Universidad Autónoma de Querétaro

Facultad de Informática

Maestría en Ingeniería de Software Distribuido

“Desarrollo de una metodología para el fortalecimiento de la elaboración de estadísticas y reportes a través de un proyecto *Data Warehouse* en la Dirección de Innovación y Tecnologías de la Información de la Universidad Autónoma de Querétaro”

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ingeniería de Software Distribuido

Presenta

José Joaquín Aguilar Guerrero

Dirigido por

M.S.I. Diego Octavio Ibarra Corona

Centro Universitario, Querétaro, Qro.

Fecha de aprobación por el Consejo Universitario (febrero 2024)

México

Desarrollo de una metodología para el fortalecimiento de la elaboración de estadísticas y reportes a través de un proyecto Data Warehouse en la Dirección de Innovación y Tecnologías de la Información de la Universidad Autónoma de Querétaro

por

José Joaquín Aguilar Guerrero

Clave RI: IFMAN-138314

Universidad Autónoma de Querétaro
Facultad de Informática
Maestría en Ingeniería de Software Distribuido

Desarrollo de una metodología para el fortalecimiento de la elaboración de estadísticas y reportes a través de un proyecto *Data Warehouse* en la Dirección de Innovación y Tecnologías de la Información de la Universidad Autónoma de Querétaro

Tesis
Que como parte de los requisitos para obtener el Grado de
Maestro en Ingeniería de Software Distribuido

Presenta
José Joaquín Aguilar Guerrero

Dirigido por
M.S.I. Diego Octavio Ibarra Corona

Sinodales

M.S.I. Diego Octavio Ibarra Corona
Presidente

Dra. Gabriela Xicoténcatl Ramírez
Secretaria

M.S.I. José Alejandro Vargas Díaz
Vocal

M.I.S.D. Carlos Alberto Olmos Trejo
Suplente

Dr. Mauricio Arturo Ibarra Corona
Suplente

Centro Universitario Querétaro, Qro.
Febrero 2024
México.

Dedicatoria

A mis padres Joaquín y Ma. Natividad por su infinito amor, paciencia, entrega e inquebrantable apoyo en cada paso.

A mi esposa Laura Sánchez, compañera de vida, aliada en cada desafío y la luz que brilla en los momentos más oscuros, gracias por tu apoyo incondicional y tu amor constante.

Agradecimientos

Mis más sinceros agradecimientos al maestro Diego Octavio Ibarra Corona por su seguimiento, atención y su invaluable apoyo en la realización de este proyecto, desempeñando el papel de director de tesis.

Agradezco a cada una de las personas que estuvieron presentes de diversas formas durante el desarrollo de la tesis, gracias por sus comentarios, observaciones, recomendaciones, consejos, llamadas de atención y, sobre todo, por su motivación.

Índice

Índice de Tablas

Índice de Figuras

Resumen

Los datos generados en los sistemas de información deben ser una fortaleza para el crecimiento de la empresas privadas o públicas por medio de la toma de decisiones oportuna y veraz, lo cual es complejo al contar con datos en diferentes fuentes de información, que han sido re-estructuradas a lo largo del tiempo con reglas del negocio ambiguas y requerimientos cambiantes; todo ello sin considerar la consistencia e integridad de los datos. Las consecuencias se presentan al tratar de generar reportes y estadísticas de manera eficaz y eficiente, teniendo que depender del personal técnico para su elaboración y del trabajo artesanal realizado para integrar la información. Por lo cual, se recomienda establecer mecanismos o metodologías que concentren, homologuen y organicen los datos dentro de un repositorio denominado *Data Warehouse (DW)*, teniendo como objetivo agilizar la generación de reportes de forma óptima. La metodología se basa en los conceptos teóricos de *Kimball* e *Inmon*, donde establecen la integración de datos en modelos dimensionales conformados por diferentes perspectivas del negocio o dimensiones, y medidas o hechos, que posibilitan la generación de reportes. Para definir el modelo dimensional, se utilizan los requisitos de información indicados por el personal directivo de la organización y la identificación de las diferentes fuentes de datos. En la migración de los datos al *DW* se establecen procesos de extracción, transformación y carga, lo cual puede ser de forma directa o a partir de una *staging area*, considerando las reglas del negocio para garantizar la calidad de los datos. El *DW* contendrá los datos necesarios para dar respuesta a los requerimientos de información con apoyo de las herramientas de explotación de datos, las cuales permiten visualizar la información en forma de gráficas, tablas dinámicas o indicadores de rendimiento. Con base a lo anterior, los resultados de aplicar una metodología clara y precisa, permiten la elaboración de reportes y estadísticas de manera automatizada, con datos íntegros, oportunos y exactos.

Palabras clave: Calidad de los datos, modelo dimensional y procesos ETL.

Abstract

The information produced by computer systems is critical for the expansion of private or public businesses. The data allows for swift and precise decision-making. However, this process can be difficult when handling information from varying sources that have changed over time with unclear business guidelines and evolving needs. This is all done while ignoring the accuracy and consistency of the data. The results are not good when attempting to create reports and statistics efficiently, relying on technical professionals for their creation and on manual labor to merge the data. It is advised to create methods or procedures that gather, systematize, and arrange the data in a Data Warehouse (DW) to simplify the report generation process. The technique is founded on Kimball and Inmon's theoretical ideas. They integrate data in dimensional patterns consisting of different business viewpoints, dimensions, and fact-powered measures that produce reports. The description of the dimensional pattern is based on the organization's management information demands and the differentiation of various data sources. During the transfer of data to the DW, we establish processes for extracting, transforming, and loading data from the staging area. We consider business rules to maintain high data quality. Using data mining tools, we can visualize the data in graphs, pivot tables, or performance indicators within the DW, allowing us to meet any information needs. Based on this information, using a straightforward and accurate approach yields automated reports and statistics that contain complete, timely, and precise data.

Keywords: data warehouse, data quality, dimensional model and ETL processes.

1. Introducción

1.1. Título del proyecto de investigación

Desarrollo de una metodología para el fortalecimiento de la elaboración de estadísticas y reportes a través de un proyecto *Data Warehouse* en la Dirección de Innovación y Tecnologías de la Información de la Universidad Autónoma de Querétaro.

1.2. Justificación

Los datos utilizados dentro en la Universidad Autónoma de Querétaro se ha convertido en un proceso que involucra esfuerzo, alto consumo en tiempos y, algunas veces, discrepancias en los resultados. Los datos mencionados son utilizados para generar reportes solicitados por distintas áreas de la Universidad, por una parte, el departamento de estadística, facultades y escuelas de bachilleres, áreas de administración central; de igual manera, áreas externas a la Universidad, como auditorias federales, auditorias estatales, la Secretaría de Educación Pública (SEP), la Comisión Estatal para la Planeación de la Educación Superior (COEPES), la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), Gobierno del Estado de Querétaro, el Servicio de Administración Tributaria (SAT), entre otras. Los reportes y/o estadísticas son solicitados durante todo el año, la mayoría de ellos son repetitivos y se entregan en formatos determinados, es por ello que los datos requieren ser completos, estandarizados, coherentes, precisos, íntegros y confiables.

Es importante mencionar que el Sistema Integral de Información Administrativa (SIIA) tiene una estructura de base de datos diseñada entre el año 2000 y 2005, lo que ha ocasionado que se establezcan soluciones poco efectivas para poder responder a las nuevas necesidades que la Universidad presenta. Como consecuencia, la obtención de reportes a partir del sistema transaccional no es óptima y puede ocasionar inconsistencias en los resultados, falta de confianza en lo obtenido, así como el uso excesivo de recursos para tratamiento de datos

manualmente. Por ejemplo, la Dirección de Desarrollo Académico solicita elaborar un reporte que muestre el promedio de la evaluación docente, por docente y programa educativo, y en términos generales, debería ser un reporte sencillo, pero al momento de realizar la integración entre la carga horaria del docente, proporcionada por la Secretaría Académica, y la asignación de materias, establecida por la Dirección de Servicios Académicos, los resultados pueden variar debido a que no se cuenta catálogos centralizados de adscripciones versus escuelas o facultades; así mismo, se puede invertir hasta un día en identificar el origen de los datos válidos y su corrección en caso de ser necesario. Aunado a lo anterior, el reporte no puede elaborarse durante el horario laboral de 08:00 hrs. a 14:00 hrs., por el consumo que pudiera representar en días álgidos de trabajo. Por lo anterior, para la elaboración y entrega del reporte conlleva un tiempo invertido de dos o tres días.

Por consecuencia, es de suma importancia la definición de un repositorio de datos centralizado que apoye los procesos de homologación de los datos; permitiendo definir las reglas del negocio para la calidad de datos, por ejemplo, enriquecer y mejorar los datos, la estandarización de formato de los datos con apoyo de catálogos predefinidos y autorizados por entidades externas, el aumento del nivel de coherencia entre las diferentes áreas de la Universidad, la eliminación de los datos duplicados y el aumento en el nivel de integridad.

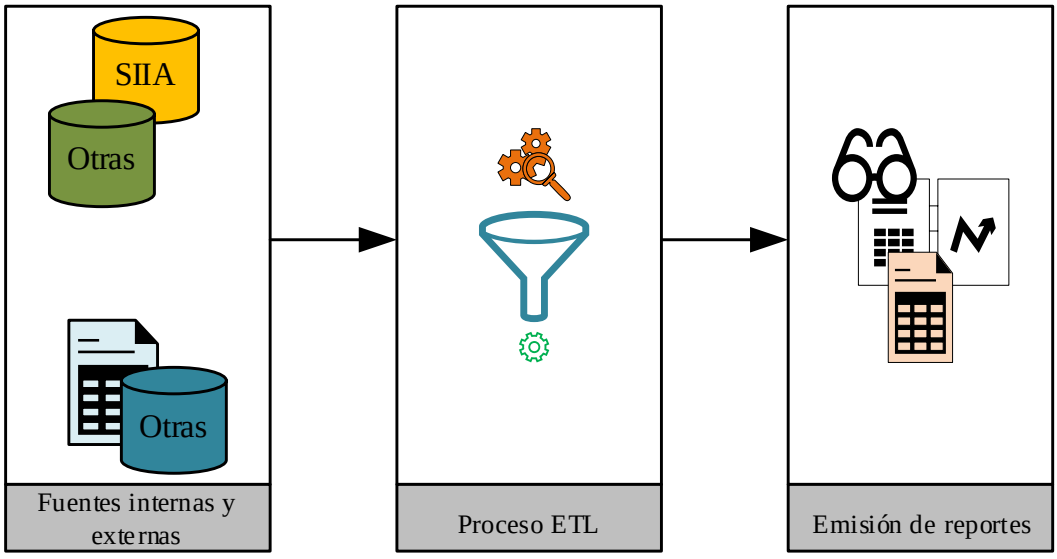
Utilizar un repositorio de datos específicamente diseñado para generar información y/o reportes, puede traer consigo, además de los puntos mencionados en el párrafo anterior, un sinfín de beneficios, como la optimización de la gestión de procesos internos, mejora en la atención al cliente (estudiantes, docentes, áreas administrativas e instituciones externas), identificación de datos duplicados en las diferentes áreas que manejan la información, disminución del trabajo manual en la elaboración de reportes, cumplimiento de normativas externas, la integración de datos con las distintas áreas en la Universidad y, lo más significativo, contar con

datos que sean relevantes y oportunos para fortalecer la elaboración de reportes, estadísticas y la toma de decisiones.

Actualmente, para obtener los reportes en la Universidad, se realizan en tres secciones principales, la primera inicia con la elección de fuentes internas y externas que darán origen al reporte, es decir, identificar qué base de datos o sistema transaccional servirá como fuente de datos; posteriormente, en la parte central se realiza la depuración, validación y agrupación de los datos, en algunos casos de forma automática y en varios otros realizando acciones manuales con el objetivo de adaptarse a los valores establecidos por las instituciones externas mediante el uso de matrices de conversión (en caso de ser necesario); por último, en la tercera sección, se establece el formato del reporte, ya sea como gráficas, tablas dinámicas, etc. (Figura 1). Como se puede apreciar, el proceso suele ser lento y con alto nivel de posibles inconsistencias.

Figura 1

Proceso actual para obtención de reportes en la UAQ



Nota. El diagrama representa el proceso actual para elaborar reportes en la Universidad Autónoma de Querétaro, dicho diagrama puede dividir en tres etapas principales, fuentes internas y externas, procesos ETL y emisión de reportes.

1.3. Descripción del problema

Derivado del crecimiento que la Universidad Autónoma de Querétaro (UAQ) ha presentado en los últimos años con relación al número de aspirantes, matrícula, egreso y procesos administrativos que permiten su operación, los sistemas transaccionales han sido modificados constantemente para adaptarse a las nuevas necesidades y reglas de negocio que la UAQ demanda. Dichos cambios afectan directamente a las estructuras de datos que dan soporte a las operaciones diarias de la Universidad; muchos de estos cambios se realizan con demasiada premura, tanto para su análisis como para su puesta en producción, lo cual ha generado que los datos almacenados presenten algún nivel de inconsistencia.

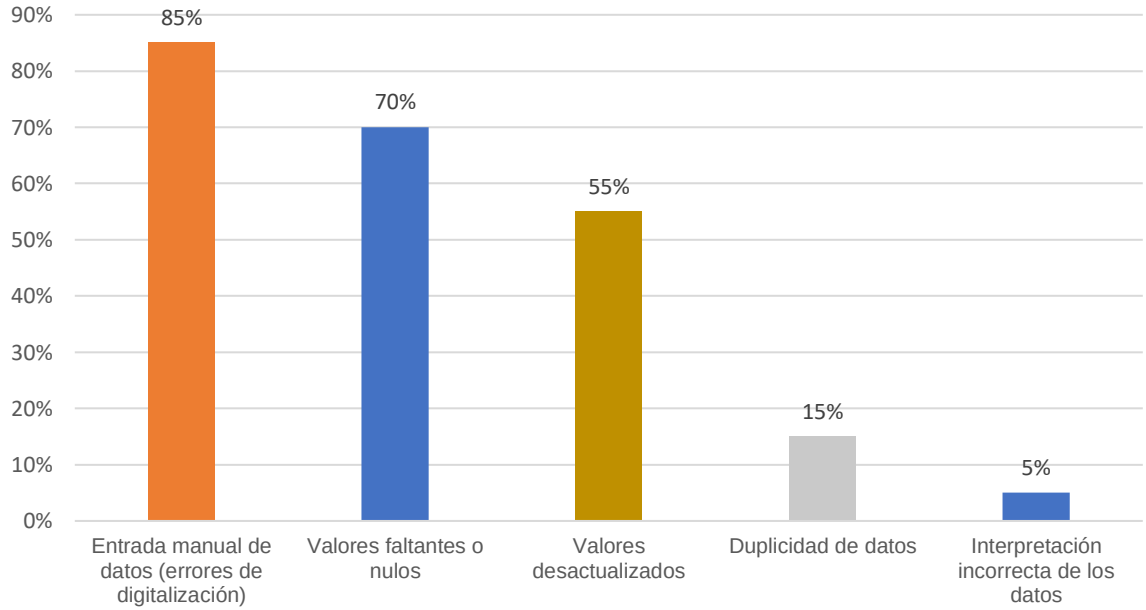
Para detectar detalladamente la problemática y, con ello, delimitar la investigación, se definieron dos instrumentos, el primero denominado *instrumento para personal responsable de reportes y estadísticas institucionales*, el cual, tenía como objetivo recabar información sobre los factores técnicos (sistemas operacionales, fuentes de datos, etc…) que influyen en la elaboración de reportes y estadísticas de la UAQ de acuerdo al personal operativo. En dicho instrumento se evaluaron las dimensiones de uso, de utilidad, de calidad de datos y su disponibilidad, con un total de 10 reactivos en escala *Likert*, aplicado durante el mes de febrero del 2022 al personal responsable de la emisión de reportes de la Institución. Se aplicó el instrumento a 20 personas de distintas áreas, por ejemplo, en la Dirección de Innovación y Tecnologías de la Información (DITI), la Dirección de Servicios Académicos (DSA), la Dirección de Recursos Humanos (DRH) y la Secretaría Académica (SAC), obteniendo el valor de alfa de *Cronbach* de 0.802, por lo tanto, el nivel de confiabilidad del instrumento es *bueno*. En seguida se muestran los resultados obtenidos.

Cada resultado del instrumento aplicado muestra que los datos maestros pueden contener errores a consecuencia de reglas de negocio mal aplicadas, duplicidad de registros, errores al capturar los datos en los sistemas transaccionales, datos faltantes, entre otros. Por esta razón, los factores

identificados que afectan la calidad de los datos utilizados en la elaboración de reportes, están determinados por errores al momento del llenado de formularios o captura de datos en los sistemas, lo que indica la presencia de errores humanos, ya sea por la falta de capacitación, la ausencia de validación en los formularios, procesos mal definidos o la falta de incentivos para fomentar la captura de datos de alta calidad, así mismo, otro factor corresponde a los valores faltantes o nulos y, por último, los valores desactualizados (Figura 2). Existen otros factores que, de igual manera, afectan la calidad de los datos, por ejemplo, contar con varios sistemas que mantienen la misma información generada de manera diferente, provocando con ello una arquitectura de datos en proceso de maduración.

Figura 2

Factores que afectan la calidad de los datos en la UAQ



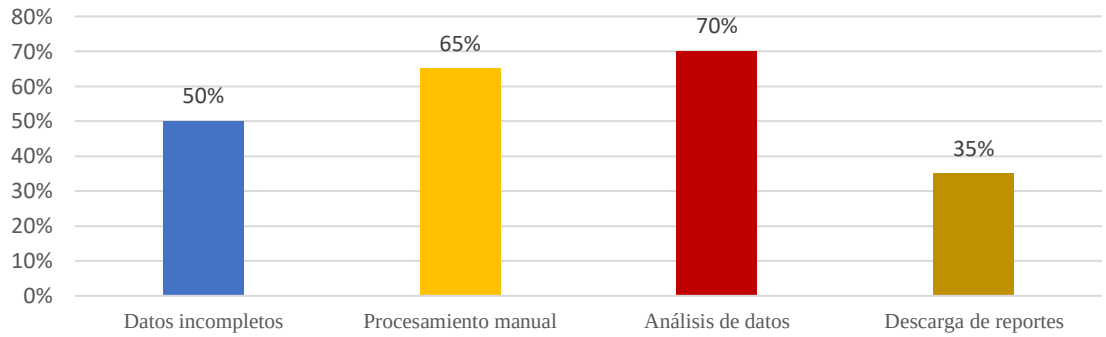
Nota. Existen cinco factores que influyen o afectan la calidad en los datos resguardados en los sistemas de bases de datos de la Universidad, por ejemplo, 85% pertenece a la inserción de registros de forma incorrecta, seguido del 70% en valores faltantes para la elaboración de algún reporte y con un 55% los valores desactualizados.

Como se mencionó anteriormente, la falta de calidad en los datos está definida por un conjunto de factores, a los cuales se le suman, en algunos casos, que la información se encuentra duplicada, incompleta, desactualizada o inexistente; lo anterior ha provocado inconsistencias en la elaboración de reportes solicitados por áreas internas o externas a la Universidad, de igual manera, trae como consecuencias una serie de impactos al momento de hacer uso de los datos, por ejemplo, duplicidad de esfuerzos, desperdicio de recursos al analizar los datos, inversión de tiempo en depuración y tratamiento de datos. En este sentido, en pláticas con los responsables de la elaboración de reportes y estadísticas, en específico, de la Dirección de Planeación, la Dirección de Servicios Académicos, la Secretaría de Finanzas, la DRH y la DITI; se concluye que, el crecimiento de la Universidad en los últimos años, tanto de estudiantes, profesores, personal administrativo, oferta académica y tipos de financiamiento; requieren de una administración eficaz y oportuna de los datos para facilitar la elaboración de reportes de la Universidad. Por ejemplo, se comentó el caso de la emisión de la estadística y reportes de la matrícula escolar que, semestre a semestre, es solicitada a la dirección de Planeación, misma que requiere el apoyo de la Dirección de Servicios Académicos, Dirección de Becas y Secretaría Académica. En cada una de las áreas se encuentra personal operativo y/o técnico que se encarga de revisar y, en su caso, depurar los datos antes de ser enviada a la Dirección de Planeación, misma que es responsable de complementar o afinar el reporte de acuerdo a los catálogos (programas de estudio, escuelas de procedencia, modalidad, etc.) solicitados por la Asociación Mexicana de Órganos de Control y Vigilancia en Instituciones de Educación Superior A.C. (AMOCVIES). El tiempo que invierten las áreas involucradas desde que reciben la petición hasta entregar el reporte a la entidad externa, puede variar entre tres o cuatro semanas (en promedio una semana por departamento). Además, intervienen aproximadamente siete personas para la revisión y depuración.

Como se puede apreciar, las instituciones externas e internas solicitan información con base en catálogos estandarizados, los cuales, en algunas ocasiones, no mantienen relación directa con los datos almacenados en la Universidad, como es el caso de los programas educativos, las escuelas de procedencia y las modalidades de los programas. Para generar los reportes se requiere un procesamiento adicional (matriz de conversión entre los datos internos y catálogos externos) de la información que puede influir en la obtención de reportes de forma rápida y consistente. Los elementos que influyen en la entrega tardía, en caso de existir, de los reportes y/o estadísticas a las áreas que la solicitan puede ser a consecuencias del procesamiento manual que cada una de las áreas involucradas en el proceso debe realizar previo a la entrega oficial del reporte, así mismo, el esfuerzo utilizado para revisar los datos que arrojan las diversas fuentes de datos y el tiempo invertido en actualizar los datos incompletos (Figura 3).

Figura 3

Factores que afectan la entrega de reportes a las entidades

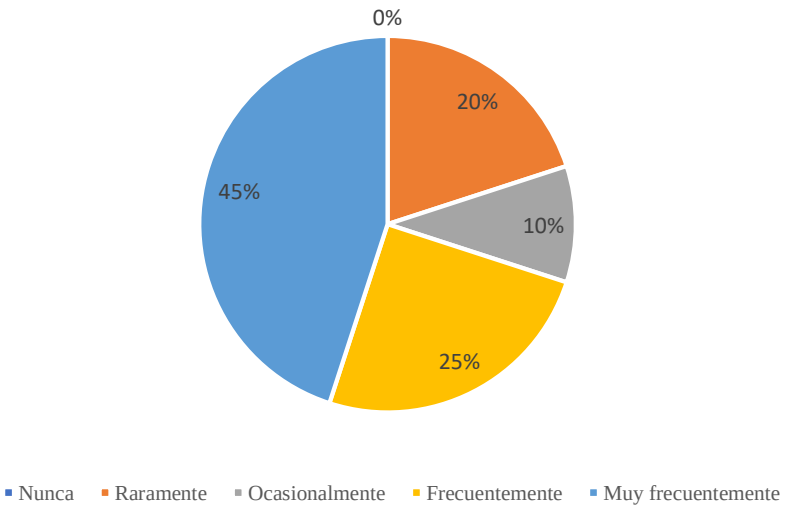


Nota. Existen cuatro factores principales que afectan la entrega de reportes en los tiempos establecidos, el primero de ellos es el análisis de datos invertido en los distintos orígenes con un 70%, seguido el procesamiento manual con un 65% referente a matrices de equivalencias y en tercer lugar los datos incompletos con un 50%, donde se tiene que completar la información antes de emitir el reporte final.

De lo anterior se deriva la importancia de homologar los datos en un repositorio creado específicamente para la generación de estadísticas y/o reportes; el repositorio deberá contener tanto catálogos externos, como internos, además de los datos institucionales, que, en conjunto, proporcionarán un panorama general y detallado sobre los indicadores solicitados a la Universidad. Otro de los factores que afecta de manera considerable la entrega de reportes y estadísticas de forma oportuna y eficiente, es la frecuencia con la cual los solicitan, durante todo el año es necesario entregar reportes y/o estadísticas de forma frecuente a las áreas internas de la Universidad para poder realizar el trabajo cotidiano, y eventualmente se solicitan reportes por auditorías externas o algún otro organismo (Figura 4).

Figura 4

Frecuencia de petición de reportes internos y externos

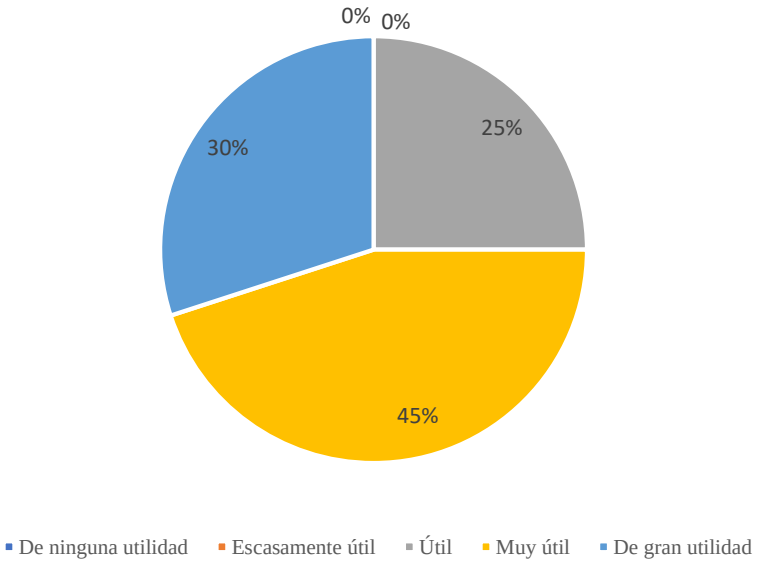


Nota. Las peticiones de reportes solicitados presentan una frecuencia del 70% durante todo el año, y un 30% de manera ocasional.

Dichos reportes contienen información basada en datos con un nivel de utilidad del 75%, esto indica que, son datos necesarios para la toma de decisiones de la Universidad, por ello la importancia de fortalecer el proceso para su elaboración (Figura 5).

Figura 5

Nivel de utilidad de los datos transaccionales

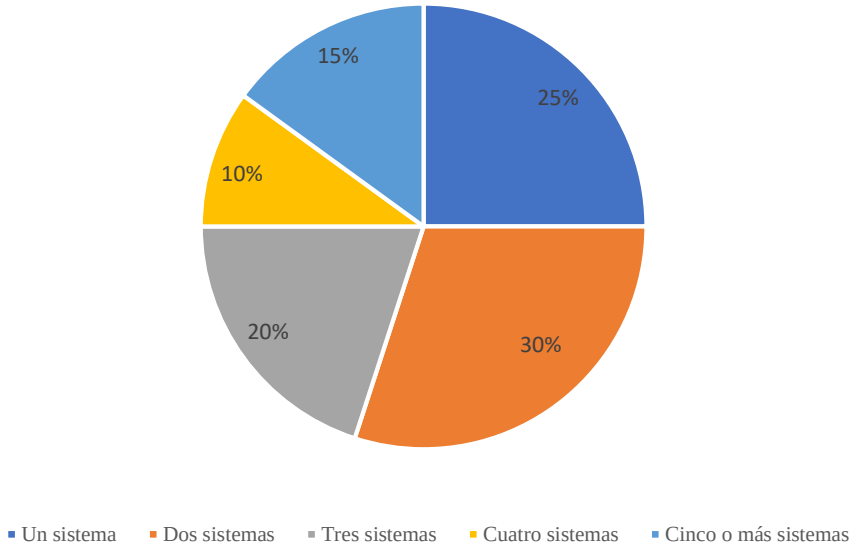


Nota. Las bases de datos institucionales almacenan los datos que dan soporte a las operaciones de la Universidad, así mismo, las encuestas indican que el 75% del personal entrevistado consideran de mucha utilidad los datos para dar respuesta a las peticiones de áreas internas y externas. Mientras que un 25% la consideran escasamente útil. Por lo tanto, faltan aplicar reglas claras que mejoren la calidad de los datos almacenados.

Así mismo, aunque represente un trabajo complejo y con inversión de tiempo, los reportes son entregados en un 80% de forma oportuna, pero no eficiente, debido a que se deben consultar más de un sistema transaccional para poder dar cumplimiento con los requisitos solicitados por las distintas áreas, el 75% del personal responsable de la emisión de reportes acceden a varios sistemas de sistemas de información, en general entre dos y cinco sistemas distintos, para estructurar su labor (Figura 6).

Figura 6

Porcentaje de sistemas transaccionales consultados



Nota. El 75% del personal responsable de la creación de reportes y estadísticas en las distintas áreas de la Universidad, requieren acceder a dos o más sistemas para poder definir un reporte de forma consistente. Solo el 25% lo realiza de una sola fuente de información.

Por otra parte, el tiempo invertido para entregar los reportes solicitados con mayor frecuencia (trabajo cotidiano) y menor frecuencia (auditorias), están en el rango de una hora hasta una semana, esto indica que los reportes y estadísticas requieren varias revisiones y ajustes antes de contar con la versión final de la misma. Como se ha comentado, se realiza trabajo manual o se contemplan validaciones poco efectivas. La tabla 1 muestra el tiempo requerido para la emisión de un reporte de acuerdo a su frecuencia de uso.

Tabla 1

Tiempo requerido para la elaboración de reportes y/o estadísticas

Frecuencia alta	Porcentaje	Frecuencia baja	Porcentaje
De 1 a 4 horas	25	Menos de 8 horas	45
De 5 a 8 horas	55	Un día	10
Dos días	15	Entre dos y tres días	30
Tres a cuatro días	0	Una semana	15
Una semana	5	Más de una semana	0
Total	**100**	**Total**	**100**

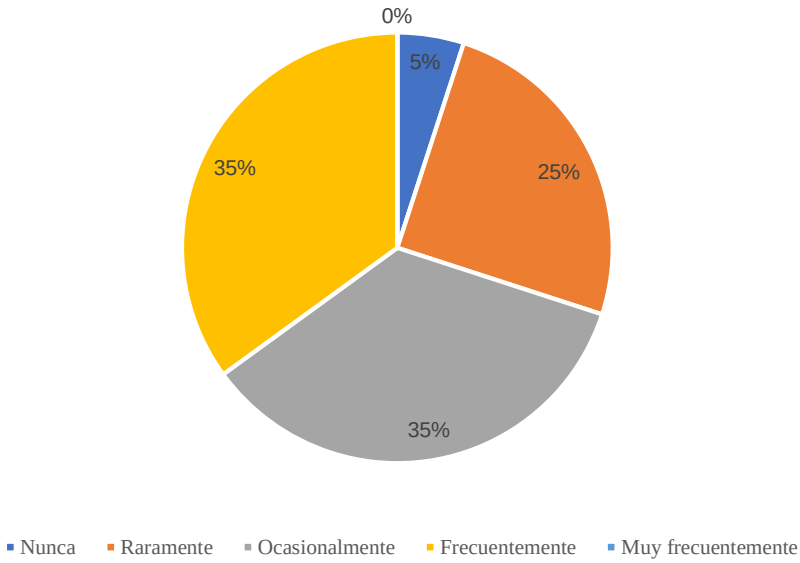
Nota. Los reportes solicitados por entidades externas a la UAQ, requieren mayor tiempo de procesamiento y atención dependiendo del trabajo manual que realizan para revisar y completar la información. Mientras que los reportes que solicitan en el trabajo cotidiano presentan un tiempo considerablemente menor, sin embargo, es factible optimizar los tiempos de elaboración.

Por último, es importante resaltar que los niveles de calidad, confiabilidad, integridad, coherencia, accesibilidad y precisión de los datos son factores que fortalecen la competitividad para cualquier empresa o institución. En este sentido, la presencia de valores faltantes, duplicados y desactualizados pueden generar como consecuencia una calidad en los datos baja, afectando directamente la elaboración de reportes y estadísticas oportunas para el análisis. Si bien, es complicado lograr la completitud en cuanto a la calidad de los datos, dentro de la Universidad se tiene una frecuencia del 70% con posibles inconsistencias menores en los datos consultados (Figura 7).

Por lo anterior, se recomienda la implementación de estrategias tecnológicas y soluciones eficaces para aumentar la integridad y nivel de pertinencia de los datos con base a la arquitectura de negocio. Por lo tanto, se refuerza la credibilidad y confianza en los resultados de los reportes y estadísticas obtenidas, además se refuerza la capacidad de la Universidad para el tomar decisiones informadas y basadas en hechos reales.

Figura 7

Frecuencia de valores con inconsistencias

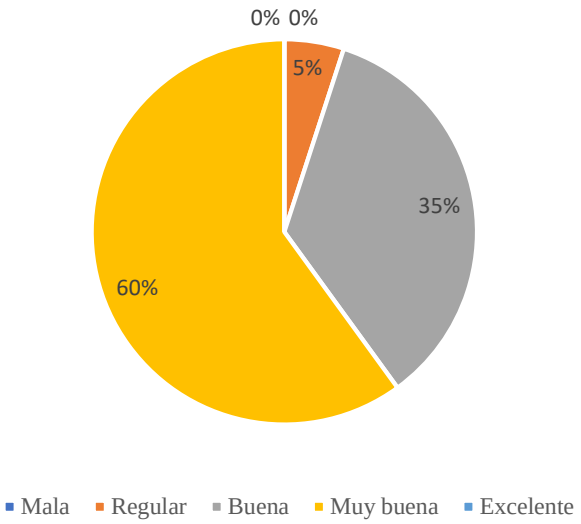


Nota. Las inconsistencias de cualquier tipo (menores o mayores), se encuentran con una frecuencia del 75% (ocasionalmente, frecuentemente y muy frecuentemente) al momento de general los reportes y/o estadísticas.

Como consecuencia es necesario realizar procesamiento manual e intervención continúa durante la fase de la elaboración de los reportes. Esto implica una inversión de tiempo significativa por el personal encargado, ya que cada reporte requiere un procesamiento minucioso y, en algunas ocasiones, incluso complejo. (tal cual se muestra en la Figura 8). Al realizar operaciones de manera manual como la identificación y corrección de valores erróneos o la integración y centralización de los diversos orígenes de datos, se convierte en acciones complejas y con mucha inversión de tiempo, con alto grado de errores humanos y sobretodo, retraso en la obtención de la información. Como se ha mencionado, se requiere buscar soluciones enfocadas a la automatización y optimización de estos procesos.

Figura 8

Nivel de confiabilidad de los datos



Nota. Ninguna persona entrevistada considera que confiabilidad de los datos es ideal o excelente, lo que indica la factibilidad para realizar mejoras en esta percepción del usuario. El nivel de confiabilidad de los datos consultados refiere un 60% como muy buena y un 35% buena.

Asimismo, se aplicó un segundo instrumento, dirigido a los *administradores de base de datos y los responsables de desarrollo de sistemas*, el instrumento consta de 13 reactivos en escala *Likert* y fue aplicado a nueve personas responsables de la parte técnica de la UAQ durante el mes de febrero del 2022, se determinó un alfa de Cronbach o nivel de fiabilidad de 0.817, lo que indica un nivel de fiabilidad del instrumento bueno. El objetivo principal del instrumento era identificar aquellos factores técnicos, específicamente la perspectiva del personal técnico, que representaban un impacto considerable en el proceso de emisión de reportes y estadísticas de manera confiable y oportuna. Al revisar los resultados de los dos instrumentos, se determinó que los factores que inciden en la calidad de los datos también afectan el proceso de generación de reportes. A partir de dichos resultados, se identificaron los siguientes problemas en el proceso:

1. Recopilación de los datos: Los datos se almacenan en distintos sistemas de información de tipo transaccional, lo cual dificulta contar con datos precisos y actualizados. De igual manera, representa esfuerzos extraordinarios para poder generar la información necesaria, por medio de herramientas extras, por ejemplo, hojas de cálculo, macros, entre otras. Es por ello, la necesidad de contar con un repositorio único de datos que contenga los datos esenciales e íntegros.
2. Tiempo invertido en el análisis de datos: Debido a que los datos son consultados en distintos orígenes de datos, y, en algunas ocasiones, datos con distinto nombre pero que hacen referencia al mismo concepto, obligan a revisar cuidadosamente los registros obtenidos en primera instancia.
3. Diccionario de datos y estandarización de catálogos institucionales: Los catálogos maestros son esenciales en cualquier sistema transaccional y se vuelven obligados al momento de realizar cualquier tipo de explotación de los datos, la Universidad requiere homologar dichos catálogos con la intención de facilitar el proceso de análisis datos. De igual manera, es necesario revisar cuales son los catálogos utilizados por las entidades externas con el objetivo de considerar la forma de agrupación de los datos. Así mismo, dentro de la Universidad se debe centralizar los catálogos para que todas las aplicaciones compartan un lenguaje común.
4. Procesamiento manual de los datos: Debido a los puntos anteriores se provoca un procesamiento manual o semiautomático para realizar agrupaciones de los datos, depuración y generación de tablas dinámicas que permiten dar respuesta lo solicitado por las entidades internas y externas de la Universidad.
5. Calidad de los datos consultados: Las características de consistencia, precisión, integridad, validez y puntualidad son elementos necesarios para la explotación de los datos mediante reportes y/o estadísticas. Dichas características se ven afectas después de analizar los datos de entrada y manipularlos con hojas de cálculo, funciones y tablas dinámicas. En

consecuencia, es necesario definir un repositorio único que contenga los datos migrados después de un proceso de calidad bien definido.

6. Reportes históricos y silos de datos: Para simular escenarios o tendencias se debe contar con datos históricos estructurados de todas las áreas de la Universidad.

Por otra parte, los Comités Interinstitucionales para la Evaluación de la Educación Superior (CIEES), reconoció a la Universidad por su calidad académica y administrativa tras una revisión que dio inicio en el año 2021 y finalizó en febrero del año 2022 con la obtención de dicha acreditación. No obstante, la visita realizada por la Comisión de Pares Académicos en modalidad presencial a la UAQ brindo la oportunidad de localizar áreas de oportunidad en los procesos administrativos a través de un informe de evaluación. Dicho informe, establece que la Universidad requiere de un sistema único de información, con información actualizada, accesible y transparente, que fortaleza la rendición de cuentas internas y externas por medio de la emisión de reportes y, por ende, facilite el proceso de toma de decisiones. El cual, puede ser constituido a partir de los distintos sistemas que contienen la información aislada en los núcleos de información, lo que disminuye la eficiencia en la elaboración de reportes, la revisión de la información y, por consecuencia, la toma de decisiones.

