Modelo de Datos para Mejorar las rutas del SITP

1

Integrantes:
Julián Andrés Díaz Rueda - 1018463059
Línea de profundización:
Business Intelligence
Presentado a:
Ingeniero Andrés Armando Sánchez Martín
Universidad Central
Facultad de Ingeniería y Ciencias Básicas
Programa de Ingeniería de Sistemas
Practica de Ingeniería de Sistemas II
Bogotá D.C.
2019 – 05

I. Contenido

II. Resumen

En este proyecto se pretende dar a entender que el SITP este ofreciendo un servicio ineficaz a los usuarios, ya que no tiene en cuenta todas las variables del entorno. Por esto se requiere unificar toda la información para crear informes que contemplen todas las opciones y así poder mejorar la calidad del transporte terrestre en la ciudad de Bogotá.

Para dichos informes se implementaran herramientas con la capacidad de encontrar información en casos que no son fáciles de detectar a simple vista, por eso se hará uso de la ciencia de los datos.

En la ciencia de datos existen conceptos muy relevantes para cualquiera que quiera explorar en los datos. La "Big Data" es uno de los más relevantes pues se quiere intentar entender mucha información que cambia rápidamente, para esto se deben modelar sistemas de bases de datos que sean capaces de organizar dicha gran cantidad de datos y por supuesto devolver resultados en tiempos diminutos.

Por eso este proyecto pretende concentrar toda la información relacionada al transporte masivo terrestre en Bogotá, modelar una base de datos para resguardar dicha información de una manera organizada y coherente y por ultimo dar informes rápidamente para que las empresas puedan tomar decisiones de manera rápida y segura con información confiable y real.

Por ultimo hay que aclarar que no es fácil consolidar procesos de este tipo en una ciudad donde la información es una fuente demasiado resguardad y no se han comenzado aplicar conceptos como OPEN DATA que lo único que buscan es dar información para que las grandes empresas pueden beneficiarse de dichas acciones y así lograr dar un mejor servicio al ciudadano promedio.

III. Introducción

Actualmente en la ciudad de Bogotá existe un medio de transporte masivo terrestre llamado SITP, este presenta muchas quejas por parte de los usuarios, pues al parecer no cumple con las necesidades que estos demandan.

Este sistema SITP tiene como meta abarcar el cien por ciento del transporte terrestre masivo de la ciudad de Bogotá y para ello debe mejorar la infraestructura de la ciudad creando portales y rutas de acceso amplias y seguras para todos los ciudadanos. Además debe cumplir su objetivo principal, que es transportar la mayor cantidad de pasajeros de manera segura y cómoda en el menor tiempo posible. Para esto el sistema debe tener muchas variables en cuenta, como el incremento de la población en los sectores más críticos de la ciudad, horarios de transporte de la ciudadanía, entre otras.

Por eso y más problemáticas que se irán presentando en este proyecto se pretende mostrar las debilidades con respecto a la recaudación de información de este sistema, haciendo la aclaración de que lo que se pretende aquí no es dar solo una crítica al sistema de transporte masivo terrestre de la ciudad de Bogotá, si no también dar una herramienta para que las empresas involucradas puedan mejorar su servicio teniendo un costo beneficio mayor al que actualmente tienen. Sin contar con el hecho de que el pasajero va a pasar de ser considerado una carga masiva a un cliente potencial para las empresas que prestan los servicios de transporte.

Se utilizaran tecnologías amplias y modernas relacionadas con la ciencia de datos donde se verán implementaciones de procesos ya existentes para fines diferentes, esto con el fin de demostrar y crear un nuevo sistema de negocios para el transporte masivo.

Este proyecto pretende dar una nueva visión tanto a empresarios como pasajeros del sistema SITP de la ciudad de Bogotá, y esta visión es el inicio para mejorar las rutas tanto en calidad como en velocidad en un sistema de transporte que deja muchas cosas malas a una ciudad que sigue en constante crecimiento.

IV. Justificación (Contexto, Problemática y oportunidad)

En Bogotá existe un medio de transporte masivo llamado SITP (Sistema Integrado de trasporte público) el cual se encarga de ofrecer las rutas para movilidad de la ciudad. Por esto se diseñó e implemento una infraestructura la cual está compuesta por vías para los servicios troncales, estos son carriles especiales que se acondicionan para el soporte de buses. También existen estaciones las cuales se construyen con la idea de darle velocidad al sistema y que sean de fácil acceso para los pasajeros, estas estaciones son los únicos puntos de parada para los servicios troncales. (Transmilenio SA, 2018).

Existen diferentes tipos de buses, los cuales se componen en buses de color rojo los cuales transitan por vías exclusivas y pueden ser articulados o biarticulados además tienen una capacidad entre 160 y 250 pasajeros. También están los alimentadores que son de color verde y sirven para trasportar a lugares aledaños a los portales de Transmilenio (Estaciones de Gran Tamaño). Los buses zonales son aquellos que se movilizan por las principales vías de la ciudad en carriles mixtos y su principal característica es que son de color azul, adicionalmente se crearon los buses de color naranja los cuales se denominan buses de servicios especiales y tienen como principal función el transporte hacia puntos específicos dentro y fuera del sistema. (Transmilenio SA, 2019)

Ahora, el mantenimiento y control de todo este sistema de transporte masivo terrestre fue asignado por la alcaldía de Bogotá a una compañía llamada TRANSMILENIO S.A, la cual decide celebrar contratos con 10 empresas que se dividen las zonas de la ciudad para poder incluir buses e infraestructura y así poder prestar el servicio. (Transmilenio SA, 2018). Para que estas empresas puedan tener un control sobre las rutas, implementaron en los buses tres elementos claves, un equipo GPS, un computador a bordo (CIBOR) el cual permite intercambiar información operativa entre el centro de control y los buses, por ultimo un sistema de comunicaciones llamado TETRA, por el cual envían y obtienen información del centro de control. (Transmilenio SA, 2013)

