Modelo Dimensional para el Registro Académico de Estudiantes de la Universidad del Cauca

Trabajo de Grado

Edison Eduardo Cerón Moreno
Daniel Alejandro Urrea Pito

Director: PhD. Martha Eliana Mendoza Becerra

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Línea de Investigación Gestión de la información y Bodega de datos
Popayán, octubre de 2018

Modelo Dimensional para el Registro Académico de Estudiantes de la Universidad del Cauca

Trabajo de Grado

Edison Eduardo Cerón Moreno
Daniel Alejandro Urrea Pito

Director: PhD. Martha Eliana Mendoza Becerra

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Línea de Investigación Gestión de la información y Bodega de datos
Popayán, octubre de 2018

Agradecimientos

Agradezco la realización de este trabajo principalmente a Dios, por permitirme haber llegado hasta este momento tan importante de mi formación. A mi familia por demostrarme su incondicional cariño y apoyo.

Gracias, a mi tutora, la doctora Martha Mendoza, por su dedicación, motivación y paciencia, fue un privilegio contar con su guía y ayuda. Gracias a mi compañero Eduardo Cerón, por su constante dedicación y esfuerzo en la realización de este trabajo.

Gracias a los usuarios, que participaron en gran parte de la realización de este trabajo y de esta forma ayudaron construir en gran medida los modelos presentados.

Finalmente, doy gracias a todas las personas que de una u otra manera, hicieron parte de este trabajo.

Daniel Alejandro Urrea P.
Popayán, Octubre de 2018

En primer lugar, quiero agradecer a Dios por permitirme llegar hasta este punto, a mi familia por apoyarme en todo momento. A mi padre y mi madre por brindarme su amor y realizar todos los esfuerzos posibles que me permitieron sacar adelante esta carrera. Gracias a ellos por todos los consejos que me permitieron seguir adelante y no decaer.

Gracias a nuestra directora Martha Mendoza, ya que gracias a su dedicación, paciencia y compromiso, se consiguió la realización de este trabajo.

Gracias a mi compañero Daniel Urrea por su dedicación, compromiso y su responsabilidad, lo cual fue importante para el desarrollo de este trabajo.

Gracias a las personas que hicieron parte del grupo de usuarios, ya que gracias a sus aportes se logró desarrollar este trabajo de la mejor manera.

Finalmente, doy gracias a todas las personas que de una u otra manera, hicieron parte de este trabajo.

Edison Eduardo Cerón M.
Popayán, Octubre de 2018.

Tabla de contenido

Índice de tablas

Índice de figuras

Capítulo I. Introducción

En este capítulo se presenta de manera detallada la motivación, problemática y justificación de este trabajo de investigación como también sus objetivos. Adicionalmente, se presenta la estructura del documento, la cual resume el contenido de cada uno de los capítulos desarrollados.

1.1. Problemática y justificación

Las instituciones educativas de nivel superior han observado durante la última década un crecimiento significativo en el número de estudiantes, debido al interés y preocupación del sector educativo por ampliar la cobertura de la oferta. Tomando como referencia los datos publicados por el Ministerio de Educación Nacional correspondientes a la cantidad de estudiantes matriculados por período académico [1], se observa en la Figura 1 que en los últimos diez años se ha generado un incremento del 93.02% en los estudiantes de pregrado a nivel nacional. Por otro lado se observa en la Figura 2 un crecimiento del 19.89% de los estudiantes de pregrado en la Universidad del Cauca en el mismo intervalo de tiempo, aunque el porcentaje de variación no se mantiene año a año se verifica una clara tendencia hacia una población estudiantil aún mayor en los siguientes años.

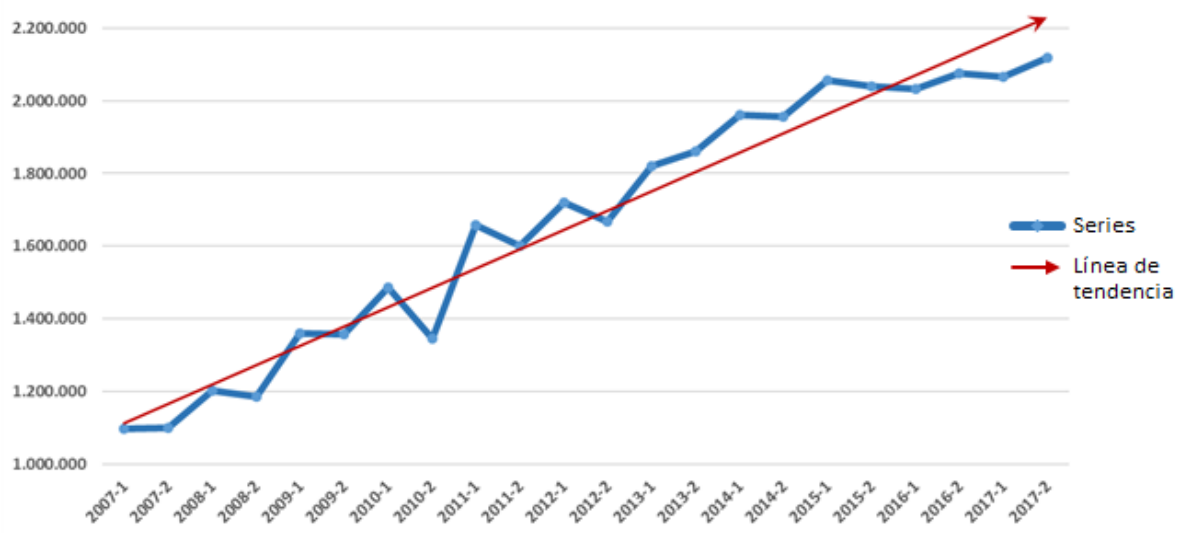


Figura 1 Cantidad de estudiantes última década a nivel nacional.

El crecimiento de esta población significa a su vez un incremento de los datos personales y académicos del estudiantado almacenados en las diferentes fuentes de información universitarias, dando la posibilidad de crear un sistema que facilite los procesos enfocados en la búsqueda y posterior análisis de información interna, permitiendo a las directivas evaluar la situación actual de la institución con el fin de prever y decidir sobre las acciones que deberán desarrollarse a futuro. A pesar de contar con esta gran cantidad de datos, hasta ahora gran parte de este tipo de instituciones no han logrado generar una consolidación de estos a nivel interno para dar soporte a un sistema que favorezca la toma de decisiones y la proyección institucional, debido a esto el horizonte de planeación se ha

visto limitado a períodos de tiempo cortos, generando así un gran inconveniente para las instituciones educativas de nivel superior [2].

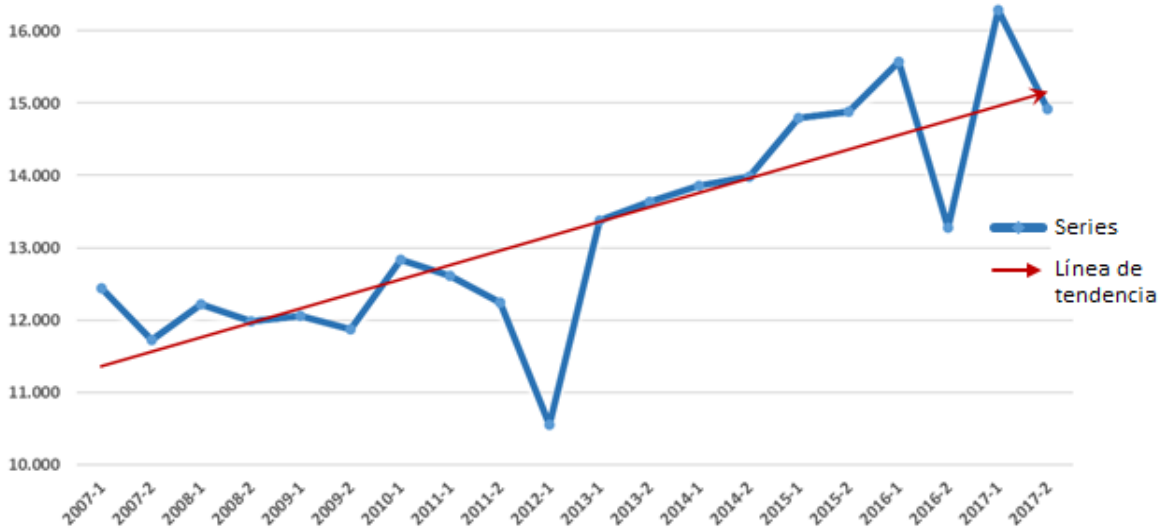


Figura 2 Cantidad de estudiantes última década en la Universidad del Cauca.

En este sentido, las bodegas de datos permiten consolidar diferentes fuentes de datos con el objetivo de facilitar las tareas de consulta y análisis, sin embargo, se han enfocado en el sector financiero y los procesos de ventas, generando impacto en la relación existente con el cliente y permitiendo la identificación de los productos más significativos. Aun reconociendo los beneficios que puede brindar una bodega de datos en los procesos de análisis y planificación, no es común su aplicación en el ámbito de la educación pública del país, esto por lo general, se debe a que las universidades del sector público se ven limitadas en su capacidad de adquisición, por lo cual no pueden costear o acceder a aquellas soluciones ofertadas por terceros. Por otro lado, la creación de una solución de bodegas de datos por parte del personal interno de la institución resulta compleja ya que en la mayoría de los casos no se cuenta con un personal capacitado en el diseño e implementación de este tipo de soluciones.

Tomando como enfoque general el desempeño del nivel académico y la continua mejora de este [3], se considera viable el diseño de un modelo dimensional para el registro académico que constituya un pilar fundamental para la creación de una bodega de datos, que permita a los directivos de la institución determinar las variables que afecten dicho desempeño [4]. No obstante, uno de los aspectos de mayor dificultad al abordar este tipo de soluciones es la creación del modelo dimensional, esto debido a que su enfoque es diferente al de un modelo relacional, contemplando elementos y relaciones particulares.

En la literatura se encuentran trabajos relacionados con las bodegas de datos [5]–[20], que modelan uno o más procesos académicos, tales como: matrícula académica, registro de notas, etc. Sin embargo, gran parte de los modelos dimensionales expuestos en estos trabajos no presentan más que un esquema general, sin explicar en detalle los componentes que los constituyen, ni los casos de diseño presentados, lo cual dificulta su entendimiento y apropiación. Además, se encontraron inconsistencias a nivel de diseño que serán explicadas en detalle en la Tabla 3 y falta de información socio-demográfica que

potencie el poder de análisis de dichos modelos. Por otro lado, se han localizado instituciones de nivel superior en el exterior que cuentan actualmente con una bodega de datos [21]–[24], aunque manifiestan la ganancia analítica que esta les ha generado, no presentan los componentes que hacen parte de la bodega (dimensiones y hechos), ni tampoco se presentan los modelos dimensionales que permitan comprenderla.

Por lo anterior, se dificulta que las universidades públicas del país puedan tomar como base estos trabajos para generar sus propios modelos dimensionales. Teniendo en cuenta el alcance de este proyecto de pregrado y que las universidades públicas colombianas manejan datos similares con respecto al registro académico, surge la siguiente pregunta de investigación: ¿Qué casos de diseño se deben tener en cuenta en el modelado dimensional del registro académico que contemple el registro de notas, control de asistencia e información socio-demográfica de los estudiantes de la Universidad del Cauca, y que pueda servir de referente para otras universidades públicas del país?

En este sentido, este trabajo de grado presenta el modelado dimensional de una bodega de datos para el registro académico (con un enfoque en los procesos de registro de notas y control de asistencia), considerando información socio-demográfica y de rendimiento académico de los estudiantes. Además, la implementación de un prototipo de bodegas de datos para el registro de notas de la facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca, que permite que las directivas de la facultad tengan contacto con este tipo de tecnología e identifiquen la utilidad de la misma como soporte para identificar problemas y definir nuevas estrategias que mejoren el rendimiento académico de los estudiantes.

1.2. Objetivos

1.2.1. Objetivo general

Proponer un modelo dimensional para el registro académico de los estudiantes de la Universidad del Cauca que incluya información socio-demográfica, que se pueda convertir en un referente para otras universidades públicas del país.

1.2.2. Objetivos específicos

- Identificar los casos de diseño que se deben tener en cuenta en el modelado dimensional del registro académico de los estudiantes que contemple registro de notas, control de asistencia e información socio-demográfica.
- Construir un prototipo de una bodega de datos para el registro de notas de la Facultad de Ingeniería Electrónica y Telecomunicaciones (FIET) basado en el modelo dimensional propuesto, que permita la visualización de los datos por medio de consultas analíticas Ad-hoc y de reportes estándar.
- Evaluar la utilidad[1] de los reportes estándar y las consultas analíticas que genere el prototipo de la bodega de datos, por medio de un test de nivel de satisfacción de los funcionarios académicos responsables del registro académico en la FIET, con base en la métrica *nivel de satisfacción*, de la ISO/IEC 25022: Medidas de Calidad en Uso, definida para la sub-característica *utilidad*.

[1] Grado en que un usuario es satisfecho cuando logra alcanzar sus objetivos planteados.

1.3. Estructura del documento

A continuación, se presentan los siete (7) capítulos que componen este trabajo de investigación, junto con una breve descripción de su contenido.

Capítulo I. Introducción: En este capítulo se presenta la problemática y la justificación que motivan la realización de este proyecto de investigación. Adicionalmente, se presentan los objetivos planteados.

Capítulo II. Marco teórico y estado del arte: Este capítulo presenta la descripción del conocimiento necesario que se requiere para comprender el problema de investigación. Así mismo, se presenta una revisión sistemática, como también un análisis de los resultados obtenidos a partir de esta. Se analizan investigaciones recientes que han formulado alguna solución de bodegas de datos dentro del ámbito académico. Finalmente, se presentan los aportes de este proyecto de investigación al tema de interés.

Capítulo III. Metodología seleccionada: En este capítulo se presenta la metodología por medio de la cual se elaboraron los modelos dimensionales y el prototipo de la bodega de datos, se dan a conocer las características que llevaron a seleccionarla, las fases que se han tomado como guía para el desarrollo de este proyecto como también una mención de las sub-fases relevantes dentro de esta.

Capítulo IV. Modelado dimensional: En este capítulo se presentan los modelos dimensionales diseñados en este proyecto, se especifican los componentes que lo constituyen, las jerarquías que se consideraron pertinentes como también los casos de diseño necesarios para su elaboración.

Capítulo V. Elaboración del prototipo propuesto: En este capítulo se presentan de manera específica las fases de desarrollo llevadas a cabo para la implementación del prototipo de bodega de datos enfocado en las necesidades analíticas de la FIET, se describen además los entregables realizados y adicionalmente las estrategias sugeridas para llevar a cabo una correcta ejecución de un proyecto de este tipo.

Capítulo VI. Evaluación del prototipo: En este capítulo se presenta la forma en la que se realizó la evaluación del prototipo. Así mismo, se presentan los objetivos, análisis de resultados, comentarios de los participantes y las acciones de mejora consideradas.

Capítulo VII. Conclusiones y trabajos futuros: Presenta las conclusiones obtenidas a partir de la realización del trabajo de investigación, las lecciones aprendidas, los aportes en la investigación realizados y posibles trabajos futuros.

Capítulo II. Marco teórico y estado del arte

En este capítulo se presentan los términos considerados dentro del marco teórico, que facilitan la comprensión del proyecto y permiten al lector familiarizarse con la temática de las bodegas de datos. Así mismo, se presenta una revisión sistemática de la literatura que permitió identificar los elementos principales y tendencias en el área de las bodegas de datos académicas, las técnicas y estrategias que se han propuesto recientemente para afrontar los diferentes retos que supone el proceso de modelado dimensional.

2.1. Marco teórico

2.1.1. Bodegas de datos

Según Inmon [25], "Una bodega de datos es una colección de datos, integrados, temáticos, no volátiles y variables en el tiempo, organizados para soportar necesidades empresariales orientadas a la toma de decisiones". A continuación, se definen estas cuatro características:

- **Integrados**: La información que se encuentra en una bodega siempre está integrada. Habitualmente, las empresas cuentan con diferentes sistemas operacionales y/o fuentes de datos externas, cada uno de ellos con sus propias bases de datos que les ayudan a soportar los diversos procesos de sus áreas de negocio. En la creación de una bodega de datos para una empresa, todos los datos de los diversos sistemas operacionales deben integrarse en una sola base de datos, por lo que las inconsistencias existentes en el momento de la integración de los diversos sistemas operacionales, deben ser eliminadas, esto implica tareas costosas de limpieza, transformación y derivación de datos [26].
- **Temáticos**: Están diseñados para ayudar a analizar los datos de un determinado tema, además los datos se encuentran relacionados entre sí, donde dicha relación se da por el proceso de negocio al que pertenecen. Por ejemplo, para saber más sobre una institución universitaria se puede construir una bodega que concentre las consultas del registro financiero, que permitiría responder preguntas del tipo: ¿Cuál ha sido el concepto de descuento más recurrente?, ¿Cuál es el promedio del valor pagado como concepto de matrícula financiera por estrato económico? La habilidad de enfocarse en un proceso prioritario hace que se considere una construcción orientada a un tema [27].
- **No Volátil**: Implica la durabilidad de los datos, estos no pueden ser modificados, ni eliminados; para realizar un buen análisis de los datos, es conveniente contar con datos solidos de diferentes períodos de tiempo para su comparación, por lo tanto, una bodega es para ser consultada y no modificada. Los datos son entonces permanentes y las actualizaciones se realizan solamente en el ingreso de datos correspondientes al último período de tiempo [28].
- **Variables en el Tiempo**: Indica la posibilidad de contar con valores diferentes de un mismo dato de acuerdo con sus cambios en el tiempo. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Los datos que se encuentran en una bodega sirven, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, la bodega se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones [28]. Cabe resaltar que una bodega de datos incluye además del componente de almacenamiento, un módulo de

visualización que permite al usuario interactuar con los datos por medio de consultas e informes [4].

Data marts: Dentro de las bodegas de datos se pueden encontrar diferentes divisiones, denominadas como datos departamentales (Data Mart), cada una de estas divisiones es enfocada usualmente en un proceso de negocio, con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Además, cada data mart debe ser representado por un modelo dimensional, por tanto la unión de todos define la estructura de una bodega de datos [4].

2.1.2. Modelado dimensional

Se considera como una técnica que ofrece al usuario una visión clara de la operación del negocio, enfocada en modelar las particularidades de los procesos que suceden en la organización. La estructura del modelo está diseñada de forma tal que facilite la comprensión a los usuarios, optimice el rendimiento de las consultas [4] y permita modificaciones de forma fácil para lograr una rápida adaptación ante los cambios en las necesidades de la información [27].

Dentro del modelado, los procesos de negocio son divididos en medidas y entorno. Las medidas son en su mayoría numéricas y se les denomina hechos. Alrededor de estos existe un contexto que describe en qué condiciones y en qué momento se registraron. Aunque el entorno se ve como un todo, existen registros lógicos de diferentes características que describen un hecho. Por ejemplo, si el hecho referido, es la nota de un estudiante, se podría dividir el entorno que rodea al hecho, en el docente que impartió la materia, el estudiante que obtuvo la nota, la materia en la que fue evaluado y la fecha en que se realizó la evaluación. A estas divisiones se le denomina dimensiones y a diferencia de los hechos que son numéricos, estos son fundamentalmente textos descriptivos [29].

Los principales componentes de un modelado dimensional son las tablas de hechos y las tablas de dimensión, que se pueden definir de la siguiente manera [4]:

- **Tablas de hechos**, representan los procesos que ocurren en la organización, son independientes entre sí (no se relacionan unas con otras). La llave de la tabla de hechos, es una llave compuesta que se forma con las llaves primarias de las tablas dimensionales a las que está unida. Se pueden distinguir dos tipos de columnas en una tabla de hechos, columnas de hechos que almacenan las medidas del negocio que se quieren controlar y las columnas llaves que forman parte de la llave de la tabla [29].

- **Tabla de dimensión**, que contiene, por lo general, una llave simple y un conjunto de atributos que describen la dimensión. Sin embargo, pueden existir atributos que representen llaves foráneas de otras tablas de dimensión. Las tablas de dimensión se relacionan con las tablas de hechos haciendo parte de la llave de un hecho**,** por tanto**,** los atributos que conforman las tablas de dimensiones también describen el hecho [29].

2.1.3. Esquemas de representación dimensional

En un esquema de representación dimensional se muestran los hechos y las dimensiones que lo conforman, entre los esquemas de representación se encuentran:

Esquema estrella: Es un tipo de modelado utilizado para la representación relacional de una bodega de datos, el cual consta de una tabla de hechos central encargada de almacenar los datos para el análisis y una serie de dimensiones relacionadas a su alrededor [4]. Es ampliamente utilizado debido a que proporciona una mejor comprensión, navegabilidad y es más cercano a como el usuario final visualiza la consulta empresarial [28]. En la Figura 3 se muestra la ejemplificación de un esquema estrella.

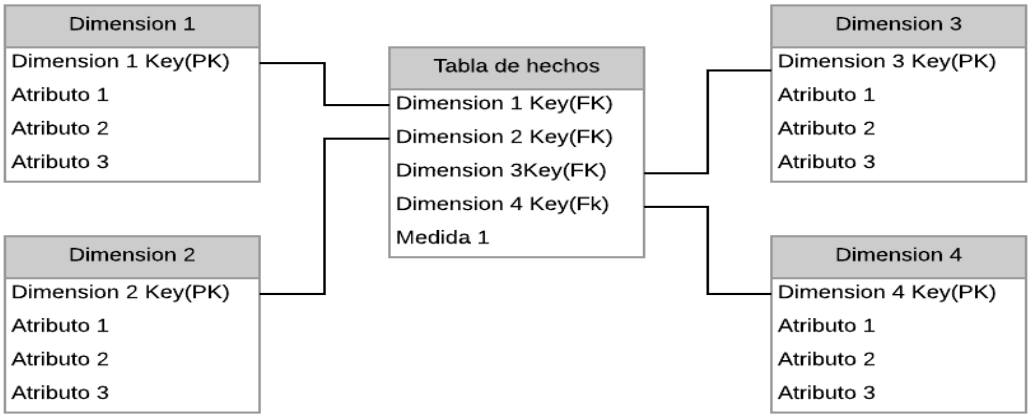


Figura 3 Esquema estrella.

Esquema de copo de nieve: Consta de una tabla de hechos que está conectada a muchas dimensiones, a diferencia del esquema estrella, estas pueden estar conectadas a otras dimensiones por medio de una relación muchos a uno. Las tablas de un esquema de copo de nieve generalmente cumplen la tercera forma normal. Cada tabla de dimensiones representa exactamente un nivel en una jerarquía. Por ejemplo, especificar la jerarquía fecha, a partir de las dimensiones: Año, mes y día [4]. Este esquema proporciona que se ocupe menos espacio de almacenamiento, pero eleva la cantidad de tablas con las que el usuario tiene que interactuar aumentando la complejidad de las consultas [28]. En la Figura 4 se muestra un esquema de copo de nieve con dos dimensiones, las cuales a su vez pertenecen a una jerarquía de tres niveles. Un esquema de copo de nieve puede tener varias dimensiones y cada dimensión puede tener varios niveles [4].

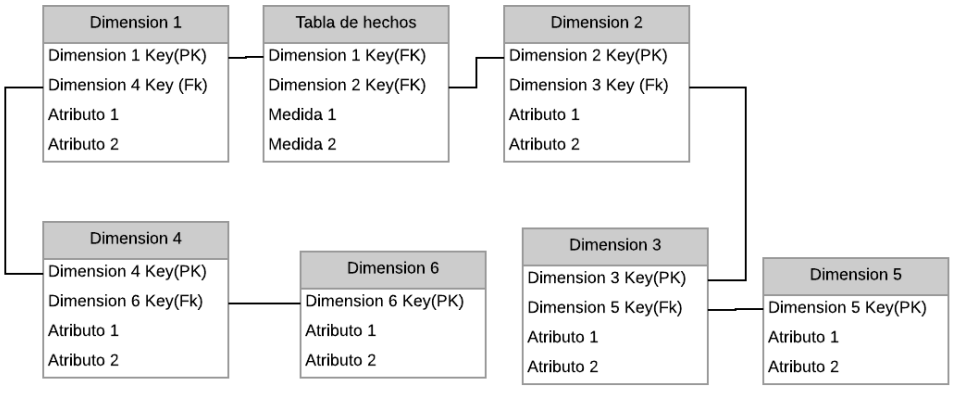


Figura 4 Esquema copo de nieve.

Esquema constelación: Se define como una combinación entre el esquema estrella y el copo de nieve, los esquemas constelación son esquemas copos de nieve en los cuales se genera un proceso de normalización en algunas de sus tablas más no en la totalidad de ellas [30]. En la Figura 5 se muestra un esquema constelación con una tabla de hechos relacionada a tres dimensiones, adicionalmente se incluye una cuarta dimensión la cual es una jerarquía dimensional compartida por la dimensión 1 y la dimensión 2.

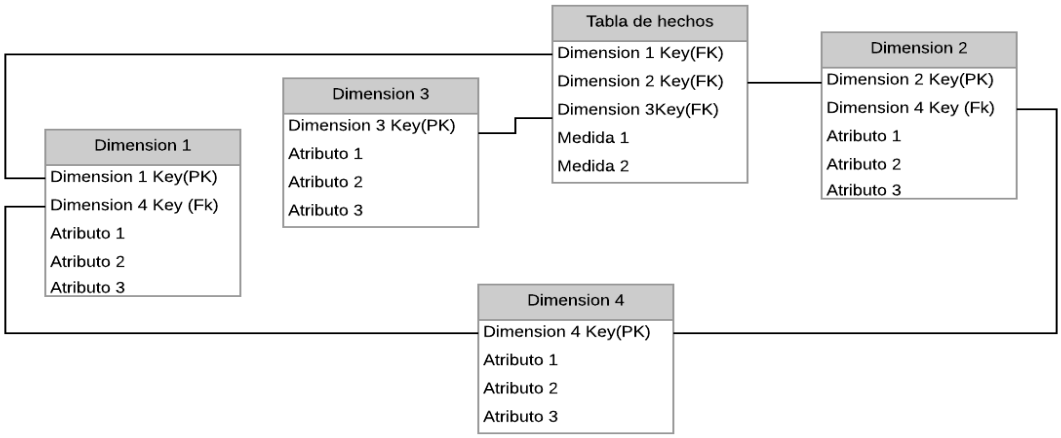


Figura 5 Esquema constelación.

2.1.4. Cubo dimensional

Permite visualizar desde diferentes perspectivas los datos que son de interés para los usuarios, estructurándolos de la siguiente manera (Ver Figura 6): los ejes son las dimensiones y en los cruces se encuentran los hechos o medidas a analizar. Es por medio de esta estructura dimensional que se puede dar soporte a las interrogantes que los usuarios tienen sobre los datos de su organización [27].

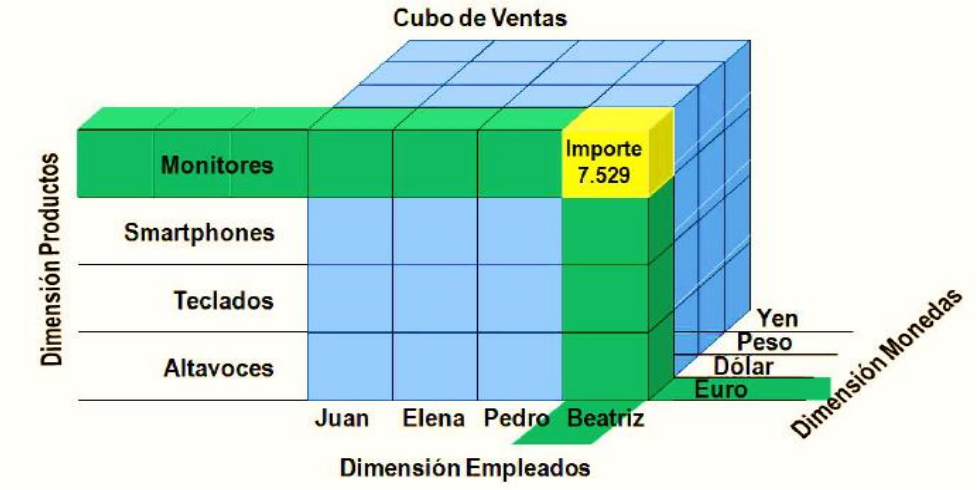


Figura 6 Cubo dimensional. Fuente: Tomado de [31]**.**

Los cubos dimensionales están enmarcados dentro de lo que se conoce como OLAP (*On-Line Analytical Processing*), definido por Ralph Kimball como la actividad general de realizar

consultas y presentar los datos almacenados en una bodega de datos [4]. Esta es clasificada según el tipo de almacenamiento [32]:

- **ROLAP:** El almacenamiento de los datos al igual que sus agregaciones[2] se lleva a cabo en motores relacionales, por lo cual el acceso a estos se da por medio de herramientas que generen SQL.
- **MOLAP:** En este caso los datos y sus agregaciones son almacenados en un formato multidimensional, por lo cual solo pueden ser accedidos por medio de MDX[3].
- **HOLAP:** Es un híbrido entre ROLAP y MOLAP, donde los datos se almacenan en bases de datos relacionales y las agregaciones en formato multidimensional.

2.1.4.1 Operaciones de cubo:

El cubo cuenta con algunas operaciones (Ver Figura 7) que permiten al usuario la mejor navegación a través de los datos, entre las cuales se destacan [4]:

- **Drill Down/Up:** Permite navegar desde los niveles de datos más resumidos (up) hasta los más detallados (down, en la figura un programa), por medio de la adición o eliminación de atributos (pertenecientes a una dimensión) en una consulta existente.
- **Slice & Dice:** Permite filtrar una consulta existente seleccionando un solo valor de una de sus dimensiones (slice, en la figura por el año 2015) o múltiples valores de varias dimensiones (dice, en la figura tres programas).

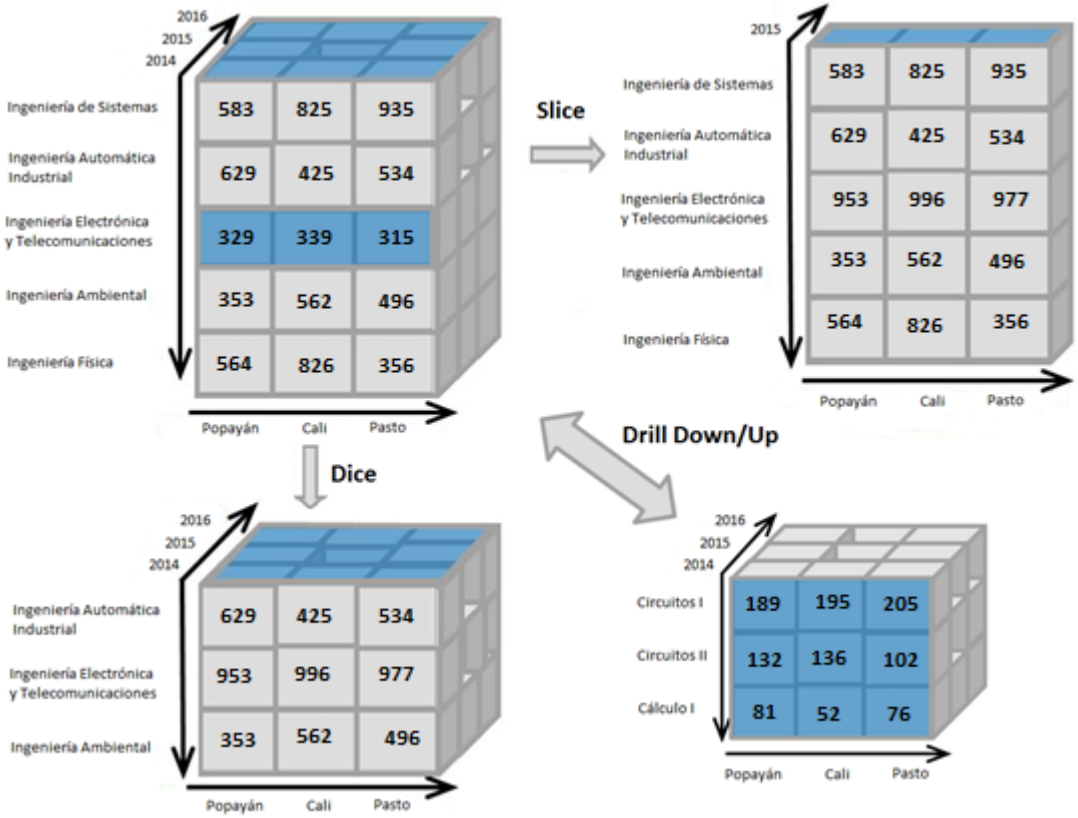


Figura 7 Operaciones sobre el cubo OLAP. Fuente: Adaptado de [33]**.**

[2] Resumenes de datos precalculados.
[3] Expresiones multidimensionales

2.1.5. Proceso ETL

Es el proceso por el cual se realiza la migración de los datos desde los sistemas transaccionales (OLTP[4]) hacia la bodega de datos relacional. Este proceso consiste de tres etapas dadas por sus siglas en inglés: Extracción (E), obtiene los datos de su ubicación de origen; Transformación (T), realiza una serie de validaciones para garantizar la calidad de los datos; y carga (L), almacena los datos en un conjunto final de tablas, en este caso las dimensiones y tablas de hechos de la bodega de datos [4].

