Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR

Jonathan Obed García Osuna
Asesorado por el Ing. Lenin Fernando Rodríguez Conde

Guatemala, septiembre de 2016

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

JONATHAN OBED GARCÍA OSUNA
ASESORADO POR EL ING. LENIN FERNANDO RODRÍGUEZ CONDE

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, SEPTIEMBRE DE 2016

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. Pedro Antonio Aguilar Polanco
VOCAL I	Ing. Angel Roberto Sic García
VOCAL II	Ing. Pablo Christian de León Rodríguez
VOCAL III	Inga. Elvia Miriam Ruballos Samayoa
VOCAL IV	Br. Raúl Eduardo Ticún Córdova
VOCAL V	Br. Henry Fernando Duarte García
SECRETARIA	Ing. Lesbia Magalí Herrera López

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Angel Roberto Sic García
EXAMINADOR	Ing. César Augusto Fernández Cáceres
EXAMINADOR	Ing. José Alfredo González Díaz
EXAMINADOR	Ing. Roberto Estuardo Ruiz Cruz
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha septiembre de 2015.

Jonathan Obed García Osuna

Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ciencias y Sistemas

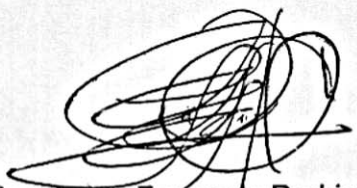
Guatemala 22 de Enero de 2016

Ingeniero
Carlos Alfredo Azurdia Morales
Coordinador del Área de Trabajos de Graduación

Respetable Ingeniero Azurdia:

Por medio de la presente me permito comunicar que el trabajo de graduación del estudiante **Jonathan Obed García Osuna** con número de carné **200915115**, cuyo título es **"UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR"** se ha completado en su totalidad, el cual he revisado y aprobado.

Atentamente.



Ing. Lenin Fernando Rodriguez Conde
Asesor de trabajo de graduación
Colegiado no. 10694

Ing. Lenin Fernando Rodríguez Conde
Col. 10694

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 3 de Febrero de 2016

Ingeniero
Marlon Antonio Pérez Türk
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Pérez:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **JONATHAN OBED GARCIA OSUNA** con carné **200915115**, titulado: **"UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR"**, y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

ESCUELA DE INGENIERÍA EN CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE INGENIERÍA EN
CIENCIAS Y SISTEMAS
TEL: 24188000 Ext. 1534

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación ***"UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR",*** *realizado por el estudiante, JONATHAN OBED GARCÍA OSUNA, aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DIRECCION DE INGENIERIA EN CIENCIAS Y SISTEMAS

Ing. Marlon Antonio Pérez Türk
Director
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 08 de septiembre de 2016

Universidad de San Carlos
de Guatemala

Facultad de Ingeniería
Decanato

DTG. 402.2016

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al Trabajo de Graduación titulado: **UTILIZACIÓN DE ANÁLISIS DE DATOS MASIVOS PARA LA REDUCCIÓN DE INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR,** presentado por el estudiante universitario: **Jonathan Obed García Osuna,** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Ing. Pedro Antonio Aguilar Polanco
Decano

USAC DECANO FACULTAD DE INGENIERÍA

Guatemala, septiembre de 2016

/gdech

ACTO QUE DEDICO A:

Dios	Por haberme alcanzado con su infinito amor y gracia, ayudándome a hacer de mí una mejor persona cada día.
Mis padres	Gonzalo García y Leticia Osuna, por ser un ejemplo en mi vida, enseñándome a luchar por alcanzar mis metas, a quienes agradezco por su confianza, paciencia y amor.
Mis hermanos	Madelaine, Keren y Jason, por haberme brindado su apoyo, amistad y cariño.
Mi abuela	Marciala Morales, por ser un ejemplo de dedicación y brindarme su incondicional amor.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	En especial a la Facultad de Ingeniería, por transmitirme valiosas enseñanzas personales y académicas, que sin duda han forjado mi futuro como profesional.
Mis amigos de la Facultad	Por brindarme de su sincera amistad y apoyo, siendo un gran ejemplo de esfuerzo a lo largo de toda la carrera.
Ing. Lenin Rodríguez	Por su importante e invaluable ayuda, ya que su paciencia y experiencia contribuyó en la realización de este trabajo de graduación.
Mi familia	Tíos, tías, primos y primas, me han brindado su apoyo durante todo este tiempo.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
ECC	Tipo de memoria RAM utilizada para la corrección y recuperación de errores.
GB	Gigabytes, compuestos por 1024 MB.
GHz	Utilizado como la frecuencia para indicar la velocidad de una unidad central de procesamiento.
GPS	Sistema de posicionamiento global, por sus siglas en inglés, que permite conocer la localización de un objeto.
HDFS	Sistema distribuido de archivos de Hadoop, por sus siglas en inglés.
KB	Kilobytes, compuestos por 1024 bytes
MB	Megabytes, compuestos por 1024 KB
NoSQL	Son sistemas de bases de datos que brindan de características distintas a las bases de datos relacionales.

RAM	Memoria de acceso aleatorio, por sus siglas en inglés, constituyendo la memoria principal y volátil de una computadora.
SQL	Lenguaje estructurado de consultas, por sus siglas en inglés, utilizado dentro de las bases de datos relacionales.
TB	Terabytes, compuestos por 1024 GB.

GLOSARIO

Actor	Toda persona, objeto o dispositivo que mantiene interacción dentro del sistema.
Alta disponibilidad	Característica de un sistema que se asocia con la seguridad de mantener en forma continua sus operaciones durante un periodo de tiempo determinado.
Apache Foundation	Organización no lucrativa que brinda soporte para proyectos de software de código abierto.
Arq. Maestro-Esclavo	Arquitectura que se caracteriza por una jerarquía centralizada, donde la mayor carga como centro control recae sobre el maestro mientras el esclavo es el encargado de realizar el procesamiento.
Arreglo	Estructura de datos que permite almacenar datos de forma secuencial.
Bases de datos	Conjunto de datos almacenados de forma persistente y organizada con la intención de ser consultada y utilizada de forma sencilla.
Bit	Unidad mínima de información, tomando únicamente los valores de cero o uno.

Byte	Conjunto de ocho bits, denominado como la unidad fundamental de información digital.
Conocimiento	Toda la información significativa que es extraída mediante el procesamiento y análisis de los datos.
Cromosoma	Una estructura informática que posee datos que simbolizan la eficiencia del medio que representan.
Escalabilidad	Capacidad de aumento o disminución de su amplitud en función del trabajo que realiza, sin afectar el sistema por completo.
Gigabit Ethernet	Ampliación del estándar Ethernet que alcanza la transmisión de un gigabit por segundo.
Hardware	Comprende los elementos físicos o tangibles de una computadora.
Horas pico	Se define como las horas, en las cuales se alcanza un volumen vehicular máximo.
Java	Lenguaje de programación de alto nivel orientado a objetos, creado en 1995.

Log	Archivo encargado de recolectar de forma continua diversos eventos que son registrados por una computadora o aplicación, tales como, estados de proceso, mensajes, errores, entre otros.
Modelo relacional	Modelo de datos basado en lógica de primer orden y en la teoría de conjuntos, el cual organiza y representa los datos por medio de relaciones.
Nodo	Denominado a cada computadora dentro de un sistema de computación distribuida.
Selección natural	Mecanismo capaz de relacionar los cromosomas y brindar de mayores probabilidades de reproducirse a los individuos más aptos.
Sistema de computación distribuido	Es el que por medio de un conjunto de computadoras interconectas realizan su trabajo de forma paralela.
Software	Constituido por programas, rutinas, procesos o servicios dentro de una computadora.

Software libre	Una denominación de software que respeta la libertad de los usuarios y la comunidad, el cual permite ejecutar, copiar, distribuir, estudiar, modificar y mejorar el software.
Tendencias	Correlaciones buscadas dentro de grandes cantidades de información para comprender el comportamiento dentro del conjunto de datos.
Tiempo real	Es el que interactúa dinámicamente con su entorno capaz de capturar o procesar datos a medida que se van generando.
Volumen vehicular	Cantidad de vehículos que transitan por una vía.

RESUMEN

Como punto de partida se describe cuáles son los conceptos fundamentales dentro del análisis de datos masivos, tratando su utilidad e importancia, iniciando desde la recolección, el tratamiento necesario para su procesamiento; detallando finalmente las múltiples técnicas de análisis que pueden ser utilizadas dentro de esta tecnología.

Realizando una descripción de cada uno de los componentes que se encuentran relacionados en el tránsito vehicular, se busca establecer un conocimiento concreto de cómo cada uno de estos contribuye y coexisten para el funcionamiento de este, haciendo mención de cómo pueden ser acoplados al análisis de datos masivos.

Por medio de la arquitectura propuesta se busca solventar la necesidad principal: la disminución de intervención humana dentro del tránsito vehicular; esto conlleva a realizar un proceso completo de análisis de datos masivos, el cual es considerado etapa por etapa, describiendo el proceso que debe ser realizado dentro de cada una de ellas, desde la recolección de datos hasta la visualización de los resultados encontrados.

Para finalizar, se analizará la factibilidad de la implementación de esta tecnología desde tres puntos: factibilidad técnica, operativa y económica, las cuales en conjunto brindarán de la capacidad para considerar la viabilidad de aplicación y utilización dela propuesta.

OBJETIVOS

General

Estudiar la factibilidad de la utilización de análisis de datos masivos para la extracción, manejo y procesamiento de información relacionada al tránsito vehicular, logrando identificar tendencias que permitan la reducción de la intervención humana.

Específicos

1. Exponer la efectividad del análisis de datos masivos y amplia utilización para optimizar la toma de decisiones.

2. Presentar los componentes que interactúan en la intervención humana con el tránsito vehicular.

3. Investigar modelos de predicción que permitan extraer de grandes cantidades de datos: recursos que ayuden en la toma de decisiones.

4. Analizar las alternativas para el almacenamiento, manejo y procesamiento de datos masivos.

5. Proponer una solución, que mediante el análisis de datos masivos, contribuya a la reducción de intervención humana en el tránsito vehicular.

INTRODUCCIÓN

Durante los últimos años, la utilidad y efectividad del análisis de datos masivos ha demostrado su importancia en la toma de decisiones basadas en correlaciones estadísticas que pueden ser extraídas de las grandes cantidades de información.

El uso de datos masivos no solo puede ser utilizado en el ámbito de los negocios, sino también haciendo uso del conocimiento extraído por el análisis a una gran fuente de datos que puede ser empleado para la utilidad dentro de la sociedad.

A menudo es cuestionada la eficiencia en la toma de decisiones dentro de la señalización del tránsito vehicular por medio de la intervención humana, suponiendo que, estas detienen la fluidez y entorpecen la facilidad para transitar dentro de un flujo vehicular. Por lo que se propone, por medio del análisis de los datos generados por el tránsito vehicular, poseer una arquitectura capaz de sostener una solución que cumpla con los requerimientos específicos de recolección, almacenamiento, procesamiento y análisis. Con la finalidad de contribuir a la reducción de la intervención humana en el tránsito vehicular y aumentar la eficiencia en la toma de decisiones, por medio del análisis de datos masivos.

1. ANÁLISIS DE DATOS MASIVOS

1.1. Datos masivos

Cada día, una gran cantidad de datos es generada, no solo de fuentes convencionales sino de aquellos que se generan en grandes cantidades y de manera muy rápida, datos que provienen de sensores de flujo de tránsito vehicular, contenidos de redes sociales, señales de GPS, transacciones bancarias, cámaras de seguridad, entre otros.

Las nuevas fuentes de datos han habilitado nuevas oportunidades para los negocios, pudiendo utilizar los grandes bancos de datos que ya poseen y lograr una utilidad por medio de técnicas de análisis de datos enfocadas a cada área.

Gracias a los grandes avances de la tecnología, actualmente se posee la capacidad de capturar, almacenar y procesar información de una manera mucho más rápida y con menor costo, lo que antes era relegado a grandes compañías con la capacidad de sufragar el costo de este tipo de análisis.

Con la creciente utilización del análisis de datos masivos se incursiona en un cambio de enfoque de búsqueda de la causalidad y enfocar solamente las correlaciones entre datos.

El uso de datos masivos representa un gran cambio de enfoque en la toma de decisiones, sin embargo, la utilización no está restringida para ser aplicada en compañías en búsqueda de ganancias, sino presenta un nuevo enfoque a la resolución de problemas.

1.1.1. Características

Para adoptar el análisis de datos masivos se debe poseer pleno conocimiento de los aspectos principales para procesar datos a gran escala, esto con el fin de organizarlos de tal manera, que generen un significado en función de las operaciones del negocio.

1.1.1.1. Volumen

La eficiencia en las redes de comunicación, como también el crecimiento en el tránsito de información ha aumentado considerablemente, por lo que los beneficios de poseer grandes cantidades de información son la principal característica para el análisis de datos masivos.

La relación proporcional entre la cantidad de información y la efectividad de predicción o solución de problemas es mejor, esto presenta nuevos retos a todas las estructuras convencionales de almacenamiento y proceso de información.

Estos grandes volúmenes han logrado que el enfoque en que habitualmente se manejaba esta información cambie, pensando en nuevas formas de almacenamiento escalable y optimizar la manera de como alcanzar la información de forma distribuida.

1.1.1.2. Velocidad

Dentro de una sociedad que busca tiempos reducidos de respuesta y un servicio casi instantáneo, la velocidad en cómo procesar esta información es vital, tanto para generar una respuesta como para almacenarla.

