# Analizador de accidentes de tráfico mediante Big Data

Eyad Abdullah
Departamento de Sistemas de
Información Facultad de Ciencias de la Computación
y la Información Universidad Rey Saud, Riad,
Arabia Saudita sfbit01_beta@live.com

Ahmed Emam
Departamento de Sistemas de
Información Facultad de Ciencias de la Computación
y la Información Universidad Rey Saud, Riad,
Arabia Saudita Universidad de Menoufia,
Menoufia, Egipto aemam@kus.edu.sa

Resumen: Los accidentes de tráfico son problemas graves que pueden causar discapacidades, lesiones e incluso la muerte. Para reducir el número de accidentes, es necesario comprender y analizar los datos de accidentes de tráfico. Casi a diario, alguien sufre accidentes de tráfico de una forma u otra, como la ralentización del tráfico debido a un accidente o una colisión en la misma vía, lo que puede dejar uno o más carriles sin disponibilidad. El ecosistema de Big Data tiene la capacidad de almacenar, manipular, analizar y extraer grandes conjuntos de datos de accidentes de tráfico, impulsando la creación de conocimiento que ayuda a los responsables de la toma de decisiones a reducir el número de accidentes. La aplicación desarrollada utiliza conjuntos de datos de tráfico real masivos, provenientes de las colisiones de tráfico de Nueva York. Esta aplicación consta de varias funciones y servicios web para analizar y visualizar la información sobre los principales accidentes de tráfico. La aplicación almacena los datos de tráfico masivos en Hadoop con un marco de computación paralela para la técnica Map-Reduce, que a su vez utiliza la interfaz de servicios web para respaldar la aplicación de minería desarrollada.

Términos del índice: Big Data, Accidentes de tráfico, Minería de datos, Hadoop y Map-Reducer.

## I. INTRODUCCIÓN

Los accidentes de tráfico tienen un gran impacto económico debido a la cantidad de lesiones y muertes que causan. Actualmente, muchos investigadores prestan especial atención a la determinación de los factores comunes que afectan significativamente a los accidentes de tráfico y a su análisis. Existen diversos enfoques aplicados para investigar este problema, como las redes neuronales artificiales, la minería de datos, la formulación lógica y los mapas difusos ART. Para lograr la mayor reducción posible de accidentes, con recursos presupuestarios limitados, es fundamental que las medidas y el análisis se basen en la investigación científica y objetiva de las causas de los accidentes [1]. El investigador especializado en minería de datos se encargó de desarrollar un método para el análisis de las causas de los accidentes de tráfico basado en la minería de datos que analizara los atributos y las causas relacionadas. Accidentes de tráfico.

Analyzer es un software escrito en lenguaje de programación C# que utiliza la biblioteca SSH [10]. Analiza accidentes de tráfico para proporcionar al usuario un conocimiento general sobre los mismos, lo que facilita la identificación de problemas y la búsqueda de soluciones. El análisis de accidentes de tráfico se realiza mediante funciones de agregación de MySQL (como SUM y COUNT) y también utiliza técnicas de minería de datos, como Mahout y Hadoop. La aplicación en sí no procesa datos.

Datos: simplemente envía y recibe comandos del servidor; el servidor se encargará de todo el procesamiento. El software utiliza las capacidades de Hadoop, Mahout y MySQL. Apache Hadoop es un framework escrito en Java para el almacenamiento y procesamiento distribuido de datos muy grandes [7]. Apache Mahout es un proyecto de Apache para producir implementaciones gratuitas de aprendizaje automático escalable y minería de datos para grandes volúmenes de datos; necesita Hadoop para funcionar. Mahout distribuye el trabajo entre varios nodos de la red para acelerar el procesamiento y admite diversas técnicas y algoritmos de minería de datos, como agrupamiento, clasificación, filtrado colaborativo y muchos otros. MySQL es un sistema de gestión de bases de datos de código abierto que admite grandes cantidades de datos [9]. Dado que el software utiliza las capacidades de estas tecnologías, tendrá una buena escalabilidad que le permitirá gestionar grandes cantidades de datos de accidentes. Hadoop, Mahout y MySQL residen en el mismo servidor y bajo el mismo sistema operativo CentOS (distribución Linux). El software se conectará a CentOS mediante el protocolo de red SSH para usar Hadoop, Mahout y MySQL. La aplicación se ha desarrollado como multihilo, donde un hilo se encarga de gestionar la interfaz de usuario y recibir las entradas del usuario, mientras que el otro se encarga de la comunicación entre...