Asimismo, la Secretaria de Finanzas de Gobierno del Estado de Querétaro, el 15 de noviembre del 2022, solicito a la Universidad la corrección de inconsistencias detectadas sobre la emisión del timbrado de los CFDI o en el entero del ISR de algunos ejercicios anteriores. Las discrepancias detectadas surgen debido a la falta de supervisión y control de los datos utilizados en el proceso de timbrado de nóminas, los cuales se encuentran distribuidos de diversas áreas de la Universidad, lo cual dificulta el análisis en tiempo y forma. Además, la ausencia de datos en algunos registros y la duplicidad de datos, provocada por la captura y procesamiento manual, son factores que, entre otros, inciden en la calidad de los datos.

La elaboración de reportes es una de las actividades más comunes dentro de la Universidad, tanto para entidades internas o externas, por ello, las áreas responsables del manejo de los datos (desarrolladores de aplicaciones, analistas de procesos, analistas de datos y responsables de procesos) concuerdan en los siguientes puntos referente al proceso actual de elaboración de reportes y estadísticas:

1. Se debe fortalecer el proceso de elaboración de reportes para su posterior automatización. A través de los años y derivado del crecimiento de la Universidad se han desarrollado diversas aplicaciones para dar soporte a los distintos procesos administrativos y académicos,
2. Definir catálogos gubernamentales (externos) e internos para su interpretación entre las áreas, de manera tal que se inicie un proceso de estandarización y alineación de datos. En el caso de los catálogos internos, establecerá un lenguaje común entre la Universidad y, para el caso de los catálogos externos se asegura cumplir con las normativas establecidas.
3. Definir herramientas flexibles que potencialicen la elaboración de reportes eficientes y oportunos son elementos clave para gestionar de forma eficiente y eficaz la información de la Universidad. Las herramientas tecnológicas deben presentar los datos por medio de gráficas, tablas e indicadores que faciliten la visualización de información, así como la identificación de patrones o tendencia.
4. Consolidar la información de las áreas o núcleos de información para mejorar el análisis, la elaboración de reportes y estadísticas. Con la integración se pretende unificar los datos almacenados en varias áreas eliminando los silos y creando una visión completa y holística de la institución.

2. Antecedentes y fundamentación teórica

2.1. Antecedentes

Las Universidades tienen una responsabilidad social muy alta, debido a que en ellas se brindan los conocimientos que fortalecen el desarrollo de un país, por ello, es importante manejar los procesos internos con eficiencia, eficacia y calidad; para apoyar el buen uso de sus recursos económicos, materiales y capital humano. La ejecución de los procesos administrativos y académicos trae como resultado la producción de grandes cantidades de datos almacenados en los diversos sistemas de información; mismos que son afectos, en primer lugar, por constantes cambios en procesos de cada administración, así mismo, el desconocimiento de las reglas del negocio por parte del personal responsable. Los puntos mencionados anteriormente generan incertidumbre y desorganización en la manera en que se almacenan los datos (Reyes y Nuñez, 2015).

En la era de la información, aplicado al ambiente educativo, prevalece la creación del conocimiento y los proceso para tomar decisiones a través de la forma en que se recopilan, estructuran y almacenan los datos para su posterior explotación, por lo cual, el desafío consiste en, a partir de los datos crudos, producir información de utilidad para la organización y sirva de base para la toma de decisiones. Por ello, se requiere el uso de herramientas y metodologías que ayuden a acceder a las fuentes de datos (sin importar los diferentes formatos y fuentes heterogéneas de las que provienen), importarlos e integrarlos en un repositorio diseñado especialmente para su posterior estudio; a lo anterior se le denomina Inteligencia de Negocios, o en inglés *Business Intelligence* (BI), (Mamani, 2018). La parte central de un proyecto BI es el almacén de datos denominado *Data Warehouse* (DW) o bodega de datos, el cual se define como un repositorio diseñado para almacenar, administrar y explota volúmenes masivos de datos. En este sentido, la implementación del DW proporciona datos actualizados y confiables, así como, analizarlos con diferentes niveles de agrupación (Fuentes et al., 2019).

El uso de los DW permite eliminar los silos de información que son generados a lo largo de la historia de una empresa o institución. Además, permite centralizar los datos, contienen los datos depurados, homologados y, en su caso, procesados, lo que facilita la utilización de información confiable y de manera eficiente. Por ejemplo, el diseño de un DW en el sector de telecomunicaciones fortaleció la toma de decisiones en la empresa a nivel estratégico, mediante cuadros de mando integrales, donde se visualizan los incidentes con los clientes, visitas, seguimientos y ubicación geográfica (Jaramillo y Pauta, 2019).

La construcción de un repositorio central de datos parte de la revisión y estudio de los requisitos del negocio con la intensión de identificar que datos deben ser migrados utilizando procesos para extraer, transformar y cargar los datos en un nuevo repositorio. Dicha migración consolida la información para su posterior explotación mediante herramientas especializadas. En el sector financiero, permitió eliminar reportes elaborados por distintos departamentos, los cuales eran elaborados manualmente con un riesgo alto de errores de recaptura de la información, en este sentido, los reportes pueden ser generados de forma eficiente y reduciendo el margen de error (Palacios, 2017).

Con los ejemplos anteriores se demuestra que el uso del DW permite analizar los datos de manera eficiente, congruente y con mayor rapidez que en los sistemas transaccionales. En este sentido, el DW es una herramienta que fortalece el estudio, seguimiento y revisiones de tendencias históricas o proyecciones de cualquier tipo de información, así como reutilizar los esquemas generados y metodologías en ambientes parecidos, es decir, de la misma rama de estudio. La tecnología de los DW otorga mecanismos para procesar grandes volúmenes de datos con rapidez en el procesamiento. Por ejemplo, el Programa para la Evaluación Internacional de Alumnos (PISA), fue un caso de estudio en el 2018, en la cual, se realizó el estudio de diversos indicadores con resultados satisfactorios utilizando un DW para analizar la cantidad de estudiantes, promedios y otras métricas, así mismo, se utilizaron dimensiones como estudiante, país, tipo de prueba , tiempo, etc., generado consigo

un modelo reutilizable en los distintos años de aplicación de la prueba (Zambrano et al., 2018).

2.2. Fundamentación teórica

Las empresas de cualquier rubro o giro almacenan gran cantidad de datos debido a las diversas transacciones realizadas durante su historia, sin embargo, solo un número reducido de ellas dan uso intensivo sobre los datos. Lo anterior, por el esfuerzo inicial que representa el tratamiento de los datos almacenados, y que posteriormente serán analizados, en este sentido radica la importancia de crear información de utilidad para la organización, dicha información permitirá ser el sustento en la emisión de reportes y estadísticas que fortalezcan la toma de decisiones fundamentada. Para lograrlo es necesario utilizar herramientas tecnológicas, diversas arquitecturas y metodologías que permitan acceder a las fuentes de datos de manera sencilla, así como, aplicar transformación y cargas masivas a un repositorio único que contendrá los datos necesarios para apoyar la explotación de la información (Lennerholt et al., 2018).

2.2.1. *Business Intelligence* (Inteligencia de negocios)

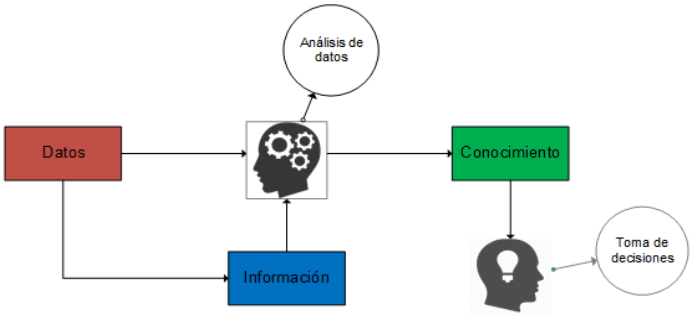
La inteligencia de negocios (BI) se define como un grupo de componentes, metodologías, software y buenas prácticas enfocadas en la información y la gestión de datos, permitiendo una mejor toma de decisiones en el nivel estratégico, táctico y operativo (Conesa y Curto, 2011). Del mismo modo, se puede considerar el BI como el conjunto de tecnologías que son el soporte a los procesos de la gestión de los datos, información, conocimiento y aprendizaje de la empresa (Rodríguez, 2019). En este sentido, BI se trata de una mezcla de mejores prácticas, habilidades gerencias y tecnológicas aplicadas en la organización para organizar, recolectar e integrar la información, mediante un conjunto de políticas o reglas de negocio y conocimiento de sus procesos administrativos. La principal característica de un proyecto BI consiste en convertir el dato en información y que esta derive en conocimiento (Muñoz et al., 2016). Por lo tanto, la inteligencia de negocios

contempla tres elementos esenciales para la organización y los procesos de elaboración de reportes, en seguida se describe cada uno de ellos:

- *Los datos*: Son considerados la materia prima de la información; son valores aislados que no poseen significado de manera individual, y las organizaciones poseen gran volumen de ellos. Además, se consideran como un conjunto de registros o transacciones que no representan nada al no tener un contexto determinado y suele estar asociado a un objeto. Sin embargo, son la base para obtención de la información. (figura 9).
- *La información:* Es considerado como la integración de varios datos de manera estructurada que brindan un propósito a los eventos al estar asociado a un contexto. Está integrada por menos cantidad de datos, pero representa mayor valor para la organización al ser la fuente principal para tomar decisiones fundamentadas que disminuyen el rango de error e incertidumbre.
- *El conocimiento*: Es la integración entre los valores y experiencias que una o un conjunto de personas tienen, y la información recibida de manera oportuna. Por tanto, el conocimiento representa el mayor valor para la empresa al ser de gran utilidad para la ejecución de acciones determinadas (Davenport y Laurence, 1998).

Figura 9

Elementos clave de una solución BI



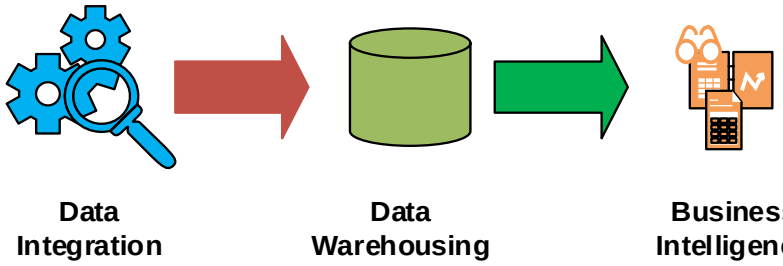
Nota. Los datos son los elementos primarios para construir cualquier sistema de información desde lo transaccional a lo estratégico.

Las soluciones tecnológicas de tipo BI, centralizan los datos de la empresa u organización que, normalmente, tienen su origen en distintas aplicaciones y diversos formatos, además de las bases de datos operacionales, archivos XML, archivos de texto plano o TXT, hojas de cálculo y otras fuentes que contengan datos relevantes para la elaboración de reportes significativos y que agreguen valor al negocio. Dichos datos fueron tratados por medio de herramientas especializadas para definir su arquitectura, modelado, limpieza y calidad; fortaleciendo con ello la versión única de la información, elaboración de reportes, producción de información histórica para cualquier nivel jerárquico de la organización, por lo anterior, la arquitectura de una solución BI consta de conceptos clave, entre ellas se encuentra: fuentes de datos, procesos ETL, DW y/o Data Marts (DM), esquemas dimensionales y herramientas para generación de reportes (Sherman, 2015). Por lo anterior, en términos generales, las soluciones de BI están conformadas por una estructura que contemplan las siguientes fases (Figura 10):

- *Data Integration*: Se trata de la unificación o integración y depuración de datos provenientes de diversas fuentes
- *Data Warehousing*: Centraliza los datos depurados
- *Business Intelligence*: Representa las visualizaciones de datos

Figura 10

Flujo base de una solución BI

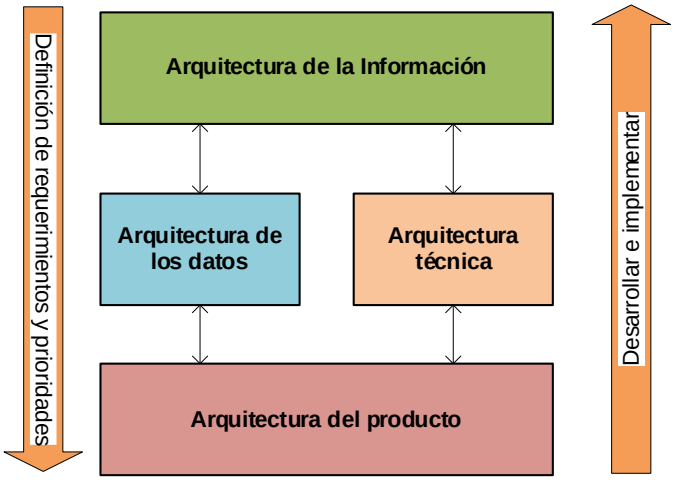


Nota. Los proyectos de BI mantienen elementos comunes sin importar el giro del negocio, siendo la parte medular el desarrollo del DW. Adaptado de *How BI, DW and DI fit together (p.15)*, de Rick Sherman, 2015, *Business Intelligence Guidebook.*

La arquitectura de una solución BI se define como la combinación de las tecnologías, un conjunto herramientas tecnológicas y procesos de negocio que facilitan la conversión de los datos en información, esta información en conocimiento y, finalmente, la aplicación de ese conocimiento en los procesos analizados, con el objetivo de generar valor a la organización y fortalecer la generación de reportes y estadísticas. (Figura 11), (Sherman, 2015). Es importante mencionar que los proyectos BI pueden fracasar a consecuencia de los procesos con poca integración con el resto de la organización, por la falta de comprensión de los orígenes de datos o por la manera incongruente de visualizar los datos; por lo cual la arquitectura BI contempla la parte tecnológica y de gestión (Testa y Malbernat, 2018).

Figura 11

Las cuatro capas de la arquitectura de una solución BI



Nota. Proporcionar información de utilidad, que genera valor y que sirva como base para la elaboración de reportes y estadísticas es un proceso complejo, aunado a esto se debe considerar la gran cantidad de datos dispersos en la organización, por ello es importante considerar el desarrollo de cada una de las etapas de la arquitectura BI. Adaptado de *The four architecture categories (p. 67)*, de Rick Sherman, 2015, *Business Intelligence Guidebook.*

2.2.2. Definir los requerimientos del proyecto BI

El primer punto para que una organización implemente una solución de tipo BI considera la revisión detallada de cada requerimiento solicitado por las áreas o departamentos que estarán involucrados. El éxito o fracaso del proyecto dependerá del trabajo realizado en esta etapa, por lo cual, es necesario considerar los siguientes aspectos:

a) Cada uno de los requerimientos debe ser claro y conciso, además descrito lo mejor posible para evitar ambigüedades
b) Enfocarse solo en los requerimientos del proyecto
c) Evitar en la medida de los posibles cambios en los requerimientos iniciales para culminar exitosamente el proyecto
d) Considerar las reglas del negocio y proceso administrativos

Aunado a lo anterior, se debe definir claramente cuál será el proceso de negocio que se desea implementar, construir un marco técnico y definir el alcance del proyecto. (Sherman, 2015).

2.2.3. Arquitectura de la información

El concepto de arquitectura de la información responde las preguntas: "¿qué?, ¿quién?, ¿dónde? y ¿por qué?" del proyecto de inteligencia de negocios:

- Definir *qué* proceso del negocio se desea desarrollar con el proyecto, además definir *qué* tipo de análisis de datos es el adecuado y, por último, *qué* tipo de decisiones son las que se espera resolver.
- Al contar con un repositorio centralizado de datos, se debe definir *quiénes* podrán acceder, por ejemplo, personal estratégico, táctico, operativo, clientes, proveedores, estudiantes, entre otros.
- En cuanto a la parte técnica establecer *dónde* estarán los datos, *dónde* se llevará a cabo el tratamiento de los datos y *dónde* se mostrarán
- Además, establecer el *porqué* de la solución BI, es decir, definir los requerimientos del negocio y técnicos.

Para establecer el panorama general en la creación de una solución de tipo BI se considera importante definir los requerimientos del área y desarrollar los puntos de la arquitectura de la información (Sherman, 2015).

2.2.4. Arquitectura de datos

En cuanto a la arquitectura de datos, considera el análisis, entendimiento y, en su caso la definición de cada elemento que constituye a los esquemas, así como sus esquemas. Además, considera cada uno de los procesos para extraer, transformar y cargar los datos en los repositorios. Así mismo, establece los pasos para atender los requerimientos provenientes de la capa de información. Esta fase comienza cuando los datos son creados en los diversos orígenes datos, es decir se ubican y analizan dichos orígenes, de los cuales se obtendrán los datos para solucionar los requerimientos de información del usuario. Posteriormente, finaliza cuando los datos se presentan a los usuarios finales por medio de herramientas visuales para su explotación, en este sentido, se crean las medidas o métricas en la tabla de hechos, las dimensiones y sus jerarquías (Sherman, 2015).

2.2.4.1. Fuentes de datos

Las fuentes de datos utilizadas en la implementación de aplicaciones BI pueden clasificarse en dos tipos: en internos y externos. Los datos internos corresponden a la información almacenada por la organización en sistemas que cuentan con base de datos transaccionales, entre dichas bases de datos se encuentran, ORACLE, SQL Server, MySQL, hojas de cálculo, archivos XML, TXT, entre otros. Por otra parte, las fuentes de datos externos, son aquellas no pertenecientes directamente a la organización incluyen textos, archivos Lenguaje de Marcas de Hipertexto del inglés *HyperText Markup Language (*HTML), bases de datos disponibles para consulta, etc. Estos datos suelen presentarse de manera heterogénea, en diversos orígenes de datos y distintos formatos, ya sea en modelos estructurados o en simples archivos. Para integrar eficazmente estos datos en una solución BI, se requiere establecer un formato homogéneo y establecer reglas que garanticen su integridad (Strand y Syberfeldt, 2020).

2.2.4.2. Calidad de los datos

Cada uno de los datos capturados a través de las plataformas tecnológicas son la materia prima para el DW, por tal motivo es necesario validar la integridad de los mismos. En algunas ocasiones es imposible mantener la calidad deseada de los datos, por ello, al migrarlos de un sistema transaccional a un sistema para el análisis de información se deben conocer las reglas y procesos del negocio, lo anterior con el objetivo de aplicar las transformaciones puntuales sobre los datos, obteniendo con ello datos consistentes y de calidad. Dentro del proceso de migración, los datos provenientes de múltiples fuentes, se recolectan y almacenan en el destino, lo cual, a consecuencia de los distintos sistemas transaccionales y formatos heterogéneos puede dificultar su extracción y depuración (Zelaya et al., 2018).

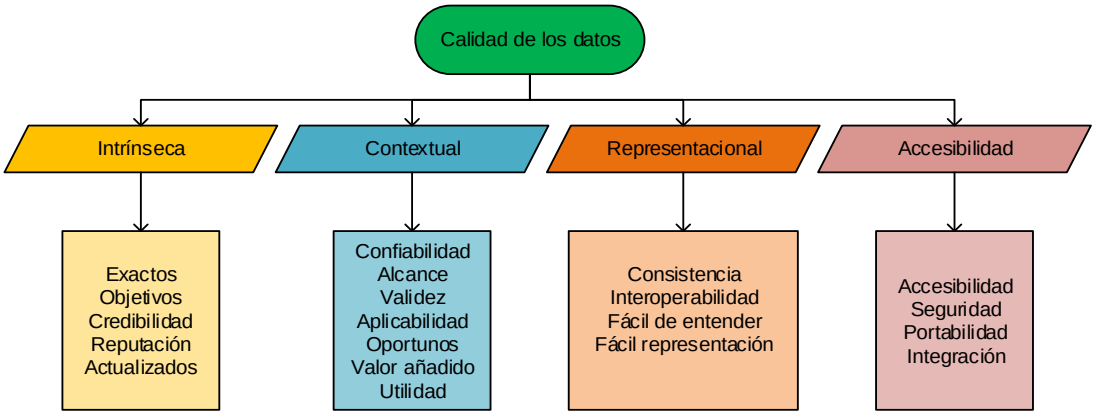
Cuando una compañía realiza proyectos orientados a mantener los datos de manera segura, confiable y actualizados, debe contar con una estrategia para la administración de sus datos, por tanto, las características esenciales para la gestión de datos es el desarrollo de procesos que controlar la calidad y aseguran que estos tienen la integridad necesaria para que la organización haga uso de ello de manera confiable. En este sentido se desarrolló un marco de referencia, a partir del análisis de varios atributos que clasifican la calidad de los datos, así como la valoración que las personas dan sobre los datos que consideran de calidad en menor o mayor medida. En el análisis se identificaron varias características que definen la calidad que poseen los datos, entre ellas se encuentran la credibilidad, el valor añadido, la interoperabilidad, la accesibilidad, etc... Estás características se agruparon en cuatro categorías; intrínseca, contextuales, representacionales y de accesibilidad (Strong et al., 1997).

Con base lo presentado anteriormente, en primer lugar, la característica *intrínseca*, se refiere a la calidad que posee los datos por sí mismos, en otras palabras, la calidad es inherente al dato. Las características principales de esta categoría abracan la precisión o exactitud, indicando que los datos son confiables y libre de errores e inconsistencias, son exentos de duplicidad y presentan sus valores

de manera correcta. Además, se consideran objetivos, al carecer de sesgos y datos parciales, aportando credibilidad y reputación por lo cual se consideran veraces y altamente confiables en términos de las fuentes. En segundo lugar, la calidad *contextual* considera que los datos deben ser relevantes o aplicables a las actividades actuales, es decir, para considerar este rubro, los datos deben ser oportunos, completos y apropiados en cuanto a la cantidad y el valor. En tercer término, la calidad de datos *representacional*, está relacionada con las plataformas tecnológicas y su capacidad de mostrar los distintos datos de manera adecuada, que sean interpretables, de fácil entendimiento, de representación concisa o compacta y consistentes. Por último, la calidad de datos de *accesibilidad* u *operacional* resaltan los aspectos como la accesibilidad, la seguridad, la personalización y la colaboración dentro de los sistemas de información de la empresa (Figura 12) (Cai y Zhu, 2015).

Figura 12

Clasificación sobre la calidad de los datos



Nota. Las características de calidad de datos establecen que los datos deben estar disponibles en el momento y contexto especifico, manejan un alto nivel de interoperabilidad o grado en que los datos pueden ser personalizados y adaptados a preferencia de los usuarios, así mismo, representan facilidad en su operación, son portables o presentan un grado en que pueden ser integrados en otras plataformas de un contexto de uso específico.

Desafortunadamente la mayoría de las organizaciones cuenta con datos e información duplicada lo cual provoca resultados irrelevantes o erróneos, además, a consecuencia del crecimiento, la falta de análisis al crear base de datos, migraciones u operaciones sobre los datos existentes y la omisión de restricciones al momento de almacenar los datos provoca el almacenamiento de datos incompletos, por lo anterior, existe una clasificación de acuerdo a las problemáticas que puedan desarrollar los datos dentro del sistemas de información (Oliveira et al., 2005). La tabla 2 y 3 muestran dicha clasificación.

Tabla 2

Causas que disminuyen la calidad de los datos de un origen

Factor	Descripción
Valores ausentes	Denominados valores NULL, por alguna razón no están presentes en el registro.
Violaciones en sintaxis	El formato preestablecido de los datos es incorrecto, por ejemplo el campo de la fecha con un formato Día/Mes/Año, si al almacenarse quedara algún registro con Año/Mes/Día, este carece de valor.
Valores antiguos	Registros que no contienen información real a la fecha.
Violaciones de intervalos	Registros que están fuera de un rango establecido. Es evidente en registros numéricos y de tipo fecha.
Violaciones de unicidad	Poco frecuente, suele darse en campos de tipo único, que por alguna razón se duplica, por ejemplo, RFC, Expediente, etc.
Violaciones de integridad referencial	Se presenta cuando la relación entre tablas se pierde a consecuencia de inserciones o modificaciones en los datos.
Datos duplicados	Los datos duplicados pueden estar representados de varias maneras, dependiendo del modelo de datos, de las restricciones y de sus características.

Nota. Las causas de una sola fuente de datos son poco frecuentes, debido a que las organizaciones desarrollan sus procesos de negocio y actividades en varios sistemas transaccionales, sin embargo, es indispensable depurar cada una de las inconsistencias antes de migrar al DW.

Tabla 3

Causas que disminuyen la calidad de los datos de múltiples orígenes

Factor	Descripción
Diferentes métricas	Las columnas almacena valores distintos dependiendo de la fuente de datos, por ejemplo, en una almacenar en metros y en otra en centímetro.
Representaciones inconsistentes	Los valores se representan de diferente forma, sin importar que signifiquen lo mismo, por ejemplo, el sexo, en un base de datos puede ser almacenada como “F”, “M” mientras que en otras “Masc”,”Fem”.
Redundancia sobre entidades	La misma entidad es representada en diferentes bases de datos, y posiblemente contienen el mismo número de atributos.
Inconsistencias sobre entidades	Registros donde no se respeta el mismo formato para diferentes bases de datos.

Nota. Los factores de múltiples orígenes de datos son más complicados de diagnosticar y controlar, debido a la cantidad de sistemas transaccionales y a la falta de documentación y control de procesos.

2.2.4.3. Staging Area

El *staging area* es el área de preparación que simplifica el tratamiento de los datos, su limpieza, homologación y consolidación de las distintas fuentes. Es una capa intermedia entre el origen de los datos y el destino teniendo como objetivo principal garantizar la calidad de los datos previos a la migración hacia el DW, esto se logra mediante la duplicación de los datos contenidos en los orígenes hacia el área de preparación, donde se aplican transformaciones específicas con base a los requerimientos del negocio. Además, se cuenta con un control completo sobre los datos, debido a que las operaciones complejas y transformaciones realizadas sobre ellos no tendrán ningún impacto en los sistemas de información de tipo transaccional (Potineni, 2021).

2.2.4.4. *Data Integration*: Proceso ETL (Extracción, Transformación y Carga)

Los procesos ETL se compone de tres elementos principales, por una parte, el proceso de extracción, enseguida las transformaciones y, por último, la carga de los datos, dichos elementos son utilizados por un origen de datos o comúnmente a partir de la existencia de múltiples aplicaciones o soluciones tecnológicas que permiten el crecimiento de las organizaciones, sin embargo, utilizan varios medios de almacenamiento y gestores de base de datos lo que implica contar con información dispersa. Las consecuencias de obtener reportes de estos sistemas operacionales dieron lugar a utilizar una fuente de datos para cualquier tipo de procesamiento, ya sea transaccional o analítico. Sin embargo, al realizar reportes que involucren historicidad y millones de registros provoca lentitud en los sistemas operacionales, incongruencia en los resultados, etc. Es por ello, la necesidad de crear repositorios para la explotación específica de la información (Montoya y Jiménez, 2018).

Los procesos ETL representan alrededor del 70% de los recursos y el tiempo requerido para construir un DW o *Staging Area*. Estos procesos ETL son utilizados para extraer datos de diferentes orígenes, aplicar transformaciones en caso de ser necesario como la limpieza y estandarización de los datos entre diferentes orígenes, para lograr una mayor calidad en los datos, y finalmente almacenarlos en el repositorio destino. Retomando lo mencionado por Bill Inmon (2005): “Un almacén de datos es una colección de datos orientados por temas, integrados, no volátiles y variables en el tiempo en apoyo a la elaboración de reportes”. En este sentido, el proceso de extracción, transformación y carga permiten integrar múltiples orígenes de datos priorizando su calidad dentro del DW. Por ejemplo, se ejecutan acciones para unificar datos de las diversas fuentes de datos, se definen agrupaciones, claves, identificadores, asimismo, se realizan conversiones entre formatos y, en general, todo aquello que permita aumentar la calidad de los datos (Souibgui et al., 2019).

El proceso de ETL se encarga de identificar, extraer y dar formato a los datos de mayor relevancia de los sistemas transaccionales, adaptando y sincronizando las diversas fuentes y plataformas tecnológicas; además se realizan acciones para depurarlos y homologarlos, respetando los esquemas de seguridad e integridad de la fuente, con el objetivo de cargarlos en un DW o DM según el diseño preestablecido; dichas herramientas son responsables de concluir exitosamente la implementación de un DW (Yulianto, 2019). Los procesos ETL inician con la identificación de los sistemas fuente u orígenes datos, por ejemplo, las hojas de cálculo, documentos de tipo CSV, texto plano TXT, bases de datos relacionales, base de datos no SQL, etc…Por lo anterior, se requiere realizar las operaciones ETL considerando los conceptos clave que el negocio ha impuesto y que son la base para definir las jerarquías, niveles de agregación y validaciones al momento de realizar la combinación, depuración y homologación de datos (Tabla 4). Al definir el nivel de granularidad del proyecto, se toman en cuenta los diversos orígenes de datos, asimismo, se debe considerar el riesgo de sobrecarga o bajo rendimiento en los sistemas de información dependientes de dichos datos, en consecuencia, se requiere establecer el momento adecuado para ejecutar el trabajo de migración y, de esta manera, consumir el mínimo de recursos en horarios que no interfieran con las operaciones cotidianas de la organización. Además, en el tratamiento de los datos es necesario considerar que algunos datos no contarán con las reglas de integridad básicas, ni con las reglas del negocio (Montoya y Jiménez, 2018).

Tabla 4

Procesos ETL - Extracción, Transformación y Carga

Proceso	Datos de entrada	Acciones	Resultado
Extracción	Orígenes de datos, hojas de cálculo, sistemas de información de tipo transaccional, TXT, entre otros.	Organización y clasificación de datos - selección	Datos crudos
Transformación	Datos con inconsistencias	Transformación, cálculos homologación, agrupación	Datos formateados, estructurados y resumidos
Carga	Datos formateados y resumidos de acuerdo a las necesidades del negocio	Inserción	Persistencia en el DW

Nota. Los procesos ETL identifican los orígenes de los datos, su recolección, el formateo o transformaciones de datos y la carga en dimensiones o nuevas tablas en el área de preparación.

Las instituciones tanto públicas como privadas experimentan un constante proceso de cambio y crecimiento, dando como resultado de esta evolución, una incorporación de diversos sistemas de información que respaldan las operaciones diarias de los procesos administrativos. Así mismo, surgen nuevas necesidades de información y una perspectiva mayor y holística de los datos, en este sentido, se requiere contar con un DW o bodega de datos que facilite a los diferentes perfiles de usuarios que requieren tomar decisiones estratégicas con base a información oportuna e integra, de acuerdo al contexto o circunstancias determinadas. El DW requiere ser poblado a través de los procesos y técnicas ETL, sin embargo, al implementar este tipo de sistemas ETL se necesitan integrar varios sistemas, entre obsoletos y modernos, debido al dinamismo de las empresas. Los procesos ETL proporcionan elementos que facilitan la selección y extracción de los datos provenientes de diversos orígenes, utilizando una serie de reglas que garantizan una mayor calidad y consistencia en los datos, además, posibilita la estandarización y homologación de los datos procedentes de las plataformas tecnológicas por medio de una serie de transformaciones. Por último, el proceso carga, consiste en

almacenar los datos en los objetos destino, listos para su posterior explotación mediante software especializado en el análisis de información (Medina et al., 2018).

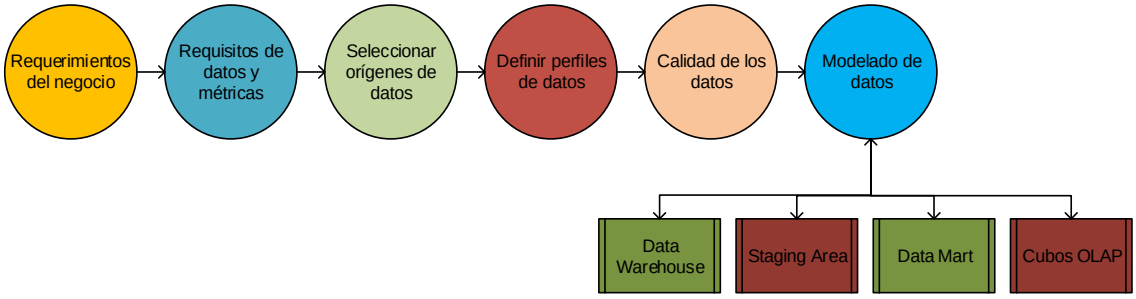
Los procesos ETL además del valor significativo que aportan a los datos por medio de la migración de las fuentes de datos origen al repositorio destino, también son responsables de:

- Realizar las transformaciones necesarias para eliminar errores y corregir datos faltantes.
- Registrar detalladamente las acciones realizadas para incrementar la calidad de los datos.
- Supervisar los flujos de datos en los sistemas de información.
- Consolidar las múltiples fuentes de datos.
- Estructurar los datos con base a los requerimientos de los usuarios finales.

La integración de datos o los procesos ETL requieren el establecimiento o definición de la arquitectura de datos para poder iniciar con esta fase. Esta arquitectura desempeña un papel fundamental al identificar las fuentes de datos, basándose en una revisión detallada de los requerimientos del negocio. A continuación, se determinan los requisitos específicos de los datos, es decir, identificar las tablas y columnas que serán utilizadas para dar respuesta a las aplicaciones BI. Además, para llegar a cabo una migración de datos exitosa, se recomienda analizar los diferentes sistemas transaccionales de origen, para comprender en detalle la estructura y sus componentes. Por último, es debe describir el modelo de datos destino, en el cual se migrarán los datos, por ejemplo, puede ser un DW, a un área de *staging* o un DM. La integración de datos implica una colaboración integral de los departamentos de la organización que están inmiscuidos sobre el proceso analizado, además de recopilar los datos de múltiples fuentes de datos y migrarlos a un repositorio centralizado. Esto se realiza con el propósito de establecer la terminología base para las columnas de las dimensiones y las métricas de las tablas de hechos (Figura 13). (Sherman, 2015).

Figura 13

Proceso de integración de datos



Nota. El proceso de integración de datos requiere un trabajo colaborativo entre las distintas áreas involucradas en el desarrollo del proyecto para fortalecer los procesos ETL, además cuentan con una visión holística del proceso de negocio. Adaptado de *Gathering data requirements (p. 281)*, de Rick Sherman, 2015, *Business Intelligence Guidebook.*

2.2.4.4.1. Extracción (*Extract*)

El primer paso del proceso ETL es denominado *extracción* de datos, el subproceso utiliza uno o diversos orígenes de datos para obtener datos históricos que permitan comprender la evolución de la organización, de igual manera analiza los datos generados durante las operaciones día, por consiguiente, los valores históricos y actuales forman parte del DW. La característica principal del proceso de extracción, es recopilar los datos generados en los diversos sistemas de información por medio de elementos que realizan la conexión a las fuentes de datos (*data sources*). Cada fuente de datos posee características propias que deben interpretarse antes de realizar el proceso de extracción de datos.

Además, a medida que la empresa o institución evoluciona, es común que los sistemas de información se encuentren dispersos en términos de lógica como físicos, lo cual genera incompatibilidades. Por ello, la fase de extracción de datos desempeña un papel fundamental al facilitar la unión y comunicación entre los diversos *data source*. Para realizar la integración se utilizan un elemento denominado mapa de dato lógico, donde se establece la relación entre los campos

de origen y los campos destino. El mapa de datos lógico contiene los metadatos necesarios para fortalecer la integridad de los datos y por lo tanto su calidad, es decir, describe cada uno de los esquemas de origen de datos con base a los requisitos establecidos por el cliente. Esta información permite la creación de perfiles de datos considerando la calidad e integridad. Al iniciar los procesos ETL, primeramente, se deben revisar, analizar y comprender los conceptos sobre los modelos dimensionales con el objetivo de identificar los campos destino, campos origen y sus características de una manera eficaz y eficiente. La tabla 5, muestra el mapa de datos lógico, el cual contiene la descripción de los orígenes de datos (*source*), incluyendo columnas y tipo de dato. Asimismo, se presentan los datos del repositorio destino y las reglas de transformación requeridas para la migración del formato original al destino (*Target*). El mapa de datos se presenta en una hoja de cálculo con sus elementos (Kimball y Caserta, 2004) y (Sabtu et al., 2017).

Tabla 5

Mapa de datos lógico de un Data Warehouse

Destino (Target)					Origen (Source)			Transformación
Tabla	Columna	Tipo de dato	Tipo de tabla	Base de datos	Tabla	Columna	Tipo de dato	
DIM_Producto	SG_Producto	Number	Dimensión	-	-	-	Number	Clave subrrogada, generada automáticamente
DIM_Producto	ID_Producto	Number	Dimensión	Ventas	Producto	PK_Producto	Number	Clave primaria de la tabla en la base de datos origen
DIM_Producto	Nombre	Varchar2(100)	Dimensión	Ventas	Producto	DescProducto	Varchar2(150)	initcap(DescProducto)
DIM_Producto	Categoria	Varchar2(50)	Dimensión	Ventas	Categoria	DescCategoria	Varchar2(100)	initcap(DescCategoria)
DIM_Producto	Sub_categoria	Varchar2(50)	Dimensión	Ventas	Categoria	DescSubCategoria	Varchar2(100)	initcap(DescSubCategoria)
DIM_Producto	Marca	Varchar2(50)	Dimensión	Ventas	Marca	DescMarca	Varchar2(100)	initcap(DescMarca)
FACT_Ventas	Dim_Cliente	Number	Hecho	DW_Ventas	Dim_Cliente	SG_Dim_Cliente	Number	-
FACT_Ventas	Dim_Tiempo	Number	Hecho	DW_Ventas	Dim_Tiempo	SG_Dim_Tiempo	Number	-
FACT_Ventas	Dim_Producto	Number	Hecho	DW_Ventas	Dim_Producto	SG_Dim_Producto	Number	-
FACT_Ventas	Importe_Total	Number	Hecho	Ventas	Detalle_Ventas	ImporteTotal	Number	PrecioUnitario * Cantidad
FACT_Ventas	Unidades_Vendidas	Number	Hecho	Ventas	Detalle_Ventas	Cantidad	Number	Sum(Cantidad)

Nota. La columna *destino* contiene la información requerida por el modelo dimensional, por ejemplo, las dimensiones, los hechos, las columnas y su tipo de dato; por otra parte, los orígenes de datos contienen la información de los registros que serán migrados al almacén de datos, por ejemplo, el nombre del esquema, tablas, columnas y el tipo de dato. Por último, la columna *transformación*, establece las reglas para realizar las migraciones, por ejemplo, una consulta de base de datos, cálculos o asignaciones directas del valor. Adaptado de *The logical data map (p. 60)*, de Kimball y Caserta, 2004), *The Data Warehouse ETL Toolkit.*

Para iniciar el proceso de extracción es importante considerar los siguientes puntos de Kimball y Caserta (2004):

- Indexar las columnas: Para reducir la afectación de rendimiento en los sistemas de información de producción, revisar que las columnas utilizadas en la cláusula WHERE contengan los índices necesarios.
- Consultar solo las columnas necesarias: Solo indicar aquellas columnas que son necesarias para la elaboración de reportes y estadísticas. E
- Uso de la cláusula DISTINCT: Al utilizar esta cláusula en una consulta el rendimiento disminuye notoriamente, utilizar solo en caso obligado.
- Los operadores SET: Al igual que el DISTINCT, provoca lentitud al ejecutar las consultas (UNION, MINUS, INTERSECT, IN, OR, NOT, <>). Por ejemplo, el operador UNION ALL es recomendable usarlo en lugar de UNION, solo revisar los valores duplicados que podrían generarse.

