

**Como requisito para la obtención del grado de**

Ingeniería de Sistemas y Computación

Universidad de los Andes

Bogotá, Colombia

Diciembre 2020

Presentado al

Departamento de Ingeniería de Sistemas y Computación

# Expansión de análisis y visualización de la infraestructura aeroportuaria y terrestre de Colombia desde la iniciativa Infraestructura visible

**Asesorado por:**

María Del Pilar Villamil

**Documento presentado por:**

Alejandro García Flores

Contenido

[Propuesta Proyecto de Grado 1](#_Toc50413608)

[1 Introducción 1](#_Toc50413609)

[1.1 Motivación: 1](#_Toc50413610)

[1.2 Descripción del problema: 1](#_Toc50413612)

[2 Descripción general 1](#_Toc50413613)

[2.1 Objetivo general del proyecto: 1](#_Toc50413614)

[2.3 Objetivos específicos del proyecto: 2](#_Toc50413615)

[2.3 Cronograma de desarrollo del proyecto: 2](#_Toc50413616)

[3 Referencias 3](#_Toc50413617)

## 0 Resumen

la iniciativa Infraestructura Visible de la Universidad de los Andes liderada por el profesor Mauricio Sánchez del departamento de Ingeniería Civil y Ambiental busca expandir el número de análisis y visualizaciones de datos que se presentan en la página web: <https://www.infraestructuravisible.org/> particularmente enfocado en la infraestructura aérea y terrestre en Colombia, ya que en estas áreas es donde se contaban con un mayor número de posibles análisis sin explorar y la iniciativa tenía el mayor interés en realizar análisis de datos. Para lo anterior, fue necesario levantar los requerimientos de la propuesta de análisis, realizar una exploración sobre los datos con los que contaba Infraestructura Visible en cuanto vías y aeropuertos, listar e inquirir sobre los análisis ya existentes y proponer, implementar y construir la visualización de los análisis aceptados por la iniciativa de acuerdo con su nivel de impacto y factibilidad en la construcción de estos dentro de los 5 meses de desarrollo del proyecto.

A través del anterior proceso se detectaron tres áreas de análisis relevantes, tráfico aéreo en Colombia, cobertura aérea de centros poblados de Colombia y tráfico y recaudo en peajes y vías de Colombia. Para cada una de ellas se realizó un estudio de la calidad de los datos, se realizaron medidas correctivas sobre los mismos de acuerdo con los problemas particulares de cada Dataset y se construyó un tablero de control para cada aérea en desplegado en una página web donde se puede observar el trabajo realizado.

## 1 Introducción

Infraestructura Visible es una iniciativa de la Universidad de los Andes que permite el libre acceso a la información sobre la infraestructura colombiana y su relación con indicadores de desarrollo socioeconómico. El objetivo principal de la iniciativa es compilar, limpiar, estudiar y mostrar los datos sobre infraestructura en Colombia que publican por ley las entidades gubernamentales al público general, de tal manera que estos datos se encuentren centralizados y sean de utilidad, tanto para la academia como para auditoria, periodismo y el público en general. Infraestructura Visible es una iniciativa de la Universidad de los Andes que consolida múltiples fuentes de datos públicos que da el Gobierno Nacional de Colombia acerca de infraestructura en sectores tales como salud, educación, transporte, energía, cárceles y demografía y su relación con indicadores económicos.

El desarrollo de este proyecto fue impulsado principalmente por la falta de recursos de la iniciativa y el deseo de complementar el trabajo de analítica ya realizado durante los últimos tres años sobre los datos públicos de infraestructura en Colombia. Estos datos ya contaban con un adelanto en cuanto a calidad de estos, organización, análisis, despliegue y visualización.

La importancia del proyecto de grado y del trabajo realizado por Infraestructura visible radica en facilitar el acceso a la información pública que existe sobre infraestructura en Colombia y dar a conocer hallazgos de impacto sobre la misma, aumentando el nivel de transparencia y contabilidad que existe sobre los proyectos de infraestructura en Colombia y como estos se relacionan con indicadores socioeconómicos del país. Durante el desarrollo del proyecto se detectaron tres áreas de especial interés para el desarrollo del proyecto desde la iniciativa; tráfico aéreo en Colombia, cobertura aérea de centros poblados de Colombia y tráfico y recaudo en peajes y vías de Colombia. Principalmente por el amplio abanico de análisis que se podían realizar desde estas áreas que no fueron exploradas previamente.

El presente proyecto de grado es presentado con una estructura previamente definida por el departamento de Ingeniería de Sistemas y Computación. Sin embargo, el proceso que se realizó para la elaboración del proyecto fue la metodología Kimball para Data Werehousing/Bussiness Intelligence.

## 2 Descripción general:

Este proyecto pretende implementar dentro del sitio web de Infraestructura Visible, o en un sitio web paralelo al sitio web oficial de Infraestructura Visible un tablero de control donde se pueda realizar exploración y análisis de datos en tres áreas de interés general para la iniciativa de Infraestructura visible. Estas son: tráfico aéreo en Colombia, cobertura aérea de centros poblados de Colombia y tráfico y recaudo en peajes y vías de Colombia. Adicionalmente, se espera que la persona que desee utilizar los tableros de control pueda generar datos interesantes que de cualquier otra forma seria una cantidad de trabajo considerable utilizando los datos que el Gobierno Nacional pública a través de la ley 1712 de 2014 o ley de transparencia al acceso a la información.