2.2. Revisión sistemática de la lectura

En esta sección se muestra el proceso definido para la realización de la revisión sistemática del estado del arte actual y los resultados obtenidos. Para la realización de esta revisión se siguieron los lineamientos bases planteados por B. Kitchenham y S. Charters [34].

2.2.1. Preguntas de investigación

Las preguntas de investigación se plantearon de forma que permiten identificar las tendencias en cuanto a los proyectos de bodegas de datos en la educación superior. Las preguntas y su motivación se muestran en la Tabla 1.

ID	Pregunta de investigación	Motivación
PI1	¿Qué trabajos e iniciativas enfocadas al proceso del registro académico se han llevado a cabo con respecto al modelado dimensional de una bodega de datos en la educación superior?	Revisar el estado del arte actual acerca del tema.
PI2	¿Cuáles son los procesos de negocio que se han modelado de manera dimensional enmarcados en el ámbito del registro académico?	Reconocer los procesos de negocio que se han tenido en cuenta en las soluciones de bodegas de datos.
PI3	¿Qué elementos contemplados dentro de un modelo dimensional deben tenerse en cuenta para llevar a cabo la construcción de una bodega de datos?	Identificar los elementos recurrentes dentro de los modelos dimensionales.

Tabla 1 Preguntas de investigación y motivación.

2.2.2. Cadena de búsqueda

Con el fin de conocer los avances generados en la actualidad en cuanto a los modelos dimensionales enfocados en los procesos de registro académico, se diseñó una cadena de búsqueda basada en las recomendaciones definidas en [34] y usando términos claves relevantes para la investigación. Se realizó una diferenciación entre la cadena según el idioma de búsqueda, las cadenas de búsqueda en su estado final se muestran a continuación:

(Data warehouse *OR* Data warehousing *OR* Business Intelligence**) AND (**Multidimensional model *OR* Multidimensional modeling *OR* Dimensional model *OR* Dimensional modeling *OR* multidimensional design *OR* dimensional design**) AND (**Academic *OR* University *OR* Higher Education**)**

[4] On-Line Transactional Processing, sistemas de información orientados a los procesos transaccionales.

(Bodegas de datos *OR* Inteligencia de negocios**) AND (**Modelo dimensional *OR* Modelo multidimensional *OR* Diseño dimensional *OR* Diseño multidimensional**) AND (**Universidad *OR* Registro académico *OR* Registro de notas**)**

2.2.3. Recursos literarios

La cadena de búsqueda presentada en el apartado anterior fue utilizada en las siguientes fuentes de búsqueda:

- Science@Direct.
- IEEE Digital Library.
- Springer.
- Literatura gris.

2.2.4. Proceso de selección de estudios

El proceso de selección de los estudios candidatos se llevó a cabo teniendo en cuenta la relevancia de los trabajos basados en su título, resumen y conclusiones. Para la selección de los estudios más relevantes se usaron los criterios de inclusión (**CI**) y exclusión (**CE**) definidos a continuación:

2.2.4.1 Criterios de inclusión y exclusión

CI1: El trabajo se encuentra en un contexto del desarrollo de bodegas de datos para instituciones educativas de nivel superior.
CI2: El trabajo propone o menciona al menos un modelo dimensional enfocado en algún proceso de negocio enmarcado dentro del ámbito del registro académico.
CE1: El trabajo no debe haber sido publicado antes del año 2004.

2.2.5. Resultados y discusión

A continuación, se muestran los resultados obtenidos, análisis y discusión para cada una de las preguntas de investigación.

2.2.5.1 Clasificación de los trabajos relacionados

En esta sección se presenta la información extraída de los trabajos relacionados, además el análisis y discusión para cada pregunta de investigación.

PI1: Trabajos o iniciativas enfocadas al proceso del registro académico que se han llevado a cabo con respecto al modelado dimensional de una bodega de datos en la educación superior.

Con la aplicación de los criterios de inclusión y exclusión se obtuvo un total de 23 trabajos (artículos, monografías, etc.) seleccionados, que se encuentran relacionados con el desarrollo de bodegas de datos para el registro académico, estos fueron clasificados por año de publicación y por proceso de negocio.

2.2.5.2 Clasificación de trabajos por año de publicación

En la Figura 8 se presenta una clasificación de la totalidad de los trabajos relacionados por año de publicación, se evidencia un incremento en cuanto al número de trabajos realizados en los años 2011 y 2015. Vale la pena resaltar que la revisión sistemática se realizó inicialmente hasta mediados del año 2017, luego se hizo una actualización del estado del arte en el mes de junio de 2018, por esta razón se puede observar un trabajo publicado a finales del año 2017 como también un trabajo relacionado en este último año.

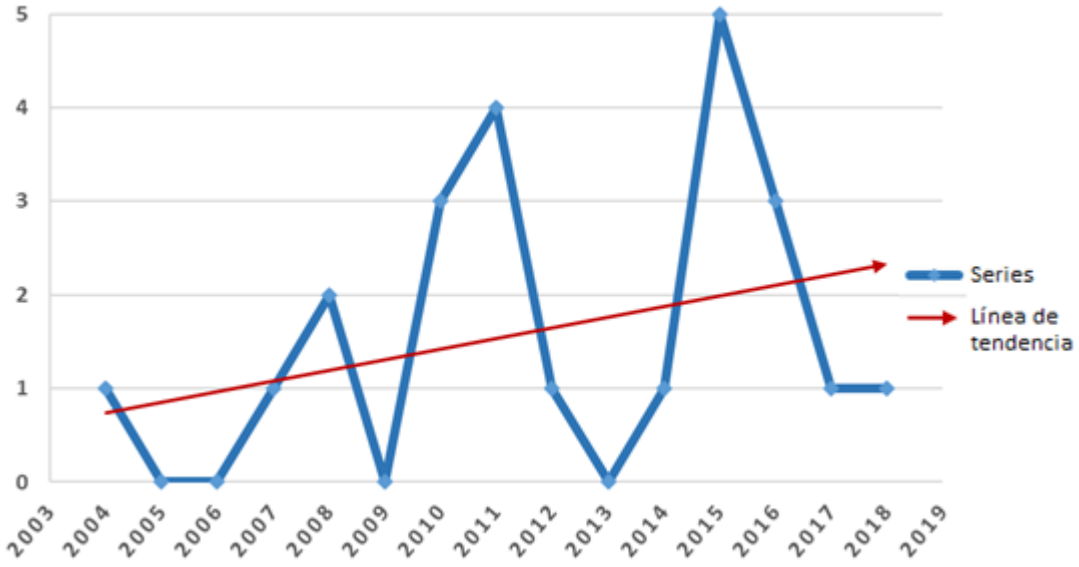


Figura 8 Número de trabajos relacionados publicados por año.

2.2.5.3 Clasificación de los estudios seleccionados por proceso de negocio

Teniendo en cuenta los trabajos seleccionados, se definieron los procesos de negocio y se clasificó cada uno de los trabajos según el proceso modelado, estos procesos son: Deserción, matrícula académica, matrícula financiera, desempeño académico (Registro de notas), retención, E-Learning, perfil estudiantil y otros. La Figura 9 muestra los resultados de esta clasificación.

PI2: Procesos de negocio que se han modelado de manera dimensional enmarcados en el ámbito del registro académico.

De los estudios seleccionados, el 56% (13 trabajos) hacen referencia al desempeño académico de los estudiantes, siendo este el tema de negocio más trabajado durante los últimos años. Los trabajos contemplan modelos enfocados en el estado final de sus evaluaciones (Aprobados-Reprobados) y en su gran mayoría a las notas obtenidas durante la vida académica del estudiante.

Figura 9 Clasificación de trabajos por proceso de negocio.

PI3: Elementos contemplados dentro de un modelo dimensional tenidos en cuenta para llevar a cabo la construcción de una bodega de datos.
Inicialmente se pueden especificar los tipos de esquema que fueron utilizados para llevar a cabo los modelos dimensionales dentro de los trabajos relacionados. Se puede apreciar que, el 91% (21 trabajos), realizó su modelo basado en el esquema estrella, mientras los 9% restantes se reparten equitativamente entre el esquema copo de nieve y el esquema constelación.

En la Tabla 2 se pueden observar aquellas dimensiones y medidas utilizadas en los trabajos relacionados, esto depende directamente del proceso de negocio como también de las fuentes de datos disponibles en el momento de la ejecución. En el apartado 2.2.5.4 se puede verificar específicamente cuales fueron las dimensiones y medidas en cada uno de los modelos dimensionales encontrados.

Dimensiones	Medidas
Dimensión relacionada con el tiempo. Dimensión relacionada con la carrera. Dimensión relacionada con la asignatura. Dimensión relacionada con el curso. Dimensión relacionada con la localización del estudiante. Dimensión relacionada con datos personales. Dimensión relacionada con el docente. Dimensión relacionada con datos demográficos. Dimensión relacionada con la economía del estudiante.	Notas parciales. Notas finales. Horas estudiadas. Asignaturas aprobadas. Asignaturas reprobadas. Asignaturas cursadas. Cantidad de alumnos. Asignaturas restantes. Créditos aprobados. Cuotas no cubiertas. Porcentaje de avance.

Tabla 2 Elementos recurrentes en los trabajos relacionados.

2.2.5.4 Detalles de trabajos relacionados
En la literatura se encuentran trabajos que proponen una bodega de datos para procesos relacionados con el registro académico. A continuación, se mencionan los trabajos más relevantes.

En el año 2004 [5] se propone un trabajo, el cual fue desarrollado en una universidad pública colombiana y estaba orientado al análisis, diseño e implementación de una bodega de datos para los procesos de planeación de dicha universidad. Este trabajo cuenta con varias tablas de hechos, pero solo dos están relacionadas con el objeto de esta investigación. *Deserción* con las dimensiones: tiempo, carrera y motivo, y las medidas cantidad, cantidadm, cantidadh y último semestre. *Matrícula financiera* con las dimensiones: tiempo, carrera y estado, y las medidas cantidad y porcentaje. Estas tablas de hechos cuentan con pocas dimensiones y no contemplan datos del estudiante, lo cual restringe el nivel de detalle de las consultas analíticas. Además, no presentan una descripción detallada de las dimensiones y medidas de los esquemas estrella. A la bodega de datos se le realizaron pruebas de funcionamiento en diferentes equipos hardware y plataformas software, y una etapa de prueba con usuarios de la oficina de planeación (No se menciona evaluación de satisfacción ni aceptación del usuario), definiendo que se debía agregar un mayor nivel de detalle a las consultas y mejorar la visualización de las mismas.

En el año 2008 [6], se desarrolló un sistema de Inteligencia de Negocios (BI por sus siglas en inglés) basado en la tecnología de bodegas de datos que utiliza herramientas del proveedor SAS, proponiendo este sistema para el proceso de negocio de matrículas de estudiantes. En la fase de diseño, plantean el modelo dimensional para la tabla de hechos *Registro Estudiantes* con las dimensiones: calendario, género, programa, carrera, categoría, grado, nacionalidad y sesión, además de la medida total estudiantes. Este modelo no contempla una dimensión estudiante y solo un tipo de información socio-demográfica (género), lo cual restringe la capacidad de análisis de los datos almacenados en la bodega de datos. Además, no presenta una explicación detallada de las dimensiones relacionadas en este modelo dimensional. Este trabajo fue implementado, especificando los procesos ETL y mostrando gráficamente algunos de los reportes desplegados en una aplicación web. Sin embargo, no se menciona la realización de algún tipo de prueba a la solución implementada, ni evaluación de los usuarios.

Una ontología[5] fue propuesta en el año 2010 [7], la cual identifica los factores más influyentes en el comportamiento académico estudiantil. La representación la realizan en un diagrama de clases, que luego es mapeado a un esquema estrella de una bodega de datos. Este esquema se enfoca en el registro de notas por medio de la tabla de hechos *Estudio* y las dimensiones: tiempo, curso, escuela, grupo, localización, programa de estudio, personal académico y estudiante. Además medidas como: nota, examen, tiempo gastado e inclusión. Se obtuvo como resultado la validación de la ontología, donde se demostró la necesidad de ésta en algunas de las universidades europeas. Se adicionaron datos como la escuela de origen, los resultados de exámenes anteriores, el tiempo trascurrido entre los exámenes, entre otros, para determinar los factores influyentes en las fallas del estudiante. A pesar de que se cuenta con una descripción general de la ontología, no se presenta información en detalle sobre los componentes del esquema estrella y la bodega de datos no llegó a ser implementada.

Con el objetivo de mejorar el análisis del rendimiento académico y generar pronósticos sobre este, en el año 2011 [8], unen las bodegas de datos con las redes neuronales artificiales. Para la creación de la bodega de datos siguieron los procesos de modelamiento conceptual (especificación de requerimientos y disponibilidad de los datos), modelado lógico e implementación física (diseño de la bodega de datos, definición del motor relacional) y la carga de los datos. Luego, la red neuronal artificial realiza la predicción del rendimiento de los estudiantes en el siguiente semestre teniendo como datos de entrada: puntaje PSU (Puntaje de selección universitaria) en lenguaje, en matemáticas, en historia, puntaje PSU de ingreso, semestre inicial, cantidad de asignaturas inscritas al comienzo del semestre y el número de asignaturas aprobadas en el mismo. En este artículo se presenta un esquema estrella conceptual sobre el registro de notas involucrando la tabla de hechos *Indicadores Alumnos*, sin hacer una descripción detallada de las dimensiones: alumno, asignatura, tiempo de avance (años de estudio), cohorte (semestre actual), oportunidad (obtenida de la red neuronal), zona geográfica y región; ni de las medidas: suma de carga académica, carga promedio, notas promedio, suma de aprobados, suma de reprobados, cantidad de alumnos y cantidad de asignaturas. Por lo anterior, el modelo queda a la libre interpretación del lector, lo que genera confusión a nivel general del diseño de la tabla de hechos. El trabajo se enfoca en su mayoría en la utilización de redes neuronales para procesos de minería de datos, la bodega de datos fue utilizada como una fuente de datos

[5] Representación de entidades, junto con sus propiedades y relaciones.

para la arquitectura de la red neuronal, por lo cual no se hace mención de resultados (beneficios, pruebas, etc.) propios de la bodega.

En ese mismo año se propone una bodega de datos [9], para el análisis del registro de notas usando técnicas de minería de datos para realizar la proyección de los estudiantes que aprobarán los cursos actuales. Este artículo presenta la estructura del sistema OLTP y del esquema estrella, el cual cuenta con la tabla de hechos *Clase* y las dimensiones: facultad, curso, horario, estudiante y salón de clase; además de las medidas: calificación, registrado y rango. Aunque se muestran los datos almacenados en cada una de las dimensiones y la tabla de hechos, no se cuenta con un detalle específico del esquema estrella, presenta pocas dimensiones y el número de atributos de éstas es pequeño. La bodega de datos fue implementada, permitiendo realizar análisis de los datos históricos almacenados. Además, por medio del modelo de minería de datos se generaron informes predictivos, pero el trabajo no especificó pruebas realizadas sobre el sistema, ni tampoco la aceptación o evaluación por parte de los usuarios finales.

También en el año 2011 [10], presentan un marco de trabajo para el desarrollo de un sistema de BI en las universidades (mencionan modelos de madurez de BI), planteando la necesidad de que las universidades tengan este tipo de repositorios para proveer un origen centralizado de la información que ayude a las estructuras organizacionales y administrativas de estas instituciones a suministrar los datos necesarios para realizar diferentes tipos de reportes, analizar las situaciones académicas basadas en datos y desarrollar planes estratégicos que ayuden a la mejora de estas instituciones. Este artículo presenta un modelo dimensional realizado bajo el esquema constelación, enfocado en la evaluación del uso de una plataforma e-learning, con tres tablas de hechos: *Utilización*, la cual cuenta con las dimensiones tiempo y curso, y las medidas número de estudiantes registrados y número de estudiantes activos. *Actividad,* con las dimensiones herramienta, persona, tiempo y curso. Y las medidas tipo de acción y duración. Y *Grado,* con las dimensiones persona y curso, y las medidas grado y terminado. En el artículo se menciona una descripción muy simple de los elementos que contempla el modelo dimensional, además tiene un escaso número de dimensiones lo cual indica una limitación en la capacidad de consulta de la bodega de datos. Finalmente, la bodega de datos fue implementada teniendo en cuenta la información que ofrece la plataforma Moodle y presenta un ejemplo de tableros de mando (dashboard) que se pueden generar para facilitar la visualización de la información. Sin embargo, no se describe ningún tipo de evaluación por parte de los usuarios finales.

Nuevamente en el año 2011 se propuso un trabajo [11] que consistió en crear un sitio web para la gestión de material didáctico que ayude a los diseñadores principiantes de bodegas de datos a reforzar los conceptos clave en este tema haciendo uso de un estudio de caso. El cual se basó en la creación de una bodega de datos para el proceso de negocio de inscripciones y matrículas de los estudiantes. Además, esta bodega de datos se creó para generar información de entrada para un sistema de minería de datos. Se presenta un modelo compuesto por dos tablas de hechos, *Matrícula*, relacionado con las dimensiones estudiante, tiempo y universidad, y la medida de matrícula; y *Predicción*, relacionado con las dimensiones estudiante, tiempo, universidad y socio-económica, además de la medida predicción. Este artículo no presenta detalle de la información relacionada con las dimensiones y solo describe en que consiste una de las medidas propuestas en una de las tablas de hechos. Además, el modelo presenta muy pocas dimensiones y la información

socio-económica presentada es muy poca, lo que limita el poder de análisis de la bodega de datos. Por último, la bodega de datos del estudio de caso no fue implementada.

En el año 2012 se llevó a cabo una consolidación de la información [12] de las fuentes de datos académicas de la Universidad Autónoma de Manizales, por medio de una bodega de datos, que en conjunto con redes neuronales y árboles de decisiones, permitiera generar conocimiento con respecto a las variables influyentes en el proceso de deserción estudiantil. Se modelaron tres esquemas estrella con el fin de observar desde diferentes perspectivas la realidad de los estudiantes; el primero de ellos enfocado en el rendimiento académico con la tabla de hechos *Estudiante Rendimiento* tiene como dimensiones relacionadas el estudiante, semestre y el programa. Como medidas: cantidad créditos inscritos, cantidad créditos aprobados, promedio académico. Por otro lado, el esquema de estados académicos relacionado con la tabla de hechos *Estudiante Estado* con las dimensiones estudiante y programa, teniendo como medida el estado y como información adicional en la tabla de hechos el semestre ingreso y el semestre retiro como dimensiones degeneradas. Finalmente, el tercer modelo enfocado en los aspectos sociales y de procedencia con la tabla de hechos *Estudiante Estado* y una posible tabla puente Estudiante Colegio, además de las dimensiones ya mencionadas, la dimensión colegio y la dimensión departamento, se puede observar que en los últimos dos modelos la tabla de hechos tiene el mismo nombre (*Estudiante Estado)*, aunque el trabajo no especifica por qué se comparte esta tabla de hechos. Estos modelos cuentan con muy pocas dimensiones, no especifican si la tabla Estudiante Colegio es una tabla puente, ni tampoco detallan los elementos que componen el modelo dimensional, dejando a libre interpretación el significado de sus componentes. Al final obtienen como resultado la construcción de la bodega, los procesos ETL y encuentran variables en la deserción estudiantil, especifican y realizan pruebas de diferentes modelos de minería. Sin embargo, no se hicieron pruebas sobre la bodega de datos, ni evaluación de los usuarios finales.

En el año 2014, se propone para la recolección de los requerimientos analíticos de una bodega de datos, hibridar el enfoque orientado a los datos con el enfoque orientado a los requerimientos [13], para así solventar las problemáticas presentadas en la integración de diferentes fuentes de datos. Los temas de negocio diseñados fueron las publicaciones de las investigaciones de los docentes y el registro académico de los estudiantes (didáctica) dentro de una institución universitaria. Los temas incluidos en el registro académico fueron: matrícula académica y financiera, notas y labor docente. En este trabajo se representa la bodega de datos por medio de un esquema copo de nieve, que contiene seis tablas de hecho, de las cuales solo tres se ajustan a los procesos de negocio de interés, a saber: *Matrícula*, con las dimensiones curso, estudiante, fecha, tipo, ciudad y sin medidas; *Examen*, que incluye las dimensiones estudiante, curso, fecha, curso de enseñanza y con las medidas calificación y cumlaude; *Distribución*, con las dimensiones curso de enseñanza, profesor y las medidas ufc (créditos formativos universitarios), horas, modo lección, modo evaluación, contrato y créditos. Este artículo no incluye una descripción detallada de las dimensiones y medidas del esquema, presenta fallos en el diseño del esquema (existe una tabla de hechos idéntica a la tabla examen) y un error de integridad referencial en la tabla de hechos distribución la cual tiene gráficamente relación con tres dimensiones, pero dentro de ésta solo se ubican dos llaves foráneas, lo cual no permite entender la totalidad del esquema. El producto final es una aplicación web que permite a los usuarios generar análisis sobre la información de la bodega de datos, haciendo mención de los usuarios y tareas que cada uno puede realizar sobre el sistema y presentando ejemplos gráficos de

las consultas ejecutadas. A pesar de ello, no nombra ningún tipo de prueba realizada por los usuarios para validar la utilidad bodega de datos.

Con el objetivo de reconocer y mejorar el Rendimiento académico de los estudiantes, se propuso en el año 2015 [14] realizar un modelo de bodegas de datos que toma como caso de estudio los estudiantes de Ingeniería de Sistemas de la Universidad Tecnológica Nacional, para determinar las características del estudiantado que influyen positiva o negativamente en su desempeño universitario. En este trabajo se presenta un esquema estrella que consta de una tabla de hechos *Alumnos* y nueve tablas de dimensiones: situación laboral (De la madre, del padre y del estudiante), residencia, procedencia, estudios secundarios, dedicación al estudio, importancia al estudio, importancia a las TICs, como medidas se consideran primeraNota, segundaNota, notaExtraordinario, situaciónFinal (aunque se menciona que las notas son obtenidas por asignaturas, en el modelo no se presenta una dimensión de este tipo ni un atributo que se relacione). El trabajo presenta las encuestas usadas para la recolección de los datos, ya que la totalidad de las dimensiones fueron cargadas por medio de estas. Finalmente discute como trabajo futuro la inclusión de modelos de minería que ayuden a determinar la tendencia de un estudiante al fracaso académico. Este trabajo no fue implementado y solo contempla la etapa de diseño de la bodega de datos y el proceso ETL.

En el mismo año 2015 [15] se llevó a cabo en la Universidad Nacional de Piura la implementación de una bodega de datos para evidenciar las problemáticas académicas de los estudiantes de la facultad de ingeniería industrial, para esto crean un modelo dimensional enfocado en las notas finales de los cursos vistos en los ciclos correspondientes de su carrera, con la tabla de hechos *HistoriaF* que presenta las dimensiones: cursos, alumnos, docentes, ciclos, notas y areaCurso. Las medidas: créditos, notas, aprobado, desaprobado, retirado, total créditos aprobados, total créditos reprobados y total créditos retirados. El trabajo detalla levemente los componentes que hacen parte del modelo, cuenta con muy pocas dimensiones y dentro de estas no incluyen un número significativo de atributos que incrementen el potencial analítico del trabajo realizado. Finalmente, se exponen los componentes ETL y una cantidad de reportes matriciales para determinar el nivel de deserción estudiantil, pero no se realizan pruebas sobre la bodega de datos ni evaluación por parte de los usuarios.

También se realizó otro trabajo en el año 2015 [16] con el objetivo de diseñar y generar un modelo de minería de datos para la identificación de patrones de comportamiento relacionados con el desempeño académico de los alumnos pertenecientes a una institución de educación media superior. También se diseñó e implementó una bodega de datos que fue poblada a través de un proceso ETL con múltiples fuentes. El diseño dimensional de la bodega presenta en un esquema estrella el cual cuenta con la tabla de hechos denominada *Dimensión Hechos* con las dimensiones: persona, socioeconómica, institución, académica y tiempo. Las medidas promedio, inteligencia emocional, coeficiente intelectual, percepción de la calidad de los servicios académicos, nivel socioeconómico y cuotas no cubiertas. En la implementación de la bodega se crearon modelos de minería de red neuronal artificial, árboles de decisión y el modelo de agrupamiento, con los cuales se determinaron variables que influyen en el desempeño académico y que pueden llevar al estudiante a la deserción, encontrando que la inteligencia emocional es uno de los aspectos determinantes en el proceso del estudiante. El trabajo está enfocado principalmente en los modelos de minería y las pruebas realizadas sobre ellos, no especifica el origen de las medidas, no cuenta con

una descripción detallada de los componentes del modelo dimensional ni tampoco con una evaluación de la bodega de datos construida.

Nuevamente en el año 2015, se propuso una bodega de datos para el análisis de diferentes factores que están involucrados en la retención estudiantil en la Universidad de Vermont [17], creando un modelo que permitiera realizar análisis para determinar los aspectos más influyentes en la retención estudiantil y de esta manera generar estrategias o planes para aumentar las tasas de retención estudiantil. El estudio se basó en las tasas de retención en los estudiantes de primer año en varias universidades de Estados Unidos, mencionando la identificación de los orígenes de información estudiantil y el proceso de limpieza de estos. En la etapa de diseño, realizaron un modelo dimensional por medio de un esquema estrella de una tabla de hechos denominada *Hecho* con las dimensiones estudiante, salón y semestre. Además de una gran cantidad de medidas entre las que se encuentran el número de As (calificación más alta), Bs, Cs, número de citaciones por vandalismo, citaciones por drogas, entre otros. Como se ha mencionado en otros trabajos, el número de dimensiones restringe la capacidad de consulta de las bodegas de datos. Por otro lado, aunque se especifican los componentes que hacen parte del modelo, no se cuenta con datos personales ni socio-demográficos de los estudiantes, lo cual es importante para la realización de mejores análisis. Finalmente la bodega de datos no fue implementada en la institución.

En el año 2016 se implementó una solución de bodega de datos [18] en la Corporación Universitaria del Caribe con el fin de analizar el fenómeno de la deserción estudiantil, teniendo como soporte el sistema de información transaccional de la institución que se encarga de los procesos académicos – administrativos. Proponiendo dos modelos dimensionales, uno enfocado en los registros de *Notas* y el otro en la *Deserción* estudiantil, en total se identificaron diez dimensiones: docente, tiempo, asignatura, estudiante, estadoCivil, programa, estrato, lugarResidencia, promedioEstudiantil y el sexo. Como medidas se ubican las notas de los diferentes cortes, la nota definitiva, asignaturas aprobadas, asignaturas cursadas. Para el modelo de deserción cuenta con las medidas matriculado, desertorPotencial y desertor. Este trabajo no incluye descripciones detalladas de los componentes de los esquemas estrellas y presenta medidas de diferente granularidad de la tabla de hechos (asignaturas aprobadas, asignaturas cursadas). La solución fue desarrollada junto con modelos de minería de datos y determinó que las variables más influyentes en el proceso de la deserción son el promedio estudiantil y el número de asignaturas cursadas por semestre. Pero no presenta pruebas de la solución ni evaluación por parte de los usuarios.

Para el año 2017 se genera una estrategia de inteligencia de negocios [19], en la cual se presenta un sistema de bodega de datos alimentado por la información académica de la Universidad Nacional y la información recolectada por el Sistema de Información de Bienestar Universitario, para permitir a la institución realizar análisis de los factores económicos, sociales, culturales y de sus incidencias con el desempeño académico del estudiante. Adicionalmente se generó la construcción de un modelo de minería de datos con el objetivo de generar predicciones de deserción académica. Dentro del modelado dimensional propuesto se encuentran dos procesos de negocio, el académico encargado del desempeño del estudiante y por otro lado el proceso de activación de servicios que contiene la información de los estudiantes que son activados en algún servicio de bienestar universitario. Estos dos procesos finalmente son resumidos en una sola tabla de hechos

denominada *Académico* la cual contiene las dimensiones estudiante, estados estudiante, tiempo, carrera, convocatoria y servicios bienestar, contando con las medidas créditos aprobados, créditos restantes, porcentaje avance, materias aprobadas, materias no aprobadas, puntaje básico de matrícula (PBM), tendencia PAPA (Promedio Académico Ponderado Acumulado), cantidad períodos, cantidad, rango PBM, rango avance y rango PAPA. Adicionalmente, el trabajo cuenta con una evaluación del cubo OLAP con los directivos del bienestar universitario. Aunque no se especifica cómo se realizó la evaluación, se mencionan comentarios positivos, sin embargo, este modelo no contempla datos socio-demográficos del estudiante y el número de dimensiones utilizadas podría ser mayor para incrementar los beneficios analíticos.

En el año 2018 se presenta el diseño de una bodega de datos [20] que permite almacenar la caracterización de los estudiantes activos como también graduados del Tecnológico de Antioquia – Institución Universitaria, haciendo un reconocimiento de las variables que pueden incidir en la permanencia o deserción del estudiantado por medio de la tabla de hechos *Caracterización* y cinco dimensiones: Tiempo, individuales, institucionales, socioeconómicas, académicas. El modelo presentado no específica medidas de la tabla de hechos. Finalmente exponen un modelo de minería de datos con el fin de potenciar la utilidad de la bodega, generando predicciones con respecto a las variables de mayor incidencia en la obtención del título académico por parte del estudiante (graduación). Aunque el modelo contempla información relevante para la resolución de necesidades analíticas, incluyendo datos socio-demográficos, no cuenta con una gran cantidad de dimensiones, lo cual limita el poder análisis de la bodega de datos. Se presentaron resultados de la bodega y la minería de datos por medio de reportes, los posibles análisis que se podrían hacer sobre estos y las predicciones llevadas a cabo con los modelos de minería, sin embargo, no se especifica un proceso de pruebas ni evaluaciones por parte de los usuarios.

En la Tabla 3, se presenta una comparación de los trabajos relacionados, mostrando el nivel de detalle de cada uno de estos. La columna *Ref.*, indica la referencia bibliográfica del trabajo relacionado, en la columna *Tipo Esq.*, se especifica el tipo de representación del modelo dimensional (Estrella o copo de nieve). El proceso de negocio que se modela en la tabla de hechos se indica en la columna *Proceso de negocio*, en la columna *Incons. Diseño*, se indica si la representación tiene inconsistencias de diseño o cuenta con pocas dimensiones; en las columnas *Det. D.* y *Det. M.*, se determina si se cuenta con una especificación detallada de las dimensiones y las medidas; incluyendo también la columna *Datos SD. Est.,* que indica si fueron tenidos en cuenta datos socio-demográficos; la columna *Impl.,* indica si la bodega de datos fue desarrollada y desplegada en la respectiva institución universitaria, el indicador /Minería resalta que ese fue el enfoque prioritario del trabajo; finalmente *Eval. con Usu.*, indica si el trabajo presenta una evaluación (con los usuarios del negocio) de la bodega de datos propuesta.

