Introducción

En la actualidad, el número de datos producidos y llevados a los sistemas de información es cada vez mayor, ofreciendo una oportunidad de uso como punto de partida para una variedad de estudios en diferentes áreas de trabajo investigativo [1]. Las técnicas de análisis de datos acompañan a los investigadores en la tarea de entender e interpretar sucesos del mundo real, labor que es difícil, pues los datos reflejan la complejidad de los fenómenos físicos y humanos que se están estudiando [2]. En este sentido, las técnicas de análisis comprenden una serie de pasos que implican la aplicación de un algoritmo a través de alguna herramienta computacional e involucra el procesamiento previo o preparación de los datos [3].

La preparación de los datos incluye tareas de recolección, compresión, limpieza, transformación, integración y carga. En la recolección y compresión de datos se pueden encontrar problemas como: fuentes de datos que difieren en su estructura, datos faltantes, datos atípicos, datos erróneos que generan conflictos y violación de validaciones, y actualización tediosa del almacenamiento [4]. Estos problemas requieren ser tratados. Ya que sin una preparación y preprocesamiento exhaustiva, los resultados de los procesos de análisis y de minería estarían limitados, puesto que esta etapa permite conocer la "forma" de los datos, para poder hacer juicios sobre su confiabilidad y calidad [5].

En la limpieza y transformación, se deben detectar inconsistencias, adecuar los datos y tomar decisiones respecto a las medidas y dimensiones necesarias. Los resultados obtenidos deben enfocarse en la calidad de los datos para garantizar que la información esté en condiciones adecuadas, ajustada a la realidad y al problema estudiado, especialmente cuando se presenta un aumento significativo en la cantidad de datos. Posterior a la preparación de los datos, se considera el diseño del esquema de almacenamiento, donde se integrarán y se cargarán. Finalizado lo anterior, se debe garantizar un conjunto de datos (dataset) confiable y listo para el procesamiento [5].

En la actualidad se encuentran herramientas que facilitan el proceso. Sin embargo, no siempre se adaptan a las relaciones y características significativas de los datos de múltiples fuentes heterogéneas [6], por lo que tampoco llegan a garantizar una integración de datos consistente [7]. Con ello, las investigaciones relacionadas con esta etapa de los procesos de análisis de datos se enfocan principalmente en técnicas para la optimización de operaciones costosas y la identificación de desafíos comunes [8]; pero dan pocas luces sobre cómo construir modelos que adopten aspectos del enfoque de dominio específico y aprovechen las particularidades de los mismos.

Los modelos de dominio específico, o en particular los enfoques de dominio específico son definidos como formas o estrategias de abordar una problemática en las cuales interviene el conocimiento propio de un campo o dominio de datos y que a través de dicho conocimiento se puede llegar a tener un mayor entendimiento del problema y por lo tanto soluciones que se adaptan mejor y que pueden llevar a resultados más ajustados [9]. Se encuentran ejemplos de aplicación en áreas como IoT [10], Lenguajes de Programación [11] [12], Ontologías [13], entre otros.

El tratamiento de datos producidos en entornos educativos, semejante a otros campos de estudio, requiere reunir los requisitos, definir las fuentes de datos y elegir la herramienta y técnicas adecuadas para un posterior análisis. La aplicación de diferentes técnicas y la realización de experimentos depende del conocimiento y dominio de los datos educativos a analizar, es importante hacer un diagnóstico inicial de estos y de la problemática abordada, teniendo en cuenta la experiencia de los interesados y expertos [14]. La organización unificada de los datos es una alternativa interesante para proporcionar información que permita análisis específicos por parte de los directivos docentes.

Así mismo, contar con los datos bajo un mismo esquema da la capacidad de apreciar de forma precisa los datos, que a menudo se distribuyen en varios software dentro de las instituciones educativas, facilitando la búsqueda de información específica, factor decisivo para la toma de decisiones en los momentos apropiados y diferencia importante para los administradores del área educativa [15]. Por ello es necesario explorar una forma precisa y efectiva de recopilación y preprocesamiento de datos educativos que permita integrar las múltiples fuentes, escalas y granularidades; teniendo

en cuenta también, que las necesidades de análisis de datos académicos son mayores y se exigen con el fin de mejorar el aprendizaje de los estudiantes y la eficacia institucional [16].

En este sentido, para la realización de este trabajo se plantearon las siguientes preguntas de investigación que ayudaron a alinear los conceptos, el proceso de diseño y desarrollo, permitiendo establecer la situación actual del entorno que recoge el trabajo:

P1 - ¿Es viable la inclusión del enfoque de dominio específico en el proceso de preparación de datos educativos, garantizando la calidad y consistencia de estos?
P2 - ¿Cuáles son las dificultades que se pueden encontrar en el diseño e implementación de un proceso de preparación de datos con enfoque de dominio específico, previo a la aplicación de EDM?
P3 - ¿Es factible aplicar el proceso de preparación de datos con enfoque de dominio específico en un entorno de educación básica y media?

Para dar respuesta a cada una de estas preguntas, este trabajo presenta el diseño y construcción de una estrategia para la inclusión del enfoque de dominio específico en el proceso de preparación de datos educativos previo a la aplicación de minería de datos. Es aplicado a un caso de estudio con datos provenientes de sistemas de información de instituciones de educación básica y media del departamento Norte de Santander en Colombia. Los datos son extraídos, tratados y llevados a un modelo de almacenamiento que permite el posterior procesamiento mediante algoritmos de minería de datos educativos. En el proceso se permite identificar las principales dificultades al tratar datos de este nivel educativo y generar conocimiento sobre el proceso para posteriores estudios.

El artículo está organizado de la siguiente manera: en la sección 1 presentó la introducción al tema de investigación, en la sección 2 se encuentran los materiales y métodos; la sección 3 incluye los resultados y análisis, donde se hace una discusión en torno a las preguntas que orientaron el trabajo y, por último, se traen a consideración las conclusiones y trabajo futuro.

Materiales y Métodos

Para el desarrollo de la estrategia se consideraron las recomendaciones dadas en [17]. Estas pautas han ayudado a orientar las investigaciones en este campo, dado que los investigadores crean y evalúan artefactos o modelos diseñados desde las TIC para resolver los problemas identificados, generalmente en un ámbito organizacional o institucional, y las preguntas guía.

