UNIVERSIDAD NACIONAL DE LA MATANZA

ESCUELA DE POSGRADO

MAESTRIA EN INFORMÁTICA

COMO MANEJA EL DATA WAREHOUSE EL PASO DEL TIEMPO

Autor: Hernán Alejandro Osores

Director: Mg. Julio Bossero

Buenos Aires, octubre *de 2023*

Agradecimientos

Agradezco a Dios por el amor y esfuerzo de mi familia que me permitieron crecer y desarrollar como persona y profesional. En especial a mis padres, mi hermano y mi esposa que supo contenerme cuando lo necesite.

Al profesor Hugo Castro por inculcar el deseo de conocimiento por el tema y su ayuda en la realización. A mi director de tesis por su dedicación y esmero por mostrarme el camino para finalizar el trabajo emprendido.

Al Sec. Académico por darme la oportunidad de llevar a la práctica los conceptos vertidos en mi tesis.

Al DIIT por la predisposición para ayudarme y brindarme información sobre la temática estudiada.

A la UNLaM por ser nuestro orgullo y brindar la posibilidad de futuro para la comunidad de La Matanza.

Resumen

Los sistemas transaccionales manejan enormes cantidades de datos organizados de forma tal que puedan ser utilizados por las aplicaciones operacionales existentes. Se basan en un modelo de entidad – relación, protección de integridad, altas, bajas, modificaciones y eficiencia en los procesos. La filosofía del Data Warehouse es diferente, se modelan los datos a partir de dimensiones y las herramientas de acceso a los datos se basan en una tecnología de procesamiento analítico, distinta al procesamiento transaccional de los sistemas operacionales. Este proceso es denominado por Kimball ETL (Extract – Transform - Load) el sistema de ETL es una actividad que no es muy visible para el usuario final, consume fácilmente el 70 por ciento de los recursos necesarios para la implementación y mantenimiento de un Data Warehouse típico. Tiene la tarea crítica de convertir el caos de datos del mundo operacional en un mundo ordenado de información. Este proceso asimila datos procedentes de tecnologías heterogéneas dentro de un entorno integrado y consistente, apto para ser consumido por los procesos de soporte de decisiones.

Una de las tareas importantes en el proceso ETL es asociar un nuevo identificador a cada registro de dimensión y evitar la dependencia de las claves definidas en las fuentes a ello Kimball lo llama clave subrogada.

Existen distintas formas de tratar los datos históricos en los sistemas de tipo Data Warehouse. El paso del tiempo en estos sistemas tiene varias maneras de ser tratado y dependerá del interés por mantener la historia en determinados datos o elementos. La naturaleza cambiante de los datos hace que debamos tomar decisiones para manejar estos cambios y las técnicas que permiten realizarlo toman particular interés en los profesionales de negocios y analistas.

En el presente trabajo veremos porque los sistemas transaccionales no pueden resolver los inconvenientes que se producen al momento de tomar decisiones. Luego veremos cómo realizar y que variantes debemos tener en cuenta en el proceso ETL para generar y actualizar los contenidos de un Data Warehouse. Se muestran las técnicas para manejar la historia en los Data Warehouse, se diseña y construye un DW para la toma

de decisiones en el área académica para la Secretaría Académica de la Universidad Nacional de La Matanza.

Índice

Índice de Figuras

Capítulo 1 Introducción

1.1 Selección del Tema y Definición del Problema

El paso del tiempo en toda organización debe ser administrado de forma eficiente con el objetivo de brindar a los directivos de las diferentes empresas información útil para tomar decisiones. Aquí es donde el especialista informático debe interpretar las necesidades para evaluar que técnica es la que mejor se adapta a la empresa en cuestión. Kimball interpreta esta problemática como dimensión de cambio lento SCD (Kimball, 2008) y Dimensión de Cambio no tan lento (Kimball, 1999) (Kimball & Ross, 2013).

1.2 Antecedentes

El tiempo es una de las variables más importantes en todo proceso. En este caso cuando nos referimos a horizonte temporal, decimos el lapso de tiempo que los datos permanecen en los sistemas de almacenamiento.

El horizonte temporal del Data Warehouse (DW) es mayor que el de los sistemas transaccionales, debe reflejar el paso del tiempo, para no perder la historia, por ejemplo, un producto cambia de denominación, una sucursal cambia de distrito. El profesional de negocios debe decidir qué hacer, si guardará la historia y con qué detalle.

Distintos tipos de dimensiones manejan en forma diferente la conservación de la historia, no hay un tipo que sea mejor que otro, esto lo debe interpretar el profesional de negocios. En principio Kimball propone tres tipos de tratamientos denominados tipo 1, tipo 2 y tipo 3 para la gestión del tiempo en un modelo dimensional, luego junto con Margy Ross sugerirán otro tipo 6 (Ross, 2000) (Kimball & Ross, 2002). A continuación, se detallan brevemente cada una de ellas.

SCD tipo 1: Sobrescribir el valor, nos limitamos a remplazar el valor antiguo del atributo en la fila de dimensión, por el valor actual. De este modo el atributo siempre refleja la asignación más reciente.

La respuesta de tipo 1 es el método más sencillo para hacer frente a la dimensión cambios en los atributos. La ventaja es que es rápido y fácil. En la tabla de dimensiones, nos limitamos a sobrescribir el valor preexistente con la asignación actual. El problema con una respuesta de tipo 1 es que perdemos toda la historia de los cambios en los atributos. Dado que la sobre escritura borra los históricos de los valores de los atributos, nos quedamos únicamente con los valores actuales. Es apropiado si el cambio de atributo es una corrección. También puede ser adecuado si no hay ningún interés en mantener la descripción anterior.

Desde el comienzo debemos determinar la importancia de mantener el valor del atributo. Con demasiada frecuencia, los equipos de proyecto utilizan un tipo 1 como la respuesta por defecto para tratar con dimensiones de cambio lento y terminan fracasando si la empresa necesita realizar un seguimiento de los cambios históricos (Kimball & Ross, 2002).

SCD tipo 2: Agregar una fila de dimensión, dijimos que uno de los objetivos del Data Warehouse es representar la historia correctamente. Un tipo 2 es la respuesta predominante para responder a este requerimiento. Cuando un valor del atributo cambia se agrega una nueva fila a la tabla de dimensiones.

Dado que el tipo 2 genera nuevas filas de dimensión, un aspecto negativo de este enfoque se puede dar cuando los cambios no son tan lentos. Por lo tanto, puede ser una técnica inadecuada para las tablas de dimensiones que ya superan el millón de filas (Kimball & Ross, 2002).

SCD tipo 3: Agregar una columna de dimensión, guarda una cantidad limitada de valores históricos de atributos seleccionados. Cuando hay un cambio, nos guardamos el valor anterior en una columna distinta, actualizando el campo con el nuevo valor (para cada campo, tendremos una tupla valor anterior, valor actual). Solo nos vamos a guardar, por tanto, los dos últimos valores.

SCD Híbrido o tipo 6: Diferentes tipos de SCD pueden ser aplicados a distintas columnas de una tabla. Con este enfoque híbrido, emitimos una nueva fila para registrar un cambio (tipo 2) y añadir una nueva columna para realizar un seguimiento de la asignación actual (tipo 3), donde cambios posteriores son manejados como tipo 1.

1.3 Justificación del Estudio

El estudio se lleva a cabo dada la importancia de los DW en la toma de decisiones, sabemos que entender y administrar la información de la empresa e instituciones es vital para tomar decisiones a tiempo y responder a las condiciones siempre cambiantes del negocio.

Se reconoce el interés de las Secretaría Académica por mejorar sus procesos a partir del análisis de los datos que se manejan de forma diaria. Por lo tanto, a partir de los problemas observados y para dar respuesta a los requerimientos del secretario Académico en el proceso de ingreso a la Universidad Nacional de La Matanza se decide desarrollar un Data Warehouse.

El proyecto es de gran utilidad ya que en un primer plano se revisan los aspectos teóricos analizando las distintas técnicas que permiten manejar la historia para mostrar a los analistas de forma clara que técnica utilizar para cubrir sus necesidades de negocio. Además, se puede considerar esta información para la construcción de un Data Warehouse como base para futuros desarrollos. En lo que se refiere al aspecto practico, se realizan los procesos necesarios para la construcción de un Data Warehouse para el ingreso de los aspirantes a la Universidad Nacional de La Matanza.

1.4 Alcances del trabajo

Este trabajo evaluara primero las limitaciones para reflejar el paso del tiempo de los sistemas transaccionales. Luego mostraremos las diferentes técnicas que presentan los Data Warehouse para administrar el paso del tiempo.

1.5 Objetivos

1.5.1 Objetivo general

El objetivo principal es establecer la incidencia del Data Warehouse en el soporte a la toma de decisiones con la construcción e implementación de un Data Warehouse para el proceso de ingresantes a la Universidad Nacional de La Matanza.

1.5.2 Objetivo Especifico

- Exponer las referencias teóricas para mejorar la toma de decisiones.
- Analizar las características de los DW.
- Determinar los indicadores para la toma de decisiones de los ingresantes a la Universidad.
- Desarrollar e implementar un Data Warehouse de ingresantes que brinde la información necesaria para lo toma de decisiones.

1.6 Hipótesis

Se pretende demostrar que la implementación de un Data Warehouse permitirá a la Secretaría Académica, analizar la información de los aspirantes del curso de ingreso a nuestra Universidad, para satisfacer sus necesidades analíticas de información.

Capítulo 2 Marco Teórico

2.1 Sistemas Transaccionales

Los sistemas transaccionales OLTP (Procesamiento de transacciones en línea, Online Transaction Processing) permiten registrar las transacciones rutinarias necesarias para conducir el negocio. Están orientadas fundamentalmente al nivel operacional de la organización.

En cuanto a sus principales características permiten la ejecución en tiempo real de un gran número de transacciones de base de datos por parte de un gran número de personas. Necesitan tiempos de respuesta muy rápidos, modifican pequeñas cantidades de datos con frecuencia y normalmente implican un equilibrio de lecturas y escrituras. Utilizan datos indexados para mejorar los tiempos de respuesta y necesitan copias de seguridad frecuentes o simultáneas de la base de datos. Requieren un espacio de almacenamiento relativamente pequeño. Normalmente se ejecutan consultas sencillas que implican solo uno o varios registros (Oracle, s.f.).

Son óptimas en el manejo de los datos que constantemente están cambiando y usualmente tienen un gran número de usuarios que están ejecutando transacciones simultáneamente y que actualizan o modifican los datos en tiempo real. En este tipo de aplicaciones la principal preocupación es la concurrencia y la atomicidad. Los controles de la concurrencia aseguran que dos usuarios no puedan cambiar o modificar el mismo dato simultáneamente o que un usuario no pueda modificar una parte de un dato antes de que otro usuario haya terminado de usarlo. La atomicidad asegura que todos los pasos relacionados en una transacción se completen satisfactoriamente como un todo. Si algún paso falla, toda la transacción se debe volver atrás (Perez Marquez, 2011).

Conformidad con ACID (Atomicity, Consistency, Isolation, Durability), los sistemas OLTP deben cumplir con las propiedades atómicas, coherentes, aisladas y duraderas (ACID) para garantizar la precisión de los datos del sistema.

Atomicidad: Una transacción es una unidad atómica de procesamiento; o se ejecuta en su totalidad o no se ejecuta en absoluto.

Conservación de la consistencia: Una transacción está conservando la consistencia si su ejecución completa lleva a la base de datos de un estado consistente a otro.

Aislamiento: Una transacción debe aparecer como si estuviera ejecutándose de forma aislada a las demás. Es decir, la ejecución de una transacción no debe interferir con la ejecución de ninguna otra transacción simultánea.

Durabilidad: Los cambios aplicados a la base de datos por una transacción confirmada deben persistir en la base de datos. Estos cambios no deben perderse por culpa de un fallo. (Elmasri & Navathe, 2007)

El historial de datos suele limitarse a los datos actuales o recientes. La alta normalización de la base de datos, por lo general, en tercera forma normal permite eliminar la redundancia de información para evitar la pérdida de integridad de los datos. Los datos administrados por los sistemas transaccionales son la fuente principal de datos para el Data Warehouse.

En la ***Imagen 1*** podemos ver el nivel operativo, es donde se agrupa la mayoría de los trabajadores y donde los procedimientos están más estructurados y definidos. Los niveles superiores (nivel táctico y estratégico) efectúan actividades más relacionadas con el control y la toma de decisiones acerca de la dirección que debe tomar la organización (Chinkes, 2008).

Además, en el nivel operativo vemos los sistemas SIU-Guaraní y SIU-Mapuche ambos desarrollados por el SIU (Sistema de Información Universitaria) que desde el año 1996 desarrolla soluciones informáticas para el sistema universitario argentino y otros organismos de gobierno, desde septiembre de 2013 es parte del CIN (Consejo Interuniversitario Nacional).

El SIU-Guaraní es un Módulo de gestión académica que permite registrar, de manera óptima y segura, las actividades de la gestión académica dentro de una institución desde que una persona se inscribe hasta que egresa.

SIU-Mapuche es un módulo de recursos humanos que permite llevar adelante la gestión digital integrada de recursos humanos en tiempo real desplegando una amplia capa de servicios entre los cuales se destacan la liquidación, la gestión impositiva actualizada y la administración de todo el personal de la institución.

Las bases de datos OLTP constituyen una gran herramienta para toda organización, pero nos encontramos con algunos inconvenientes a la hora de brindar información a los niveles superiores de la pirámide. Chinkes en su libro (Chinkes, 2008) lo llama "Los tres problemas del OLTP". Los describe de la siguiente manera:

a) Problema 1: Información no integrada
   Los datos de la organización se encuentran muchas veces dispersos y no integrados. Son bases de datos que tienen diseños conceptuales y físicos que difieren entre sí.

b) Problema 2: Inadecuados tiempos de respuesta
   La toma de decisiones se construye a partir del acceso y procesamiento de importantes volúmenes de datos. Se realizan de grandes porciones de la base de datos en forma interactiva. En la medida en que las bases tienen grandes cantidades de datos, se hace muy difícil obtener adecuados tiempos de respuesta para este tipo de consultas.

c) Problema 3: Imposibilidad de realizar consultas *ad doc* amigables
   Los tomadores de decisiones necesitan alternativas amigables de acceso a los datos. Es decir que puedan hacerlo sin tener mucho conocimiento sobre el uso de la tecnología, siendo esta lo más transparente posible. El problema entonces es que los usuarios no cuentan con herramientas para explorar los datos y obtener información para la toma de decisiones en forma amigable para necesidades *ad doc*.