En este momento existe un trabajo importante realizado sobre veinte diferentes fuentes de datos en extracción, transformación y calidad de datos. Existe igualmente una cantidad importante de análisis implementados con estos datos a través de una infraestructura de BI que resulta en un visualizador de datos que permite cruzar diferentes variables que existen en los datos. Estos análisis en su mayoría son de un carácter sencillo, puesto que el mayor esfuerzo se a invertido en construir bases de datos de calidad. Sin embargo, diferentes limitaciones en los recursos disponibles, ha limitado expandir el número de análisis que se pueden realizar sobre los datos por medio de nuevos desarrollos.

Por este motivo, el profesor Mauricio Sánchez director de Infraestructura Visible nos ha propuesto desarrollar sobre los datos existentes nuevos análisis que arrojen información desconocida y útil para aquellas personas que consultan el sitio web de Infraestructura Visible.

### 2.1 Objetivos:

Los objetivos del proyecto son:

* Realizar un estudio exploratorio y de calidad de la información en las bases de datos de infraestructura visible en las áreas de interés, tales como aeropuertos, tráfico aéreo, vías, puentes y peajes.
* Proponer nuevos análisis de la infraestructura aérea y vial en Colombia a partir de los hallazgos del estudio anterior.
* Generar un tablero de control que permita visualizar los análisis desarrollados.

### 2.2 Antecedentes.

Previa a la realización de este proyecto en específico, el proyecto de grado que genero la iniciativa de Infraestructura Visible fue elaborado en 2014 por Fernando Patiño y Nicolás Villa, asesorados por Mauricio Sánchez para el departamento de Ingeniería Civil y Ambiental de la Universidad de los Andes. Este era un prototipo en el que se basó el posterior desarrollo de la iniciativa, y donde se realizaron las primeras recolecciones de datos.

Posteriormente, se realiza la implementación y desarrollo de la iniciativa, en donde se recopila información de 12 fuentes de información pública en 9 sectores distintos con 21 subsectores específicos. Los sectores son: vías, energía, demografía, aeropuertos, puertos, educación, salud, hidrocarburos y cárceles.

Adicionalmente, se encontró que el Gobierno de México tiene una plataforma centralizada con un propósito similar a Infraestructura Visible en la página <https://datos.gob.mx/> en donde se pueden encontrar datos abiertos de todos los sectores sociales y económicos de México, incluyendo proyectos de infraestructura. Sin embargo, esta no cuenta con un visualizador de información o análisis específicos de la información.

### 2.3 Identificación del problema y su importancia

Actualmente, la iniciativa de infraestructura visible cuenta con dos bases de infraestructura aérea de Colombia, y una internacional. En la primera se describen todos los aeropuertos de Colombia con diferentes propiedades de infraestructura tales como la longitud de la pista, ancho de la pista, tipo de aeropuerto y clase de aeropuerto. En la segunda base de datos se lleva el registro de todos los vuelos en donde al menos un aeropuerto de Colombia participa entre 2004 y 2018 compilado de carácter mensual por aerolínea, tipo de aeronave, ruta, número de pasajeros que abordaron, carga abordo, carga ofrecida y numero de sillas en la aeronave. En la tercera base de datos se tiene un registro de la ubicación geográfica, siglas y altitud de todos los aeropuertos del mundo.

Estas bases de datos se utilizan en más de 30 análisis y visualizaciones en la página web de Infraestructura Visible. Sin embargo, ninguna de las visualizaciones y análisis cruza la información de más de una tabla, limitándose a representaciones básicas de la información que se encuentra en cada una de las tablas de la base de datos. Del mismo modo, en la página web no es posible interactuar con la información para ver detalles más específicos de información, realizar filtros por ubicación geográfica o alguna variable específica, por lo que una persona que desee un mayor nivel de introspección sobre la información necesita bajar los datos y elaborar sus propias gráficas y conclusiones de manera independiente. Esto también sucede en el área de infraestructura vial de Colombia, que cuenta con 5 bases de datos diferentes.

Al ser uno de los objetivos principales de la iniciativa poder brindar a sus usuarios una experiencia que facilite la comprensión y exploración de los datos abiertos de infraestructura en Colombia, se espera que con la implementación de los tableros de control y análisis escogidos se descubra con mayor facilidad información que de otra forma le llevaría una cantidad de tiempo considerable a cualquier persona consultar o construir de manera independiente. Con esto el proyecto pretende publicar información relevante sobre la infraestructura aérea y vial de Colombia que de otra forma seria de difícil acceso y que requiere conocimiento técnico especializado para construirse.

## 3 Diseño y especificaciones

### 3.1 Definición del problema:

La iniciativa Infraestructura Visible desea expandir el número de análisis y visualizaciones de datos que tienen disponibles en su página web en las áreas de infraestructura y tráfico terrestre y aéreo en Colombia. Específicamente, desean aprovechar las diferentes fuentes de datos que ya han trabajado, incluir nuevas fuentes y desplegar por medio de un tablero de control análisis y visualizaciones de datos que exploren información cruzada entre las diferentes tablas que tienen a su disposición.

### 3.2 Especificaciones:

### 3.2.1 Requerimientos Funcionales:

* El tablero de control debe poder consultarse en línea.
* La información utilizada para la construcción del tablero de control debe poder descargarse desde la página web desplegada.
* El origen de los datos debe estar en un formato común fácil de editar y actualizar.
* Los datos deben poder actualizarse agregando nuevos registros a los archivos originales.
* Los tableros de control deben poder filtrar la información de manera interactiva

### 3.2.2 Requerimientos no funcionales:

* Se debe describir la metodología usada para realizar los análisis de tal manera que sea replicable de manera externa.
* Se debe informar al usuario cuales son las fuentes de los datos originales, y la fecha de consulta de estos.

