datos.	No se menciona.	Ataques o incidentes a direcciones IP.	No se menciona.	Patrón 5Ws para reducir significativamente los datos.	Finalizado.
[30]	Los ambientes HR proporcionan una mejora en la visualización de grandes cantidades de información.	Ambientes HR.	No se menciona.	Visualización 2D, 3D, panorámica de 320 grados.	Reporte sobre ambientes HR.	Finalizado.

Tabla 2.1 Tabla comparativa de los trabajos relacionados.

Como se observa en la tabla anterior, el análisis visual de *Big Data* es de vital importancia en los diferentes campos donde se requiera obtener conocimiento que ayude a la toma de decisiones, por lo que el presente trabajo representa una gran oportunidad para ayudar a los científicos de datos en la selección de técnicas de visualización para los conjuntos de datos *Big Data*.

Es importante mencionar que, al ser un campo nuevo el análisis visual de *Big Data*, hoy en día no hay muchos reportes técnicos o marcos referenciales [16]–[18] que ayuden a determinar cuáles son las mejores técnicas de visualización para los conjuntos de datos de *Big Data*. En los análisis existentes no consideran una revisión exhaustiva de la literatura por medio de la que se analicen y clasifiquen los conjuntos de datos y las técnicas de visualización para *Big Data*.

En [16] los autores plantean una revisión analítica de seis métodos de visualización en *Big Data* (*Treemap* – Mapas de árbol, *circle packing* – embalaje circular, *sunburst* – resplandor solar, *circular network diagram* – diagrama de redes circular, *parallel coordinates* – coordenadas paralelas y *streamgraph* – diagrama de flujo), sin embargo, solo plantean tres características de *Big Data* (volumen, variedad y velocidad) sin profundizar en otras características (como veracidad y valor), además de no contemplar más métodos o técnicas de visualización.

Lidong Wang et al. [17] también presentan una revisión de la visualización en *Big Data*, mostrando información similar a la presentada en [16], pero abarcando temas como: desafíos en la visualización de *Big Data* (pérdida de información, ruido visual, entre otros) y algunas herramientas para la visualización de grandes datos (Pentaho, JasperReports, Tableau, entre otros). Sin embargo, no realizan un análisis más profundo para conocer y determinar las mejores opciones para la realización de un análisis visual de *Big Data*.

Finalmente, en [18] se muestra un análisis comparativo de herramientas, tanto comerciales como de código abierto, para la visualización de *Big Data*, pero no plantean una revisión exhaustiva de las técnicas de visualización utilizadas por estas herramientas, solo mencionan las técnicas que se utilizan para realizar un análisis de *Big Data*.

Por tal motivo, en esta investigación se llevó a cabo un análisis exhaustivo en la literatura para conocer las técnicas de visualización más utilizadas en el campo de *Big Data*, así como los conjuntos de datos utilizados, realizar las visualizaciones pertinentes para determinar cuáles son las mejores al momento de analizar *Big Data*.

2.3. Propuesta de solución

En esta sección se presenta la propuesta que solucionó la problemática planteada en anteriores puntos.

Se realizó un análisis exhaustivo de las diferentes tecnologías y herramientas para el procesamiento, análisis y visualización de grandes datos, así como también de la metodología a utilizar para el desarrollo de este proyecto.

La selección de estas herramientas se realizó de acuerdo a las necesidades pertinentes del proyecto y con el fin de cumplir los objetivos de la investigación. La propuesta de solución seleccionada incluye como plataforma base a Cloudera QuickStarts VMs para el trabajo con *Big Data*, dentro de esta plataforma se seleccionó como herramienta de almacenamiento de *Big Data* a Hadoop, en específico su sistema de archivos HDFS (*Hadoop Distributed File System*, Sistema de archivos distribuidos de Hadoop), a Apache Pig como herramienta de procesamiento y análisis de los datos, una metodología de análisis visual (*"Visual Analytics*

Methodology") y por último, también se incluye a R como herramienta de visualización. Todo esto se ilustra en la Tabla 2.2.

Aspecto	Propuesta
Herramienta de almacenamiento	Sistema de archivos de Hadoop (HDFS)
Herramienta de procesamiento	Apache Pig
Metodología	*Visual Analytics Methodology*
Herramienta de visualización	R

Tabla 2.2 Propuesta de solución.

Trabajar con Hadoop/MapReduce requiere de una instalación y configuración muy compleja de manejar y entender, así como de la programación de trabajos MapReduce para el procesamiento de grandes datos, por lo que se decidió estudiar plataformas que facilitaran este trabajo. Este estudio llevo al descubrimiento de Cloudera, la cual es una plataforma escalable, flexible e integrada que facilita la gestión de grandes volúmenes de datos. Permite implementar y administrar Apache Hadoop y proyectos relacionados para manipular y analizar los datos, manteniéndolos seguros y protegidos. Es una máquina virtual que integra todo el ecosistema de Apache Hadoop, incluyendo los elementos principales que proporcionan un procesado de datos fiable y escalable [31].

Dicha máquina virtual proporciona el sistema de archivos de Hadoop (HDFS), del cual se hará uso para el almacenamiento de los conjuntos de datos *Big Data*. Se contempló el uso de PostgreSQL o Cassandra para esta tarea, sin embargo, al momento de analizar el flujo de trabajo, se determinó que no era necesario el uso de estas herramientas para el almacenamiento de los datos, puesto que era suficiente contar con los datos en el sistema de archivos de Hadoop.

Dentro de las herramientas de procesamiento de grandes datos, se decidió utilizar Apache Pig. Esta herramienta es una abstracción de MapReduce que proporciona Cloudera, por lo que facilita mucho el trabajo con los datos. Trabajar al mismo nivel que MapReduce hace que las labores de programación de los trabajos de análisis y procesamiento de datos sean demasiado complejas y muy difíciles de realizar. Pig proporciona un lenguaje basado en SQL que es más entendible y amigable para el programador, además de que cuenta con un proceso transparente que transforma los programas Pig en trabajos MapReduce para Hadoop [32].

También en esta investigación se utilizó la metodología de análisis visual [33], ya que dicho procedimiento proporciona los pasos necesarios para realizar el análisis visual de los datos de manera efectiva, como se muestra en la Figura 2.1; además, de acuerdo a la literatura revisada, no se cuenta con alguna otra metodología para la realización de dicho análisis.

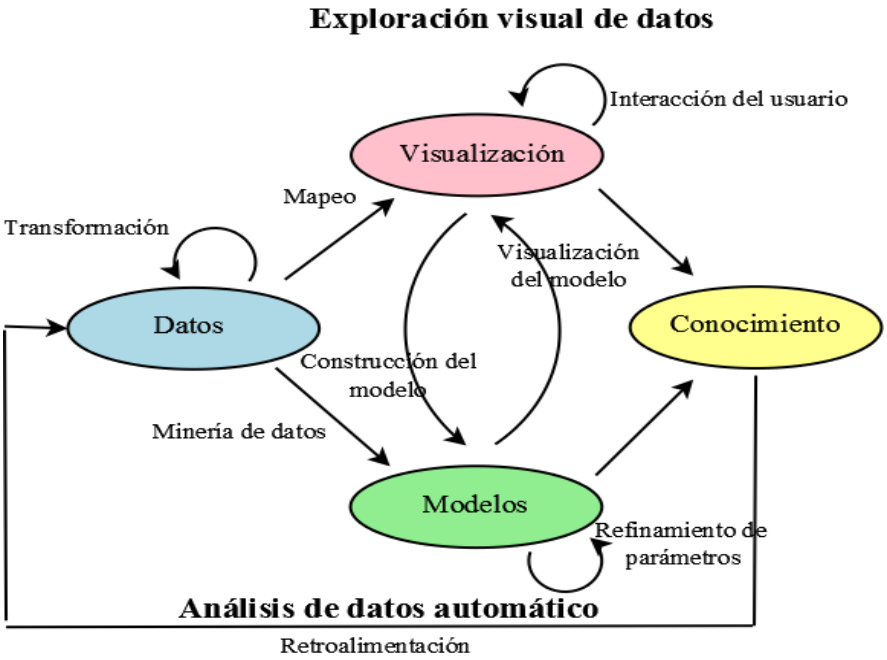


Figura 2.1 Metodología para el análisis visual.

Finalmente, se seleccionó a R como herramienta de visualización. R es un lenguaje y entorno para computación y gráficos estadísticos, inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland. R provee una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas de estadística clásica, análisis de series temporales, clasificación, agrupación, entre otras), técnicas gráficas y es altamente extensible, además es una herramienta *off-line* que facilita la interacción e intercambio de los datos [34]. Por otra parte, R también proporciona una integración con Hadoop para la lectura de los datos que están almacenados en HDFS, lo cual es perfecto para este proyecto de investigación.

Con la propuesta de solución planteada en los párrafos anteriores, lo cual llevó a la selección de las herramientas y plataforma a utilizar en esta investigación, se llevó a cabo el desarrollo de la solución a la problemática presentada y así procesar y visualizar los conjuntos de datos *Big Data*.

Capítulo 3. Aplicación de la Metodología

Con base en la problemática presentada en el punto 1.2 y a los objetivos planteados en el apartado 1.3 del capítulo uno respectivamente, es necesario realizar el análisis comparativo de las técnicas de visualización que se ocupan en *Big Data*, con el fin de determinar cuáles son las más idóneas para el análisis de grandes conjuntos de datos. Es por ello que en este capítulo se describen las actividades realizadas para cumplir con el objetivo de esta investigación.

3.1. Análisis de técnicas de visualización y conjuntos de datos para *Big Data*

En esta sección se presenta el proceso de análisis que se llevó a cabo a los diferentes trabajos relacionados con las técnicas de visualización y conjuntos de datos para *Big Data*, así como el resultado del mismo, con el cual se conocieron interesantes tópicos relacionados en el campo de *Big Data*, por ejemplo, las editoriales que más trabajos tienen en visualización de grandes datos, los dominios más utilizados, entre otros.

3.1.1. Proceso de análisis

Para realizar el proceso análisis de las técnicas de visualización que se emplean en *Big Data*, se siguieron los pasos que se muestran en la Figura 3.1.

Como se observa en la figura, se realizó la búsqueda de artículos en las principales editoriales de investigación, las cuales son: ACM, IEEE, Springer y Elsevier, también se tomaron en cuenta otras editoriales siempre y cuando la revista tuviera un factor de impacto.

Una vez ingresado en los repositorios correspondientes, se realizó la búsqueda de los trabajos de acuerdo a los siguientes criterios de selección: palabras clave de búsqueda *"Big Data visualization"* y publicaciones entre 2010 y 2017. Se descartaron artículos de otros idiomas que no fueran en inglés, además de tesis de maestría y/o doctorado.

Después de descargar los artículos seleccionados, se procedió a analizar cada uno de ellos buscando los conjuntos de datos que se ocupaban dentro del artículo, así como las técnicas de visualización empleadas sobre dichos conjuntos de datos. Se agruparon las técnicas conforme a su clasificación y los conjuntos de acuerdo al dominio de los datos. Cabe mencionar que los dominios utilizados en esta clasificación se seleccionaron de acuerdo con *"Research Trends"*

de Elsevier [35], los cuáles son: (1) Ciencias de la Computación, (2) Ingeniería, (3) Matemáticas, (4) Negocios, Gestión y Contabilidad, (5) Física y Astronomía, (6) Bioquímica, Genética y Biología Molecular, (7) Ciencias Sociales, (8) Ciencia de los Materiales, (9) Medicina, (10) Ciencias de la Decisión, (11) Multidisciplinario, (12) Artes y Humanidades.

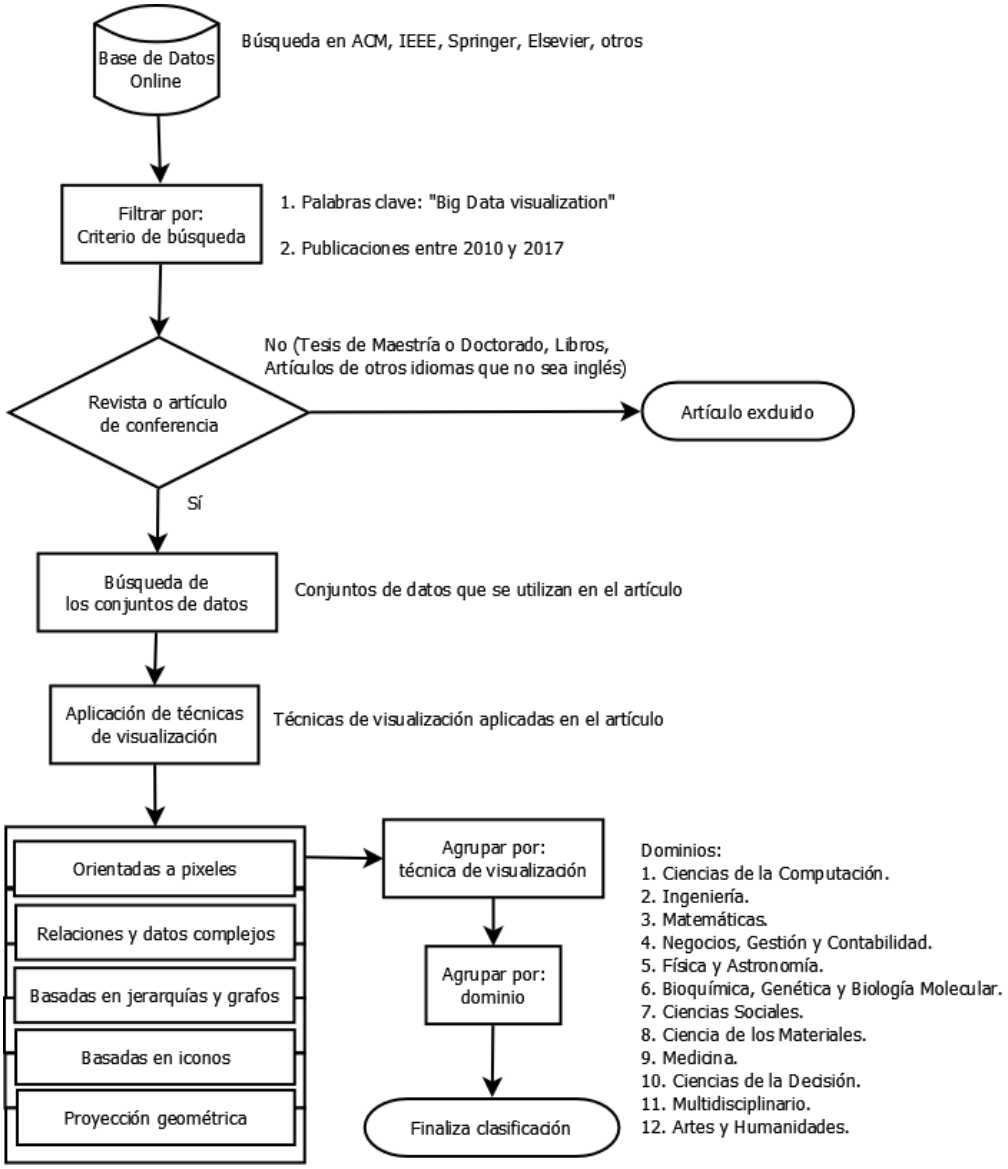


Figura 3.1 Metodología de análisis

3.1.2. Resultado del análisis

En las siguientes figuras se presenta el resultado del análisis planteado en el punto anterior. Se describen importantes tópicos que se tomaron en cuenta para la realización del análisis comparativo de técnicas de visualización y conjuntos de datos para *Big Data*.

En la Figura 3.2 se aprecia cómo aumentó el interés por el análisis visual en el campo de *Big Data* en los últimos seis años, esto se debe por ejemplo, al aumento en el interés de analizar y visualizar grandes cantidades de datos provenientes de sensores, ya sea de tráfico o geográficos [36]–[41]. En este análisis se tomó en cuenta un total de 83 artículos en este campo de investigación.

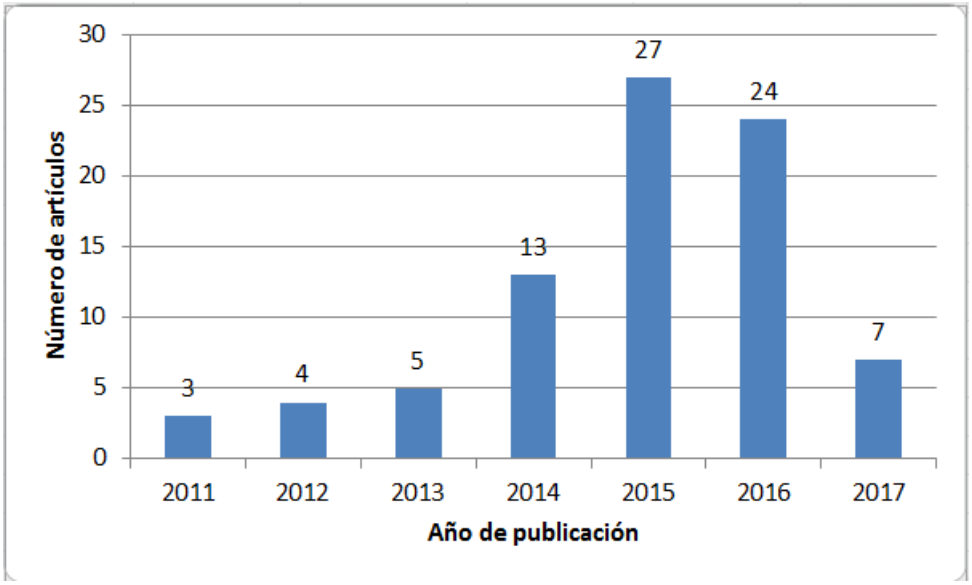


Figura 3.2 Distribución de artículos por año de publicación.

Es importante conocer cómo se distribuyeron dichos artículos de acuerdo a su editorial de publicación, la cual se muestra en la Figura 3.3. En dicha figura se observa que la editorial IEEE es donde se concentran la mayoría de los trabajos relacionados con la visualización de *Big Data*.

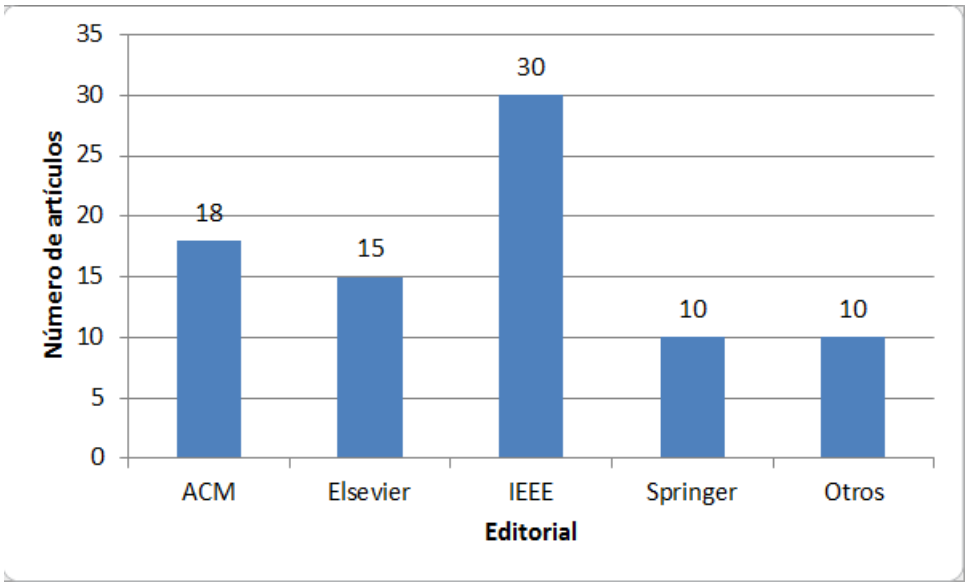


Figura 3.3 Distribución de artículos por editorial.

Todos los conjuntos de datos utilizados en los artículos de visualización de *Big Data* se clasificaron de acuerdo a la lista de dominios presentada en el punto 3.1.1. La Figura 3.4 muestra esta clasificación, donde se observa que la mayoría de los conjuntos de datos son referentes a Ciencias de la Computación [41]–[55] seguido por el dominio Multidisciplinario [36], [56]–[68].

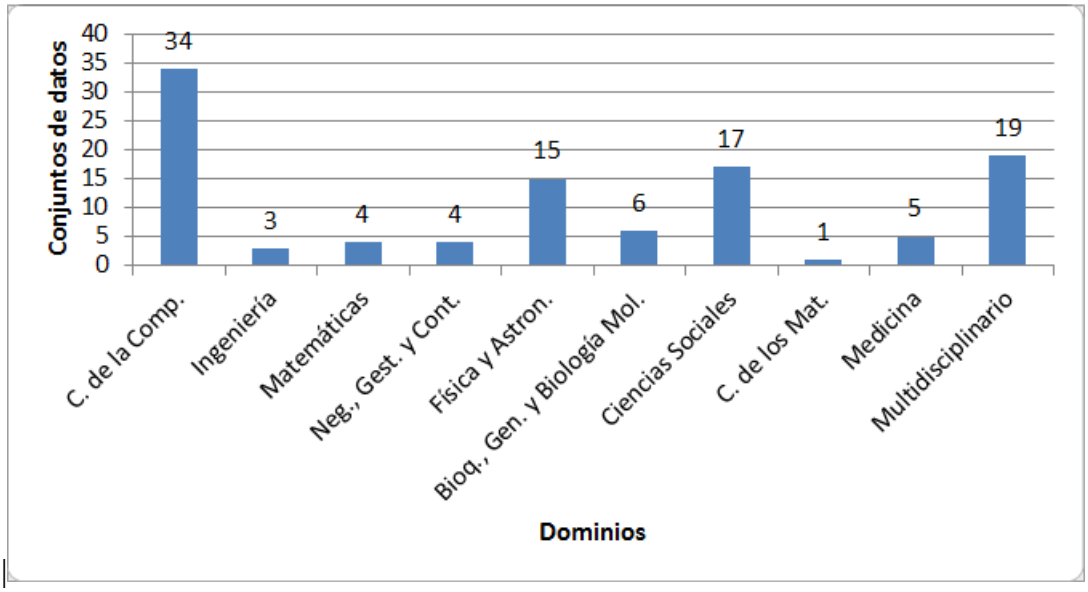


Figura 3.4 Clasificación de los conjuntos de datos de acuerdo a su dominio.

También es importante considerar la distribución de las editoriales de los artículos con base en el dominio de los conjuntos de datos que utiliza. En la Figura 3.5 se visualiza esta distribución en la cual se aprecia que la editorial ACM junto con la editorial Springer tiene el mayor número de conjuntos de datos relacionados con Ciencias de la Computación, así mismo se observa que la editorial IEEE posee diez conjuntos de datos del dominio Multidisciplinario.

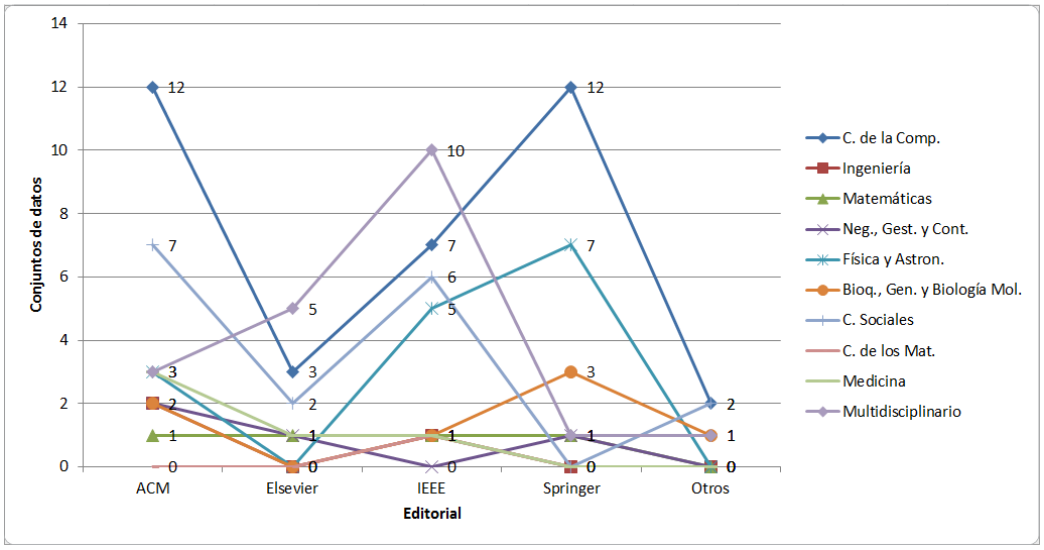


Figura 3.5 Clasificación de los dominios por editorial.

Como se mencionó anteriormente, las técnicas de visualización se clasificaron en: Basadas en iconos, Basadas en jerarquías y grafos, Orientadas a pixeles, Proyección geométrica y Relaciones y datos complejos, por lo que también fue necesario distribuir los conjuntos de datos de acuerdo con la técnica de visualización aplicada en sus respectivos artículos. La Figura 3.6 presenta dicha clasificación, mostrando que las técnicas más utilizadas son las de Relaciones y datos complejos y de Proyección geométrica. Dentro de las técnicas de Relaciones y datos complejos se encuentra que las más utilizadas son *Spatial Information Flow* [36]–[38], [40], [53], [61], [62], [66]–[74] y Clustergram [40], [42], [75]–[78], mientras que dentro de Proyección geométrica, las técnicas más implementadas son los gráficos de líneas [39], [41], [43], [47], [61], [62], [70], [79], de dispersión [47], [57]–[59], [62], [79], [80] y de barras [43], [47], [49], [52], [60], [62], [64], [68], [69], [72], [79], [81]–[83].

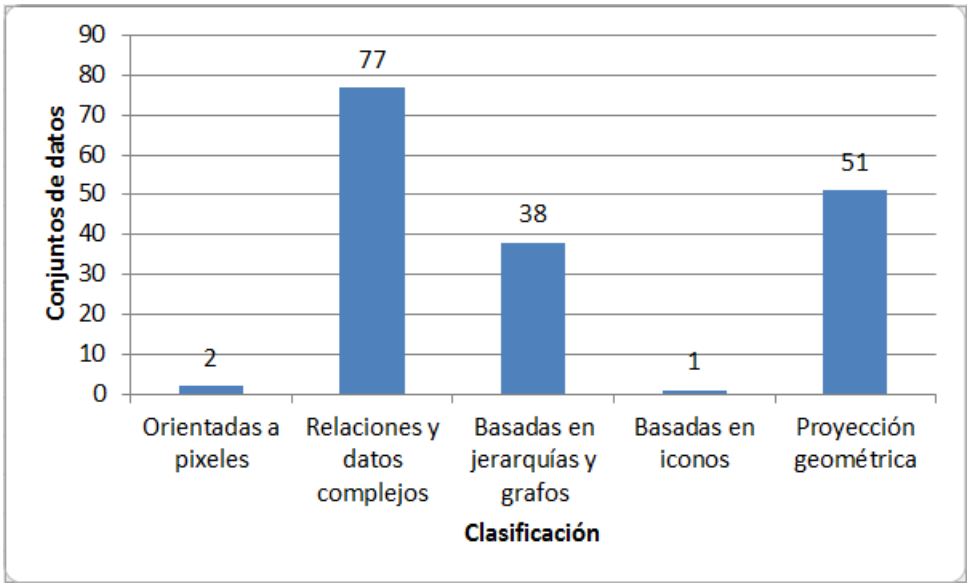


Figura 3.6 Clasificación de técnicas de visualización.

También es importante conocer en qué editoriales se concentran los trabajos de acuerdo con la técnica de visualización empleada. La Figura 3.7 muestra esta compilación, donde se observa que la editorial ACM contiene el mayor número de conjuntos de datos en donde se aplica técnicas de Proyección geométrica. Por otra parte, se aprecia que en la editorial IEEE junto con ACM, se concentran mayormente los conjuntos de datos en donde se implementan técnicas de Relaciones y datos complejos.

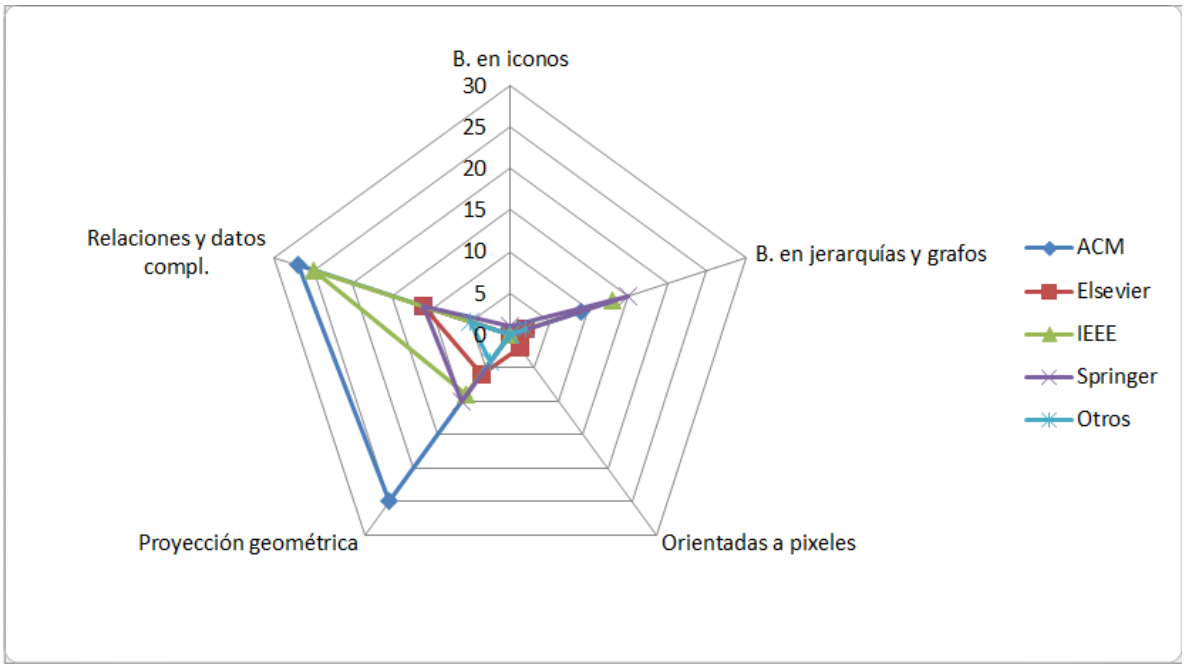


Figura 3.7 Distribución de las técnicas de visualización de acuerdo a la editorial.

Asimismo es importante catalogar las técnicas de visualización que se aplican a los conjuntos de datos de acuerdo con el dominio de los mismos. La Figura 3.8 muestra esta clasificación, en donde se observa que para los conjuntos de datos de Ciencias de la Computación se aplican en su mayoría técnicas Basadas en jerarquías y grafos [45], [46], [49], [51], [54], mientras que para conjuntos de Ciencias Sociales y Física y Astronomía se aplican técnicas de Relaciones y datos complejos [40], [42], [70], [75], [80], [84]–[88].

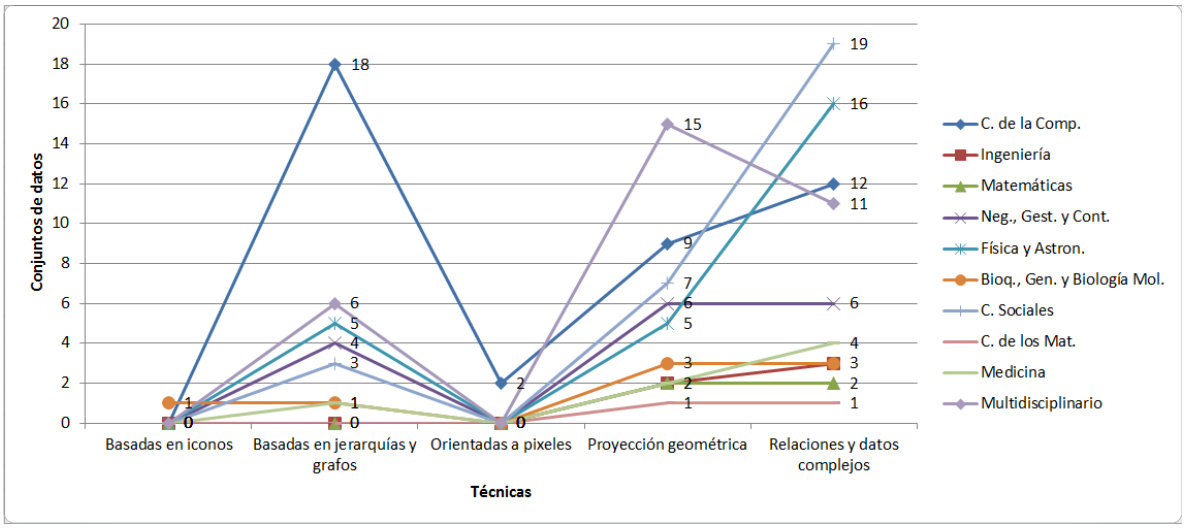


Figura 3.8 Clasificación de técnicas de visualización de acuerdo al dominio.

Finalmente, se expone la distribución de las herramientas y lenguajes de programación que se encontraron en esta revisión de la literatura. La Figura 3.9 presenta la cantidad de artículos que usaron cada herramienta y lenguaje, mostrando que la plataforma de Hadoop/MapReduce es la preferida a la hora de realizar análisis de *Big Data*, dado los beneficios que ésta proporciona al momento de procesar grandes cantidades de datos. Por otra parte, se observa que lenguajes de programación como Python y Java están presentes en el desarrollo de sistemas referentes a *Big Data*. La sección de "Otros" se refiere a herramientas y lenguajes que sólo se mencionaron en un artículo.

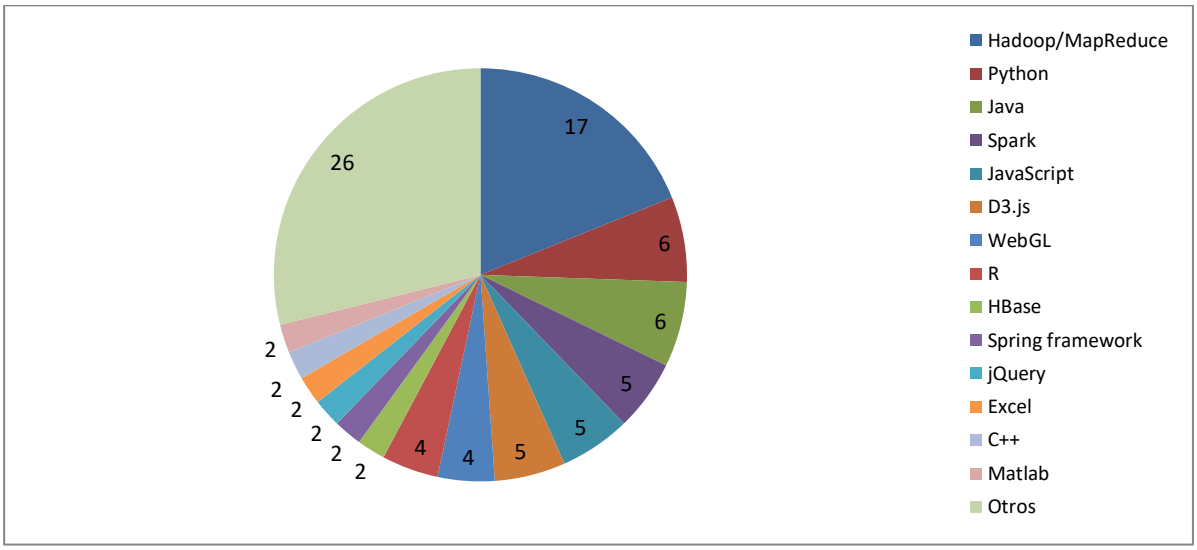


Figura 3.9 Distribución de herramientas y lenguajes de programación.

3.2. Selección de las técnicas de visualización

Con el análisis realizado en el apartado 3.1, se procede a seleccionar las técnicas de visualización más utilizadas en *Big Data* en los 108 conjuntos de datos que se encontraron. La Figura 3.10 muestra la distribución de las técnicas de visualización con base en su clasificación y el número de veces que se aplicaron en los conjuntos de datos.

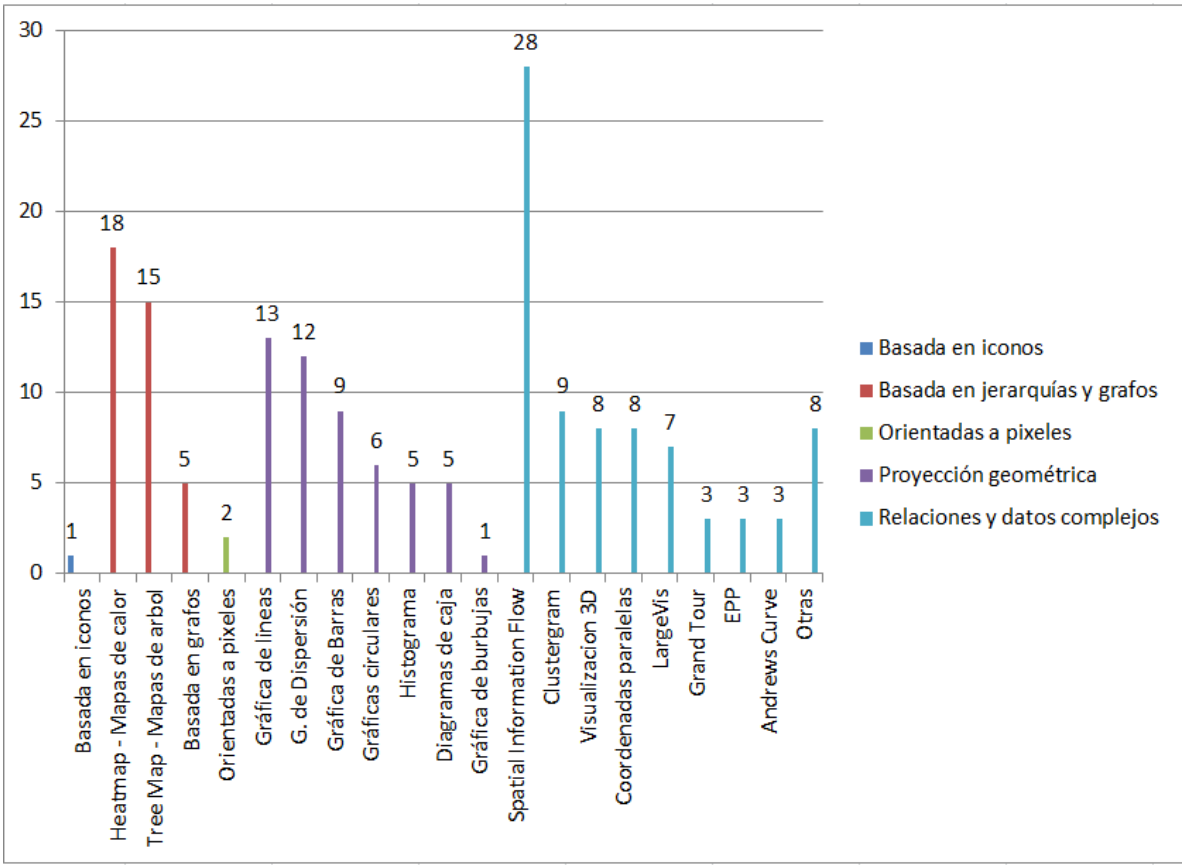


Figura 3.10 Clasificación de técnicas de visualización.

Se procedió a seleccionar las técnicas de visualización más utilizadas de cada clasificación, dando como resultado la Tabla 3.1.

Clasificación	Técnica de visualización
Basada en iconos	Descartado
Basada en jerarquías y grafos	*Heatmap* / *Treemap*
Orientadas a pixeles	Descartado
Proyección geométrica	Gráfica de líneas / de dispersión / de barras
Relaciones y datos complejos	*Spatial Information Flow* / Clustergram / Coordenadas paralelas

Tabla 3.1 Selección de las técnicas de visualización.

Las técnicas Basada en iconos y Orientadas a pixeles se descartaron en el análisis comparativo, ya que solo se utilizaron en un total de tres conjuntos de datos como se aprecia en la Figura 3.10, por lo que no representa algo relevante para incluirlo en el análisis comparativo.

3.3. Determinación de los conjuntos de datos

Después del proceso de análisis que se realizó en el punto 3.1, se procedió a seleccionar los conjuntos de datos más utilizados en esta revisión. Estos conjuntos de datos se clasificaron de acuerdo al modelo de 3V's de *Big Data*, dado que para que un conjunto de datos se considere *Big Data* debe cumplir con al menos dos de las tres V (Volumen, Velocidad y Variedad) [16]. La Figura 3.11 presenta esta clasificación en donde se observa que no todos los conjuntos de datos encontrados cumplen con la condición antes mencionada, específicamente 55 conjuntos de datos solo cumplen con una o ninguna V. Por consiguiente los otros 53 conjuntos de datos son los que importan para esta investigación.

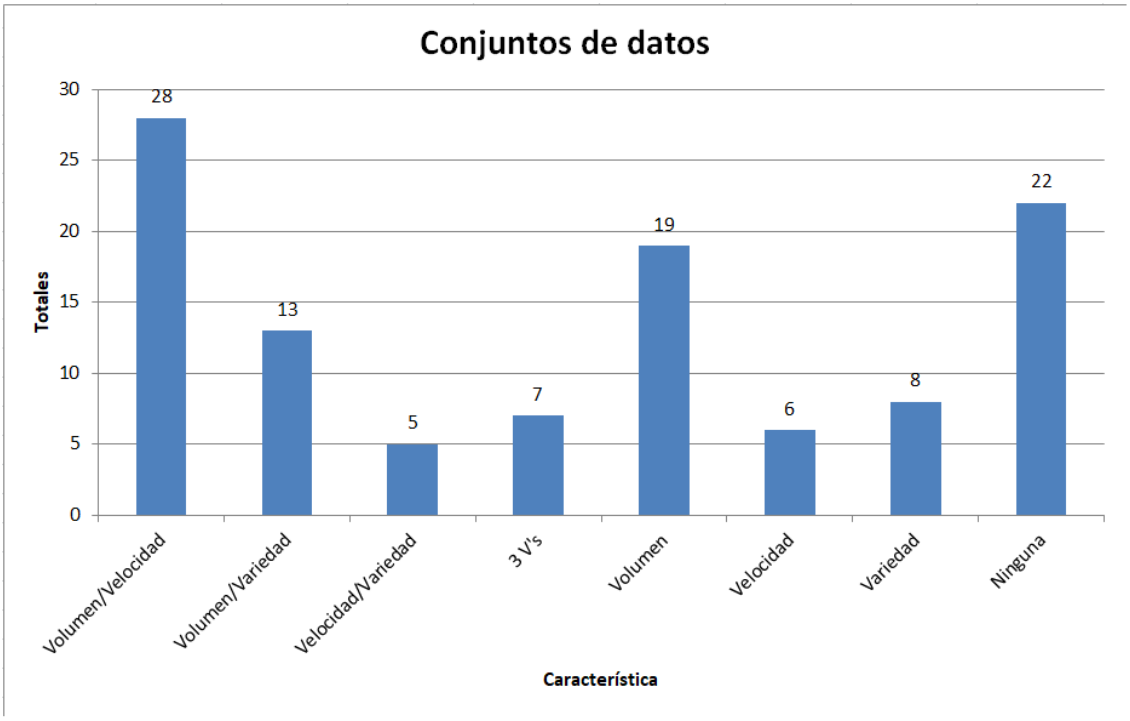


Figura 3.11 Clasificación de los conjuntos de datos de acuerdo al modelo 3V.

La selección de los conjuntos de datos se aprecia en la Tabla 3.2, los cuales se eligieron con base en la disponibilidad, importancia y uso dentro de los artículos revisados. Estos conjuntos de datos son los que se utilizaron para implementar las técnicas de visualización seleccionadas y así realizar el análisis comparativo.

Se descartó la combinación de las características Velocidad/Variedad debido a que de los cinco conjuntos de datos encontrados, ninguno está disponible, por lo que fue imposible

obtener algún conjunto de datos de dicha combinación. Sin embargo, se decidió utilizar otro conjunto de datos que contempla las tres V's.

Característica	Conjunto de datos
Volumen/Velocidad	Principales exportaciones e importaciones por país [89]
Volumen/Variedad	Artículos de Wikipedia [90]
Velocidad/Variedad	Descartado
3 V's	Información sobre el blog LiveJournal [91] / YFCC100M - Imágenes y videos de Flickr extraídos por Yahoo! [92]

Tabla 3.2 Selección de conjuntos de datos.

Es importante mencionar que el uso de estos conjuntos de datos está sujeto a que sea posible utilizarlos con las herramientas mencionadas en el apartado 2.3, en dado caso que estos datos no sean compatibles con dichas herramientas, se hará uso de *Benchmarks* para *Big Data* para la generación de datos sintéticos que sirvan para la construcción de las gráficas correspondientes.

De forma general, los *Benchmarks* son herramientas que sirven para medir el rendimiento de un sistema o uno de sus componentes, ejecutando programas o conjunto de programas en una máquina, con el objetivo de estimar el rendimiento de un elemento concreto [93].

Con el surgimiento de sistemas *Big Data*, es necesario que estas herramientas también evolucionen hacia esa dirección, dando como resultado plataformas como BigBench [94] o BigDataBench [95]. Dentro de este contexto, estas plataformas se desarrollaron con el fin de abordar los problemas de probar y comparar dichos sistemas, y así medir su rendimiento, eficiencia y rentabilidad.

El modelo de negocio de BigBench es la venta de productos de diferentes proveedores. Contempla un modelo de datos y un generador de datos sintéticos que abordan los aspectos de volumen, variedad y velocidad de los sistemas *Big Data*, generando datos estructurados, semi-estructurados y no estructurados. Este generador de datos diseñado para BigBench es completamente escalable [96].

Por otra parte, BigDataBench es otro *Benchmark* para *Big Data* que también contiene un generador de datos, denominado BDGS (*Big Data Generator Suite*, suite de generador de grandes datos) que ayuda en la generación de conjuntos de datos con características de volumen, variedad, velocidad y veracidad, mientras se emplean modelos de datos para

capturar y preservar las características importantes de los datos reales durante la generación de los mismos [97].

3.4. Desarrollo del ambiente para *Big Data*

Como se mencionó en la propuesta de solución, específicamente en el punto 2.3, se hizo uso de la plataforma Cloudera para el ambiente de *Big Data*, esto se debe a las facilidades que presta dicha plataforma para la administración de Hadoop/MapReduce y que proporciona diversas herramientas para el procesamiento de los datos, como lo es Apache Pig.

Dado que Cloudera es una máquina virtual, solo se requiere de la instalación de Oracle VM VirtualBox para que se instale. La Figura 3.12 muestra las características proporcionadas en la instalación de la máquina virtual de Cloudera.

Cloudera no requiere de configuraciones complejas para que funcione Hadoop. Desde el momento que está iniciando la máquina virtual, los servicios de Hadoop empiezan a ejecutarse para que el servicio esté disponible al momento de iniciar Cloudera, la Figura 3.13 muestra los servicios de Hadoop iniciando al mismo momento de ejecutar la máquina virtual.

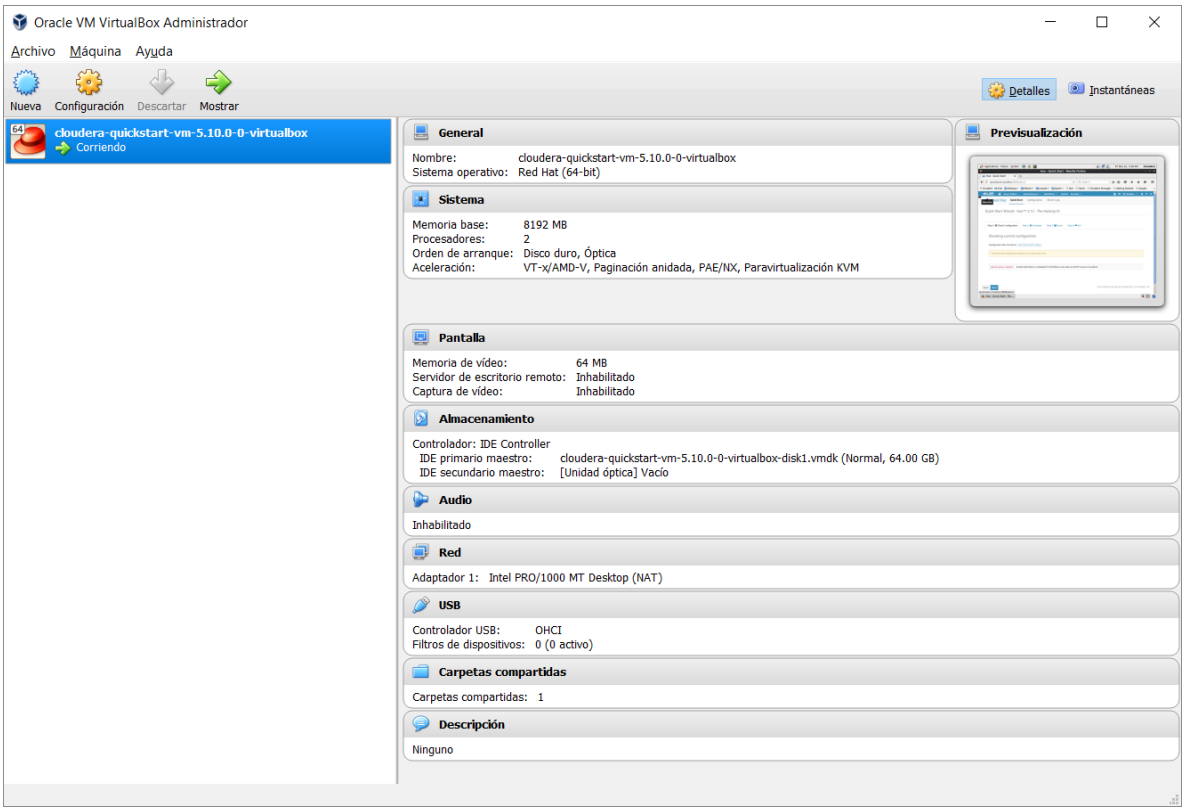


Figura 3.12 Características de Cloudera.

```
Using config: /etc/zookeeper/conf/zoo.cfg
Starting zookeeper ... STARTED
Starting abrt daemon:                                      [  OK  ]
starting datanode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-datanode-quicksta
rt.cloudera.out
Started Hadoop datanode (hadoop-hdfs-datanode):            [  OK  ]
starting journalnode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-journalnode-qu
ickstart.cloudera.out
Started Hadoop journalnode:                                [  OK  ]
starting namenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-namenode-quicksta
rt.cloudera.out
Started Hadoop namenode:                                   [  OK  ]
starting secondarynamenode, logging to /var/log/hadoop-hdfs/hadoop-hdfs-secondar
ynamenode-quickstart.cloudera.out
Started Hadoop secondarynamenode:                          [  OK  ]
Started Hadoop httpfs (hadoop-httpfs):                     [  OK  ]
starting historyserver, logging to /var/log/hadoop-mapreduce/mapred-mapred-histo
ryserver-quickstart.cloudera.out
Started Hadoop historyserver:                              [  OK  ]
starting nodemanager, logging to /var/log/hadoop-yarn/yarn-yarn-nodemanager-quic
kstart.cloudera.out
Started Hadoop nodemanager:                                [  OK  ]
starting proxyserver, logging to /var/log/hadoop-yarn/yarn-yarn-proxyserver-quic
kstart.cloudera.out
```

Figura 3.13 Iniciando los servicios de Hadoop al momento de ejecutar Cloudera.

Para una mayor facilidad en la administración de Hadoop, Cloudera proporciona una herramienta gráfica llamada Hue (*Hadoop User Experience*, Experiencia de Usuario Hadoop), la cual es una interfaz vía Web de código abierto que permite el acceso y manipulación de todo el ecosistema de Hadoop, incluyendo el acceso a Apache Pig y al sistema de archivos HDFS. La Figura 3.14 muestra la página inicial de Hue.

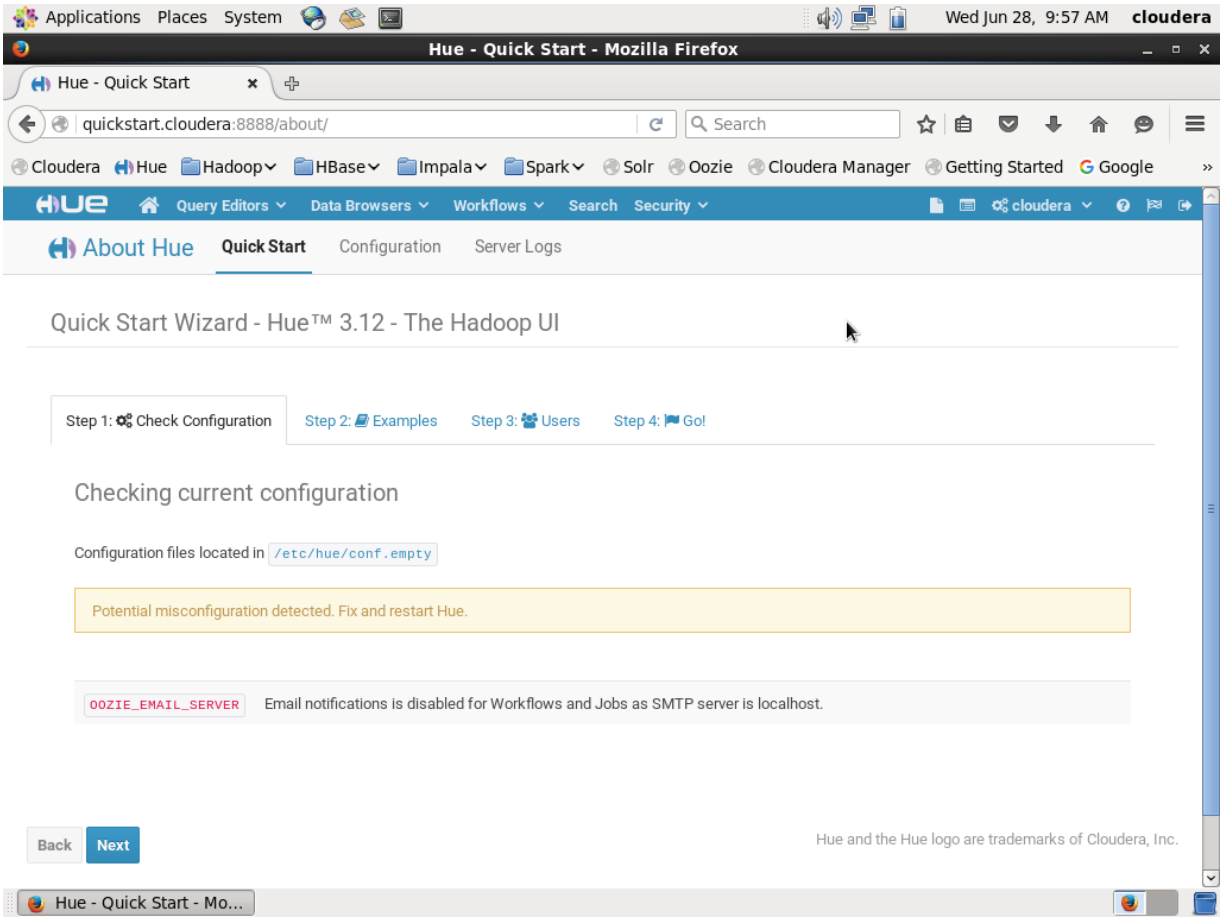


Figura 3.14 Página principal de Hue.