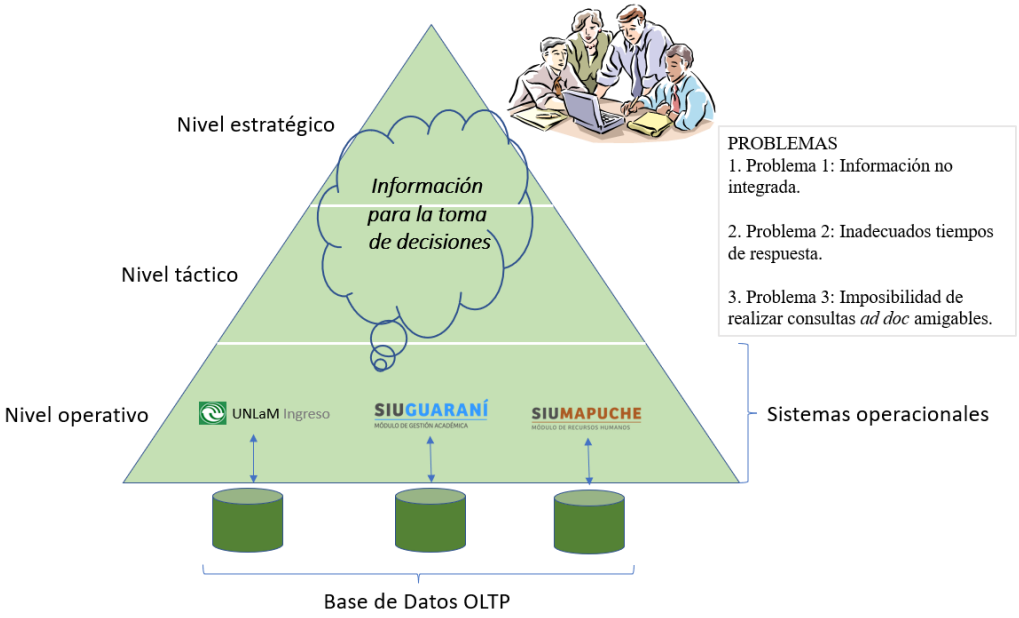


Imagen 1 Sistemas operacionales y los tres problemas.

Fuente: (Chinkes, 2008) modificado con sistemas operacionales de la UNLaM.

2.2 Que es un Data Warehouse

Un Data Warehouse es una base de datos orientada al análisis de la información histórica contenida en ella. Dependiendo de las necesidades de análisis de la organización puede almacenarse desde unos meses hasta varios años de información. El modelo que soporta la información contenida se encuentra diseñado, estructurado e implementado con la finalidad y propósito del análisis y navegación de los datos. Se entiende por navegación o drilling de los datos, la posibilidad de ver información correspondiente a diferentes contextos o entornos, por ejemplo, analizar las ventas anuales y poder abrirlas por sucursal, después analizar más en detalle una sucursal para ver cómo se discriminan las ventas por cada producto (MicroStrategy LATAM South, 2006).

La ***Imagen 2*** muestra por un lado los sistemas operacionales basados en los comandos INSERT, UPDATE, DELETE y luego de un proceso de cargas masivas obtenemos el Data Warehouse, basado en comandos SELECT para agregar datos para informes.

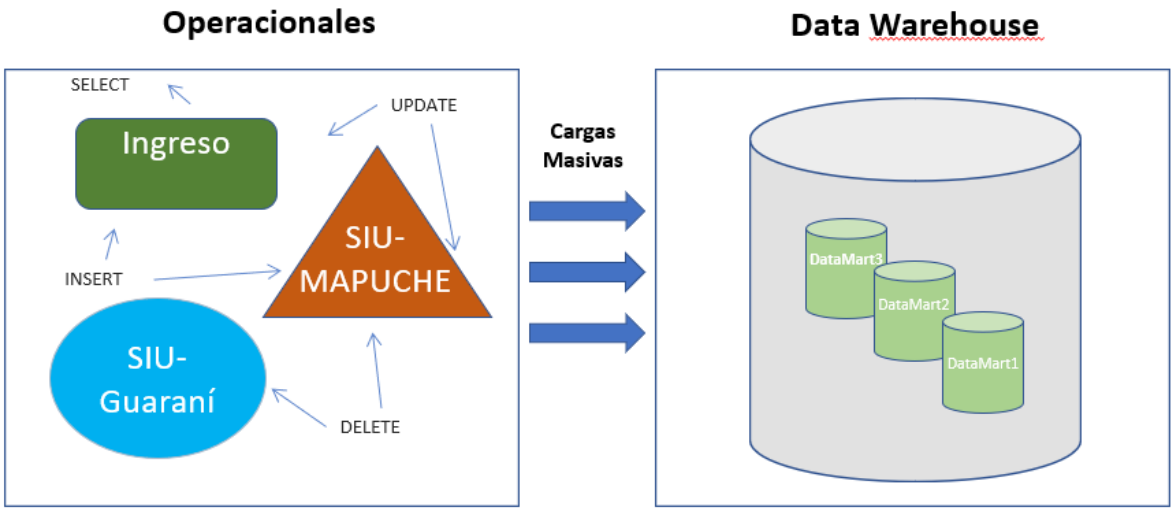


Imagen 2 Sistemas operacionales y Data Warehouse.

Fuente: Elaboración Propia.

Según la implementación seleccionada, los datos son almacenados en forma relacional RDBMS (Relational DataBase Management System, Sistema de gestión de bases de datos relacionales) respetando ciertos estándares a nivel de definición y consulta de datos o en formato multidimensional, las bases de datos multidimensionales son arquitecturas propietarias definidas por cada proveedor que son frecuentemente actualizadas desde base de datos relacionales. Típicamente los usuarios tienen sólo permisos de lectura (read-only) sobre el DW (MicroStrategy LATAM South, 2006).

Comúnmente se dice que los Data Warehouse son fuentes secundarias de información pues no generan información por sí mismos, sino que son actualizados desde sistemas fuentes existentes internamente en la organización (sistema de ventas, sistema presupuestario, etc.) o sistemas externos de información (datos meteorológicos, información de la competencia, cotizaciones de la bolsa, etc.). Por todo esto los DW son identificados como ambientes OLAP, procesamiento analítico en línea, en contraposición a los ambientes transaccionales clásicos OLTP (Bedell, 1997).

El Data Warehouse maneja un gran volumen de datos debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes y áreas en un solo lugar centralizado. Presenta información sumarizada y agregada desde múltiples versiones, y maneja información histórica. Organiza y almacena los datos que se necesitan para realizar consultas y procesos analíticos, con el propósito de responder a preguntas complejas y brindarles a los usuarios finales la posibilidad de que, mediante una interfaz amigable, intuitiva y fácil de utilizar, puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos (Bernabeu, 2009).

Las cualidades de un Data Warehouse son imposibles de saldar en un típico ambiente operacional. Son bases de datos más voluminosas, que surgen de la necesidad de analizar la información para saber qué ocurre y tomar decisiones estratégicas. Si tenemos en cuenta que el proceso de toma de decisiones consiste en encontrar hechos destacados, explicarlos en términos de negocio y tomar las decisiones correspondientes necesitamos una nueva estructura, una base de datos con la estructura adecuada. En este punto podemos decir que un DW es una base de

datos separada de los sistemas transaccionales e independiente de ellos, diseñada para apoyar la toma de decisiones en las organizaciones (Castro, 2014).

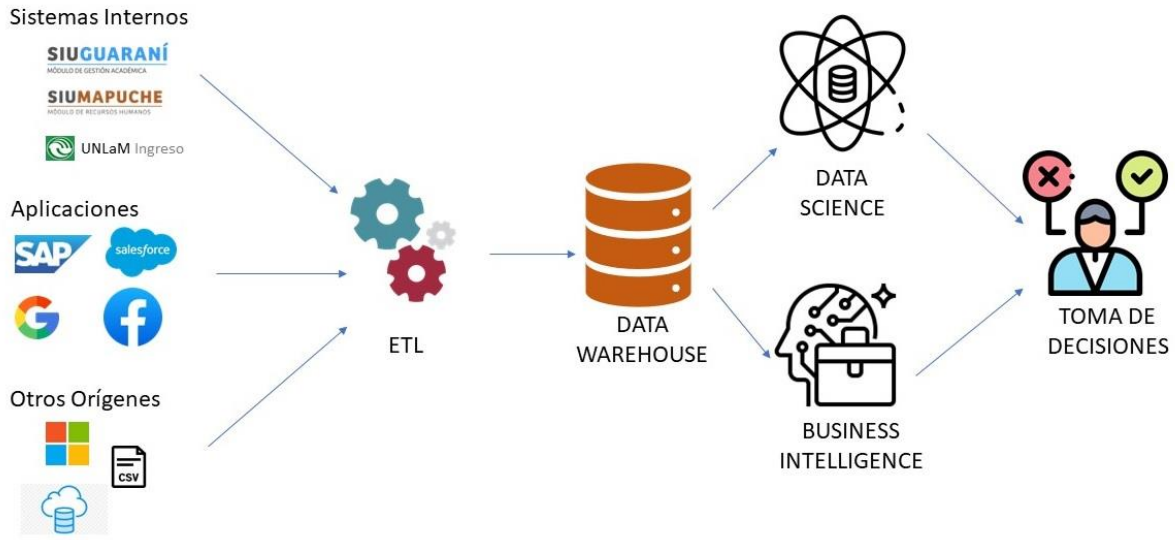


Imagen 3 Data Warehouse.

Fuente: Elaboración Propia.

Kimball, define brevemente un Data Warehouse como una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis (Kimball R. , 1996). Otra definicion clasica lo define como una colección de datos no volátil, integrada y variable en el tiempo, orientada al negocio que tiene como objetivo ser una herramienta para la toma de decisiones (Inmon, 2005).

Veamos las caracteristicas que enuncia Inmon en detalle.

2.2.1 Orientada al negocio

Organiza y presenta los datos desde la perspectiva de los conceptos que maneja la empresa (fecha, franja horaria, producto, sucursal y ventas). Los datos tienen el nivel de detalle y la estructura que necesitan los que toman decisiones. Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

En la ***Imagen 4*** se muestra cómo en el ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. Por ejemplo, una aplicación de ingreso de órdenes puede acceder a los datos sobre clientes, productos y cuentas. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación. En el ambiente DW se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, éstos pueden ser clientes, productos, proveedores y vendedores. Para una universidad pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc. La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el DW. Las principales áreas de los temas influyen en la parte más importante de la estructura clave. En el DW se excluye la información que no será usada por el proceso de sistemas de soporte de decisiones, mientras que la información de las orientadas a las aplicaciones contiene datos para satisfacer de inmediato los requerimientos funcionales y de proceso, que pueden ser usados o no por el analista de soporte de decisiones. Otra diferencia importante está en la interrelación de la información. Los datos operacionales mantienen una relación continua entre dos o más tablas basadas en una regla comercial que está vigente. Las del DW miden un espectro de tiempo y las relaciones encontradas en el DW son muchas. Muchas de las reglas comerciales (y sus correspondientes relaciones de datos) se representan en el DW, entre dos o más tablas.

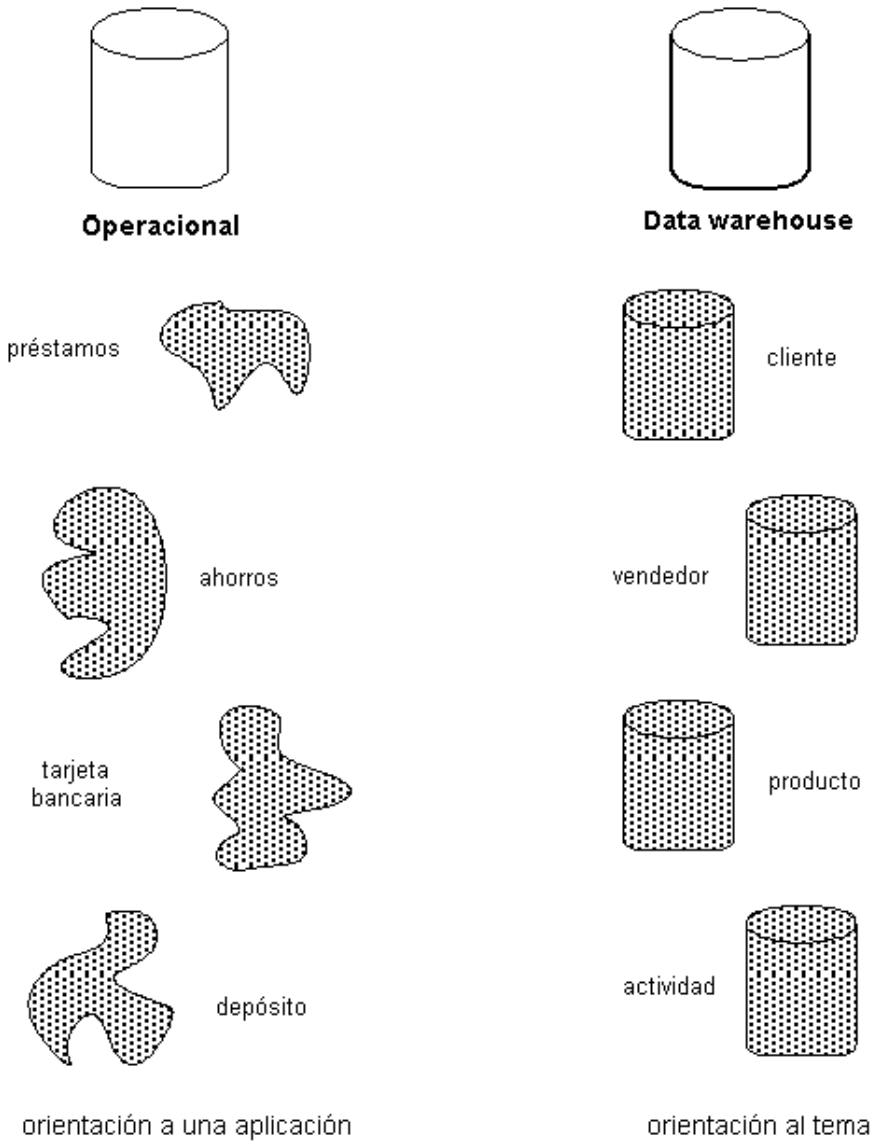


Imagen 4 Organización de los datos en sistemas Operacionales y Data Warehouse.