### 3.2.3 Niveles de aceptación:

* Es necesario que la visualización de los datos en los tableros de control sea clara, contenga colores fáciles de identificar, los datos estén correctamente etiquetados y sea presentable.

### 3.3 Restricciones:

Las restricciones con las que cuenta el proyecto son:

* en primer lugar, la solución de tableros de control implementada tuvo que ser de muy bajo costo, de tal manera que sea sostenible en el tiempo para la iniciativa.
* En segundo lugar, durante el desarrollo del proyecto el estudiante no tuvo acceso a el código fuente de la página web, ni a la documentación de esta debido a negociaciones de la Iniciativa con el desarrollador original, por lo que los tableros de control tuvieron que ser implementados en un sitio web paralelos al de Infraestructura Visible.
* En tercer lugar, todos los datos adicionales a los ya existentes en la página web de infraestructura visible y metodologías utilizadas a lo largo del proyecto tuvieron que ser consultadas para su inclusión con el equipo encargado de la iniciativa y aprobadas por ellos para cumplir con sus estándares de calidad.
* Datos adicionales utilizados en el desarrollo del proyecto deben provenir de entidades públicas y estos deben ser de libre acceso para cualquier persona o si son desarrollados por el estudiante durante el transcurso de proyecto deben poder ser públicos en la entrega final junto a la metodología utilizada para construir los datos.
* El proyecto debe desarrollarse dentro de un semestre académico.

## 4 Desarrollo del diseño

### 4.1 Recolección de la información:

Para poder realizar el proyecto se realizaron ocho reuniones quincenales a lo largo del semestre a partir del 31 de agosto de 2020, el primer mes, por medio de estas reuniones se buscó entender el negocio del cliente, definir sus necesidades y requerimientos para la implementación del trabajo. Durante el segundo mes se exploraron las diferentes fuentes de datos y se realizaron propuestas para los análisis a desarrollar en el proyecto. Para la realización del proyecto se utilizaron las bases de datos disponibles en la página web de Infraestructura Visible durante el periodo de agosto a diciembre del 2020 correspondiente a las siguientes tablas:

* Aeropuertos del mundo.
* Aeropuertos (Colombia).
* Aeropuertos Histórico de datos.
* Tabla Peajes.

Desde el sitio web de la Agencia Nacional de Infraestructura (ANI) se utilizaron los datos de tráfico y recaudo de peajes entre 2014 y 2018 que correspondían a las siguientes tablas de Excel consultadas por última vez en diciembre de 2020:

* Resumen Consolidado Mensual por estación de Peaje de Tráfico, TPD y Tráfico x Tarifa 2018
* Resumen Consolidado Mensual por estación de Peaje de Tráfico, TPD y Tráfico x Tarifa 2017
* Resumen Consolidado Mensual por estación de Peaje de Tráfico, TPD y Tráfico x Tarifa 2016
* Resumen Mensual por estación de Peaje de Tráfico, TPD y Recaudo 2015
* Resumen Mensual por estación de Peaje de Tráfico, TPD y Recaudo 2014

Desde el sitio web del Instituto Nacional de Vías (INVIAS) se utilizaron las tablas CSV disponibles desde su API de datos abiertos correspondientes a la base de datos de Peajes y Vías en Colombia.

Desde el DANE se utilizó la Tabla de Cabeceras municipales y Centros poblados – Colombia en formato CSV, consultada por última vez en noviembre de 2020.

### 4.2 Alternativas de diseño

Para crear el producto final, que es un tablero de control ***en la nube*** se dividió el problema en 5 capas: Recolección, integración, almacenamiento de datos, análisis y presentación. De cada una de estas capas de consideraron diferentes opciones que permitieran la implementación del producto final con los requerimientos del cliente. Las alternativas de diseño se dividieron en dos posibles soluciones, de las cuales se escogió la segunda:

1. La primera opción consistía en utilizar una Máquina Virtual en una plataforma de web servicies tal como Google Cloud Services, AWS o Azure para crear una infraestructura tradicional de BI en un servidor único. En la capa de recolección los datos se almacenarían dentro de la maquina virtual en forma de archivos CSV. Para la integración de los datos se utilizarían scripts en Apache Kafka o similares para recuperar datos externos por medio de APIs regularmente que aplicarían transformaciones sobre los datos para volverlos usables y los aplicarían a la solución de almacenamiento. Posteriormente, para implementar la capa de almacenamiento se utilizaría PosgreSQL o similares para crear un Data Mart que almacenaría todos los datos. A continuación, para la capa de análisis y presentación se utilizaría un software de BI tal como Pentaho, Tableu o Power BI para diseñar la visualización de los datos en los tableros de control y que estos fueran accesibles por medio de puertos, finalmente utilizar un proveedor de dominios para configurar la página web final que mostraría los tableros de control directamente sobre el software de BI. No se escogió esta opción por las siguientes razones: mantener una Maquina Virtual activa dentro de cualquier plataforma conduce a un costo eventual, en AWS solo se ofrece el servicio gratuito por un año, en GCS el costo mínimo es de tres dólares mensuales para la maquina con menor especificación y en Azure el costo es de 0.0075 centavos de dólar por hora de uso, ninguno es gratuito, por lo que la segunda opción era más favorable.
2. La segunda opción consistió en utilizar infraestructura de plataformas de web servicies que permitiera realizar todo o una parte importante del proceso desde sus soluciones integradas. Dentro de este grupo de soluciones se examinaron tres posibilidades: Amazon Web Services, Azure y Google Cloud Services. Debido al ‘‘Free Tier’’ que ofrece Google Cloud Services se decidió implementar la solución en esta plataforma ya que no se espera que el sitio web para el proyecto supere el número límite de operaciones gratuitas.