El uso de esta interfaz gráfica también facilita el trabajo de carga de archivos a HDFS, ya no se requiere del uso de línea de comandos, simplemente se carga el archivo como cualquier sistema conocido vía Web. La Figura 3.15 muestra la carga de datos al sistema de archivos HDFS.

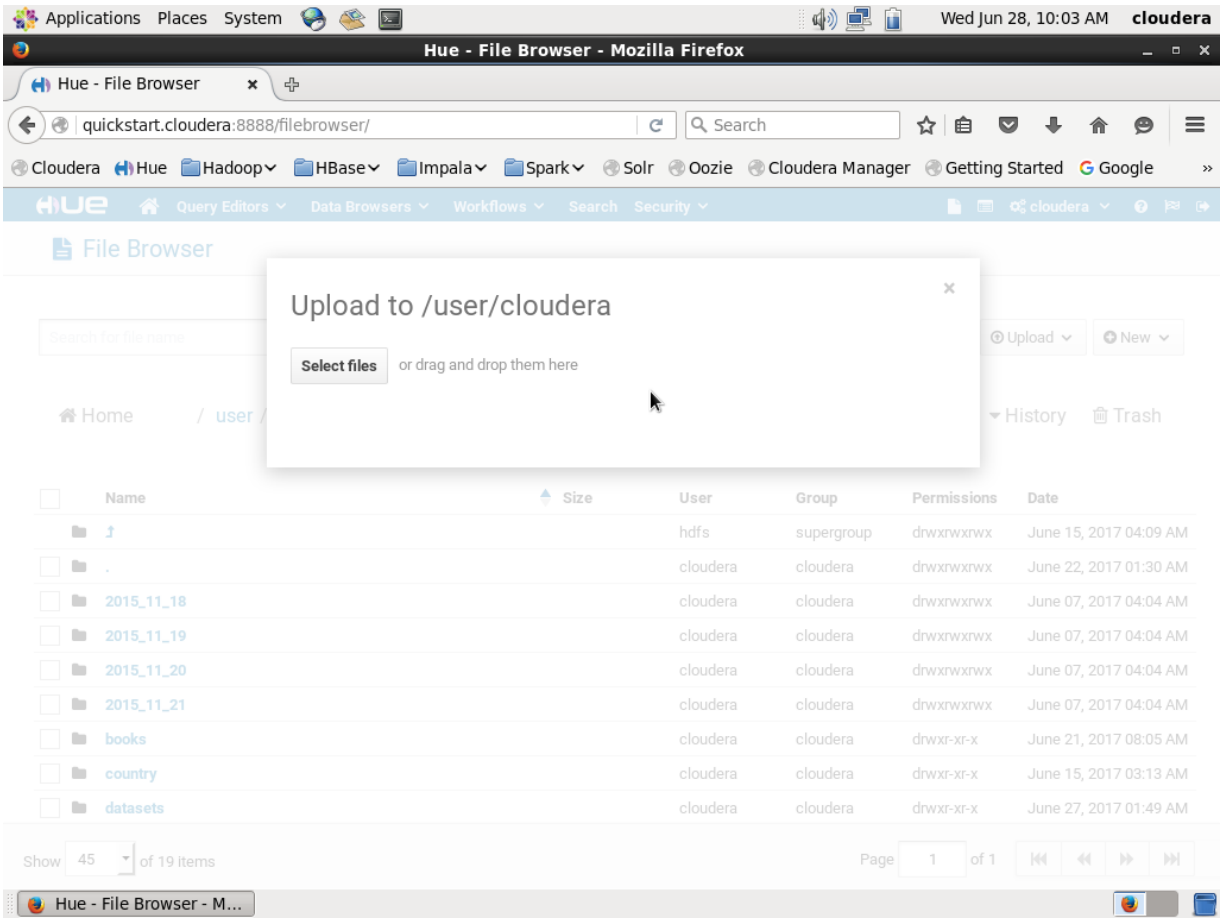


Figura 3.15 Interfaz gráfica para carga de datos al sistema HDFS.

En este punto ya se tiene la plataforma base para el almacenamiento y procesamiento de *Big Data*, por lo cual ahora se requiere la instalación de R para la visualización de los datos. Junto con R también se incluye la instalación de RStudio para un mejor manejo de la herramienta. Ya con la instalación de dichas herramientas, es necesario realizar la integración entre R y Hadoop para que R sea capaz de acceder al sistema de archivos HDFS, lo cual se logra con la instalación de los paquetes rmr2 y rhdfs [98] con sus respectivas dependencias, además de la configuración de las variables de entorno necesarias (HADOOP_HOME, HADOOP_CMD y HADOOP_STREAMING). La Figura 3.16 expone la interfaz gráfica de RStudio con las variables configuradas correctamente y los paquetes de integración con Hadoop cargados y listos para el trabajo entre estas dos plataformas.

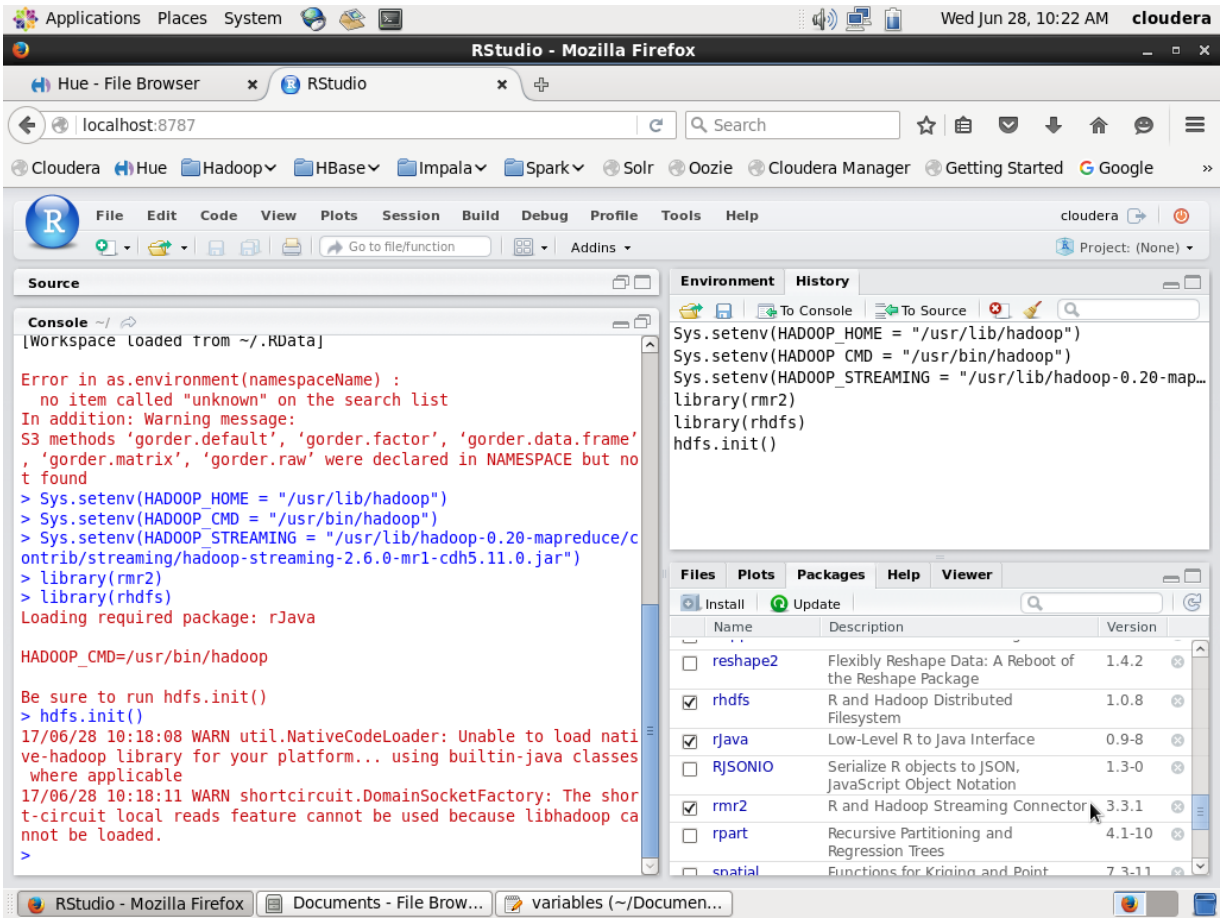


Figura 3.16 Integración entre R y Hadoop.

Con la instalación de Cloudera, R y la integración entre Hadoop y R, se concluye la parte de construcción del ambiente para análisis, procesamiento y visualización de *Big Data*, lo que indica que el sistema está listo para la realización de las pruebas pertinentes en el análisis comparativo.

3.5. Determinación del marco de evaluación

Es necesario determinar los puntos que se tienen que evaluar al momento de realizar las diferentes visualizaciones de los conjuntos de datos, esto permitirá tener una visión clara al momento de hacer el análisis comparativo.

Después de haber analizado los diferentes artículos para conocer las técnicas de visualización más utilizadas en *Big Data* así como los conjuntos de datos, se determinaron los siguientes puntos para el marco de evaluación:

- **Número de variables procesadas.-** Corresponde al número total de variables que se están procesando en el/los conjuntos de datos.
- **Número de tuplas procesadas.-** Corresponde al número total de registros que se están procesando en el/los conjuntos de datos.
- **Número de variables visualizadas.-** Presenta el número total de variables utilizadas para generar la gráfica correspondiente.
- **Número de tuplas visualizadas.-** Presenta el número total de tuplas que se utilizaron para generar la gráfica correspondiente.
- **Curva de aprendizaje.-** Se refiere al grado de complejidad (Fácil, Medio y Difícil) que constituye la construcción del modelo completo, incluyendo desde el análisis, procesamiento y hasta visualización de los datos.
- **Tiempo de construcción.-** Indica el tiempo máximo que requiere un desarrollador para la construcción del modelo completo.
- **Tiempo de visualización.-** Indica el tiempo promedio que requiere R para la generación de la gráfica desde que se introduce el comando indicado. El promedio se toma de 10 iteraciones.
- **Facilidad de interpretación.-** Muestra el grado de complejidad (Fácil, Medio, Difícil) al momento de interpretar los datos en la gráfica generada.
- **Volumen.-** Indica si el conjunto de datos tiene la característica de volumen.
- **Velocidad.-** Indica si el conjunto de datos tiene la característica de velocidad.
- **Variedad.-** Indica si el conjunto de datos tiene la característica de variedad.
- **Veracidad.-** Indica si el conjunto de datos tiene la característica de veracidad.

Con la lista de puntos mencionados anteriormente, se tiene listo el marco de evaluación que permitirá valorar las implementaciones de las técnicas de visualización y así realizar el análisis comparativo.

3.6. Proceso para la implementación de las técnicas de visualización

En este punto, ya con el ambiente para el procesamiento de los datos, la herramienta para la visualización de los mismos y el marco de evaluación definido, se procede a plantear el

proceso para la implementación de las técnicas de visualización. La Figura 3.17 muestra de forma general, el enfoque para trabajar con el análisis y visualización de grandes datos, haciendo uso de las herramientas antes mencionadas.

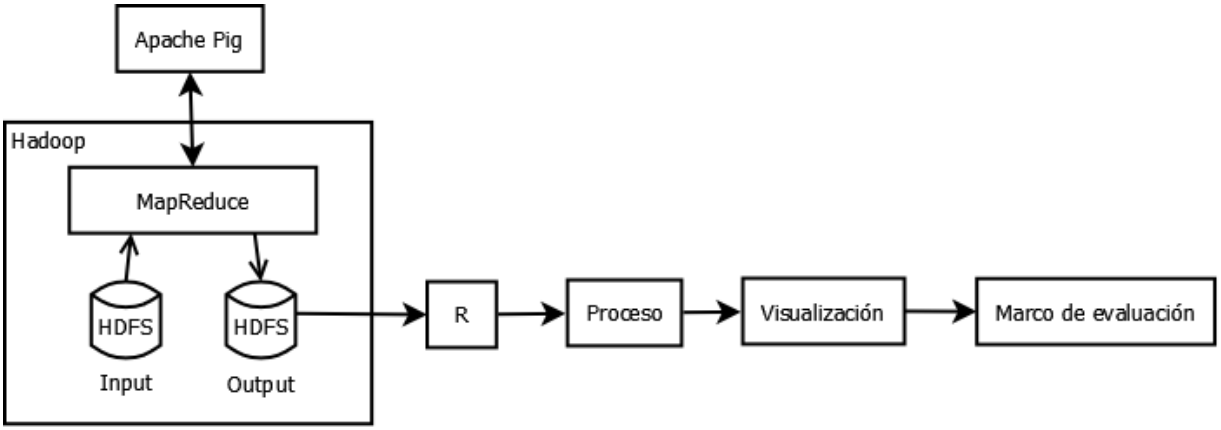


Figura 3.17 Esquema de trabajo para el análisis y visualización de *Big Data*.

Cada conjunto de datos se almacenó en HDFS, posteriormente con la programación de los *scripts* en Apache Pig, se procesaron para resumir la gran cantidad de datos posible, estos datos se resumen dependiendo de la técnica de visualización que se quiera implementar y de la información que se desee mostrar. Los *scripts* de Apache Pig se transforman en trabajos MapReduce, lo cual facilita el trabajo con esta tecnología y que finalmente almacenan el resultado en un nuevo archivo en HDFS.

Con este nuevo archivo almacenado en HDFS, ahora se programa un *script* en R para su visualización. Cada *script* tiene los comandos necesarios para implementar cada una de las técnicas de visualización seleccionadas (gráfica de líneas, de barras, de dispersión, *Heatmap*, *Treemap*, Clustergram, coordenadas paralelas y *Spatial Information Flow*). Una vez procesada esta información por R, se genera la gráfica correspondiente para finalmente, recopilar la información necesaria para el marco de evaluación.

Es importante mencionar que la implementación y análisis de las técnicas de visualización para *Big Data* se hicieron sobre la máquina virtual Cloudera, instalada sobre una computadora con sistema operativo Windows™ 10 Home, procesador Intel® Core™ i5 de sexta generación a 2.3 GHz, memoria RAM de 12 GB y tarjeta gráfica integrada Intel® HD Graphics 520.

Capítulo 4. Resultados

El objetivo principal de esta investigación es la realización de un análisis comparativo de técnicas de visualización para *Big Data*, por lo que en este capítulo se presentan cuatro conjuntos de datos *Big Data* a los cuales se les aplicó las técnicas de visualización seleccionadas. Se muestra el procesamiento de los datos así como de las gráficas resultantes. Finalmente se expone una tabla comparativa de estos conjuntos de datos de acuerdo al marco de evaluación propuesto en el apartado 3.5 del capítulo tres, con lo cual se logra dicho análisis comparativo.

Es importante mencionar que, de los conjuntos de datos que se habían seleccionado anteriormente, solo se logró trabajar con el de "exportaciones e importaciones por país", los otros conjuntos de datos requerían del uso de otras herramientas para manipular la información, por lo que se tomó la decisión de descartarlos. Sin embargo, se trabajó con datos generados por medio de *Benchmarks* de *Big Data*, lo cual garantiza que los datos cumplan con las características principales de *Big Data*.

4.1. *Dataset* 1: Exportaciones e importaciones por país

Este primer conjunto de datos trata sobre las principales exportaciones e importaciones de varios países, la información está representada por año, código del país, código del producto, valor de exportación, valor de importación, entre otros, además cuenta con 5,108,305 registros, por lo que se cataloga que este conjunto de datos tiene las características de volumen y velocidad por la creciente generación de información. También se contempla la característica de veracidad, puesto que son datos reales de cada país presentado. La Figura 4.1 muestra los primeros registros de este conjunto de datos ya cargado en el sistema de archivos HDFS.

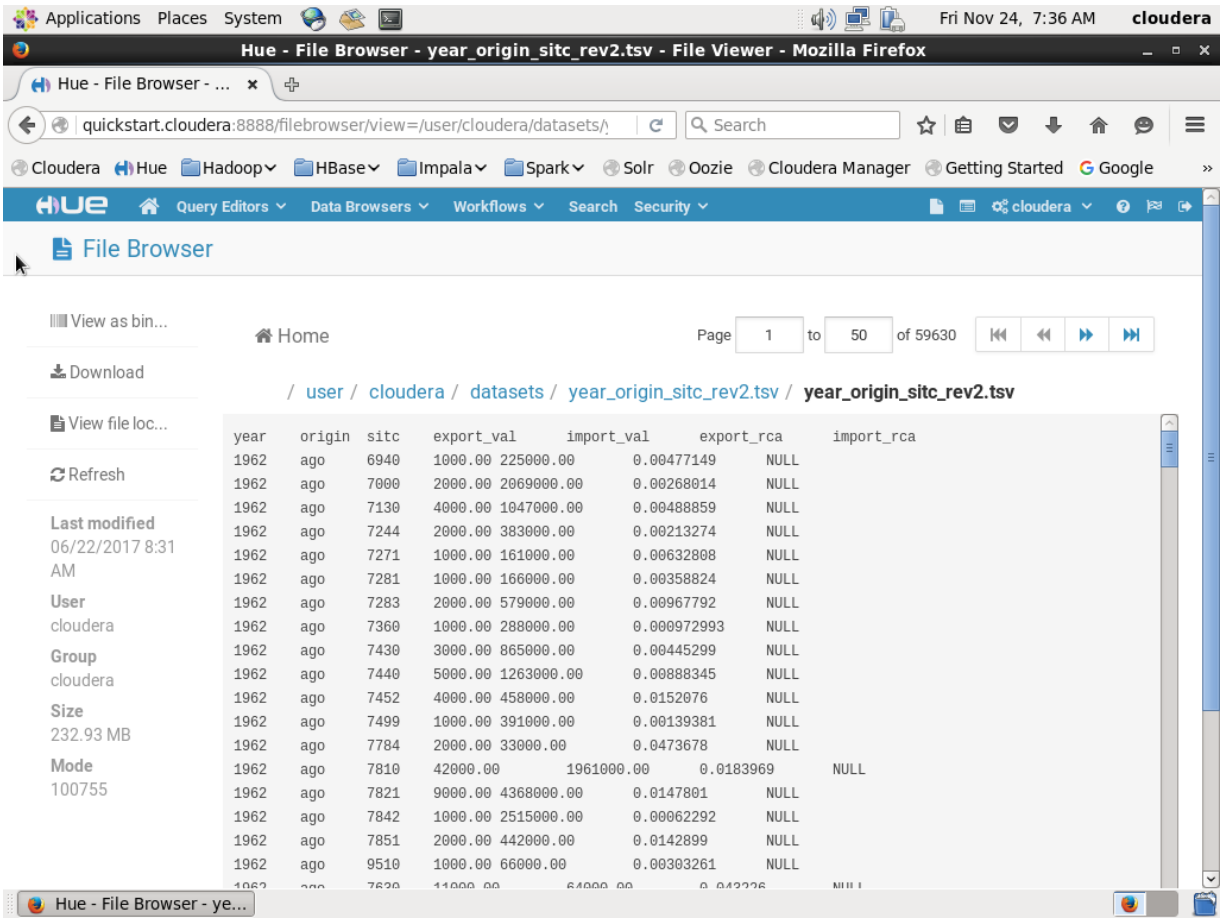


Figura 4.1 Conjunto de datos "exportaciones e importaciones por país".

Para empezar a trabajar con la generación de las gráficas, es necesario procesar primero los datos, lo cual conlleva a trabajar con Apache Pig para esta tarea. La Figura 4.2 muestra el *script* utilizado para resumir los datos.

La función principal de este *script* es la de agrupar la información de acuerdo al país y año, una vez agrupada, se procede a realizar la suma de sus exportaciones e importaciones, guardando el resultado en un nuevo archivo en HDFS. La Figura 4.3 presenta el archivo resultante que será entrada para R.

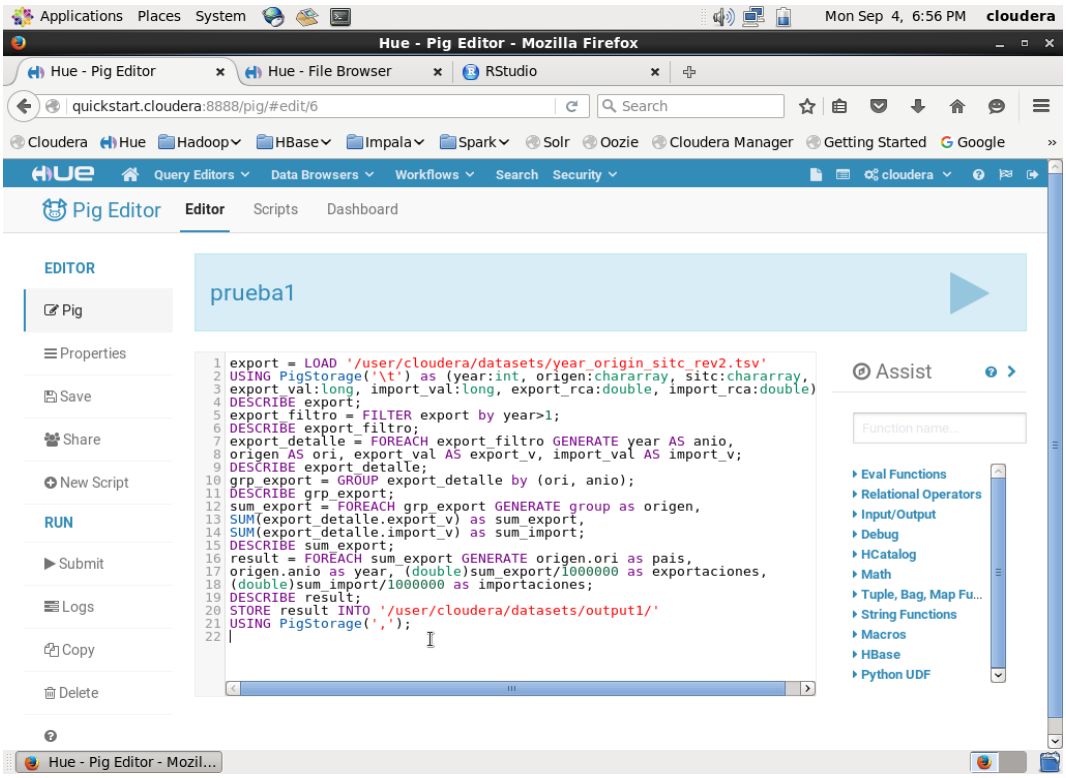


Figura 4.2 *Script* 1 de Apache Pig para el conjunto de datos 1.

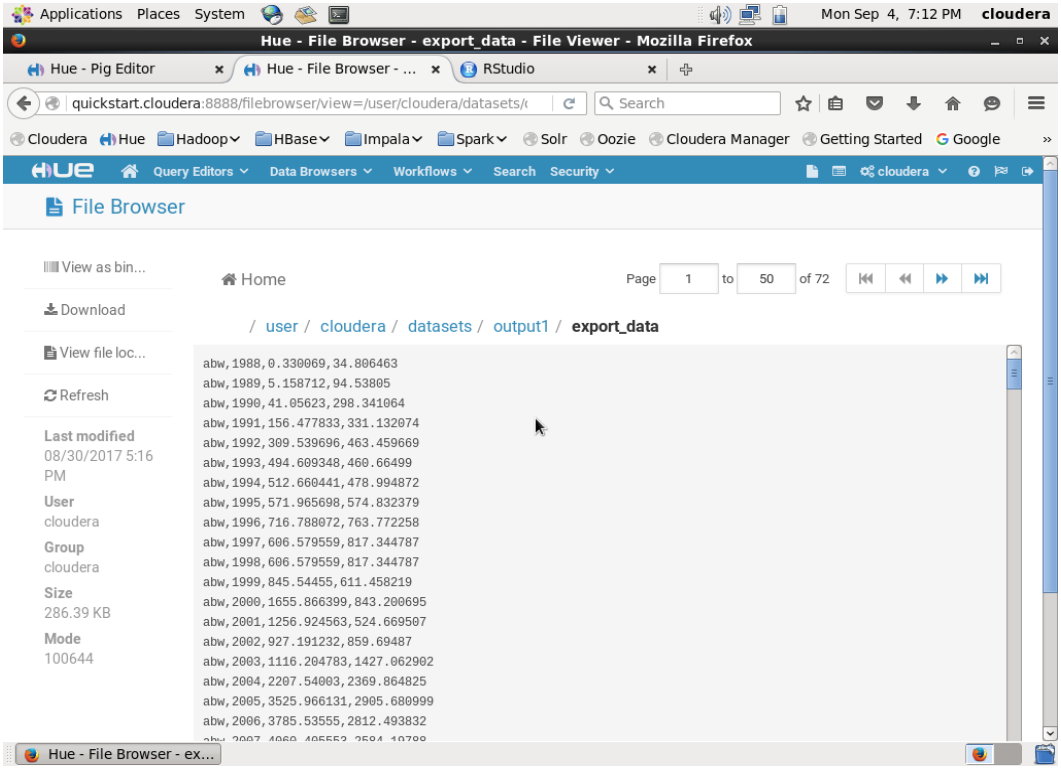


Figura 4.3 Archivo resultante del *script* 1 para el conjunto de datos 1.

Una vez generado el archivo resumido del conjunto de datos original, ahora se procede con la generación del *script* en R para visualizar los datos. En este caso, el *script* de R lee el archivo de HDFS, lo carga a un conjunto propio y, posteriormente, toma la información relacionada con el acrónimo “mex” para visualizar la información de México, finalmente, genera la gráfica de líneas correspondiente. La Figura 4.4 presenta la gráfica generada, con la cual se observa el crecimiento de las exportaciones de México a lo largo del tiempo, principalmente se muestra un crecimiento exponencial después de la década de 1990’s, salvo antes de entrar a los 2010’s, pero se vuelve a recuperar el crecimiento en las exportaciones.

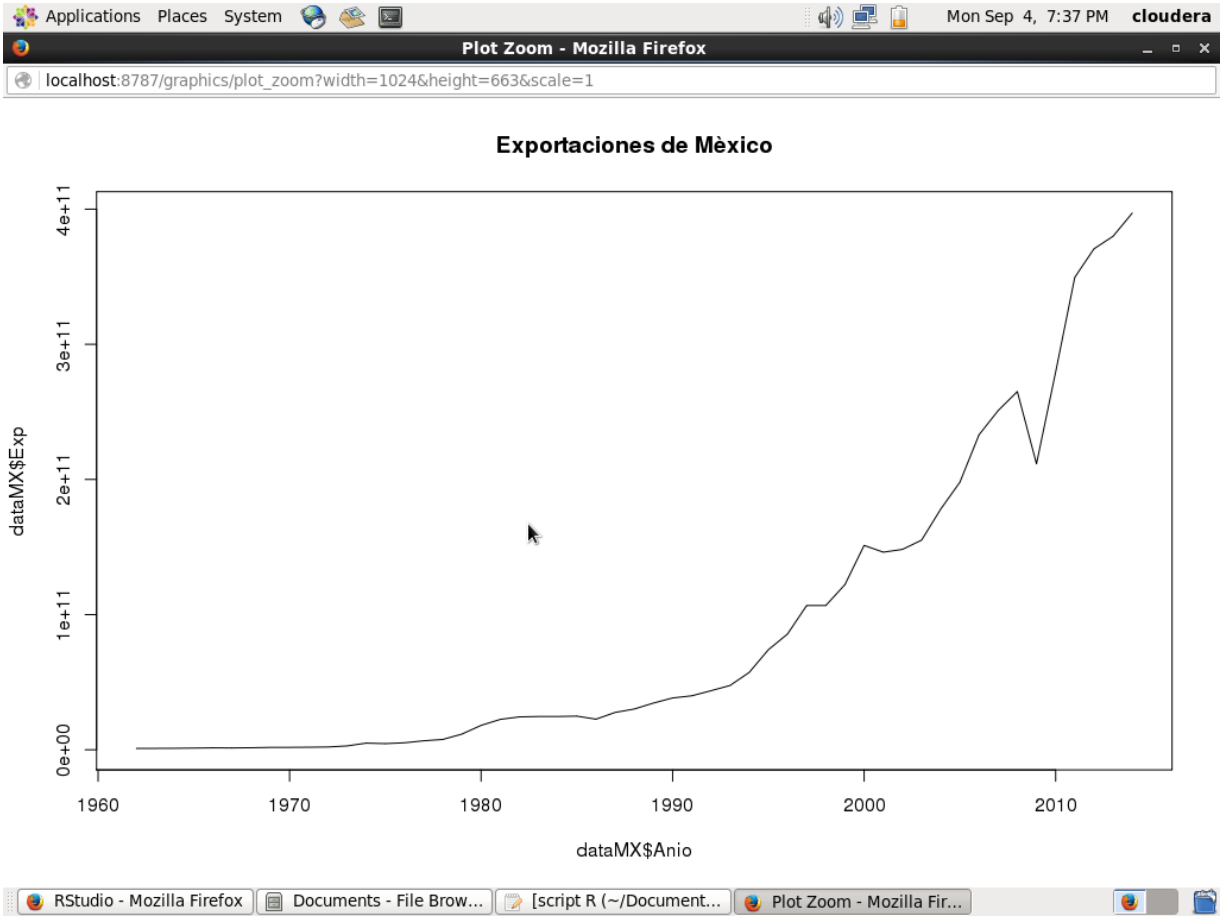


Figura 4.4 Gráfica de líneas para el conjunto de datos 1.

La siguiente gráfica a generar es la gráfica de barras, que de igual forma se tomará como datos de entrada el archivo resultante del *script* 1 de Apache Pig (Figura 4.2). Para el *script* en R se toma la misma premisa que el *script* anterior, salvo que ahora se requiere que los datos estén

en un objeto de tipo matriz. La Figura 4.5 presenta la gráfica resultante, en la cual no se logra apreciar muy bien los primeros datos en comparación con la gráfica de líneas.

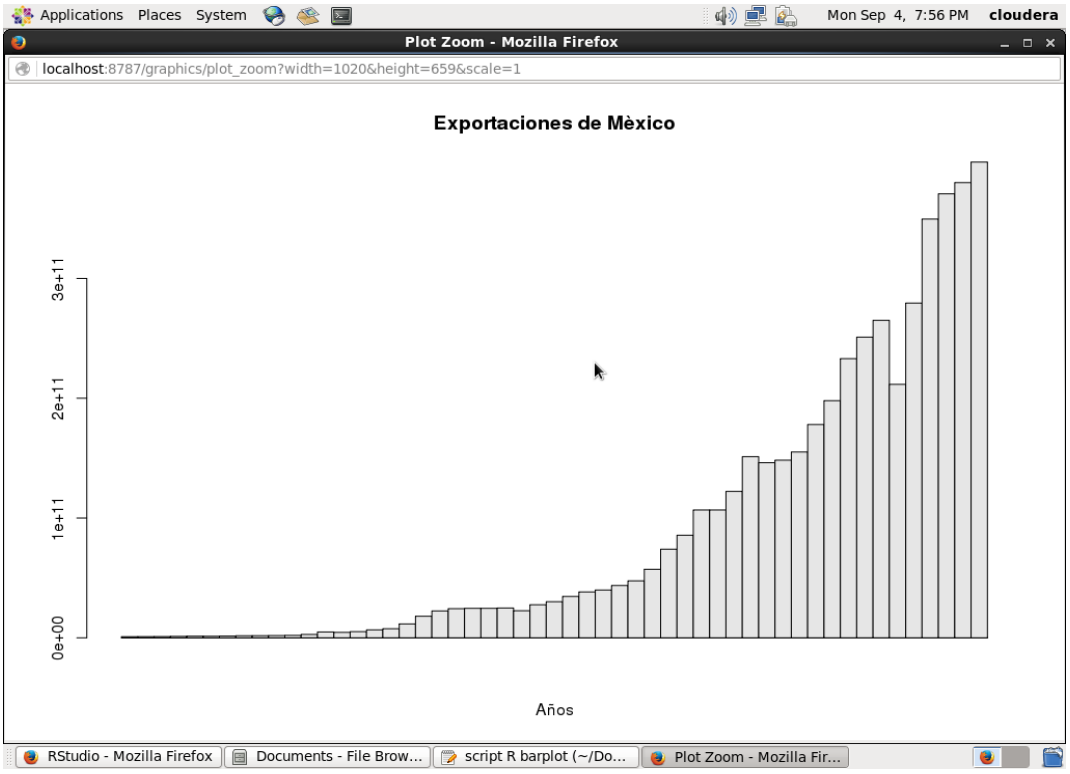


Figura 4.5 Gráfica de barras para el conjunto de datos 1.

Ahora toca el turno de generar la gráfica de dispersión y en la cual se siguen tomando los mismos datos y la misma premisa de los *scripts* anteriores. La finalidad de presentar la misma información en diferentes tipos de gráficas es con el propósito de comprender en qué tipo de gráfica se aprecian de mejor manera los datos. La Figura 4.6 muestra la gráfica correspondiente, en este caso la distancia que hay entre los puntos corresponde al crecimiento en el valor de las exportaciones de México. También en la Figura 4.7, se agrega una gráfica de dispersión con las demás variables del conjunto de datos, lo cual permite comparar las variables entre ellas mismas, destacando que entre las exportaciones e importaciones hay un crecimiento casi equitativo, permitiendo que se dibuje una diagonal casi perfecta entre estas variables.

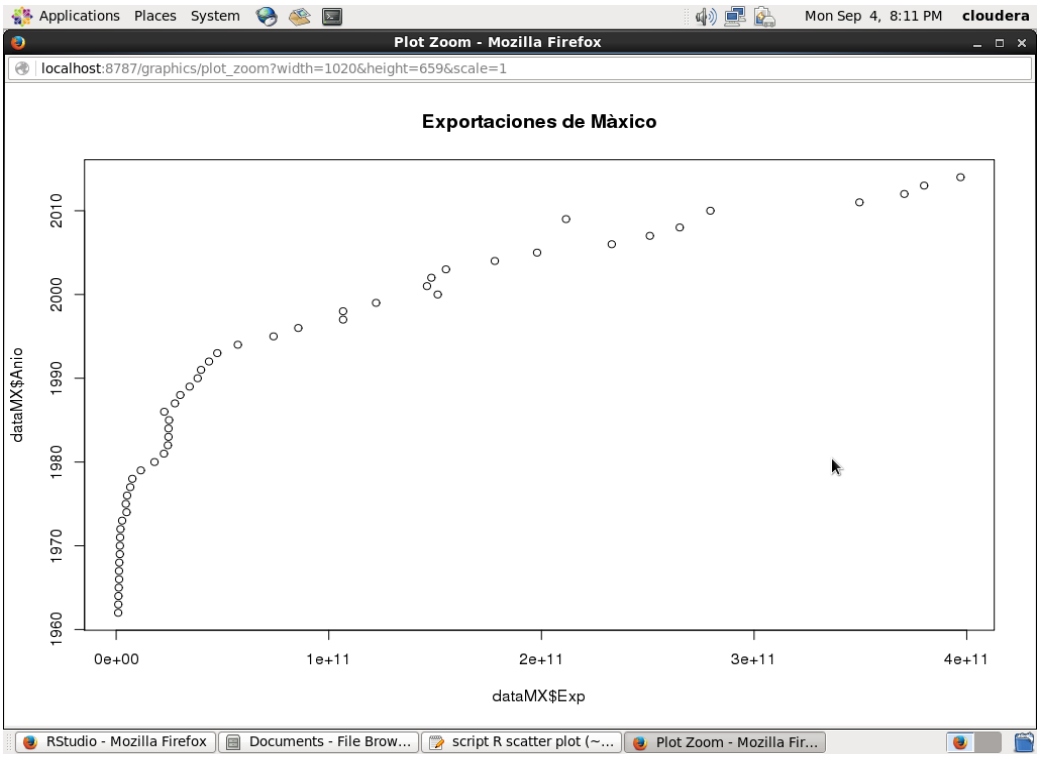


Figura 4.6 Gráfica de dispersión 1 para el conjunto de datos 1.

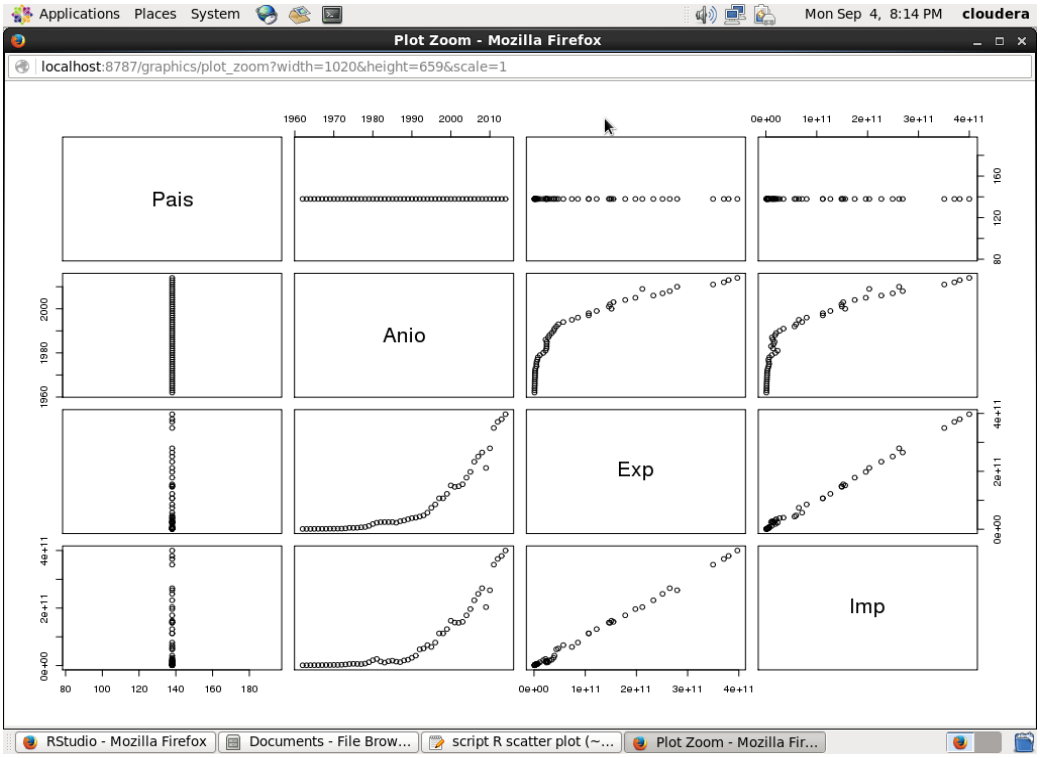


Figura 4.7 Gráfica de dispersión 2 para el conjunto de datos 1.

Para la gráfica *Heatmap*, se decidió tomar solo las exportaciones e importaciones del año 2014, por lo que se volvió a ejecutar un nuevo *script* en Apache Pig como se muestra en la Figura 4.8, dando como resultado un nuevo archivo almacenado en HDFS, el cual se aprecia en la Figura 4.9.

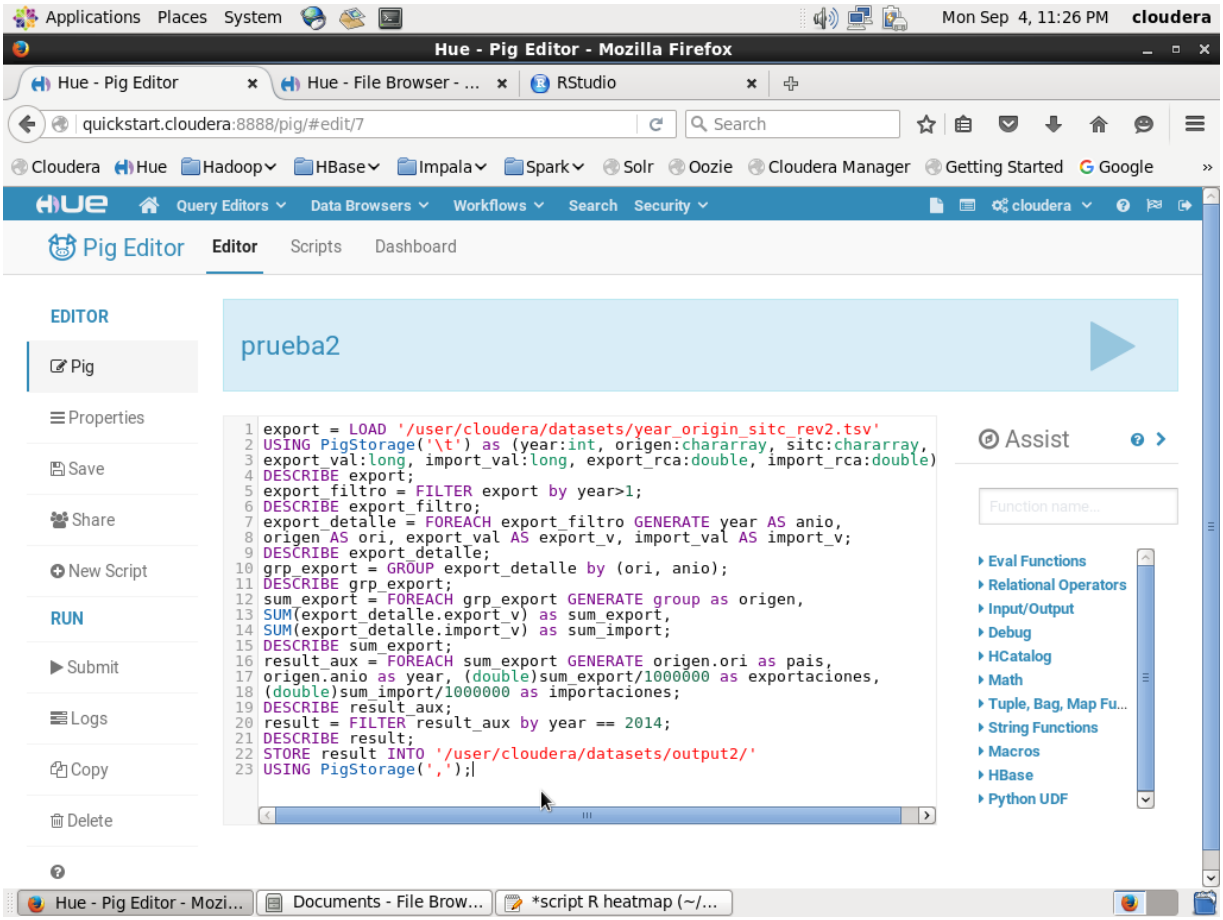


Figura 4.8 *Script* 2 de Apache Pig para el conjunto de datos 1.

En este caso, la gráfica *Heatmap* permite comparar las exportaciones e importaciones de todos los países del conjunto de datos. En la Figura 4.10 se muestra la gráfica generada, puesto que son demasiados registros no se logra apreciar de buena manera los datos, por lo que se decide presentar los primeros 50 registros en la Figura 4.11. En esta figura, el color rojo representa el menor valor, dando a entender, por ejemplo, que para el país Afganistán (afg), su valor de exportaciones es menor al valor de sus importaciones.

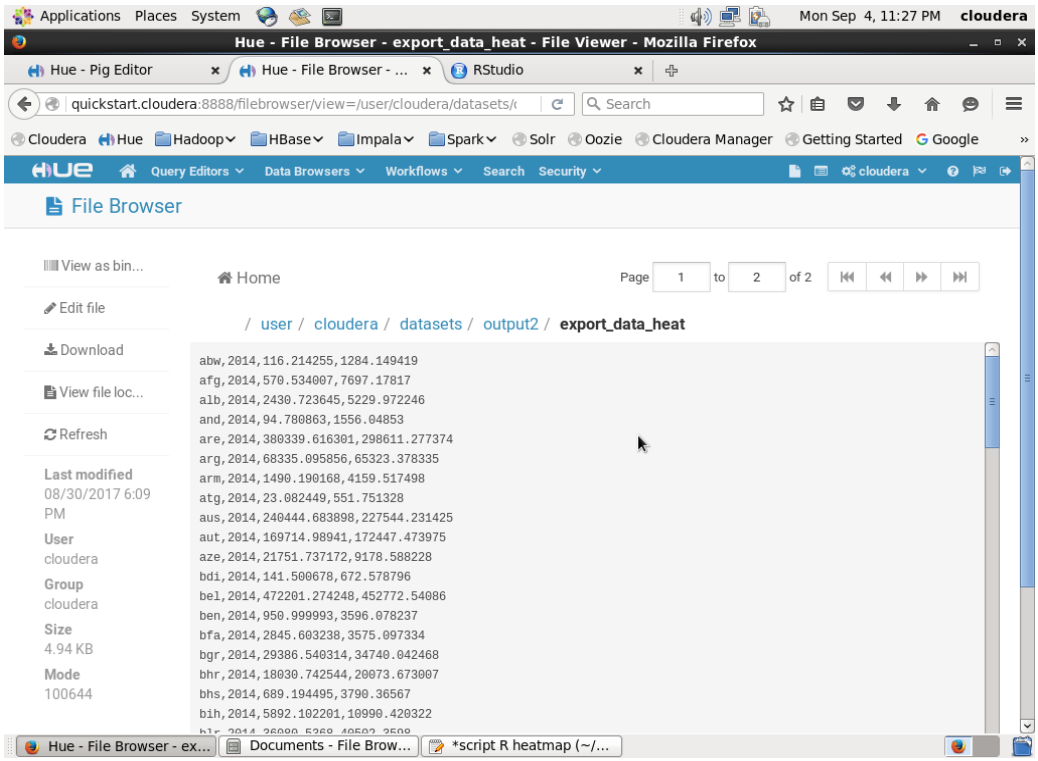


Figura 4.9 Archivo resultante del *script* 2 para el conjunto de datos 1.

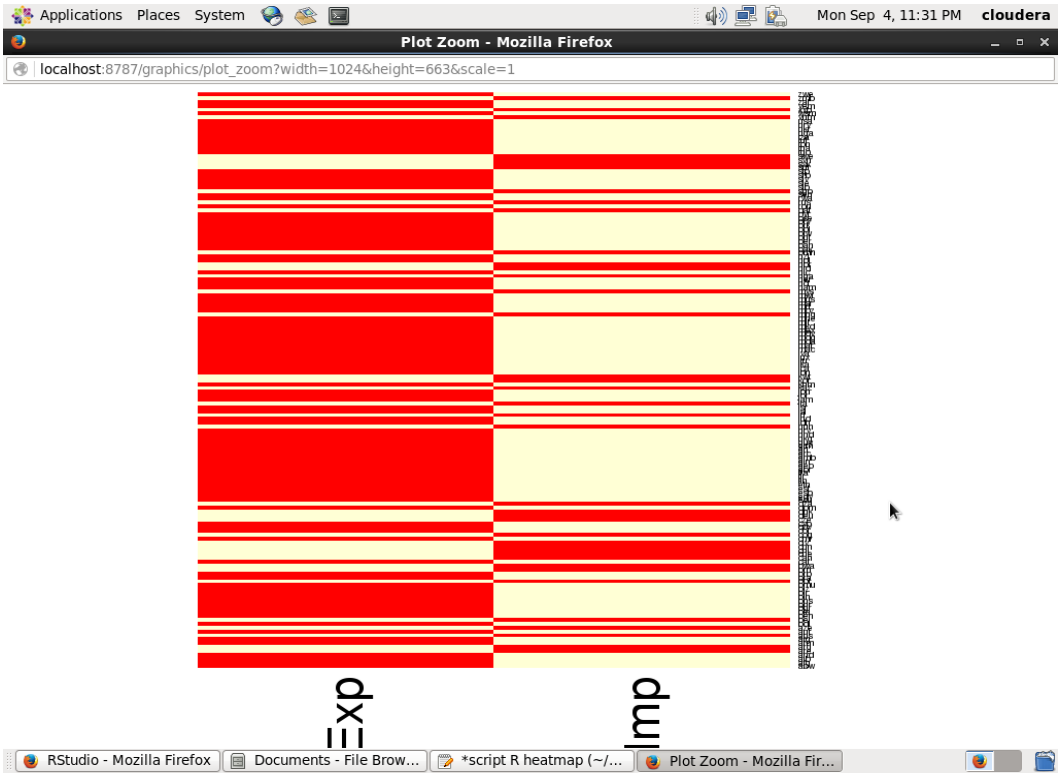


Figura 4.10 Gráfica *Heatmap* 1 para el conjunto de datos 1.

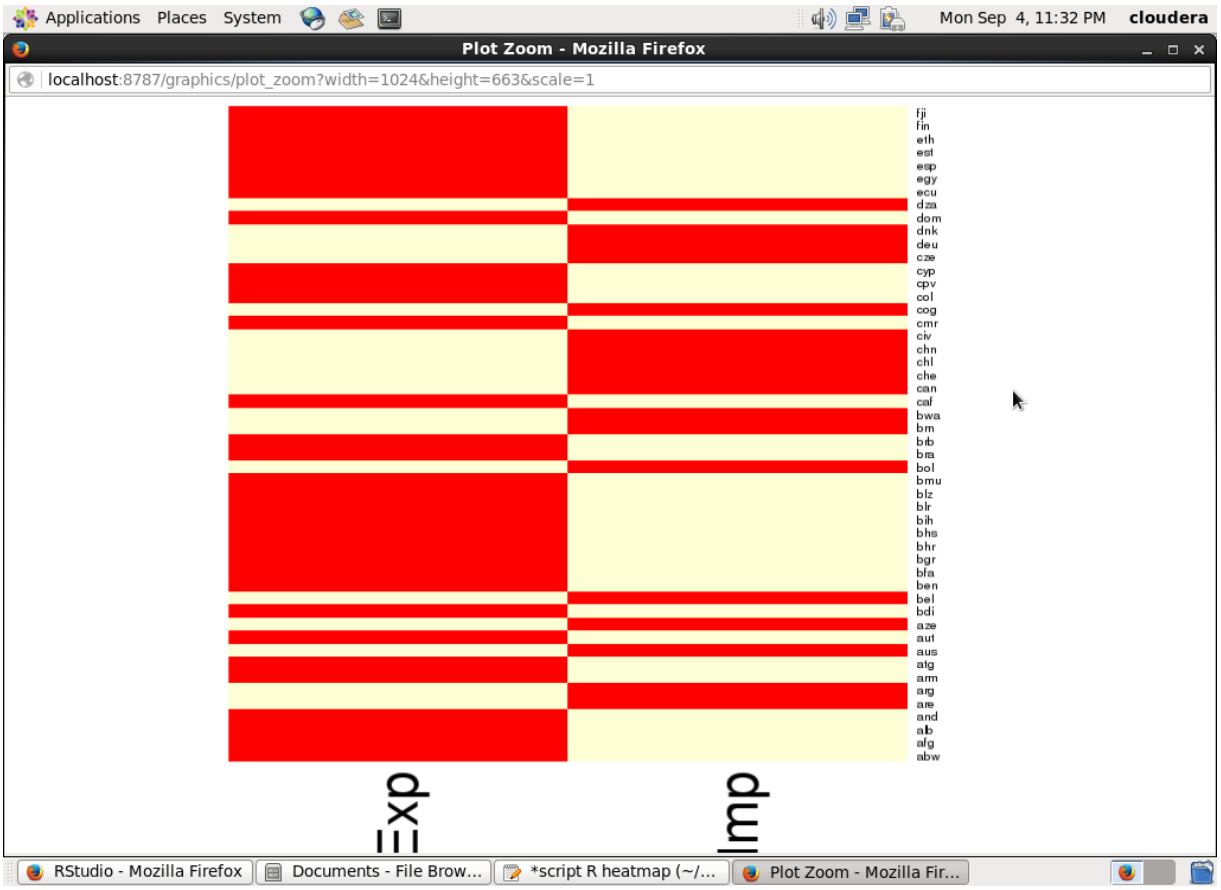


Figura 4.11 Gráfica *Heatmap* 2 para el conjunto de datos 1.

De nueva cuenta, se decidió realizar otro procesamiento al conjunto de datos original, en este caso, se indicó en el *script* de Apache Pig que se sumaran todas las exportaciones e importaciones a lo largo de todos los años que contenían en el conjunto de datos, dando como resultado un archivo que solo contenía el código de país, el valor total de exportaciones y el valor total de importaciones. La Figura 4.12 muestra el *script* de Apache Pig y la Figura 4.13 muestra el archivo resultante en HDFS.