Ref.	Tipo Esq.	Proceso de negocio	Incons. Diseño	Det. D.	Det. M.	Datos SD Est.	Impl.	Eval. con Usu.
[5]	Esquema estrella	Deserción, Matrícula financiera	Si (Pocas dimensiones)	No	No	No	Si	No

[6]	Esquema estrella	Matrícula académica	No presenta información estudiante	No	No	No	Si	No
[7]	Esquema estrella	Registro de notas	No	No	No	No	No	No
[8]	Esquema estrella	Registro de notas	No	No	No	No	Si / Minería	No
[9]	Esquema estrella	Registro de notas	No	No	No	No	Si / Minería	No
[10]	Esquema constelación	E-Learning	Si (Pocas dimensiones)	Si	Si	No	Si	No
[11]	Esquema estrella	Inscripciones, Matrículas académicas	Si (Pocas dimensiones)	No	No	Si	No	No
[12]	Esquema estrella	Registro de notas, Deserción	Si (Pocas dimensiones)	No	No	No	Si	No
[13]	Copo de Nieve	Matrícula académica, Matrícula financiera, Registro de notas, Labor docente	Si (Diseño)	No	No	No	Si	No
[14]	Esquema estrella	Registro de notas	No	Si	Si	No	No	No
[15]	Esquema estrella	Registro de notas	Si (Pocas dimensiones)	Si	Si	No	Si	No
[16]	Esquema estrella	Registro de notas	No	No	No	Si	Si / Minería	No
[17]	Esquema estrella	Retención	Si (Pocas dimensiones)	Si	Si	No	No	No
[18]	Esquema estrella	Registro de notas, Deserción	No	No	No	No	Si / Minería	No
[19]	Esquema estrella	Registro de notas, Deserción	Si (Pocas dimensiones)	Si	Si	No	Si / Minería	Si
[20]	Esquema estrella	Deserción	Si (Pocas dimensiones)	No	No	Si	Si / Minería	No

Tabla 3 Comparación de trabajos relacionados.

Como se puede apreciar en la Tabla 3, en la última década se han llevado a cabo trabajos relacionados con las bodegas de datos de registro académico en el ámbito universitario. En la mayoría, no presentan descripciones detalladas de los esquemas, dejando un nivel de subjetividad en la interpretación de las dimensiones y medidas involucradas, lo cual hace muy difícil adquirir conocimiento de estos trabajos y apropiarlo en otras instituciones educativas. Aunque los trabajos [10], [15], [17] y [19] cuentan con descripciones de dimensiones y medidas, presentan inconsistencias de diseño por la baja cantidad de dimensiones involucradas [4], lo cual limita el poder de análisis de este tipo de soluciones. Por otro lado se pueden encontrar trabajos como el [14], que no presentan este tipo de problemáticas, sin embargo, carecen de datos socio-demográficos del estudiante, lo cual impide que los usuarios puedan realizar análisis con respecto a este tipo de información.

2.3. Aportes

Este trabajo de grado genera conocimiento para la comunidad académica con respecto al modelo dimensional y los casos de diseño presentados en una bodega de datos para el registro académico (registro de notas y control de asistencia) que incluye información socio-demográfica en el ámbito de una universidad pública. Este modelo se podrá tomar como base por otras universidades públicas colombianas, teniendo en cuenta la similitud en los datos que manejan este tipo de instituciones.

A nivel de aplicación, la FIET cuenta con un prototipo de bodega de datos enfocado en el registro de notas de los estudiantes de la FIET, permitiendo a los usuarios finales realizar consultas analíticas para desarrollar a futuro nuevas estrategias que mejoren el rendimiento y la retención de los estudiantes.

Capítulo III. Metodología seleccionada

Para realizar la ejecución de este trabajo, se utilizó la **Metodología de desarrollo de bodega de datos para empresas MiPymes** (MBD) [35], esta metodología es dirigida a equipos de desarrollo pequeños que cuentan con conocimientos básicos sobre el desarrollo de las bodegas de datos pero sin experiencia práctica en la construcción de dichos sistemas, por lo cual se adapta muy bien a las necesidades de este proyecto, por otra parte cuenta con características particulares que facilitan el proceso de desarrollo en el tiempo planteado, considerando una cantidad reducida de artefactos y presentando una definición detallada de las actividades involucradas.

3.1. Ciclo de vida

La metodología MBD tiene en cuenta aspectos tanto de la metodología propuesta por Ralph Kimball como también los encontrados en la metodología Ágil método de desarrollo de sistemas dinámicos para bodegas de datos (DSDM DW), por lo cual presenta un enfoque iterativo e incremental con una activa participación del cliente, sin dejar de lado las actividades estructuradas propuestas por Kimball para la correcta ejecución de un proyecto de este tipo.

Las fases que componen la metodología MBD son las observadas en la Figura 10: Iniciación, Planeación, Análisis, Desarrollo, Mantenimiento y Crecimiento, y Gestión del Proyecto.

Figura 10 Ciclo de vida. Tomado de [35]**.**

A continuación, se genera una descripción de cada una de las fases pertenecientes al ciclo de vida, detallando el trabajo llevado a cabo durante cada una de estas.

3.2. Fase de iniciación

El objetivo principal de esta fase es identificar los procesos de negocio con mayor impacto y viabilidad para la empresa con el fin de establecer el de mayor favorabilidad para iniciar el proyecto. Aunque la metodología involucra al cliente, ya que este es quien tiene el mayor conocimiento acerca del negocio, debido a la naturaleza académica del proyecto, durante

la elaboración del anteproyecto de trabajo de grado se seleccionó como prioritario el proceso de Registro académico que incluye el registro de notas y el control de asistencia. Sin embargo, durante esta fase se identificaron otros procesos de negocio cuyo análisis se presenta en el Anexo A.

3.3. Fase de planeación

Al igual que la fase de iniciación, esta fase fue realizada durante la elaboración del anteproyecto, por lo cual no se tuvo una activa participación del cliente. Esta fase desarrollada una vez por cada uno de los procesos de negocio seleccionados (Registro académico), tiene como prioridad definir los objetivos y los límites del proyecto, organizar la forma en la cual se llevarán a cabo las tareas y definir a los responsables de ejecutarlas (Especificación de roles). Los artefactos relacionados con esta fase se encuentran en el Anexo B.

3.4. Fase de análisis y diseño

Esta fase tiene como finalidad la identificación de los requerimientos del proceso de negocio seleccionado, para posteriormente con base en ellos producir el modelo dimensional y definir la forma en la que se cargarán los datos en la bodega. Esta fase se divide en dos sub-fases: Recolección de requerimientos y Diseño.

3.4.1. Sub-fase de recolección de requerimientos

En esta sub-fase se hace necesaria la participación del cliente, ya que es él quien conoce las reglas de negocio y facilita al equipo la comprensión de los requerimientos comunicados por los usuarios entrevistados para así priorizarlos de forma correcta.

En este sentido, teniendo en cuenta el conocimiento con respecto al proceso de registro académico y la disponibilidad de tiempo con la que contaba, se decidió trabajar con el Coordinador del programa de Ingeniería de Sistemas como el cliente representante de los usuarios. Por otro lado, se hizo necesaria la elección de usuarios que tuvieran conocimiento del proceso de negocio seleccionado, como también usuarios cercanos al sistema de información de la Universidad, los usuarios que participaron del proyecto son:

Analista-Administrativo:

- Coordinador del programa de Ingeniería de Sistemas.
- Técnico administrativo – Vinculado a la Decanatura de la Facultad de Ingeniería Electrónica y Telecomunicaciones.
- Decano de la facultad de Ingeniería Electrónica y Telecomunicaciones.
- Coordinadora del programa de Ingeniería Automática Industrial – Vinculada al proceso de acreditación institucional.
- Docente del programa de Ingeniería de Sistemas – Vinculado a la división de tecnología.

Encargados del sistema de información:

- Ingenieros vinculados a la división de tecnología – Desarrolladores SIMCA.

Esta fase fue ejecutada a finales del año 2017, por lo cual es posible que los roles mencionados anteriormente sean ocupados en la actualidad por otros funcionarios.

Una vez contactados los usuarios, se llevó a cabo la creación de los cuestionarios, definiendo un cuestionario diferente para aquellos usuarios ubicados en un nivel gerencial,

otro para los que desempeñan un rol analítico-administrativo y finalmente un cuestionario para los usuarios encargados del sistema de información. Por cuestiones de disponibilidad no fue posible realizar entrevistas a ningún usuario con nivel gerencial como el vicerrector académico de la universidad, ya que cuentan con un alto nivel de ocupación, lo cual dificultó una posible reunión.

Posteriormente se realizaron las entrevistas, con las cuales fue posible reconocer los requerimientos analíticos relevantes, los cuales fueron resumidos para finalmente ser priorizados en conjunto con el representante de los usuarios, en donde se asignaron valores entre alto, medio o bajo para representar la importancia del requerimiento. Adicionalmente el equipo de desarrollo decidió incluir una priorización por viabilidad del requerimiento, esta indica la existencia digital de los datos necesarios para el cumplimiento de un requerimiento en particular.

A continuación, se muestran en la Tabla 4 las solicitudes de análisis con mayor prioridad y viabilidad, el resto de la lista priorizada, junto con los diferentes cuestionarios pueden ser encontrados en el Anexo C.

Proceso de negocio	Tema analítico	Solicitud de análisis	Comentario	Prioridad	Viabilidad
Registro académico	Desempeño académico	Estado del estudiante por materia	Cantidad de aprobados o reprobados, junto con la repetición en la que se cursaba (Ej. 25 Aprobado – R0)	Alto	Alto
		Nivel del estudiante por materia	Promedio de nota que el estudiante obtuvo, con una etiqueta que lo agrupe por rango (Ej. Alto – 4,8)	Alto	Alto
	Mortalidad académica	Reprobados por materia	Cantidad de estudiantes que reprobaron con respecto a la cantidad de estudiantes que cursaron (Ej. 10/30)	Alto	Alto
		Estudiantes con varias materias reprobadas	Estudiantes que han reprobado un cierto número de materias en el período	Alto	Alto

Tabla 4 Priorización de los requerimientos.

3.4.2. Sub-fase de diseño

En esta sub-fase se crearon los modelos dimensionales necesarios para satisfacer los requerimientos identificados en la sub-fase anterior, todo lo relacionado al proceso de modelado se especifica en el Capítulo IV.

Por otro lado, esta sub-fase involucra la creación del diseño físico de la bodega de datos relacional, como también un mapeo de origen-destino con el fin de facilitar el proceso de carga de datos a la bodega, esto es especificado en detalle en el Capítulo V.

3.5. Fase de desarrollo

Durante esta fase se busca definir la arquitectura técnica del proyecto, construir la bodega de datos con base en los requerimientos y los modelos producidos en la fase de análisis y diseño, poner en producción la solución y brindar la capacitación con respecto al manejo

del sistema a los usuarios finales. Esta fase de desarrollo se compone de cinco sub-fases: Definición de la arquitectura, desarrollo del Back-Room, desarrollo del Front Room, integración y despliegue.

Como se mencionó durante la descripción inicial de la metodología, esta tiene un componente iterativo e incremental, específicamente en las sub-fases de desarrollo (Back-Room, Front-Room e Integración). Una descripción detallada de esta fase y lo generado en ella puede ser encontrado en el Capítulo V.

3.5.1. Sub-fase de definición de la arquitectura

En esta sub-fase se debe definir la arquitectura técnica e infraestructura que se utilizarán durante el desarrollo de la bodega de datos. La definición de la arquitectura técnica se considera indispensable ya que proporciona un marco general de trabajo que da claridad al equipo de desarrollo acerca de la forma en la cual cada uno de sus integrantes debe participar en el proyecto para obtener el producto final.

Durante esta sub-fase se generó el plan de arquitectura técnica, el cual es dividido en cuatro partes. La primera de ellas es el desarrollo de la arquitectura a alto nivel, en donde se define de forma gráfica la descripción de la arquitectura general de la bodega de datos. En la segunda parte, se detallan los principales servicios con los que cuenta la bodega de datos creada. En la tercer parte, se describe el plan de seguridad utilizado y se definen los productos elegidos especificando las diferentes herramientas necesarias para la construcción de la solución. Finalmente en la cuarta parte, se genera una descripción de la infraestructura en la cual se desplegó la solución, especificando los requisitos hardware con los que se deben cumplir para una correcta ejecución de la bodega de datos.

Con el fin de evaluar las herramientas utilizadas y poder hacer la selección de las mismas, se generaron matrices de comparación, en donde se confrontan los productos candidatos para cada uno de los componentes de la arquitectura (ETL, DBMS[6], OLAP y Reportes). En esta matrices, presentadas en las tablas 5, 6, 7, y 8 se especifican las características consideradas junto a su peso (Importancia dada por el equipo), la puntuación por característica que cada uno de los productos obtuvo (Evaluada entre 0 y 100, teniendo en cuenta la documentación propia de cada herramienta) y la columna *Total* de cada herramienta, la cual representa la multiplicación de las dos anteriores. Las herramientas seleccionadas son: Oracle, SQL Server Integration Service, SQL Server Analisys Service y SQL Server Reporting Service, ya que presentaron el mayor puntaje total en cada área. Cabe mencionar que cada característica evaluada contó a su vez con una serie de subcaracterísticas, la descripción de estas se detalla en el Anexo E.

Es importante aclarar que Oracle fue considerado dentro de la comparación mostrada en la Tabla 5 ya que la institución cuenta con una licencia (por lo cual no se hace necesario un gasto adicional), pero debido a los altos costos para sus demás tecnologías, fue descartado para las demás comparaciones [36].

En la Figura 11 se presenta la arquitectura de la bodega de datos, junto a las herramientas seleccionadas. El plan de arquitectura completo puede ser encontrado en el Anexo D.

[6] Sistema de gestión de bases de datos.

Figura 11 Arquitectura de la bodega de datos.

3.5.2. Sub-fase de Back-Room

En esta sub-fase se creó el diseño físico de la bodega de datos relacional, con lo cual se consideran aspectos como la definición de los estándares de nombrado, la construcción del modelo físico de la bodega de datos, la estimación del tamaño de la bodega de datos y la construcción de los planes de indexación, agregación y particionamiento; una vez diseñada la bodega de datos, se implementó en el DBMS seleccionado. Por otro lado, se desarrollaron los procedimientos de extracción, transformación y carga, con el fin de asegurar la consistencia de los datos incluidos en la población de la bodega de datos. Toda la información en detalle sobre esta sub-fase puede ser encontrada en el Capítulo V.

3.5.3. Sub-fase de Front-Room

En esta sub-fase se busca definir y desarrollar las aplicaciones de usuario final, para esto se identifican los reportes candidatos los cuales son obtenidos a partir de los requerimientos finales adquiridos en la fase de análisis y diseño, para posteriormente ser priorizados en conjunto con el representante de los usuarios en base a su importancia con respecto al proceso de negocio. Una vez obtenida la lista de reportes priorizados, se debe diseñar la estrategia de navegación la cual le permite al usuario encontrar rápidamente la información necesaria para sus labores de análisis, los reportes elegidos son desarrollados bajo un estándar general, con el cual se asegura que los formatos, las representaciones gráficas y la distribución de la información sea idéntica en cada uno de los reportes y permita al usuario una rápida identificación de lo visualizado.

Al igual que la sub-fase de Back-Room, esta sub-fase es descrita con profundidad en el Capítulo V, en donde se especifican los enfoques de implementación utilizados, como también las aplicaciones de usuario generadas a lo largo del proyecto.

Característica	Peso	Oracle	Sql Server	PostgreSql	Mysql	Total Oracle	Total Sql Server	Total PostgreSql	Total Mysql
Costo	10	100	0	100	100	10	0	10	10
Documentación	5	100	100	100	100	5	5	5	5
Potencia del lenguaje utilizado	10	100	80	60	80	10	8	6	8
Tamaño de base de datos	10	100	100	80	80	10	10	8	8
Memoria	10	100	100	20	80	10	10	2	8
Número de usuarios conectados concurrentemente	10	100	100	60	60	10	10	6	6
Compatibilidad con diversos sistemas operativos	5	100	80	100	60	5	4	5	3
Particiones y comprensión de datos	10	100	100	100	100	10	10	10	10
Seguridad	10	100	100	80	60	10	10	8	6
Map Reduce	5	60	60	20	20	3	3	1	1
APIs y otros métodos de acceso	5	80	100	100	60	4	5	5	3
Métodos de replicación	5	100	60	60	100	5	3	3	5
Lenguajes de Programación Soportados	5	100	40	40	40	5	2	2	2
Puntaje Total	100					**97**	80	71	75

Tabla 5 Motor de Bases de datos Relacional.

Característica	Peso	Sql Server Analysis Services Enterprise	Pentaho Schema Workbench Enterprise	Iccube Enterprise	Sql Power Architect Enterprise Edition	Total SSAS	Total Psw	Total IcCube	Total Spae
Costo	**20**	100	60	20	20	20	12	4	4
Documentación	**15**	100	48	84	60	15	7,2	12,6	9
Casos Diseño	**30**	100	83	88	83	30	24,9	26,4	24,9
Funcionalidad	**20**	100	80,2	76,9	72,8	20	16,04	15,38	14,56
Características Técnicas	**15**	57	55	100	50	8,55	8,25	15	7,5
Puntaje Total	100					**93,55**	68,39	73,38	59,96

Tabla 6 Motor OLAP.

Característica	Peso	Sql Server Reporting Service Enterprise Edition	Pentaho Report Designer Enterprise Edition	Sql Power Wabit Enterprise Edition	Jasperreport Enterprise Edition	Total SSRS	Total Pentaho	Total Sql Power	Total Jasper
Costo	15	100	60	80	20	15	9	12	3
Documentación	15	100	48	28	40	15	7,2	4,2	6
Conexiones	15	100	60	60	120	15	9	9	18
Funcionalidades	20	92	88,8	72,2	72	18,4	17,76	14,44	14,4
Gráficos	20	100	100	60	100	20	20	12	20
Características Técnicas	15	95	80	100	85	14,25	12	15	12,75
Puntaje Total	100					**97,65**	74,96	66,64	74,15

Tabla 7. Herramientas de reportes.

Característica	Peso	Sql Server Integration Services Enterprise	Talend For Data Integration	Pentaho Data Integration (Kettle)	Clover Etl Community	Total SSIS	Total Talend	Total Pent	Total Clover
Precio	20	100	20	20	40	20	4	4	8
Documentación	15	100	100	80	60	15	15	12	9
Características técnicas	15	98	85	85	100	14,7	12,75	12,75	15
Componentes de transformación	35	93	95	91,4	88,2	32,55	33,25	31,99	30,87
Funcionalidades	15	100	100	100	100	15	15	15	15
Puntaje Total	100					**97,25**	80	75,74	77,87

Tabla 8 Herramientas ETL.

3.5.4. Sub-fase de integración

En esta sub-fase se integra el Front-Room junto con el Back-Room, con el fin de obtener un producto completo, se realizan las pruebas necesarias para comprobar la consistencia de los datos (Para mayor detalle ver Capítulo V) y posteriormente se genera la reunión con el representante de los usuarios para obtener la realimentación que llevará a la mejora de lo desarrollado. En el caso de que el usuario genere observaciones que consideren ajustes, se debe generar una nueva iteración, regresando a las sub-fases de Back-Room y Front-Room para poder modificar o incluir lo mencionado por el usuario.

En este caso en particular la prueba realizada para verificar la aceptación del proyecto se realizó con base en la métrica *nivel de satisfacción*, de la ISO/IEC 25022: Medidas de Calidad en Uso, definida para la subcaracterística *utilidad*. Debido a que los resultados de las pruebas fueron positivos no fue necesario generar una nueva iteración por lo cual se pasó directamente a la sub-fase de despliegue. Este proceso de aceptación involucró tanto al representante de los usuarios como también a dos usuarios adicionales elegidos debido al rol desempeñado en la institución, la evaluación realizada del sistema junto con la aceptación de los usuarios es descrita en profundidad en el Capítulo VI.

3.5.5. Sub-fase de despliegue

En esta sub-fase se debe poner la bodega a disposición de los usuarios, debido a que el alcance de este proyecto es académico y no una iniciativa de la administración de la universidad, no fue posible realizar el despliegue en un servidor de la división de tecnologías de la información de la Universidad del Cauca. Por lo tanto, este proceso se realizó en un equipo ubicado en una de las salas de informática disponibles de la Facultad de Ingeniería Electrónica y Telecomunicaciones, para que el Decano de la FIET pudiera verificar que el proyecto quedaba configurado en un equipo de la universidad. Adicionalmente en esta sub-fase se especifica una actividad en la cual se deben capacitar a aquellos usuarios que accederán al sistema, dicho proceso de capacitación en conjunto con los demás detalles del proceso de despliegue son descritos con detalle en el Capítulo V.

3.6. Fase de mantenimiento y crecimiento

En esta fase se genera un proceso de seguimiento al prototipo de la bodega de datos puesto en producción, además se plantea la constante evaluación de crecimiento de la bodega, como también los procedimientos de mantenimiento y control que aseguren un correcto funcionamiento a lo largo del tiempo. Debido al alcance del proyecto, esta fase no se realizó, porque el prototipo de la bodega de datos no quedó en funcionamiento.

3.7. Fase de gestión del proyecto

Esta fase se mantiene a lo largo de todo el proyecto y tiene como objetivo el realizar un seguimiento constante del progreso y de los resultados obtenidos. La metodología específica como opcional la actividad de administrar el log de cambios y recomienda su utilización para proyectos de gran tamaño o de alcance variable. Debido a que el proyecto se planteó desde un inicio con un alcance bien definido solo se mantuvo la activa participación del gerente del proyecto para comparar los avances que se iban obteniendo con lo que se había considerado inicialmente en la etapa de planeación. Adicionalmente, se llevaron a cabo reuniones semanales para informar los avances obtenidos y los problemas presentados en cada una de las fases del proyecto. Al finalizar el proyecto se ejecutó una reunión en la cual se documentaron las lecciones aprendidas, las cuales pueden ser encontradas en el Capítulo VII.

Capítulo IV. Modelado dimensional

El modelado dimensional es un elemento fundamental del proyecto, ya que es el pilar de la etapa de desarrollo y de él depende directamente la satisfacción de los requerimientos recolectados. A continuación, se presentan los modelos dimensionales diseñados en este proyecto, junto con una descripción de sus dimensiones y hechos, también se especifican los casos de diseño identificados en estos modelos.

Los esquemas de los modelos dimensionales expuestos en este capítulo son considerados según el marco teórico como esquemas constelación, sin embargo, dentro del proyecto han sido tratados como esquema estrella, por lo cual de ahora en adelante serán nombrados de esa manera.

4.1. Matriz bus

El proyecto inicialmente estaba enfocado en los modelos dimensionales relacionados al proceso de registro académico (primer objetivo), que consideran la información relevante con respecto a las notas y las faltas de los estudiantes. Sin embargo, como algo adicional, se incluyeron otros modelos que se muestran en la matriz bus (ver Tabla 9), en la cual las filas representan los modelos que se agrupan por procesos de negocio y en las columnas las dimensiones involucradas en estos procesos. La relación entre cada modelo y las dimensiones utilizadas en él, se verifica en la intersección de la matriz, en la cual los casos afirmativos de la relación son marcados con un "✓".

Al finalizar la etapa de diseño se lograron crear satisfactoriamente diez modelos dimensionales, agrupados en cinco procesos de negocio diferentes y donde fue necesaria la inclusión de veinticinco dimensiones. Tres dimensiones fueron consideradas con múltiples vistas (Juego de roles: Ver 4.2.1.4), localización con una vista para la residencia, el lugar de nacimiento y el lugar de procedencia tanto para el estudiante, como para el docente; el programa al igual que el pensum, tienen una vista para el programa/pensum del estudiante y una para el programa/pensum de la materia, ya que un estudiante puede matrícular materias que se encuentren registradas en un programa diferente al que pertenece. La relación en la matriz bus se genera cuando al menos una de las vistas es necesaria en el modelo.

		Estudiante	*Localización*	*Demografía*	*Datos económicos*	*Semestre*	*Indicadores*	*Programa*	*Materia*	*Adicional materia*	*Pensum*	*Docente*	*Adicional docente*	*Estado Curso*	*Período académico*	*Componente*	*Supletorio*	*Fecha*	*Salón*	*Franja*	*Tipo descuento*	*Prueba área*	*Estado cancelación*	*Tipo homologación*	*Proyecto*	*Dificultades*
Registro académico	*Notas semestral*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓											✓
	Notas por componente	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓									✓
	Control de asistencia	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			✓	✓	✓						✓

Registro financiero	*Descuento semestral*	✓	✓	✓	✓	✓	✓	✓			✓				✓						✓					✓
	Totalidad financiera	✓	✓	✓	✓	✓	✓	✓							✓											✓
Pruebas Saber	*Pruebas saber 11° y Saber pro*	✓	✓	✓	✓		✓	✓							✓							✓				✓
Procesos académicos	*Matrícula*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓								✓			✓
	Homologación	✓	✓	✓	✓	✓	✓	✓	✓						✓									✓		✓
	Proyecto de grado	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓		✓										✓	✓
Estados	*Estados estudiantes*	✓	✓	✓	✓	✓		✓			✓				✓											✓

Tabla 9 Matriz bus.

Los modelos de *Notas,* fueron considerados tanto por el equipo, como por los usuarios, como los de mayor impacto, es por esto que en la siguiente sección se realizará una profundización sobre estos modelos, mientras que los demás podrán ser encontrados en el Anexo E.

4.2. Modelos dimensionales

Para obtener los modelos dimensionales, se siguió el método de diseño de una tabla de hechos propuesto por Kimball (desarrollado por cada esquema estrella), la cual se encontraba constituida por los siguientes pasos:

- Definir el grano, esto significa identificar lo que representará cada uno de los registros almacenados en la tabla de hechos. Generalmente se escoge lo más bajo o granular como sea posible ya que de esa forma es más fácil dar respuesta a consultas que no se han considerado e incluir nueva información.
- Seleccionar las dimensiones, considerando que el grano de las dimensiones debe coincidir con el grano de la tabla de hechos.
- Adicionalmente se deben agregar las medidas consideradas para cumplir con las necesidades analíticas, acorde a la granularidad de la tabla de hechos.
- Construir el modelo, definiendo las relaciones que existen entre sus diferentes componentes, aquellas consideradas particulares ocasionaron la inclusión de diferentes casos de diseño los cuales pueden ser observados en detalle en las secciones 4.2.1.4 y 4.2.2.4.

4.2.1. Modelo de notas semestral

El modelo de notas semestral responde a las necesidades analíticas identificadas con los usuarios finales, enfocadas en el desempeño académico de los estudiantes.

4.2.1.1 Granularidad de la tabla de hechos

El grano se definió como la nota definitiva por materia a nivel semestral obtenida por cada uno de los estudiantes, por lo cual se especificó la granularidad de la tabla de hechos como snapshot periódico, ya que la nota definitiva se registra en un período de tiempo específico, en este caso el período académico.

4.2.1.2 Selección de dimensiones

Para este modelo fueron seleccionadas quince dimensiones, las cuales enmarcan el contexto de la nota obtenida por el estudiante. Dentro de estas dimensiones se consolidaron datos personales, demográficos, académicos y aspectos relacionados a la situación económica del estudiante. Con respecto a las materias cursadas se incluyeron aspectos que especifican la intensidad de la materia, el plan de estudios al que pertenece, entre otros. Finalmente se estableció la relación entre las notas obtenidas por el estudiante y los docentes que participaban en el proceso de enseñanza, por lo cual se adicionaron datos del docente como sus datos personales, el nivel académico, datos relacionados al proceso de contratación con el cual se generó su vínculo a la institución, entre otros.

En el proceso de diseño se tuvo en cuenta los componentes encontrados en los trabajos relacionados, los cuales fueron mencionados en la sección 2.2.5.4. Dentro de estos componentes se hallaron algunos atributos considerados de utilidad por el equipo de desarrollo y mencionados por parte de los usuarios para la resolución de sus necesidades analíticas, pero que no se encuentran almacenados en el sistema de información con el que cuenta la Universidad del Cauca (Sistema integrado de matrícula y control académico-SIMCA). Sin embargo, se tomó la decisión de incluir está información en el modelo dimensional, para que la universidad los pueda contemplar en futuras modificaciones de SIMCA. Aunque no fue posible la inclusión de dichos atributos en las etapas posteriores al diseño del modelo dimensional. Estos atributos se encuentran en diferentes dimensiones (Estudiante, dificultades, semestre y demografía datos económicos), además la dimensión *Dificultades* está compuesta en su totalidad por atributos inexistentes en el sistema transaccional.

A continuación, se describen algunas de las dimensiones (las demás se encuentran en Anexo E). Para cada una de ellas se nombran sus atributos (los atributos subrayados son aquellos considerados solo en la fase del modelado), se presenta una breve descripción y un ejemplo de los valores que pueden tener almacenados dentro de ellos, además se mencionan las jerarquías definidas, y las carpetas que fueron creadas con el fin de facilitar al usuario la navegación por la bodega de datos.

Dimensión Estudiante		
Información personal y universitaria del estudiante.		
Atributos	**Descripción**	**Ejemplo**
Identificador del estudiante (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2, 3.

Identificador del lugar de nacimiento (Lugar_Nacimiento_ID)	Llave foránea, proveniente de la dimensión “Localización”, referencia el lugar de nacimiento del estudiante.	1, 2,3.
Identificador del lugar de procedencia (Lugar_Procedencia_ID)	Llave foránea, proveniente de la dimensión “Localización”, referencia el lugar de procedencia del estudiante.	1, 2,3.
Nombre completo	Nombre completo del estudiante, contemplando los nombres y apellidos.	Pablo Pérez, Andrés Camilo Uribe.
Teléfono	Número de teléfono del estudiante.	8308899, 8403455.
Código del estudiante (Código)	Código asignado por la universidad, este es una secuencia de dígitos única para cada estudiante.	104612012322,104713017888.
Correo institucional (Correo)	Correo electrónico asignado por la universidad, este es único para cada estudiante.	User2@unicauca.edu.co, user3@unicauca.edu.co.
Tipo de Identificación	Tipo de documento de identidad que tiene el estudiante.	Cédula de ciudadanía, Cédula de Extranjería.
Número de documento de identidad (Número Identidad)	Número de identificación del estudiante como ciudadano.	1061726238,10789543.
Edad actual	Edad actual del estudiante.	18, 19,20.
Dirección actual	Dirección de residencia del estudiante.	Calle 140 # 100 – 30.
Indicador de ingreso (Indicador Ingreso)	Indicador que permite verificar la forma en la que el estudiante ingreso a la universidad.	Normal, Zona Marginal, Isleño, Costa Pacífica.
Indicador de procedencia de programa regional (Indicador Regional)	Indicador que permite verificar si el estudiante inicio en un programa regional y posteriormente realizo un traspaso a un programa regular.	Si proviene programa regional, No proviene programa regional.
Período de ingreso	Período en el que el estudiante ingreso al programa.	2012-I, 2013-I.
Libreta militar	Número de libreta militar del estudiante.	1061726238,10789543.
Celular	Número de celular del estudiante.	3102324567, 3016785432.
Factor RH	Factor RH y grupo sanguíneo del estudiante.	B+, O- .
Pensum actual	Pensum que actualmente se encuentra cursando el estudiante.	114,100.
Énfasis	Énfasis que ha sido elegido por el estudiante (Solo aplicado a algunos programas).	Telemática, Telecomunicaciones
Género	Identificación sexual del estudiante.	Masculino y Femenino.

