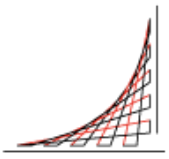
Modelo de análisis de datos para simular y visualizar el destino de los usuarios de un sistema de Buses de Transito Rápido (BTR)

Diego Esteban Valencia Salamanca

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría Gestión de Información
Bogotá D.C., 19 de julio de 2021



Modelo de análisis de datos para simular y visualizar el destino de los usuarios de un sistema de Buses de Transito Rápido (BTR)

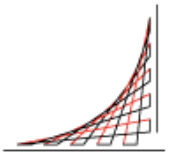
Diego Esteban Valencia Salamanca

Trabajo de investigación para optar al título de
Magíster en Gestión de Información

Director
Luis Daniel Benavides Navarro
Doctor en informática

Jurados
Fabiola del Toro Osorio
Dante Conti

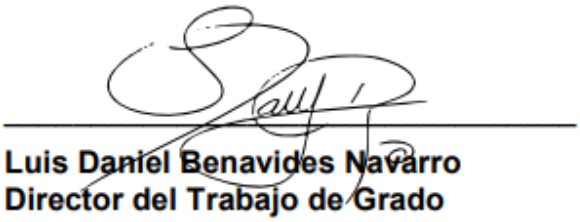
Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría en Gestión de Información
Bogotá D.C., 19 de julio de 2021



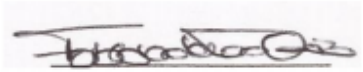
Publicado en 2021 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

PÁGINA DE ACEPTACIÓN

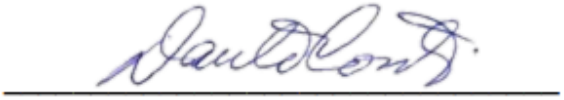
El trabajo de grado de maestría titulado “Modelo de análisis de datos para simular y visualizar el destino de los usuarios de un sistema de Buses de Transito Rápido (BTR)” presentado por DIEGO ESTEBAN VALENCIA SALAMANCA, cumple con los requisitos establecidos y recibe nota aprobatoria para optar al título de Magíster en Gestión de información.



Luis Daniel Benavides Navarro
Director del Trabajo de Grado



Fabiola del Toro Osorio
Jurado



Dante Conti
Jurado

Bogotá, D.C., 19 de julio de 2021

Resumen

Las ciudades inteligentes buscan aprovechar las tecnologías de la información y comunicación para analizar los datos recolectados de las actividades cotidianas de los ciudadanos. Una de esas actividades (y fuente grande de información) es el transporte público.

A nivel mundial, una solución de transporte público son los buses de transito rápido (BTR), principalmente para las ciudades en crecimiento. Su implementación es rápida y económica, versus otras opciones, por lo que es una gran iniciativa de solución a la movilidad.

La limitante para analizar estos sistemas se presenta cuando se observa que la información que se recolecta proviene únicamente del momento en el que el usuario ingresa, por lo que no hay un seguimiento de su trayecto y su destino. El presente trabajo propone un modelo para estimar el destino de los usuarios, y un artefacto para visualizar los resultados, basándose en la información histórica ya recolectada (tomando el caso del sistema Transmilenio, de Bogotá, Colombia). El desarrollo del modelo supone dos retos: 1. transformar los datos que genera el sistema, de tal forma que pueda encontrarse la estimación esperada, y, 2. encontrar una forma visual de presentar dicha solución a los interesados.

La base del modelo se sustenta en el supuesto de que el usuario tiene un comportamiento recurrente, por lo que el primer paso es identificar las estaciones de mayor recurrencia en dos momentos diferentes del día. Una vez adoptado este supuesto, y como solución del primer reto, los datos deberán ser agrupados a nivel de usuario, lo que permitirá reducir considerablemente el tamaño de las bases a procesar; posteriormente, las transformaciones darán pautas para realizar análisis con otros enfoques diferentes al del usuario. Por el lado de las visualizaciones, primero debe entenderse al público objetivo y definir el valor que esta información puede significar para él. El apoyo gráfico, se sustentará en esta definición para recurrir a los atributos preatentivos correctos y elegir el tipo de visualización adecuado.

Una vez se tengan los resultados del modelo de simulación y de visualización, el artefacto permitirá validar la información con casos reales y evaluar trabajos futuros a partir de los datos procesados y el cruce de nuevas fuentes.

Abstract

Smart cities seek to take advantage of information and communication technologies to analyze the data collected from the daily activities of citizens. One of those activities (and a great source of information) is public transport.

Globally, a public transport solution is the bus rapid transit (BRT) system, mainly for growing cities. Its implementation is faster and cheaper, compared to other options, making it a great mobility solution initiative.

The limitation to analyze these systems occurs when it is observed that the information that is collected comes only from the moment in which the user enters, so there is no tracking of their journey and their destination. The present work proposes a model to estimate the destination of the users, and an artifact to visualize the results, based on the historical information already collected (taking the case of the Transmilenio system, from Bogotá, Colombia). The development of the model involves two challenges: 1. transforming the data generated by the system, in such a way that the expected estimate can be found, and 2. finding a visual way to present said solution to the interested parties.

The basis of the model assumes that the user has a recurring behavior, so the first step is to identify the stations with the highest recurrence at two different times of the day. Once this assumption has been adopted, and as a solution to the first challenge, the data must be grouped at the user level, which will considerably reduce the size of the databases to be processed; later, the transformations will give guidelines to carry out analysis with other approaches. On the visualization side, we must first understand the target audience and define the value that this information can mean for them. The graphic support will be based on this definition to select the correct preattentive attributes and choose the appropriate type of visualization.

Once the results of the simulation and visualization model are available, the artifact will allow to validate the information with real cases and evaluate future work from the processed data and the crossing of new sources.

CONTENIDO

1 INTRODUCCIÓN 6

1.1 OBJETIVOS 7
1.2 ALCANCE 7

2 MARCO TEÓRICO Y ESTADO DEL ARTE 8

2.1 CIUDADES INTELIGENTES 8
2.1.1 Sistemas de movilidad en ciudades inteligentes 9
2.2 SISTEMAS DE BUSES DE TRANSITO RÁPIDO 11
2.2.1 Sistema de transporte masivo, Transmilenio 11
2.3 PROCESO ETL 12
2.4 VISUALIZACIONES 13
2.4.1 ¿Qué es la visualización de datos? 14
2.4.2 Tipos de visualizaciones 14
2.4.3 Diseño de visualizaciones 16

3 MODELOS DE EXTRACCIÓN, SIMULACIÓN Y ANÁLISIS DE DATOS 21

3.1 MODELO ETL 21
3.2 MODELO SIMULACIÓN DESTINO 24
3.3 MODELO DE VISUALIZACIÓN 25

4 RESULTADOS Y CONTRIBUCIÓN 27

4.1 ARQUITECTURA DE DATOS 27
4.2 MODELO DE SIMULACIÓN 29
4.3 MODELO DE VISUALIZACIÓN 31

5 CONCLUSIONES Y RECOMENDACIONES 37

6 REFERENCIAS BIBLIOGRÁFICAS 38

7 ANEXOS 41

FIGURAS

Figura 1. Modelo de Plataforma de Smart City 10
Figura 2. Transmilenio. (2020). Tarjeta TuLlave en el datáfono de recaudo. Recuperado de https://www.transmilenio.gov.co 12
Figura 3. González, O. (2021). ¿Qué es un ETL? Recuperado de https://www.appvizer.es 13
Figura 4. Kirk, A. (2019). Data Visualisation: A Handbook for Data Driven Design. Recuperado de https://book.visualisingdata.com 14
Figura 5. Pasos para comunicar los datos. Fuente propia 18
Figure 6. Ejemplos de atributos preatentivos en texto. Fuente propia 19
Figura 7. Few, S. (2012). Preattentive attributes. Show me the numbers. Recuperado de https://laptrinhx.com 19
Figura 8. Ejemplo grafico de barras horizontales. Fuente propia 25
Figura 9. Gráfico de usuarios por hora. Fuente propia 26
Figura 10. Ejemplo de filtros. Fuente propia 26
Figura 11. Modelo estrella de datos. Fuente propia 27
Figura 12. Arquitectura de aplicaciones del modelo. Fuente propia 29
Figura 13. Ejemplo reducción registros. Fuente propia 30
Figura 14. Modelo estrella de la simulación. Fuente propia 31
Figura 15. Página "Principal" del dashboard. Fuente propia 32
Figura 16. Página "Por tarjeta" del dashboard. Fuente propia 33
Figura 17. Página "Por Origen" del dashboard. Fuente propia 34
Figura 18. Página "Por Destino" del dashboard. Fuente propia 34

TABLAS

Tabla 1. Metadatos de las tablas entregadas por el sistema 21
Tabla 2. Selección de columnas, proceso ETL 22
Tabla 3. Columnas base de salida 'Flujo' 22
Tabla 4. Columnas base de salida 'Conteo' 23
Tabla 5. Columnas base de salida 'O-D' 23
Tabla 6. Columnas base de salida 'Matriz' 23
Tabla 7. Metadatos tablas de salida del proceso ETL 28
Tabla 8. Aplicación modelo de simulación 30

ANEXOS

Anexo 1. Plano de estaciones y portales de Transmilenio 41
Anexo 2. Link artefacto de visualización y bases de datos 42

1 INTRODUCCIÓN

En el marco de las ciudades inteligentes, se establecen lineamientos que buscan integrar las necesidades de los ciudadanos con los servicios que ofrece la ciudad, dicha integración se sustenta fuertemente en la recolección de datos que se produce por la interacción de ambos agentes, lo cual permite tener una comprensión holística de la ciudad, y eventualmente gestionar proyectos, minimizar tiempos, reducir costos y aprovechar las tecnologías de la información y comunicación. (Departamento Nacional de Planeación de Bogotá, 2020).

La recolección y análisis de grandes volúmenes de datos se convierte en un tema trascendental para generar valor y poner a disposición herramientas de interacción con la ciudadanía. Uno de los sistemas a analizar en recolección de datos, es el sistema de transporte público, ya que es el resultado de las interacciones recurrentes causadas por la cotidianidad de los ciudadanos, lo que evidentemente genera grandes volúmenes de datos. Específicamente, una solución de transporte en las ciudades emergentes y grandes ciudades, se encuentran los sistemas de buses de transito rápido (BTR) los cuales consisten en buses (generalmente articulados) que se mueven a través de un carril, con estaciones y portales de uso exclusivo para los vehículos y los usuarios del sistema.