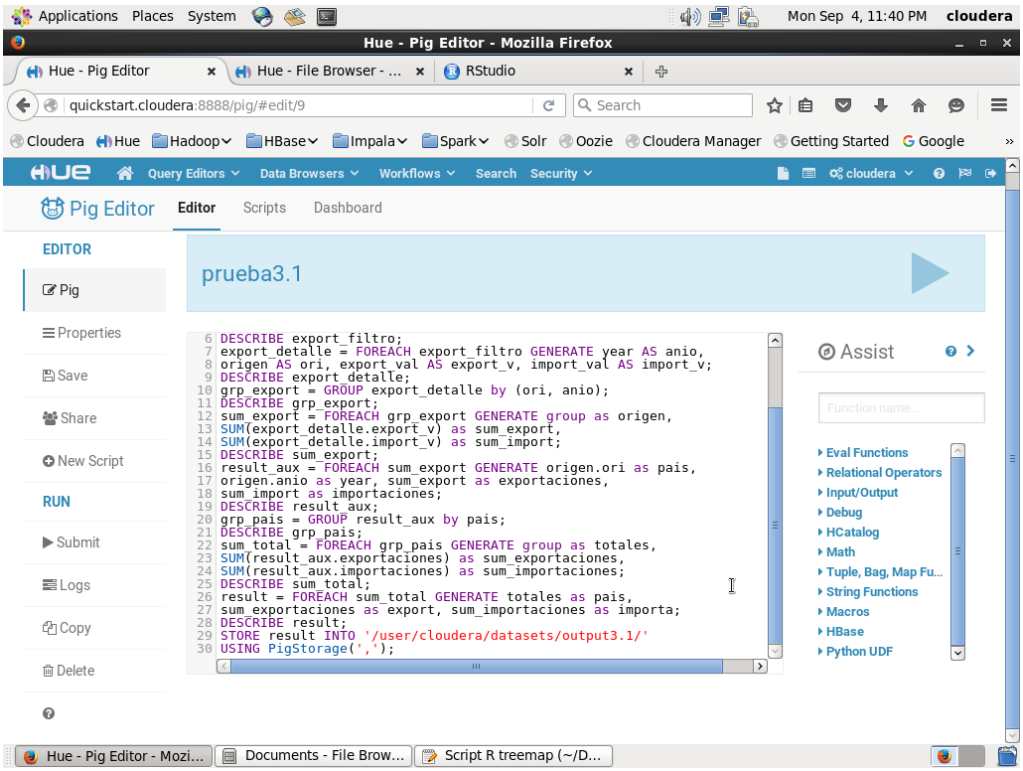


Figura 4.12 *Script* 3 de Apache Pig para el conjunto de datos 1.

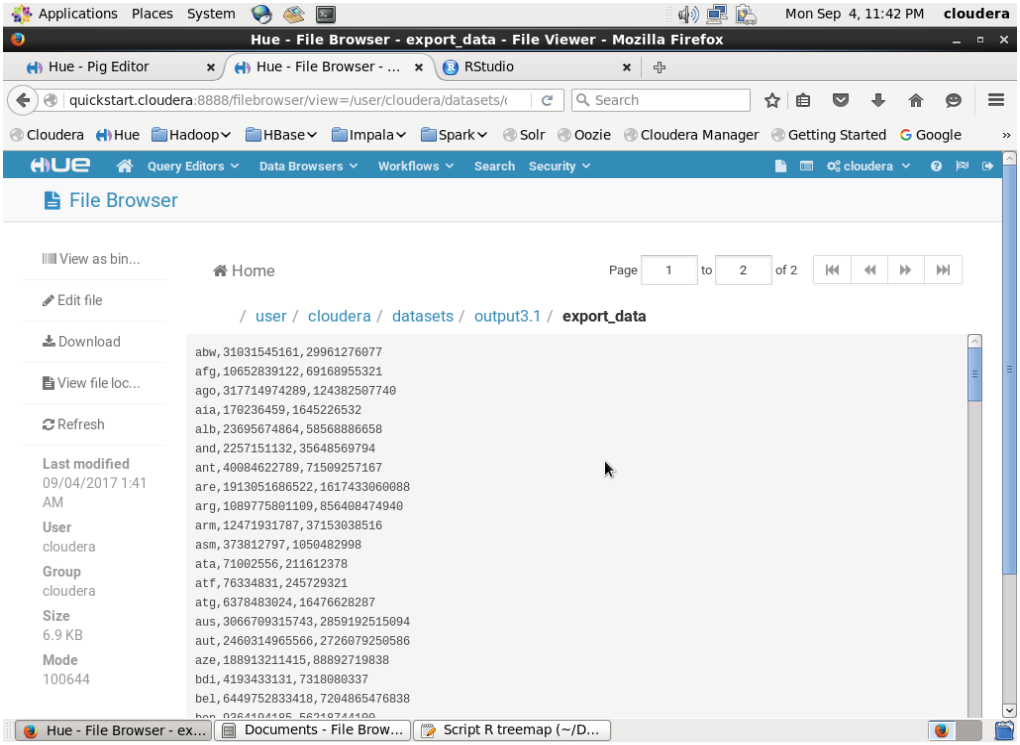


Figura 4.13 Archivo resultante del *script* 2 para el conjunto de datos 1.

Teniendo este nuevo archivo, se procede a generar la gráfica *Treemap* mediante el *script* en R. Para este *script* se toma en cuenta el valor total de las exportaciones. La Figura 4.14 muestra la gráfica resultante, en la que se observa que países como Estados Unidos (usa), Alemania (deu), China (chn) y Japón (jpn) son los principales exportadores del mundo.

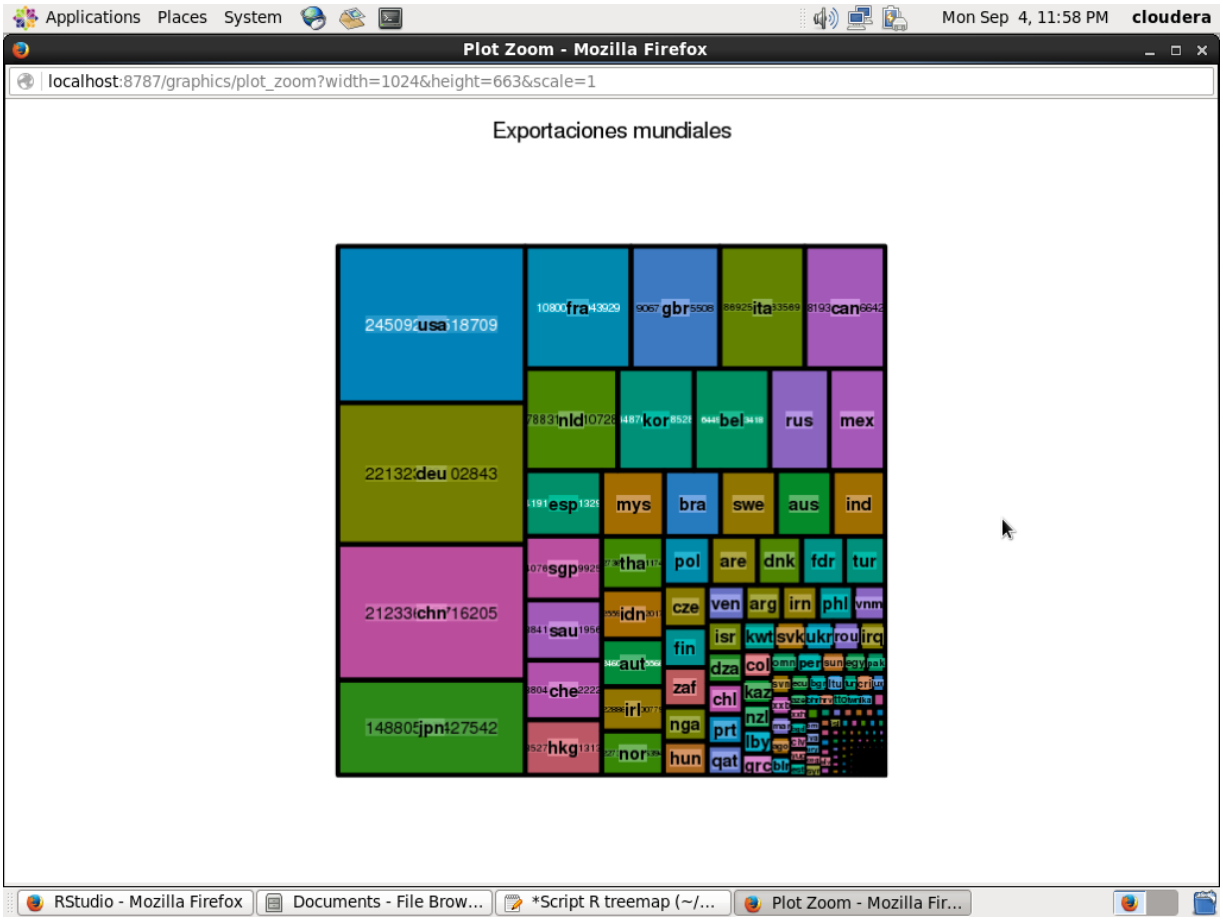


Figura 4.14 Gráfica *Treemap* para el conjunto de datos 1.

Ahora se procede a analizar cómo se comportan los datos a la hora de formar cúmulos mediante el algoritmo *k-means* y mediante la técnica de visualización de Clustergram. Para esto se toma el archivo HDFS del *script* anterior, se aplica el algoritmo *k-means* al conjunto de datos y posteriormente se grafica mediante Clustergram para conocer el comportamiento de los datos de dos a ocho grupos. La Figura 4.15 presenta la gráfica Clustergram, en la que se observa cómo la mayoría de los datos se concentran en el primer grupo y solo los de mayor valor son los que cambian entre diferentes grupos, estos datos de mayor valor corresponden a los países que mayor exportación hacen.

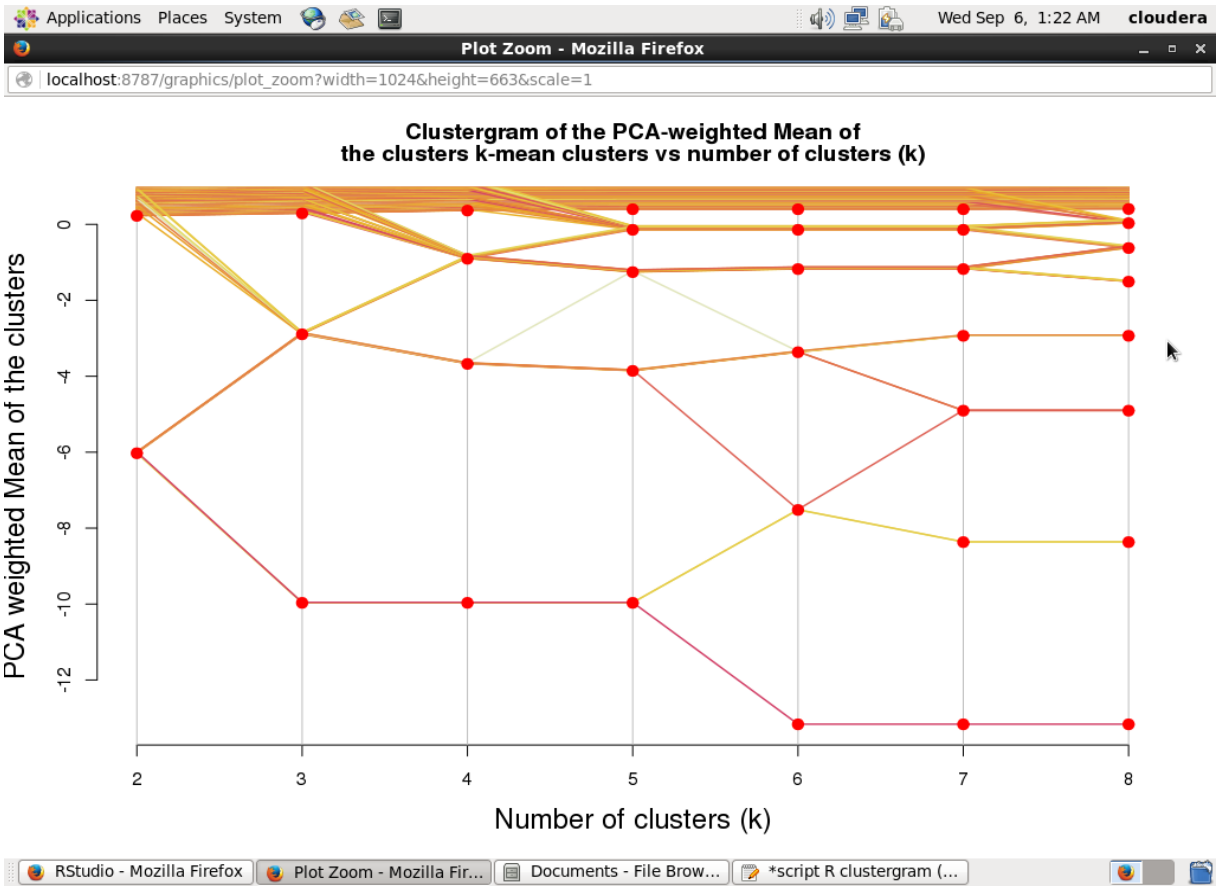


Figura 4.15 Gráfica Clustergram para el conjunto de datos 1.

Otra técnica implementada es la de coordenadas paralelas, para esto, se genera el *script* en R correspondiente. Para este *script*, se toman los datos de las exportaciones e importaciones de los países de México y Estados Unidos, con el fin de realizar una comparación de sus valores a través de la técnica de coordenadas paralelas. La Figura 4.16 presenta la gráfica correspondiente, en la cual se observa claramente que Estados Unidos sobrepasa por mucho los valores de las exportaciones e importaciones de México.

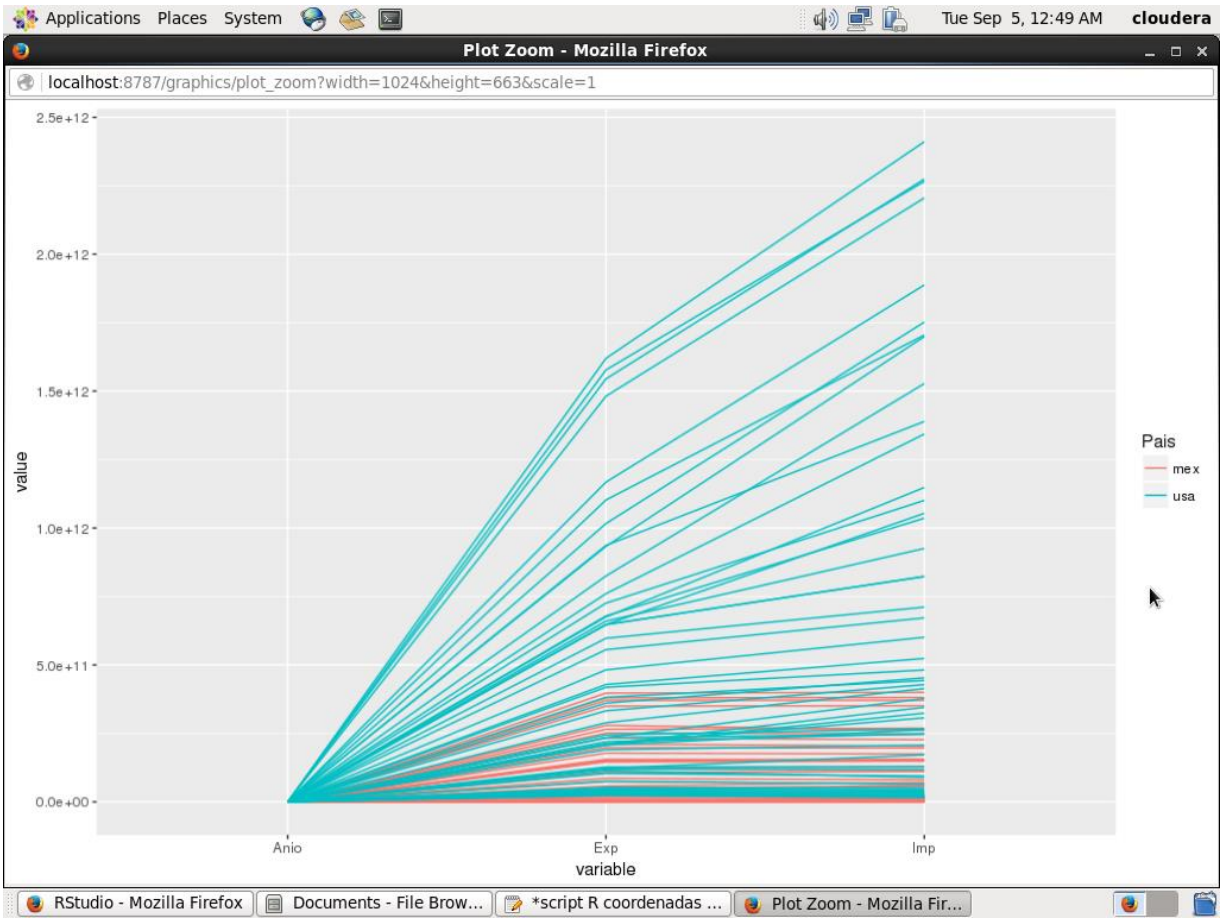


Figura 4.16 Gráfica de coordenadas paralelas para el conjunto de datos 1.

Finalmente para este conjunto de datos, falta implementar la técnica de *Spatial Information Flow*. El objetivo de esta técnica es la representación de los datos mediante el uso de mapas geográficos, por lo que requiere que los datos a graficar contengan los puntos de latitud y longitud, para así determinar los puntos en el mapa correspondiente. El conjunto de datos original no contiene dicha información, sin embargo, con la ejecución del último *script* de Apache Pig, al archivo resultante contiene solo unos 229 registros, lo cual hace factible que el ingreso de estos datos faltantes se lleve a cabo de manera manual. La Figura 4.17 muestra el archivo ya con los datos necesarios.

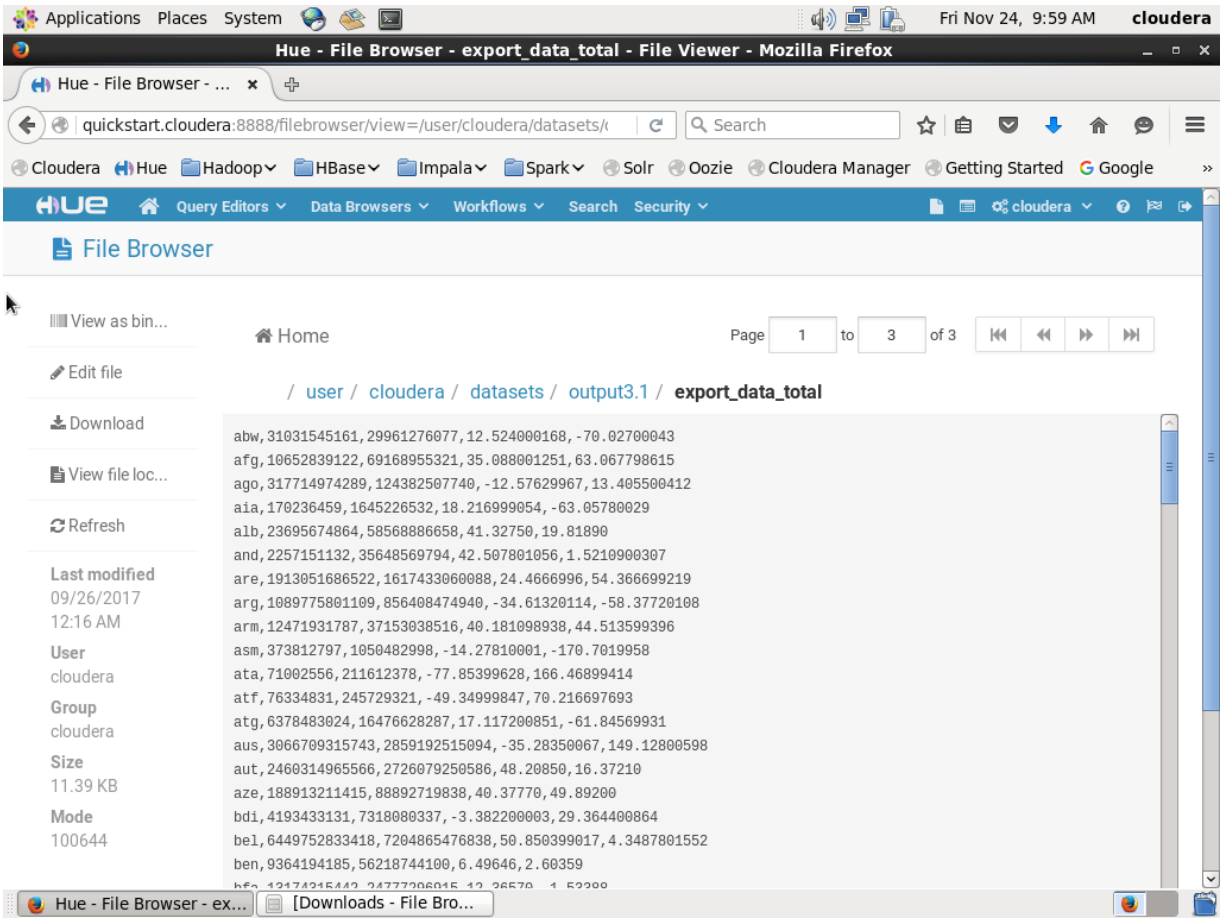


Figura 4.17 Archivo con latitud y longitud.

Una vez que ya se tiene el archivo completo, se procede a escribir el *script* en R para realizar la gráfica. En este *script* se agregan algunas funciones para que resalten los marcadores en color verde, naranja y rojo, indicando los de exportación mayor, exportación media y exportación menor, respectivamente. La Figura 4.18 presenta la gráfica resultante, en la cual se aprecian los datos de una manera más agradable y atractiva al momento de realizar algún análisis.

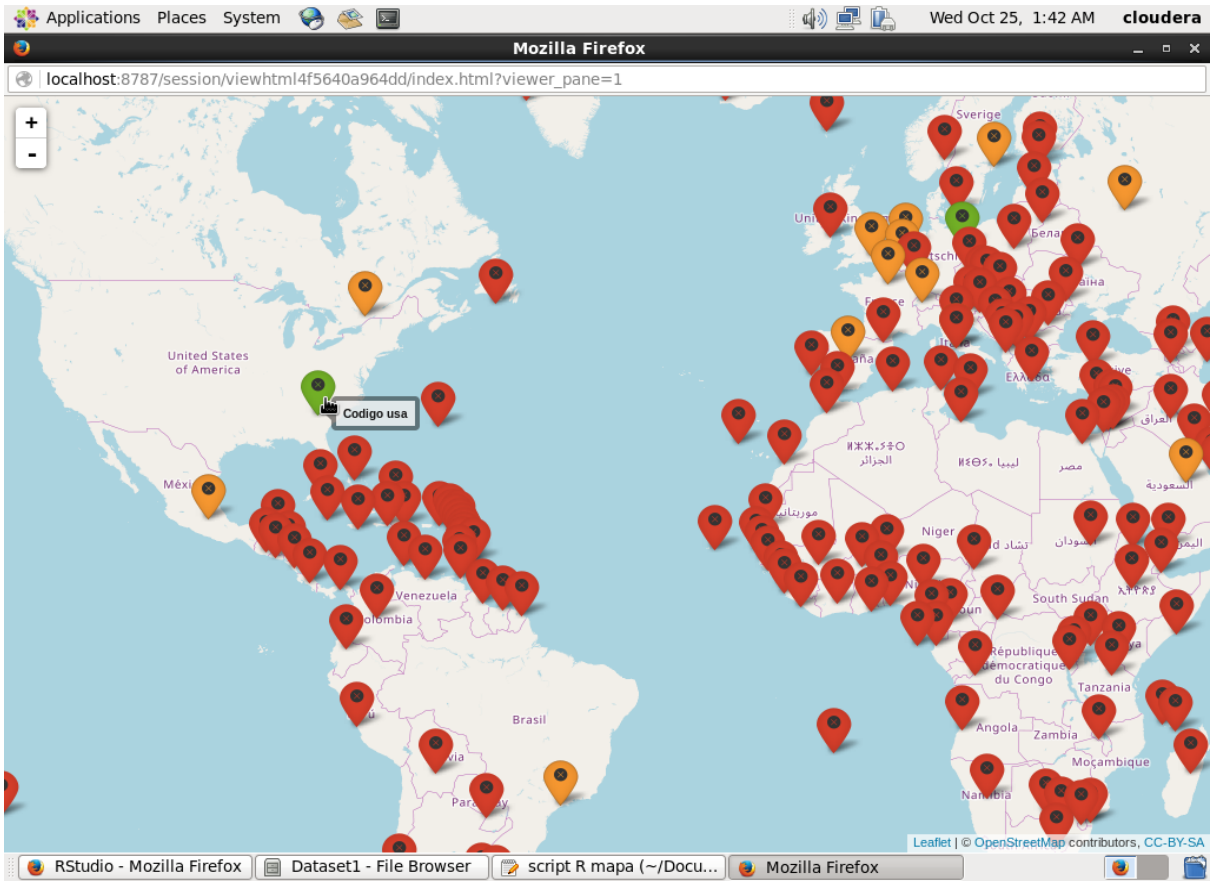


Figura 4.18 Gráfica de *Spatial Information Flow* para el conjunto de datos 1.

4.2. *Dataset* 2: Flujo de clics en páginas Web de ventas

El segundo conjunto de datos a utilizar contiene información sobre el flujo de clics que se realizan en páginas Web de ventas de productos, contenidos en los archivos de registro del historial Web. Este conjunto de datos se generó a partir del *benchmark* llamado BigBench. De acuerdo con [96], parte de los datos semi-estructurados que genera la herramienta son los flujos de clics de diferentes páginas Web, por lo que se decidió tomar este conjunto de datos para la implementación de las técnicas de visualización. Al ser datos generados por el *benchmark*, corresponden a datos sintéticos que no contienen la característica de veracidad como en el anterior conjunto de datos, sin embargo, sí cumplen con las características volumen, velocidad y variedad.

El conjunto de datos cuenta con seis variables, las cuáles son: id del registro, fecha y hora al momento de realizar el clic, código del producto en dado caso que se esté dando clic, código de la página Web, y código de usuario si está registrado. También cuenta con 1,692,441 registros, lo que afianza la característica de volumen, en cuanto a velocidad se refiere a qué tan rápido se están generando los datos al tratarse de archivos de registro. Por último, la característica de variedad se cumple al tratarse de datos semi-estructurados. La Figura 4.19 muestra los primeros registros de este conjunto de datos cargados el sistema HDFS.

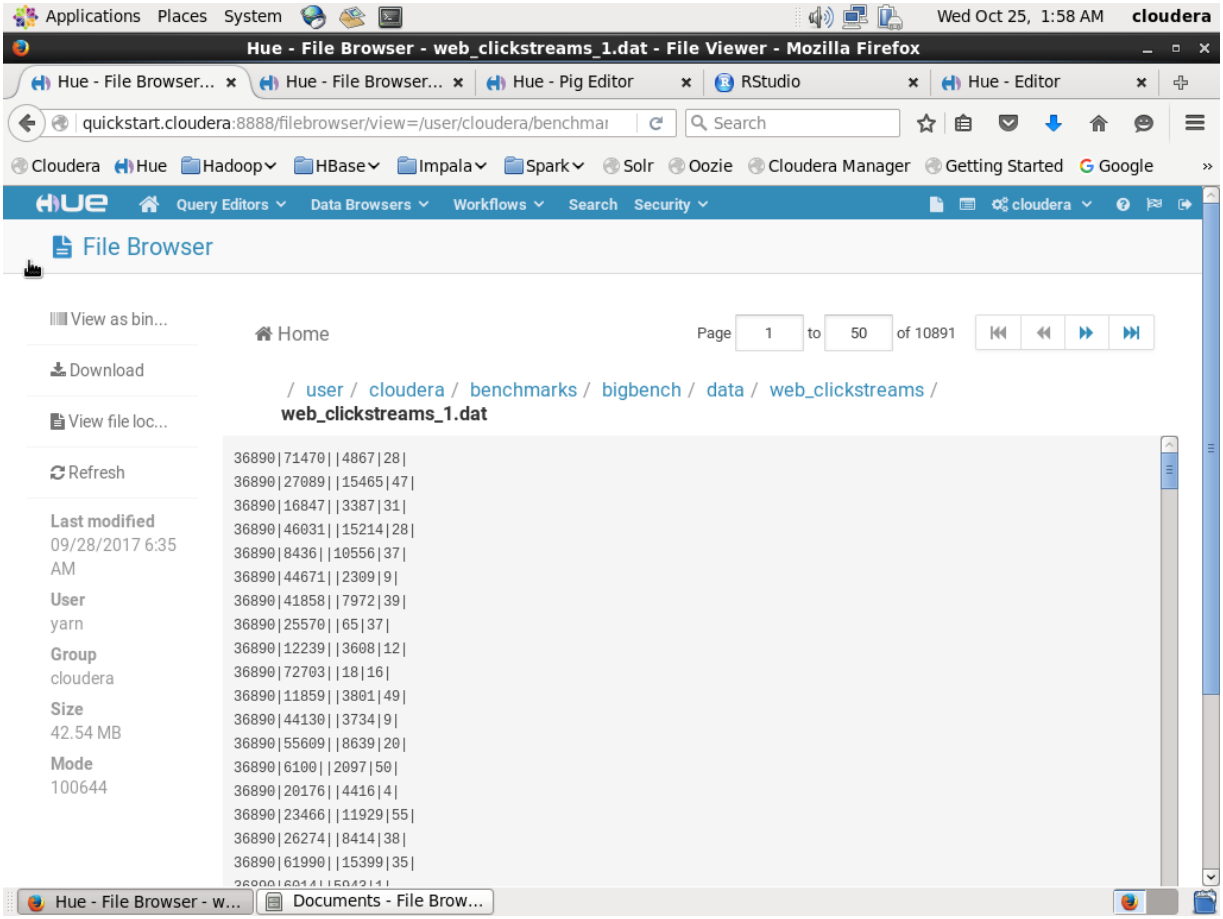


Figura 4.19 Conjunto de datos "Flujo de clics de páginas Web**".**

Siguiendo el flujo de trabajo del primer conjunto de datos, ahora es necesario procesar los datos mediante Apache Pig y así resumir la información para posteriormente visualizarla. En este primer *script* para el conjunto de datos en cuestión, la función principal es agrupar todos los datos por medio del código de usuario y así contabilizar los clics que han dado cada uno de estos usuarios. La Figura 4.20 muestra el *script* y la Figura 4.21 muestra el archivo resultante

almacenado en HDFS, en donde se observan todos los códigos de usuarios registrados y los clics que han dado a lo largo del tiempo.

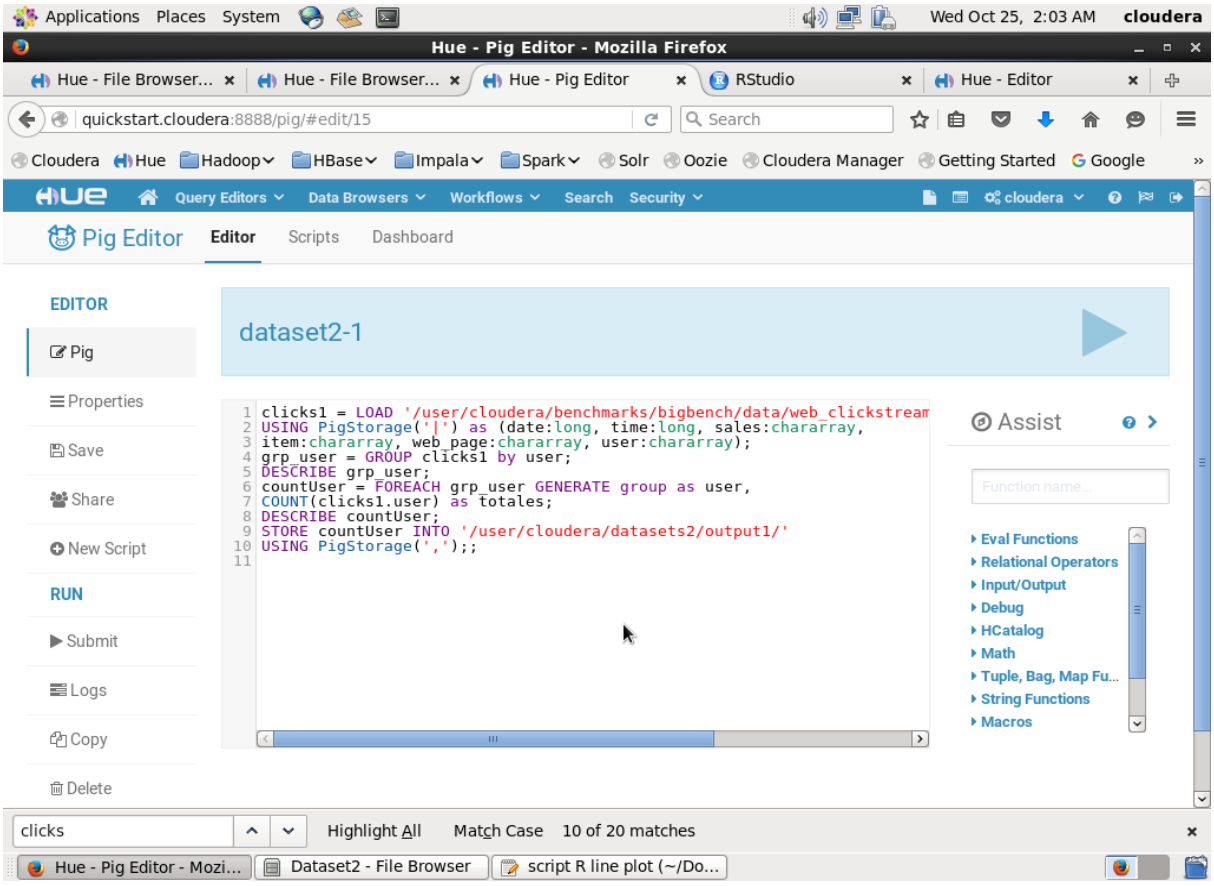


Figura 4.20 *Script* 1 de Apache Pig para el conjunto de datos 2.

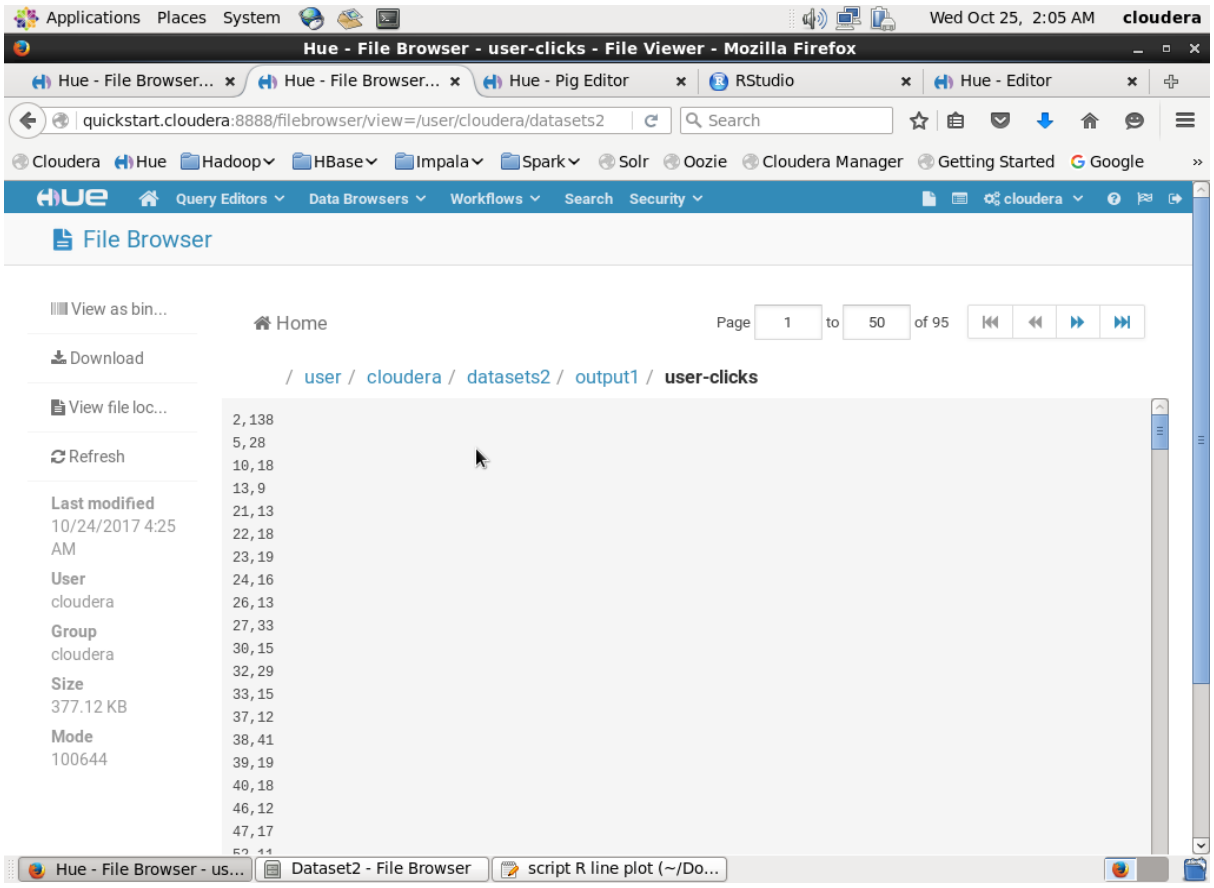


Figura 4.21 Archivo resultante del *script* 1 para el conjunto de datos 2.

Ya teniendo el archivo resumido del conjunto de datos original, ahora toca generar el *script* en R para construir la gráfica. Como primera técnica, se va a generar la gráfica de líneas, el *script* de R toma los datos de HDFS y los carga a un conjunto propio, el cual al solo tener dos variables, estas se grafican sin ningún otro procesamiento. La Figura 4.22 muestra la gráfica resultante. De esta gráfica se observa que aún a pesar de ya ser datos resumidos, el conjunto de datos sigue siendo demasiado grande para comprender de buena forma la gráfica, ya que son alrededor de 100,000 usuarios los que están registrados, sin embargo, si solo se grafican los primeros 100 registros se aprecia de una mejor manera, como se expone en la Figura 4.23.

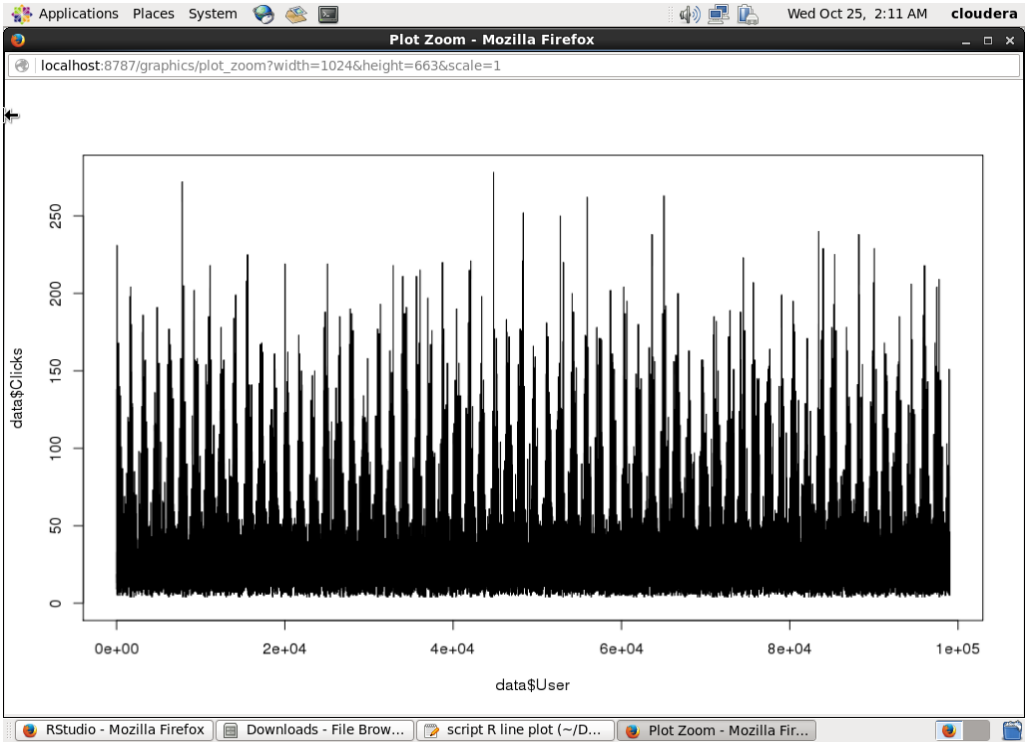


Figura 4.22 Gráfica de líneas para el conjunto de datos 2.

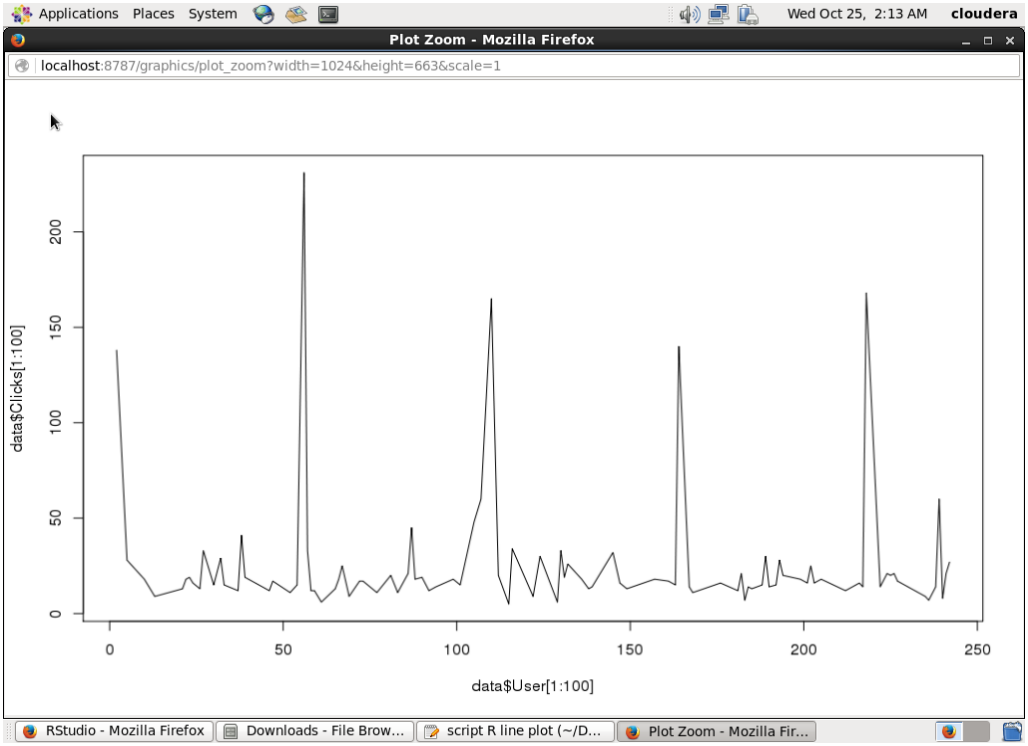


Figura 4.23 Gráfica de líneas con 100 registros para el conjunto de datos 2.

Ahora toca el turno de implementar la gráfica de barras, que de la misma forma se toman los datos del primer *script* de Apache Pig para este conjunto de datos. Para este caso, los datos que se toman de HDFS se transforman a un objeto de tipo matriz para visualizarse en R. La Figura 4.24 presenta la gráfica resultante, de la cual se observa que tiene el mismo problema que la técnica anterior al ser demasiados datos, por lo que también se toma la decisión de solo graficar los primeros 30 registros para una mejor visualización, esto se expone en la Figura 4.25.

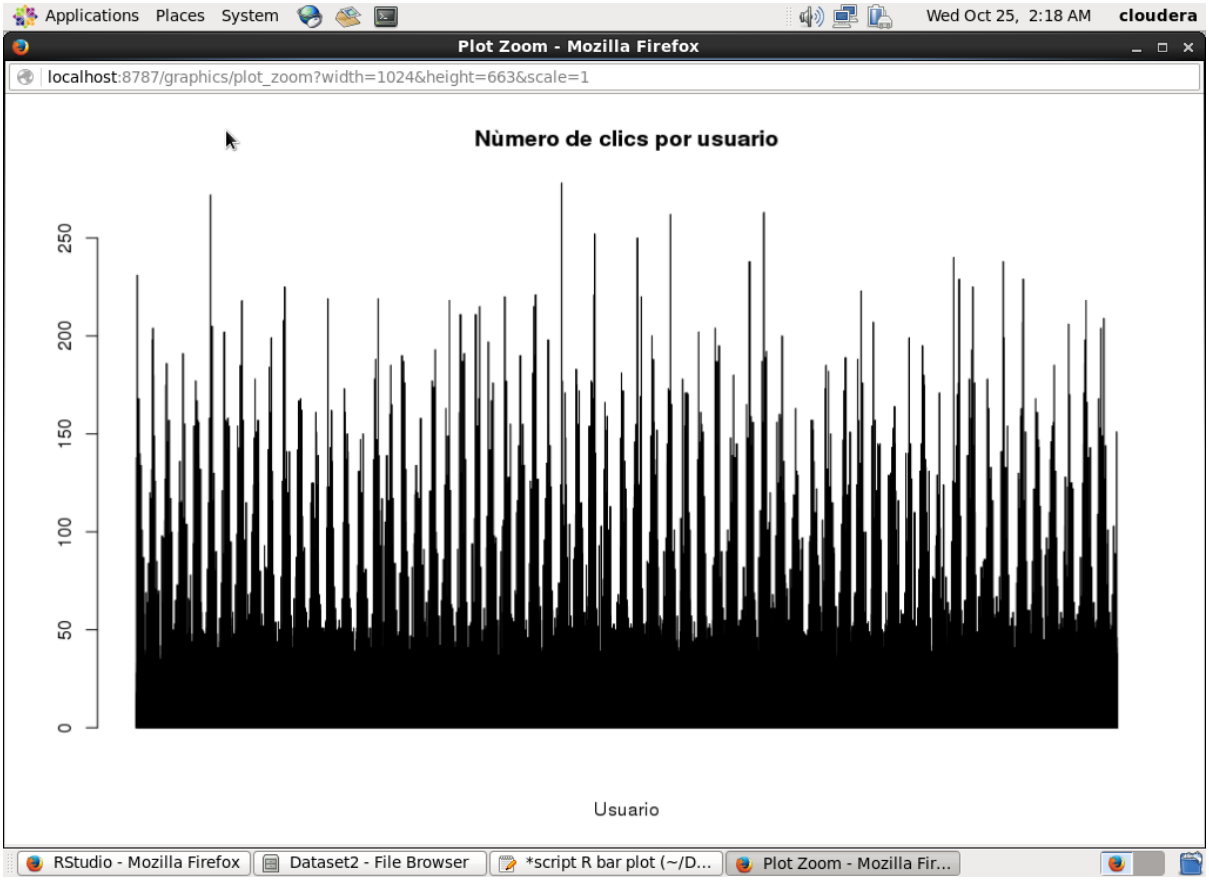


Figura 4.24 Gráfica de barras para el conjunto de datos 2.

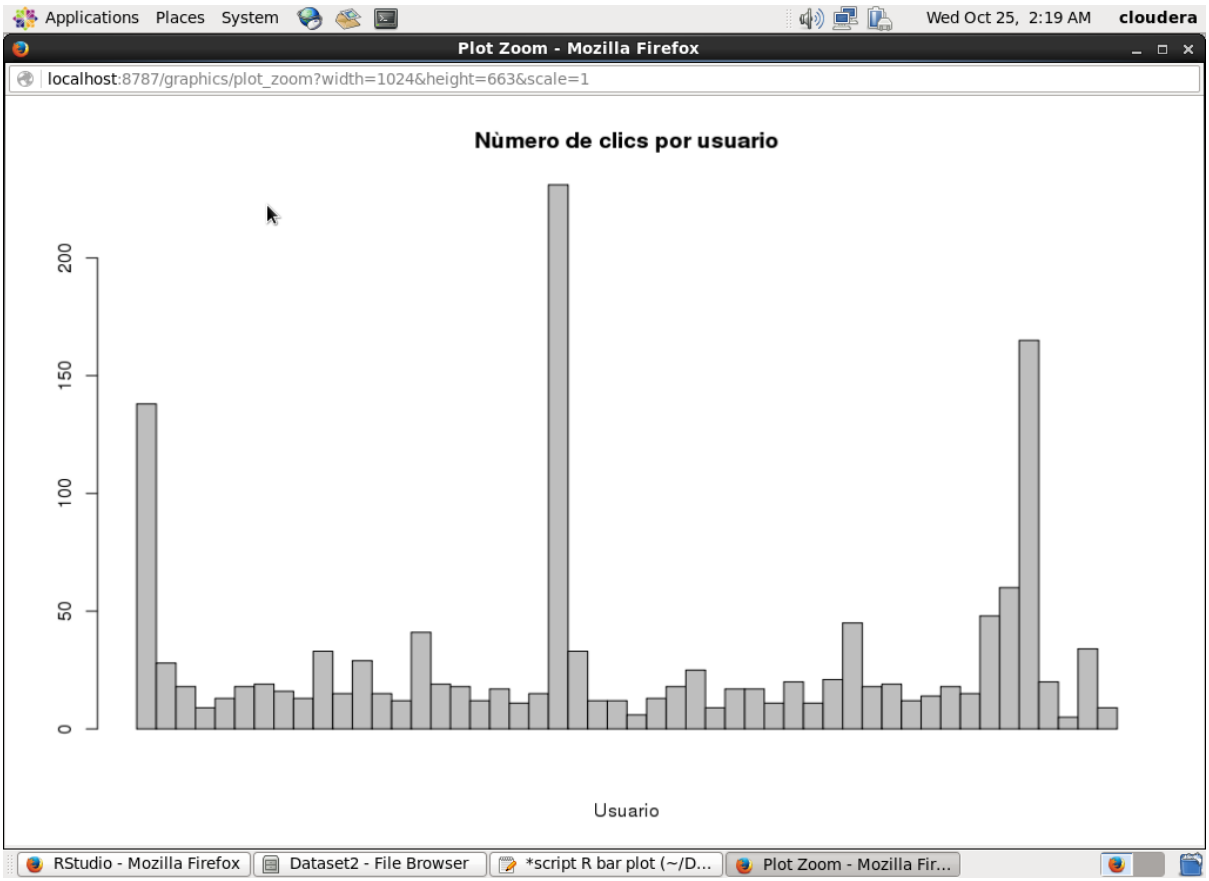


Figura 4.25 Gráfica de barras con 30 registros para el conjunto de datos 2.

Se procede a implementar la técnica de gráfica de dispersión, que de la misma forma que la técnica anterior, se toman los datos del primer *script* de Apache Pig. En este caso, la Figura 4.26 presenta la gráfica resultante. Para esta gráfica se logra observar de mejor manera la cantidad de datos que se están graficando, ya que cada registro corresponde a un solo punto, sin embargo, aún se dificulta la apreciación de los datos como las otras dos técnicas de visualización. De igual forma, la Figura 4.27 expone la gráfica de dispersión pero ahora con solo 100 registros.

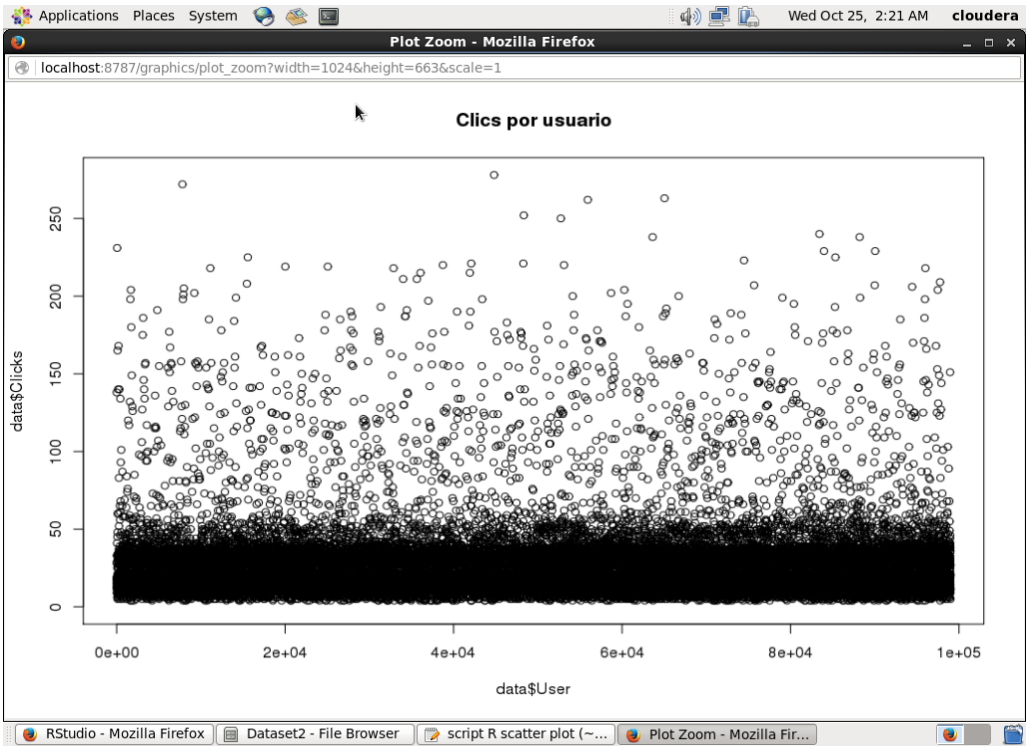


Figura 4.26 Gráfica de dispersión para el conjunto de datos 2.