Con el objetivo de fortalecer la revisión y comprensión de las fuentes de datos y el análisis de los resultados dentro del dominio educativo, se definió para el trabajo un alcance de tipo exploratorio y descriptivo; con este alcance se pretende hacer uso de un diseño secuencial, en el cual se tiene una primera etapa donde se exploran los datos y se realiza una comprensión cualitativa del fenómeno o contexto educativo. La implementación y validación de la estrategia se hizo siguiendo las fases comunes de un proceso de preparación de datos y en particular, se incluyó, el enfoque de dominio específico para capturar los elementos del dominio a considerar. Los pasos seguidos comprenden:

(1) Diseño y construcción de la estrategia para preparación de los datos educativos. En esta etapa se realizó la concepción de la estrategia para la preparación orientada a los datos del dominio educativo, para ello, se siguen cuatro fases: *entendimiento de los datos, extracción y filtrado, transformación y carga.* Estas fases serán detalladas posteriormente. **(2)** Reconocimiento y selección de datos. En esta etapa se presenta la descripción de las fuentes de datos, las cuales provienen de un caso de estudio de educación básica (primaria y secundaria) y media del departamento Norte de Santander en Colombia, la aplicación a datos de este nivel educativo constituye uno de los principales retos y aportes de este trabajo. **(3)** Aplicación de la estrategia y revisión de resultados. En esta última etapa, se presentan los datos procesados y cargados, las métricas aplicadas, así como los resultados y discusión en general.

Diseño y Construcción de la Estrategia para Preparación de los Datos Educativos Usando Enfoque de Dominio Específico

Los datos del contexto educativo pueden estar archivados desde acciones previas (p.ej. los datos de calificaciones en la trayectoria de un grupo de estudiantes), o ser datos generados en una plataforma (p.ej. las interacciones en una plataforma de educación virtual). Una estrategia para el proceso de preparación de datos debe permitir abordar los datos de gran variedad de fuentes y formatos, garantizando que sean útiles para el procesamiento con las técnicas de análisis y/o algoritmos de minería de datos. Este trabajo se concentró en diseñar una estrategia que permita llevar datos educativos a un formato operable, haciéndolos accesibles para las siguientes etapas de un proceso de minería y garantizando su calidad y confiabilidad.

El proceso de preparación, como se aprecia en la Figura 1 está compuesto por cuatro fases representadas en las barras verticales, existe una barra horizontal que es paralela e interviene en todo el proceso, se trata del conocimiento propio del dominio de datos educativos, el cual es alimentado por diferentes abstracciones tomadas del ecosistema educativo, como es el conocimiento de los expertos, los estándares educativos, la taxonomía de datos educativos, la legislación educativa, los currículos, el contexto académico, los perfiles de estudiante, las políticas institucionales, entre otros. A continuación, se hace una descripción de cada una de las fases.

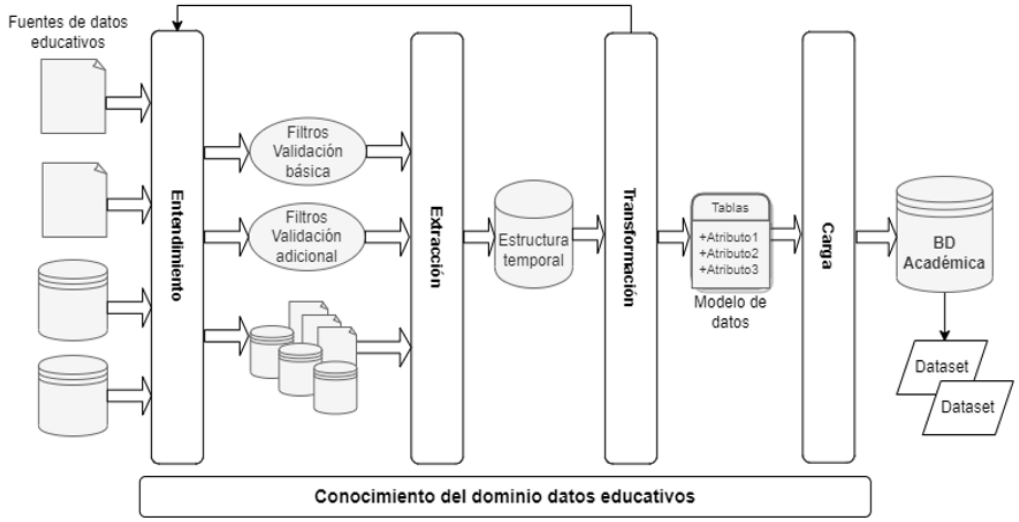

Figura 1. Estrategia de preparación de los datos educativos con enfoque de dominio específico.

Fase de entendimiento de los datos. Las entradas de esta fase corresponden a las fuentes de datos educativos que pueden tener diferente estructura y formato, pueden ser: archivos planos, bases de datos, logs, entre otros. Una vez se reciben los datos e identifican las estructuras, formatos y tipo, se empieza a hacer uso del conocimiento del dominio para entender las relaciones existentes y los problemas a atender. En este punto, uno de los aspectos fundamentales es contar con la ayuda de los expertos o propietarios de los datos, porque son ellos quienes tienen a priori el conocimiento de los sistemas que están generando la información. Las salidas de esta fase son los filtros que se usarán en la siguiente tarea: los filtros o rangos básicos y adicionales. Además, las fuentes de datos, que no sufren ninguna transformación o procesamiento, pero de las cuales se hace una copia para conservar el original para posteriormente realizar comparaciones y tener una trazabilidad de las alteraciones realizadas.

Fase de extracción y filtrado. La fase de extracción y filtrado comprende dos actividades principales, la validación básica y adicional. En la primera se reciben los datos originales en su estructura, se examinan y detectan posibles problemas por medio de los filtros de validación básica, estos son extraídos de acuerdo con la documentación de los datos, por ejemplo, la escala de calificaciones, el número de dígitos de un documento de identidad, entre otros. En la segunda se aplican filtros de validación adicional o compuesta, estos pueden revisar la relación entre dos o más atributos y requieren conocimiento del dominio que no necesariamente está a priori en la documentación del sistema o fuente de datos. La salida son los datos extraídos después del filtrado y llevados a una estructura temporal.