Alineado con las políticas de ciudades inteligentes, los datos generados por un sistema BTR deberían ser procesados y comunicados de tal forma que se pueda mejorar la planeación, la gestión y el control del sistema; para esto, se vuelve indispensable generar modelos que simulen numérica y gráficamente el funcionamiento del mismo. El procesamiento de grandes volúmenes de datos y una respuesta visual rápida, se convierten en una solución para lograr dicho fin dentro del sistema; a esto se le suma el aprovechamiento de los datos históricos (ya recolectados) y las técnicas de visualización utilizadas.

El presente trabajo se enmarca en los sistemas BTR exponiendo un artefacto que permita analizar los datos recolectados para simular el trayecto de los usuarios, tomando como base su comportamiento rutinario. Para esto se propone un modelo de extracción, transformación y limpieza de los datos históricos, con el cual se crean modelos para la simulación y visualización de las trayectorias, obteniendo como resultado una herramienta con 4 tableros dinámicos que permiten analizar el sistema de forma global (cantidad de usuarios, cantidad de viajes, filtros de fecha y hora) y entrar en detalle a analizar el comportamiento de las estaciones y el trayecto de los usuarios (Estaciones de origen y destino por usuario, destinos más probables por origen, y orígenes más probables por destino). Este análisis se enmarca en el sistema Transmilenio (BTR de la ciudad de Bogotá, Colombia), con datos de febrero a octubre del 2019.

Para dar solución al reto planteado, se estructura el presente trabajo de la siguiente forma: En el marco teórico y estado del arte (capitulo 2) se profundiza sobre los sistemas BTR (caso del Transmilenio), se conceptualiza sobre los procesos ETL y sobre el diseño de visualizaciones. En los modelos de extracción, simulación y análisis de datos (capítulo 3), se presenta el proceso ETL aplicado a las bases entregadas por el sistema, la metodología utilizada para la simulación de estimación de destinos y el proceso para la definición y construcción de artefactos de visualizaciones. En los resultados y contribución (capítulo 4), se muestra la arquitectura de datos y aplicaciones utilizada en el proceso, junto con los resultados del modelo de simulación y la herramienta de visualización. Y, por último, en conclusiones y recomendaciones (capítulo 5), se muestran las conclusiones a las que se llegaron durante el desarrollo del trabajo y se proponen recomendaciones para trabajos futuros sustentadas en los entregables.

1.1 OBJETIVOS

Objetivo General

Implementar un modelo de análisis de datos que permita predecir y visualizar el destino de los usuarios de un sistema de Buses de Transito Rápido (BTR) por medio del análisis histórico de viajes en el sistema.

Objetivos específicos

- Extracción, transformación y limpieza de datos históricos de viajes en Transmilenio.
- Modelo para la simulación de trayectorias por medio del análisis histórico de datos.
- Modelo para la visualización de trayectorias y viajes de usuarios en Transmilenio.
- Validación del modelo por medio de la construcción de un artefacto para la visualización dinámica de las trayectorias históricas.

1.2 ALCANCE

El presente trabajo se enmarca en el diseño y elaboración de artefactos para la manipulación y entendimiento de los usuarios de un sistema BTR; si bien el análisis permitiría tomar decisiones para el negocio como cambios en la operación logística o incluir variables financieras, el proyecto toma como interesado y beneficiario a los ciudadanos que utilizan el servicio de transporte público. El impacto que se busca generar esta dirigido al ciudadano recurrente del sistema, esperando mejorar el uso del mismo y la toma de decisiones que mejoren su experiencia

2 MARCO TEÓRICO Y ESTADO DEL ARTE

En el presente apartado se introduce la importancia de los tableros de control (Dashboards) en el análisis y la gestión de datos para ciudades inteligentes, enfocando la explicación en los procesos de extracción, transformación, limpieza y visualización para los sistemas de transporte público, específicamente los sistemas BTR. Gran parte estará enfocada en la definición de las visualizaciones hablando de atributos preatentivos y buenas prácticas para la creación de gráficos.

2.1 Ciudades inteligentes

En el documento de ciudades inteligentes presentado por el equipo de alta consejería distrital de TIC en 2018, se muestran los modelos conceptuales y las buenas prácticas en ciudades inteligentes a nivel internacional, tomando como ejemplos: el modelo de KPMG, el modelo de Mckinsey Global Institute (MGI), el modelo del banco interamericano de desarrollo (BID) y el modelo del parlamento europeo. En los cuatro casos, se presentan los elementos básicos que sustentan el modelo de ciudad inteligente, coincidiendo todos en que el último elemento es la interfaz de comunicación (la presentación de los datos) destacando el uso de dashboards, KPI's y datos abiertos a través de portales web y aplicaciones móviles. (Cruz, G., Gamboa, J., Martínez, S., González, B., Gacha, L., 2018).

Esto evidencia que la presentación de los datos toma un papel relevante cuando se toma al ciudadano como público objetivo. Dado que se asume un conocimiento básico (no técnico) por parte del ciudadano, la comunicación y entrega de la información debe ser lo más sencilla posible, buscando el mayor entendimiento por parte de él.

Urban data and city dashboards

Kitchin & McArdle (2016) proponen seis preguntas a la hora de definir los objetivos de los 'City dashboards'. Específicamente hay dos que cuestionan el uso de los paneles de control:

1. ¿Qué tan comprensibles y usables son los dashboards?
2. ¿Cuál es la aplicación y el valor de estos?

La comprensibilidad y usabilidad de los dashboards, se define con el diseño de visualizaciones y el conocimiento del público objetivo. La aplicación y el valor está más asociada a los datos y a la información valiosa que de estos pueda extraerse. Para dar solución a ambas preguntas, específicamente en el caso del transporte público, se debe tener en cuenta dos atributos característicos de los datos y de los tableros de control: el primero es cantidad, haciendo referencia al número de registros recolectados, lo que propone un reto a la hora de procesarlos con la

finalidad de encontrar una forma de disminuirlos sin afectar los resultados; el segundo atributo está asociado a la parte visual, al uso adecuado de las herramientas disponibles para mostrar información geográfica y filtrar la información de tal forma que se resalten los resultados relevantes.

Para el diseño de tableros geoespaciales que monitoreen el comportamiento de las ciudades, se proponen dos estilos de interfaces gráficas de usuarios, estilo de una sólo página o estilo de profundización. El primero (una sola página), se enfoca en mostrarle al usuario todos los indicadores operacionales al mismo tiempo, el *London citydashboard* es un ejemplo de estilo de una sola página; por otro lado, el estilo de profundización se enfoca en la creación de menús y jerarquías que permitan ver la información en distintos niveles, lo que no se podría mostrar en una sola visualización, el *Dublin dashboard* es un ejemplo de este estilo (Jing, C., Du, M., Li, S., Liu, S. 2019).

2.1.1 Sistemas de movilidad en ciudades inteligentes

Según Boyd Cohen (2018), unos de los indicadores a evaluar en la movilidad de una ciudad inteligente, es el transporte eficiente (enmarcado en la categoría de movilidad). Dicha eficiencia se logra a través del uso adecuado de los datos que genera el sistema, desarrollando "[…] sistemas flexibles de información y de toma de decisiones para operar distintos modos de transporte en tiempo real impactando positivamente en el ahorro de tiempo de los usuarios y la mejora en la eficiencia de desplazamientos" (F. Perez et al, s.f).

Por otro lado, el crecimiento desacelerado de las ciudades enfrenta a la movilidad a tener que ajustarse a las condiciones demográficas de dicho crecimiento. La solución no puede ser desarrollar un nuevo sistema de transporte que reemplace al anterior (por el elevado costo económico que esto conllevaría), por este motivo, se deben pensar en soluciones cuidadosas que aprovechen los recursos actuales y permitan guiar el crecimiento de la ciudad y el desarrollo social.

Las soluciones se deben adaptar a las necesidades y expectativas del cliente (ya sean los ciudadanos o las empresas prestadoras del servicio). Sin embargo, algunas de las soluciones se enfocan en desarrollos de software (como la creación de aplicaciones móviles), o en implementaciones de hardware (como el uso de sensores o cámaras). Estas soluciones requieren grandes inversiones económicas, en tiempo y/o en infraestructura. La implantación de tecnología requiere inversión en la aplicación de artefactos, por lo que debe haber una financiación grande en la obtención de las herramientas necesarias.

Sin importar si se implementan o no nuevas tecnologías, también debe haber una evolución en los procesos de transformación y análisis de los datos recolectados para encontrar soluciones. Dentro del modelo de plataforma de Smart City propuesto por el ayuntamiento de Rivas Vaciamadrid (figura 1) se evidencia que una de las capas se enfoca en el conocimiento, dentro de la cual se da tratamiento al proceso ETL y al análisis de los datos.