La importancia de la velocidad de proceso dentro de grandes cantidades de información proviene cuando el tiempo de respuesta debe ser mínimo en función de tomar una decisión en segundos, las cuales afecten detectando fraudes en transacciones bancarias, recomendaciones a clientes y hasta juegos en línea.

Al poseer un patrón de crecimiento parecido al del volumen, el flujo de datos sobreviene un aspecto más a considerar, cuando los datos de entrada son demasiado rápidos para almacenarse completamente, el análisis se debe realizar sobre el propio flujo de entrada.

Cuando el flujo de datos presenta una cantidad aplastante de información, se opta por descartar mucha de esa información, esperando que, lo que fue descartado no posea datos útiles para la toma de decisiones.

1.1.1.3. Variedad

Debido a la inmensa variedad de cómo se genera actualmente la información, esta es una de las características más importantes a tomar en cuenta para el análisis masivo de datos, tomando en cuenta que dicha información puede ser, desde un archivo de texto hasta un video dentro de una red social.

Cuando la información generada posee intervención humana siempre será propensa a contener errores e inconsistencias, sin embargo, aun tratándose de comunicación entre computadoras, la información puede volverse confusa y aumentar la dificultad de su integración.

Uno de los procesos más importantes dentro del análisis de datos masivos, es tomar esta variada información y realizar extracciones que tengan un significado para su interpretación por humanos o como una entrada estructurada para una aplicación.

El proceso de reestructuración de la información genera pérdida de datos, sin embargo, como uno de los principios de este tipo de análisis es mantener tanta información como se pueda.

1.1.2. Tipos de datos masivos

Para aumentar la comprensión y definición del análisis de datos masivos, es pertinente comprender la variabilidad en estructura que comprende las grandes cantidades de información.

1.1.2.1. Datos no estructurados

Se caracterizan por provenir de fuentes distintas y no tener una estructura estandarizada con la que sea posible almacenarse, como tradicionalmente se hace.

Este tipo de información usualmente posee interacción humana, lo que dificulta la forma de cómo ser interpretada y procesada, uno de los mejores ejemplos de este tipo de datos es la que se genera dentro de las redes sociales.

1.1.2.2. Datos estructurados

Son todos aquellos que provienen de fuentes tradicionales, tales como: transacciones bancarias, registros de compras entre otras, estos tipos de datos también representan datos masivos.

Aunque puede verse disminuido su valor por el abrumador volumen de toda la información que es generada, representa un gran activo para una organización.

1.1.2.3. Datos multiestructurados

Este tipo de datos presenta gran variedad en tipo de formatos, el cual puede ser generado por la interacción entre máquinas y humanos, dando como resultado un conjunto de información con una multiestructura.

Uno de los ejemplos que mejor ilustran este tipo de estructura es la colección de actividades de un humano dentro de un sitio web, la cual puede incluir información de texto, imágenes, tiempo de estadía en páginas acompañados de transacciones de compras.

1.1.3. Aplicaciones

La aplicación de análisis de datos masivos se puede considerar en un vasto rango de posibilidades, desde razones puramente humanitarias, como detener un brote de una posible epidemia hasta motivos económicos, buscando mejorar ganancias dentro de cualquier tipo de comercio.

Los datos masivos han sido utilizados para la lucha contra el crimen, donde por medio de este tipo análisis puede predecir la probabilidad que un criminal reincida y vuelva a la cárcel.

La personalización de ofertas a clientes, permitiendo aumentar la retención de clientes enviando ofertas especiales, específicamente a sus necesidades antes que decidieran abandonar un servicio.

Por medio de búsquedas, brindar un indicador de brotes de resfriado, pudiendo predecir la actividad de un virus basándose en la predicción mediante un modelo lineal.

1.2. Tratamiento de datos

Este proceso representa una parte fundamental dentro del análisis de datos masivos, el gran volumen de los datos impone características específicas para la recolección de datos.

1.2.1. Métodos de recolección de datos

La importancia de cómo se deben recolectar los datos para este tipo de análisis, recae en la metodología utilizada, una estrategia que sea capaz llevar a un análisis integral del foco de investigación y alcanzar por medio de un conjunto de reglas y procedimientos la efectividad de la información recolectada.

Un aspecto importante dentro del análisis de datos masivos es que no discrimina la procedencia de datos según su fuente, ya sea una fuente primaria que brinda interacción con un humano, o bien, una fuente secundaria como registros históricos.

1.2.1.1. Observación

Uno de los métodos más utilizados para la recolección de datos es la observación, brinda la comodidad de realizar un registro visual de los eventos pertinentes a las condiciones fijadas para el objetivo de estudio.

La recolección de datos por medio de este método permite recolectar información tanto cualitativa como cuantitativa, sin embargo, este debe poseer una exhaustiva planificación y así, asegurar la confiabilidad y veracidad de los datos recolectados.

- Características
    - Debido a la intervención humana puede presentar errores e inconsistencias dentro de los datos que fueron recolectados, lo que conlleva a realizar recopilaciones más estrictas y desechar muchos datos por considerarse no útiles.
    - Por medio de este método es posible realizar la recolección de los datos de manera no intrusiva, es decir, que el sujeto de investigación no es consciente que la información relacionada está relacionada a él.

- La observación es una de las técnicas de investigación y recopilación de datos más aceptadas por todo el ámbito científico, para describir comportamientos individuales y en masa.

1.2.1.2. Instrumento de medición

Buscando la recopilación de datos más certera, se utilizan instrumentos que permiten no solo minimizar los errores en la intervención humana, sino agilizar la recopilación de datos de una forma más estandarizada.

Dichos instrumentos son herramientas para un investigador que se pueden presentar desde formularios o listas de opciones hasta sensores que recogen información para el cambio climático.

- Características

  - Este método permite recopilar datos de una forma más estandarizada, logrando crear una estructura y así facilitar el análisis sobre grandes cantidades de datos sin incurrir en una previa extracción o limpieza de datos.

  - El costo de conservar estos instrumentos en funcionamiento durante largos periodos de tiempo puede ser alto, no solo de funcionamiento sino también de mantenimiento.

  - Por medio de estos métodos muchos datos son recolectados, sin poseer interacción humana, usualmente sensores que recolectan grandes cantidades de datos y realizar toma de decisiones con base en ellos.

1.2.1.3. *Crowdsourcing*

Un término que en los últimos años ha tomado mucha fuerza, que no es más que la externalización de actividades a multitudes que cooperan para alcanzar un objetivo en conjunto.

Este tipo de recolección presenta un nuevo enfoque a los que tradicionalmente se tenían, presenta una nueva alternativa a la captura de datos por medio de multitudes que activamente brindan información que puede ser capturada y ser procesada.

- Características

    - Una de las características importantes de este tipo de recolección de datos es la completa intervención humana, lo que genera datos con cualquier tipo de estructura pudiendo dificultar la extracción del significado para la investigación.

    - Al compartir información se puede incurrir en atravesar la barrera de la privacidad, dando como resultado la sensación de espionaje o que los datos proporcionados pueden ser vendidos y obtener ganancias con base en ellos.

    - Este tipo de recolección necesita de una herramienta que sea capaz de recibir la información que sea proporcionada, y transmitirla de manera adecuada al lugar donde pueda ser procesada apropiadamente.

1.2.2. Procesamiento de datos

Tras finalizar el proceso de recolección de datos, hay aspectos que deben ser identificados para tener datos que representen información confiable que se pueda analizar y obtener conocimiento significativo de ella. Entre estos se encuentran los que se describen a continuación:

1.2.2.1. Frecuencia de recolección

La recolección de datos conlleva tener planes de la frecuencia para ser recogidos, la cual será representada por la utilización que se le dará y la importancia de poseer datos actuales y realizar el proceso de análisis.

Una de las variables más importantes para la recolección de datos es el costo que se genera durante los periodos de recolección, esto implica poseer planes adecuados para realizarlo y minimizar todo tipo de costes en los que se incurre para realizar este proceso, sin embargo, todos estos datos deben ser lo más precisos y abundantes que sea posible.

Las frecuencias pueden resultar muy variadas, dependiendo del sujeto de investigación, pudiendo ser capturados de forma diaria, como bitácora de actividades, hasta recopilar datos de forma anual, como los utilizados en estudios demográficos, o bien de forma variable, sin poseer un estándar de tiempo en una y otra recolección.

1.2.2.2. Cantidad y calidad de los datos

La calidad en la información recolectada es de suma importancia para procesarla y obtener resultados certeros, lo que conlleva un aumento en la utilización de recursos, incrementando los costos del análisis.

Dentro de los métodos tradicionales para el análisis de la información, todos los datos recolectados deben ser estrictamente adquiridos y lograr obtener una muestra confiable y veraz, lo que reduce el número de la muestra eliminando la aleatoriedad dentro de la información recabada.

Para el efecto de los datos masivos, más datos es mejor, debido a las grandes cantidades de información que se obtiene, la calidad de esta es muy variada, sin embargo, por el nivel masivo de información esta permite detectar correlaciones que aumentan la probabilidad de precisión al tomar decisiones.

En contraste con los métodos tradicionales, los datos masivos no permiten hacer exclusión dentro de la cantidad de información que se recolecta, sin importar el nivel de calidad que esta tenga, debido a que el enfoque principal en la utilización de recursos se focaliza dentro del procesamiento de los datos, mas no en su recolección como objetivo principal.

1.3. Modelos predicción

El conocimiento dentro de la información recolectada debe ser buscado por metodologías que permitan alcanzar el mayor provecho de esta, para el proceso de análisis a grandes volúmenes de información de encuentran los que se describen a continuación:

1.3.1. Análisis de regresión

Son utilizados para modelar la relación entre dos variables, logrando realizar predicciones sobre el comportamiento de estas variables, dichos modelos son creados con base en las ecuaciones matemáticas que describen la conducta entre las variables de estudio.

Se define como variable dependiente aquella que su valor depende o es modificado por otra, y variable independiente a la que no necesita de ninguna otra interacción para describir su valor.

Los modelos matemáticos buscan describir y determinar por qué la variable dependiente es alterada al realizarse un cambio dentro de la variable independiente, esto implica que solo se utilizarán los datos que contribuyen a obtener los resultados deseados.

Existen dos tipos de modelos de regresión, en los que la diferencia principal está en la variable independiente: regresión simple que cuenta con una sola variable independiente y regresión múltiple que posee más de una variable independiente.

1.3.2. Análisis de correlación

Al igual que el análisis de regresión, este busca la relación entre dos variables, no obstante, se enfoca en la fuerza o grado con el que las variables están asociadas una con la otra basándose, principalmente en inferencias estadísticas.

El análisis de correlación no intenta explicar el porqué de dicha asociación, sino con el que es suficiente, ya que no se busca la causalidad sino se enfoca en cómo el comportamiento de la variable dependiente puede ser descrito por los cambios dentro de la variable independiente.

Las correlaciones son buenas cuando no se cuenta una gran cantidad de datos, pero su verdadera utilidad es demostrada al contar con grandes cantidades de datos de donde extraer las asociaciones.

Las correlaciones pueden dividirse en fuertes, quiere decir, que cuando la variable independiente cambie es altamente probable que la variable dependiente sufra un cambio, y una correlación débil indica lo contrario, si la variable independiente cambia es poco probable que se vea reflejado en la dependiente.

1.4. Técnicas para el análisis de datos masivos

Con el crecimiento de los datos se ha creado la necesidad de innovar las técnicas de análisis que ya se poseían con anterioridad, logrando una unión entre el estudio de la estadística por medio del uso de probabilidades y la importancia de correlaciones y el poder de la informática para procesar grandes cantidades de datos y obtener un significado de ellos.

1.4.1. Minería de datos

Es una técnica que permite extraer información de grandes volúmenes de datos, la cual permite detectar tendencias o relaciones que no pueden ser detectadas de forma tradicional, ya sea por la complejidad de los datos o por su gran volumen.

Utilizando como materia prima toda la información recolectada, que al ser tratados confieren un conocimiento que puede ser extraído por medio de modelos estadísticos, para ser utilizados en la optimización de toma de decisiones.

1.4.1.1. Etapas de la minería de datos

Como cualquier metodología, esta conlleva una serie de etapas que usualmente son utilizadas para alcanzar el significado de los datos que serán analizados.

1.4.1.1.1. Definir el problema

Como primer paso es importante entender qué es lo que se busca solucionar por medio de esta técnica de análisis, definir un ámbito y todas las métricas que servirán como guía para el estudio de los datos que se evaluarán.

1.4.1.1.2. Preparación de los datos

Dependiendo de la fuente de donde los datos hayan sido obtenidos dentro de esta etapa se procederá a realizar la limpieza de datos, realizando acciones como: eliminación de datos que se consideren inservibles, estandarización de formato, estructuración de la información, entre otros.

1.4.1.1.3. Generar modelos

Como siguiente paso es imprescindible crear un modelo que contenga los algoritmos necesarios para obtener la información útil al tratar los datos, el modelo debe de corresponder a los datos recolectados para tener el máximo provecho.

1.4.1.1.4. Evaluar los modelos

Luego de obtener la información útil por medio del tratamiento de los datos, es de suma importancia realizar una verificación, y que la información obtenida sea relevante para el problema que se busca solucionar y que no contenga errores que puedan afectar las decisiones futuras

1.4.1.1.5. Implementar modelos

Al haber verificado y poseer un modelo funcional dentro del entorno delimitado por el problema, la información obtenida del procesamiento de los datos, por medio de un modelo, brinda una clara ventaja en la optimización de toma de decisiones.