Fase de transformación. La entrada de esta fase es una estructura temporal con los datos filtrados, esta estructura existe para permitir validar y separar los datos que pasaron los filtros y los que no cumplieron las condiciones. En este punto, se verifica con el conocimiento del dominio, para determinar si los datos que no pasaron el filtro requieren una transformación para llegar al almacenamiento, o si por el contrario se detecta la necesidad de modificar el filtro o si finalmente se determina que el dato está errado y debe ser retirado. Esta fase es de refinamiento, porque puede existir tanto transformación de los datos como transformación de los filtros, es por ello, que en el diseño se establece que desde esta fase se puede regresar a la fase de entendimiento de los datos, incluso los filtros pueden requerir modificación. La salida de esta fase es el modelo de datos y los datos listos para ser cargados en el motor de base de datos.

Fase de carga. Esta es la última fase del proceso, consiste en tomar los datos que ya han sido filtrados y están limpios en la estructura temporal y llevarlos al modelo de datos para ser cargados en una base o bodega de datos académica. El conocimiento del dominio puede también influir en esta fase de carga, por ejemplo, en la definición del orden de los datos para llevarlos al almacenamiento, dependiendo de las relaciones establecidas en el modelo de datos. Finalmente, la salida de esta fase será la base o bodega de datos poblada.

Reconocimiento y Selección de Datos

Para mostrar una aplicación práctica de la estrategia de preparación de datos diseñada, se tomaron datos del sistema educativo colombiano, particularmente de educación básica (primaria y secundaria) y media del departamento Norte de Santander.

Descripción del origen de los datos. Colombia cuenta con un sistema educativo que incluye cinco niveles de educación: inicial, preescolar, básica, media y superior [18]. La educación inicial y preescolar tienen como propósito que los niños aprendan a convivir, a integrarse y a jugar con otros. El currículo de formación formal empieza en la educación básica, la cual se divide en dos ciclos, la primaria que comprende de los grados 1 a 5 y la secundaria que comprende de los grados 6 a 9. La educación media consta de dos grados, 10 y 11, finalmente, la educación superior se divide en técnica, tecnológica y universitaria con una duración variable. El sistema educativo colombiano está liderado por el Ministerio de Educación Nacional (MEN), organismo que define las políticas y propósitos educativos del País y que administra una parte de los datos generados en los establecimientos educativos. En los Departamentos y Municipios se encuentran las secretarías de educación departamentales y municipales, quienes tienen a cargo la gestión directa de los sistemas de información y conjuntos de datos generados en los entornos académicos de los niveles educativos previos a la formación superior.

Selección de los datos. Para el caso de estudio en que se aplicó la estrategia, se contó con datos recolectados desde dos fuentes y relacionados con instituciones públicas de educación básica y media del departamento Norte de Santander. Por un lado, los datos relacionados con las matrículas e información socioeconómica de los estudiantes para los años escolares de 2014 a 2019, los cuales son administrados por la Secretaría de Educación Departamental (SED) de Norte de Santander (desde ahora BD1). La segunda fuente corresponde a los datos de las calificaciones finales para las diferentes asignaturas cursadas por los estudiantes en cada uno de los grados (de transición a 11°) de tres Instituciones Educativas (IE) del mismo departamento y para los años de 2013 a 2018 (desde ahora BD2).

Recolección de los datos. Los datos de la BD1 se recolectan a través del sistema integrado de matrícula SIMAT, una herramienta que permite organizar y controlar el proceso de matrícula en todas sus etapas. La BD1 contiene datos que pueden ser descritos en cinco categorías de información: identificación de la institución educativa, identificación del estudiante, ubicación geográfica, datos socioeconómicos y datos de situación académica. Los datos de la BD2 son recolectados en cada una de las instituciones educativas, estas contratan de forma independiente el proveedor para sus sistemas de información académica, por lo que los esquemas de almacenamiento pueden diferir al igual que la forma como se presentan los reportes. La BD2 contiene los datos: institución y sede educativa, jornada, grado, curso, código y nombre del estudiante, calificación final para cada asignatura y estado; este último atributo corresponde a una etiqueta que deja clasificar a los estudiantes entre promovidos y reprobados.

Adicionalmente, cabe resaltar que, para el caso de la BD1, el formato en el que se recibieron los datos fue una planilla con macros de Excel. Mientras que, para BD2, se recibieron en formato PDF, se realizó un proceso de transformación a Excel y luego a archivos planos (CSV). Esta última tarea fue bastante tediosa y requirió de una revisión exhaustiva, dado que la conversión de formato generó algunos inconvenientes principalmente por temas de membretes y campos combinados.

Aplicación de la Estrategia y Revisión de Resultados

Como se mencionó en la subsección 2.1, la estrategia de preparación comienza con el entendimiento de los datos, aspecto que se cubrió en paralelo con la selección y recolección de los datos. A continuación, se presenta la aplicación de las fases de filtrado y extracción, transformación y carga. En cada una se va comentando, para el caso de estudio particular, la forma cómo se incluyó el enfoque de dominio específico.

Descripción de los filtros. Para la selección y aplicación de los filtros en las fuentes de datos, se hizo un trabajo de revisión de la documentación otorgada por la SED, en este caso, correspondió al diccionario de datos de la BD1. Algunos ejemplos de los filtros empleados se presentan en la Tabla 1. Para todos los atributos se realizó un filtro de validación básica, pero también se construyeron algunos filtros de validación adicionales en los cuales se incluía la revisión de más de un atributo y en las que existían relaciones de dependencia, por ejemplo, la relación departamento y municipio debe estar acorde con la codificación definida por el DANE (Departamento Administrativo Nacional de Estadística).