Tipo institución	Indica el tipo de institución educativa de procedencia de la que egreso el estudiante como bachiller.	Técnica, Acelerada.
Institución privada	Indica el carácter de la institución educativa de procedencia.	Pública, Privada.
Institución educativa de procedencia (Institución procedencia)	Institución educativa en la que el estudiante finalizo sus estudios como bachiller.	Colegio Don Bosco, Colegio José Eusebio Caro.
Grupo étnico	Etnia a la cual pertenece el estudiante.	No aplica, Pueblo Indígena, Comunidad Negra.
Pueblo-Comunidad	Pueblo indígena o comunidad negra a la cual pertenece el estudiante.	Coyaima, Desano.
Capacidad excepcional	Indica si el estudiante cuenta con una capacidad o talento excepcional.	Cuenta con alguna capacidad excepcional, No cuenta con alguna capacidad excepcional.
Bachiller indígena	Indica si el estudiante obtuvo un título de bachiller indígena.	Aplica como bachiller Indígena, No aplica como bachiller indígena.
Andrés bello	Indica si el estudiante obtuvo la distinción Andrés bello al presentar las pruebas Saber 11.	Si Andrés Bello, No Andrés Bello.
Tipo de admisión	Indica el tipo de admisión con la cual el estudiante ingreso a la universidad.	Prueba Saber 11(Icfes), Prueba Interna.
Puesto prueba de admisión	Puesto que el estudiante ocupó en la prueba que le permitió el ingreso a la Universidad, registra el puesto dependiendo del atributo 'Tipo de admisión'.	1, 2,3.
Tipo Ingreso	Indica como fue el ingreso del estudiante.	Especial, Normal.
Elección programa	Indica si el estudiante ingreso al programa por voluntad propia.	Elección propia, Elección impuesta.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
No contiene jerarquías.			

Carpeta	**Descripción**	**Campos**
Datos del estudiante	Información personal y universitaria del estudiante.	Bachiller Indígena, Énfasis, Genero, Andrés Bello, Factor RH.
Grupo Étnico	Información del grupo étnico del estudiante.	Grupo étnico, Pueblo-Comunidad.

Indicadores	Información de diversos indicadores que se plantean para esta dimensión.	Indicador ingreso estudiante, Tipo de admisión, Tipo Ingreso, Indicador de procedencia de programa regional.
Institución de Procedencia	Información concerniente a la institución de cual procede el estudiante.	Tipo institución, Institución privada, Institución educativa de procedencia.

Tabla 10 Dimensión estudiante.

Dimensión Localización		
Es una dimensión con juego de roles, para el lugar de nacimiento, el lugar de procedencia y el lugar de residencia tanto del estudiante como del docente.		
Atributos	**Descripción**	**Ejemplo**
Identificador de la localización (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
País	Nombre del país.	Colombia, Chile.
Departamento	Nombre del departamento.	Cauca, Nariño, Cundinamarca.
Municipio	Nombre del municipio.	Popayán, Cali, Pasto.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
Jerarquía localización	Jerarquía compuesta por el País, departamento y municipio.	Colombia -> Cauca -> Popayán, Colombia -> Nariño -> Pasto	Completa y estricta.

Carpeta	**Descripción**	**Campos**
No contiene carpetas.		

Tabla 11 Dimensión localización.

Dimensión Demografía Datos Económicos		
Información demográfica del estudiante que varía en el tiempo y que está relacionada a la posición económica del estudiante.		
Atributos	**Descripción**	**Ejemplo**
Identificador de la demografía (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Estrato	Estrato económico al cual pertenece el estudiante.	Estrato 1, Estrato 2, Estrato 3.
Número de hermanos	Número de hermanos que tiene el estudiante.	1 hermano, 2 hermanos, 3 hermanos.
Posición hermanos	Posición que el estudiante ocupa en su grupo de hermanos.	Primer Hijo, Segundo Hijo, Tercer Hijo.
Vivienda propia	Indica si el estudiante cuenta o no con una vivienda propia.	Tiene vivienda propia, No tiene vivienda propia.

Desplazado	Indica si el estudiante ha abandonado su vivienda por causas de la guerra.	Es desplazado, No es desplazado.
Madre Cabeza	Indica si el estudiante es una madre cabeza de familia.	Es madre cabeza, No es madre cabeza.
Situación laboral	Indica si el estudiante está trabajando.	Trabaja, No trabaja.
Rango horas trabajadas	Indica el rango de horas trabajadas en el que se encuentra el estudiante.	0, 0-2, 2-4, 4-6,6-8.
Dependencia económica	Indica si el estudiante depende económicamente de su acudiente.	Depende económicamente, No depende económicamente.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
No contiene jerarquías.			

Carpeta	**Descripción**	**Campos**
Carpeta Demografía Económica	Información correspondiente a la mayoría de atributos de la dimensión excepto el atributo Posición hermanos.	Estrato, Número de hermanos, Vivienda propia, Desplazado, Madre Cabeza.

Tabla 12 Dimensión demografía datos económicos.

Dimensión Datos Demográficos		
Otro tipo de información demográfica relevante del estudiante que varía en el tiempo.		
Atributos	**Descripción**	**Ejemplo**
Identificador de la demografía (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Rango de edad	Rango de edad en el que se encuentra el estudiante.	18-20 años, 21-22 años.
Estado civil	Estado civil del estudiante.	Soltero, Casado, Unión Libre.
Acudiente	Indica si el estudiante tiene o no un acudiente registrado.	Tiene acudiente, No tiene acudiente.
Discapacidad	Indica el tipo de discapacidad que tiene el estudiante.	No tiene, Auditiva, Visual.
Deportista	Indica si el estudiante es un deportista de alto rendimiento.	Es deportista, No es deportista.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
No contiene jerarquías.			

Carpeta	**Descripción**	**Campos**
Carpeta Demografía	Información relevante para el análisis de esta dimensión.	Discapacidad, Estado civil, Rango de edad.

Tabla 13 Dimensión datos demográficos.

Dimensión Dificultades		
Registros binarios considerados útiles relacionados a dificultades personales que puedan afectar emocionalmente al estudiante.		
Atributos	**Descripción**	**Ejemplo**
Identificador del Indicador (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2, 3.
Problemas de conducta	Indica si el estudiante tiene problemas de vandalismo, alcoholismo o drogadicción.	Tiene problemas de conducta, No tiene problemas de conducta.
Vandalismo	Indica si el estudiante tiene problemas de vandalismo.	Con problemas de vandalismo, Sin problemas de vandalismo.
Alcoholismo	Indica si el estudiante tiene problemas de alcoholismo.	Con problemas de vandalismo, Sin problemas de alcoholismo.
Drogadicción	Indica si el estudiante tiene problemas de drogadicción.	Con problemas de vandalismo, Sin problemas de drogadicción.
Calamidades	Indica si el estudiante ha tenido calamidades personales, como por ejemplo la muerte de alguien cercano o experiencias traumáticas.	Tiene calamidades, No tiene calamidades.
Padres separados	Indica si los padres del estudiante son separados.	Padres separados, Padres juntos.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
No contiene jerarquías.			

Carpeta	**Descripción**	**Campos**
Problemas de conducta	Atributos relacionados a los problemas de conducta que puede tener el estudiante.	Problemas de conducta, Vandalismo, Alcoholismo y Drogadicción.

Tabla 14 Dimensión dificultades.

Dimensión Docente		
Información personal e institucional del docente.		
Atributos	**Descripción**	**Ejemplo**
Identificador del docente (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Identificador del lugar de nacimiento (Lugar_Nacimiento_ID)	Llave foránea, proveniente de la dimensión “Localización”,	1, 2,3.

	referencia el lugar de nacimiento del docente.	
Identificador del lugar de procedencia (Lugar_Procedencia_ID)	Llave foránea, proveniente de la dimensión "Localización", referencia el lugar de procedencia del docente.	1, 2,3.
Nombre Completo	Indica los nombres y apellidos de los docentes que trabajan en la universidad del cauca.	Luis Pérez, Francisco Zemanate.
Tipo de documento	Para aquellos docentes extranjeros, los cuales no tienen una cedula de ciudadanía.	Pasaporte, cédula de extranjería.
No. Doc. Identidad	Número de documento de identidad.	10567432, 11789888.
Departamento	Indica el departamento de la institución universitaria, al que pertenece un docente.	Departamento de Sistemas, Departamento de Telemática.
Facultad	Indica la facultad a la que pertenece el docente.	Facultad Ingeniería Electrónica y Telecomunicaciones, Facultad Ingeniería Civil.
Sigla Facultad	Indica la abreviatura de la facultad a la que pertenece el docente.	FIET.
Género	Identificación sexual del docente.	Masculino y Femenino.
Correo	Correo electrónico institucional que se encuentra relacionado al docente.	User1@unicauca.edu.co, user2@unicauca.edu.co.
Edad Actual	Indica la edad actual que tiene el docente.	30, 31, 27.
Estado	Estado actual el docente en la institución universitaria.	Activo o Inactivo.
Teléfono	Indica el número telefónico personal del docente.	8456277, 8765432.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
Jerarquía Nacimiento Docente	Jerarquía compuesta por el lugar de nacimiento del docente País, Departamento y Municipio.	Colombia-> Cauca-> Popayán.	Completa y estricta.
Jerarquía Procedencia Docente	Jerarquía compuesta por el lugar de procedencia del docente País, Departamento y Municipio.	Colombia-> Valle del Cauca-> Cali.	Completa y estricta.

Carpeta	Descripción	Campos
Datos Facultad	Información del docente relacionada con la facultad.	Departamento, Facultad, Sigla Facultad.

Tabla 15 Dimensión docente.

Dimensión Estado Curso		
Atributos relacionados con diferentes características de la relación estudiante-curso.		
Atributos	**Descripción**	**Ejemplo**
Identificador de estado curso (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Repetición	Indica la repetición en la que se encuentra el estudiante en el curso.	R0, R1, R2.
Estado	Indica si el estudiante aprobó o reprobó el curso.	Aprobado, Reprobado.
Rango Nota	Indica el rango específico en el que se encuentra la nota final del estudiante en el curso.	Muy bajo: 0.0-2.0, Bajo: 2.0-3.0.
Rango de faltas	Indica el rango específico en el que se encuentra la cantidad de faltas del estudiante en el curso.	0-6 faltas, Más de 18, “Reprobada por faltas”.
Tipo de nota	Indica la manera en la que se generó la nota del curso para el estudiante, considerando notas normales las que han sido evaluadas a lo largo del semestre, hasta notas generadas por procesos de homologación, validación o cancelación.	Homologación, Habilitación, Cancelación.
Hacinamiento	Indica si el grupo tuvo un hacinamiento en algunos de sus salones.	Grupo con hacinamiento de estudiantes, Grupo sin hacinamiento de estudiantes.
Compartida	Indica si el grupo o curso fue compartido entre docentes o fue dictado por un solo docente durante todo el período.	Grupo Compartido Por Docentes, Grupo Dictado Por Un Solo Docente.

Jerarquías	Descripción	Ejemplo	Tipo
Jerarquía Estado	Jerarquía compuesta por Estado, Tipo de nota y repetición.	Aprobado -> Normal -> R0	Completa y estricta.

Carpeta	Descripción	Campos
Rangos	Información relacionada a rangos manejados dentro de esta dimensión.	Rango Nota, Rango de faltas.

Tabla 16 Dimensión estado curso.

4.2.1.3 Selección de medidas

En esta sección se presenta el listado de las medidas consignadas en la tabla de hechos del modelo, clasificadas por el tipo de medidas. Se consideraron las medidas para las notas

de las materias que cuentan para el promedio (Ej: Cálculos, Físicas) y materias que no cuentan para el promedio (Ej: Materias ofertadas por el Programa de Formación en Idiomas).

Medidas básicas. Aquellas que son extraídas directamente de la bodega de datos relacional, estos son:

- Nota: Indica la nota final, para las materias que cuentan para el cálculo del promedio semestral y tienen una nota cuantitativa. Esta medida tendrá un valor numérico, de lo contrario tendrá un valor nulo.
- Nota no promedio: Indica la nota final, para las materias que no cuentan para el promedio semestral y que tienen una nota cuantitativa. Esta medida tendrá un valor numérico, de lo contrario tendrá un valor nulo.
- Número faltas: Indica el número de faltas totales que tuvo un estudiante en una materia.
- Recuento: Es un conteo de filas recuperadas.

En las tablas 17 y 18 se presentan las medidas básicas y calculadas con función de agregación[7], respectivamente. Como se puede observar existen medidas creadas únicamente con la finalidad de ser utilizadas en la creación de otras, y no se consideran útiles por si solas como indicadores, ya que no aportan información útil para la toma de decisiones, como es el caso de los recuentos.

Adicionalmente se presenta la Tabla 19, con las medidas derivadas con una función de cálculo. En todas las tablas se indica la visibilidad de las medidas, la cual especifica si el usuario podrá verla o no.

Medidas Básicas	Función de agregación	Visibilidad
Nota	Suma	No visible
Nota no promedio	Suma	No visible
Número faltas	Suma	No visible
Recuento	Conteo	Visible

Tabla 17 Medidas básicas semestrales.

Medidas derivadas o calculadas. Aquellas que necesitan una operación adicional en el cubo.

- Recuento notas: Es utilizada para calcular el “Promedio Notas”.
- Promedio notas: Representa la media de las notas finales de los estudiantes.
- Recuento notas no promedio: Es utilizada para calcular el “Promedio Notas No Promedio”.
- Promedio notas no promedio: Representa la media de las notas finales de los estudiantes en aquellos cursos que no cuentan para el promedio semestral.

[7] Permite calcular a partir de varias filas un solo valor (Ej: suma, conteo, conteo distinto.)

- Promedio número faltas: Permite realizar análisis para verificar la media de faltas teniendo en cuenta cualquier nivel de agrupación.
- Cantidad estudiantes: Indica el número de estudiantes distintos registrados en la tabla de hechos.
- Cantidad docentes: Indica el número de docentes distintos.
- Relación docente estudiantes: Indica la relación entre los docentes y los estudiantes, con el fin de identificar la cantidad de estudiantes que se encuentran a cargo de los docentes.

Medidas Derivadas	**Función de agregación**	**Visibilidad**
Recuento notas	Conteo de filas no nulas	No visible
Recuento notas no promedio	Conteo de filas no nulas	No visible
Cantidad estudiantes	Conteo distinto de estudiantes	Visible
Cantidad docentes	Conteo distinto de docentes	Visible

Tabla 18 Medidas derivadas con función de agregación.

Medidas Derivadas	**Fórmula de cálculo**	**Visibilidad**
Promedio notas	Nota/ Recuento notas	Visible
Promedio notas no promedio	Nota no promedio/ Recuento notas no promedio	Visible
Promedio número faltas	Número faltas/ Recuento	Visible
Relación docente estudiantes	Cantidad docentes/ Cantidad estudiantes	Visible

Tabla 19 Medidas derivadas con fórmula de cálculo.

4.2.1.4 Especificación de los casos de diseño

Esta sección está orientada a la pregunta de investigación del proyecto, especificando los casos de diseño identificados durante el diseño del modelado dimensional de notas semestral. A continuación se explican los cinco casos de diseño identificados:

Subdimensiones: Permiten representar las relaciones de uno a muchos entre dimensiones, este caso de diseño se incluyó debido a la relación existente entre los estudiantes con la dimensión *Localización*, ya que una gran cantidad de estudiantes están relacionados a una misma localización. Se puede observar en la Figura 12 que en la dimensión estudiante se hace necesaria dos llaves foráneas de la subdimensión localización (para procedencia y nacimiento). Este caso de diseño también se presenta en la relación entre la localización y la dimensión docente en la Figura 17.

Juego de roles: Es una múltiple relación entre dos tablas. Se puede visualizar en la Figura 13 una relación doble entre la dimensión *Programa* y la tabla de hechos, para considerar la información con respecto al programa al que pertenece el estudiante y el programa en la cual se encuentra adscrita la materia. Este caso de diseño también es utilizado en la relación entre la dimensión *Pensum* (pensum del estudiante y pensum de la materia) y la tabla de hechos. Además, en la relación de la dimensión *Localización* y la dimensiones *Estudiante*/*Docente* (localización de nacimiento y localización de procedencia).

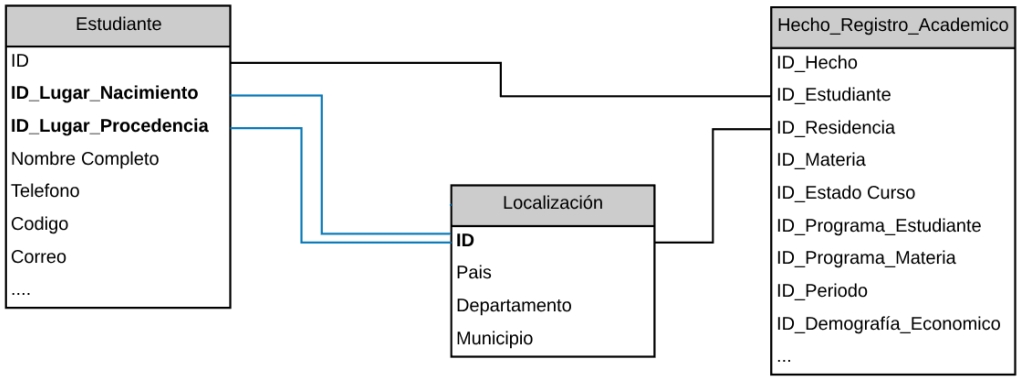


Figura 12 Caso de diseño subdimensión.

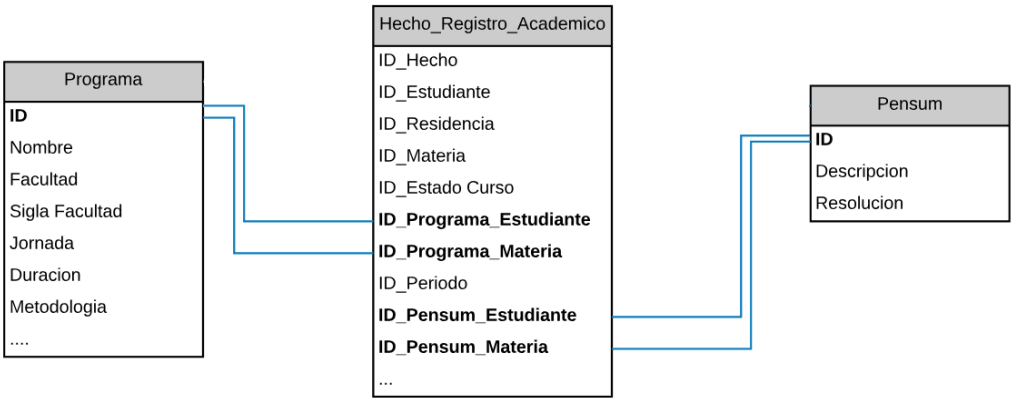


Figura 13 Caso de diseño juego de roles.

Minidimensión: Es una dimensión compuesta por atributos que cambian en el tiempo y de los cuales se considera necesario almacenar su histórico. Como se observa en la Figura 14 se diseñó una dimensión *Datos Demográficos* para agrupar información demográfica del estudiante. Los atributos como el género y el grupo étnico se mantuvieron en la dimensión del estudiante, ya que estos fueron considerados como invariantes. A diferencia de las subdimensiones, las minidimensiones si tienen una relación directa con la tabla de hechos. Las demás minidimensiones son: *Semestre*, *Dificultades*, *Demografía Datos Económicos*, *Datos Adicionales Docente*, *Datos Adicionales Materia*, como también la dimensión *Indicadores* y las *Localización* de residencia tanto del estudiante como también del docente, ya que estas localizaciones nacen inicialmente como atributos del estudiante/docente, pero pueden cambiar en el tiempo, a diferencia del lugar de nacimiento y procedencia.

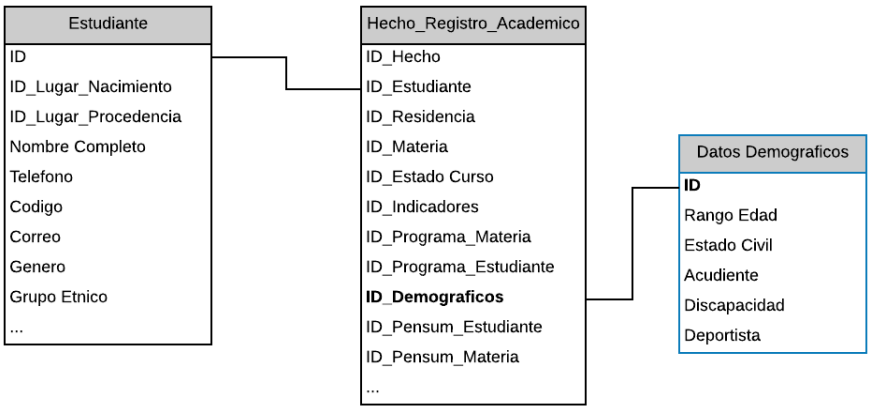


Figura 14 Caso de diseño minidimensión.

Dimensión multivaluada y tabla puente: Es utilizada para representar una relación muchos a muchos entre una dimensión y la tabla de hechos. Se incluyó en el diseño para reflejar aquellos cursos compartidos, es decir cursos que fueron dictados por más de un docente. En este caso, la dimensión multivaluada es la dimensión *Docente* (Ver Figura 15), debido a los múltiples valores que pueden estar relacionados a un mismo hecho. La relación entre la dimensión y la tabla de hechos se hace por medio de la tabla puente *Materia-Docente*, la cual almacena las llaves de las dimensiones relacionadas (Dimensión *Docente* y sus minidimensiones), la llave de la tabla de hechos y la medida *Conteo docentes*.

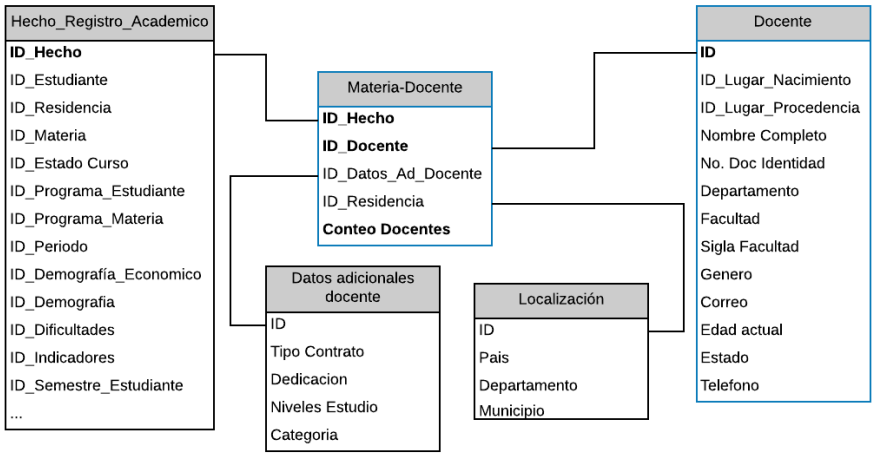


Figura 15 Caso de diseño dimensión multivaluada y tabla puente.

Dimensiones basura: Combinan varios indicadores y atributos de baja cardinalidad en una sola tabla de dimensiones en lugar de modelarlos como dimensiones separadas o medidas de la tabla de hechos, por lo general estos atributos no se encuentran relacionados entre sí. Este tipo de dimensión se crea para reducir el tamaño de la tabla de hechos y simplificar la navegación del usuario por el modelo dimensional. En la Figura 16 se muestran las dimensiones consideradas como dimensiones basura, las cuales tienen una relación directa con la tabla de hechos, estos atributos se agruparon en diferentes dimensiones, con el objetivo de facilitar al usuario los procesos de análisis, cabe resaltar que estas dimensiones a su vez se consideran como minidimensiones, ya que se originan inicialmente en la dimensión *Estudiante*.

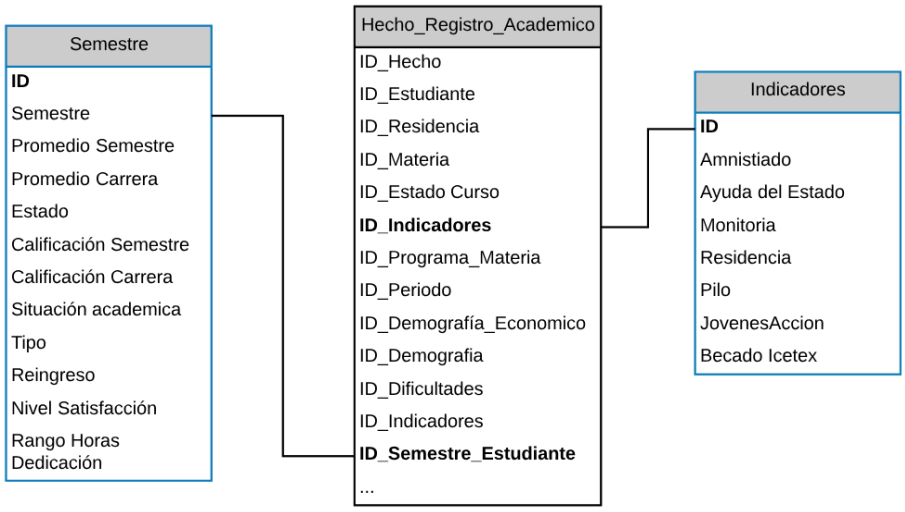


Figura 16 Caso de diseño dimensión basura.

4.2.1.5 Modelo dimensional completo

En la Figura 17 se presenta el modelo dimensional de notas semestral con las dimensiones y las relaciones mencionadas con anterioridad. La dimensión *Dificultades* como se mencionó anteriormente no es incluida en las etapas posteriores al diseño del modelo dimensional. En el modelo se aprecian todas las medidas involucradas, cabe resaltar que la medida *Conteo docentes*, no se encuentra en la tabla de hechos, ya que esta se calcula como un recuento distinto de la llave foránea del docente, la cual se encuentra en la tabla puente. Adicional a esto se observa la inclusión de una llave primaria para la tabla de hechos, ya que se recomienda la creación de ésta en los modelos que cuenta con una tabla puente [4], debido a que en la tabla puente se deben incluir las llaves de las entidades relacionadas, si no se tiene una llave primaria única para la tabla de hechos, tendrían que incluirse todas las llaves foráneas de la tabla de hechos (llave primaría compuesta).

Con este modelo se cumple con los requerimientos mencionados por los usuarios a nivel de notas semestrales (Ver Tabla 4, para mayor detalle ver Anexo C). Por ejemplo, la consulta “Cantidad de reprobados junto con la repetición en la que se encontraba cursando” es ejecutada con la medida *Conteo estudiantes* y los atributos de *Estado* y *Repetición* encontrados dentro de la dimensión *Estado Curso*.

4.2.2. Modelo de notas por componente

El modelo por componentes surge de la necesidad de monitorear el desempeño académico de los estudiantes en los componentes (Cortes) en los que se divide la nota semestral de una materia. A pesar de que las necesidades analíticas mencionadas por los usuarios no hacían énfasis en este tipo de información, se consideró útil la generación del modelo ya que por medio de este se podrían identificar aspectos como:

- Diferenciación entre el desempeño académico de los estudiantes entre los componentes considerados como parciales y finales.
- Relación entre la cantidad de supletorios llevados a cabo en un componente y el desempeño académico.

- Identificación de los estudiantes que se encuentran en matrícula condicional (tercera repetición) y en el transcurso de los componentes parciales llevan como reprobada la materia.

4.2.2.1 Granularidad de la tabla de hechos

El grano se definió como la nota definitiva por materia a nivel del componente, obtenida por cada uno de los estudiantes. Al igual que en el modelo presentado con anterioridad, se especificó la granularidad de la tabla de hechos como snapshot periódico, en donde el período de tiempo considerado es el componente.

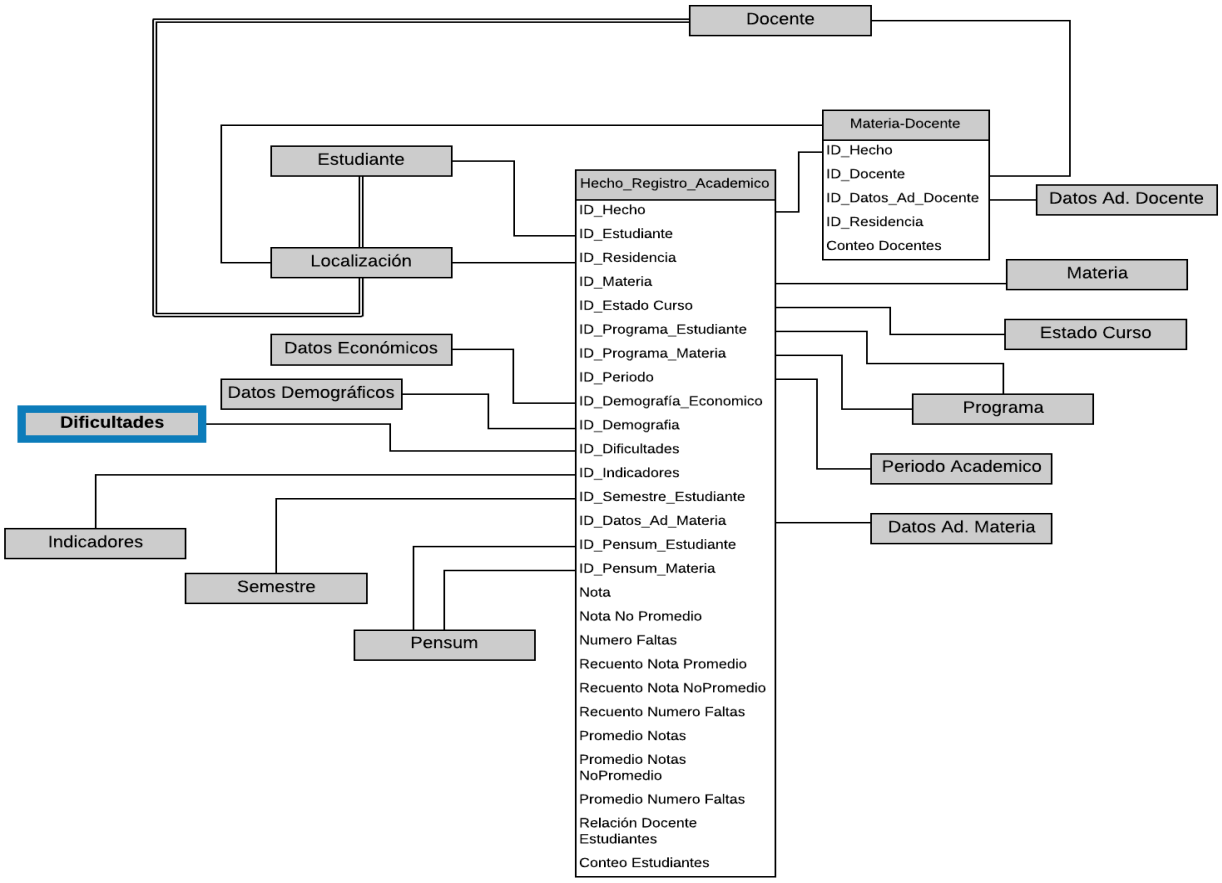


Figura 17 Modelo dimensional de notas semestral.

4.2.2.2 Selección de dimensiones

En este modelo fueron consideradas dieciséis dimensiones, entre las cuales se encuentran catorce dimensiones conformadas con el modelo de notas semestral, es decir dimensiones que son comunes a ambos modelos. Como se puede observar en la Tabla 9 las dimensiones exclusivas para este modelo son las relacionadas con los supletorios y los componentes, por lo cual serán las únicas detalladas en este apartado.

Dimensión Componente		
Almacena los componentes que se realizan en un período académico. Estos componentes son la agrupación de varias configuraciones o evaluaciones y en el contexto de la universidad son reconocidos como cortes.		
Atributos	**Descripción**	**Ejemplo**
Identificador del componente (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Descripción	Indica la descripción/nombre que el docente registro para el componente.	Nota 70%, Nota 30%.
Tipo	Indica si el componente es un componente parcial o final, siendo el componente final el que genera el cierre oficial de la materia en el sistema.	Parcial y Final.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
Jerarquía Componente	Jerarquía compuesta por Tipo y Descripción.	Parcial -> Nota 70 %.	Completa y estricta.

Carpeta	**Descripción**	**Campos**
No contiene carpetas.		

Tabla 20 Dimensión componente.