Para implementar un transporte masivo terrestre se deben tener varios factores en cuenta, el crecimiento poblacional y estructural son algunos puntos de los cuales más influyen. En los últimos años en Bogotá ha aumentado la población y por ende su infraestructura (Palau, 2013), haciendo que controlar dichas fuentes sea algo muy complejo, pues estos datos no se encuentran en tiempo real y las empresas de transporte no tienen la infraestructura para cuantificar estas variables. En la actualidad

las empresas se centran en el control de los buses para aplicar un plan con una frecuencia de abastecimiento a las estaciones y en caso de llegar a un tope, el centro de control es el encargado de tomar las decisiones sobre qué rutas son las más importantes a satisfacer, es decir las decisiones son tomadas por personas bajo las variables que tienen a su disposición.

Esto genera un gran problema, el mal diseño y planificación de las rutas de transporte, donde las variables mencionadas no son parte del proceso para la toma de decisiones, lo cual produce un efecto negativo para las personas que utilizan el servicio donde el mal funcionamiento que incluye retraso en buses y deterioro de las estaciones son las principales quejas de los usuarios. Además la sobrepoblación en portales como buses hace que la gente tenga una percepción negativa sobre el sistema.

Para mejorar esta perspectiva en los usuarios y en verdad notar un cambio en la mejora del funcionamiento del sistema del transporte masivo en Bogotá se deben tener en cuenta todas las variables posibles y para esto se han creado modelos o sistemas para la optimización y control de transportes, por ejemplo el "Modelo computarizado para toma de decisiones en el transporte aéreo de pasajeros", el cual se basa en la aeronáutica de Colombia y puede ayudar en el aprovechamiento de los recursos para implementarlos de forma más eficiente (González Rivera, Cristancho, & A, 1997). Además existen modelos donde se considera al usuario como cliente y esto hace que se implementen cosas como "definición de la calidad del servicio" haciendo que se obligue a las empresas a prestar un servicio de calidad para el transporte de pasajeros (Fernández & Sergio, 2011) , pero a pesar de implementar modelos y hacer sentir al usuario como un cliente importante se necesita poder recoger la información que no se tiene en consideración, por eso en Londres se implementó un modelo donde Transport for London (TfL) decidió dar la información de manera gratuita, es decir entregar a los desarrolladores los datos de aglomeración, datos de señales de tráfico, datos de los tranvías, datos de los ingresos en las estaciones, entre otras fuentes de información para que estas empresas o desarrolladores diseñaran nuevas aplicaciones logrando crear más de 500, las cuales entregan al usuario y Tfl las mejores rutas disponibles según la aplicación utilizada. (AWS, 2016).

Ahora supongamos que se logra recopilar toda esta información de la ciudad de Bogotá en un almacén de datos donde la información se extrae, transforma y carga. Así se podrá crear un almacén de información que sirva en el sistema integrado de transporte masivo de esta ciudad, el cual se integrará con el sistema actual de buses y centros de control para que puedan tener una mejor visión de la movilidad. Esto con el

fin de poder tomar decisiones considerando todas las variables necesarias para optimizar y mejorar el sistema de transporte, sabiendo que el mismo sistema puede ofrecer las mejores rutas sin que los centros de control tengan que estudiar las soluciones más óptimas, pues este almacén de datos muestra la mejor solución.

V. Base de Conocimiento

1. SITP significa Sistema integrado de transporte público, es un sistema que se encarga de ofrecer e implementar el transporte masivo en la ciudad de Bogotá.

    - Estructura Organizacional.

      El sistema de transporte SITP está estructurado bajo la imagen de Transmilenio S.A, esta es una empresa que se encarga de contratar empresas que cuentan con la flota para poblar las rutas necesitadas por la ciudad de Bogotá.

      Para esto Transmilenio S.A decidió hacer licitaciones tanto para las empresas operadoras de buses troncales como para los buses zonales. Las empresas prestadoras del servicio de transporte son 10 para solo servicios troncales y 9 para servicios zonales, es decir que Transmilenio ha tenido que hacer 19 contratos para poder reunir al 98% de los transportadores en la ciudad de Bogotá. (Transmilenio S.A, 2016)

    - Planeación y generación de servicios.

      El sistema de Transmilenio tiene un sistema de control para los buses, este sistema está basado en un centro de mando que permite la distribución de las rutas para las estaciones, esto se hace con el fin de controlar la velocidad, la frecuencia, los horarios y las rutas de los vehículos. Según Transmilenio S.A esto permite una prestación adecuada del servicio en cada uno de los recorridos de las rutas.

      Para lograr este funcionamiento adecuado los buses deben estar equipados con un equipo GPS (Sistema de posicionamiento Global), que permite saber la ubicación del bus, también los buses están equipados con un computador (CIBOR) que permite enviar y recibir información con el centro de control para así lograr el cumplimiento de cada uno de los buses con sus rutas y tiempos planificados. Por último estos tienen un sistema de comunicación llamado (TETRA, Terrestrial Trunked Radio), el cual sirve para enviar información entre el centro de control, buses y personal de inspección.

Tener esta comunicación entre los buses y los entes de control del sistema constituyen el sistema de creación y manejo de rutas dentro del sistema, bajo la modalidad de toma de decisiones por el personal que compone este centro de control; Estos son los encargados en enviar las rutas a las estaciones más congestionadas con la información que poseen. (Transmilenio SA, 2013)

2. Datos:

   "Información dispuesta de manera adecuada para su tratamiento por una computadora." (Definición diccionario de la asociación de academias de la lengua española).