El usuario (cliente) y el servidor, de modo que, si la aplicación espera los resultados del servidor, no se bloqueará (a diferencia de un solo subproceso) y permanecerá activa en todo momento. Además, el software informará al usuario del progreso de la operación en ejecución y le permitirá cancelarla en cualquier momento.
Esto es muy importante si la operación tarda minutos u horas en completarse debido al gran tamaño de los datos. Gracias a la arquitectura multihilo del software, tendrá una buena disponibilidad.

## II. ESTUDIOS DE INVESTIGACIÓN ANTERIORES

En el artículo de Chong [1], se afirma que la aplicación de técnicas de minería de datos para modelar registros de accidentes de tráfico puede ayudar a comprender las características del comportamiento de los conductores, el estado de la carretera y las condiciones meteorológicas que se relacionan causalmente con la diferente gravedad de las lesiones. Al mismo tiempo, ayudará a los responsables de la toma de decisiones a formular mejores políticas de control de la seguridad vial. El autor analizó los datos de accidentes automovilísticos del GES de 1995 a 2000 e investigó el rendimiento de redes neuronales, árboles de decisión, máquinas de vectores de soporte y un árbol de decisión híbrido con técnicas de redes neuronales para la predicción.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.187

Gravedad de las lesiones de los conductores en accidentes de tráfico
desarrolló un conjunto de experimentos con un predictor
clasifica las lesiones fatales y no fatales. El único d
estudio fue que el conjunto de datos utilizado no proporcionó suficiente
velocidad real y otra información como
Yang et al [2] utilizaron el enfoque de red neuronal
de conducción más seguros que tienen menos posibilidades de accidente
Lesiones en un accidente de tráfico. En el artículo
ilustró cómo los métodos estadísticos basados en
sobre datos para el período reciente, m
modelar las muertes por accidentes de tráfico comparando el mod
utilizando gráficos dirigidos a un modelo, basado en
de muestra. En el artículo de investigación de Rui [4], la a
Los factores de la carretera que juegan un papel vital en los accidentes
linealidad, la pendiente, la combinación lineal y la superficie de la carretera.
También mencionó los factores del vehículo, como
técnico de los vehículos, el factor motor, la dirección, el frenado,
factores eléctricos. La au
de datos de la base de datos de accidentes de tráfico de
analizó los datos de accidentes de tráfico utilizando
estadísticos que respaldan el método de investigación sobre
de investigación de Shi [5]
comprender la tendencia del flujo de tráfico en las carreteras es fundamental.
colisiones es importante para reducir el impacto de los accidentes de tráfico.
propuso un método para construir series de tiempo d
datos de flujo de tráfico cuando se produjeron accidentes. Para evitar el defecto de no
considerar la deriva lineal en el dominio del tiempo.
se realizó DFT para extraer características
temporales originales. La tendencia del flujo de tráfico podría
bien mediante el análisis de agrupamiento. El caso s
Harbin con datos reales demostró la viabilidad del proyecto.
conjuntos de datos reales extraídos de colecciones de registros
Beijing-Harbin (G1) entre Harbin y
2010 a julio de 2011. Datos de flujo de tráfico transformados en
temporales mediante el Modo de Transmisión Celular
utilizó la Transformada de Fourier Discreta (DFT) para extraer las características más
importantes, luego aplicar la agrupación en clústeres.
comprender la tendencia del flujo de tráfico en las carreteras.
minería de información de tráfico es una práctica típica
Data" con datos de tráfico superiores a 1,5 PB,
almacenamiento supera los 5 PB, y se producen aproximadamente
día en Pekín. El autor afirmó que el tráfico...
utilizará datos históricos de tráfico y descubrirá transacciones.
mientras se aplican reglas de asociación comunes a través de
minería de datos, el autor desarrolló funciones paralelas.
datos de tráfico distribuidos y bloqueados basados en el marco
contienen detección de accidentes y
tendencias de tráfico. Los servicios en la nube desarrollados son...
capas para soportar almacenamiento distribuido, paralelo
computación y servicios personalizados para datos de tráfico masivos
conjuntos de datos de tráfico masivos, incluyendo floa
automóviles, teléfonos móviles y autobuses. La minería de datos de tráfico desarrollada
se basó en la técnica de computación en la nube y se construyó.