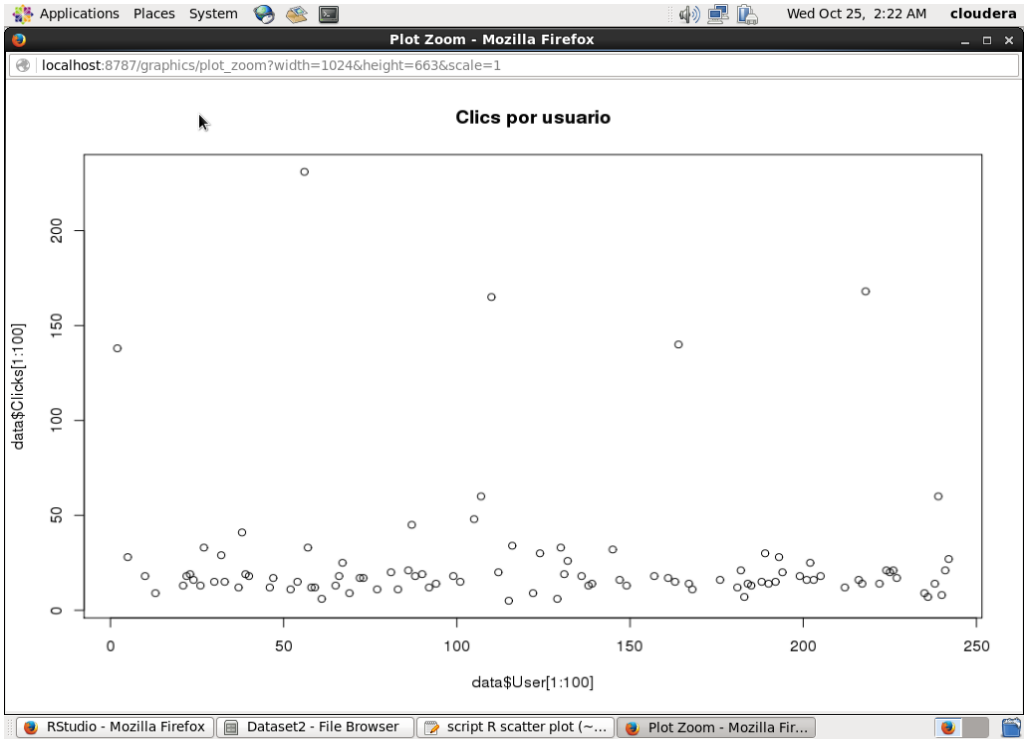


Figura 4.27 Gráfica de dispersión con 100 registros para el conjunto de datos 2.

Para las siguientes técnicas de visualización, se decidió volver a procesar el conjunto de datos original para agregarle dos variables más al archivo que se va a visualizar. Como se mencionó al inicio de este apartado, cada registro que tenga asociado algún código de producto indica que ese clic termina en alguna compra, mientras que los otros solo son clics de navegación. Siguiendo esta premisa, se programó otro *script* de Apache Pig para conocer esta información para cada usuario registrado. El *script* agrupa la información por medio del código de usuario, posteriormente se contabiliza para cada usuario su total de clics, de la misma forma se contabiliza si tiene asociado o no un código de producto, con esto se logra obtener la información deseada. La Figura 4.28 muestra el *script* antes mencionado y la Figura 4.29 muestra los primero registros del archivo resultante almacenado en HDFS, en donde se despliega el código de usuario, el total de clics, los clics que terminaron en venta y los clics que solo son de navegación.

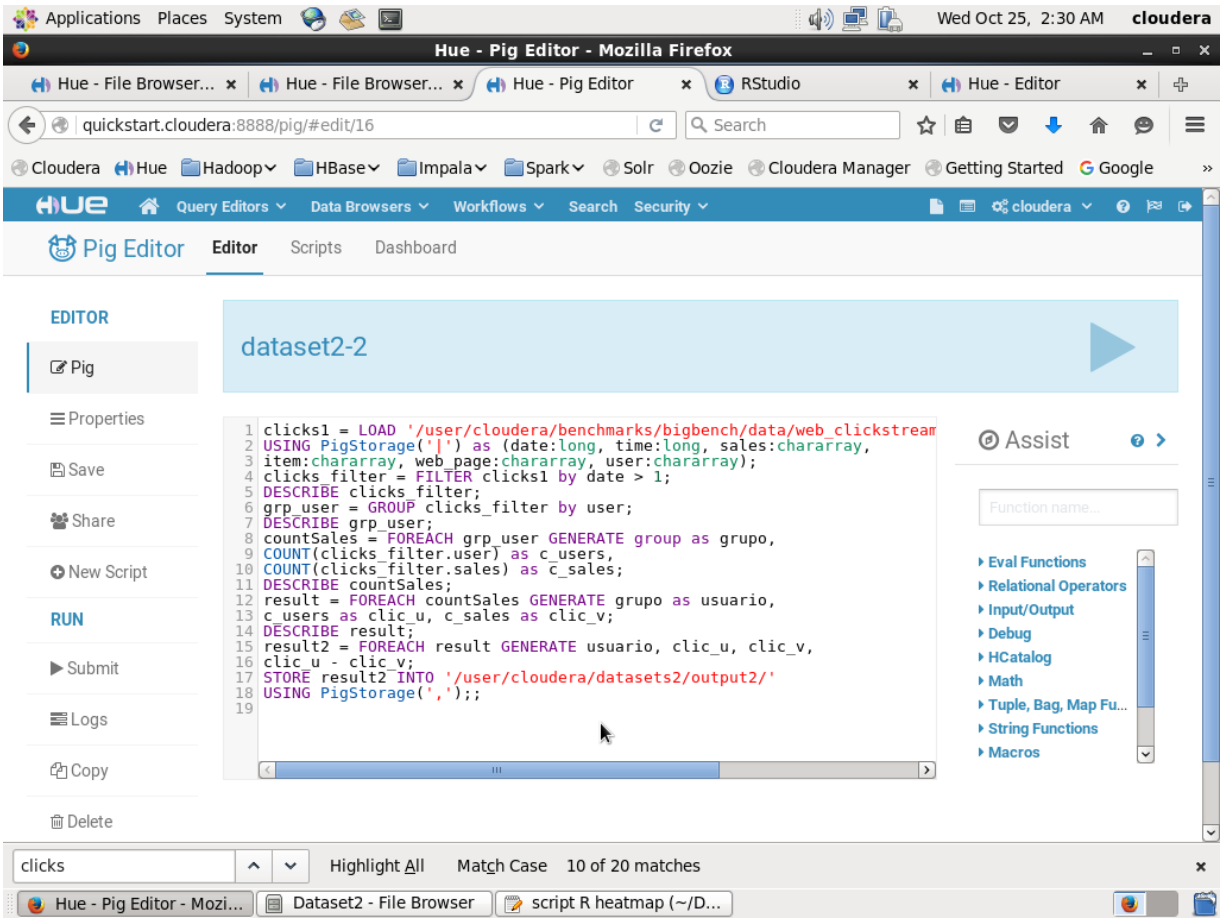


Figura 4.28 *Script* 2 de Apache Pig para el conjunto de datos 2.

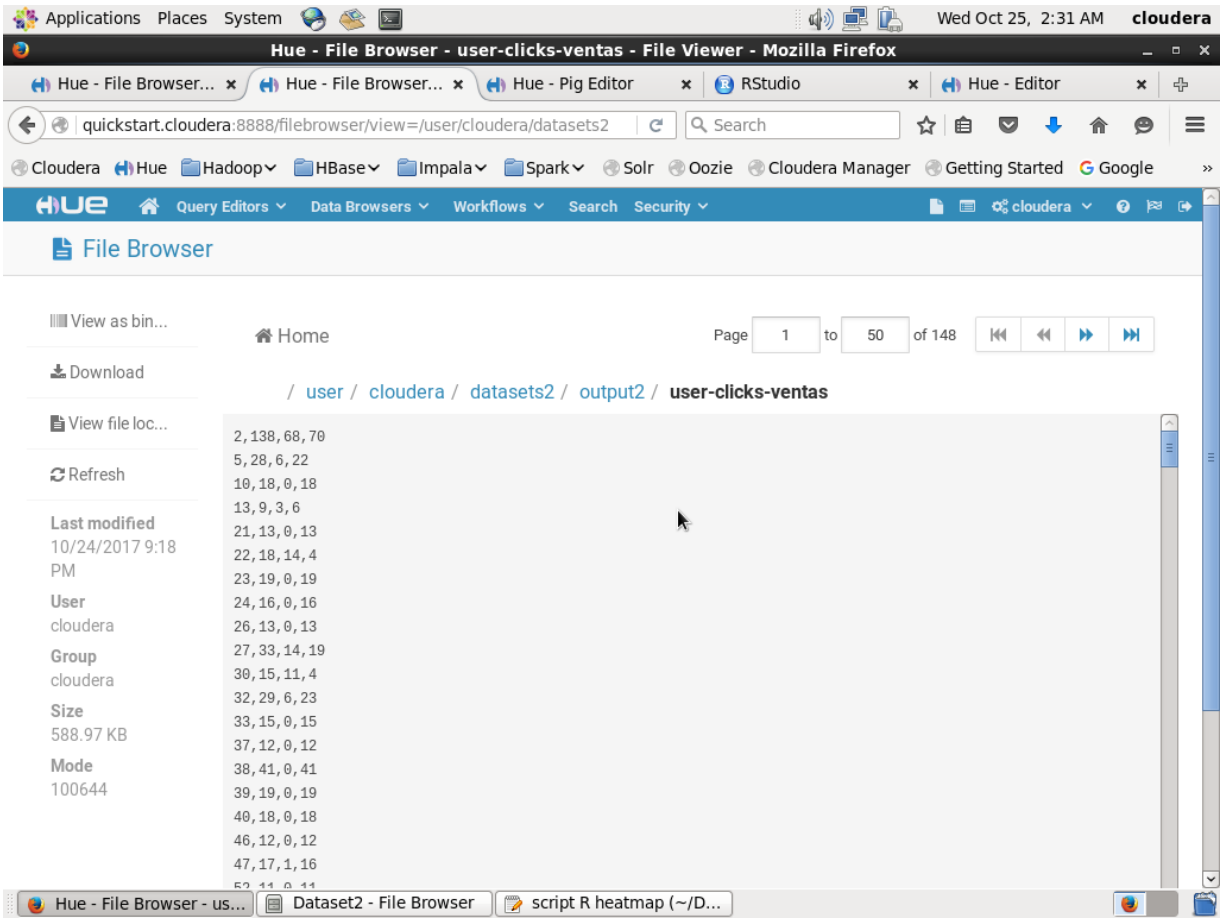


Figura 4.29 Archivo resultante del *script* 2 para el conjunto de datos 2.

Tomando la información antes presentada, se procede a implementar la técnica de visualización *Heatmap* en R. La función principal del *script* es tomar los datos que se encuentran en HDFS, adecuarlos a un objeto de tipo matriz para así implementar la visualización. La Figura 4.30 presenta la gráfica generada, el color rojo representa los valores menores, lo cual da a entender que los usuarios dan más clics cuando están navegando y cuando se deciden a comprar algo, normalmente van ya al producto que desean sin dar tantos clics. Esta gráfica también muestra que los datos son demasiados para ser vistos claramente en pantalla, por lo que en la Figura 4.31 se presentan los primeros 50 registros.

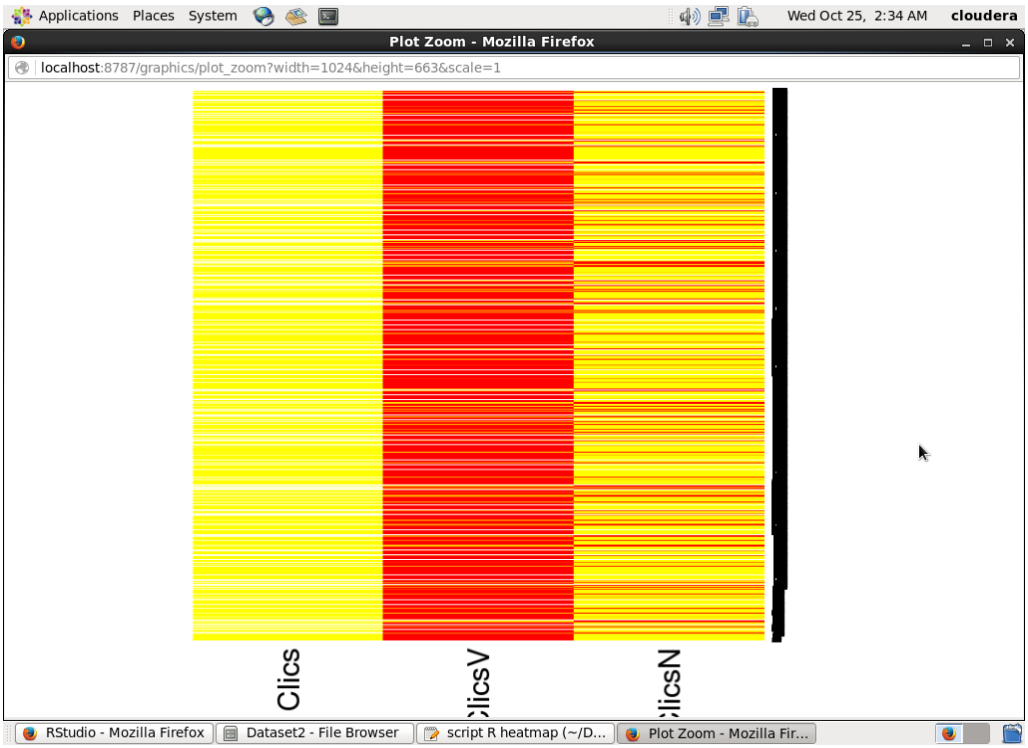


Figura 4.30 Gráfica *Heatmap* para el conjunto de datos 2.

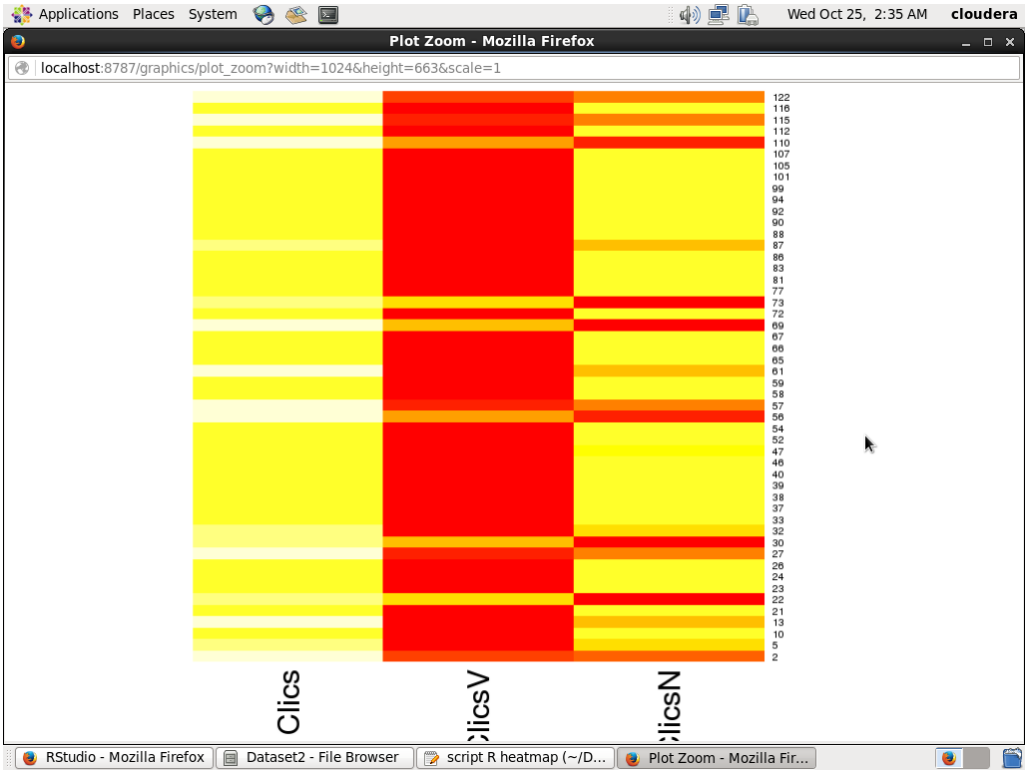


Figura 4.31 Gráfica *Heatmap* con 50 registros para el conjunto de datos 2.

Con la misma información del conjunto de datos anterior, se procede a implementar la técnica de visualización *Treemap*. El *script* de R tiene la función de leer los datos de HDFS y visualizarlos sin ninguna transformación previa, sin embargo, al tener demasiados registros no se logra visualizar por completo, las características de la máquina impiden que se ejecute por completo el comando, por lo que se tomó la decisión de visualizar los primeros 1,000 registros representados en la Figura 4.32. En esta figura se observa que el usuario 56 es el que mayor número de clics realiza, seguido de los usuarios 110 y 326.

Figura 4.32 Gráfica *Treemap* para el conjunto de datos 2.

Mediante la técnica de visualización Clustergram, se conoce cómo se comportan los datos al momento de formarlos en cúmulos, esto es, aplicando el algoritmo *k-means*. Esta técnica se aplicó a los datos con los que se viene trabajando y así conocer su comportamiento. El *script* en R para esta visualización se observa la aplicación del algoritmo *k-means* a los datos que se encuentran en HDFS para posteriormente visualizarlos con el comando correspondiente. La

Figura 4.33 presenta la gráfica Clustergram, en esta figura se analiza cómo se van moviendo los datos a través de los diferentes grupos (dos a ocho grupos).

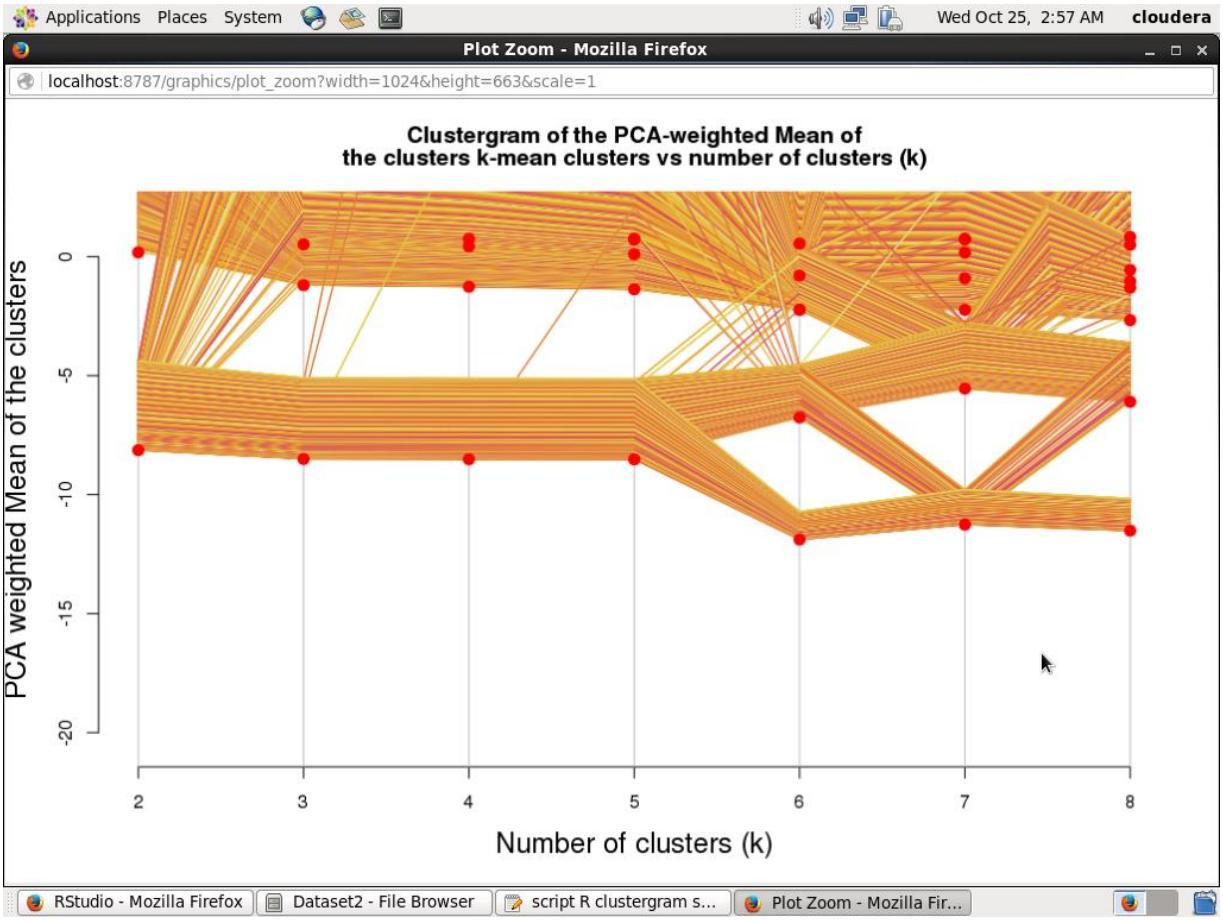


Figura 4.33 Gráfica Clustergram para el conjunto de datos 2.

Para la implementación de la técnica de visualización de coordenadas paralelas, es necesario crear el *script* en R correspondiente. La función de este *script* es tomar los datos que se encuentran en HDFS, formar tres conjuntos de datos con la información de los usuarios 2, 56 y 110, estos conjuntos de datos se comparan aplicando la técnica de coordenadas paralelas. La Figura 4.34 presenta la gráfica resultante en la cual se observa dicha comparación. También se aplica esta técnica a todo el conjunto de datos, como se contempla en la Figura 4.35, de donde se obtiene que el patrón de comportamiento en la mayoría de los usuarios es dar el menor número de clics al momento de realizar alguna compra.

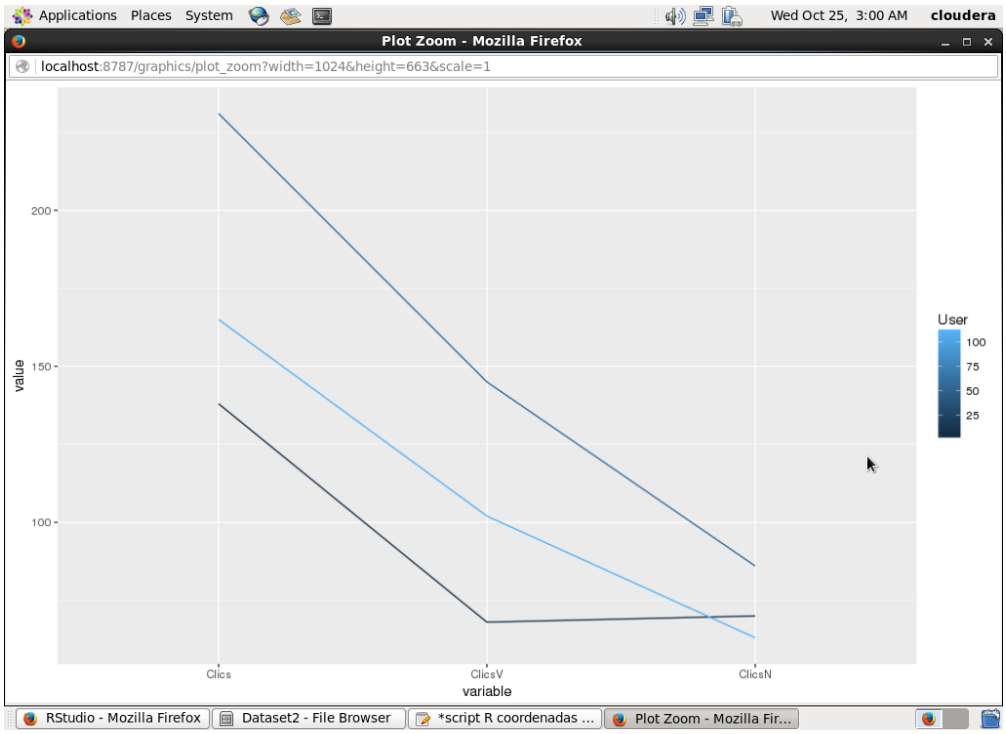


Figura 4.34 Gráfica de coordenadas paralelas con tres registros para el conjunto de datos 2.

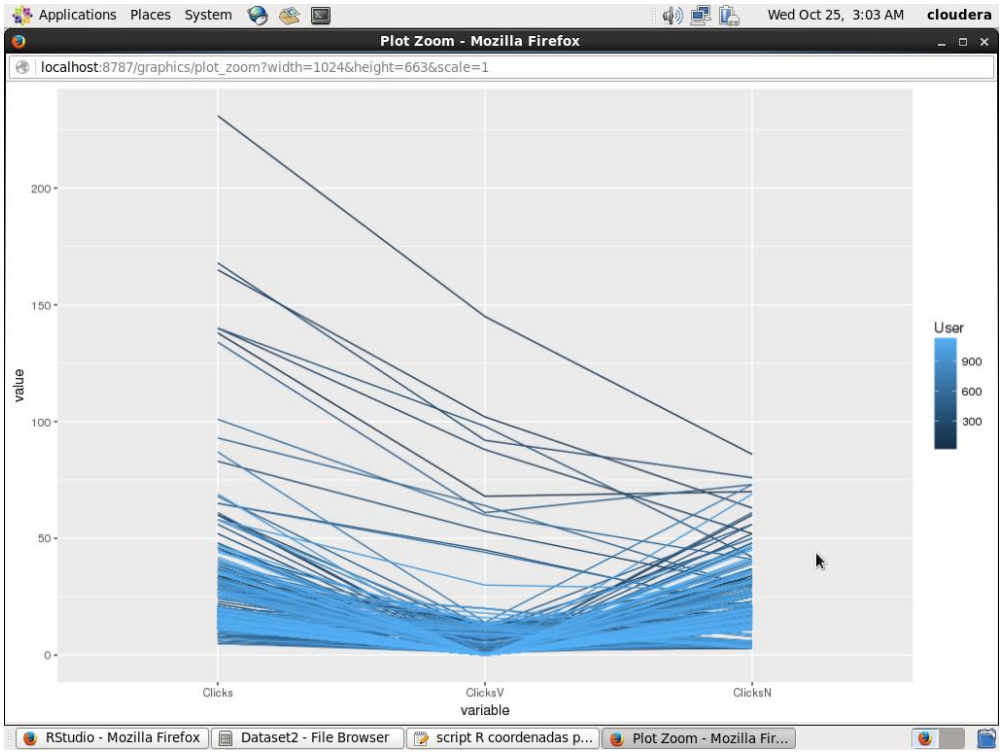


Figura 4.35 Gráfica de coordenadas paralelas para el conjunto de datos 2.

Para finalizar con este conjunto de datos, falta implementar la técnica de *Spatial Information Flow*. Como se mencionó anteriormente, para esta técnica se requiere de la información de latitud y longitud, pero este conjunto de datos no contempla dicha información, sin embargo, dentro de los datos generados por BigBench, se encuentra otro conjunto de datos con la información de los clientes, entre los que destaca el campo de lugar de nacimiento. El *script* de Apache Pig, que se muestra en la Figura 4.36, toma la información resumida que ya se había generado anteriormente, también toma la información de los clientes y las "une" mediante el campo de código de usuario, una vez unida la información, se procede a agrupar los datos por medio del lugar de nacimiento, contabilizando los usuarios de cada país, con esto, se tiene un archivo mucho más resumido con la información de país y el número de usuarios registrados. A este archivo de sólo 185 registros, se le agrega la información de latitud y longitud de cada país y así ya es posible visualizar la información. El archivo completo se muestra en la Figura 4.37.

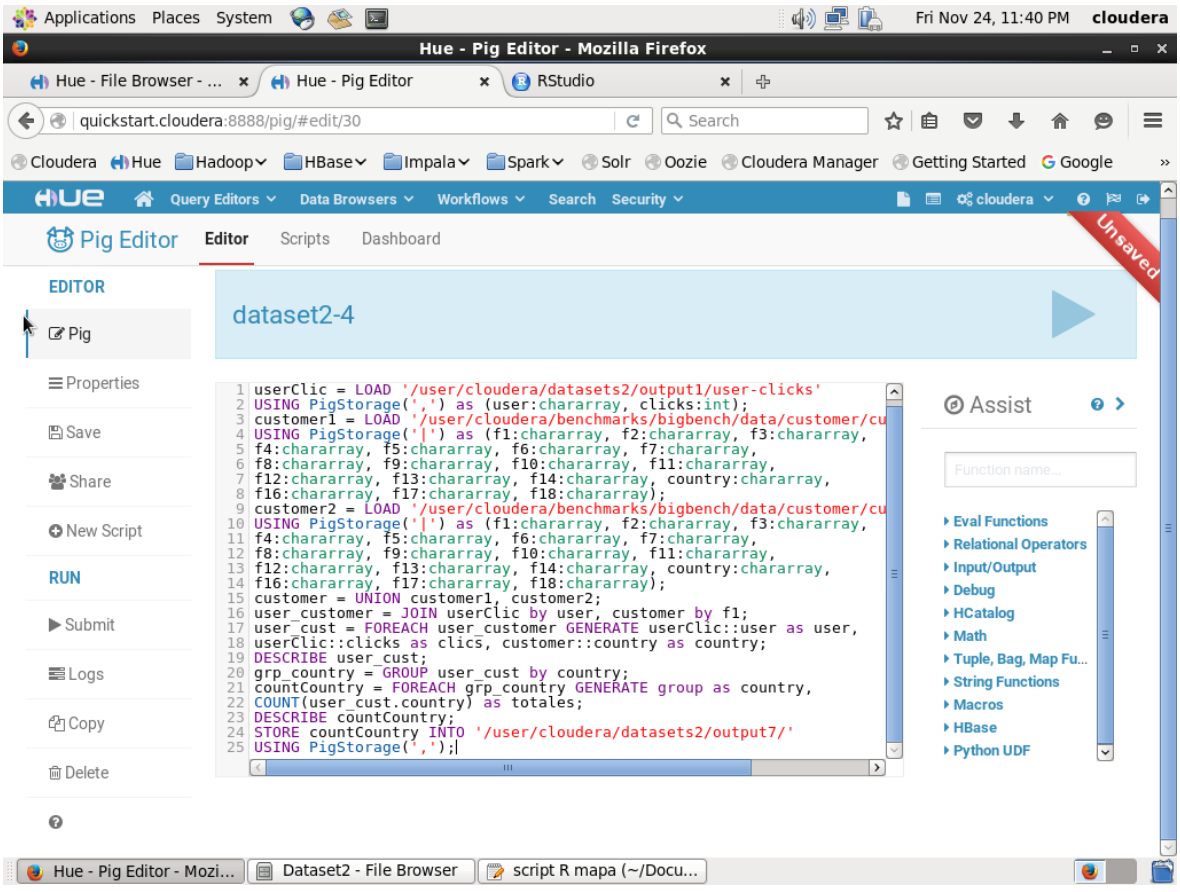


Figura 4.36 *Script* 3 de Apache Pig para el conjunto de datos 2.

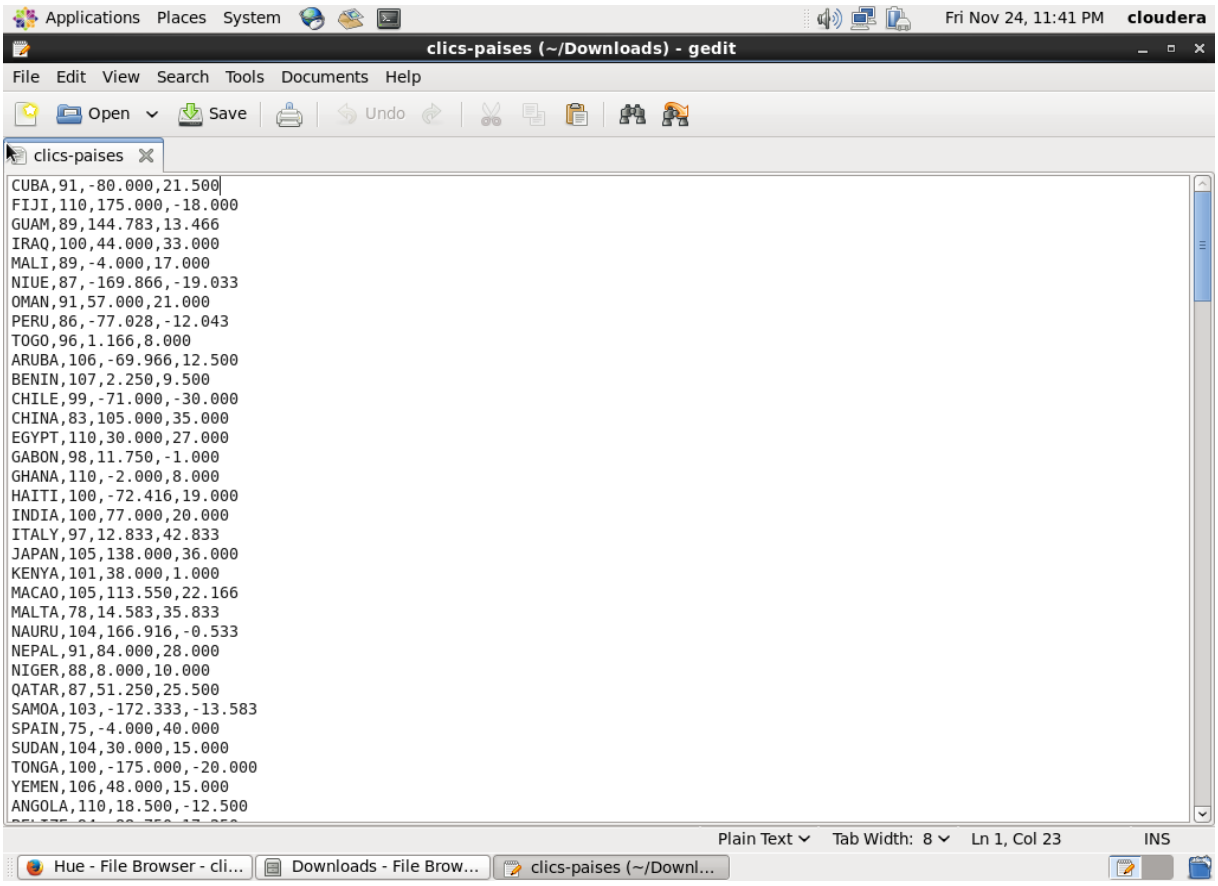


Figura 4.37 Archivo con latitud y longitud del conjunto de datos 2.

El *script* en R para *Spatial Information Flow* se le agrega una función para mostrar en verde los países que tienen más de 100 usuarios registrados, en naranja los países que tienen entre 90 y 100 usuarios y en rojo los países que tienen menos de 90 usuarios. La Figura 4.38 muestra la visualización en el mapa geográfico.

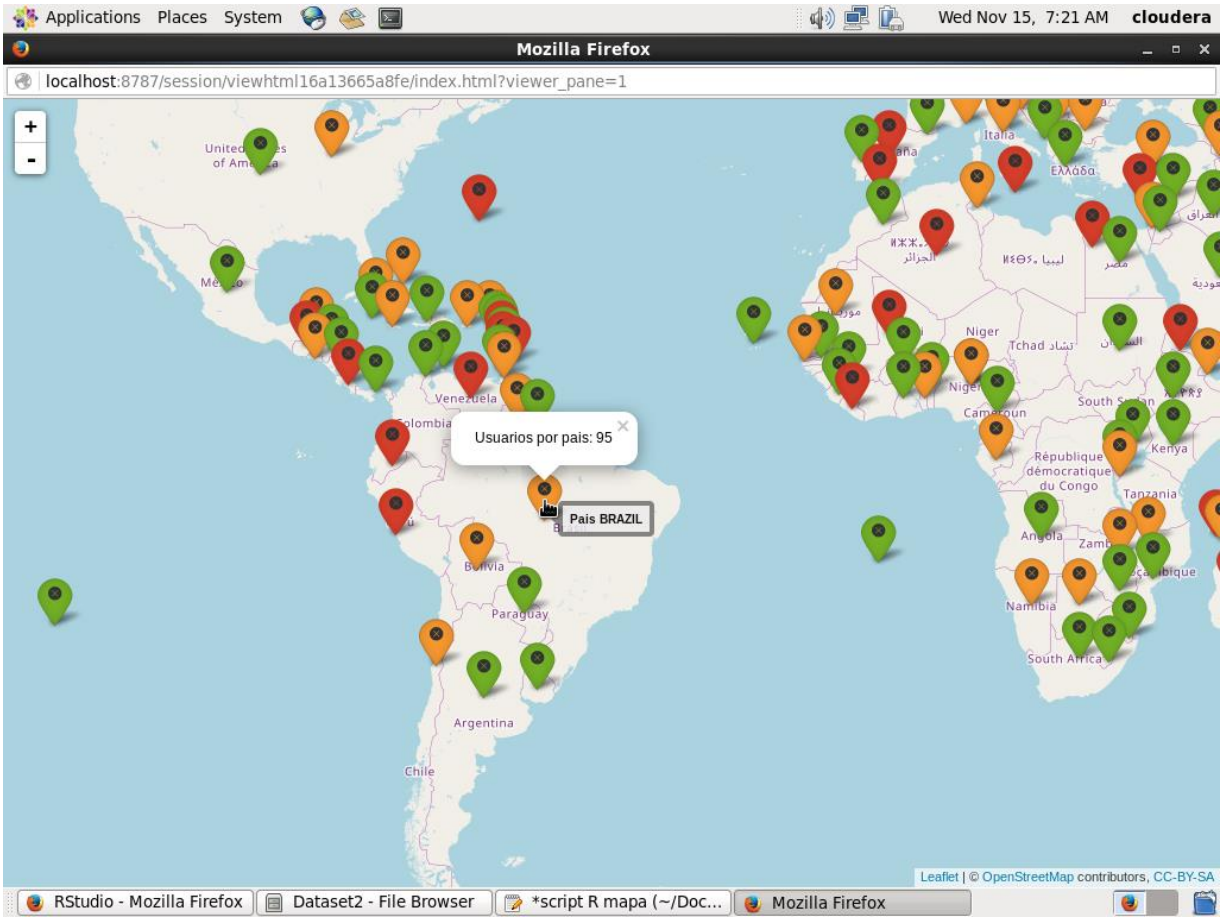


Figura 4.38 Gráfica de *Spatial Information Flow* para el conjunto de datos 2.

4.3. *Dataset* 3: Calificación de productos

Este conjunto de datos contiene información de las diferentes calificaciones que dan los usuarios a diversos productos ofrecidos por distintos proveedores. Contempla información como clave del registro, fecha y hora de la calificación, la calificación otorgada, clave del usuario, clave del producto, comentarios, entre otros. Esta información proviene de diferentes fuentes, usuarios registrados, usuarios invitados, usuarios registrados que hicieron la compra del producto, entre otros. De acuerdo a [96], este conjunto de datos se clasifica como datos no estructurados, por lo que cumple con la característica de variedad, sin embargo, al ser un conjunto de 89,991 registros no se considera que tenga volumen. Por otra parte, sí contempla la característica de velocidad, dado que los datos se generan constantemente por los usuarios.

Al ser un conjunto de datos sintéticos, debido a que los genera la herramienta BigBench, no se considera que los datos tengan veracidad. En la Figura 4.39 se aprecian los primeros registros de este conjunto de datos, almacenado en HDFS para su procesamiento.

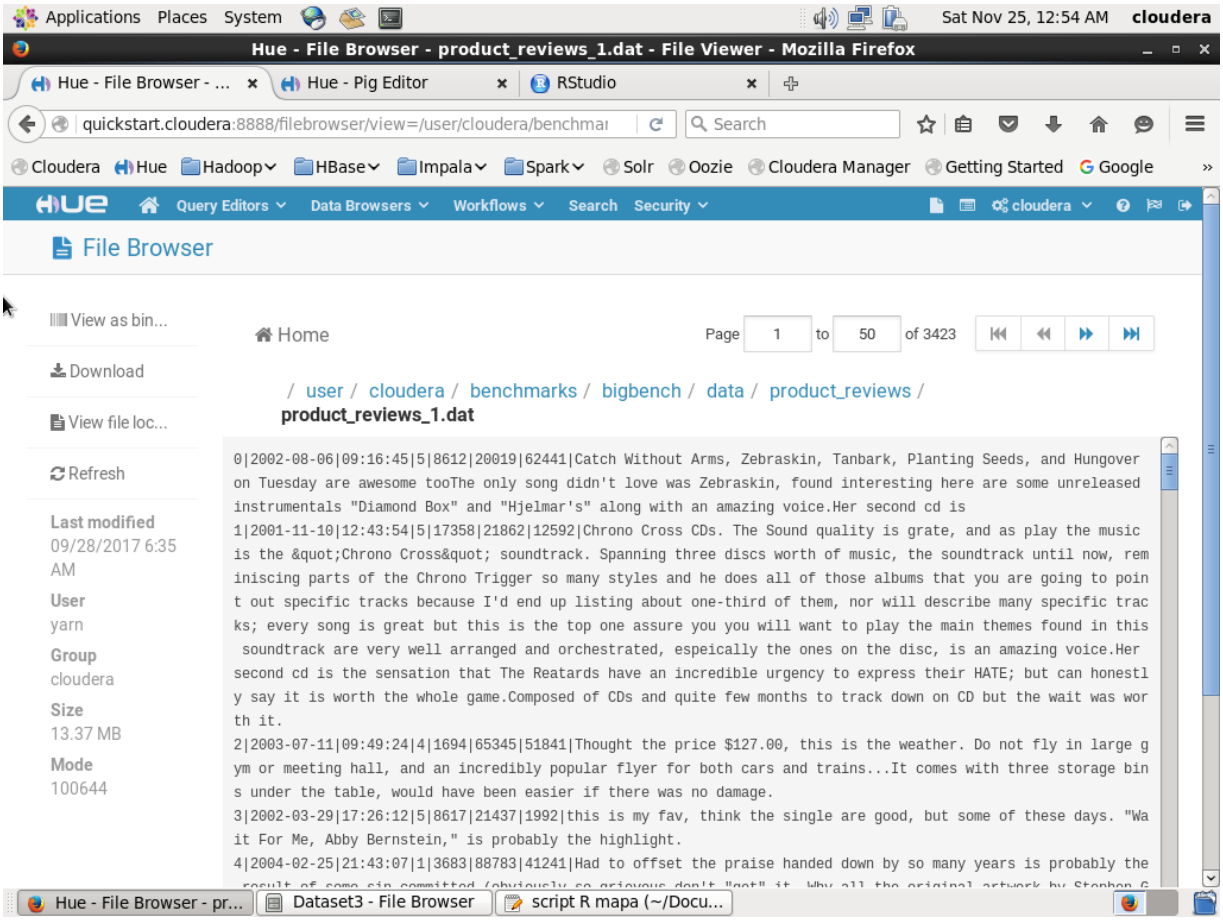


Figura 4.39 Conjunto de datos "Calificación de productos".

Como primer tratamiento de esta información, se creó un archivo de Apache Pig con la función principal de contabilizar todas las calificaciones. Cabe mencionar que BigBench genera el conjunto de datos en cuatro archivos, por lo que es necesario unir estos archivos al momento de ejecutar el *script* de Pig. Posteriormente, se agrupa la información por medio de la calificación otorgada y al mismo tiempo se contabiliza dicha información y se almacena el archivo resultante en HDFS, la Figura 4.40 muestra el *script* antes mencionado. Por otra parte, la Figura 4.41 muestra el resultado de la ejecución del *script*, en el cual se observa que se ha resumido el conjunto de datos original a solo cinco registros, que es la calificación otorgada y el total de veces que han registrado dicha calificación.

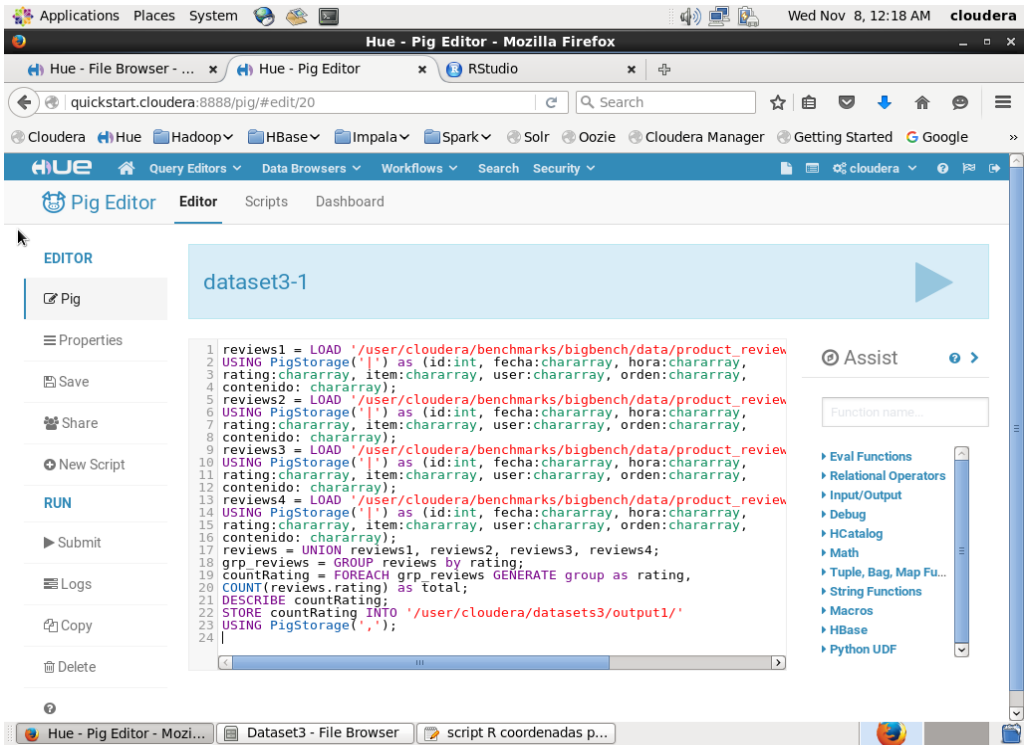


Figura 4.40 *Script* 1 de Apache Pig para el conjunto de datos 3.

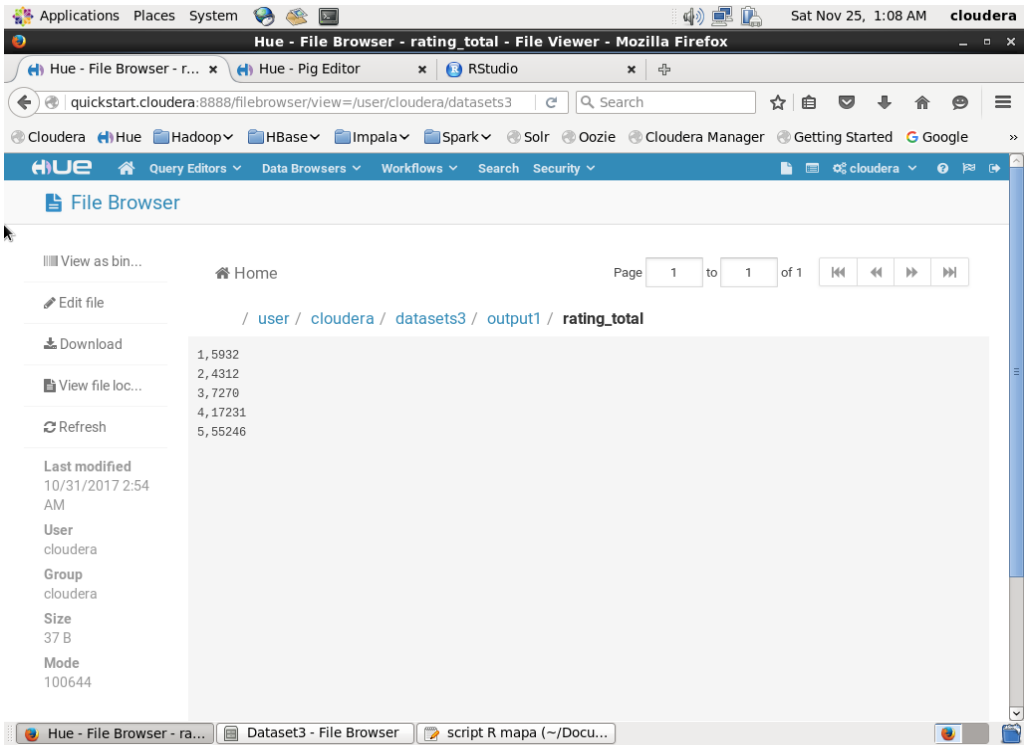


Figura 4.41 Archivo resultante del *script* 1 para el conjunto de datos 3.

Con este archivo resumido, se empieza con la implementación de las técnicas de visualización, específicamente con la gráfica de líneas. El *script* en R toma el conjunto de datos de HDFS y lo visualiza con el comando correspondiente. La gráfica resultante se presenta en la Figura 4.42, de la cual se observa que la mayoría de los usuarios está calificando de manera positiva los productos, de forma concreta con valores de cinco.

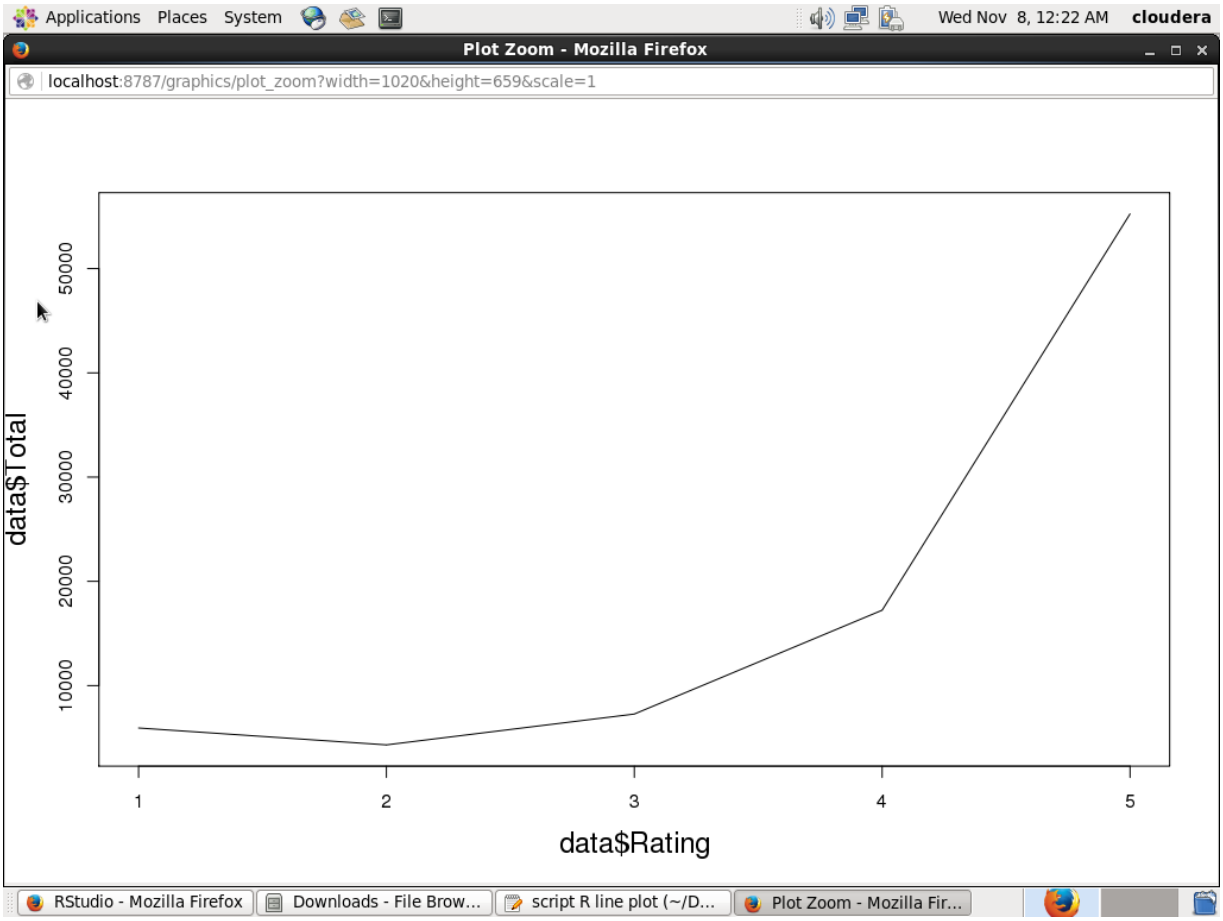


Figura 4.42 Gráfica de líneas para el conjunto de datos 3.

Con esta misma información, se implementa la técnica de visualización de gráfica de barras. El *script* en R, de la misma forma que el anterior, toma la información de HDFS y la transforma en un objeto de tipo matriz para que sea visualizado por el comando, en la Figura 4.43 se expone la gráfica resultante y a pesar de ser la misma información, en esta gráfica se aprecia de mejor manera los datos.

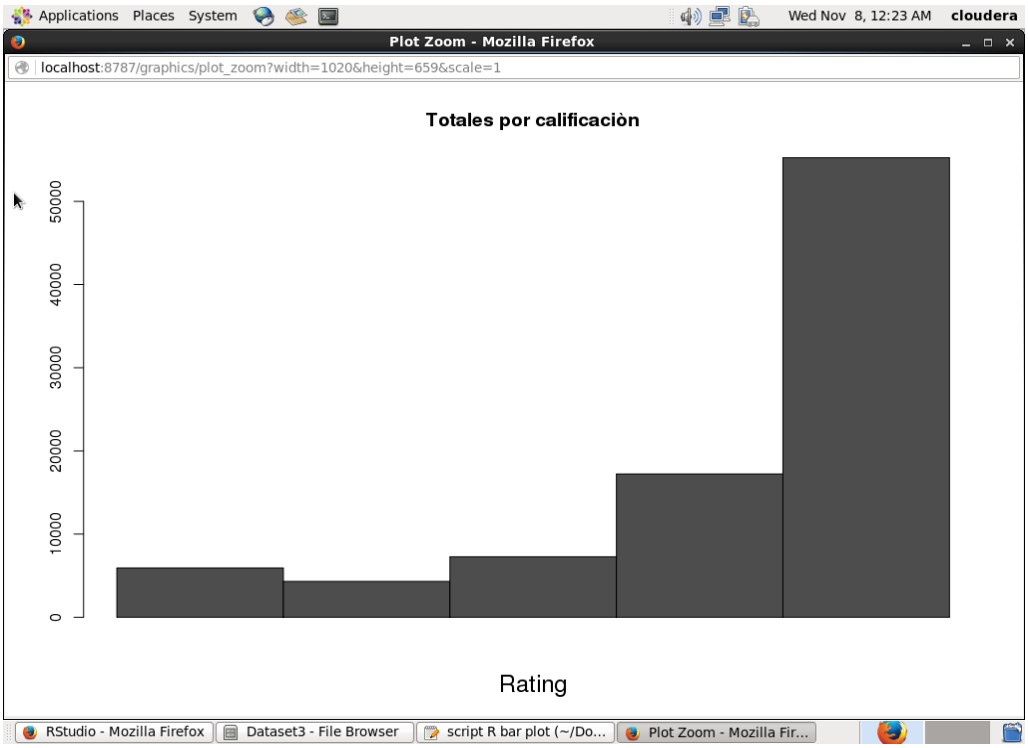


Figura 4.43 Gráfica de barras para el conjunto de datos 3.