1.4.1.2. Aplicaciones

La minería de datos permite su aplicación en una amplia variedad de entornos, como en la predicción de eventos promedio de hallazgos de patrones o tendencias repetitivas y tomar decisiones con base en eventos que no han sucedido.

La agrupación demográfica por medio de afinidades descritas dentro de los datos y brindar una atención personalizada para cada uno de los interesados.

Es posible realizar una optimización de procesos que al mismo tiempo aumenten la calidad de un producto, lo cual repercute en un alza de ventas para cualquier empresa que utilice este tipo de análisis para la creación de sus productos.

1.4.2. Algoritmos genéticos

Basados en las teorías postuladas por Charles Darwin, los algoritmos genéticos son métodos adaptables que imitan organismos vivos en la forma en cómo una generación evoluciona por medio de la selección natural o la supervivencia del más apto.

Estos métodos permiten alcanzar soluciones óptimas por medio de iteraciones y evolución de las generaciones, siendo la generación siguiente más óptima o más cercana a los resultados esperados, utilizando de manera eficiente la información histórica de las generaciones para crear por cada iteración una mejorada.

1.4.2.1. Elementos de un algoritmo genético

Debido a que están basados en cómo un organismo evoluciona, las operaciones o elementos más importantes dentro de la metodología están relacionadas ampliamente con la terminología biológica.

1.4.2.1.1. Población

Es un conjunto de seres vivos que habitan en un determinado lugar, los cuales proveen de diversidad a la generación, que inicialmente puede ser creada aleatoriamente o iniciar con una generación histórica que brinde un acercamiento óptimo a la solución buscada.

1.4.2.1.2. Selección

Dentro de esta operación se realiza la selección de dos cromosomas dentro de la población, los que serán encargados de reproducirse y generar una nueva descendencia. Basado en la selección natural, entre más apto sea uno de los elementos será seleccionado más veces para reproducirse.

1.4.2.1.3. Cruce

Tal y como sucede en los organismos vivos, luego de haber seleccionado los cromosomas necesarios para la reproducción, se realiza un cruce entre estos con el fin de obtener una descendencia que sea más apta para la supervivencia, obteniendo cada vez una generación mejor.

1.4.2.1.4. Mutación

Una mutación es una variación aleatoria dentro de un cromosoma de la población, basados en una probabilidad muy pequeña esta operación puede reacondicionar una generación y lograr alcanzar una solución óptima en un menor número de iteraciones.

1.4.2.2. Aplicaciones

La aplicación más importante de este tipo de método es la utilización en la optimización de recursos para alcanzar un objetivo, esto debido a sus características evolutivas permite obtener una mejora a la solución anterior, hasta encontrar la que mejor se ajuste a las necesidades.

1.4.3. Aprendizaje automático

El aprendizaje de máquina o automático es una disciplina de la inteligencia artificial que trata de la construcción de algoritmos capaces de realizar predicciones o tomar decisiones, basados en un modelo estadístico antes que seguir de forma rígida las instrucciones programadas explícitamente.

Los principales usos de este tipo de inteligencia artificial es construir un sistema capaz de colaborar con expertos y disminuir la interacción humana logrando automatizar procesos mediante la generalización de resultados a través de un algoritmo.

1.4.3.1. Representación

Define como serán interpretados los datos y las acciones que se deben realizar mediante la descripción de una particular entrada, realizándose mediante el uso ejemplos para describir la forma de aprendizaje y definir la forma de operar de un algoritmo.

1.4.3.2. Generalización

Es la capacidad que posee un aprendiz, en este caso un programa computacional, de realizar nuevas decisiones con base en los ejemplos que ya han sido descritos para su aprendizaje, el principal objetivo es que las elecciones o decisiones que sean tomadas, sean generalizadas de acuerdo su experiencia.

1.4.3.3. Tipos de algoritmos

Para la realización del aprendizaje se hace uso de tres principales algoritmos: aprendizaje supervisado, no supervisado y por refuerzo, los cuales describen diferentes técnicas para estimular el aprendizaje mediante una entrada de datos que detalle el comportamiento deseado.

1.4.3.3.1. Aprendizaje supervisado

Provee al aprendiz de un conjunto de datos detalladamente identificados para describir la forma en cómo identificar cada entrada y el resultado esperado, por medio de este conjunto de datos el aprendiz preparará y establecerá reglas de identificación, creando dentro la experiencia para tomar decisiones con base en el conocimiento adquirido.

1.4.3.3.2. Aprendizaje no supervisado

Por medio de este tipo de algoritmo se busca encontrar relaciones implícitas dentro de las entradas, que no presentan una estructura fija, al únicamente poseer entradas, es necesario que el aprendiz sea capaz de categorizar e identificar por medio de detección de patrones subyacentes.

1.4.3.3.3. Aprendizaje por refuerzo

Este algoritmo de aprendizaje cae dentro de los dos extremos anteriores, ya que su entorno provee de poca retroalimentación para predecir sus resultados con base en una entrada, pero no cuenta con una precisa medida para detectar los errores, se considera que ese algoritmo permite al aprendiz adquirir experiencia en base de prueba y error.

1.4.3.4. Aplicaciones

La utilización del aprendizaje automático ha sido de gran utilidad dentro de las herramientas para detección de patrones, como ejemplo, los motores de búsqueda, pudiendo predecir y mostrar los resultados que el usuario desea, aun si la entrada provista no es lo suficientemente explícita.

El procesamiento de lenguaje natural ha sido beneficiado por medio de esta técnica de inteligencia artificial, ya que es posible automatizar la realización de un análisis de morfología y sintaxis, para lograr extraer la información, clasificación y agrupamiento de textos.

Para las tiendas que operan en línea, los sistemas de recomendaciones han causado un impacto en cómo predecir los gustos de un cliente logrando presentarlas de forma personalizada, basándose en rastros en su web, compras anteriores o hasta similitudes con otros compradores, todo esto basado en la experiencia adquirida a través de patrones de comportamiento.

2. INTERVENCIÓN HUMANA EN EL TRÁNSITO VEHICULAR

2.1. Introducción

Con el crecimiento del volumen de vehículos, el tránsito generado es cada vez más engorroso y difícil de soportar, lo que ha provocado muchas soluciones alternativas con la finalidad de disminuir tal afluencia y agilizar el movimiento vehicular.

Muchas de las acciones tomadas para agilizar la creciente cantidad de vehículos, han sido con base en conocimientos empíricos sobre como el tránsito vehicular usualmente se desarrolla.

Existen diversos componentes que están particularmente relacionados con la intervención humana dentro del tránsito vehicular, y cómo cada uno de estos contribuye en la toma de decisiones, logrando beneficiar o perjudicar la agilización del flujo vehicular.

2.2. Agente de tránsito

Un agente de tránsito será definido como cualquier persona civil que tiene a su cargo la responsabilidad de brindar un servicio de calidad para la dirección, control y administración del tránsito vehicular, el cual permita aumentar el flujo vehicular por medio de acciones que están basadas en normas vigentes.

Dentro de las funciones básicas se encuentran:

- Supervisar y regular el cumplimiento de las normas de tránsito vehicular vigentes.
- Reportar incumplimientos presentadas hacia las normas de tránsito vehicular de forma escrita.
- Soporte a la señalización e infraestructura en función de su cumplimiento.
- Estructuración de operativos para garantizar la seguridad vial como: revisión con alcoholímetro, puestos de registro, entre otros.
- Formular nuevas normas que impulsen el aumento del flujo vehicular.

Cada una de estas funciones realizadas por los agentes de tránsito son reflejadas en la fluidez en el flujo vehicular, sin embargo, estas conllevan la toma de decisiones, que no siempre puede ser óptima, tomando en cuenta la intervención humana, existe un margen de error que puede afectar negativamente la fluidez del tránsito vehicular.

2.3. Señales de control de tránsito

Las señales de control de tránsito o como son conocidos comúnmente, los semáforos, son mecanismos de regulación de flujo de tránsito vehicular que, por medio de diferentes tipos de indicaciones permite ceder el paso a vehículos de manera secuencial.

Más que el control del flujo vehicular, estos mecanismos proveen seguridad para los vehículos, logrado a través de la asignación secuencial de intervalos de tiempo de paso que permiten el ordenamiento en diferentes intersecciones.

El flujo vehicular es influenciado por las decisiones que estos mecanismos toman en la asignación de tiempos para cada intervalo de paso, el impacto de cada una de las decisiones tomadas repercute en la fluidez del tránsito vehicular, por lo que la selección en los intervalos de tiempo deben ser las más óptimas y acertadas.

2.3.1. Funciones básicas

Dentro de las funciones esenciales para cada uno de estos mecanismos se encuentran:

- Por medio de asignaciones de tiempo de paso de forma secuencial, ceder el paso a un flujo vehicular.
- Regular la velocidad de la circulación de los vehículos y lograr mantener un flujo contante.
- Proporcionar el correcto ordenamiento vehicular.
- Reducir la cantidad de accidentes de tránsito vehicular y su gravedad.
- Dirigir la circulación dentro de los carriles.

2.3.2. Clasificación

Como función fundamental en el flujo del tránsito vehicular, un dispositivo de señales de control debe proveer de intervalos de tiempo para la asignación de paso a cada corriente vehicular, por lo que son clasificados en la forma como estos son asignados.

2.3.2.1. Tiempos fijos

Las señales de control de tránsito clasificadas dentro de tiempos fijos, poseen la peculiaridad de brindar un intervalo de tiempo, ya sea predeterminado o de acuerdo a una programación establecida, la cual le dicte la cantidad de tiempo asignada a cada flujo vehicular.

Estos tipos de mecanismos son de mayor utilidad donde los flujos vehiculares se mantienen de manera estable y su cambio con el tiempo es de poca variabilidad, por lo cual, ceder el paso a distintos flujos vehiculares dentro de una intersección se realiza de forma secuencial.

Usualmente son utilizados, también como un control presincronizado que ayuda a disminuir la interacción humana por medio de condiciones ya establecidas que permiten un flujo vehicular contante sin causar demoras o congestionamientos.

La eficacia de esta clasificación radica en las condiciones mencionadas, ya que dentro de un entorno cambiante que requiere de la modificación de intervalos de tiempo en cómo se concede el paso, se presentarán dificultades que no podrán ser resueltas.

2.3.2.2. Tiempos variables

Las señales de control de tránsito clasificadas dentro de tiempos variables poseen las características necesarias para diferenciar sus intervalos de tiempo en los que cede o detiene el paso dentro de un flujo vehicular, realizado por medio de sensores que registran información dentro del tránsito vehicular.

Estos mecanismos son usualmente utilizados en donde el tránsito vehicular no sigue un cambio estático o lineal, sino variable, causando irregularidades en los flujos vehiculares, lo cual requiera modificaciones en los intervalos de tiempos según sea demandado por los cambios en el tránsito vehicular.

Las modificaciones a los intervalos de tiempo son realizados por medio de dispositivos que son capaces de captar información relacionada al tránsito vehicular, lo cual le permite tomar decisiones en cuál es la mejor estrategia para ceder el paso o detenerlo.

Cuando estos mecanismos se encuentran operando en las condiciones adecuadas, su eficiencia se demuestra durante periodos en los que el volumen vehicular es bajo, ya que no causa demoras significativas.

2.3.3. Sensores

Dispositivos capaces de percibir cambios en su entorno y generar datos que permiten registrar cualquier cambio que proporcione información suficiente para tomar una decisión en cómo secuencialmente define el tiempo de paso de una vía.

2.3.3.1. Tipos de sensores

Existe una gran variedad de sensores que pueden ser implementados para capturar información del tránsito vehicular, sin embargo, con la finalidad de capturar datos útiles tomar decisiones relacionados a este, se presentan los que a continuación se describen:

2.3.3.1.1. Sensores de presión

Este tipo de sensor es accionado por medio de las ruedas del vehículo, lo cual genera un registro de cuántos vehículos transitan por ese tramo de vía, lo cual brinda un conteo del volumen vehicular.

Sin embargo, este tipo de sensor puede brindar información poco precisa si no se contempla la posibilidad de que un vehículo puede, no solo pasar por él, sino detenerse sobre lo que puede volver el dispositivo ineficaz.

2.3.3.1.2. Sensores magnéticos

Este tipo de dispositivo es capaz de registrar cambios dentro de su entorno por medio de alteraciones dentro de un campo magnético, el cual permite realizar el conteo necesario dentro del flujo vehicular.

Este tipo de sensor es de mayor utilidad en calles angostas, donde es baja la posibilidad de obstáculos que se interpongan y puedan causar información errónea, logrando brindar de información precisa para ser procesada.

2.3.3.1.3. Sensores de radar

Poseyendo el funcionamiento usual de un radar, son capaces de detectar un vehículo en movimiento por medio de un campo de energía de microondas, lo que permite minimizar la información errónea que puede ser recopilada.

Al ser capaz de registrar eventos que posean un mínimo de velocidad, elimina la posibilidad de recopilar información incorrecta cuando un vehículo está estacionado o debe disminuir su velocidad, incluso cualquier objeto que pueda causar una influencia dentro de su campo de energía.

2.3.4. Datos a recopilar

Para tomar las decisiones óptimas y agilizar el flujo vehicular, es necesario contar con los datos necesarios y precisos para procesarlos y tomar la mejor decisión basada en toda la información recopilada.

La mayor parte de los datos necesarios pueden ser medidos por sensores que aumenten la eficacia al brindar información exacta con un margen de error significativamente bajo. La información recolectada ayudará a mejorar la toma de decisiones para la selección de intervalos de tiempo.

