Diseño de data warehouse en la dirección de cobertura de la secretaria de educación distrital de Bogotá (SED)

Ivan Fabricio Aponte Diaz

Diana Marcela Benítez Cera

Gustavo Horacio Triana Lozada

Universidad EAN

Facultad de ingeniería

Maestría en inteligencia de negocios

Bogotá, Colombia

15 Octubre 2022

Diseño de data warehouse en la dirección de cobertura de la secretaria de educación distrital de Bogotá (SED)

Ivan Fabricio Aponte Diaz

Diana Marcela Benítez Cera

Gustavo Horacio Triana Lozada

Trabajo de grado presentado como requisito para optar al título de:

Magister en Inteligencia de Negocios

Directora

Sandra del Pilar Forero Poveda

Modalidad

Trabajo Dirigido

Universidad EAN

Facultad de ingeniería

Maestría en inteligencia de negocios

Bogotá, Colombia

15 Octubre 2022

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 15 Octubre 2022

Dedicatoria

A nuestras familias con mucho cariño, quienes son un ejemplo de empeño, dedicación, superación y quienes con su apoyo incondicional y su confianza nos han ayudado siempre a cumplir nuestras metas.

Ivan F. Aponte Diaz, Diana M. Benitez Cera y Gustavo H. Triana Lozada

Agradecimientos

Agradezco primero a mi familia por incentivarme a ser mejor persona día a día, dedico está maestría a mis padres que me dieron la vida, educación, y sobre sus consejos. Mi mama por su ejemplo de superación, fortaleza y dedicación, a mi papa por el apoyo, consejos y motivación a lo largo de mi vida, a mi hermana Tatiana, por su respaldo y admiración, a Camilo Montañez por su apoyo constante en mi vida personal y laboral y a todas y cada una de las personas que han hecho parte de este proceso.

Iván Fabricio Aponte Diaz

Agradecimientos

Dios sigue siendo bueno, el nunca incumple una promesa y por eso agradezco a el con todo mi corazón el desarrollo de esta tesis, a mi mamá, mi esposo, mi hija y toda mi familia que me apoyaron en este camino pero también agradezco a una persona que ha sido mi maestro, mi jefe y quién con mucha paciencia ha estado conmigo en este camino Álvaro Pabón. Esto ha sido un camino duro y todos ellos fueron las bases para poder seguir y cumplir mis sueños y objetivos.

Diana Marcela Benítez Cera

Agradecimientos

Dedico la presente tesis a mi madre por haber sido el gran ejemplo a seguir, por haber estado siempre en cada paso, por llenarme de valores, por guiarme a convertirme en la persona que soy, a mi esposa e hija quienes me han acompañado en este gran reto, brindándome todo su apoyo y siendo la motivación para continuar en los procesos de aprendizaje, a Dios por ayudarme a culminar mis objetivos y bendecirme al conformar un maravilloso hogar.

Gustavo Horacio Triana Lozada

Resumen

La secretaria desde hace ya varios años ha venido presentando inconvenientes con la cantidad de requerimientos recibidos con respecto a las variables de su actividad, el principal inconveniente es no contar con un repositorio centralizado de la información y esto causa que en cada requerimiento se repita una y otra vez los procesos de limpiezas y preparación de la data para dichos requerimientos, el siguiente proyecto tiene como fin la propuesta de un modelo de limpieza de la información y la propuesta del desarrollo de un data warehouse en la dirección de cobertura en la secretaria de educación distrital de Bogotá que permita darle solución a la problemática.

La propuesta da su inicio con la recepción de la data que recibe la secretaria y que llegaría a un archivo compartido que posteriormente será leído por un código en R y pasado por un proceso de limpieza hasta dejar la data disponible para el proceso de almacenamiento.

Los principales resultados del anterior proyecto se dan en la reducción de tiempos de respuesta a cada uno de los requerimiento, esto porque la data siempre estaría consolidada, limpia, preparada y disponible para cualquier requerimiento.

Palabras clave: Limpieza, Código R, almacenamiento, Data warehouse, respuesta.

Abstract

The secretary for several years now has been presenting a number of inconveniences with the requirements received with respect to the variables of its activity, the main drawback is not having a centralized repository of information and this causes a repetition of a once again the processes of cleaning and preparing the data for those requirements, the following project has as objective the proposal of an information cleaning model and the proposal of the development of a data warehouse in the direction of coverage in the secretary of Bogotá district education that allows solving the problem.

The proposal begins with the reception of the data received by the secretary and that would reach a shared file that will later be read by a code in R and passed through a cleaning process until the data is available for the storage process.

The main results of the previous project are given in the reduction of response times to each of the requirements, this because the data would always be consolidated, clean, prepared and available for any requirement.

Keywords: Cleaning, R Code, storage, Data warehouse, response.

Tabla de contenido

1. Lista de figuras

2. Lista de tablas

3. Introducción

Para poder entender la propuesta y desarrollo del presente trabajo de grado, se realiza una descripción de los antecedentes de la Secretaría de Educación del Distrito, conocer sus funciones permite entender como la solución puede ayudar a la operatividad del día a día de la SED, así como también explicar el problema específico que se busca solucionar con la propuesta planteada.

3.1 Antecedentes

La Secretaría de Educación del Distrito (SED) es la rectora de la educación inicial (preescolar), básica (primaria y secundaria) y media en Bogotá, de acuerdo con el Decreto 330 de 2008 mediante el cual se reestructuró la entidad. La secretaría tiene por objeto orientar y liderar la formulación y ejecución de políticas, planes y programas para garantizar el derecho a la educación y asegurar a la población el acceso al conocimiento y la formación integral. (Distrito, Secretaría de Educación del Distrito, 2021)

Como pilar misional está el de promover la oferta educativa en la ciudad para garantizar el acceso y la permanencia de los niños, niñas y jóvenes en el sistema educativo, en sus distintas formas, niveles y modalidades; la calidad y pertinencia de la educación, con el propósito de formar individuos capaces de vivir productiva, creativa y responsablemente en comunidad.

Como líder en el proceso de gestión de la cobertura educativa en el Sistema Educativo Oficial de Bogotá se encuentra la Dirección de cobertura que se encarga de la implementación de estrategias y acciones para garantizar el acceso y permanencia escolar a toda la población escolar, como Búsqueda Activa, Unidades Móviles de Atención, Matriculatones, Gratuidad Educativa, implementación de modelos y estrategias educativas flexibles para poblaciones de especial protección constitucional, entre otras, además de los programas

encaminados a bajar la deserción escolar e implementación de la Ruta de Acceso y Permanencia. (Distrito, Secretaría de Educación del Distrito, 2021)

De manera permanente ingresan consultas o requerimientos al área de cobertura que superan el número de hasta mil radicados por mes, siendo caracterizados por peticiones como:

- Identificar la cantidad de estudiantes matriculados por distintas variables
- Conocer la cantidad de estudiantes matriculados en una localidad con una condición especial
- Cantidad de estudiantes por género y localidad para un periodo
- Cantidad de estudiantes matriculados de población migrante por las 20 localidades del distrito
- Cantidad de estudiantes por tipo de establecimiento educativo para un periodo especifico
- Identificar la población matriculada por cada uno de los colegios de cualquier localidad
- Cantidad e estudiantes matriculados para la población con discapacidad visual en una jornada especifica
- Cantidad de estudiantes matriculados por localidad especifica entre rangos de edades (Distrito, Secretaría de Educación del Distrito, 2021)

El Ministerio de Educación Nacional (MEN) en aras de controlar la información de las instituciones y sus estudiantes implementó el sistema de matrícula estudiantil de educación básica y media (SIMAT) donde recopila todos sus datos, cada Secretaria de Educación Distrital (SED) es responsable de la administración correspondiente a la base estudiantil y las consultas o requerimientos donde se analizan datos referentes a distintas variables como son zona, localidad, institución educativa, tipo de institución educativa, nivel escolar, mes, jornada, grado, género, población migrante, condición migratoria, tipo de discapacidad, etnias, capacidad y/o talento excepcional. La respuesta a una solicitud demanda

verificar la información para cada variable en distintas bases de datos (locales independientes) lo cual genera reprocesos y pérdidas de tiempo. En conclusión, la SED no cuenta con un repositorio de almacenamiento centralizado donde pueda almacenar la información que se genera por medio del SIMAT por lo que se requiere de un esfuerzo mancomunado de tiempo y capacidad humana que, si se lograra condensar, permitiría disminuir los tiempos de respuestas, así mismo contar con una base actualizada y que responda a las necesidades inmediatas que tiene los usuarios. (Ministerio de Educación Nacional, 2021)

3.2 Descripción del problema

- Por años, el área de cobertura de la SED ha venido alimentando bases de datos robustas, que lo que pretenden es disponer de toda la información detallada de cada una de las variables antes mencionadas. Además de los problemas de almacenamiento, también se requiere de un proceso de extracción y limpieza que permita cargar la información con características óptimas -ordenadas, caracterizadas y sin errores de tipeo, para que se puedan resolver las peticiones de manera automática, pues en la actualidad, el acceso a la información se hace de forma independiente (almacenamiento mes a mes); esta situación acrecienta los procesos para llegar al resultado esperado, de tal manera que se requiere de más esfuerzo humano y desgaste logístico que se puede minimizar al contar con una herramienta que contenga todas las bondades de visualización de la información de manera amigable y entendible, facilitando dar respuesta a las solicitudes, al reducir los tiempos de consulta y de respuesta (10 a 15 días). Beneficiando tanto al emisor: Haciendo la búsqueda de la información más rápida y sencilla sin importar cuántas variables se deban consultar. Y es por eso que se pretende diseñar un modelo de almacenamiento y limpieza de datos que permita una visualización efectiva de la información.

3.3 Pregunta de investigación

¿Cómo mejorar y automatizar los procesos de limpieza y almacenamiento de datos de la dirección de cobertura de la SED para dar una respuesta más eficaz y efectiva de los requerimientos?

4. Objetivos de la investigación

4.1 Objetivo general

Diseñar un modelo de almacenamiento y limpieza de datos que permita una visualización efectiva de la información que responda a las necesidades y que permita obtener la información necesaria para dar la respuesta deseada por el usuario.