Dimensión Supletorio		
Almacena los diferentes motivos por los cuales el estudiante ha optado por la presentación de un examen supletorio. En esta dimensión se consideran únicamente los supletorios presentados, ya que no existe información digital con respecto a los supletorios solicitados que no llegaron a presentarse.		
Atributos	**Descripción**	**Ejemplo**
Identificador del supletorio (ID)	Llave primaria sustituta de la dimensión, generada a partir de una secuencia auto-incremental.	1, 2,3.
Descripción	Descripción especifica del supletorio, que indica el motivo por el cual se realizó el supletorio.	Motivo enfermedad, Motivo viaje.

Jerarquías	**Descripción**	**Ejemplo**	**Tipo**
La dimensión no cuenta con jerarquías ya que solo se cuenta con un atributo dentro de esta.			

Carpeta	Descripción	Campos
No contiene carpetas.		

Tabla 21 Dimensión supletorio.

4.2.2.3 Selección de medidas

En esta sección se presenta el listado de las medidas consignadas en la tabla de hechos del modelo, clasificadas por el tipo de medidas. Se considerarán medidas tanto para las materias que afectan el promedio académico del estudiante (Ej. Cálculos, Físicas) como para las que no lo afectan (Ej. Materias ofertadas por el Programa en Formación de Idiomas).

Existen dos tipos de notas en la institución: las notas obtenidas en los componentes, las cuales son registradas por cada corte; y las notas definitivas, las cuales son obtenidas a partir de las notas componentes (cada componente tiene un porcentaje de peso en esta nota). Por esto en el modelo se ha considerado la inclusión de notas componentes y de notas parciales (que indicarán el valor que tienen la nota componente en la nota definitiva). La nota parcial es obtenida del producto entre la nota del componente por su respectivo porcentaje, por ejemplo: si un estudiante obtiene una nota de 3.5 en el primer corte, el cual representa el 35% de la nota definitiva, su nota parcial será de 1.225.

Medidas básicas. Aquellas que son extraídas directamente de la bodega de datos relacional, estas son:

- Nota componente: Indica la nota del corte, considerado para materias que afectan el promedio.
- Porcentaje notas promedio: Se relaciona con la “Nota componente”, esta medida indica el porcentaje que se le dio al componente en Simca.
- Nota parcial: Indica el valor que tiene la nota componente en la definitiva de las materias que afectan el promedio.
- Nota componente no promedio: Representa la nota de un corte, considerado para materias que no afectan el promedio.
- Porcentaje notas no promedio: Se relaciona con la “Nota componente no promedio”, esta medida indica el porcentaje que se le dio al componente en Simca.
- Nota parcial no promedio: Indica el valor que tiene la nota componente en la nota definitiva de las materias que no afectan el promedio.

De manera similar al anterior modelo, este cuenta con las medidas *conteo estudiantes*, *conteo docentes* y *relación docentes estudiantes*.

En las tablas 22 y 23 se presentan las medidas básicas y calculadas con función de agregación, respectivamente, adicionalmente se presenta la Tabla 24, con las medidas derivadas con una función de cálculo. En todas las tablas se indica la visibilidad de las medidas, la cual especifica si el usuario podrá verla o no.

Medida Básicas	Función de agregación	Visibilidad.
Nota componente	Suma	No visible

Nota componente no promedio	Suma	No visible
Porcentaje notas promedio	Suma	No visible
Porcentaje notas no promedio	Suma	No Visible
Nota parcial	Suma	No Visible
Nota parcial no promedio	Suma	No visible

Tabla 22 Medidas básicas componentes.

Aunque las medidas “Nota parcial” y “Nota parcial no promedio” son consideradas como básicas, necesitan un cálculo previo, este cálculo puede ser encontrado con detalle en la sección 5.3.

Medidas derivadas o calculadas. Aquellas que necesitan una operación adicional en el cubo.

- Recuento nota componente: Conteo de las notas componentes pertenecientes a una materia que afecta el promedio.
- Promedio nota componente: Indica el promedio de las notas componentes, solo considerado para las materias que afectan el promedio.
- Promedio nota parcial: Indica el promedio de las notas parciales, solo considerado para las materias que afectan el promedio.
- Recuento nota componente no promedio：Indica el promedio de las notas parciales en las materias que no afectan el promedio.
- Promedio nota componente no promedio: Indica el promedio de las notas componentes de las materias que no afecta el promedio
- Promedio nota parcial no promedio: Promedio de las notas parciales para materias que no afectan el promedio.
- Conteo supletorios: Indica el número total de supletorios realizados en el componente.

Medidas Derivadas	**Función de agregación**	**Visibilidad**
Recuento nota componente	Conteo de filas no nulas	No visible
Recuento nota no promedio	Conteo de filas no nulas	No visible
Conteo supletorios	Conteo.	Visible

Tabla 23 Medidas derivadas con función de agregación componentes.

Medidas Derivadas	Fórmula de cálculo	Visibilidad
Promedio nota componente	Nota componente /Recuento nota componente	Visible
Promedio nota parcial	Nota parcial/Recuento nota componente	Visible
Promedio nota componente no promedio	Nota componente no promedio/Recuento nota componente no promedio	Visible
Promedio nota parcial no promedio	Nota parcial no promedio/Recuento nota componente no promedio	Visible

Tabla 24 Medidas derivadas con fórmula cálculo componentes.

4.2.2.4 Especificación de los casos de diseño

Esta sección está orientada a la pregunta de investigación del proyecto, especificando los casos de diseño identificados durante el diseño del modelado dimensional de notas componentes. A continuación se explican los cinco casos de diseño identificados:

Debido a la similitud entre los dos modelos por enfocarse en el mismo proceso de negocio, el único caso de diseño diferente a los expuestos anteriormente es una dimensión multivaluada y la tabla puente, el cual se describe a continuación.

Dimensión multivaluada y tabla puente: Permite almacenar los diferentes supletorios realizados por el estudiante para el mismo componente. En este caso, la dimensión multivaluada es la dimensión *Supletorio* (Ver Figura 18), debido a los múltiples valores que pueden estar relacionados a un mismo hecho. La relación entre la dimensión y la tabla de hechos se hace por medio de la tabla puente *Puente-Supletorio*, la cual almacena las llaves de la dimensión *Supletorio*, la llave de la tabla de hechos y la medida *Conteo supletorios*.

4.2.2.5 Modelo dimensional completo

En la Figura 19 se presenta el modelo dimensional de notas por componente con las dimensiones y relaciones mencionadas con anterioridad. Además del caso de diseño de dimensión multivaluada y tabla puente, la otra diferencia con el modelo de notas semestrales, es la relación directa de la dimensión *Docente* con la tabla de hechos, debido a que las clases durante un componente son dictadas por un único docente.

El modelo cumple con los requerimientos solicitados a nivel de notas por componente. Por ejemplo, la consulta “Conteo de supletorios por tipo de componente” se realiza con la medida *Conteo supletorios* de la tabla Puente-Supletorio y el tipo de componente (Parcial, Final) de la dimensión *Componente*.

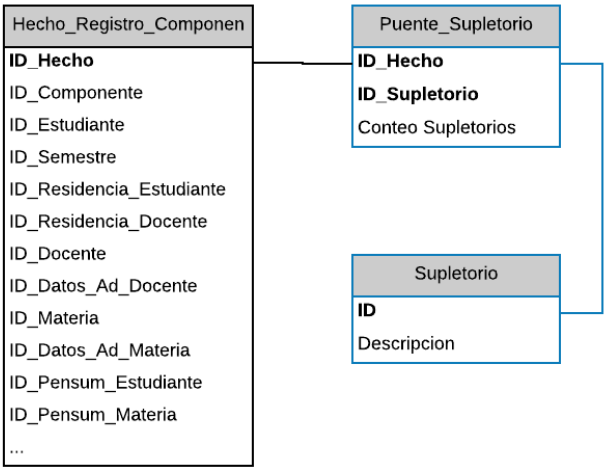


Figura 18 Caso de diseño dimensión multivaluada y tabla puente supletorio.

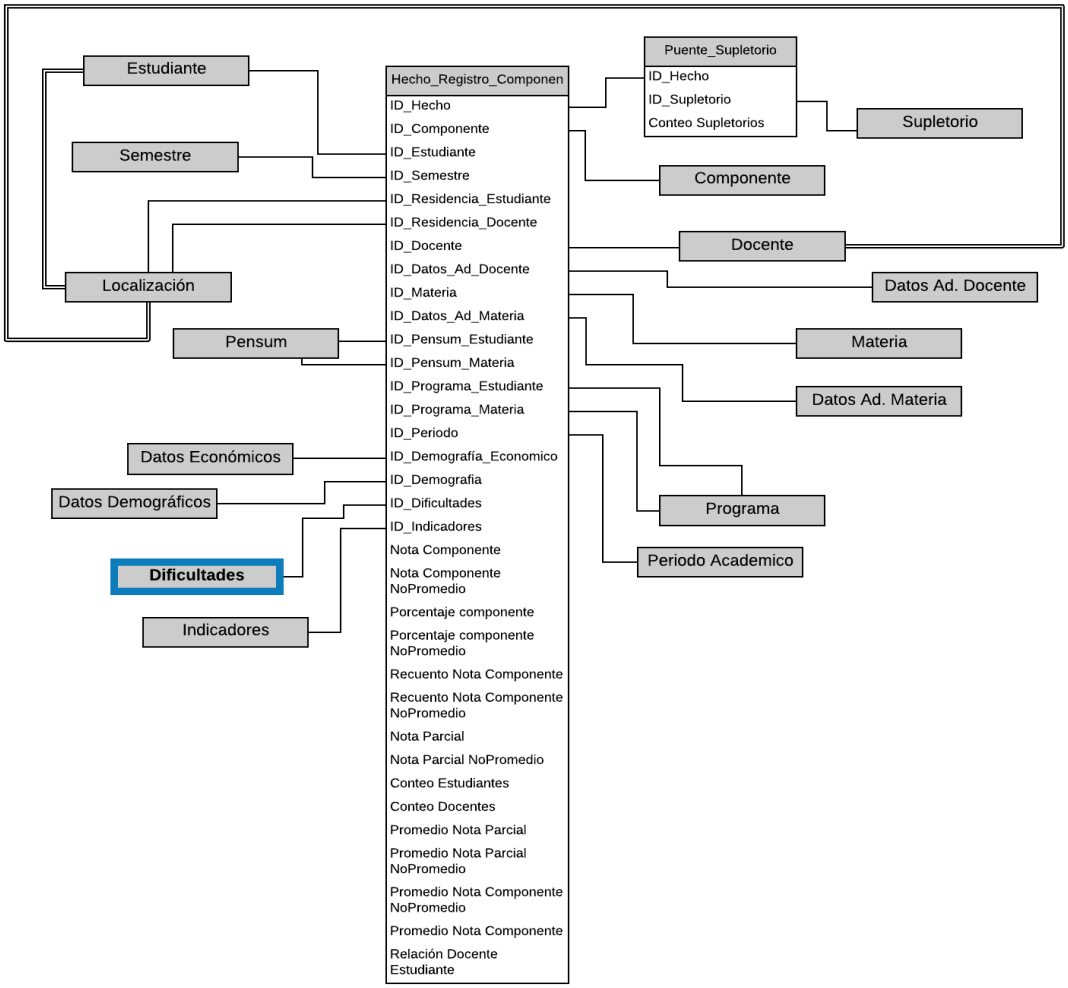


Figura 19 Modelo dimensional de notas por componente.

Capítulo V. Elaboración del prototipo propuesto

En este capítulo se presentan las actividades realizadas para el desarrollo del prototipo, relacionado con los modelos dimensionales de notas propuestos en el Capítulo IV. Por cada esquema se describe la creación de la bodega datos relacional, el proceso de Extracción, Transformación y Carga (ETL *sus siglas en inglés*), construcción del cubo OLAP con sus correspondientes jerarquías y demás componentes, e informes generados.

5.1. Esquema notas definitivas

A continuación, se presentan todos los elementos que se tuvieron en cuenta en la construcción del prototipo con respecto a este esquema.

5.1.1. Desarrollo Back-Room

Para el desarrollo de esta fase, se utiliza el DBMS de Oracle y el Servicio de Integración de SQL Server, mencionados en la sección 3.5.1, debido a esto los planes de indexación y particionamiento tuvieron en cuenta las restricciones y características del DBSM.

5.1.1.1 Diseño físico de tablas

Esta actividad es muy importante porque de esta depende la creación de un repositorio sólido y capaz de gestionar los elementos físicos del disco para poder generar consultas con buen rendimiento. Los modelos lógico y físico de la bodega de datos se hicieron en la herramienta Oracle SQL Developer Data Modeler, ya que como se menciona en la metodología [35], es recomendable hacer uso de este tipo de herramientas para generar el diseño físico de tablas.

A continuación, se presentan los elementos tenidos en cuenta para la creación del diseño físico de tablas.

Estándares de nombrado para la bodega de datos relacional

Los estándares de nombrado es una parte importante para la identificación de las tablas y atributos, ya que permite un buen entendimiento del esquema relacional. Gran parte de los nombres de atributos estuvieron basados en los nombres de las columnas origen de la base de datos del sistema OLTP (SIMCA), ya que permite a la división de tecnología una mejor familiarización con la bodega de datos, para un futuro despliegue y mantenimiento del sistema. Para la estandarización de nombrado de la bodega de datos se tuvo en cuenta lo siguiente:

- Los nombres de las tablas y atributos deben contener el símbolo "_" cuando haya un espacio.
- Los nombres de las tablas deben ser claros de acuerdo a la información almacenada en la dimensión o tabla de hechos, dependiendo del caso.
- Los nombres de las llaves primarias y foráneas deben contener al final de la cadena la palabra "ID" separada por un "_".
- Las tablas de hechos comenzarán con la palabra "Hecho".

Modelo físico de datos y tamaño de la bodega de datos.
El modelo físico de datos se creó a partir de los estándares de nombrado y del modelo dimensional propuesto en el Capítulo IV, y su resultado se presenta en el Anexo F. Dentro del diseño físico cabe destacar que se crearon una serie de columnas extras en algunas tablas de dimensiones como el atributo "CHECKSUM", los cuales fueron utilizados durante el proceso ETL y que no tienen significado alguno para el negocio.

Debido a que el prototipo no cuenta con datos reales, ya que por cuestiones de confidencialidad no se pudo tener acceso a estos, no se puede realizar una estimación real del tamaño de la bodega de datos. Sin embargo, para realizar la estimación de la cantidad de datos, se recomienda tener claro que algunas tablas crecerán continuamente en el tiempo como la tabla estudiante, la tabla puente y la tabla de hechos. Adicionalmente, se recomienda verificar el crecimiento de los estudiantes en cada período académico y el número de materias promedio que matriculan, ya que estos son los factores que determinan en gran medida el número de registros en la tabla de hechos. De igual manera la estimación del tamaño de los índices no puede ser definida, ya que el tamaño de estos depende de la cantidad de datos.

Plan de indexación inicial
En este apartado se presentan las recomendaciones sobre el plan de indexación del esquema relacional de la bodega de datos, teniendo en cuenta la guía [37] que ofrecen los desarrolladores de Oracle Database.

Normalmente, la indexación en una base de datos transaccional, se realiza con el objetivo de mejorar el rendimiento en las funciones DML[8], por el contrario, las bodegas de datos se enfocan en la realización de consultas más que en procesos transaccionales, por lo cual es importante tener un buen plan de indexación que permita la optimización de los tiempos de respuesta y así la satisfacción del usuario.

Según [37], los índices más apropiados para las bodegas de datos son los mapa de bits, ya que estos tienen un espacio de almacenamiento considerablemente bajo con respecto a los índices de Árbol B, por lo tanto se ajustan para el desarrollo de las bodegas de datos que generalmente cuentan con una gran cantidad de datos históricos y donde los índices pueden llegar a ser demasiados grandes. También es necesario tener en cuenta que el uso de índices de mapa de bits, según [38], es comúnmente utilizado en sistemas donde no existe una actualización o eliminación de datos frecuente.

A continuación, se muestra el plan de indexación propuesto para el esquema en mención:

Tabla de Hechos
Debido a que la tabla de hechos, por lo general, contiene una gran cantidad de datos se hace necesario realizar una indexación de esta. Según [37], una columna de la tabla de hechos es candidata a utilizar un mapa de bits, siempre y cuando:
1. Todos los valores distintos de dichas columnas estén presentes en cien o más filas de la tabla de hechos.
2. Adicionalmente, se debe cumplir una de las siguientes condiciones:

[8] Lenguaje de manipulación de datos.

- La columna indexada estará restringida en las consultas (a las que se hace referencia en la cláusula WHERE de una sentencia SQL).
- La columna indexada es una llave foránea para una tabla de dimensiones.

Todos los atributos presentados en la Tabla 25 cumplen con la condición número 2, por ser llaves foráneas en la tabla de hechos. Además, se observa que existen varias columnas que tienen un número elevado de valores distintos que cumplen con la condición 1, sin embargo la condición establece que deben ser todos. Debido a que pocos atributos cumplían con esta condición, se decidió establecer un límite porcentual, donde las columnas con un porcentaje mayor o igual a 80% sean consideradas candidatas a utilizar un índice de mapa de bits (los atributos resaltados en "*negrilla*"). Para las demás columnas considerando su alta cardinalidad, no se aplica ningún tipo de índice. En cuanto a la clave primaria simple, se debe crear un índice de Árbol B debido a su unicidad y por ende a su alta cardinalidad (no se recomienda aplicar un índice de mapa de bits para esta columna, ya que el tamaño del índice depende directamente de la cardinalidad [38]). Por último, las medidas no contarán con índices, ya que, a pesar de que puedan cumplir con la condición 1**,** no cumplen con la condición 2**,** ya que no se encontrarán como restricciones en las clausulas WHERE y no son llaves foráneas.

Atributo	Número de valores distintos	Número de valores distintos en 100 o más filas	Relación
dem_economicos_id	1241	777	62,61%
datos_ad_materia_id	50	44	**88,00%**
demografia_id	142	80	56,34%
estado_curso_id	1132	292	25,80%
estudiante_id	7000	0	0,00%
indicadores_id	22	13	59,09%
materia_id	491	469	**95,52%**
pensum_est_id	5	5	**100,00%**
pensum_mat_id	6	5	**83,33%**
período_id	29	25	**86,21%**
programa_estudiante_id	5	5	**100,00%**
programa_materia_id	6	5	**83,33%**
residencia_id	2	2	**100,00%**
semestre_estudiante_id	318	160	50,31%

Tabla 25 Candidatos para bitmap, hechos semestrales.

Tabla Puente

La tabla puente es similar a una tabla de hechos, ya que contiene varias llaves foráneas y una gran cantidad de datos, por lo cual se recomienda realizar la indexación teniendo en cuenta las mismas condiciones presentadas anteriormente. Todas las columnas de esta tabla puente son llaves foráneas, por lo tanto se realizó la verificación solo de la condición 1 y sus resultados son presentados en la Tabla 26. Los atributos resaltados en "*negrilla*" son candidatos para crearles índices de mapa de bits. Por otro lado, para la columna "hecho_id", debido a su alta cardinalidad, se sugiere aplicarle un índice de Árbol B.

Atributo	Número de valores distintos	Número de valores distintos en 100 o más filas	Relación
hecho_id	226721	0	0,00%
docente_id	916	774	**84,50%**
residencia_docente_id	2	2	**100,00%**
contratacion_id	12	12	**100,00%**

Tabla 26 Candidatos para bitmap, tabla puente semestrales.

Dimensiones

De igual manera se hace necesario realizar un plan de indexación para las tablas de dimensiones, ya que cuenta con una llave primaria simple (llave sustituta), en este caso se recomienda utilizar un índice de Árbol B, de la misma forma se recomienda aplicar este tipo de índices para columnas con valores únicos o casi únicos como es el caso del *código* en la tabla de dimensión *Estudiante*. Finalmente para los atributos o columnas de baja de cardinalidad (pocos valores distintos) y por los cuales se hagan consultas más frecuentes, se recomienda hacer uso de índices de mapa de bits como es el caso del *género* en la tabla de dimensión *Estudiante*, *condición académica* en la dimensión *Semestre*, etc. Adicionalmente, en este modelo se encuentra el caso de diseño denominado subdimensión presentado en el Capítulo IV. Para estos casos según [37], se sugiere indexar por medio de los índices de combinación de mapa de bits que mejoran considerablemente el rendimiento de las consultas, ya que al crearlos, estos almacenan un mapa de bits relacionando los registros de la tabla de hechos con el atributo indexado, un ejemplo del mapa de bits puede ser observado en Tabla 27. Las columnas a indexar serán país y departamento, que pertenecen a la dimensión Localización. La columna municipio no se indexa debido a su alta cardinalidad.

Registros Hechos	Departamento 'Cauca'	Departamento 'Nariño'	……
Registro 1	1	0	
Registro 2	0	1	
Registro 3	1	0	
Registro 4	1	0	
Registro 5	1	0	
…..			

Tabla 27 Ejemplo de mapa de bits.

Por último se recomienda, en las tablas de hecho, puente y dimensiones, hacer uso de índices simples, ya que se desconoce que atributos o columnas serán consultados con mayor regularidad por parte de los usuarios finales. En el Anexo F, se encontrará la definición de los índices de la tabla de hechos y las tablas de dimensiones.

Plan de particionamiento

Las bodegas de datos contienen, por lo general, un gran volumen de datos, es por esto que es necesario el particionamiento para permitir una mejor escalabilidad, además de una facilidad en la administración de las tablas con gran cantidad de datos. Adicionalmente, el particionamiento permite realizar de una manera más rápida procesos como el cargue de

datos, indexación, etc. En este apartado se presenta la recomendación para el particionamiento de la tabla de hechos, ya que esta contiene la mayoría de los datos en la bodega, sin embargo, en este prototipo no fue implementado, ya que en la versión Express de Oracle Database no es posible acceder a esta funcionalidad. Adicionalmente como se menciona en [32], una tabla es candidata a ser particionada cuando almacena 100 millones de filas o diez gigabytes de datos, por lo cual la tabla de hechos y puente no cumplen con esta condición como se observa en la Tabla 28.

Según [32], el particionamiento de las tablas debe estar definido por períodos de tiempo, en este caso, se recomienda particionar la tabla de hechos por períodos académicos. Por lo tanto, es necesario crear una llave sustituta para la tabla dimensión *Período* con un significado claro (ejemplo: para el primer período de verano del año 2012 la clave sustituta sería 201203), además se espera que la información de esta tabla sea utilizada frecuentemente por los usuarios finales.

Plan de agregación
Las agregaciones son utilizadas en consultas y reportes, estas almacenan un pre-cálculo de las medidas en el nivel más resumido de los datos, por esta razón las consultas pueden mejorar considerablemente ya que no tienen que obtener dichos cálculos en tiempo de ejecución. Como se menciona en [32], la agregación de tablas se puede realizar a nivel de la bodega relacional o a nivel del cubo OLAP. Sin embargo, si la versión del ambiente de desarrollo permite crear y mantener las agregaciones, es recomendable gestionarlas con el motor OLAP (SSAS). Adicionalmente se debe considerar que el almacenamiento en una base de datos multidimensional es mucho más pequeño que en una base de datos relacional. Las agregaciones para el prototipo se presentan en la sección 5.3.1.

5.1.1.2 Implementación de la bodega de datos relacional
En este apartado se presenta el tamaño de las tablas e índices para el esquema en mención y su implementación. Cabe destacar que el tamaño calculado para la bodega de datos del prototipo es un aproximado porque no se cuenta con la información real respecto a los datos de la Universidad del Cauca.

Tamaño de tablas e índices
En la Tabla 28, se presentan las tablas e índices particulares para este esquema. Solo se presentan los índices de Árbol B propuestos en el plan de indexación, debido a que el prototipo fue desarrollado en la versión Express Oracle Database, en la cual no está habilitada la funcionalidad para crear índices de mapa de bits. El tamaño de las demás tablas de dimensiones de este esquema se presenta en el Anexo F.

Objeto de la bodega de datos	Tipo	Número de filas	Tamaño en disco (MB)
ESTADO_CURSO	TABLA	4416	0,875
ESTADO_CURSO_PK	ÍNDICE	NO APLICA	0,1875
HECHO_REGISTRO_ACADEMICO	TABLA	226721	17
HECHO_REGISTRO_ACADEMICO_PK	ÍNDICE	NO APLICA	4
MATERIA_DOCENTE	TABLA	291860	7
MATERIA_DOCENTE_INDEXHECHO	ÍNDICE	NO APLICA	6

Tabla 28 Tamaño del esquema notas definitivas.

Creación de tablas e índices
La creación de las tablas e índices para este esquema se realizó teniendo en cuenta el diseño físico de tablas presentado en la sección 5.1.1.1. Además de la creación de las tablas, índices, llaves primarias, etc., también se creó un tablespace[9], para almacenar de forma independiente todo lo relacionado a este prototipo de bodegas de datos. Por último, el script fue ejecutado sobre un cliente de base de datos, en este caso se utilizó Oracle SQL Developer. En el Anexo F, se encuentra la definición de la bodega de datos haciendo uso de un script que contiene las respectivas sentencias DDL.

5.1.1.3 Desarrollo proceso ETL
En este apartado se presentarán las fuentes de datos, recomendaciones para el proceso ETL e inconvenientes presentados durante dicho proceso.

Fuentes de Datos
La selección de las fuentes de datos, tuvo en cuenta el tipo de datos almacenados y la posibilidad de tener acceso a estos. Por lo anterior, se tomó solo el Sistema Integrado de Matrícula y Control Académico (SIMCA). Aunque este sistema hace uso de diversos repositorios o bases de datos en su funcionamiento, solo se pudo tener acceso a: "ACADEMICO" y "USRINSCRIPCIONES". Adicionalmente, se tuvo en cuenta archivos Excel generados por el equipo de desarrollo.

En la base de datos "ACADEMICO" se almacena los datos correspondientes a la vida académica de los estudiantes y su información personal, por lo tanto, esta base de datos fue muy importante como insumo para este proceso. Los datos contenidos en esta base de datos correspondiente a los docentes, es relativamente poca, ya que esta se encuentra en detalle en el sistema de Recursos Humanos, sin embargo, este sistema no es gestionado por la División de Tecnologías de la Universidad del Cauca, lo cual imposibilitó el acceso.

La base de datos "USRINSCRIPCIONES" contiene los datos obtenidos en el proceso de inscripción del estudiante. Los datos utilizados de este repositorio en el desarrollo de este proceso, fueron relativamente pocos, pero son útiles para realizar distintos análisis de la población estudiantil.

De las bases de datos mencionadas anteriormente, solo se pudo obtener la estructura de estas, los datos no pudieron ser obtenidos por cuestiones de confidencialidad. Debido a lo anterior, las estructuras obtenidas se implementaron en equipos propios y posteriormente se poblaron las tablas necesarias con datos ficticios por medio de procedimientos almacenados, y de esta manera construir un proceso ETL que pueda ser útil para la universidad.

Proceso ETL
Para el desarrollo de este proceso, se tuvo como punto de partida las fuentes de datos descritas anteriormente y la implementación de la bodega de datos relacional mencionada en el apartado 5.1.1.2. La herramienta para la realización de este proceso es SQL Server Integration Services mencionada en la Tabla 8. Dicha herramienta maneja elementos llamados paquetes, los cuales contienen una serie de componentes de control para realizar

[9] Espacio de almacenamiento donde se guardan los objetos de una base de datos.

todo el proceso del flujo de datos. Adicionalmente para la creación de estos paquetes se siguió el diseño del mapa origen-destino de los datos presentado en el Anexo G.

Para todas las tablas de este esquema se crearon dos tipos de paquetes: el paquete inicial y el paquete incremental. El paquete inicial permite el cargue de los datos teniendo en cuenta toda la información histórica, mientras que el paquete incremental permite verificar si existe algún nuevo registro (para el caso de las dimensiones y la tabla de hechos) o actualización de estos (solo para las dimensiones).

A continuación se presentaran las descripciones del flujo datos de cada paquete relacionado a las tablas de dimensiones *Estudiante y Docente*, y la tabla puente. La descripción de los paquetes de cargue de datos que no son presentados en este apartado, se encuentran en el Anexo G.

Tabla de dimensión Estudiante
Paquete de cargue inicial: El cargue de datos para la dimensión *Estudiante* se puede observar en la Figura 20.

- Extracción: La mayoría de los datos provienen del repositorio “ACADEMICO”, como la información personal y académica del estudiante, además otro tipo de información como: colegio de procedencia, lugar de nacimiento, entre otros. En la base de datos “USRINSCRIPCIONES” se encuentra información como: tipo de admisión del estudiante, puesto en la prueba de admisión, premios del estudiante (ej: Andrés Bello), entre otros. Esta información se extrae utilizando una consulta que combina la información de las bases de datos mencionadas, de esta manera se reduce en gran medida el uso de componentes, lo cual podría generar complejidad a la hora de mantenimiento al paquete. El componente denominado *Fuente SIMCA*, es el encargado de ejecutar la consulta mencionada.

- Transformación: Se realiza una conversión de los tipos de datos con el fin de evitar problemas en el momento del cargue en el destino. La limpieza de datos se enfoca principalmente en evitar los valores nulos y reemplazarlos por valores más significativos para los usuarios (Ej.: Cuando el Grupo étnico viene nulo del sistema OLTP pasa a la bodega como *No registra*). Las búsquedas tanto del lugar de procedencia como de nacimiento, tienen como objetivo transformar el código del municipio del sistema OLTP en la llave sustituta de la dimensión *Localización.* El componente *Creación código hash* consiste en la generación de un número único (sirve como insumo para la actualización y adición de filas en el paquete de cargue incremental) a partir de los valores de las columnas que pueden cambiar en el tiempo. Por último se genera la llave sustituta por medio del componente *Creación llave sustituta*.

- Cargue: El destino de los datos es la tabla de la dimensión *Estudiante* de la bodega de datos relacional, en la configuración del destino se asocia cada una de las columnas del flujo de datos con las columnas de la tabla destino.

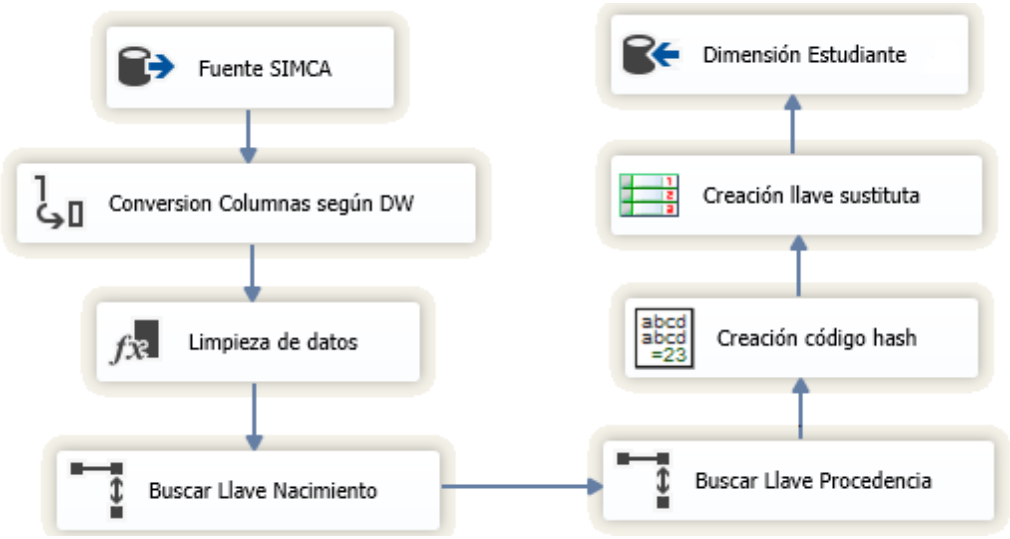


Figura 20 Cargue inicial estudiante.