Tabla I. Ejemplo de algunos de los filtros aplicados

ATRIBUTO	FILTRO - VALIDACIÓN BÁSICA	FILTRO - VALIDACIÓN ADICIONAL
Document_type Corresponde al tipo de documento de identidad, para este caso los aceptados en Colombia	1 - Cédula de Ciudadanía 2 - Tarjeta de Identidad 3 - Cédula de Extranjería o Identificación de Extranjería 5 - Registro Civil de Nacimiento 6 - Número de Identificación Personal (NIP) 7 - Número Único de Identificación Personal (NUIP) 8 - Número de Identificación establecido por la SED 9 - Certificado Cabildo 99 - No definido	La combinación tipo, número y lugar de expedición de documento debe ser única a nivel nacional. Si el tipo es 1(CC) y 2(TI) debe ser único por número de documento.
Social_level Corresponde al estrato socioeconómico, es un tipo de estratificación social que se usa en Colombia para agrupar según los ingresos familiares	0 - Estrato 0 1 - Estrato 1 2 - Estrato 2 3 - Estrato 3 4 - Estrato 4 5 - Estrato 5 6 - Estrato 6 9 – NA	9 es para identificar los casos especiales que no tienen estrato, puede ser por pertenecer a comunidades especiales (resguardos indígenas), ciudadanos extranjeros o no registraron el campo al momento de la matrícula
Exceptional_capabilities Corresponde a características o capacidades especiales que hacen sobresalir o resaltar algún estudiante	Filtro inicial: 1 - Superdotado 2 - Con talento científico 3 - Con talento tecnológico 4 - Con talento subjetivo 9 - No Aplica	Filtro después de modificación SIMAT en 2016: 1 - Capacidades excepcionales 2 - Talento científico 3 - Talento tecnológico 4 - Talento subjetivo/Artístico 5 - Talento atlético/deportivo 6 - Doble excepcionalidad 9 - No Aplica
Academic_situation_previous_year Condición de aprobación o reprobación del año escolar inmediatamente anterior	0 - No estudió en la vigencia anterior 1 - Aprobó 2 - Reprobó 8 - No culminó estudios	
student_status_previous_year Indica el estado del estudiante para el año escolar anterior en términos relacionados con deserción o traslados entre instituciones	3 - Desertó 5 - Trasladado a otra institución educativa 8 - Otro motivo de retiro 9 - No Aplica	Si Academic_situation_previous_year es igual a 8, el valor de la condición debe ser 3 o 5. Si Academic_situation_previous_year es diferente 8, el valor de la condición debe ser 9

Cabe resaltar que, para algunos de los atributos, estos filtros debieron ser transformados dado que cuando se iniciaba el filtrado se encontraban datos atípicos, sin explicación evidente y se debía recurrir a los propietarios de la fuente. Un ejemplo de esto sucedió con los códigos para las capacidades excepcionales (Exceptional_capabilities), para este atributo se encontró que en el año 2016 estos cambiaron y se empezaron a ingresar al sistema de matrícula algunos adicionales, pero no se encontraban documentados en el diccionario de datos inicial, este hallazgo permitió rescatar datos que podían haber sido borrados por considerarse atípicos.

Transformaciones. La estrategia de preparación de datos fue implementada en el lenguaje de programación Python, con la ayuda de Jupiter Notebook (https://jupyter.org/). Para la manipulación de los datos se utilizaron algunas librerías, una de ellas fue Pandas, que permite hacer el tratamiento de datos provenientes de diferentes estructuras utilizando el concepto de Dataframes para almacenar y manipular. Además, las librerías: Numpy para trabajar vectores y matrices multidimensionales con operaciones y funciones matemáticas y estadísticas; Matplotlib que permite hacer visualizaciones simples y Pickle para la serialización de objetos. El proceso de implementación se inició con un preprocesamiento y construcción de Dataframes para identificar mediante los encabezados, los diferentes atributos de interés y que pasarían a la fase de filtrado.

Seguidamente, se hicieron algunas exploraciones por medio de visualizaciones de muestras de los datos para, junto con el conocimiento del dominio, entender cómo se encontraban organizados los mismos, identificar posibles datos atípicos, faltantes, problemas de separación de columnas y de allí surgieron los

filtros iniciales. Jupiter Notebook permite que se pueda realizar esa primera visualización interactiva de los datos, que a la vez puede ser compartida con los expertos para que junto con los programadores se definan los filtros necesarios. Con el uso de Pandas (funcionalidad filters) se implementaron los filtros y se hizo la verificación de estos para llegar a la estructura de almacenamiento temporal. A partir de la estructura temporal se identifica la necesidad de modificar los filtros, luego de que los expertos revisan los datos de cada uno de los atributos, haciendo un refinamiento hasta obtener el conjunto de filtros que permitan extraer los datos que se quieran llevar al modelo con exactitud e integridad. Cuando los filtros ya presentan un nivel de refinamiento adecuado, se hace su aplicación y las transformaciones requeridas para llegar a la fase de carga.

Modelo de datos y carga. Para hacer la carga de los datos se diseñó un modelo de datos relacional con seis tablas: student, socioeconomic, academic, institution, centre y grades, (ver Figura 2). Posteriormente se creó la estructura en el motor de bases de datos PostgreSQL. Los datos fueron cargados siguiendo una secuencia definida para garantizar que las llaves primarias se poblaran en el orden requerido.

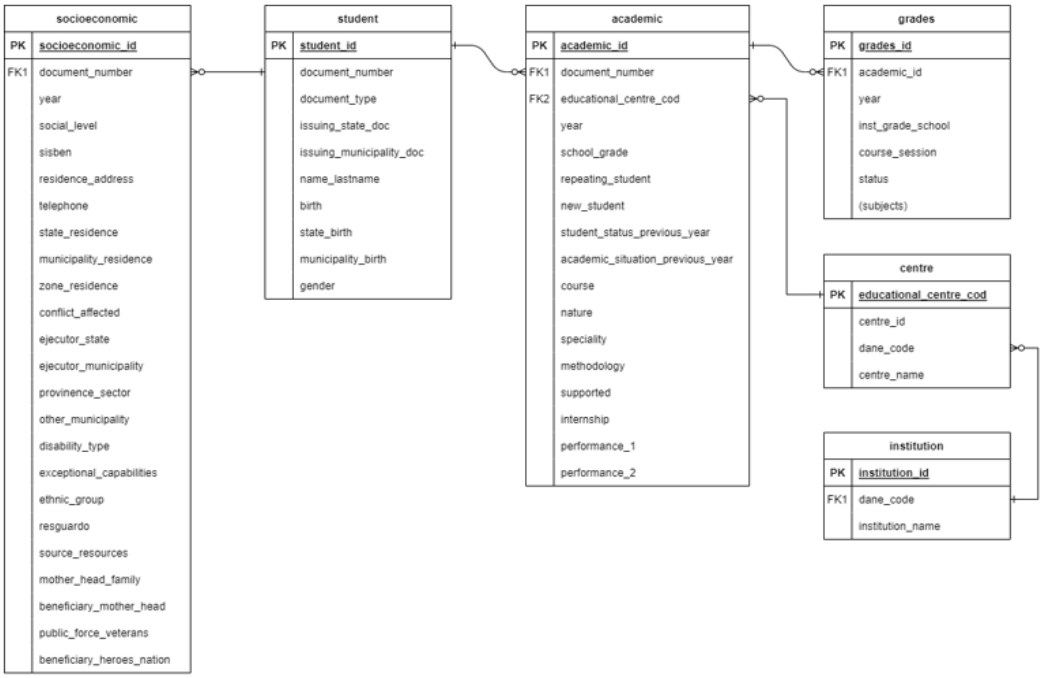

Figura 2. Modelo de datos.