Ahora se procede a implementar la gráfica de dispersión para este conjunto de datos. La función principal del *script* en R es tomar la información de HDFS e implementarla con el comando de gráfica de dispersión. La Figura 4.44 presenta la gráfica resultante, la cual resulta ser una gráfica sencilla y fácil de entender, esto se debe a que la información que se está visualizando es muy pequeña en comparación con las gráficas de los anteriores conjuntos de datos, en donde había la necesidad de presentar los primeros 50 o 100 registros, pero para este caso no es necesario aplicar este "recorte".

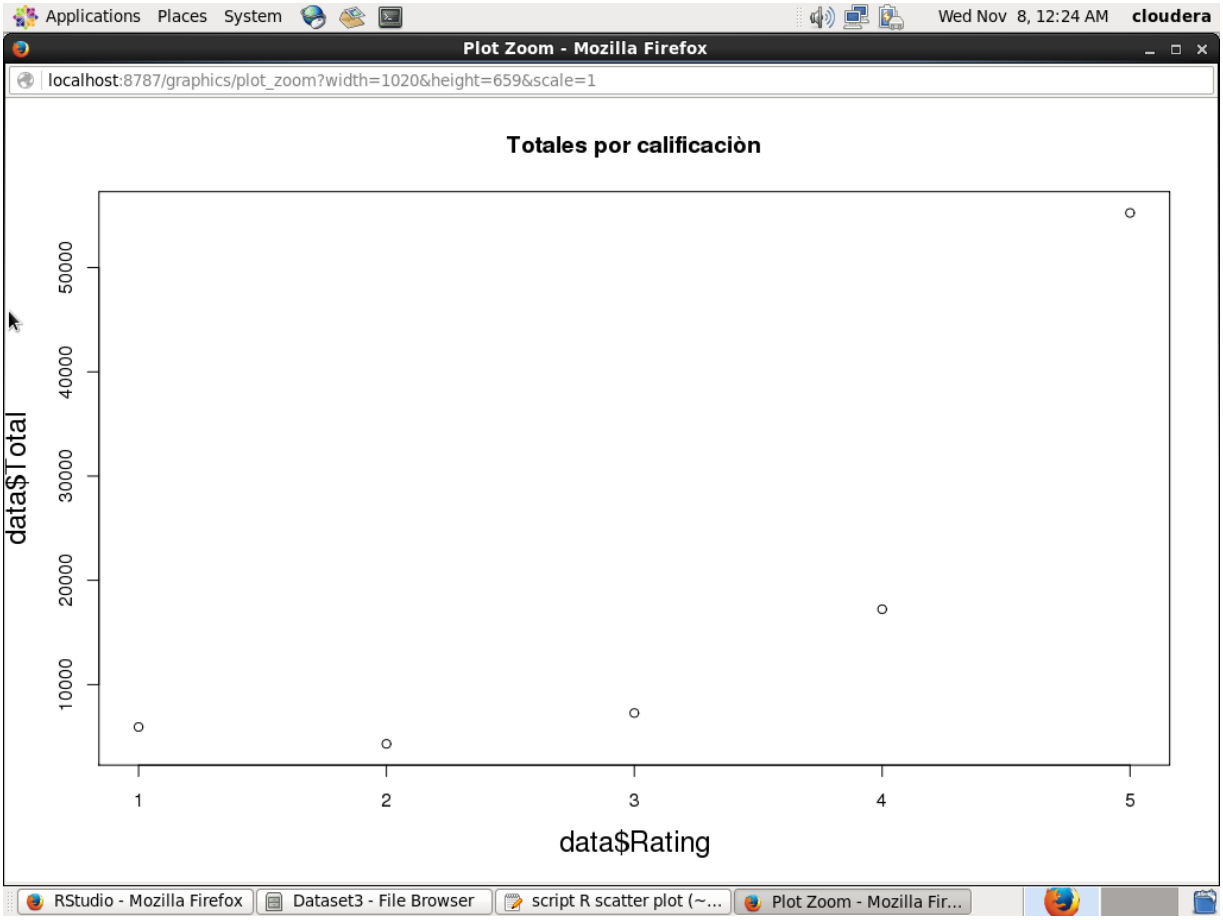


Figura 4.44 Gráfica de dispersión para el conjunto de datos 3.

Cambiando el análisis del conjunto de datos original, ahora se desea conocer las calificaciones que ha dado cada usuario. Para esto se genera un nuevo *script* de Apache Pig, como se muestra en la Figura 4.45, en el cual, se agrupa la información original mediante el código de usuarios, posteriormente se contabiliza para cada usuario las veces que ha registrado una calificación de uno, dos, tres, cuatro o cinco, todo esto se almacena en un archivo en HDFS que se muestra en la Figura 4.46.

Ahora toca el turno de implementar la técnica de visualización *Heatmap* con los datos generados en la ejecución del último *script*. Se genera el *script* en R para esta implementación. Este *script* toma los datos de HDFS para visualizarlos con el comando correspondiente.

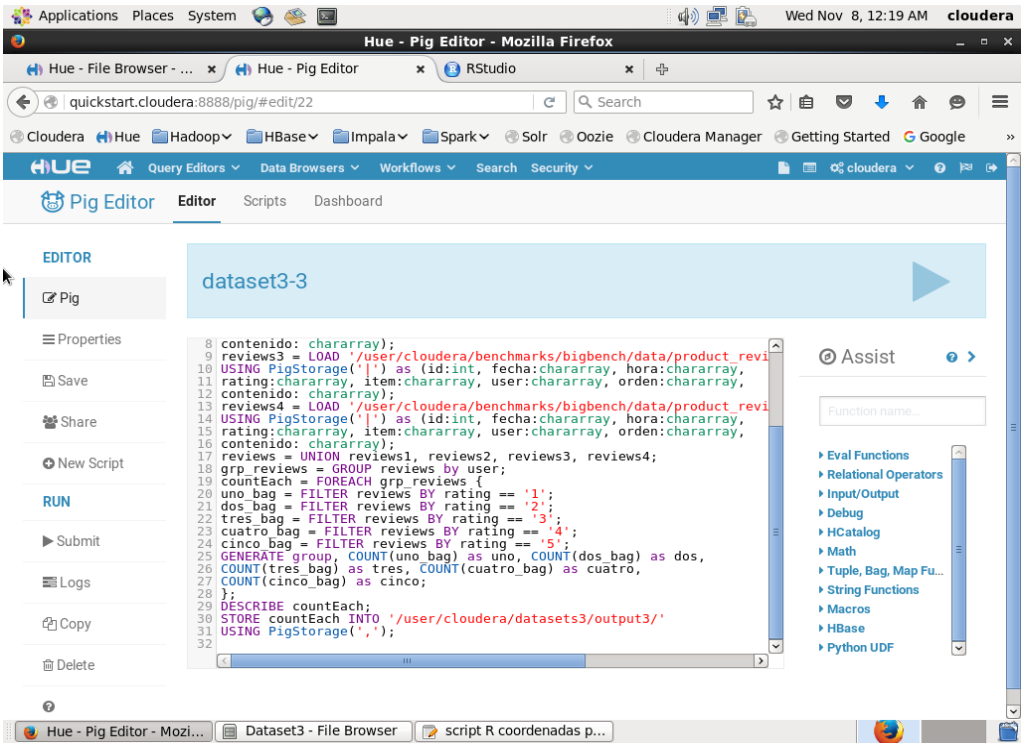


Figura 4.45 *Script* 2 de Apache Pig para el conjunto de datos 3.

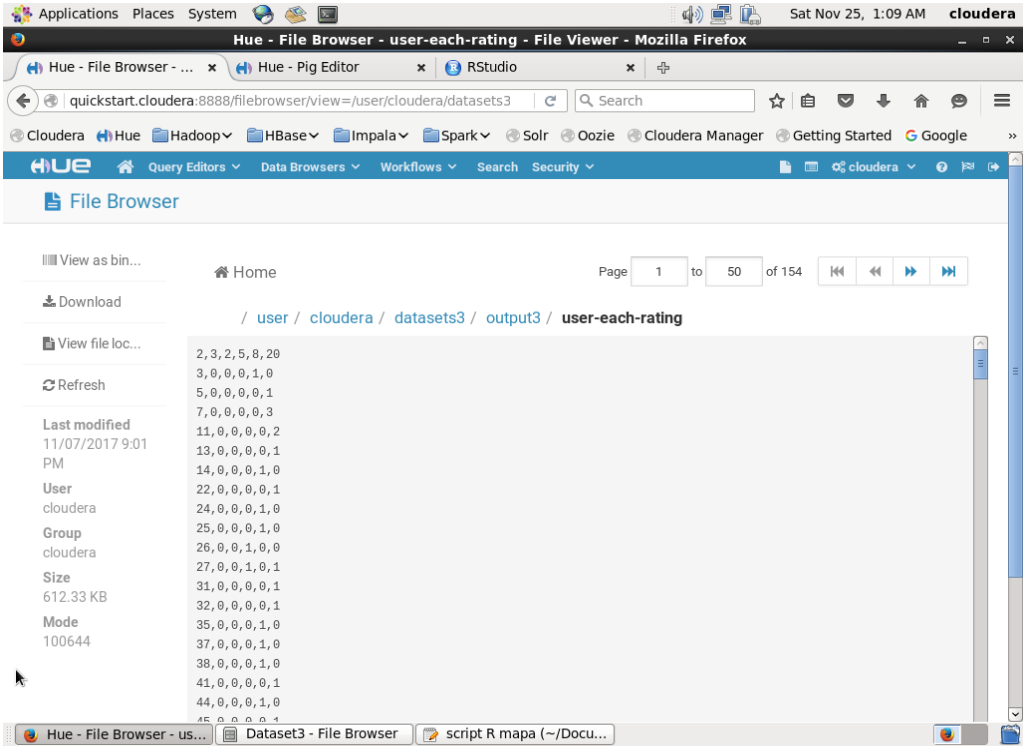


Figura 4.46 Archivo resultante del *script* 2 para el conjunto de datos 3.

Con la implementación del *script* en R se genera la gráfica *Heatmap* que se aprecia en la Figura 4.47. En esta gráfica, los colores naranja implican un valor menor, los colores amarillo un valor intermedio y los colores blancos representan un número mayor, todo esto referente al número total de calificaciones por usuario, sin embargo, al ser demasiados registros, no se logra apreciar de manera clara los códigos de usuario para conocer qué usuarios están calificando en mayor número con cinco. En este caso es necesario presentar los primero 50 registros para tener una visión más clara, esto se expone en la Figura 4.48. Esta modificación hace una variación en los colores presentados, ahora se da a entender que los colores rojizos corresponden a los valores menores, los colores amarillos a los valores intermedios y los colores blancos siguen siendo los de mayor valor. Con esta nueva gráfica se observa, por ejemplo, que el usuario 2 tiene un mayor número de calificaciones en cinco respecto a las otras calificaciones.

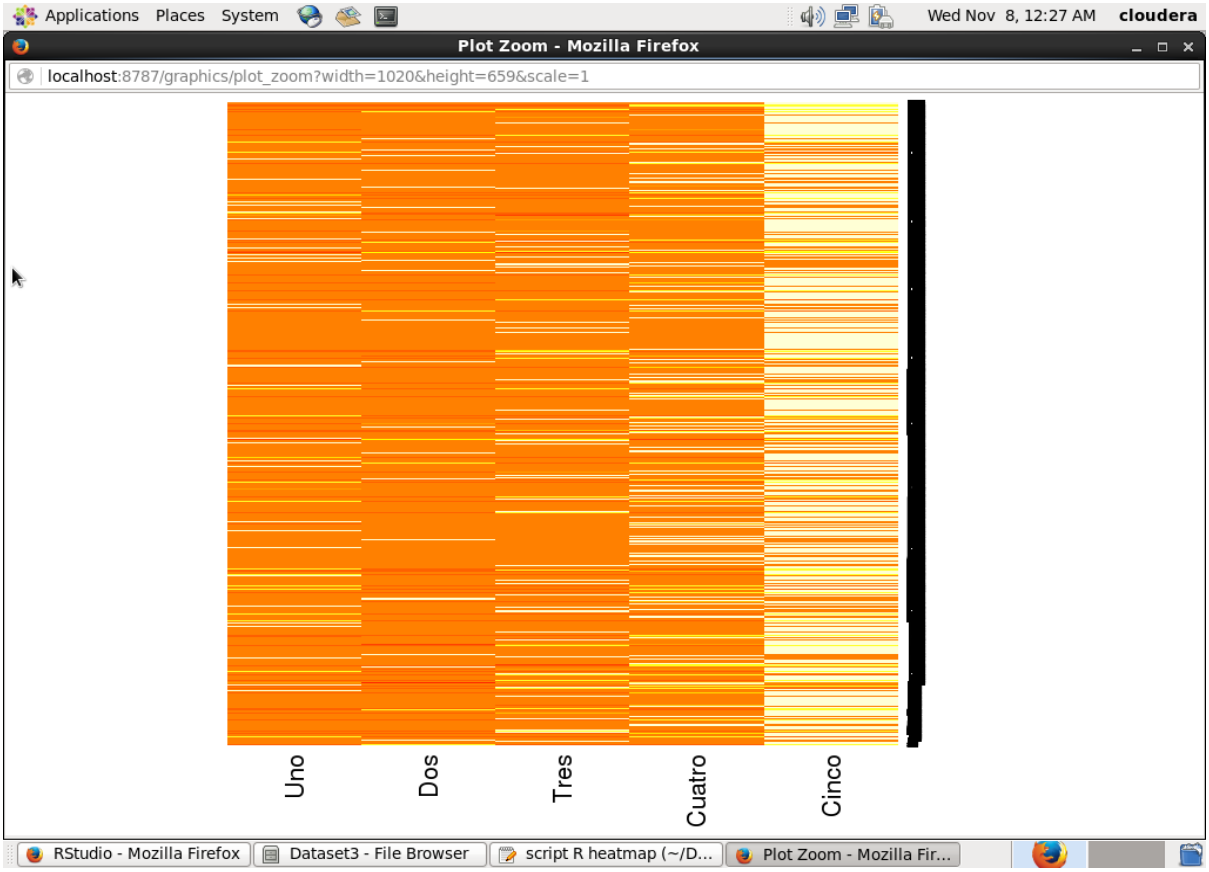


Figura 4.47 Gráfica *Heatmap* para el conjunto de datos 3.

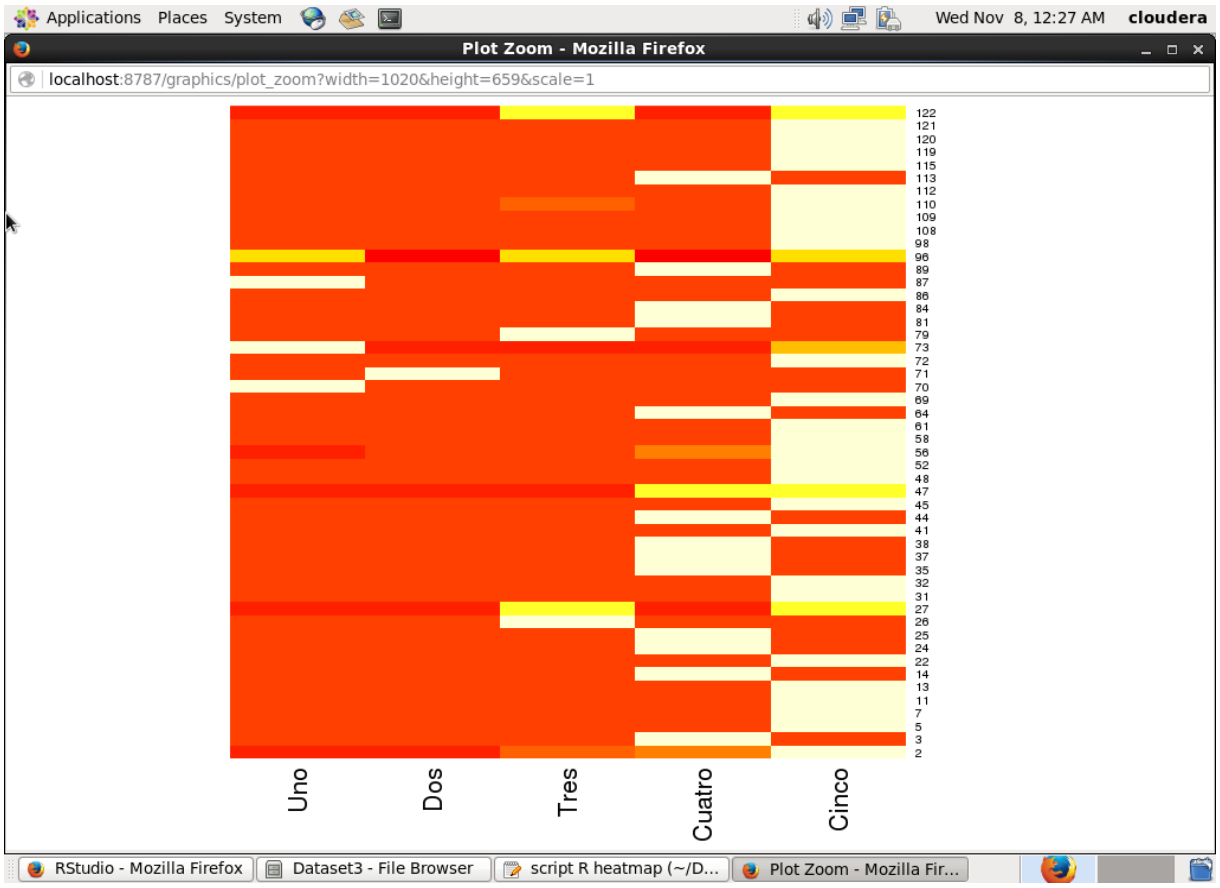


Figura 4.48 Gráfica *Heatmap* con 50 registros para el conjunto de datos 3.

Para la siguiente técnica de visualización, la cual es la gráfica *Treemap*, es necesario volver a procesar los datos originales. Ahora se desea conocer los usuarios y el total de sus calificaciones, ya no de forma separada como el conjunto de datos anterior, si no de forma global. La Figura 4.49 muestra el *script* de Apache Pig para lograr este objetivo.

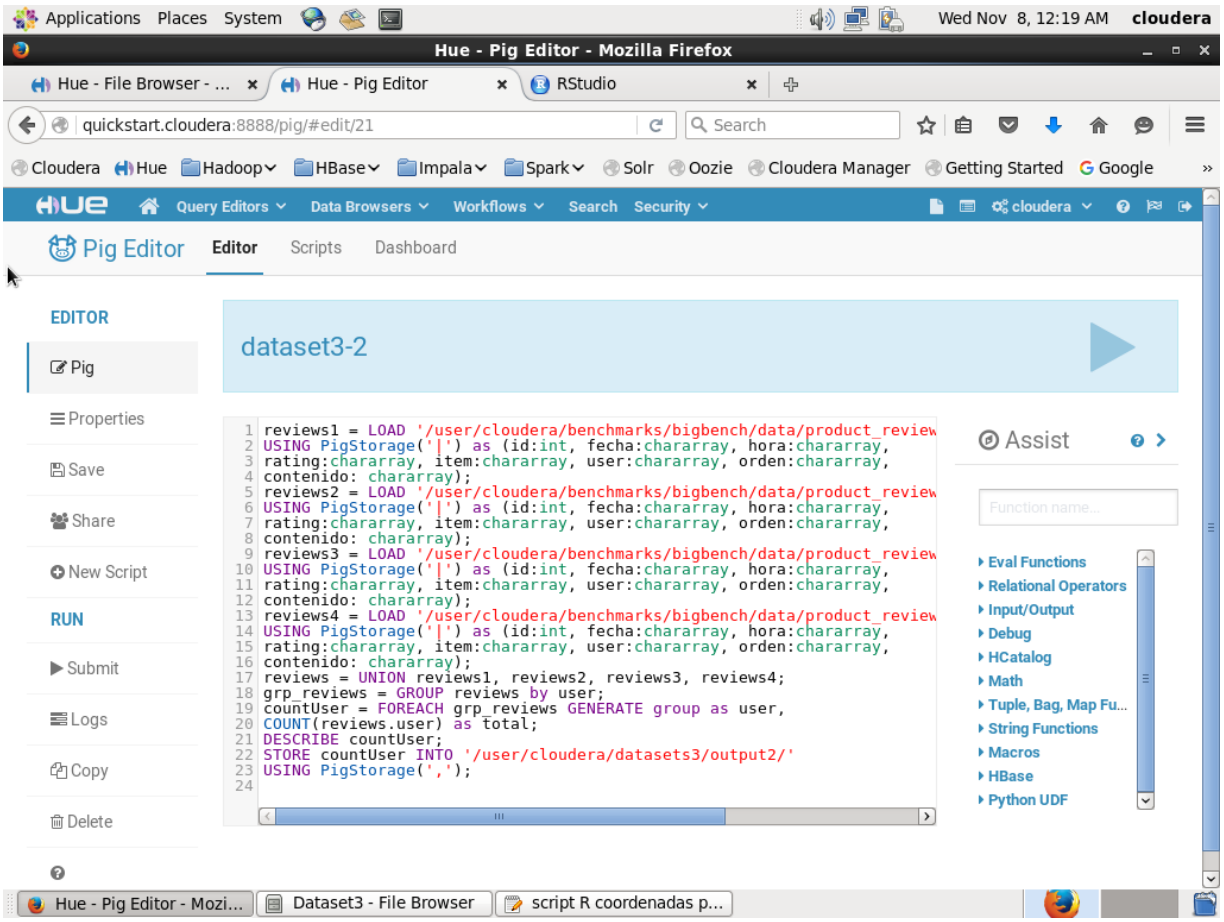


Figura 4.49 *Script* 3 de Apache Pig para el conjunto de datos 3.

El archivo resultante de este *script* se muestra en la Figura 4.50, el cual solo tiene dos variables, el código de usuario y el número total que ha registrado calificaciones. Con esta información ahora sí es posible construir la gráfica *Treemap* para determinar qué usuarios son los que tienen mayor número de calificaciones registradas. Cabe mencionar que la visualización total de este conjunto de datos no se logra concretar, debido a que son demasiados registros, sin embargo, se muestran los primeros 1,000 registros. La gráfica *Treemap* resultante se presenta en la Figura 4.51, de la cual se deduce que los usuarios 326, 218 y 56 son los usuarios que han registrado más calificaciones que los demás, también dando a entender que son los usuarios más activos.

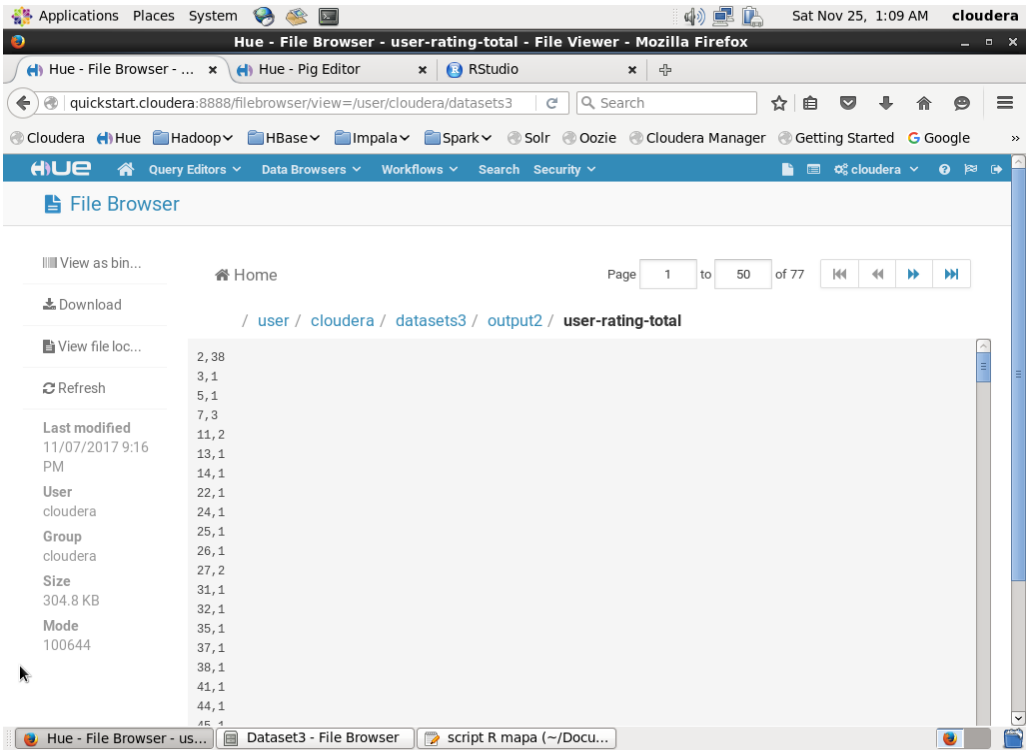


Figura 4.50 Archivo resultante del *script* 3 para el conjunto de datos 3.

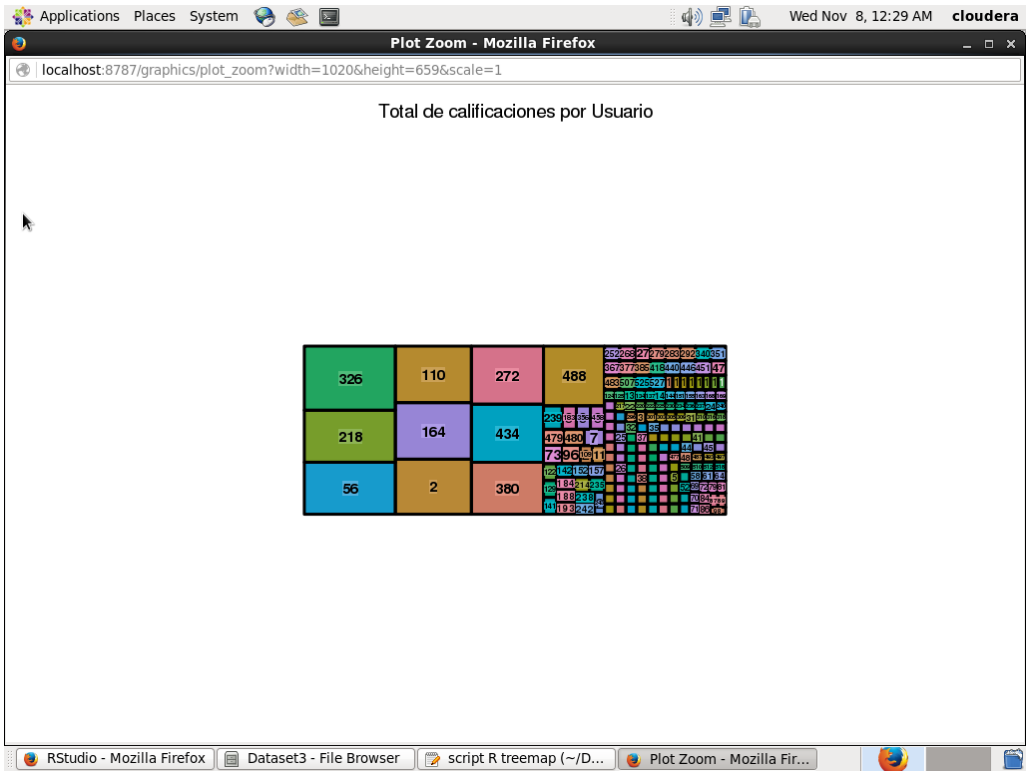


Figura 4.51 Gráfica *Treemap* para el conjunto de datos 3.

Con la misma información de la anterior técnica de visualización, se procede a implementar la técnica de Clustergram para conocer cómo se comportarían los datos si se agrupan en diferentes cúmulos. Como ya se ha mencionado antes, esta técnica hace uso del algoritmo *k-means* dentro del *script* en R para el agrupamiento de los datos. La Figura 4.52 es la gráfica resultante. Al ser demasiados datos, no se logra apreciar de buena forma toda la información, sin embargo, sí se logra apreciar algunos cambios de los datos a través de los diferentes grupos que se están visualizando.

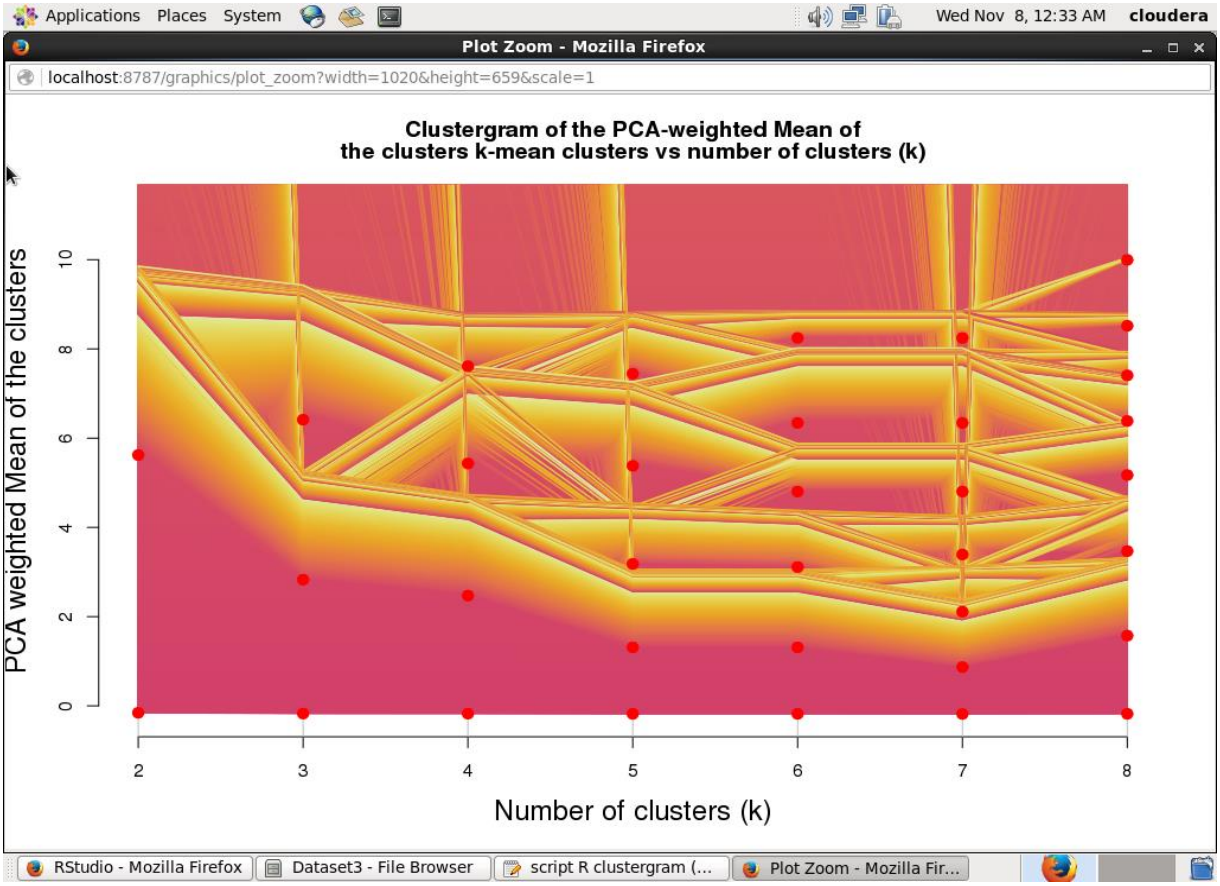


Figura 4.52 Gráfica Clustergram para el conjunto de datos 3.

Ahora se tomará la información que se mostró en la gráfica *Heatmap* para ser usada en la gráfica de coordenadas paralelas, esto con el fin de realizar la comparación entre las diferentes calificaciones de los usuarios. El *script* en R toma los datos de HDFS, de los cuales, se toman como muestra a los usuarios 2, 56 y 110 y se les genera la gráfica correspondiente que se presenta en la Figura 4.53. También se realiza la gráfica para todo el conjunto de datos, dando

como resultado la Figura 4.54, de la cual se observa los cambios en las diferentes calificaciones de los usuarios, presentando una mayor cantidad en la calificación cinco.

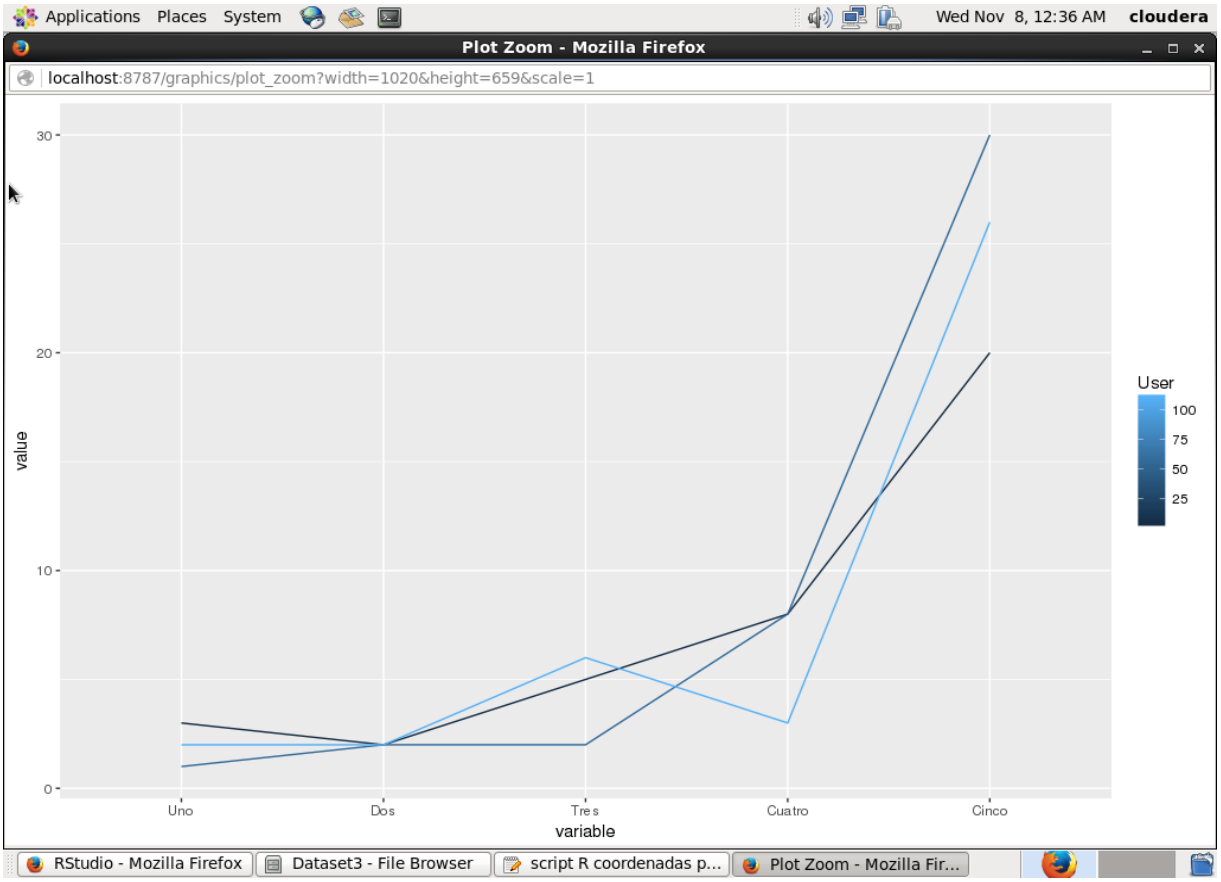


Figura 4.53 Gráfica de coordenadas paralelas con tres registros para el conjunto de datos 3.

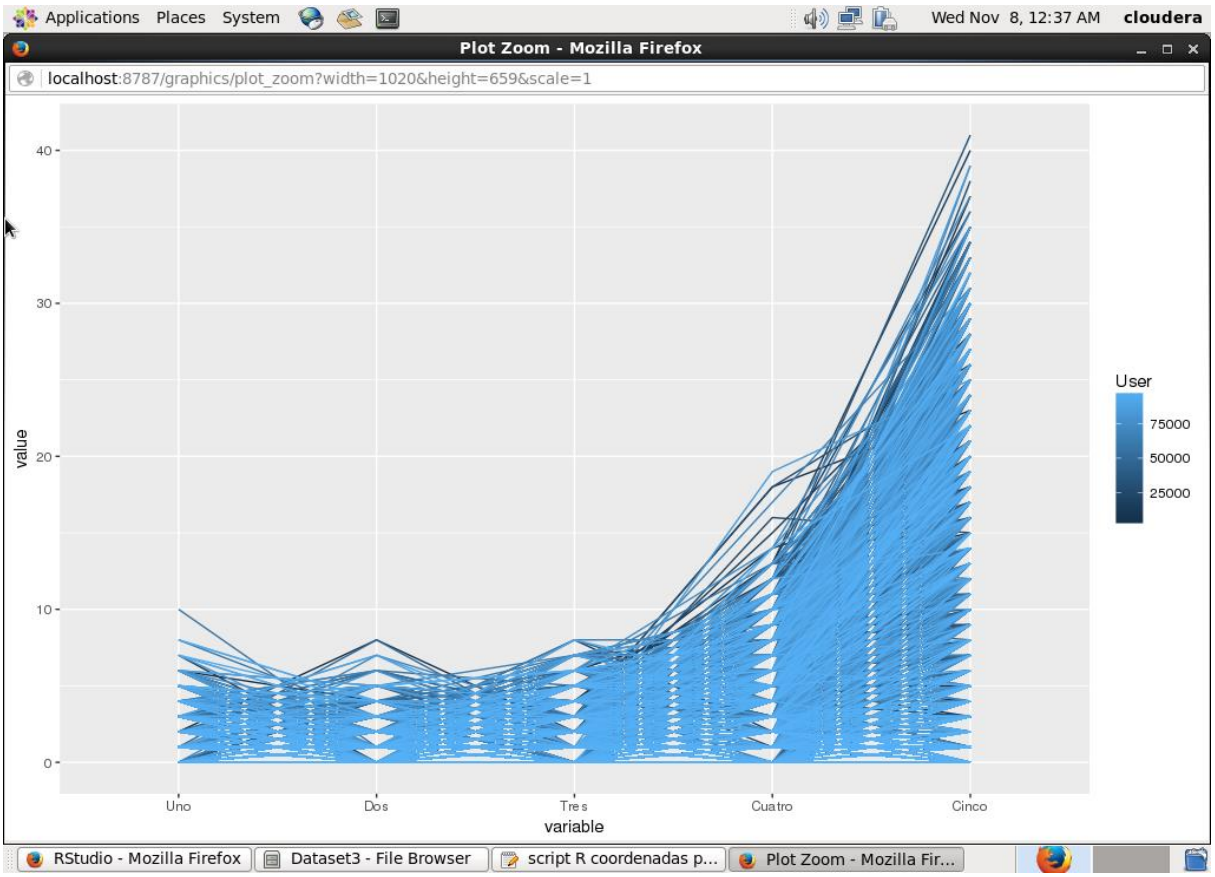


Figura 4.54 Gráfica de coordenadas paralelas para el conjunto de datos 3.

Como última técnica de visualización a implementar en este conjunto de datos, se tiene a *Spatial Information Flow* y en este caso, se aplicará el mismo flujo de trabajo que se aplicó en el conjunto de datos de flujos de clics, esto es, consultar y unir la información del conjunto de datos de clientes a través del código de usuario, agrupar los datos mediante el país de nacimiento y, para este caso, sumar las calificaciones de los usuarios. Con esto se da a conocer la suma de los puntajes que ha dado cada país a los diferentes productos. La Figura 4.55 muestra el *script* de Apache Pig que se acaba de explicar. El resultado de este *script* es un archivo con 185 registros, del cual se modifica manualmente para agregarle la latitud y longitud de cada país y así cumplir con los datos necesarios para implementar esta técnica de visualización. La Figura 4.56 muestra el archivo con los datos necesarios.

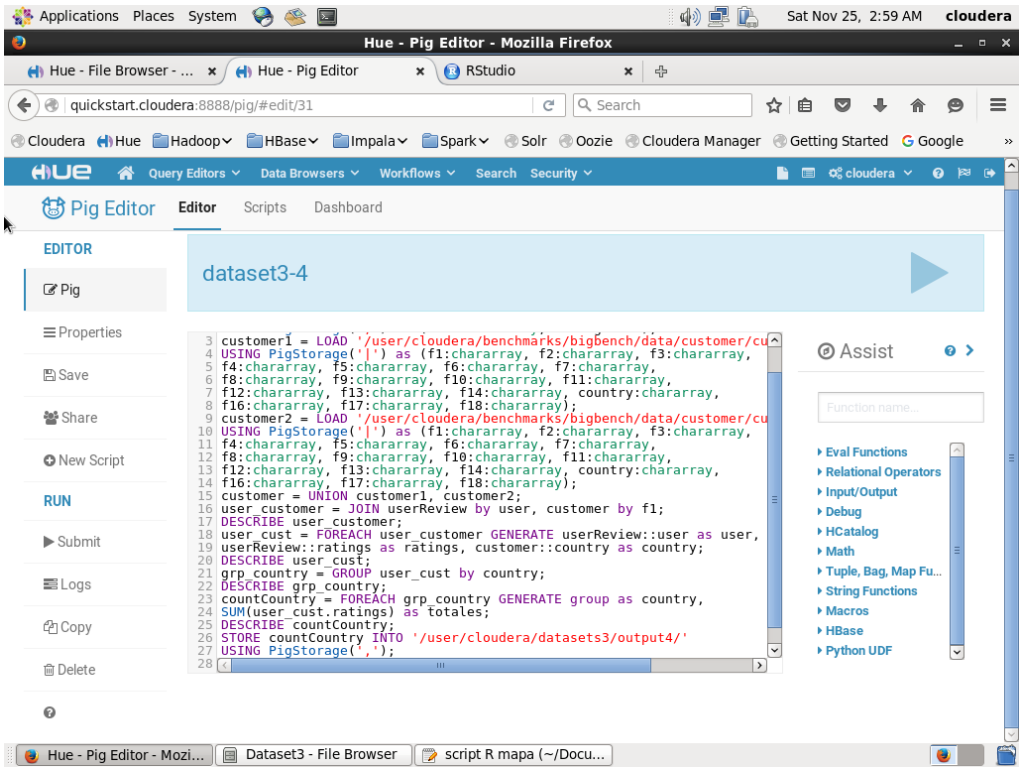


Figura 4.55 *Script* 4 de Apache Pig para el conjunto de datos 3.

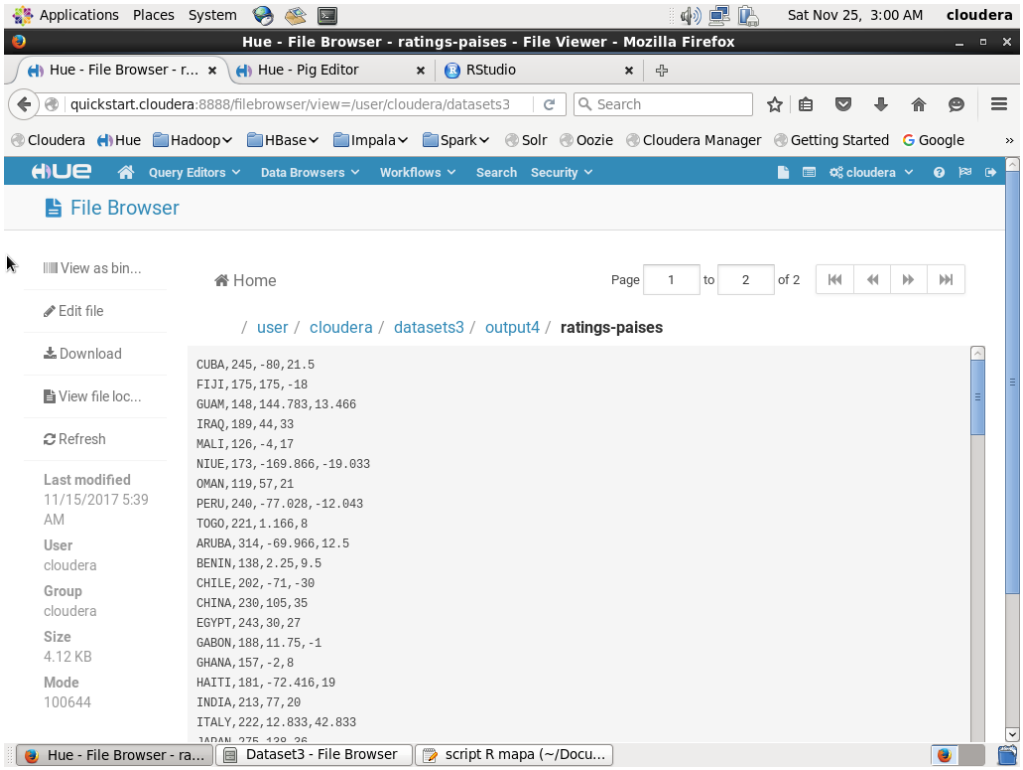


Figura 4.56 Archivo con latitud y longitud del conjunto de datos 3.

Finalmente es turno de implementar la técnica de *Spatial Information Flow* una vez que ya se tienen los datos completos. El *script* en R contiene una función para mostrar en verde los países que han otorgado más de 200 puntos de calificación, en naranja los países entre 150 y 200 puntos y en rojo los países que han otorgado menos de 150 puntos de calificación. Esta gráfica se muestra en la Figura 4.57 donde, por ejemplo, el país de Paraguay solo ha otorgado 116 puntos de calificación entre sus usuarios registrados, por lo que su marcador se ilumina en color rojo.

Figura 4.57 Gráfica de *Spatial Information Flow* para el conjunto de datos 3.

4.4. *Dataset* 4: Calificación de películas

Este conjunto de datos se extrajo de la herramienta BDGS dentro del *benchmark* BigDataBench. Se trata de datos extraídos de la página Web MovieLens [99], la cual se encarga de que sus usuarios califiquen diversas películas en un rango de 0.5 a 5 puntos y mediante su sistema de recomendación, ofrecer a sus usuarios películas afines con sus gustos. Este conjunto de datos contempla tres características de *Big Data*, como es volumen, variedad y veracidad. Contiene volumen, ya que el conjunto posee 20,000,264 registros repartidos en 4 variables (clave de usuario, clave de la película, calificación y tiempo). Contempla variedad, ya que de acuerdo con [97], estos datos se consideran semi-estructurados. Finalmente tiene veracidad al tratarse de datos reales extraídos de una página Web funcional. En la Figura 4.58 se muestran los primeros registros de este conjunto de datos almacenado en HDFS.

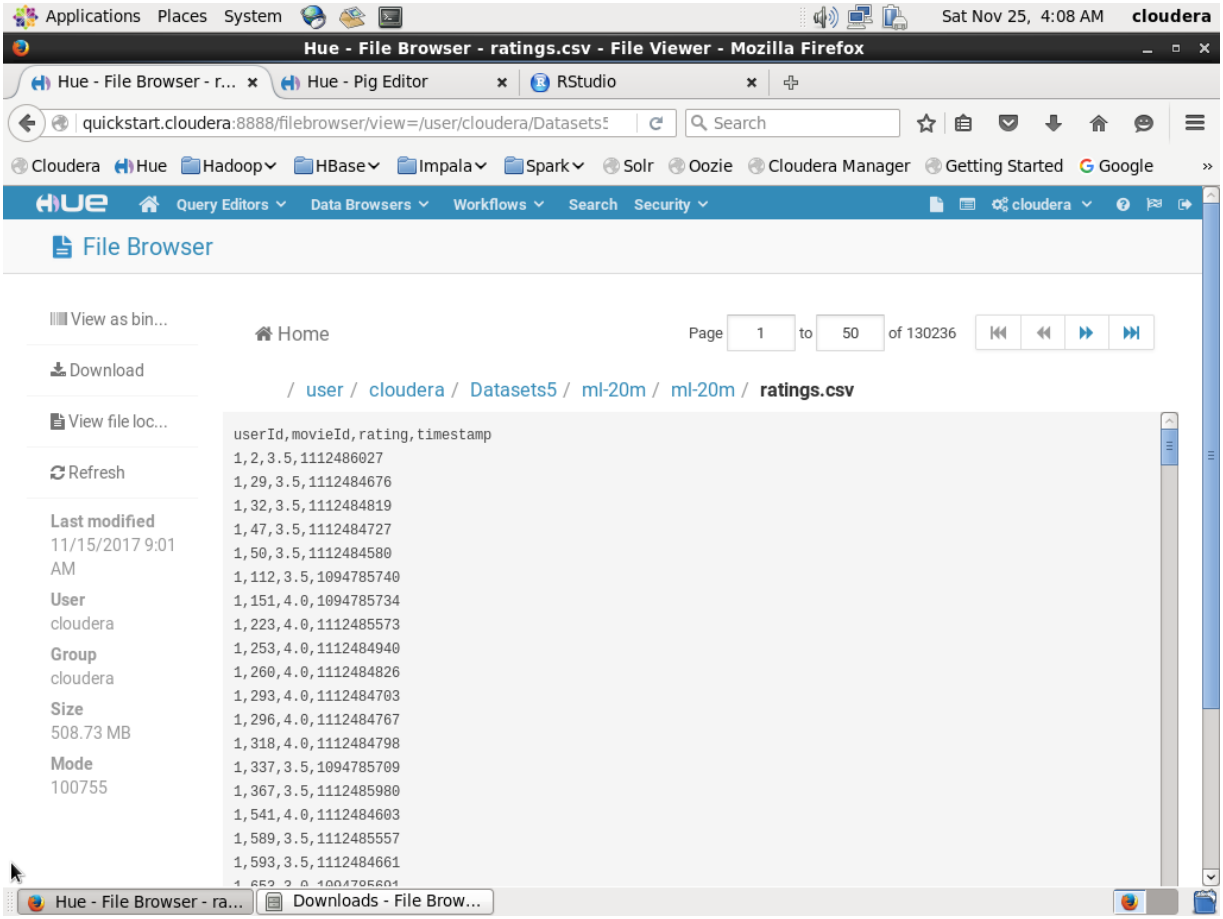


Figura 4.58 Conjunto de datos "Calificación de películas".

En primera instancia se desea conocer cómo están distribuidas las calificaciones, es decir, conocer el número total de registros por cada calificación otorgada. Para cumplir con esta tarea, es necesario generar el *script* en Apache Pig, como se muestra en la Figura 4.59. Este *script* se encarga de agrupar toda la información de acuerdo a la calificación otorgada, posteriormente se contabiliza cada calificación y se almacena en HDFS para su posterior visualización. El archivo resultante se muestra en la Figura 4.60, este archivo solo tiene dos variables (calificación y número de total de dichas calificaciones) y tan solo 10 registros.

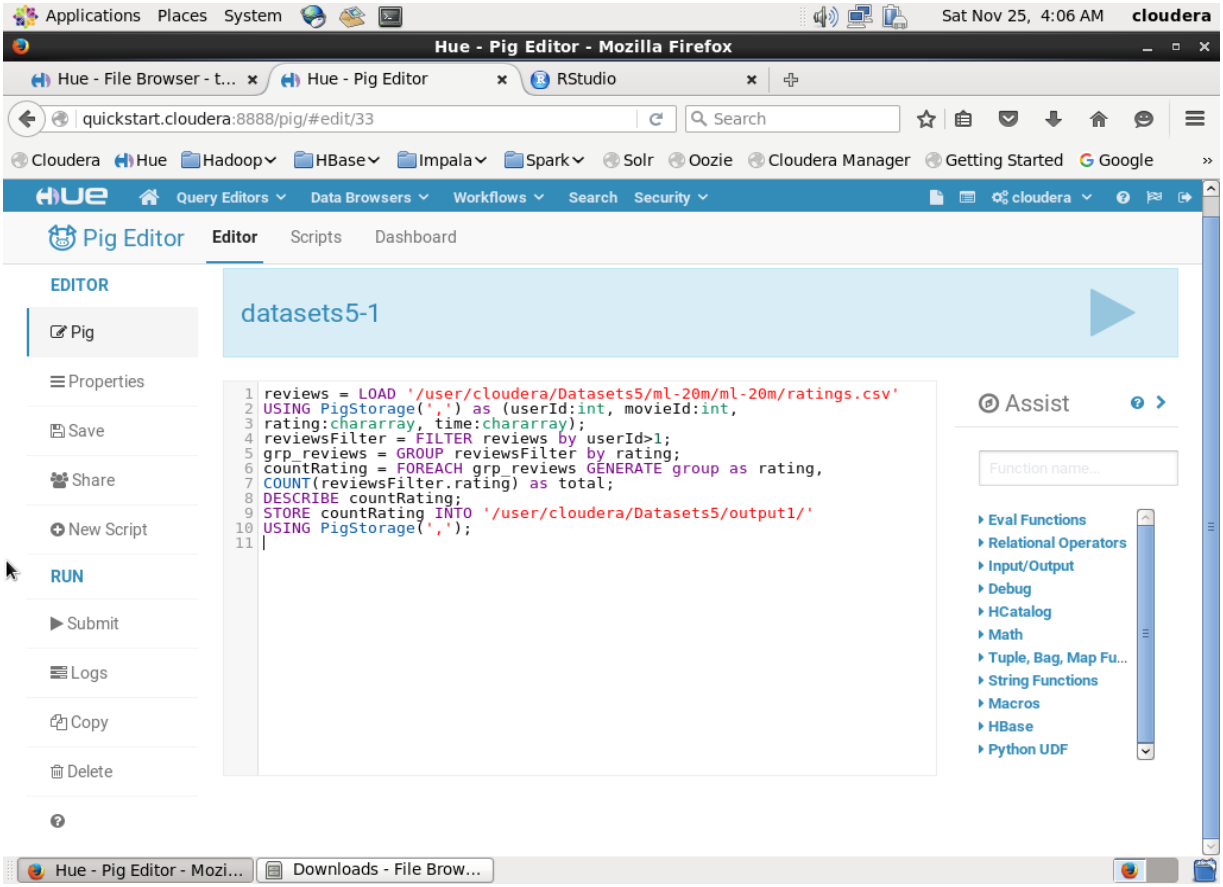


Figura 4.59 *Script* 1 de Apache Pig para el conjunto de datos 4.