Fuente: (Inmon, 2005)

2.2.2 Integrada

Es uno de los aspectos mas importantes de un Data Warehouse. Se construye a partir de fuentes de datos heterogéneas. Bases de datos relacionales, archivos planos, hojas de cálculo y documentos impresos. Se unifican denominaciones, codificaciones y formatos.

Con los años los diseñadores de las diferentes aplicaciones han tomado sus propias decisiones en cuanto a cómo una aplicación debe ser construida. El estilo y las decisiones personales del diseñador de la aplicación se pueden ver de distintas maneras. En las diferencias en la codificación, las estructuras de clave, características físicas, las convenciones de nombres.

Cuando se integran datos de diferentes fuentes hay que contemplar cuestiones como la estandarización de codificaciones. Por ejemplo, los diseñadores de aplicaciones han elegido para codificar el campo género de diferentes maneras. Un diseñador representa el género como una “m” y una “f”. Otro diseñador representa el género como un “1” y un “0”. Otro diseñador representa el género como una “x” e “y”, y otro diseñador representa el género como "masculino” y "femenino". No importa tanto cómo llega el género al Data Warehouse. “M” y “F” son probablemente tan buenos como cualquier representación. Lo que importa es que independientemente de la fuente de donde provenga, el género debe llegar al Data Warehouse en un estado integrado y uniforme. Por lo tanto, cuando el género se carga en el Data Warehouse desde la aplicación en la que se ha representado que el género no sea un formato “M” y “F”, los datos deben convertirse al formato del Data Warehouse.

Cualquiera que sea la cuestión de diseño, los datos tienen que ser almacenados en el Data Warehouse de una manera singular, globalmente aceptable, incluso cuando los sistemas operacionales almacenen los datos de manera diferente.

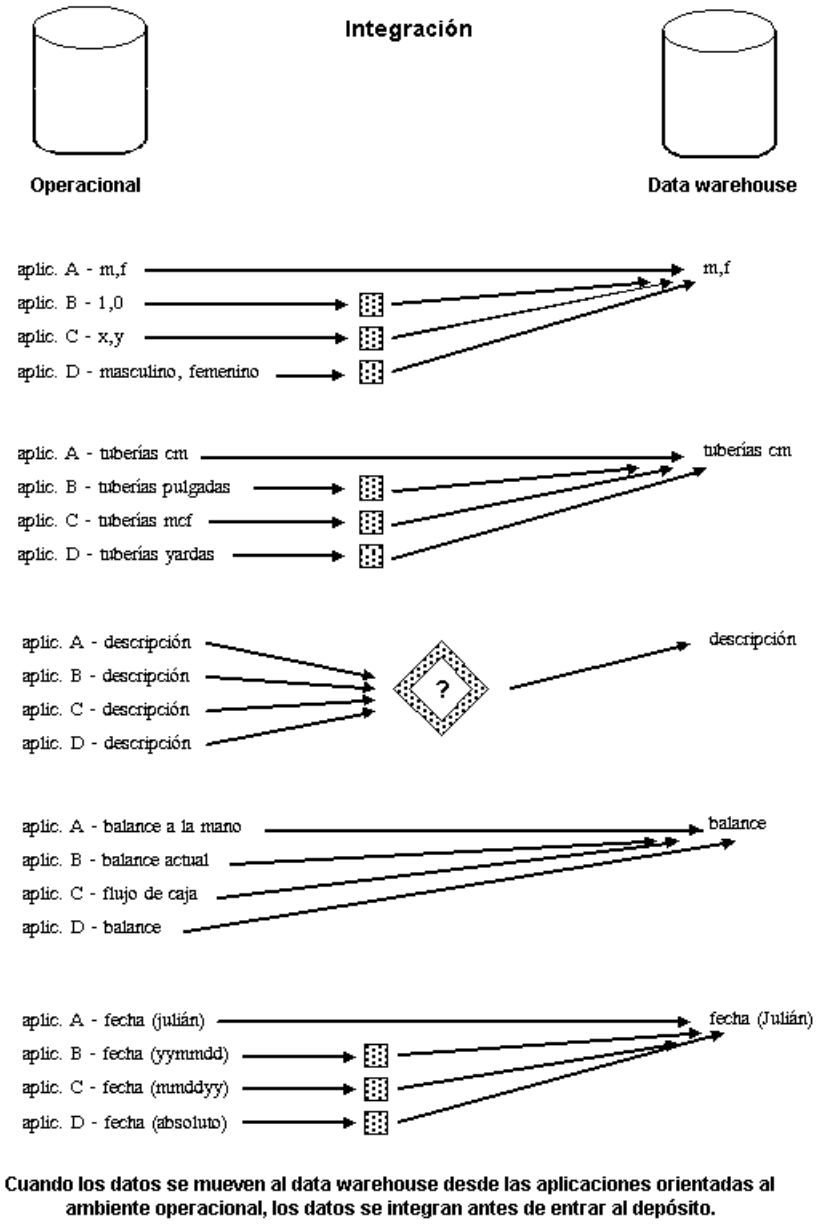


Imagen 5 Integración de los datos.

Fuente: (Inmon, 2005)

2.2.3 Variable en el tiempo

El horizonte temporal del Data Warehouse es más amplio que el de los sistemas transaccionales. La vida útil de los datos, el tiempo es parte implícita de la información contenida en un DW. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el DW sirve, entre otras

cosas, para realizar análisis de tendencias. Por lo tanto, el DW se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

El horizonte de tiempo para los datos que se encuentran dentro de un Data Warehouse es significativamente mayor que el de los sistemas operacionales. Un horizonte de tiempo de 60 a 90 días es normal para los sistemas operacionales; un horizonte de tiempo de 5 a 10 años es normal en un DW. Como resultado de esta diferencia en horizontes de tiempo, el DW contiene mucha más historia que cualquier otro entorno.

Se puede recuperar datos históricos de 3 meses, 6 meses, 12 meses, o incluso más antiguos en un Data Warehouse. Esto contrasta con un sistema de transacciones, donde se mantiene a menudo sólo los datos más recientes. Por ejemplo, un sistema de transacción puede tener la dirección más reciente de un cliente, un DW puede contener todas las direcciones asociadas a un cliente.

La fecha es un dato fundamental, todos los datos en el Data Warehouse son asociados con un período de tiempo específico, marcación temporal. La clave de los datos operativos puede o no contener algún elemento de tiempo, como año, mes, día, etc. La clave del DW siempre contiene algún elemento de tiempo. La incorporación del elemento de tiempo puede tomar muchas formas, tales como una marca de tiempo en cada registro, una marca de tiempo para toda una base de datos, y así sucesivamente.

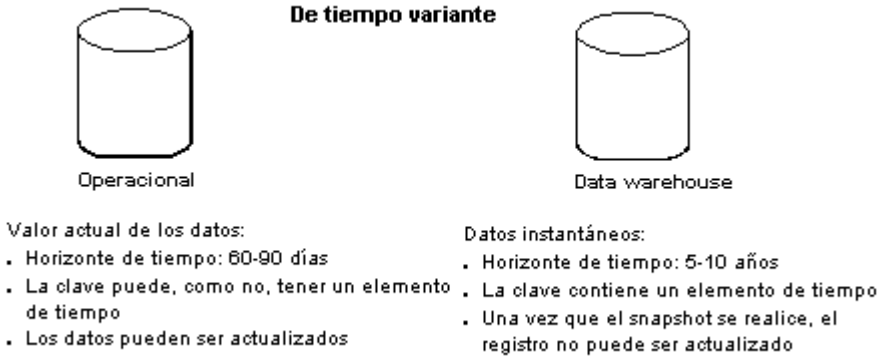


Imagen 6 Horizonte temporal sistemas Operacionales y Data Warehouse.

Fuente: (Inmon, 2005)

2.2.4 No Volatil

En el Data Warehouse los datos no se modifican. Un DW existe para ser leído, y no modificado. La información es por tanto permanente, significando la actualización del DW la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

El Data Warehouse se renueva, Los datos permanecen intactos entre renovaciones. Sólo existen dos operaciones carga y acceso. Los datos son estables en el DW. Se puede agregar más datos, pero los datos existentes no son removidos.

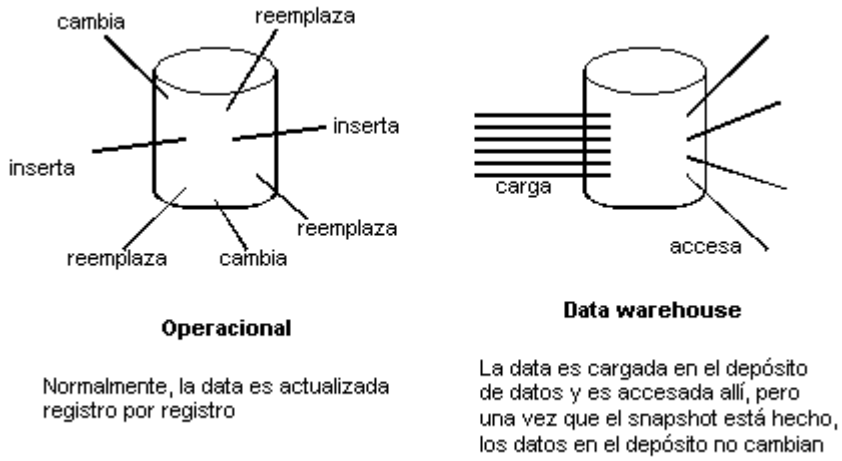


Imagen 7 Volatilidad de los datos.

Fuente: (Inmon, 2005)

La no volatilidad de los datos muestra que los datos operacionales se acceden y manipulan de un registro a la vez. Los datos se actualizan en el ambiente operacional como algo habitual, pero los datos del Data Warehouse muestran un conjunto muy diferente de características. Los datos son cargados (por lo general en masa) y accedidos, pero no se actualizan. Cuando se cargan los datos, se carga una foto estática (snapshot). Cuando se producen cambios posteriores, se cargan fotos nuevas. Al hacerlo, un historial de datos se mantiene en el DW.

2.2.5 Metodologías de Diseño y Construcción de un Data Warehouse

Tanto William H. Inmon como Ralph Kimball son los autores más importantes en lo que refiere a Data Warehouse y debemos aclarar que hay diferencias en la metodología que utilizan para el diseño y construcción. Para comprender la mayor diferencia entre estas dos metodologías, debemos explicar la idea de Data Mart (DM). Un DM es un repositorio de información, similar a un DW, pero orientado a un área o departamento específico de la organización (por ejemplo, Compras, Ventas, RRHH, etc.), a diferencia del DW que cubre toda la organización, es decir la diferencia fundamental es su alcance.

Desde el punto de vista arquitectónico, la mayor diferencia entre los dos autores es el sentido de la construcción del Data Warehouse, esto es comenzando por los DM o ascendente (Bottom-up, Kimball) o comenzando con todo el DW desde el principio, o descendente (Top-Down, Inmon).

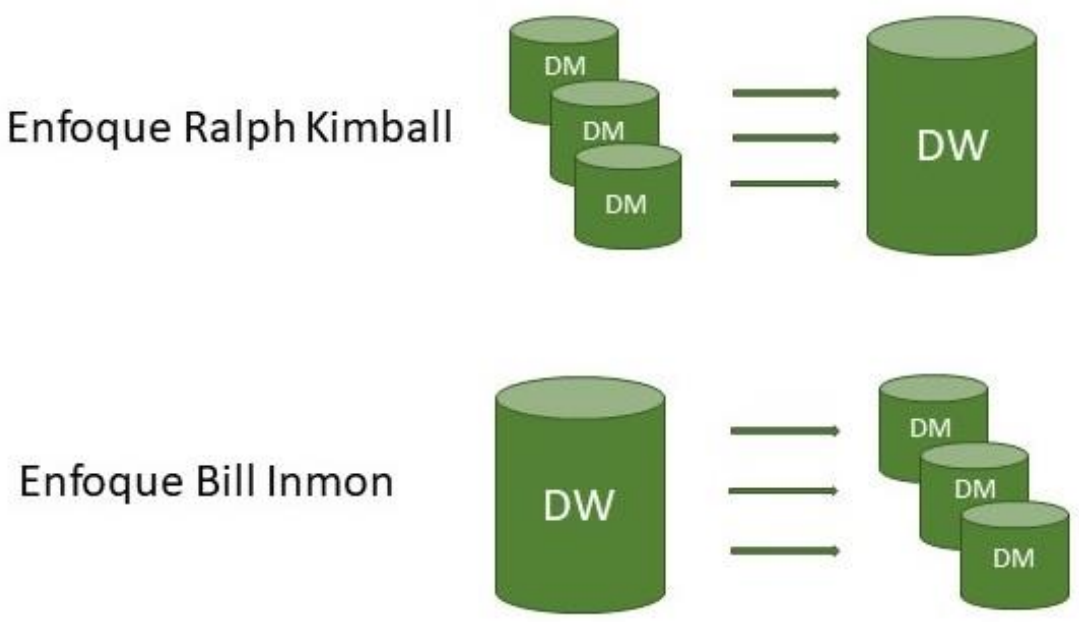


Imagen 8 Diferencia de enfoque Kimbal y Inmon.

Fuente: Elaboración Propia.

Con respecto a los modelos de datos, la metodología de Inmon se basa en conceptos bien conocidos del diseño de bases de datos relacionales, la información se almacena en la tercera forma normal. Al contrario de la de Kimball donde la información siempre se almacena en el modelo dimensional es decir en un modelo desnormalizado (Rivadera, 2010). Podemos tomar como ejemplo un esquema estrella donde la tabla de hechos está en tercera forma normal, no tiene filas repetidas, pero las tablas de dimensiones están en segunda forma normal porque todos los productos de una misma familia llevan como atributo el nombre de la familia.