Después de tener todos los datos de BD1 y BD2 pre-procesados se realizó la transformación y carga de los mismos en la base de datos, para poder hacer esto, se identificaron los atributos comunes entre las dos fuentes, los cuales son referentes al centro de estudio: educational_centre_cod, centre_name, institution_ name, y referentes al estudiante: name, document_number, birth y adicionalmente el atributo de año: year. Con estas informaciones se realiza una función "join'' y los datos son cargados. En primer lugar, se pobló la base de datos con los provenientes de BD1 donde se encontraban los datos socioeconómicos de los estudiantes y posteriormente los datos de BD2, referentes a las calificaciones. Se cargaron primero los datos de BD1 y luego los de BD2 por medio de los atributos **educational_centre_cod, document_number, year, academic_id. El atributo educational_centre_cod** es un atributo de alto interés en la BD2 pues contiene tanto información del centro o sede de estudio como de la institución. Lo anterior fue implementado por

medio de un script en Python, herramienta que permitió incluir las necesidades del dominio en el proceso de ETL.

Como resultado de esto se cargaron un total de 887.874 registros provenientes de la BD1 y 11.744 registros provenientes de la BD2, estos registros corresponden al periodo 2014 – 2018 que incluyen la información de matrículas de un total de 290.259 estudiantes provenientes de todo el departamento. Cabe anotar que el número de registros de calificaciones (BD2) es menor porque solo se trabajó con una muestra de tres IEs para este tipo de datos. De estos registros cargados no todos fueron intervenidos, al igual que algunos de los atributos no requirieron de filtros adicionales, solo filtros de rango básico.

Para algunas de las instituciones educativas se contaba con los datos de BD2 referentes al año 2013, pero estos no fueron cargados en la base de datos final ya que para la BD1 (datos socioeconómicos) no se logró tener acceso a este periodo, solo fueron suministrados a partir de 2014. Para la BD1 se recibieron inicialmente un total de 64 atributos, sin embargo, después de ser analizados se conservaron y cargaron solo 51 atributos, dado que los restantes no habían sido diligenciados para la mayoría de los estudiantes, lo cual los dejaba con la mayoría de los datos faltantes. Una de las transformaciones se dio al atributo status (estado) de la BD1 que correspondía a registros de tipo categóricos (aprobado, reprobado) se hizo una transformación a un tipo booleano (0 para reprobado, 1 para aprobado). De la misma forma, para los atributos de la BD2 que responden a dos posibilidades (SI – NO) se les transformó a boleando (0 para no y 1 para sí). Lo anterior con el fin de facilitar el almacenamiento en la base de datos y posterior procesamiento.

Resultados y Análisis

Para hacer la validación del ajuste de la estrategia para el proceso de preparación de datos incluyendo el enfoque de dominio específico, se realizó el procesamiento de los datos del caso de estudio y se lograron llevar al almacenamiento los datos provenientes de las diferentes fuentes y estructuras. A continuación, se describen los resultados en términos de métricas y análisis con relación a las preguntas orientadoras del estudio.

Métricas

La aplicación de las métricas incluyó dos momentos, en el inicial se hizo una evaluación de la calidad de los datos cargados haciendo uso de un indicador de confianza. Posteriormente, se aplicaron una serie de métricas que se definen a partir de la norma ISO/IEC 25012:2008 con relación a la calidad inherente a los datos, cuando se habla de inherente se hace referencia a las características de calidad de los datos que les dan el potencial para suplir algunas necesidades de los usuarios a través de su uso bajo un dominio específico [19]. De acuerdo con esto, se decidió aplicar métricas inherentes de exactitud, completitud y consistencia de los datos.

Confianza. Este indicador se define como la relación entre la cantidad de registros que ingresan sin errores y el total de registros entrantes por cada atributo filtrado [20]. El resultado del cálculo del indicador de confianza para algunos de los atributos principales presentes en la base de datos muestra que, atributos como document_type, school_grade, disability_type y academic_situation_previous_year tienen una confianza de 0.99 es decir, del 99%. Sin embargo, se presentan otros atributos como course_session con 93%, conflict_affected con 82,9%, dane_code y educational_centre_cod con 0,0%. Los datos correspondientes a course_

session eran de tipo categórico (mañana, tarde y sábados), no obstante, se recibieron cadenas de caracteres con presencia de errores de digitación, se tenía diferentes formas de escribir: algunas veces todo en mayúsculas, otras todo en minúscula o combinación de estas, este hallazgo se complementa con una métrica de exactitud que se mencionara más adelante. En cuanto a conflict_affected se encontraban registros fuera de los valores aportados por el diccionario de datos. El caso de dane_code y educational_centre_cod, los cuales muestran una confianza de 0%, se debe a problemas con la exportación de los archivos que se recibieron de la fuente, dado que las columnas correspondientes a estos códigos fueron corrompidas y presentaban datos atípicos, por lo cual sus valores se tuvieron que rescatar a partir del nombre de la institución y de la sede, haciendo uso de una tabla de codificación externa del DANE. En relación con los datos de calificaciones presentes en la BD2, todas se encontraban en el rango válido (0-5), este rango es el patrón de calificaciones que usan las instituciones del caso de estudio. Por lo cual para estos atributos la confianza es del 100%, esto puede ser explicado por el hecho de tener un mayor cuidado en las instituciones con estos valores, dado que son los entregados a los estudiantes y acudientes y con los cuales se establece el estado de aprobación o reprobación del año escolar. En la Figura 3 se presenta un resumen del indicador de confianza.