2.3.4.1. Volumen vehicular

Una de las variables por registrar es el volumen vehicular que transita por cada vía, esto permitirá conocer la cantidad exacta de vehículos que usan la intersección considerada y cómo deben ser variados los tiempos para otorgar el paso al flujo vehicular por medio de la información recopilada.

2.3.4.2. Circulación tránsito oblicuo

Dentro de un entorno donde se presente una intersección transversal, es necesario medir la densidad vehicular que transita por cada una de las vías y conocer las condiciones que causen una variación el tiempo para detener u otorgar el paso a un flujo vehicular.

2.3.4.3. Volumen en horas pico

Uno de los factores más importantes a registrar es el volumen vehicular dentro de las horas pico, es decir, cuando el flujo vehicular se encuentre en su punto más alto, la información recopilada permitirá conocer el comportamiento del tránsito vehicular en este entorno.

2.3.4.4. Velocidad promedio

Como factor importante en el tránsito vehicular, la velocidad promedio de un vehículo es algo que debe ser considerado, ya que este factor demuestra lo libre que se encuentra una vía y puede ser utilizada como una variable de estudio.

2.3.4.5. Tiempo

Dentro de cualquier estudio donde sean registrados datos, la variable de tiempo es muy importante, pues brinda una sólida base de referencia con cada una las mediciones realizadas y provee la capacidad de llevar un registro que ofrezca información histórica que permita conocer comportamientos pasados para predecir futuros.

2.4. Educación vial

Comúnmente se cree que la educación vial está centrada en el cumplimiento de las señales de tránsito vehicular, respetar los semáforos y conducir adecuadamente un vehículo, sin embargo, la educación vial trata también la forma de cómo un peatón debe conducirse, desde la utilización de pasarelas hasta la utilización del paso de cebra.

La educación vial es una base de conocimiento que brinda a un conductor, pasajero o peatón las necesarias habilidades y hábitos que permitan el completo desenvolvimiento por medio de la compresión y respeto a las normas y leyes vigentes de tránsito vehicular.

La falta de educación vial es uno de los valores sociales que son poco inculcados, ya sea por ignorancia o por falta de respeto. Esto más que provocar cualquier tipo de accidente vial atenta contra la integridad y calidad de vida de cualquier peatón, conductor o pasajero.

2.5. Intervención humana

Se definirá como intervención humana, cualquier acción que provenga de un agente de tránsito con la intención de agilizar el flujo vehicular, específicamente la interacción con sistemas de señalización de control de tránsito.

Debido a las múltiples variabilidades dentro del tránsito vehicular, muchas de las decisiones en cambio de fase de un semáforo son hechas por intervención de un agente de tránsito, esto conlleva a contemplar un margen de error generado por dicha interacción.

2.5.1. Riesgos

La intervención humana dentro del tránsito vehicular por medio de los agentes de tránsito, presenta riesgos que deben ser evaluados para poseer un conocimiento integral de lo que conlleva este tipo de actividades.

La mayor parte estos agentes que intervienen con las señales de control de tránsito vehicular o se convierten en extensiones de estas, están usualmente a la intemperie, frente al flujo vehicular, lo cual presenta propensos a accidentes o acciones que atenten contra su integridad.

2.5.2. Recursos

Como se puede inferir, dentro de la intervención humana para el tránsito vehicular se encuentran dos actores que son los encargados de controlar los flujos vehiculares, principalmente dentro de intersecciones y realizar las acciones necesarias para aumentar la fluidez dentro de estas, esos actores son: los dispositivos de señalización de control de tránsito como recurso primario para coordinar los flujos vehiculares; y los agentes de tránsito que interfieren en la mejora de los intervalos de tiempo para la otorgación de paso o detención de una vía.

2.5.3. Costos

Debido a la intervención de un agente de tránsito, hay costos implícitos que deben ser considerados, tales como, el equipo que debe ser utilizado para desenvolverse adecuadamente y la remuneración por la realización de su trabajo.

El costo de los dispositivos de control de tránsito, ya sea de su operación o mantenimiento, representan un factor a tomar en cuenta para su implementación, sin embargo, los costos de operación pueden ser controlados o limitados al tiempo en que los sensores estén en funcionamiento.

2.5.4. Eficiencia

Es una de las características más cuestionadas al hablar de la intervención humana en la modificación de las señalizaciones de control de tránsito, ya que se cree que, en lugar de ayudar a que el flujo vehicular sea más fluido lo hace más lento y engorroso.

Esto se debe a que cuando la densidad de tránsito vehicular aumenta, las señalizaciones que controlan el flujo vehicular son mayormente ignoradas, lo que conlleva a recurrir a la intervención humana que impacta en la fluidez dentro del flujo vehicular.

3. ARQUITECTURA PARA ANÁLISIS DE DATOS MASIVOS

3.1. Introducción

El análisis de datos masivos se basa, fundamentalmente, en tres premisas, aunque no nuevas, con un enfoque distinto debido a las características intrínsecas de esta tecnología, lo que ha hecho que sean adaptadas para cada una de las necesidades y proveer de nuevas soluciones.

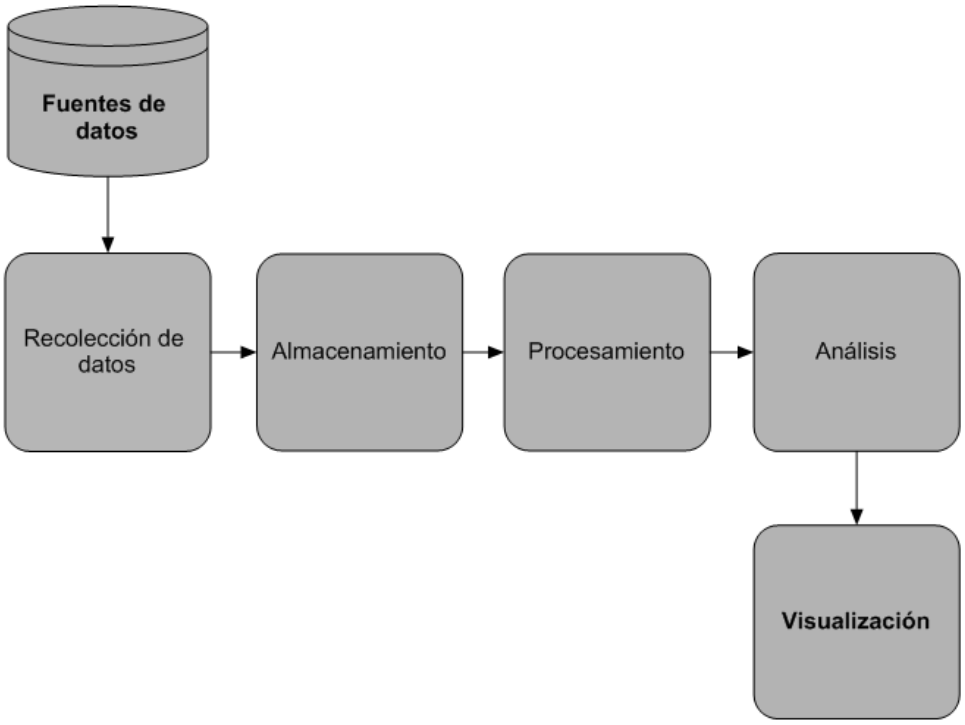
Capturar, almacenar y analizar: por medio de cada una de estas etapas se logra adquirir conocimiento que se encuentra escondido dentro de volúmenes de datos, de modo que la utilización de tecnologías tradicionales resulta tan costos como lento para procesar tal magnitud de información.

Justificada por cada una de estas bases se propone una arquitectura capaz de brindar una solución por medio del análisis de datos masivos, que contribuya a la reducción de la intervención humana en el tránsito vehicular.

Debido a la manera de cómo los datos son tratados y el proceso que conlleva, se propone una arquitectura en capas, las cuales son: recolección de datos, almacenamiento, procesamiento, análisis y visualización.

La arquitectura propuesta conforma una básica y completa solución al caso de uso en particular, pues cada uno de sus componentes contribuye a solventar el objetivo en común.

Figura 1. **Arquitectura para análisis de datos masivos**



Fuente: elaboración propia, empleando programa ClickCharts.

3.2. Arquitectura propuesta

La arquitectura propuesta conforma una básica y completa solución, pues cada uno de sus componentes contribuye a solventar el objetivo en común, estos son:

3.2.1. Fuentes de datos

Como componente principal para esta arquitectura, los datos son la base principal, y debido a que se puede capturar en grandes cantidades; provee de una gran fuente de información de la cual pueden obtenerse grandes ventajas.

El crecimiento en el volumen de los datos acompañado de una velocidad abrumadora en su generación y provenientes de diversas fuentes, ha generado la necesidad de cambios en lo que convencionalmente era utilizado, abriendo nuevas oportunidades a nuevas tecnologías.

Los datos utilizados serán los capturados por medio de los sensores posicionados a los alrededores de las señales de control de tránsito, estos datos serán utilizados para ser procesados y analizados.

Los datos capturados por medio de los sensores serán los descritos anteriormente, con los cuales se podrán realizar los cálculos necesarios para aumentar la eficiencia en las tomas de decisión y disminuir la necesidad de la intervención humana.

3.2.2. Recolección de datos

Como primer etapa, luego de que los datos son capturados, se debe realizar la comunicación entre los sensores que son los encargados de realizar todo el proceso de capturado de datos y transmitirlos hacia donde serán almacenados.

Dependiendo de las necesidades que se presenten, la recolección de datos puede realizarse dividiéndose en dos grandes grupos: por lotes y en tiempo real; los cuales brindan una solución dependiendo los usos de los datos y la frecuencia de utilización.

3.2.2.1. Recolección por lotes

Este tipo de recolección de datos se caracteriza por interactuar con la fuente de datos de manera periódica, de modo que pueda ser recolectada información nueva que no ha sido migrada.

Se realiza una transmisión de datos, cuando se realiza una conexión a la fuente de datos, ya sea una base de datos o algún sistema de ficheros, estos son transportados para ser almacenados, solamente los que han sufrido cambios desde la última conexión.

La ventaja principal que presenta este tipo de recolección de datos, es que los recursos no son utilizados de manera continua, esto ayuda a reducir el número de fallas y costos por la utilización de los recursos relacionados a su funcionamiento.

Sin embargo, debido a que su conexión con la fuente de datos es de forma periódica, los datos que son recolectados no pueden ser procesados ni monitoreados en tiempo real.

3.2.2.2. Recolección en tiempo real

Para este tipo de recolección de datos se realiza una conexión continua con la fuente datos, de modo que la información se transfiere de forma continua desde la fuente de datos al recolector.

Debido a las características de este tipo de recolección de datos, la información es migrada al momento que la fuente de datos realiza su detección o generación, la información es transmitida a donde será almacenada para su procesamiento.

La principal ventaja para este tipo de recolección de datos, es contar con datos actuales, lo que provee de un flujo constante de información útil para escenarios donde la monitorización de información en tiempo real es de gran importancia.

Sin embargo, la cantidad de recursos utilizados por este tipo de recolección de datos es alta, esto lo hace propenso a fallas y altos costos de funcionamiento y mantenimiento.

3.2.3. Almacenamiento

Para la capa de almacenamiento es importante que cumpla con características que son de utilidad para su futuro procesamiento, por lo que existen varias alternativas que cumplen con cada una de las necesidades que presenta cada implementación.

La importancia de esta capa reside en la capacidad de almacenar grandes volúmenes de datos para que puedan ser procesados, sin importar la estructura que estos posean, y lograr un tiempo de respuesta corto, logrando mostrar pronto los resultados buscados.

Existen, principalmente dos grupos: los sistemas de archivos que permiten el almacenamiento de forma distribuida y las bases de datos, las cuales se pueden dividir en bases de datos relacionales y bases de datos NoSQL.

3.2.3.1. Sistemas de ficheros distribuidos

Los sistemas de ficheros distribuidos constituyen una parte fundamental de una arquitectura para el análisis de datos masivos, ya que muchas otras herramientas están construidas o funcionan sobre este tipo de almacenamiento.

La característica principal de esta forma de almacenamiento es la escalabilidad, lo cual permite variar su tamaño dependiendo de las necesidades que se presentan sin comprometer el rendimiento completo del sistema.

La idea detrás de este tipo de almacenamiento es poseer un conjunto de computadoras, llamadas nodos, interconectadas en una red, las cuales permiten distribuir los ficheros de forma completa o por medio de bloques de datos.

Este sistema de archivos se divide en un sistema lógico, el cual permite visualizar los ficheros de tal manera que parecieran estar en una misma computadora y el sistema físico, que es el encargado de distribuir los ficheros dentro de los nodos interconectados.

3.2.3.2. Bases de datos relacionales

Las bases de datos ha sido la forma de almacenamiento más usado a lo largo de muchos años, ya que su robustez y facilidad de uso permite poseer información almacenada de forma estructurada por medio de relaciones descritas dentro de un modelo.

Presentan grandes ventajas en el manejo de persistencia e integridad de la información que se almacena dentro de ellas, contando con información veraz que no fue alterada de manera aleatoria, esto gracias a una estructura descrita por medio de tablas, relaciones, claves, entre otros.

Una de las principales herramientas que han hecho de este tipo de almacenamiento una forma fácil y rápida de usar es el lenguaje SQL, el cual permite crear consultas hacia la base de datos con un lenguaje muy sencillo, parecido al lenguaje humano, pudiendo ser utilizado por usuarios no expertos.