Por último, un Data Warehouse contiene información histórica para visualizar tendencias y efectuar comparaciones. Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios, contiene información consolidada para acelerar la respuesta a las consultas con un mayor soporte de información que resultan en obtener decisiones más rápidas; así también, la gente de negocios adquiere mayor confianza en sus propias decisiones y las del resto, y logra un mayor entendimiento de los impactos de sus decisiones. Convierte los datos operacionales en información relacionada y estructurada, que genera el conocimiento necesario para la toma de decisiones. Esto permite establecer una base única del modelo de información de la organización, que puede dar lugar a una visión global de la información en base a los conceptos de negocio que tratan los usuarios. Además, aporta una mejor calidad y flexibilidad en el análisis del mercado, y del entorno en general

(Bigatti & Grasso, 2008).

2.2.6 Entorno del Data Warehouse

El objetivo de cualquier entorno de almacenamiento de datos es publicar los datos correctos y hacerlos fácilmente accesibles para los responsables de la toma de decisiones. Los dos componentes principales de este entorno son la puesta en escena y la presentación. El área de presentación (Staging Area) consiste en el soporte y procesos de extracción, transformación y carga (ETL[1]). Una vez que los datos se preparan adecuadamente, se cargan en el área de presentación (o entrega) donde se utilizan una variedad de aplicaciones de consulta, generación

[1] Se explica en detalle en el punto 2.3

de informes, inteligencia comercial y análisis para sondear, analizar y presentar datos en infinitas combinaciones (Ross M. , 2004).

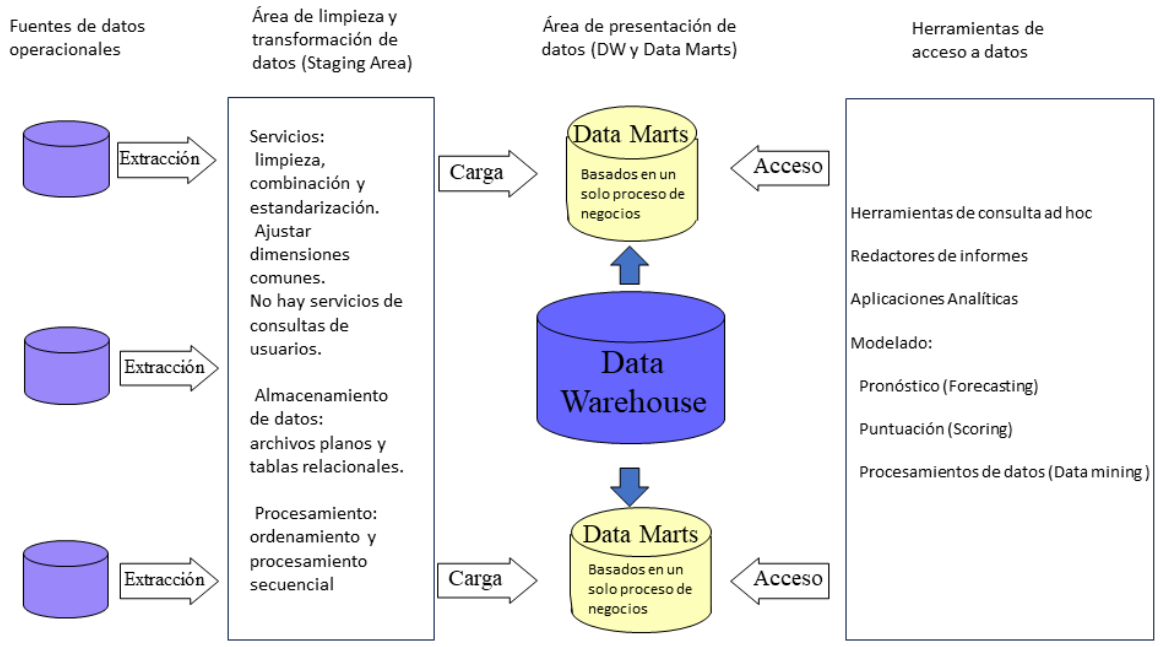


Imagen 9 Elementos Básicos de un Data Warehouse.

Fuente: Traducción de (Kimball & Ross, 2002)

2.2.7 Introduciendo el Modelo Dimensional

En los sistemas transaccionales Utilizamos el Modelo de entidad-relación ER que es una técnica de diseño lógico que busca eliminar la redundancia en los datos.

En los sistemas de soporte a la decisión utilizamos el Modelo dimensional que es una técnica de diseño lógico para estructurar datos de modo que sea intuitivo para los usuarios de negocios y brinde un rendimiento rápido de consultas (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008) Es lo que hace que un Data Warehouse sea una base de datos orientada al negocio (Castro, 2014).

En cuanto a los beneficios del modelo tenemos la comprensión, porque es más fácil para un usuario de negocios que el modelo del típico sistema de origen normalizado dado que la información se agrupa en categorías comerciales o dimensiones que tienen sentido para la gente de negocios. El rendimiento de las consultas debido a que nuestro modelo es desnormalizado. Por último, la flexibilidad a los cambios con nuevos datos.

Los modelos dimensionales almacenados en una plataforma de base de datos relacional se denominan típicamente ***esquemas en estrella***; Los modelos dimensionales almacenados en estructuras de procesamiento analítico en línea (OLAP) multidimensional se denominan ***cubos*** (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008).

Cuando los datos se cargan en un cubo OLAP, se almacenan e indexan utilizando formatos y técnicas diseñadas para datos dimensionales. Las agregaciones de rendimiento o las tablas de resumen pre calculadas a menudo son creadas y administradas por el motor de cubo OLAP. En consecuencia, los cubos ofrecen un rendimiento de consulta superior debido a los cálculos previos, las estrategias de indexación y otras optimizaciones. Los usuarios comerciales pueden obtener mayor o menor detalle agregando o eliminando atributos de sus análisis con un rendimiento excelente sin generar nuevas consultas (Kimball & Ross, 2013)

Podemos visualizar en la ***Imagen 10*** que las 3 dimensiones: Materia, Tiempo y Carrera forman un cubo. Si asumimos que la materia x, se promociona en el tiempo y, en la carrera z, el cruce de los valores x, y, z a lo largo de las abscisas determinan el valor del dato, es decir, definen el hecho de que ocurrió una promoción. Nótese que cada porción del cubo de la figura representa las promociones de una materia, de una carrera determinada en un tiempo establecido. Una consulta para el cubo descripto podría ser la cantidad de promocionados de la materia Contabilidad del año 2019 en la carrera Contador público.

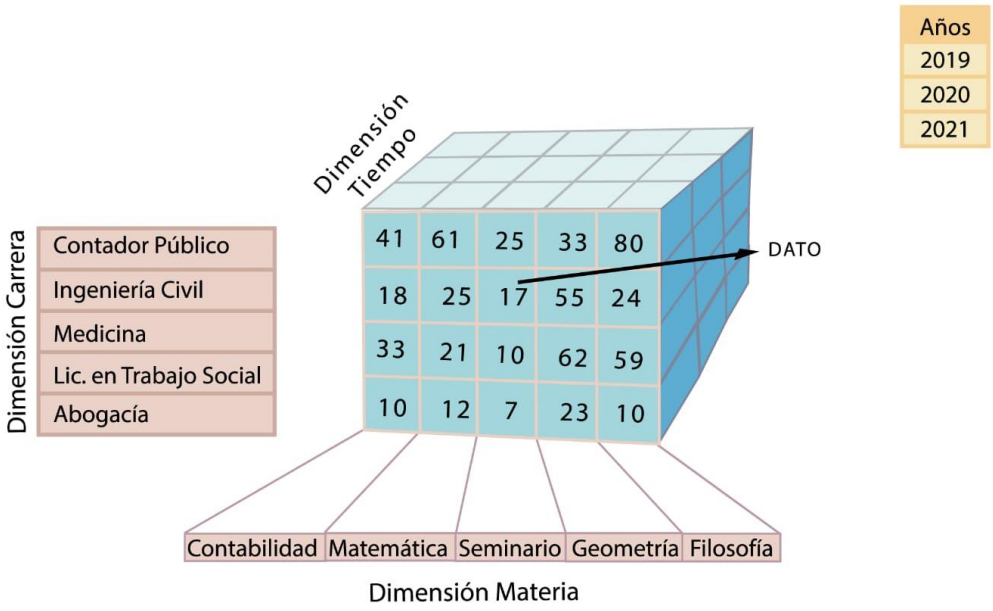


Imagen 10 Cubo con Información Académica.

Fuente: Elaboración Propia.

Según (Chinkes, 2008), dice que la problemática del negocio se modela para luego ser estudiada a partir de hechos y dimensiones. Un hecho es, por ejemplo, una venta. Pero esta venta solo tiene valor analítico una vez que esta puesta en un contexto. El contexto para los hechos de negocio está dado por las dimensiones. Ejemplos de dimensiones son tiempo, cliente, producto, etc.

Las tablas de hechos almacenan las medidas de rendimiento generadas por las actividades o eventos comerciales de la organización (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008).

Las tablas de dimensiones son complementos integrales de una tabla de hechos. Las tablas de dimensiones contienen el contexto textual asociado con un evento de medición de procesos comerciales. Describen el "quién, qué, dónde, cuándo, cómo y por qué" asociado con el evento (Kimball & Ross, 2013).

2.3 El proceso ETL

El termino ETL significa la ejecución de tres pasos principales en la construcción de un *Data Warehouse* que son la Extracción (*Extraction*), Transformación (*Transformation*) y Carga (*Loading*) de datos (Kimball & Caserta, 2008).

Se definen como procesos de extracción a aquellos requeridos para obtener los datos que permitirán efectuar la carga del Modelo Físico acordado. Así mismo, se definen como procesos de transformación los procesos para convertir o recodificar los datos fuente a fin de poder efectuar la carga efectiva del Modelo Físico. Por otra parte, los procesos de carga de datos son los procesos requeridos para poblar el Data Warehouse.

Todas estas tareas son altamente críticas pues tienen que ver con la materia prima del Data Warehouse: los datos. La desconfianza y pérdida de credibilidad del Data Warehouse serán resultados inmediatos e inevitables si el usuario choca con información inconsistente. Es por ello que la calidad de los datos es un factor determinante en el éxito de un proyecto de DW. Es en esta etapa donde deben sanearse todos los inconvenientes relacionados con la calidad de los datos fuente.

Kimball propone un plan de 10 etapas para desarrollar un ETL

1- Elaborar un plan de alto nivel, un esquema de una página del flujo origen-destino.
2- Probar, elegir e implementar una herramienta ETL, existe una multitud de herramientas ETL disponibles en el mercado de Data Warehouse. Sirven para una variedad de funciones. Algunas son particularmente buenas para extraer información de sistemas de origen específicos.
3- Desarrollar estrategias para gestionar dimensiones, manejo de errores y otros procesos.
4- Drill - Down (subdividir el problema en subconjuntos) para tablas de destino, gráficamente bosquejar cualquier restructuración o transformación de datos complejo y desarrollo preliminar de la secuencia de trabajo.
5- Construir y probar la carga de tablas de dimensión históricas.

6- Construir y probar la carga de la tabla de hechos históricas, incluyendo búsqueda y sustitución de clave subrogada.
7- Construir y probar los procesos de carga incremental de la tabla de dimensiones.
8- Construir y probar los procesos de carga incremental de la tabla de hechos.
9- Construir y probar la carga de la tabla agregada y procesos OLAP.
10- Diseñar, construir y probar la automatización del sistema ETL (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008).

Como se comentó, la etapa de ETL consume el 70% de las necesidades de recursos para el desarrollo y mantenimiento de un sistema Data Warehouse. Además, estos procesos no son solamente un mero traspaso de información de un sistema a otro. Son mucho más, pues pueden dar un valor significativo a los datos (Kimball, 2008) (Kimball & Caserta, 2008).

Desarrollar el sistema ETL es complejo porque tantas restricciones externas ejercen presión sobre su diseño: los requerimientos del negocio, la calidad de los datos en los sistemas de origen, el presupuesto y habilidades del personal disponible. Sin embargo, puede ser difícil de apreciar por qué el sistema ETL es tan complejo y consume muchos recursos.

Muchos diseñadores argumentan que el proceso ETL depende muchas veces de la fuente, de la idiosincrasia de los datos, de los lenguajes de programación y herramientas ETL disponibles, de las habilidades del personal etc. el "depende" es peligroso porque se convierte en una excusa para hacer su propio sistema de ETL. Quizá este tipo de enfoque de diseño era apropiado hace algunos años, cuando todo el mundo estaba tratando de entender la tarea del ETL, pero con el beneficio de miles de Data Warehouse exitosos, Kimball elaboro un conjunto de buenas prácticas para la elaboración del ETL donde enumera 34 subsistemas que van a servir a la mayoría de los Data Warehouse (Kimball & Ross, 2013)

2.4 Claves subrogadas

Las tablas se relacionan con otras tablas mediante una relación de clave primaria o de clave foránea. Las relaciones de claves primarias y foráneas se utilizan en las bases de datos relacionales para definir relaciones de muchos a uno entre tablas. Esto se refiere a la cardinalidad que es el número de entidades con la cual otra entidad puede asociarse mediante una relación. Los sistemas operacionales utilizan el concepto de clave primaria (primary key) y se utiliza para identificar de forma única un registro en una tabla. Se requiere una identificación única para cada registro porque no hay otra forma de encontrar un registro sin la posibilidad de encontrar más de un registro, si no se utiliza el identificador único. Las claves foráneas (foreign keys) son las copias de las claves primarias creadas en tablas secundarias para formar el lado opuesto del enlace en una relación entre tablas, estableciendo una relación de base de datos relacional. Una clave foránea define la referencia para cada registro en la tabla secundaria, haciendo referencia a la clave principal en la tabla principal (Powell, 2006).

Hugo Castro (2014) menciona que para conectar la tabla de hechos a las de dimensión tiene que verificarse la integridad referencial entre la tabla de hechos y las tablas de dimensión. En la tabla de hechos, cada dimensión tiene una clave foránea (foreign key) que apunta a la fila que corresponde en la tabla de dimensión. En la tabla de dimensión, esa clave tiene que ser una clave primaria (primary key). ¿Cuál es esa clave? Tenemos dos opciones. La primera opción la clave provista por los sistemas fuente (ej. código de artículo, código de cliente). Se la llama clave natural, clave del negocio, clave operativa o clave inteligente (Castro, 2014).