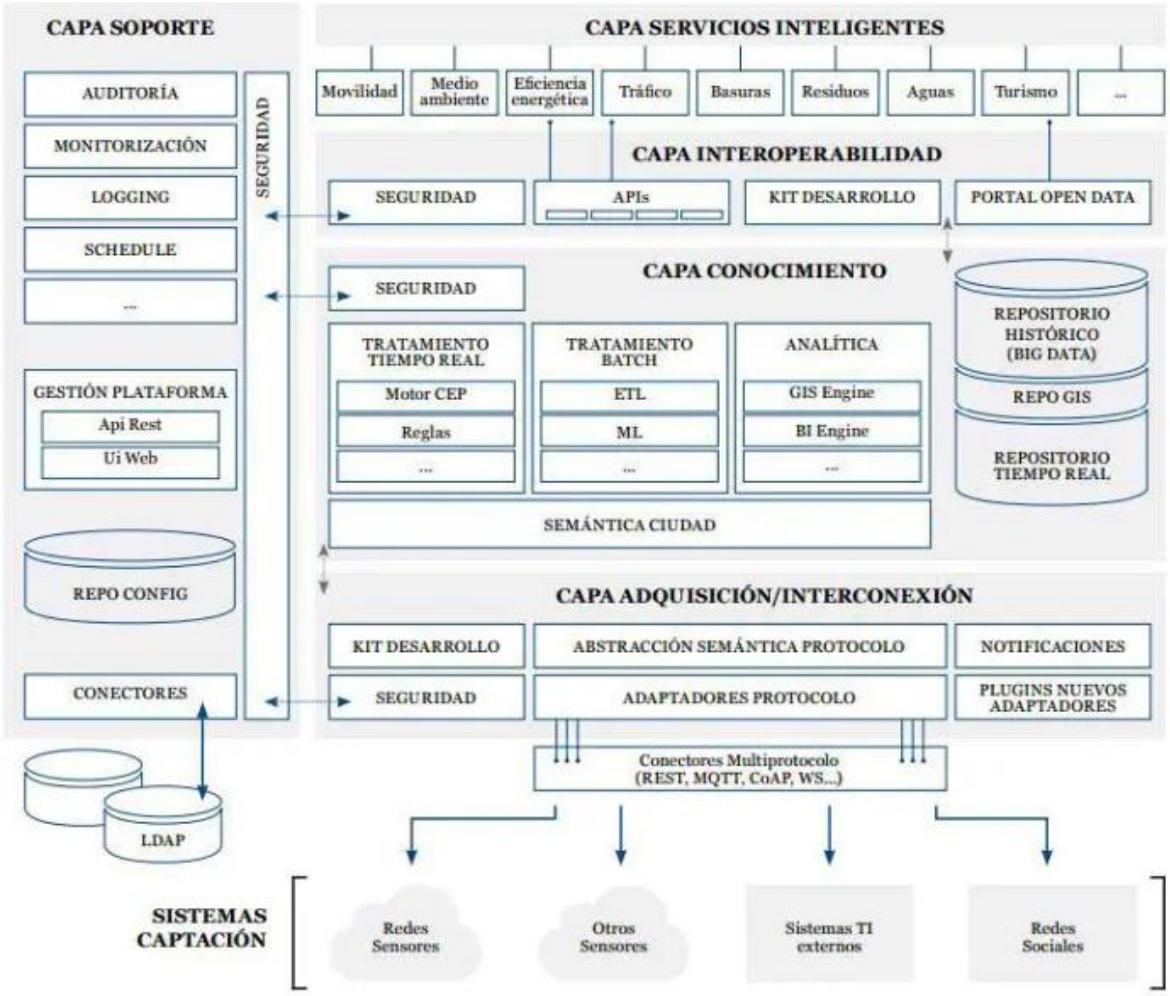


Figura 1. Ayuntamiento de Rivas Vaciamadrid (2021). Modelo de Plataforma de Smart City. Recuperado de: https://ovacen.com/

Por lo anterior, la transformación que se le dé a los datos recolectados por un sistema de transporte, debe ser un factor relevante en el análisis de sistemas eficientes de movilidad en las ciudades inteligentes, procurando aprovechar los recursos actuales y buscando la mayor extracción de información posible.

2.2 Sistemas de buses de transito rápido

El sistema BTR nace como solución de movilidad para el crecimiento de las ciudades, los problemas de congestión vehicular y la informalidad del transporte público. Como lo explica la federación internacional de los trabajadores del transporte (ITF, 2021) el sistema se sustenta en carriles destinados a buses articulados y/o biarticulados), con una infraestructura de estaciones, intersecciones, terminales y carreteras diseñadas exclusivamente para su adecuado uso, donde ningún otro vehículo particular puede acceder.

Una de las características del sistema es la definición de las paradas, las rutas y los horarios de los buses que transitan en él. En la literatura se encuentran tres tipos de rutas: aquellos que se detienen en todas las estaciones, aquellos que se detienen en ciertas estaciones, y aquellos que buscan abarcar la mayor cantidad de distancia entre cada parada. De igual forma, según los horarios, se llegan a activar buses para suplir la demanda. Todo esto sin contar con los buses complementarios o la integración de otros sistemas.

Otra de las características es el sistema de pago, si bien en algunos aún se da la opción de pagar con efectivo, como ocurre en el sistema de La paz, Bolivia (LaPaz Bus. 2021)., también existe la opción de prepago, y cómo lo asegura la empresa Volvo (s.f) “el prepago de los pasajes es fundamental para alcanzar la eficiencia”. El prepago se realiza a través de tarjetas inteligentes, en las cuales se recarga un saldo inicial el cual se va descontando con cada ingreso al sistema hasta, eventualmente, volver a recargarse.

2.2.1 Sistema de transporte masivo, Transmilenio

El Transmilenio (TM) es el sistema de transporte público masivo de la ciudad de Bogotá, Colombia, el cual nace en el año 2000 como una solución de buses articulados (y posteriormente la inclusión de biarticulados) para la crisis que se vivía en el sistema de transporte público.

A la fecha, el sistema está conformado por una red de troncales, que cuenta con 9 portales y 138 estaciones en su trayecto, para la parada exclusiva de los buses. Anexo 1. Los buses que recorren la red son articulados y biarticulados con capacidad para 160 y 250 personas, respectivamente[1]. Las estaciones y los portales están diseñados para tener puertas en ambos costados (doble dirección) para el acceso a los buses. El acceso para las personas es mediante registradoras con

[1] Existen otros buses en el sistema, como los alimentadores y los duales, que pueden ingresar en la red de troncales, sin embargo, no se tienen en cuenta en el presente trabajo.

torniquetes, ubicados en los ingresos peatonales (puede ser uno o dos ingresos para cada estación).

Para el ingreso a las estaciones, el usuario debe adquirir una tarjeta inteligente llamada “TuLlave”, la cual es recargable e irá descontando el saldo de cada ingreso al sistema. Su utilización se limita únicamente al momento del ingreso.

Figura 2. Transmilenio. (2020). Tarjeta TuLlave en el datáfono de recaudo. Recuperado de https://www.transmilenio.gov.co

Una vez que el usuario ingresa al sistema, debe seleccionar una de las rutas disponibles en la estación que abordó, bien sea una ruta que lo lleve a la estación destino, o una que le permita hacer transbordo en otra estación.

Las rutas están establecidas en dos formatos: servicios expresos y servicios ruta fácil. Según la página oficial del Transmilenio, los servicios expresos, son aquellos que están diseñados para ir a destinos específicos a la mayor velocidad posible, lo que implica no detenerse en todas las estaciones en la que pasa, por el contrario, los servicios de ruta fácil son aquellos que “se detienen en todas las estaciones a lo largo del recorrido”.

Una vez el usuario llega a su estación de destino, abandona el sistema sin ningún registro de salida.

2.3 Proceso ETL

Es un proceso de integración en tres etapas, en el que se buscan unificar fuentes de datos para convertirlas a un formato estándar que se pueda cargar en un sistema de almacenaje.

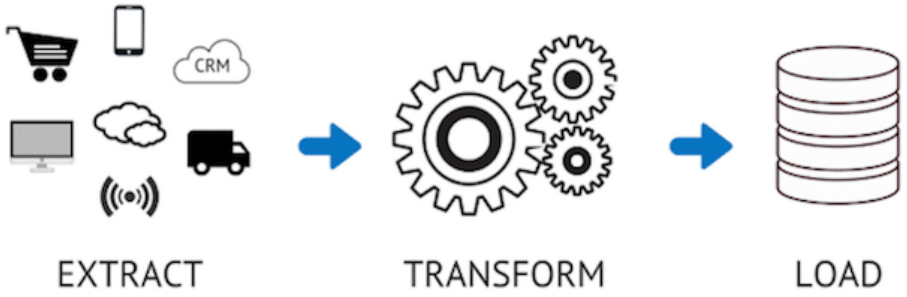


Figura 3. González, O. (2021). ¿Qué es un ETL? Recuperado de https://www.appvizer.es

La **extracción** recopila los datos provenientes de las diferentes fuentes de recolección (ej., sistemas transaccionales) para dirigirlos a un sistema que aplique una serie de reglas que los **transforme** en un nuevo formato para **cargarlos** a un sistema de almacenaje que permita su aprovechamiento (análisis y visualización).

Algunos de los beneficios del proceso ETL son:

- Limpieza y homologación de datos
- Mejora en la velocidad de respuesta para análisis
- Disponibilidad para la visualización
- Accesibilidad a los datos
- Acercamiento al entendimiento del negocio

2.4 Visualizaciones

“Cuanta más información manejamos, más difícil resulta filtrar lo que es verdaderamente importante” (como se citó en Knaflic, 2017), y es que definir cómo mostrar la información recolectada por un sistema transaccional que registra, por día, millones de filas en una tabla, representa un reto para cualquier analista.

Ahora bien, el problema no radica sólo en la cantidad de datos que se quieren mostrar, también el público objetivo y el mensaje que se quiere presentar influyen en la definición visual. Ya sea con una infografía o un dashboard, en una presentación o en un artículo web; la forma de presentar la información dependerá de múltiples variables.

2.4.1 ¿Qué es la visualización de datos?

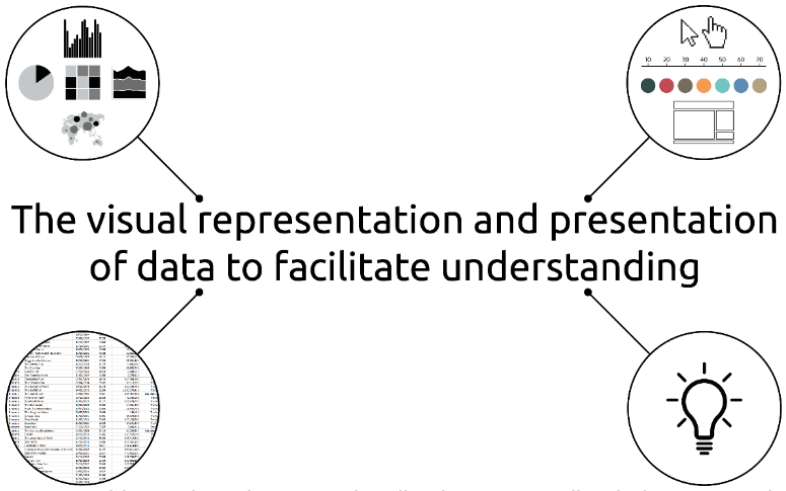


Figura 4. Kirk, A. (2019). Data Visualisation: A Handbook for Data Driven Design. Recuperado de https://book.visualisingdata.com

Para Kirk (2019) el elemento fundamental de las visualizaciones son los **'datos'**, generalmente, estos datos (texto o número) vienen almacenados en forma tabular donde cada fila representa un registro con diferentes variables (columnas). Para relacionar estos datos entre ellos y poder contextualizarlos, es necesario tomar **'the representation'**, con lo que se define la forma (gráficos) en la que se presentarán los datos para lograr la mejor percepción visual. **'Presentation'** es consecuente a la representación, pues en este elemento se definen componentes gráficos complementarios, como el tamaño (de figuras y textos), el color, anotaciones, fondos, etc. Por último, **'understanding'** se complementa con el pensamiento del consumidor final, quien sufre un proceso de entendimiento de tres etapas.

Por otro lado, Attardi (2016) afirma que "La visualización de datos tiene dos objetivos generales: reducir y revelar los datos" (p.18). La idea de reducir la cantidad de registros almacenados, cumple la función de no sobrecargar la capacidad técnica de los sistemas y poder agrupar los datos en características similares, permitiendo crear categorías (dimensiones) y métricas (hechos). La segunda idea propuesta por Attardi —revelar los datos— busca entregar un formato sencillo y entendible para el usuario final, dependiendo del tipo, o los tipos de visualización elegidos.