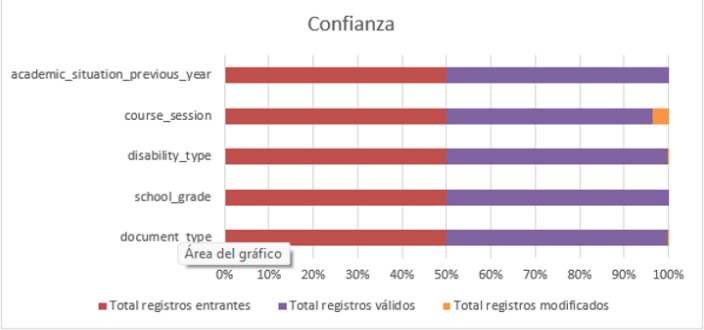

Figura 3. Indicador Confianza.

Exactitud. Hace referencia a la representación de forma correcta del valor que los atributos de un concepto pueden tomar en un contexto específico de uso. Para el caso de estudio se decidió aplicar dos métricas de esta característica, se midió la Exactitud Sintáctica de los Datos, la cual se aplicó para los atributos categóricos institution_name, centre_name, y course_sesion y el Riesgo de Inexactitud del conjunto de datos, el cual se aplicó a todos los atributos. La Exactitud Sintáctica de Datos, se define como la relación X=A/B, donde A es la cantidad de datos con errores sintácticos para un atributo categórico y B es la totalidad de registros para dicho atributo. En el caso de los tres atributos analizados con esta métrica se obtuvieron los siguientes resultados: institution_name 17.6%, centre_name 5.58% y course_sesion 3.18%. Esto quiere decir que cada uno de los porcentajes mostrados corresponde a la cantidad de registros que presentaban errores sintácticos. Se midió esta métrica para los atributos mencionados, dado que eran los atributos categóricos para los cuales se podía tener una referencia sintáctica. En el caso del Riesgo de Inexactitud del Conjunto de Datos se define como la relación X=A/B, donde A es la cantidad de datos atípicos para un atributo dado, y B es el total de registros para dicho atributo. Para esta métrica se destacan los siguientes resultados: dane_code 100%, educational_centre_cod 100%, conflict_affected 17,1%.

Completitud. Hace referencia a la presencia total de valores para todos los atributos e instancias esperadas en un contexto específico de uso. En otras palabras, la completitud es definida como el porcentaje de registros o campos existentes o diligenciados. La completitud, para este caso, fue evaluada desde dos métricas, la

Completitud de Registros y la Completitud de Atributos. La Completitud de Registros es la relación X = A/B donde A es la cantidad de datos no nulos y B es la cantidad de datos esperados. Esta métrica fue medida para todos los atributos recibidos y se destacan los resultados encontrados para: institution_name 49,03%, centre_name 49,03%, sisben 96,13%, ejector_state 95,21%, ejector_municipality 95,21%, performance_1 100%, performance_2 100%. Por su parte, la Completitud de Atributos es la relación entre la cantidad de atributos con datos nulos y el total de atributos recibidos de las fuentes. Esta fue medida sobre cada una de las BD, encontrando: 31,25% para BD1 y 63,67% para BD2. Se muestra en la Figura 4 el indicador de completitud.

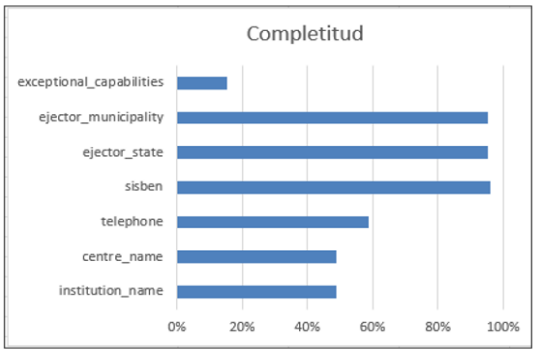

Figura 4. Indicador Completitud.

Consistencia. Hace referencia a la condición de los datos libres de contradicciones y coherentes en un contexto específico de uso. Para este caso se tomaron en cuenta dos métricas, la Consistencia del Formato de los Datos y el Riesgo de Inconsistencia de los datos. La primera fue aplicada para los archivos recibidos de parte de las fuentes, teniendo como base de formato el primer año recibido, tanto para las calificaciones (BD2) como para los archivos de matrícula (BD1). La segunda métrica se aplicó sobre los atributos "document_number" y "name". La consistencia del formato de datos se define como la relación entre el número de ítems de datos donde el formato de todas las propiedades es consistente en diferentes archivos de datos y el número de ítems de datos para los cuales la consistencia de formatos debe estar definida. Para la BD2 se hizo la revisión por institución educativa, por año y se tomó como referencia base del formato el primer año. Para la BD1 se hizo la revisión por año, siendo 2014 el año base en cuanto al formato. Se encontró que 100% para BD1 y 0% para los datos de BD2, dado que todos los archivos difieren del primer año recibido. Para el Riesgo de Inconsistencia de los Datos, definido como la relación entre el número de ítems de datos donde existe duplicación en valores y el número de ítems de datos considerados, se encontró para los atributos "document_number" 67,31% y "name" 68%. Esto cuando se realiza una evaluación de los atributos aislados, pero teniendo en cuenta otros atributos, como 'name','birth','state_birth','municipality_birth, el valor de la métrica baja a 5,31% de estudiantes duplicados.

Análisis

Para continuar con el análisis y discusión del desarrollo de esta experiencia se traen a colación las tres preguntas que orientaron el trabajo y se describe la respuesta a la que se logró llegar por medio de la aplicación de la estrategia de preparación.

P1 - ¿Es viable la inclusión del enfoque de dominio específico en el proceso de preparación de datos educativos, garantizando la calidad y consistencia de estos?

Por medio de la aplicación de la estrategia al caso de estudio se encontró viable la inclusión del enfoque de dominio específico en la preparación de datos provenientes de un ambiente educativo. Un primer momento y uno de los más importantes fue conocer los datos y entender el funcionamiento del sistema educativo objeto de estudio a través del contacto con los expertos y de la exploración de la información/documentación recibida, dado que esto sentó las bases para proponer una solución adaptada a las necesidades del contexto.