Sin embargo, cuando la información que se necesita almacenar no proviene de una fuente que brinde los datos de forma estructurada, una estructura fija es poco funcional, lo cual puede crear la necesidad de cambios se dificulta dentro de un sistema de bases de datos relacional.

La velocidad en la consulta sobre las bases de datos está directamente relacionada con los índices, estos se vuelven cada vez más complicados de crear y de mantener con el crecimiento del volumen de datos.

3.2.3.3. Bases de datos NoSQL

Debido al a necesidad de procesar información sin una estructura fija y almacenar grandes volúmenes de datos, la base de datos NoSQL surgió como repuesta a los primeros problemas relacionados con los datos masivos.

Estas no almacenan la información siguiendo algún modelo relacional, por lo que proveen de una alta flexibilidad en la forma en que la información es grabada, no dependiendo de una estructura fija.

Unos de los mayores puntos a favor es que permiten la escalabilidad horizontal de una forma más simplificada, pudiendo añadir nuevos nodos sin repercutir en problemas con índices, ya que no hace uso de estos.

Existen diferentes tipos, ya que sin depender de una estructura fija es posible que sean almacenadas de diferentes formas, entre las cuales están: orientadas a documentos, orientadas a columnas, de clave valor y de grafo.

Cada uno de estos tipos de almacenamiento es eficaz para un escenario específico dentro de las cuales sus características son completamente útiles, cada uno de ellos posee sus propias ventajas y desventajas sobre el otro.

Aunque cada una es útil por separado, es posible aprovechar las capacidades de ellas, pudiendo utilizar como base un almacenamiento tal como un sistema de ficheros distribuidos o una base de datos NoSQL y la velocidad de consulta sobre una base de datos relacional con la información de los datos ya procesados.

3.2.4. Procesamiento

Como siguiente paso, luego de poseer almacenada la información, es necesario procesarla para explotar la información que se encuentre inmersa dentro de todos los datos recolectados.

Dentro de esta etapa se busca tomar todos los datos en su estado puro y por medio de funciones que operan sobre estos datos poder realizar todo tipo de acciones de agregación, las cuales permiten extraer la información de todo lo que fue capturado.

Luego de que la información ha sido procesada, puede utilizarse para realizar el respectivo análisis o almacenarse nuevamente para un análisis futuro, actualmente se cuentan con dos tipos de proceso de información: por lotes y en tiempo real.

3.2.4.1. Proceso por lotes

Este tipo de proceso de información se caracteriza por no realizarse de modo continuo y procesar la información por lotes, lo que provee de una sola escritura y muchas lecturas.

El impacto dentro de su rendimiento es importante, ya que la velocidad de respuesta aumenta al procesar volúmenes de datos de gran tamaño, algo que puede ser contrarrestado agregando más poder de procesamiento.

3.2.4.2. Proceso en tiempo real

El proceso en tiempo real es de mucha utilidad cuando es necesario realizar una monitorización continua de todos los eventos que son registrados por sensores dentro de logs u otras formas de captura de información.

Este tipo de procesamiento debe poseer recursos de alto poder de cómputo para cumplir con la demanda de trabajo que se tiene al estar realizando las tareas necesarias para el tratamiento de la información.

Cada uno de los tipos de procesamiento son útiles por sus características en escenarios específicos, sin embargo, estas características pueden ser sumadas y tomar la ventaja del proceso por lotes y utilizar datos históricos capturados y el proceso en tiempo real para tratar los datos en el instante que se generaron.

3.2.5. Análisis

La etapa de análisis está muy relacionada con la capa de procesamiento, ya que estas dos proveen de herramientas para extraer conocimiento de los datos, que con métodos convencionales seria costoso y conllevaría a un largo tiempo de espera por resultados.

Dentro de esta etapa es donde se hacen uso de los modelos estadísticos en búsqueda de correlaciones que puedan aportar un significado; debido a los grandes volúmenes de datos, los métodos estadísticos utilizados tradicionalmente se vuelven obsoletos.

Una característica más de esta capa, es brindar del conocimiento oculto dentro de los datos capturados. Mediante técnicas de análisis de datos avanzadas es posible detectar tendencias o patrones sobre las cuales basar decisiones.

Como parte de una arquitectura, es donde se debe emplear más recursos para su implementación, ya que todas las decisiones tomadas con base en el análisis de datos masivos recaen sobre estas dos capas, por lo que la técnica a utilizar para el análisis y el modelo estadístico deben ser cuidadosamente elaborados para su uso específico.

3.2.6. Visualización

Para la etapa final de la arquitectura, luego de haber recorrido cada una de las etapas de recolección, almacenamiento, procesamiento y análisis, es necesario presentar ese conocimiento que ha sido obtenido a lo largo del proceso.

Todos los datos que fueron recolectados, que ahora poseen un significado son traducidos a tiempos para ceder el paso, que por medio de señales de control de tránsito, controlar de una manera más eficaz el flujo vehicular y así disminuir la intervención humana dentro de este.

3.3. Hadoop

Es un proyecto de Apache Foundation, el cual ha tomado mucho auge debido a la necesidad de procesamiento de datos masivos, ya que provee de las herramientas para enfrentar los nuevos retos que este enfoque representa.

El propósito principal es dotar de un marco de trabajo donde un conjunto de herramientas, aplicaciones y entornos Java permitan crear sistemas capaces de brindar una solución por medio de computación escalable y distribuida utilizando modelos de programación poco complejos.

Por medio de la computación distribuida y escalable, Hadoop permite realizar análisis sobre grandes cantidades de datos, ya sea en una computadora individual, o en miles de computadoras interconectadas que realizan el procesamiento y almacenamiento de forma local.

Hadoop se caracteriza por brindar alta disponibilidad a nivel de aplicación, esto permite controlar y detectar los errores dentro de los nodos que lo componen, lo que aumenta la tolerancia a fallos del sistema completo.

Hadoop está construido básicamente sobre dos módulos, un sistema de archivos distribuido como capa de almacenamiento y una capa de procesamiento constituido por un entorno de computación distribuida.

3.3.1. Hadoop Distributed File System

El sistema de archivos distribuido de Hadoop (HDFS, por sus siglas en inglés) está diseñado para almacenar grandes archivos de entrada de datos sobre hardware de bajo costo, sobre el cual funcionan muchas herramientas de Hadoop.

La idea principal bajo HDFS es reducir los tiempos de escritura; por lo que no son permitidas las escrituras aleatorias, más bien se trabaja bajo un patrón de una escritura y muchas lecturas, por lo que luego de que los datos han sido movidos o copiados desde la fuente puede ser leído muchas veces para su análisis.

HDFS permite la escalabilidad horizontal simplificada, ya que es posible añadir más nodos en cuanto sea necesario sin detener el sistema completo, esto también aumenta la tolerancia de fallos y disminuye la pérdida de datos. Debido a su factor de replicación la caída o desconexión de uno o más nodos no constituye pérdida total de datos.

La arquitectura de HDFS está basada en el patrón de maestro-esclavo, donde el maestro es llamado NameNode, que es el encargado de gestionar la logística de acceso a los bloques de datos entre los DataNodes, que funcionan como esclavos siendo estos los encargados de almacenar dichos bloques de datos.

3.3.2. Hadoop MapReduce

Es un modelo de programación orientado a procesar grandes volúmenes de datos de forma distribuida y paralela, el cual se divide fundamentalmente en las funciones *map* y *reduce*.

Este entorno está implementado de tal forma que, un programador este encargado de escribir solamente el algoritmo, pues este mismo se encarga de realizar la gestión de cada uno de sus procesos y fases, aun si alguno falla.

Al igual que HDFS, MapReduce tiene un patrón de arquitectura maestro-esclavo, donde el nodo maestro es llamado JobTracker, que es el que se encarga de gestionar la planificación de MapReduce y el nodo esclavo llamado TaskTracker que es el encargado de realizar ejecución de todo el proceso de MapReduce.

3.3.2.1. Etapas de MapReduce

El proceso de MapReduce se realiza por medio de etapas, las cuales constituyen todo el procesamiento de la información, desde la realización del mapeo de la información hasta su reducción.

3.3.2.1.1. Mapeo

Durante la fase de mapeo se realiza la ejecución del algoritmo descrito por el programador que creó la función de mapeo, esta se ejecuta de forma distribuida dentro de nodos llamados *mappers*.

Al finalizar de procesar la información de entrada, la función de mapeo genera pares de clave-valor como salida al mapeo, estas son las que sirven como entrada a las siguientes fases, la salida de los *mappers* debido a que son resultados intermedios no son escritos en HDFS, solamente son escritos los resultados del mapper en HDFS, cuando no se utilice la función de reducción.

3.3.2.1.2. Mezcla

Toma como entrada la salida de la función de mapeo, con el objetivo de realizar una mezcla de manera local de los resultados obtenidos por el *mapper*, generando una lista de clave-valor.

3.3.2.1.3. Ordenamiento

Dentro del ordenamiento se toma la lista de clave-valor de cada uno de los *mappers* y se realiza la ordenación con base en la clave de cada par, de tal forma que la salida a esta fase es una lista de claves asociados a una lista de valores.

Cabe destacar que este ordenamiento se realiza de forma aleatoria, basándose únicamente en las claves, por lo que si se necesita un ordenamiento específico debe ser implementado explícitamente.

3.3.2.1.4. Combinación

La fase de combinación puede ser realizada opcionalmente, ya que su función principal es realizar una reducción de carga de trabajo a la función de reducción, por lo que esta puede ser reutilizada para combinación siempre y cuando cumpla con ser asociativa y conmutativa.

3.3.2.1.5. Partición

Al igual que la combinación, la etapa de partición puede ser realizada opcionalmente dentro del proceso de MapReduce, ya que su función es realizar una ordenación y delegar a cuál Reducer se entrega la salida de la etapa anterior, debe ser utilizada si se necesita realizar una partición distinta a la que se hace por defecto.

3.3.2.1.6. Reducción

Como fase final se realiza la función de reducción, llamada *reducer*, al igual que los *mappers*, los *reducers* pueden ser más de uno, según sea necesario o sea descrito en su ejecución.

Dentro de esta fase se realiza la reducción de los resultados encontrados dentro de la etapas anteriores, dando como resultado final un listado de pares clave-valor, los cuales son el producto de todo el proceso que conlleva la realización de MapReduce, estos son escritos directamente en HDFS.

3.3.3. Herramientas

Hadoop cuenta con muchas herramientas y aplicaciones que son utilizadas para distintos propósitos, según sean las necesidades, por lo que se realizará una descripción de las herramientas que serán empleadas para satisfacer las exigencias para la arquitectura propuesta.

3.3.3.1. Recolección de datos: Flume

Es una herramienta dentro del ecosistema de Hadoop, que permite transportar grandes cantidades de datos de forma distribuida de manera eficiente y confiable, ya sea una o muchas fuentes hacia su almacenamiento de forma centralizada.

Además de posibilitar el transporte de datos, Flume permite realizar agregación y realizar una personalización a sus fuentes de datos logrando que sea adaptada a cualquier fuente de datos sin importar la estructura que estos contengan.

3.3.3.1.1. Evento

Es la unidad de datos que utiliza Flume para ser transportados, el cual está constituido por datos y metadatos, estos pueden poseer un encabezado, dichos eventos son transportados desde donde son originados hasta su destino final.

3.3.3.1.2. Agente

Es un proceso independiente de Flume, compuesto de fuentes, canales y sumideros, con los cuales provee de la capacidad de recibir almacenar y trasladar los eventos hacia su próximo destino, cada agente debe tener como mínimo uno de cada uno de sus componentes para realizar su función de manera adecuada.

- Fuente: es una implementación de una interfaz que está a la escucha de eventos para consumirlos y almacenarlos dentro de un canal, cabe destacar que los eventos son escritos en el canal de forma transaccional y no se pierde ninguno, a menos que el canal esté configurado explícitamente de esa manera.
- Canal: es el encargado de almacenar todos los eventos provenientes de la fuente, cada evento escrito en el canal no es removido hasta que el sumidero finaliza una transacción asegurando que ningún evento sea perdido antes de ser almacenado.
- Sumidero: todos los eventos grabados en el canal son removidos hacia un almacenamiento por medio de transacciones, lo que asegura que cada evento sea eliminado del canal hasta que haya sido escrito en el siguiente destino.

3.3.3.1.3. Cliente

Es un componente más de Flume, que permite recolectar los eventos donde son originados, para luego realizar la entrega hacia los agentes, cada cliente usualmente opera desde donde se originan los datos, esto permite utilizar un cliente para distintas fuentes de datos y uno solo para recolectar los eventos.

3.3.3.2. Almacenamiento: HBase

Por medio de HBase es posible realizar escrituras o lecturas de forma aleatoria o en tiempo real, permitiendo almacenar grandes tablas de datos dentro de un conjunto de computadoras de bajo costo.

HBase es una base de datos NoSQL, la cual está orientada a columnas, por lo que todos los datos que se desean almacenar dentro de esta, son celdas agrupadas en columnas, a las cuales pueden ser agregadas columnas como fuese necesario durante se encuentra en funcionamiento.

Esta base de datos permite almacenar todos los datos de forma distribuida, logrando dividir y fragmentar cada uno dentro de múltiples nodos, debido a la fácil escalabilidad permite añadir más nodos, los que son agregados automáticamente aumentando, también, la tolerancia a fallos.

Como sistema de almacenamiento utiliza HDFS lo que la hace completamente compatible con Hadoop; además posee interfaces ya implementadas para realizar trabajos de procesamiento con MapReduce y tablas de HBase.