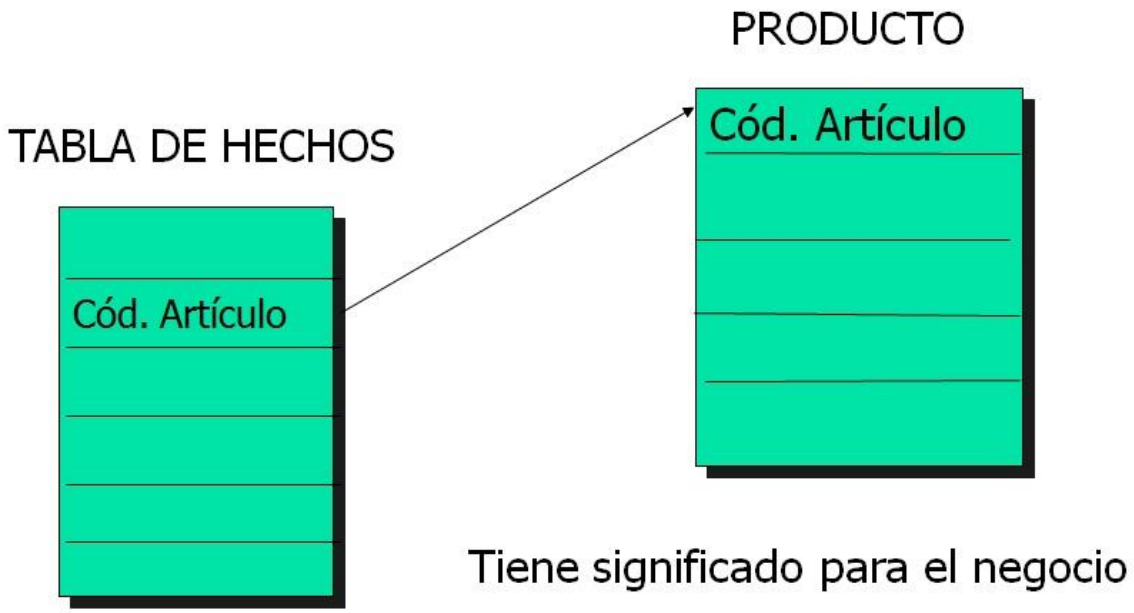


Imagen 11 Clave operativa o del negocio.

Fuente: (Castro, 2014)

Además, Castro (2014) advierte que las desventajas de una clave inteligente son que incluye lógica del negocio (ej. parte del código de artículo es el código de proveedor). Requiere el uso conjunto de 2 o más campos para identificar unívocamente a la fila (ej. código de artículo, fecha de vigencia). Es de longitud considerable (ej. alfanumérico de 15 o más posiciones), ocupa mucho espacio en la tabla de hechos. Los códigos son reutilizados en los sistemas fuente. La estructura o longitud puede cambiar con el tiempo. La forma de identificar un elemento cambia con el tiempo.

La clave Inteligente tiene dos funciones, aportar conocimiento sobre el negocio y conectar la tabla de hechos con una tabla de dimensiones.

La segunda opción que responde nuestra pregunta es generar dentro del ámbito del Data Warehouse una clave numérica sin significado para el negocio (número entero asignado en forma secuencial). Se la llama clave artificial, clave entera o ***clave subrogada*** (Castro, 2014).

Las claves subrogadas son números enteros que se asignan secuencialmente para identificar las filas de una tabla de dimensión. También llamadas clave artificial, clave entera, clave no

natural. Son claves que se mantienen en el ámbito del DW en lugar de las claves naturales extraídas de los sistemas de datos de origen. Tiene la única función de conectar la tabla de dimensiones con la tabla de hechos, sirven para identificar a cada instancia o entidad en la tabla de dimensiones (Ballard, Farrell, Gupta, Mazuela, & Vohnik, March 2006).

Una clave subrogada es una clave artificial o sintética que se utiliza como un sustituto de una clave natural. En realidad, una clave subrogada en un Data Warehouse es algo más que un sustituto de una clave natural. En un DW, una clave subrogada es una generalización necesaria de la clave de producción natural y es uno de los elementos básicos de diseño del DW. Cada unión entre tablas de dimensiones y tablas de hechos en un entorno de DW debe basarse en claves subrogadas, no claves naturales (Kimball & Ross, 2002).

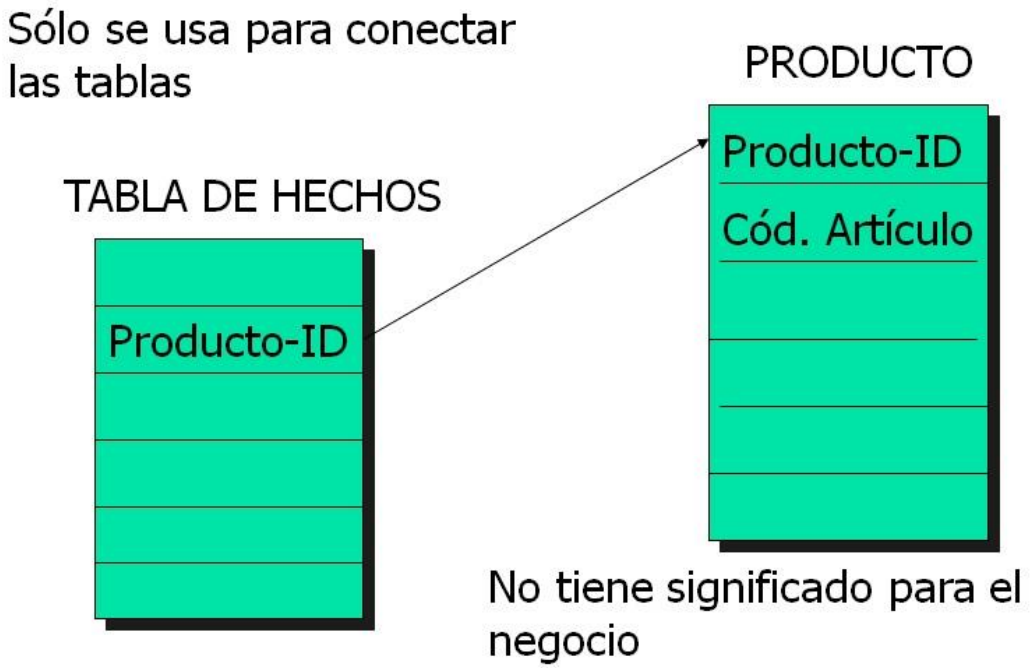


Imagen 12 Clave artificial o subrogada.

Fuente: (Castro, 2014)

Veamos algunas razones para utilizar claves subrogadas en un ambiente Data Warehouse. Las tablas de datos, en diversos sistemas de origen OLTP, pueden utilizar diferentes claves para la misma entidad. También puede ser posible que una sola clave esté siendo utilizada por

diferentes instancias de la misma entidad. Esto significa que los clientes deben ser representados con la misma clave en diferentes sistemas OLTP.

Por lo general, cuando el administrador del Data Warehouse encuentra un cambio en la descripción de un registro de dimensión como producto o cliente, la respuesta correcta es la emisión de un nuevo registro de dimensión. Pero para hacer esto, el DW debe tener una estructura de claves más generales.

Todavía hay más razones para usar claves subrogadas. Una de las más importantes es la necesidad de codificar conocimiento incierto. Puede que tenga que suministrar una clave de cliente para representar una operación, pero tal vez no sepa a ciencia cierta quién es el cliente. Esto sería una ocurrencia común en una situación de venta donde las transacciones en efectivo son anónimas, como la mayoría de las tiendas de comestibles. ¿Cuál es la clave del cliente para el cliente anónimo? Tal vez debamos introducir una clave especial que identifique a este cliente anónimo.

Podemos ahorrar espacio de almacenamiento considerable con claves subrogadas de valores enteros. Supongamos que tenemos una gran tabla de hecho con mil millones de filas de datos. En dicha tabla, cada byte desperdiciado en cada fila es un gigabyte de almacenamiento total. La ventaja de una clave de número entero de cuatro bytes es que puede representar más de 2 mil millones de valores diferentes. Eso es suficiente para cualquier dimensión. Comprimimos todos nuestros ID de clientes, nuestro stock de productos de mantenimiento y todos nuestros campos de fecha a claves de cuatro bytes. Esto ahorra muchos gigabytes de almacenamiento total (Kimball & Ross, 2002).

Para crear claves subrogadas, cada vez que vemos una clave natural en el flujo de datos de entrada, se debe buscar el valor correcto de la clave subrogada y reemplazar la clave natural con la clave subrogada. Debido a que este es un paso significativo en el proceso diario de extracción y transformación en el área de preparación de datos, tenemos que ajustar nuestras técnicas para que esta búsqueda sea simple y rápida.

La sustitución de claves grandes y naturales y claves compuestas con claves subrogadas enteras, mejoran el rendimiento. Los requisitos de almacenamiento se reducen, y las búsquedas basadas en índices son más simples. Inicialmente, puede ser más rápido para poner en práctica un modelo tridimensional usando claves operativas, pero las claves subrogadas sin duda darán sus frutos en el largo plazo (Kimball & Ross, 2002).

Por supuesto, se requiere un poco de esfuerzo para asignar y administrar claves subrogadas, pero no es tan complejo como parece en principio. Tendremos que establecer y mantener una tabla de referencia cruzada en el área de limpieza y transformación de datos (Staging Area) que se utiliza para sustituir la clave subrogada apropiada en cada fila de la tabla de hechos y dimensiones.

Uno de los principales beneficios de las claves subrogadas es que amortiguan el ambiente del Data Warehouse de los cambios operacionales. Las claves subrogadas permiten al equipo del DW mantener el control del entorno en lugar de ser condicionado por reglas operacionales para generar, actualizar, eliminar, reciclar, y reutilizar códigos de producción. En muchas organizaciones los códigos históricos operativos (por ejemplo, números de cuentas inactivas o códigos de productos obsoletos) son reasignados después de un período de inactividad. Si los números de cuenta son reciclados después de 12 meses de inactividad, los sistemas operacionales no pierden nada porque sus reglas de negocio prohíben que sus datos estén por mucho tiempo.

El Data Warehouse, por otra parte, conservara los datos durante años. Las claves subrogadas proporcionan al DW un mecanismo para diferenciar estas dos instancias separadas del mismo número de cuenta operacional. Si nos basamos únicamente en códigos de operación, también son vulnerables a la superposición de claves en el caso de una adquisición o consolidación de los datos. Las claves subrogadas permiten al equipo DW integrar datos de múltiples orígenes de sistemas operacionales, incluso si carecen de claves de origen consistentes.

Los modeladores a veces son reacios a renunciar a sus claves naturales, debido a que quieren navegar por la tabla de hechos basados en el código de operación evitando al mismo tiempo la unión con la tabla de dimensiones. Cabe recordar, que las tablas de dimensiones son nuestros puntos de entrada a los hechos. Si del quinto al noveno carácter de la clave operativa identifica al fabricante, el nombre del fabricante debe ser incluido como un atributo de tabla de dimensiones. En general, queremos evitar la incorporación de inteligencia en las claves del Data Warehouse, porque las suposiciones que hacemos pueden cambiar con el tiempo. Del mismo modo, las consultas y solicitudes de acceso a datos no deben tener ninguna dependencia incorporada en las claves (Kimball & Ross, 2002).

Finalmente (Kimball & Ross, 2002) afirman que no es aconsejable el uso de claves concatenadas o compuestas para tablas de dimensiones. No podemos crear claves subrogadas uniendo varias claves naturales o mediante la combinación de la clave natural con una marca de tiempo. Además, debemos evitar múltiples uniones entre las tablas de dimensión y de hechos porque tiene un impacto negativo en el rendimiento.

2.5 Dimensión de cambio lento

Una dimensión de cambio lento o SCD, por sus siglas en inglés (Slowly Changing Dimensions), es una dimensión donde uno o varios de sus atributos puede cambiar esporádicamente. En un modelo dimensional, los atributos de la tabla de dimensión no son fijos. Normalmente cambian lentamente durante un período de tiempo, aunque también pueden cambiar rápidamente. El equipo de diseño de modelado dimensional debe involucrar a los profesionales de negocio para ayudarles a determinar una estrategia de manejo de cambios para capturar los cambios de los atributos de dimensión. Esto describe qué hacer cuando un atributo dimensional cambia en el sistema de origen. Una estrategia de manejo de cambios implica el uso de una clave subrogada (suplente) como clave principal para la tabla de dimensión (Ballard, Farrell, Gupta, Mazuela, & Vohnik, March 2006).

El manejo de cambios en los datos dimensionales a través del tiempo puede ser complejo, los atributos dimensionales rara vez permanecen estáticos. La dirección del cliente puede cambiar, representantes de ventas van y vienen, y las empresas introducen nuevos productos para reemplazar los viejos.

Distintos tipos de estrategias manejan en forma diferente la conservación de la historia, no hay un tipo que sea mejor que otro, el profesional de negocios debe elegir cual prefiere.

2.5.1 Tipo 0

A esta técnica no se le ha dado un número de tipo en el pasado. Pero ha estado presente desde el comienzo de la SCD. Con el tipo 0, el valor del atributo de dimensión nunca cambia, por lo que los hechos siempre se agrupan por este valor original. Es apropiado para cualquier atributo con la etiqueta "original", como la puntuación de crédito original del cliente. También se aplica a la mayoría de los atributos de una dimensión fecha.

El método Tipo 0 es pasivo. Gestiona los cambios dimensionales y no se realiza ninguna acción. Los valores permanecen como estaban en el momento de insertar el registro. En determinadas circunstancias, la historia se conserva con un tipo 0 (Kimball & Ross, 2013).

2.5.2 Tipo 1

Con el tipo 1, cuando el valor del atributo cambia, sobrescribimos el valor, nos limitamos a remplazar el valor antiguo del atributo en la fila de dimensión, por el valor actual. De este modo el atributo siempre refleja la asignación más reciente.

ID	Código Producto	Descripción	…
154	267894	Yogur Dietético	…

Imagen 13 Tabla de dimensión Producto (enero de 2007).

Fuente: (Castro, 2014).

Cuando se utiliza un enfoque de tipo 1, se sustituye el valor antiguo Yogur Dietético, con el nuevo valor de Yogur BC, como se muestra en la Tabla siguiente.