2.2.4.4.2. Transformación (*Transform*)

Los datos históricos y actuales, requieren ser procesados para su posterior carga al DW o *staging area*. Entre las acciones principales de este sub-proceso denominado transformación se deben considerar la medida de los atributos, agregando límite a los valores para estandarizar los datos, así mismo, homologar los nombres de las columnas y tablas; una acción indispensable es elegir el origen de datos con alto nivel de disponibilidad, integridad y confianza, para garantizar en mayor grado la calidad de los datos; por último, seleccionar únicamente aquella información que es relevante para el negocio (Contreras, 2018).

La limpieza e integración de datos son los elementos centrales en los procesos ETL, denominada etapa de transformación, esta fase se encarga de manipular los datos considerados de baja calidad y poco confiables para garantizar su ingreso al repositorio de datos. En la transformación se realizan operaciones de filtrado de datos, estandarización, depuración y agrupación de los datos, por ejemplo, se definen reglas para transformar valores nulos o descartarlos al momento

de realizar la migración del origen al destino, algunos otros ejemplos como calcular valores con alguna constante u operaciones entre columnas, unión de datos de varias fuentes, división de columnas, asignación de constantes o valores por defecto. Los datos sin inconsistencias son migrados de forma transparente, mientras que los datos erróneos, deben pasar por una serie de políticas o reglas de migración antes de proceder con su migración. (Kimball y Caserta, 2004).

En esta fase de transformación, los datos migrados pueden contener un gran número de errores o inconsistencias que pueden resumirse en las siguientes:

- Longitud inconsistente de campos: Dependiendo del origen de los datos, la misma columna puede contener longitudes variantes, es por ello, que deben homologarse el tamaño y definir uno en común que contemple la mayoría de los casos.
- Descripción inconsistente de campos: Algunas columnas pueden tener el mismo nombre y almacenar datos diferentes en cada fuente de datos, por ejemplo, el campo dirección, en un origen de datos puede almacenar calle y número, y en otro origen de datos almacenar estado, municipio, localidad y CP.
- Distintas codificaciones para el mismo término: Es común encontrar este tipo de columnas, debido a poco uso de catálogos institucionales que garanticen un valor único, por ejemplo, estado de nacimiento, institución de origen, etc…
- Valores nulos: En caso de encontrar valores nulos y que son necesarios para el proceso de migración, se debe tomar la decisión del valor por default a utilizar.

Por otra parte, el objetivo de la limpieza y estandarización de los datos es reducir los errores en los datos y mejorar la calidad, por ello las transformaciones más utilizadas en los procesos ETL pueden resumirse en, de acuerdo a (Trujillo et al., 2009):

- Generador de claves: Para garantizar una clave única en la tabla de dimensión, es necesario crear claves a partir de claves compuestas de los distintos orígenes de datos.
- Conversión de datos: En este caso, pueden existir columnas que significan lo mismo con valores distintos, por ejemplo, el sexo, en una fuente de datos pueden ser almacenarlo como masculino y femenino, en otra fuente de datos como M o F; en este caso se debe homologar cada una de las columnas para una única presentación de los datos.
- Conversión: Al migrar los datos, es esencial estandarizar los campos de unidades de medida, fechas, precios, entre otros, de acuerdo a las reglas del negocio y el objetivo final de los datos.
- Filtrado: Estas operaciones son utilizadas para migrar únicamente aquellos valores que cumplan con las características y calidad requerida para su uso en el DW.
- Unión y combinación: Permite combinar filas de distintas fuentes de datos en una única fila, de acuerdo a las reglas del negocio.
- Agregación: Toma un conjunto de filas de datos y genera una única fila utilizando una función de agregación.
- Valores atómicos: Consiste en crear varias columnas a partir de una. Por ejemplo, el campo “nombre_completo”, es necesario dividirlo en tres columnas para cumplir con la normalización, dividiendo la comuna en: nombre_persona, apellido_paterno y apellido_materno.

2.2.4.4.3. Carga (*Load*)

Esta fase está ligada directamente con las reglas del negocio en donde se define el tipo de carga o migración a realizar, en algunos casos será necesario sobrescribir los datos en cada una de las tablas; mientras que, en otros, bastará con migrar únicamente los datos nuevos o que han sufrido algún cambio en los sistemas origen. La tercera fase del proceso ETL recopila los datos procesados de la etapa de transformación para cargarlos en el repositorio destino, es necesario considerar que

al igual que pueden existir varios repositorios origen también pueden existir varios repositorios destinos diferentes. En este sentido, la calidad de los datos se garantiza al aplicar las restricciones que el repositorio destino solicite, por ejemplo, los campos obligatorios, valores estandarizados, disparadores, integridad referencial, etc.

Además, el proceso de carga puede desarrollarse principalmente de dos formas distintas; la primer de ella consiste en la acumulación simple, que incluye un compendio de las transacciones realizadas durante un periodo de tiempo determinado, mismas que serán migradas como una única transacción al almacén de datos destino. Este enfoque representa la manera más simple para cargar los datos, sin embargo, al tratarse de una gran cantidad de información puede verse afectada tras una interrupción en la carga a consecuencia de un corte de luz, fallo del disco, etc… Por otro lado, está la carga de tipo escalonada, la cual agrupa la información por fechas, jerarquías o niveles de granularidad. De esta manera, en caso de existir algún fallo, es sencillo localizar los datos inconsistentes y volver a ejecutar el proceso (Lépez, 2019).

2.2.5. *Data Warehouse* y *Data Marts*

Como se ha mencionado con anterioridad, la mayoría de las organizaciones y entidades, tanto públicas y como privadas utilizan los datos resguardados dentro de los sistemas de información de tipo transaccional con el objetivo de elaborar reportes y estadísticas, sin embargo, debido a que dichas organizaciones se encuentran en constante cambio, la calidad y el volumen de datos contenidos en las bases de datos provocan una ineficiente obtención de reportes fiables y oportunos. El objetivo de este tipo de sistemas, consiste en gestionar eficaz y eficientemente las operaciones diarias, por ejemplo, pagos de nómina, gestión de inventarios, evaluación docente, contabilidad, control de aspirantes y matrícula escolar, etc., es por ello que, es impráctico la generación de reportes sobre estos sistemas.

La diferencia entre las dos formas de almacenamiento radica en la disposición de los datos y su estructura de almacenamiento (Tabla 6): por una parte, la cuestión operativa, también llamada sistemas transaccionales u OLTP (*ON-Line*

Transaction Processing) de la empresa, los datos están orientados al manejo de las transacciones diarias y son caracterizados por el gran número de inserciones, actualización y eliminación de los registros. El principal enfoque de los sistemas OLTP se centra en el procesamiento de tareas en tiempo real, además de mantener los datos íntegros provenientes de múltiples accesos y ambientes; por otra parte, en el contexto de los datos para la elaboración de reportes, estadísticas y el proceso de toma de decisiones, comúnmente denominado sistemas de procesamiento analítico (OLAP - *On-line Analytical Processing*) los datos se orientan al tema, es decir, conceptos que tengan relevancia en la organización, los sistemas OLAP se caracterizan por el volumen reducido de transacciones, las consultas son complejas e involucra agregaciones, datos históricos, almacenamiento en esquemas multidimensionales, usualmente en esquemas estrella, (Morales et al., 2016).

Tabla 6

Los sistemas OLTP vs OLAP

Propiedad	Base de datos OLTP	Base de datos OLAP
Fuente de los datos	Sistemas operacionales.	Datos consolidados; los datos provienen de diferentes fuentes OLTP, hojas de cálculo, etc.
Propósito de los datos	Para controlar y ejecutar las tareas fundamentales del negocio.	Para ayudar con la planificación, resolución de problemas y soporte a la decisión.
Tipo de datos	Procesos actuales del negocio.	Vistas multidimensional de diversas actividades del negocio.
Inserts y Updates	Inserciones y actualizaciones cortos y rápidos iniciados por los usuarios finales.	Trabajos por lotes de larga duración, actualización periódica de los datos.
Consultas	Consultas relativamente estandarizadas y simples que regresan pocos registros.	Consultas complejas con agregaciones.

Velocidad de procesamiento	Normalmente muy rápido.	Depende de la cantidad de datos involucrados, consultas complejas pueden tomar muchas horas. La velocidad de procesamiento se mejora mediante la creación de índices.
Espacio requerido	Puede ser relativamente pequeño si se archivan los datos históricos.	Grande, debido a la existencia de estructuras de agregación y datos históricos.
Diseño de base de datos	Altamente normalizado con muchas tablas	Normalmente des-normalizadas con esquemas en estrella.
Backup y recuperación	Backup dependiendo de las necesidades del negocio, pueden ser diariamente, semanal, etc.	Solo se recuperan los datos periódicamente de los sistemas OLTP. Los backups no son muy comunes.

Nota. Los sistemas OLTP brindan los elementos necesarios para la construcción de sistemas OLAP. Los dos tipos de sistemas tienen como objetivo principal la optimización constante de las operaciones de la organización.

El personal responsable del nivel estratégico, como los directivos o personal responsable del análisis de la información y de la toma de decisiones, requieren estructuras de base de datos especializadas con el objetivo de crear reportes y estadísticas confiables que no afecten el rendimiento de los sistemas transaccionales. En este sentido, estas estructuras centralizan los datos con la mayor calidad, estandarizados y homologados, mismos que pasaron por una serie de procesos antes de ser migrados. Por ejemplo, datos sobre clientes, estudiantes, proveedores, empleados, productos, pagos, entre muchos otros. Aproximando de esta manera, el concepto de un *Data Warehouse* (DW) (Pozo, 2004).

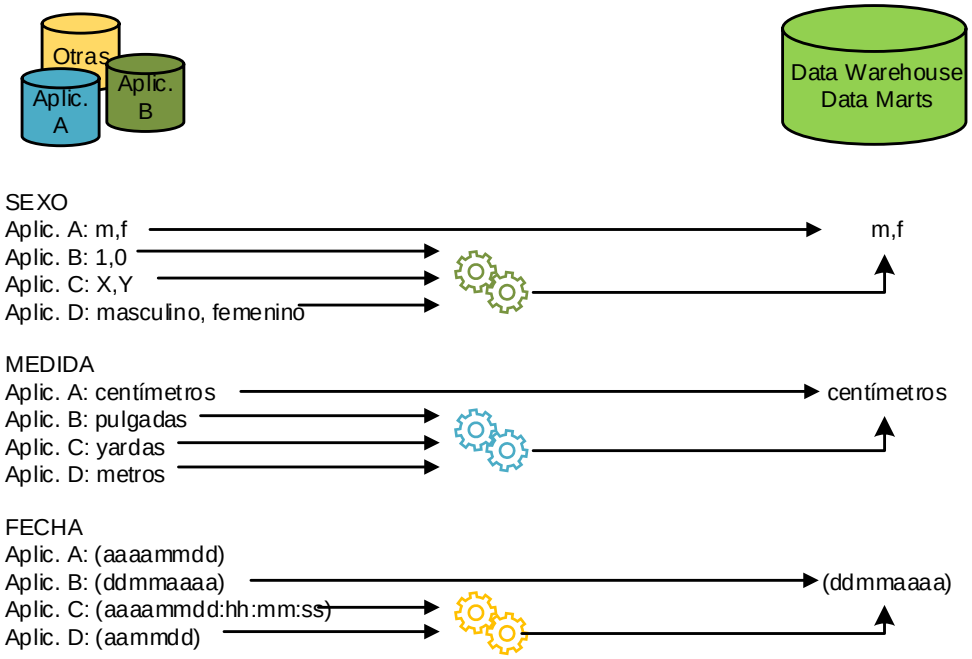
Los términos sobre un DW fueron originalmente propuestos por Bill Inmon (2005), quién acuño el termino como *almacén de datos*. De acuerdo a Bill Inmon los DW son bases de datos que posee ciertas características como, datos integrados, orientados a temáticas, variante en el tiempo, no volátil, consistente y construida a

partir de múltiples fuentes de datos. Además, almacena los datos para para facilitar una acceso eficiente, de igual manera, fortalece la emisión de reportes y la toma de decisiones a nivel empresarial (Inmon, 2005); en términos de la definición se obtienen las siguientes características como lo son, orientado a temas, integrado, variante en el tiempo y no volátil (Zangana, 2018):

- Integrado: El DW almacena datos provenientes de diversos sistemas de información, tanto de fuentes internas como externas de la organización. Tendiendo como objetivo asegurar un nivel de calidad, disponibilidad e integridad de los datos. En este sentido, para llenar el DW se crean reglas y políticas que facilitan el proceso de migración, por ejemplo, se estandarizan nombres y terminología de la empresa, homologación de valores dentro de las columnas y se unifican descripciones o valores de los registros. Es importante considerar que los datos almacenados en los diversos orígenes de datos pueden usar diferentes terminologías o términos para representar el mismo valor (Figura 14). Por ejemplo, la columna “género” puede tener representaciones como {F, M}, {1, 0}, {X, Y}, {Masculino, Femenino}, entre otros. Por ello, se deben homologar los datos para garantizar el proceso de migración realizado entre las fuentes de datos y el DW. Asimismo, por distintas circunstancias de las organizaciones, las unidades de medida pueden contar con diferencias, por ejemplo, la ubicación geográfica de la empresa X almacena los grados en escala *celsius* mientras que la empresa Y en *fahrenheit*, lo cual requiere la aplicación de transformaciones en los datos para su estandarización. Otro aspecto importante en la característica de *integración*, se refiere a las reglas del negocio, en las cuales se deben aplicar nomenclaturas a los nombres de procesos y objetos para comunicarlas de forma adecuada y sin ambigüedades. Al definir un DW como integrado, se consideran los dos aspectos mencionados anteriormente, el primero la parte técnica de la arquitectura de los datos y la parte administrativa, como son las reglas del negocio o la arquitectura de la información (Inmon, 2005) y (Sherman, 2015).

Figura 14

Integración de los datos en los DW



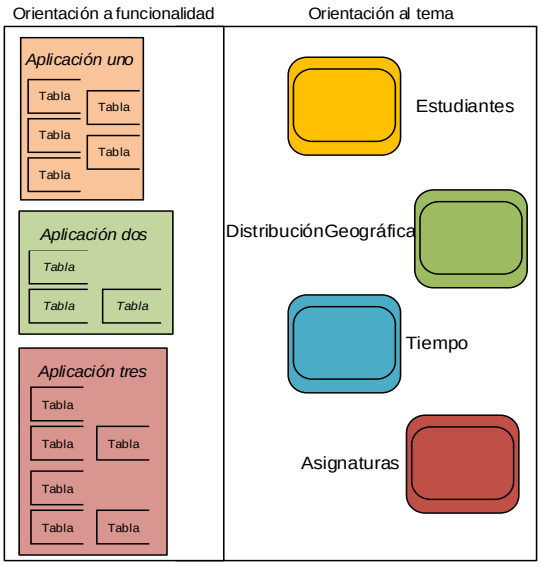
Nota. La característica de integración del DW asegura que, sin importar la fuente de datos, los procesos de transformación extraerán correctamente el dato solicitado, es decir, sin importar el diseño de la aplicación el resultado será el mismo debido a que la información almacenada en el DW está contenida en un modelo globalmente aceptable y singular. Adaptado de *To properly move data from the existing systems environment to the data warehouse environment, it must be integrated (p.73)*, de William H. Inmon, 2005, *Building the Data Warehouse.*

- Orientado a temáticas: El DW organiza los datos de manera específica, de acuerdo a las áreas o procesos determinados de la organización, lo que facilita la comprensión y el acceso a los usuarios quienes están familiarizados con los términos y procedimientos del negocio, al ser los responsables de dicho proceso (Figura 15). En consecuencia, el DW se estructura con base a objetos abstractos, tales como estudiantes, proveedores, clientes, pagos, clases y docentes, por ejemplo, todos los datos generales de los alumnos se centralizan

y consolidan en una única tabla del DW, en otra se almacena las carreras y calificaciones, es decir, el DW hace referencia a todo el proceso de estudiantes. Consolidar los datos en grupos brinda un rendimiento óptimo al ejecutar cada una de las consultas solicitadas por el usuario, esto debido a la centralización de los datos en una misma ubicación, evitando la generación de consultas complejas. Por otro lado, los sistemas transaccionales, hacen uso de bases de datos relacionales cuyo propósito es la funcionalidad sobre las actividades diarias de toda la organización, por ejemplo, los procesos de inscripciones de los estudiantes, el control de préstamos de libros de una biblioteca, la gestión de los inventarios, pago de nóminas, etc. (Inmon, 2005).

Figura 15

Diseño orientado a la funcionalidad y orientación al tema



Nota. Los ambientes DW excluyen los datos que no serán utilizados o no son relevantes para la elaboración de reportes y estadísticas, por otro parte, existen sistemas orientados a los procesos funcionales de la empresa o aplicativos, los cuales contiene todos los datos necesarios para satisfacer los requerimientos funcionales y de proceso. Adaptado de *An example of a subject orientation of data (p.30)*, de William H. Inmon, 2005, *Building the Data Warehouse.*

- Variante en el tiempo: Una característica esencial del DW es su capacidad para almacenar información histórica, lo que permite el análisis de tendencias y evoluciones a lo largo del tiempo. Es por ello, que el DW muestra la historia de la organización a través de estructuras de datos dimensionales que posibilitan el almacenamiento de los diversos valores a lo largo del tiempo. En consecuencia, los datos serán consultados desde el primer registro almacenado en el DW hasta la actualización más reciente. A diferencias de los sistemas transaccionales, donde los datos almacenados corresponden a un periodo de tiempo relativamente corto o estado actual de las actividades del negocio, mientas que el DW almacena los registros históricos de cinco a diez años o más. (Inmon, 2005).
- No volátil: Los datos del DW son almacenados para ser consultados y, pocas veces, podrán ser modificados o eliminados. Es decir, una vez migrada la información, para mantener consistencia en los datos no debería haber eliminaciones ni actualizaciones de datos previamente cargados. Por lo cual son valores permanentes y no volátiles; únicamente se permite la inserción de registros nuevos sin alterar los existentes.

Por otra parte, Ralph Kimball y Caserta (2004) es uno de los pioneros en los conceptos de DW. Considera a los *almacenes de datos* como un lugar donde se centralizan aquellos datos que representan o tienen algún valor para la organización. Para alimentar el DW se utilizan los procesos ETL, con el propósito fundamental de extraer, limpiar, transformar y cargar los datos provenientes de un origen o fuente de datos a una estructura dimensional. Dicho modelo permitirá realizar las consultas con un rendimiento mayor a los sistemas transaccionales; además de facilitar la revisión minuciosa de los datos y su impacto en la toma de decisiones. Asimismo, el almacenamiento exitoso de los datos dentro del DW depende de tres puntos fundamentales: centrarse en el negocio, estructura dimensional de los datos y desarrollo iterativo (Kimball et al., 2008). El DW contiene datos históricos que derivan de datos transaccionales de múltiples fuentes, es por

ello que, no se deben mezclar las aplicaciones de tipo BI con las actividades diarias de la parte operativa, lo que permite consolidar y estandarizar datos de varias fuentes.

Una alternativa más en las definiciones del DW, refiere que el DW es un sistema donde periódicamente se identifican y centralizan los datos provenientes de diversas fuentes u orígenes de datos a un repositorio único de datos dimensional. Este repositorio almacena datos históricos que podrán ser consultados posteriormente en actividades analíticas (Rainardi, 2008).

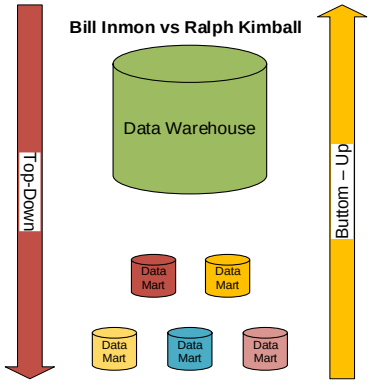
Los principales autores del término DW son *Inmon* y *Kimball*, cada uno defiende una postura o perspectiva distinta quienes sobre la forma de desarrollar un proyecto de esta naturaleza. Por un lado, el autor *Bill Inmon* utiliza la metodología denominada descendente o top-*down*, en la cual se consideran todas las áreas responsables de la generación de datos de la organización para el diseño del almacén de datos. La metodología *top-down*, define el DW completo de la organización para después iniciar con el diseño de los DM particulares para cada unidad. Por otra parte, Kimball establece una metodología llamada ascendente o *bottom-up* para definir la estructura de un almacén de datos, en el cual primero se crean los DM de las unidades de la organización y al combinar éstos modelos dimensionales se da origen al DW. El modelo dimensional es desnormalizado y se representa por un esquema en estrella (Figura 16).

La implementación de un proyecto de tipo DW, puede realizarse a partir de las dos metodologías anteriores, es por ello que, la elección entre estas dos metodologías depende totalmente del tiempo que se tenga asignado para el proyecto y del presupuesto otorgado. Por un lado, la metodología descendente implica una inversión considerable en términos de tiempo y recursos económicos al tratar de visualizar todos los procesos de negocio de la organización. Por otra parte, dicha metodología ascendente ofrece resultados en un periodo de tiempo más corto al centrarse en un proceso de negocio específico, facilitando el análisis del mismo (Foster y Godbole, 2016).

Un *data mart* (DM) se constituye como un conjuntos de datos históricos que provienen de uno o múltiples sistemas de base de datos. Tanto del DM como el DW proporcionan una perspectiva integral del estado de la organización y facilitan el análisis de datos a través de los modelos dimensionales. Además, organizan los datos considerando un conjunto de reglas establecidas durante la definición de los procesos ETL. El DM y el DW, son repositorios de datos utilizados en la elaboración de consultas, además, se realizan operaciones de carga masiva y, solo en caso obligado, se realizan eliminaciones o actualizaciones de datos ya migrados. Sin embargo, la diferencia principal entre los enfoques radica en la amplitud analizada: por un lado, el DW debe analizar cada uno de los procesos que forman para te de la organización, mientras el DM se enfoca en un proceso especifico de misma. En este sentido, poseen las mismas características y conceptos, por lo tanto, las definiciones utilizadas en este documento aplicarán para ambos términos. (Kimball y Ross, 2013).

Figura 16

Metodología de Bill Inmon y Ralph Kimball



Nota. La metodología *Top-Down* fue propuesta por Inmon, y establece la creación de un DW corporativo para después diseñar los DM, mientras que Kimball establece la creación de DM y, en consecuencia, el diseño del DW utilizando todos los DM mediante la metodología *Buttom-Up*. Adaptado de *Simplified illustration of the independent data mart "architecture." (p.27)*, de William H. Inmon, 2013, The *Data Warehouse Toolkit.*

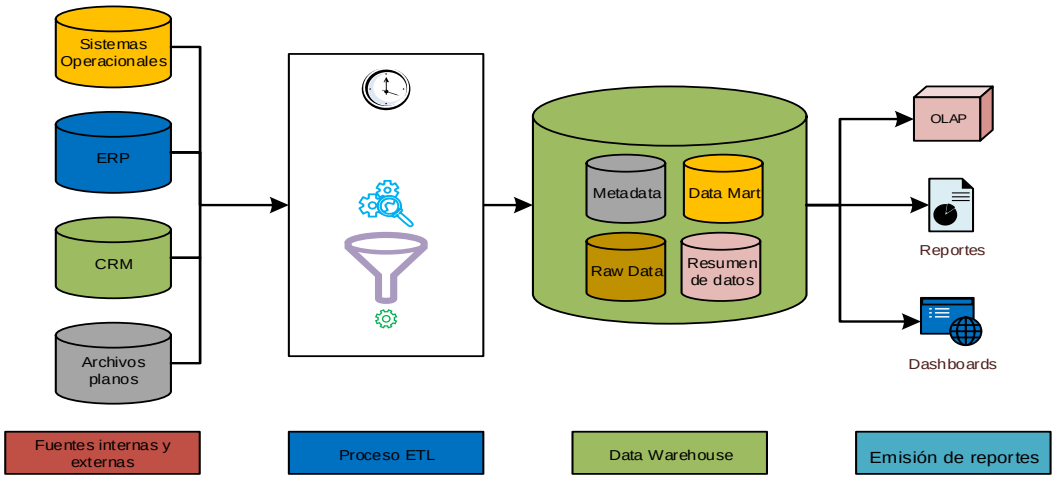
Las dos perspectivas implican almacenar los datos en una estructura única o repositorio centralizado, para su posterior uso y explotación con apoyo de herramientas especializadas para generar reportes, informes y estadísticas. Además, ambas metodologías emplean los conceptos de los procesos ETL. No obstante, la principal diferencia la define el alcance del proyecto, por esta razón es difícil determinar cuál de los dos enfoques debería ser utilizado en un proyecto BI, debido a que ambas metodologías operan en contextos distintos. Por lo tanto, en la elaboración de esta tesis y considerando que la Universidad es una institución de tamaño medio en cuanto a la cantidad de datos trabajados, se optó por seguir la metodología de *Kimball* con el fin de obtener resultados en un periodo más corto y con costos reducidos.

2.2.5.1. Arquitectura del *Data Warehouse*

Un DW o DM están compuestos por una arquitectura que se caracteriza por una serie de elementos que operan de manera conjunta, entre ellos se encuentra la posibilidad de crear estructuras dimensionales para el almacenamiento de los datos, además permite definir los flujos de comunicación entre los orígenes de datos y los repositorios destino. Asimismo, dentro de la arquitectura del DW se definen los procesos de migración y homologación aplicables a los datos para contar con una mayor calidad e integridad de la información. Por último, se establece una capa de presentación o visualización de los datos utilizadas por los directivos o personal responsable del análisis de los reportes. Por otra parte, para diseñar el modelo de base de datos, existen diversas implementaciones que se pueden llevar a cabo con base a los requerimientos de la organización y de las reglas del negocio. Por ejemplo, el modelo propuesto por *Kimball* (2013), establece la definición de múltiples *data mart*, que pueden ser poblados por una o varias fuentes de datos (Figura 17). Como se puede observar, el uso de alguna arquitectura en la construcción de un DW, apoya a la calidad de los datos al tenerlos disponibles en un repositorio, dichos datos fueron recopilados, transformados, almacenados y entregados al personal encargado de los procesos de emisión de reportes y estadísticas.(Sherman, 2015).

Figura 17

La arquitectura de un Data Warehouse de Kimball



Nota. La arquitectura de *Kimball* está conformado por cuatro componentes utilizados en el entorno de DW / BI: orígenes de datos de sistemas transaccionales; proceso para extraer, transformar y cargar los datos; modelo dimensional; por último, las herramientas de explotación de datos. Adaptado de *Core elements of the Kimball DW/BI architecture (p.19)*, de Ralph Kimball, 2013, The *Data Warehouse Toolkit.*

A continuación, se describen los componentes de la arquitectura de un DW de acuerdo a *Kimball*:

- En la sección 2.2.4 Arquitectura de los datos, se abordó el tema sobre las fuentes de datos utilizados en las soluciones de inteligencia de negocios, Dicho componente es responsable del flujo de captura de registros transaccionales del negocio.
- En la sección 2.2.5 *Data Integration*: Se abordó a detalle el proceso para extraer, transformar y cargar datos de diversos orígenes a un repositorio destino.
- Modelo de datos: En proyectos o soluciones de inteligencia de negocios se utilizan modelos dimensionales, siendo el de mayor uso el modelo en estrella, dicho modelo brinda la posibilidad de organizar los datos en estructuras de datos

simples, así como almacenar y facilitar su disponibilidad para poder acceder a ellos por medio de aplicaciones especializadas o por el personal técnico. En la sección 2.2.7 se profundizará en detalle el Modelo dimensional.

- Por último, las herramientas tecnológicas de BI: Utilizan los datos de las estructuras de base de datos con el propósito de elaborar reportes, presentando los resultados de manera simple. Por ejemplo, en las aplicaciones BI es común realizar consultas *ad hoc* o aplicar técnicas avanzada sobre minería de datos. En la sección 2.2.1 *Business Intelligence* (Inteligencia de negocios), se presenta a detalle la información.

2.2.6. Esquemas dimensionales

El modelo dimensional es una técnica utilizada para representar los elementos de las soluciones de BI, en específico se elaboran estructuras de datos para el DW. Los modelos dimensionales representan de forma gráfica la dimensión o tabla de datos de un área específica de la empresa, por lo cual, se facilita el acceso y rendimiento sobre los datos, tanto al personal técnico como a las aplicaciones. Para realizar el modelo dimensional se consideran conceptos o elementos clave, por ejemplo, las denominadas tablas de hechos o tablas de medidas, las dimensiones o contextos sobre el cual se analizarán las medidas y, los atributos de cada tabla. En el caso de las tablas de hechos se almacenan las transacciones que normalmente son sumativas, es decir valores numéricos. Por otro lado, las dimensiones, establecen el contexto por medio de la incorporación de jerarquías, nivel de granularidad y los atributos que proporcionan información detallada sobre quien, realizada la acción, la ubicación donde se realiza, el porqué de la acción y en específico qué acción. La representación de los modelos dimensionales y sus elementos se realiza por medio de distintos esquemas, por ejemplo, el modelo en estrella, el modelo en copo de nieve y el modelo en constelación (Sherman, 2015).

La construcción de un modelo dimensional depende de los requerimientos de la organización, tanto técnicos como administrativos, en el caso del proyecto desarrollado en la Universidad se optó por el modelado dimensional en estrella (*Star*

Squema). El modelo en estrella únicamente define una cantidad de tablas reducida, a comparación de los modelos relacionales, debido a que des-normaliza los datos en tablas de hechos y dimensiones, lo cual aumenta el rendimiento al momento de realizar consultas que traen miles de registros. (Japal, 2021).

El siguiente esquema es conocido como el esquema de copo de nieve (*snowflake schema*), y se origina con base a la normalización de la información almacenada en las tablas de dimensiones. Esta normalización trae como consecuencia, en primer punto, la creación de nuevas dimensiones conectadas entre sí al reducir la redundancia de los datos, en segundo lugar, afecta al rendimiento de las consultas ejecutadas sobre las tablas que almacenan grandes cantidades de registros. Este esquema, del mismo modo que los modelos en estrella son representados por una tabla central (*fact table*) y varias dimensiones que rodean a la tabla central. La tabla de hechos almacena los atributos de tipo medidas o métricas que la organización requiere analizar, mientas que las dimensiones almacenan los contextos sobre los cuales se desean analizar las medidas.

La última clasificación de los esquemas dimensionales, se trata del modelo en constelación (*Constellation Schema*. Se trata de una versión del modelo en estrella extendido, es decir, este modelo los atributos de las dimensiones pasan a formar parte de una nueva entidad o dimensión, mismos que serán reutilizados por otros almacenes de datos de la organización. Esta práctica ofrece la posibilidad de reducir el espacio de almacenamiento físico, así como la disminución de la redundancia de los objetos y se definen los mecanismos de migración de datos para una sola dimensión que será utilizada por varios proyectos. (Iqbal et al., 2020).

2.2.6.1. Tabla de hechos

Los hechos son representados con tipos de datos numéricos y se definen como las medidas realizadas a alguna actividad, registro o evento de un área o proceso de negocio de la empresa. Algunos ejemplos de hechos son, el total de ventas, total de gastos, las calificaciones de un estudiante, promedios, impuestos, evaluaciones, entre otros. Asimismo, los hechos se utilizan para realizar conteos, sumas,

porcentajes, promedios y, en general, cualquier operación matemática o estadística. En este sentido las medidas o métricas son almacenados en las tablas de hechos o *fact table*, dichas medidas tienen el objetivo fortalecer, por un lado, la emisión de reportes y estadísticas de la organización, y, en consecuencia, el proceso para tomar decisiones informadas. Es por ello, que los responsables de dichas actividades deben definir cuáles serán los hechos que se deseen almacenar en el DW. Las tablas de hechos están compuestas por dos tipos de atributos, el primero son las llaves foráneas y el segundo las medidas o métricas (*measures*); además, los hechos pueden ser analizados en contextos específicos por medio de las tablas de dimensiones. (Sherman, 2015).

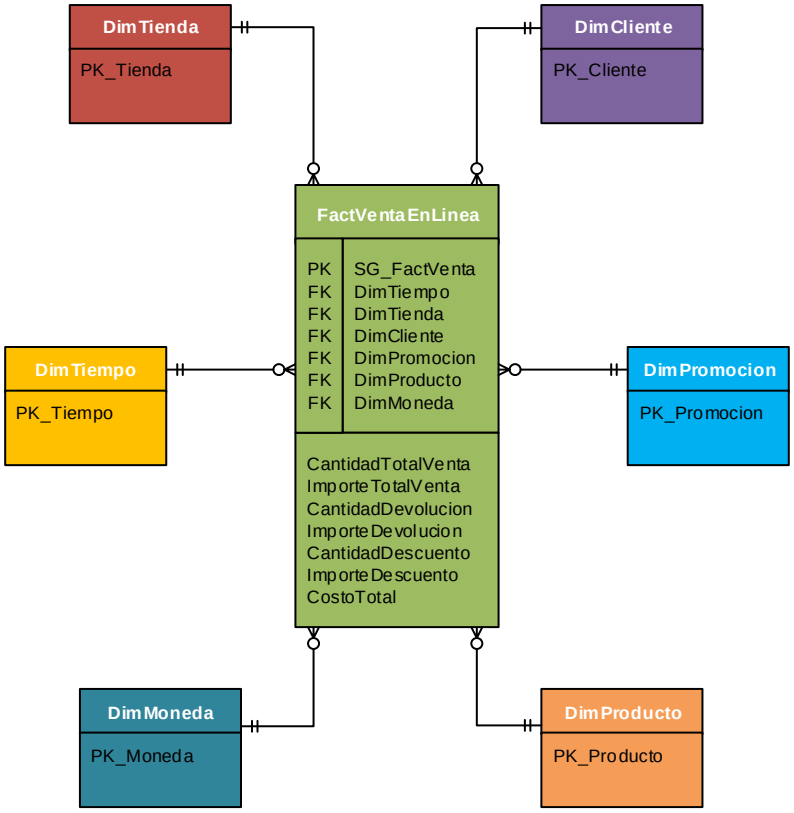
Como parte de las columnas que conforman la tabla de hechos se encuentran las claves o llaves foráneas que permiten establecer la relación con las llaves primarias de las dimensiones correspondientes, el tipo de relación definida con las tablas de hechos y las tablas de dimensiones es considerado de uno a muchos y mantienen las mismas características que los modelos relacionales (Figura 18). La combinación de las llaves foráneas, permite establecer de forma única la clave principal o primaria compuesta de la tabla de hechos, sin embargo, en los casos donde esta combinación no sea suficiente para identificar de forma única la fila almacenada, se debe definir una clave subrogada (*Surrogate key*) de tipo entero para realizar las consultas de manera eficiente. El manejo de las llaves foranes y llaves primarias, son consideradas mejores prácticas al momento de diseñar un DW, de igual manera, se sugiere indexar las columnas que se considere necesario y evitar el trabajo con valores nulos. Los puntos anteriores retoman importancia al evidenciar que el 90% de los datos almacenados en un proyecto de DW corresponden a las tablas de hechos. (Sherman, 2015).

Como se mencionó en párrafos anteriores el componente número dos dentro de las tablas de hechos, corresponde a los hechos en si o medidas utilizadas para realizan los cálculos necesarios del negocio, por ejemplo, el promedio de un conjunto de calificaciones, la suma de ISR en una empresa, los costos de ventas,

los costos de compras a proveedores, entre otros. Las medidas, son de vital importancia para el DW, debido a que de ellas depende el nivel de granularidad o detalle utilizado para almacenar los datos en el epositorio, dando origen al nivel de granularidad del DW, dicho nivel tiene como límite los valores almacenados en los orígenes de datos, por ello, como una buena práctica, se recomienda almacenar el máximo nivel de detalle capturado en el proceso de negocio. (Kimball y Ross, 2013).

Figura 18

Tabla de hechos en el modelo multidimensional



Nota. En la figura se muestra un hecho (evento) denominado venta por internet, las medidas son cantidad total de la venta, importe total de la venta, cantidad de devoluciones, importe de devoluciones, cantidad de descuento, importe de descuento y costo total; el nivel de granularidad se centra en cada producto adquirido por un cliente en una tienda determinada y fecha (llaves foráneas).

2.2.6.2. Tablas de dimensiones

Otro componente del modelo dimensional, son las tablas de dimensiones, las cuales, a diferencia de las tablas de hechos compuestas por llaves y medidas con valores numéricos, están formadas por atributos descriptivos que proporcionan el mecanismo o la manera de analizar los datos de acuerdo al ámbito o contexto de la empresa, añadiendo diferentes perspectivas. Lo anterior se logra al restringir, filtrar y etiquetar los resultados en las consultas. Las dimensiones establecen el "quién", "qué", "dónde" y "porqué" del modelo dimensional, además agrupan los atributos en categorías o temas que definen la organización lógica de los datos, dichos atributos deben ser descriptivos para que los directivos o personal de negocios puedan comprenderlo con facilidad y evitar la presencia de los valores nulos.