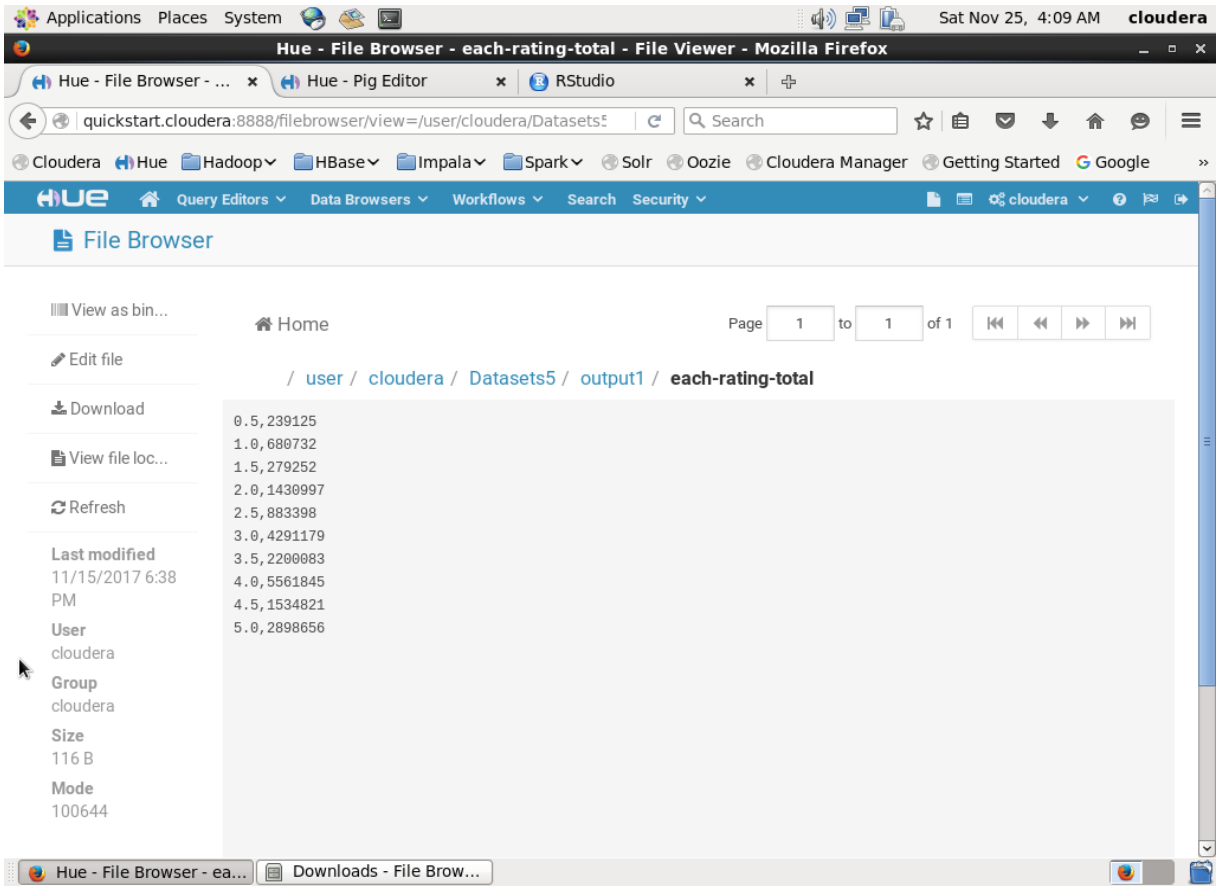


Figura 4.60 Archivo resultante del *script* 1 para el conjunto de datos 4.

Después de haber resumido más de 20 millones de registros, se procede con la implementación de las primeras técnicas de visualización. En primer lugar se implementa la gráfica de líneas, generando el *script* en R. Este *script* toma los datos de HDFS y los visualiza con el comando correspondiente, generando la gráfica que se muestra en la Figura 4.61, de la cual se observa que el mayor número de calificaciones otorgadas son de 3 y 4 puntos, manteniendo un promedio intermedio en las películas revisadas.

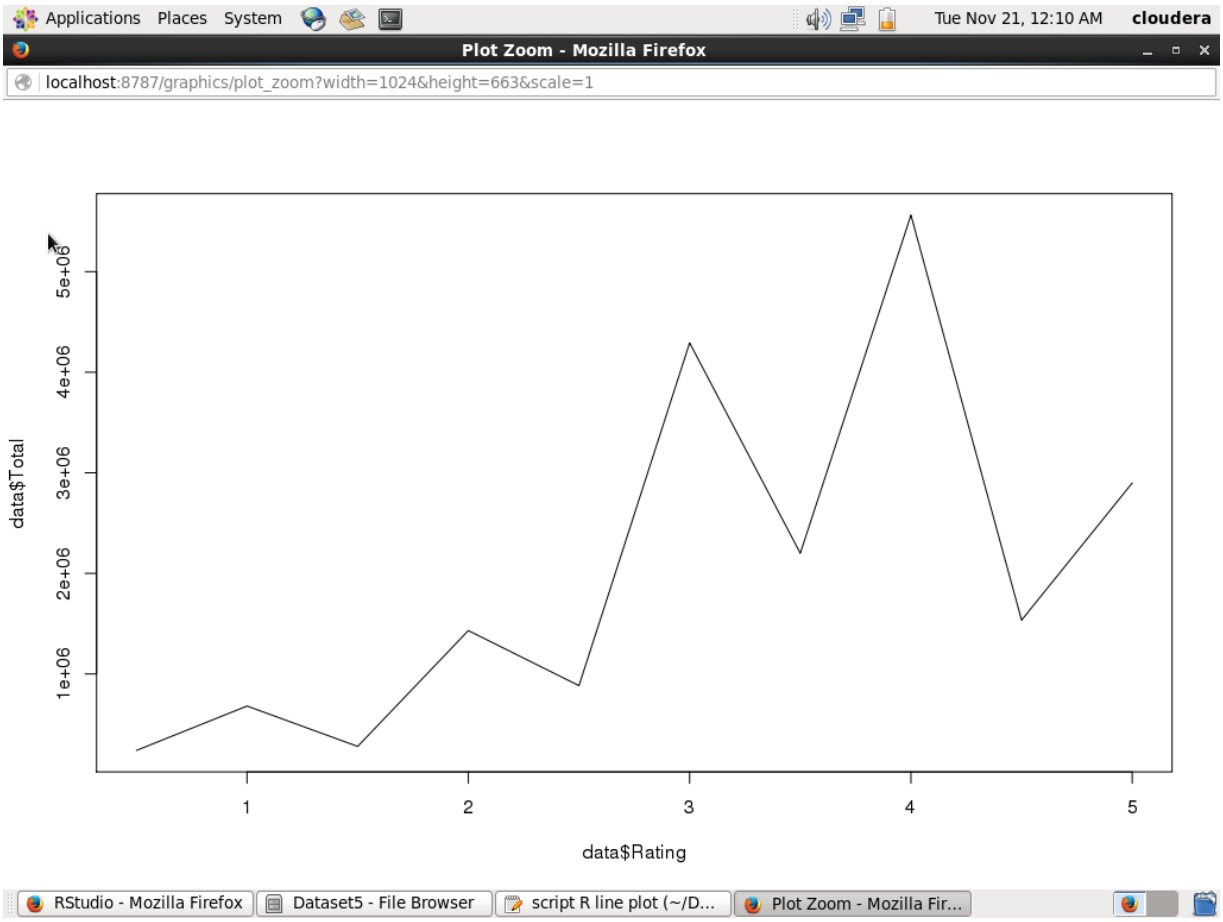


Figura 4.61 Gráfica de líneas para el conjunto de datos 4.

Ahora es el turno de implementar la gráfica de barras para esta misma información. El *script* de R toma de igual forma los datos de HDFS y los transforma en un objeto de tipo matriz para visualizarlos con el comando correspondiente. La Figura 4.62 se muestra la gráfica resultante. De esta gráfica se observa de una manera más clara la información, donde se muestra que hay muy pocas calificaciones de 0.5 y 1.5, dando a entender que hay un mayor número de aceptación en las películas evaluadas.

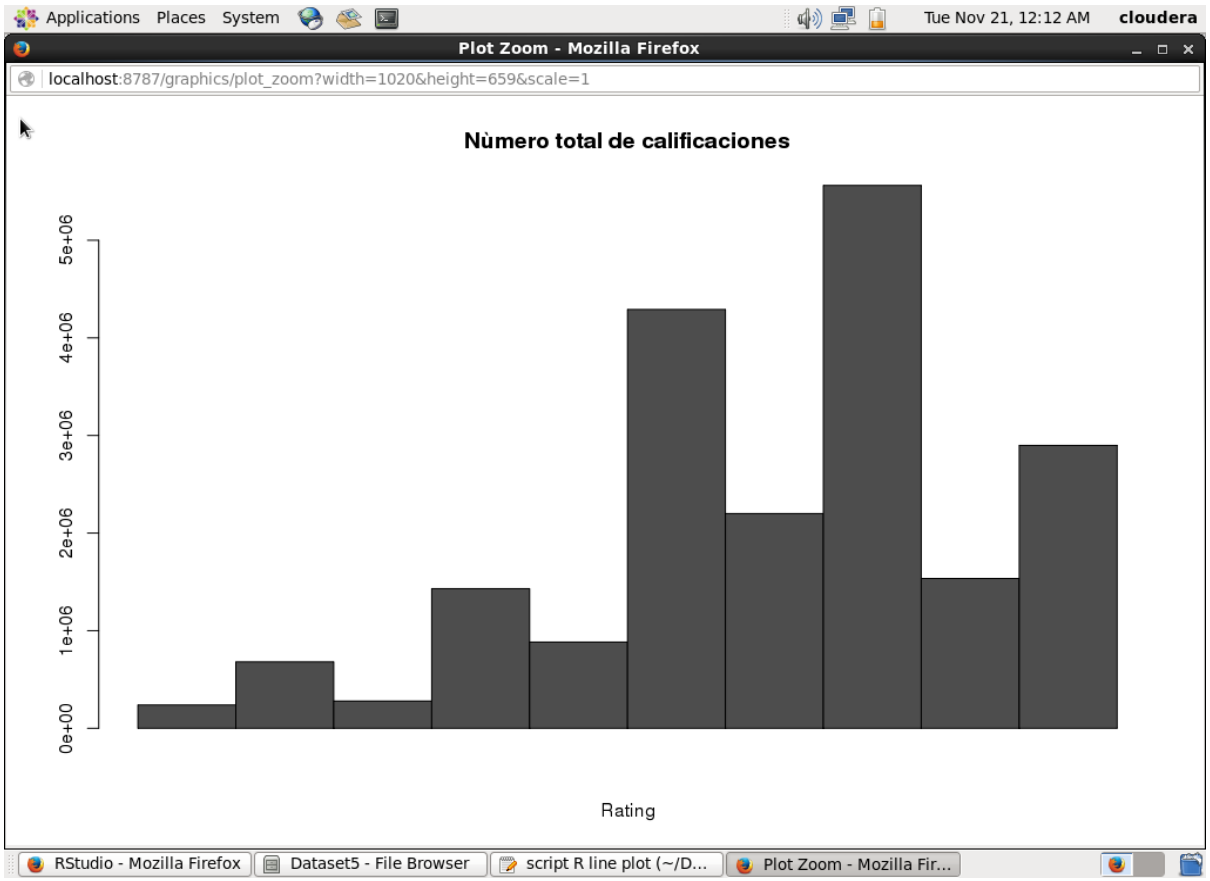


Figura 4.62 Gráfica de barras para el conjunto de datos 4.

Finalmente con el uso de esta misma información, se procede a implementar una gráfica de dispersión. Esta gráfica se realiza con el *script* de R. La gráfica como tal se observa en la Figura 4.63. Como sucedió en el conjunto de datos anterior, al ser datos muy resumidos, la gráfica resultante suele ser una gráfica muy sencilla, que es fácil de entender y analizar.

Es importante recordar que el visualizar la misma información en diferentes técnicas de visualización, es con el fin de observar el nivel de interpretación en los datos, deduciendo si es factible el uso de una gráfica de líneas, una gráfica de barras o una gráfica de dispersión.

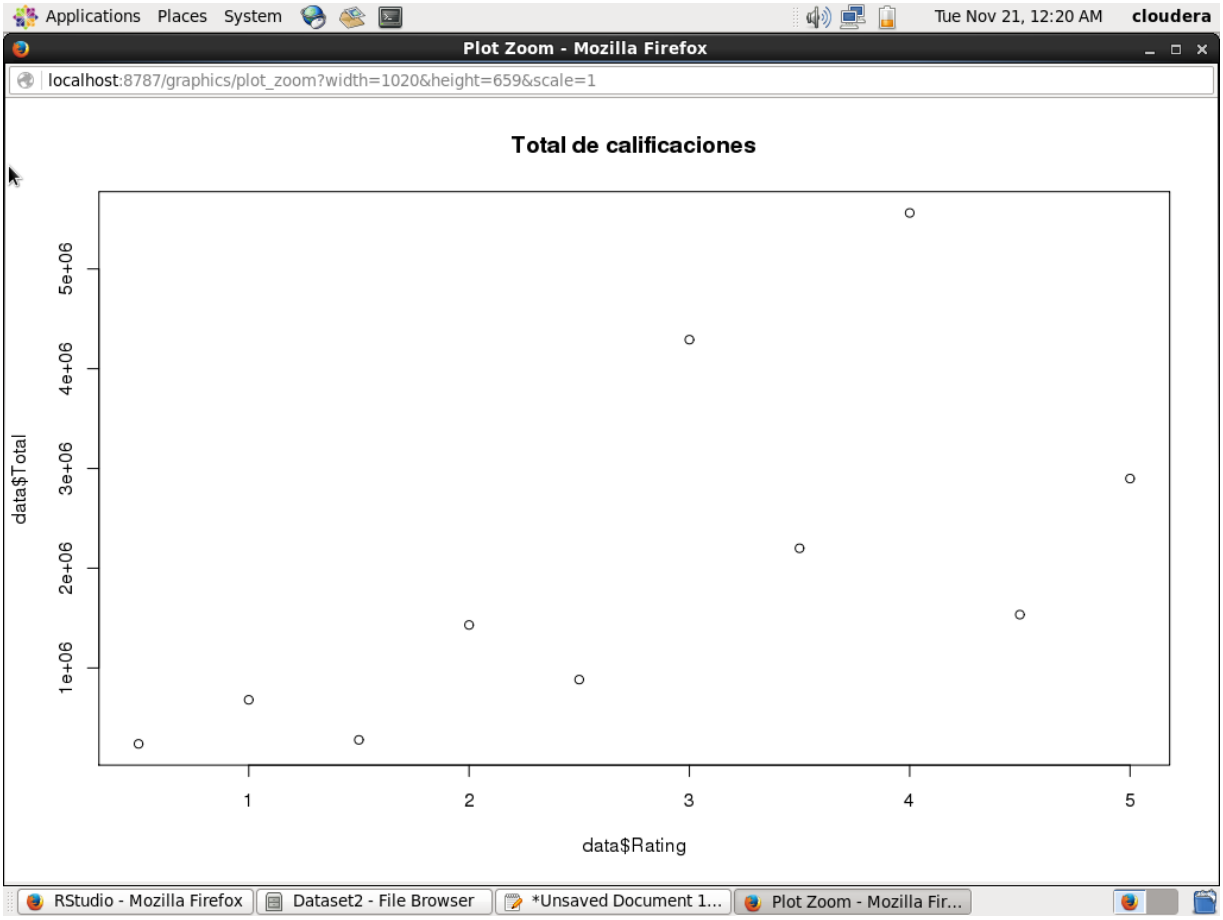


Figura 4.63 Gráfica de dispersión para el conjunto de datos 4.

Para seguir con la implementación de la técnica *Heatmap*, es necesario cambiar la presentación de la información. Es por ello que se requiere separar la información de acuerdo a las calificaciones que dio cada usuario, conocer el número total de la calificación 0.5, 1, 1.5, 2, entre otros, por cada usuario. El *script* que realiza esto se muestra en la Figura 4.64, mientras que en la Figura 4.65 se muestra al archivo resultante almacenado en HDFS.

Ya con la información separada, ahora es turno de implementar la técnica *Heatmap*. Para esto se genera un *script* en R con el objetivo de extraer la información de HDFS y ejecutar el comando correspondiente para la creación de la gráfica.

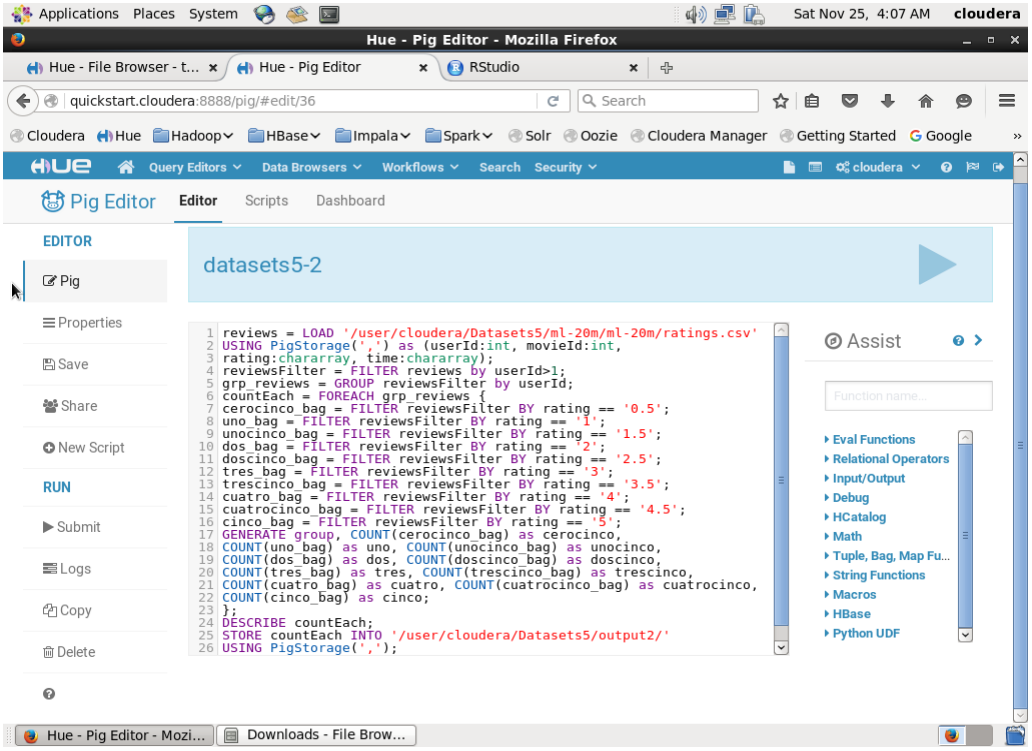


Figura 4.64 *Script* 2 de Apache Pig para el conjunto de datos 4.

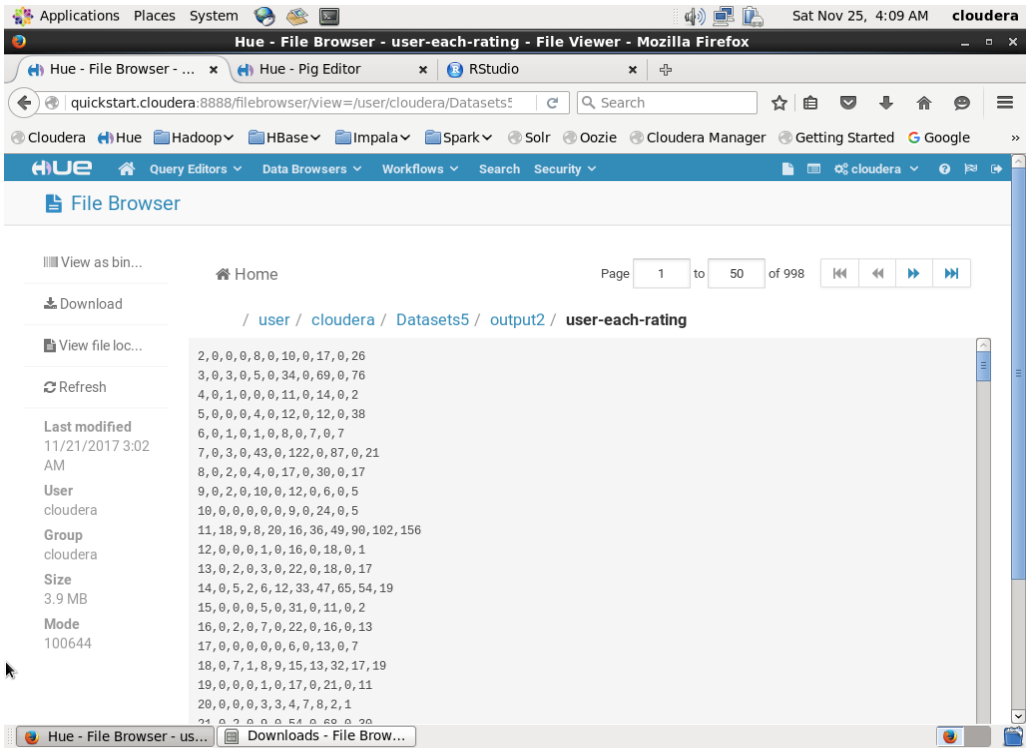


Figura 4.65 Archivo resultante del *script* 2 para el conjunto de datos 4.

Al ejecutar el *script* de R, se genera la gráfica de la Figura 4.66. De esta gráfica se observa que los colores naranjas y rojizos indican valores menores a los que son representados con colores claros y amarillos, con lo que se deduce que la mayoría de las calificaciones se concentran en valores de tres y cuatro. También se muestra que al ser demasiados registros de usuarios, es difícil deducir la clave del usuario, por lo que en la Figura 4.67 solo se representan los primeros 50 registros, lo que hace que cambien las tonalidades de color. Para esta gráfica, ahora solo se aprecian colores rojizos para los valores menores y se siguen manteniendo los colores claros y amarillos para valores mayores.

Figura 4.66 Gráfica *Heatmap* para el conjunto de datos 4.

Figura 4.67 Gráfica *Heatmap* con 50 registros para el conjunto de datos 4.

Nuevamente se necesita realizar cambios en el conjunto de datos original para implementar la técnica de *Treemap*. Para este caso se desea conocer cuántas calificaciones ha dado cada usuario, lo cual se logra con el *script* que se muestra en la Figura 4.68, en donde se agrupan los datos originales por medio de la clave de usuario, para posteriormente contabilizar cada calificación de dicho usuario y almacenar el resultado en un archivo HDFS. Este archivo se presenta en la Figura 4.69.

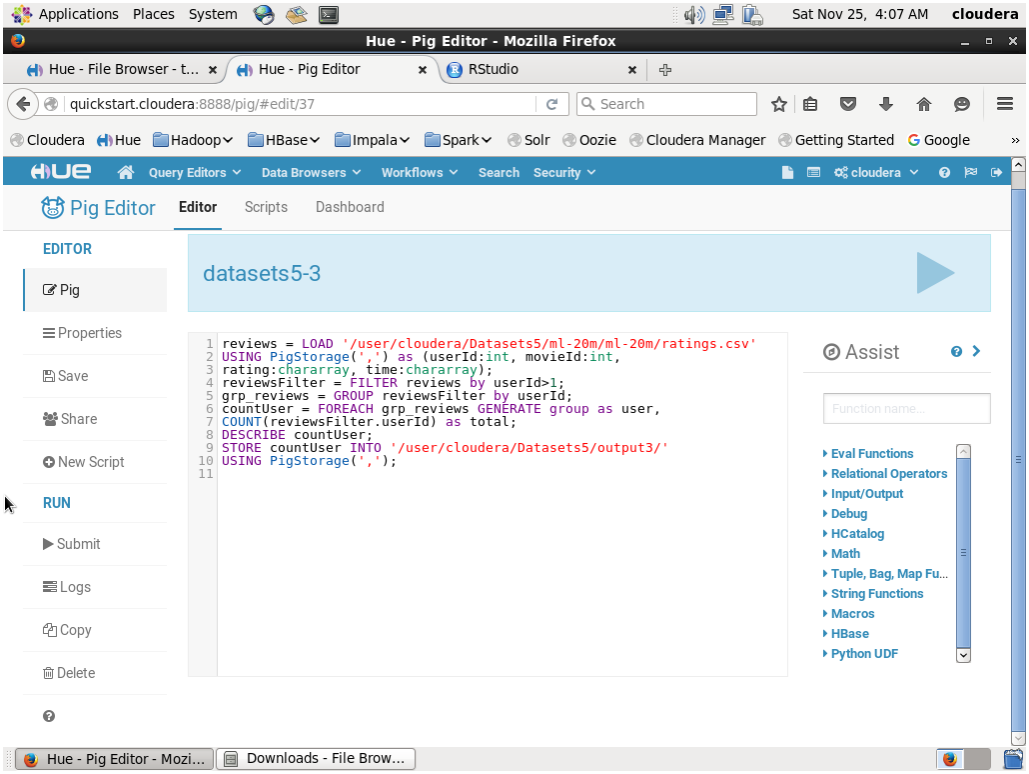


Figura 4.68 *Script* 3 de Apache Pig para el conjunto de datos 4.

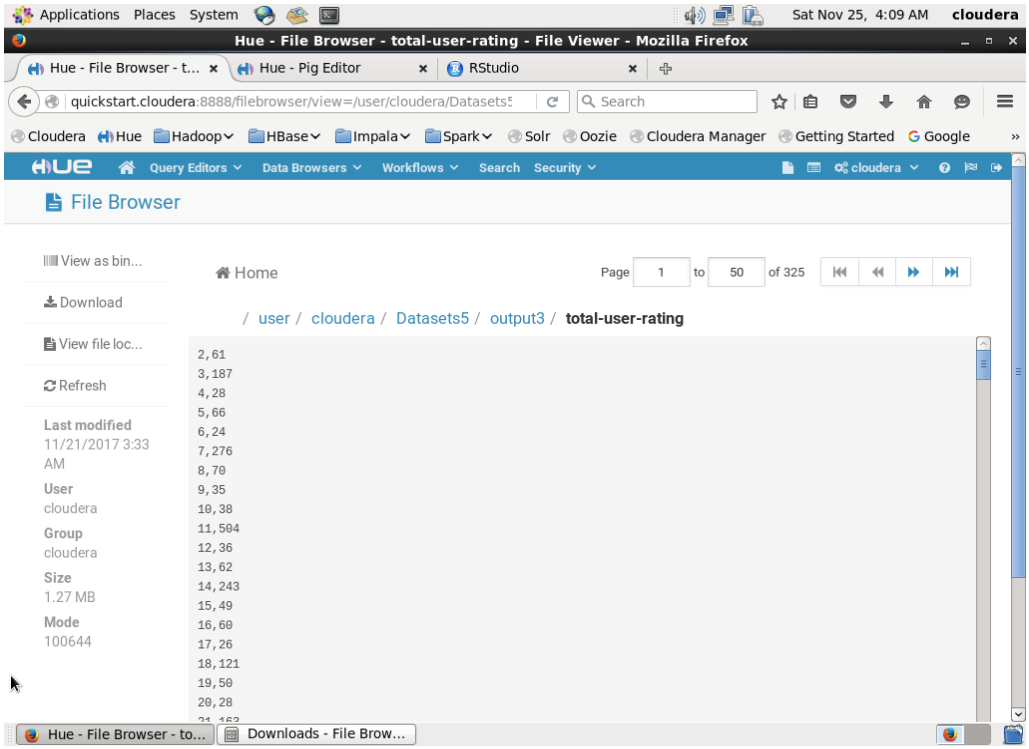


Figura 4.69 Archivo resultante del *script* 3 para el conjunto de datos 4.

Con el archivo resultante, se procede a generar el *script* en R para la técnica *Treemap*, tomando los datos de HDFS y aplicando el comando correspondiente. Para este caso no se logra visualizar todo el conjunto de datos, sin embargo y de la misma forma que en conjuntos anteriores, se procede a visualizar los primeros 1,000 registros en la Figura 4.70, de donde se aprecia que los usuarios 903, 741, 982 y 156 son los que han registrado más calificaciones en el sitio Web de MovieLens.

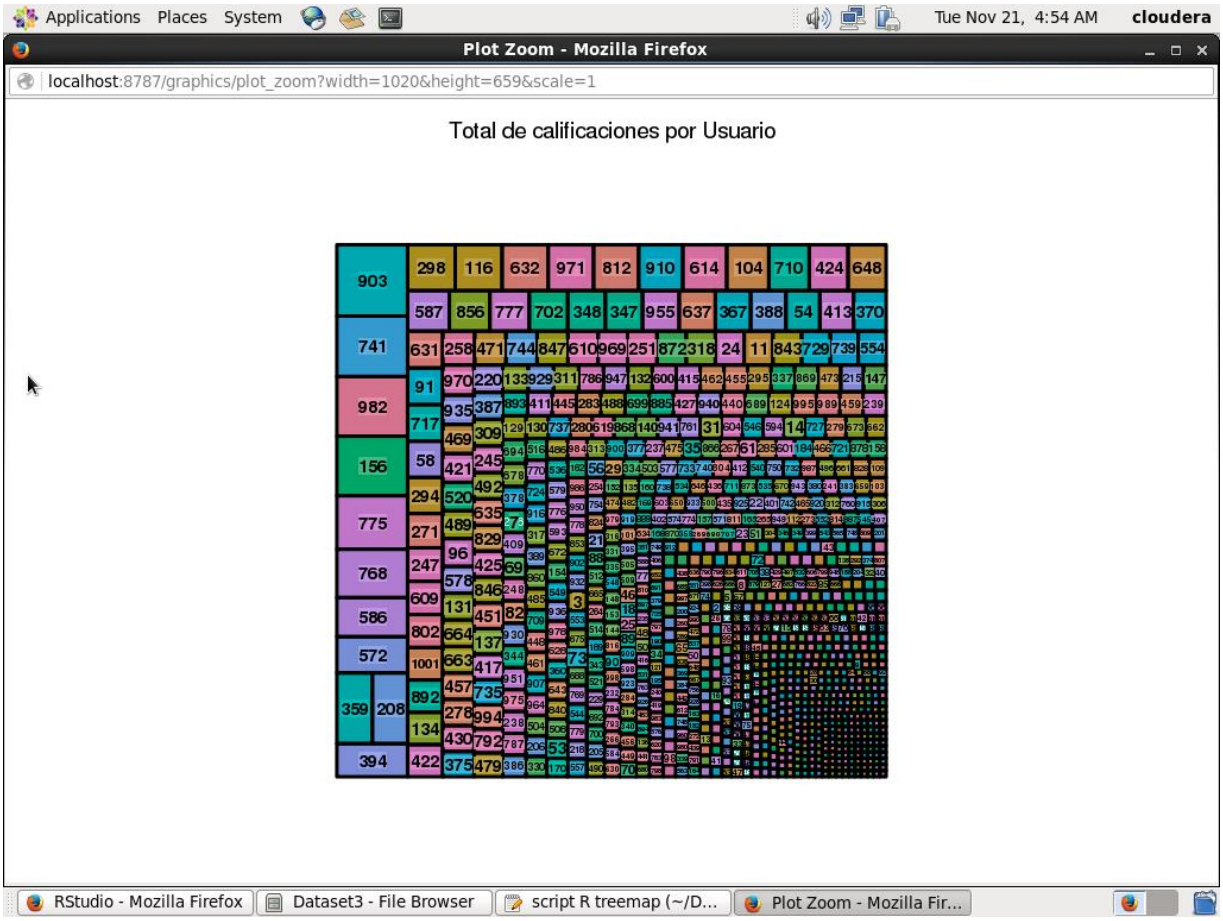


Figura 4.70 Gráfica *Treemap* para el conjunto de datos 4.

Es turno de implementar la técnica de visualización Clustergram usando el mismo conjunto de datos de la técnica de *Treemap*. Con la técnica de Clustergram se conocerá como se comportarían los datos al momento de aplicar el algoritmo *k-means* para generar de dos a ocho grupos. Se genera el *script* de R para implementar Clustergram y dando como resultado la gráfica de la Figura 4.71. En esta gráfica se logra apreciar la enorme cantidad de datos, por lo

que se dificulta mucho la lectura de los mismos, aunque se alcanza a observar algunos intercambios de datos entre los diferentes grupos.

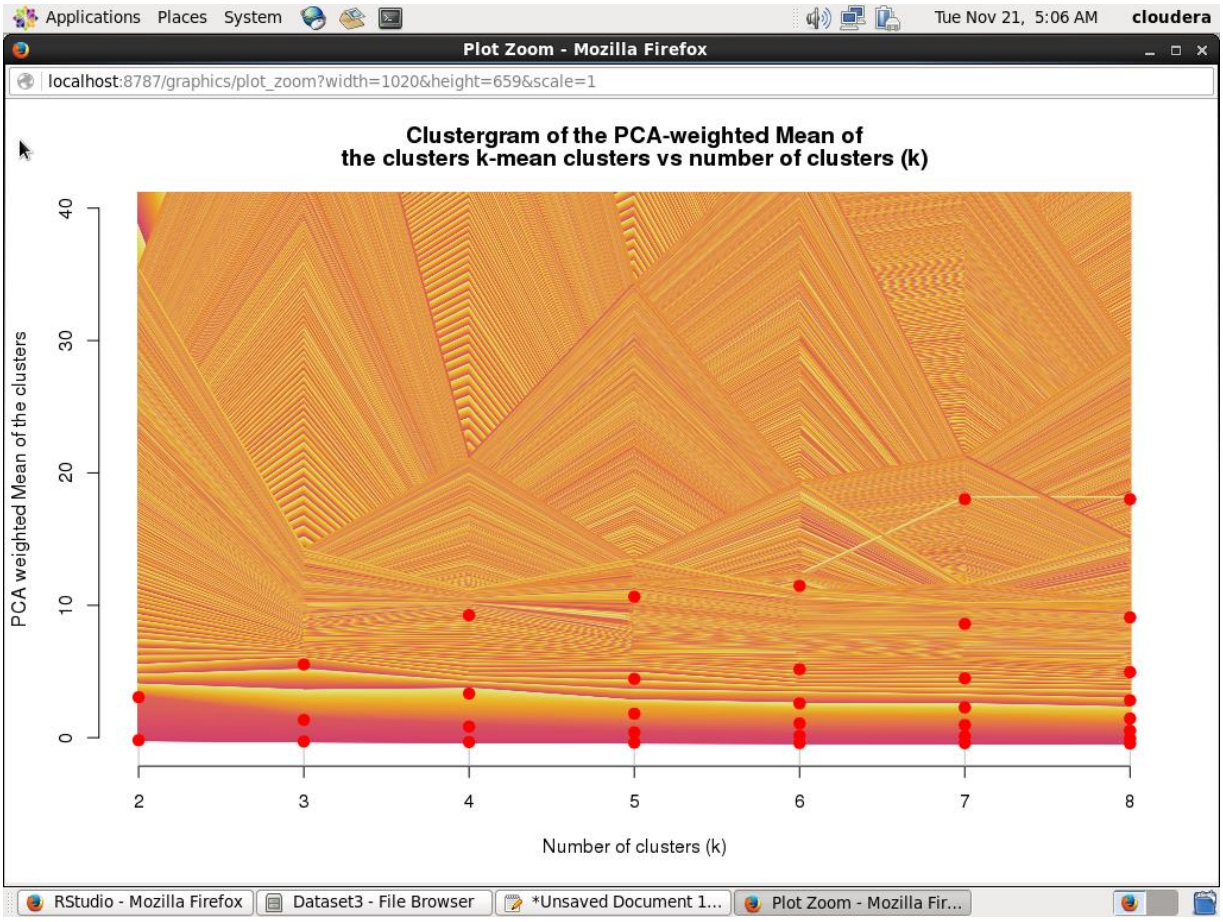


Figura 4.71 Gráfica Clustergram para el conjunto de datos 4.

Finalmente toca el turno de implementar la técnica de visualización de coordenadas paralelas, para este caso se empleará el conjunto de datos que utilizó la técnica de *Heatmap*. Se crea el *script* en R para lograr este objetivo, tomando los datos de HDFS, de los cuales se toma la información de los usuarios 2, 3149 y 9872 para compararlos en esta técnica. La Figura 4.72 muestra la gráfica de coordenadas paralelas de estos tres usuarios. También se graficó el total de registros del conjunto de datos, como se aprecia en la Figura 4.73, de la cual se observa que la mayoría de los usuarios tiene un promedio de 1,000 calificaciones y solo pocos sobrepasan dicho número para sobresalir en la gráfica.

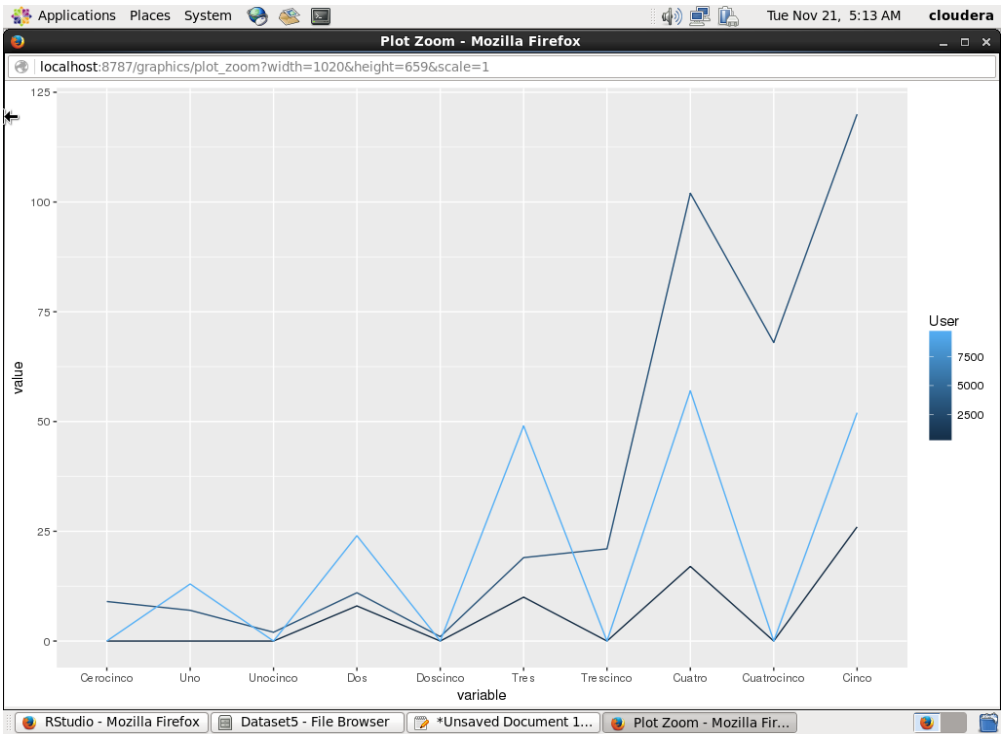


Figura 4.72 Gráfica de coordenadas paralelas con tres registros para el conjunto de datos 4.

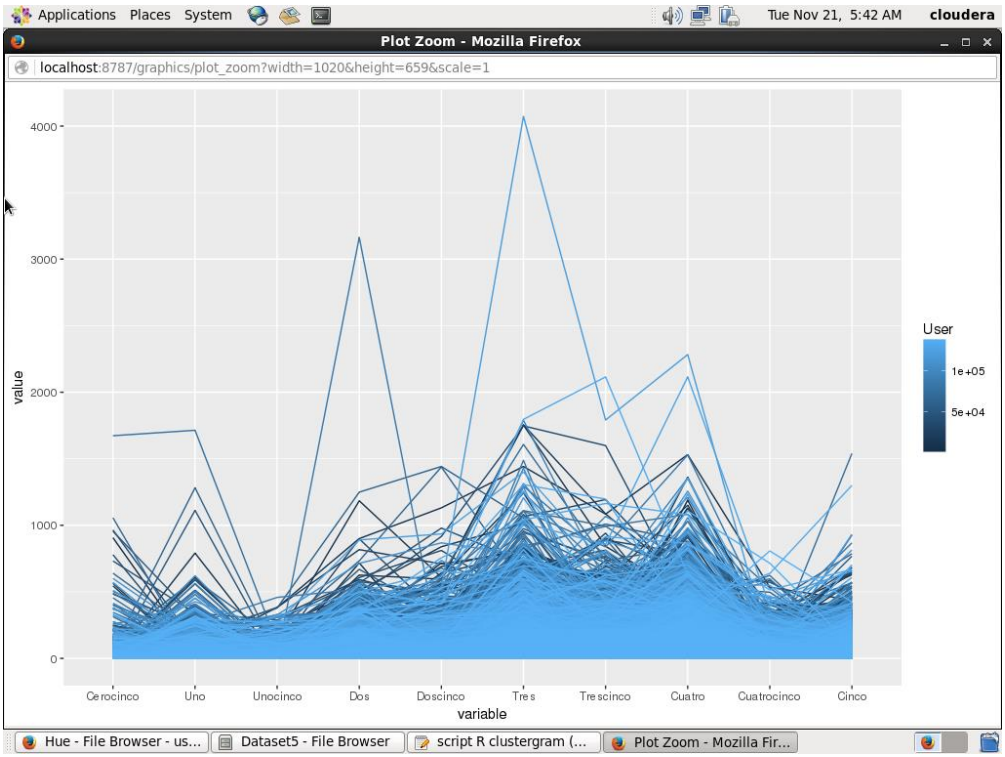


Figura 4.73 Gráfica de coordenadas paralelas para el conjunto de datos 4.

En este conjunto de datos no se aplicó la técnica de *Spatial Information Flow*, ya que por las políticas de privacidad de la página Web de MovieLens, mantiene los datos de sus usuarios de manera anónima, solo proporcionando el código de cada usuario, los demás datos no están incluidos en el conjunto de datos que proporciona la página Web.

4.5. Análisis comparativo de las técnicas de visualización

La Tabla 4.1 muestra el análisis comparativo de las técnicas de visualización implementadas al primer conjunto de datos.

Dataset 1 - Exportaciones e importaciones	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Número de variables procesadas	7	7	7	7	7	7	7	7
Número de tuplas procesadas	5,108,305	5,108,305	5,108,305	5,108,305	5,108,305	5,108,305	5,108,305	5,108,305
Número de variables visualizadas	2	2	2	3	3	3	4	5
Número de tuplas visualizadas	53	53	53	149 / 50	250	250	106	229
Curva de aprendizaje	Fácil	Fácil	Medio	Difícil	Difícil	Difícil	Medio	Difícil
Tiempo de construcción	60 mins.	50 mins.	60 mins.	90 mins.	90 mins.	120 mins.	120 mins.	180 mins.
Tiempo de visualización	0.75 seg.	0.74 seg.	0.75 seg.	0.74 seg.	12.37 seg.	0.97 seg.	1.64 seg.	2.23 seg.
Facilidad de interpretación	Fácil	Fácil	Fácil	Medio	Fácil	Medio	Medio	Fácil
Volumen	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Velocidad	OK	OK	OK	OK	OK	OK	OK	OK
Variedad	--	--	--	--	--	--	--	--
Veracidad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok

Tabla 4.1 Análisis comparativo de técnicas de visualización del primer conjunto de datos.

La Tabla 4.2 muestra el análisis comparativo de las técnicas de visualización implementadas al segundo conjunto de datos.

Dataset 2 - Flujo de clics	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Número de variables procesadas	6	6	6	6	6	6	6	24
Número de tuplas procesadas	1,692,441	1,692,441	1,692,441	1,692,441	1,692,441	1,692,441	1,692,441	1,741,941

Dataset 2 - Flujo de clics	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Número de variables visualizadas	2	2	2	4	2	2	4	4
Número de tuplas visualizadas	43,600 / 100	43,600 / 30	43,600 / 100	43,600 / 50	1,000	43,600	43,600 / 3	185
Curva de aprendizaje	Fácil	Fácil	Fácil	Medio	Medio	Medio	Medio	Difícil
Tiempo de construcción	90 mins.	90 mins.	90 mins.	120 mins.	120 mins.	180 mins.	120 mins.	180 mins.
Tiempo de visualización	0.98 seg.	10.55 seg.	11.87 seg.	14.57 seg.	16.23 seg.	11.19 seg.	51.71 seg.	1.94 seg.
Facilidad de interpretación	Medio	Medio	Medio	Medio	Medio	Medio	Medio	Fácil
Volumen	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Velocidad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Variedad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Veracidad	--	--	--	--	--	--	--	--

Tabla 4.2 Análisis comparativo de técnicas de visualización del segundo conjunto de datos.

La Tabla 4.3 muestra el análisis comparativo de las técnicas de visualización implementadas al tercer conjunto de datos.

Dataset 3 - Calificación de productos	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Número de variables procesadas	8	8	8	8	8	8	8	26
Número de tuplas procesadas	89,991	89,991	89,991	89,991	89,991	89,991	89,991	139,491
Número de variables visualizadas	2	2	2	6	2	2	6	4
Número de tuplas visualizadas	5	5	5	39,376 / 50	1,000	39,376	39,376 / 3	185
Curva de aprendizaje	Fácil	Fácil	Fácil	Medio	Difícil	Medio	Medio	Medio
Tiempo de construcción	60 mins.	60 mins.	60 mins.	120 mins.	120 mins.	120 mins.	120 mins.	180 mins.

Dataset 3 - Calificación de productos	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Tiempo de visualización	0.52 seg.	0.51 seg.	0.51 seg.	14.63 seg.	22.44 seg.	8.61 seg.	46.19 seg.	1.92 seg.
Facilidad de interpretación	Fácil	Fácil	Fácil	Medio	Medio	Difícil	Medio	Fácil
Volumen	--	--	--	--	--	--	--	--
Velocidad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Variedad	OK	OK	OK	OK	OK	OK	OK	OK
Veracidad	--	--	--	--	--	--	--	--

Tabla 4.3 Análisis comparativo de técnicas de visualización del tercer conjunto de datos.

La Tabla 4.4 muestra el análisis comparativo de las técnicas de visualización implementadas al cuarto conjunto de datos.

Dataset 4 - Calificación de películas	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Número de variables procesadas	4	4	4	4	4	4	4	NA
Número de tuplas procesadas	20,000,264	20,000,264	20,000,264	20,000,264	20,000,264	20,000, 264	20,000,264	NA
Número de variables visualizadas	2	2	2	11	2	2	11	NA
Número de tuplas visualizadas	10	10	10	138,492 / 50	1,000	138,492	138,492 / 3	NA
Curva de aprendizaje	Fácil	Fácil	Fácil	Medio	Medio	Medio	Medio	NA
Tiempo de construcción	80 mins.	80 mins.	80 mins.	140 mins.	140 mins.	150 mins.	160 mins.	NA
Tiempo de visualización	0.69 seg.	0.6 seg.	0.62 seg.	66.67 seg.	27 seg.	44.08 seg.	186.77 seg.	NA
Facilidad de interpretación	Fácil	Fácil	Fácil	Medio	Medio	Difícil	Difícil	NA
Volumen	Ok	Ok	Ok	Ok	Ok	Ok	Ok	NA
Velocidad	--	--	--	--	--	--	--	NA
Variedad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	NA

Dataset 4 - Calificación de películas	Gráfica de líneas	Gráfica de barras	Gráfica de dispersión	*Heatmap* / Mapa de calor	*Treemap* / Mapa de árbol	Cluster gram	Coordenadas paralelas	*Spatial Information Flow*
Veracidad	Ok	Ok	Ok	Ok	Ok	Ok	Ok	NA

Tabla 4.4 Análisis comparativo de técnicas de visualización del cuarto conjunto de datos.

En las tablas anteriores se recapituló la información de todas las técnicas de visualización implementadas en los cuatro conjuntos de datos, de acuerdo a los puntos propuesto en el marco de evaluación.

Con estas tablas se observa que las técnicas de visualización de proyección geométrica (gráfica de líneas, de barras y de dispersión) son las que llevan menor tiempo de construcción y son fáciles de interpretar siempre y cuando no haya un gran número de datos a graficar.

Por otra parte, las técnicas basadas en jerarquías y grafos (*Heatmap* y *Treemap*) son excelentes para observar cómo se comportan las variables dentro del mismo conjunto de datos y con *Treemap* se aprecia el crecimiento de los valores para cada registro, como lo fue el caso de los conjuntos de datos de las exportaciones o el de calificaciones de películas.

En cuestión con las técnicas de relaciones y datos complejos, suelen ser las más complicadas de realizar y las que llevan mayor tiempo en visualizar los datos, sin embargo, resultan muy provechosas si se quiere conocer diferentes tipos de información, por ejemplo, para descubrir información relacionada con lugares geográficos, la implementación de *Spatial Information Flow* resulta muy efectiva. Para descubrir el comportamiento de los datos al momento de aplicar algoritmos como *k-means*, Clustergram resulta ser una excelente opción. También si se desea comparar variables dentro del mismo conjunto de datos y se buscan alternativas para la técnica *Heatmap*, coordenadas paralelas resulta ser una opción viable.

Capítulo 5. Conclusiones y recomendaciones

5.1. Conclusiones

Big Data se popularizó en los últimos 20 años debido a la creciente explosión en la generación y almacenamiento de los datos, es por esto que requiere analizarse para extraer información útil que le sirva a empresas y a investigadores en la toma de decisiones. El análisis visual de *Big Data* es un campo nuevo que representa una nueva forma de entender los datos y que ha emergido como una herramienta para descubrir conocimiento, pero que al mismo tiempo representa una tarea desafiante.

El propósito de esta investigación fue realizar un análisis comparativo de técnicas de visualización para *Big Data*, dada la poca literatura que existe al respecto de este tema, sin embargo, para llegar a cumplir este objetivo, primero se requirió analizar las técnicas de visualización que más se ocupan en el análisis de *Big Data*, así como también de los principales conjuntos de datos. Esto dio como resultado que las técnicas más utilizadas son las gráficas de líneas, de barras, de dispersión, mapas de calor, mapas de árbol, clustergram, coordenadas paralelas y *Spatial Information Flow*.

Fue de vital importancia implementar estas técnicas de visualización en conjuntos de datos *Big Data* que tuvieran la combinación de cuatro características, las cuáles son: volumen, velocidad, variedad y veracidad. Estos conjuntos de datos se obtuvieron gracias al análisis previo y además de la utilización de *benchmark*s para *Big Data*. Estos *benchmark*s como BigBench o BigDataBench garantizan que los datos generados cumplan con al menos dos de estas características, asegurando la integridad de las pruebas realizadas en esta investigación.

Es importante mencionar que el uso de la plataforma Cloudera, facilitó trabajar con Hadoop/MapReduce, ya que dicha plataforma ofrece al usuario el ecosistema perfecto para este tipo de trabajos, proporcionando también herramientas como Apache Pig que ayudan en el procesamiento de los datos que se encuentran almacenados en el sistema de archivos de

Hadoop, comúnmente llamado HDFS, transformando estos *script* en trabajos MapReduce de forma transparente para el programador.

La integración que se logra entre R y Hadoop, hizo que la implementación de las visualizaciones no fuera tan complicada, ya que R fue capaz de acceder a los datos que se encontraban en HDFS para posteriormente visualizarlos con los comandos apropiados, los cuales al principio resultaron confusos, pero con el paso del tiempo resultan ser sencillos de manejar.

Una vez implementadas todas las técnicas de evaluación a los cuatro conjuntos de datos con los que se trabajó, se logró realizar el análisis comparativo con el marco de evaluación propuesto. Con este análisis se concluye que las gráficas de líneas, de barras y de dispersión, son perfectas cuando se quiere presentar datos ya resumidos, como fue el caso de los conjunto de datos de calificaciones de productos y calificaciones de películas. Por otra parte, técnicas como *Heatmap* y coordenadas paralelas son perfectas para cuando se desea comparar variables dentro del mismo conjunto de datos, el primero con una mayor facilidad de interpretación si no son demasiados datos. *Treemap* es ideal para conocer de manera gráfica los máximos valores dentro del conjunto de datos. Clustergram es la mejor opción si se desea averiguar el comportamiento de los datos cuando se les aplica el algoritmo de agrupamiento *k-means*. Finalmente, *Spatial Information Flow* resulta ser de las mejores opciones si se cuentan con los datos necesarios para su visualización en un mapa geográfico, como sucedió en el conjunto de datos de exportaciones e importaciones de los países.

Por último, se concluye que la información presentada en esta investigación, el análisis comparativo de las técnicas de visualización para *Big Data*, dependerá del investigador que esté consultando la información. Para algunos será mejor aplicar una sencilla gráfica de barras o para otros será mejor visualizar los datos en un mapa de calor, todo dependerá del tipo y cantidad de información que se esté procesando y analizando para resumir grandes datos.

5.2. Recomendaciones

El objetivo general de esta investigación se cubrió en su totalidad, sin embargo, queda como recomendación para trabajos futuros, el profundizar más en los temas de tecnologías de *Big Data*, así como la realización de pruebas con otro tipo de conjuntos de datos.

Día a día surgen nuevas tecnologías para el manejo de *Big Data*, sería interesante implementar las técnicas de visualización en estas nuevas herramientas, como sería Apache Spark para el procesamiento de datos o D3.js y Gephi para la visualización de los datos.

También sería interesante probar nuevos tipos de conjunto de datos, por ejemplo con datos de tipo imagen, videos, audios, archivos XML, entre otros. Descubrir la complejidad que llevaría procesar, analizar y visualizar estos tipos de datos.