Paquete de cargue incremental: El flujo de datos construido para este paquete (Ver Figura 21) es igual que el del cargue inicial hasta el componente denominado *Creación código hash.* Sin embargo, este paquete incremental cuenta con una segunda fuente de datos que tiene como tabla origen la dimensión *Estudiante* de la bodega de datos. Esta segunda fuente permite realizar la combinación con la fuente del sistema OLTP descrita en el paquete inicial, posteriormente se realiza una verificación por medio del componente denominado *Verificación filas nuevas y actualizadas,* el cual determina que filas deben ser agregadas como nuevas a la dimensión y cuáles deben ser actualizadas. Para el caso de las filas que deben ser actualizadas, se utiliza un componente que ejecuta una sentencia DML de actualización, por otro lado, para las filas nuevas por medio del componente denominado *Último valor llave primaria,* se obtiene el valor siguiente en la secuencia de la llave sustituta de la dimensión, con el fin de evitar problemas de integridad referencial. El destino de los datos es la tabla de dimensión *Estudiante* de la bodega de datos relacional y su configuración se realiza de la misma manera que en el paquete de cargue inicial.

Tabla de dimensión Localización

Paquete de carga inicial: En la Figura 22, se presenta el flujo de datos para este paquete, que se explica a continuación:

- Extracción: La dimensión *Localización* tiene una granularidad a nivel de *municipio*, por esta razón la fuente de datos utiliza una consulta que realiza una combinación entre las tablas países, departamentos y municipios por medio de la integridad referencial, además estas tablas se encuentran en la base de datos “ACADEMICO”.

- Transformación: Se validan algunas columnas para hacer la limpieza de los datos. Seguido a esto, se realiza la construcción de la llave sustituta y el código hash (sirve como insumo para la actualización y adición de filas en el paquete de cargue incremental).

- Cargue: El destino de los datos se configura teniendo en cuenta que la tabla de dimensión *Localización* almacenará la información proveniente del flujo.

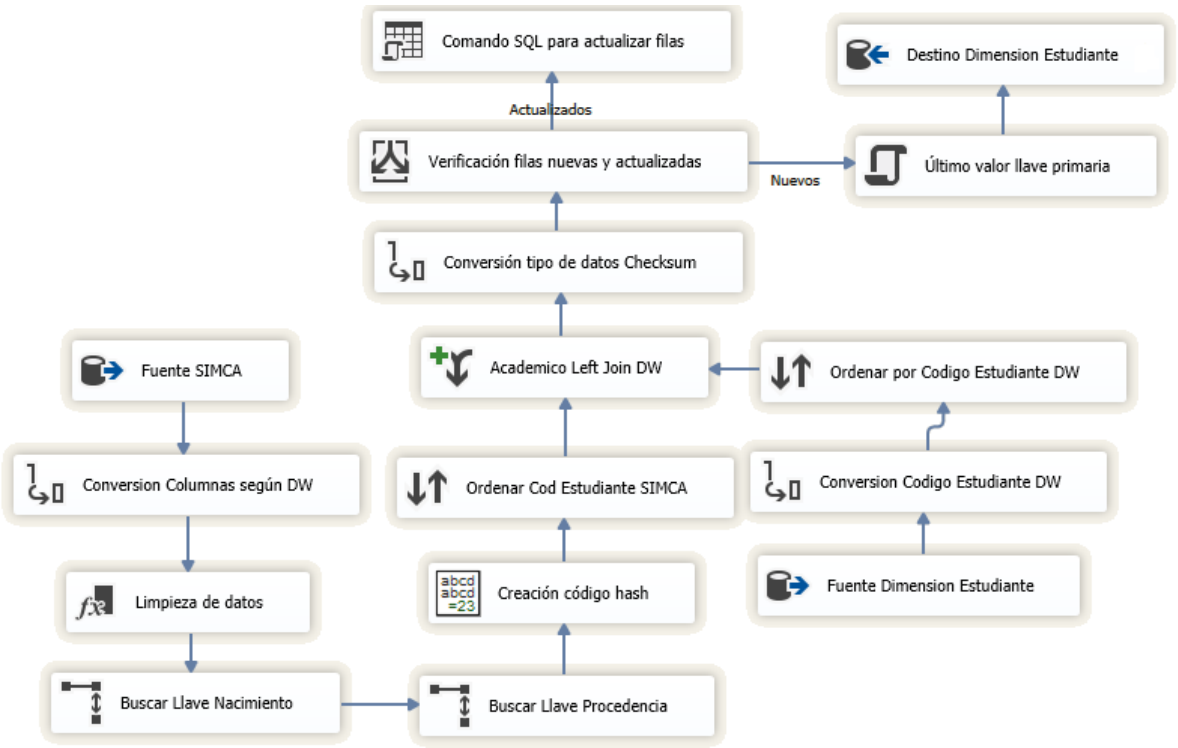


Figura 21 Cargue incremental estudiante.

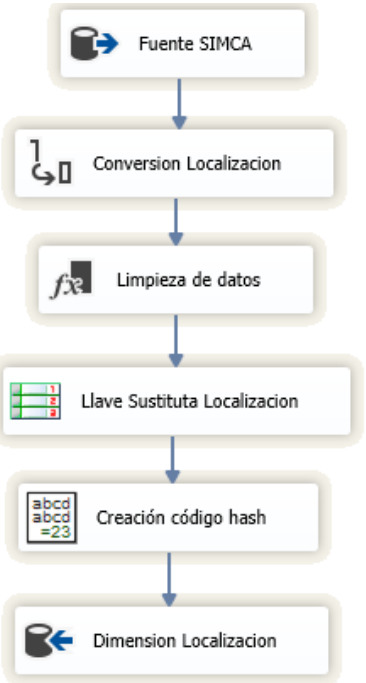


Figura 22 Cargue inicial Localización.

El paquete de cargue incremental: Este paquete (Ver Figura 23) tiene el mismo flujo de datos del cargue inicial hasta el componente denominado *Creación código hash,* de igual manera, como se realiza para el cargue incremental de la dimensión *Estudiante,* se configura la fuente de datos de la bodega relacional, se genera una combinación de dichas fuentes, para posteriormente verificar los nuevos registros y los que deben ser actualizados, y finalmente se configura el destino.

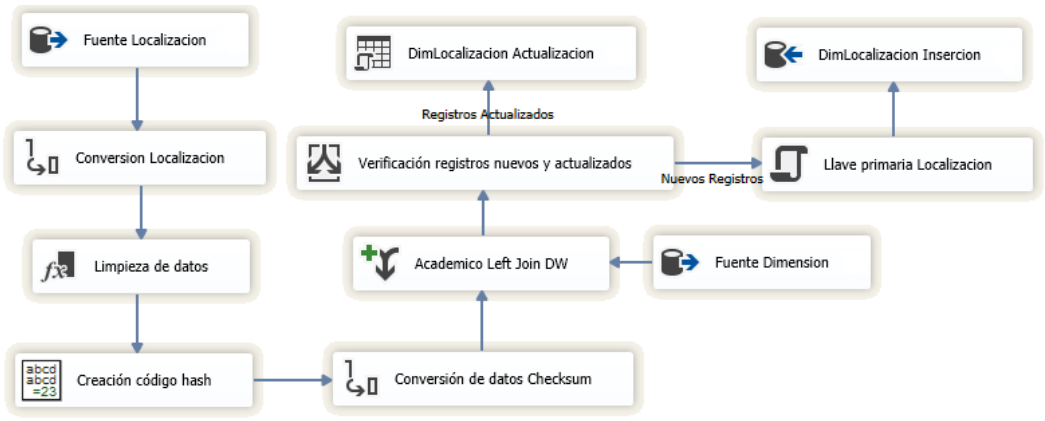


Figura 23 Cargue incremental Localización.

Tabla de dimensión Docente

Paquete de cargue inicial: Este paquete de cargue de datos es particular con respecto a los demás paquetes, ya que en la tabla de dimensión *Docente* se presentan atributos tipo dos[10] [4]. Estos tipos de atributos debe tener una serie de columnas auxiliares para poder ser gestionados en el cargue de datos incremental. A continuación se describen tales atributos:

a. Actual: Determina si el registro es actual, aquellos registros que no sean el actual, tendrán un valor de expirado.
b. Fecha Efectiva: Permite verificar cuando se añadió el registro a la tabla de dimensión.
c. Fecha Expiración: Fecha hasta cuando tiene vigencia dicho registro.
d. Checksum Tipo 2: Almacena el código hash que permite identificar las columnas que pueden cambiar en el tiempo y que se encuentran relacionadas a la vinculación del docente con la facultad y el departamento.
e. Checksum Tipo 1: Permite verificar el cambio de los atributos de la tabla exceptuando los atributos relacionados al tipo dos.

En la Figura 24 se presenta el flujo de datos para este paquete, que se explica a continuación:

- Extracción: El componente de fuente de datos ejecuta una consulta que extrae la información personal e institucional del docente.

- Transformación: La limpieza de los datos, al igual que en el cargue inicial de estudiante, evita los valores nulos, aunque en este paquete, adicionalmente, se asignan los valores

[10] Para algunos atributos que cambian en el tiempo, se debe agregar una nueva fila en la dimensión.

predeterminados para las columnas utilizadas en el cargue incremental de los atributos de tipo dos. En este paquete, a diferencia de los demás, se crearon dos componentes para la creación del código hash, el primero denominado *Creación código hash atributos tipo 1,* el cual genera dicho código partiendo de las columnas que cambian con el tiempo pero que deben sobrescribirse (Celular, Teléfono, entre otros); y el segundo, denominado *Creación código hash atributos tipo 2,* el cual genera el identificador de las columnas que cambian con el tiempo pero que necesitan un registro histórico (Departamento, Facultad y Sigla Facultad).

- Cargue: La tabla destino es la dimensión *Docente* de la bodega de datos relacional.

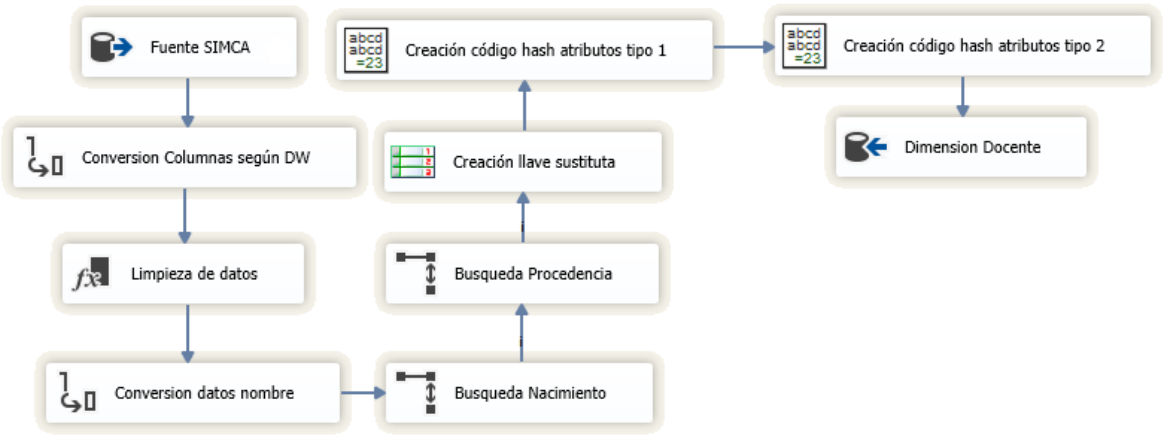


Figura 24 Cargue inicial docente.

Paquete de cargue incremental: Este paquete (Ver Figura 25) tiene el mismo flujo de datos que el cargue inicial hasta el componente denominado *Creación código hash atributos tipo 2,* después se realiza una combinación de fuentes de datos, entre la información del sistema OLTP y la tabla de dimensión *Docente.* Por medio del componente denominado *Verificación de actualización y nuevos,* se verifica que filas son nuevas para la dimensión, cuales deben actualizarse y cuales deben generar un nuevo registro por el cambio de alguno de los atributos tipo dos, dicha verificación se realiza por medio de los códigos hash generados anteriormente. Para los registros que son determinados como nuevos, se sigue el flujo normal como en el cargue inicial, sin embargo, la creación de la llave sustituta obtiene el último valor de la dimensión. En la actualización de los atributos de tipo dos, el componente denominado *Actualización columnas auxiliares,* permite asignar los valores que tendrán las columnas auxiliares, luego de modificar dichas columnas, el componente de control denominado *Actualización Fechas Tipo 2,* ejecuta la sentencia DML para la actualización de tales columnas en la bodega de datos relacional, finalmente, para este caso se utilizó el componente denominado *Unión de todo,* el cual permite combinar varias entradas para generar una salida con el fin de evitar la replicación del flujo de datos. La actualización de los demás atributos se realiza de igual manera como en el paquete de cargue incremental de estudiante.

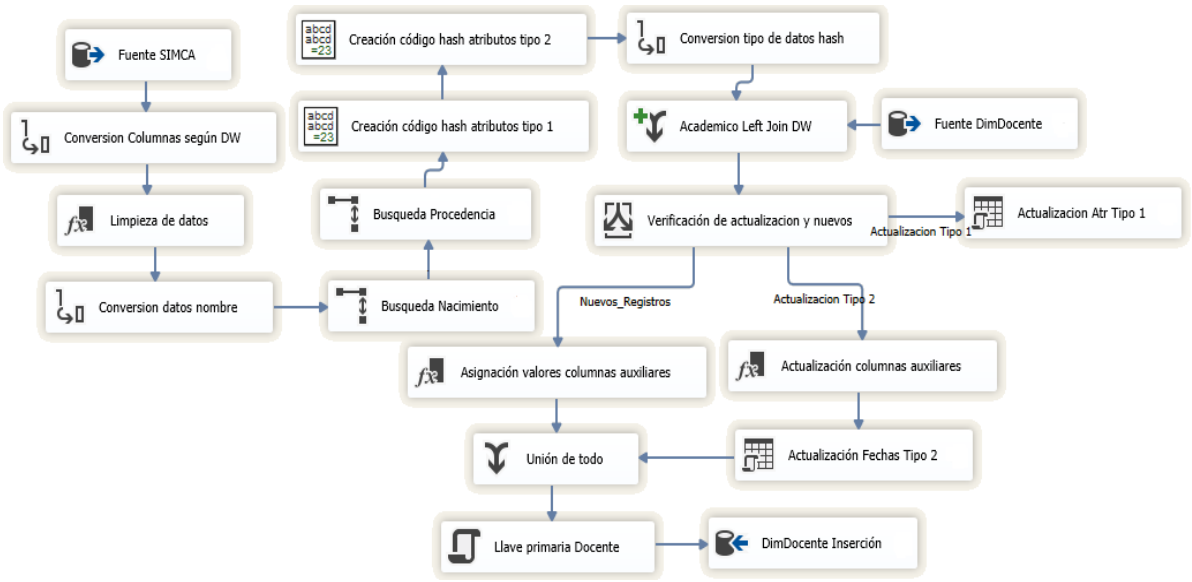


Figura 25 Cargue incremental docente.

Tabla Puente Materia – Docente

Paquete de cargue inicial: Para la ejecución de este paquete se debe tener cargada la tabla de hechos de Notas definitivas, para evitar problemas de integridad referencial, ya que esta tabla contiene una llave foránea que hace referencia a la tabla de hechos. En la Figura 26, se presenta el flujo de datos para este paquete.

- Extracción: Para este paquete fueron configuradas varias fuentes de datos, la primera permite extraer de la bodega de datos los grupos que fueron dados en cada período académico junto con la llave sustituta de la tabla de hechos; la segunda fuente de datos consiste en una consulta que extrae de la base de datos "ACADEMICO", los grupos que han sido dictados por los docentes, además de su información profesional, laboral y de residencia. El componente *Fuente Académico Grupos Docentes* es el encargado de ejecutar esta consulta; y las demás fuentes de datos son archivos Excel que contiene información más significativa para el usuario, ya que la información de la segunda fuente de datos no son más que valores numéricos (Ej.: permite convertir un "1" de la segunda fuente de datos en un "Doctorado" almacenado en los archivos Excel).

- Transformación: Luego de obtener la información de la primera fuente de datos, se realizan dos búsquedas para encontrar el identificador de la materia y el período en la base de datos "ACADEMICO", posteriormente se puede identificar el código del grupo, con el cual se relaciona cada docente. El componente *DW Join Academico,* realiza la combinación de la primera fuente de datos con la segunda, de esta manera se relaciona el identificador de la tabla de hechos con cada docente. Se realizan diferentes combinaciones con las fuentes Excel, además de una limpieza de los datos para evitar valores nulos, una vez hecho esto, se realizan las búsquedas de las llaves sustitutas de las dimensiones *Docente, Localización (Residencia) y Datos Adicionales Docente.*

- Cargue: La tabla destino es la tabla Materia-Docente de la bodega de datos relacional.

Paquete de carga incremental: Este paquete contiene el mismo flujo de datos que el paquete de cargue inicial, la diferencia se encuentra en la configuración del componente *Fuente Hecho Dw,* el cual contiene una cláusula WHERE que permite restringir los datos por período de tiempo. Por lo tanto, para este paquete el flujo se presenta en la Figura 26.

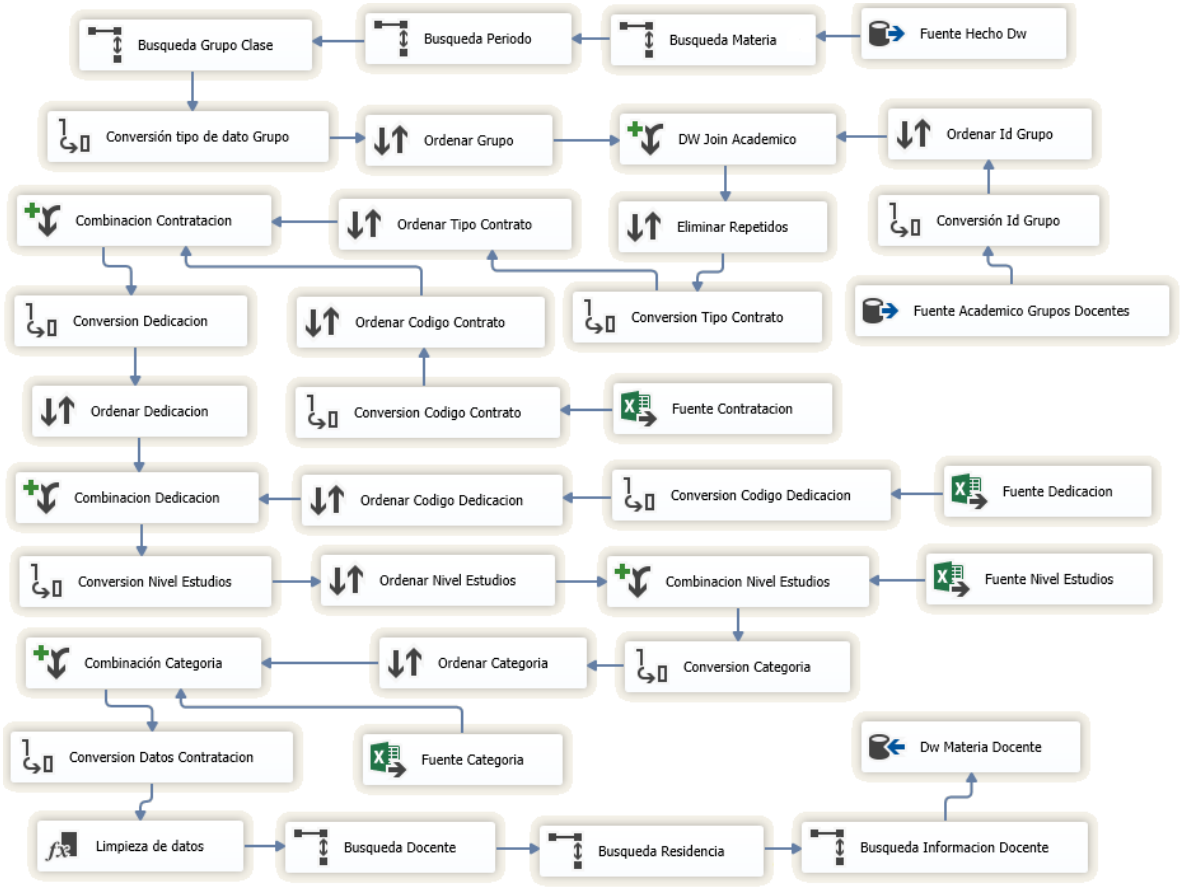


Figura 26 Cargue inicial puente materia-docente.

Recomendaciones

- Cuando se desean obtener datos de múltiples fuentes se debe considerar la construcción de vistas materializadas, ya que estas optimizan los tiempos de ejecución de los procesos de cargue y por lo general son utilizadas para en el proceso ETL de las tablas de hechos.
- La creación de la llave sustituta debe ser generada en el proceso ETL y no con la utilización de disparadores implementados en el DBMS, esto permite realizar un control y seguimiento sobre estas llaves desde el mismo paquete, de esta manera se asegura desde el proceso ETL la integridad referencial de los datos, además de la optimización del tiempo de creación.

A lo largo del proceso se presentaron problemáticas en el desarrollo del proyecto, estas son presentadas en la Tabla 29 junto a las recomendaciones que se consideran útiles para las respectivas soluciones.

Problemática	Recomendación de solución
Debido a la estructura del sistema OLTP, en el cual la información se encontraba distribuida en múltiples tablas, la realización del proceso ETL por medio de controles, específicamente la *Combinación de mezcla*, encargada de hacer una unión entre dos tablas u orígenes de datos elevaba el tiempo de ejecución del proceso.	Utilizar consultas SQL sobre los controles dados por SSIS, preferiblemente cuando se desean realizar procesos ETL con una gran cantidad de tablas y datos, ya que se realizaron pruebas con ambas formas y los tiempos de ejecución de las consultas fueron mejores.
Para la actualización o registro de nuevas filas en las dimensiones, se consideró inicialmente la comparación de todas las columnas. Al identificar un cambio en alguna de estas, se comprobaba si la fila había tenido alguna modificación, por otro lado, se comparaban aquellas columnas consideradas únicas, para determinar la inserción de los nuevos registros, esto adicionaba complejidad al flujo de datos, además de tomar un gran tiempo de ejecución.	Con la inclusión del componente *Checksum*, se generó un código hash que permite la identificación de cada fila, de esta forma los tiempos de ejecución en los procesos de actualización se redujeron considerablemente.
La carencia de índices en el sistema OLTP, se reflejaba en el tiempo de extracción de los datos. En la tabla de hechos como también en la tabla puente se consideraron tiempos de ejecución del flujo de datos demasiado altos.	Preferiblemente las fuentes de datos (OLTP) deben considerar un plan de indexación, ya que esto permite una extracción más rápida de los datos en los respectivos paquetes.

Tabla 29 Problemáticas y soluciones ETL.

5.1.1.4 Población y validación de datos

La población de los datos en la bodega relacional debe tener un orden específico en la ejecución de cada uno de los paquetes mencionados en la sección 5.1.1.3. Dado que algunas tablas de dimensiones como Estudiante y Docente contienen claves externas de otra dimensión, estas tablas de dimensiones deben poblarse después de las tablas a las que hace referencia la clave externa. Teniendo en cuenta lo anterior, la ejecución del cargue de los datos comienza con la tabla de dimensión Localización, después se realizó una ejecución paralela de los demás paquetes de las tablas de dimensiones y finalmente se ejecutó el paquete de cargue de datos de la tabla de hechos y la tabla puente, respectivamente.

La validación de los datos se realizó por medio de consultas, tanto en la bodega de datos relacional como en el cubo OLAP, comprobando que no existiesen datos nulos y que todas las notas estuviesen dentro del rango permitido.

5.1.2. Desarrollo Front-Room

El desarrollo del Front-Room está enfocado en la visualización de los datos procesados en la etapa del Back-Room. Las herramientas por medio de las cuales se hizo la presentación de los datos a los usuarios finales fueron: Excel y el servicio de reportes de SQL Server.

5.1.2.1 Identificación y priorización de reportes candidatos

La priorización de los reportes candidatos se basó en los requerimientos de los usuarios, los cuales se encuentran en el Anexo C. Fueron implementados los reportes de mayor prioridad, en la Tabla 30 se presentan los reportes más relevantes con los cuales se realizaron las pruebas de satisfacción.

Reporte	Prioridad	Usuario	Tipo
Cantidad de aprobados y reprobados junto con la repetición en la que se cursaba.	Alto	Decano FIET.	Reporte
Cantidad de estudiantes en cada período académico agrupados por desempeño académico.	Alto	Decano FIET.	Reporte
Cantidad de estudiantes admitidos por período académico discriminados por tipo de ingreso.	Alto	Coordinadora de PIAI.	Reporte
Cantidad de estudiantes que pierden un cierto número de materias.	Alto	Decano FIET.	Reporte
Cantidad de profesores inscritos por departamento.	Medio	Coordinador de PIS.	Reporte
Estudiantes vinculados a monitorias por período académico.	Alto	Coordinador del PIS y PIAI.	Reporte
Profesores del departamento adscritos respecto con dedicación de tiempo completo, medio tiempo y cátedra, según nivel de formación.	Alto	Coordinador del PIS y PIAI.	Consulta Ad-Hoc.
Número de profesores adscritos a la facultad y el departamento, por categorías académicas establecidas en el escalafón.	Alto	Coordinador del PIS y PIAI.	Consulta Ad-Hoc.
Desempeño académico de los estudiantes discriminados por carácter de la institución de procedencia y el tipo de ingreso a la universidad.	Alto	Decano FIET.	Consulta Ad-Hoc.

Lista de materias con cantidad de estudiantes reprobados discriminados por repetición y nivel académico.	Alto	Decano FIET.	Consulta Ad-Hoc.

Tabla 30 Priorización reportes del esquema de notas definitivas.

5.1.2.2 Diseño de la estrategia de navegación

Los reportes creados se organizaron en carpetas para que los usuarios pudieran encontrarlos de una manera más fácil en el servidor de reportes. Dado que los usuarios cuentan con un cargo específico en la universidad, se optó por organizar los reportes en carpetas, una para los reportes del decano y la demás para los reportes referentes de los coordinadores de programa de la FIET.

La estrategia de navegación en los reportes y las consultas Ad-Hoc se apoyó en las técnicas del "Drill-Down". Para presentar la información en los reportes, se hizo uso de una tabla de datos y de uno o más gráficos. En las consultas Ad-Hoc, el usuario podrá observar los atributos agrupados por carpetas, con el fin de que pueda identificar rápidamente la información relevante de cada dimensión.

5.1.2.3 Desarrollo de estándares para aplicaciones de usuario final

En el apartado 5.1.2.1 se presentaron dos tipos de reportes: los reportes creados con el servicio de reportes SQL Server y las consultas Ad-Hoc. En este apartado se explica los formatos utilizados para cada tipo de reporte.

Los reportes tuvieron un esquema sencillo, en el cual su cabecera cuenta con el logo de la Universidad del Cauca seguido del título del reporte (ver reporte gráfico en el Anexo H). El título fue establecido de una forma entendible para los usuarios y teniendo cuenta el contenido presentado en el reporte. El cuerpo del reporte estuvo compuesto por una tabla, la cual contiene la información, además de uno o varios gráficos.

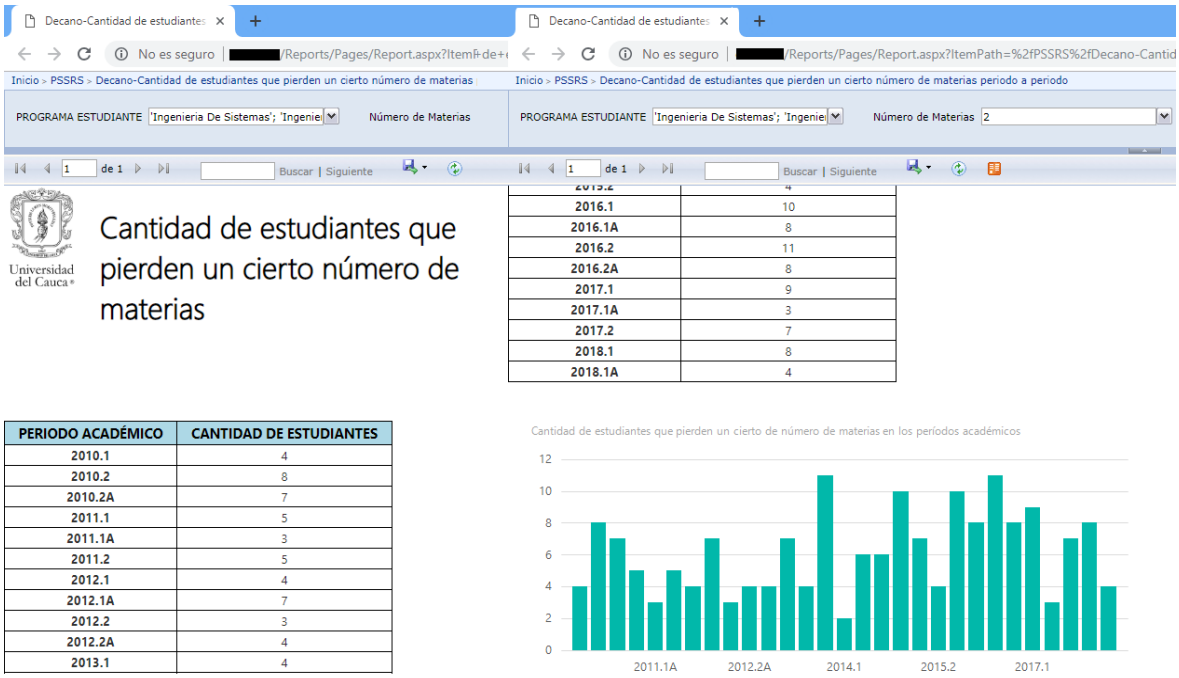
Las consultas Ad-Hoc se basaron en la tabla dinámica de Excel, por esta razón el formato general para este tipo de consultas estuvo apoyado en esta forma de visualización matricial. La creación de las consultas Ad-Hoc y su formato depende de la necesidad del usuario.

5.1.2.4 Selección enfoque de implementación

El enfoque de implementación seleccionado para el acceso a los datos, más específicamente a los reportes, fue a través de una herramienta basada en la Web. Este tipo de herramienta permitió a los usuarios acceder a las carpetas mencionadas en el apartado 5.1.2.2, a través de una dirección URL. Por otro lado el acceso a los datos para las consultas Ad-Hoc se hizo a través de la herramienta de escritorio Excel.

Ejemplos de reportes accedidos a través de la URL

En la Figura 27, se presenta el reporte denominado "Cantidad de estudiantes que pierden un cierto número de materias", el cual obtiene la información accediendo directamente a la bodega de datos relacional. En la Figura 28, se presenta el reporte denominado "Cantidad de estudiantes en cada período académico agrupados por desempeño académico", este reporte obtiene la información accediendo directamente al cubo OLAP.



2016.1	10
2016.1A	8
2016.2	11
2016.2A	8
2017.1	9
2017.1A	3
2017.2	7
2018.1	8
2018.1A	4

PERIODO ACADÉMICO	CANTIDAD DE ESTUDIANTES
2010.1	4
2010.2	8
2010.2A	7
2011.1	5
2011.1A	3
2011.2	5
2012.1	4
2012.1A	7
2012.2	3
2012.2A	4
2013.1	4

Figura 27 Reporte relacional desde PC.

	Superior: 4.6-5.0	Alto: 4.0-4.5	Basico: 3.0-3.9	Bajo: 2.0-2.9	Deficiente: 0.0-
2010.1	1002	1110	1029	1014	1032
2010.2	1031	1056	1025	1043	1061
2011.1	1018	1125	1045	1050	1046
2011.2	1067	1026	1046	1012	1019
2012.1	989	1024	1021	1038	1023
2012.2	1051	1006	1067	1022	1089
2013.1	1008	1067	1039	1051	990
2013.2	1047	1037	1070	1069	1079
2014.1	995	1073	1000	1053	1006
2014.2	1057	1071	996	1011	1049
2015.1	1094	1090	1072	1074	1046
2015.2	1013	1094	1086	1033	1042
2016.1	1083	1059	1048	1091	1046
2016.2	1055	1025	1002	1081	1047
2017.1	1051	1032	998	1091	1052
2017.2	1108	998	1025	1028	1037
2018.1	1006	1066	1056	1129	1054

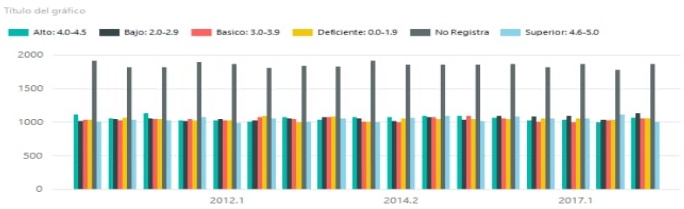


Figura 28 Reporte OLAP desde móvil.