Dentro de Google Cloud Services, para la capa de recolección de datos se utilizo Google Storage, que funciona como un Bucket o repositorio de datos, se utilizó Google BigQuery como Data Mart y centro de ETLs ya que este permite consultar las fuentes externas directamente y aplicar operaciones programadas sobre los datos, se utilizó Google Data Studio como capa de análisis para crear los tableros de control y, finalmente, se creo un sitio web HTML hosteado en Github con los tableros de control embebidos en el código para crear la fachada.

## 5 Implementación

Para el desarrollo del proyecto se utilizó una aproximación a una iteración de la metodología Kimball para el ciclo de vida de DW/BI. En esta metodología la primera etapa se realiza la definición de los requerimientos del negocio, en la segunda etapa se realiza el modelado dimensional en conjunto con el diseño de la arquitectura técnica y se escogen los productos a utilizar, en la tercera etapa se desarrollan las ETLs y los tableros de control, y finalmente, en la última etapa se realiza el despliegue del producto. (Kimball, et al. 2008)

### 5.1 Descripción de la implementación

Durante la primera fase del proyecto, en donde se definieron los requerimientos del negocio, se realizaron 4 reuniones en un periodo de 8 semanas en donde, en primer lugar, se escogieron las áreas de interés del proyecto, en segundo lugar, se exploraron los análisis y datos ya existentes en Infraestructura Visible, y en tercer lugar se propusieron nuevos análisis y fuentes de datos para realizarlos. Las áreas de interés escogidas por el equipo de Infraestructura Visible fueron las siguientes:

1. Infraestructura aérea de Colombia.
2. Infraestructura Vial de Colombia.
3. Infraestructura marítima de Colombia.

Para cada una de estas áreas de interés se exploraron las bases de datos y análisis ya disponibles en la pagina web de infraestructura visible de tal forma que el proyecto aprovechara los datos ya existentes y propusiera nuevos análisis a partir de estos datos.

Exploración de datos: Aeropuertos.

Dentro de la página web de Infraestructura Visible se encontraban tres bases de datos sobre infraestructura y tráfico aéreo en Colombia, estos son:

* Aeropuertos del mundo: esta base de datos cuenta con ubicación geográfica, altitud, longitud, latitud y abreviaciones IATA para 9301 aeropuertos.
* Aeropuertos (Colombia): esta base de datos cuenta con nombre del aeropuerto, municipio en donde se encuentra ubicado, departamento en donde se encuentra ubicado, categoría del aeropuerto (aeródromo, aeropuerto regional, aeropuerto nacional o aeropuerto internacional), propietario del aeropuerto, empresa que explota el aeropuerto, altitud, longitud, latitud, abreviaciones colombianas, abreviaciones IATA, ancho de la pista, longitud de la pista, peso bruto máximo operacional, resolución de construcción, fecha de construcción, clase del aeropuerto, y tipo de aeropuerto (Aerocivil, fumigación, militar, privado o público) para 886 aeropuertos en Colombia.
* Aeropuertos\_histórico: esta base de datos es un registro histórico de los vuelos realizados desde o hacia aeropuertos en Colombia entre 2004 y 2018. La granularidad de la base datos es de carácter mensual, y cada registro contiene el mes, año, nombre de la ruta, aeropuerto de origen, aeropuerto de destino, nombre del modelo del avión utilizado para la ruta, aerolínea que realizo los vuelos, tipo de vuelo (Charter, Regular, Taxi o Adicionales), tipo de trafico (Internacional, Nacional o Externo), número de vuelos realizados, numero de sillas total para los vuelos realizados, carga ofrecida para los vuelos realizados, pasajeros que viajaron y generaron ingresos a la aerolínea y carga a bordo de los vuelos.

En total se encontraron 30 análisis y visualizaciones diferentes en la página de Infraestructura Visible que hacían uso de esta información, de los cuales solo 5 pertenecían a la tabla de histórico de vuelos en donde mostraban el total de vuelos, sillas disponibles, pasajeros a bordo, carga ofrecida y carga abordo para la totalidad de los vuelos desde el 2004 hasta el 2018. Uno de los hallazgos mas importantes de esta etapa fue determinar que en ninguno de estos análisis se cruzó información entre las tablas, por lo que existe un potencial considerable de nuevos análisis y visualizaciones útiles realizando este cruce de datos.

Exploración de datos: puertos

Dentro de la página web de Infraestructura Visible se encontraban dos bases de datos sobre infraestructura y tráfico marítimo en Colombia, estos son:

* Sociedades portuarias (2018): esta base de datos cuenta con nombre de la sociedad portuaria, Zona (Ciudad), Latitud, Longitud, Tipo (Público, Privado), Departamento, Municipio, Código del departamento, Código del municipio, Toneladas de carbón comercializadas en el último año, Toneladas de contenedores comercializadas en el último año, Toneladas comercializadas de materiales generales en el último año, Toneladas comercializadas de granel líquida (químicos, petróleo, combustibles entre otros) en el último año, Toneladas comercializadas de granel sólido en el último año, Toneladas de otros productos no nombrados anteriormente en el último año y Toneladas totales comercializadas en el último año para 55 registros.
* Histórico de puertos: esta base de datos es un registro histórico anual de toneladas de productos comercializados entre el 2013 y 2018, cada registro contiene nombre de la sociedad portuaria, Zona (Ciudad), Latitud, Longitud, Tipo (Público, Privado), Departamento, Municipio, Código del departamento, Código del municipio, Año del registro, Toneladas comercializadas de materiales generales, Toneladas comercializadas de granel líquida (químicos, petróleo, combustibles entre otros), Toneladas comercializadas de granel sólido, Toneladas de otros productos no nombrados anteriormente, Toneladas totales comercializadas para 280 registros.