ID	Código Producto	Descripción	…
154	267894	Yogur BC	…

Imagen 14 Tabla de dimensión Producto (junio de 2007).

Fuente: (Castro, 2014).

En este caso la actualización modifica todos los hechos anteriores, por lo que parecería que la descripción siempre fue Yogur BC. Si necesitamos conservar la historia exacta, un enfoque de tipo 1 puede funcionar para cambiar el nombre de la descripción, como en este caso, pero no para mantener la historia.

Es muy importante involucrar a los usuarios de negocio al implementar el enfoque de tipo-1. Aunque es el método más sencillo de implementar, puede no ser el más apropiado para todas las dimensiones que cambian lentamente ya que puede desactivar el seguimiento de la historia empresarial en determinadas situaciones.

La respuesta de tipo 1 es el método más sencillo para hacer frente a los cambios en los atributos. La ventaja es que es rápida y fácil. En la tabla de dimensiones, nos limitamos a sobrescribir el valor preexistente con la asignación actual. El problema con una respuesta de tipo 1 es que perdemos toda la historia de los cambios en los atributos. Dado que la sobre escritura borra los históricos de los valores de los atributos, nos quedamos únicamente con los valores actuales. Es apropiado si el cambio de atributo es una corrección. También puede ser adecuado si no hay ningún interés en mantener la descripción anterior. Desde el comienzo debemos determinar la importancia de mantener el valor del atributo. Con demasiada frecuencia, los equipos de proyecto utilizan un tipo 1 como la respuesta por defecto para tratar con dimensiones de cambio lento y terminan fracasando si la empresa necesita realizar un seguimiento de los cambios históricos (Kimball & Ross, 2013).

2.5.3 Tipo 2

Agregar una fila de dimensión, dijimos que uno de los objetivos del Data Warehouse es representar la historia correctamente. Un tipo 2 es la respuesta predominante para responder a este requerimiento. Cuando un valor del atributo cambia se agrega una nueva fila a la tabla de dimensiones.

Como vemos en el siguiente ejemplo, los nuevos hechos apuntan a la nueva fila de Yogur BC y los hechos anteriores continúan apuntando a la fila anterior. Después del 1 de junio de 2007 el producto tendría el ID 542 para reflejar el cambio en la descripción del yogur.

ID	Código Producto	Descripción	FechaInicio	FechaFin	FilaActual
146	264894	Yogur Dietético	15/2/2007	31/5/2007	No
542	264894	Yogur BC	1/6/2007	31/12/9999	Si

Imagen 15 Tabla de dimensión Producto (después del 1 de junio de 2007).

Fuente: (Castro, 2014)

Cuando una tabla de dimensiones incluye atributos de tipo 2, debe incluir algunas columnas para administrar las filas como se muestra en el ejemplo. La fecha de inicio y fecha de fin se

refiere al momento en que los valores de atributo de la fila tendrán validez. Podemos observar que un mínimo de tres columnas adicionales debe añadirse a la fila de dimensión con el tipo 2.

1) Fecha de inicio: es la fecha que entró en vigencia el registro actual
2) Fecha de fin: fecha en la cual el registro dejo de estar en vigencia
3) Fila actual: es el que indica si el registro actual es el vigente. Es de utilidad para obtener de forma rápida los registros actuales.

Las fechas de inicio y fin son necesarias en el sistema ETL, ya que necesita saber qué clave subrogada es válida cuando se cargan filas de hechos históricos. Cuando un nuevo producto se carga en la tabla de dimensiones, la fecha de fin se establece al 31 de diciembre de 9999. Al evitar un valor NULL en la fecha de fin, se puede utilizar de forma fiable un comando BETWEN para encontrar las filas de dimensión que estaban en vigencia en una fecha determinada.

Dado que el tipo 2 genera nuevas filas de dimensión, un aspecto negativo de este enfoque se puede dar cuando los cambios no son tan lentos. Por lo tanto, puede ser una técnica inadecuada para las tablas de dimensiones que ya superan el millón de filas (Kimball & Ross, 2002) .

2.5.4 Tipo 3

Agregar una columna de dimensión, guarda una cantidad limitada de valores históricos de atributos seleccionados. Cuando hay un cambio, nos guardamos el valor anterior en una columna distinta, actualizando el campo con el nuevo valor (para cada campo, tendremos una tupla valor anterior, valor actual). Solo nos vamos a guardar, por tanto, los dos últimos valores.

ID	Apellido	Nombre	ProvinciaOriginal	ProvinciaActual
92365	Aguirre	Mónica	Buenos Aires	Córdoba

Imagen 16 Tabla de dimensión Cliente SCD tipo 3.

Fuente: (Castro, 2014)

En comparación con SCD tipo 2, la SCD tipo 3 no aumenta el tamaño de la tabla, ya que la nueva información se actualiza mientras que todavía mantiene parte de la historia. Sin embargo,

la SCD Tipo 3 se utiliza raramente en la práctica, ya que modifica la estructura de las tablas de dimensiones (agrega más columnas). Adicionalmente, SCD tipo 3 no es capaz de mantener toda la historia cuando se cambia un atributo más de una vez, ya que mantiene sólo los valores originales y los actuales del atributo cambiado. Los valores intermedios se pierden.

El enfoque tipo 3 se utiliza típicamente sólo si hay una necesidad limitada de preservar y describir con precisión la historia. Un ejemplo es cuando alguien se casa y deseamos mantener el apellido original de la persona.

Las técnicas que cambian lentamente son adecuadas para la mayoría de las situaciones. Sin embargo, muchas veces necesitamos variaciones hibridas que se basan en estos conceptos básicos para atender análisis más sofisticados. Antes de utilizar esta técnica, debemos recordar la necesidad de equilibrar el poder analítico contra la facilidad de uso, debemos mantener el equilibrio entre la flexibilidad y la complejidad.

2.5.5 Tipo 6

Con este enfoque híbrido, se emite una nueva fila para capturar el cambio (tipo 2) y añadir una nueva columna para rastrear la asignación actual (tipo 3), donde los cambios posteriores son manejados como una respuesta de tipo 1. El nombre de este enfoque combinado tipo 6 se debe a que tanto la suma y producto de 1, 2, y 3 es igual a 6 (Ross, Slowly Changing Dimension Types 0, 4, 5, 6 and 7, 2013).

La técnica de tipo 6 tiene un atributo incorporado que es un valor alternativo de un atributo normal de tipo 2 en la dimensión base. Por lo general, un atributo de este tipo es simplemente una alternativa de tipo 3, pero en este caso se sobrescribe sistemáticamente el atributo siempre que el atributo se actualiza.

Clave Producto	SKU(NK)	Descripción Producto	Historico Nombre Departamento	Actual Nombre Departamento	Fila Fecha Vigencia	Fila Fecha Expiración	Indicador de Fila Actual
12345	ABC922-Z	IntellKidz	Educación	Educación	01/01/2012	31/12/9999	Actual

Imagen 17 Fila original en la dimensión Producto.

Fuente: Traducción de (Kimball & Ross, 2013).

Clave Producto	SKU(NK)	Descripción Producto	Historico Nombre Departamento	Actual Nombre Departamento	Fila Fecha Vigencia	Fila Fecha Expiración	Indicador de Fila Actual
12345	ABC922-Z	IntellKidz	Educación	Estrategia	01/01/2012	31/12/2012	Expirado
25984	ABC922-Z	IntellKidz	Estrategia	Estrategia	01/01/2013	31/12/9999	Actual

Imagen 18 Fila de dimensión de producto después de reasignar el departamento.

Fuente: Traducción de (Kimball & Ross, 2013).

Clave Producto	SKU(NK)	Descripción Producto	Historico Nombre Departamento	Actual Nombre Departamento	Fila Fecha Vigencia	Fila Fecha Expiración	Indicador de Fila Actual
12345	ABC922-Z	IntellKidz	Educación	Pensamiento Critico	01/01/2012	31/12/2012	Expirado
25984	ABC922-Z	IntellKidz	Estrategia	Pensamiento Critico	01/01/2013	03/02/2013	Expirado
31726	ABC922-Z	IntellKidz	Pensamiento Critico	Pensamiento Critico	04/02/2013	31/12/9999	Actual

Imagen 19 Fila de dimensión de producto después de la segunda reasignación del departamento.

Fuente: Traducción de (Kimball & Ross, 2013).

2.6 Dimensión de cambio no tan lento

Se presenta cuando la tabla de dimensiones tiene gran cantidad de filas, los atributos cambian con cierta frecuencia o hay un aumento desmedido de la cantidad de filas.

Una dimensión que cambia rápidamente puede tener un aumento mayor si se utiliza el enfoque de tipo 2. Consideremos la posibilidad de un escenario para una dimensión de cliente que tiene 100 000 filas. Supongamos que estamos manejando los cambios para este cliente utilizando el método de tipo 2. Supongamos además que en un año se producen un promedio de 10 cambios para cada cliente.

Por lo tanto, en un año el número de filas se incrementará a 100 000 x 10 = 1 000 000. Para algunas empresas, esto puede ser un número pequeño de manejar incluso usando un tipo 2. Esto significa que, para algunos, la dimensión del cliente puede ser una dimensión de cambio lento.

Ahora supongamos que una tabla de clientes tiene 10 millones de filas. Imagine el mismo escenario donde, en promedio, 10 cambios se producen para un cliente cada año. Esto significa que al final del año, la tabla crecerá a aproximadamente 100 millones de filas, es un enorme crecimiento. Tal dimensión de cliente puede ser considerada como una dimensión de crecimiento rápido. Entonces, el manejo de una dimensión de crecimiento rápido usando un enfoque Tipo-2 no es factible (Kimball, 1999).

Para identificar la razón por la cual una dimensión cambia rápidamente, hay que buscar atributos que tienen valores que varían continuamente, tales como la edad, puntuación de la prueba, tamaño, peso, historial de crédito, el estado de cuenta de cliente o los ingresos.

La peor combinación de un enfoque SCD es una dimensión muy grande (tal como una dimensión de cliente de varios millones de fila) donde el promedio de registro cambia muchas veces por año.

No existe una regla fija que diga que el enfoque SCD de crear un nuevo registro dimensional ya no es práctico. Cuando una tabla de dimensión se vuelve demasiado grande y está cambiando con tanta frecuencia que la administración y el rendimiento de las consultas son cada vez más complejos, entonces es hora de hacer algo (Kimball, 1999).

2.6.1 Tipo 4

Un método adecuado para el manejo de las dimensiones que cambian muy rápido es dividir los atributos que cambian rápidamente en una o más dimensiones separadas, llamado mini-dimensiones. La tabla de hechos tendría entonces dos claves foráneas, una para la tabla de dimensiones y otra para los atributos que cambian rápidamente. Estas tablas de dimensiones se asocian entre sí cada vez que se inserta una fila en la tabla de hechos (Ballard, Farrell, Gupta, Mazuela, & Vohnik, March 2006).

Clave Demográfica	Rango de edad	Puntuación Frencuencia de Compra	Nivel de Ingresos
1	21-25	Bajo	< $30.000
2	21-25	Medio	< $30.000
3	21-25	Alto	< $30.000
4	21-25	Bajo	$30.000 - $39.999
5	21-25	Medio	$30.000 - $39.999
6	21-25	Alto	$30.000 - $39.999
...	...	...	...
8	26-30	Bajo	< $30.000
9	26-30	Medio	< $30.000
10	26-30	Alto	< $30.000
...	...	...	...

Imagen 20 Tipo 4 filas de ejemplo mini dimensión.

Fuente: Traducción de (Kimball & Ross, 2013).

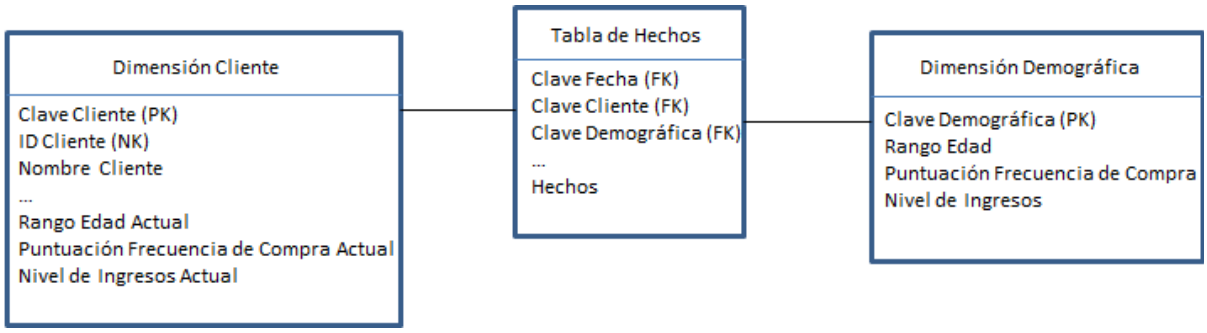


Imagen 21 Vista de la tabla de hechos y dimensiones Tipo 4.

Fuente: Traducción de (Kimball & Ross, 2013).

2.6.2 Tipo 5

Volvamos a la mini-dimensión de tipo 4. Una mejora para esta técnica es agregar una clave actual a la mini-dimensión como un atributo de la dimensión primaria. Esta referencia es un atributo de tipo 1, sobrescrito con cada cambio. No se necesita seguir este atributo como un tipo 2 porque entonces estaríamos capturando cambios volátiles en la dimensión y evitar este crecimiento fue una de las motivaciones originales de tipo 4.

La técnica de tipo 5 es útil si necesita un número de perfil actual en ausencia de métricas en la tabla de hechos o quiere enlazar hechos históricos basados en el perfil actual del cliente. Deberíamos representar la dimensión primaria y la mini-dimensión externa como una tabla simple en el área de presentación.

Para minimizar la confusión de los usuarios y el potencial error, el atributo actual en esta dimensión de roles debería tener un nombre de columna distinto para distinguirlos. Incluso con un etiquetado único, tenga en cuenta que la presentación de los usuarios con dos caminos para acceder a los datos demográficos puede ofrecer una mayor funcionalidad y complejidad de lo que algunos pueden manejar.