Cabe resaltar que la experiencia reportada introduce la preparación de los datos previa a la minería de datos educativos, se logra garantizar que los datos lleguen con calidad y consistencia al paso siguiente. Para llegar al desarrollo de esta estrategia se hizo un proceso de revisión de literatura y se tuvieron en cuenta experiencias y hallazgos previos por autores que han trabajado en EDM. El enfoque de dominio específico permite enriquecer el proceso de preparación de datos en la medida se incluye la experticia en la detección de problemas e inconsistencias por parte de expertos calificados. Se reconoce que el experto debe estar activamente involucrado en el diseño de los filtros y la validación de los datos.

P2 - ¿Cuáles son las dificultades que se pueden encontrar en el diseño e implementación de un proceso de preparación de datos con enfoque de dominio específico, previo a la aplicación de EDM?

En particular, cuando se hizo la implementación para el caso de estudio surgieron inconvenientes como los que se resumen a continuación: (1) Las fuentes no se "hablaban" directamente entre sí, por lo que existió una gran dificultad para unir los datos. (2) El estudiante puede tener más de un documento de identidad a lo largo de su vida escolar. (3) En la BD2 se dificulta el seguimiento del estudiante porque solo se tiene el nombre de este para hacer la identificación. Adicionalmente, el nombre del estudiante tiene formato diferente en las dos BDs. En la BD1, aparece como un campo único, mientras que en BD2 viene como nombre1, nombre2, apellido1, apellido2; y se requirió hacer una unión para poder hacer la posterior comparación y unificación entre BDs. (4) Algunos estudiantes tienen el mismo nombre y en la BD2 este era el identificador que se podía usar. (5) Si el estudiante cambia de institución educativa en un mismo año escolar, al no contar con el documento en la BD2, se dificulta encontrar un nuevo registro en la matrícula (BD1). (6) Errores en la digitalización de los datos, por ejemplo, edades superiores a 100 años. (7) A lo largo de los años, algunos códigos institucionales y de la sede han cambiado y se tuvo que reasignar los códigos haciendo una transformación a la nueva codificación. (8) La presencia de documentos de identidad extranjeros es otro reto, puesto que presentan formato diferente y pueden cambiar más de una vez a lo largo de la vida escolar por su carácter temporal.

Es de resaltar el inconveniente dado por la no existencia de un atributo de identificación único para los estudiantes entre las diferentes fuentes. A pesar de que el número de documento del estudiante parecía en un principio ser un identificador único, se verificó posteriormente que los estudiantes podían cambiar de documento de identidad a lo largo de los años presentes en los datos, incluso más de una vez. Por lo anterior, este atributo por sí solo no era suficiente para poder registrar al estudiante en la base de datos; teniendo que recurrir a realizar una combinación de atributos para tener una forma compuesta de identificar a cada estudiante, esto se realizó por medio del nombre completo, la fecha y municipio de nacimiento.

P3 - ¿Es factible aplicar el proceso de preparación de datos con enfoque de dominio específico en un entorno de educación básica y media?

Si es factible, posterior a superar los problemas mencionados, con los filtros definidos y depurados, el proceso de carga en el modelo de datos se llevó a cabo y el almacenamiento podrá ser actualizado de acuerdo con la aparición de nuevos datos. En el momento con la base de datos poblada se puede hacer la selección de dataset para pruebas de acuerdo con los análisis que se deseen realizar, por ejemplo, clasificación de acuerdo con el estado de aprobación-reprobación, clustering por calificaciones o características socioeconómicas, trayectorias escolares, entre otras. La construcción de un modelo de dominio específico cobra importancia para permitir el uso de conocimiento del experto en la validación de la calidad y coherencia de los datos. Las herramientas de preparación de datos genéricas son importantes, pero no se puede dejar de lado a los expertos, porque a pesar de la herramienta ofrecer los medios no tienen cómo detectar aspectos particulares y/o críticos. En concreto, el enfoque de dominio específico fue aprovechado al proveer una guía la preparación de datos para los interesados en la minería de datos educativos. Esto por medio de construcciones basadas en el conocimiento recolectado de la literatura, el contexto y compartido por los expertos, facilitando el análisis para llegar a una interpretación significativa de la información rescatable.

De esta experiencia surgen algunas recomendaciones para los propietarios de las fuentes, directivos y encargados de las políticas de administración de los datos del caso de estudio. En primer lugar, definir una identificación única a lo largo de la vida escolar del estudiante permitiría la eliminación de registros duplicados innecesarios, inconsistencias en la información y mejoraría la comunicación entre instituciones educativas, SED y el Ministerio de Educación Nacional. Así mismo, contar con un sistema de gestión de calificaciones que sea común para las instituciones educativas del departamento, permitiría a la SED hacer un mejor control y análisis de los datos, al igual que evitaría que se pierdan los históricos de esta información que es de alta importancia para el soporte de decisiones y para la generación de políticas de mejora de los procesos de enseñanza-aprendizaje. Además, se hace necesario garantizar la custodia de esta información y protegerla dada su sensibilidad.

Conclusiones y Trabajo Futuro

La tarea de preparación de datos no es un proceso lineal, es un proceso iterativo, un ciclo. En el desarrollo de esta experiencia, pasando de la fase de filtrado a la carga, se fueron descubriendo problemas y se requirió volver al refinamiento y entendimiento para encontrar la solución.

Utilizar datos reales para la validación de la estrategia contribuyó tanto en el ámbito investigativo como en la aplicación del conocimiento en un caso existente y abrió la posibilidad a hacer pruebas posteriores con otro tipo de datos del contexto. Uno de los propósitos de este documento es servir como una guía práctica para profesionales e investigadores que deseen realizar experimentación con técnicas de EDM en instituciones educativas de básica y media en Colombia.

Existe la necesidad de adelantar trabajos que tomen los datos de otros niveles diferentes a la educación superior [21], como, por ejemplo, de educación básica y media, siendo este un diferencial de la experiencia relatada en este artículo. El diseño de un modelo de almacenamiento para los datos producidos en educación básica (primaria y secundaria) y media dará un gran beneficio a los funcionarios de las instituciones educativas públicas/gubernamentales, al obtener una versión única de la información escolar que está siendo producida en su ecosistema.