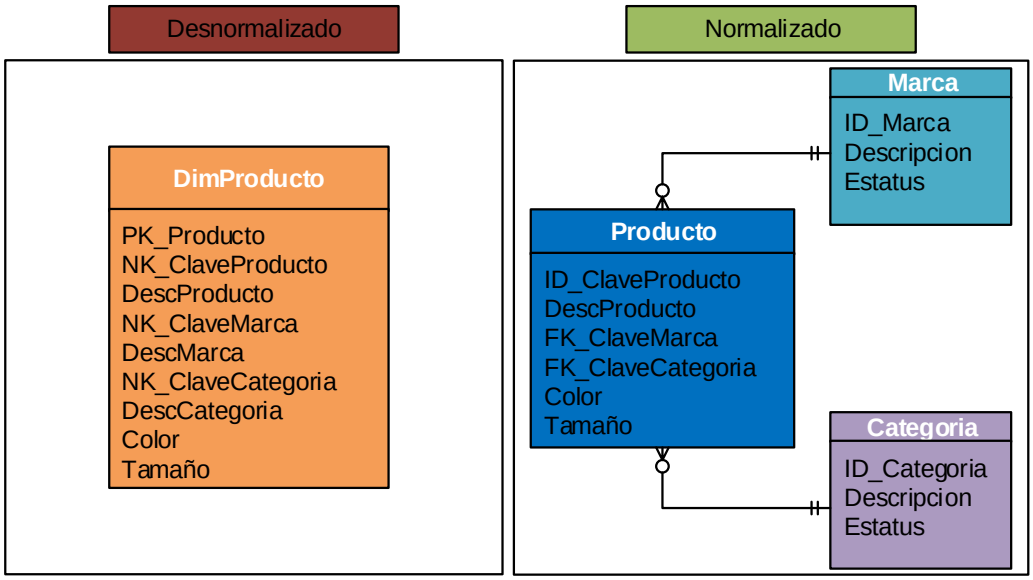
Las tablas de dimensiones, en la mayoría de los casos, crean relaciones jerárquicas que permiten la agrupación de datos en una misma tabla. Por ejemplo, si los productos se agrupan en marcas y categorías, representaría una relación jerárquica dentro de una misma tabla de dimensión producto, que incluirá, los atributos descriptivos del producto, la marca y categoría. En el ejemplo anterior, se observa que los registros de marca y categoría causaran redundancia en los datos, en los modelos multidimensionales es aplicable la des-normalización para agilizar las consultas realizadas sobre el DW (Figura 19). Otros ejemplos de jerarquías, pueden ser las agrupaciones geográficas, el tiempo (los años se componen en semestres, meses, semanas, etc.). El objetivo principal de las jerarquías es profundizar del nivel inferior al superior dentro de la dimensión. Por ejemplo, para visualizar las ventas acumuladas en un año o profundizar en el semestre o incluso aun nivel más detallado como la semana (Kimball et al., 2008).

Las tablas de dimensiones, además de contener valores descriptivos también hacen uso de llaves primarias o claves únicas, normalmente de tipo de dato entero, de la misma manera que las *fact table*. No obstante, en las tablas de dimensiones se utilizan claves subrogadas (*surrogate key*), cuyo uso se traduce en un mejor rendimiento debido su tamaño e índice. Además de la clave subrogada, como una

práctica recomendada, se sugiere agregar una clave natural (*natural key*) o clave del negocio (*business key*) en las dimensiones, para hacer referencia a la llave primaria del origen de los datos. (Sherman, 2015)

Figura 19

Des-normalización de tablas para crear las dimensiones



Nota. La DimProducto fue creada a partir de tres tablas de un modelo relacional, la tabla producto contiene una relación con marca y categoría, al pasar a un modelo dimensional es necesario des-normalizar los datos para eficientar las consultar realizadas sobre el DW.

2.2.6.3. Proceso de diseño del modelo dimensional

Los modelos dimensionales están conformados por una serie de tablas des-normalizadas, por lo tanto, son pocas tablas relacionadas con el proceso del negocio analizado. Las tablas del modelo dimensional categorizan los datos con el objetivo de agilizar el acceso al personal responsable de la emisión de reportes de manera eficaz. En este sentido, para poder construir el modelo se deben seguir una serie de pasos de acuerdo a (Kimball y Ross, 2013)

1. Seleccionar el proceso de negocio (*Business Process*): El primer paso corresponde a la identificación del proceso de negocios donde se implementará el DW. Dicho proceso es definido por las actividades y tareas desarrolladas en conjunto y cuyo propósito es cumplir los objetivos estratégicos o tácticos de las unidades de negocios. Por ejemplo, la gestión de pedidos, la facturación electrónica, el registro de pagos, la atención a llamadas de servicio, control de registros de estudiantes, los procedimientos médicos, el pago de nóminas, el timbrado de nóminas, entre otros. Cada uno de los procesos de negocio generan un conjunto de datos que serán almacenados en bases de datos transaccionales (OLTP). En este contexto, para iniciar la construcción del modelo dimensional, es necesario proporcionar una descripción detallada del proceso de negocio y las reglas aplicables en las actividades.
2. Definir el nivel de granularidad o detalle: Una vez establecido el contexto del proceso de negocio y los requerimientos a resolver, se define el nivel de detalle con el cual trabajará el modelo dimensional. El nivel de granularidad representa cada transacción realizada en el proceso de negocio y se almacena dentro de la de hechos. Por ejemplo, los niveles de granularidad de un proceso de registro de calificaciones puede ser la calificación asignada a los estudiantes, las inscripciones o reinscripciones que cada estudiante ha tendido. En otros procesos de negocio pueden ser las ventas de productos a un cliente, el manejo de cuentas bancarias, los ingresos netos, entre otros. El nivel de granularidad establecido al inicio del proyecto definirá el tipo de reportes que se podrán analizar, por ello, es importe analizar adecuadamente a que detalle se desea conocer la información.
3. Identificar las dimensiones: Al contar con los puntos uno y dos, se procede a identificar las dimensiones. Al definir las dimensiones es importante realizarlo con base al contexto sobre el cual se desean analizar los hechos de los procesos de negocio, normalmente se puede agilizar este proceso al responder las preguntas: quién, qué, dónde, cuándo, por qué y cómo, Por

ejemplo, la identificación de las dimensiones de tiempo, ubicación, artículos, proveedores, clientes, vendedores, entre otros.

4. Identificar los hechos: Finalmente, se crea la tabla de hechos utilizando las llaves foráneas de las dimensiones y agregando cada una de las métricas descritas en el proceso de negocio y de las cuales se deba realizar algún análisis minucioso (Kimball y Ross, 2013).

2.2.7. Herramientas para generación de reportes

Las herramientas BI han estado presentes durante décadas, no obstante, las capacidades de profundidad, funcionalidad y alcance continúan en constante evolución. A continuación, se presenta los diversos estilos analíticos de la solución BI (Sherman, 2015):

1. Reportes empresariales y gestión de alertas: Utilizados para generar reportes estáticos con formato predeterminado, accesibles para un gran número de personas.
2. Reportes a la medida (*Ad-hoc*): Permite la elaboración de reportes avanzados y dinámicos con características específicas para cada usuario.
3. Análisis predictivos o estadísticos: Los utilizan usuarios especializados del negocio como analistas y tomadores de decisiones, quienes, por medio de las funciones estadísticas y matemáticas puede encontrar patrones, correlaciones y proyecciones en los datos almacenados.
4. Cuadros de mando: Permite visualizar indicadores de forma gráfica, en tablas dinámicas, tablas estáticas y realizar *drill down & drill up* sobre los elementos establecidos.

3. Hipótesis

Al desarrollar una metodología para la elaboración de un sistema de estadísticas y reportes por medio de un *Data Warehouse*, se podrá disminuir el tiempo de elaboración y entrega de reportes, generando resultados oportunos y de calidad.

4. Objetivos

Objetivo general

Desarrollar una metodología para la elaboración de estadísticas y reportes con base en las teorías de *Inmon* y *Kimball* para concentrar y organizar los datos de diferentes fuentes, así como disminuir el tiempo de elaboración y entrega de reportes a entidades internas y externas de la Dirección de Innovación y Tecnologías de la Información de la UAQ, mediante un *Data Warehouse*.

Objetivos específicos

1. Identificar y/o definir los reportes y estadísticas de mayor impacto para la Dirección de Innovación y Tecnologías de la Información de la UAQ.
2. Definir las fuentes de datos internas (base de datos transaccionales de la UAQ) y externas (catálogos de instituciones gubernamentales).
3. Describir el modelo de base de datos multidimensional (DW) con base en los reportes del objetivo uno.
4. Diseñar los Procesos de Extracción, Transformación y Carga de datos transaccionales en un DW que facilite la explotación de la información.
5. Aplicar la propuesta para la elaboración de reportes y estadísticas basado en el DW definido.
6. Analizar los resultados obtenidos al utilizar la metodología diseñada con el DW.

5. Material y métodos o metodología

El proyecto de tesis se realizó en la Universidad Autónoma de Querétaro, utilizando la arquitectura y ciclo dimensional de *Kimball*. Durante el desarrollo, se generaron diversos DM con el objetivo de centralizar los datos de mayor relevancia en los procesos de elaboración de reportes y estadísticas de las distintas áreas de la UAQ. Estos datos se procesaron por medio de los procesos ETL, es decir, con elementos de extracción, transformación y carga de datos. Es importante mencionar que se utilizaron datos de prueba para proteger la información confidencial de la

Universidad y cumplir con las leyes y normas vigentes sobre la protección de datos personales. Sin embargo, la cuestión teórica, metodológica y práctica resulto funcional y podrán ser reutilizada en entornos similares.

5.1. Métodos

En el desarrollo del proyecto de tesis se utilizaron los métodos de investigación que a continuación se describen:

Descriptivo: Fue utilizado para recopilar, analizar y curar el contenido sobre los procesos, características, acciones, actores y tareas relacionadas con las soluciones de inteligencia de negocios, el DW, la arquitectura propuesta por *Kimball*, la arquitectura de *Inmon*, la arquitectura de datos e información, así como otras áreas afines con la transformación de los datos en información a través de reportes y estadísticas. Se llevó a cabo una búsqueda minuciosa de cada término para comparar las distintas vertientes de los autores y, de esa manera, filtrar la información significativa para la metodología utilizada en el proyecto, además de poder analizar los resultados obtenidos durante su ejecución.

Comparativo: Se revisaron y analizaron las características de los procesos involucrados en las arquitecturas y metodologías base de *Kimball* e *Inmon,* lo cual facilito la identificación de la opción óptima a las necesidades de la Universidad.

5.2. Técnicas

Para recabar la información de los reportes y estadísticas de las distintas áreas que intervinieron en la investigación, se aplicaron las siguientes técnicas:

- Observación directa: Esta técnica permitió revisar directamente el proceso que el responsable del negocio realizaba para obtener los reportes y estadísticas.
- Encuestas: Apoyaron en la recopilación de información de los usuarios involucrados y del entorno en el que se aplicó el proyecto.

- Entrevistas: Se utilizaron para obtener requerimientos confiables y sólidos para el DW.
- Curado de contenidos: Facilitó el tratamiento de la información y su representación en el proyecto de investigación.

5.3. Población y muestra

Las universidades, ya sean de carácter públicas o privadas, tienen una función de vital importancia sobre la gestión del conocimiento y la capacitación y formación de profesionistas e investigadores que serán responsables de innovar e impulsar a la sociedad. Es por ello que el esfuerzo continuo de las instituciones de educación les permite llevar a cabo esta misión histórica. En particular, las Universidades públicas adoptan posturas de crítica sobre los distintos conocimientos y, en general, sobre la organización social, además, estas instituciones también se someten a un riguroso escrudiño público constante, tanto en los modelos académicos como en la estructura administrativa. Las revisiones son parte integral para cualquier institución pública, y la UAQ, al recibir financiamiento de tipo federal, estatal y recursos generados internamente (propios), la rendición de cuentas es un compromiso ineludible. En este sentido, se contempla que las Universidades publicas generen reportes y estadísticas precisas sobre la comunidad estudiantil, docente y administrativa. Estos datos son esenciales para abordar las necesidades de cada entidad que lo requiera a fin de transparentar el trabajo realizado.

El caso de estudio se realizó en la Universidad Autónoma de Querétaro (UAQ), no obstante, es aplicable a cualquier institución de nivel superior pública. La UAQ, se encuentra ubicada en Av. 5 de Febrero esquina con Calle Hidalgo s/n, Centro Universitario "Cerro de las Campanas", Querétaro, Qro. C.P. 76010 Teléfono: 442 192 1200. De acuerdo Universidad Autónoma de Querétaro [UAQ] (2022), “La UAQ es un organismo público descentralizado del Estado, dotado de autonomía, personalidad jurídica y patrimonio propio, líder en investigación y programas educativos reconocidos”.

La misión de la Universidad Autónoma de Querétaro, con base a lo estipulado en UAQ (2022), es “ser una Institución de Educación Media Superior y Superior de carácter pública, autónoma, socialmente responsable, con libertad de cátedra y funciones en docencia, investigación, vinculación y extensión. Su compromiso social se establece con el desarrollo integral de Querétaro, México y con el reconocimiento a nivel internacional. Así mismo, la visión indica que, la UAQ como una Institución pública de Educación Media Superior y Superior, mantiene su autonomía como pilar fundamental en el ejercicio y promoción de sus funciones sustantivas, de su calidad académica, su compromiso y responsabilidad social en un espacio cultural plural y de libre pensamiento. Se destaca por ser una administración eficiente, austera, desconcentrada y transparente en el uso de sus recursos con rendición de cuentas, al establecer las perspectivas de la obligatoriedad, gratuidad y universalidad. Cuenta con Unidades Académicas y Administrativas certificadas y acreditadas, así como, con claras directrices y procedimientos para su aplicación en los ámbitos de organización de la docencia, investigación, extensión, vinculación y la administración de los recursos humanos y económicos”.

La UAQ cerró el periodo 2021 - 2022 con una planta de 2 mil 521 docentes que participan en 235 programas educativos de los diferentes niveles (bachillerato, TSU, licenciatura, especialidad, maestría y doctorado) de forma escolarizada y no escolarizada, atendiendo una matrícula de 33,739 estudiantes en sus 13 facultades y una escuela de bachilleres. Además, con un total de 2 mil 270 administrativos activos distribuidos en los distintos campus de la Universidad (UAQ, 2022).

Para cumplir con la misión y visión, la Universidad debe contar con información puntual e integra. Para ello existen diversos departamentos que fortalecen la automatización de procesos como la DITI y, los núcleos de información o receptores de información a cargo de la Dirección de Planeación y Gestión Institucional (DPGI), Dirección de Servicios Académicos (DSA), la Dirección de

Desarrollo Académico (DDA), la Dirección de Recursos Humanos (DRH) y la Coordinación de Información y Estadística (UAQ, 2022).

La Dirección de Innovación y Tecnologías de la Información (DITI), mantiene una estructura organizacional de cuatro coordinaciones para solventar las necesidades tecnologías de la UAQ. Una de dichas áreas, es la Coordinación General de Sistemas (CGS) “responsable de planear, organizar, dirigir y controlar las actividades requeridas para el desarrollo de software institucional, enfocado en la automatización de procesos de la Universidad. Así mismo, establecer los mecanismos para el soporte y mantenimiento de aplicaciones desarrolladas. Tiene como principal objetivo desarrollar e implementar proyectos de software que gestionen eficientemente los recursos universitarios y proporcionen información significativa para la planificación, evaluación y toma de decisiones”. (Universidad Autónoma de Querétaro (UAQ), 2023)

La DITI y CGS fue parte medular para la investigación, debido al alcance y responsabilidades con las que cuenta dentro de la UAQ. Por otra parte, se encuentran los núcleos de información o las áreas generadoras de los datos, por ejemplo, la Dirección de Planeación y Gestión Institucional (DPGI), que, con apoyo de la Coordinación de Información y Estadística (CIE), tienen como visión “contar con un sistema de planeación universitaria que le permita el uso eficiente y sustentable de los recursos humanos, financieros y materiales para promover un crecimiento sostenido de sus funciones sustantivas”; además, tienen como misión “generar y proporcionar datos e indicadores oportunos, suficientes y pertinentes para la planeación, la toma de decisiones y la evaluación institucional, que permitan a la institución ofrecer educación de calidad”. (Universidad Autónoma de Querétaro (UAQ), 2023)

La Dirección de Servicios Académicos (DSA), es un núcleo más de información, en este caso, los datos académicos de la UAQ. La DSA tiene como misión “brindar apoyo mediante un servicio de calidad a los estudiantes universitarios en todas las acciones administrativas referentes al ingreso,

permanencia y egreso, llevando el control del historial académico de los alumnos a lo largo de su estancia como estudiantes de los programas de estudio que ofrece la Universidad Autónoma de Querétaro en los niveles de bachillerato, licenciatura y posgrado". Así mismo, la Dirección de Desarrollo Académico, cuyo objetivo es "contribuir al fortalecimiento del proceso enseñanza-aprendizaje en los Programas Educativos de la Institución, cuenta con los programas de formación, evaluación y promoción docente; así como los de orientación educativa e institucional de tutorías". Por último, la Dirección de Recursos Humanos (DRH), "tiene a su cargo la administración del personal, gestión de asuntos laborales, contractuales, nominales, de servicios y de capacitación, promoviendo el desarrollo integral de cada uno de empleados de la UAQ" (UAQ, 2022).

Las áreas mencionadas en párrafos anteriores fueron parte medular en la elaboración del proyecto de investigación, de ellas surgen los datos en primera instancia y, a su vez, son las principales áreas que demandan información oportuna e integra para la toma de decisiones. Las estadísticas, informes y/o reportes solicitados por áreas internas o dependencias externas, recaen en gran medida sobre dichas Direcciones y Coordinaciones. Además, la CGS, llevó a cabo mejoras en las aplicaciones existentes para aumentar la calidad de los servicios ofrecidos. Así mismo, la CIE, logró acceder a catálogos centralizados para facilitar la elaboración de indicadores e información.

5.4. Metodología

La investigación se basó en las áreas de la especificadas en el apartado 5.1, de las cuales se eligieron los reportes y estadísticas que se consideraron podrían formar parte del *Data Warehouse*. Con lo anterior, se diseñó la metodología para fortalecer la emisión de reportes y estadísticas con base en datos de mayor calidad y centralizados. Al finalizar el proyecto, se identificaron y evaluaron cuantitativamente los resultados al aplicar la nueva metodología vs la forma anterior, para determinar si el uso del *Data Warehouse* impactó y fortaleció de forma positiva la generación de los reportes y estadísticas de la UAQ.

Con base a los instrumentos aplicados previamente para conocer el estado de los datos en los distintos núcleos de información y el nivel de complejidad para la elaboración de reportes y estadísticas en general, así como las recomendaciones y observaciones emitidas por distintos organismos externos de la Universidad, se definió una metodología que apoyará la elaboración de reportes y estadísticas con base a un repositorio centralizado de información. Esto permitió dar respuesta oportuna a las peticiones solicitadas.

La metodología desarrollada en la investigación fue con base en la arquitectura de *Kimball* descrita en la sección 2.2.6.1., donde se identifican tres características principales (Foster y Godbole, 2016), (Kimball y Ross, 2013):

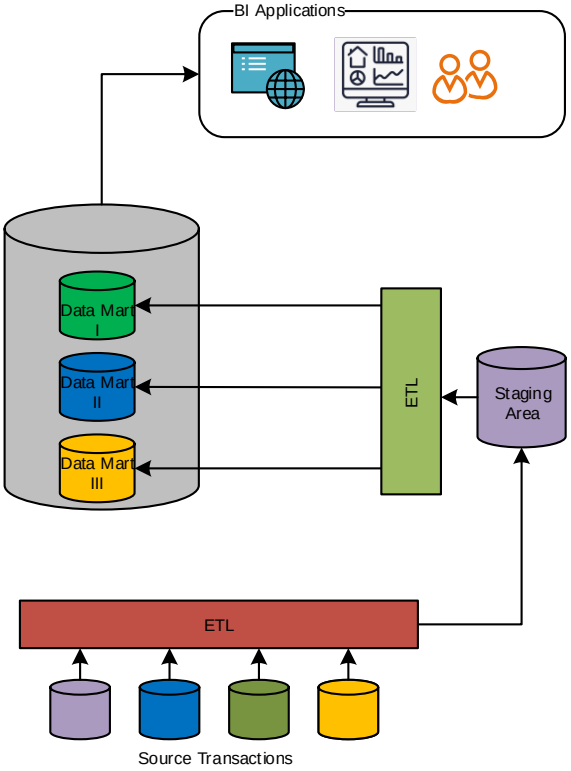
- DW base: Es la arquitectura base de *Kimball*, la arquitectura consiste en migrar los datos provenientes de los orígenes de datos directamente al DW.

- DW con *Staging Area* (Área de ensayo, pruebas o preparación): a partir de la arquitectura de *Kimball*, se agrega un área de preparación de datos, en la cual las organizaciones pueden mejorar la calidad de los datos provenientes de los sistemas de información de tipo transaccional (fuentes de datos), antes de migrarlo al DW, con el propósito de homologar, estandarizar y aumentar la calidad de los datos.

- DW con *Staging Area* y *Data Mart*: Para simplificar la tarea de analizar los procesos de toda la organización y con ello crear el DW, se definió una arquitectura para que las organizaciones dividieran el DW en unidades manejables de manera simple. Es por ello que surgen los *data mart*, para analizar áreas específicas del negocio. Para el caso del proyecto, se utilizó esta arquitectura (Figura 20).

Figura 20

Arquitectura con Staging Area y Data Mart



Nota. La arquitectura híbrida permitió crear una fuente principal de datos limpios e integrados llamada *Staging Area*, lo que facilitó la migración a los DM o al DW. Además, la explotación de la información se desarrolló a partir de consultas simples o con la implantación del portal BI. Adaptado de *Hybrid architecture with 3NF structures and dimensional Kimball presentation area (p.30)*, de Ralph Kimball, 2013, *The Data Warehouse Toolkit.*

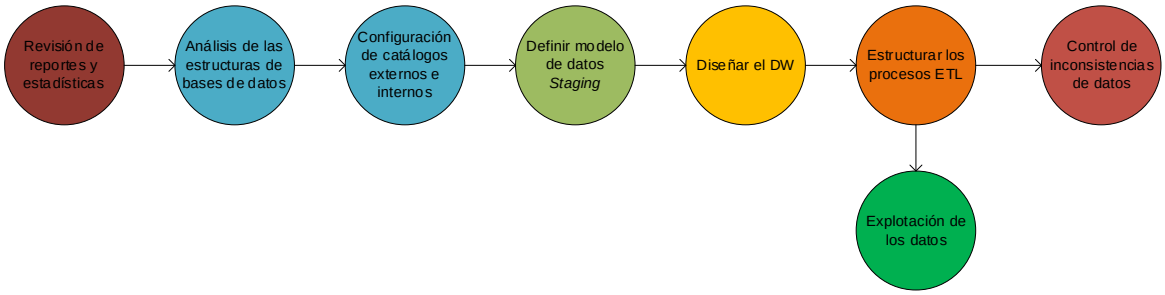
5.4.1. Caso de estudio

En seguida se describe detalladamente la metodología utilizada (Figura 21):

1. El proceso inició con el análisis de los reportes y estadísticas solicitados en ese momento por las diferentes instituciones gubernamentales y facultades de la UAQ, con el objetivo de identificar aquellos atributos que formarían parte del *Data Warehouse*.
2. Se estudiaron y comprendieron las estructuras de datos de las áreas de la DITI, la DSA y la DRH. Se identificaron los nombres de las tablas, las funciones que desempeñaban, las relaciones con otros esquemas y las reglas del negocio aplicables con la finalidad de localizar los datos necesarios para la implementación del *Data Warehouse,* de acuerdo al área temática analizada.
3. Se obtuvieron los datos externos (catálogos de datos externos) después de investigar el tipo de reportes solicitados por las diferentes entidades.
4. Se diseñó la estructura de datos para el área de *staging* (pre-procesamiento de datos).
5. Se diseñaron las dimensiones, niveles y tablas de hecho que contenían los datos provenientes del área de *staging* (Modelo multidimensional - DW).
6. Se diseñó la arquitectura de flujo de datos lógica y física.
7. Se definieron los procesos ETL, primeramente, para migrar la información al área de *staging* y posteriormente para el *Data Warehouse*.
8. En este punto se contaba con los reportes que permitieron identificar cuáles eran los datos inconsistentes dentro de las bases de datos de la UAQ para ser enviados a las personas encargadas de administrarlos.
9. Tras almacenar los datos en el *Data Warehouse*, se procedió a realizar pruebas utilizando herramientas para elaboración de reportes, tomando como base los reportes y estadísticas previamente analizados.

Figura 21

Elementos de la metodología propuesta



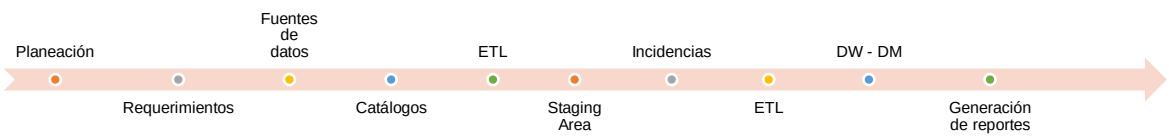
Nota. El diagrama representa el proceso desarrollado para eficientar y fortalecer los procesos de elaboración de reportes y estadísticas de la UAQ.

5.4.1.1. Revisión de reportes y estadísticas: Reportes del área administrativa

El primer punto de la metodología propuesta implicó revisar y analizar los reportes y estadísticas del área administrativa, académica, desarrollo docente y financiera (Figura 22). Asimismo, se llevó a cabo una planeación previa del proyecto, donde se definió el objetivo general o propósito del proyecto. Además, se describieron los objetivos específicos, la definición del alcance, el personal involucrado en el proceso del negocio, los departamentos o subáreas responsables y las tareas asociadas. Como parte del alcance, y, para definir el nivel de riesgo y las métricas, se consideraron las características de la calidad de los datos con base al proceso analizado. (Kimball et al., 2008).

Figura 22

Etapas de la metodología para generación de reportes UAQ



Nota. En la imagen se muestran las distintas etapas de la metodología implementada para la elaboración de reportes y estadísticas de la Universidad, partiendo de la planeación del proyectos hasta la generación de reportes.

Es común que los requerimientos seleccionados afecten las decisiones tomadas a lo largo del diseño e implementación del DW, por lo tanto, se definieron en compañía del personal involucrado (responsables del proceso del negocio). La comprensión correcta de los requisitos contribuyó a definir adecuadamente el alcance del proyecto, el modelado de los datos, las reglas de transformación, etc. En este sentido, fue necesario comprender a nivel macro el alcance general del proyecto DW, después, a nivel micro, un *Data Mart*, en el cual, se establecieron las necesidades de un proceso de negocio y la interacción con las demás áreas del DW, el personal involucrado y las fuentes de datos relacionadas. Los requerimientos se obtuvieron a través de entrevistas o sesiones con los involucrados del proceso de negocio y siempre respondieron a la pregunta ¿qué necesitamos hacer? (Silva Peñafiel, 2018).

En los siguientes puntos se muestra un ejemplo de la estructura del primer punto de la metodología:

- Nombre del proyecto: *Data Mart* para el control de nóminas y timbrado de la DRH - Dirección de Recursos Humanos.
- Objetivo general: Implementar un DM con los datos necesarios para solventar los reportes solicitados por entidades externas e internas de la UAQ
- Objetivos específicos:
  - ✓ Identificar y analizar los requerimientos de información.
  - ✓ Identificar los roles y responsabilidades de cada departamento asociado.
  - ✓ Identificar las fuentes de información y depuración de datos.
  - ✓ Migrar los datos al *staging area.*
  - ✓ Definir el modelo multidimensional.
  - ✓ Migrar los datos a las dimensiones y tabla de hecho.
  - ✓ Generar los reportes con alguna herramienta de BI.
- Alcance del proyecto: Para fines del proyecto, se consideraron los reportes solicitados al área administrativa, en específico a la Dirección de Recursos Humanos, como núcleo de información base. Los reportes y/o estadísticas

emitidas por DRH eran solicitadas por entidades externas, por ejemplo, el Servicio de Administración Tributaria (SAT), Auditoría Superior de la Federación (ASF), Gobierno del Estado de Querétaro, entre otras dependencias, así mismo, dentro de las áreas internas se encuentra la Rectoría, la Secretaría de la Contraloría y la Secretaria de Finanzas. La información requerida fue sobre el pago de nómina y el Comprobante Fiscal Digital por Internet (CFDI); en específico datos generales del personal, fuentes de financiamiento, ISR enterado al SAT y ISR declarado en los CFDI.

- Departamentos responsables:
  - ✓ Dirección de Recursos Humanos
  - ✓ Coordinación de Nóminas
  - ✓ Coordinación de Selección, Gestión y Control de Personal
  - ✓ Coordinación de Análisis Fiscal – Contable de nóminas
  - ✓ Dirección de Innovación y Tecnologías de la Información
  - ✓ Coordinación de General de Sistemas
- Personal involucrado: Para fines de la investigación no fue requerido listar al personal.
- Acciones y tareas asociadas

Tabla 7

Metodología para generación de reportes - Listado de departamentos y actividades asignadas

Departamento	**Acciones**
Coordinación de Nóminas	✓ Área encargada de la emisión de pagos y descuentos de las diferentes nóminas.
Coordinación de Selección, Gestión y Control de Personal	✓ Área encargada de recibir la documentación necesaria para el proceso de contrataciones y bajas.

	✓ Administración y resguardo de los expedientes de los empleados.
Coordinación de Análisis Fiscal – Contable de nóminas	✓ Realizaba la revisión y estructura de los informes trimestrales de nómina. ✓ Realizaba la revisión de expedientes de empleados. ✓ Realizaba el análisis de información de nóminas.
Coordinación de General de Sistemas	✓ Generaba reportes *ad-hoc.* ✓ Realizaba la revisión y mejoras en las aplicaciones involucradas. ✓ Realizaba el tratamiento de la información en caso de ser necesario.

Nota. Identificar las áreas y actividades involucradas en el proyecto de un DM permitió llevar el control detallado de los reportes desarrollados. En el ejemplo se observan las distintas áreas, tanto operativas como técnicas, que daban respuesta a las peticiones de información.

- Listado de requerimientos
  - ✓ Datos generales del personal que laboraba en la UAQ al momento del análisis (Tabla 8)
  - ✓ Suma de *xml* emitidos (vigentes) y total de pagos realizados por periodo y tipo de nómina (regular o extraordinaria)
  - ✓ Suma de importes netos de los *xml* vigentes por periodo (mensual)
  - ✓ Suma de Impuesto Sobre la Renta (ISR) con una periodicidad mensual (Tabla 9)
  - ✓ Suma de ISR con una periodicidad mensual y clave de nómina.
  - ✓ Distribución de ISR por fuente de financiamiento u origen el recurso: recursos propios o recurso federal (Tabla 11)

- ✓ Distribución de ISR por centro de gasto
- ✓ Distribución de ISR por tipo de personal (administrativo, docente, asimilado)
- ✓ Suma de importes netos por fuente de financiamiento interno mensual
- ✓ Listado de *Universal Unique Identifier* (*UUID*) o folio fiscal con RFC de empleado (a), clave de nómina, estatus del *UUID* (vigente, cancelado), ISR y mes de pago (Tabla 10).

Tabla 8

Ejemplo de datos generales del empleado solicitados por la DRH

No.	Clave	RFC	Adscripción	Celular	Domicilio	Correo institucional	Correo personal	Fec. de ingreso	Fec. baja	Tipo
1										
2										
3										
n										

Nota. Identificar los datos requeridos en los reportes permitió definir las dimensiones, métricas, jerarquías y hechos del DM. La tabla presenta las columnas requerida sobre los empleados.

Tabla 9

Ejemplo de suma de registros por año y mes de pago

Año	Mes	Total de CFDI	ISR Timbrado	Total de pagos	ISR Faltante	% Avance
2022	Enero	3500	$ 6,846,863.00	3600	$ 503,000.00	93.16%
2022	Febrero	4000	$ 8,846,863.00	4002	$ 18,000.00	99.80%
2022	Marzo	3800	$ 7,846,863.00	3800	$ -	100.00%
…	…	…	…	…	…	…
	Totales	11300	$ 23,540,589.00	11402	$ 521,000.00	97.83%

Nota. Las agrupaciones y jerarquías son comunes dentro del diseño de un DM, la tabla muestra los totales de pagos, totales de *XML* generados y el ISR correspondiente, donde se obtuvo el porcentaje de avance por mes.

Tabla 10

Ejemplo del nivel de granularidad de los datos

UUID	Clave empleado	RFC	Clave de nómina	ISR	Año	Mes
738e8fb9	F - 012345	ROOX550629AAA	F123456	$522,142.33	2022	11
c6ae8c60	H - 000779	GEGX930725BBB	H123456	$ 313.80	2022	1
ec74b64a	H - 000254	JEQX930725CCC	H123452	$ 313.80	2022	1

Nota. El nivel de granularidad de los datos permitió obtener información de un solo empleado o realizar agrupaciones anuales, por clave de nómina, etc.

Tabla 11

Ejemplo de reporte por origen del recurso

Año	Mes	Fuente de Financiamiento	ISR
2022	Enero	Federal	$ 503,000.00
2022	Enero	Estatal 1	$ 2,800,123.00
2022	Enero	Estatal 2	$ 1,000.00
2022	Enero	Propio	$ -

Nota. Las agrupaciones de los datos se realizaron de acuerdo al nivel de granularidad que tenía el DM, en la tabla se muestran las agrupaciones por origen del recurso, año y mes de pago.

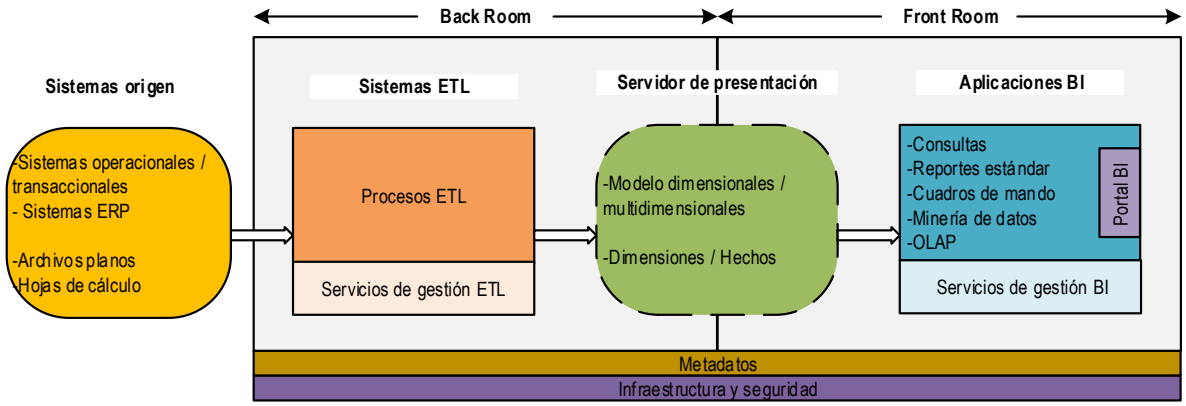
5.4.1.2. Estructuras de base de datos: Identificación de fuentes de datos

La fase siguiente de la metodología empleada durante el proyecto de tesis (Figura 21 y 22) implicó un análisis de las estructuras de las bases de datos y la identificación de los orígenes de datos, que facilitaron la recopilación de información para el DM, tomando como base los requerimientos solicitados por el usuario. En el ciclo de vida dimensional de *Kimball,* este proceso equivale al diseño de la arquitectura técnica del DM o DW, en esta etapa, únicamente se tomó como referencia el *back room* para migrar los datos desde el origen hasta el área de preparación (*staging area*).

Los entornos DW requirieron la integración de diversas tecnologías, respondiendo a la pregunta: ¿Cómo haremos el DW?, la arquitectura identificó los componentes necesarios para cada fase del DW, desde la adquisición de los datos hasta las soluciones BI o de inteligencia de negocios. La arquitectura brindó soporte al *back room* (adquisición de datos), el cual incluyó desde la búsqueda y transformaciones necesarias de los datos; y al *front room* (acceso a datos) encargado de recibir los datos del *back room* y entregarlos a los usuarios finales. Las dos capas (*back room* y *front room*) se comunican a través de los metadatos, los cuales almacenan, entre otras cosas, un conjunto de información de control acerca del DW, el contenido, los sistemas fuentes y el proceso de carga (Figura 23).

Figura 23

Arquitectura técnica de un Data Warehouse



Nota. Dentro del ciclo de vida dimensional de *Kimball*, se presentaron dos procesos principales, el primero de ellos denominado *back room* se encargó del análisis de los orígenes de datos para almacenar los datos en el DM; por otra parte, el *front room*, indicó la manera de presentar la información a los usuarios finales. Adaptado de *Relational DBMS architecture showing big data extensions (p.529)*, de Ralph Kimball, 2013, *The Data Warehouse Toolkit.*

El *back room* permitió definir la infraestructura tecnológica para visualizar y presentar los datos a los responsables del proceso. El objetivo principal era migrar los datos de los sistemas fuentes hacia el modelo dimensional, utilizando los

procesos ETL y las transformaciones que se consideraron pertinentes para la capa de presentación. En esta capa, la arquitectura contempló los siguientes aspectos:

- Ubicación de los orígenes de datos internas: los sistemas de información identificados en el proceso de negocio, el área de preparación de datos, archivos XML, TXT, entro otros.
- El flujo de los procesos ETL utilizadas durante el proceso de migración.

Por otro lado, se utilizó el *front room* para que los usuarios contarán con herramientas que facilitaran el análisis de los datos almacenados en el DM. Por lo tanto, la metodología se concibió estratégicamente para incorporar una capa intermedia que conectara a los usuarios con los datos provenientes del *back room*. Esta estructura garantizo el acceso a los almacenes de datos (*Data Mart*) a través de servicios de navegación, acceso, seguridad, monitoreo, administración de consultas, reportes estándar, etc. (Kimball et al., 2008), (Silva Peñafiel, 2018).

El análisis del requerimiento realizado en las etapas previas fue de gran utilidad para identificar las fuentes de datos y los elementos utilizados en los procesos ETL. Estos requerimientos describieron cuales serían los objetos a los que se necesitaba acceder para obtener la información requerida. Asimismo, fue necesario considerar tanto la migración histórica y como la actual. En el caso de la DRH, los datos a migrados correspondieron a los ejercicios 2018, 2019, 2020, 2021 y 2022; donde la información se encontraba en diversos formatos y en varias bases de datos operacionales. Asimismo, en los años 2018 al 2020, los datos se recabaron a partir de archivos XML, en hojas de cálculo y en bases de datos transaccionales; por último, en los años 2021 y 2022, los datos se identificaron en las bases de datos de la Universidad. La siguiente lista describe a detalla las fuentes de datos definidas:

- Archivos XML: Estos archivos contenían la información que había sido timbrada por el SAT, y que fue necesario realizar la descarga de los mismos para su posterior procesamiento. La estructura de los archivos correspondía a la establecida en la guía de llenado de los CFDI versión

3.3. Los nodos contenidos en el archivo XML fueron determinados con base al tipo de contratación del empleado (Figura 24).