3.3.3.2.1. Modelo de datos

Posee un modelo de datos orientado a columnas, dentro del cual los conceptos fundamentales, en cómo esta base de datos almacena cada uno de los datos, son los siguientes:

- Tablas: básicamente, consiste en múltiples filas y columnas, sin embargo, el concepto más cercano a su funcionamiento es como un arreglo de clave-valor multidimensional.
- Fila: consiste en una clave de fila y un grupo de columnas asociadas a un valor, cada una de estas se encuentran ordenadas por medio de su clave, esta puede ser constituida por casi cualquier tipo de datos, desde una cadena hasta una representación binaria como una estructura de datos.
- Columnas: son representadas por pares de una familia y un calificador de columna, los cuales están separadas por el símbolo de dos puntos.
- Celdas: consiste en una combinación de fila, columna y versión, la cual contiene un valor, el cual es considerado un arreglo binario ininterrumpido.
- Versión: es considerado un identificador para un valor determinado, este valor es el que se encuentra dentro de una celda, por defecto, este es una marca de tiempo que es generada cuando el valor es grabado, sin embargo, esta puede ser cambiada de ser necesario.
- Familia de columnas: son utilizadas para agrupar columnas de manera física, es decir, que todos los miembros de la familia de columnas son almacenadas juntas dentro del sistema de archivos.
- Calificador de columna: cada uno de estos es para realizar la creación de índices dentro de una familia de columnas.

- Regiones: una tabla es dinámicamente dividida en forma horizontal por regiones, las cuales contienen un conjunto de filas de una tabla, estas son administradas por un RegionServer implementados dentro de un DataNode en HDFS.

3.3.3.2.2. ZooKeeper

Es el nombre de un servicio que provee de mantenimiento a la configuración y sincronización de los diferentes nodos, dentro de un sistema distribuido, el cual provee de interfaces altamente personalizables para alcanzar y mantener persistencia de datos.

HBase hace uso de esta interfaz, por el que realiza una coordinación de servicios dentro de los nodos en el sistema, permitiendo también, la interconexión con el sistema de archivos para su almacenamiento y comunicación con MapReduce, para procesar los datos dentro de la base de datos.

3.3.3.3. Procesamiento: MapReduce

Para la capa de procesamiento se utilizará el entorno de MapReduce que brinda la capacidad de ser adaptado a cualquier algoritmo que sea descrito dentro de su programación y crear la extracción necesaria para realizar un análisis sobre los datos procesados.

Dentro de esta capa se tomarán los datos almacenados para que puedan ser procesados y ordenados de forma que todos pasen a través de un tratamiento con funciones de limpieza, eliminación, filtrado y sustitución de datos, logrando obtener una salida con la mayor precisión para la capa de análisis.

3.3.3.4. Análisis: Mahout

Diseñado e implementado para ser utilizado mayormente sobre Hadoop, Mahout provee de osas herramientas creadas para realizar aprendizaje automático de forma automática, por medio de sistemas escalables y distribuidos, haciendo que la compleja computación se vea disminuida al realizarse de forma paralela.

Por medio de los algoritmos para el aprendizaje automático de forma escalable que Mahout provee, es posible tomar grandes volúmenes de datos realizando un análisis sobre estos y detectar patrones o tendencias, brindando una serie de recomendaciones sobre las cuales basar la toma de decisiones.

3.3.3.4.1. Algoritmos

Debido al extenso número de problemas que pueden ser tratados por medio de la utilización de aprendizaje automático, Mahout provee de distintos algoritmos que permiten solventar cada una de estas necesidades, sin embargo, están basados en dos algoritmos centrales, estos se describen a continuación:

- Filtrado colaborativo: provee de recomendaciones personalizadas por medio de técnicas de filtrado o aplicación de modelos, que realizan el descubrimiento de conocimientos dentro de la información recibida en tiempo real o de la que se posee ya almacenada.
- Agrupación por clasificación: como su nombre lo indica, este algoritmo realiza la agrupación de objetos según sea útil dentro de un conjunto de datos buscando similitudes e identificar relaciones dentro de ellos.

3.3.3.4.2. Tendencias

Al poseer grandes cantidades de datos, como objetivo principal se tiene identificar las tendencias o patrones dentro de estos, debido a las correlaciones descubiertas es posible predecir comportamientos futuros y adecuar acciones con base en el conocimiento adquirido por medio de estas.

Aunque las tendencias o patrones no pueden ser descritos con anterioridad por completo, principalmente se buscarán aquellos que permitan tomar acciones que ayuden a identificar el comportamiento del tránsito vehicular, entre las más importantes en función del tiempo se encuentran:

- Volumen vehicular: encontrar tendencias en el volumen vehicular permitirá identificar las acciones necesarias para otorgar el paso a una vía mediante las correlaciones obtenidas relacionadas al tiempo y conocer la diferente cantidad de vehículos en determinada hora.
- Flujo vehicular: identifica la prioridad de una vía dentro de una intersección, logrando describir las acciones de paso o cese sobre una de ellas, este comportamiento permitirá tomar acciones futuras de una manera más eficiente.

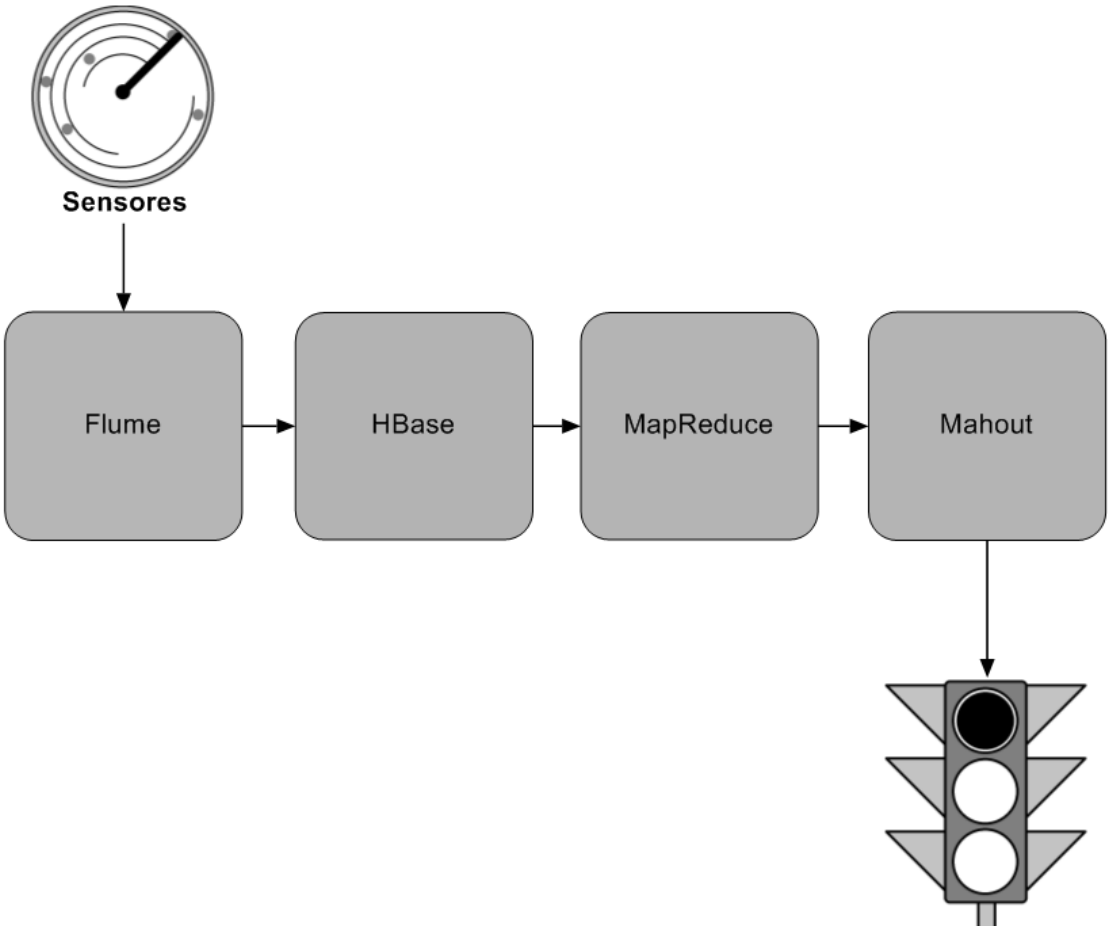
3.4. Proceso de análisis de datos masivos

Tomando como base la arquitectura para el análisis de datos masivos y la arquitectura propuesta para el específico caso de uso, y utilizando las herramientas ya descritas para realizar todo procesamiento de los datos hasta alcanzar resultados, se obtiene:

- Como capa inicial todos los datos serán capturados por medio de los sensores, que brindarán la fuente para el procesamiento, análisis y resultados.
- Por medio de Flume se recolectarán todos los datos capturados por los sensores y serán trasladados hacia donde serán almacenados.
- HBase permitirá que los datos sean almacenados en tiempo real y la posibilidad de que se encuentren grabados de forma distribuida para su procesamiento.
- El procesamiento será realizado por MapReduce, realizando la toma de todos los datos que ya han sido almacenados y realizar todo el tratamiento de estos y prepararlos para el análisis.
- Mahout realizará los algoritmos de análisis por medio de filtrado colaborativo para encontrar patrones dentro de los resultados de los datos ya procesados.
- Por último, por medio de las recomendaciones generadas dentro de los algoritmos de aprendizaje automático se toman las decisiones para los tiempos de paso para cada flujo vehicular mostrado por medio de las señales de control de tránsito.

En la figura 2 se muestra el diseño de la arquitectura propuesta.

Figura 2. **Arquitectura propuesta**



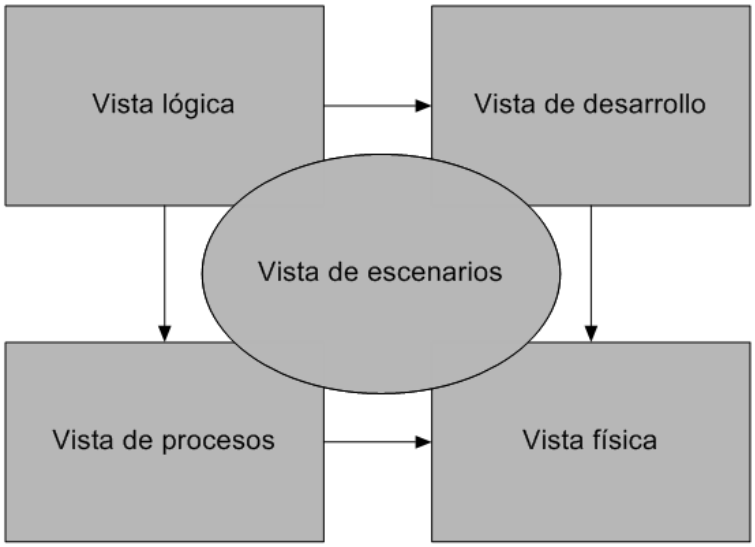
Fuente: elaboración propia, empleando programa ClickCharts.

3.5. Modelo 4 + 1 vistas

Una vista de arquitectura representa uno o más modelos que detallan la arquitectura de un sistema, brindando una completa descripción desde un enfoque en particular.

El modelo 4 + 1 vistas permite documentar un sistema desde distintas perspectivas, permitiendo comprender de manera sistemática la arquitectura de dicho sistema, tanto a desarrolladores, arquitectos, usuarios finales o al resto del equipo. Esta se describe con el siguiente diagrama:

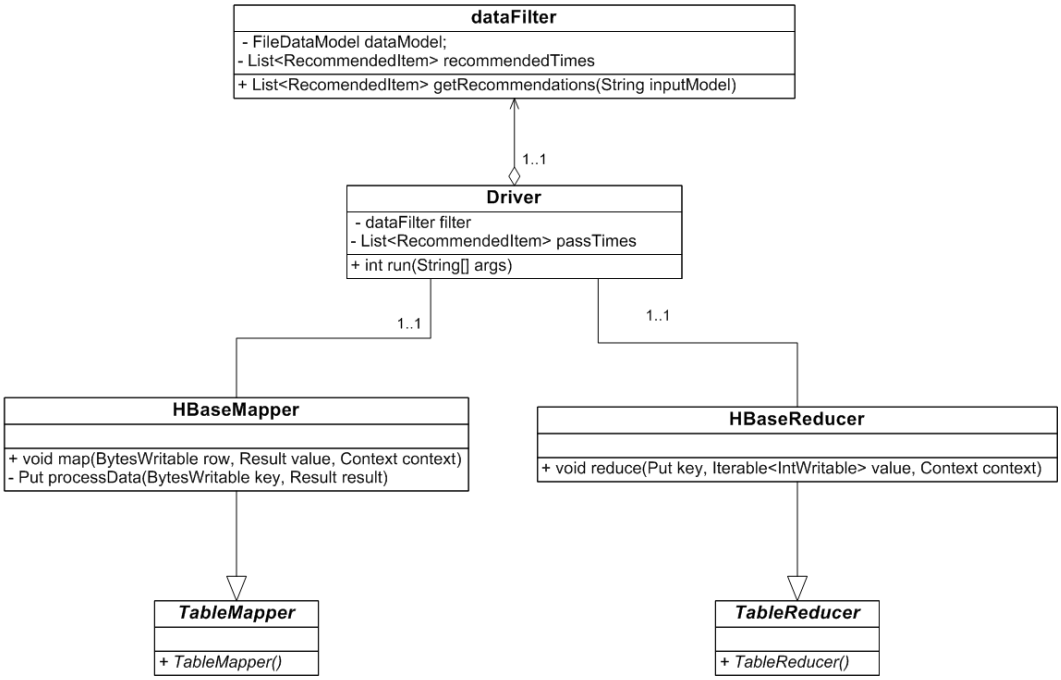
Figura 3. **Modelo 4 + 1 vistas**



Fuente: elaboración propia, empleando programa ClickCharts.