4.2 Objetivos específicos

- Establecer en la literatura los referentes teóricos necesarios para el diseño de un data warehouse.
- Realizar análisis situacional de la dirección de cobertura de la SED que permita conocer su estado actual y posibles oportunidades de mejora.
- Proponer la normalización de las bases de datos para garantizar la integridad de la información al interior de la dirección de cobertura de la SED.
- Diseñar data warehouse para cargar la información que ya ha pasado por un proceso de normalización.

5. Justificación

El presente trabajo de investigación pretende alcanzar una efectiva limpieza y transformación de la información al interior de la dirección de cobertura de la Secretaria de Educación Distrital, gracias a la implementación de un diseño de almacenamiento, que le dará a la entidad beneficios importantes reduciendo los y tiempos de capacidad de respuesta a los requerimientos que se encuentran como radicados; es decir, ya ingresaron al área de cobertura y se identificaron como requerimientos. Teniendo en cuenta la finalidad de la SED que es la de orientar y liderar la formulación y ejecución de políticas, planes y programas para garantizar el derecho a la educación y asegurar a la población el acceso al conocimiento y la formación integra, la entidad estaría como secretaria del distrito, abriendo una vitrina eficiente en capacidad de respuesta con un esquema como el que se pretende alcanzar.

Analizando el alcance esperado del proyecto, la entidad logrará eficiencia y eficacia al descentralizar la información, permitiendo la reducción de tiempos de respuesta a los radicados, sin perder de vista en ningún momento que lo que prima en el manejo de los datos, corresponde a la seguridad y mantenerlos blindados en su calidad, así mismo la dirección de cobertura podrá acceder a una información histórica y compacta que permitirá identificar patrones basados en el comportamiento de la información.

Teniendo en cuenta el papel que juega el área de cobertura en el actuar diario de la secretaria de educación en cuanto a la implementación de estrategias y acciones que garantizan el acceso y la permanencia a toda la población escolar. lograr la implementación de este proyecto permitirá cubrir muchas de las necesidades en cuanto a la respuesta de las solicitudes que se reciben de distintos actores.

Este proyecto consideró como punto de partida la importancia de tener acceso a la información, refiriéndose a la disposición y manipulación de las bases de datos (repositorios) en dónde se concentran de manera compacta todos los radicados. Además, este proyecto estaría respaldado para la puesta en marcha, ejecución y validación de resultados por la alta dirección de la entidad y el área beneficiada.

La Secretaria de Educación Distrital por ser parte del sector central de la administración distrital, en cabeza de la Alcaldía Mayor, opera como rectora de la educación inicial (preescolar), básica (primaria y secundaria) y media en Bogotá, pretende estar en innovación permanente en cada una de las áreas que la componen, así que implementar un modelo que facilite y optimice los procesos de respuestas a usuarios y la comunicación entre las distintas oficinas y direcciones de la entidad, siempre será bienvenido y apoyado por parte de las mesas directivas responsables, así mismo, el proyecto dispondrá de los recursos necesarios para la ejecución y avance, acceso a las bases de datos, un software de soporte y todas la herramientas tecnológicas que se demanden. Luego que se evidencien resultados positivos en la implementación de este modelo, será posible replicarlo en las demás dependencias y convertirse en un esquema a seguir en toda la entidad.

También se pretende tener un impacto en la sostenibilidad en la SED ya que es un tema de importancia a nivel global y que impacta a todos los sectores. Entre otros beneficios, se puede disminuir el gasto en papelería innecesaria al tener toda la información ordenada y disponible.

En la secretaria de educación el volumen de los datos se encuentra creciendo de forma exponencial, por lo tanto es necesario realizar una gestión eficaz que permita garantizar el uso de los recursos de almacenamiento y que estos datos se almacenen de forma segura, cumpliendo con las políticas de la SED y la normativa gubernamental, pues, debido a la pandemia se incrementó el trabajo remoto por lo que el traslado de los archivos se ha convertido en perdidas ocasionales de

información esto abre la posibilidad del almacenamiento en la nube que se comportan muy bien en todas las zonas geográficas, el tiempo y los usuarios. Adicionalmente garantiza que los datos estén a salvo de amenazas externas, errores humanos y fallos del sistema.

Con las amenazas tanto internas como externas, la seguridad del almacenamiento es más importante que nunca para una estrategia de gestión pues garantiza la protección y la disponibilidad permitiendo la accesibilidad de los datos a los usuarios autorizados y protegiéndolos contra el acceso no autorizado, estos datos corresponden a toda la información personal de la totalidad de la población escolar del distrito convirtiéndolos en datos sensibles, de esta manera se podrán compartir datos agregados que ayudarán a agilizar los procesos de respuesta en las peticiones de la ciudadanía en general.

6. Marco teórico

A continuación, se describe la información que permitió estructurar el modelo de almacenamiento y limpieza de datos en la dirección de cobertura de la Secretaría de Educación Distrital de Bogotá (SED).

Referentes Conceptuales:

- **Fuente de Datos.** Una base de datos es un conjunto de datos almacenados en memoria externa que están organizados mediante una estructura de datos. Cada base de datos ha sido diseñada para satisfacer los requisitos de información de una empresa u otro tipo de organización. (Marqués, 2009)
- **Recolección de Datos.** La etapa de recolección de datos se refiere a la obtención de los datos. En esta etapa, el sistema se conecta a las diferentes fuentes de información para extraerlos datos que luego se han de almacenar, procesar, analizar y visualizar. (Joyanes, 2019)
- Extracción de Datos.
  - Data profiling (subsistema 1): consiste en la exploración de los datos para verificar su calidad y si cumple los estándares conforme los requerimientos.
  - Change data capture (subsistema 2): detecta los cambios para refinar los procesos ETL y mejorar su rendimiento.
  - Sistema de extracción (subsistema 3): permite la extracción de datos desde la fuente de origen a la fuente destino. (Curto Diaz, 2016) Esta etapa engloba diferentes acciones, que van desde la extracción de datos de diferentes fuentes, el análisis o chequeo de estos datos, realizar una criba de los datos analizados para rechazar duplicidades y/o datos erróneos o sin valor y preparar los datos obtenidos para el siguiente proceso de transformación. (Lopez, 2018)

- **Transformación – Limpieza de Datos:**
    - Data Cleaning (subsistema 4): implementa los procesos de calidad de datos que permite detectar las incoherencias de calidad.
    - Rastreo de eventos de errores (subsistema 5): captura todos los errores que proporcionan información valiosa sobre la calidad de datos y permiten la mejora de estos.
    - Creación de dimensiones de auditoría (subsistema 6): permite crear metadatos asociados a cada tabla. Estos metadatos permiten validar la evolución de la calidad de los datos.
    - Reduplicación (subsistema 7): eliminar información redundante de tablas importantes como cliente o producto. Requiere cruzar múltiples tablas en múltiples sistemas de información para detectar el patrón que permite identificar cuándo una fila está duplicada.
    - Conformación (subsistema 8): permite identificar elementos equivalentes que permiten compartir información entre tablas relacionadas. (Curto Diaz, 2016)

  Se aplica un conjunto de reglas de unificación de datos básicos para transformar los datos desde el origen al destino. Esto incluye la conversión de los datos medidos a la misma dimensión, usando las mismas unidades, para que más adelante se puedan unificar. Una vez transformados los datos, es necesario realizar una serie de operaciones de depuración. Esta etapa es una de las más importantes, ya que garantiza la calidad de los datos por tratar (Joyanes, 2019)
- **Técnicas de Limpieza de Datos:** El proceso de limpieza de datos consiste en erradicar de una base de datos las anomalías y errores existentes en ella con el fin de que en el momento de esta ser utilizada sea más fácil su manipulación. Algunas técnicas para limpiar datos se toman de las diferentes necesidades que presenta cada una de las bases de datos con

las que se desea trabajar. Existen distintas técnicas que se plantean referente a este problema que procederemos a explotar el día de hoy:

- **Métodos Estadísticos:** Los métodos del cálculo de la media, la desviación estándar y rango, basados en el teorema de Chebyshev y considerando un intervalo de confianza para cada campo, sirven para determinar valor es excepcionales en campos y registros de datos. Aunque pueden generar muchos falsos positivos son rápidos y simples y pueden también ser combinados con otros métodos. (Beatriz E. – Ramiro Perez – 2009)
- **Algoritmo Soundex:** Utilizado para codificar palabras a partir de su sonido y realizar búsquedas de manera que no se igualen cadenas de caracteres. Los códigos de este algoritmo empiezan con la primera letra de la palabra seguida de un código de tres dígitos que representan las primeras consonantes. Para calcular un código Soundex se eliminan los espacios, puntuación, acentos y otras marcas, se eliminan a su vez cualquiera de los caracteres A, E, I, O, U, H, W, Y así como la segunda letra de los caracteres duplicados y la segunda letra de los caracteres adyacentes con el mismo número Soundex, convertir los caracteres en las posiciones 2 a 4 en un número, y, por último, en cualquier posición no usada agregar ceros. El código Soundex ha sido utilizado con los datos de los censos de los Estados Unidos.
- **Sustitución de los valores nulos o vacíos:** El problema fundamental en esta sustitución es que la aparición de los valores nulos en una base de datos puede venir dada por dos razones fundamentales: datos omitidos o valores inaplicables para dichos registros. En el primer caso la sustitución del nulo es posible, en el segundo no se pudiera realizar, porque falsearía la información (Martinez, 2003). Este es entonces el primer problema por resolver: ¿qué valores pueden ser sustituidos? A esta interrogante no se han encontrado respuestas satisfactorias en la literatura consultada. Otro problema para considerar es que en muchos sistemas de bases de datos no se utilizan el valor NULO para indicar información ausente, sino que

sencillamente se deja en blanco, algunos gestores sustituyen este blanco por CERO (si estamos en presencia de un campo numérico) o por CADENA VACÍA (en caso de datos alfabéticos. En el caso de valores numéricos se puede lograr que se mantengan las medidas de tendencia central: media, moda, mediana, varianza, desviación típica. En todos los casos se busca la medida de tendencia central que se trate en el conjunto de datos que se tiene sin considerar los valores nulos o vacíos y luego se sustituyen estos por valores tales que hagan que estas medidas de tendencia central no cambien.