- Archivos TXT: Los archivos contenían la estructura para creación de los XML que posteriormente serían timbrados, es decir, contenían los datos requeridos por el SAT de acuerdo a la guía de llenado de la versión 3.3. Estos archivos TXT se utilizaron como insumo para los sistemas responsables del timbrado (Figura 25).
- Hojas de cálculo: Dichas hojas contenían el detalle de los CFDI timbrados por el SAT (Tabla 12).
- Bases de datos transaccionales: Almacenaban información general de los pagos efectuados a los empleados y la distribución ejercida de la fuente de financiamiento.

Figura 24

Data Source – Archivos XML

```
<cfdi:Comprobante Version="3.3" Serie="Q_____1" Folio="1675" Fecha="2021-02-02T05:26:20" Sello="F89ppY ==" FormaPago="99" NoCertificado="0004807" Certificado="MIIGDTCCAHiBRM=" SubTotal="36303.69" Descuento="10849.26" Moneda="MXN" Total="25454.43" TipoDeComprobante="N" MetodoPago="PUE" LugarExpedicion="76010">/>
<cfdi:Conceptos>
        <cfdi:Concepto ClaveProdServ="8405" Cantidad="1" ClaveUnidad="ACT" Descripcion="Pago de nómina" ValorUnitario="36303.69" Importe="36303.69" Descuento="10849.26" />
</cfdi:Conceptos>
<cfdi:Complemento>
        <nomina12:Nomina Version="1.2" TipoNomina="O" FechaPago="2020-12-16" FechaInicialPago="2020-12-16"  FechaFinalPago="2020-12-31" NumDiasPagados="15.000" TotalPercepciones="36303.69" TotalDeducciones="10849.26" TotalOtrosPagos="0.00">
                <nomina12:Emisor RegistroPatronal="E____-107"> <nomina12:EntidadSNCF OrigenRecurso="IM" MontoRecursoPropio="18569.69" />
                </nomina12:Emisor>
                <nomina12:Receptor Curp="A_3" NumSeguridadSocial="1_6" FechaInicioRelLaboral="2014-05-11" Antigüedad="P346W" TipoContrato="01" Sindicalizado="No" TipoJornada="99" TipoRegimen="02" NumEmpleado="10007" Departamento="DXX___C" Puesto="5001408" RiesgoPuesto="1" PeriodicidadPago="04" SalarioBaseCotApor="755.35" SalarioDiarioIntegrado="1208.56" ClaveEntFed="QUE" />
                <nomina12:Percepciones TotalSueldos="3630.69" TotalGravado="1526.96" TotalExento="2103.73">
                        <nomina12:Percepcion TipoPercepcion="001" Clave="_0" Concepto="CODIGO _0" ImporteGravado="10593.78" ImporteExento="0.00" />
                        <nomina12:Percepcion TipoPercepcion="038" Clave="_1" Concepto="CODIGO _1" ImporteGravado="845.44" ImporteExento="1130.55" />
                        <nomina12:Percepcion ---------- … ---------- />
                </nomina12:Percepciones>
                <nomina12:Deducciones TotalOtrasDeducciones="8167.50" TotalImpuestosRetenidos="2681.76">
                        <nomina12:Deduccion TipoDeduccion="002" Clave="_1" Concepto="CODIGO _1" Importe="2681.76" />
```

```
                    <nomina12:Deduccion TipoDeduccion="001" Clave="_2" Concepto="CODIGO _2"
                    Importe="487.42" />
                    <nomina12:Deduccion ---------- … ---------- />
                </nomina12:Deducciones>
                <nomina12:OtrosPagos>
                    <nomina12:OtroPago TipoOtroPago="002" Clave="_4" Concepto="CODIGO _4"
                    Importe="0.00">
                        <nomina12:SubsidioAlEmpleo SubsidioCausado="0.00" />
                    </nomina12:OtroPago>
                </nomina12:OtrosPagos>
            </nomina12:Nomina>

    </cfdi:Complemento>
    </cfdi:Comprobante>
```

Nota. El contenido del archivo XML fue parcial y los datos presentados fueron de carácter representativo; sin embargo, mantuvieron la estructura original del SAT. Del contenido del XML, se obtuvieron los valores necesarios para poder complementar la información y realizar las depuraciones requeridas.

Figura 25

Data Source – Archivos TXT

```
DC|3.3|Q[redacted]1|726|2020-03-08T12:53:30|99|17151.31|354.30|MXN||16797.01|N|PUE|76010||||
EM|UAQ510111MQ9|UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
CNE|603||E[redacted]||IM|6638.26
RC|M[redacted]|M[redacted]|P01
CNR|O|2019-01-14|2019-01-01|2019-01-15|15.000|11404.91|354.30|5746.40|M[redacted]|
        1[redacted]|2004-02-01|P780W|01|Sí|99|02|5281|F[redacted]|10601|1|04|||13.90|22.24|QUE
CN|84111505|1|ACT|Pago de nómina|17151.31|17151.31|354.30
MI|i[redacted]|r[redacted]|
CNP|11404.91|||0.00|11404.91||||||||||
NPD|1|038|  1|CODIGO   1|0.00|483.77
NPD|2|038|  2|CODIGO   2|0.00|96.44
[redacted]
CND|354.30|
NDD|001|  2|CODIGO   2|68.89
N[redacted]
N[redacted]
N[redacted]
N[redacted]
NOP|999|  9|CODIGO   9|5746.40|0.00|||
NOP|002|  4|CODIGO   4|0.00|203.31|||
```

Nota. La estructura de los archivos TXT, era conforme a lo solicitado por los sistemas responsables del timbrado de nómina. Los TXT contenían los datos iniciales para el timbrado de nómina, por lo cual era importante obtener la información y compararla con la información timbrada en caso de datos faltantes

Tabla 12

Data Source – Hojas de cálculo

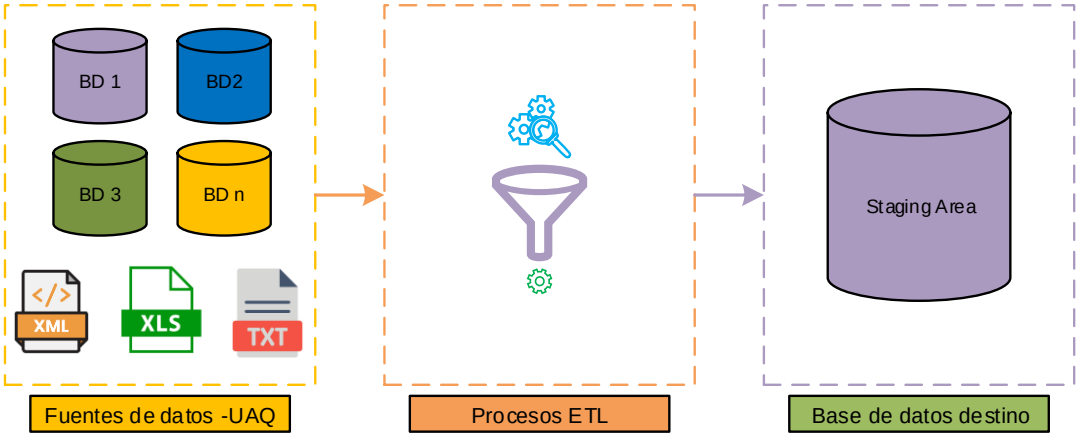
Folio Fiscal	RFC Receptor	Fecha Timbrado	Total	Serie	Folio	Versión
2bee8539	HE----IA	2019-02-25T13:47:37	$ 3,045.64	S--01141	270	3.3
dce955fd	RA----27	2019-02-25T13:47:36	$ 997.89	S--01141	531	3.3
165c3b04	GO----6A	2019-02-25T13:47:36	$ 2,393.45	S--01141	182	3.3
84ffc4a9	RA----C0	2019-02-25T13:47:35	$ 2,118.55	S--01141	487	3.3
bc5ad1ca	AE----E8	2019-02-25T13:47:35	$ 3,539.41	S--01141	292	3.3
4a58e8eb	RE----N1	2019-02-25T13:47:34	$ 718.89	S--01141	226	3.3
59fa6c95	VE----N2	2019-02-25T13:47:34	$ 2,469.26	S--01141	336	3.3
69d931a8	TE----Q6	2019-02-25T13:47:33	$ 1,581.28	S--01141	248	3.3

Nota. Las hojas de cálculo contenían la información de los CFDI generados en los sistemas de los proveedores del timbrado. Sin embargo, no contenían la información necesaria para obtener los reportes solicitados, estos archivos permitieron conciliar la información de los CFDI versus los pagos de los empleados.

Los datos se encontraban distribuidos en diversas fuentes de datos (*data source),* cada una aportando información complementaria y necesaria para la generación de los reportes y estadísticas solicitadas por la DRH. Por ello, fue se elaboraron estrategias para consolidar los datos, como parte del proceso, fue necesario migrar los datos al *staging area* para procesarla y realizar las validaciones correspondientes con el propósito de aumentar los datos de calidad (Figura 26). Esta área se convirtió en un espacio clave para concentrar los datos procesados después de aplicarles una serie de validaciones y procesos de limpieza. En este contexto, el objetivo de esta fase fue garantizar la calidad, la relevancia y la coherencia de los datos.

Figura 26

Identificación de fuentes de datos



Nota. En la tercera fase de la metodología propuesta, previo al análisis y revisión de los requerimientos, se identificaron las fuentes de datos para ser tratadas mediante procesos ETL y migradas al *staging area*.

5.4.1.3. Configuración de catálogos externos e internos

Los reportes y estadísticas de la UAQ utilizaban catálogos base para estandarizar los resultados con las áreas internas de la Universidad y, en su caso, con dependencias externas que ya contaban con información homologada y formatos predefinidos. Estos catálogos son inventarios detallados de datos que posibilitan, la estandarización los datos en los sistemas de información, la definición de agrupaciones de información, la disminución de los errores de capturas manual y realizar búsquedas rápidas, entre otros beneficios.

En el caso específico, los catálogos empleados corresponden con el SAT, mismos que se integraron en los sistemas transaccionales para adaptar las estructuras de datos internas a las solicitudes realizadas por dicha entidad. Algunos de los catálogos utilizados se indican en la tabla 13.

Tabla 13

Relación de catálogos externos para el área administrativa

Catálogo	Descripción
c_FormaPago	Distintos medios de pago por los cuales se paga el servicio prestado. Por ejemplo, tarjeta de débito, cheque, efectivo, entre otros.
c_CodigoPostal	Permitió estandarizar la búsqueda de direcciones fiscales de los empleados.
c_PeriodicidadPago	Definió la periodicidad de los pagos realizados al empleado o prestador de servicios.
c_Banco	Contenía el listado de los bancos nacionales y extranjeros.
c_Colonia	Contenía el listado de las colonias nacionales, indicando el CP al que corresponde.
c_Estado	Contenía el listado de estados nacionales y algunos extranjeros, muestra la clave del país al que pertenece.
c_Municipio	Contenía el listado de los municipios, muestra la clave del estado al que pertenece.
c_OrigenRecurso	Indicaba el origen de recurso con el cual se realizaba el pago al empleado, por ejemplo recurso propio, recurso federal y mixtos.
c_TipoPercepcion	Contenía el listado de percepciones de acuerdo al SAT
c_TipoDeduccion	Contenía el listado de deducciones de acuerdo al SAT

Nota. Los catálogos se definieron de acuerdo a los datos solicitados por el SAT para el timbrado de la nómina y generación de reportes.

La implementación de los catálogos permitió la reestructuración de los orígenes de datos, específicamente las bases de datos de tipo transaccionales, que almacenaban los registros referentes al proceso de pagos, la información de los empleados y el pago de nóminas. Como resultado de la implementación de catálogos se actualizaron los sistemas transaccionales y algunos reportes que seguían los mecanismos de las fuentes de datos previas. Además, la adecuación de estos catálogos a la Universidad, facilitó la creación de múltiples reportes estandarizados, que, al contar con datos íntegros, se obtenían con la calidad necesaria para los procesos de obtención de reportes y estadísticas.

De igual manera, la Dirección de Recursos Humanos contaba con otros catálogos internos, con los cuales se realizaban las clasificaciones de los datos para crear los reportes establecidos en las secciones anteriores. Por motivos de seguridad de la información Institucional, en la tabla 14 de presentan únicamente ciertos datos de prueba.

Tabla 14

Relación de catálogos internos para el área administrativa

Catálogo	Descripción
c_ClaveNomina	Almacenaba las claves de nómina que se realizan en cada tipo de periodicidad de pago.
c_FuenteFinancimiento	Indicaba la fuente de financiamiento para control interno de la Universidad.
c_CentroGasto	Almacenaba las claves y nombres de los departamentos de la UAQ
c_TipoPersonal	Contenía el listado de los tipos de personal que laboran en la Universidad, por ejemplo, Docente, Administrativo, etc.

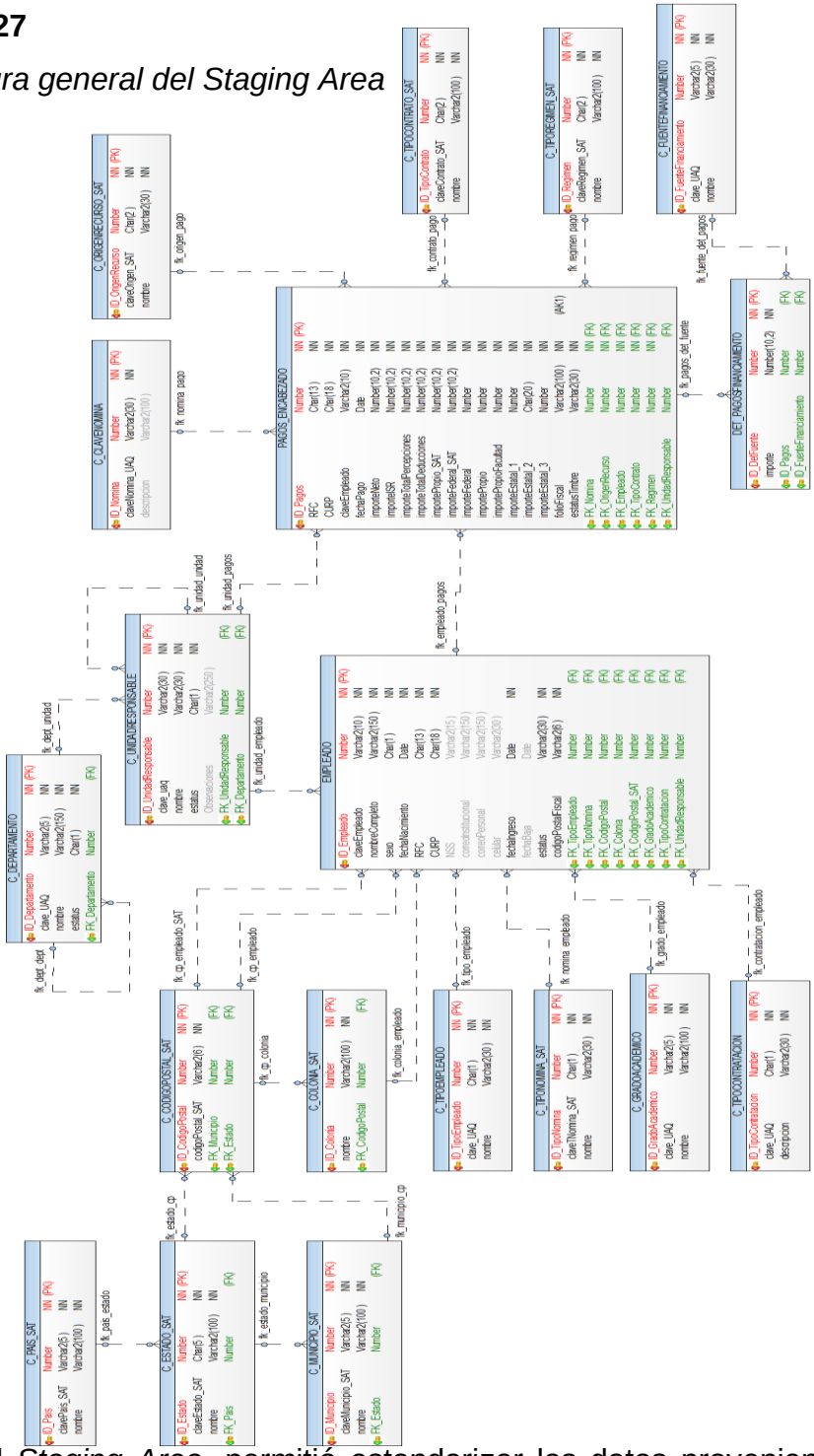
Nota. Los catálogos internos establecieron los alcances para la elaboración de los reportes y estadísticas que las áreas de la Universidad requerían para su análisis.

5.4.1.4. Definir el modelo del *Staging Area* y proceso ETL

En la *sección 5.4.1.2* se hizo referencia a múltiples fuentes de datos que resguardaban los datos empleados en la elaboración de los reportes del área administrativa (DRH). Al tratarse de varias fuentes de datos generadas en distintos momentos y con especificaciones de requerimientos cambiantes, fue común encontrar inconsistencias en los datos. Por ello, fue necesario crear un *staging area* con el propósito de centralizar, homologar la información y realizar depuraciones previas a la migración al DM. Con base a lo anterior, se definió una estructura de base datos flexible, normalizada y respetando la integridad en cada una de las tablas (Figura 27).

Figura 27

Estructura general del Staging Area



Nota. El *Staging Area*, permitió estandarizar los datos provenientes de múltiples orígenes de datos, además de normalizar los datos utilizados en los reportes y estadísticas de la Universidad. La imagen muestra el *staging area* para la DRH.

Por otra parte, los procesos ETL definidos para la extracción de los datos, las transformaciones realizadas y la carga de los datos, facilitaron la transferencia de datos desde las diversas fuentes hasta el almacén institucional. Durante el proceso se consolidaron los datos procedentes de diversas fuentes, las cuales presentaban estructuras y formatos variables. Además, tomando en cuenta que los procesos ETL representan el 70% del tiempo y esfuerzo en la construcción del DM se definieron las siguientes fases para llevar a cabo la actividad:

- La primera fase del proceso ETL consistió en identificar las fuentes de datos para realizar la extracción.
- La fase dos consistió en aplicar las transformaciones necesarias a los datos para asegurar su calidad mediante filtros, conversiones, cálculos, definición de claves, etc…
- En la fase tres, se creó un repositorio central (*staging área* o DM) de datos a partir de la integración d los diversos orígenes de datos.
- Por último, la fase cinco comprendió el poblado de la base de datos (Duque Méndez et al., 2016)

En el área de preparación de los datos, perteneciente a la fase de integración, se definieron dos etapas principales: la carga inicial y las actualizaciones. Cada una de las etapas se desarrolló considerando la menor afectación a los sistemas transacciones de la universidad. La carga inicial fue realizada considerando la mayor cantidad de datos históricos, estableciendo periodos de carga con base a las necesidades de información. Además, al ser el primer acercamiento a las fuentes de datos, se llevaron a cabo tareas indispensables como la limpieza de los datos y la mejora de su calidad. En algunas ocasiones, se permitió establecer modificaciones directamente a las fuentes de datos para evitar con ello la replicación de inconsistencias. Dependiendo de los distintos orígenes de datos y de la calidad definida, el proceso de carga inicial representó una lógica compleja para realizar algunas migraciones. Por lo cual, estas migraciones se realizaron con el apoyo de

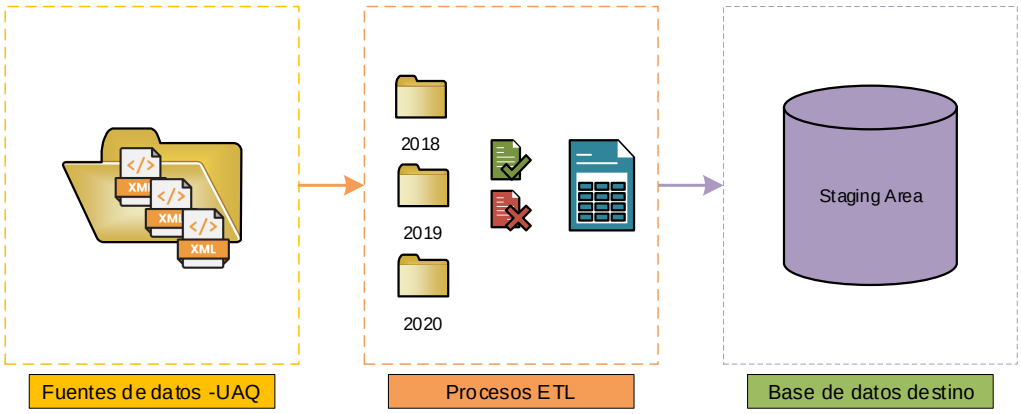
los responsables de procesos administrativo y el personal técnico responsable soporte al proceso.

La carga inicial se realizó con base al ejercicio fiscal de la universidad, del 2018 al 2020; posteriormente, la actualización de información se realizó de forma mensual, de acuerdo al origen de los datos del 2021 al 2023. En este sentido, la carga inicial del 2018 al 2020 obtuvo la información de los archivos XML, archivos TXT, hojas de cálculo y los esquemas de base de datos de la UAQ. Las fuentes de datos seleccionadas en la carga inicial de los datos se describen en los siguientes puntos:

- Archivos XML: Para realizar el proceso de carga inicial fue necesario utilizar un aplicativo para leer los archivos XML mediante el lenguaje de programación *Python* llamado *lectorXML.py*. Los archivos XML contenían algunos de los datos y estructura requerida por el *staging area* para su migración. La figura 28 muestro las actividades realizadas en esta etapa.

Figura 28

Carga inicial de archivos XML



Nota: Los archivos XML se consideraron la principal fuente de datos del *staging area*. Para realizar la integración, se codificó un programa para leer cada uno de los XML y cargarlos a un repositorio destino.

Una de las actividades de inicio consistió en la descarga masiva de los archivos XML almacenados en los servidores del SAT. Posteriormente, se seleccionaron los archivos a migrar de acuerdo a la fecha de pago del XML. A continuación, se ejecutó el programa *lectorXML.py* obteniendo como salida una hoja de cálculo, la cual contenía los campos necesarios para almacenarse en un repositorio destino mediante el *software* de *Pentaho Data Integration*. Durante esta etapa, se prestó total atención a la calidad de los datos, entre algunas acciones realizadas, se encuentra la asignación de valores por defecto en caso de que el valor no existiera, revisión de fechas, claves de nómina y estatus del timbre, por ejemplo, en los importes se asignó el valor de 0 (cero) y únicamente se consideró aquellos timbres con estatus *vigente*. En general, los archivos XML por sí mismos contenían información previamente depurada, lo que resultó en una alta calidad de los datos. Se muestra un parte del programa lectorXML.py en la Figura 29.

Figura 29

Carga inicial de archivos XML utilizando lectorXML.py

```
from bs4 import BeautifulSoup
from xml.dom import minidom
import pandas as pd
import os
from tqdm import tqdm

dfdiList = []
pathD = 'D://RUTA//NOMINAS_2018'

contenido = os.listdir(pathD)
for i in tqdm(contenido):
    with open(pathD+"//"+i, 'r', encoding="UTF-8-*") as f: data = f.read()
    bs_data = BeautifulSoup(data, 'xml')
    bs_Nombre = bs_data.find('cfdi:Receptor')
    Nombre = bs_Nombre.get('Nombre')
    bs_Rfc = bs_data.find('cfdi:Receptor')
    Rfc = bs_Rfc.get('Rfc')
    if(bs_data.find('nomina12:Receptor')):
        bs_NumEmpleado = bs_data.find('nomina12:Receptor')
        NumEmpleado = bs_NumEmpleado.get('NumEmpleado')
        bs_Curp = bs_data.find('nomina12:Receptor')
        Curp = bs_Curp.get('Curp')
        bs_NumSeguridadSocial = bs_data.find('nomina12:Receptor')
        NumSeguridadSocial = bs_NumSeguridadSocial.get('NumSeguridadSocial')
        bs_TipoContrato = bs_data.find('nomina12:Receptor')
        TipoContrato = bs_TipoContrato.get('TipoContrato')
        bs_Sindicalizado = bs_data.find('nomina12:Receptor')
        Sindicalizado = bs_Sindicalizado.get('Sindicalizado')
        bs_Departamento = bs_data.find('nomina12:Receptor')
        Departamento = bs_Departamento.get('Departamento')
    else:
        NumEmpleado = "" Curp = "" NumSeguridadSocial = "" TipoContrato = "" Sindicalizado = ""  Departamento = ""
    if(bs_data.find('cfdi:Comprobante')):
        bs_Total = bs_data.find('cfdi:Comprobante')
        Total = bs_Total.get('Total')
        bs_SubTotal = bs_data.find('cfdi:Comprobante')
        SubTotal = bs_SubTotal.get('SubTotal')
```

```
        bs_Serie = bs_data.find('cfdi:Comprobante')
        Serie = bs_Serie.get('Serie')
        bs_FormaPago = bs_data.find('cfdi:Comprobante')
        FormaPago = bs_FormaPago.get('FormaPago')
        bs_Folio = bs_data.find('cfdi:Comprobante')
        Folio = bs_Folio.get('Folio')
    else:
        Total = "0" SubTotal = "0" Serie = "0" FormaPago = "0" Folio = "0"
    if(bs_data.find('nomina12:Nomina')):
        bs_FechaPago = bs_data.find('nomina12:Nomina')
        FechaPago = bs_FechaPago.get('FechaPago')
        bs_FechaInicialPago = bs_data.find('nomina12:Nomina')
        FechaInicialPago = bs_FechaInicialPago.get('FechaInicialPago')
        bs_FechaFinalPago = bs_data.find('nomina12:Nomina')
        FechaFinalPago = bs_FechaFinalPago.get('FechaFinalPago')
        bs_TotalDeducciones = bs_data.find('nomina12:Nomina')
        TotalDeducciones = bs_TotalDeducciones.get('TotalDeducciones')
        bs_TotalPercepciones = bs_data.find('nomina12:Nomina')
        TotalPercepciones = bs_TotalPercepciones.get('TotalPercepciones')
    else:
        OrigenRecurso = "" MontoRecursoPropio = "0"
    if(bs_data.find('tfd:TimbreFiscalDigital')):
        bs_uuid = bs_data.find('tfd:TimbreFiscalDigital')
        uuid = bs_uuid.get('UUID')
        bs_FechaTimbrado = bs_data.find('tfd:TimbreFiscalDigital')
        FechaTimbrado = bs_FechaTimbrado.get('FechaTimbrado')
    else:
        uuid = "" FechaTimbrado = ""
    if(bs_data.find('nomina12:Percepciones')):
        bs_TotalSueldos = bs_data.find('nomina12:Percepciones')
        TotalSueldos = bs_TotalSueldos.get('TotalSueldos')
        bs_TotalGravado = bs_data.find('nomina12:Percepciones')
        TotalGravado = bs_TotalGravado.get('TotalGravado')
        bs_TotalExcento = bs_data.find('nomina12:Percepciones')
        TotalExcento = bs_TotalExcento.get('TotalExento')
    else:
        TotalSueldos = "0" TotalGravado = "0" TotalExcento = "0"
    if(bs_data.find('nomina12:Deducciones')):
        bs_TotalOtrasDeducciones = bs_data.find('nomina12:Deducciones')
        TotalOtrasDeducciones = bs_TotalOtrasDeducciones.get('TotalOtrasDeducciones')
        bs_TotalImpuestosRetenidos = bs_data.find('nomina12:Deducciones')
        TotalImpuestosRetenidos = bs_TotalImpuestosRetenidos.get('TotalImpuestosRetenidos')
    else:
        TotalOtrasDeducciones = "0" TotalImpuestosRetenidos = "0"
    if (bs_data.find('nomina12:Deduccion', {'TipoDeduccion':'002'})):
        bs_Importe = bs_data.find('nomina12:Deduccion', {'TipoDeduccion':'002'})
        if(bs_Importe.get('Importe') is not None): Importe = bs_Importe.get('Importe')
    else:
        Importe = "0"

    persona = [NumEmpleado,Nombre,Rfc,Curp,NumSeguridadSocial,TipoContrato,Sindicalizado,Departamento,Total,SubTotal,Serie,
               FormaPago,Folio,FechaPago,FechaInicialPago,FechaFinalPago,TotalDeducciones,TotalPercepciones,OrigenRecurso,
               MontoRecursoPropio, uuid, FechaTimbrado, TotalSueldos, TotalGravado, TotalExcento, TotalOtrasDeducciones,
               TotalImpuestosRetenidos, Importe]


df = pd.DataFrame(dfdiList,columns =['NumEmpleado', 'Nombre', 'Rfc', 'Curp', 'NumSeguridadSocial', 'TipoContrato', 'Sindicalizado',
          'Departamento', 'Total','SubTotal','Serie',  'FormaPago', 'Folio', 'FechaPago', 'FechaInicialPago', 'FechaFinalPago',
          'TotalDeducciones', 'TotalPercepciones','OrigenRecurso','MontoRecursoPropio', 'uuid', 'FechaTimbrado', 'TotalSueldos',
          'TotalGravado','TotalExento','TotalOtrasDeducciones','TotalImpuestosRetenidos','Importe'])

df.to_csv("D://RUTA//XMLSalida.csv")
```

Nota. El programa *lectorXML.py* permitió obtener información de un conjunto de archivos XML contenidos en una carpeta y generar como salida un archivo CSV con los campos requeridos en las tablas del *staging area*.

Después de completar la lectura de los archivos XML, el programa *lectorXML.py* generó un archivo CSV, el cual contenía los datos extraídos de los XML. Estos datos fueron migrados al *staging area*, complementando los datos con los almacenados en la base de datos Universitaria. Los archivos CSV estaban compuestos por campos como, el número de empleado, nombre, RFC, CURP, NSS, tipo de contrato, sindicalizado, departamento, serie, folio interno, total, subtotal, forma de pago, fecha de pago, fecha inicial de pago, total exento, total deducciones, total percepciones, origen del recurso, monto del recurso propio, folio fiscal, fecha de timbrado, total de sueldos, total gravado, total impuestos retenidos, la fecha final de pago e importe. Cada una de las columnas mantuvo una relación con el *staging area*. En la tabla 15 se muestra las equivalencias entre las tablas transaccionales y el *staging area* diseñada en el proyecto.

Con el apoyo del mapa de datos lógico comenzó la fase de implementación de los procesos ETL, teniendo como objetivo la migración de los datos con base a los requerido en la figura 27. En la etapa de los procesos ETL, se recabaron los datos de varios orígenes y características específicas para ser migrados al repositorio centralizado o almacén de datos, garantizando un enfoque sistemático y preciso. En este último paso, se presentaron múltiples alternativas para migrar los datos, por ejemplo, una de ellas implicaba ejecutar consultas SQL entre las bases de datos involucradas, del mismo modo se utilizaron hojas de cálculo pre-cargadas y mediante herramientas especializadas que facilitaron la migración de múltiples orígenes de datos. En el caso particular del proyecto, la tabla “empleado” fue poblada mediante la aplicación *Pentaho Data Integrator*, la cual es una herramienta versátil y potente para el manejo de los datos y su tratamiento. La figura 30 proporciona una visión general de algunos elementos clave utilizados durante el proceso de migración. Sin embargo, debido a la complejidad del proceso y a la naturaleza de información, solo se muestra una parte representativa del proceso para salvaguardar la seguridad de los datos institucionales.

Tabla 15

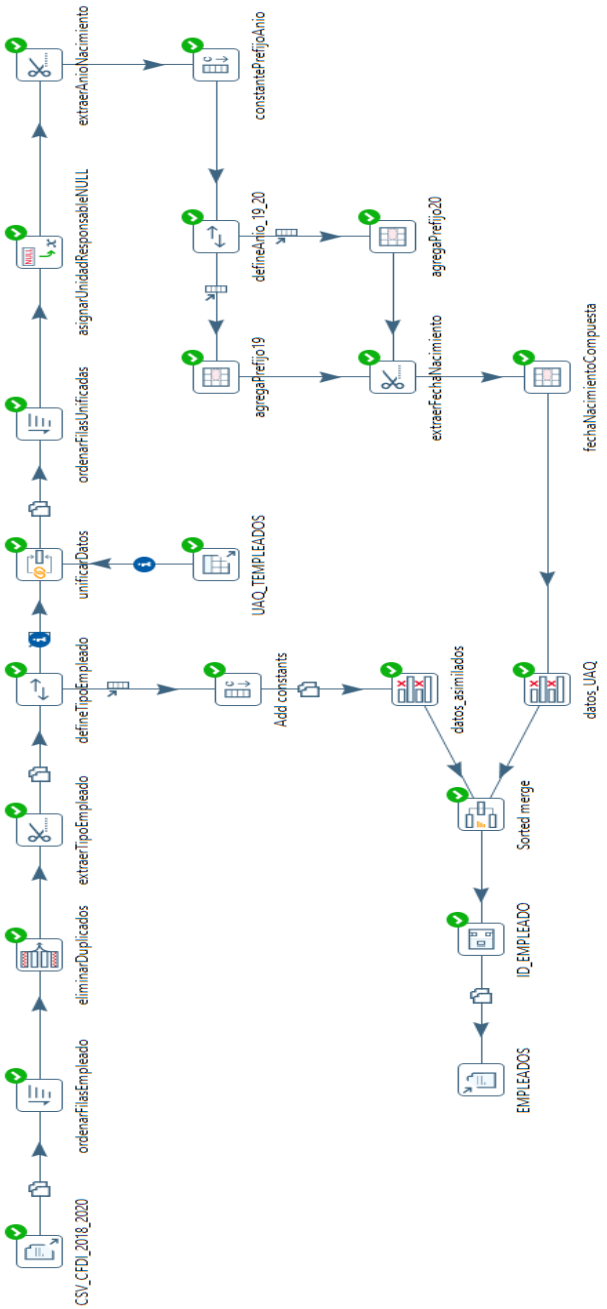
Mapa de datos lógico de la tabla empleados del Staging Area

Origen (Source)				Destino (Target)			
Fuente	Columna	Tipo de dato	Tabla	Columna	Tipo de dato	Transformación
CSV_TimbradoSAT	NumEmpleado	String	T_Empleado	claveEmpleado	varchar2(10)	Se validaban n caracteres para los empleados de base, honorarios, eventual y jubilados. En el caso de personal asimilado iniciaban con la primer letra del departamento al que pertenecen.
	Nombre	String		nombreCompleto	varchar2(150)	-
	Rfc	String		RFC	char(13)	Los valores *null*, no fueron migrados. Se generó un reporte para revisión posterior.
	Curp	String		CURP	char(18)	Los valores *null*, no fueron migrados. Se generó un reporte para revisión posterior.
	NumSeguridadSocial	String		NSS	varchar2(15)	Aplico únicamente para empleados de base, eventual y jubilados. En caso de haber contenido algún valor null en dichas nóminas, el registro no fue migrado.
	sexo	varchar2(1)		sexo	char(1)	-
	fecha_nacimiento	date		fechaNacimiento	date	En caso de valores *null*, se tomó el dato del RFC o CURP
	correo_uaq	varchar2(80)		correoInstitucional	varchar2(150)	Únicamente se aceptaron con domino @uaq.mx
	correo_personal	varchar2(80)		correoPersonal	varchar2(150)	Se validó que contendría un @
	celular	varchar2(15)		celular	varchar2(30)	-
	fecha_ingreso	date		fechaIngreso	date	-
	fecha_baja	date		fechaBaja	date	-
	estatus_contratacion	varchar2(1)		estatus	varchar2(30)	-
c_codigo_postal_SAT	cp_fiscal	varchar2(6)		codigoPostalFiscal	varchar2(6)	Se aceptaron valores *null* para los años 2018 - 2022, al no haber sido obligatorio por el SAT
c_tipo_empleado	tipo_empleado	varchar2(1)		FK_TipoEmpleado	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_tipo_nomina_SAT	tipo_nomina_SAT	char(1)		FK_TipoNomina	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_codigo_postal_SAT	cp	varchar2(6)		FK_CodigoPostal	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_colonia_SAT	colonia	varchar2(100)		FK_Colonia	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_codigo_postal_SAT	cp_fiscal	varchar2(6)		FK_CodigoPostal_SAT	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_grado_academico	grado_academico	varchar2(3)		FK_GradoAcademico	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_tipo_contratacion	nomina	varchar2(1)		FK_TipoContratacion	number	Los valores *null*, no fueron migrados. Se generó reporte de incidencia.
c_departamento	departamento	varchar2(3)		FK_UnidadResponsable	number	En caso de valores *null*, se agregó como constante DRH

Nota. El mapa de datos lógico proporcionó una visión general de los orígenes de datos utilizados en los procesos ETL. Para llenar la tabla empleado se utilizaron los datos obtenidos tras la lectura de los archivos XML, de igual manera, se necesitaron datos de las tablas transaccionales de la Universidad y los catálogos correspondientes para estandarización de los datos.

Figura 30

Proceso ETL utilizando Pentaho Data Integrator



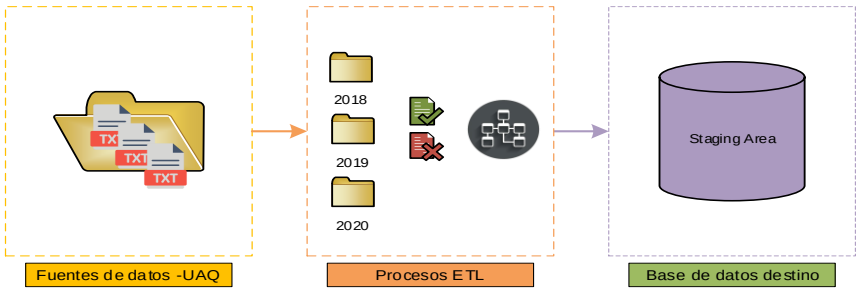
Nota. Para realizar los procesos ETL se utilizó la herramienta *Pentaho Data Integrador*, por ejemplo, se empleó esta herramienta para migrar la información contenida en un archivo CSV y cargar la tabla empleado, para ello, se definió un flujo de trabajo con distintas transformaciones de acuerdo a la tabla 15.