   "Los datos son la materia prima que emplea el diseño de la información" (Alcalde, 2015)

   En pocas palabras los datos son la representación de un determinado atributo o variable, la descripción codificada de un hecho o suceso.

   - Tipos de Datos

     En la informática cuando hablamos de tipo de dato o tipo nos referimos a un atributo de la naturaleza del dato que se va a procesar. Esto es delimitar o restringir los datos, definir los valores y operaciones que se pueden realizar con esos tipos de datos. Algunos tipos de datos son: Caracteres, Caracteres Unicode, Numéricos. Booleanos.

   - Fuentes de Datos

     En informática se conoce como fuente de información cualquier cosa que pueda ser representado con una señal analógica o digital. El objetivo es procesar, almacenar o transmitir la información.

3. Ciencia de datos.

   La ciencia de datos es un conjunto de métodos científicos que permite a los investigadores emplear técnicas automatizadas para obtener información relevante sobre un conjunto de datos determinado.

- Ramas con las que se desarrolla la ciencia de datos:

  Estas son áreas como la estadística, la minería de datos, el aprendizaje automático y la analística predictiva. La idea de estas ramas en generar es el de recopilar, procesar y extraer valores de las bases de datos; comprender y comunicar conclusiones a los investigadores; crear soluciones basas en datos que aumente los beneficios y reduzcan los costos del problema planteado y por ultimo esta ciencia es aplicable a cualquier área que se quiera investigar, pues es una herramienta de investigación y compresión. (Molinar, y otros, 2017)

- Modelos o Mecanismos

  Existen varios modelos de ciencia de datos como el marketing, gobernanza, Big Data, etc.

  En nuestro caso el modelo que nos interesa es "Big Data" ya que cuando nos referimos a este término hablamos de un conjunto de datos. Pero la idea de la Big Data es trabajar con conjuntos o combinaciones de datos en grandes volúmenes, complejidad alta y gran velocidad de crecimiento, esto lo que hace es dificultar la captura de la información para su análisis y comprensión. Estos modelos implementan sus propias técnicas para resolver los problemas en los cuales se implementan.

- Procesos

  La ciencia de datos tiene procesos para ordenar, mantener y extraer los datos, existen algunos como la encriptación y tokenizacion de Datos, Arquitectura de Datos Unificada, Implementación de Proceso de Datos en equipos (TSDP), entre otros.

  Se mencionan estos tres ya que se quiere enfatizar en el hecho de que es importante primero organizar la información (Tokenizar), unificar datos (Arquitectura) e implementar con metodologías ya comprobadas (TSDP) (Team Data Science Process)

  "El proceso de ciencia de datos en equipo (TDSP) es una metodología de ciencia de datos ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente." (Microsoft, 2019)

- Open Data

Cuando hablamos de "Open Data" se hace referencia a información que cualquiera puede acceder o tratar libremente en cualquier parte del mundo, pero para que un dato se considere Open Data debe cumplir ciertas características:

- Disponibilidad y acceso: Los datos deben mostrarse en un formato muy común para sé que facilite su procesamiento.
- Reutilización y redistribución: Los datos deben ser de uso libre y no tener restricciones para cruzar con otros datos.
- Participación universal: Cualquier persona interesada pueda acceder a los datos y utilizarlos de la manera que prefieran sin restricciones.

Estas tres características se cumplen para datos en ciudades inteligentes o Smart cities. (AWS, 2016) (US Fed News Service, 2016)

4. Bases de Datos

   Se conoce como una colección de información organizada para poder acceder rápidamente a los fragmentos de datos que necesite.

   Una base de datos se compone de una información importante de un mismo contexto, y debe estar organizada de una manera sistemática para recuperar, analizar y transmitir los datos. En la actualidad las personas en sus actividades diarias se ponen en contacto con cualquier tipo de base de datos.

5. Tipos de bases de datos:

   Las bases de datos se clasifican según sus características principales:

   - Según su variabilidad, esto se refiere a la forma de recuperar y preservar los datos.
     - Bases de datos estáticas.
     - Bases de datos dinámicas.
     - Según su contenido, según la naturaleza de los datos.
     - Bibliográficas.
     - De texto completo.
     - Directorios.
     - Especializadas.

- Arquitectura de Sistemas de Base de Datos (ANSI)

  Esto es una arquitectura basada en tres niveles (ANSI – American National Standard Institute - Standards Planning and Requirements Committee) y la idea principal es la de separar los programas de aplicaciones y los datos, el manejo de vistas por el usuarios y la forma de almacenar esquemas de la base de datos. Existen tres niveles los cuales son:

  - Nivel Interno: este describe la estructura física de almacenamiento.
  - Nivel Conceptual: describe la estructura de toda la base de dato y utiliza elementos lógicos como entidades, atributos y relaciones.
  - Nivel Externo a de vista: se encarga de mostrar la visión que se tiene de la base de datos a los usuarios.

- Existe un concepto llamado independencia de los datos el cual se puede definir bajo dos tipos:

  - La independencia lógica: es donde se pueden modificar los esquemas conceptuales sin tener que cambiar los esquemas externos de la base de datos.
  - La independencia física: es donde se pueden modificar los esquemas internos sin tener la necesidad de modificar los esquemas conceptuales, por ejemplo reorganizar entidades de la base de datos.

6. Bases De Datos Relacionales

   Una base de datos relacional es la unión de elementos de datos con relaciones entre ellos. Estos datos se organizan en tablas compuestas por columnas y filas.

7. Bases De Datos No Relacionales o NoSQL

   Estas bases de datos son diseñadas específicamente para modelos de datos que tiene esquemas flexibles, son útiles ya que son fáciles de desarrollar y poseen un excelente rendimiento ya que pueden incluir clave-valor para aumentar su velocidad de búsqueda.