nts. El autor
modelo clasificado que
El inconveniente de este
información sobre la
estado de la carretera.
ch para detectar patrones
causando muerte y e
por Roh [3], dirigió
gráficos, construidos
Puede ser útil para
ds especificado y
sobre pronósticos fuera
El autor afirmó que
condiciones climáticas.
s la calidad del estado
la conducción y los
Hor extrajo 100 conjuntos
ciudad de fa, y luego
Análisis de datos
, Afirmó que
La prevención de
abolladuras. El artículo
Datos que utilizan
uras de datos de series
Entonces se entendería
Un estudio realizado en
Método propuesto sobre
Se realizó la autopista
y Lalinhe de enero de
el método (CTM) y se
Método de análisis para
Yu [6] cree que la
Aplicación cal "Big
y la capacidad de
800 GB de datos cada
La minería de datos
Modo de transporte
Mediante técnicas de
ciones que procesan
Map-Reduce, que
y predicción de
sed en un RTIC-C (4
El sistema de
Conjunto de datos de
t con un masivo

Base de almacenamiento de datos de tráfico en Hadoop
para diversos tipos de m
mecanismo Map-Reduce y un resto
soportar aplicaciones de minería de terceros

III. ARCO PROPUESTO ARQUITECTURA

La solicitud propuesta será
(MySQL), y en caso de que
técnica de minería de datos, extraerá los datos
y los pasará a la siguiente base de datos.
Luego se utiliza para analizar los datos extraídos
Lo que se ha utilizado es el conjunto de datos
[11], que se utiliza para simular la operación.
seleccionados del conjunto de datos.

op con un marco de computación paralela
Aplicaciones mineras basadas en el
Interfaz de servicios web tful para
iones.

Trabajaré en la base de datos operacional
Se solicita a la aplicación que aplique la
requeridos de la base de datos operativa
Para Hadoop, Mahout es un dato. La
base de datos operativa...
de colisiones de tráfico de Nueva York
Base de datos operativa. Atributos

| Nombre del atributo Atrib | Descripción de ute |
|---|---|
| Hora La hora del accidente | nt en (24 horas) |
| Municipio La división administrativa | Latitud |
| de la isión La latitud de la acci | dent |
| Longitud de la longitud de la ac | |
| accidente On_Street_Name La calle del accidente | ent |
| Personas_Lesionadas Número total de personas | y herido |
| Personas_Muertos Número total de personas | Causa |
| del asesino Causas de ese accidente | ___ |
| Vehicle_Types Tipos de vehículos para los que se utiliza un | Accidente puede contener uno o varios |
| tipos de vehículos. | vehículos separados por coma |

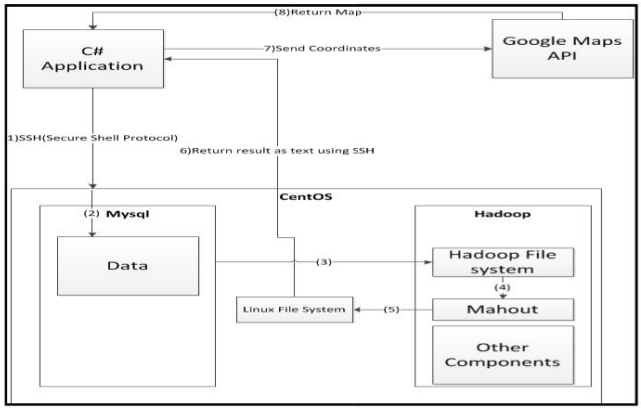


Figura 1: Arquitectura del p

Desafortunadamente, el conjunto de datos en
redundantes, y no es no
datos a MySQL, los registros...
los atributos importantes que no se pueden corregir. Los atributos que describen lo
mismo se agrupan en un solo atributo y sus

Tabla 1 para ver los atributos que se utilizan.
utilizados no se incluyen en la tabla). Véase
proyecto.