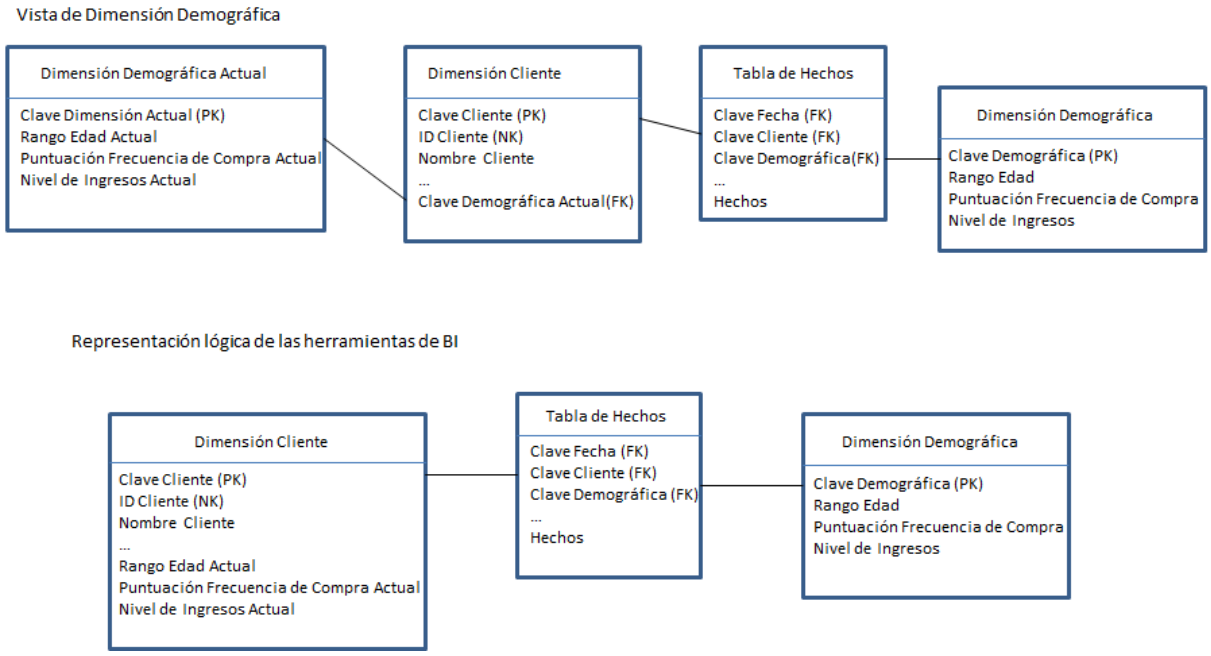


Imagen 22 Vista de la tabla de hechos y dimensiones Tipo 5.

Fuente: Traducción de (Kimball & Ross, 2013).

2.6.3 Tipo 7

Surge como respuesta a la pregunta si la técnica SCD tipo 6 sería apropiada para soportar las perspectivas actuales e históricas de 150 atributos en una tabla de dimensiones grande.

En esta técnica híbrida final, la clave natural en la dimensión (asumiendo que es duradera) se incluye como una clave foránea en la tabla de hechos, además de la clave subrogada para el tipo 2.

Si la clave natural es difícil de manejar o ha sido reasignada, se debe utilizar una clave sobrenatural permanente (clave durable, persistente y que no cambia) independiente en su lugar. La dimensión de tipo 2 contiene atributos históricamente precisos para el filtrado y agrupación basado en los valores efectivos de cuando ocurrió el hecho. La clave permanente se une a una dimensión de tipo 1 sólo con los valores actuales. Los títulos de las columnas de esta tabla deben ser precedidos de "actual" para reducir el riesgo de confusión en el usuario.

Se pueden usar estos atributos de dimensión para resumir o filtrar hechos basados en el perfil actual, independientemente de los valores de los atributos en vigencia cuando se produjo el evento. Este enfoque ofrece la misma funcionalidad que el tipo 6. Aunque la respuesta de tipo 6 genera más columnas de atributos en una sola tabla de dimensiones, este enfoque se basa en dos claves foráneas de la tabla de hechos. Tipo 7 siempre requiere menos esfuerzo en el ETL debido a que la tabla actual de atributos de tipo 1 podría fácilmente ser entregada a través de una vista de la tabla de dimensiones de tipo 2, limitado a las filas más actuales. El costo adicional de esta última técnica es la columna adicional realizada en la tabla de hechos, sin embargo, las consultas sobre la base de valores de los atributos actuales se filtran en una tabla de dimensiones más pequeña.

Utilizada para soportar informes de cómo-era y como-es. La tabla de hechos se puede acceder a través de una dimensión modelada tanto como una dimensión de tipo 1, que muestra sólo los valores de los atributos más actuales, o como una dimensión de tipo 2 que muestra los perfiles históricos. La misma tabla de dimensiones permite ambas perspectivas. Tanto la clave permanente y la clave subrogada primaria de la dimensión se colocan en la tabla de hechos.

Para la perspectiva de tipo 1, la bandera actual de la dimensión se restringe a ser actual, y la tabla de hechos se une a través de la clave permanente. Para la perspectiva de tipo 2, la bandera actual no está restringida, y la tabla de hechos se une a través de la clave primaria subrogada. Estas dos perspectivas se desplegarían como vistas separadas a las aplicaciones de BI (Kimball & Ross, 2013).

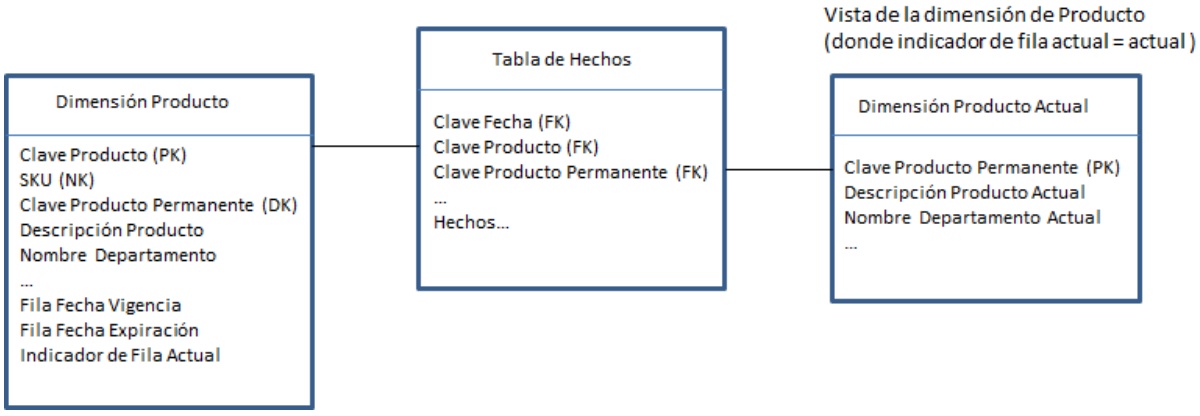


Imagen 23 Vista de la tabla de hechos y dimensiones Tipo 7.

Fuente: Traducción de (Kimball & Ross, 2013).

Clave Producto	SKU(NK)	Clave Producto Permanente	Descripción Producto	Nombre Departamento	...	Fila Fecha Vigencia	Fila Fecha Expiración	Indicador de Fila Actual
12345	ABC922-Z	12345	IntellKidz	Educación	...	1/1/2012	31/12/2012	Expirado
25984	ABC922-Z	12345	IntellKidz	Estrategia	...	1/1/2013	30/6/2013	Expirado
31726	ABC922-Z	12345	IntellKidz	Pensamiento Critico	...	1/7/2013	31/12/9999	Actual

Imagen 24 Filas en la Dimensión Producto.

Fuente: Traducción de (Kimball & Ross, 2013).

Clave Producto	SKU(NK)	Clave Producto Permanente	Descripción Producto Actual	Nombre Departamento Actual	...
12345	ABC922-Z	12345	IntellKidz	Pensamiento Critico	...
25984	ABC922-Z	12345	IntellKidz	Pensamiento Critico	...
31726	ABC922-Z	12345	IntellKidz	Pensamiento Critico	...

Imagen 25 Filas en la vista de Dimensión de Producto Actual.

Fuente: Traducción de (Kimball & Ross, 2013)

Capítulo 3 Planteamiento del problema

3.1 Descripción del problema

La Universidad Nacional de La Matanza (UNLaM) utiliza un sistema informático para administrar los datos de los aspirantes del curso de ingreso llamado "UNLaM gestión de ingreso", la modalidad de ingreso a nuestra casa es por medio de la aprobación de tres o cuatro asignaturas, de acuerdo con el Departamento académico de interés, según lo establecido por la Resolución HCS Nro. 50/2015. Dicho sistema registra datos de los aspirantes e información referente a su paso como ingresante, como materias, calificaciones obtenidas y aulas de cursada.

La información que se requiere es obtenida a través de consultas a la base de datos y elaboración de reportes que significan un alto costo para los desarrolladores y administradores de base de datos. Este procedimiento en muchos casos no llega a cubrir las necesidades analíticas de la Secretaría Académica.

El problema que voy a abordar tiene como objetivo permitir realizar un análisis y seguimiento de los aspirantes en su fase de ingreso. Esto permitirá tener información necesaria para analizar, por ejemplo, porcentaje de aprobados por materia, cantidad de inscriptos, porcentaje de desaprobados etc. Este tipo de análisis ayudara a la Secretaría Académica a tener una mirada más enfocada en determinadas variables que se pueden mejorar para brindar un mejor ingreso y/o sugerir a los distintos representantes educativos hacer hincapié en determinadas temáticas que hacen a la mejora de nuestro sistema educativo.

La razón de mi elección se debe a que existen sistemas Data Warehouse como el SIU-Wichi para los alumnos en las carreras universitarias que se alimentan del SIU-Guaraní[2] además de otras soluciones particulares que se realizaron puntualmente para visualizar información analítica que el SIU-Wichi no suministraba[3], pero hasta el momento se carece de este tipo de

[2] SIU - Sistema de Información Universitaria https://portal.comunidad.siu.edu.ar/el-siu
[3] Implementación de un Data Warehouse para la toma de decisiones en el Área Académica. https://repositoriocyt.unlam.edu.ar/handle/123456789/307

tecnologías para el análisis del ingreso. Nuestra Universidad alberga alumnos principalmente del Partido de La Matanza donde las realidades son muy diversas y es imprescindible contar con datos de todas las instancias del alumno. Dadas nuestras características sociales la importancia de contar con una universidad y el efecto socializador que significa para los alumnos de menores recursos principalmente.

Deberíamos ser capaces de articular la educación con los gobiernos provinciales y nacionales de acuerdo con las problemáticas de nuestros lugares y poder fijar políticas específicas para problemas concretos. No es posible realizarlo sin voluntad política y es imprescindible, pero tampoco podemos hacer esta tarea sino contamos con información para el análisis.

Por lo tanto, contar con información en tiempo y forma para que tanto desde Rectorado como la SA y Departamento de Alumnos pueda analizarla mejorara la toma de decisiones. Entonces podemos decir que la información adecuada, en el plazo adecuado, para las personas adecuadas redundara en mejores decisiones.

La SA es un área vital en nuestra institución porque tiene, entre otras tareas, la de organizar y coordinar todo lo que concierne a los alumnos y docentes. Nuestras instituciones educativas no tendrían razón de ser sin estos dos actores fundamentales. Podemos decir entonces, que la importancia del área en la estrategia de mejora en la gestión es muy grande. En la actualidad no podemos estar ajenos o desconocer las tecnologías de la información IT y las ventajas que brindan para cumplir con nuestros objetivos. Por ello, necesitamos trabajar en conjunto con la Secretaría de Informática y Comunicaciones (SIC), para poder realizar con éxito el proyecto.

Tanto la Comisión Nacional de Evaluación y Acreditación (CONEAU), como los distintos programas de mejora de la enseñanza requieren información para sus procesos de evaluación. Acorde a esta necesidad la Universidad Nacional de La Matanza reconoce la importancia de contar con un repositorio que permita de una manera práctica realizar el análisis de la información.

A partir de lo expuesto se decide utilizar una arquitectura acorde a las necesidades que permita dar respuesta a los requerimientos de los usuarios. Dicha arquitectura es denominada OLAP, que es una técnica por la cual los datos de un almacén de datos son empleados por usuarios de distintas áreas para satisfacer sus necesidades analíticas.

El proyecto se basa en la construcción de un Data Warehouse que es una base de datos orientada al análisis de la información histórica contenida en ella, que permitirá de una manera práctica, a través de un archivo Excel, realizar el análisis de la información. Además, utilizo Power BI para realizar un tablero de control (dashboard) que es la mejor manera para que los usuarios vean sus datos. Como veremos los tableros de control de Power BI son más atractivos visualmente, interactivos y personalizables.

Los resultados que espero obtener con el desarrollo del proyecto son los siguientes:

- Facilitar la información necesaria para la toma de decisiones en el contexto educativo y, en este caso particular, con los aspirantes del curso de ingreso.

- Colaborar para mejorar la calidad de las decisiones tomadas. Es vital para el proyecto convertir la información en conocimiento para el área académica.

- Brindar una herramienta eficiente y eficaz para analizar políticas a mediano y largo plazo que permitan mejorar los procesos y servicios de nuestra institución.

- En muchos casos, para realizar un informe necesitamos dedicar un tiempo considerable porque debemos buscar datos en distintas fuentes, ordenarlos, darle un sentido. Entonces, es uno de los objetivos del proyecto, reducir los tiempos en la elaboración de informes para el análisis de la información.

- Proporcionar un DW capaz de mantener un histórico de datos para el análisis.

- Brindar una solución práctica, que incentive el uso de DW, para el análisis de los alumnos a lo largo de su vida académica.

3.2 Abordaje de elementos estructurales, legales, reglamentarios relacionados con la iniciativa propuesta

Para la realización del proyecto necesitamos trabajar con datos relativos a los aspirantes, por tratarse de datos sensibles, se deberá notificar al Secretario Académico, a fin de autorizar la utilización de los datos de los aspirantes pertenecientes a la instancia de ingreso. Además, se encuentran protegidos por la ley nacional 25326 de protección de los datos personales[4]. Por otra parte, el Ingeniero en Informática tiene el derecho y la obligación de guardar secreto profesional respecto de la información obtenida en el transcurso del ejercicio de la profesión más allá del cumplimiento de la normativa vigente en materia de protección de datos de carácter personal.