2.4.2 Tipos de visualizaciones

Una gráfica de barras es igual de válida que una gráfica de líneas, todo depende del mensaje que se quiera expresar. Los tipos de visualización no se diferencian por la forma del gráfico, por esta razón en este apartado se presentan dos metodologías

para tipificar las visualizaciones: la taxonomía propuesta por Valero, Marín y Català (2014), y, la visualización estática y dinámica propuesta por Attardi (2016).

Taxonomía de visualizaciones

Valero et al. (2014) presentan una aproximación a una taxonomía de la visualización de datos con ocho tipos de visualizaciones:

1. **Espacial:** Técnica más usada. Hace referencia a gráficos de barras, líneas, tortas, puntos, etc. Donde se busca comparar la información aprovechando las dimensiones del espacio
2. **Tabular:** Cualquier visual que compare y/o relacione linealmente la información. Normalmente segmentada por categorías.
3. **Posicional:** Visualización donde el dato adquiere significado al acercarse o alejarse de un punto de referencia.
4. **Topográfica:** Representa información sobre mapas territoriales (fenómenos geográficos, ciudades, carreteras, etc.). Son los usados en los Sistemas de información geográfica (SIG).
5. **Teledinámica:** Enfocado a los datos analizados en tiempo real, para evidenciar su cambio.
6. **Miscelánea:** Aquella que muestra de forma simultánea multitud de gráficos.
7. **Drag and Drop:** Enfocada a una interacción mayor con el usuario, buscando la elección y arrastre de los elementos.
8. **Identificación aumentada:** Acompañada de imágenes fotográficas, representa los objetos identificados en dicha imagen.

La taxonomía de Valero et al., categoriza las visualizaciones basándose principalmente en los elementos gráficos utilizados en ella. Attardi, por su parte, las define según la interacción que tiene el usuario con ellas.

Visualización dinámica y estática

La diferenciación que propone Attardi se fundamenta principalmente en el cambio de los datos en el tiempo, en otras palabras, si la visualización actualiza los datos en un marco temporal se denominará dinámica, en caso contrario será estática.

Dentro del grupo de visualizaciones estáticas, se resaltan las infografías como una forma de presentar los resultados dado que su intención es explicar una historia específica con un conjunto de datos específicos (Carmack, 2015).

Por su parte, las visualizaciones dinámicas presentan actualizaciones en el conjunto de datos. Es independiente si la actualización es en tiempo real o en un marco de tiempo mayor, lo importante es que haya un delta de información.

En cuanto a los elementos visuales (gráficos) utilizados en ambos tipos de visualizaciones, pueden ser los mismos, permitiendo incluso al usuario interactuar con la información, normalmente a través de filtros.

2.4.3 Diseño de visualizaciones

Previo a la elaboración de las visualizaciones, Knaflic (2017) propone establecer el tipo de análisis a realizar (exploratorio o explicativo); argumentando que el análisis exploratorio hace referencia a la comprensión de los datos y al discernimiento de lo que puede ser relevante o interesante destacar; es en este análisis donde se plantean hipótesis, se manipulan los datos y se proyectan escenarios; es aquí donde se realiza la ciencia de datos. Por otro lado, en el análisis aclaratorio se busca comunicar los resultados del análisis exploratorio, destacando la información importante sin entrar en detalle de todo el proceso previo, es aquí donde se deben exponer los resultados finales de los experimentos sin obligar al usuario final a repetirlos.

Ahora bien, en un análisis exploratorio se permite el uso de visualizaciones estáticas, se usan más números y el actor principal es la información (Knaflic, 2017, P. 29) no el usuario final, pues se asume que el usuario que está manipulando los datos tiene un conocimiento técnico previo, que le otorga el derecho de dicha manipulación.

Por otro lado, el análisis aclaratorio es el que más tiempo de desarrollo conlleva. Requiere una evaluación profunda para llegar a un resultado que permita enviar un mensaje claro a la audiencia. Si bien en la presentación de resultados el exponente (creador de la visualización) puede estar presente para aclarar dudas y profundizar en la explicación, en la mayoría de los casos, y con el avance tecnológico, la información se distribuye de forma exponencial sin permitirle al creador tener el control sobre esta, es por esta razón que se deben apoyar las visualizaciones con las mejores herramientas para entregar un buen mensaje.

Este apartado se centra en el análisis aclaratorio.

Definición del público

En el análisis aclaratorio se usan como base tres preguntas ¿Quién? ¿Qué? ¿Cómo?, en ese respectivo orden. En el *Quién* se busca conocer y definir con el mayor detalle posible al público objetivo al que irá dirigida la explicación; preguntas

sobre información previa, antecedentes, prejuicios y demás que pueda tener el público, servirán para encontrar dicho detalle. El *Qué,* se refiere a la acción del público; qué se busca que ellos hagan con la visualización y qué se espera que abstraigan de aquella interacción. Por último, el *cómo,* busca definir el apoyo visual necesario para respaldar la historia que se va a contar. (Knaflic, 2017)

Una vez se llega al punto del apoyo visual, es necesario contemplar el proceso de entendimiento del espectador cuando se enfrenta cara a cara con las visualizaciones. Este proceso se soporta inicialmente en las etapas de entendimiento de Kirk.

Etapas de entendimiento

El espectador de las visualizaciones atraviesa tres etapas en el proceso de entendimiento: percepción, interpretación y comprensión. (Kirk, A. 2019. P. 22):

- **Percepción:** se define como la capacidad de identificar qué está viendo el público o espectador, en cuanto a formas, colores, letras, etc. En sí, es el reconocimiento básico de lo que está observando.
- **Interpretación:** se refiere a la capacidad del espectador de entender el significado de lo observado, es decir, poder relacionar la información con algún contexto lógico.
- **Comprensión:** se otorga principalmente a la capacidad intelectual del espectador de darle un significado propio a la información, lo cual conlleva al entendimiento final del mensaje.

Varios factores afectan considerablemente la interpretación visual del público, cada espectador —según su contexto— podría tener un entendimiento propio y diferente al de los demás. Esta es la razón por la que se debe guiar a la audiencia a través de la historia.

Contar la historia

Para presentar la información (análisis aclaratorio), Knaflic (2017) propone seis pasos para “[…] comunicarse de manera mejor con los datos”:

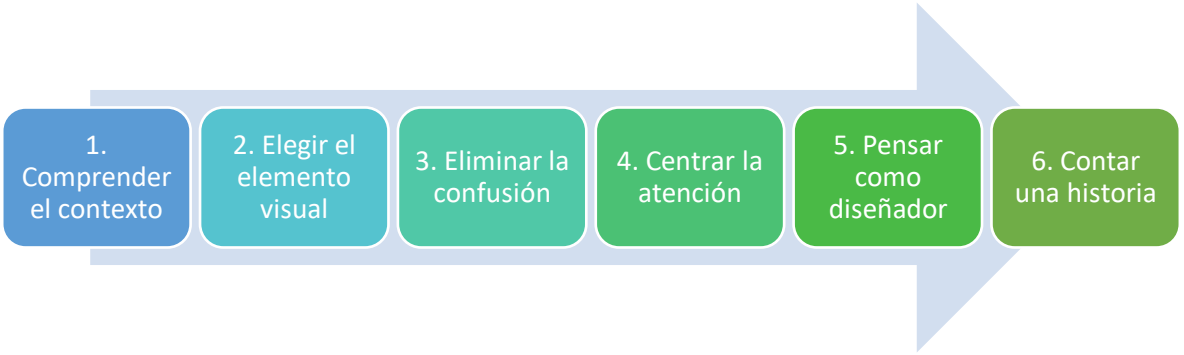


Figura 5. Pasos para comunicar los datos. Fuente propia

1. Comprender el contexto: Ya se hizo referencia anteriormente sobre este paso. Es aquí donde se busca conocer al público objetivo con el mayor detalle posible, para comprender hacía dónde debe dirigirse la historia.
2. Elegir el elemento visual apropiado: Dentro de un océano de gráficos, los más usuales son, Texto simple, tabla, mapas de calor, dispersión, líneas, gráfica de pendiente, columnas, columnas apiladas, cascada, barras, barras apiladas y cuadrantes. No existe uno correcto, todo depende de la presentación.
3. Eliminar confusión: Quitar elementos distractores. Si bien la cantidad de elementos que podrían omitirse en un objeto visual es elevada (lo que también depende del mensaje que se quiera transmitir), hay elementos que generalmente "sobran" en una visualización o que deben tenerse en cuenta para crear armonía. Las siguientes son algunas recomendaciones para la construcción de visualizaciones:

    - No usar elementos 3D, la lectura se puede distorsionar causada por la dimensionalidad de los objetos.
    - Tener un orden visual soportado en la forma en la que el ser humano lee, de izquierda a derecha y de arriba abajo. Lectura en Z.
    - Alinear los bordes de los objetos visuales y de texto. No hacerlo podría generar distorsión en la lectura e incomodidad en el plano general.
    - Usar estratégicamente los contrastes, principalmente si algo se quiere resaltar, en general los colores deberían ir en armonía, principalmente cuando se quiere hacer uso de colores empresariales.
    - Usar espacios en blanco. Este es uno de los atributos preatentivos, que se presentarán más adelante. Es comparable al uso de pausas en una exposición oral, con lo que se busca generar una pausa intencional al espectador con la intención de crear expectativa y dar espacio para entendimiento o cuestionamiento.
    - Eliminar bordes de gráficos, alineado a buscar la mayor limpieza posible de los elementos visuales. Si no hay aporte informativo del elemento visual, este debería omitirse.

- Eliminar líneas de cuadricula con la misma finalidad anterior de mantener la mayor limpieza de la visualización. Aplicable en gráficos y en tablas, en estas últimas se soporta la decisión con los principios de Gestalt.
- Eliminar puntos de datos, recomendado principalmente en los gráficos de líneas para suavizar la línea, sin crear interrupciones para el ojo en la lectura.
- Manejo adecuado de etiquetas, haciendo foco en dos puntos: eliminar las etiquetas de los ejes y etiquetando directamente los datos. Esta recomendación se alinea con la de eliminar las líneas de cuadricula en gráficos permitiendo una lectura más limpia y directa.
- Usar el color de forma coherente, buscando principalmente la armonía entre los colores, evitando la saturación extrema lo que podría dificultar la lectura y distraer sobre el mensaje que se quiere emitir.