Solicitud propuesta.
incluye muchos valores faltantes y
Normalizado. Después de cargar los
Se eliminarán los valores faltantes de
valores se separan.
Separado por una coma. Consulte la
En el análisis de datos (los atributos no
también la Fig. 1 para la arquitectura del

IV. SOLICITUD PROPUESTA OPERACIÓN

Cuando el usuario ejecuta la aplicación
al servidor usando SSH y un

, La aplicación se conectará con la
sesión establecida.

CentOS. La Fig. 2 muestra la interfaz principal de la aplicación. Esta ofrece seis funciones de análisis diferentes. Los resultados del análisis se convierten en vista tabular, vista gráfica y vista de mapa (según la función). La vista tabular muestra el resultado como una tabla. La vista gráfica muestra el resultado como un gráfico.

El eje x contiene los valores comunes y el eje y la frecuencia de estos valores. La vista de mapa extrae las coordenadas GPS de los resultados y las representa en un mapa (mediante la API de Google Maps). El usuario puede analizar los datos de tráfico de una división administrativa específica (municipio) o de las divisiones administrativas en su conjunto.

Discutiremos estas funciones con mucho detalle.

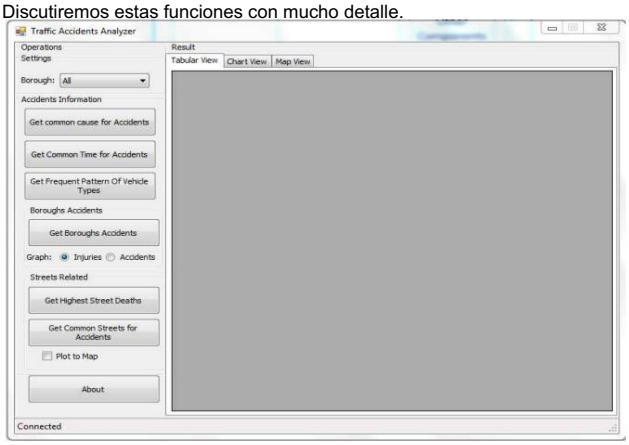


Figura 2: Interfaz principal de la aplicación propuesta.

Con base en los estudios previos, la aplicación propuesta puede analizar los siguientes tipos de datos: información de accidentes, accidentes de tránsito y relacionados con la vía. Cada dirección constará de varias funciones, cada una de las cuales realizará una tarea específica.

A. Causa común de accidentes Función

Esta función “Obtener causa común de accidentes” ejecutará la consulta SQL, que aplicará la función agregada COUNT en MySQL para obtener las causas comunes de los accidentes y ordenar los resultados en orden descendente según el número de accidentes.

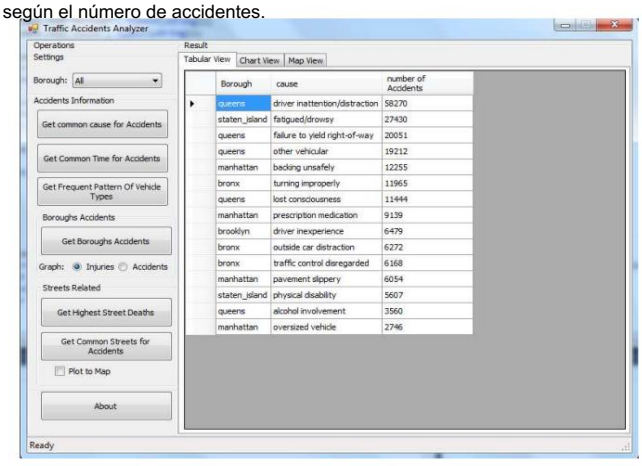


Figura 3: Vista tabular de “Obtener causa común de accidentes”

El resultado final se formateará en una tabla que tiene tres columnas: Municipio, Causa y Número de Accidentes como se muestra en la Fig. 3. Al mismo tiempo, la Fig. 4 muestra la distribución de las principales causas de accidentes para el período de tiempo aguanieve.