Al discutir la exploración de datos y posibles análisis realizables con la información sobre puertos con el equipo de Infraestructura visible se decidió priorizar los datos de vías e infraestructura aérea por lo que no se continuó trabajando en la información de puertos ni se realizaron implementaciones con esta información.

Exploración de datos: Vías

Dentro de la página web de Infraestructura Visible se encontraban dos categorías de información para vías en Colombia, la primera categoría correspondía a carreteras de Colombia y la segunda categoría correspondía a vías férreas. Sobre carreteras se encontraban los siguientes datos:

* Tabla concesiones: esta base de datos cuenta con datos para 47 concesiones con Nombre del proyecto, Generación, Ola, Nombre de la concesión, Nit, Numero del contrato, Dirección concesionario, Teléfono, página web, departamento, valor del contrato, vigencia, fecha firma, fecha estimada de entrega, longitud concesionada, longitud de doble calzada, longitud de calzada sencilla, longitud segunda calzada, mejoramiento y o rehabilitación del contrato, longitud tercer carril, interventoría, viaducto contratado, intervención túneles contratados, sitios críticos contratados.
* Tabla Peajes: esta base de datos contaba con 173 entradas de peajes con código de la vía, Nombre, Responsable, Costo por categoría, Sector, Administrador, Ubicación en lenguaje natural, Teléfono fijo, Teléfono Móvil, costo por eje adicional, Nombre de las Categorías, Latitud, Longitud, Departamento y código departamento.
* Tabla Puentes: esta base de datos contaba con 3055 registros de puentes con nombre, Carretera, Punto de referencia del puente, administrador, numero de luces, longitud de luz mínima, longitud de luz máxima, longitud total del puente, longitud del tablero del puente, ancho del andén derecho e izquierdo, departamento, municipio, código departamento y código municipio.
* Tabla Túneles: esta base de datos contaba con 73 registros para túneles con nombre del túnel, código del proyecto, proyecto, generación, estado, latitud, longitud, departamento, municipio, código departamento, código municipio.
* Archivo Geojson Red primaria: este archivo es un modelo de carreteras de las vías primarias de Colombia, utiliza el formato JSON para describir la geometría geográfica de las vías principales del país, este archivo también contiene el nombre de las vías y una id único para cada una.
* Archivo Geojson vías de cuarta generación: este archivo es un modelo de carreteras de las vías de cuarta generación de Colombia, utiliza el formato JSON para describir la geometría geográfica de las vías, este archivo también contiene el nombre de las vías y una id único para cada una.

De la información explorada en el proyecto, la información de vías fue la que genero mayor interés por parte del cliente. Por esta razón, se dedicó gran parte del tiempo del proyecto a explorar esta información y sus posibles análisis. Se propuso en ese momento la posibilidad de actualizar estas bases de datos con información mas reciente desde INVIAS y la ANI para complementar los datos y analizar el tráfico y recaudos en peajes de las vías de Colombia a partir de los datos de estas dos entidades públicas.

Exploración de datos INVIAS:

* Red vial de Colombia: INVIAS, al igual que infraestructura Visible, cuenta con un archivo GeoJSON que se puede consultar desde su API publica, en este se encuentran representadas las vías primarias de Colombia. Al revisar este archivo se realizo un estudio comparativo con el archivo de Infraestructura Visible y se encontraron las siguientes diferencias entre ambos archivos:
  + Archivo GeoJSON - Red Primaria y secundaria INVIAS – septiembre 2019:
    - Número de tramos: 618
    - Longitud total (km): 19816.59
    - Longitud promedio por tramo (km): 32.065
    - Longitud Máxima en un tramo (km): 194.47
    - Longitud Mínima en un tramo (km): 0.011
    - Adicionalmente, el archivo GeoJSON de INVIAS venia acompañado de una tabla CSV con la siguiente información: Codigo del tramo de via, Codigo de la vía a la que pertenece el tramo, ID del punto de referencia Inicial, ID punto de referencia Inicial, Codigo del territorio, Nombre del tramo, nombre del sector, Codigo Administrador, Numero de calzadas, Fuente de la información, Fecha de la consulta, Fecha de Actualización, Observaciónes de invias, Estado de la via, Observaciones IGAC.
  + Red vial Primaria Infraestructura Visible:
    - Número de tramos: 923
    - Longitud total (km): 15771.18
    - Longitud promedio por tramo (km): 17.11
    - Longitud Máxima en un tramo (km): 101.52
    - Longitud Mínima en un tramo (km): 0.51

Adicionalmente, se realizaron comparaciones similares para los archivos CSV disponibles en INVIAS para puentes y peajes contra los archivos de Infraestructura Visible, las diferencias se describen a continuación:

Puentes:

* Puentes – INVIAS: Ultima actualización junio 2017
  + 3266 registros.
  + A diferencia del archivo de Infraestructura Visible, este archivo contaba con la siguiente información adicional: Longitud, Latitud, Referencial al material de construcción, año de construcción, año de reconstrucción, fecha de inspección, id ruta.
* Puentes – Infraestructura Visible:
  + 3056 registros.
  + A diferencia del archivo de Invias este archivo contaba con el Nombre del puente, el departamento, municipio, código del departamento y código del municipio.