- **Métodos de Agrupamiento:** Estos métodos implementan algoritmos de agrupamiento o clusterizacion usando alguna distancia, para identificar excepciones en los registros de los datos. (Beatriz E. – Ramiro Perez – 2009)
- **Metodología para la limpieza de datos:** Galhardas (2001) propuso una técnica la cual consiste en dividir los datos de manera lógica y física. En la parte lógica se encuentran las llaves y las técnicas de normalización y en la parte física, se separan los datos por medio de sentencias SQL realizando agrupaciones de datos con el fin de clasificarlos de manera efectiva para concretar una consistencia entre los estos.
- **Métodos basados en patrones:** un patrón es definido por un conjunto de registros que tienen características o comportamientos similares en un p% de campos en el conjunto de datos, donde p es el valor definido por el usuario (90 frecuentemente). (Beatriz E. – Ramiro Perez – 2009)
- **Parsing :** Se utiliza en la limpieza para la detección de errores sintácticos. Un “parser” para una gramática G es un programa que decide, dada una cadena si es o no un elemento del lenguaje definido por la gramática. En el contexto de los compiladores para lenguaje de programación las cadenas son los programas; en limpieza de datos las cadenas pueden representar tuplas de datos de una instancia de una relación o valores de atributos de un dominio dado. (Beatriz E. – Ramiro Perez – 2009)

- **Documentos.** Es importante tener los siguientes conceptos y procesos claros y así lograr una mejor comprensión de la información a presentar:
  - Producción de documentos, comprende los aspectos de origen, creación y diseño documento, conforme al desarrollo de actividades y a las funciones oficiales propias de cada dependencia, por lo cual estas deben tener el formato normalizado.
  - Recepción de documentos, se debe verificar que estén completos, que sean pertinentes y sean competencia de la entidad, para efectos de su radicación, registró y distribución con el fin de dar inicio a los trámites correspondientes.
  - Distribución de documentos, es en el proceso archivístico el cual se ordenan los documentos de cada unidad administrativa de acuerdo con las series, sub series y tipos de documentales para los respectivos expedientes.
  - Trámite de documentos, comienza con el flujo de comunicaciones recibidas y enviadas y la aplicación de los principios básicos archivísticos de procedencia y orden original, con miras a la conformación de las series documentales. El ciclo inicia en el área de correspondencia donde se asigna el número de radicación consecutiva a las comunicaciones oficiales. (Lavalle, 2018)
- **Carga.** En esta etapa o fase, los datos ya están preparados para ser cargados en su almacén. Esta acción puede darse directamente acumulando toda la información, cargando todos los datos obtenidos directamente al Data Warehouse, o bien la carga puede realizarse en distintos niveles de información o granularidad; esta técnica se denomina rolling. (Lopez, 2018)
- **Almacenamiento.** Un almacén de datos siempre está orientado hacia la información relevante de la organización. Se diseña para consultar eficientemente información relativa a las actividades (ventas, compras, producción...) básicas de la organización. (Perez, 2015)

- **Modelos de Estructura de base de datos.** Los modelos nos darán una representación gráfica de la manera como se estructura los datos, esto nos servirán como guía para conocer la forma del almacenamiento y posterior la manera como se darán las consultas de estos; existen varios modelos que estaremos explotando en este apartado:
  - **Modelo Relacional:** Representa la base de datos como una colección de relaciones. Informalmente, cada una de estas relaciones se parece a una tabla de valores o, de forma algo más extensa, a un fichero plano de registros. Por ejemplo, la base de datos de ficheros es similar a la representación del modelo relacional. (Ramez E. 2007) En este modelo los datos se estructuran lógicamente en forma de relaciones (tablas) y su objeto fundamental es mantener la independencia de la estructura lógica respecto al modo de almacenamiento y ante cualquier otra característica de tipo físico. . (Elizabeth R. – Oscar E. Jose N. – 2019). El modelo relacional representa la segunda generación de los SGBD. En él, todos los datos están estructurados a nivel lógico como tablas formadas por filas y columnas, aunque a nivel físico pueden tener una estructura completamente distinta. Un punto fuerte del modelo relacional es la sencillez de su estructura lógica. (Mercedes M. 2011)
  - **Modelo Orientado a objetos.** Una base de datos orientada a objetos está compuesta de objetos y clases de objetos unidas mediante diversos mecanismos de abstracción. Un objeto es un paquete de datos y procedimiento. Los atributos de un objeto contienen los datos. (Paul B. 2014). En este modelo, a diferencia de otros programas de este tipo los objetos deben ser persistentes; es decir que cuando la ejecución de las aplicaciones termina, los objetos de la base deben seguir existiendo. Otra diferencia radica en que mientras la BDR representa relaciones mediante llaves foráneas, en este modelo se incluyen relaciones en la definición de los objetos con los que se relaciona; es decir para que las

relaciones se lleven a cabo, se debe designar un atributo interno de cada objeto para esta función.(Elizabeth R. – Oscar E. Jose N. – 2019)

- **Data Warehouse.** Un almacén de datos (Data Warehouse) es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización, con independencia de cómo se vayan a emplear, posteriormente, por los diferentes usuarios. (Joyanes, 2019)
- **La Nube.** Ofrece un método ligero y ágil para acceder a las aplicaciones de BI, ya que una de sus grandes ventajas reside en que las aplicaciones de Inteligenciade Negocios en la nube tienen carácter ubicuo, y se puede acceder a ellas desde múltiples dispositivos (especialmente móviles), navegadores web, en cualquier lugar y en cualquier momento que se pueda necesitar. (Joyanes, 2019)
- **Seguridad de los Datos.** La mejor alternativa que ha de seguir una organización o empresa para migrar sus servicios a la nube es la elección de un proveedor de cloud computing fiable y de calidad, que ofrezca todos los servicios de Inteligencia de Negocios que quiera la empresa y que cumpla rigurosamente todas las normativas nacionales e internacionales de protección de datos y de privacidad. (Joyanes, 2019)

Referentes Conceptuales:

- **Modelo de éxito de un data warehouse.** Este trabajo de grado hecho en una universidad de Puerto Rico busca explicar los factores que se deben contemplar en la creación e implementación de un data warehouse, donde terminan concluyendo que la mayor parte del éxito recae en la calidad de los datos y la estructura organizacional de la empresa. (Vicente, 2012)
- **Creación de una data warehouse a una pyme teniendo en cuenta el concepto de inteligencia de negocios.** Este trabajo de grado desarrollado en la universidad EAFIT de Medellín lo que busca es proponer la creación de un data warehouse en pymes del sector hotelero en Colombia para poder aprovechar las ventajas de la inteligencia de negocios tales como la

toma de decisiones enfocada en la información y que pueda ser automatizada basada en el mercado y las necesidades de los posibles clientes. (Cardona & Arevalo, 2012)

- **Data warehouse: marco de calidad.** En este trabajo de grado de una universidad de Madrid, España, el estudiante planteo una estructura de 4 bloques que se deben tener en cuenta para que la creación de un data warehouse tenga cierto orden y coherencia, siempre basado en las buenas prácticas y objetivos de la inteligencia de negocios. (Gutierrez, 2012)

7. Marco institucional

7.1 Secretaría de Educación del Distrito (SED)

Fue creada mediante el Acuerdo 26 del 23 de mayo de 1955, del Concejo de Bogotá hace parte del sector central de la Administración Distrital, en cabeza de la Alcaldía Mayor, tiene por objeto orientar y liderar la formulación y ejecución de políticas, planes y programas para garantizar el derecho a la educación y asegurar a la población el acceso al conocimiento y la formación integral. Según el portal de función pública, hay 53.992 empleados en la secretaria de educación distrital con el contrato de servidor público al mes de mayo de 2022 (publica, 2022).

7.2 Referentes estratégicos

La entidad presenta como misión promover la oferta educativa en la ciudad para garantizar el acceso y la permanencia de los niños, niñas y jóvenes en el sistema educativo, en sus distintas formas, niveles y modalidades; la calidad y pertinencia de la educación, con el propósito de formar individuos capaces de vivir productiva, creativa y responsablemente en comunidad.

De igual manera ubica su visión en garantizar el derecho a la educación de los niños, niñas y jóvenes de la ciudad, a través de colegios distritales modernos, humanos e incluyentes y de un proceso de formación democrático, participativo, permanente, personal, cultural y social (educación, s.f.). Según el Ideario Ético del Distrito (2007), las entidades distritales deben cumplir sus funciones bajo la solidaridad, equidad, el respeto, la vocación de servicio, probidad, el trabajo en equipo y la responsabilidad.

Dispone su portafolio de servicios con el fin de Formular, orientar y coordinar las políticas y planes del Sector Educación, en concordancia con el Plan de Desarrollo Distrital, el Plan Sectorial de Educación, el Acuerdo 257 de 2006 y las demás normas legales del orden nacional.

Así mismo, se enfoca en Desarrollar estrategias que garanticen el acceso y permanencia de los niños, niñas y jóvenes en el sistema educativo, así como la

pertinencia, calidad y equidad de la educación en sus diferentes formas, niveles y modalidades.

7.3 Estructura organizacional

La Secretaría de Educación Distrital consta de la oficina del secretario, oficinas de control y asesorías, y cuatro subsecretarias, haremos énfasis en la subsecretaria de Acceso y Permanencia a la cual pertenece la Dirección de Cobertura que es es la dependencia para la cual se presenta el proyecto.