Como se mencionaba en párrafos anteriores, en la obtención de la información de los años 2018 al 2020, fue necesario utilizar los archivos XML con el objetivo de identificar exactamente qué información se encontraba timbrada de cada empleado, de igual manera, se corroboró la información con apoyo de los archivos TXT, previamente generados para el proceso de timbrado. En este sentido, los datos se encontraban dispersos por la ausencia de reglas del negocio y el seguimiento dado hasta ese momento. La falta de reglas de negocio, en varias ocasiones representa la acumulación de datos desorganizados y difíciles de entender; por ello, se realizaron procesos de verificación de la información entre la parte técnica y el usuario final con el propósito de aumentar la calidad de los datos.

- Archivos TXT: Para migrar los datos del origen de dato de tipo TXT, se creó una aplicación web especializada, está aplicación web tomaba el archivo general como entrada y extraía ciertos datos que eran cargados en la base de datos. La aplicación web, fue desarrollada para agilizar la migración de los datos y para asegurar que la información ingresada fuese la correcta. La figura 31 indica el proceso realizado en esta fase.

Figura 31

Carga inicial de archivos TXT



Nota. El contenido de los archivos TXT, permitió validar la integridad de la información fiscal (timbres de nómina). Para realizar el proceso se revisó la estructura de los archivos TXT, con base en ello, la aplicación lee el contenido de los archivos y los deposita en la base de datos destino.

Al completar las etapas de integración realizada sobre los datos contenidos en cada archivo TXT, se procedió a generar las vistas con los datos requeridos para

discernir qué pagos habían sido efectivamente timbrados y cuáles no. En este sentido, se procedió a una revisión minuciosa de la información con apoyo de los reportes proporcionados por el proveedor externo del timbrado de nómina, estos reportes se encontraban en formato de hojas de cálculo y abarcaban el periodo comprendido entre 2018 y 2020. La integración con las hojas de cálculo fue necesaria debido a que los datos del timbrado experimentaban actualizaciones periódicas, de acuerdo a cada ciclo de pago, por ejemplo, quincena a quincena. Dentro de esta dinámica implicaba que algunos registros podían ser cancelados o incluso re- timbrados en el transcurso del tiempo, por lo tanto, con el objetivo de garantizar datos íntegros, de calidad y factibles se procedió a cotejarlos con los reportes generales emitidos desde la plataforma del proveedor. Dicha medida fue realizada para confirmar que los datos reflejaban la realidad de lo efectivamente timbrado y los pagos faltantes de timbrar.

- Hojas de cálculo: La migración de los datos desde hojas de cálculo represento un proceso relativamente sencillo y eficaz, ya que se importaron directamente utilizando un cliente SQL y la herramienta *pentaho data integration*.
- Bases de datos transaccionales: La etapa final involucró la incorporación de los registros provenientes de los sistemas de bases de datos transaccionales que contenían los detalles de los pagos efectivamente efectuados. Para fines del proyecto, se mostraron algunas de las vistas realizadas para estandarizar la información en conjunto con los archivos XML, TXT y hojas de cálculo, con la finalidad de lograr una visión integral y estandarizada de los datos. La tabla 16 describe algunos de los objetos SQL implementados sobre la base de datos de la UAQ, los cuales contenían información de consulta para otros procesos institucionales.

Tabla 16

Muestra de los objetos de base de datos

Ubicación	Tipo de objeto	Nombre de objeto	Descripción
Base Institucional	Vista materializada	nrh.view_impuestos	Contenía datos del 2015 al 2023 referentes al ISR efectivamente pagado a cada empleado, por clave y tipo de nómina.
Base Institucional	Vista materializada	fac.view_enominas	Agrupaba la información de aquellos empleados recibieron más de un pago en el mismo período de pago. Se obtuvo el total pagado e ISR.
Base Institucional	Vista materializada	fac.view_timbres_excel	De acuerdo a los reportes del proveedor de facturación, durante los años 2018 al 2020, se obtuvo el resumen de los pagos timbrados durante dichos años por cada empleado.
Base Institucional	Vista materializada	fac.view_timbres_xml	De acuerdo a los XML y archivos TXT durante los años 2018 al 2020, se obtuvo el resumen de los pagos timbrados durante dichos años por cada empleado.
Base Institucional	Vista	nrh.view_epagos	Resumen general de los datos a timbrar
Base Institucional	Vista	nrh.view_dpagos	Detalle de cada pago realizado a los empleados. Contenía los conceptos de pago del SAT
Base Institucional	Vista	fac.v_timbres	Detalle de los pagos timbrados exitosamente del 2021 a una fecha determinada.
Base Institucional	Vista	fac.v_timbres_error	Contiene los pagos que no pudieron ser timbrados por alguna inconsistencia en los datos o cálculos. Fue necesaria una revisión personalizada. Del 2021 a una fecha determinada
Base Institucional	Vista	fac_epagos	Pagos disponibles para su timbrado del 2021 a una fecha determinada. Contenía datos como el número de pago la fecha de emisión, el monto total, clave del empleado, número de seguro social, salario base, la fecha de inicio y fin del período de pago, etc.
Base Institucional	Vista	fac_timbrado	Encabezado de los pagos timbrados, almacenaba la información global del 2021 a una fecha determinada.
Staging area	tabla	empleado	Información general de los empleados. Dentro de los campos de incluía el nombre completo del empleado, número de identificación personal (por ejemplo, CURP), fecha de nacimiento, domicilio fiscal, correo electrónico, entre otros.
Staging area	tabla	pagos_encabezado	Contenía la información de cada pago realizado a los empleados en los periodos definidos, así mismo, los importes agrupados por percepciones, deducciones y origen de recurso de acuerdo a la UAQ.
Staging area	tabla	det_pagosfinanciemiento	Indicaba el detalle de los pagos por fuente de financiamientos de acuerdo al SAT.
Staging area	tabla	varios catalogos	La figura 27, muestra varios catálogos utilizados para homologar la información almacenada en las tablas del *staging area*.

Nota. Los objetos descritos son una muestra de cómo se llevó el control del timbrado fiscal y los reportes requeridos por la Universidad y las instancias externas. Las vistas y tablas creadas provienen de distintas fuentes de información previamente analizadas y pobladas con procesos ETL.

5.4.1.5. Diseño del modelo dimensional: Data Warehouse

La siguiente fase de la metodología para fortalecer la generación de reportes y estadísticas en la Universidad consistió en el diseño del modelo dimensional en forma de estrella, implementado por medio de un *Data Warehouse* o *Data Mart*. Este modelo resultó ser el más adecuado de acuerdo a lo descrito en la sección de esquemas dimensionales. En dicha sección se estableció que la estructura de datos dimensional se considera como una estrategia para diseñar bases de datos, misma

que se utiliza para implementar soluciones de tipo BI, teniendo como objetivo el poder analizar una amplia cantidad de datos utilizando las tablas de hechos, métricas, jerarquías y dimensiones.

Los hechos y dimensiones se representan por medio de un modelo conceptual en estrella, donde se establece que las tablas de hechos se conectan con las dimensiones correspondientes a la temática o proceso de negocio. En este sentido, las dimensiones contienen la información o los elementos sobre los cuales se desea visualizar los hechos, es decir, bajo qué contexto se muestran las métricas de la tabla de hechos. Las dimensiones son principalmente características descriptivas de los datos. Por ejemplo, las dimensiones de ventas de una tienda pueden ser el tiempo, la ubicación, los artículos y los clientes. La dimensión de tiempo puede contener el día, la semana, el mes o el trimestre; la dimensión de ubicación se puede analizar por país, estado o ciudad; la dimensión de artículo describe las categorías, marcas y líneas del artículo; por último, la dimensión cliente almacenaría datos generales como su ubicación y comportamiento de compras.

Por otra parte, las tablas de hechos, son los elementos centrales del modelo dimensional y sus columnas representan los eventos o transacciones que el negocio requiere medir, por ejemplo, ventas, ingresos, impuestos, promedios, etc. Continuando con el ejemplo anterior, las tablas de hechos contendrían información como los precios de venta de los artículos, la cantidad vendida, descuentos aplicados, etc…A diferencia de las dimensiones, las tablas de hechos no contienen información descriptiva, cada fila representa una transacción individual.

En el caso específico del proyecto de investigación, el modelo dimensional fue definido, en primera instancia, al identificar el requerimiento del proceso de negocio, descrito en la sección revisión de reportes y estadísticas: Reportes del área administrativa, referente al proceso de modelado de un *Data Mart* para el control de nóminas y timbrado de la Dirección de Recursos Humanos. En dicho apartado, se cumplió con la primera fase del modelo dimensional. En resumen, se identificó el ¿por qué? de la creación del *Data Mart* a nivel general, para ello se llevaron a cabo

diversas técnicas que facilitaron la obtención e identificación de los requisitos de información de la DRH, entre ellas, entrevistas, cuestionarios de apoyo y observaciones directas a las actividades involucradas. Los requerimientos se encuentran en la sección revisión de reportes y estadísticas: Reportes del área administrativa.

El segundo paso consistió en definir el máximo nivel de detalle, especificidad o nivel de granularidad en los datos. La importancia fundamental de este punto radicaba en la identificación de las medidas posibles y los contextos sobre los cuales se analizaron los datos. En este paso, se respondió a la pregunta ¿cuánto?, es decir, los reportes proporcionados se veían afectados o beneficiados con base al nivel de detalle de los datos almacenados en el DM, por ejemplo, ¿cuánto se ha vendido este mes?, considerando un nivel de granularidad alto, el resultado sería únicamente el total de las ventas registradas en el mes sin mayor información. Por otro lado, al utilizar un nivel de granularidad bajo, se obtendría el total del mes y se podría realizar el análisis a nivel de cada venta individualmente. Por lo anterior, en la implementación llevada a cabo, se adoptó el nivel de granularidad de acuerdo a los requerimientos obtenidos en el primer paso. Se buscaba alcanzar el nivel de detalle máximo para lograr una comprensión precisa de los datos, sin recurrir a réplicas exactas de las bases de datos transaccionales, lo complicaría su gestión y comprensión. En este sentido, se obtuvo información anual, semestral, trimestral, mensual, de los pagos de cada empleado.

Considerando los requerimientos de las secciones previas, se estableció el nivel de granularidad para las medidas (hechos) y perspectivas (dimensiones), las cuales se describen a continuación:

- Medidas
    - Total de pagos: Representaba la cantidad de pagos emitidos para cada empleado.
    - Total de timbres: Representaba la cantidad de pagos timbrados hasta el momento para cada empleado.

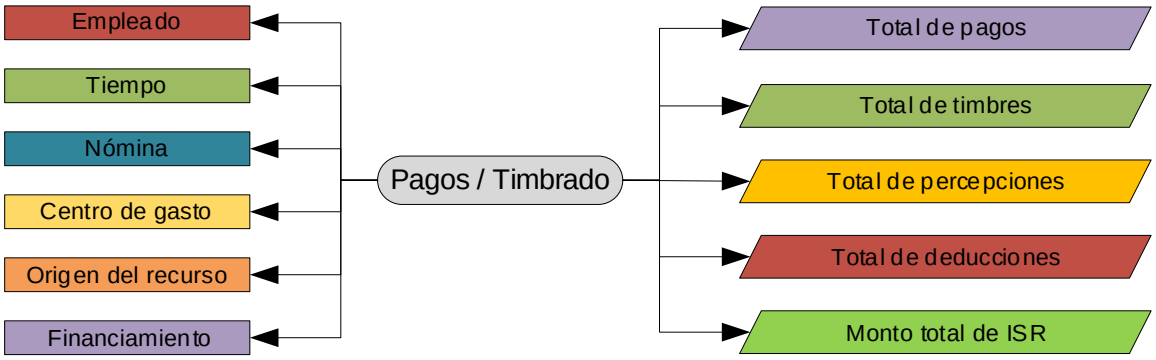
- Monto total de ISR enterado (timbrado): Suma del ISR para cada timbre emitido.
- Total de percepciones: Suma el nodo de percepciones para cada timbre emitido.
- Total de deducciones: Suma el nodo de deducciones para cada timbre emitido.

- Perspectivas
  - Nómina: Clave de nómina o serie, tipo de nómina (ordinaria y extraordinaria) y tipo contratación (honorarios, base, eventual, etc.…)
  - Fuente de financiamiento interna: La forma de pago que estableció la Universidad, recursos propios de rectoría, recursos propios de facultades, estatal1, estata2 y estatal3
  - Origen del recurso SAT: Clasificación del SAT, por ejemplo, recurso propio, federal o mixto.
  - Tiempo: Periodicidad de los pagos, las revisiones se dieron a nivel anual, semestral, trimestral, mensual, quincenal y semanal.
  - Ubicación: Agrupación de la Universidad, se definió como el centro de gasto o departamento donde laboró el empleado.
  - Empleado: Perspectiva principal del proceso de nómina y timbrado fiscal, son quienes reciben el pago.

A partir de las medidas y perspectivas identificadas en los puntos anteriores, se logró desarrollar una versión preliminar del modelo dimensional mediante un modelo conceptual. Este modelo ofreció una vista completa del proceso analizado, por una parte, la estructura de los datos, y, de igual manera, las relaciones o la cardinalidad existente entre ellos, lo que permitió visualizar los objetivos estratégicos del proyecto de manera clara y detallada. La representación gráfica para el modelo conceptual se visualiza en la figura 32.

Figura 32

Modelo conceptual del proyecto



Nota. El modelo conceptual del proyecto, permitió observar de forma sencilla las posibles dimensiones, métricas, niveles de agregación y jerarquías que formaron parte del DM.

Una vez concluida la revisión de los requerimientos de información de la organización, la definición del nivel de granularidad del proyecto y el establecimiento de un primer modelo conceptual, la siguiente etapa consistió en identificar las dimensiones que formarían parte del DW o DM. Las dimensiones se crearon con base a los requerimientos de información y el nivel de detalle o de granularidad deseado para consultar los datos del DM. Por consiguiente, las dimensiones son categorías o perspectivas utilizadas para entender el contexto sobre el cual se presentan los datos, es decir, son valores descriptivos que provienen de una o varias tablas del modelo relacional. Los componentes de una dimensión son los siguientes:

- atributos: Son las características que definen a la dimensión, por ejemplo, el nombre, marca, precio, son atributos de una dimensión de producto.
- jerarquías: Agrupan y organizan los atributos en niveles de detalle, por ejemplo, la dimensión tiempo incluía las jerarquías de año, semestre, trimestre, mes, semana, etc… Dichas jerarquías permitieron realizar un análisis detallado a distintos niveles de granularidad.

Las dimensiones responden a una serie de preguntas en relación al negocio, por ejemplo, *¿cuándo?*, *¿dónde?* y *¿quién?*. En el caso de la pregunta *¿cuándo?*, indica que el DM o DW, debe tener una dimensión que permita conocer *cuando* se realiza determinada acción, para ello se utilizan dimensiones de tiempo que, dependiendo de las reglas del negocio, establecen el contexto de medida ideal para los reportes. Así mismo, la pregunta *¿dónde?*, indica que la acción o transacción se ejecutó en algún *lugar*, por tanto, debe existir una dimensión para poder identificarlo, por ejemplo, la dimensión geográfica, almacenaría el país, estado, ciudad, CP, etc… Por último, la pregunta *¿quién?*, especifica la persona o entidad ejecutora de la acción, por ejemplo, el *empleado*, que pude tener atributos como nombre, departamento, fecha de contratación, etc… En este sentido, las dimensiones definidas respondieron los cuestionamientos anteriores.

En el proyecto de investigación se identificaron una serie de dimensiones y jerarquías que permitieron analizar la información en diferentes periodos de tiempo y de manera detallada para cada empleado, desde el ISR pagado en cada recibo como el ISR global de una nómina completa o el ISR anual de toda la Universidad. Las dimensiones definidas se observan en la tabla 17, de igual manera se describe la utilidad de cada dimensión y las diferentes jerarquías.

Tabla 17

Definición de dimensiones, jerarquías y atributos

Dimensión	Descripción general	Columna/atributo	Descripción	Jerarquía
dim_empleado	Almacenaba los datos correspondientes al personal de las distintas nóminas de la UAQ, por ejemplo, docentes, administrativos, asimilados, etc…	BK_claveEmpleado	Clave de negocio - empleado	
		nombreCompleto	Nombre completo	-
		sexo	Sexo, pude ser masculino o femenino	-
		fechaDeNacimiento	Fecha de nacimiento	-
dim_ubicacion	Almacenaba los datos generales del lugar donde laboran los empleados.	BK_campus	Clave de negocio - campus	
		campusPlantel	Campus donde se encontraba la ubicación del empleado	Nivel 1
		tipoDepartamento	Nivel de estudios predominante en el campus	Nivel 2
		BK_departamento	Clave de negocio – departamento	
		nombreDepartamento	Departamento o área asignada al empleado, como facultad de ingeniería,	Nivel 3

			departamento de recursos humanos	
		BK_unidadResponsable	Clave de negocio – unidad responsable	
		unidadResponsable	Ubicación donde laboró el empleado	Nivel 4
dim_tiempo	Estableció la temporalidad de la información almacenada en el DW o DM.	anio	Año donde se realizó la transacción o acción	Nivel 1
		trimestre	Trimestre donde se realizó la transacción o acción	Nivel 2
		mes	Mes donde se realizó la transacción o acción	Nivel 3
		fecha	Fecha donde se realizó la transacción o acción	Nivel 4
dim_financiamiento	Indicó la fuente de financiamiento y/o el origen del recurso destinado para los pagos de los empleados.	BK_claveFinanciamiento	Clave de negocio - financiamiento	
		tipoFinanciamiento	Tipo de financiamiento, puede ser interno (UAQ), externo (Gobierno)	Nivel 1
		origenDelRecurso	Descripción del tipo de financiamiento, por ejemplo, Federal, Propio, etc…	Nivel 2
dim_nomina	Contenía la información respecto a las nóminas de la Universidad.	tipoNomina	De acuerdo al catálogo del SAT, fue ordinaria o extraordinaria	Nivel 1
		periodicidad	Indicó el periodo de tiempo de pago de la nómina, semanal, quincenal, mensual y anual.	Nivel 2
		regimen_SAT	Tipo de régimen del empleado de acuerdo al catálogo del SAT.	Nivel 3
		descripcionNomina	Estableció el tipo de contratación: asimilados, honorarios, eventual, quincenal, jubilado y pensionado	Nivel 4
		serie	Identificador de cada nómina	Nivel 5
dim_categoria	Indicó la categoría o puesto del empleado.	tipoEmpleado	Indicó si el empleado es docente o administrativo	Nivel 1
		dedicacion	Indicó si un empleado es de tiempo completo, medio tiempo o profesor por asignatura	Nivel 2
		BK_clavePuesto	Clave del puesto	Nivel 3
		puesto	Descripción del puesto o categoría el empleado	-
dim_gradoAcademico	Información sobre los grados académicos del empleado.	nivelEducativo	Nivel educativo, por ejemplo, educación básica, media superior y superior	Nivel 1
		BK_claveGrado	Clave de negocio - grado	-
		gradoAcademico	Grado académico, como licenciatura, maestría, etc…	Nivel 2

Nota. En el DM implementado para la Dirección de Recursos Humanos de la UAQ, se utilizaron varias dimensiones para analizar y comprender mejor los datos utilizados en los sistemas de pagos y timbrado de nómina; aportando una visión completa y detallada de los distintos contextos de la Universidad.

Las dimensiones, jerarquías y atributos del DW o DM definidos para la DRH de la UAQ, permitieron generar una gran cantidad de reportes y estadísticas con información de calidad que fue soporte para la toma de decisiones basadas en datos íntegros. Con base a los requerimientos establecidos en secciones anteriores, las dimensiones fueron creadas para brindar información a detalle, además, se realizaron análisis minuciosos sobre los pagos emitidos a los empleados y al proceso de integración del ISR enterado al SAT. Lo anterior fortaleció la entrega de reportes en tiempo y forma a las distintas entidades que así lo requirieron.

El siguiente paso en la definición del modelo dimensional consistió en identificar los hechos, medida y métricas que permitirían responder a la pregunta *¿qué?*, es decir, las tablas de hechos contienen datos numéricos utilizados para medir el rendimiento del algún proceso de negocio, y de esta manera, realizar el análisis de datos correspondiente a cada requerimiento de información. Dichas medidas de tipo numéricas fueron almacenadas en las tablas de hechos, las cuales son diseñadas para contener datos cuantitativos que serán analizados en el DW o DM, por ejemplo, las ventas, cantidades, ingresos, montos, ISR, calificaciones, etc… Las tablas de hecho, están conformadas por las claves foráneas de las dimensiones y las medidas.

De acuerdo a los requerimientos de información del proyecto y al modelo dimensional plasmado en la figura 32, las medidas utilizadas corresponden a cada pago emitido a los empleados de la Universidad, los timbres fiscales emitidos correspondientes a los pagos, el importe ISR enterado, las percepciones y deducciones emitidas en los recibos de pago. Las dimensiones creadas permitieron analizar un panorama completo y preciso de los requerimientos de información solicitadas por la DRH. La tabla 18 muestra las columnas, descripciones y atributos de la tabla de hechos implementadas en la investigación.

Tabla 18

Definición de la tabla de hechos utilizada en la investigación

Tabla de hechos	Descripción general	Columna/atributo	Descripción
fact_pagos	Tabla de hechos definida para el análisis del proceso de negocio de pagos de nómina y timbrado fiscal.	dim_empleado	Llave foránea referenciada a la dimensión del empleado.
		dim_ubicacion	Llave foránea referenciada a la dimensión de ubicación.
		dim_tiempo	Llave foránea referenciada a la dimensión de tiempo
		dim_financiamiento	Llave foránea referenciada a la dimensión del tipo de financiamiento.
		dim_nomina	Llave foránea referenciada a la dimensión de nómina.
		dim_categoria	Llave foránea referenciada a la dimensión de categoría del empleado.
		dim_gradoAcademico	Llave foránea referenciada a la dimensión de grado académico del empleado.
		fechaDePago	Fecha de pago de nómina.
		importeNeto	Importe neto pago al empleado.
		totalPercepciones	Importe de ingresos que el empleado recibe.
		totalDeducciones	Importe de descuentos realizados sobre el salario bruto.
		importeISR	Impuesto de ISR descontado en el recibo.
		importePropio	Clasificación interna de la UAQ
		ImporteEstatal1	Clasificación interna de la fuente de financiamiento UAQ
		importeEstatal2	Clasificación interna de la fuente de financiamiento UAQ
		importeEstatal3	Clasificación interna de la fuente de financiamiento UAQ
		importeFederal1	Clasificación interna de la fuente de financiamiento UAQ
		importeFederal2	Clasificación interna de la fuente de financiamiento UAQ
		PTC	Especifica si el momento del pago el empleado era docente de tiempo completo, el valor a almacenar es 1 ó 0
		PROMEP	Especifica si el momento del pago el empleado contaba con PROMEP, el valor a almacenar es 1 ó 0
		SNI	Especifica si el momento del pago el empleado era miembro del SNI, el valor a almacenar es 1 ó 0
		folioFiscal	El valor a almacenar es 1 ó 0 en caso de que el pago ya cuente con timbre fiscal o no.

Nota. La tabla de hechos del proyecto engloba los datos de vital importancia para la Universidad, como es el proceso de pago de la nómina. Las dimensiones y medidas asignadas permitieron realizar análisis detallados sobre los datos contenidos en las bases de datos Institucionales, esto con base a las descripciones de cada columna.

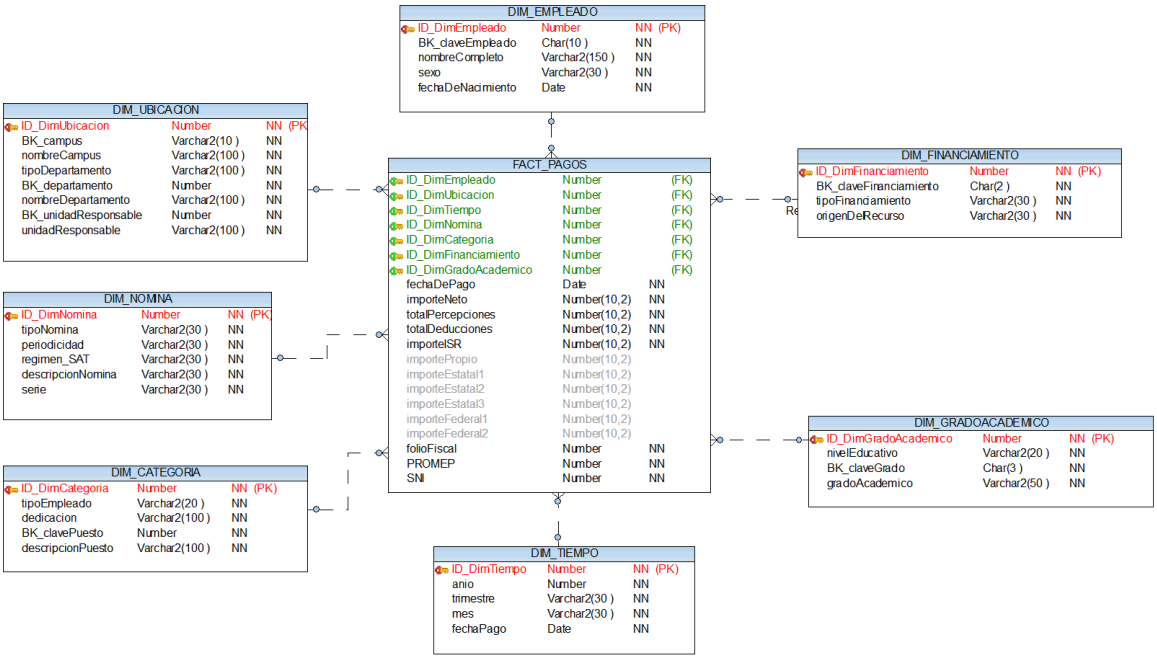
La tabla de hechos construida para el análisis sobre los pagos y timbres fiscales de la DRH, abarcó varias perspectivas y medidas que, en conjunto, no solo atendieron las peticiones solicitadas en el requerimiento inicial, sino que además enriquecieron los procesos para tomar decisiones en las áreas involucradas del proyecto. Al utilizar la tabla de hechos y las distintas dimensiones se logró identificar las normas y reglas del negocio, así como también una visión holística de los datos almacenos en los sistemas de información. En general, se pudieron responder a múltiples cuestionamientos, tales como: ¿Cuánto es el ISR pagado a nivel universidad en un tiempo determinado?, ¿Cuáles son los departamentos de mayor ISR para la Universidad?, ¿Cuál ha sido el crecimiento de nómina durante los años analizados?, ¿Cuánto se ha pagado a los distintos tipos de empleados en un tiempo determinado?, ¿Cuál ha sido el IRS enterado por departamento?, ¿Cuál es el importe faltante por timbrar en un tiempo determinado?, entre muchas otras interrogantes.

Con la definición de las tablas de dimensiones, jerarquías, niveles y tabla de hechos, se construyó el modelo dimensional del proyecto, siendo la parte esencial de cualquier solución BI, debido a la trascendencia que se tiene cuando se analiza a detalle las estructuras de datos en correspondencia con los requerimientos específicos de información solicitada por la DRH. Esto representó cambios drásticos en algunos de los aplicativos y bases de datos institucionales, además brindo la posibilidad de contar con un panorama más detallado y profundo sobre los procesos de negocio, las reglas y las áreas con las que se relacionan.

En cuanto a su representación gráfica, se emplearon modelos lógicos, herramientas que realizan representaciones detalladas de las estructuras de datos utilizadas en la construcción del DW o DM. Estos modelos indican de manera precisa los componentes de cada dimensión, la tabla hechos y sus relaciones, brindando una visión general sobre los registros almacenados. Las tablas de dimensiones y tabla de hechos definidas previamente en el proyecto de la DRH se representan en la figura 33.

Figura 33

Modelo Dimensional del proyecto



Nota. El modelo dimensional de la DRH permitió obtener información detallada de los pagos y estatus de los timbres fiscales de cada trabajador. La información se revisó a partir de distintas perspectivas, por ejemplo, categoría, grado académico, ubicación, etc…, lo cual representó un análisis de datos con mayor calidad y claridad.

5.4.1.6. Procesos ETL en el modelo dimensional

Los procesos ETL ofrecen un conjunto de herramientas y componentes que simplifican las actividades para extraer, transformar y cargar los datos en distintos repositorios o destinos. Estas herramientas permiten una gestión eficaz, segura y confiable sobre el flujo de los datos, a partir de fuentes de datos hasta el almacenamiento en el repositorio destino. Estas etapas del proceso ETL se describen a continuación:

En la etapa de extracción, brindan conectores a diversos orígenes de datos o fuentes, entre ellos se encuentran las bases de datos relacionales, las hojas de

cálculo, archivos txt, entre otros…, de igual manera se aplican filtros y transformaciones preliminares para reducir la falta de calidad de los datos, mejorando la integridad y consistencia requerida en las siguientes etapas. Las herramientas o componentes utilizados en la etapa de transformación, permiten, mediante interfaces gráficas o lenguaje de programación, realizar agregaciones, cálculos complejos, combinar múltiples fuentes de datos, filtrado de datos, conversiones de formatos, entre otras. Dichas transformaciones son definidas a partir de las reglas del negocio para garantizar la consistencias e integridad de datos. Por último, en la etapa de carga, se pueden establecer diversos tipos, como la incremental, donde se recude el número de datos por migrar y, por consecuencia, el tiempo de ejecución. Así mismo, gestiona los errores o inconsistencias detectadas durante la migración para su posterior tratamiento (Kimball y Ross, 2013).

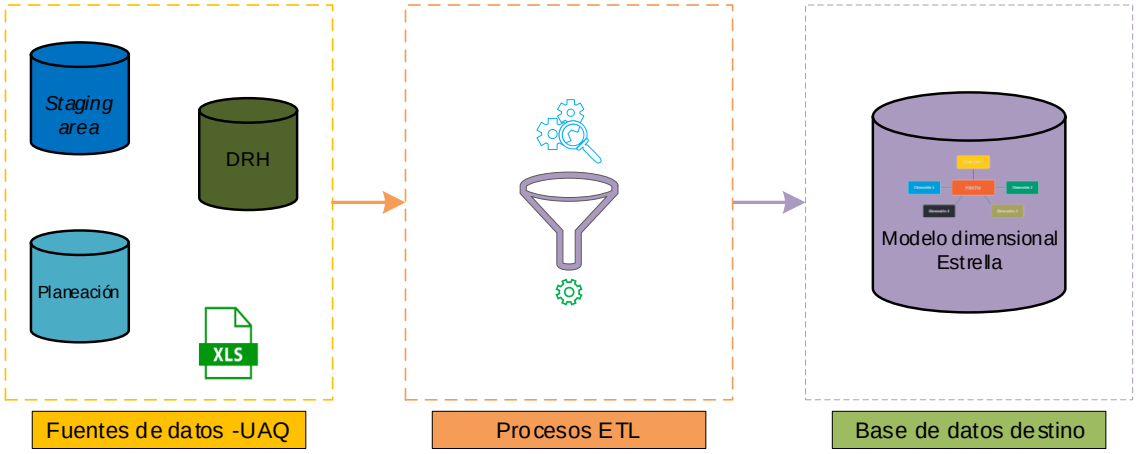
Como se menciona anteriormente, los proceso ETL pueden ser empleados para migrar la información a cualquier destino, en el contexto del proyecto se aplicaron en el *staging area* y en el modelo dimensional definido en la sección anterior. Las dimensiones y la tabla de hecho fueron pobladas utilizando las tablas del *staging area* y algunas de las bases de datos institucionales, lo cual garantizo la integridad de los datos almacenos en el repositorio y, por consecuencia, la generación de reportes y estadísticas requeridas por la DRH. De esta manera, los procesos ETL desempeñaron un papel fundamental en la migración de los datos con la calidad necesaria para el estudio, análisis y la de toma de decisiones de la Universidad. La tabla 19 muestra un fragmento del mapa de datos lógico utilizado en el proyecto, el cual facilitó la visualización de forma detallada de los orígenes de datos, su destino y algunas de las transformaciones requeridas antes de migrar los datos, en el caso particular del modelo dimensional del proyecto, los procesos ETL se emplearon para cargar las dimensiones y los hechos.

Por lo anterior, el llenado de las dimensiones y la tabla de hechos se realizó de manera eficaz y eficiente, en primer lugar, al definir la carga inicial de datos y, posteriormente, la definición del proceso de actualización. Los procesos de carga

inicial, normalmente se ejecuta en una sola ocasión o cuando existen cambios en el modelo dimensional, como la definición de nuevas dimensiones y/o métricas. El proceso implica consultar las fuentes de datos que contienen los datos históricos para actualizar las tablas de dimensiones y las tablas de hechos, después de identificar que datos serán migrados es indispensable aplicar las transformaciones de acuerdo a las reglas del negocio, de acuerdo a lo especificado en la tabla 19, así mismo, en la Figura 34 se muestran los elementos identificados durante los procesos de carga inicial del modelo dimensional de la Dirección de Recursos Humanos.

Figura 34

Carga inicial del modelo dimensional



Nota. La carga inicial o carga de valores históricos contempló diversos elementos, por ejemplo, la identificación de las fuentes de datos, la aplicación de transformaciones sobre los datos y, por último, el almacenamiento de datos en el modelo dimensional.

Tabla 19

Mapa de datos lógico del modelo dimensional

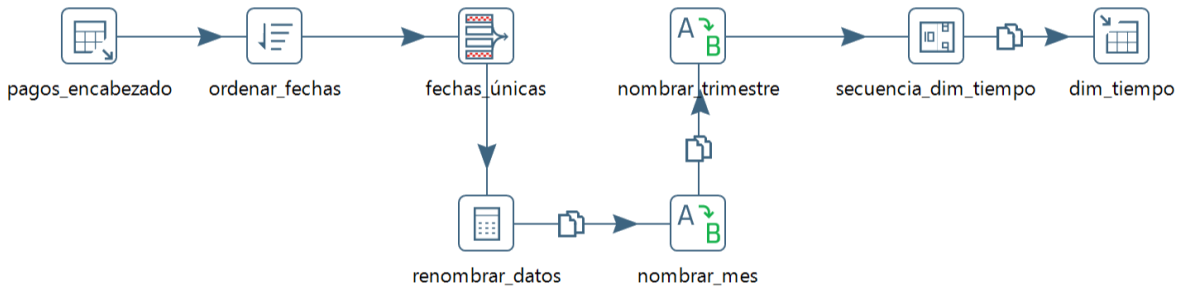
Origen (Source)				Destino (Target)			
Fuente	Columna	Tipo de dato	Dimensión / Hecho	Columna	Tipo de dato	Transformación
-	-	-		ID_DimTiempo	number	El valor se obtuvo de la secuencia: SEQ_DIM_TIEMPO
				anio	number	Se obtuvo del año de la fecha de pago a partir del 2018
				trimestre	varchar2(30)	Se definió la integración de los trimestres a partir del 2018
pagos_encabezado	fechaPago	date	dim_tiempo	mes	varchar2(30)	Nombre del mes a partir del 2018
				fecha	date	Se obtuvo de la fecha de pago a partir del 2018
dim_empleado	ID_DimEmpleado	number		ID_DimEmpleado	number	-
dim_financiamiento	ID_DimFinanciamiento	number		ID_DimFinanciamiento	number	-
dim_ubicacion	ID_DimUbicacion	number		ID_DimUbicacion	number	-
dim_tiempo	ID_DimTiempo	number		ID_DimTiempo	number	-
dim_nomina	ID_DimNomina	number		ID_DimNomina	number	-
dim_gradoAcademico	ID_DimGradoAcademico	number		ID_DimGradoAcademico	number	-
dim_categoria	ID_DimCategoria	number		ID_DimCategoria	number	-
	fechaPago	date		fechaDePago	date	-
	importeNeto	number(10,2)		importeNeto	number(10,2)	-
	importeTotalPercepciones	number(10,2)		totalPercepciones	number(10,2)	-
	importeTotalDeducciones	number(10,2)		totalDeducciones	number(10,2)	-
	importeISR	number(10,2)		importeISR	number(10,2)	-
	importe_propioSAT	number(10,2)	fact_pagos	importePropio_SAT	number(10,2)	Los importes pueden ser NULL, dependiendo del origen del recurso
	importe_federalSAT	number(10,2)		importeFederal_SAT	number(10,2)	asignado a cada pago.
pagos_encabezado	importe_propio	number(10,2)		importePropio_UAQ	number(10,2)	
	importeEstatal_1	number(10,2)		importeEstatal1_UAQ	number(10,2)	
	importeEstatal_2	number(10,2)		importeEstatal2_UAQ	number(10,2)	
	importeEstatal_3	number(10,2)		importeEstatal3_UAQ	number(10,2)	
	importePropioFacultad	number(10,2)		importePropio_Facultad	number(10,2)	
	importeFederal1	number(10,2)		importeFederal1_UAQ	number(10,2)	
	importeFederal2	number(10,2)		importeFederal2_UAQ	number(10,2)	
	estatusTimbre	varchar2(30)		folioFiscal	number	De acuerdo al rango de fechas el valor a almacenar es 1 ó 0
	fk_tabulador	number		PTC	number	De acuerdo al rango de fechas el valor a almacenar es 1 ó 0
	fecha_ini, fecha_fin	number		PRODEP	number	De acuerdo al rango de fechas el valor a almacenar es 1 ó 0
acad_promep	fecha_ini, fecha_fin	number		SNI	number	De acuerdo al rango de fechas el valor a almacenar es 1 ó 0

Nota. El modelo dimensional definido para el Proyecto de la DRH contempla varias dimensiones como tiempo, empleado, nomina, ubicación, entre otras, para lograr obtener distintas perspectivas de las medidas solicitadas. El mapa de datos lógico permitió localizar de forma clara y sencilla las fuentes de datos y su destino.

La carga inicial de los datos depende de las necesidades de información de la organización, en el caso del proyecto, se identificaron las fuentes de datos en las bases de datos institucionales: las bases de datos de la DRH y los esquemas de la Dirección de Planeación. Además, gran parte de la información se localizó en el *staging area*, mientras que algunos catálogos se encontraron únicamente en hojas de cálculo. En la migración de los datos al modelo dimensional, se utilizaron consultas directas a las bases de datos institucionales y la herramienta *pentaho data integration*, por ejemplo, la figura 35 presenta los elementos que fueron seleccionados para actualizar la dimensión de tiempo.