5.2. Esquema notas por componentes

En este apartado se presentan las actividades desarrolladas con respecto al esquema de componentes, el cual presenta un mayor nivel de granularidad, aumentando el número de registros en la tabla de hechos con respecto a la tabla de hechos del esquema de notas definitivas.

5.2.1. Desarrollo Back-Room

Las herramientas seleccionadas para este esquema fueron las mismas que para el esquema anterior, ya que los dos esquemas (Notas definitivas y Notas componentes) conforman el prototipo completo.

5.2.1.1 Diseño físico de tablas

De igual manera que para el esquema de notas definitivas, este esquema se apoyó en el diseño físico de tablas propuesto en el Anexo F.

Estándares de nombrado para la bodega de datos relacional.

Los estándares de nombrado para este esquema son los mismos expuestos en el apartado 5.1.1.1.

Modelo físico de datos y tamaño de la bodega de datos.

Igualmente que en el esquema de notas definitivas, este modelo físico de datos se apoyó en el modelo dimensional propuesto en el Capítulo IV y en los estándares de nombrado.

Para la estimación de tamaño a corto y largo plazo de este esquema, es necesario tener en cuenta las recomendaciones presentadas en la sección 5.1.1.1, además del número de componentes promedio que se crean para los cursos (normalmente un curso tiene dos o tres componentes, pero no existe restricción con respecto al número de componentes máximo).

Plan de indexación inicial

De la misma forma como se menciona para el esquema de notas definitivas, el plan de indexación se basó en [37]. Para la indexación de este esquema también se recomienda hacer uso de los índices de mapa de bits, pero no se descarta el posible uso de los índices de Árbol B.

Tabla de hechos

La tabla de hechos de este esquema contiene una gran cantidad de datos, por lo anterior, esta tabla de hechos es un punto crítico para el rendimiento de las consultas, de esta manera esta tabla debe tener una indexación adecuada que permita una optimización en el rendimiento de las consultas.

Como se mencionó en el apartado del plan de indexación de la sección 5.1.1.1, para que los atributos en una tabla de hechos sean candidatos a utilizar índices de mapa de bits deben cumplir con la condición 1 y una de las condiciones del numeral 2 mencionadas en dicha sección. Todos los atributos en la Tabla 31, cumple con condición 2, por ser llaves foráneas en la tabla de hechos. Al igual que en el esquema anterior, se determinaron los atributos candidatos por medio del límite porcentual mayor o igual a 80%, de esta manera los atributos resaltados en “negrilla” son candidatos a utilizar este tipo de índices. Para los campos que no cumplieron con dichas condiciones se optó por no establecerles ningún tipo

de índice. En cuanto a la clave primaria simple, se debe crear un índice de Árbol B debido a su unicidad y por ende a su alta cardinalidad. Por último, las medidas no contarán con índices, ya que, a pesar de que puedan cumplir con la condición 1**,** no cumplen con la condición 2**,** ya que no se encontrarán como restricciones en las clausulas WHERE y no son llaves foráneas.

Atributo	Número de valores distintos	Número de valores distintos en 100 o más filas	Relación
componente_id	4	4	100,00%
contratacion_id	12	12	100,00%
dato_ad_materia_id	50	47	94,00%
dem_economicos_id	1241	1133	91,30%
demografia_id	142	86	60,56%
docente_id	916	796	86,90%
estudiante_id	7000	3691	52,73%
indicadores_id	22	16	72,73%
materia_id	491	471	95,93%
pensum_est_id	5	5	100,00%
pensum_mat_id	6	6	100,00%
período_id	29	28	96,55%
programa_estudiante_id	5	5	100,00%
programa_materia_id	6	6	100,00%
residencia_docente_id	2	2	100,00%
residencia_estudiante_id	2	2	100,00%
semestre_id	318	194	61,01%

Tabla 31 Candidatos mapa de bits, notas componentes.

Tabla puente

Dado que esta tabla funciona solo como conexión entre la tabla de hechos y dimensiones, se recomienda no hacer uso de una llave primaria la cual implicaría un índice de Árbol B compuesto y esto conllevaría a la generación de un índice con un tamaño muy grande, incluso podría llegar a ocupar más espacio que los mismos datos, por lo cual podría afectar el rendimiento de las consultas. Sin embargo, debido a que las tablas puente están compuestas por llaves foráneas, se hizo uso de la evaluación de las condiciones expuestas en el apartado 5.1.1.1, por lo tanto, en la Tabla 32, se presentan las columnas que cumplen con limite porcentual propuesto (resaltadas en “negrilla”), y por ende las candidatas a utilizar índice de mapa de bits.

Atributo	Número de valores distintos	Número de valores distintos en 100 o más filas	Relación
hecho_id	107716	0	0,00%
supletorio_id	1	1	100,00%

Tabla 32 Candidatos mapa de bits, puente componentes.

Para la indexación de las dimensiones Componente y Supletorio que son particulares a este esquema, se recomienda indexar de la misma manera como se presentó en 5.1.1.

Plan de particionamiento
Como se mencionó en el esquema de notas definitivas, se recomienda realizar el particionamiento por períodos académicos, por lo cual en el cargue de datos se debe crear una llave sustituta para la dimensión período (ejemplo: para el primer período del año del año 2012 la clave sustituta sería 201201) con un significado claro.

Plan de agregación
Debido a que las necesidades analíticas de los usuarios no se enfocaban en este esquema, hubo una cantidad de reportes limitada, por lo cual no se generó un plan de agregación para este esquema, adicional a esto, se desconocen las consultas frecuentes, por lo tanto, no se puede generar una optimización basada en el uso.

5.2.1.2 Implementación de la bodega de datos relacional

En este apartado se presentará el tamaño de las dimensiones y tabla de hechos de este esquema con sus respectivos índices. Además, se explicará la forma en la que fueron creados.

Tamaño de tablas e índices
De igual manera que en el esquema de notas definitivas, el tamaño de este esquema sirve como referencia para un sistema de este tipo en un ambiente de producción. El tamaño de las tablas e índices particulares a este esquema se presenta en la Tabla 33. Las demás tablas e índices son presentadas en el Anexo F.

Objeto de la bodega de datos	Tipo	Número de filas	Tamaño en disco (MB)
COMPONENTE	TABLA	4	0,0625
COMPONENTE _PK	ÍNDICE	NO APLICA	0,0625
SUPLETORIO	TABLA		0,0625
SUPLETORIO_PK	ÍNDICE	NO APLICA	0,0625
HECHO_COMPONENTES	TABLA	634913	56
HECHO_COMPONENTES _PK	ÍNDICE	NO APLICA	11
PUENTE_SUPLETORIO	TABLA	107716	2
PUENTE_INDEX_HECHO	ÍNDICE	NO APLICA	2

Tabla 33 Tamaño del esquema notas por componentes.

Como se puede observar en la Tabla 33, el tamaño de la tabla de hechos depende del detalle en la granularidad.

Creación de tablas e índices
La creación de tablas e índices de este esquema, se realizó de la misma forma como se presenta en el apartado 5.1.1.2.

5.2.1.3 Desarrollo proceso ETL

El proceso ETL fue desarrollado, al igual que en el esquema de notas definitivas, utilizando las mismas herramientas.

Fuentes de Datos
Las fuentes de datos utilizadas para este esquema fueron las mismas que en el esquema de notas definitivas, mencionadas en el apartado 5.1.1.3. Sin embargo, la única diferencia radica en que este proceso tuvo en cuenta tablas adicionales de la base de datos "ACADEMICO", como la tabla notas componentes y notas configuración.

Proceso ETL
El proceso ETL para este esquema, al igual que para el esquema de notas definitivas, fue desarrollado por medio de componentes denominados paquetes.

Debido que para el desarrollo del prototipo se tuvo en cuenta el concepto de dimensión conformada, en este apartado no se realizará la descripción de tales dimensiones, ya que la mayoría son descritas en el Anexo G. Sin embargo, en esta sección se describirán los cargues para la dimensión *Componente y* la tabla de hechos. A continuación, se describe cada uno de los paquetes desarrollados:

Tabla de dimensión componente
Paquete de cargue inicial: En la Figura 29 se presenta el flujo de datos correspondiente a este paquete.

- Extracción: La información fue extraída desde una sola tabla de la base de datos "ACADEMICO", la cual contiene la información que describe el porcentaje del componente y el tipo (nota final o parcial).

- Transformación: El componente *Transformación Tipo Componente,* permite modificar los valores de la columna tipo con información más significativa al usuario (Ej: el valor F se reemplaza con Final). La creación del código hash permite generar el identificador con respecto a las columnas que describen el componente y el tipo, con el fin de determinar en el cargue incremental solo los nuevos registros. La llave sustituta para cada registro se genera por medio del componente denominado *Creación llave sustituta.*

- Cargue: El destino de los datos es la dimensión *Componente* de la bodega de datos relacional.

Paquete de cargue incremental: El flujo de datos correspondiente a este paquete se presenta en la Figura 30.

El flujo de datos es el mismo al del cargue inicial, específicamente hasta el componente denominado *Creación código hash,* una vez hecho esto, se realiza una búsqueda utilizando dicho código, con el fin de determinar las filas que no coinciden con ningún registro de la dimensión *Componente,* las cuales serán los nuevos datos a ingresar. Adicionalmente, se obtiene el último valor de la llave sustituta para evitar inconvenientes de integridad referencial. El destino de los datos es el mismo presentado en el paquete de cargue inicial.

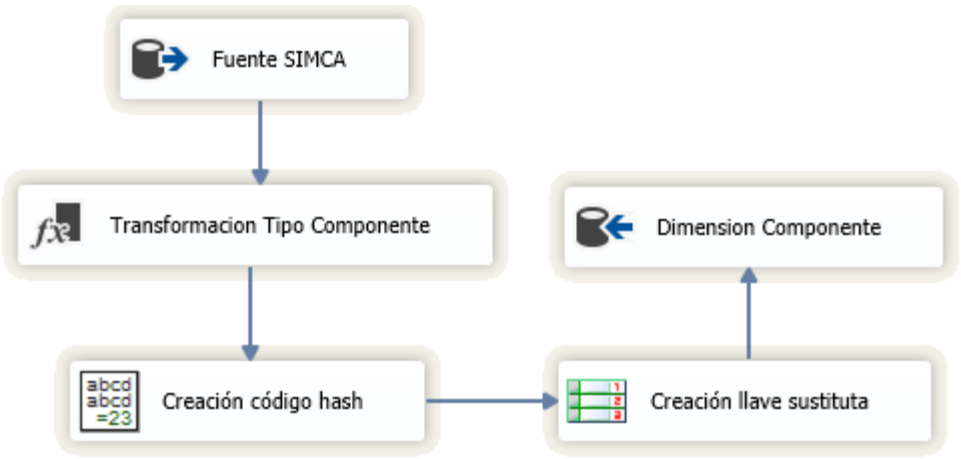


Figura 29 Cargue inicial componentes.

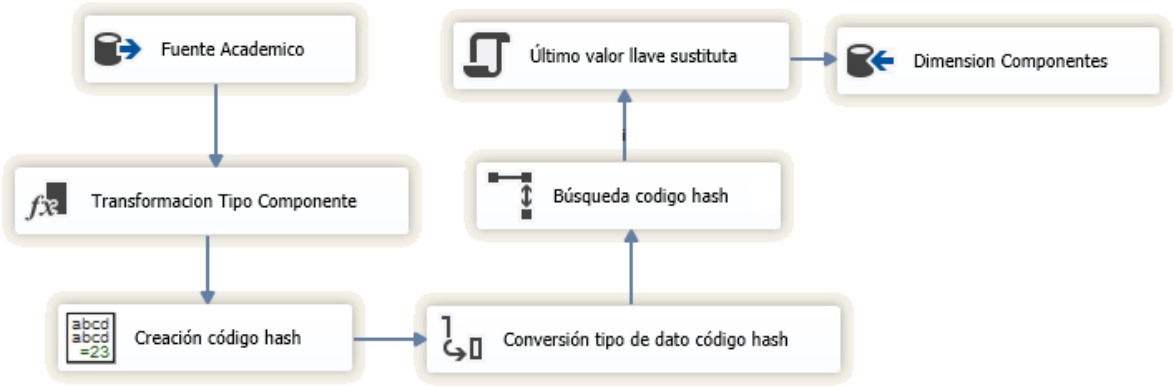


Figura 30 Cargue incremental componente.

Tabla de hechos notas componentes

Paquete cargue inicial: En este paquete (Ver Figura 31), de igual manera que para la tabla de hechos del esquema de notas definitiva, se hizo uso de las vistas materializadas ofrecidas por el DBMS seleccionado para el prototipo.

- Extracción: El componente *Fuente SIMCA* extrae la información de la vista materializada, lo que evita ejecutar consultas muy complejas, de esta manera optimiza el rendimiento en el proceso de ejecución del flujo de datos del paquete.

- Transformación: Para cada llave foránea de la tabla de hechos se creó un componente de búsqueda que tiene como dato de entrada la información obtenida en la vista materializada, esto con el fin de obtener la llave sustituta de cada tabla de dimensión. El componente denominado *Llave sustituta* genera un valor único para cada registro de la tabla de hechos.

- Cargue: El destino de los datos es la tabla de hechos, la configuración del destino relaciona cada llave obtenida de las búsquedas con las columnas de la tabla.

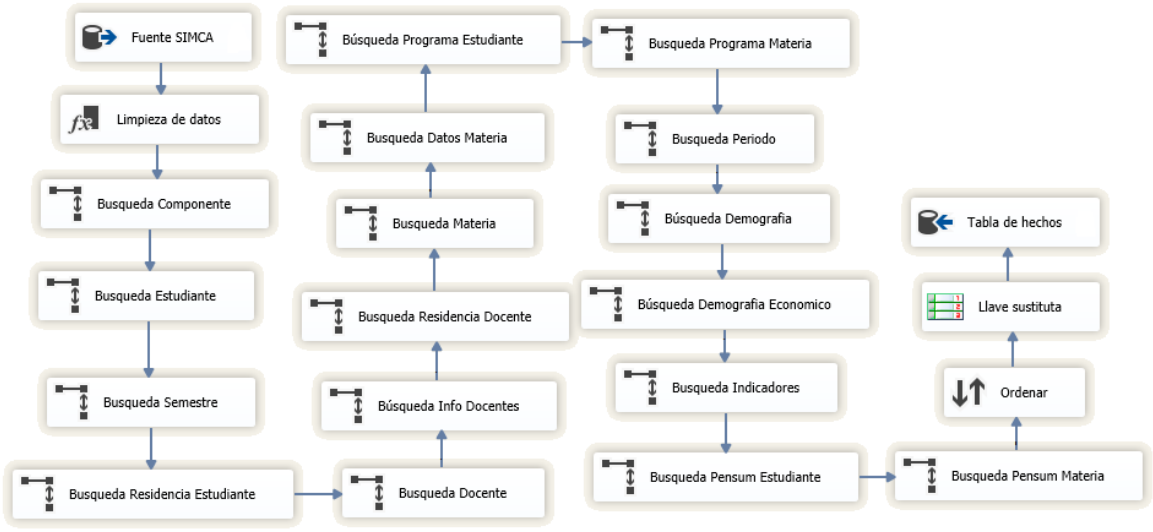


Figura 31 Cargue inicial hecho componentes.

Paquete de cargue incremental: Este paquete es el mismo del paquete inicial, las diferencias son, la restricción de los datos en el componente fuente por medio de una clausula WHERE, y la creación de la llave sustituta que permite obtener el último valor de la llave primaria de la tabla de hechos.

5.2.1.4 Población y validación de datos

La población y validación de los datos se realizó de la misma forma como describió en la sección 5.1.1.4.

5.2.2. Desarrollo Front-Room

En este apartado se presentan las actividades desarrolladas en esta subfase.

5.2.2.1 Identificación y priorización de reportes candidatos

Se desarrollaron reportes de prioridad baja, debido a que la mayoría de necesidades analíticas estuvieron enfocadas en el esquema de notas definitivas. Además el usuario que presentó una necesidad analítica relacionada a este esquema, no contó con la disponibilidad de tiempo para poder realizar las respectivas pruebas. Por lo anterior, en la Tabla 34 se presentan los reportes, uno requerido por un usuario y el otro considerado relevante por el equipo de desarrollo.

Reporte	Prioridad	Usuario	Tipo
Cantidad de supletorios por corte en cada materia. Diferenciados por la razón inicial.	Baja	Antiguo técnico administrativo decanatura FIET.	Reporte
Desempeño académico de los estudiantes diferenciando componentes parciales y finales a través de todos los períodos académicos.	Baja	Equipo de desarrollo.	Consulta Ad-Hoc.

Tabla 34 Priorización de reportes esquema de notas componentes.

5.2.2.2 Diseño de la estrategia de navegación

La estrategia de navegación establecida para los reportes y las consultas Ad-Hoc de este esquema fue la misma presentada en la sección 5.1.2.2. Aunque, los reportes fueron agrupados en una sola carpeta denominada “Reportes por corte”, de esta forma se pueden identificar los reportes que pertenecen a este esquema. Igualmente, que en el desarrollo del Front-Room en el esquema de notas definitivas, se utilizaron las técnicas de “Drill-Down” en los reportes estándar y las consultas Ad-Hoc.

5.2.2.3 Desarrollo de estándares para aplicaciones de usuarios final

Los estándares tenidos en cuenta para los reportes y consultas Ad-Hoc relacionados a este esquema, fueron realizados de igual manera como se especificó en la sección 5.1.2.3.

5.2.2.4 Selección enfoque de implementación

El enfoque de implementación de los reportes y las consultas Ad-Hoc, fue el mismo propuesto en la sección 5.1.2.4.

5.3. Cubo OLAP

Para el desarrollo de la bodega de datos en este trabajo de grado, se ha seguido la metodología descrita en el Capítulo III. Sin embargo, en esta no se presenta una descripción clara sobre la construcción del cubo OLAP. Por esta razón en este apartado se realizará una especificación completa de este componente del prototipo.

Para el tipo almacenamiento del cubo se recomienda utilizar MOLAP, ya que este permite comprimir los datos y de esta manera evitar mucho almacenamiento en disco, como se menciona en [32].

La creación del cubo fue realizada sobre la herramienta SQL Server Analysis Services, configurando inicialmente los siguientes componentes:

- Origen de datos: Permite establecer la conexión a la bodega de datos relacional.
- Vista de origen de datos: Presenta una visión general de las tablas que conforman la bodega de datos relacional, adicionalmente dentro de este se pueden adicionar columnas calculadas.

El cubo OLAP se creó con los dos esquemas, almacenando las medidas de cada esquema (a partir de este momento denominados grupos de medidas), se consideró un solo cubo, ya que facilita el mantenimiento y el plan de seguridad basado en roles se define una sola vez. Las relaciones entre los grupos de medidas y las dimensiones deben ser configuradas, por esto, en el desarrollo de este cubo se establecieron diferentes tipos de relaciones:

- Normales: Relación típica entre un grupo de medida y una dimensión, es decir, cuando una dimensión está relacionada directamente con la tabla de hechos.
- Varios a varios: Permite la implementación a nivel dimensional de la tabla puente, esta es establecida como un grupo de medida. Para este tipo de relación fue necesario crear una dimensión que hiciera referencia a la tabla de hechos, esto con el fin de poder establecer la conexión de dimensiones con los grupos de medida de la tabla puente.
- Referenciada: Permite la implementación del caso de diseño subdimensión, en el cual una dimensión accede al grupo de medida por medio de una dimensión intermedia.

Estas relaciones permitieron establecer el vínculo entre cada dimensión y el grupo de medida. En los esquemas de notas definitivas y de componentes, se presenta en ambos una tabla puente, por esta razón se crearon dos grupos de medidas adicionales. Para el esquema de notas definitivas se creó el grupo de medida que hace referencia a la tabla materia-docente en la bodega relacional, y para el esquema de notas componentes el grupo de medida fue creado a partir de la tabla puente supletorio.

Luego, se procedió a crear las jerarquías y/o carpetas que permiten la agrupación de los atributos para una mejor organización y por ende un mejor entendimiento para el usuario. La diferencia entre las jerarquías y las carpetas radica principalmente en el orden que pueden tomar los atributos para la navegación de los datos. En las jerarquías se establece un orden predeterminado que no puede ser modificado, sin embargo, en las carpetas se deja el establecimiento del orden al usuario. La mayoría de dimensiones tienen agrupados sus atributos por carpetas, algunos fueron estructurados en jerarquías, teniendo en cuenta que el orden de navegación siempre es el mismo, por ejemplo: la navegación de la dimensión localización siempre comenzará desde la información más agrupada (país) hasta la más detallada (municipio).

Los miembros calculados, por lo general, hacen uso del lenguaje MDX (Expresiones multidimensionales), sin embargo, en este prototipo para algunos de estos no fue necesario este tipo de expresiones. A continuación, se presenta una descripción general de los miembros calculados agrupados por grupo de medida:

- Hecho registro académico: En este grupo de medidas los miembros calculados se enfocaron principalmente en la creación de promedios, ya sea de notas o de faltas, además, en este grupo de medida se utilizaron sentencias MDX simples para crearlos.

- Hecho semestral docente: La medida que permite saber el número de docentes, se generó sobre este grupo de medida, ya que la llave foránea del docente se encuentra almacenada en la tabla materia-docente.

- Hecho semestral estudiante: En este grupo de generó la medida conteo de estudiantes, haciendo uso de la función de agregación de un conteo distinto.

- Hecho componentes: La mayoría de miembros calculados en este grupo de medida, están enfocados a la generación de promedios. El promedio de la nota componente se realiza con un cálculo simple utilizando sentencias MDX, Sin embargo, la nota parcial (como también la nota parcial no promedio) fue calculada de diferente manera, ya que se creó una columna derivada (denominado cálculo con nombre) sobre la vista de origen de datos, la cual es el producto de la nota componente por su respectivo porcentaje. Aunque esta medida puede ser calculada por medio de MDX, se prefirió el cálculo con nombre, debido a que, este se evalúa cuando se procesa el cubo, por lo tanto es almacenado en la base de datos multidimensional, mientras que la consulta MDX se realiza en tiempo de ejecución, lo que afectaría considerablemente el rendimiento.

Normalmente se recomienda hacer uso de cálculos MDX siempre y cuando no sean tan complejos, en tal caso se sugiere verificar posibles cálculos en la vista de origen de datos.

Adicionalmente, en el cubo OLAP contiene elementos que permiten la mejor visualización y organización de las dimensiones y grupos de medidas. Estos elementos son:

- Traducciones: Permite establecer nombres para los grupos de medida, medidas, dimensiones, atributos y perspectivas, en un idioma en específico (en este caso, Colombia), para mejorar el entendimiento de los usuarios.

- Perspectivas: Filtran la visualización de los componentes de la bodega de datos. En este caso, se crearon dos: la primera permite visualizar los grupos de medida correspondiente a las notas definitivas con sus respectivas dimensiones; y la segunda, para los grupos de medidas y dimensiones relacionadas a las notas por componente.

5.3.1. Agregaciones

Las agregaciones generadas en el cubo OLAP se realizaron teniendo en cuenta las pautas presentadas en [39]. Estas se enfocaron en mejorar el rendimiento de los reportes implementados en este prototipo. Aunque, ya que este cubo no se desplegó en un ambiente de producción no fue posible realizar una optimización real basada en el uso.

Cada grupo de medida puede contar con uno o más diseños de agregación, los cuales permiten crear diversas agregaciones que estarán definidas por una serie de atributos. El diseño de agregaciones se basa en seleccionar un número correcto de agregaciones teniendo en cuenta el costo/beneficio de estas [39]. Al definir muchas agregaciones el rendimiento puede mejorar considerablemente, pero se ocupa más espacio de almacenamiento, afectando el tiempo de procesamiento del cubo. Por lo anterior, en [40], se recomienda que “Las agregaciones deben tener un tamaño menor a un tercio del tamaño de la tabla hechos”. Según [39], las agregaciones se determinan como una posible combinación del producto cartesiano de todos los atributos en el cubo, por esta razón el número de agregaciones podría llegar a ser muy grande en cubos que cuenten con un gran número de atributos. Sin embargo, SSAS considera solo un pequeño grupo de estas combinaciones posibles, indicando: “Un diseño de agregación eficaz generalmente contiene decenas o cientos de agregaciones, no miles”.

Para ayudar a los desarrolladores en la creación del diseño de agregaciones, SSAS utiliza un algoritmo que permite realizar un estudio costo/beneficio de las agregaciones, brindando dos tipos de diseños: Diseño de agregación, para crear agregaciones basadas en el diseño del cubo y la distribución de datos; y de Optimización basado en el uso, que permite crearlas teniendo en cuenta las consultas más frecuentes por los usuarios, esto lo hace por medio de un registro de consultas.

Aunque la configuración de uso de agregación de los atributos puede ser realizada por medio de un asistente que ofrece SSAS, no se recomienda, ya que el algoritmo no identifica el uso agregación para cada atributo, es por esto que el diseño de agregaciones se realizó de forma manual, especificando para cada atributo la propiedad del uso de agregación [39]:

- Completo: Esta configuración se debe aplicar cuando se tenga la certeza que un atributo será utilizado en muchas de las consultas (Ej: Período académico). No se recomienda aplicar este tipo de uso de agregación a atributos con alta cardinalidad, ya que los atributos configurados como *Completo* son tenidos en cuenta para todas las

agregaciones, por lo tanto el tamaño de cada agregación sería demasiado alto, lo cual generaría costos de almacenamiento.

- Ninguno: Con esta configuración ninguna agregación puede incluir tal atributo, esta se aplica a aquellos atributos que no son incluidos regularmente en las consultas (Ej: Libreta militar).

- Sin restricciones: Aquellos atributos configurados con este valor, son evaluados por el algoritmo de SSAS al momento de crear las agregaciones, para determinar si el atributo se reconfigura como *Completo,* como *Ninguno* o como *Predeterminado,* aquí se deben ubicar aquellos atributos pertenecientes a una jerarquía normalmente usada en las consultas, por ejemplo, para los atributos de la jerarquía de localización, aunque esta configuración también puede ser usada para atributos que sean incluidos ocasionalmente en las consultas, en [39] no se recomienda incluir más de diez atributos por dimensión en esta configuración, ya que el algoritmo tomaría demasiado tiempo para clasificarlos.

- Predeterminado: El diseñador aplica una regla predeterminada basada en el tipo de atributo y dimensión. Esta es utilizada cuando se desconozca la frecuencia de uso del atributo, pero deba ser considerado para algunas consultas. Si no se realiza el uso de agregación de forma manual, SSAS configura todos los atributos como *Predeterminados.*

Luego de establecer el uso de agregación para cada uno de los atributos, es necesario tener en cuenta el tamaño del cubo. Ya que el cubo creado para el prototipo es pequeño (la tabla de hechos cuenta con miles de valores y no millones o más), es posible generar un incremento en el rendimiento entre 20% y el 30%, como se menciona en [39]. Adicionalmente se hizo una simulación para optimizar las consultas utilizando el asistente basado en el uso, esta simulación consistió en consultar recurrentemente los reportes estándar y posteriormente aplicar dicha optimización. Finalmente, en el Anexo I, se puede observar la clasificación de los atributos de cada dimensión y la descripción de la configuración necesaria en la herramienta para realizar la optimización basada en el uso.

5.4. Integración

La integración, que corresponde a la unión de los componentes creados en la fase de Back-Room y el Front-Room se hizo por medio de una solución en SQL Server, compuesta de la definición cubo OLAP, los paquetes del proceso ETL y los reportes estándar. Las sentencias para la creación de la bodega de datos relacional, no fueron almacenadas en la solución (se almacenaron en archivos aparte).

Al final de esta fase, según la metodología [35], el representante de los usuarios debe firmar una carta de aceptación, en este caso este artefacto fue reemplazado por los resultados obtenidos en la evaluación del prototipo, la cual es presentada en el Capítulo VI.

5.5. Despliegue

El despliegue es la subfase en la cual se pone a disposición el sistema a los usuarios, sin embargo, para este proyecto no se realizó un despliegue en un ambiente real, ya que los datos manejados en este prototipo no fueron reales como se mencionó anteriormente y además no fue posible la gestión de un servidor para poder ejecutar esta subfase. Como

se menciona en la sección 3.5.5, el sistema fue instalado en un equipo de una sala de cómputo de la FIET, por lo cual algunas actividades de esta fase no fueron desarrolladas.

5.5.1. Verificar la preparación de los equipos para llevar a cabo la instalación

La verificación del equipo estuvo basada en determinar los requisitos mínimos de este y la compatibilidad con las herramientas utilizadas en el ambiente de desarrollo. Teniendo en cuenta lo anterior, las características del equipo a nivel de hardware y software, son las siguientes: El equipo cuenta con una capacidad en el disco de duro de 500 Gigabytes (GB), la memoria física o RAM con un tamaño de 8 GB, un procesador doble núcleo AMD A10, velocidad del procesador de 2,8 Ghz y marca del equipo denominada Lenovo. A nivel de software el equipo cuenta con un sistema operativo Windows 10 Professional, SQL Server 2014, SQL Server Data Tools 2015, Oracle en su versión 11g y Microsoft Excel 2013.

5.5.2. Diseño de una estrategia de capacitación de usuarios

La estrategia de capacitación a los usuarios, consistió en una explicación del objetivo de una bodega de datos, la información almacenada en el prototipo y los posibles análisis a generar. Posteriormente, se presentó una explicación de la estructura del cubo OLAP por medio de Excel realizando una descripción de las medidas y las dimensiones, generando una consulta para mostrar el funcionamiento. Adicionalmente se presentó el funcionamiento de los reportes estándar, haciendo una explicación de la forma en la que se tiene acceso a estos, las carpetas en las que se encuentran organizados, los parámetros tenidos en cuenta en cada reporte y la organización estructural de los mismos. El resultado de la capacitación se presenta en el Capítulo VI.

5.5.3. Definir una estrategia de soporte a usuarios

Debido a que el sistema no fue desplegado en un servidor y no quedó en funcionamiento, esta actividad no fue realizada.

5.5.4. Definir una estrategia de liberación

De igual manera que para la actividad presentada en la sección anterior, esta es omitida ya que el sistema no fue puesto en producción en un ambiente real.

5.5.5. Evaluar la disposición para el despliegue

A pesar de que el sistema no fue desplegado en un ambiente real, la evaluación de la disposición fue determinada en las características del equipo de la sala mencionadas en la sección 5.5.1, esto con el fin de reconocer el cumplimiento de todas las condiciones necesarias para llevar a cabo el despliegue.

5.5.6. Capacitación de los usuarios

La capacitación de los usuarios se basó en la estrategia planteada en la sección 5.5.2, esta actividad se realizó en la misma reunión programada para la ejecución de las pruebas de satisfacción, presentadas en el Capítulo VI, esto debido a la disponibilidad de tiempo de los usuarios.

5.6. Mantenimiento y crecimiento

Como se mencionó en el Capítulo III, esta fase fue omitida ya que el sistema no fue desplegado en un ambiente real, por lo tanto no se pueden ejecutar las actividades concernientes a esta fase.

Capítulo VI. Evaluación del prototipo

En este capítulo se presenta la explicación de la evaluación realizada al prototipo generado en este proyecto, el cual se menciona en el Capítulo V. Por medio de esta evaluación, se cumple con el objetivo número tres mencionado en el Capítulo I.