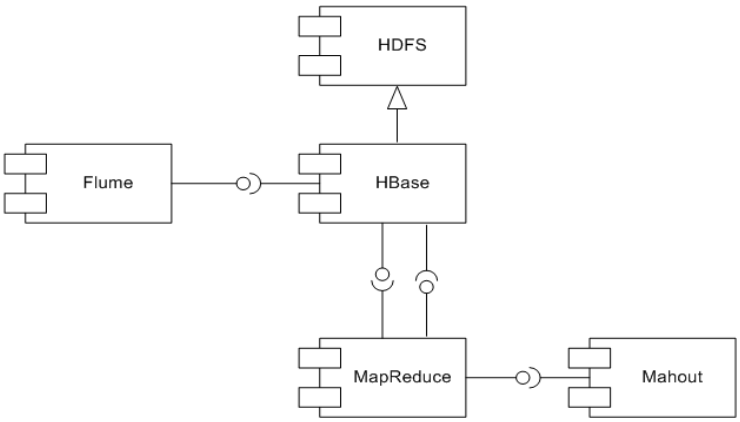
En las figuras de la 4 a la 8 se describe cada vista propuesta por el modelo.

Figura 4. **Vista lógica**



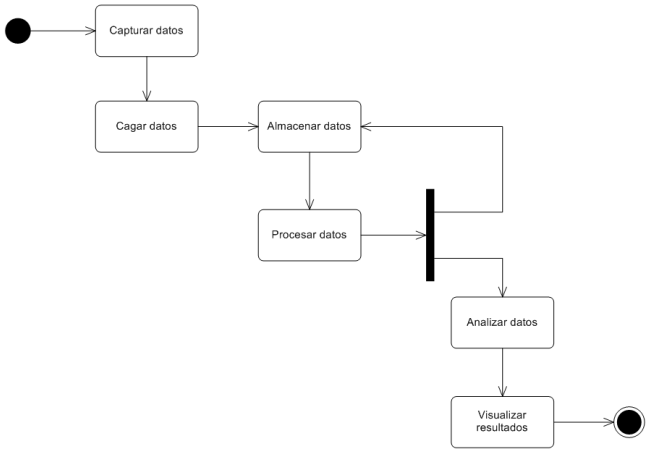
Fuente: elaboración propia, empleando programa ClickCharts.

Figura 5. **Vista de desarrollo**



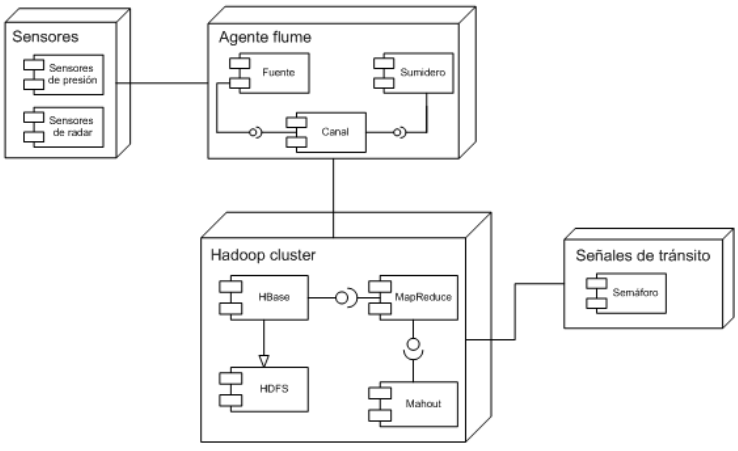
Fuente: elaboración propia, empleando programa ClickCharts.

Figura 6. **Vista de procesos**



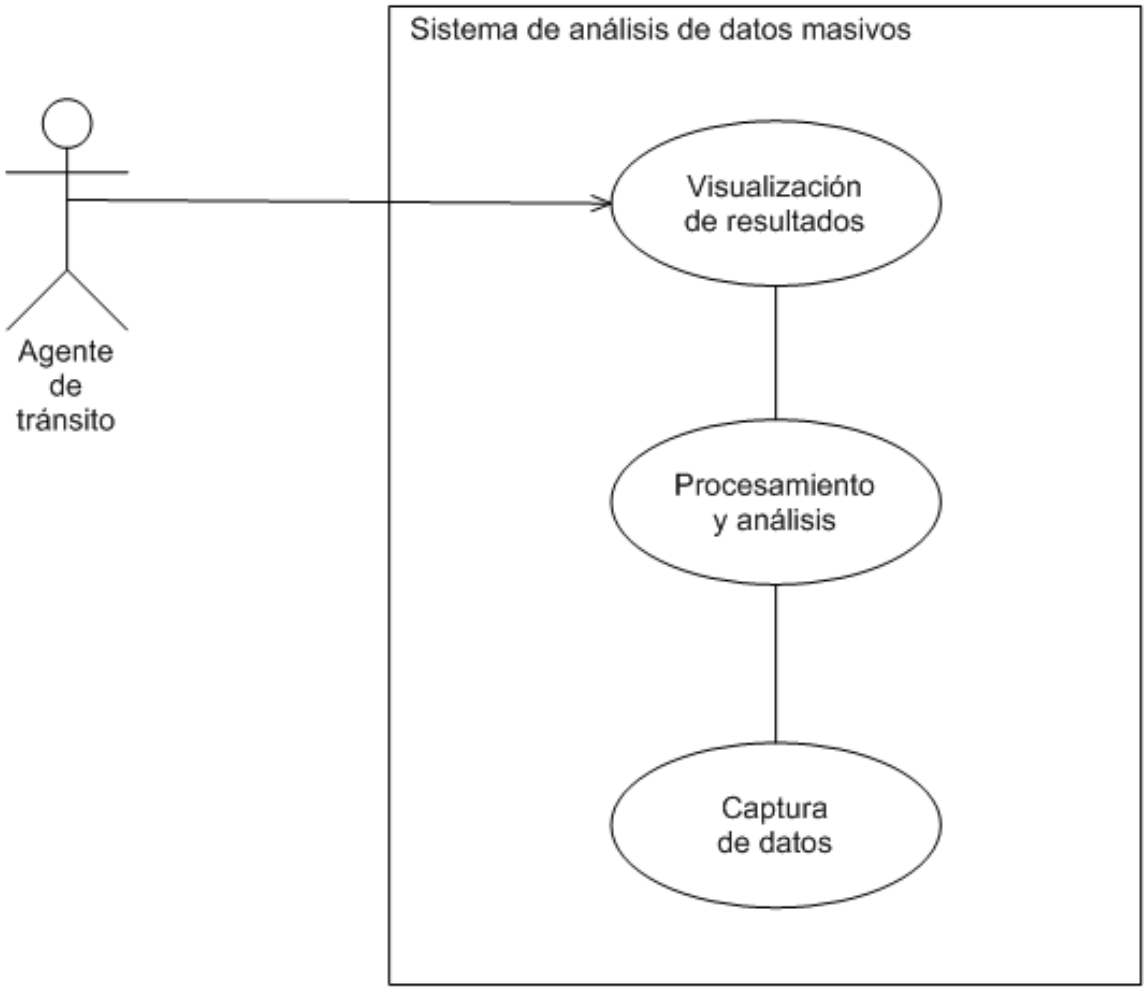
Fuente: elaboración propia, empleando programa ClickCharts.

Figura 7. **Vista física**



Fuente: elaboración propia, empleando programa ClickCharts.

Figura 8. **Vista de escenarios**



Fuente: elaboración propia, empleando programa ClickCharts.

4. ANÁLISIS DE FACTIBILIDAD

4.1. Introducción

Se determinará la factibilidad de que, por medio de la arquitectura propuesta se puede alcanzar el objetivo de contribuir a la disminución de intervención humana dentro del tránsito vehicular, utilizando como herramienta tecnológica el análisis de datos masivos.

Se realizará el análisis por medio de tres aspectos de importancia: factibilidad técnica, factibilidad operativa y factibilidad económica, por los cuales se alcanzará el conocimiento necesario para saber si el uso de esta tecnología es viable para alcanzar los objetivos planteados.

El análisis será realizado dentro de un ámbito tecnológico, donde podrán ser evaluadas cada una de las ventajas y desventajas que posee la implementación de esta herramienta, y así determinar su viabilidad dentro del entorno de su implementación.

4.2. Factibilidad técnica

Para este aspecto se busca describir todos los componentes necesarios para realizar la implementación de esta solución, brindando un amplio criterio de cada uno de los aspectos importantes dentro del punto de vista técnico.

4.2.1. Hardware

Se define a todo componente tangible dentro de un sistema de computación, por lo que será detallado cada uno de ellos dentro de la arquitectura propuesta.

- Sensores: como parte inicial y fundamental de la solución propuesta, estos sensores serán los encargados de realizar la captura de los datos para su posterior procesamiento, cada uno de los sensores deben cumplir con las especificaciones necesarias que permitan realizar la tarea de forma continua y brindar de un flujo continuo de datos en tiempo real. Los sensores propuestos se describen en la tabla 1.

Tabla I. **Sensores necesarios**

Tipo de sensor	Funcionamiento
Sensor de presión	Capturar los datos relacionados al volumen vehicular y obtener datos en diferentes horarios sobre la densidad de vehículos.
Sensor de radar	Capturar los datos sobre las intersecciones entre dos flujos vehiculares y la velocidad promedio de ellos.

Fuente: elaboración propia.

- Computadoras: debido a que la solución se basa en los fundamentos de una computación distribuida y paralela, no es necesario que cada nodo cuente con alto poder de procesamiento, sino que por medio de escalabilidad horizontal, aumentar ese poder de procesamiento como conjunto. Los requerimientos mínimos se describen en la tabla 2.

Tabla II. **Requerimientos mínimos por nodo**

Procesador	2 cuatro núcleos de 2.5 GHz
Memoria	16 GB ECC RAM
Almacenamiento	4 TB discos SATA
Conexión de red	Gigabit Ethernet

Fuente: WHITE, Tom. *Hadoop: The Definitive Guide* p. 296.

- Semáforos: como parte de la visualización, estas señales de control de tránsito serán las encargadas de presentar los resultados del análisis que se le ha realizado a los datos recolectados, por lo que será necesario utilizar dispositivos de tiempo variable, el cual será modificado según sea necesario.

4.2.2. Software

Cada uno de los componentes utilizados dentro de la arquitectura propuesta para el análisis de datos masivos, se encuentra bajo licencias de software libre, lo que brinda cada una de las características intrínsecas de estos.

Sin embargo, existen compañías que brindan un entorno completo para la utilización de estas herramientas, facilitando un soporte especializado sobre cada una de sus implementaciones, dando un valor extra, lo cual genera un costo.

4.2.3. Infraestructura

Uno de los puntos importantes en estas tecnologías es lograr mantener una infraestructura capaz de sostener todo el flujo de datos dentro de los nodos que se encuentran interconectados, por lo que es de suma importancia poseer una correcta interconexión entre cada uno de los componentes del sistema.

Una forma de mitigar este inconveniente y delegarlo a terceros es utilizando la computación en la nube, dejando el mantenimiento de una infraestructura en las manos de proveedores por medio de acuerdos de servicio y únicamente prestar atención a la administración de cada uno de los componentes del sistema.

4.2.4. Recurso humano

Es necesario contar con el personal con los conocimientos técnicos para la implementación y administración de cada uno de los componentes y contar, también, con las cualidades técnicas para desarrollar los algoritmos para el procesamiento y análisis de los datos de entrada.

El recurso humano debe reflejar que forma parte de un equipo multidisciplinario que permita alcanzar los objetivos por medio de diversas disciplinas más allá del ámbito tecnológico, por medio de la aplicación de técnicas de estadística y modelos de predicción que permitan optimizar la toma de decisiones.

4.3. Factibilidad operativa

Se busca demostrar los cambios a nivel operativo que esta herramienta conllevará, cuáles de estos deberán ser realizados para entrar en funcionamiento, identificando los principales beneficios, procesos y actividades que se alcanzarán con su implementación.

4.3.1. Beneficios alcanzados

Una de las principales razones para implementar el análisis de datos masivos para el tránsito vehicular es que brinda muchos beneficios, desde la reducción de costos hasta la optimización de toma de decisiones, entre los más importantes se pueden mencionar los siguientes:

- Reducción de errores y aumento de precisión: brinda la capacidad de tomar decisiones basadas en correlaciones, lo cual disminuye la selección empírica de tiempos y aumenta la precisión en cómo son seleccionados, lo que repercute en la fluidez del tránsito vehicular.
- Automatización de procesos manuales: el proceso de selección de tiempos de paso era arbitrariamente elegido por agentes de tránsito, esta tecnología permitirá que el proceso de selección de tiempos sea decidido autónomamente por el sistema encargado, por medio de un análisis de los datos recopilados.
- Reducción de costos: debido a la posibilidad de funcionar de manera autónoma, permitirá reducir el costo de contar con agentes dentro de intersecciones dando el paso a diferentes las diferentes vías de forma continua.

- Integración sistemática: una de las características más complicadas de alcanzar, pero no imposibles, es contar con la integración de cada una de las señales de control de tránsito y realizar decisiones óptimas de manera sistemática.