Como trabajo futuro, se plantea integrar nuevas fuentes de datos, por ejemplo, los resultados de la evaluación externa de la calidad educativa, que en Colombia es realizada a través de las pruebas SABER (3°, 5°, 9° y 11°) por el Instituto Colombiano para la Evaluación de la Educación – ICFES [22]. El ICFES cuenta con una plataforma de datos abiertos que permite acceso a los datos para los investigadores, sin embargo, los datos se encuentran anonimizados y se tendría que evaluar la opción para poderlos integrar con las otras fuentes.

Agradecimientos

Al programa de Formación de Capital Humano de Alto Nivel para el Departamento de Norte de Santander en el marco de la Convocatoria N°753 de Colciencias.

Referencias

[1] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017, doi: 10.1109/ACCESS.2017.2696365.

[2] K. Kasemsap, "Knowledge discovery and data visualization: theories and perspectives," *Int. J. Organ. Collect. Intell*., vol. 7, no. 3, 2017, Accessed: Jul. 28, 2020. [Online]. Available: https://www.igi-global.com/article/knowledge-discovery-and-data-visualization/182757.

[3] M. Couceiro and A. Napoli, "Elements about exploratory, knowledge-based, hybrid, and explainable knowledge discovery," in *ormal Concept Analysis. ICFCA 2019. Lecture Notes in Computer Science*, vol. 11511, Cristea D., Le Ber F., and Sertkaya B., Eds. Springer Cham, 2019, pp. 3–16, DOI:10.1007/978-3-030-21462-3_1

[4] B. Oliveira and O. Belo, "On the specification of extract, transform, and load patterns behavior: A domain-specific language approach," *Expert Syst*., vol. 34, no. 1, p. e12168, Feb. 2017, doi: 10.1111/exsy.12168.

[5] T. Costello and L. Blackshear, "What Is ETL?," in *Prepare Your Data for Tableau,* Apress, 2020, pp. 1–3.

[6] R. Wang et al., "Review on mining data from multiple data sources," *Pattern Recognit. Lett*., vol. 109, pp. 120–128, Jul. 2018, doi: 10.1016/J.PATREC.2018.01.013.

[7] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective," *IEEE Trans. Knowl. Data Eng*., vol. 33, no. 4, pp. 1328–1347, Apr. 2021, doi: 10.1109/TKDE.2019.2946162.

[8] V. Debroy, L. Brimble, and M. Yost, "NewTL: Engineering an extract, transform, load (ETL) software system for business on a very large scale," in *Proceedings of the ACM Symposium on Applied Computing*, Apr. 2018, pp. 1568–1575, doi: 10.1145/3167132.3167300.

[9] F. Bellifemine, G. Fortino, R. Giannantonio, R. Gravina, A. Guerrieri, and M. Sgroi, "SPINE: a domain-

specific framework for rapid prototyping of WBSN applications," *Softw. Pract. Exp*., vol. 41, no. 3, pp. 237–265, 2011, doi: 10.1002/spe.998.

[10] G. Desolda, C. Ardito, and M. Matera, "Empowering end users to customize their smart environments: Model, composition paradigms, and domain-specific tools," *ACM Trans. Comput. Interact*., vol. 24, no. 2, pp. 1–52, Apr. 2017, doi: 10.1145/3057859.

[11] A. Iung et al., "Systematic mapping study on domain-specific language development tools," *Empir. Softw. Eng*., vol. 25, no. 5, pp. 4205–4249, Sep. 2020, doi: 10.1007/S10664-020-09872-1/TABLES/9.

[12] M. Beg, R. A. Pepper, and H. Fangohr, "User interfaces for computational science: A domain specific language for OOMMF embedded in Python," *AIP Adv*., vol. 7, no. 5, p. 056025, Feb. 2017, doi: 10.1063/1.4977225.

[13] P. Selvaraj, V. K. Burugari, D. Sumathi, R. K. Nayak, and R. Tripathy, "Ontology based Recommendation System for Domain Specific Seekers," in *Third International conference on I-SMAC (IoT in Social*, Mobile, Analytics and Cloud), 2019, pp. 341–345, doi: 10.1109/i-smac47947.2019.9032634.

[14] J. Samuelsen, W. Chen, and B. Wasson, "Integrating multiple data sources for learning analytics—review of literature," *Res. Pract. Technol. Enhanc. Learn*., vol. 14, no. 1, pp. 1–20, Dec. 2019, doi: 10.1186/S41039-019-0105-4/TABLES/7.

[15] M. S. Mussa, S. C. Souza, E. F. S. Freire, R. G. Cordeiro, and H. R. M. Hora, "Business intelligence in education: an application of pentaho software," *Rev. Produção e Desenvolv*., vol. 4, no. 3, pp. 29–41, 2018, doi: 10.32358/rpd.2018.v4.274.

[16] G. G. W. Mhon and N. S. M. Kham, "ETL Preprocessing with Multiple Data Sources for Academic Data Analysis," Feb. 2020, doi: 10.1109/ICCA49400.2020.9022824.

[17] J. vom Brocke, A. Hevner, and A. Maedche, "Introduction to Design Science Research," pp. 1–13, 2020, doi: 10.1007/978-3-030-46781-4_1.

[18] Ministerio de Educación Nacional, "Sistema educativo colombiano," 2020. https://www.mineducacion.gov.co/portal/Preescolar-basica-y-media/Sistema-de-educacion-basica-y-media/233839:Sistema-educativo-colombiano (accessed May 01, 2020).

[19] M. A. Fernández Sáenz, "Desarrollo de un modelo de calidad de datos aplicado a una solución de inteligencia de negocios en una institución educativa : Caso Lambda," Pontificia Universidad Católica del Perú, 2018.

[20] N. D. Duque-Méndez, E. J. Hernández-Leal, A. Pérez Zapata, A. Arroyave Tabares, and D. Espinosa Gómez, "Modelo para el proceso de extracción, transformación y carga en bodega de datos. Una aplicación con datos ambientales," *Cienc. e Ing. Neogranadina*, vol. 26, no. 2, pp. 95–109, 2016.

[21] G. Jayashree and C. Priya, "Comprehensive Guide to Implementation of Data Warehouse in

Education," in *Intelligent Computing and Innovation on Data Science*, vol. 118, Springer, Singapore, 2020, pp. 1–8.

[22] Instituto Colombiano para la Evaluación de la Educación (Icfes), "Portal Icfes," 2019. https://www.icfes.gov.co/web/guest/funciones-icfes (accessed Aug. 30, 2019).