Figura 35

Proceso ETL para dimensión de tiempo



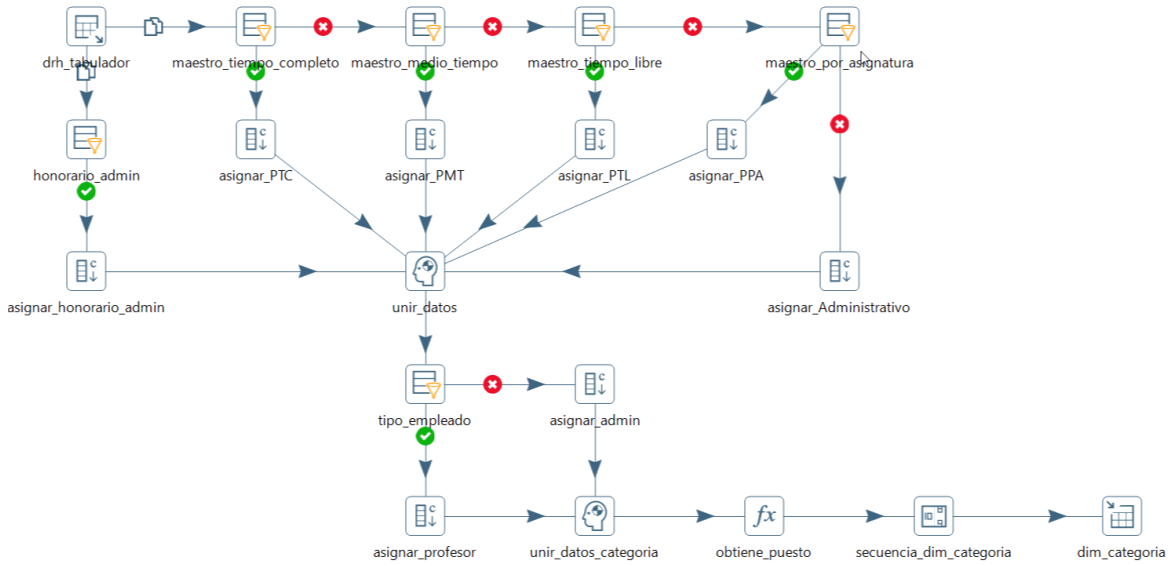
Nota. El uso de *Pentaho Data Integration* facilitó la migración de una base de datos operacional a un modelo dimensional. La dimensión tiempo requirió consultar las fechas de pago de cada recibo generado en al DRH para poder definir los distintos niveles, por ejemplo, año, trimestre, mes y la fecha de pago.

Por otra parte, en la dimensión de “categoría” fue necesario un enfoque especial debido a sus diferentes niveles y la cantidad de fuentes de datos asociadas. Esta dimensión contenía la clasificación del personal en categorías, por ejemplo, el personal administrativo y el personal docente. Específicamente, la categoría de docente presentaba subniveles que incluían docentes de tiempo completo, medio tiempo, tiempo libre y por asignatura. Para garantizar una migración eficiente de los datos se ejecutaron los procesos ETL, de acuerdo a la Figura 36. Este proceso

aseguró que los datos se transfirieran y transformaran adecuadamente, permitiendo una gestión eficiente de la dimensión de categoría en el modelo dimensional.

Figura 36

Proceso ETL para dimensión de categoría



Nota. La dimensión categoría permitió categorizar a los empleados de acuerdo a la actividad que desempeñaba cada persona y el tipo de contratación con el que contaba. El proceso ETL dio inicio con la fuente de datos, posteriormente se clasificó al personal, se asignaron puestos y las secuencias para el llenado de la dimensión.

Al finalizar la carga inicial de cada una de las dimensiones en el *data warehouse*, el siguiente paso consistió en actualizar la tabla de hechos. Esta tabla se consideró fundamental en el modelo dimensional, ya que almacenaba los valores cuantitativos, métricas y/o medidas utilizadas en el análisis, la generación de reportes y estadísticas. Dichas métricas se definieron con base en los requerimientos del negocio y tuvieron como objetivo principal proporcionar información clave para la elaboración de reportes y estadísticas. Las tablas de hechos mantienen una relación con las dimensiones por medio de las claves foráneas correspondientes, las cuales se trataban de identificadores únicos para cada dimensión y se utilizaron para establecer las relaciones y vínculos entre los

datos de la tabla de hechos y las tablas de dimensiones asociadas. La figura 37 muestra parte de la consulta utilizada para realizar la migración de la información, asimismo, la figura 38 ilustra el proceso ETL utilizado en la migración de datos a la tabla de hechos.

Figura 37

Consulta SQL para obtener datos de los distintos orígenes

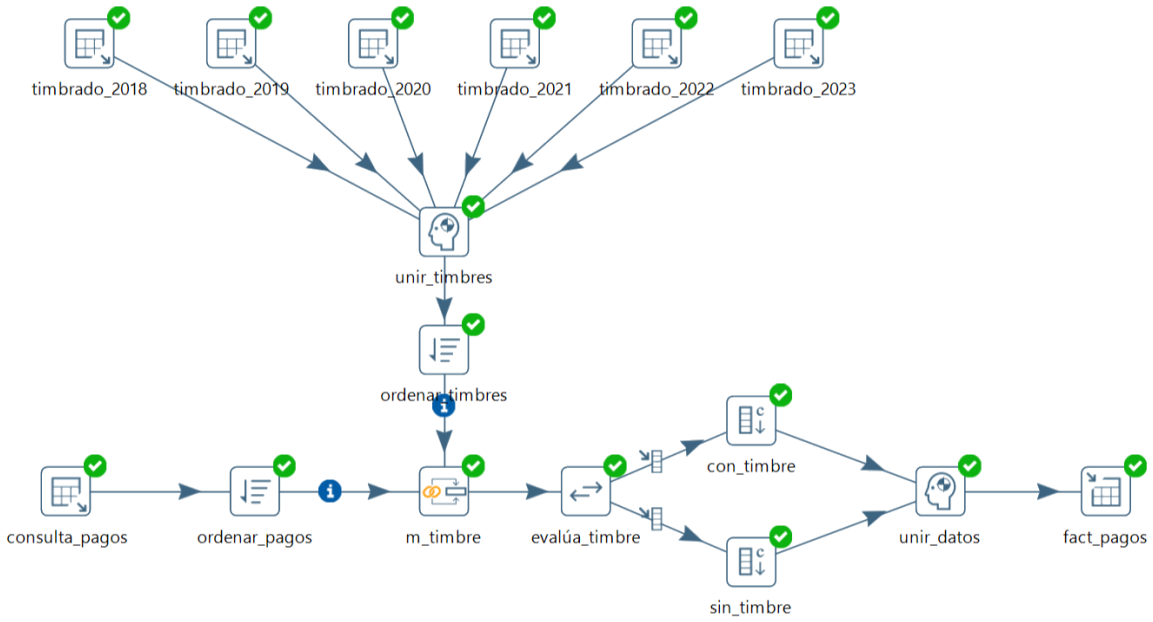
```
SELECT
  de.ID_Dim_Empleado, du.ID_Dim_Ubicacion, dt.ID_Dim_Tiempo, dm.ID_Dim_Nomina
, dc.ID_Dim_Categoria, df.ID_Dim_Financiamiento, dg.ID_Dim_GradoAcademico, t.fechaPago
, t.importeNeto, t.importeTotalDeducciones, t.importeTotalPercepciones, t.importeISR
, t.importePropio, t. importeEstatal1, t.importeEstatal2, t.importeEstatal3
, t.importeFederal1, t.importeFederal2
FROM staging.pagos_encabezado t
INNER JOIN drh.trabajadores p ON t.claveEmpleado = p.claveEmpleado
INNER JOIN dw.dim_empleado de ON t.claveEmpleado = de.claveEmpleado
AND p.claveEmpleado = de.claveEmpleado
INNER JOIN staging.c_unidadresponsable ur ON ur.clave_UAQ = t.FK_unidadResponsable
INNER JOIN staging.c_departamento d ON ur.FK_departamento = d.ID_departamento
INNER JOIN staging.c_campus c ON c.ID_Campus = ur.FK_campus
INNER JOIN dw.dim_ubicacion du ON du.bk_campus = c.ID_Campus
          AND du.bk_departamento = d.id_departamento
          AND du.bk_unidad_responsable = ur.id_unidadresponsable
INNER JOIN dw.dim_tiempo dt ON dt.fecha_pago = t.fechaPago
INNER JOIN dw.dim_nomina dm ON dm.serie = t.FK_nomina
INNER JOIN drh.tabulador tab ON NVL(t.FK_Puesto,1) = tab.clavePuesto
INNER JOIN dw.dim_categoria dc ON dc.clavepuesto = NVL(t.FK_Puesto,1)
                  AND DECODE(
                           (CASE
                                    WHEN NVL(t.FK_Puesto,1) BETWEEN 1 AND 10 THEN 'D'
                                    WHEN NVL(t.FK_Puesto,1) BETWEEN 30 AND 40 THEN 'D'
                                ELSE 'A'
                          END),'D','Profesor','Administrativo') = dc.tipoempleado
INNER JOIN dw.dim_financiamiento df ON  (CASE
                                    WHEN (NVL(t.importePropio,0) > 0
                                             OR NVL(t.importeEstatal1,0) > 0
                                             OR NVL(t.importeEstatal2,0) > 0
                                             OR NVL(t.importeEstatal3,0) > 0)
                                             AND NVL(t.importeFederal1,0) = 0
                                             AND NVL(t.importeFederal2,0) = 0 THEN 'IP'
                                    WHEN (NVL(t.importePropio,0) = 0
                                             AND NVL(t.importeEstatal1,0) = 0
                                             AND NVL(t.importeEstatal2,0) = 0
                                             AND NVL(t.importeEstatal3,0) = 0)
                                             AND (NVL(t.importeFederal1,0) > 0
                                             OR NVL(t.importeFederal2,0) > 0) THEN 'IF'
                                    ELSE 'IM'
                                             END) = df.bk_ClaveFinanciamiento
```

```
INNER JOIN dw.dim_gradoacademico dg
ON dg.bk_clavegrado = NVL(p.gradoacademico,'NA')
WHERE TO_CHAR(t.fechaPago,'YYYY-MM-DD')
BETWEEN '2018-01-01' AND '2023-06-30';
```

Nota. Alimentar la tabla de hechos requirió la unión de distintas tablas del negocio y las dimensiones creadas previamente. La consulta SQL muestra una parte del proceso realizado para migrar los datos. En esta consulta, se aplicaron algunos filtros y se realizaron cálculos con el propósito de garantizar la integridad de los datos, por ejemplo, el tratamiento de valores nulos, la asignación de valores por defecto, entre otros.

Figura 38

Proceso ETL para actualizar la tabla de hechos



Nota. Los procesos ETL utilizados en la tabla de hechos requirió la unión de las dimensiones y las tablas transaccionales del negocio o el *staging area*. Se recomendó utilizar consultas SQL como entrada de datos y realizar los cálculos y transformaciones dentro del *Pentaho Data Integration*. La métrica de timbrado se obtuvo desde diferentes orígenes de datos, como se describe en secciones previas.

En resumen, para el proceso de carga inicial del DW se realizó una revisión detallada y completa de las fuentes de datos, dicha revisión se hizo con el propósito de fortalecer las características sobre la calidad los datos. De igual manera, se evitó cargar valores nulos, anómalos o que presentarán algún tipo de inconsistencia en el origen de datos (*data source*), lo cual implicó aplicar condiciones y restricciones definidas en los requerimientos. Asimismo, se migraron los registros en las tablas de dimensiones y en la tabla de hechos. Para realizar la carga inicial se utilizó la herramienta *Pentaho Data Integration* y se aplicaron varias consultas *SQL*, tanto en las dimensiones como en la tabla de hecho.

En cuanto a los pasos de actualización periódica de los datos se definieron un conjunto de reglas, políticas y estratégicas con apoyo del personal involucrado, de igual manera, se utilizaron consultas *SQL* y el software de *Pentaho Data Integration*. En seguida se describen las acciones consideradas para el proceso:

- La migración de los datos correspondientes al pago de nóminas se recomendó realizarlo con máximo 36 horas después de la emisión del pago al empleado, dependiendo de la periodicidad de la nómina.
- Los reportes de inconsistencias de datos se enviaron después de cada carga al DW, para su resolución parcial o total, dependiendo de la problemática.
- Las aplicaciones que recabaron los datos iniciales se actualizaron, agregando nuevas validaciones para fortalecer la calidad de los datos.
- El timbrado de nómina se realizó máximo 72 horas después de la emisión del pago.
- El proceso de carga de datos al DW, se ejecutó entre las 01:00 y 03:00 horas para no afectar las operaciones y transacciones realizadas por los usuarios dentro de la base de datos.

5.4.1.7. Control de inconsistencias en los datos del DW

La implementación exitosa del *Data Warehouse,* dependió en gran medida del nivel de calidad e integridad de los datos, debido a que el DW almancena la historia del proceso de timbrado y pagos de la DRH considerando aquellos datos esenciales

para la emisión de reportes y estadísticas. En la sección 2.2.4.2, se describieron las características principales para identificar el grado de calidad en los datos. Al aplicar los principios de calidad de datos en el proceso de negocio, del cual se llevó a cabo el DW, surgieron distintas problemáticas a lo largo del desarrollo del proyecto, en consecuencia, se generaron las siguientes observaciones:

- Campus faltantes en los catálogos y en las tablas transaccionales: Al realizar consultas y análisis de los registros de nóminas por campus, se identificaron campus en los datos transaccionales que no se encontraban registrados en el catálogo de campus. Lo cual demostró que existían inconsistencias en la integridad referencial.
- RFC incorrectos, nulos, o duplicados: En la revisión del RFC, se encontraron casos que no cumplían con el formato requerido, al ser un dato indispensable para la emisión del timbre fiscal fue necesario revisar los expedientes físicos de aquellos empleados que no se logró contactar por otro medio.
- CURP incorrectos, nulos, o duplicados: En la revisión del CURP, se encontraron casos que no cumplían con el formato requerido. Dado que este dato es indispensable para la emisión del timbre fiscal, se optó por revisar los expedientes físicos de aquellos empleados que no se logró contactar por otro medio.
- Unidades responsables nulas en los registros timbrados: Está inconsistencia en los registros dificultó la identificación oportuna de la entidad o área donde laboraba físicamente el empleado, generando problemas en los informes y reportes.
- Claves de puestos con valor nulo: La inconsistencia se presentó únicamente con el personal administrativo de honorarios, lo cual permitió la corrección y asignación manual de una clave de puesto.
- Tipo de empleado (docente o administrativo) nulos o inconsistentes: Afectaba la categoría y clasificación adecuada de los empleados.
- Tipos de incapacidades con valores incorrectos: Durante el proceso, se identificó la falta de un registro oportuno de las incapacidades de los

empleados. Para abordar este problema, se agregaron las validaciones adecuadas en los sistemas de información correspondientes.

- Registros de pagos incompletos o en hojas de cálculo: La emisión de cheques por cualquier circunstancia se consideraba como requerido en la DRH, sin embargo, fue necesario automatizar el proceso captura de los datos.
- Errores en los datos al realizar migraciones entre sistemas: Algunas de las nóminas requirieron ser migrar para centralizar los datos y realizar los registros contables-presupuestales, sin embargo, se encontraron valores truncados, campos vacíos o con formato incorrecto.

Para elevar las características de integridad de los datos almacenados o durante su captura, se establecieron una serie de mejores prácticas, por ejemplo, la implementación de un control efectivo sobre las inconsistencias de datos en la DRH, específicamente en la emisión de pagos de nómina y timbrado fiscal. Estas mejores prácticas, diseñadas de acuerdo a las reglas del negocio de la Universidad, se focalizaron en la optimización constante de los procesos administrativos y técnicos. Las siguientes estrategias muestran algunos ejemplos que se adoptaron al realizar el proyecto de investigación:

- Implementar reglas y políticas de validación: Para reforzar la calidad y precisión de los datos en la DRH, se llevaron a cabo diversas comprobaciones de integridad referencial, que aseguraban mantener las relaciones entre los datos. Además, se llevaron a cabo análisis sobre conjuntos de datos muestra, utilizado consultas SQL, para verificar los cálculos realizados. Asimismo, se aplicaron validación y estandarización de formatos de fechas y rangos de valores para los campos que así lo requirieron.
- Desarrollar controles automatizados: Se actualizaron varios sistemas de información, por ejemplo, se aplicaron controles de inconsistencia agregando validaciones en los formularios de registro para prevenir y detectar las inserciones y/o actualizaciones de datos incorrectos. Por ejemplo, se

incluyeron campos obligatorios para garantizar que se capture los datos necesarios. También se aplicaron verificaciones de formato para validar campos como el RFC y CURP. Además, se verificaron rangos de fechas y claves de campos con base a catálogos. Las validaciones incluyeron mensajes de error claros y descriptivos para que el usuario final contará con funcionalidades técnicas y poder realizar la corrección de forma oportuna. Algunos ejemplos de los sistemas de información actualizados en esta etapa son: el expediente digital del empleado, cálculo de nóminas, timbrado de nóminas, constancia SAT y contratos.

- Comparación de las fuentes de datos: Al contar con diversos orígenes de datos para el pago y timbrado de nómina fue necesario hacer validaciones que permitieran conocer la integridad de los datos. Por ejemplo, comparar los importes totales de ISR, valores nulos entre cada origen de datos y buscar discrepancias en las fechas. De igual manera se realizaron varias consultas SQL, vistas materializadas y la creación del *staging area* para disminuir posibles inconsistencias entre los orígenes de datos.
- Monitoreo continúo del proceso de nómina y timbrado: Se definieron reportes mensuales para corroborar que los datos almacenados en el DW y las bases de datos transaccionales contarán con la calidad necesaria. En este sentido, se agregaron reportes a las aplicaciones de timbrado que permitieron visualizar la información faltante y hacer comparativos de los importes totales.
- Capacitación y concientización al personal operativo: Se aplicaron instrumentos que facilitaron la identificación de la problemática en la calidad de los datos, detectando que el 85% pertenece a la inserción de registros de forma incorrecta, por ello, con las actualizaciones realizadas en las aplicaciones disminuyo considerablemente el porcentaje, además, se capacitó constantemente a los usuarios operativos y se explicó la importancia de reportar cualquier incidencias de las aplicaciones para poder corregir y disminuir la entrada incorrecta de datos.

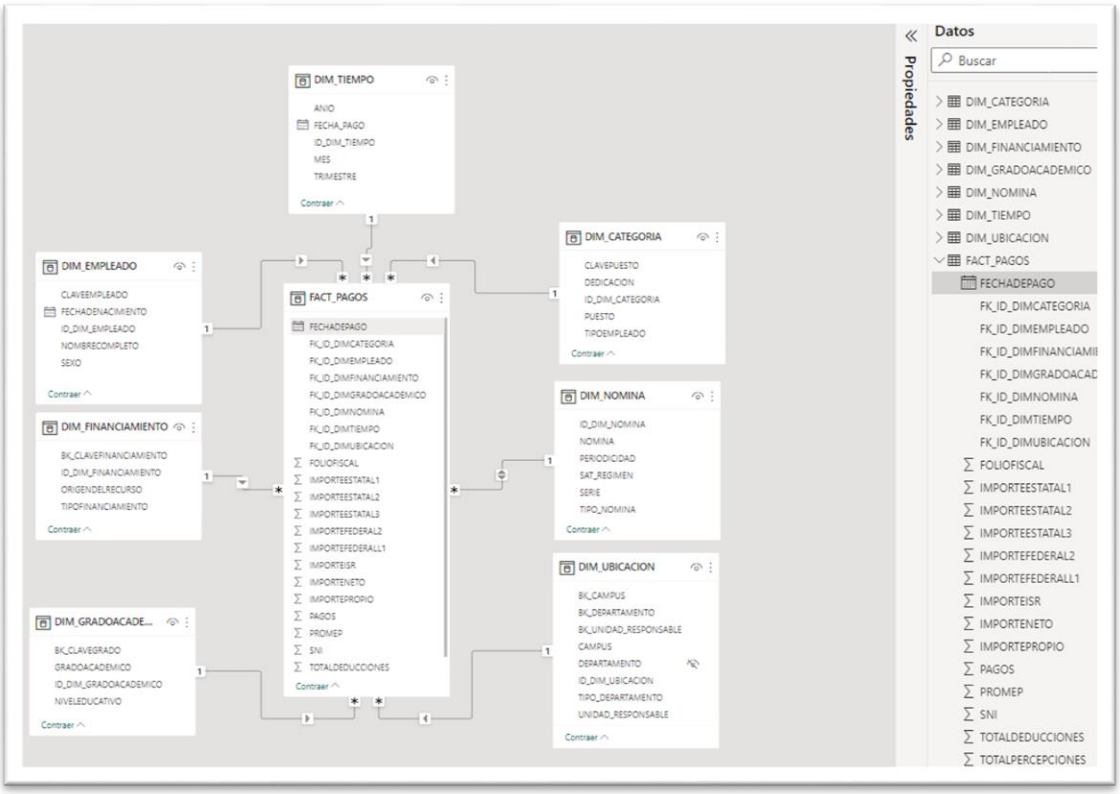
5.4.1.8. Explotación de los datos

La etapa o fase final de la metodología para el fortalecimiento en la generación de reportes y estadísticas de la UAQ, consistió en la capa de presentación o visualización de datos utilizando tablas, filtros de información, gráficas de barras, graficas de líneas, gráfico circular, entre otros. En esta etapa se pueden utilizar herramientas diseñadas para la explotación de modelos dimensionales. Por ejemplo, *Tableau*, *Microsoft Power BI*, *Python*, *QlikView*, etc. Dichas aplicaciones son llamadas herramientas para el análisis de los datos utilizadas para fortalecer la emisión de reportes, estadísticas apoyando la de toma de decisiones a nivel empresarial. En el proyecto se desarrollaron elementos con apoyo de la implementación de *Microsoft Power BI*, sin embargo, los modelos dimensionales pueden ser explotados por cualquier otra herramienta especializada.

Los reportes definidos en la *sección 1.1.1.1. Revisión de reportes y estadísticas: Reportes del área administrativa*, se realizaron con apoyo de *Microsoft Power BI*, se obtuvieron resultados óptimos para su consulta, elaboración y entrega a los usuarios finales. *Microsoft Power BI*, ofreció una serie de opciones de visualización de los datos, permitió la conexión al modelo dimensional y la aplicación de filtros de información. En la sección de “datos” y la sección “Vista Modelo” se mostraron las dimensiones y tabla hechos, así mismo, las jerarquías y configuración general sobre los campos. La figura 39 indica la estructura y el diagrama conceptual del proyecto.

Figura 39

Vista modelo del Data Warehouse



Nota. El modelo conceptual del proyecto se mostró con la tabla de hechos al centro rodeada de las dimisiones definidas y su estructura para cada columna.

A continuación, se describen e ilustran algunos de los reportes diseñados. Por políticas y normas de seguridad de la información y las leyes de protección de datos personales, los datos presentados son únicamente de ejemplo y no representan la información real de la Universidad. Uno de los reportes solicitados, fue el total de los XML emitidos (vigentes) y el total de pagos por periodo y tipo de nómina, siendo estos valores los almacenados en las tablas de dimensiones y tabla de hechos descritas en secciones anteriores, la Figura 40 muestra la tabla de resultados obtenidos a lo largo del tiempo.

Figura 40

Reporte de pagos vs timbres por año

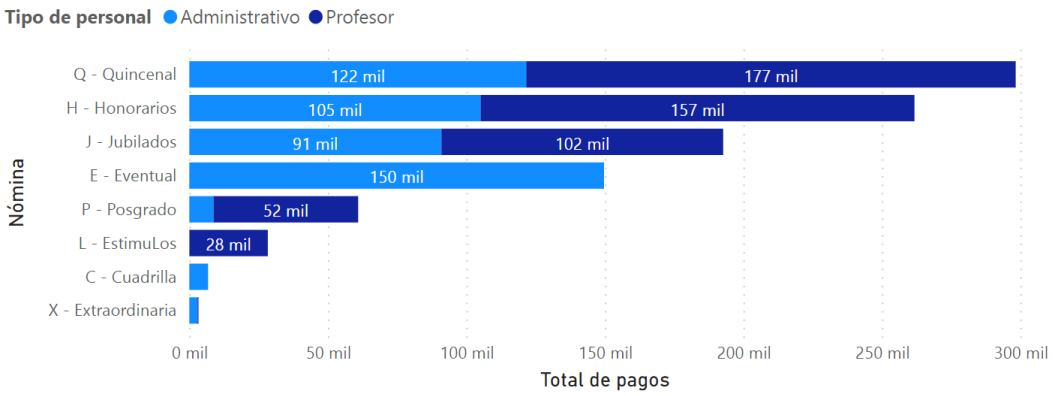
Año	Total de pagos	Total de timbres	Diferencia	Tipo de nómina
2018	157.721	119.378	38.343	Ordinaria
2019	173.968	133.589	40.379	Ordinaria
2020	183.902	141.048	42.854	Ordinaria
2021	191.961	149.547	42.414	Ordinaria
2022	204.032	159.266	44.766	Ordinaria
2023	92.311	67.430	24.881	Ordinaria
Total	**1.003.895**	**770.258**	**233.637**	

Nota. Las visualizaciones de tipo tabla, permitieron generar reportes de manera sencilla y ágil, en el reporte se observan el total de pagos emitidos y el total de timbres por nómina y año de pago.

Es importante considerar la forma de presentar la información contenida en el modelo dimensional a los usuarios finales, debido a que el análisis de reportes y estadísticas se basa en la comprensión oportuna, clara y concisa de los datos. En este sentido, se evalúo el total de datos a presentar y que usuarios accederían a ella, en el caso de los informes o reportes con múltiples datos, se recomendó utilizar visualizaciones que agruparan los datos relevantes en gráficas o resúmenes ejecutivos, lo que permitiría identificar patrones o tendencias de manera eficiente. Así mismo, algunos usuarios solicitaban acceder a la información de manera detallada, en respuesta se sugirió mostrar tablas y filtros de información para una comprensión eficaz y eficiente. En general, la presentación de los datos se realizó con base al perfil del usuario, por ejemplo, para aquellos que requerían una visión general rápida, se diseñaron visualizaciones intuitivas. La figura 41 muestra una visualización con los totales de pagos emitidos por tipo de personal, utilizando una gráfica de barras.

Figura 41

Reporte de total de pagos por tipo de nómina



Nota. La gráfica de barras simple, permitió representar gran cantidad de información de manera visible ofreciendo al usuario final un panorama general de los datos. Se muestra lo pagos totales por tipo de personal de la DRH.

Los elementos expuestos en las figuras anteriores presentaron los datos de manera estática y limitada por un diseño predefinido. Sin embargo, existen otras formas de visualizaciones que permitieron un análisis profundo de los datos, presentado una visión general de los datos con opción de descender o ascender entre los niveles jerárquicos establecidos en el modelo dimensional. Este proceso es conocido como *drill down* y *drill up*, y permitió navegar en la información desde lo general a lo particular, además brindaron mayor interactividad y flexibilidad para explorar la información con mejor comprensión para los directivos.

El elemento principal utilizado en los procesos *drill down* y *drill up*, son las jerarquías y sus niveles, los cuales se definieron al momento de diseñar las dimensiones y el nivel de agregación del proyecto, a continuación, se enlistan algunas jerarquías definidas en las dimensiones:

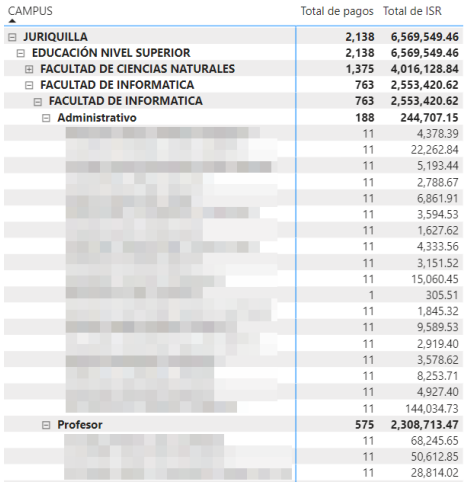
- fecha de pago: se utilizó para obtener con los niveles de año, que a su vez fue dividido en trimestres, en meses y, en caso de ser necesario por día,
- tipo de empleado: con los niveles de tipo de empleado y dedicación,
- nivel académico: con los niveles de nivel educativo y grado académico,

- nóminas: con los niveles de tipo de nómina, nómina y serie,
- centro de trabajo: con los niveles de campus, tipo departamento, unidad responsable y departamento.

Las jerarquías se visualizaron de dos formas distintas: la primera, en forma de gráfica y la segunda, en una tabla de datos. La elección entre ambas dependió de la cantidad de información a analizar y los niveles involucrados. La figura 42 ejemplifica los niveles de agregación de los datos en la dimensión de centro de trabajo por medio de una tabla o matriz de datos.

Figura 42

Reporte del total de pagos e ISR por medio de jerarquías



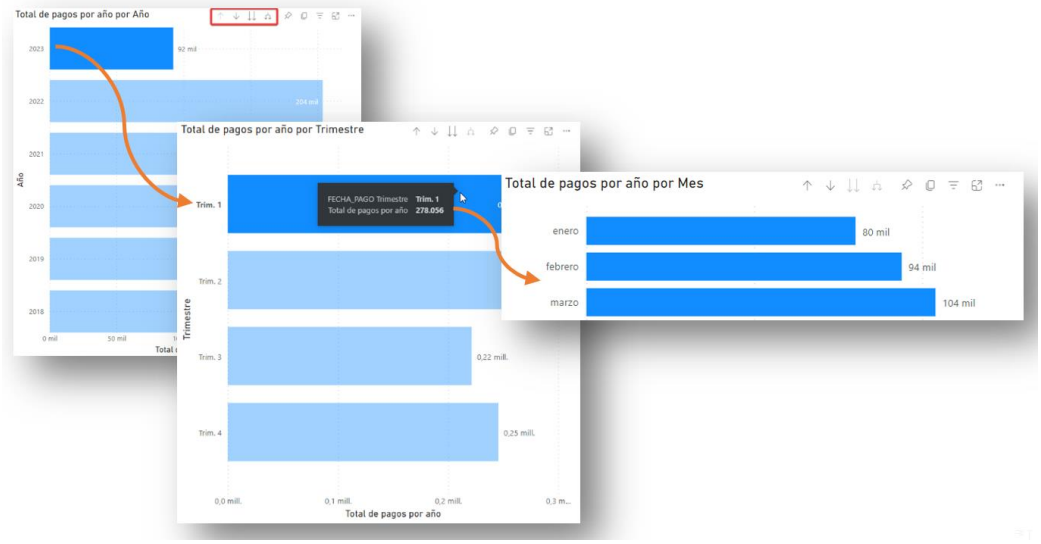
CAMPUS	Total de pagos	Total de ISR
JURIQUILLA	2,138	6,569,549.46
EDUCACIÓN NIVEL SUPERIOR	2,138	6,569,549.46
FACULTAD DE CIENCIAS NATURALES	1,375	4,016,128.84
FACULTAD DE INFORMATICA	763	2,553,420.62
FACULTAD DE INFORMATICA	763	2,553,420.62
Administrativo	188	244,707.15
[redacted]	11	4,378.39
[redacted]	11	22,262.84
[redacted]	11	5,193.44
[redacted]	11	2,788.67
[redacted]	11	6,861.91
[redacted]	11	3,594.53
[redacted]	11	1,627.62
[redacted]	11	4,333.56
[redacted]	11	3,151.52
[redacted]	11	15,060.45
[redacted]	1	305.51
[redacted]	11	1,845.32
[redacted]	11	9,589.53
[redacted]	11	2,919.40
[redacted]	11	3,578.62
[redacted]	11	8,253.71
[redacted]	11	4,927.40
[redacted]	11	144,034.73
Profesor	575	2,308,713.47
[redacted]	11	68,245.65
[redacted]	11	50,612.85
[redacted]	11	28,814.02

Nota. La tabla muestra el total de pagos e ISR utilizando una jerarquía que inicia con el campus del empleado, el tipo de departamento, unidad responsable y departamento donde se encontraba físicamente el empleado. Así mismo, se agregó el nivel de tipo de empleado y se desgloso el listado de empleados.

La figura anterior representaba un conjunto de datos con un nivel de agregación alto y utilizaba distintas dimensiones, por lo cual, fue necesario desarrollar una tabla de datos que permitiera realizar las funciones de *drill down* y *drill up*. Por otra parte, la figura 43 muestra una gráfica con las mismas funcionalidades aplicado en la jerarquía de fechas de pago.

Figura 43

Reporte de total de pagos con nivel de agregación de tiempo



Nota. El tipo de presentación utilizado en el reporte ofreció una visión rápida y detallada de los pagos realizados por año, trimestre y mes; los cuales formaban parte de la jerarquía de fechas de pago. La información de cada nivel ascendente o descendente se pudo realizar por medio de las opciones enmarcadas en color naranja.

La explotación de los datos dependió en gran medida del tipo de reporte que el usuario final deseaba visualizar en el *dashboard* y la cantidad de datos o agrupaciones realizadas. En este sentido, al contar con el modelo dimensional fue posible realizar una amplia variedad de reportes y estadísticas que permitieron contar con información oportuna e integra para su análisis. Así mismo, se proporcionaron elementos y estructuras de datos optimizadas que redujeron en gran medida el tiempo en la elaboración y procesamiento de la información, por medio de jerarquías, niveles de agregación, agrupaciones y medidas.

6. Resultados y discusión

La combinación sobre las tecnologías de la información, el software especializado y los procesos administrativos (personas, recursos, actividades, medición y control), fomenta el desarrollo de sistemas de información cada vez más robustos, tanto en el nivel operativo, táctico y estratégico. Los tres niveles organizacionales impactan de manera significativa en el crecimiento y desarrollo del área administrativa y académica de la Universidad, por lo tanto, es esencial mantener una formación continua sobre los procesos y su interrelación con los sistemas de información, los cuales se desarrollan e implementan de acuerdo a los requerimientos específicos de la Universidad.

El personal del nivel operativo utiliza los sistemas de información de tipo operacional, es decir, los sistemas donde se realizan las actividades o tareas cotidianas que requiere la Universidad para su operación diaria. Estas actividades incluyen el registro de estudiantes, el alta de materias, la gestión de horarios, la evaluación docente, la gestión de becas, el registro de calificaciones, los proyectos de investigación, las tutorías, el pago de nómina, el timbrado de nómina, los registros contables y presupuestales, el control de activos y la gestión de compras. Para el funcionamiento adecuado se requiere que el personal cuente con la capacitación y el conocimiento necesario para llevar a cabo sus funciones, así mismo, al tratarse de sistemas a nivel transaccional, el personal debe tener pleno conocimiento de los datos que ingresan en los formularios de registro para asegurar un alta calidad e integridad en los datos; de igual manera, en apoyo a la correcta captura de los datos, los sistemas deben considerar las validaciones necesarias en cada formulario de registro; por último, los objetos de las bases de datos requieren contar con la implementación de las reglas de normalización para proteger los datos que serán ingresados.

El personal que desempeña sus labores en el nivel táctico, utilizan sistemas de información que les brindan reportes e informes donde evalúan el desempeño de sus departamentos y facultades o escuelas de bachilleres. Estos informes o reportes son elaborados con base a los datos capturados por el personal del nivel

operativo y ofrecen una visión general de los objetivos planteados por el área administrativa. El personal del nivel táctico es responsable de la capacitación y fortalecimiento de su área en cuanto al manejo adecuado de los datos ingresados en los sistemas de información, debido al impacto que pueden traer en los tres niveles organizacionales.

Por último, en el nivel estratégico se encuentra el personal de alta gerencia, rectoría y secretarios. Los sistemas de información implementados en este nivel son fundamentales para el análisis de los datos, representados mediante reportes y estadísticas que dan soporte al proceso para tomar decisiones certeras, apoyando la ejecución de los objetivos estratégicos, así como la planificación a mediano y largo plazo. Estos sistemas permiten recopilar, depurar, estandarizar y analizar la información proveniente de los demás sistemas transaccionales, con el objetivo de realizar análisis detallados y, con base en ello, diseñar planes estratégicos de acuerdo al contexto externo e interno de la Universidad para anticipar cambios y mitigar los riesgos. El impacto de las decisiones tomadas en este nivel afecta a toda la organización, por ello, los informes, reportes, cuadros de mando e indicadores deben ser entregados de forma oportuna, precisa, clara y con la calidad necesaria.

Como se puede observar los tres niveles están relacionados y se basan los datos capturados en el nivel operativo, lo cual fue parte de la problemática detectada al aplicar el instrumento sobre los factores técnicos que influyen en la elaboración de reportes y estadísticas de la Universidad. En general, cada uno de los resultados demuestran inconsistencias en los datos contenidos en los sistemas de información, como la duplicidad de registros, las reglas de negocio mal aplicadas, los errores en la captura de datos, la ausencia de validaciones en los sistemas de información del nivel operacional, una capacitación y concientización al usuario de manera poco efectiva. Además, las distintas áreas de la Universidad no contaban con herramientas para generar reportes y estadísticas eficientemente, obteniendo reportes directamente de las bases de datos transacciones. De igual manera, no se tenía una visión global de los datos o un repositorio especifico que garantice la integridad de los datos, por el contrario, existía una fragmentación de los sistemas

de información y en consecuencia de los datos, lo que se traduce en versiones distintas de la verdad, obteniendo resultados diferentes dependiendo del departamento que elaborara el informe o reporte. Así mismo, la necesidad de obtención de reportes bajo estas circunstancias ocasionaba la elaboración manual por medio de varias herramientas que resultaban en procesos lentos, duplicaban los esfuerzos y ocasionaban errores y falta de confiabilidad en ellos.

Las soluciones o proyectos de BI contribuyen significativamente a la mejora de los procesos institucionales involucrados en el proyecto, debido a que el personal de diferentes niveles de la organización realizó actividades de mejora continua en sus respectivas áreas. Por lo cual, el uso de estos proyectos debe formar parte de las estrategias de la Universidad, para optimizar los recursos, monitorear los objetivos y fortalecer el proceso de toma de decisiones por medio del acceso a reportes, estadísticas e informes que muestren el estatus actual del proceso analizado. Con apoyo de estos conceptos se logra un trabajo colaborativo entre los distintos niveles, corrigiendo de fondo los factores que influyen en la integridad de los datos contenidos en las bases de datos transaccionales y dimensionales.