7.4 Organigrama de la SED:

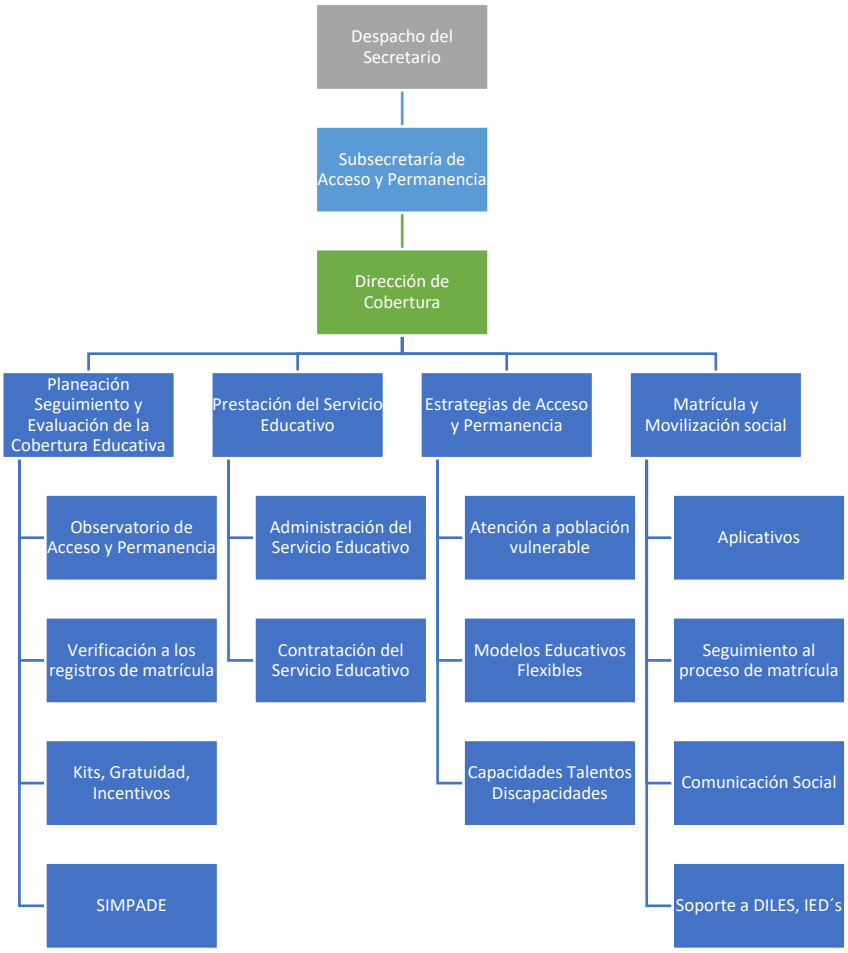

Figure 1 Organigrama de la SED Fuente Elaboración propia

- Elaboración propia con base en el organigrama de la Secretaría de Educación del Distrito 2021
- La SED consta de cinco oficinas, tres de asesorías y dos de control. Las oficinas de asesoría son de planeación, de comunicación y prensa, y jurídica, y su función principal es asesorar al secretario de educación en la formulación de los planes, programas y proyectos que requiera la Secretaría para el cumplimiento de sus funciones. Las oficinas de control tienen el propósito de medir y evaluar la eficiencia, eficacia y economía de los demás controles, asesorando a la alta dirección en la continuidad del proceso administrativo, la reevaluación de los planes establecidos y en la

introducción de correctivos necesarios para el cumplimiento de las metas u objetivos previstos.

- Las subsecretarías son cuatro, la de integración interinstitucional, la de calidad y pertinencia, la de acceso y permanencia y la de gestión institucional. Cada subsecretaria se divide en distintas direcciones, pero todas se encargan de recopilar los datos, propuestas y planes que se proponen desde las direcciones hacia el secretario de educación, quien hace llegar dichas propuestas a la alcaldía mayor de Bogotá y desde la alcaldía, las discuten en los debates ante el Concejo de Bogotá.

La subsecretaria de “Acceso y Permanencia” en la dirección de cobertura es la dependencia en donde se centra la ejecución del proyecto, partiendo de la necesidad del diseño de un modelo de almacenamiento y limpieza de datos en la dirección de cobertura de la Secretaría de Educación Distrital de Bogotá (SED).

7.5 Productos o servicios ofertados por el área de cobertura

El área de cobertura, realiza su aporte a la subsecretaria desde la identificación de la demanda de la población en el sistema educativo oficial del Distrito Capital y la oferta de sus establecimientos educativos, en los niveles de preescolar, básica y media, Promueve la gestión de la cobertura educativa, identificando las particularidades del territorio, desarrollando instrumentos y estrategias que contribuyen a fortalecer la Ruta de Acceso y Permanencia Escolar, así mismo establece los lineamientos, criterios y procedimientos para implementar el proceso anual de gestión de la cobertura educativa, realiza el seguimiento y evaluación de las condiciones de acceso y permanencia, de igual manera realiza seguimiento y gestión del uso de los Sistemas de Información relacionados con el proceso de gestión de la cobertura, así como definir y aplicar los criterios para la inscripción, evaluación y selección de los colegios privados registrados en el banco de oferentes.

En la dirección de cobertura se reciben requerimientos de información de toda índole, por mencionar algunos como: clases de colegios, tipo de discapacidad, tipos de etnias, capacidades excepcionales, población migrante, población víctima, nivel escolar, edades, genero, grado, localidad, institución educativa, sedes educativas, entre otras; algunas de estas solicitudes se resuelven con respuestas en donde la fuente de datos nos registra la información como se evidencia en la figura 2.

En la figura 2, muestra un ejemplo actual y real en caso de que se intentara responder a un requerimiento de la ciudadanía sin hacer ningún tipo de cambio o ajuste en las bases de datos de la SED. Se evidencia que se crea sin tener en cuenta parámetros básicos de estandarización de la información necesarios tales como:

- Escribir en la primera variable el nombre completo que corresponde a Localidad.
- Los nombres de las localidades deben tener la primera letra en mayúscula.
- Las demás letras deben ser en minúsculas y con tildes.
- Para las variables de la identificación de género debería verse el nombre completo “Femenino” y “Masculino” en vez de solo “F” y “M”.

Estos escenarios son algunos de los ejemplos donde la información no tiene un estándar y los cuales son recurrentes en el registro de la información contenida en las bases de datos.

LOCALIDAD	F	M	TOTAL GENERAL
ANTONIO NARIÑO	4.266	4.856	9.122
BARRIOS UNIDOS	5.681	5.717	11.398
BOSA	51.814	53.037	104.851
CHAPINERO	1.443	1.616	3.059
CIUDAD BOLIVAR	45.482	47.853	93.335
ENGATIVA	29.803	30.815	60.618
FONTIBON	11.178	11.940	23.118
KENNEDY	52.674	55.569	108.243
LA CANDELARIA	1.318	1.445	2.763
LOS MARTIRES	5.727	4.215	9.942
PUENTE ARANDA	11.241	9.965	21.206
RAFAEL URIBE	29.383	26.440	55.823
SAN CRISTOBAL	23.026	25.442	48.468
SANTAFE	4.340	4.513	8.853
SUBA	36.362	37.175	73.537
SUMAPAZ	404	442	846
TEUSAQUILLO	1.470	1.555	3.025
TUNJUELITO	15.940	17.355	33.295
USAQUEN	11.792	12.547	24.339
USME	33.659	35.516	69.175
TOTAL GENERAL	**377.003**	**388.013**	**765.016**

Figure 2 Ejemplo base de datos de la SED Fuente: SED

8. Diseño metodológico

8.1 Tipo de Investigación

En el presente estudio se propuso una solución a la situación identificada en la SED, basados en un proceso de indagación con el personal involucrado en los diferentes procesos. La investigación descriptiva ha sido escogida como modelo de estudio en esta investigación debido a que se centra en describir y conocer más a fondo la problemática, basados en las experiencias del usuario y así abordar posibles soluciones.

La investigación describe la propuesta relacionada con el diseño de un modelo de almacenamiento y limpieza de datos que permita mejorar los tiempos de respuesta a las solicitudes al interior de la dirección de cobertura de la SED.

La falta de elementos y capacitación en este campo hacen del manejo de la base de datos de la dirección de cobertura de la SED un problema bastante complejo por lo que los elementos de indagación planteados fueron basados en la problemática presentada, en la necesidad de sus operaciones y dando respuesta a la forma más conveniente y más eficiente de suplir sus requerimientos.

8.2 Análisis externo:

Para el análisis se utilizó el instrumento PESTEL teniendo en cuenta que en este proyecto en particular se necesitó de tener un contexto externo de la situación actual de la Secretaria de Educación Distrital que permitiera verificar que la solución propuesta tuviera validez y un verdadero impacto para los actores externos que interactúan con la SED.

Factores Externos	Locales	Nacionales	Internacionales
Políticos	Aprovechamiento de la política de datos públicos y seguridad de los datos de la alcaldía de Bogotá.	Aprovechamiento de la política de datos públicos y seguridad de los datos del gobierno nacional.	Los modelos de migración de otros países en materia de transformación digital.
Económicos	La migración de los datos y los costos asumidos	La dificultar de aprobación de recursos por ser una entidad gubernamental.	La gran variedad de desarrollo de almacenamiento en la nube, con cobros por consumo.
Sociales	La falta de cultura y conocimiento de almacenamiento y manejo de datos	La falta de cultura y conocimiento de almacenamiento y manejo de datos	El buen desarrollo internacional y ofertas en almacenamiento
Tecnológicos	La expansión de ofertas de oportunidades tecnológicas debido a la pandemia.	Las oportunidades dadas por el gobierno nacional en pro de la transformación digital.	Los grandes desarrollos de la industria europea y EE. UU..
Jurídicos	La falta de cultura y seguridad en el manejo de los datos	Los beneficios que trae la transformación y control de los datos	

Ambientales	Mejoras del medio ambiente en reducción de papelería.	Mejoras del medio ambiente en reducción de papelería.	Mejoras del medio ambiente en reducción de papelería.

Table 1 PESTEL de la SED Elaboración propia en base a la situación de la SED

8.3 Análisis Interno:

Para el análisis interno se utilizó la herramienta EFI, la cual permitió conocer las fortalezas que se tenían al interior de la dirección de cobertura para estar seguros de que se iban a tener las herramientas necesarias para cumplir con el objetivo y también conocer los puntos de mejora que se debían tener en cuenta en caso de que fueran a afectar la viabilidad de implementar la solución.

Factor Crítico de Éxito	Valor	Calificación	Calificación Ponderada
Fortalezas			
Buena Fuente de datos	0,15	4	0,6
Disponibilidad de la información	0,10	4	0,4
Beneficios de políticas de datos públicos	0,09	4	0,36
Debilidades			
Escasez de recursos	0,15	2	0,3
Poca viabilidad para almacenamiento interno.	0,09	4	0,36

No se encuentra en búsqueda de transformación digital.	0,05	2	0,1
Valor Ponderado			2,12

Table 2 EFI de la dirección de cobertura de la SED Elaboración propia en base a la dirección de cobertura de la SED

Al revisar y analizar el resultado de la EFI de 2,12, se puede reafirmar que la dirección de cobertura de la SED tiene muchos puntos de mejora en cuanto a la búsqueda de soluciones digitales que permitan un trabajo con la información mucho más fácil y organizado.