Peajes:

* Peajes – INVIAS: Ultima actualización septiembre 2020
  + 178 registros.
  + Contiene URLs a fotos de los peajes.
* Peajes – Infraestructura Visible:
  + 173 registros.

Del estudio se concluyó que los archivos de INVIAS contenían información más actualizada para peajes, puentes y vías, además, el archivo de vías contaba con tramos de las vías secundarias de Colombia que no aparecían en los archivos de Infraestructura Visible. Por este motivo, se decidió utilizar estos archivos como fuente primaria de datos de vías para el proyecto.

Exploración de datos ANI:

Se decidió explorar los datos de recaudo y tráfico de peajes de la Agencia Nacional de Infraestructura (ANI) para determinar si era posible o no diseñar visualizaciones y realizar análisis de estos datos para Infraestructura Visible en conjunto con los datos de INVIAS. Al revisar el listado de datos públicos de la ANI encontramos siete archivos de Excel que compilaban información de recaudo y trafico de peajes del 2014 al 2020 de manera anual, de los cuales el archivo de 2019 estaba dañado y el archivo de 2020 tenia un formato diferente, con la información incompleta, por lo que se decidió eximir estos dos archivos del análisis de datos. Todos los archivos de la ANI contaban con los datos de los peajes organizados por concesión a la cual el peaje pertenecía y los datos de tráfico y recaudo de cada mes para los peajes que estaban a cargo de la ANI durante el año en particular. Al terminar la exploración de datos determinamos que si era posible realizar un análisis a partir de estos, por lo que proseguimos con la siguiente etapa del proyecto.

Como conclusión de la primera fase del proyecto se proponen los siguientes análisis a realizar:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tema analítico** | **Análisis requeridos o inferidos/requerimientos analíticos** | **Procesos de negocio** | **Datos requeridos** |
| **Comportamiento de los peajes y las vías en Colombia** | Análisis de los recaudos hechos en peajes y sus variaciones en el tiempo por peaje especifico, ruta a la que pertenece el peaje y departamentos en Colombia. | Recaudo de peajes y Trafico en las vías primarias de Colombia | \* Peajes en Colombia (Ubicación geográfica, Concesión, cobro por categoría, código de vía, Departamento al que pertenece) \* Vías de Colombia (Trazo geográfico por puntos de referencia con sus coordenadas geográficas, nombre del trazo, código de la vía) \* Información de recaudo y tráfico de los peajes (De carácter mensual, por concesión y peaje) \* Información derivada (Trafico y recaudos por ruta, departamento) |
| Análisis del tráfico en Colombia y sus variaciones en el tiempo por peaje especifico, ruta a la que pertenece el peaje y departamentos en Colombia. |
| Análisis del costo de viajar a través de la vía primaria de Colombia y sus variaciones en el tiempo por categoría de vehículo y Ruta. |
| **Cobertura y comportamiento de vuelos y aeropuertos en Colombia** | Análisis de la cobertura aérea por centro poblado y departamento. Mostrar Numero de centros poblados por categoría de cobertura y porcentaje de coberturas por tipo de departamento. | Cobertura, Trafico y capacidad de carga aérea en aeropuertos de Colombia | \* Aeropuertos de Colombia (Siglas, Código IATA, nombre, departamento, categoría, coordenadas geográficas, tipo) \* Centros Poblados de Colombia (Nombre, departamento, coordenadas geográficas) \* Información de vuelos (De carácter mensual, con capacidad de carga, capacidad máxima de pasajeros, pasajeros transportados, carga transportada) \* Información derivada (Matriz de distancia entre centros poblados y aeropuertos, tipo de cobertura por centro poblado y distancia a aeropuerto más cercano por categoría) |
| Análisis del tráfico de pasajeros aéreos Colombia, sus variaciones en el tiempo por aeropuerto específico y departamentos en Colombia. Mostrar número de pasajeros transportados, capacidad máxima de pasajeros y porcentaje de desaprovechamiento de capacidad de pasajeros por aeropuerto y región geográfica a través del tiempo. |
| Análisis del tráfico de carga aéreo en Colombia, sus variaciones en el tiempo por aeropuerto específico y departamentos en Colombia. Mostrar número de vuelos de carga, cantidad de carga transportada, porcentaje de desaprovechamiento de capacidad de carga y capacidad máxima de carga por aeropuerto y región geográfica a través del tiempo. |

Durante la segunda fase del proyecto, se escogió a Google Cloud Servicies como la plataforma en la que se desarrollaría el proyecto, esto principalmente gracias a su ‘‘Free Tier’’ que nos permitiría tener un desarrollo y despliegue gratuito de los tableros de control sobre una arquitectura en la nube. Para la capa de recolección de datos se utilizó Google Storage, que funciona como un Bucket o repositorio de datos, se utilizó Google BigQuery como Data Mart y centro de ETLs ya que este permite consultar las fuentes externas directamente y aplicar operaciones programadas sobre los datos, se utilizó Google Data Studio como capa de análisis para crear los tableros de control y, finalmente, se creó un sitio web HTML hosteado en Github con los tableros de control embebidos en el código para crear la fachada.