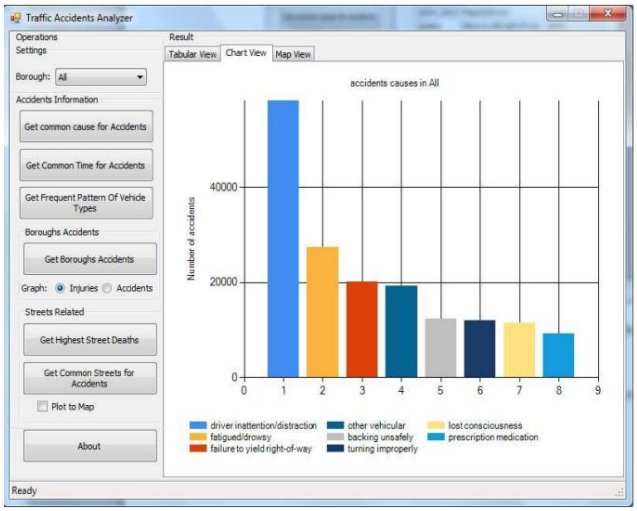


Figura 4: Distribución gráfica de causa común de accidentes.

B. Función de tiempo común para accidentes

El objetivo de esta función es obtener la hora común de accidentes. Ejecutará la consulta SQL que aplicará la función de agregación COUNT de MySQL para obtener la hora común de accidentes y ordenar los resultados en orden descendente según la hora de los accidentes, como se muestra en la Fig. 5.


Figura 5: Distribución gráfica para tiempo común de accidentes.

C. Función del tipo de vehículo que participa con frecuencia en accidentes

Esta función aplica la técnica de minería de datos FPgrowth, utilizando Mahout para extraer los patrones frecuentes de los tipos de vehículos.

que aparecen en accidentes y luego ordena el resultado en orden descendente. El resultado de esta función es una tabla con dos atributos: Patrones de tipos de vehículos y Accidentes (número de accidentes). El algoritmo para la función designada funcionará de la siguiente manera: 1. Eliminar los archivos antiguos de resultados de Mahout de Hadoop.

2. Elimine los datos de entrada antiguos y los archivos de resultados de CentOS.
3. Elimine los datos de entrada antiguos y los archivos de resultados de Hadoop.
4. Ejecute una consulta SQL para contar la cantidad de registros en la base de datos MySQL (para calcular el soporte para el paso 8).
5. Ejecute la consulta SQL para exportar los valores del atributo 'vehicle_types' sólo en un archivo CSV.
6. Ejecute el comando “sed” de Linux para eliminar las comillas del Archivo CSV.
7. Cargue el archivo en Hadoop. Consulte la Figura 7. 8. Aplique FPgrowth usando Mahout con el método "mapreduce" y un soporte del 1% (este paso puede tardar unos minutos).
9. Descargue los resultados del sistema de archivos Hadoop al archivo CentOS sistema.
10. Utilice el comando “cat” de Linux para mostrar los resultados.
11. Analice el resultado en vista tabular y vista de gráfico.

Los resultados brutos de Mahout se muestran en la Fig. 6 y la representación gráfica de la función Obtener patrón frecuente de tipos de vehículos se presenta en la Fig. 7.

```
Input Path: patterns/frequentpatterns/part-r-00000
Key class: class org.apache.hadoop.io.Text Value Class: class org.apache.mahout.
Max Items to dump: 50
Key: SUV: Value: ([SUV],81192), ([passenger_vehicle, SUV],39750), ([SUV, unknown
Key: bicycle: Value: ([bicycle],5224), ([passenger_vehicle, bicycle],2515)
Key: bus: Value: ([bus],7712), ([passenger_vehicle, bus],3543)
Key: large_vehicle: Value: ([large_vehicle],8132), ([passenger_vehicle, large_ve
Key: livery_vehicle: Value: ([livery_vehicle],4981)
Key: other: Value: ([other],12045), ([passenger_vehicle, other],5383), ([SUV, ot
Key: passenger_vehicle: Value: ([passenger_vehicle],159087), ([passenger_vehicle
Key: pick-up_truck: Value: ([pick-up_truck],6314), ([passenger_vehicle, pick-up_
Key: small_vehicle: Value: ([small_vehicle],7605), ([passenger_vehicle, small_ve
Key: taxi: Value: ([taxi],16150), ([passenger_vehicle, taxi],6533), ([SUV, taxi]
Key: unknown: Value: ([unknown],29151), ([passenger_vehicle, unknown],16598), ([
Key: van: Value: ([van],14037), ([passenger_vehicle, van],6251), ([SUV, van],344
Count: 12
```