A la hora de darle los valores en la herramienta EFI, se hizo en conjunto con uno de los integrantes del grupo de trabajo, el cual trabaja directamente en la dirección de cobertura de la SED y tiene contacto con el procesamiento diario que se lleva a cabo con la información, por lo cual los valores se asignaron con conocimiento de primera mano y de la percepción real de uno de los empleados de la SED.

8.4 Población, muestra y ficha técnica

- **Población:** La población para la investigación son los empleados del equipo de planeación, seguimiento y evaluación de la dirección de cobertura de la SED; los cuales son los encargados de trabajar con la información por lo cual, conocen los principales puntos de mejora de los procesos que se llevan a cabo día a día. Por otro parte, se tiene a los usuarios que son quienes instauran los requerimientos por lo cual terminan recibiendo los datos de forma organizada.
- **Muestra:** Considerando la accesibilidad a las poblaciones, se tomó como muestra a los 2 empleados encargados del almacenamiento, control y visualización de los datos y 30 consumidores de la información. El tamaño de esta muestra se tomó con base en la disponibilidad de tiempo de los trabajadores de la dirección de cobertura, sin dejar de lado que era necesario tener una cantidad significativa de comentarios por parte de los

trabajadores para obtener unos resultados de calidad a la hora de proponer una mejora a los procesos que llevan a cabo en el día a día.

8.5 Identificación de las variables

Las variables identificadas para la investigación son las siguientes:

- **Volumen de los Registros presentados mensualmente - Estimado:** Esta variable se refiere a la cantidad de registros que recibe la SED por parte de los colegios con la información de cada alumnos que se encuentran inscritos en los colegios oficiales de Bogotá. Se evaluó la cantidad de registros mensuales que ingresa producto de las operaciones diarias.
- **Número de Solicitudes de información recibidas:** Esta variable se refiere a las solicitudes que hacen los ciudadanos a través de los diferentes canales de comunicación que tiene la SED, donde solicitan información de los colegios oficiales de Bogotá. Se evaluó el promedio de solicitudes externas recibidas en cuanto consolidados y datos agrupados.
- **Frecuencia de la información:** Esta variable busca definir la cantidad de solicitudes versus una unidad de tiempo para ver qué tan frecuente ocurren las solicitudes por parte de la ciudadanía. Se conoce la frecuencia con la que se reciben las solicitudes.
- **Beneficios del almacenamiento:** Con esta variable se buscó enfocarse en el beneficio de tiempo que se va a obtener implementando la solución. Se conocen los beneficios de la automatización de los procesos.
- **Capacitación de personal:** Con esta variable se mide el conocimiento en procesos de almacenamiento y procesamiento de grandes cantidades de data por parte de los empleados de la SED. Se conocen los grados de conocimiento del personal involucrado.

8.6 Instrumento de medición:

Esta es una investigación de tipo descriptivo, por lo cual se realiza una medición de las necesidades de la SED, basada en las necesidades directas que tienen los

empleados de la SED. La ventaja que ofrece el instrumento de medición es que tiene en cuenta la realidad del día a día de sus labores, lo que permite conocer a fondo las falencias, causas y puntos de mejora, y de esta forma proponer una solución al interior de la SED de una forma más acertada, atacando problemas reales.
Para la medición del análisis interno se determinó utilizar el modelo de encuestas; se muestra como anexo 1 el instrumento elaborado y ejecutado.

- **Validación del instrumento de medición:** Se utilizó la herramienta V de Aiken durante este proceso de validación, se consultaron varios expertos, entre los cuales están, Luis Forero Analista datos económicos, Alejandra Niño Coordinadora de proyectos, Juan Huérfano Profesional especializado, Camilo Akle Coordinador de proyectos, Gower Chacon Coordinador de Testing de Software. Teniendo en cuenta sus comentarios, se hicieron los ajustes necesarios en el instrumento de medición. La herramienta antes de ajustes puede ser encontrada en el anexo 1 y después de ajustes como anexo 2. Los ajustes fueron en la pregunta 2, 7, 10, 11, 15, 16 y 19 enfocados en tema de redacción y de la forma en que se formulaban las preguntas para que fueran más claras para las personas que respondieron el instrumento de medición. Como anexo 3 se encuentra el archivo donde se encuentra la herramienta V de Aiken.

9. Desarrollo

Con el fin de facilitar el almacenamiento y la limpieza de datos para permitir una visualización efectiva de la información, se utilizó la herramienta R para normalizar la data almacenada en las bases de datos de la Secretaría de Educación Distrital de Bogotá. Debido al alto flujo de información que se almacena en el Área de Cobertura de la SED y a los altos números de búsqueda de información, se hace necesaria la implementación de una herramienta que haga la normalización de los datos, para poder organizar y visualizar de una manera óptima la data. Todo este proceso se hizo con el fin de garantizar un diseño de base de datos con éxito, debido a que si no se realiza puede haber inexactitudes de los sistemas de bases de datos, ralentización de los procesos o ineficacia en las operaciones. La normalización permite también comprobar si las bases de datos existentes garantizan la integridad de los datos necesarios para cumplir con los estándares de calidad y con lo solicitado por los usuarios.

El proceso de normalización de datos se hace necesaria pues la información que estaba almacenada en la base de datos de la división de cobertura de la Secretaría de Educación Distrital estaba desorganizada, con caracteres desiguales y con caracteres que no pertenecían a las casillas a las que estaban asignadas, por ende, se hace necesaria la normalización de las variables, para así ajustar los valores medidos en diferentes escalas respecto a una escala común.

9.1 Librerías

En primer lugar, se corrieron las librerías de R para obtener todas las funciones que se necesitan, fueron los siguientes:

- library(readxl) : Es un paquete diseñado para hacer una sola tarea, la cual es importar hojas de Excel a R.
- library(tidyverse): Es un paquete diseñado para ciencia de datos, ayuda en todo el proceso de importar, transformar, visualizar, modelar y comunicar la información, comparten nombre y estructuras comunes.

- library(lubridate): El paquete ayuda con la estructura de las fechas y horas, además de realizar operaciones aritméticas para resolver problemas de tiempo.
- library(openxlsx): carga en memoria la librería que permite leer archivos Excel

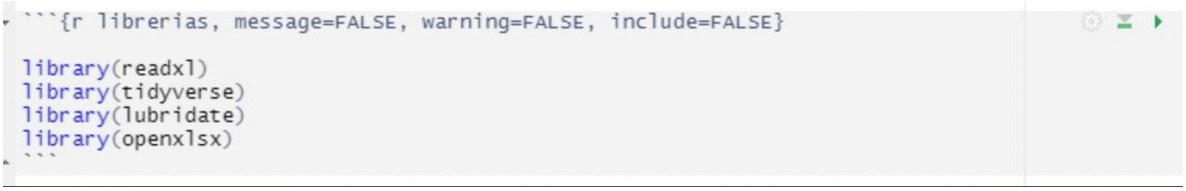

Figure 3 Librerías usadas en R Fuente: Elaboración propia

9.2 Ejecutar el código

Se tuvo que correr el código para escoger el archivo con la función: **file.choose()**, con ella se pedirá que se seleccione el archivo. Con la función **read_excel()**, se leerá un archivo de Excel con la extensión ".xlsx", tal y como se muestra en la figura 4.

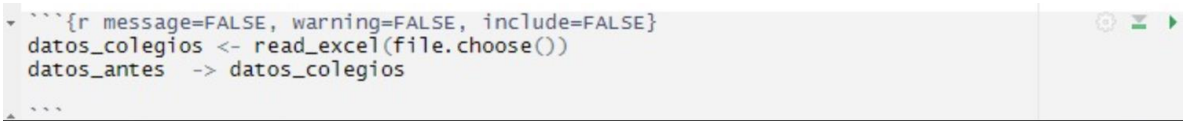

Figure 4 Forma en que se toma el archivo en Excel desde R. Elaboración propia

Sin importar el tamaño del archivo este proceso no toma más de 2 minutos, después, en la parte de leer datos carga un Excel con todos los cambios que se requieran (en este caso se debían hacer cambios en las categorías de zona, localidad, clase y nombre de la escuela).

Una vez el archivo esté atado a R, es necesario que R identifique las pestañas donde debe hacer el cambio. Para que sea más gráfico y entendible, en la figura 5 y 6 se evidencian las pestañas del Excel donde están las pestañas.

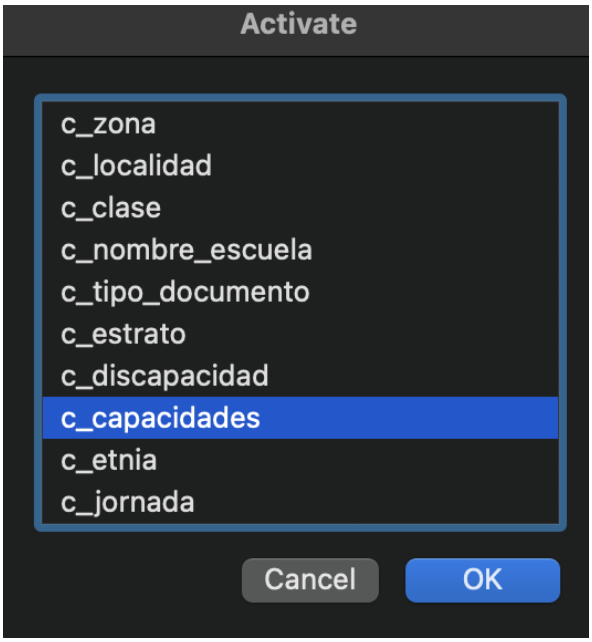

Figure 5 Primeras pestañas Excel en R. Elaboración propia

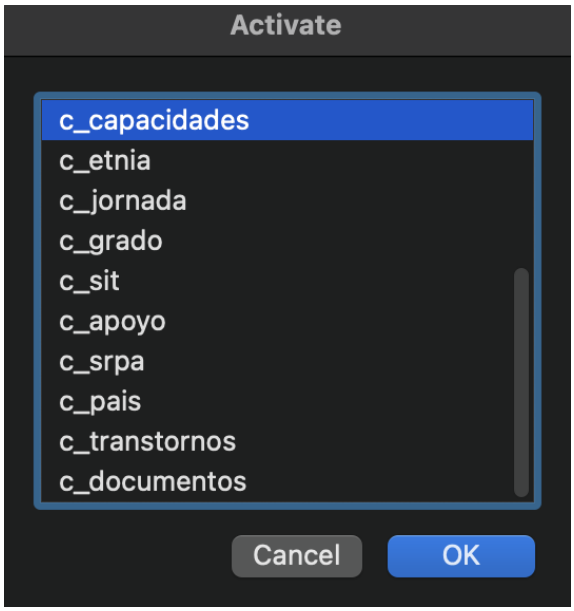

Figure 6 Segundas pestañas Excel en R. Elaboración propia

Y en las figuras 7,8 y 9 se muestra también el código en R, donde se pide que se haga la lectura de dichas pestañas.

```
```{r leer datos, message=FALSE, warning=FALSE, include=FALSE}

datos_zona           <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_zona")

datos_localidad      <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_localidad",
                                   col_types = c("numeric","text"))

datos_clase          <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_clase")

datos_nombre_escuela <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_nombre_escuela")

datos_tipo_documento <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_tipo_documento")

datos_estrato        <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_estrato")
```

Figure 7 Lectura en R de las pestañas en Excel. Elaboración propia