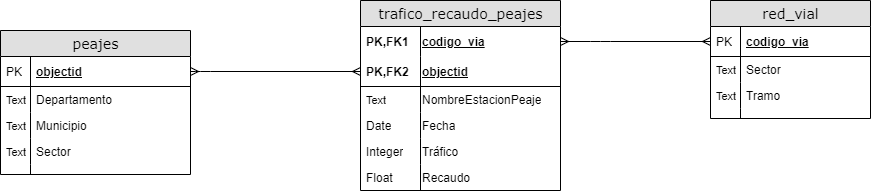
Adicionalmente, se continuaron las reuniones quincenales con el equipo de Infraestructura Visible para trabajar la construcción de los archivos que conformarían las tablas de la base de datos final del proyecto. El primer paso dentro de estas reuniones fue construir un compilado de todos los archivos de Peajes, este primer archivo era una tabla de Excel fácil de entender de tal manera que se pudiera realizar una revisión inicial de la validez de los datos en conjunto al equipo de infraestructura visible:



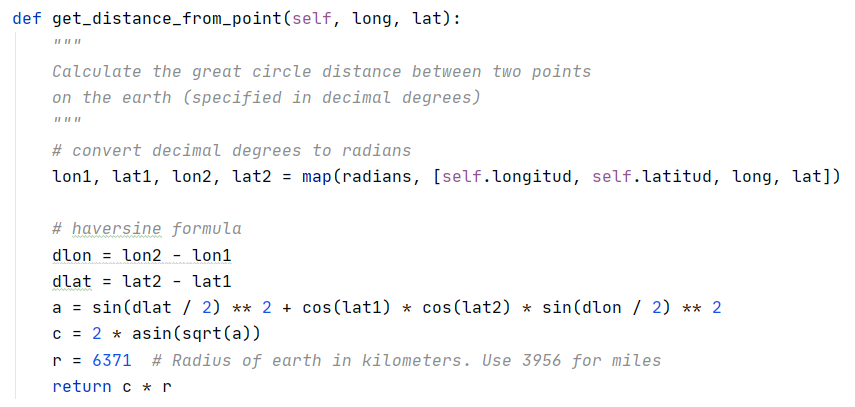
Al realizar esta primera versión del compilado, encontramos que existen 18 entradas en los datos que tienen problemas de duplicación. Estos corresponden a peajes con el mismo nombre, pero distintas concesiones, para resolver el problema, se decidió dejar el nombre de la última concesión vigente en los datos y se consulto la base de datos de concesiones en conjunto al equipo experto de Infraestructura visible para corregir conflictos en las cifras de tráfico y recaudo, de tal manera que coincidieran con los datos reportados por la ANI.

Como se esperaba realizar un JOIN de esta base de datos con la base de datos de peajes de INVIAS, se hace una primera aproximación entre ambos archivos, en donde se encontraron 14 peajes con diferentes nombres entre ambas bases de datos. Con ayuda del equipo de infraestructura visible, se corrigen estos nombres no idénticos a los que aparecen en la base de datos de INVIAS, ya que estos se utilizarían como referencia. Al finalizar el cruce de datos, encontramos que hay 51 peajes en la base de datos de INVIAS que no tienen datos de trafico y recaudo, esto se debe a que estos no están administrados por la ANI, por lo que no fue posible encontrar los datos para estos. Adicionalmente, encontramos dos peajes que no existen en la base de datos de INVIAS, pero que la ANI si reporto datos históricos de trafico y recaudo. Estos son los peajes de Ponedera y Rio Seco, debido a esto se excluyeron del análisis final. Finalmente, se propone una estructura fácil de procesar para datos históricos dentro de BigQuery.

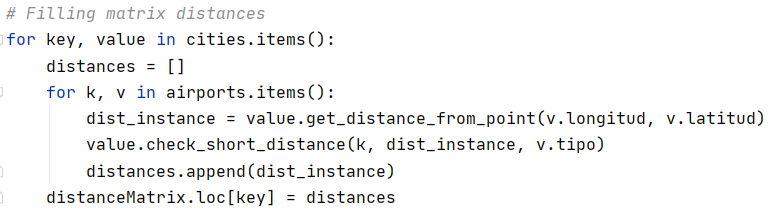
La estructura final de la base de datos para vías, tráfico y recaudo fue la siguiente:



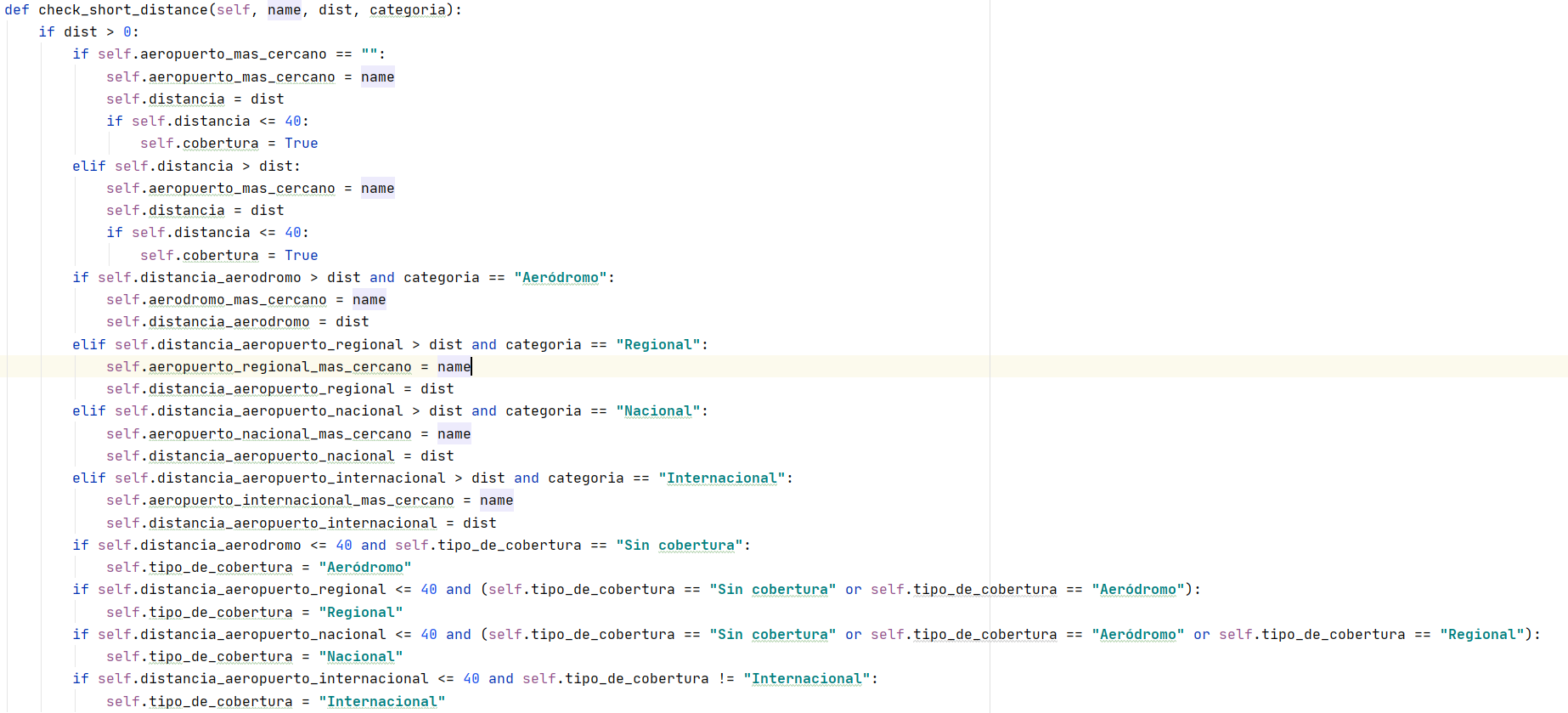
Simultáneamente, se trabajó en la construcción de un modelo que calculara la cobertura de los aeropuertos para centros poblados. Para esto, el equipo de Infraestructura Visible determino que la cobertura de los aeropuertos era de 40 Kilómetros de radio alrededor de la ubicación del aeropuerto. Con esto en cuenta, se desarrollo un script que a partir de la base de datos de centros poblados del DANE y la base de datos de Aeropuertos de Colombia, en donde ambos contaban con coordenadas geográficas, se calcula la distancia entre los dos puntos sobre la tierra usando la fórmula de Haversine.



A partir de estas distancias se calcula una matriz entre todos los centros poblados de Colombia y todos los aeropuertos de Colombia:

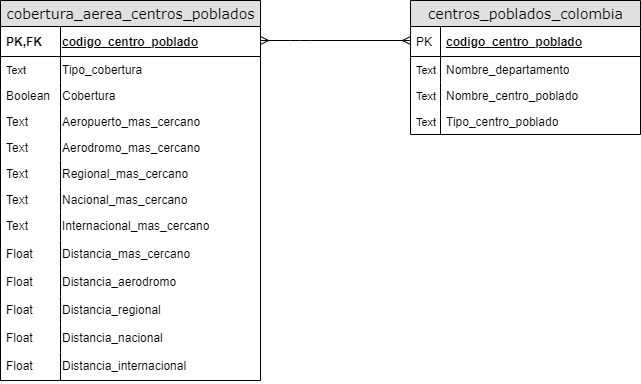


Y finalmente se verifican los datos de cobertura que nos interesan, en este caso, cual es el aeropuerto mas cercano y su distancia. Si la distancia es menor a 40km entre un centro poblado y un aeropuerto se determina que el aeropuerto cubre el centro poblado. Al finalizar este primer análisis, se propone hacer un análisis adicional usando la categoría del aeropuerto utilizando la misma metodología. Por lo que para cada centro poblado se calcula el aeropuerto mas cercano de cada categoría, y se determina el tipo de cobertura que tiene un centro poblado de acuerdo con la mejor categoría de aeropuerto que lo cubre.



De esto se genera un archivo CSV que contiene la información de distancia y cobertura por categoría. Este archivo se utilizará como archivo final dentro de la base de datos.

La estructura final de la base de datos para cobertura de aérea de centros poblados fue la siguiente:



Para los análisis de trafico aéreo

### 5.2 Resultados esperados

## 6 Validación

## 7 Conclusiones

## 8 Referencias

[1] KIMBALL, Ralph, ROSS, Margy. “The Data Warehouse Toolkit: the definitive guide to dimensional modeling". Third Edition. John Wiley & Sons, Inc, 2013.

2] KIMBALL, Ralph; CASERTA, Joe. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data”. Wiley, 2004

[3] KIMBALL, Ralph, REEVES, Laura, ROSS, Margy, THORNTHWAITE, Warren. “The Data Warehouse Lifecycle Toolkit”. John Wiley & Sons, Inc., 2008.