4.3.2. Actores del sistema

- Agente de tránsito: son los indicados de velar por el bienestar del tránsito vehicular, y contribuir con la reducción de percances, basándose en sus decisiones para aumentar la fluidez y conducta del flujo vehicular.
- Tránsito vehicular: representa el flujo de vehículos sobre los cuales es realizado el análisis y detectar tendencias o patrones que permitan aumentar la fluidez de estos por medio de los datos generados por ellos.
- Arquitectura propuesta: esta interactuará con el nuevo entorno, desde la captura de datos en tiempo real hasta la presentación de los resultados en forma de tiempos de paso, contribuyendo a la reducción de la intervención de los agentes con el tránsito vehicular.

4.3.3. Resistencia al cambio

La implementación de esta nueva herramienta deberá representar un cambio de pensamiento, ya que debido a la autonomía de esta tecnología, no se presentará de forma continua una figura de autoridad visible, como lo simboliza un agente de tránsito, por lo que la educación vial jugará un papel muy importante en el recibimiento y aceptación de este sistema.

4.4. Factibilidad económica

Tal y como su nombre lo indica será evaluada la factibilidad desde el punto de vista económico, pudiendo detallar los recursos necesarios, tanto para su implementación como para el posterior mantenimiento, luego de que este esté en funcionamiento, finalizado con una comparativa entre el costo que conllevará su realización y los beneficios que este proveerá.

4.4.1. Costos de implementación

- Hardware: una de las ventajas de utilizar la computación distribuida, es contar con nodos que no posean alto poder de procesamiento individual, sino en conjunto aumentar el procesamiento en forma paralela, por lo que el costo de cada uno de los nodos disminuye significativamente en relación a hardware proveniente de compañías que brindan este servicio.
- Software: gracias a las licencias sobre las cuales fueron creados, el software utilizado no presenta ningún costo de adquisición, cabe mencionar que existen compañías que brindan estas soluciones en conjunto, realizando un cobro por soporte técnico y utilización de algunas herramientas de las cuales son propietarios.
- Infraestructura: tal y como se mencionó anteriormente, para la implementación de una infraestructura, existe la posibilidad de poseerla completa de forma local, la cual debe ser mantenida y administrada por personal capacitado, sin embargo, por medio de la computación en la nube, la infraestructura puede ser contenida parcialmente dentro de esta, reduciendo costos inherentes que la misma conlleva, teniendo únicamente de forma local la captura de datos y visualización de información.

- Recursos humanos: para la implementación de esta solución será necesario contar con personal calificado en diferentes áreas para hacer que la solución propuesta sea una realidad y lograr transformar datos capturados dentro del tránsito vehicular, para la optimización en la toma de decisiones.

4.4.2. Costos de mantenimiento

- Hardware: los costos de mantenimiento de hardware serán los sufragados para los sensores y las señales de control de tránsito, realizando una revisión continua a cada uno de sus componentes de manera preventiva y así disminuir el riesgo de fallos.
- Software: los posibles costos que pueden ser representados son los relacionados con soporte técnico, debido a su licencia de distribución gratuita.
- Infraestructura: la computación en la nube provee de servicios elásticos, es decir, que el aumento de poder de procesamiento sea aumentado o disminuido según sea utilizado, también es importante mencionar el mantenimiento que debe ser realizado a la interconexión entre los sensores, estos son los encargados de capturar los datos, y las señales de control de tránsito para la visualización de la información obtenida.
- Recurso humano: como cualquier otro sistema de información, debe ser adecuado con la finalidad de mantener un funcionamiento óptimo de cada uno de los componentes, brindando las actividades preventivas, con el objetivo de disminuir la posibilidad de un fallo, y de algún imperfecto en los componentes del sistema.

4.4.3. Relación costo-beneficio

La implementación de soluciones tecnológicas conlleva costos relacionados, sin embargo, por lo que es preciso determinar de forma clara como dicha solución puede mitigar los costos, en función de los beneficios que provee y así determinar su viabilidad.

Luego de haber determinado todos los costos que estarán relacionados con la implementación de análisis de datos masivos en el ámbito de tránsito vehicular, los beneficios proporcionados por dicha solución son los que deben ser discutidos.

Como objetivo principal, la implementación de esta solución permitirá disminuir la intervención humana dentro de tránsito vehicular, el cual estará basado en la optimización de la toma de decisiones para el cálculo de tiempos de paso para cada flujo vehicular.

Esto disminuirá la exposición del agente de tránsito al flujo vehicular, logrando evitar daños físicos y mantener la integridad de cada uno de los agentes que encuentren sus actividades profesionales relacionadas al control de estos flujos de tránsito vehicular.

Como punto final, la disminución de la intervención del agente de tránsito con las señales de control, permitirá la reducción de costos generados por él, ya que no será necesario que este permanezca de forma continua haciendo cambios manuales en la señalización o gestionando el control del flujo vehicular.

CONCLUSIONES

1. La implementación de una arquitectura para el análisis de datos masivos, dentro del entorno de tránsito vehicular, proporciona la capacidad para analizar el comportamiento de este, logrando identificar tendencias dentro del volumen y flujo vehicular, las cuales permiten tomar acciones más eficientes para la asignación de tiempos de paso o cese sobre una vía, haciendo viable la reducción de la intervención humana por medio de la utilización del análisis de datos masivos.

2. Los datos masivos han representado un gran cambio en la forma en cómo tratar la información, este cambio de enfoque trajo consigo desafíos que pudieron ser resueltos mediante la creación y aplicación de nuevas tecnologías, siendo estas capaces de recolectar, almacenar, procesar y analizar grandes volúmenes de datos de forma más eficiente y rápida logrando encontrar un valor significativo para el negocio.

3. Al poseer grandes cantidades de datos, provenientes de diferentes fuentes y de calidad no completamente confiable, los modelos estadísticos convencionales no proveen de mucha seguridad, no obstante, un análisis capaz de revelar una correlación dentro de estos datos obviando la causalidad subyacente y preponderando la relación estadística entre los valores en los datos lo hace más adecuado para su utilización con grandes volúmenes de datos

4. Las nuevas tecnologías han representado una gran avance para los datos masivos; procesamiento, almacenamiento y demás actividades que anteriormente estaban relegadas a compañías con grandes cantidades de capital para sufragar estos gastos, han llegado a estar alcance de compañías que haciendo uso de tecnologías como Hadoop han alcanzado, también, los beneficios de los datos masivos a un costo menor.

RECOMENDACIONES

1. La arquitectura propuesta fue basada para satisfacer la necesidad de la aplicación del análisis de datos masivos para el caso de uso en particular, por lo tanto se recomienda el estudio de diferentes herramientas que pueden ser de mejor aprovechamiento en requerimientos específicos distintos al caso estudiado.

2. Las fuentes de datos estudiadas fueron únicamente las generadas por el tránsito vehicular en sí mismo, sin embargo, para enriquecer una recolección de datos se recomienda aumentar las fuentes de donde estos son extraídos como el clima o la información generada dentro de las redes sociales.

3. Como foco de estudio únicamente se utilizaron los vehículos, pero es importante mencionar que, existen más opciones que pueden ser analizadas y para poder realizar un análisis holístico de un sistema de control de tránsito es recomendable incluir las demás interacciones como ciclistas o peatones.

4. La tecnología se mantiene en constante reinvención, muchas necesidades surgen creando una exigencia de cambios e innovación, debido a esto es recomendable estar a la vanguardia de cómo la tecnología ayuda a satisfacer las emergentes necesidades.

BIBLIOGRAFÍA

1. Apache Flume. *Flume 1.5.2 User Guide.* [en línea]. <http://flume.apache.org/FlumeUserGuide.html>. [Consulta: 2 de noviembre de 2014].

2. Apache Hadoop. [en línea]. <http://hadoop.apache.org/>. [Consulta: 30 de noviembre de 2014].

3. Apache HBase. [en línea]. <http://hbase.apache.org/>. [Consulta: 4 de noviembre de 2014].

4. Apache Mahout. [en línea]. <http://mahout.apache.org/>. [Consulta: 5 de diciembre de 2014].

5. Apache Zookeeper. [en línea]. <http://zookeeper.apache.org/>. [Consulta: 4 de noviembre de 2014].

6. *Determinación de la factibilidad.* [en línea]. <http://www.itlp.edu.mx/publica/tutoriales/desproyectos/tema3_1.htm>. [Consulta: 25 de noviembre de 2014].

7. *Elaboración e implementación de una maqueta prototipo de semáforo inteligente para la intersección de dos avenidas*. Quito, Perú: Escuela Politécnica Nacional. 2013. 203 p.

8. FRÍAS-NAVARRO, Dolores. Universidad de Valencia. *Inferencia estadística III*. [en línea]. <http://www.uv.es/~friasnav/CorrelacionRegresion.pdf>. [Consulta: 10 de abril de 2014].

9. INGERSOLL, Grant. IBM Developer Works. *Apache Mahout: Aprendizaje escalable con máquina para todos.* [en línea]. <http://www.ibm.com/developerworks/ssa/library/j-mahout-scaling/>. [Consulta: 7 de diciembre de 2014].

10. JAIN, Kalpit. Wag Mobile Inc. *Big Data and Hadoop.* [Versión Kindle]. [Consulta: noviembre 2014].

11. KRUCHTEN, Philippe. *Planos Arquitectónicos: El Modelo de 4+1 Vistas de la Arquitectura de software*. IEEE. 1995.

12. LOZANO, Jose A. Universidad del País Vasco. *Algoritmos genéticos*. [en línea]. <http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/temageneticos.pdf>. [Consulta: 25 de marzo de 2014].

13. MAYER-SCHÖNBERGER, Viktor. *Big data: La revolución de los datos masivos*. Madrid: Turner Publicaciones, 2013. 304 p.

14. MCCLARY, Dan. Dr. Dobb's. *Acquiring Big Data Using Apache Flume.* [en línea]. <http://www.drdobbs.com/database/acquiring-big-data-using-apache-flume/240155029>. [Consulta: 2 de noviembre de 2014].

15. Ministerio de Transporte - Colombia. *Semáforos - Capítulo 7*. [en línea]. <https://www.mintransporte.gov.co/descargar.php?idFile=4288>. [Consulta: 12 de septiembre de 2014].

16. MITCHERLL, Tom. *Machine Learning*. Portland: McGraw-Hill, 1997. 423 p.

17. MORROS, Robert Serrat. *Big Data - Análisis de herramientas y soluciones*. Barcelona: Facultat d'Informàtica de Barcelona – UPC, 2013. 126 p.

18. Municipalidad de la ciudad de Guatemala. *Educación vial - ¡Juntos podemos salvar vidas!* [en línea]. <http://www.muniguate.com/index.php/rb/1307>. [Consulta: 18 de agosto de 2014].

19. _____. *Policía Municipal de Tránsito*. [en línea]. <http://mu.muniguate.com/index.php/component/content/article/1-emetra/15-pmt>. [Consulta: 18 de agosto de 2014].

20. NORVIG, Peter. *Inteligencia artificial un enfoque moderno*. Madrid: Pearson Educación, S. A., 2004. 1179 p.

21. O'Reilly Media, Inc. *Big Data Now: 2012 Edition*. Sebastopol: O'Reilly, 2012. 123 p.

22. PONTE, Francisco. Universidad de Carabobo. [en línea]. <http://www.cid.uc.edu.ve/fponte/ejemplo/factib.pdf>. [Consulta: 22 de noviembre de 2014].

23. PULIDO, Francisco Javier. *BigData & Hadoop (III) - Zookeeper*. [en línea]. <http://www.franciscojavierpulido.com/2013/09/bigdata-hadoop-iii-zookeeper.html>. [Consulta: 4 de noviembre de 2014].

24. RODRÍGUEZ-PIÑERO, Piedad Tolmos. Universidad de Valencia. *Introducción a los algoritmos genéticos y sus aplicaciones*. [en línea]. <http://www.uv.es/asepuma/X/J24C.pdf>. [Consulta: 25 de marzo de 2014].

25. *Sensores y software para semáforos inteligentes en 5 intersecciones de la avenida Luis Gonzales*. Lambayeque: Escuela Profesional de Ingeniería de Sistemas - Universidad Nacional Pedro Ruiz Gallo, 2013. 31 p.

26. Sitio Big Data*. Qué es HBase*. [en línea]. <https://sitiobigdata.com/que-es-hbase/>. [Consulta: 4 de noviembre de 2014].

27. SLUTSKY, Anton. Philadelphia Area Java Users' Group. *Hadoop + Mahout.* [en línea]. <http://phillyjug.files.wordpress.com/2013/03/hadoopmahout.pdf>. [Consulta: 7 de diciembre de 2014].

28. TARIQ, Mohammad. Blogspot cloudfront. *Hadoop Herd: When to use What...* [en línea]. <http://cloudfront.blogspot.in/2013/04/hadoop-herd-when-to-use-what.html>. [Consulta: 30 de noviembre de 2014].

29. The Apache Software Foundation - Blogging in Action. *Apache Flume - Architecture of Flume NG.* [en línea]. <https://blogs.apache.org/flume/entry/flume_ng_architecture>. [Consulta: 2 de noviembre de 2014].

30. Universidad Politécnica de Cartagena. *TEMA 4: Regresión y correlación.* [en línea]. <http://metodos.upct.es/Asignaturas/Diplomatura/Introduccion_estadistica/2008_2009/material_didactico/apuntes/TEMA4REGRESIONYCORRELACION.pdf>. [Consulta: 7 de diciembre de 2014].

31. WHITE, Tom. *Hadoop: The Definitive Guide*. Sabastopol: O'Reilly, 2012. 621 p.