4. Centrar la atención donde nos interesa: Con el uso adecuado de los atributos preatentivos tanto en gráficos como en texto. La finalidad de estos atributos es aprovecharlos en dos vías, dirigir la atención hacia un punto específico y la creación de jerarquías para dirigir el mensaje.

Este es un ejemplo de **atributo preatentivo** utilizando negrita	Este es un ejemplo de *atributo preatentivo* utilizando cursiva	Este es un ejemplo de atributo preatentivo utilizando subrayado
Este es un ejemplo de atributo preatentivo utilizando color	Este es un ejemplo de atributo preatentivo utilizando tamaño	Este es un ejemplo de atributo preatentivo utilizando espacios

Figure 6. Ejemplos de atributos preatentivos en texto. Fuente propia

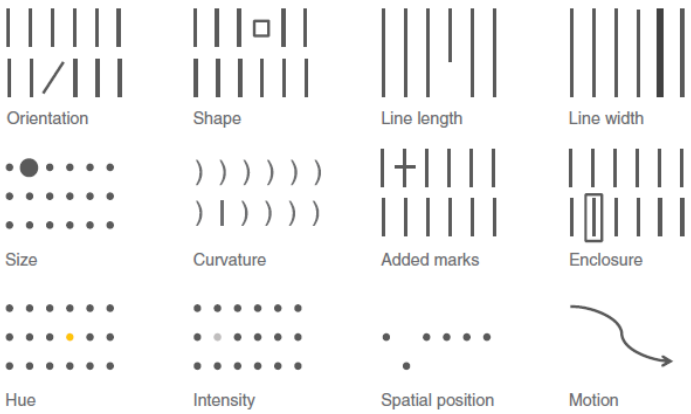


Figura 7. Few, S. (2012). Preattentive attributes. Show me the numbers. Recuperado de https://laptrinhx.com

5. Pensar como un diseñador: Aduciendo a “los 8 sombreros del diseño de visualización de datos” de Kirk, A (2019), el sombrero de diseñador es el encargado de asegurar la armonía entre la forma y la función, haciendo foco en los detalles visuales.
   En este punto deben tomarse todos los conceptos previamente explicados y aplicarlos de mejor forma para enviar el mensaje que se quiere. Se analizan tres elementos:
   - Ofrecimientos estimulares, con el aprovechamiento de los atributos preatintivos buscando destacar lo importante y creando jerarquías
   - Accesibilidad para el usuario, es decir que la mayor cantidad de personas puedan entender (sin dejar a un lado el foco de público objetivo). El mensaje debe ser legible, nítido y con lenguaje sencillo.
   - Estética, asociado a las recomendaciones para eliminar la confusión, se busca armonía en el color, alineación (principios de Gestalt) y uso de espacios en blanco (principios preatentivos).

6. Contar una historia: Etapa final donde toma forma y sentido todo lo anterior. Se busca guiar al usuario a través de una historia con inicio, nudo y desenlace.

3 MODELOS DE EXTRACCIÓN, SIMULACIÓN Y ANÁLISIS DE DATOS

La metodología se presenta en tres partes, el modelo ETL de las bases entregadas por el sistema, el modelo para simular el destino de los usuarios, y, el modelo de definición de los gráficos y la construcción de los tableros.

3.1 Modelo ETL

El modelo de extracción, transformación y carga de datos parte desde el punto de que el sistema comparte un archivo por cada día de operación, lo que en este caso corresponde a 272 archivos independientes, los cuales en una primera instancia deben ser unificados para minimizar la cantidad de procesos, al unificar los archivos por mes, el primer filtro de columnas seleccionadas se aplica únicamente a 9 bases. De ahí en adelante, y según la base final que se quiera obtener, el segundo pasó más importante es la agrupación de los registros para minimizar la cantidad de filas permitiendo disminuir los tiempos de transformación de las tablas.

Para iniciar el modelo, se presentan los metadatos de las tablas antes del proceso (bases de entrada), seguida por el proceso de transformación.

Tabla 1. Metadatos de las tablas entregadas por el sistema

Campo	Descripción
Fecha de Clearing	**Desconocimiento de significado
Fecha de Transacción	Fecha y hora de registro de transacción
Day Group Type	**Desconocimiento de significado
Hora Pico SN	Categorización de hora pico o valle
Fase	Fase en la que se construyó la estación
Emisor	Emisor de la tarjeta utilizada para el ingreso
Operador	**Desconocimiento de significado
Línea	Zona sobre la que se encuentra la estación de ingreso
Estación	Estación que registra la transacción (de ingreso)
Acceso de Estación	Entrada de la estación en la que se registró el ingreso
Dispositivo	**Desconocimiento de significado
Tipo de Tarjeta	Tarjeta utilizada para el ingreso
Nombre de Perfil	**Desconocimiento de significado
Número de Tarjeta	Número de la tarjeta de cada usuario
Tipo de Tarifa	**Desconocimiento de significado

Saldo Previo a Transacción	Saldo en pesos (COP), antes de realizar la transacción de ingreso
Valor	Costo de la transacción (de ingreso)
Saldo Después de Transacción	Saldo después de restar el saldo previo menos el valor

Transformaciones iniciales:

a) Dado que las bases iniciales corresponden a las transacciones de cada día del año, la primera transformación que se aplicó fue la unión de las bases por mes, seguido de la elección de las columnas (*Fecha de transacción, Estación y Número de tarjeta)*.
b) Se continuó con la división de la columna *fecha de transacción,* en cinco columnas: año, mes, día, hora, minutos y se aproximaron los minutos a la décima más baja.

Tabla 2. Selección de columnas, proceso ETL

Estacion	Numero de Tarjeta	Año	Mes	Dia	Horas	Minutos
(02502) Terminal	1010000129936860	2019	9	30	14	30
(02502) Terminal	1010000147499891	2019	9	30	14	30
(02502) Terminal	1010000138084496	2019	9	30	14	30
(02502) Terminal	1010000129202479	2019	9	30	14	30

Para la base de flujo:

c) Se realizó un conteo de registros, agrupando por *estación, año, mes, día, hora, minutos.*

Tabla 3. Columnas base de salida 'Flujo'

Estacion	Año	Mes	Dia	Hora	Count*(Numero de Tarjeta)
(02000) Cabecera Autopista Norte	2019	10	1	10:0	781
(02000) Cabecera Autopista Norte	2019	10	1	10:10	810
(02000) Cabecera Autopista Norte	2019	10	1	10:20	806
(02000) Cabecera Autopista Norte	2019	10	1	10:30	643

Para la base de conteo:

d) Se crearon dos columnas binarias (*am* y pm), evaluando si la transacción es hecha antes del mediodía (0), o después (1).

e) Se realizó un conteo de registros para las columnas *am* y *pm*, agrupando por *estación* y *número de tarjeta*.

Tabla 4. Columnas base de salida 'Conteo'

S Estacion	L Numero de Tarjeta	D Sum(am)	D Sum(pm)
(02000) Cabecera Autopista Norte	1010000000000638	1	0
(02000) Cabecera Autopista Norte	1010000000000711	2	0
(02000) Cabecera Autopista Norte	1010000000004622	0	1
(02000) Cabecera Autopista Norte	1010000000012401	2	1

Para la base de O-D:

f) Se identificó la *estación* con el conteo máximo en la columna *am (Origen)*, y la estación con el conteo máximo en la columna *pm (Destino);* ambos cálculos a nivel de *número de tarjeta*. **En este paso quedaron identificadas las estaciones origen y destino del usuario.**

g) Se eliminaron los registros que no están identificados como máximos.

h) Se unificaron las columnas *am* y *pm*, dejando una sólo columna *O-D* con registros Origen o Destino.

Tabla 5. Columnas base de salida 'O-D'

S Estacion	L Numero de Tarjeta	I OD
(02000) Cabecera Autopista Norte	1010000000000943	7
(02000) Cabecera Autopista Norte	1010000000004622	7
(02000) Cabecera Autopista Norte	1010000000008045	5
(02000) Cabecera Autopista Norte	1010000000008144	7

Nota: Por efectos de practicidad, se manejó formato numérico en la columna OD, donde 7 significa *destino* y 5 significa *origen.*

Para la base de matriz:

i) Se realizó un pivote de la columna *O-D,* a nivel de *número de tarjeta*. Teniendo como resultado dos columnas (*Estación origen* y *Estación destino*) con el nombre/código de la estación correspondiente.

Tabla 6. Columnas base de salida 'Matriz'

L Numero de Tarjeta	S 5+First(Estacion)	S 7+First(Estacion)
1010000000000042	(03012) SUBA - CALLE 95	?
1010000000000620	(02105) Calle 142	(09119) Calle 57
1010000000000711	(02502) Terminal	(09110) Avenida Jimenez
1010000000000737	(10002) Av. Primero de Mayo	(04104) Avenida 68

El proceso se realizó con el programa Knime. Los resultados de cada base se exportaron en archivos CSV, almacenándolos en el disco duro del PC para luego ser cargados en la versión desktop de Power BI, donde se realizaron las visualizaciones.

3.2 Modelo simulación destino

El modelo de simulación se sustenta en la hipótesis de que las personas que utilizan el sistema tienen un comportamiento rutinario todos los días. Se asume que todas las mañanas el usuario se dirige a la estación del sistema más cercana a su hogar, y en las tardes se dirige a la estación más cercana a su trabajo; esto, sumado a la jornada laboral que inicia en la mañana y finaliza en horas de la tarde, permitiría deducir que la estación que registra el ingreso del usuario en las tardes es su estación destino de las mañanas; y viceversa, la estación destino de las tardes será la que registre el ingreso del usuario en las mañanas.

La simulación propone que, haciendo un conteo por usuario de las veces que ingresó al sistema durante un rango de tiempo, se pueden conocer las estaciones más recurrentes en horarios de la mañana y de la tarde, de esta forma se define cuál es su estación origen en cada momento del día (mañana y tarde), y dependiendo del punto de vista, la estación contraria será su destino.

Para cada usuario, asumiendo un origen i y destino j donde se registra el conteo de ingresos R al sistema, la estimación del origen O y el destino D estará definida por:

$$O = Max \sum R_i$$

$$D = Max \sum R_j$$

Donde:

- $\boldsymbol{R}$ = ingresos del usuario al sistema
- $\boldsymbol{i}$ = estaciones de ingreso en las mañanas
- $\boldsymbol{j}$ = estaciones de ingreso en las tardes
- $\boldsymbol{O}$ = la estación i con mayor número de ingresos R
- $\boldsymbol{D}$ = la estación j con mayor número de ingresos R

Cabe aclarar que la metodología no busca usar modelos estadísticos para predecir con exactitud la información, lo que busca es usar la información histórica para identificar comportamientos recurrentes del usuario asumiendo el mismo comportamiento futuro.