Dada la importancia de este nuevo tema de *Big Data*, es recomendable seguir en esta línea de investigación para aportar la mayor cantidad de información posible y así facilitar el análisis visual de grandes datos.

Productos académicos

Conferencias internacionales

Luis Eder Velázquez Peña, Lisbeth Rodríguez Mazahua, Giner Alor Hernández, Beatriz Alejandra Olivares Zepahua, Silvestre Gustavo Peláez Camarena.
"Revisión de técnicas de visualización para Big Data". XXX Congreso Nacional y XVI Congreso Internacional de Informática y Computación ANIEI 2017 (CNCIIC-ANIEI 2017).
Estado: Presentado.

Luis Eder Velázquez Peña, Lisbeth Rodríguez Mazahua, Giner Alor Hernández, Beatriz Alejandra Olivares Zepahua, Silvestre Gustavo Peláez Camarena, Isaac Machorro Cano.
"Visualización de Big Data: Revisión de técnicas y conjuntos de datos".
Applications in Software Engineering. Proceedings of the 6th International Conference on Software Process Improvement (CIMPS 2017), IEEE Press, ISBN: 978-1-5386-3230-7, pp. 91-99.
Estado: Publicado.

Proyectos de investigación

Análisis comparativo de técnicas de visualización para Big Data. Convocatoria de Apoyo a la Incorporación de nuevos Profesores de Tiempo Completo 2016, Programa para el Desarrollo Profesional Docente (PRODEP), tipo superior. Estado: Aprobado.

Referencias

[1] ASALE y RAE, “Diccionario de la lengua española - Edición del Tricentenario”, *Diccionario de la lengua española*. [En línea]. Disponible en: http://dle.rae.es/. [Consultado: 01-oct-2016].

[2] David J. Hand, Heikki Mannila, y Padhraic Smyth, *Principles of Data Mining*. The MIT Press, 2001.

[3] Ian H. Witten, Eibe Frank, y Mark A. Hall, *Data Mining - Practical Machine Learning Tools and Techniques*, Third Edition. Elsevier, 2011.

[4] James Manyika *et al.*, Big Data*: the Next Frontier for Innovation, Competition and Productivity*. McKinsey & Company, 2011.

[5] Min Chen, Shiwen Mao, y Yunhao Liu, “Big Data: A Survey”, *Mobile Netw Appl*, vol. 19, pp. 171–209, ene. 2014.

[6] Lisbeth Rodríguez Mazahua, Cristian Aarón Rodríguez Enríquez, José Luis Sánchez Cervantes, Jair Cervantes, Jorge Luis García Alcaraz, y Giner Alor Hernández, “A general perspective of Big Data: applications, tools, challenges and trends”, *The Journal of Supercomputing*, vol. 72, núm. 8, pp. 3073–3113, ago. 2015.

[7] Jinson Zhang, Mao Lin Huang, y Zhao-Peng Meng, “Visual Analytics for BigData Variety and Its Behaviours”, *Computer Science and Information Systems*, vol. 12, núm. 4, pp. 1171–1191, Diciembre 2015.

[8] “The 7 V’s of Big Data”, *Impact Radius*, 07-abr-2016. [En línea]. Disponible en: https://www.impactradius.com/blog/7-vs-big-data/. [Consultado: 06-abr-2017].

[9] E. McNulty, “Understanding Big Data: The Seven V’s”, *Dataconomy*, 22-may-2014. .

[10] “Las 7 V del Big data: Características más importantes”, *IIC*, 28-jun-2016. .

[11] M. Ali-ud-din Khan, Muhammad Fahim Uddin, y Navarun Gupta, “Seven V’s of Big Data understanding Big Data to extract value”, presentado en Conference of the American Society for Engineering Education, 2014.

[12] Aziz Nasridinov y Young-Ho Park, “Visual Analytics for Big Data using R”, presentado en 2013 IEEE Third International Conference on Cloud and Green Computing, 2013, pp. 564–565.

[13] Zeqian Shen, Jishang Wei, Neel Sundaresan, y Kwan-Liu Ma, “Visual Analysis of Massive Web Session Data”, presentado en IEEE Symposium on Large Data Analysis and Visualization 2012, Seattle, WA, USA, 2012, pp. 65–72.

[14] Jiawei Han, Micheline Kamber, y Jian Pei, *Data Mining: Concepts and Techniques*, Third Edition. Morgan Kaufmann, 2012.

[15] Joachim D. Pleil, Matthew A. Stiegel, Michael C. Madden, y Jon R. Sobus, "Heat map visualization of complex environmental and biomarker measurements", *Chemosphere*, vol. 84, pp. 716–723, Abril 2011.

[16] Evgeniy Yur'evich Gorodov y Vasiliy Vasil'evich Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data", *Journal of Electrical and Computer Engineering*, vol. 2013, ene. 2013.

[17] Lidong Wang, Guanghui Wang, y Cheryl Ann Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress", *Digital Technologies*, vol. 1, núm. 1, pp. 33–38, 2015.

[18] Leishi Zhang *et al.*, "Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems", presentado en IEEE Conference on Visual Analytics Science & Technology 2012, Seattle, Washington, USA, 2012, pp. 173–182.

[19] Cheng-Long Ma, Xu-Feng Shang, y Yu-Bo Yuan, "A Three-Dimensional Display for Big Data Sets", presentado en 2012 International Conference on Machine Learning and Cybernetics, Xian, 2012, pp. 1541–1545.

[20] Shinnosuke Takeda, Aimi Kobayashi, Hiroaki Kobayashi, Saori Okubo, y Kazuo Misue, "Irregular Trend Finder: Visualization tool for analyzing time-series big data", presentado en IEEE Symposium on Visual Analytics Science and Technology 2012, Seattle, WA, USA, 2012, pp. 305–306.

[21] Cheng Daniel, Schretlen Peter, Kronenfeld Nathan, Bozowsky Neil, y Wright William, "Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis", presentado en 2013 IEEE International Conference on Big Data, pp. 2–4.

[22] Ackermann, Klaus y D. Angus, Simon, "A Resource Efficent Big Data Analysis Method for the Social Sciences: the case of global IP activity", presentado en 14th International Conference on Computational Science, pp. 2360–2369.

[23] A. Steed, Chad *et al.*, "Big data visual analytics for exploratory earth system simulation analysis", *Computers & Geosciences*, vol. 61, pp. 71–82.

[24] Fan Bao y Jia Chen, "Visual framework for big data in d3.js", presentado en 2014 IEEE Workshop on Electronics, Computer and Applications, 2014, pp. 47–50.

[25] Cagatay Turkay, Aiden Slingsby, Helwig Hauser, Jo Wood, y Jason Dykes, "Attribute Signatures: Dynamic Visual Summaries for Analyzing Multivariate Geographical Data",

IEEE Transactions on Visualization and Computer Graphics, vol. 20, núm. 12, pp. 2033–2042, 2014.

[26] Johanna Beyer, Ali Al-Awami, Narayanan Kasthuri, Jeff W. Lichtman, Hanspeter Pfister, y Markus Hadwiger, "ConnectomeExplorer: Query-Guided Visual Analysis of Large Volumetric Neuroscience Data", *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, núm. 12, pp. 2868–2877, Diciembre 2013.

[27] Jaegui Choo y Haesun Park, "Customizing Computational Methods for visual Analytics with Big Data", *IEEE Computer Graphics and Applications*, vol. 33, núm. 4, pp. 22–28, jul. 2013.

[28] Johanna Beyer *et al.*, "Exploring the Connectome: Petascale Volume Visualization of Microscopy Data Streams", *IEEE Computer Graphics and Applications*, vol. 33, núm. 4, pp. 50–61, jul. 2013.

[29] Conglei Shi, Yingcai Wu, Shixia Liu, Hong Zhou, y Huamin Qu, "LoyalTracker: Visualizing Loyalty Dynamics in Search Engines", *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, núm. 12, pp. 1733–1742, Diciembre 2014.

[30] Khairi Reda *et al.*, "Visualizing Large, Heterogeneous Data in Hybrid Reality Environments", *IEEE Computer Graphics and Applications*, vol. 33, núm. 4, pp. 38–48, jul. 2013.

[31] "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].

[32] "Welcome to Apache Pig!" [En línea]. Disponible en: https://pig.apache.org/. [Consultado: 24-nov-2017].

[33] Jörn Kohlhammer, Daniel Keim, Margit Pohl, Giuseppe Santucci, y Gennady Andrienko, "Solving Problems with Visual Analytics", *Procedia Computer Science*, vol. 7, pp. 117–120, 2011.

[34] "R: The R Project for Statistical Computing". [En línea]. Disponible en: https://www.r-project.org/. [Consultado: 16-oct-2016].

[35] Gali Halevi y Henk Moed, "The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature", *Research Trends*, vol. 30, pp. 3–6, sep. 2012.

[36] Mingjin Zhang *et al.*, "TerraFly GeoCloud: An Online Spatial Data Analysis and Visualization System", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, núm. 3, p. 24, may 2015.

[37] Song Gao, Linna Li, Wenwen Li, Krzysztof Janowicz, y Yue Zhang, "Constructing gazetteers fromvolunteered Big Geo-Data based on Hadoop", *Computers, Environment and Urban Systems*, vol. 61, pp. 172–186, ene. 2017.

[38] Siqi Wu, Luca Morandini, y Richard O. Sinnott, "SMASH: A Cloud-based Architecture for Big Data Processing and Visualization of Traffic Data", presentado en IEEE International Conference on Data Science and Data Intensive Systems, Sydney, NSW, Australia, 2015, pp. 53–60.

[39] Ari Wibisono, Wisnu Jatmiko, Hanief Arief Wisesa, Benny Hardjono, y Petrus Mursanto, "Traffic big data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD)", *Knowledge-Based Systems*, vol. 93, pp. 33–46, feb. 2016.

[40] Jie Li, Zhao-Peng Meng, Mao-Lin Huang, y Kang Zhang, "An interactive visualization approach to the overview of geoscience data", *Journal of Visualization*, vol. 20, núm. 3, pp. 433–451, 2016.

[41] Amit Kumar Bhardwaj y Maninder Singh, "Data mining-based integrated network traffic visualization framework for threat detection", *Neural Computing and Applications*, vol. 26, núm. 1, pp. 117–130, 2014.

[42] Anastasios Drosou, Ilias Kalamaras, Stavros Papadopoulos, y Dimitrios Tzovaras, "An enhanced Graph Analytics Platform (GAP) providing insight in Big Network Data", *Journal of Innovation in Digital Ecosystems*, vol. 3, núm. 2, pp. 83–97, dic. 2016.

[43] Aqsa Hameed, Saqib Ali, Rodger Les Cottrell, y Bebo White, "Applying Big Data Warehousing and Visualization Techniques on PingER Data", presentado en 3rd International Conference on Big Data Computing, Applications and Technologies, Shanghai, China, 2016, pp. 67–72.

[44] Jinson Zhang, Wen Bo Wang, Mao Lin Huang, Liang Fu Lu, y Zhao-Peng Meng, "Big Data Density Analytics using Parallel Coordinate Visualization", presentado en 17th International Conference on Computational Science and Engineering, Chengdu, China, 2014, pp. 1115–1120.

[45] Yalong Yang, Kang Zhang, Jianrong Wang, y Quang Vinh Nguyen, "Cabinet Tree: an orthogonal enclosure approach to visualizing and exploring big data", *Journal of Big Data*, dic. 2015.

[46] Huiqi Xu, Zhen Li, Shumin Guo, y Keke Chen, "CloudVista: Interactive and Economical Visual Cluster Analysis for Big Data in the Cloud", *Proceedings of the VLDB Endowment*, vol. 5, núm. 12, pp. 1886–1889, ago. 2012.

[47] Uwe Jugel, Zbigniew Jerzak, Gregor Hackenbroich, y Volker Markl, “Faster Visual Analytics through PixelPerfect Aggregation”, *Proceedings of the VLDB Endowment*, vol. 7, núm. 13, pp. 1705–1708, ago. 2014.

[48] Miguel Molina-Solana, David Birch, y Yi-ke Guo, “Improving data exploration in graphs with fuzzy logic and large-scale visualisation”, *Applied Soft Computing*, vol. 53, pp. 227–235, abr. 2017.

[49] Alexandre Perrot, Romain Bourqui, Nicolas Hanusse, Frederic Lalanne, y David Auber, “Large Interactive Visualization of Density Functions on Big Data Infrastructure”, presentado en Symposium on Large Data Analysis and Visualization, Chicago, Il, USA, 2015.

[50] Carson K. Leung, Vadim V. Kononov, Adam G.M. Pazdor, y Fan Jiang, “PyramidViz: Visual Analytics and Big Data Visualization of Frequent Patterns”, presentado en 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, Auckland, New Zealand, 2016, pp. 913–916.

[51] Rhema Linder y Eunyee Koh, “Quarry: Picking From Examples to Explore Big Data”, presentado en 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, Republic of Korea, 2015, pp. 1869–1874.

[52] Mike Barnett *et al.*, “Stat!: an interactive analytics environment for big data”, presentado en ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 2013, pp. 1013–1016.

[53] Juha K. Laurila *et al.*, “The Mobile Data Challenge: Big Data for Mobile Computing Research”, presentado en Mobile Data Challenge by Nokia Workshop, 2012.

[54] Eduard Glatz, Stelios Mavromatidis, Bernhard Ager, y Xenofontas Dimitropoulos, “Visualizing big network traffic data using frequent pattern mining and hypergraphs”, *Computing*, vol. 96, núm. 1, pp. 27–38, ene. 2014.

[55] Jian Tang, Jingzhou Liu, Ming Zhang, y Qiaozhu Mei, “Visualizing Large-scale and High-dimensional Data”, presentado en 25th International Conference on World Wide Web, Montréal, Québec, Canada, 2016, pp. 287–297.

[56] Jinson Zhang y Mao Lin Huang, “2D Approach Measuring Multidimensional Data Pattern in Big Data Visualization”, presentado en 2016 IEEE International Conference on Big Data Analysis (ICBDA), Hangzhou, China, 2016.

[57] José M. Martínez-Martínez, Pablo Escandell-Montero, Emilio Soria-Olivas, José D. Martín-Guerrero, y Antonio J. Serrano-López, “A new visualization tool for data mining techniques”, *Progress in Artificial Intelligence*, vol. 5, núm. 2, pp. 137–154, may 2016.

[58] Bo Wu y B. M. Wilamowski, “An Algorithm for Visualization of Big Data in a Two-Dimensional Space”, presentado en 41st Annual Conference of the IEEE Industrial Electronics Society, Yokohama, Japan, 2015.

[59] Hari Bhaskar Sankaranarayanan, Gaurav Agarwal, y Viral Rathod, “An Exploratory Data Analysis of Airport Wait times Using Big Data Visualisation Techniques”, presentado en 2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions, Bangalore, India, 2016, pp. 324–329.

[60] Ricardo Kramer de Oliveira Barros y Giuliano Araujo Bertoti, “An Information Visualization Tool for Data Journalism”, presentado en Brazilian Symposium on Human Factors in Computing Systems, Cuiaba, Brazil, 2012, pp. 41–42.

[61] Xiao Luo *et al.*, “Analysis on spatial-temporal features of taxis’ emissions from big data informed travel patterns: a case of Shanghai, China”, *Journal of Cleaner Production*, vol. 142, pp. 926–935, ene. 2017.

[62] Cesar A. Hidalgo, “Big Data Visualization Engines for Understanding the Development of Countries, Social Networks, Culture and Cities”, presentado en 25th ACM conference on Hypertext and social media, Santiago, Chile, 2014, pp. 3–3.

[63] Enrico G. Caldarola, Antonio Picariello, y Antonio M. Rinaldi, “Big Graph-based Data Visualization Experiences: The WordNet Case Study”, presentado en 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 2015.

[64] Alun Evans, Javi Agenjo, y Josep Blat, “Combined 2D and 3D Web-based visualisation of on-set big media data”, presentado en IEEE International Conference on Image Processing, Quebec City, QC, Canada, 2015, pp. 1120–1124.

[65] Patricia Conde-Clemente, Gracian Trivino, y Jose M. Alonso, “Generating automatic linguistic descriptions with big data”, *Information Sciences*, vol. 380, pp. 12–30, feb. 2017.

[66] Yi Gu, Huaiguang Jiang, Yingchen Zhang, Jun Jason Zhang, Tianlu Gao, y Eduard Muljadi, “Knowledge Discovery for Smart Grid Operation, Control, and Situation Awareness – A Big Data Visualization Platform”, presentado en 2016 North American Power Symposium (NAPS), Denver, CO, USA, 2016, pp. 1–6.

[67] Wenwu Tang y Wenpeng Feng, “Parallel map projection of vector-based big spatial data: Coupling cloud computing with graphics processing units”, *Computers, Environment and Urban Systems*, vol. 61, núm. Part B, pp. 187–197, ene. 2017.

[68] Xiaoming Li, Zhihan Lv, Baoyun Zhang, Weixi Wang, Shengzhong Feng, y Jinxing Hu, “WebVRGIS Based City Bigdata 3D Visualization and Analysis”, presentado en Pacific Visualization Symposium, 2015.

[69] Roberto Tardio, Alejandro Mate, y Juan Trujillo, “An Iterative Methodology for Big Data Management, Analysis and Visualization”, presentado en IEEE International Conference on Big Data, 2015, pp. 545–550.

[70] Jeremy W. Crampton *et al.*, “Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoWeb”, *Cartography and Geographic Information Science*, vol. 40, núm. 2, pp. 130–139, 2013.

[71] Seokyeon Kim *et al.*, “Big Data Visual Analytics System for Disease Pattern Analysis”, presentado en International Conference on *Big Data* Applications and Services, Jeju Island, Republic of Korea, 2015, pp. 175–179.

[72] Shahid Mahmud, Rahat Iqbal, y Faiyaz Doctor, “Cloud enabled data analytics and visualization framework for health-shocks prediction”, *Future Generation Computer Systems*, vol. 65, pp. 169–181, dic. 2016.

[73] Julia Kokina, Dessislava Pachamanova, y Andrew Corbett, “The role of data visualization and analytics in performance management: Guiding entrepreneurial growth decisions”, *Journal of Accounting Education*, vol. 38, pp. 50–62, mar. 2017.

[74] Cleverson Ledur, Dalvan Griebler, Isabel Manssour, y Luiz Gustavo Fernandes, “Towards a Domain-Specific Language for Geospatial Data Visualization Maps with Big Data Sets”, presentado en 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications, Marrakech, Morocco, 2015.

[75] Zhaowei Tan, Changfeng Liu, Yuning Mao, Yunqi Guo, Jiaming Shen, y Xinbing Wang, “AceMap: A Novel Approach towards Displaying Relationship among Academic Literatures”, presentado en International Conference Companion on World Wide Web, Montréal, Québec, Canada, 2016, pp. 437–442.

[76] Jillian Aurisano, Khairi Reda, Andrew Johnson, Elisabeta G Marai, y Jason Leigh, “BactoGeNIE: a large-scale comparative genome visualization for big displays”, *BMC Bioinformatics*, vol. 16, núm. 11, jul. 2015.

[77] Tam Thanh Nguyen y Insu Song, “Centrality Clustering-Based Sampling for Big Data Visualization”, presentado en International Joint Conference on Neural Networks, Vancouver, BC, Canada, 2016, pp. 1911–1917.

[78] Guangchen Ruan y Hui Zhang, “Closed-loop Big Data Analysis with Visualization and Scalable Computing”, *Big Data Research*, vol. 8, pp. 12–26, jul. 2017.

[79] Ali Salehi *et al.*, “SensorDB: a virtual laboratory for the integration, visualization and analysis of varied biological sensor data”, *Plant Methods*, 2015.

[80] Ahmed Eldawy, Mohamed F. Mokbel, y Christopher Jonathan, “HadoopViz: A MapReduce framework for extensible visualization of big spatial data”, presentado en 32nd International Conference on Data Engineering (ICDE), Helsinki, 2016, pp. 601–612.

[81] Usman Iqbal *et al.*, “Cancer-disease associations: A visualization andanimation through medical big data”, *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 44–51, abr. 2016.

[82] Shujaat Hussain y Sungyoung Lee, “Visualization and descriptive analytics of wellness data through Big Data”, presentado en The Tenth International Conference on Digital Information Management, Jeju, South Korea, 2015, pp. 69–71.

[83] Omesh Kumar y Abhishek Goyal, “Visualization: A novel approach for big data analytics”, presentado en Second International Conference on Computational Intelligence & Communication Technology, Ghaziabad, India, 2016, pp. 121–124.

[84] Hua Fang y Zhaoyang Zhang, “An Enhanced Visualization Method to Aid Behavioral Trajectory Pattern Recognition Infrastructure for Big Longitudinal Data”, *IEEE Transactions on Big Data*, ene. 2017.

[85] Matthias Nielsen y Kaj Grønbæk, “PivotViz: Interactive Visual Analysis of Multidimensional Library Transaction Data”, presentado en 15th ACM/IEEE-CS Joint Conference on Digital Libraries, Knoxville, Tennessee, USA, 2015, pp. 139–142.

[86] Carolin Helbig, Hans-Stefan Bauer, Karsten Rink, Volker Wulfmeyer, Michael Frank, y Olaf Kolditz, “Concept and workflow for 3D visualization of atmospheric data in a virtual reality environment for analytical approaches”, *Environmental Earth Sciences*, vol. 72, núm. 10, pp. 3767–3780, nov. 2014.

[87] F. J. Villanueva, C. Aguirre, A. Rubio, D. Villa, M. J. Santofimia, y J. C. López, “Data stream visualization framework for smart cities”, *Soft Computing*, pp. 1671–1681, 2015.

[88] Kayoko Yamamoto, “Visualization of GIS Analytic for Open Big Data in Environmental Science”, presentado en International Conference on Cloud Computing and Big Data, Shanghai, China, 2015, pp. 201–2018.

[89] “OEC - Data Sources”. [En línea]. Disponible en: http://atlas.media.mit.edu/en/resources/data/. [Consultado: 24-may-2017].

[90] “Wikipedia:Database download”, *Wikipedia*. 08-may-2017.

[91] "SNAP: Network datasets: LiveJournal social network". [En línea]. Disponible en: https://snap.stanford.edu/data/soc-LiveJournal1.html. [Consultado: 24-may-2017].

[92] "Webscope | Yahoo Labs". [En línea]. Disponible en: https://Webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67. [Consultado: 24-may-2017].

[93] Melyssa Barata, Jorge Bernardino, y Pedro Furtado, "Survey on Big Data and Decision Support Benchmarks", presentado en International Conference on Database and Expert Systems Applications, 2014, vol. 8645, pp. 174–182.

[94] "Big Data Benchmark for Big Bench: Big Bench Workload Development", 10-nov-2017. [En línea]. Disponible en: https://github.com/intel-hadoop/Big-Data-Benchmark-for-Big-Bench.

[95] "BigDataBench | A Big Data Benchmark Suite, ICT, Chinese Academy of Sciences". [En línea]. Disponible en: http://prof.ict.ac.cn/. [Consultado: 10-dic-2017].

[96] Ahmad Ghazal *et al.*, "BigBench: Towards an Industry Standard Benchmark for Big Data Analytics", presentado en ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 2013, pp. 1197–1208.

[97] Zijian Ming *et al.*, "BDGS: A Scalable Big Data Generator Suite in Big Data Benchmarking", presentado en Third and Fourth Workshop on Big Data Benchmarking, Xi'an, China, 2013, pp. 138–154.

[98] "RHadoop", 01-nov-2017. [En línea]. Disponible en: https://github.com/RevolutionAnalytics/RHadoop.

[99] F. Maxwell Harper y Joseph A. Konstan, "The MovieLens Datasets: History and Context", *ACM Transactions on Interactive Intelligent Systems (TiiS) - Regular Articles and Special issue on New Directions in Eye Gaze for Interactive Intelligent Systems (Part 1 of 2)*, vol. 5, núm. 4, ene. 2016.

Anexos

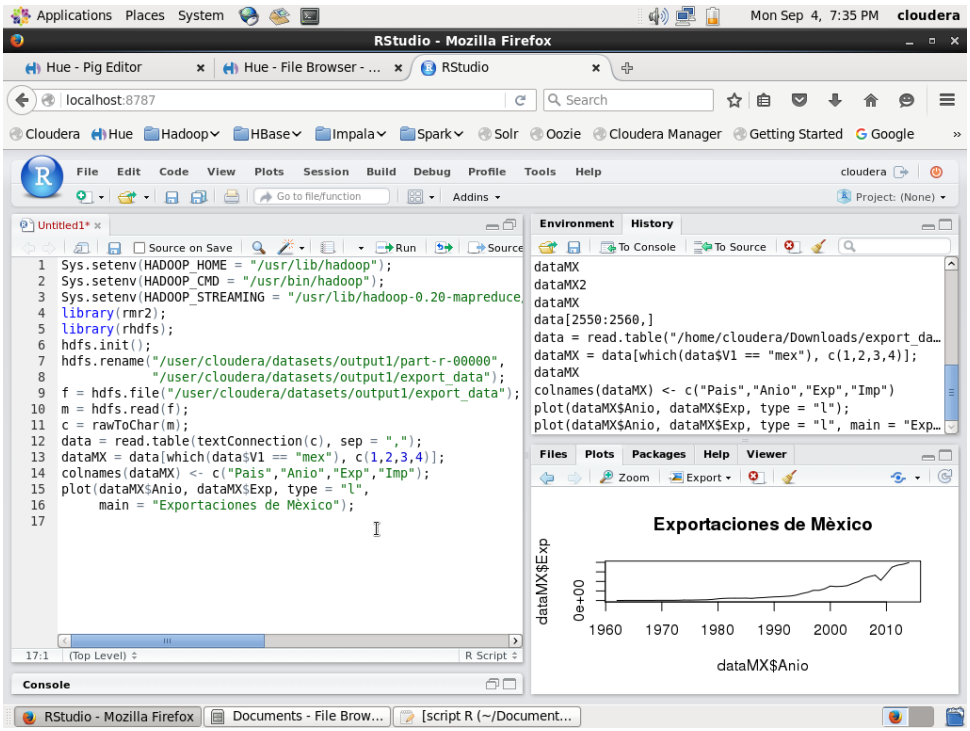


Figura A.1 *Script* R de gráfica de líneas para el conjunto de datos 1.

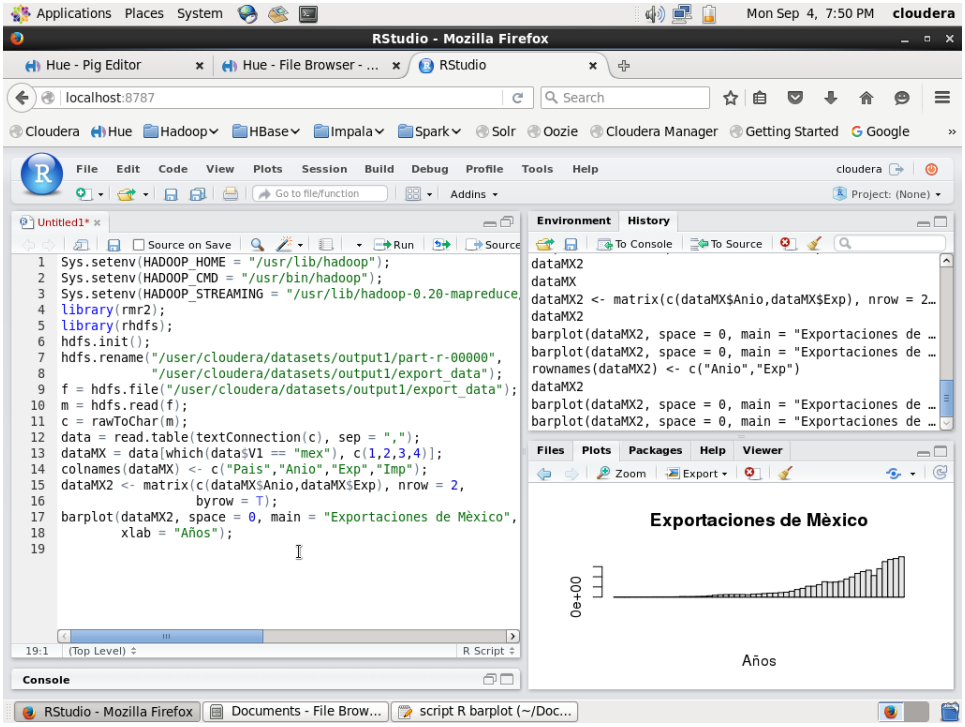


Figura A.2 *Script* R de gráfica de barras para el conjunto de datos 1.

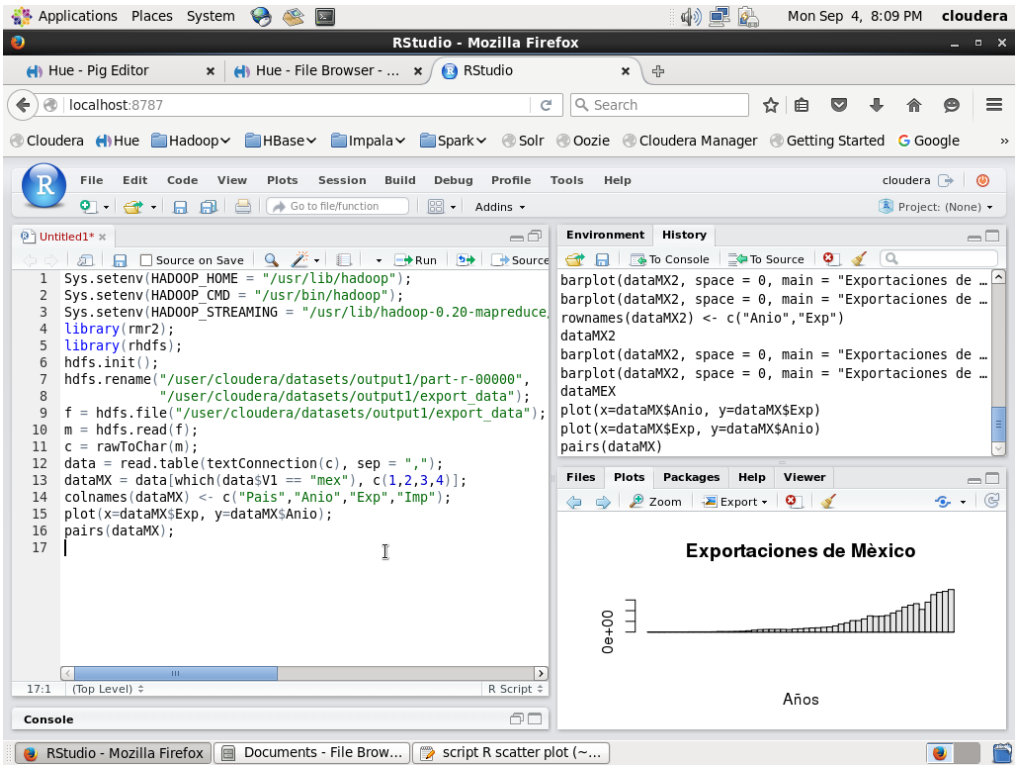


Figura A.3 *Script* R de gráfica de dispersión para el conjunto de datos 1.

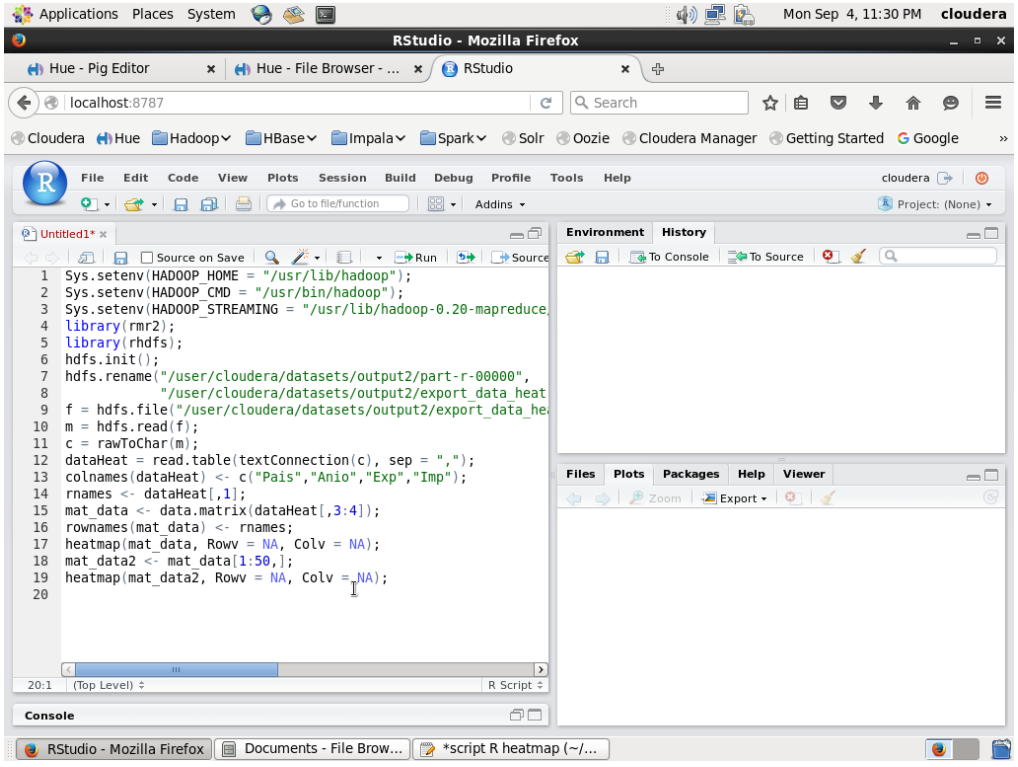


Figura A.4 *Script* R de *Heatmap* para el conjunto de datos 1.

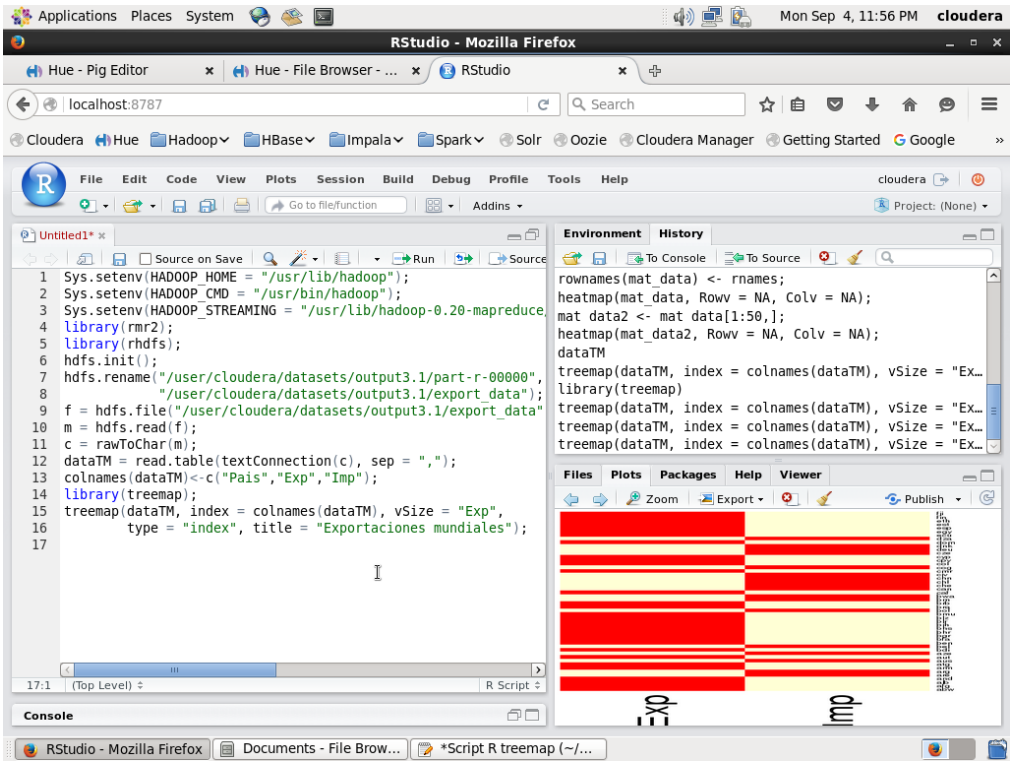


Figura A.5 *Script* R de *Treemap* para el conjunto de datos 1.

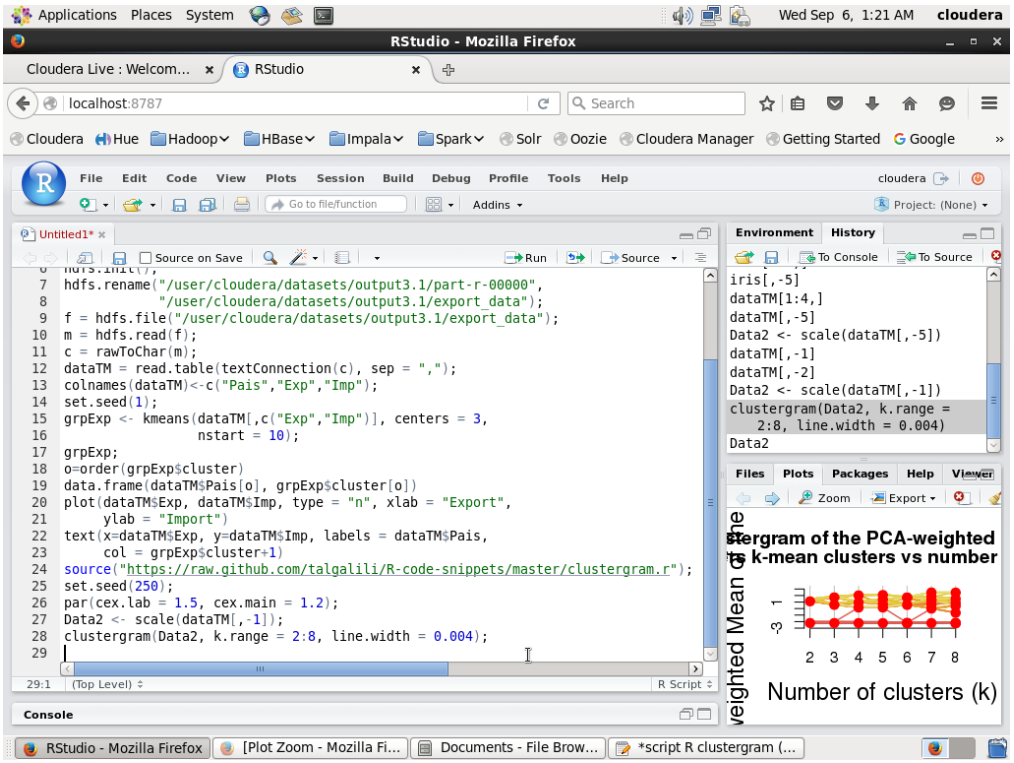


Figura A.6 *Script* R de Clustergram para el conjunto de datos 1.

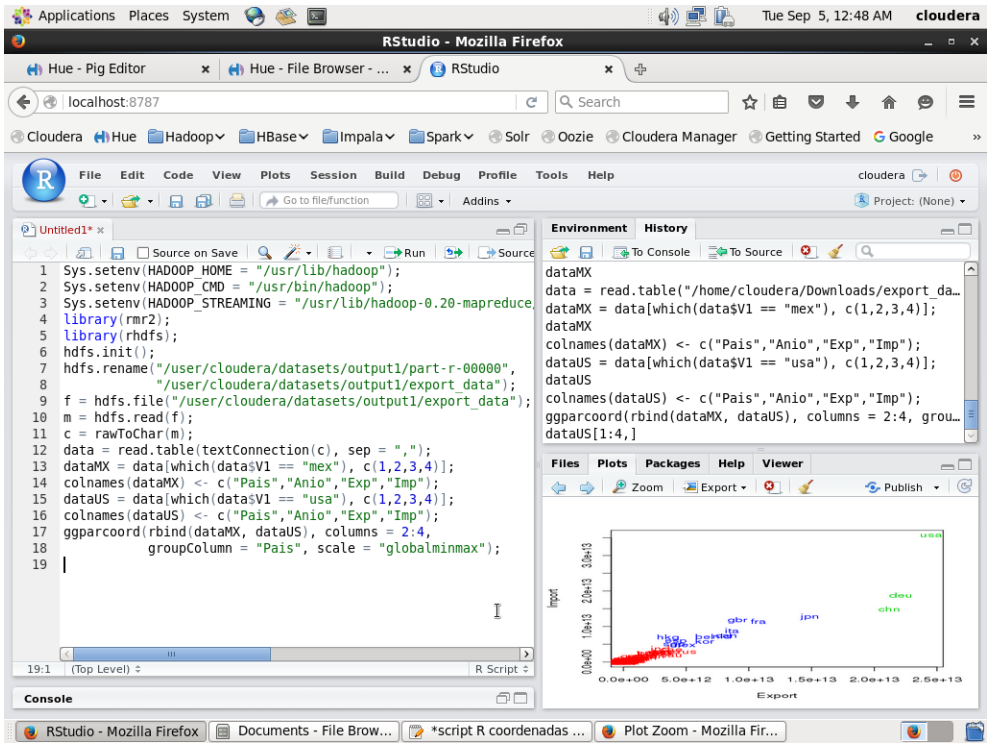


Figura A.7 *Script* R de coordenadas paralelas para el conjunto de datos 1.

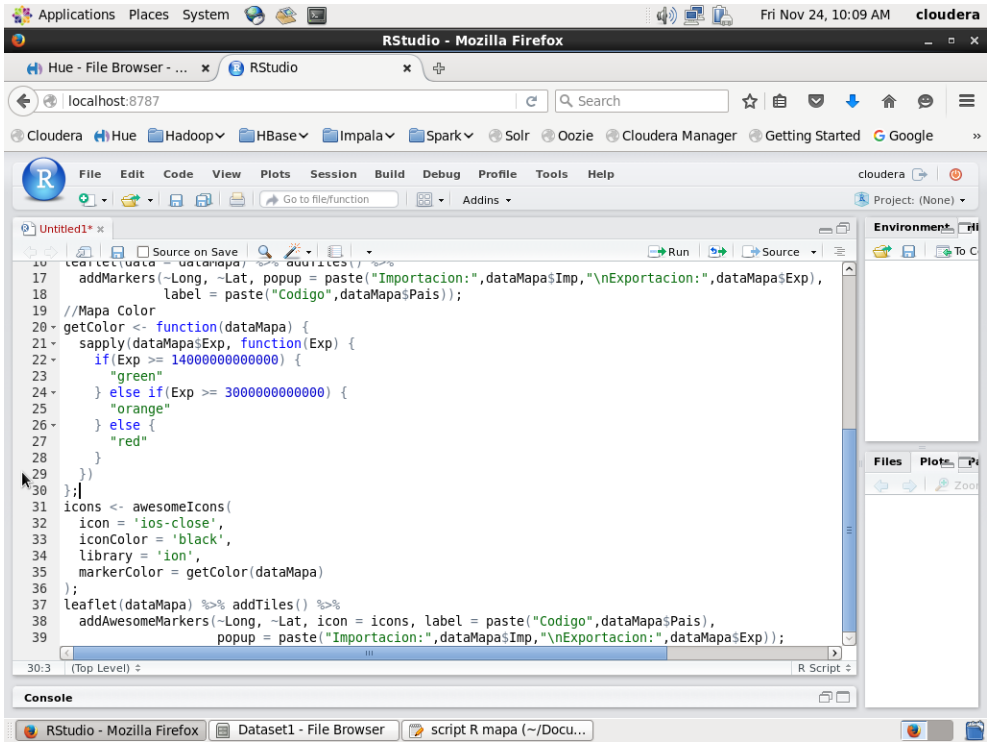


Figura A.8 *Script* R de *Spatial Information Flow* para el conjunto de datos 1.

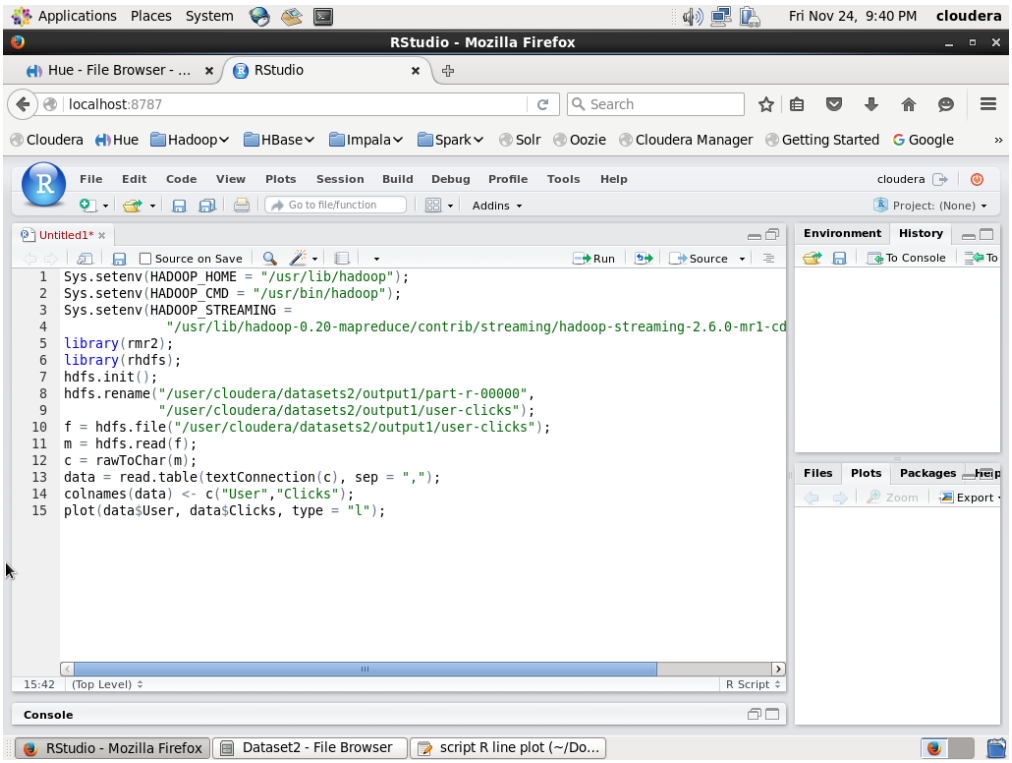


Figura A.9 *Script* R de grafica de líneas para el conjunto de datos 2.

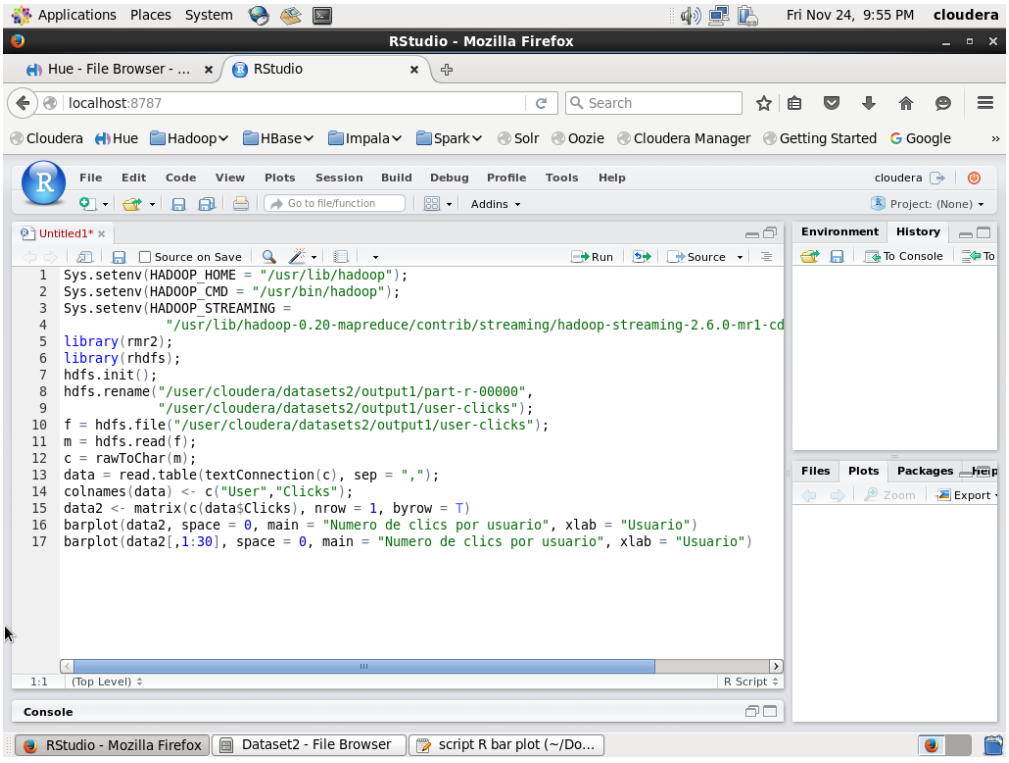


Figura A.10 *Script* R de gráfica de barras para el conjunto de datos 2.

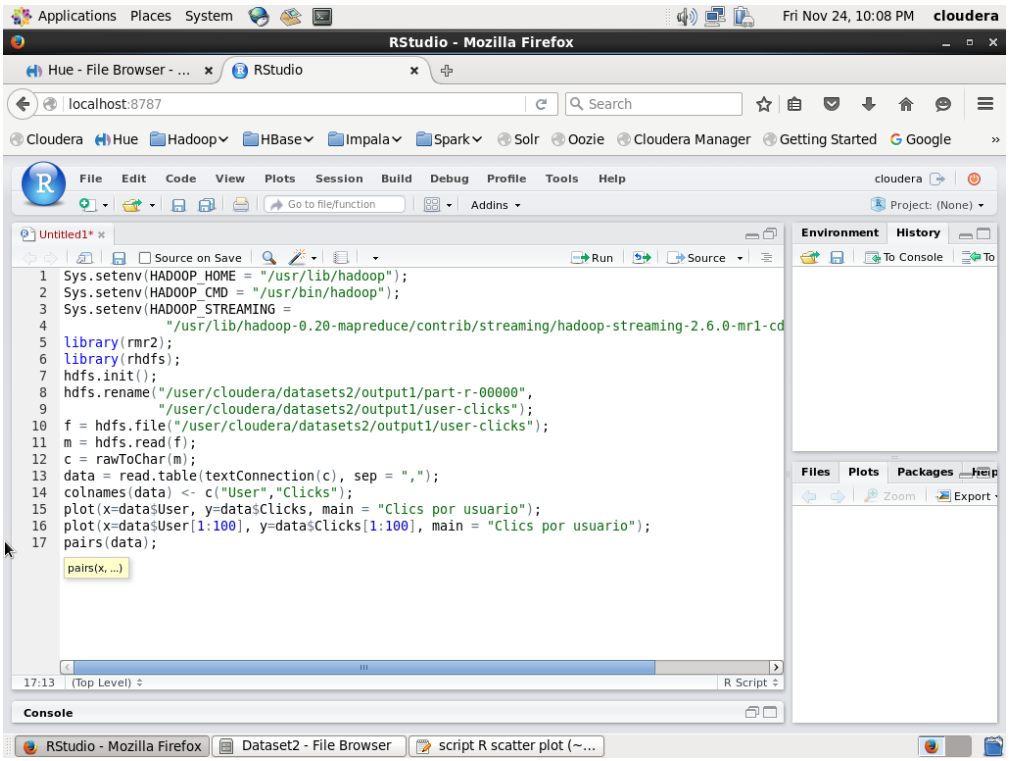


Figura A.11 *Script* R de gráfica de dispersión para el conjunto de datos 2.

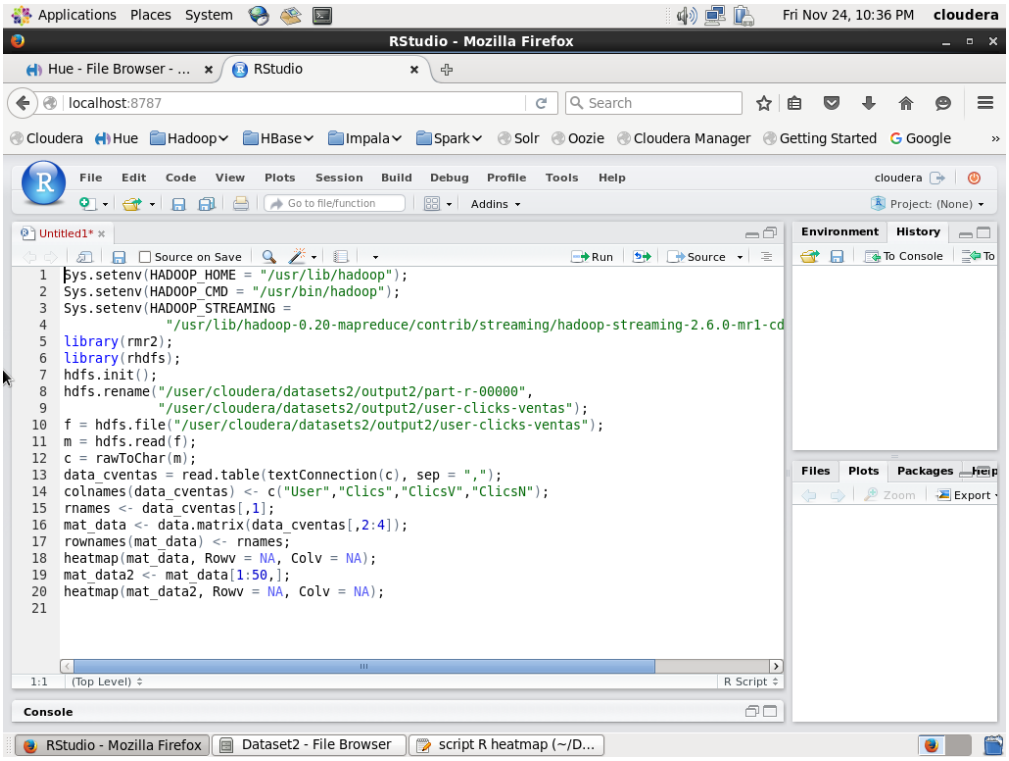


Figura A.12 *Script* R de *Heatmap* para el conjunto de datos 2.