   Tipos de Bases de Datos No Relacionales

   Clave –Valor: Son bases de datos altamente divisibles y se pueden escalar horizontalmente, regularmente se utilizan para tecnologías publicitarias, juegos o Iot (Internet de las cosas). (Ejemplo las historias de Snapchat).

   Documentos: Los datos se representan como objetos o documentos tipo Json y se facilita el almacenamiento y consulta de datos, son flexibles a la hora de crear el modelo de datos pues se puede recibir información semiestructurada y jerárquica. Un ejemplo son los perfiles de usuarios ya que pueden ir evolucionando con el tiempo. MongoDB es un motor de Bases de datos poderoso para un desarrollo flexible e iterativo de este tipo de bases de datos.

   Gráficos: Su principal función es facilitar la ejecución de aplicaciones que funcionan con grupos de datos altamente relacionados. Las redes sociales son un ejemplo de la implementación de este tipo de bases de datos.

   Memoria: Se utilizan para guardar información en tiempo real, tiendas online, lo cual hace que tengan tiempos de respuesta en microsegundos, además pueden soportar grandes picos de tráfico en cualquier momento.

   Buscar: Las aplicaciones generan registros para ayudar a los desarrolladores a solucionar problemas de búsqueda, este tipo de base de datos NoSql ayuda a proporcionar visualizaciones en tiempos real y análisis de datos, esto se logra cuando se indexan las registros en las métricas semiestructuradas de la información que se quiere buscar.

8. OLTP

OLTP (On Line Transaction Processing) en sus siglas en inglés, en español significa procesamiento de transacciones en línea. OLTP es un proceso que facilita y administra aplicaciones transaccionales. Y se basan en arquitecturas cliente - servidor

OLTP se ha utilizado para referirse a la transformación de como un sistema responde a las peticiones de usuarios, esta tecnología se utiliza en cualquier tipo de aplicaciones en la actualidad.

Una transacción es un proceso donde se utiliza el commit para validar o el rollback para invalidar un proceso atómico, las operaciones que incluyen las bases de datos OLTP son inserción, modificación y borrado de datos.

Características de las Bases de datos OLTP:

- El acceso de los datos debe estar optimizado.
- Los datos se deben estructurar.
- Los formatos de los datos no deben ser uniformes.
- Se limitan a datos a historiales recientes.

9. OLAP

   Las bases de datos que tienen sistemas OLAP (On Line Analytical Processing), procesos de analítica en línea están orientadas al procesamiento analítico. Por lo general se implementan grandes cantidades de datos y de allí se logra sacar información relevante, informes complejos, comportamiento de los consumidores, tendencias, este sistema es típico de los datamarts.

   Características de los sistemas OLAP:

   - El acceso de los datos usualmente es solo de lectura. Se hacen muy pocas inserciones, modificaciones o borrados.
   - Los datos se deben estructurar bajo los parámetros de negocios ya establecidos.
   - Estos sistemas se alimentan de sistemas operacionales ya existentes, con procesos (ETL).

10. Tipos de Persistencia OLAP.

    Existen tres tipos de persistencia para las bases de datos OLAP, para entender estos tres conceptos es necesario entender que los cubos, las dimensiones y las jerarquías son la base para navegar en una base de datos multidimensional.

    Un principio importante en los sistemas OLAP, es que los tiempos de respuestas obtenidos deben ser consistentes, es decir se debe calcular por adelantado la información. Actualmente se sabe que los sistemas de administración de bases de datos relacionales se pueden utilizar para OLAP y gracias a esto surge el nombre ROLAP (OLAP relacional) y por ende, también está la contra parte, que es el uso de bases de datos multidimensionales (MOLAP), y cada una de estos sistemas tiene sus ventajas y desventajas.

    - Sistemas MOLAP (OLAP Multidimensional)

      Los sistemas MOLAP se desempeñan mejor que los sistemas ROLAP pero tienen la gran desventaja de que tienen problemas de escalabilidad.

La arquitectura MOLAP usa base de datos multidimensionales para proporcionar un análisis, la idea de utilizar estas bases de datos multidimensionales es mostrar un análisis que se pueda ver en varias dimensiones. Para esto hay que hablar de una arquitectura de dos niveles. La base de datos multidimensional que se encarga del manejo, acceso y obtención de datos; por otro lado está el motor analítico, el cual tiene el nidel de aplicación y aquí es donde se encarga la ejecución de los requerimientos OLAP y este motor analítico también posee el nivel de presentación en donde se proporciona una interface para que los usuarios finales puedan ver los análisis.

- Sistemas ROLAP (OLAP Relacionales)

  Los sistemas ROLAP son mucho más fáciles de escalar pero con la desventaja de que desempeñan de una manera menos eficiente, esta arquitectura o sistema tiene como base acceder a los datos almacenados en un datawarehouse y su premisa es que las capacidades OLAP si se pueden utilizar en bases de datos relacionales.

  En este sistema existen tres niveles, El nivel de base de datos usa bases detos relacionales para el manejo de datos, el nivel de aplicación es el motor donde se ejecutan las consultas multidimensionales y el motor ROLAP tiene la funcionalidad de analítica.

  Por lo general esta arquitectura accede directamente a los datos del datawarehouse y puede utilizar optimizaciones en el modelo de datos para mejorar los tiempos de consulta.

- Sistemas HOLAP (OLAP Hibrida)

  Como su nombre lo dice, es la unión de dos tecnologías basadas en arquitecturas OLAP, en este caso combina las arquitecturas ROLAP y MOLAP, HOLAP utiliza lo mejor de cada una de las dos tecnologías, por el lado del desempeño se basa en MOLAP pero para poder tener un modelo de datos sostenible emplea la arquitectura ROLAP para poder hacer un modelo con mejor escalabilidad.