```
datos_discapacidad   <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_discapacidad")

datos_capacidades    <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_capacidades")

datos_etnia          <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_etnia")

datos_jornada        <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_jornada")

datos_grado          <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_grado")

datos_sit            <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_sit")

datos_apoyo          <- read_excel("datos_cambios.xlsx",
                                   sheet = "c_apoyo")
```

Figure 8 Lectura en R de las pestañas en Excel. Elaboración propia
```

```
datos_srpa          <- read_excel("datos_cambios.xlsx",
                                  sheet = "c_srpa")

datos_pais          <- read_excel("datos_cambios.xlsx",
                                  sheet = "c_pais")

datos_transtornos   <- read_excel("datos_cambios.xlsx",
                                  sheet = "c_transtornos")

datos_documentos    <- read_excel("datos_cambios.xlsx",
                                  sheet = "c_documentos")
```

Figure 9 Lectura en R de las pestañas en Excel. Elaboración propia

Cuando se realizó todo el procedimiento anterior y los cambios señalados, se cruza la tabla de **datos_colegios** y la de los cambios por medio del **codigo_dane** para el cambio de la zona. Para cruzar dos tablas se debe usar la función **left_join()** usando el código DANE como unión, luego con la función **mutate** se le asigna a la columna ZONASDP una condición **ifelse()**, que si después del cruce esta no encontró el código DANE escribe “urbana” en ZONASDP sino, que escriba “rural”, y luego el **select(-c(Zona)** eliminará el cruce para que quede con las mismas columnas. Como evidencia se adjuntan las figuras 10, 11, 12, 13, 14 y 15 del código en R del proceso descrito, junto con los demás ajustes necesarios:

```
```{r cambio de zona, message=FALSE, warning=FALSE, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_zona, by = "CODIGO_DANE") %>%
  mutate(ZONASDP = ifelse(is.na(Zona),"Urbana","Rural")) %>%
  select(-c(Zona))

```

```{r Cambios localidad, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_localidad, by = "NUMERO_LOCALIDAD") %>%
  mutate(NOMBRE_LOCALIDAD.x = NOMBRE_LOCALIDAD.y) %>%
  rename(NOMBRE_LOCALIDAD = NOMBRE_LOCALIDAD.x) %>% select(-NOMBRE_LOCALIDAD.y)

```

```{r Cambios clase, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_clase, by = "CLASE") %>%
  mutate(CLASE = CLASEC) %>%
  select(-CLASEC)

```
```

Figure 10 Cambios en Zona, Localidad, Clase. Elaboración propia

```
```{r Cambios escuelasCondicionC}

datos_colegios <- datos_colegios %>% left_join(datos_nombre_escuela, by = "CODIGO_DANE") %>%
  mutate(NOMBRE_ESTABLECIMIENTO_EDUCATIVO.x = NOMBRE_ESTABLECIMIENTO_EDUCATIVO.y) %>%
  select(-NOMBRE_ESTABLECIMIENTO_EDUCATIVO.y) %>%
  rename(NOMBRE_ESTABLECIMIENTO_EDUCATIVO = NOMBRE_ESTABLECIMIENTO_EDUCATIVO.x )


```

```{r cambio tipo documento, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_tipo_documento, by = "TIPO_DOCUMENTO")
%>%
  mutate(`@2_TIPO_DOCUMENTO.DESCRIPCIONCAMPO.x` = `@2_TIPO_DOCUMENTO.DESCRIPCIONCAMPO.y`) %>%
  select(-`@2_TIPO_DOCUMENTO.DESCRIPCIONCAMPO.y`) %>%
  rename(`@2_TIPO_DOCUMENTO.DESCRIPCIONCAMPO` = `@2_TIPO_DOCUMENTO.DESCRIPCIONCAMPO.x`)
```

```{r Estrato, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_estrato, by = "ESTRATO") %>%
  mutate(ESTRATO = ESTRATOC) %>%
  select(-ESTRATOC)
```
```

Figure 11 Cambios en Escuela, Tipo de documento, Estrato. Elaboración propia

```
```{r genero, include=FALSE}
datos_colegios <- datos_colegios %>% mutate(GENERO = case_when(GENERO == "F" ~ "Femenino",
                                                               GENERO == "M" ~ "Masculino"))

```

```{r Discapacidad, include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_discapacidad, by = "TIPO_DISCAPACIDAD")
%>%
  mutate(`@4_TIPO_DISCAPACIDAD.DESCRIPCIONCAMPO.x` =
`@4_TIPO_DISCAPACIDAD.DESCRIPCIONCAMPO.y`) %>%
  select(-`@4_TIPO_DISCAPACIDAD.DESCRIPCIONCAMPO.y`) %>%
  rename(`@4_TIPO_DISCAPACIDAD.DESCRIPCIONCAMPO`= `@4_TIPO_DISCAPACIDAD.DESCRIPCIONCAMPO.x`)

```

```{r Capacidad,include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_capacidades, by = "CAP_EXC") %>%
  mutate(`@5_CAP_EXC.DESCRIPCIONCAMPO.x` = `@5_CAP_EXC.DESCRIPCIONCAMPO.y`) %>%
  select(-`@5_CAP_EXC.DESCRIPCIONCAMPO.y`) %>%
  rename(`@5_CAP_EXC.DESCRIPCIONCAMPO`= `@5_CAP_EXC.DESCRIPCIONCAMPO.x`)
```
```

Figure 12 Cambios en género, discapacidad, capacidad. Elaboración propia

```r
```{r Etnia, include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_etnia, by = "ETNIA") %>%
  mutate(NOMBRE = NOMBREC) %>%
  select(-NOMBREC)
```

```{r Jornada, include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_jornada, by = "TIPO_JORNADA") %>%
  mutate(`@6_TIPO_JORNADA.DESCRIPCIONCAMPO.x` = `@6_TIPO_JORNADA.DESCRIPCIONCAMPO.y`) %>%
  select(-`@6_TIPO_JORNADA.DESCRIPCIONCAMPO.y`) %>%
  rename(`@6_TIPO_JORNADA.DESCRIPCIONCAMPO`= `@6_TIPO_JORNADA.DESCRIPCIONCAMPO.x`)
```

```{r Grado, include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_grado, by = "GRADO") %>%
  mutate(NOM_GRADO = NOM_GRADOC, NIVELESCOLARIDAD = NIVELESCOLARIDADC) %>%
  select(-c(NOM_GRADOC,NIVELESCOLARIDADC))

```
```

Figure 13 Cambios Etnia, Jornada, Grado. Elaboración propia

```r
```{r Sit, include=FALSE}
datos_colegios <- datos_colegios %>% left_join(datos_sit, by = "SIT_ACAD_ANO_ANT") %>%
  mutate(SIT_ACAD_ANO_ANT = SIT_ACAD_ANO_ANT_C) %>%
  select(-c(SIT_ACAD_ANO_ANT_C))
```

```{r apoyo, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_apoyo, by = "APOYO_ACADEMICO_ESPECIAL")
%>%
  mutate(DESCRIP_APOYO_ACADEMICO = DESCRIP_APOYO_ACADEMICOC) %>%
  select(-c(DESCRIP_APOYO_ACADEMICOC))

```

```{r SRPA, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_srpa, by = "SRPA") %>%
  mutate(DESCRIP_SRPA = DESCRIP_SRPAC) %>%
  select(-c(DESCRIP_SRPAC))

```
```

Figure 14 Cambios en Situación academica, apoyo, SRPA. Elaboración propia

```r
```{r Pais, include=FALSE}

datos_colegios <- datos_colegios %>% left_join(datos_pais, by = "PAIS_ORIGEN") %>%
  mutate(NOMBRE_PAIS = NOMBRE_PAISC) %>%
  select(-c(NOMBRE_PAISC))

```

```{r transtornos, include=FALSE}

datos_colegios <- datos_colegios %>%
  left_join(datos_transtornos, by = "TRASTORNOS_ESPECIFICOS DEL APRENDIZAJE") %>%
  mutate(DESCRIP_TRASTORNOS = DESCRIP_TRASTORNOSC) %>%
  select(-c(DESCRIP_TRASTORNOSC))

```
```

Figure 15 Cambios en país y transtornos. Elaboración propia

La mayoría de estos cambios se hacen de la misma forma exceptuando la edad y el rango, debido a que estos usan un tipo de condicional llamado **case_when()**, que realiza cambios dependiendo de la condición, como lo podemos ver en la figura 16, en donde el valor del Rango cambiará dependiendo de la Edad: cuando la edad sea 7, caerá dentro de la condición donde la edad sea mayor o igual a 6 o igual a 10, así que se asignará la etiqueta “6 a 10 años”.

```r
```{r edad, include=FALSE}

datos_nacimiento <- datos_colegios$FECHA_NACIMIENTO %>% as.Date(format = "%d/%m/%Y")
datos_colegios <- datos_colegios %>% mutate(Edad = as.period(interval(start =
datos_nacimiento, end = today()))$year)

datos_colegios <- datos_colegios %>% mutate(Rango = case_when(Edad <= 5 ~ "Hasta 5 años",
                                                              6 <= Edad & Edad <=10 ~ "6 a 10 años",
                                                              11 <= Edad & Edad <= 14 ~ "11 a 14 años",
                                                              15 <= Edad & Edad <= 16 ~ "15 a 16 años",
                                                              Edad >= 17 ~ "17 años y más"))

```
```

Figure 16 Cambio en Edad y Rango. Elaboración propia

Por último, con la función **write_xlsx()** se escribirá un archivo de Excel que contenga todos los cambios solicitados, y junto con la función **paste0**, se crea un nombre con la fecha del día, tal y como se evidencia en la figura 17.

```r
```{r Export to Excel}
fecha <- paste0("ETL_Base_cambios_",today(),".xlsx")
write.xlsx(datos_colegios,fecha,overwrite = FALSE)
```
```

Figure 17 Creación archivo final en el Excel desde R. Elaboración propia

9.3 Resultados después de ejecutar el código

Todo el código en R anteriormente mencionado se puede encontrar en el anexo 4 del presente trabajo de grado, así como también en el anexo 5 el archivo de Excel donde están los parámetros para los cambios de los datos, en el anexo 6 el archivo en Excel de los datos sin cambiar y en el anexo 7 el archivo en Excel que genero el código en R con los datos normalizado.