Figura 6: Resultados brutos de Mahout.

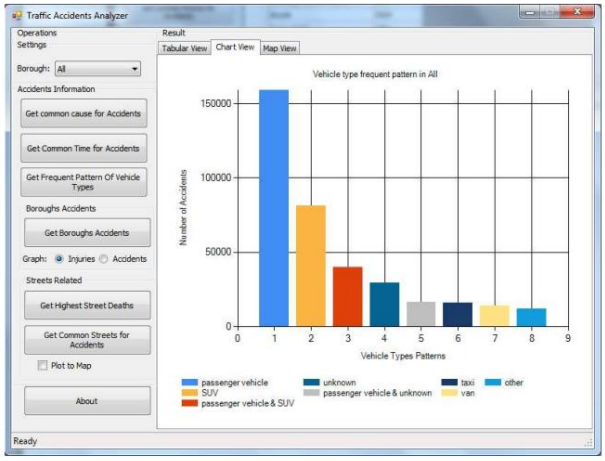


Figura 7: Representación gráfica del patrón de obtención frecuente del vehículo.

D. Función de Accidentes de Distritos

Esta función ejecuta la consulta SQL, que utiliza las funciones de agregación COUNT y SUM, para contar el número de accidentes en cada distrito. Posteriormente, suma el total de lesiones y muertes de dicho distrito. Ofrece dos opciones para crear los gráficos: una para lesiones y otra para muertes. Consulte la Figura 8 para más detalles.

E. Función "Obtener el mayor número de muertes en calles". Esta función ejecuta la consulta SQL que utiliza la función de agregación SUM. El resultado contiene los nombres de las calles y el total de muertes ocurridas en cada una, en orden descendente. Los resultados finales de esta función se pueden representar gráficamente en una vista de gráfico o de mapa, como se muestra en la Fig. 9.

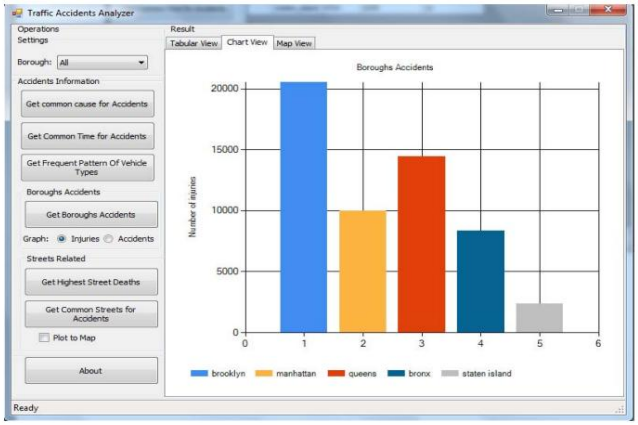


Figura 8: Vista de gráfico para la función Obtener accidentes de distritos.


Figura 9: Vista del mapa para la función Obtener la cifra más alta de muertes en las calles.

F. Función de Calles Comunes para Accidentes

Esta función ejecuta una consulta SQL, que utiliza la función de agregación COUNT, para contar el número de calles.

El resultado contiene los nombres de las calles y el número de accidentes ocurridos en cada calle, ordenados en orden descendente como se muestra en la Figura 10.