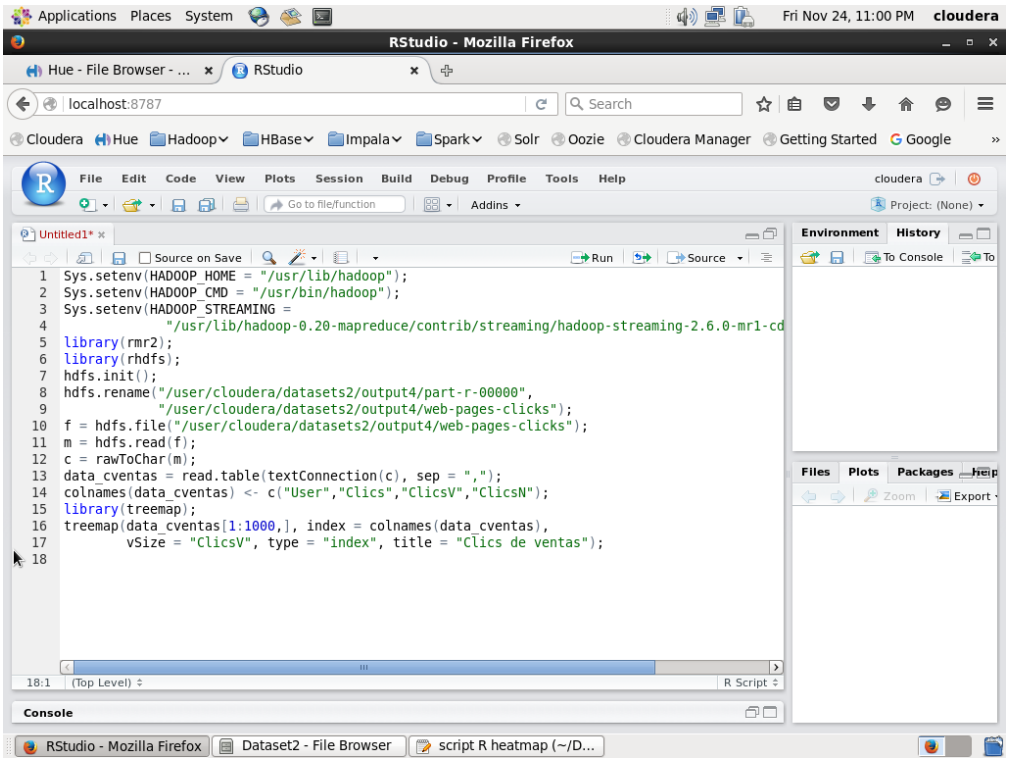


Figura A.13 *Script* R de *Treemap* para el conjunto de datos 2.

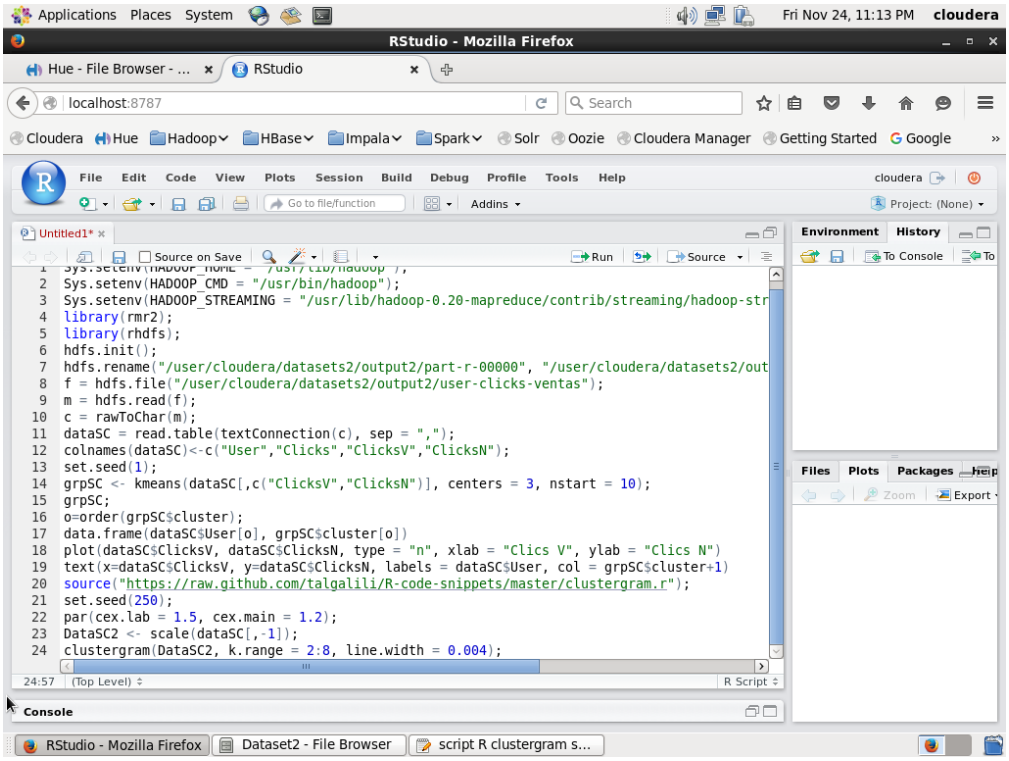


Figura A.14 *Script* R de Clustergram para el conjunto de datos 2.

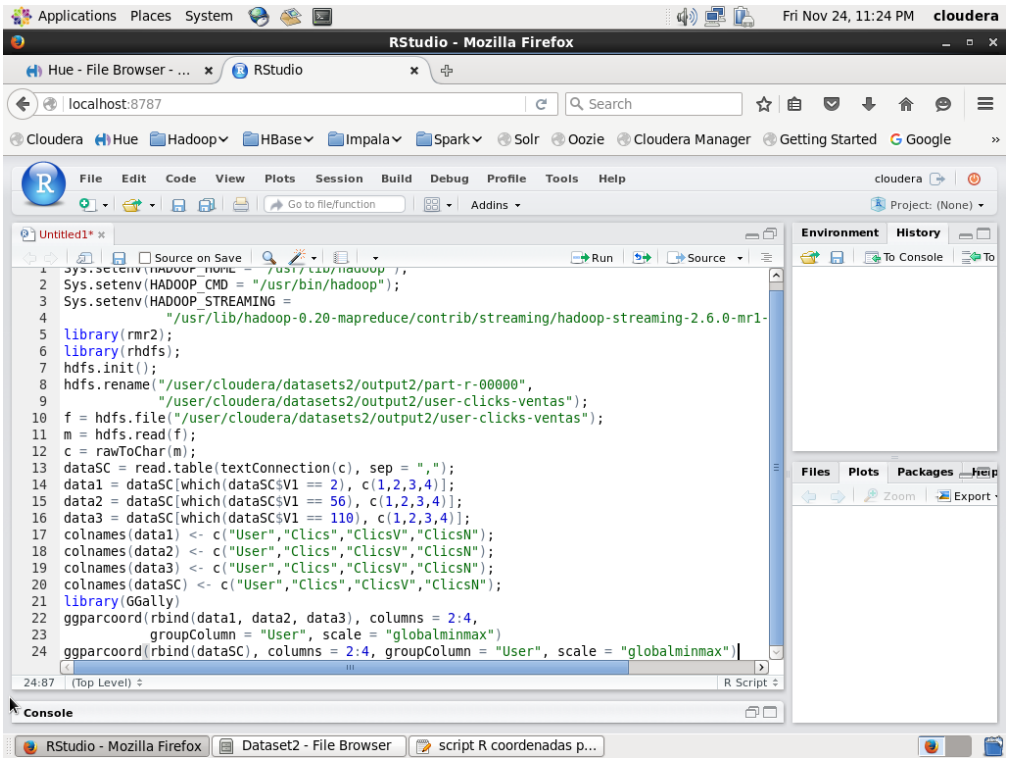


Figura A.15 *Script* R de coordenadas paralelas para el conjunto de datos 2.

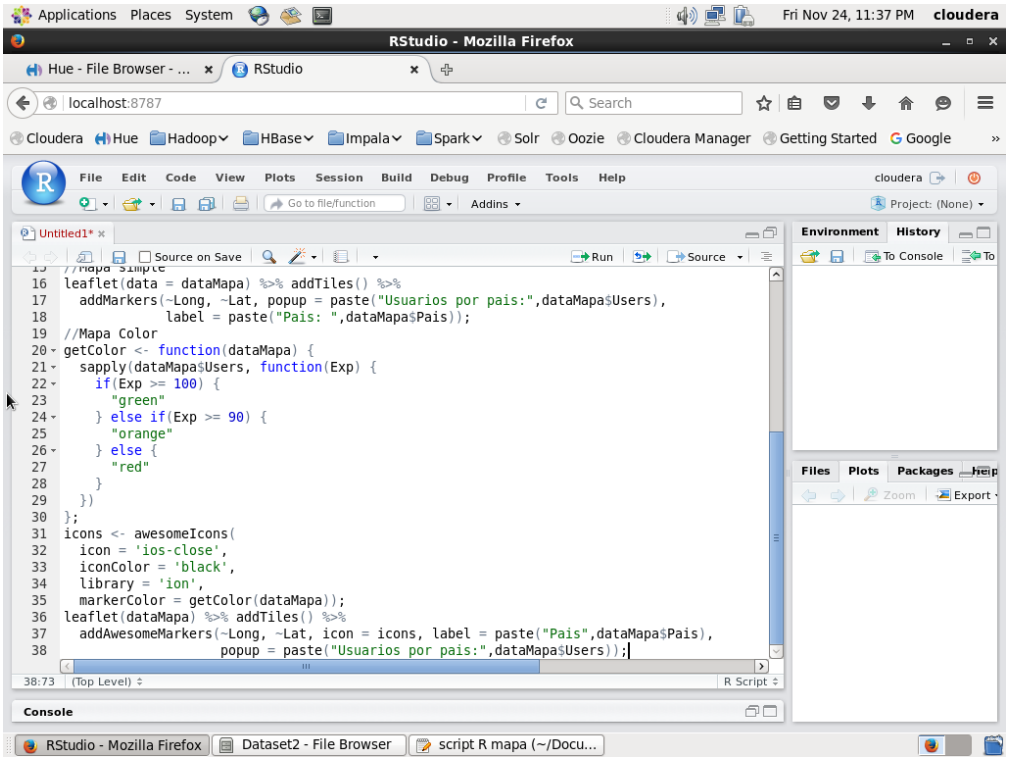


Figura A.16 *Script* R de *Spatial Information Flow* para el conjunto de datos 2.

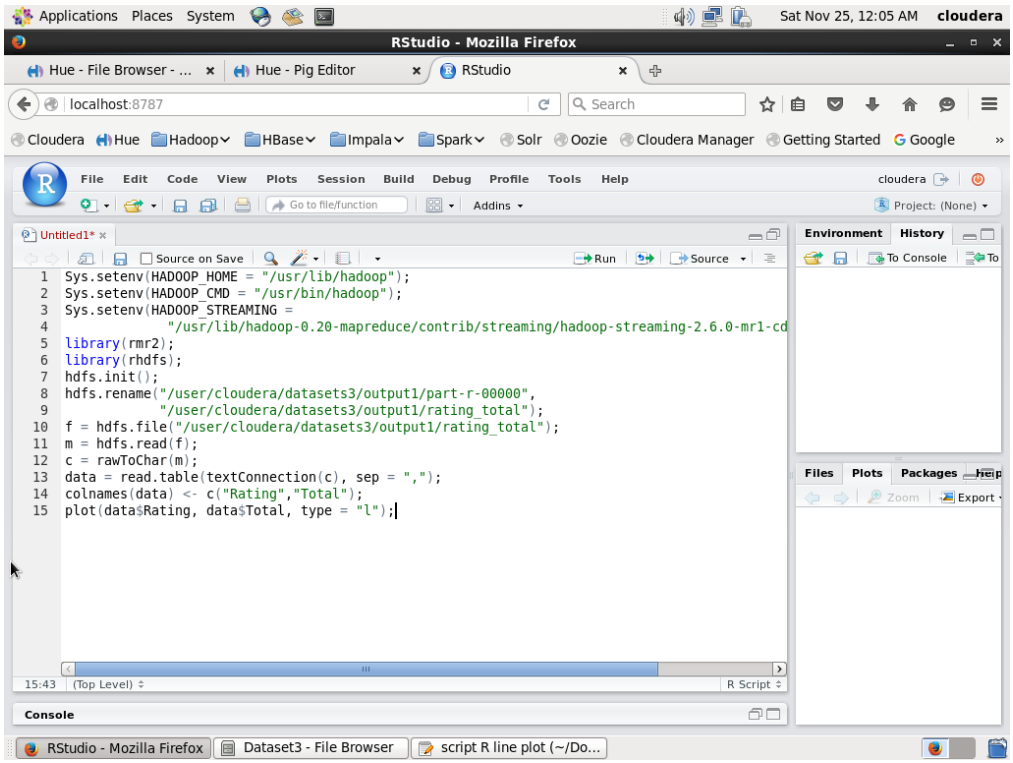


Figura A.17 *Script* R de grafica de líneas para el conjunto de datos 3.

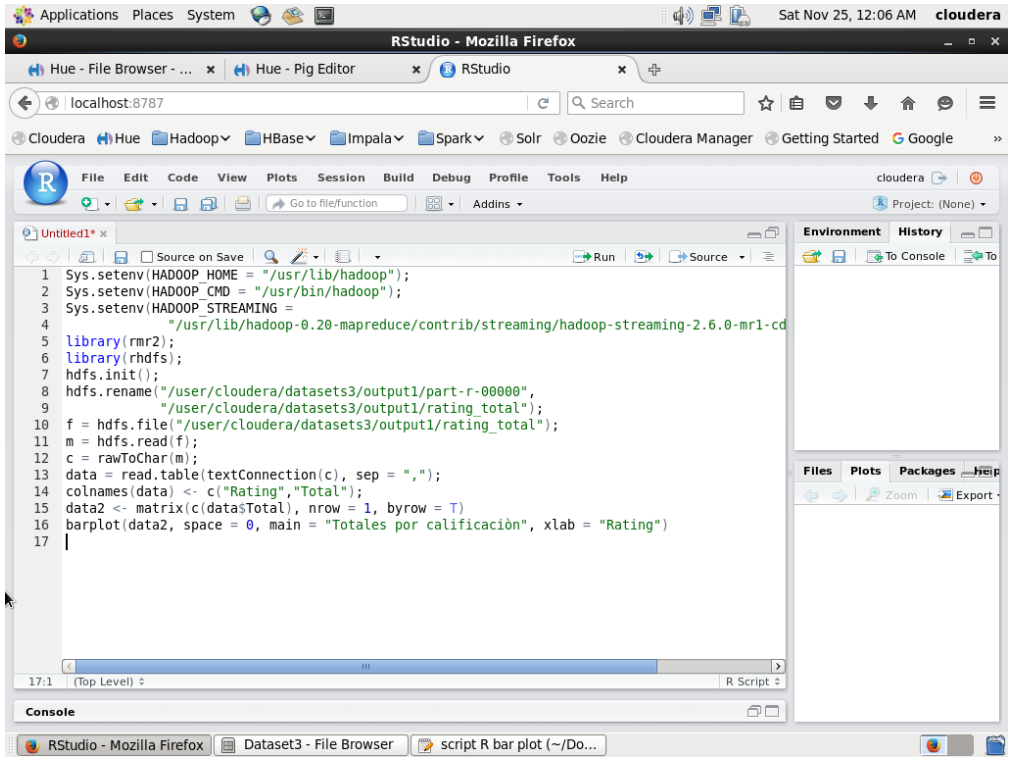


Figura A.18 *Script* R de gráfica de barras para el conjunto de datos 3.

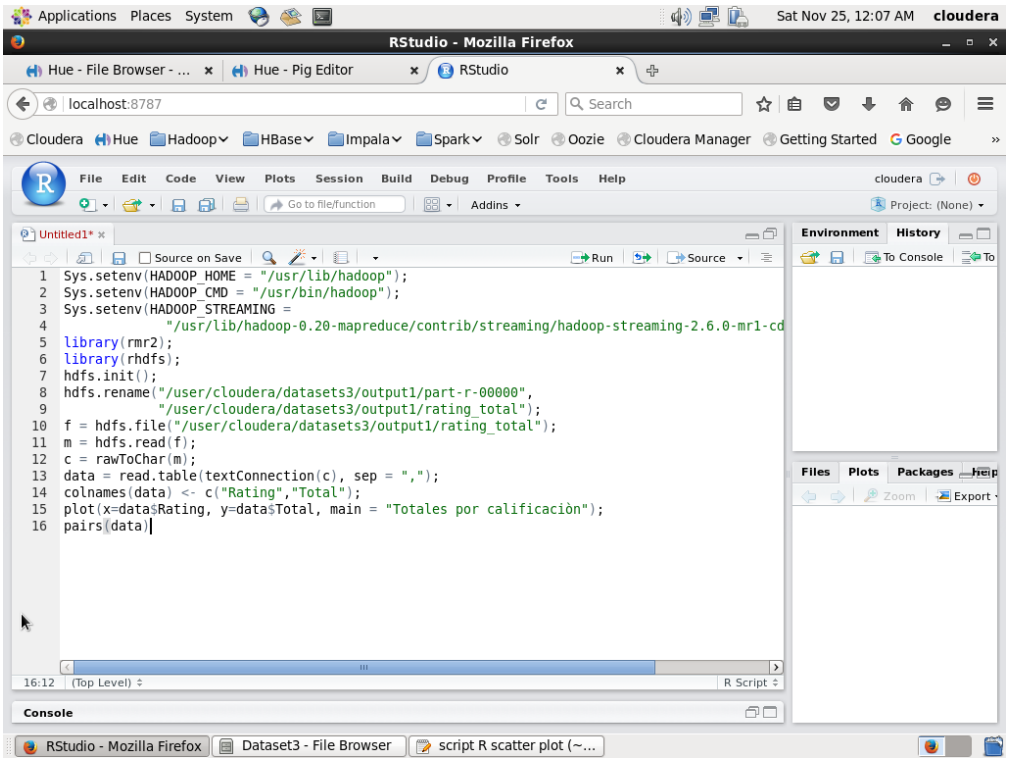


Figura A.19 *Script* R de gráfica de dispersión para el conjunto de datos 3.

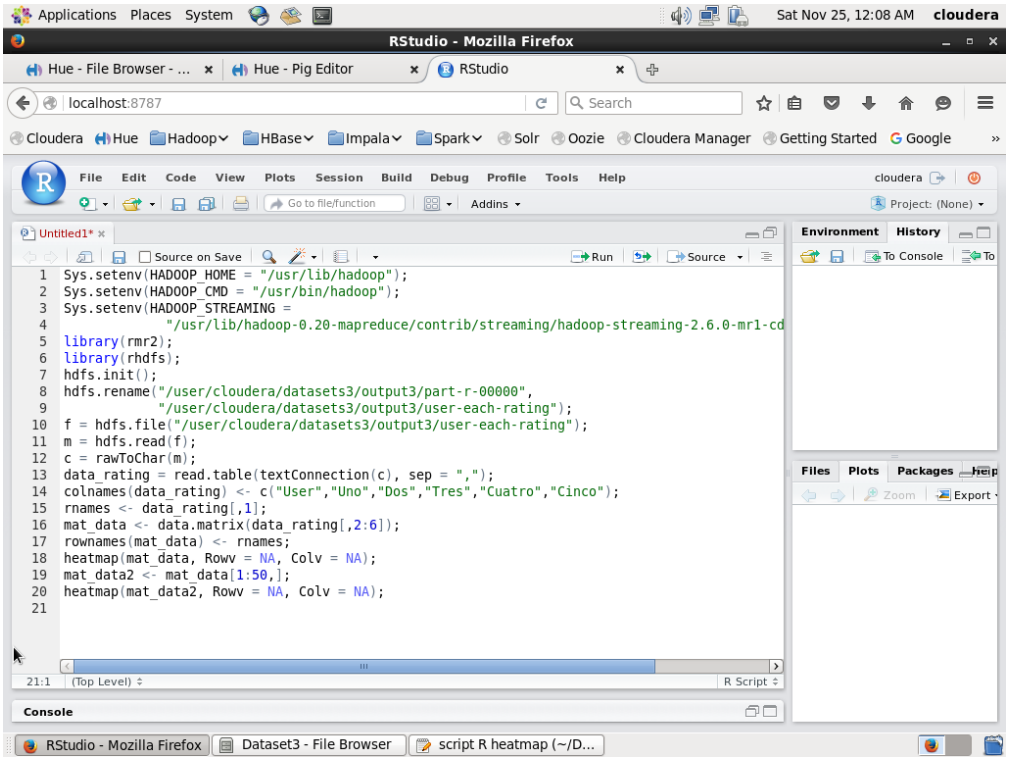


Figura A.20 *Script* R de *Heatmap* para el conjunto de datos 3.

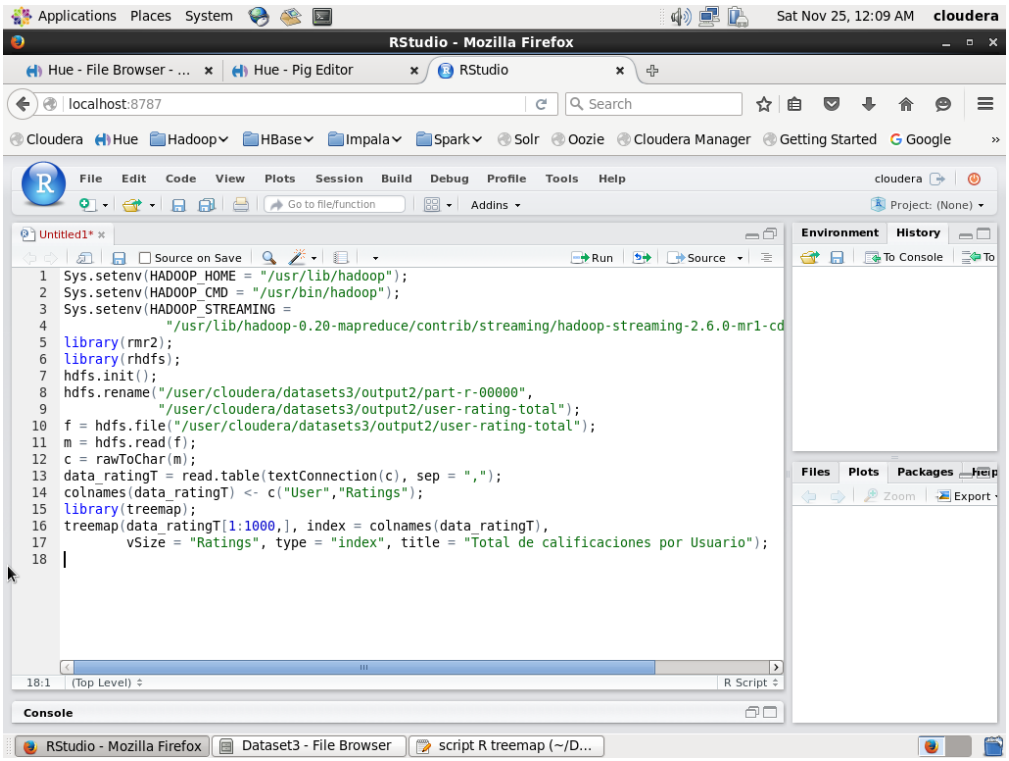


Figura A.21 *Script* R de *Treemap* para el conjunto de datos 3.

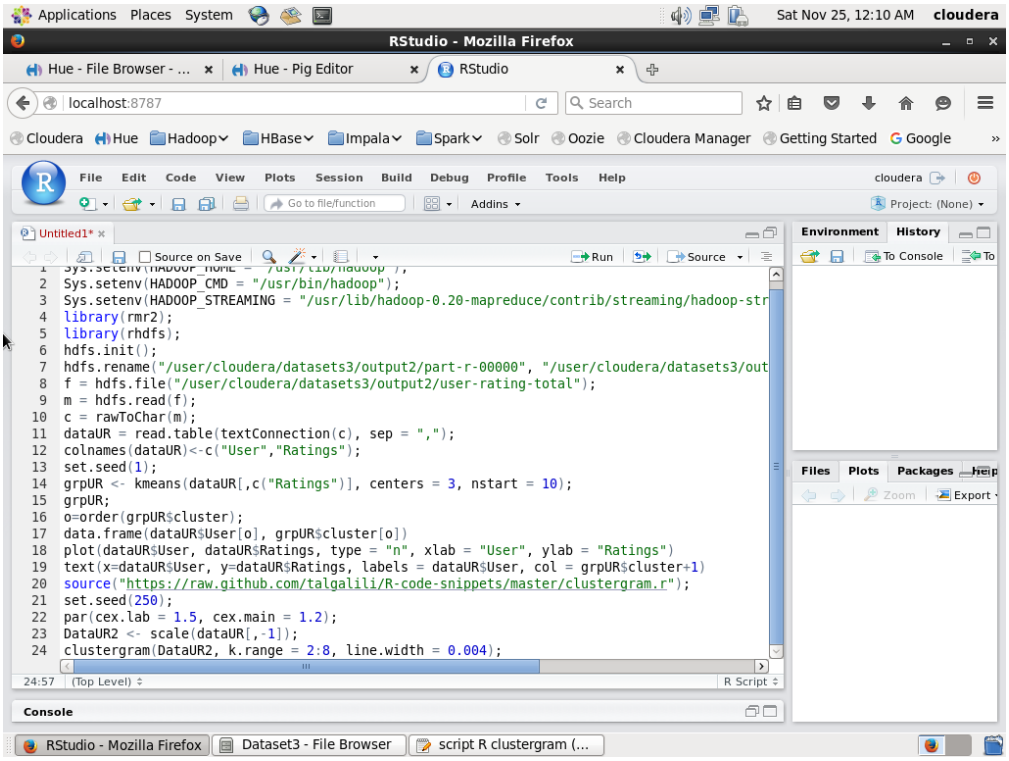
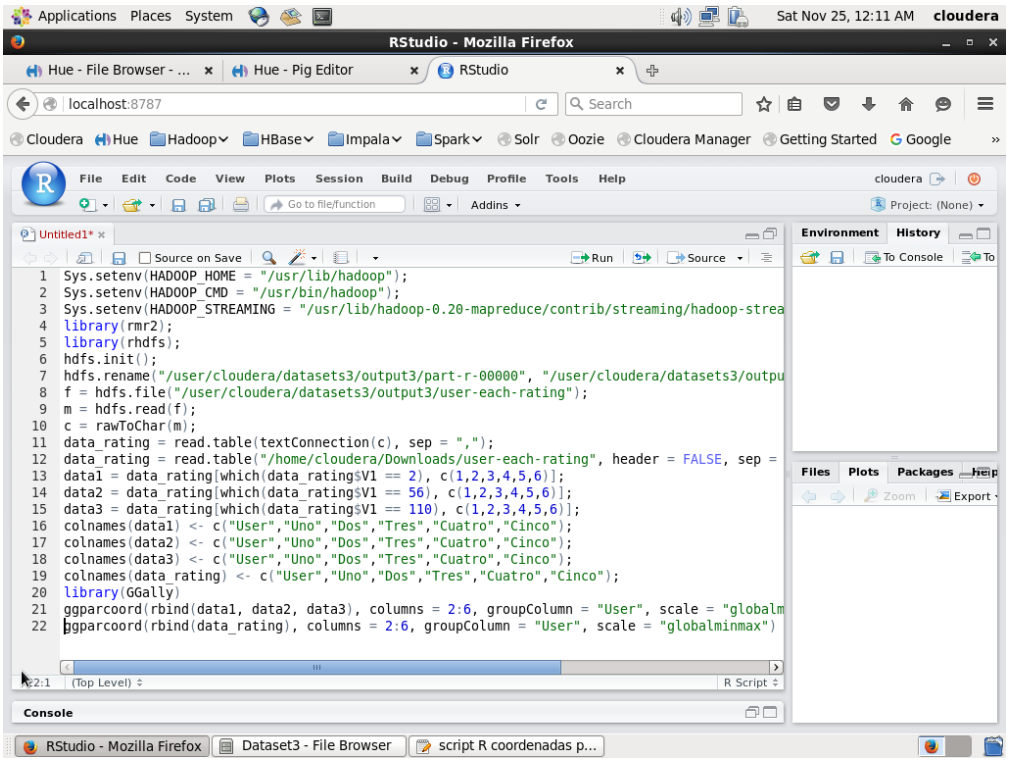
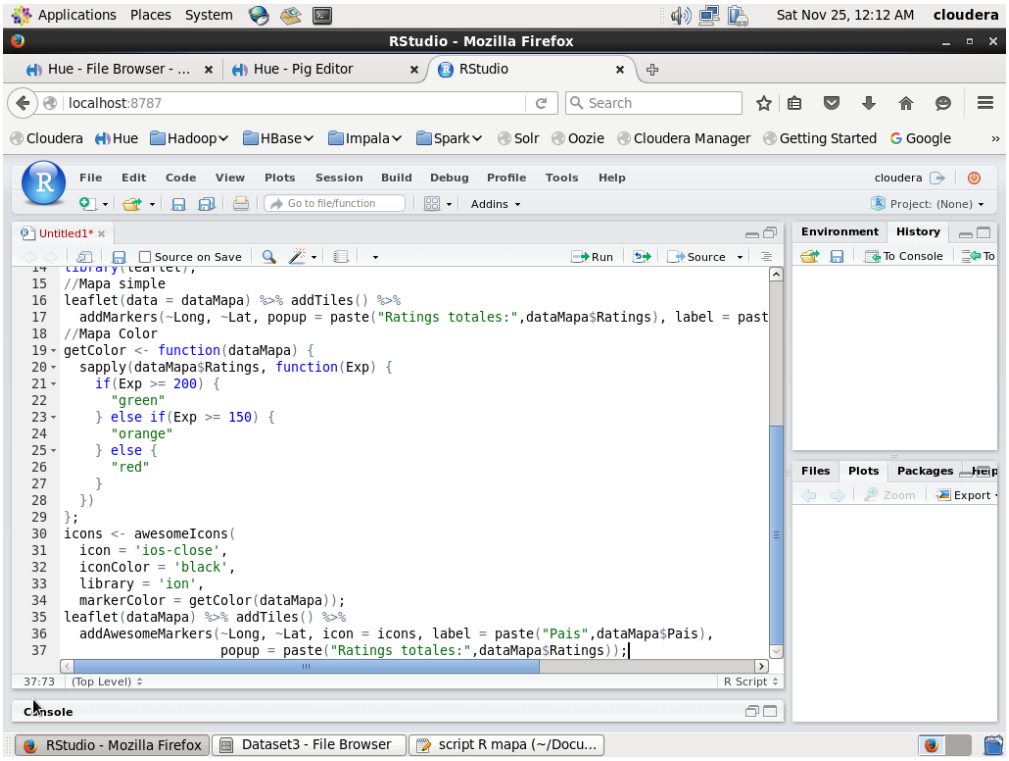


Figura A.22 *Script* R de Clustergram para el conjunto de datos 3.




```
1  Sys.setenv(HADOOP_HOME = "/usr/lib/hadoop");
2  Sys.setenv(HADOOP_CMD = "/usr/bin/hadoop");
3  Sys.setenv(HADOOP_STREAMING = "/usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-strea
4  library(rmr2);
5  library(rhdfs);
6  hdfs.init();
7  hdfs.rename("/user/cloudera/datasets3/output3/part-r-00000", "/user/cloudera/datasets3/outpu
8  f = hdfs.file("/user/cloudera/datasets3/output3/user-each-rating");
9  m = hdfs.read(f);
10 c = rawToChar(m);
11 data_rating = read.table(textConnection(c), sep = ",");
12 data_rating = read.table("/home/cloudera/Downloads/user-each-rating", header = FALSE, sep =
13 data1 = data_rating[which(data_rating$V1 == 2), c(1,2,3,4,5,6)];
14 data2 = data_rating[which(data_rating$V1 == 56), c(1,2,3,4,5,6)];
15 data3 = data_rating[which(data_rating$V1 == 110), c(1,2,3,4,5,6)];
16 colnames(data1) <- c("User","Uno","Dos","Tres","Cuatro","Cinco");
17 colnames(data2) <- c("User","Uno","Dos","Tres","Cuatro","Cinco");
18 colnames(data3) <- c("User","Uno","Dos","Tres","Cuatro","Cinco");
19 colnames(data_rating) <- c("User","Uno","Dos","Tres","Cuatro","Cinco");
20 library(GGally)
21 ggparcoord(rbind(data1, data2, data3), columns = 2:6, groupColumn = "User", scale = "globalm
22 ggparcoord(rbind(data_rating), columns = 2:6, groupColumn = "User", scale = "globalminmax")
```

Figura A.23 *Script* R de coordenadas paralelas para el conjunto de datos 3.




```
15 //Mapa simple
16 leaflet(data = dataMapa) %>% addTiles() %>%
17   addMarkers(~Long, ~Lat, popup = paste("Ratings totales:",dataMapa$Ratings), label = past
18 //Mapa Color
19 getColor <- function(dataMapa) {
20   sapply(dataMapa$Ratings, function(Exp) {
21     if(Exp >= 200) {
22       "green"
23     } else if(Exp >= 150) {
24       "orange"
25     } else {
26       "red"
27     }
28   })
29 };
30 icons <- awesomeIcons(
31   icon = 'ios-close',
32   iconColor = 'black',
33   library = 'ion',
34   markerColor = getColor(dataMapa));
35 leaflet(dataMapa) %>% addTiles() %>%
36   addAwesomeMarkers(~Long, ~Lat, icon = icons, label = paste("Pais",dataMapa$Pais),
37                     popup = paste("Ratings totales:",dataMapa$Ratings));
```

Figura A.24 *Script* R de *Spatial Information Flow* para el conjunto de datos 3.

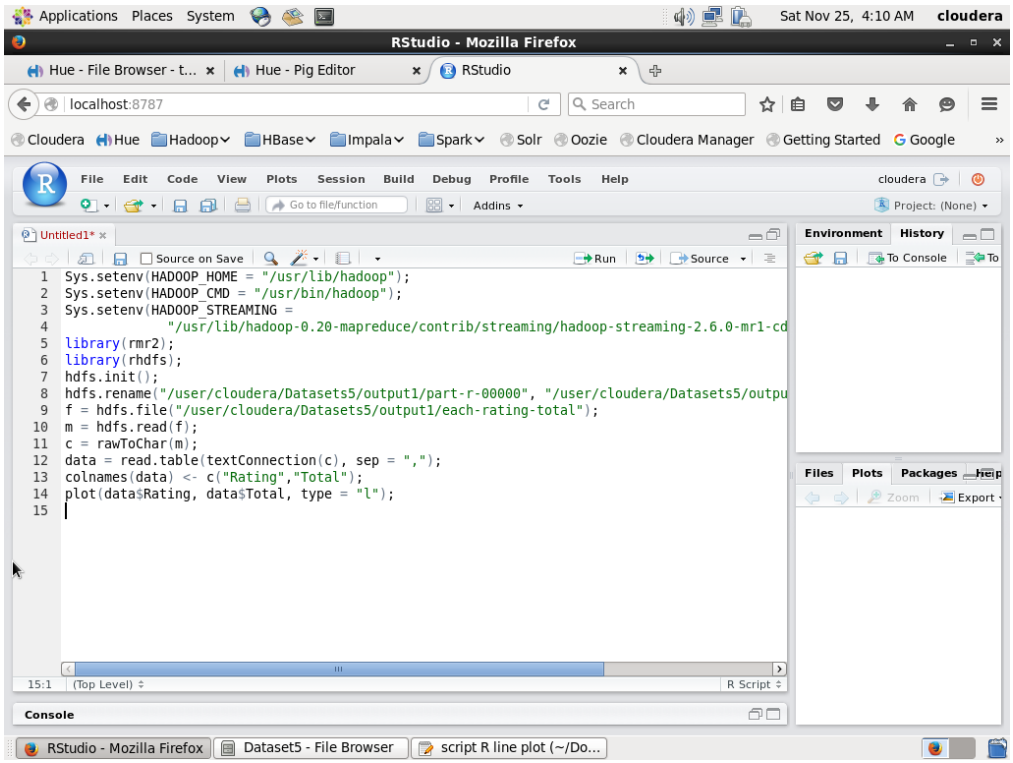


Figura A.25 *Script* R de grafica de líneas para el conjunto de datos 4.

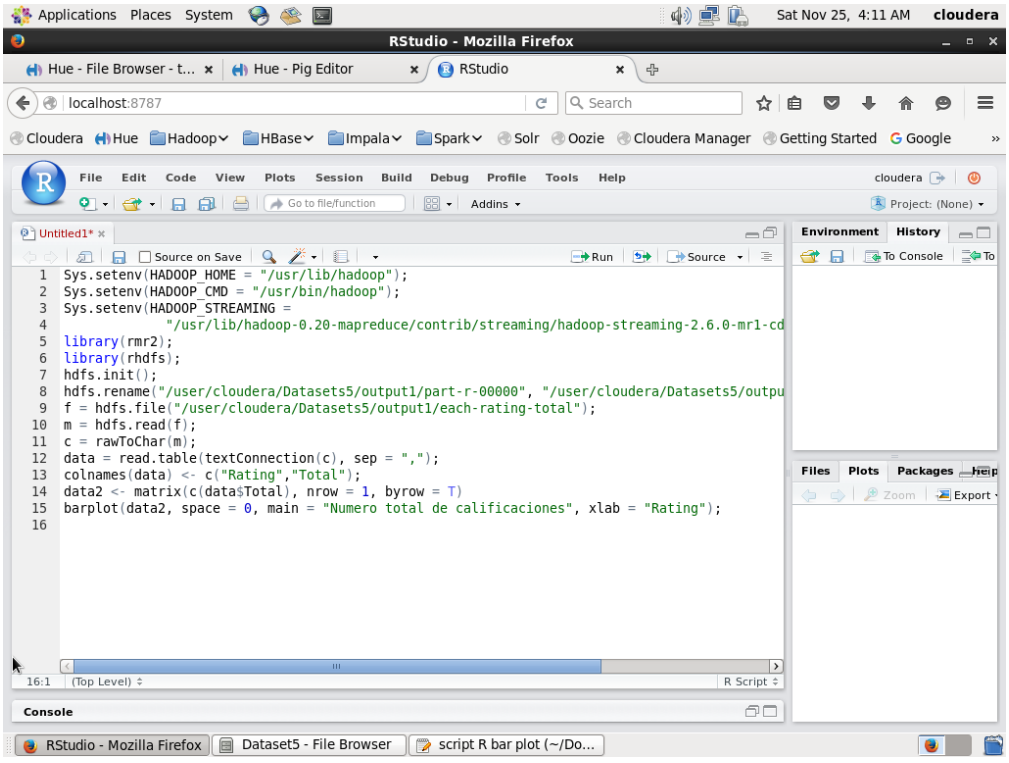


Figura A.26 *Script* R de gráfica de barras para el conjunto de datos 4.

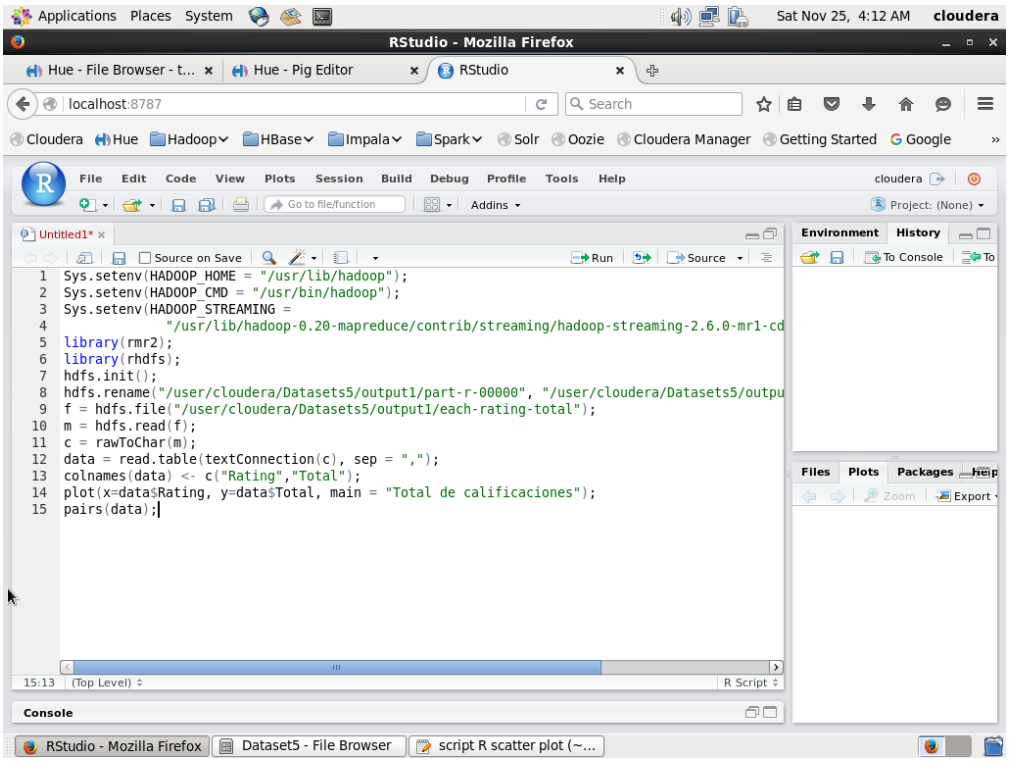


Figura A.27 *Script* R de gráfica de dispersión para el conjunto de datos 4.

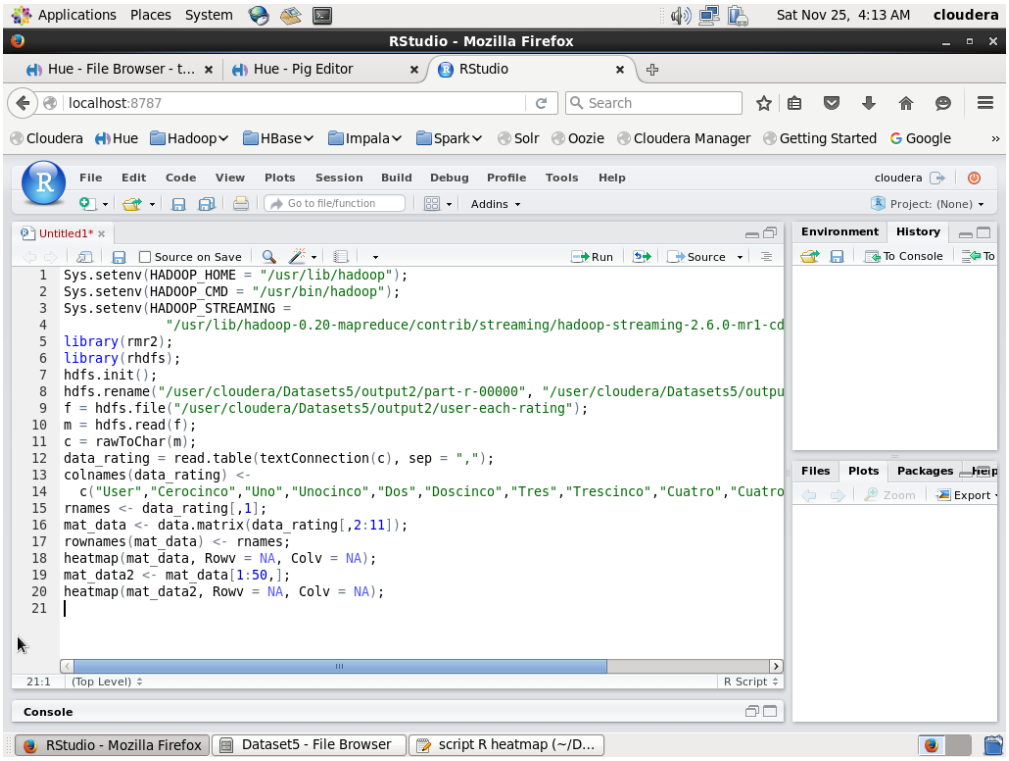
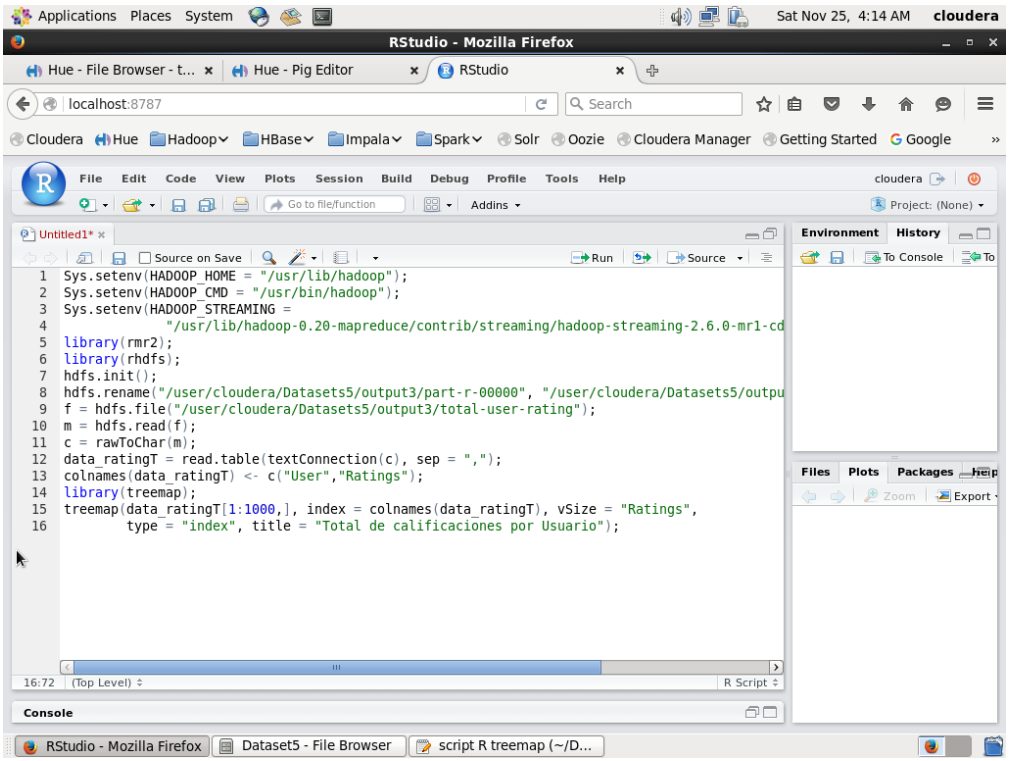
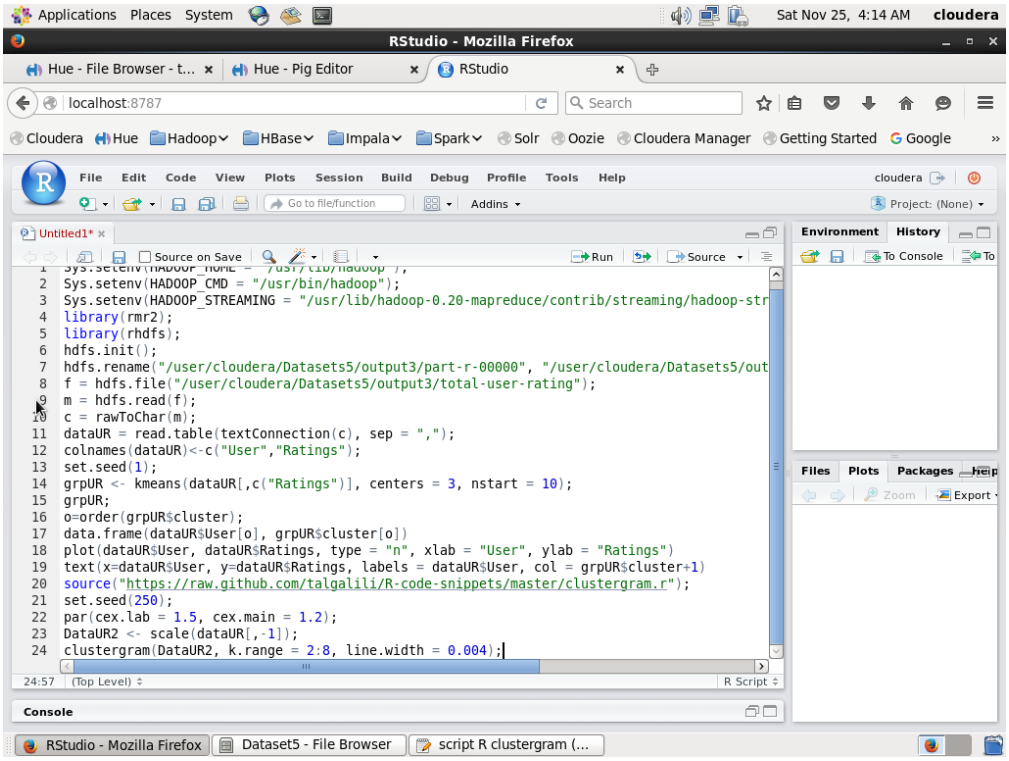


Figura A.28 *Script* R de *Heatmap* para el conjunto de datos 4.




```
Sys.setenv(HADOOP_HOME = "/usr/lib/hadoop");
Sys.setenv(HADOOP_CMD = "/usr/bin/hadoop");
Sys.setenv(HADOOP_STREAMING =
             "/usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-2.6.0-mr1-cd
library(rmr2);
library(rhdfs);
hdfs.init();
hdfs.rename("/user/cloudera/Datasets5/output3/part-r-00000", "/user/cloudera/Datasets5/outpu
f = hdfs.file("/user/cloudera/Datasets5/output3/total-user-rating");
m = hdfs.read(f);
c = rawToChar(m);
data_ratingT = read.table(textConnection(c), sep = ",");
colnames(data_ratingT) <- c("User","Ratings");
library(treemap);
treemap(data_ratingT[1:1000,], index = colnames(data_ratingT), vSize = "Ratings",
        type = "index", title = "Total de calificaciones por Usuario");
```

Figura A.29 *Script* R de *Treemap* para el conjunto de datos 4.




```
Sys.setenv(HADOOP_CMD = "/usr/bin/hadoop");
Sys.setenv(HADOOP_STREAMING = "/usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-str
library(rmr2);
library(rhdfs);
hdfs.init();
hdfs.rename("/user/cloudera/Datasets5/output3/part-r-00000", "/user/cloudera/Datasets5/out
f = hdfs.file("/user/cloudera/Datasets5/output3/total-user-rating");
m = hdfs.read(f);
c = rawToChar(m);
dataUR = read.table(textConnection(c), sep = ",");
colnames(dataUR)<-c("User","Ratings");
set.seed(1);
grpUR <- kmeans(dataUR[,c("Ratings")], centers = 3, nstart = 10);
grpUR;
o=order(grpUR$cluster);
data.frame(dataUR$User[o], grpUR$cluster[o])
plot(dataUR$User, dataUR$Ratings, type = "n", xlab = "User", ylab = "Ratings")
text(x=dataUR$User, y=dataUR$Ratings, labels = dataUR$User, col = grpUR$cluster+1)
source("https://raw.github.com/talgalili/R-code-snippets/master/clustergram.r");
set.seed(250);
par(cex.lab = 1.5, cex.main = 1.2);
DataUR2 <- scale(dataUR[,-1]);
clustergram(DataUR2, k.range = 2:8, line.width = 0.004);
```

Figura A.30 *Script* R de Clustergram para el conjunto de datos 4.

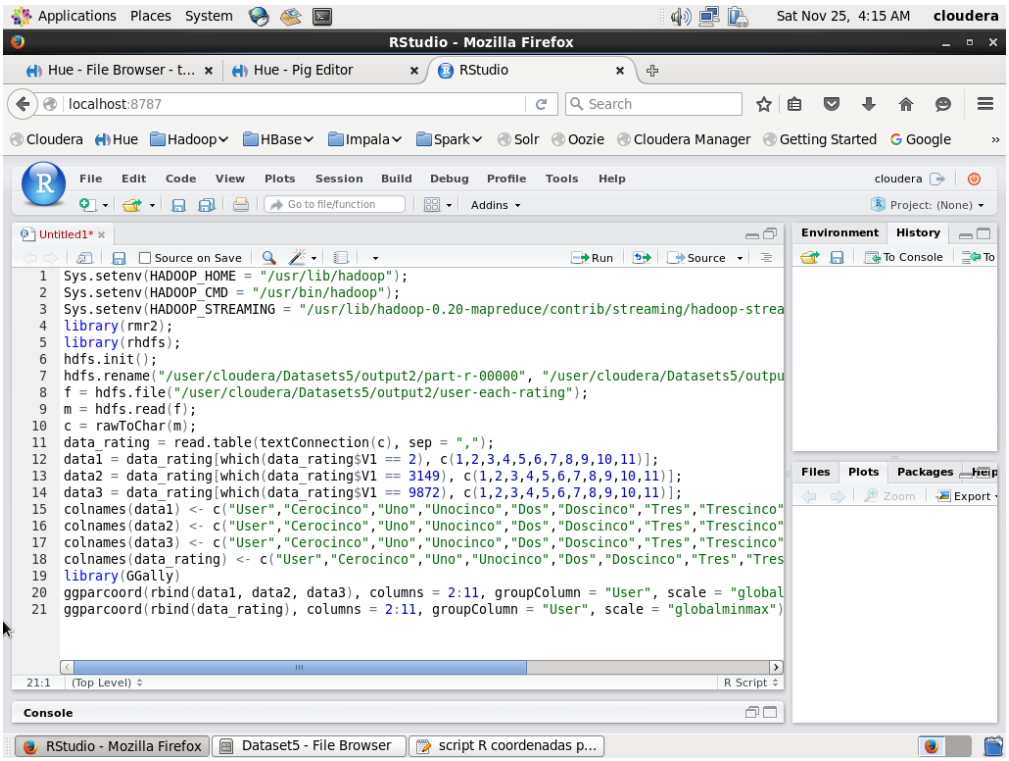


Figura A.31 *Script* R de coordenadas paralelas para el conjunto de datos 4.