Basados en una de las respuestas de la Secretaria de Educación Distrital a un requerimiento de un ciudadano, podemos hacer la comparativa del resultado sin el código en R vs el resultado usando el código en R.

En la figura 18, se muestra la respuesta hecha por uno de los trabajadores de la dirección de cobertura con la información extraída de las bases de datos de forma manual y ajustadas parcialmente para que cumplan con los criterios de calidad.

LOCALIDAD	F	M	TOTAL GENERAL
ANTONIO NARIÑO	4.266	4.856	9.122
BARRIOS UNIDOS	5.681	5.717	11.398
BOSA	51.814	53.037	104.851
CHAPINERO	1.443	1.616	3.059
CIUDAD BOLÍVAR	45.482	47.853	93.335
ENGATIVÁ	29.803	30.815	60.618
FONTIBÓN	11.178	11.940	23.118
KENNEDY	52.674	55.569	108.243
LA CANDELARIA	1.318	1.445	2.763
LOS MÁRTIRES	5.727	4.215	9.942
PUENTE ARANDA	11.241	9.965	21.206
RAFAEL URIBE URIBE	29.383	26.440	55.823
SAN CRISTÓBAL	23.026	25.442	48.468
SANTA FE	4.340	4.513	8.853
SUBA	36.362	37.175	73.537
SUMAPAZ	404	442	846
TEUSAQUILLO	1.470	1.555	3.025
TUNJUELITO	15.940	17.355	33.295
USAQUÉN	11.792	12.547	24.339
USME	33.659	35.516	69.175
TOTAL GENERAL	**377.003**	**388.013**	**765.016**

Figure 18 Respuesta actual SED. Fuente SED

En la figura 19 podemos ver el mismo requerimiento de información, pero después de haber ejecutado el código en R, por lo cual la información se muestra mucho más limpia.

Localidad	Femenino	Masculino	Total general
Antonio Nariño	4.266	4.856	9.122
Barrios Unidos	5.681	5.717	11.398
Bosa	51.814	53.037	104.851
Chapinero	1.443	1.616	3.059
Ciudad Bolívar	45.482	47.853	93.335
Engativá	29.803	30.815	60.618
Fontibón	11.178	11.940	23.118
Kennedy	52.674	55.569	108.243
La Candelaria	1.318	1.445	2.763
Los Mártires	5.727	4.215	9.942
Puente Aranda	11.241	9.965	21.206
Rafael Uribe Uribe	29.383	26.440	55.823
San Cristóbal	23.026	25.442	48.468
Santa Fe	4.340	4.513	8.853
Suba	36.362	37.175	73.537
Sumapaz	404	442	846
Teusaquillo	1.470	1.555	3.025
Tunjuelito	15.940	17.355	33.295
Usaquén	11.792	12.547	24.339
Usme	33.659	35.516	69.175
Total general	**377.003**	**388.013**	**765.016**

Figure 19 Respuesta SED después de código en R. Elaboración propia

También podemos comparar las mejoras a niveles de tiempo de respuesta a las solicitudes de los ciudadanos, en la figura 20 se muestra el flujo de acciones y tiempos de dichas acciones que hoy en día ocurren cuando un ciudadano instaura una petición ante la SED.

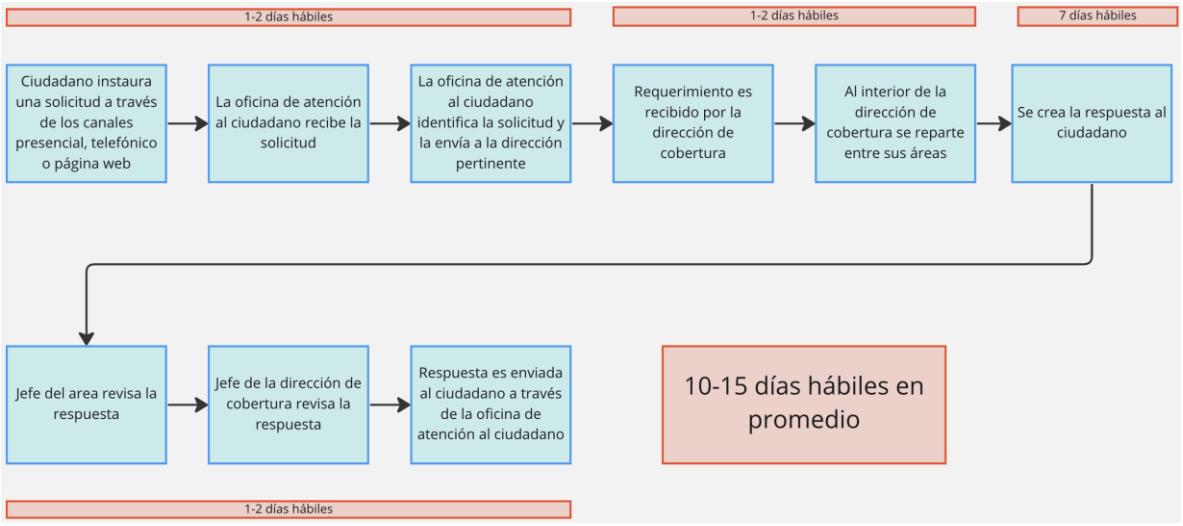

Figure 20 Flujo de una solicitud ante la SED. Elaboración propia

En caso de que se implemente la solución propuesta, los tiempos de respuesta podrían disminuirse de 10 a 15 días máximo a tan solo 6 a 9 días máximo, tal y como se muestra en la figura 21.

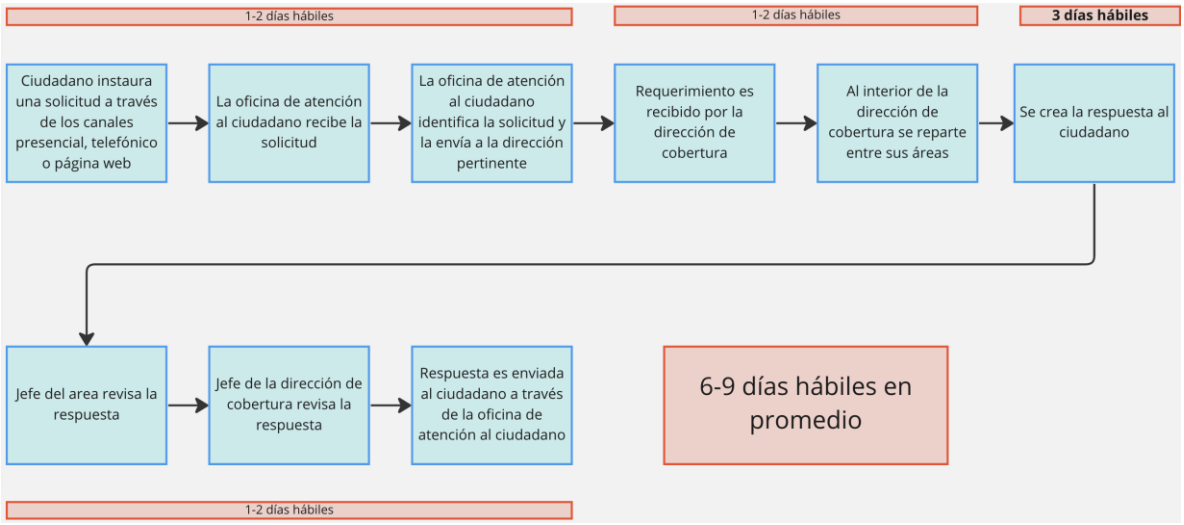

Figure 21 Flujo en caso de usar la solución propuesta en la SED. Elaboración propia

Otro beneficio que nos traería la implementación de la solución podríamos agregar la seguridad de los datos ya que la información no estaría disponible en los exceles sino que las consultas se realizarían desde la fuente de almacenamiento quien deja registro de consulta, otro punto bastante beneficioso es la centralización de información y la disponibilidad de ella; antes de la implementación para dar respuesta tocaba consolidar una y otra vez la información con esta implementación tenemos la información siempre almacenada y disponible.

9.4 Selección y propuesta del Data Warehouse

Eventualmente, la información almacenada al interior de la Secretaría Distrital de Educación debe ser analizada para obtener una visión crítica y poder tomar decisiones basados en los datos. Para el almacenamiento de los datos se selecciona un almacenamiento en nube por su pago por consumo las mejoras en

su almacenamiento de tipo columnar lo que permite un mejor procesamiento, evitar costos de mantenimiento y reducción de tiempos en procesos como modelado de datos, este tipo es el que mejor se ajusta a las necesidades de la Secretaría de Educación, específicamente en la dirección de cobertura, donde su uso principal es el de reportería. La estructura escogida para el Data Warehouse es:

- Nivel inferior: Se propone un servidor de base de datos donde se van a almacenar las bases de datos en Excel de los diferentes colegios de Bogotá junto con sus estudiantes para que después pueda ser extraída y transformada para los reportes.
- Nivel medio: En este nivel va a estar el código en R para hacer la transformación y limpieza de la data para que todo quede normalizado para su posterior entrega a los solicitantes de las bases. El almacenamiento columnar permitirá mejor procesamiento para las consultas realizadas por los reportes finales que les llegan a los ciudadanos.
- Nivel superior: En este nivel ya es donde se obtiene el reporte solicitado por los ciudadanos a forma de Excel.