11. Escalabilidad Horizontal y Vertical.

    La escalabilidad es la capacidad del software para adaptarse a las necesidades del negocio a medida que crece la informacion. Existen dos formas de pensar la escalabilidad y estas son la horizontal y la vertical.

- Escalabilidad Vertical: Es cuando se opta por mejorar el hardware por uno más potente, como un disco duro más grande, más memoria o un procesador mejor, si se mira rápidamente, el esfuerzo de hacer esta escalabilidad es mínimo pues solo se debe migrar el sistema al nuevo hardware.
- Escalabilidad Horizontal: Este sin duda es la mejor opción para tener un buen sistema escalable, pero es mucho más complejo pues requiere un nivel de abstracción amplio. Esta escalabilidad requiere tener servidores conocidos como nos y la finalidad es repartir el trabajo en varios nodos, Por ejemplo para aplicaciones Java se tienes servidores de aplicaciones como WebLogic o Jboss.

12. Latencia:

    La latencia es lo que tarda llegar un dato de un punto a otro, en pocas palabras es el valor que mide el desfase temporal entre un servidor y el cliente.

13. ETL (Extract,Transform y Load)

    Como su nombre lo indica ETL son los procesos que se encargar de extraer, transformar y cargar los datos en una base de datos, para entender las claramente este concepto hay que tener claro que los ETL permiten mover datos desde múltiples fuentes a las bases de datos denominadas datawarehouse.

    Fases de un ETL:

    - Extraer: Esta fase se encarga de sacar los datos desde una o varias fuentes de información o sistemas.
    - Transformación: Esta fase se encarga de formatear y limpiar los datos extraídos cuando sea necesario.
    - Carga: Esta fase se encarga de colocar los datos en grandes bases de datos como los datawarehouse con el de analizarlos.

14. DataWareHousing

    Como su nombre lo dice es una técnica para el almacenamiento de datos, está consiste en recopilar y gestionar datos de diversas fuentes para proporcionar información. Esta técnica mezcla tecnologías y componentes que permiten un aprovechamiento de los datos.

    La mayor importancia de esta técnica es poder crear un proceso para transformar datos en información relevante para que los usuarios la utilicen de manera apropiada y así generar una diferencia.

15. DataWareHouse

    Un almacén de datos como su nombre lo indica, es un depósito donde se consolida la información de una o más fuentes de datos, estas fuentes de datos pueden ser bases de datos relacionales o NoSql.

    Los datos que se depositan en estos almacenes pueden ser estructurados, semi-estructurados o datos no estructurados y la forma de consultar los datos de un datawarehouse es utilizando herramientas de Business Intelligence, clientes SQL o Hojas de Cálculo.

    El objetivo principal de crear un datawarehouse es el de poder analizar a los clientes de manera más integral y así asegurar que se han considerado toda la información disponible para una mejor toma de decisiones.

    Tipos de Almacenamiento de datos:

    - Entrerprise DataWareHouse: Es un almacén de datos centralizado donde proporciona un apoyo a las decisiones de toda la empresa, la idea es unificar la información de la empresa de forma organizada y representarla según la clasificación de la empresa.
    - Almacén de datos operacionales (ODS): Se utiliza cuando ni los sistemas OLAP y los almacenes de datos soportan las necesidades de la empresa; Un ODS se actualiza en tiempo real y es muy utilizado para actividades de rutina o con aplicativos que deben mantener una latencia baja.
    - DataMart: Es una subconjunto de almacenes de datos, es decir un área de un empresa puede tener un DataMart donde se centra en una línea específica del negocio.

    Etapas Generales de datawarehouse.

    - Base de datos operacional fuera de línea: En esta etapa se copian los datos al servidor. Esto con el fin de no afectar el rendimiento del sistema de donde se están obteniendo los datos.

- Almacén de datos fuera de línea: Aquí los datos se deben actualizar regularmente desde la base de datos operativa, se asignan y se transforman los datos del datawarehouse para cumplir con sus objetivos.
- Almacén de datos en tiempo real: En esta etapa, los almacenes de datos se deben actualizar cada vez que se realice una transacción en la base de datos operativa.
- Almacén de datos integrado: En esta etapa, los almacenes de datos se deben actualizar continuamente cuando el sistema operativo realice una transacción y luego el datawarehouse debe generar y devolver una transacción al sistema operativo.

16. Metodología para implementar un Datawarehouse.

Para implementar un datawarehouse se den tener tres pasos que a continuación de describen:

- Estrategia Empresarial: Debemos identificar las técnicas, arquitectura y herramientas con las cuales vamos a trabajar, además debemos identificar los atributos y dimensiones del negocio empresarial al cual vamos a tratar.
- Entrega por Fases: La implementación debe realizarse pos fases según los atributos encontrados en el negocio, se puede comenzar implementando las entidades del negocio de manera independiente luego se debe comenzar a integrar dichas áreas entre sí.
- Prototipo iterativo: Un datawarehouse debe ser desarrollado y probado de manera iterativa para así asegurar la persistencia de la información recolectada.

17. Herramientas para administrar un datawarehouse:

- Administrador de carga o componente frontal: se asocia con todas las operaciones de carga de datos, aquí también se manejan las transformaciones para preparar datos que ingresaran al almacén de datos.
- Administrador de almacenes: Este realiza las opciones como el análisis de datos para garantizar la consistencia, tiene a cargo funciones como des normalización y agregaciones, transformación y fusión de datos.

- Administrador de consultas: Aquí se realizan las operaciones operativas, esto quiere decir que se gestionan las consultas y la programación para la ejecución de las mismas.