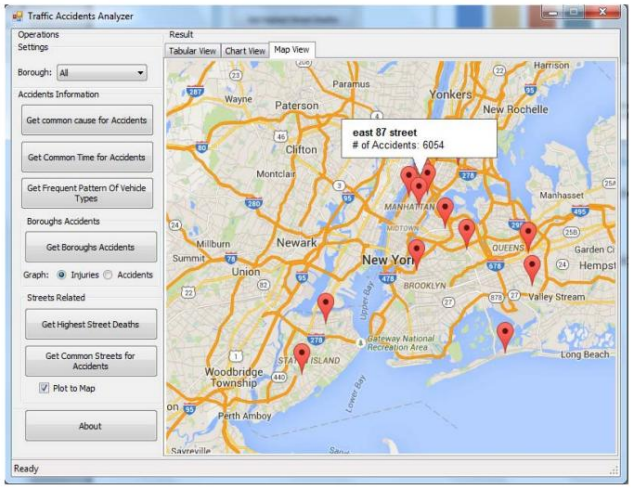


Figura 10: Vista del mapa para la función Obtener calles comunes para accidentes.

V. ANALIZADOR DE ACCIDENTES DE TRÁFICO COMO SERVICIO WEB

Se ha creado una versión mejorada de la aplicación propuesta como servicio web, que puede ser invocada por numerosos clientes, como PHP, ASP, teléfonos móviles o cualquier dispositivo o aplicación que pueda enviar solicitudes POST, o que utilice SOAP, un protocolo de comunicación con servicios web. Esta versión mejorada permite utilizar las capacidades de Hadoop y Mahout en línea y de forma remota. La figura 11 muestra la versión de la arquitectura mejorada para la aplicación propuesta basada en servicio.

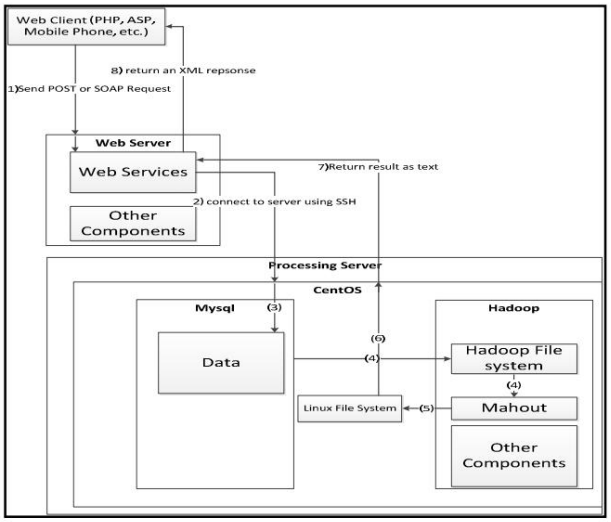


Figura 11: Arquitectura de aplicaciones basada en servicios.

La figura 12 representa una captura de pantalla de páginas web (HTML) que utilizan JavaScript (AJAX) para conectarse al servidor PHP, que a su vez se conecta al servicio web y al que se accede mediante una computadora y un iPhone.

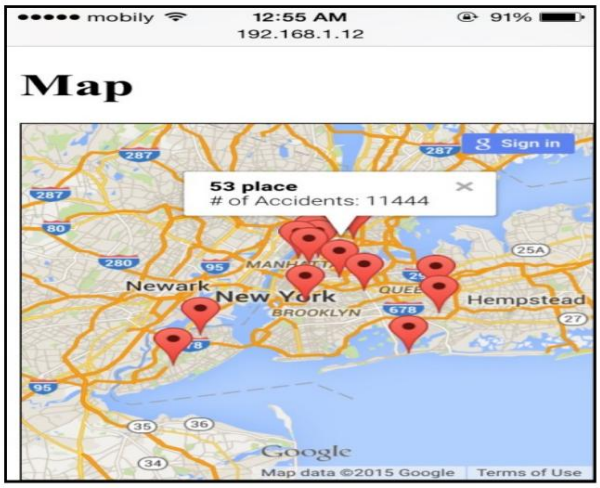


Figura 12: Vista del mapa para la función Obtener calles comunes para accidentes

VI. CONCLUSIÓN

Los accidentes de tráfico son problemas graves que pueden causar discapacidades, lesiones e incluso la muerte. Para reducir su número, es necesario comprenderlos y analizarlos.