3.3 Modelo de visualización

Para el modelo de visualización se tomó como base el hecho de que al dashboard se cargan 4 bases de diferente granuralidad, lo que se deriva en diferentes niveles de información con comportamiento dinámico distinto, haciendo referencia específicamente a los filtros aplicados. Como segundo paso, se tienen en cuenta las características de diseño de visualizaciones explicadas anteriormente, para facilitar la interfaz de los tableros buscando encontrar una comprensión e interacción eficiente por parte del usuario.

A continuación, se presentan las consideraciones que se tuvieron en cuenta para la definición de elementos visuales y la creación de los dashboards:

a) En vista de que se crearon cuatro bases con diferentes finalidades, el tablero se realizó imitando la estructura del dashboard de Dublin, donde no se busca tener todos los indicadores en una misma página, sino que se busca presentar la información categorizada por temas.

b) Se decidió darles doble funcionalidad a los gráficos geográficos: como fondo, ambientado el Dashboard, y como resultado dinámico de lo que está filtrado.

c) Los colores se definen con la bandera de Bogotá (amarillo y rojo) buscando un contraste para resaltar la información. En ocasiones se hace uso del azul como un segundo atributo de resaltado.

d) Se usan únicamente barras horizontales aprovechando su ventaja en la forma de lectura (lectura en Z), para una interpretación más rápida.

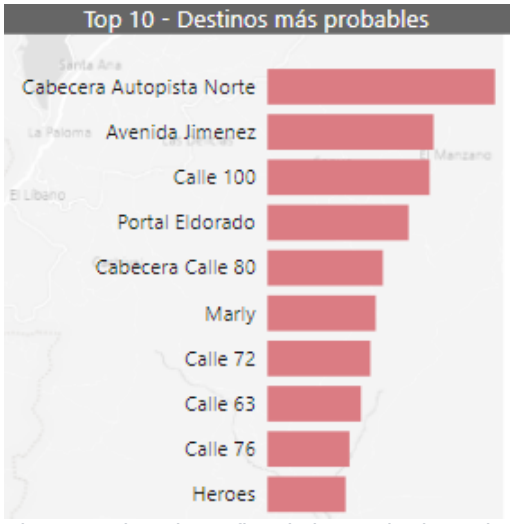


Figura 8. Ejemplo grafico de barras horizontales. Fuente propia

e) La gráfica de usuarios por hora es la única que se decide usar con barras verticales, debido a que no se requiere tener una lectura detallada de cada barra. La intención de este gráfico es mostrar todo el panorama y enfocar la atención del usuario en un punto específico, por lo que aquí se utiliza la herramienta de atributo preatentivo resaltando con otro color las barras de mayor longitud. Adicionalmente, la lectura temporal se lee más fácil de izquierda a derecha, que de arriba hacia abajo.

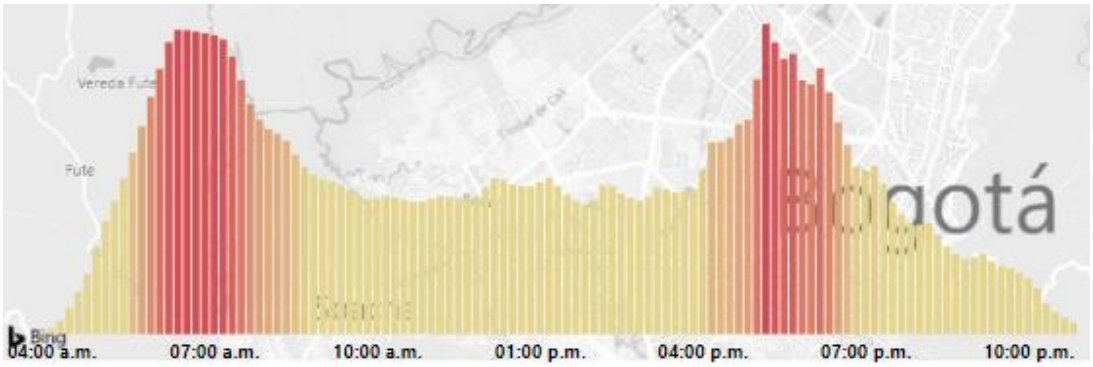


Figura 9. Gráfico de usuarios por hora. Fuente propia

f) Para los filtros se usó el criterio de cantidad de registros, si hay muchas opciones para filtrar se decide usar un menú desplegable de tal forma que no tuviera que ajustarse una visualización larga que ocupara una gran área dentro del gráfico. Por otro lado, si las opciones son pocas o de fácil acomodación se utilizan botones en forma de cuadrícula, como se hizo con los meses y los días (aquí destaca el principio de proximidad de Gestalt, dando interpretación de fecha a todo el conjunto).

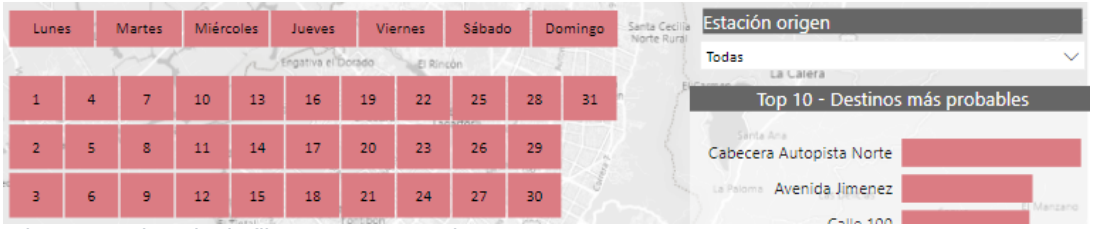


Figura 10. Ejemplo de filtros. Fuente propia

4 RESULTADOS Y CONTRIBUCIÓN

En el presente apartado se presentan los resultados obtenidos basados en los objetivos planteados. Inicialmente se presentan las arquitecturas resultantes del proceso ETL: la arquitectura de datos mostrando el modelo dimensional propuesto y la arquitectura de aplicaciones utilizada para el proceso. En el segundo apartado se presenta el modelo de simulación para predecir el destino de los usuarios basado en el análisis histórico de datos. Y por último, se presentan los tableros que conforman la herramienta para visualizar la dinámica de los viajes del sistema.

4.1 Arquitectura de datos

Cada base de entrada corresponde a un día del año con un promedio de dos millones de registros, donde cada registro corresponde al ingreso de un usuario al sistema, por lo que las bases correspondientes a un mes tendrán aproximadamente sesenta millones de registros. El peso aproximado de los archivos es de 500Mb (fines de semana entre 130Mb y 300Mb), por lo tanto, el peso total de los nueve meses es aproximadamente de 130.000Mb (130Gb).

Teniendo en cuenta lo anterior, se pensó en generar un muestreo de los datos, sin embargo, con el programa Knime se logró procesar toda la información sin tener la necesidad de filtrarla. Como resultado final, después del proceso ETL, la base con mayor detalle quedó con un peso de 3.400Mb (3,4Gb).

Las 4 bases resultantes del proceso ETL, más la base de ubicación de las estaciones, se relacionaron en un modelo estrella, dejando como centro la base de estaciones y como llave primaria el código de la estación.

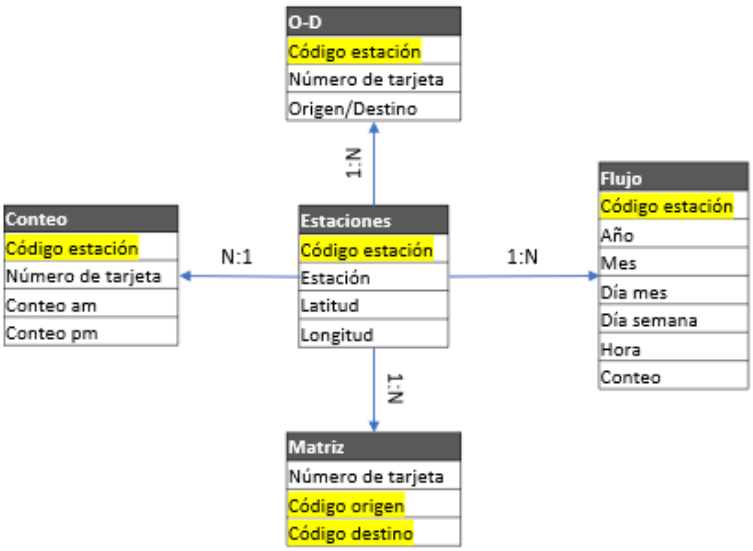


Figura 11. Modelo estrella de datos. Fuente propia

En la siguiente tabla se muestran los metadatos de las bases de salida del proceso ETL, utilizadas como entrada en el modelo.

Tabla 7. Metadatos tablas de salida del proceso ETL

Base	Campo	Descripción	Ejemplo
Conteo	Código estación	Código de la estación de ingreso	2101
	Número de tarjeta	Número de la tarjeta de cada usuario	1010 0001 4552 9030
	Conteo am	Sumatoria de usuarios que ingresan antes del medio día	45
	Conteo pm	Sumatoria de usuarios que ingresan después del medio día	38
O-D	Código estación	Código de la estación de ingreso	2101
	Número de tarjeta	Número de la tarjeta de cada usuario	1010 0001 4552 9030
	Origen/Destino	Variable dicotómica	Origen o Destino
Matriz	Número de tarjeta	Número de la tarjeta de cada usuario	1010 0001 4552 9030
	Código origen	Código de la estación de origen	2101
	Código destino	Código de la estación destino	2300
Flujo	Código estación	Código de la estación	2101
	Año	Año del registro	2019
	Mes	Mes del registro (2 a 10)	4
	Día mes	Día del mes (1 a 31)	15
	Día semana	Día de la semana (lunes a domingo)	Martes
	Hora	Hora:minuto, aproximado a la décima más baja	11:20:00 o 16:40:00
	Conteo	Sumatoria de todos los usuarios que ingresan al sistema	621
Estaciones	Código estación	Código de la estación	2101
	Estación	Nombre de la estación	Calle 161
	Latitud	Número latitudinal de la estación	4,74154193
	Longitud	Número longitudinal de la estación	-74,04796835

Por su parte, las aplicaciones utilizadas fueron:

- Knime (plataforma gratuita para minería de datos), para desarrollar todo el modelo ETL. Su interfaz gráfica no requiere conocimientos de programación y cuenta con gran variedad de opciones de transformación de datos. Por otro lado, la herramienta tiene la capacidad de usar gran parte de la memoria RAM del computador, lo que le permite procesar gran cantidad de datos en corto tiempo versus otras opciones gratuitas.
- Power BI (plataforma gratuita de Microsoft para inteligencia de negocios), es la herramienta que se encuentra liderando el cuadrante mágico de Gartner 2021 para plataformas de análisis y business intelligence (Gartner, 2021), por esta razón, sumado al hecho de ser una herramienta gratuita, y la facilidad de interfaz y capacidad de procesamiento, se selecciona como la herramienta para el diseño de las visualizaciones.