Por lo anterior, se desarrolló una metodología para fortalecer la elaboración de reportes y estadísticas con apoyo de un proyecto *Data Warehouse*, implementado en la DITI. La metodología consta de varias fases que abarcan la definición de requerimientos de las áreas involucradas, la revisión de reportes y estadísticas que generan actualmente, el análisis de las estructuras de datos origen, la configuración de catálogos internos y externos para homologar los datos, la definición del *staging area*, el diseño del modelo dimensional a través de un DW, la definición de los procesos ETL, la explotación de los datos, el control y tratamiento de inconsistencias de datos encontrados.

Durante la implementación de la presente metodología, basada en la visión de *Kimball* e *Inmon*, se cumplieron los objetivos específicos y generales, además se dio respuesta a las necesidades de información solicitados por la Dirección de Recursos Humanos con la opción de obtener otros reportes con un nivel de granularidad muy alto y de manera oportuna. En primer lugar, se recabo la

información de los requerimientos mediante entrevistas realizadas a los responsables de los procesos administrativos, coordinadores y directores de áreas para contar con la visión detallada de los reportes, estadísticas y preguntas que requerían resolver al finalizar el proyecto. En secciones previas se describen cada uno de los reportes solicitados por la DRH. Además, se identificaron las fuentes de datos internas y externas que fueron la base para crear el *staging area*. La tabla 20 muestra las herramientas tecnológicas utilizadas en las fases de la metodología son.

Tabla 20

Herramientas tecnológicas del proyecto

Fase de la metodología	**Tipo**	**Herramienta**
1.- Revisión de reportes y preguntas	-	-
2.- Análisis de las estructuras de datos	DB relacionales, hojas de cálculo, XML y TXT	*ORACLE, Python* y *Pentaho Data Integration*
3.- Configuración de catálogos externos e internos	Bases de datos relacionales y hojas de cálculo	*ORACLE y Pentaho Data Integration*
4.- Definir el modelo del *Staging Area*	DB relacional	*ORACLE y Pentaho Data Integration*
5.- Diseño del modelo dimensional: DW	DB relacional	*ORACLE y Pentaho Data Integration*
6.- Procesos ETL	Técnicas ETL	*Pentaho Data Integration y Python*
7.- Control de inconsistencias en los datos	Aplicativos internos	*ORACLE*
8.- Explotación de los datos	Análisis de datos	*Microsoft Power BI*

Nota. El listado de las herramientas utilizadas en el proyecto permitió desarrollar cada fase de la metodología de manera unificada y escalable. En caso de requerirse es posible reemplazarlas de acuerdo a las características del proyecto.

Como se mencionó anteriormente, los proyectos que involucran la definición de un almacén de datos centralizado y, en consecuencia, la inteligencia de negocios, repercutirá en todas las áreas y personas involucras en el proceso analizado, es por ello, que, durante la fase de definición del *staging area*, fue necesario realizar modificaciones y reestructuraciones en diversas aplicaciones que

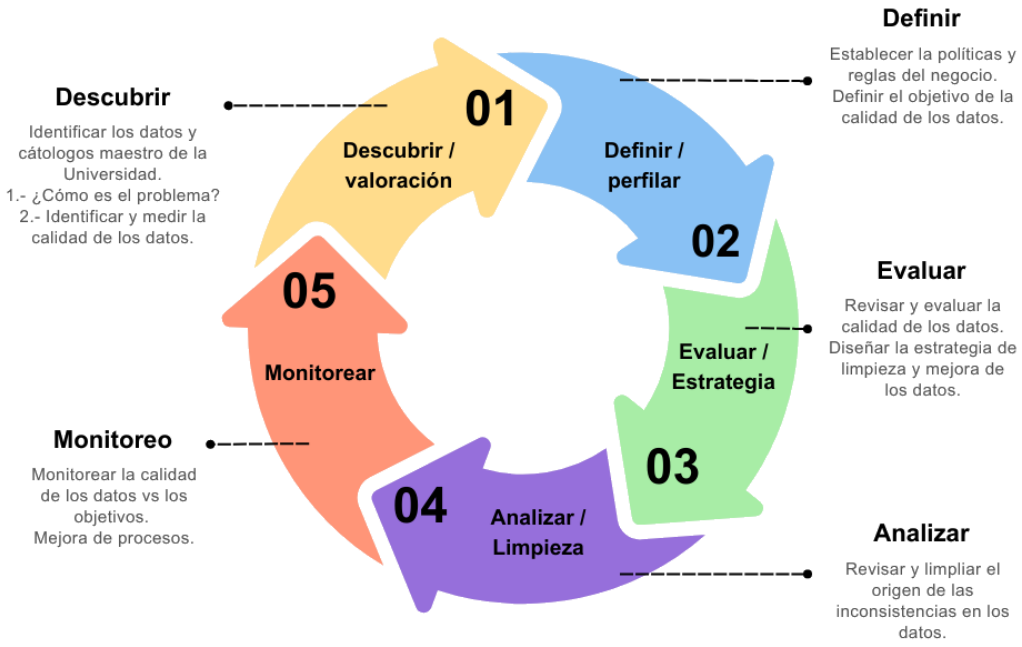
se encuentran en producción, por ejemplo, sistemas de nómina, expediente digital, gestión de constancias SAT y timbrado de nómina. Cada una de estas aplicaciones cumplían con las funciones establecidas en el momento de su desarrollo, sin embargo, se sumaron validaciones en los formularios de registros, se aplicaron mecanismos de integración de sistemas, se aumentaron reglas de integridad en las bases de datos y se agregaron nuevas funcionalidades en la aplicación de timbrado. Lo anterior para apoyar a los usuarios finales en la ejecución de sus funciones y validaciones en los orígenes de datos principales.

Los hallazgos de las primeras cuatro etapas de la metodología propuesta, se basaron en la calidad e integridad de los datos generados en los distintos sistemas de información y la manera de centralizar esos datos para robustecer la emisión de reportes y estadísticas. La integración y homologación de los datos heterogéneos identificados en las fuentes de dato, por ejemplo, en los sistemas transaccionales, las bases de datos internas y externas, hojas de cálculo, *XML* y *TXT*, brindaron una visión holística y una única *versión de la verdad* lo que mejoró el proceso para emitir reportes y estadísticas utilizando datos de calidad, estructurados y estandarizados. La figura 44 representa las acciones realizadas en la revisión de calidad de los datos.

La implementación de la metodología impactó de manera positiva a la parte técnica y administrativa de la Universidad, es decir, en la parte técnica, se actualizaron o depuraron algunas estructuras de datos en los sistemas transaccionales. Mientras que la parte administrativa, se logró una mejora en la cultura organizacional, al generar conciencia sobre la importancia de la calidad y consistencia de la información en la Universidad. Así mismo, el trabajo colaborativo entre los departamentos involucrados, ya que los datos provenían de diferentes departamentos.

Figura 44

Ciclo de vida de la calidad de los datos



Nota. La calidad de los datos brinda la oportunidad de generar reportes y estadísticas oportunos y eficientes, por ello se realizaron los procesos ETL del proyecto considerando las distintas etapas del ciclo.

Los procesos ETL definidos en el proyecto permitieron migrar los datos crudos de las fuentes de datos al *staging area* y/o al DW aplicando las reglas definidas, con lo cual se garantizó que los datos almacenados contaban con la calidad necesaria para su explotación, es decir, los datos presentaban integridad, consistencia y exactitud debido a las reglas de negocio aplicadas en las transformaciones utilizadas para corregir inconsistencias en los datos, por ejemplo, valores duplicados, cálculos erróneos, valores faltantes, etc. Así mismo, la disponibilidad de los datos es mayor debido a la actualización periódica y programada que brindan las herramientas ETL, ya sea como carga inicial o actualización de los datos sin importar las múltiples fuentes de datos. En consecuencia, los reportes y estadísticas consultadas permitieron una toma de

decisiones informada, logrando una eficiencia operativa reduciendo la necesidad de intervención manual por parte del personal técnico.

Por otra parte, el diseño del modelo dimensional y su desarrollo como repositorio destino de los datos migrados por medio de los procesos ETL brindaron numerosos beneficios y ventajas, tanto a la DRH – Dirección de Recursos Humanos como en la Universidad. Con el DW, la Universidad cuenta con una herramienta capaz de fortalecer la generación de reportes y estadísticas, basado en un almacén de datos o repositorio que contiene datos íntegros, consolidados y centralizados sobre los pagos y timbres generados a lo largo del tiempo. Así mismo, se reduce el tiempo requerido para elaborar informes y se permite el monitoreo continúo de las operaciones relacionadas con los timbres elaborados para cada pago, lo que se traduce en la optimización del proceso y fortalece la toma de decisiones asertiva.

Los valores históricos permitieron analizar la evolución de los pagos de acuerdo al centro gastos, a la fuente de financiamiento, al tipo de empleado, a las nóminas y al nivel académico del personal involucrado, contemplado un nivel de calidad alto. Es decir, se logró acceder a cualquier reporte donde intervengan los elementos antes mencionados, además de realizaron análisis avanzados de los datos en diferentes niveles de agregación utilizando las jerarquías disponibles en el modelo y sus métricas, permitiendo el estudio detallado de patrones y tendencias que influyen en los pagos y timbres de la nómina.

Otro aspecto a resaltar de la implementación del DW fue la obtención de una visión pormenorizada del estado actual de los datos almacenados en los diversos sistemas de información de tipo operacionales o transaccionales, proporcionando una perspectiva holística sobre los procesos administrativos, el control y seguimiento de los responsables de áreas, las transacciones ejecutadas por el personal encargado y la funcionalidad de los sistemas de información. Al centralizar los datos se logró un panorama completo del proceso de negocio, lo que permitió la detección de inconsistencias o problemas en los datos. Estos puntos permitieron realizar mejoras en los distintos sistemas de información, replanteando la forma del

monitoreo del proceso y de las acciones del personal operativo, así como la estandarización y homologación de catálogos maestros utilizados por la institución.

La fase final del proyecto implicó la utilización de los datos almacenados en el DW. El modelo implementado puede ser empleado de diversas formas, que abarcan desde consultas directas a las tablas mediante el lenguaje SQL, el desarrollo de soluciones personalizadas para presentar los datos en informes, tablas y gráficos, o con apoyo de herramientas especializadas en la inteligencia de negocios. Para este proyecto en particular, se realizaron consultas directas al repositorio y el software *Microsoft Power BI*.

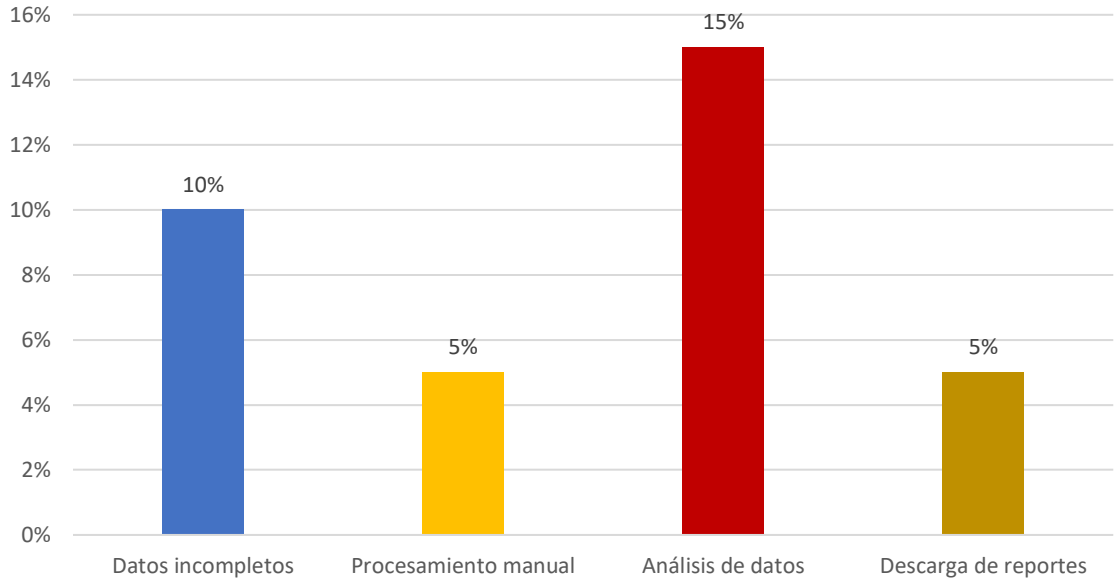
Los resultados en la investigación describen una mejora considerablemente alta sobre los factores que influyen en la calidad de los datos dentro de la Universidad (descritos en la figura 2). La implementación de la metodología descrita en secciones previas, tomando como base el DW, ha logrado fortalecer la elaboración de estadísticas y reportes en la Universidad, específicamente en la DRH y la DITI. Por lo cual, se lograron identificar varias áreas de oportunidad en los procesos realizando mejoras en los sistemas de información transaccionales para disminuir las inconsistencias de entradas manuales o errores de digitalización en un 80%, por ejemplo, validaciones en los formularios, uso de catálogos estandarizados, desarrollo de nuevas funcionalidades, capacitación y concientización a los usuarios. De igual manera, con dichas mejoras disminuyeron en un 95% los valores faltantes, nulos o duplicados en las columnas principales. Mientras que se definieron mecanismos y aplicaciones para la actualización constante de la información, por ejemplo, se recabo el 90% de los datos fiscales requeridos para la nueva versión del timbrado 4.0 con apoyo del expediente digital y recepción de la constancia SAT.

La hipótesis central de la investigación sostiene que al desarrollar una metodología para la elaboración de un sistema de estadísticas y reportes por medio de un *Data Warehouse*, se podrá disminuir el tiempo de elaboración y entrega de reportes, generando resultados oportunos y de calidad, lo cual es indicado con los resultados alcanzados y, por lo tanto, la hipótesis queda demostrada de forma afirmativa. En lo que respecta a los factores o elementos que afectan la entrega de

reportes oportunamente a las entidades externas e internas, de acuerdo a la figura 3, entre ellos se encuentra el tiempo invertido en analizar los datos de las distintas fuentes de datos, lo que representaba un 70% del tiempo, seguido del procesamiento manual y uso de matrices de equivalencia de los datos con un 65%, en tercer lugar el tratamiento de los datos incompletos indicado con el 50%, por último el tiempo destinado para la descarga de los reportes por un 30%. Con la implementación de la metodología, por medio del DW, y la herramienta de explotación de los datos, se estima que estos factores disminuyeron considerablemente en el proceso de pagos de nómina y timbrado, de acuerdo a la figura 45.

Figura 45

Resultados de la implementación de la metodología propuesta

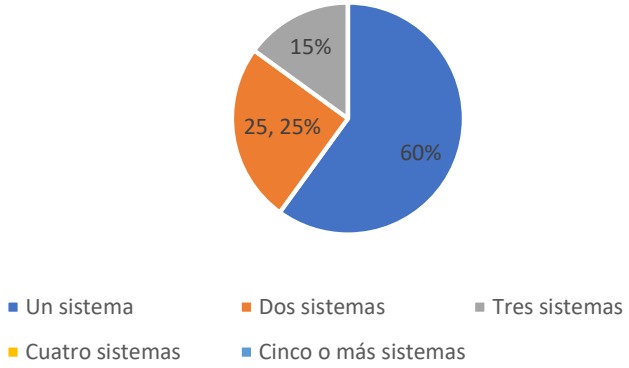


Nota. Con la implementación de la metodología se logró disminuir la percepción de los usuarios finales al momento de emitir los reportes solicitados por entidades externas e internas. En la gráfica se ilustra que la revisión y cotejo de datos disminuyó de un 70% a un 15%, el procesamiento manual de un 65% a un 5%, mientras que la revisión de datos incompletos de un 50% a 10% y la descarga de reportes de un 30% a un 5%.

Al analizar el porcentaje de sistemas de información consultados por el personal de la DRH y otras áreas de la Universidad, de igual manera disminuyo tras implementar la metodología propuesta, inicialmente varios usuarios requerían consultar los datos en más de tres sistemas de información distintos, con las mejoras aplicadas se puede realizar en una sola fuente de información (figura 46).

Figura 46

Porcentaje de fuentes de datos consultadas con la metodología

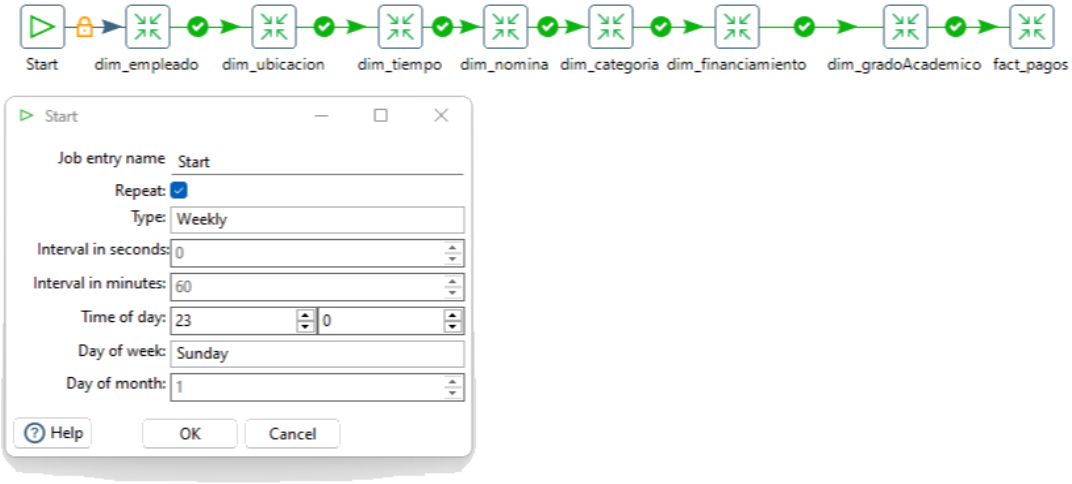


Nota. Las ventajas del DW es contar con información centralizada para su explotación. Con la implementación de la metodología los usuarios finales consultas en una sola fuente de información pasaron de un 25% al 60%.

Otro de los factores analizados en la investigación fue el tiempo utilizado para entregar la información a las entidades externas e internas, con base a lo descrito anteriormente en las figuras 45 y 46, los reportes se generan en cuestión de segundo una vez que los datos se encuentran migrados en el DW, de igual manera, al contar con procesos ETL automatizados (figura 47) la corrección de incidencias es mucho más rápida al momento de homologar los datos. La elaboración de los reportes y estadísticas, dependiendo de la complejidad del mismo, requería entre cuatro horas a una semana, lo que involucraba mayor tiempo de procesamiento y trabajo manual. Con apoyo de *Microsoft Power BI*, los reportes son generados en un tiempo no mayor a una hora, entre la migración de la información y la descarga del reporte.

Figura 47

Proceso ETL programado para su ejecución semanalmente

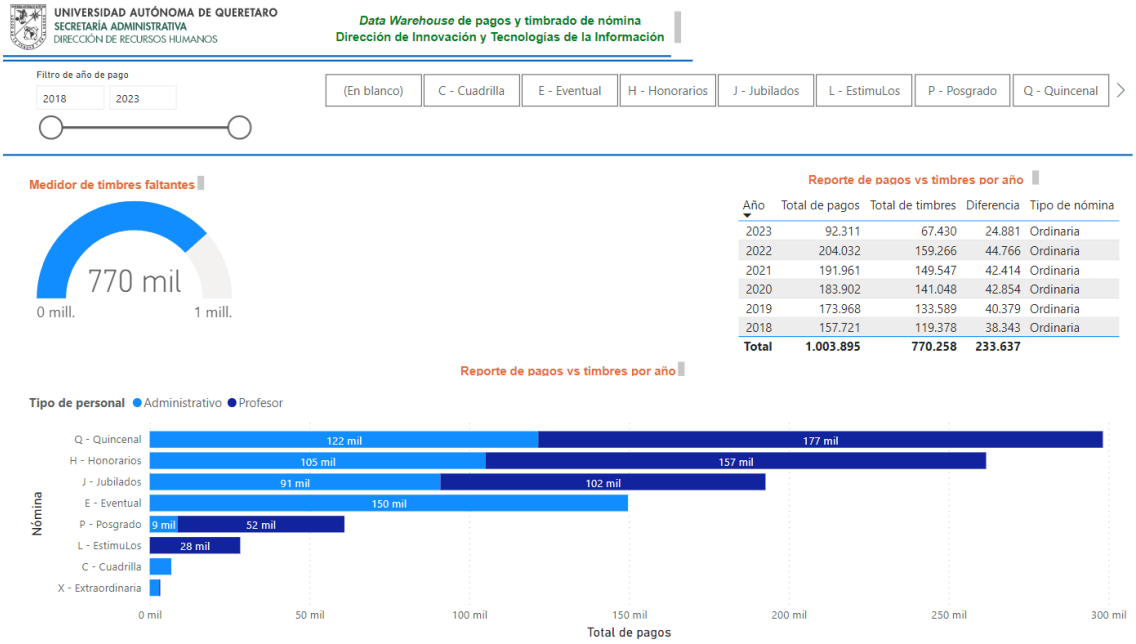


Nota. La migración de información se realiza con base a un *Job* programado durante cada semana. En caso de ser necesario se puede ejecutar manualmente en el momento que se requiera, el *Job* inicializa cada transformación de las tablas de dimensiones y la tabla de hechos del proyecto.

El *software* de explotación permite la incorporación de nuevo reportes e informes que estén dentro de las tablas de dimensiones y tabla de hechos del proyecto, lo que incorpora gran flexibilidad y optimización de tiempo para generar los reportes y compartirlos con el personal autorizado. De manera similar, como se mencionó previamente en este documento, los resultados demuestran que, tras aplicar la metodología propuesta, los datos cuentan con la calidad necesaria evitando la ausencia de valores, duplicidad y valores desactualizados, en comparación con las inconsistencias encontradas por medio de la elaboración de reportes tradicional donde excedía el 70% de inconsistencia encontradas en algunos casos. En la Figura 48 se muestra la interface principal de los reportes generados a partir de la metodología.

Figura 48

Dashboard principal de pagos y timbrado



Nota. El portal de consulta permite visualizar y descargar los datos que en ese momento se encuentran en el repositorio. La representación de los datos se realiza mediante gráficas y tablas dinámicas para facilitar el analisis y explotación de los datos.

Los hallazgos del proyecto de tesis coinciden con lo descrito por Fuentes et al. (2019), quienes sostenían que para contar con datos actualizados, confiables y poderlos analizar desde distintas perspectivas era necesario el uso de un repositorio que centralice los datos, denomina DW. Así mismo, los resultados obtenidos coinciden con lo referido por Jaramillo y Pauta (2019), donde menciona que el uso de un DW permite eliminar los silos de información generados a lo largo de la historia de la organización, obtenido con ello, centralización de los datos, la depuración, la homologación y datos confiables. Lo cual estaba relacionado con los factores que influyen en la entrega tardía de los reportes y el re-trabajo realizado para su obtención.

La implementación de una metodología cuya la parte central es un DW, permite la eliminación del trabajo manual, el cual tiene un alto grado de riesgo al recapturar la información de acuerdo a lo descrito por Palacios (2017), los resultados demostraron que al aplicar las ocho fases propuestas inicialmente reducen notoriamente el riesgo de elaborar reportes y estadísticas incongruentes, desactualizados y fuera de tiempo. Las fases consisten en la revisión de reportes, el análisis de las estructuras de bases de datos utilizadas en la obtención de los reportes, la configuración de posibles catálogos internos y externos, la definición de un área de pre-procesamiento de datos, el diseño dimensional en estrella (*Data Warehouse*), la definición y desarrollo de los procesos ETL, el control de las inconsistencias detectas y, por último, la presentación de datos almacenados en el DW a los responsables o personal gerencial por medio de reportes o estadísticas.

7. Conclusiones

La elaboración de reportes, informes y/o estadísticas es una actividad cotidiana en las organizaciones actuales, dichos reportes son solicitados por entidades internas y externas a la empresa, por lo cual deben ser elaborados con precisión, oportuna y eficientemente, de ahí la importancia de la metodología desarrollada en la Universidad, la cual fortaleció y estandarizo el proceso. La metodología se basó en las teorías de *Inmon* y *Kimball* para concentrar y organizar los datos provenientes de diferentes fuentes a través de un repositorio denominado *Data Warehouse*, con el propósito de agilizar la generación y entrega de reportes. Las perspectivas de estos autores han servido como fundamento para el desarrollo de proyectos de tipo BI, considerando como parte fundamental de las teorías la integración del DW, así mismo, la facilidad y velocidad con la que se obtienen los reportes utilizando los datos almacenados en los modelos dimensionales.

La metodología planteada considera diferentes fases que abarcan desde la definición de requerimientos hasta la capa de presentación de los datos almacenados en el DW. De esta manera, la implementación de la metodología facilita la centralización y organización de los datos provenientes de múltiples

orígenes a un repositorio único, permitiendo acceder a datos íntegros y con la calidad necesaria para ser utilizados en los reportes y estadísticas. Esto se logra al identificar los requerimientos o necesidades de información que tienen mayor impacto para el negocio o institución y, de las cuales, se requiere un análisis detallado a partir de distintas perspectivas o contextos. Al mismo tiempo el estudio de los orígenes de datos, tanto en estructura como en contenido, fortalece el entendimiento de las reglas del negocio y la elaboración de planes de acción en caso de encontrar inconsistencias en los datos almacenados.

La fase central de toda solución BI se enfoca en el análisis y diseño del modelo dimensional, donde se definen las medidas y el contexto sobre los cuales se generarán los reportes y estadísticas. Para realizar la migración de los datos a las tablas de dimensiones y tabla de hechos del modelo dimensional comúnmente se utilizan procesos ETL que garantizan la calidad de los datos cargados en el DW al implementar las reglas del negocio, estandarizando y homologando los datos provenientes de distintos orígenes. En este punto radica la importancia de trabajar en equipo con los integrantes del proceso de negocio, pues son ellos los que facilitan la definición de catálogos maestros, definen las reglas o normas que se deben considerar al encontrar inconsistencias en los datos y las acciones a tomar.

La reducción de los tiempos al momento de generar y entregar los reportes, se logra al implementar metodologías que centralicen los datos en modelos dimensionales con apoyo de procesos ETL que disminuyen considerablemente el trabajo manual y la intervención del personal técnico. Los datos contenidos en el repositorio contienen la estructura necesaria para ser explotados por herramientas especializadas que ofrecen componentes visuales y de configuración relativamente sencilla, lo que facilita la generación de reportes y estadísticas oportunas, eficientes y consistentes para fortalecer la toma de decisiones.

El listado de requerimientos, los reportes analizados, las estadísticas referentes a los procesos de pago de nómina, el timbrado de nómina, el *staging area* y el modelo dimensional en estrella construido en el proyecto, son procesos similares en cualquier institución de nivel superior pública, por ello, pueden ser

considerados para implementación similares, agregando o quitando columnas a las dimensiones y tabla de hechos. De igual manera, el manejo de catálogos como medio de homologación de los datos es un elemento crucial que puede ser utilizado como mejores prácticas en cualquier organización. Sin embargo, los procesos ETL serán distintos para cada institución, esto debido a que cada una de ellas maneja desarrollos propios.

La aportación de esta tesis radica en la creación y aplicación de un enfoque o metodología que permite la centralización de los datos para su posterior análisis y explotación, considerando el uso de capas intermedias como lo es la estandarización de los datos por medio de catálogos y la definición del área de preparado de datos o *staging area*, dicha metodología puede ser replicable en cualquier institución superior pública que requiera construir un repositorio de datos o DW para la Dirección de Recursos Humanos, específicamente en los procesos de pagos y timbrado de nómina. De igual manera, se brinda al personal una manera simple de observar la conversión de datos en información y la evolución de la información hacia el conocimiento, lo cual permite contar con información estratégica verídica y confiable que puede medirse y monitorear su comportamiento a lo largo de la organización.

La implementación de la metodología ofrece un panorama amplio para explotar los datos contenidos sobre las diversas fuentes de datos institucionales, provenientes de los procesos administrativos y académicos. Posibilitará la generación de reportes o informes detallados y actualizados, proporcionado al personal directivo de la institución un panorama preciso y oportuno del proceso analizado. Dicha información fortalecerá la gestión estratégica de la Universidad, además de analizar los datos históricos, la identificación de posibles patrones, el análisis de tendencias a lo largo del tiempo y la posibilidad de realizar comparativos entre la misma área y las demás.

Por otra parte, las limitaciones reconocidas en la investigación fueron al momento de identificar claramente los requerimientos de información del cliente, entre las auditorías externas y los reportes internos, el tiempo invertido en definir las

dimensiones fue extenso y con cambios constantes a lo largo de la implementación, de igual manera el desconocimiento sobre las fuentes de datos origen y la poca o nula documentación sobre las reglas del negocio y los procesos administrativos dificultaron la ejecución del proyecto en un lapso menor. Al tratarse de una institución cambiante, pueden surgir nuevos elementos no contemplados en el análisis inicial e incluso puede darse el caso de no contar con algunos datos.

La única constante en los proyectos de *software* es el cambio, y, este proyecto no fue la excepción, por ello, el trabajo colaborativo y la comunicación fueron fundamentales para llevar a buen término la implementación de la metodología y el uso de un repositorio que centralice los datos de la organización. Parte de los desafíos superados consistieron en la consolidación e integración de los datos priorizando su calidad, debido a la interpretación o valor que cada departamento les daba al momento de su captura y el seguimiento correspondiente, además, el establecimiento de controles de calidad basados en las reglas de negocio no documentadas o con visiones encontradas entre las áreas fue un desafío importante. En cuanto a las lecciones aprendidas se destaca la importancia de la colaboración de los directivos y personas operativo responsable de los procesos del negocio para definir y dar seguimiento al proyecto, por otra parte, la flexibilidad y disposición del equipo técnico para realizar cambios de manera iterativa al descubrir nuevas necesidades.

La implementación de una metodología robusta con apoyo del elemento principal de una solución BI denominado *Data Warehouse*, conlleva diversas complicaciones que deben ser resueltas con el apoyo y soporte de los directivos, dichas complicaciones pueden ser tanto de carácter técnico como de gestión de personal. Sin embargo, es indispensable recordar que el objetivo principal de un proyecto de esta naturaleza es el fortalecimiento de la elaboración de estadísticas y reportes que contengan datos con la calidad necesaria para su entrega oportuna a entidades externas o internas mediante visualizaciones detalladas y de fácil comprensión, que sean útiles para su análisis y fortalezcan la toma de decisiones estratégicas.

8. Bibliografía o Referencias

Cai, L., y Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*(November). https://doi.org/10.5334/dsj-2015-002

Conesa, J., y Curto, J. (2011). *Introducción al Business Intelligence* (S. A. El Ciervo 96 (ed.)).

Contreras, R. (2018). *CMMI DEV 2 Como base del proceso ETL para la generación de conocimiento en la inteligencia de negocios.*

Davenport, T. H., y Laurence, P. (1998). Working knowledge: how organizations manage what they know. *Choice Reviews Online*, *35*(09), 35-5167-35–5167. https://doi.org/10.5860/choice.35-5167

Duque Méndez, N. D., Hernández Leal, E. J., Pérez Zapata, Á. M., Arroyave Tabares, A. F., y Espinosa, D. A. (2016). Model for the Extraction , Transformation and Load Process. *Ciencia e Ingeniería Neogranadina*, *26*(2), 95–109.

Foster, E. C., y Godbole, S. (2016). Database systems: A pragmatic approach. En *Database Systems: A Pragmatic Approach*. https://doi.org/10.1007/978-1-4842-1191-5

Fuentes, D. A., Soraca, J. A., Cobos, C. A., Mendoza, M. E., y Gómez, L. C. (2019). Una Revisión de Bodegas de Datos para Educación Superior. *Revista Ibérica de Sistemas y Tecnologias de Información*, *33*, 309–322.

Inmon, W. H. (2005). Building the Data Warehouse: Timely. Practical. Reliable. En *Wiley Publishing, Inc.* (Vol. 13, Número 401).

Iqbal, M. Z., Mustafa, G., Sarwar, N., Wajid, S. H., Nasir, J., y Siddque, S. (2020). A Review of Star Schema and Snowflakes Schema. *Communications in Computer and Information Science*, *1198*(August), 129–140. https://doi.org/10.1007/978-981-15-5232-8_12

Japal, A. (2021). *Designing a Lecturer ' s Performance Data Warehouse Model Using Star Scheme*. 73–76.

Jaramillo, A. M., y Pauta, S. L. (2019). Diseño de un modelo físico de Data Warehouse para la gestión de incidencias para una empresa de telecomunicaciones, aplicando la metodología Hefesto. *Polo del Conocimiento*, *4*(7), 95. https://doi.org/10.23857/pc.v4i7.1026

Kimball, R., y Caserta, J. (2004). *The Data Warehouse ETL Toolkit, Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*.

Kimball, R., Reeves, L., Ross, M., y Thornthwaite, W. (2008). The Data Warehouse Lifecycle Toolkit Table of Contents. *Architecture*, 1–405.

Kimball, R., y Ross, M. (2013). *The Data Warehouse Toolkit The Definitive Guide to Dimensional Modeling*.

Lennerholt, C., van Laere, J., y Söderström, E. (2018). Implementation challenges of self service business intelligence: A literature review. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2018-Janua*, 5055–5063. https://doi.org/10.24251/hicss.2018.631

Lépez, H. R. (2019). Técnica de extracción, transformación y carga de datos de estaciones meteorológicas. *Revista de Ciencia y Tecnología*, *32*, 28–32. https://doi.org/10.36995/j.recyt.2019.32.005

Mamani, Y. (2018). Business Intelligence: herramientas para la toma de decisiones en procesos de negocio. *ResearchGate*, *March*, 0–6.

Medina, F., Fariña, F., y Castillo, R. (2018). Data mart to obtain indicators of academic productivity in a university | Data mart para obtención de indicadores de productividad académica en una universidad. *Ingeniare*, *26*, 88–101.

Montoya, D. M., y Jiménez, J. A. (2018). Caracterización de elementos para la creación de una herramienta computacional para la gestión del conocimiento en las organizaciones. En *Mg. Jovany Sepúlveda Aguirre* (Número October).

Morales, A., Cuevas, R., y Martínez, J. M. (2016). Analytical Processing with Data Mining. *RECI Revista Iberoamericana de las Ciencias Computacionales e Informática*, *5*(9), 22–43. http://www.reci.org.mx/index.php/reci/article/view/40/176

Muñoz, H., Osorio Mass, R. C., y Zúñiga Pérez, L. M. (2016). Inteligencia de los negocios. Clave del Éxito en la era de la información. *Clío América*, *10*(20), 194. https://doi.org/10.21676/23897848.1877

Oliveira, P., Rodrigues, F., y Galhardas, H. (2005). A Taxonomy of Data Quality Problems. *2nd Int. Workshop on Data and Information Quality*, *April 2014*, 219–233.

Palacios, G. S. C. (2017). *Un sistema de ayuda a la decisión para instituciones financieras del sector de la economía popular y solidaria: Un enfoque basado en conceptos de data warehouse.*

Potineni, P. (2021). *Oracle ® Database Data Warehousing Guide 10.2* (Vol. 2, Número December).

Pozo, J. C. (2004). *Diseño de un sistema de información, bajo un enfoque de inteligencia de negocios, para el proceso de toma de decisiones. Caso: Empresa Diafoot*.

Rainardi, V. (2008). Building a data warehouse: With examples in SQL server. En *Building a Data Warehouse: With Examples in SQL Server*. https://doi.org/10.1007/978-1-4302-0528-9

Reyes, Y. D., y Nuñez, L. M. (2015). La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico. *L*, *3*(2), 63–73.

Rodríguez, J. (2019). Business Intelligence (BI)/Inteligencia de negocios •. *Cómo hacer inteligente su negocio: business intelligence a su alcance*, *2014*, 101–116.

Sabtu, A., Azmi, N. F. M., Sjarif, N. N. A., Ismail, S. A., Yusop, O. M., Sarkan, H., y

Chuprat, S. (2017). The challenges of extract, transform and load (ETL) for data integration in near real-time environment. *Journal of Theoretical and Applied Information Technology*, *95*(22), 6314–6322.

Sherman, R. (2015). *Business Intelligence Guidebook: From Data Integration to Analytics 1st Edition, Kindle Edition*.

Silva Peñafiel, G. E. (2018). Análisis de metodologías para la implementación de un data warehouse aplicado a la toma de decisiones del Instituto Nacional de Patrimonio Cultural Regional 3. *Repositorio de Tesis - PUCE*, 233. http://repositorio.pucesa.edu.ec/handle/123456789/2367%0Ahttp://repositorio.pucesa.edu.ec/bitstream/123456789/2367/1/76540.pdf

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., y Yahia, S. Ben. (2019). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Strand, M., y Syberfeldt, A. (2020). Using external data in a BI solution to optimise waste management. *Journal of Decision Systems*, *29*(1), 53–68. https://doi.org/10.1080/12460125.2020.1732174

Strong, D. M., Lee, Y. W., y Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804

Testa, S., y Malbernat, L. R. (2018). *Madurez de los entornos tecnológicos empresariales.* 251–254.

Trujillo, J. C., Mazón, J. N., y Pardillo, J. (2009). *Diseño y explotación de datos. Conceptos básicos de modelado multidimensional*.

Universidad Autónoma de Querétaro (UAQ). (2023). *No Title*. https://www.uaq.mx/

Yulianto, A. A. (2019). Extract Transform Load (ETL) Process in Distributed Database Academic Data Warehouse. *APTIKOM Journal on Computer Science and Information Technologies*, *4*(2), 61–68.

https://doi.org/10.11591/aptikom.j.csit.36

Zambrano, C. del C., Rojas, D. F., y Salcedo, P. A. (2018). A method for analyzing data from standardized educational tests using data warehouse and triangulation. *Formacion Universitaria*, *11*(4), 3–14. https://doi.org/10.4067/S0718-50062018000400003

Zangana, H. M. (2018). Developing Data Warehouse for Student Information System (IIUM as a Case Study). *International Organization of Scientific Research*, *20*(1), 9–14. https://doi.org/10.9790/0661-2001020914

Zelaya, E., Enciso, L., y Quezada, P. A. (2018). International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC) Enfoque de arquitectura empresarial en las organizaciones de gestión de datos Focus of enterprise architecture in the organizations of data management. *Ijisebc*, *5*(2), 07–17.