18. Herramientas de acceso:
    - Informes de datos.
    - Herramientas de desarrollo de aplicaciones.
    - Herramientas OLAP.
    - Herramientas de minería de datos.
19. Herramientas de almacenamientos de datos:
    - MarkLogic.
    - Oracle
    - Amazon RedShift.
    - Pentaho
20. Business Intelligence

    "Se entiende por Business Intelligence al conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización". (Curto Diaz & Conesa Caralt, 2010)

21. Minería de Datos

    La minería de datos busca patrones ocultos, válidos y útiles dentro de grandes conjuntos de datos.

    Los resultados obtenidos de la minería de datos se pueden utilizan en marketing o detección de fraudes, entren otros.

22. Procesos de implementación de minería de datos:
    - Entendimiento de negocio:
    - Entender las necesidades y objetivos del negocio.
    - Hacer balance del contexto actual.
    - Definir los objetivos de la minería de datos del proyecto.
    - Una implementación detallada del plan de minería de datos.

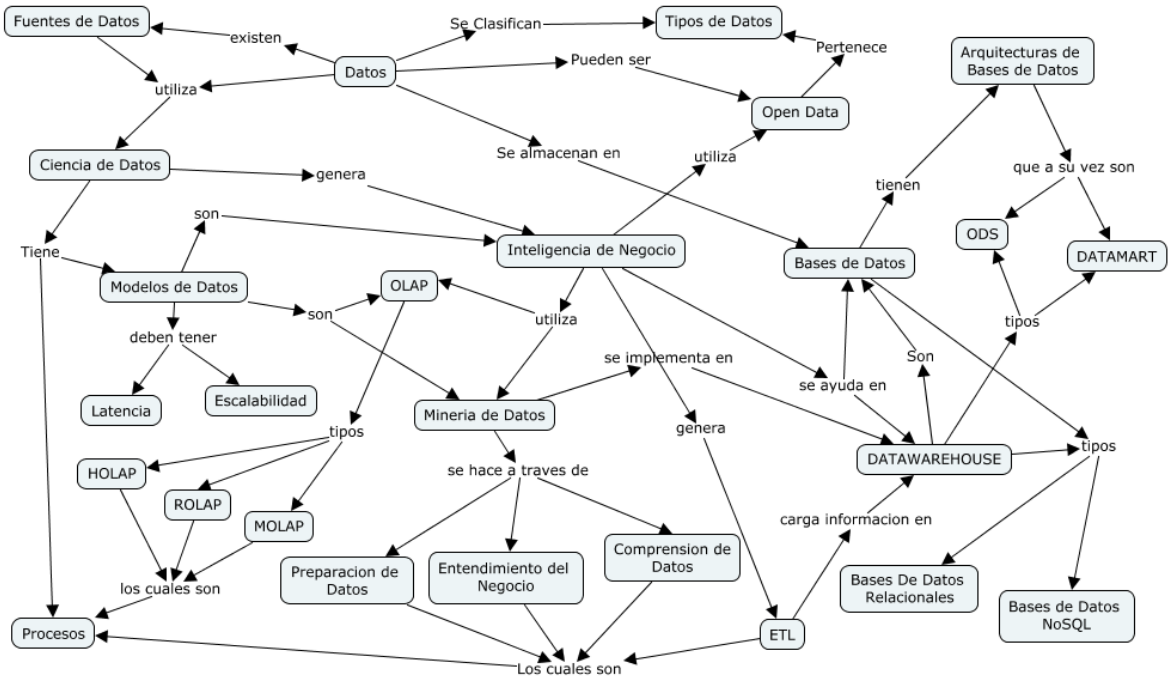
23. Compresión de datos:
    - Recopilar datos de múltiples fuentes de datos.
    - Integración de datos de las múltiples fuentes.
    - Utilizar metadatos para reducir los erros en el proceso de integración.
    - Buscar propiedades de los datos integrados para responder los objetivos planteados en el proceso de entendimiento de negocio mediante consultas.
    - Determinar la calidad de los datos para ver si son confiables.
24. Preparación de datos:

    Esta fase consume el 90% del proyecto ya que aquí es donde se deben validar que los datos son consistentes.

    - Transformación de datos:
    - Suavizado.
    - Agregación.
    - Generalización.
    - Normalización.
    - Construcción de atributos.
25. Modelado: (Modelos matemáticos para determinar patrones)
    - Evaluación.
    - Despliegue.
26. Técnicas de minería de datos.
    - Clasificación: Sirve para recuperar información importante y relevante sobre datos y metadatos para poder clasificarlos.
    - Agrupación: Técnica que sirve para extraer datos que son similares entre sí.
    - Regresión: Sirve para identificar la relación entre variables y así poder calcular su probabilidad.
    - Reglas de asociación: Asocia dos o más elementos para encontrar patrones ocultos en estos conjuntos de datos.
    - Detección externa: Es una técnica que se encarga de observar conjunto de datos que no coinciden con un patrón esperado.

- Patrones Secuenciales: Ayuda a identificar tendencias similares durante un periodo de tiempo determinado.
- Predicción: Combina las técnicas anteriores con el fin de analizar eventos o instancias correctas y así predecir acciones futuras.

VI. Mapa de Co-Relación de Conocimientos



VII. Estado del Arte

Toma de Decisiones: Alcances de la inteligencia de negocio, tener claridad a la hora de tomar decisiones sobre problemas de transporte terrestre, esto con el fin de ver utilidad en los informes realizados.

Cliente Potencial: Cambiar la visión de la empresa sobre los pasajeros, mostrar que estos son un cliente potencial, esto con el fin de incentivar a las empresas a utilizar inteligencia de negocio como alternativa para satisfacer las necesidades de los pasajeros.

Técnicas de inteligencia de negocio: Las herramientas como OLAP o minería de datos son técnicas que ayudan a entender claramente la inteligencia de negocio de una empresa, estas se deben aplicar en un almacén de datos que se tiene que implementar de una forma iterativa según la metodología seleccionada.