Ambas plataformas se alimentan por archivos csv. Con Knime se inicia el proceso ETL y se finaliza con la carga de los datos a Power BI. La arquitectura de las aplicaciones es la siguiente:

Figura 12. Arquitectura de aplicaciones del modelo. Fuente propia

4.2 Modelo de simulación

El modelo de simulación se sustenta en las bases *O-D* y *Matriz* (explicadas anteriormente). Ambos conjuntos de datos contienen la misma información estructurada de forma diferente: la base *O-D* repite el registro de número de tarjeta variando el valor de la columna OD (Origen/destino), mientras la base *Matriz* muestra por cada registro dos columnas, el origen y el destino. Ambas bases permitieron, eventualmente, elaborar diferentes elementos visuales para ver la información en otras vías.

Como se observa en la imagen, ambas tablas (O-D y Matriz) reducen los datos de forma significativa. En este ejemplo aislado de un usuario, se lograron disminuir 344 registros a 2, en el caso de la tabla O-D; y a 1, en el caso de la tabla Matriz.

Estación	Total viajes	Am	Pm
Cabecera Autopista Norte	16	13	3
Calle 100	2	1	1
Calle 106	95	0	95
Calle 127	17	1	16
Calle 146	187	179	8
Calle 45	1	1	0
Calle 72	1	0	1
Calle 85	2	0	2
Centro Comercial Santa Fe	1	0	1
Corferias	1	0	1
Marly	1	0	1
MazurÃ®n	1	1	0
MOVISTAR ARENA	1	0	1
Pepe Sierra	6	0	6
Portal Eldorado	2	0	2
Salitre El Greco	8	0	8
Terminal	2	2	0
Total	**344**	**198**	**146**

Tabla O-D

Numero de Tarjeta	Estación	OD
1010000036443729	Calle 106	Destino
1010000036443729	Calle 146	Origen

Tabla Matriz

Numero de Tarjeta	Origen	Destino
1010000036443729	Calle 146	Calle 106

Figura 13. Ejemplo reducción registros. Fuente propia

Retomando la ecuación explicada anteriormente, se tiene:

$$O = Max \sum R_i$$

$$D = Max \sum R_j$$

- $\boldsymbol{R}$ = ingresos del usuario al sistema
- $\boldsymbol{i}$ = estaciones de ingreso en las mañanas
- $\boldsymbol{j}$ = estaciones de ingreso en las tardes
- $\boldsymbol{O}$ = la estación i con mayor número de ingresos R
- $\boldsymbol{D}$ = la estación j con mayor número de ingresos R

Tabla 8. Aplicación modelo de simulación

i	R
Cabecera Autopista Norte	13
Calle 100	1
Calle 127	1
Calle 146	179
Calle 45	1
Mazuren	1
Terminal	2

O = Calle 146
D = Calle 106

j	R
Cabecera Autopista Norte	3
Calle 100	1
Calle 106	95
Calle 127	16
Calle 146	8
Calle 72	1
Calle 85	2
Centro Comercial Santa Fe	1
Corferias	1
Marly	1
Movistar Arena	1
Pepe Sierra	6
Portal El Dorado	2
Salitre El Greco	8

El aporte no sólo se da en la reducción del tamaño de las bases; al poder darle una nueva categoría a las estaciones, nacen otras opciones para analizar la información: aislando una estación origen, se pueden evaluar los destinos más recurrentes que tienen los usuarios que ingresan a dicha estación; y de igual forma en el caso contrario, aislando el destino se pueden evaluar los orígenes más comunes de los usuarios que descienden en dicha estación.

4.3 Modelo de visualización

Como se mencionó anteriormente se utilizó power bi para el diseño del dashboard. Los archivos CSV generados en el proceso ETL se conectaron a power bi completando el proceso ETL, cargando directamente los datos en el programa.

En la opción de *Modelo,* al interior de Power BI, se hizo el relacionamiento de las tablas en el modelo estrella explicado en el capítulo de metodología, de este modo las dimensiones se centralizan en la tabla de *Estaciones*, dejando como llave primaria el código de las estaciones.

Figura 14. Modelo estrella de la simulación. Fuente propia

A continuación, se describe cada una de las pestañas con su funcionamiento y finalidad:

Principal: Esta pestaña se diseña con la finalidad de visualizar la concentración de usuarios a lo largo del día en intervalos de 10 minutos. La información puede ser filtrada por mes, día del mes o día de la semana, y estación. Al aplicar filtro de estación se filtran los 10 destinos con mayor probabilidad de destino para dicha estación.

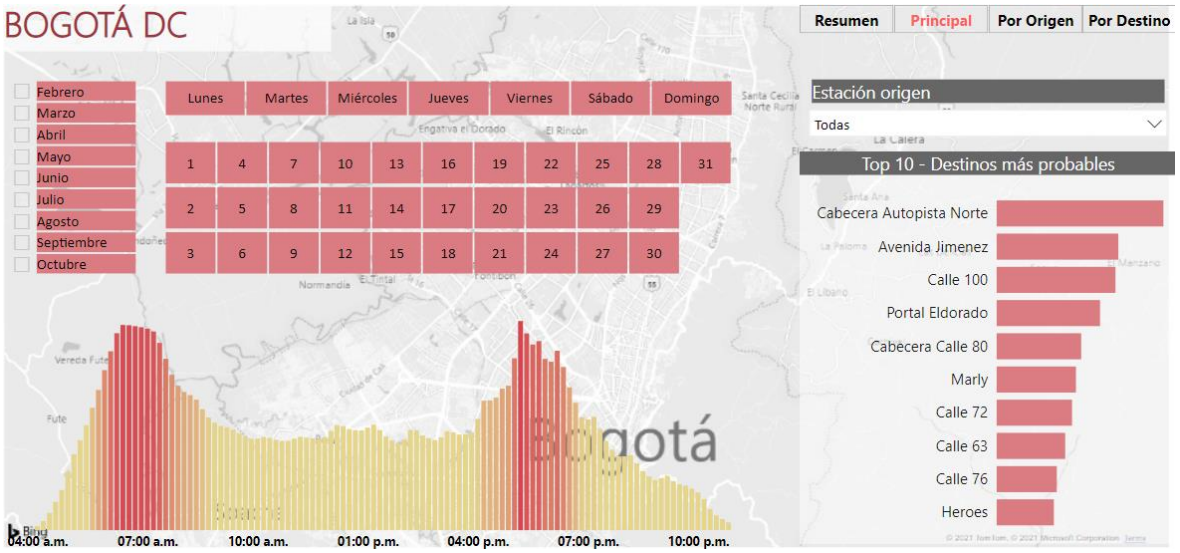


Figura 15. Página "Principal" del dashboard. Fuente propia

Por tarjeta: Esta pestaña se diseña para ubicar geográficamente las estaciones Origen y Destino que se estimaron con el modelo. Aquí se aplica en su totalidad el modelo de simulación de destino, permitiendo visualizar por número de tarjeta la información histórica. Al filtrar el número de tarjeta, se realiza el conteo total de viajes, el conteo de estaciones visitadas (con su respectivo listado) y se visualizan geográficamente las estaciones de origen y destino estimadas por el modelo.

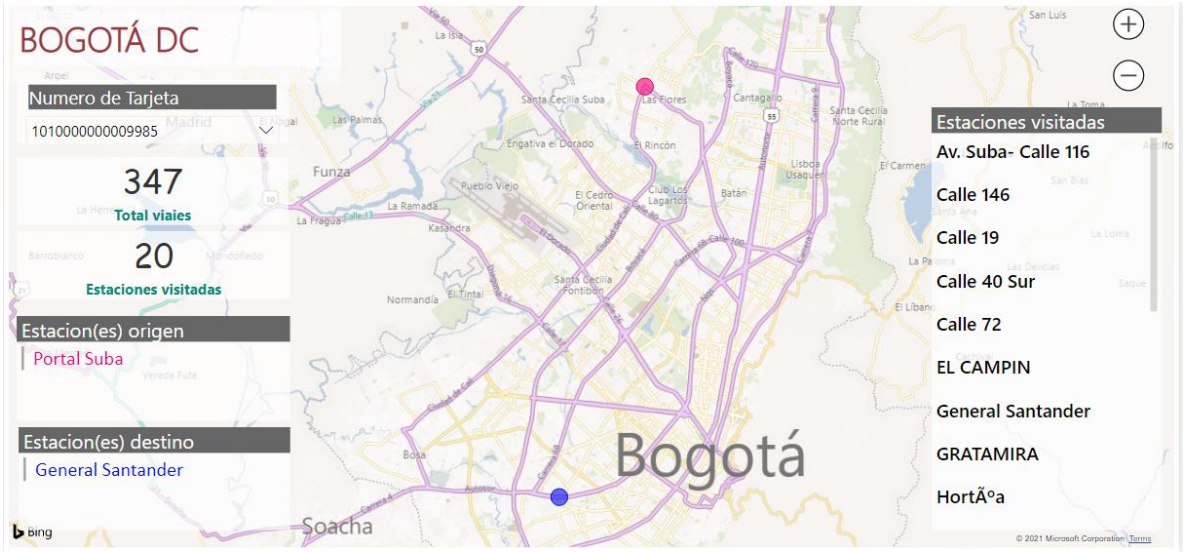


Figura 16. Página "Por tarjeta" del dashboard. Fuente propia

Por Origen/Por destino: Esta pestaña se diseñó con la finalidad de hacer uno de los análisis que surgen a raíz del modelo. Cuando se definen los orígenes y destinos de los usuarios es posible cambiar el enfoque y centrar el análisis en las estaciones. En este punto, la cualidad de Origen/Destino pasa a ser una categoría de filtro para la estación, en otras palabras, es posible filtrar una estación dándole el atributo de Origen y según el número de usuarios que se movilizan en esa estación, determinar el top de los destinos mas recurrentes por dichos usuarios. La visualización cuenta con filtro de Origen (para filtrar la estación) arrojando como resultado: el flujo de usuarios que se movilizan por mes y por día de la semana, y la ubicación geográfica de los 20 destinos más probables para los usuarios que ingresan a esa estación. También es posible filtrar por hora (con rangos de 10 minutos), sin embargo, este filtro aplica únicamente a los flujos.