6.1. Selección de normativa de evaluación de calidad

La calidad de software es un factor muy importante en el desarrollo del producto software, ya que debe haber una "Concordancia del software producido con los requerimientos explícitamente establecidos, con los estándares de desarrollo prefijados y con los requerimientos implícitos no establecidos formalmente, que desea el usuario" [41]. Normalmente, las evaluaciones realizadas a un producto software verifican la concordancia entre lo entregado y lo requerido por el usuario. Sin embargo, el interés del proyecto es evaluar la utilidad de las bodegas de datos en el ámbito académico, ya que estas no son comúnmente utilizadas en instituciones educativas de nivel superior a nivel nacional.

En la literatura se encuentran diversas normativas para medir la calidad, para la selección de esta se tuvieron en cuenta las normativas mencionadas en [42], dado que se busca realizar una evaluación del producto más que del proceso, se descartaron normativas como CMMI (Capability Maturity Model Integration), ISO 15504 (SPICE), etc. Cabe destacar que la elección de la normativa también se estableció teniendo en cuenta la completitud de la misma, por lo tanto las normas ISO/IEC 9126 e ISO/IEC 14598 son descartadas porque actualmente se cuenta con una norma (ISO 25000) que mejoró las inconsistencias presentadas en estas y surgió como una evolución de las mismas [43].

La norma seleccionada fue la ISO/IEC 25000, sin embargo, como se requiere evaluar la utilidad del prototipo es necesario utilizar, específicamente, la norma ISO/IEC 25022 que permite evaluar el producto en base al uso. Partiendo del objetivo número tres (mencionado en el Capítulo I), la subcaracterística seleccionada para la evaluación es la "*Utilidad*", que permite determinar la capacidad en la que se puede aprovechar el producto software por parte de los usuarios. Dentro de esta subcaracterística, se encuentran diversas medidas, aunque se optó por evaluar solo con la medida que permite identificar la *satisfacción del usuario con las características del sistema*, ya que por cuestiones de alcance (el proyecto no pasa a la etapa de producción), no es posible utilizar las demás medidas observadas en la Tabla 35.

Satisfacción	
Subcaracterística	Nombre de las medidas
Utilidad	• **Satisfacción con las características del sistema.** • Uso discrecional[11]. • Utilización de características del sistema. • Proporción de usuarios que se quejan. • Proporción de quejas de usuarios sobre una característica en particular.

Tabla 35 Medidas y subcaracterísticas de "Satisfacción". Fuente: Adaptada de [44]**.**

[11] Cantidad de usuarios potenciales que eligen utilizar el sistema.

6.2. Objetivo de la evaluación

La evaluación tiene como finalidad verificar la importancia que puede tener un sistema de bodega de datos en la Universidad del Cauca y adicionalmente brindar una perspectiva de la utilidad de este tipo de soluciones a las demás instituciones universitarias del país.

6.3. Ejecución de la evaluación

La norma ISO/IEC 25022 presenta qué aspectos se evalúan en cada característica, sin embargo no define la forma de evaluar. Para el prototipo, como se recomienda en [45], se definió esta evaluación a través de cuestionarios, los cuales están compuestos por tres secciones: la primera presenta al usuario el objetivo de la evaluación junto con la información relevante de cada usuario; la segunda sección enfocada en los reportes estándar, donde se presenta al usuario las correspondientes formas matriciales en las que el usuario calificará cada pregunta correspondiente a cada uno de los reportes; y finalmente la tercera centrada en la evaluación de las consultas analíticas o Ad-Hoc, que de manera similar a los reportes, presenta la forma de calificación para cada una de estas consultas. En la parte final del cuestionario se presenta un apartado para que el usuario pueda firmar y de esta manera tener un registro de que la evaluación si se realizó. Por último, la escala utilizada para la calificación de la evaluación fue la de Likert (Ver Tabla 36), debido a que esta es utilizada comúnmente para cuestionarios y encuestas de investigación.

Valor numérico	Concordancia
5	Totalmente de acuerdo
4	Parcialmente de acuerdo
3	Indiferente
2	Algo en desacuerdo
1	Totalmente en desacuerdo

Tabla 36 Escala de Likert para calificación del cuestionario.

Los usuarios (decano y coordinadores de programa) con los cuales se hizo la recolección de requerimientos (sección 3.4.1) fueron seleccionados para el proceso de evaluación. Por tal razón, fue necesario programar una reunión anticipada con los mismos. Como se mencionó en la estrategia de capacitación de los usuarios presentada en el Capítulo V, antes de realizar la ejecución de las pruebas se hizo la correspondiente capacitación sobre el prototipo. El proceso de evaluación consistió en presentar (a cada usuario) cada uno de los reportes descritos en el cuestionario, por cada uno de estos, el usuario inició la calificación de las preguntas.

En la mayoría de las pruebas que se realizaron a los usuarios, cada uno expuso recomendaciones, las cuales se presentan en la sección 6.4.

6.4. Análisis de resultados obtenidos de la evaluación

El proceso de evaluación estuvo basado en [45]. La manera de obtener el resultado de la evaluación estuvo definida por el número de preguntas en cada escala posible de calificación, presentada en forma porcentual. Sin embargo, para el puntaje total de la evaluación se tuvo en cuenta solo los porcentajes relacionados a las calificaciones "*Totalmente de acuerdo*" y "*Parcialmente de acuerdo*". El valor deseado para los resultados de la evaluación es 100%. Finalmente se debe tener en cuenta que el resultado de la

evaluación debe indicarse de manera cualitativa dependiendo de la escala de medición obtenida. En la Tabla 37 se realiza una clasificación de estas con su respectivo grado de satisfacción y nivel de puntuación, aunque esta tabla fue levemente modificada para tener una referencia porcentual.

Escala de medición	Niveles de puntuación	Grado de satisfacción
80,75% - 100%	Cumple con los requisitos	Muy satisfactorio
50,00% - 80,74%	Aceptable	Satisfactorio
20,75% - 40,90%	Mínimamente aceptable	Insatisfactorio
0,00% - 20,74%	Inaceptable	

Tabla 37 Grado de satisfacción según niveles de puntuación. Fuente: Adaptada de [45]**.**

A continuación se presentarán los resultados de la evaluación aplicada por cada uno de los usuarios seleccionados (todos los cuestionarios tuvieron 19 preguntas).

- Decano Facultad de Ingeniería Electrónica y Telecomunicaciones (FIET)

Como se puede observar en la Tabla 38, este usuario tuvo un porcentaje del 84,21% de respuestas satisfactorias con respecto a los reportes y consultas Ad-Hoc. Por lo anterior y teniendo la Tabla 37 como referencia, se puede deducir que para este usuario su nivel de puntuación es "*Cumple con los requisitos*" y su grado de satisfacción es "Muy *Satisfactorio*". Mostrando que para este usuario es muy útil el desarrollo de una solución de este tipo.

Escala de calificación	Número de preguntas	Porcentaje con respecto al número total de preguntas en el cuestionario
Totalmente de acuerdo	14	73,68%
Parcialmente de acuerdo	2	10,53%
Indiferente	3	15,79%
Algo en desacuerdo	0	0,00%
Totalmente en desacuerdo	0	0,00%
	Evaluación final	84,21%

Tabla 38 Evaluación correspondiente al usuario con el cargo de decano.

- Coordinador(a) Programa Ingeniería de Sistemas (PIS).

Para este usuario se obtuvo el porcentaje de preguntas satisfactorias (Ver Tabla 39), el cual fue del 78,95%, con un nivel de puntuación "*Aceptable*" y un grado de satisfacción de "*Satisfactorio*", lo cual permite interpretar que para este usuario es útil un sistema de este tipo, aunque se deben realizar mejoras para un mejor aprovechamiento del mismo.

Es importante mencionar que el usuario consideró las consultas Ad-Hoc como un insumo de gran importancia para sus labores, sin embargo, la información visualizada en los reportes estándar debe tener una organización más simple (menor sobrecarga de información) para su mejor entendimiento.

Escala de calificación	Número de preguntas	Porcentaje con respecto al número total de preguntas en el cuestionario
Totalmente de acuerdo	10	52,63%
Parcialmente de acuerdo	5	26,32%
Indiferente	4	21,05%
Algo en desacuerdo	0	0,00%
Totalmente en desacuerdo	0	0,00%
	Evaluación final	78,95%

Tabla 39 Evaluación correspondiente al usuario con el cargo de coordinador del PIS.

- Coordinador(a) Programa Ingeniería Automática Industrial (PIAI)

Teniendo en cuenta los resultados de la Tabla 40, el nivel de satisfacción fue de "*Cumple con los requisitos*" y el grado de satisfacción "*Muy satisfactorio*", los cuales fueron obtenidos por el porcentaje de preguntas calificadas como "*Totalmente de acuerdo*" y "*Parcialmente de acuerdo*". Por lo tanto, el porcentaje de preguntas satisfactorias fue del 100%, permitiendo deducir que a este usuario le parece realmente útil este tipo de sistemas. Además según lo comentado por dicho usuario, este tipo de información almacenada en el sistema podría ayudar en los procesos de acreditación institucional.

Escala de calificación	Número de preguntas	Porcentaje con respecto al número total de preguntas en el cuestionario
Totalmente de acuerdo	16	84,21%
Parcialmente de acuerdo	3	15,79%
Indiferente	0	0,00%
Algo en desacuerdo	0	0,00%
Totalmente en desacuerdo	0	0,00%
	Evaluación final	100%

Tabla 40 Evaluación correspondiente al usuario con el cargo de coordinador del PIAI.

Como se observa, la evaluación con cada usuario fue distinta, por lo cual se optó por tener una visión global de los resultados. Encontrando un porcentaje de preguntas satisfactorias del 87,72% (Ver Tabla 41), que muestran que el sistema "*Cumple con los requisitos*" y es considerado "*Muy Satisfactorio*".

Escala de calificación	Número de preguntas	Porcentaje con respecto al número total de preguntas en el cuestionario
Totalmente de acuerdo	40	70,18%
Parcialmente de acuerdo	10	17,54%
Indiferente	7	12,28%
Algo en desacuerdo	0	0,00%
Totalmente en desacuerdo	0	0,00%
	Evaluación final	87,72%

Tabla 41 Resultado evaluación del prototipo.

A pesar de clasificar en una buena puntuación, los usuarios realizaron algunas recomendaciones, estas son presentadas en la siguiente lista de ítems:

- El nombre de los reportes debe ser más claro, de tal manera el usuario entienda qué tipo de información es la que contiene este.
- Los reportes deben poder hacer una especialización de los datos, es decir, deben poder acceder al nivel de detalle requerido, como, por ejemplo: el código del estudiante, el nombre de las materias, etc.
- Creación de un manual en el que se explique la forma como se llevan a cabo las consultas Ad-Hoc en Excel, desde la conexión hasta la realización de las consultas.
- En los reportes se debe evitar la agrupación de mucha información, ya sea en las filas o columnas de la tabla, dado que esto puede llegar a confundir a los usuarios.
- El tipo de gráfico utilizado en los reportes debe ser el adecuado para representar la información contenida en la tabla.

Adicional a las recomendaciones de mejora, se presentan algunos de los comentarios positivos obtenidos en esta etapa:

- El sistema permite reconocer información valiosa con respecto a los factores influyentes en el desempeño académico de los estudiantes, con lo cual se pueden realizar estrategias para acompañar a aquellos estudiantes que presenten falencias.
- Al contar con un sistema de este tipo, se eliminan los intermediarios entre la información y los usuarios finales, ya que no sería necesaria la petición de los informes a los funcionarios de la división de tecnologías.
- La detección temprana de estudiantes con promedios académicos bajos, con posibilidades de balanceo, permite generar estrategias que reduzcan la deserción estudiantil.
- El seguimiento semestre a semestre de los estudiantes que han ingresado bajo un caso especial, permite verificar la situación actual de los mismos y generar estrategias para brindar un mejor servicio a esta población.

Los cuestionarios de evaluación aplicados a cada usuario, se pueden encontrar en el Anexo J.

Capítulo VII. Conclusiones y trabajo futuro

En esta monografía se presentan una serie de modelos dimensionales basados en las técnicas propuestas por [4] que pretenden apoyar las necesidades analíticas con respecto al desempeño académico del estudiantado en la Universidad del Cauca y ofrecer estos modelos a las demás universidades públicas del país. Para llevar a cabo el desarrollo de este proyecto se definieron un conjunto de objetivos, los cuales fueron logrados a partir de la realización de las actividades presentadas en cada uno de los capítulos anteriores.

7.1. Análisis de los objetivos de investigación

A continuación, se presentan cada uno de los objetivos propuestos junto con un resumen de los capítulos que demuestran el cumplimiento de los objetivos de este trabajo:

- Identificar los casos de diseño que se deben tener en cuenta en el modelado dimensional del registro académico de los estudiantes que contemple registro de notas, control de asistencia e información socio-demográfica.

  En el Capítulo IV. Modelado dimensional, se describen de manera detallada los principales componentes dentro de un modelo dimensional enfocado en el proceso de registro de notas, considerando dentro del mismo, información con respecto al control de asistencia y la información socio-demográfica recolectada por la institución en los últimos años. Adicionalmente se presenta una especificación de los casos de diseño identificados en los mismos, que permite su adaptación en otros procesos de negocio u otras instituciones educativas de nivel superior.

- Construir un prototipo de una bodega de datos para el registro de notas de la Facultad de Ingeniería Electrónica y Telecomunicaciones (FIET) basado en el modelo dimensional propuesto, que permita la visualización de los datos por medio de consultas analíticas Ad-hoc y de reportes estándar.

  El Capítulo V. Elaboración del prototipo propuesto, describe el desarrollo del prototipo, especificando aspectos importantes para la realización de una solución de este tipo, presentando el proceso que permitió la construcción de cada uno de los elementos que constituyen una bodega de datos. Adicionalmente se exponen algunas recomendaciones que permitan enriquecer el sistema desarrollado con respecto a la optimización de los tiempos de respuesta al usuario (planes de agregación e indexación) y la mejora en la administración de los datos (planes de particionamiento). Así mismo, se presentan las actividades involucradas en el proceso ETL, junto a las problemáticas encontradas y las recomendaciones generadas. Por otra parte, se detallan las actividades realizadas para la construcción de las aplicaciones de usuario final.

- Evaluar la utilidad de los reportes estándar y las consultas analíticas que genere el prototipo de la bodega de datos, por medio de un test de nivel de satisfacción de los funcionarios académicos responsables del registro académico en la FIET, con base en

la métrica *nivel de satisfacción*, de la ISO/IEC 25022: Medidas de Calidad en Uso, definida para la sub-característica *utilidad*.

El Capítulo VI. Evaluación del prototipo, describe el proceso llevado a cabo para la realización de la evaluación del prototipo propuesto, a través de reuniones con los usuarios finales, quienes calificaron el prototipo haciendo uso de cuestionarios, los cuales contenían preguntas con respecto a los reportes estándar y las consultas ad-hoc. Los resultados obtenidos en esta etapa permitieron (por medio de la métrica de utilidad descrita en la ISO/IEC 25022) reconocer la utilidad de una solución de este tipo en el ámbito académico, específicamente en la FIET. Adicionalmente se presentan comentarios positivos, como también recomendaciones de mejora, las cuales fueron tenidas en cuenta para la realización de la versión final del prototipo.

- **Objetivo general:** Proponer un modelo dimensional para el registro académico de los estudiantes de la Universidad del Cauca que incluya información socio-demográfica, que se pueda convertir en un referente para otras universidades públicas del país.

  A partir del cumplimiento de los objetivos específicos descritos anteriormente, se consiguió definir un modelo que da soporte al proceso del registro académico dentro de la Universidad del Cauca. Para la definición del modelo dimensional, se tuvieron en cuenta aspectos internos de la institución, como también aspectos considerados relevantes obtenidos en el Capítulo II, abarcando la información personal, académica y socio-demográfica del estudiante. De esta manera, se da cumplimiento también al objetivo general propuesto para este trabajo.

7.2. Conclusiones

A continuación, se presentan las principales conclusiones realizadas a partir de este trabajo:

- Después de realizar un análisis de la literatura a través del enfoque de revisión sistemática, se identificó un conjunto de propuestas de modelos dimensionales, que abordaban diferentes procesos de negocio y de los componentes que generalmente se incluyen dentro de un modelo dimensional del proceso de registro académico. Sin embargo, se pudo observar que hasta el momento no existen propuestas que involucren la identificación de casos de diseño de registro académico de estudiantes de universidades.

- Para la elaboración de un modelo dimensional que involucre diferentes casos de diseño, es necesario el conocimiento detallado del proceso de negocio, por lo cual, trabajar con una metodología enfocada en el usuario, facilita la identificación de los aspectos particulares del negocio, permitiendo al equipo de desarrollo reconocer los casos de diseño que deben ser incluidos dentro del modelado dimensional.

- Para el desarrollo de un proceso ETL es necesario realizar un estudio exhaustivo de las fuentes de datos, el cual se facilita con la ayuda de los funcionarios encargados del sistema OLTP.

- El proceso de capacitación se vio beneficiado, al ser realizado con directivos de la institución que trabajan en su día a día con la herramienta Excel.

- Con base en la evaluación realizada y en los comentarios positivos recibidos durante la misma, se reconoce el poder analítico de los modelos dimensionales creados y la utilidad que la bodega de datos puede brindar en una institución educativa de nivel superior, facilitando los procesos de análisis de la información y contribuyendo de gran manera en la toma de decisiones.

- El poder analítico con el que cuenta una solución de bodegas de datos, es directamente proporcional a la cantidad de datos que se encuentren dentro la institución. En este sentido, el no contar con un sistema OLTP que contenga información completa con respecto al registro académico, reduce el potencial de análisis de la bodega de datos, por lo cual, aunque se cumplieron las expectativas de los usuarios, el sistema de bodegas de datos podría ser mejorado en un futuro.

7.3. Lecciones aprendidas

Durante la realización de este trabajo, surgieron diferentes situaciones, las cuales pueden ser tenidas en cuenta para la realización de trabajos futuros. Estas situaciones son descritas a continuación:

- Aunque los requerimientos analíticos abarcaron los diferentes aspectos considerados dentro del proceso de registro académico, la elaboración de uno o más grupos focales podrían enriquecer la recolección de dichos requerimientos, ya que tener diferentes puntos de vista agilizaría el reconocimiento de las necesidades analíticas de la institución.

- Para el estudio de las diferentes fuentes de datos, es de vital importancia la participación de los encargados de las mismas, de esta manera se obtiene mayor claridad y entendimiento, facilitando el diseño del proceso ETL, puntualmente en la etapa de extracción de datos.

- El uso de estándares de nombrado, perspectivas y estrategias de navegación, otorga un mayor entendimiento de las aplicaciones de usuario final, facilitando los procesos de capacitación llevados a cabo con los usuarios.

- Para la definición de un modelo dimensional fue importante realizar previamente una revisión sistemática del estado del arte, que permitiera no sólo saber el estado actual de la literatura, sino que también permitiera identificar los elementos importantes dentro de un modelo dimensional para la academia, lo cual posibilita la creación de un modelo dimensional más flexible y completo. Además, la realización de una revisión sistemática hizo posible la identificación brechas y de esta manera realizar aportes importantes a la comunidad académica.

- La indexación de las tablas relacionales en las bodegas de datos permite la optimización de las consultas, reducción de los tiempos de ejecución de las mismas, lo cual mejora la fluidez del usuario con el sistema. Por otro lado, contar con un buen plan de indexación en el sistema OLTP, mejora los tiempos en la extracción del proceso ETL.

7.4. Trabajos futuros

Como continuación de este trabajo de grado quedan diversas líneas de trabajo abiertas: algunas relacionadas directamente con este trabajo y otras líneas generales pueden ser retomadas posteriormente o ser opción de trabajos futuros.

- **Actualizar el estado del arte.** Realizar una actualización de la revisión realizada, adicionando temas de interés que complementen la propuesta o aborden temas de investigación relacionados.

- **Realizar la recolección de requerimientos mediante grupos focales.** Al realizar la recolección de requerimientos por medio de entrevistas con los usuarios de forma individual, no permitió a los usuarios discutir las necesidades analíticas, para construir un modelo dimensional con mayor potencial analítico. Por esta razón, se hace necesario realizar una actualización de los requerimientos recolectados por medio de grupos focales.

- **Implementación de nuevos modelos.** Realizar una profundización en aquellos modelos propuestos que no fueron incluidos dentro del prototipo, que permita adicionar nuevos componentes dentro los mismos. Además, generar nuevos modelos enfocados en los procesos de negocio que no se tuvieron en cuenta en la realización de este trabajo, esto con el fin de consolidar este tipo de tecnologías en la totalidad de procesos de la institución.

- **Implementación con herramientas con versiones para producción y datos reales.** Para la realización de análisis útiles para la Universidad, es necesaria la implementación del prototipo con herramientas en una versión que permita desplegar en un ambiente real el sistema de bodegas de datos, además de obtener los datos reales de las fuentes seleccionadas. Para la ejecución del proceso ETL, no se consideran grandes cambios, ya que este proceso se implementó basado en la estructura del sistema OLTP.

- **Evaluación de la calidad del producto.** Poner en producción el prototipo para realizar una evaluación que contemple todas las métricas de satisfacción (de la subcaracterística utilidad) consignadas en la ISO/IEC 25022, permitiendo obtener el nivel de satisfacción en esta etapa de producción, de esta forma afirmar la importancia de este tipo de soluciones en la Universidad del Cauca. Además, con esta evaluación se puede generar un referente para las demás instituciones educativas de nivel superior de carácter público del país.

- **Minería de datos.** Con el fin de identificar las variables que influyen en gran medida en el rendimiento académico de los estudiantes, es importante la inclusión de modelos de minería de datos. En conjunto con la información consolidada dentro de la bodega de datos, permitiría realizar un análisis con respecto al estado actual de los estudiantes en la institución, además de análisis predictivos, con esto se facilitaría la generación de estrategias correctivas y preventivas.

7.5. Contribuciones de la investigación

A continuación, se describen las principales contribuciones realizadas por este trabajo de grado a la comunidad académica:

- La realización de la revisión sistemática de la literatura relacionada con el diseño de modelos dimensionales enfocados en procesos universitarios, permitió identificar los principales trabajos relacionados sobre el tema. Sin embargo, se evidenció que aún existen trabajos con inconsistencias de diseño, carencias de información e inexistencia de casos de diseño.

- La completitud de los modelos dimensiones propuestos, junto con la identificación de los casos de diseño que permiten manejar relaciones particulares entre los datos, permite que estos puedan ser tomados como referentes para otras universitarias públicas del país.

- El seguimiento de la metodología seleccionada en conjunto con las recomendaciones dadas, facilita a las instituciones universitarias del país la correcta implementación de soluciones de este tipo.

7.6. Contribuciones en la Universidad del Cauca

Dado que el proyecto fue orientado al desarrollo de modelos dimensionales enfocados en los procesos manejados por la institución, como también en el cumplimiento de lo requerido por los usuarios, la Universidad del Cauca cuenta con el pilar de una bodega de datos (modelo dimensional). Por otra parte, el diseño del proceso ETL fue realizado basándose en la estructura del sistema integrado de matrícula y control académico, la apropiación de este por parte de la institución facilita el cargue de la bodega de datos para así llevarla a una etapa de producción.

Referencias bibliográficas

[1] "SNIES - Estadísticas." [Online]. Available: https://www.mineducacion.gov.co/sistemasinfo/Informacion-a-la-mano/212400:Estadisticas. [Accessed: 16-Jul-2018].

[2] Ministerio de Educación Nacional de Colombia, "Documento metodológico MIDE Universitario 2017," p. 28, 2018.

[3] Ministerio de educación nacional, "Colombia Aprende | La red del conocimiento." [Online]. Available: http://aprende.colombiaaprende.edu.co/mide. [Accessed: 09-Sep-2018].

[4] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modelling*. 2011.

[5] U. Francisco and D. E. P. Santander, "Análisis diseño e implementación de un Datawarehouse para el apoyo de la toma de decisiones en los componentes academicos de la oficina de planeación de la Universidad Francisco de Paula Santander," *200.93.148.28*, pp. 1–329, 2004.

[6] A. Ta'a, M. S. A. Bakar, and A. R. Saleh, "Academic Business Intelligence System Development Using SAS® Tools," *SAS Glob. Forum 2008*, no. 2006, pp. 1–14, 2008.

[7] P. Tanuska, O. Vlkovic, A. Vorstermans, and W. Verschelde, "The proposal of ontology as a part of University data warehouse," *ICETC 2010 - 2010 2nd Int. Conf. Educ. Technol. Comput.*, vol. 3, pp. 21–24, 2010.

[8] C. Zambrano Matamala, D. Rojas Díaz, and K. Carvajal Cuello, "Analisis De Rendimiento Academico Estudiantil Usando Data Warehouse Y Redes Neuronales," *Rev. Chil. Ing.*, vol. 19, pp. 369–381, 2011.

[9] K. G. Akintola, A. O. Adetunmbi, and O. S. Adeola, "Building Data Warehousing and Data Mining From Course Management Systems : A Case Study of Federal University of Technology (FUTA) Course Management Information Systems," *Inf. Technol. People-Centred Dev. (ITePED 2011)*, no. ITePED, pp. 1–11, 2011.

[10] M. Muntean, A. R. Bologa, R. Bologa, and A. Florea, "Business intelligence systems in support of university strategy," *Proc. WSEAS/IASME Int. Conf. Educ. Technol.*, pp. 118–123, 2011.

[11] H. K. Kunta, P. K. Eswar, and P. Harini, "Data Warehousing-A Case-Based Courseware," *Int. J. Recent Trends Eng. Technol. Nov 2011*, vol. 6, no. 1, pp. 162–166, 2011.

[12] J. E. Gutierrez, "Descubrimiento de conocimientos en la base de datos académica de la Universidad Autónoma de Manizales aplicando redes neuronales," pp. 1–57, 2012.

[13] F. Di Tria, E. Lefons, and F. Tangorra, "Academic data warehouse design using a hybrid methodology," *Comput. Sci. Inf. Syst.*, vol. 12, no. 1, pp. 135–160, 2014.

[14] D. Martinez, M. Karanik, M. Giovannini, and N. Pinto, "Estudio de perfil de rendimiento académico: Un abordaje desde data warehousing," *Rev. Tecnol. y Cienc.*, pp. 89–99, 2015.

[15] C. A. Cardoza Timana, "Elaboración De Un Data Mart Para Evidenciar El Retraso Académico En Los Alumnos De Pregrado De La Fii-Unp," pp. 1–132, 2015.

[16] J. Á. Hernandez Cedano and J. A. Castro, "Modelo de minería de datos para identificación de patrones que influyen en el aprovechamiento academico," no. 612, 2015.

[17] E. Bates, "UVM Big Data? Aggregating Campus Databases and Creating a Data Warehouse to Improve Student Retention Rates at the University of Vermont," pp. 1–37, 2015.

[18] L. C. Canchila Tapias and J. A. Sánchez Bohórquez, "Análisis de la Deserción Estudiantil de Cecar Utilizando Herramientas de Inteligencia de Negocios con Licencia Libre," pp. 1–125, 2016.

[19] J. R. Cuadrado Montaña and E. León Guzmán, "Modelo para la gestión de indicadores y análisis de permanencia estudiantil de usuarios de bienestar universitario de la Universidad Nacional de Colombia," pp. 1–82, 2017.

[20] R. A. Suárez Urrego, C. Jiménez Ramírez, and J. A. Restrepo Morales, "Modelo de gestión para el análisis de los factores de deserción en el Tecnológico de Antioquia," pp. 1–97, 2018.

[21] N. Y. University, "University Data Warehouse Plus," *New York University*, 2016. [Online]. Available: https://www.nyu.edu/employees/resources-and-services/administrative-

services/university-data-warehouse-plus.html. [Accessed: 16-Jul-2018].

[22] Universidad de Pittsburgh, “University Data Warehouse | Information Technology | University of Pittsburgh.” [Online]. Available: https://www.technology.pitt.edu/services/university-data-warehouse. [Accessed: 16-Jul-2018].

[23] Instituto Tecnológico de Massachusetts, “Data Warehouse | Information Systems & Technology.” [Online]. Available: http://ist.mit.edu/warehouse. [Accessed: 16-Jul-2018].

[24] Universidad de Alberta, “Strategic Analysis and Data Warehousing - University of Alberta.” [Online]. Available: https://www.ualberta.ca/vice-president-finance/audit-and-analysis/about-audit-and-analysis/strategic-analysis-and-data-warehousing. [Accessed: 16-Jul-2018].

[25] S. Chaudhuri *et al.*, *Building the Data Warehouse, Fourth Edition*, vol. 13, no. 401. 2005.

[26] L. Zepeda S., “Metodología para el Diseño Conceptual de Almacenes de Datos,” 2008.

[27] R. A. ESPINOSA, “Data Warehouse Para La Gestión De Lista De Espera Sanitaria,” *Univ. Politécnica Madrid*, p. 148, 2008.

[28] I. Ingrid, P. Solano, M. Martha, E. Mendoza, D. T. Completo, and F. De Ingeniería, “Modelo Dimensional De Bodegas De Datos Adaptable a Empresas Mipymes De Ventas Al Detal,” no. 2002, pp. 1–7, 2011.

[29] C. Trujillo, “Modelo Multidimensional,” 2006.

[30] IBM, “IBM Knowledge Center - Esquemas de constelación.” [Online]. Available: https://www.ibm.com/support/knowledgecenter/es/SS9UM9_9.1.2/com.ibm.datatools.dimensional.ui.doc/topics/c_dm_schema_starflake.html. [Accessed: 09-Sep-2018].

[31] Anthony Araujo, “Desarrollo de un Cubo Olap ~ Code Botic.” [Online]. Available: http://codebotic.blogspot.com/2015/12/cubo-olap-cinema-i.html. [Accessed: 18-Jul-2018].

[32] J. Mundy and W. Thornthwaite, “The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset (Google eBook),” p. 720, 2011.

[33] Wikimedia, “Operaciones OLAP.” [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/f/ff/OLAP_slicing.png. [Accessed: 02-Oct-2018].

[34] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3,” *Engineering*, vol. 45, no. 4ve, p. 1051, 2007.

[35] M. E. Mendoza, “Mbd 1.0 - Metología de desarrollo de bodegas de datos para micro, pequeñas y medianas empresas,” no. 1, pp. 85–101, 2010.

[36] O. Corporation, “Oracle Technology Global Price List Software Investment Guide,” pp. 1–13, 2018.

[37] Oracle, “Oracle® Database Data Warehousing Guide 11g Release 2 (11.2),” vol. 2, no. July, p. 530, 2013.

[38] Vivek Sharma, “Índice de mapa de bits vs. índice de árbol B: Cuándo usar cuál,” 2014. [Online]. Available: https://www.oracle.com/technetwork/es/articles/sql/indices-mapa-de-bits-y-arbol-b-2393445-esa.html. [Accessed: 17-Sep-2018].

[39] E. Melomed, I. Gorbach, A. Berger, and P. Bateman, “Microsoft SQL Server 2005 Analysis Services,” no. February, 2008.

[40] Microsoft, “Analysis Services Performance Guide for SQL Server 2012 and 2014,” vol. 2012, no. May, p. 111, 2014.

[41] R. S. Pressman, *Ingeniería del software. Un enfoque práctico*. 2013.

[42] L. A. Espinel, N. J. Acosta, and J. L. García, “Estándares para la calidad de software,” *Tecnol. Investig. y Acad.*, vol. 5, no. 1, pp. 75–84, 2017.

[43] ISO/IEC, “NORMAS ISO 25000.” [Online]. Available: https://iso25000.com/index.php/normas-iso-25000. [Accessed: 01-Oct-2018].

[44] Subcomité 7 Copyright © 2016 - 2018 All Rights Reserved, “NC ISO/IEC 25022 SQuaRE – Medición de la calidad en el uso.” [Online]. Available: https://subcomite7.cubava.cu/2017/11/23/nc-isoiec-25022-square-medicion-de-la-calidad-en-el-uso/#.W5xTG85KjIW. [Accessed: 14-Sep-2018].

[45] E. A. Chisaguano Balseca, “Evaluación de Calidad de Productos Software en Empresas de desarrollo de Software aplicando la Norma ISO/IEC 25000,” 2014.