Actualización de Datos: Implementación de las técnicas para la actualización de datos, con el fin de mantener un almacén de datos funcional y coherente.

Almacenamiento de Datos: Consolidación de datos como la información Open Sourse, captación de información y persistencia en los datos sobre las rutas de transporte terrestre seleccionadas, esto con el fin de mostrar que la información despreciada si es relevante para la toma de decisiones de la empresa.

Informes de transporte en Colombia: Obtener informes con minería de datos u OLAP en la demanda de transporte en Colombia.

VIII. Pregunta y Objetivos

Pregunta Generadora:

¿Cómo mejorar el diseño y planificación de las rutas del SITP de Bogotá, consolidando la información generada por distintas fuentes en un sistema que extraiga, transforme y almacene los datos para aplicar procesos de analítica?

Objetivo General:

Modelar un almacén de datos (DataWareHouse) que sirva para consolidar la información, entregando cubos de datos con los cuales se van a diseñar y planificar rutas de transporte SITP.

Objetivos Específicos

- Identificar las fuentes de datos relevantes sobre el tráfico, desplazamiento y servicios de transporte de la ciudad de Bogotá clasificando la información en sus tipos de datos, con el fin de definir los requerimientos de acceso y uso de la información.
- Implementar Inteligencia de negocio para consolidar la información identificada en un almacén de datos, para poder cargar la información se deben utilizar técnicas, procedimientos y arquitecturas relacionadas con la ciencia de los datos.
- Crear información útil para el SITP. Dichos informes serán dados bajo los parámetros de técnicas establecidos a la hora de implementar la inteligencia de negocio en el datawarehouse.
- Validar que los datos entregados sirvan para diseñar posibles rutas de transporte, comparando la documentación identificada en la fase uno, con la información resultante del proyecto.

IX. Alcance y Limitaciones

A. Alcances:

- El proyecto entregara **inteligencia de negocio** para rutas identificadas y establecidas.
- La información **identificada** está basada en el sistema de transporte SIPT de la ciudad de Bogotá.
- Cubos OLAP, los cuales **contendrán** información relevante y organizada.
- El proyecto entregara información para que las rutas puedan mejorar los tiempos de respuesta en las estaciones.
- Creación de DataWareHouse con las tecnologías seleccionadas en la fase de identificación y planeación.
- Mejorar la calidad del servicio para el usuario final.

B. Limitaciones:

- Los Datos NO son fuentes Open Data para acceder fácilmente.
- Inversión del proyecto tiene un costo muy alto para todo el sistema.
- El tiempo estimado para el desarrollo del proyecto es de 36 semanas, las cuales se distribuyen en los tiempos establecidos por la materia de práctica de ingeniería en el plan de estudios de ingeniería de sistemas de la universidad central.
- Por el tiempo establecido para el proyecto se seleccionara una cantidad de rutas específicas las cuales quedan a decisión del ponente del proyecto.
- Los resultados serán visibles a través del administrador de datos que se seleccione al momento de especificar las tecnologías a utilizar.

X. Metodología

Teniendo en cuenta los tiempos límites para realizar el proyecto se debe crear una metodología en la cual se pueda realizar todas las actividades dentro del tiempo establecido. Por esto se debe implementar la metodología correcta, con la idea de lograr los objetivos específicos ya planteados anteriormente. Para este proyecto donde se quiere mejorar las rutas de transporte se utilizará una metodología secuencial donde la recolección, el análisis de datos, la implementación y el despliegue son sus fases principales.

El siguiente diagrama muestra el orden de las tareas y como se van a ir resolviendo:

Fase 1: Identificación

Actividades:

1. Verificar si existe la BI (Inteligencia de negocio)
2. Datos Accesibles.
3. Selección de tecnologías (Pentaho)
4. Selección de Arquitectura de Base de Datos.
5. Selección de Modelado BD y Escalabilidad.
6. Selección de Servidor de Aplicaciones.

Entregables:

1. Documento con la información recolectada y seleccionada

Fase 2: Obtención

Actividades:

1. Aparición de procesos formales para toma de decisiones.
2. Configuración de entorno.
3. Se estable un proceso de obtener la información para la toma de decisiones basado en las OLTP.

Entregables:

1. Informes con los procesos basados en OLTP

Fase 3: Creación

Actividades:

1. Crear un almacén de datos
2. Almacén de datos crece y el informe se formaliza
3. Se comienza a precargar la información para los procesos OLAP

Entregables:

1. Informes con los procesos basados en OLTP
2. Documentación del almacén de datos creado.

Fase 4: Diseño

Actividades:

1. Deshacerse de los informes de OLTP. 2. Comenzar con el Despliegue de OLAP.
2. Comenzar a utilizar el sistema en los procesos de toma de decisiones.
3. Business Intelligence se formaliza y se crea la necesidad de establecer nuevos procesos de negocio como DataMart.

Entregables:

1. Informes con los procesos basados en OLTP
2. Documentación del almacén de datos creado.
3. Consultas OLAP
4. Documentos validando la comparación entre los documentos OLTP y OLAP

XI. Cronograma

Actividades	Inicio	Periodos											
		1	2	3	4	5	6	7	8	9	10	11	12
Practica de Ingeniería 3 - Identificación													
Verificar si existe la BI	1												
Datos Accesibles													
Pentaho													
Arquitectura BD													
Modelado BD y Escalabilidad													
Servidor de Aplicaciones													
Practica de Ingeniería 3 – Obtención													
Aparición de procesos formales													
Configuración de entorno													
Proceso información OLTP													
Practica de Ingeniería 4 – Creación													
Crear un almacén de datos													
AD Crece y el informe se formaliza													
Precargar la información OLAP													
Practica de Ingeniería 5 – Diseño													
Comenzar con el Despliegue de OLAP													
Comparar OLTP contra OLAP													
Utilizar OLAP en toma de decisiones													
BI se formaliza y aparecen DataMart													

XII. Conclusión y Proyección de su Práctica

Párrafo que describe la proyección de su proyecto y como planea desarrollarlo en las practicas futuras.