Dado que los datos sobre accidentes de tráfico se generan casi a diario y su tamaño aumenta rápidamente en todas las dimensiones, se ha vuelto esencial la urgente necesidad de una aplicación que gestione eficazmente este crecimiento y analice los accidentes de tráfico. Hoy en día, el ecosistema de Big Data permite almacenar, manipular, analizar y extraer grandes conjuntos de datos de accidentes de tráfico e impulsar la creación de conocimiento que puede ayudar a los responsables de la toma de decisiones a reducir el número de accidentes. Algunos investigadores prestan mucha atención al análisis de grandes conjuntos de datos de accidentes de tráfico mediante el enfoque de Big Data. Este estudio presenta una herramienta de aplicación muy importante para el uso de Big Data en el almacenamiento, la integración y el análisis de accidentes de tráfico mediante Mahout Data Mining como parte del ecosistema de Big Data. La aplicación desarrollada utiliza conjuntos de datos de tráfico reales y de gran tamaño, provenientes de las colisiones de tráfico de Nueva York, como fuente de datos. Esta aplicación consta de varias funciones y servicios web para analizar y visualizar la información más importante sobre accidentes de tráfico. La aplicación almacena los datos de tráfico masivos en Hadoop con un marco de computación paralela para su procesamiento y extracción basado en la técnica Map-Reduce, y posteriormente utiliza una interfaz de servicios web para respaldar la aplicación de minería desarrollada.

REFERENCIAS

[1] Miao Chong, Ajith Abraham2 y Marcin Paprzycki1, “Análisis de accidentes de tráfico utilizando paradigmas de aprendizaje automático”, Informatica Journal Vol. 29, págs. 89–98, 2005.

[2] Yang, WT, Chen, HC y Brown, DB, Detección de patrones de conducción más seguros mediante un enfoque de redes neuronales. ANNIE '99 para las Actas de Diseño de Sistemas de Ingeniería Inteligente, Redes Neuronales, Programación Evolutiva, Sistemas Complejos y Minería de Datos, vol. 9, págs. 839-844, noviembre de 1999.

[3] Roh JW, Bessler DA y Gilbert RF, Muertes por accidentes de tránsito, modelo de Peltzman y gráficos dirigidos, Accident Analysis & Prevention, vol. 31, números 1-2, págs. 55-61, 1998.

[4] Rui Tian, Zhaosheng Yang y Maolei Zhang, “Método de análisis de causas de accidentes de tráfico basado en minería de datos”, Conferencia internacional sobre inteligencia computacional e ingeniería de software (CiSE) de 2010, págs. 1-4, 10-12 de diciembre de 2010, DOI:10.1109/CISE.2010.5677030, 2010.

[5] An Shi, Zhang Tao, Zhang Xinming y Wang Jian, “Evolución del análisis del flujo de tráfico en caso de accidentes en carreteras mediante minería de datos temporales”, Quinta Conferencia Internacional sobre Diseño de Sistemas Inteligentes y Aplicaciones de Ingeniería, 2014, DOI 10.1109/ISDEA.2014.109, págs. 454-457, 2014.

[6] Jianjun Yu, Fuchun Jiang y Tongyu Zhu, “RTIC-C: Un sistema de big data para la minería masiva de información de tráfico”, Conferencia internacional sobre computación en la nube y big data de 2013, DOI 10.1109/CLOUDCOM-ASIA.2013.9, págs. 395-402, 2014.

[7] http://en.wikipedia.org/wiki/Apache_Hadoop última visita julio de 2015. [8] http://en.wikipedia.org/wiki/Apache_Mahout, recuperado el 6 de mayo de 2015. [9] http://www.tesora.com/myth-4-mysql-cannot-handle-large-volumes-data-specially-queries-joins-and-aggregations-i-must/ recuperado el 10 de junio de 2015.

[10] https://sshnet.codeplex.com/ recuperado el 10 de mayo de 2015. [11] https://data.cityofnewyork.us/NYC-BigApps/NYPD-Motor-Vehicle-Collisions/h9gi-nx95. recuperado el 6 de julio de 2015.