El objetivo de este Data Warehouse es que sea sencillo, nada complejo para que pueda llegar a ser aceptado e implementado por la SED, la cual no cuenta con un presupuesto amplio para desarrollar nuevas iniciativas. Teniendo en cuenta lo anterior, se propone un Data Warehouse sencillo, que cumpla con las necesidades básicas que permitan suplir las necesidades que tiene el departamento de cobertura con el almacenamiento, transformación y reporte de las bases solicitadas por los ciudadanos con la información de los estudiantes de los colegios en la ciudad de Bogotá.

Entre las opciones consideradas para el almacenamiento está Google con Bigquery y Amazon con Redshift de las más populares en el mercado escogiendo Google Bigquery porque además de ser columnar nos permite realizar anidamiento de datos lo que nos ayuda en reducción de costos y además Google

tienen un servicio llamado Cloud Run por medio de cual utilizando contenedores Docker nos permitiría poner en producción y de manera automática el proceso de limpieza de datos.

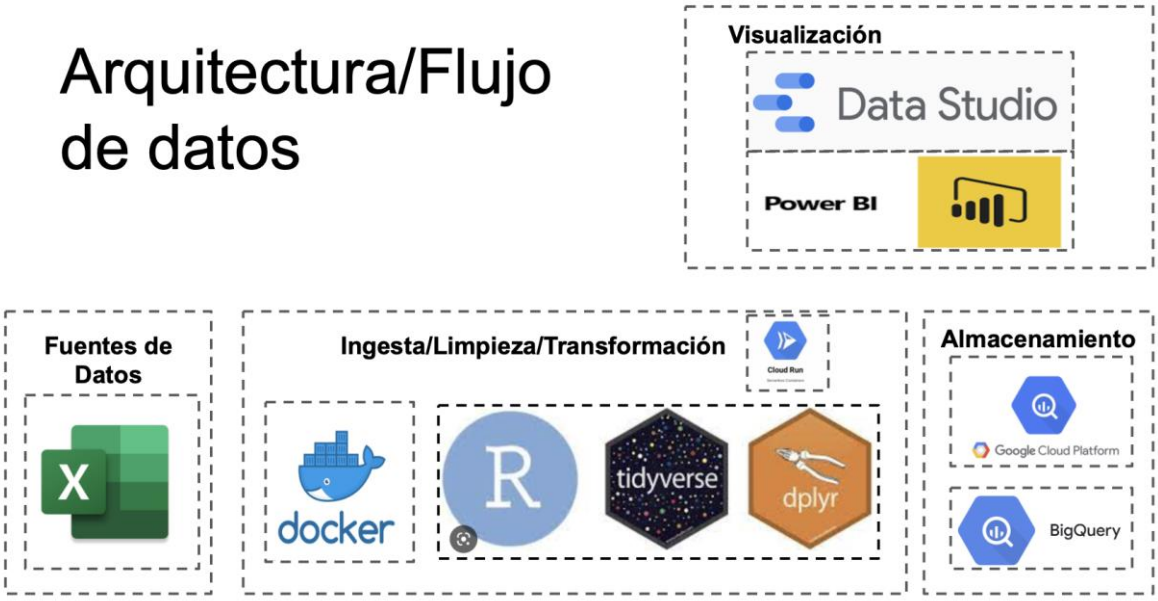

Figure 22 Arquitectura de los datos. Elaboración propia

10. Conclusiones

- La implementación de un Data Warehouse es una necesidad para la optimización de los procesos que se llevan a cabo al interior de la Secretaria Distrital de Educación, pero con un repositorio de información donde la data este parametrizada y sea fácil de extraer, suple las necesidades de la dirección de cobertura.
- Implementar únicamente el código en R para la normalización de la información representa el cambio más grande a la hora de disminuir tiempos y cantidad de procesos que debe hacer el equipo de la dirección de cobertura para resolver una solicitud de la ciudadanía.
- La literatura existente sobre normalización de data y almacenamiento es suficiente para poder implementar el proyecto de forma exitosa, teniendo en cuenta que se contempló un almacenamiento simple y el código en R está listo para su uso.
- Gracias al conocimiento interno de la SED, específicamente en la dirección de cobertura, se pudo conocer un punto crítico de la operación, donde se pudo proponer una solución viable enfocada en el área de business intelligence.
- Al ser una propuesta, queda a consideración de la Secretaria de Educación Distrital si se implementa la solución completa, una parte de ella o ninguna.

11. Recomendaciones

- Para futuros trabajos en la Secretaria Distrital de Educación de Bogotá se deberían contemplar el trabajo en conjunto de más áreas para que más recursos puedan a llegar a ser asignados al desarrollo e implementación.
- Este mismo trabajo se debería poder analizar en unos años, cuando la SED y en general las demás entidades gubernamentales estén más dispuestas a adoptar las nuevas tecnologías.
- Con una mayor cantidad de información, se podría llegar a considerar un data warehouse con muchas más capacidades de procesamiento y con funciones más especializadas que permitan hacer reportes y análisis complejos.
- Se puede llegar a derivar un trabajo donde se analice la información almacenada para poder obtener tendencias de los estudiantes de Bogotá y de esta forma proponer programas que mejores el sistema de educación en la ciudad.

12. Referencias

Distrito, S. d. (22 de Agosto de 2021). *Secretaría de Educación del Distrito*. Obtenido de FUNCIONES Y DEBERES: https://www.educacionbogota.edu.co/portal_institucional/nuestra-entidad/funciones-y-deberes

Distrito, S. d. (22 de Agosto de 2021). *Secretaría de Educación del Distrito*. Obtenido de MISIÓN - VISIÓN : https://www.educacionbogota.edu.co/portal_institucional/nuestra-entidad/mision---vision

Distrito, S. d. (22 de Agosto de 2021). *Secretaría de Educación del Distrito*. Obtenido de MANUAL DE FUNCIONES : https://www.educacionbogota.edu.co/portal_institucional/nuestra-entidad/manual-de-funciones

Myers, P. (11 de Noviembre de 2020). *Microsoft*. Obtenido de Arquitectura de la solución de BI en el centro de excelencia: https://docs.microsoft.com/es-es/power-bi/guidance/center-of-excellence-business-intelligence-solution-architecture

educación, S. d. (s.f.). *Secretaria de educación*. Obtenido de Secretaria de educación: https://www.educacionbogota.edu.co/portal_institucional/nuestra-entidad/

Ardila, J. M. (12 de Septiembre de 2021). *Bogotá le apuesta a formación de talentos en ciencia, tecnología e innovación* . Obtenido de Bogotá: https://bogota.gov.co/gobierno-abierto-de-bogota/bogota-le-apuesta-la-innovacion-y-tecnologia

Marqués, M. (2009). *Bases de Datos.* Obtenido de https://elibro-net.bdbiblioteca.universidadean.edu.co/es/ereader/bibliotecaean/51645?page=10

Joyanes, L. (2019). *Inteligencia de negocios y analítica de datos.* Obtenido de Alfaomega: https://www.alphaeditorialcloud.com/reader/inteligencia-de-negocios-y-analitica-de-datos-1?location=90

Curto Diaz, J. (2016). *Introducción al business intelligence.* Obtenido de https://elibro-net.bdbiblioteca.universidadean.edu.co/es/ereader/bibliotecaean/101030?page=127

Lopez, Y. (2018). *Business Intelligence.* Obtenido de https://elibro-net.bdbiblioteca.universidadean.edu.co/es/ereader/bibliotecaean/124393?page=105

Galhardas, H. (2001). *Declarative Data Cleaning: Languaje, model and algorithms.*

Martinez, E. (2003). *Sistema para la limpieza de datos.* Guadalajara, México.

Lavalle, R. &. (2018). *Mejoramiento al sistema de gestión documental de la secretaria de educación distrital, Santa Marte.* Colombia.

Perez, M. (2015). *Business Intelligence: Tecnicas, herramientas y aplicaciones.* Obtenido de Alfaomega.

Pulido, E. (. (2019). *Base de Datos.* Obtenido de https://elibro-net.bdbiblioteca.universidadean.edu.co/es/ereader/bibliotecaean/121283?page=29

Ministerio de Educación Nacional. (22 de Agosto de 2021). *Ministerio de Educación Nacional*. Obtenido de SIMAT: https://www.mineducacion.gov.co/1759/w3-article-168883.html?_noredirect=1

publica, F. (23 de Mayo de 2022). *Función pública*. Obtenido de Función pública: https://www.funcionpublica.gov.co/dafpIndexerBHV/?find=FindNext&query=+&dptoSeleccionado=&entidadSeleccionado=6121&munSeleccionado=&tipoAltaSeleccionado=&bloquearFiltroDptoSeleccionado=&bloquearFiltroEntidadSeleccionado=&bloquearFiltroMunSeleccionado=&blo

Vicente, I. C. (16 de 10 de 2012). *SciELO*. Obtenido de Modelo de éxito de un data warehouse: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-921X2013000100011

Cardona, C., & Arevalo, V. (2012). *Repository EAFIT.* Obtenido de CREACION DE UNA DATAWAREHOUSE A UNA PYME TENIENDO EN CUENTA EL CONCEPTO DE INTELIGENCIA DE NEGOCIOS: https://repository.eafit.edu.co/bitstream/handle/10784/2756/Carlos_Cardona_Viviana_Arevalo_2012.pdf?sequence=12&isAllowed=y

Gutierrez, P. M. (2012). *Universidad Carlos III de Madrid.* Obtenido de DATA WAREHOUSE: MARCO DE CALIDAD.: https://core.ac.uk/download/pdf/30046568.pdf

13. Anexos

- Anexo 1. Encuesta para la medición interna sin ajustes https://forms.gle/KZqHctD4mekvB3xm9
- Anexo 2. Encuesta para la medición interna después de ajustes https://forms.gle/sAKSE8d8w9ccfX7i7
- Anexo 3. Herramienta V de Aiken 5 Formato_validacion_V_de_Aiken consolidada.xlsx
- Anexo 4. Código en R para la estandarización Warehouse dirección de cobertura (SED).rmd
- Anexo 5. Archivo en Excel donde están los parámetros para los cambios de los datos datos_cambios.xlsx
- Anexo 6 el archivo en Excel de los datos sin cambiar Base Cambios.xlsx
- Anexo 7 el archivo en Excel que genero el código en R con los datos normalizado. ETL_Base_cambios.xlsx