Conclusiones

- Los modelos de BI no son fáciles de implementar, ya que requieren un nivel de abstracción alto.
- La BI es una forma de obtener información en cualquier empresa en la cual se maneje grandes volúmenes de datos que sean difíciles de estructurar.
- La BI siempre debe estar en constante modelamiento ya que la información va ir cambiando según los sucesos del negocio.
- La BI es una técnica útil y eficiente para mejorar los procesos de la empresa que lo implemente.

Lecciones aprendidas y experiencia

1. El proyecto pretende ser una base sólida para la implementación de una tesis de grado.
2. Lo desarrollado sirve de plataforma para incentivar la mejora del transporte terrestre en Bogotá con tecnologías de última generación.
3. Culminar con las cinco prácticas de ingeniería con un proyecto estable y duradero para la universidad.
4. Insignia para demostrar que la inteligencia de negocio puede ser implementada en cosas diferentes a los campos empresariales.

XIII. Bibliografía y Referencias

Alcalde, I. (2015). Visualización de la información : de los datos al conocimiento. Barcelona: Editorial UOC.

AWS. (17 de Abril de 2016). https://aws.amazon.com/. Obtenido de https://aws.amazon.com/: https://aws.amazon.com/es/blogs/publicsector/new-transport-for-london-open-data-sets-available/

Ballesteros Silva, P. P., Valencia Bonilla, M. B., & Hernandez amariles, j. d. (2015). Diseño, desarrollo y validación del sistema de información de transporte y mensajería de audifarma SA (SITA). Scientia et Technica, 345-351.

Curto Diaz, J., & Conesa Caralt, J. (2010). Introducción al business intelligence. Barcelona: Editorial UOC.

Fernández, G., & Sergio, c. (2011). La calidad en el transporte público de pasajeros. Madrid: Asociación Española de Normalización y Certificación (AENOR).

González Rivera, C., Cristancho, Q., & A, J. (1997). Modelo computarizado para toma de decisiones en el transporte aéreo de pasajeros. Bogotá: Bogotá Universidad Central.

Hurtado Tarazona, A., Hernández Ospina, M., & Miranda Ruiz, L. (2014). Gestión de grandes proyectos urbanos en espacios metropolizados : los sistemas integrados de transporte masivo en Colombia. Bogotá: Bogotá Universidad Piloto de Colombia.

Mallach, Efrem, 1942-. (2000). Sistemas de soporte de decisiones y data warehouse. Nueva York: Singapur McGraw-Hill.

Microsoft. (26 de 02 de 2019). https://docs.microsoft.com. Obtenido de https://docs.microsoft.com: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Molinar, a., Mario, C., Espinoza, M., Pedro, Ocho, L., & Ileana. (2017). Evaluación de destinos turísticos mediante la tecnología de la ciencia de datos. Estudios y perspectivas en turismo, 286-305.

Palau, J. J. (2013). Análisis del transporte masivo y la movilidad en Bogotá . Universidad & Empresa, 15-23.

Ricardo, C. (2009). Bases de datos. McGraw-Hill Interamericana.

Ruiz, J. A., & Simeón, R. (2011). Análisis y síntesis de sistemas de ingeniería para la preparación y toma de decisiones bajo criterios múltiples. Revista de Matemática: Teoría y Aplicaciones, 67-91.

Sánchez Gutiérrez, J. (2012). El sistema de transporte público en español: una perspectiva interregional. Cuadernos de Economía, 195-228.

SIE7E. (23 de 02 de 2019). http://www.sieteprogramacion.com/. Obtenido de http://www.sieteprogramacion.com/: http://www.sieteprogramacion.com/

Transmilenio S.A. (07 de Septiembre de 2016). https://www.transmilenio.gov.co. Obtenido de https://www.transmilenio.gov.co: https://www.transmilenio.gov.co/publicaciones/146276/operadores_del_sitp/

Transmilenio SA. (13 de Septiembre de 2013). https://www.transmilenio.gov.co. Obtenido de https://www.transmilenio.gov.co: https://www.transmilenio.gov.co/publicaciones/146195/sistema_de_control/

Transmilenio SA. (23 de agosto de 2018). https://www.transmilenio.gov.co. Obtenido de https://www.transmilenio.gov.co: https://www.transmilenio.gov.co/publicaciones/146180/infraestructura/

Transmilenio SA. (18 de Diciembre de 2018). https://www.transmilenio.gov.co. Obtenido de https://www.transmilenio.gov.co: https://www.transmilenio.gov.co/publicaciones/146194/empresas_operadoras/

Transmilenio SA. (25 de enero de 2019). https://www.transmilenio.gov.co. Obtenido de https://www.transmilenio.gov.co: https://www.transmilenio.gov.co/publicaciones/151051/buses-de-transmilenio/

US Fed News Service. (2016). HOW THE WORLD BANK AND LONDON ARE HELPING OTHER CITIES WITH THEIR TRANSPORT SYSTEMS. US Fed News Service, Including US State News .

Varela, J., Arias Rodríguez, J. E., Cotos Yáñez, J. M., Sordo, y., & Triñanes Fernández, J. A. (2002). Sistema de apoyo a la toma de decisiones para el despliegue de medios de defensa contra incendios forestales. Geofocus: Revista Internacional de Ciencia y Tecnología de la Información Geográfica.