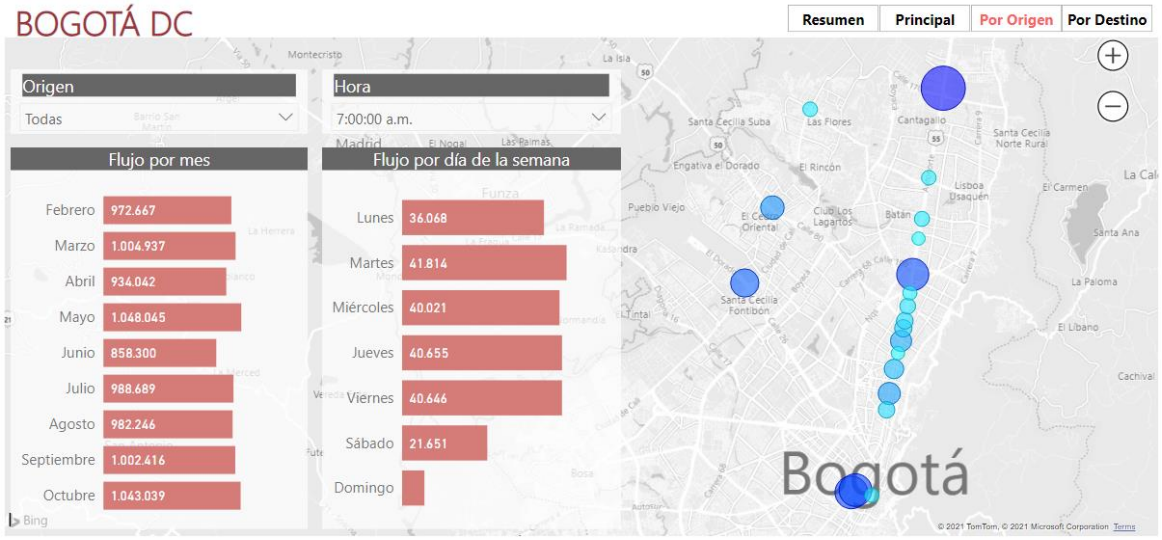


Figura 17. Página "Por Origen" del dashboard. Fuente propia

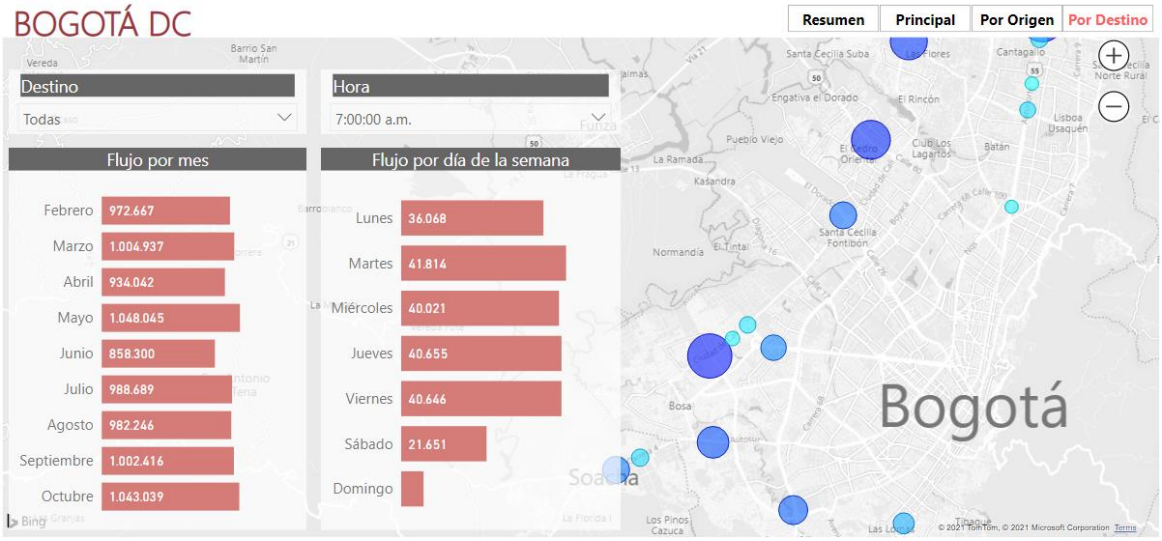


Figura 18. Página "Por Destino" del dashboard. Fuente propia

Contribución

Si bien se han propuesto trabajos para registrar la salida de los usuarios en el sistema Transmilenio, todos requieren una inversión económica, tecnológica, de infraestructura y de dinámica de uso. En ocasiones el sistema ha realizado pruebas para determinar el destino de los usuarios entregando códigos de barras al ingreso de la estación con la esperanza de que el usuario lo entregue en la estación de salida, sin embargo, esta prueba está sujeta a variables que podrían no mostrar una información real, adicional al costo que conlleva.

Por lo anterior, la aplicación del modelo contribuiría a un ahorro económico, ya que se aprovecha la información que ya recolecta el sistema y no requiere ninguna inversión adicional en tecnología o infraestructura.

Los datasets puros y transformados, se cargan en la web (Anexo 6) para su aprovechamiento por parte de la ciudadanía y el desarrollo de trabajos futuros. Adicionalmente, se carga el dashboard (excluyendo la pestaña *Por tarjeta*, por motivos de seguridad de datos) para el uso de los usuarios del sistema.

5 CONCLUSIONES Y RECOMENDACIONES

El proceso de visualización inicia antes de crear el primer gráfico, desde la transformación de los datos y definición del público objetivo. Lo primero permitirá estructurar los datos en un modelo eficiente que mejore las capacidades de procesamiento, y lo segundo ayudará a definir el modelo visual que permita responder preguntas de negocio.

Para la ejecución del proceso ETL, es clave la elección de las aplicaciones a utilizar de tal forma que logre un proceso eficiente en términos del tiempo de procesamiento. En el presente trabajo se inició el almacenamiento y transformación en Azure datastudio, sin embargo, al cambiar de aplicación a Knime, los tiempos de transformación disminuyeron en más de un 50%.

Para realizar un modelo de simulación y visualización, como el del presente trabajo, se recomienda: 1. Suficientes datos para encontrar recurrencias (al menos tres meses, con usuarios que tengan más de 24 transacciones mensuales), 2. definición de usuario a quién va dirigido el análisis, y 3. elección correcta de herramientas para procesar y comunicar los resultados.

La validación del sistema se pudo hacer con casos reales, filtrando los códigos de las tarjetas de usuarios del sistema encontrando dos resultados: algunos usuarios no son recurrentes en el uso del sistema por lo que los resultados no eran significativos (en este grupo entran los usuarios que adquirieron la tarjeta de forma tardía con respecto a los datos analizados); por otro lado, los usuarios recurrentes vieron reflejado su historial de viajes validando razonablemente las estaciones propuestas como origen y destino.

Como fruto de este trabajo podrían proponerse otras líneas de investigación, por ejemplo, combinando información de capacidad de estaciones y buses para evaluar saturación y capacidad máxima del sistema, o combinando la información de rutas y las estaciones que se utilizan como transbordos, permitiendo tomar decisiones de negocio como el alquiler de más o menos buses a los operadores.

Otra posibilidad de trabajo futuro estaría enfocada al procesamiento de los datos en tiempo real, dado que ya se tiene una estimación del destino del pasajero, podría identificarse su ingreso al sistema, o incluso predecir la hora a la que va a ingresar, y preparar a los buses para prestar un servicio eficiente en cuanto al trayecto estimado de los usuarios.

6 REFERENCIAS BIBLIOGRÁFICAS

Attardi, M. (2016). *Análisis de diseño de visualización interactiva de información* (tesis de maestría). Universidad Politécnica de Valencia. Valencia, España.

Carmack, J. (2015, 07, 10). *Throwdown: data visualization vs. infographics.* Recuperado de: https://visage.co

Cohen, B. (2018). *Blockchain cities and the smart city wheel.* Recuperado de: https://medium.com/iomob

Cruz, G., Gamboa, J., Martínez, S., González, B., Gacha, L. (2018). *Bogotá, Ciudad Inteligente.* Recuperado de: https://bogota.gov.co

Departamento Nacional de Planeación (2020). *Documento de lineamientos de política de ciudades inteligentes.* Recuperado de: https://www.dnp.gov.co

Federación Internacional de los Trabajadores de Transporte. (2021). Sistemas de bus de tránsito rápido (BRT). Recuperado de: https://www.itfglobal.org

Few, S. (2012). *Show me the numbers*. Recuperado de https://laptrinhx.com

González, O. (2021, 02, 21). *Procesa exitosamente tus datos gracias al ETL.* Recuperado de: https://www.appvizer.es

Jing, C., Du, M., Li, S., Liu, S. (2019). *Geospatial dashboards for monitoring smart city performance.* Universidad Ryerson. Toronto, Canada.

Kirk, A. (2019). *Data visualisation: A handbook for data driven design.* Londres, Inglaterra: Sage.

Kitchin, Rob & McArdle, Gavin. (2016). *Urban data and city dashboards: Six key issues.*

Knaflic, C. N. (2017). *Storytelling con datos.* Madrid, España: Anaya multimedia.

LaPaz Bus. (2021). *Tarifas.* Recuperado de: http://www.lapazbus.bo

Ovacen (2021). *Smart city: qué es, cómo funcionan, ventajas y desventajas de las Smart cities.* Recuperado de: https://ovacen.com/

Perez, F., Velázquez, G., Fernández, V., Dorao, J. (s.f.) *Movilidad inteligente.* Madrid, España: Mincotur.

Transmilenio. (2021). *Transmilenio S.A.* Recuperado de: https://www.transmilenio.gov.co/

Valero, J., Marín, B., Català, J. (2014). *Aproximación a una taxonomía de la visualización de datos.* Revista Latina de Comunicación Social, 69, pp. 486 a 507.

Volvo. (2021). *Sistema de autobuses de tránsito rápido.* Recuperado de: https://www.volvobuses.co

7 ANEXOS

Anexo 1. Plano de estaciones y portales de Transmilenio

Anexo 2. Link artefacto de visualización y bases de datos

https://app.powerbi.com/view?r=eyJrIjoiMzc5ODUxOTYtNGYyMS00OTU1LWJiY2QtMThkNWNmOTEzNGRmIiwidCI6IjUwNjQwNTg0LTJhNDAtNDIxNi1hODRiLTliM2VlMGYzZjZjZiIsImMiOjR9&pageName=ReportSectionc5157a1f0647063ba1a4