La Secretaría de Informática y Comunicaciones cumple con la normativa vigente de registro ante el Registro Nacional de Bases de Datos, perteneciente a la Agencia de Acceso a la Información Pública tal como establece la ley 25326 en su art. 3. Dicho organismo garantiza el efectivo ejercicio del derecho de acceso a la información pública y a la protección de los datos personales.

3.3 Parámetros humanos, financieros, físicos de la iniciativa

Requerimientos Personales: conocimientos informáticos, bases de datos OLTP (On-Line Transactional Processing) y OLAP (On-Line Analytical Processing), habilidades para analizar requerimientos, diseñar modelos. Actitudes y competencias para gestionar los recursos, trabajar en equipo, buena comunicación e iniciativa para alcanzar los objetivos.

[4] Ver Anexo

Requerimientos institucionales: coordinar con el área Académica la obtención de las distintas fuentes de datos de los aspirantes, coordinar con la SIC la sesión de la infraestructura necesaria para el desarrollo del proyecto como herramientas y posibilidad de consultas con docentes especializados en el tema.

Para la iniciativa se necesitará un desarrollador, un administrador de base de datos y un analista. En cuanto a los requerimientos técnicos necesitamos, un servidor con sistema operativo Windows Server 2016 o superior y Microsoft SQL Server 2014 para el ambiente de desarrollo. La misma configuración para el ambiente de producción que será utilizado una vez que el proyecto ha sido aprobado para ser puesto en marcha.

Ambiente local para el desarrollador, con una PC con procesador i7 memoria RAM 8,00 GB, Microsoft Visual Studio 2019. Microsoft Project 2010 o superior.

Con respecto a los usuarios, necesitaran Microsoft Office Excel 2013 o superior y un navegador web con conexión a internet.

Capítulo 4 Solución

4.1 Solución propuesta

En este trabajo considero que la metodología de Kimball Bottom-Up mostrada en el capítulo 2 es la mejor opción porque se adapta a la naturaleza de nuestra universidad. Ya que plantea que podemos implementar pequeños DM en áreas específicas de las mismas, en este caso la Secretaría Académica, con pocos recursos y luego ir integrándolos en un gran Data Warehouse. Esta metodología nos permite tener una solución a mediano plazo a partir del DM que requieren un diseño de menor complejidad que un repositorio de datos global. Además, la implementación inicial suele ser menos costosa en términos de hardware y otros recursos.

La construcción de un DW tiene una complejidad que debemos organizar para poder garantizar el éxito de nuestro objetivo. La solución debe ser dirigida por un proceso eficiente y eficaz que brinde una estrategia a seguir. La metodología de Ralph Kimball que se muestra en la ***Imagen 26***, me permite simplificar dicha complejidad por lo que considero la metodología adecuada a seguir.

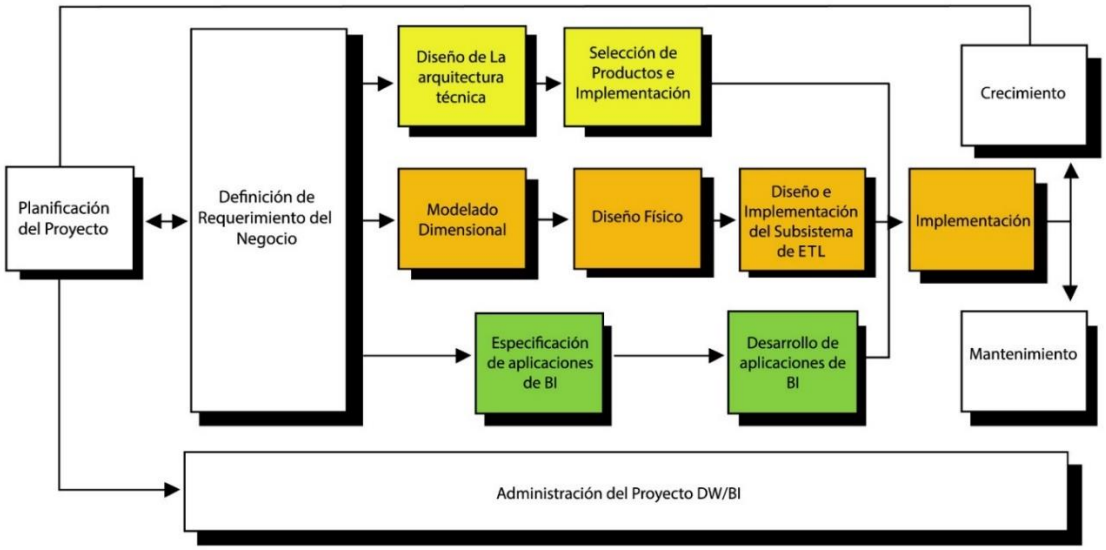


Imagen 26 Metodología de Kimball.

Fuente: Traducción de (Kimball & Ross, 2013)

Podemos ver en la ***Imagen 26*** que los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto.

Podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas:

- Tecnología (Camino Superior). Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.
- Datos (Camino del medio). En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de ETL para cargar el DW.
- Aplicaciones de Inteligencia de Negocios (Camino Inferior). En esta ruta se encuentran tareas en las que diseñamos y desarrollamos las aplicaciones de negocios para los usuarios finales (Kimball & Ross, 2013).

4.2 Planificación del Proyecto

El primer punto de la metodología es la planificación del proyecto. En este proceso se determinan las actividades, el alcance, los objetivos, los Stakeholders[5] y se planifica la gestión de riesgos.

Kimball en (Kimball & Ross, 2013) describe tres factores en orden de importancia que podemos utilizar para tener éxito en nuestro DW. El primero y más crítico es tener un patrocinador o referente que tenga una visión clara del impacto potencial del DW en la organización. Deben ser líderes políticamente astutos que puedan convencer a sus pares para que apoyen el esfuerzo. El segundo factor de preparación es tener una motivación sólida y convincente para abordar la iniciativa del DW. Este factor está directamente ligado con el patrocinador. El tercer factor al evaluar la preparación es la viabilidad. Hay varios aspectos de la viabilidad, incluida la viabilidad técnica y de recursos, pero la viabilidad de los datos es el más crucial. ¿Está recopilando datos reales en sistemas fuente operativos reales para respaldar los requisitos? La viabilidad de los datos es una preocupación importante porque no existe una solución a corto plazo si aún no está recopilando datos de origen razonablemente limpios con la granularidad correcta.

Dada la envergadura del trabajo considero que están cubiertos los dos primeros factores debido a que el patrocinador o referente del proyecto es el Secretario Académico además de ser el usuario principal. De surgir un proyecto más grande, es decir nuevos DM, se deberán revisar estos criterios. Con respecto al último factor, nuestros datos de origen son viables y contamos con la granularidad necesaria para el proyecto.

Si bien en el capítulo 3 se detallaron los parámetros necesarios, utilizaremos este espacio para dar una mayor descripción técnica.

4.2.1 Etapas y tareas

Definimos las etapas y tareas con una breve descripción que luego son expuestas en el Gantt de la ***Imagen 27***.

Proyecto: Data Warehouse Académicas

[5] Stakeholder es una persona que tiene una participación directa o indirecta en la organización.

Requerimientos: Elaboración del documento surgido a partir de las entrevistas con el Secretario Administrativo donde se establecen los requerimientos para determinar el alcance del proyecto.

Entrevistar al Secretario Académico

Validar requerimientos con el Secretario Académico

Definir el alcance del proyecto

Configuración de las herramientas: Preparar el ambiente de desarrollo con las herramientas necesarias para el proyecto. Visual Studio requiere descargar las extensiones de Integration Services (SSIS), necesarias para realizar el ETL y Analysis Services (SSAS) para realizar los cubos.

Instalar SQL Server 2014

Instalar Visual Studio 2019 configurar SSIS y SSAS

Modelo Dimensional: Realización de los diagramas estrella con las dimensiones de la tabla de hechos.

Diseño Físico: Realizar las tablas con sus respectivos campos y relaciones en SQL Server. **Carga de Datos, Diseño y Desarrollo ETL:** Realizar la extracción de los datos desde MySQL, realizar las transformaciones necesarias para la carga en las tablas SQL Server.

Extraer los datos de la Base de Datos MySQL

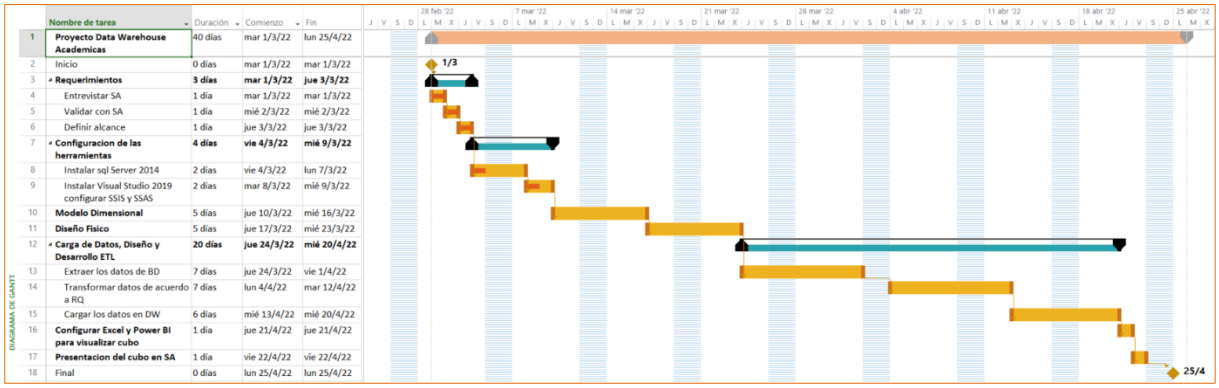
Transformar datos de acuerdo con los requerimientos RQ

Cargar los datos en DW

Configurar Excel y Power BI para visualizar cubo: Configurar Excel para conectar con los cubos. Implica administrar permisos de SQL Server y permisos Power BI para visualización.

Presentación del cubo en SA: Mostrar la visualización de los cubos resultantes al Secretario Académico.

A partir de las etapas y tareas establecidas realizamos el diagrama de Gantt para visualizar su desarrollo y los tiempos estimados.



	Nombre de tarea	Duración	Comienzo	Fin
1	**Proyecto Data Warehouse Academicas**	40 días	mar 1/3/22	lun 25/4/22
2	Inicio	0 días	mar 1/3/22	mar 1/3/22
3	**Requerimientos**	**3 días**	**mar 1/3/22**	**jue 3/3/22**
4	Entrevistar SA	1 día	mar 1/3/22	mar 1/3/22
5	Validar con SA	1 día	mié 2/3/22	mié 2/3/22
6	Definir alcance	1 día	jue 3/3/22	jue 3/3/22
7	**Configuracion de las herramientas**	**4 días**	**vie 4/3/22**	**mié 9/3/22**
8	Instalar sql Server 2014	2 días	vie 4/3/22	lun 7/3/22
9	Instalar Visual Studio 2019 configurar SSIS y SSAS	2 días	mar 8/3/22	mié 9/3/22
10	**Modelo Dimensional**	5 días	jue 10/3/22	mié 16/3/22
11	**Diseño Fisico**	5 días	jue 17/3/22	mié 23/3/22
12	**Carga de Datos, Diseño y Desarrollo ETL**	**20 días**	**jue 24/3/22**	**mié 20/4/22**
13	Extraer los datos de BD	7 días	jue 24/3/22	vie 1/4/22
14	Transformar datos de acuerdo a RQ	7 días	lun 4/4/22	mar 12/4/22
15	Cargar los datos en DW	6 días	mié 13/4/22	mié 20/4/22
16	**Configurar Excel y Power BI para visualizar cubo**	1 día	jue 21/4/22	jue 21/4/22
17	**Presentacion del cubo en SA**	1 día	vie 22/4/22	vie 22/4/22
18	Final	0 días	lun 25/4/22	lun 25/4/22

Imagen 27 Cronograma de trabajo.

Fuente: Elaboración Propia.

4.2.2 Consideraciones sobre el tamaño del sistema

De acuerdo con la infraestructura necesaria considero a la solución como pequeña de acuerdo con las siguientes características:

Volumen de datos: de 100 GB a 1 TB

Complejidad de análisis e informes: entre simple y media debido a que se utilizaran declaraciones SELECT relativamente sencillas y en algunos casos consultas que incluyen agregaciones o muchas uniones.

Número de usuarios: 10 en total y 3 a 5 concurrentes.

Disponibilidad: Se trata de la capacidad del DW a ser accesible y utilizable por los usuarios autorizados cuando estos lo requieran. Supone que la información pueda ser recuperada en el momento en que se necesite, evitando su pérdida o bloqueo. El servidor utilizado contara con backups full los días lunes e incremental diario. Al ser virtuales se tiene la posibilidad de realizar snapshot[6] con 8 puntos de restauración. CTI (Centro de Tecnología Informática) además cuenta con un plan de recuperación de desastres o DRP el cual permite reanudar el trabajo rápidamente después de un incidente no planificado por medio de una réplica de los servidores en otra ubicación. Con lo que respecta a las bases de datos, se realizaran backups full diarios. Por ello estimo que ante algún incidente con el servidor o motor de base de datos el tiempo necesario para restablecer el servicio no será mayor a 8 horas.

[6] Conservan el estado y los datos de una máquina virtual en el momento que crea dicha snapshot

Si bien es de suma importancia la utilización del DW no la considero critico o de alta disponibilidad, en casos de inactividad por cuestiones de mantenimiento como trabajos de carga, actualizaciones y demás tareas que puedan ser planificadas, se efectuaran fuera de los horarios habituales de trabajo en UNLaM.

4.2.3 Elementos que componen el proyecto

A continuación, en la ***Imagen 28*** vemos un diagrama con las fases de del proceso de transformación para nuestra solución. En primer lugar, se realizará la operación de extracción, transformación y carga (ETL) desde nuestra fuente de datos origen (MySQL), situadas en el área operacional, a una base de datos que se encuentra en el área de integración (SQL Server), utilizando para ello paquetes de los Servicios de Integración (SQL Server Integration Services o SSIS), la cual realiza también tareas de depuración de datos. Luego pasaremos a la fase de construcción del cubo, que desarrollaremos utilizando las herramientas de los servicios de análisis (SSAS). Finalmente, llegaremos a la fase de acceso a los cubos por parte de los usuarios finales, Microsoft Excel y Power BI.

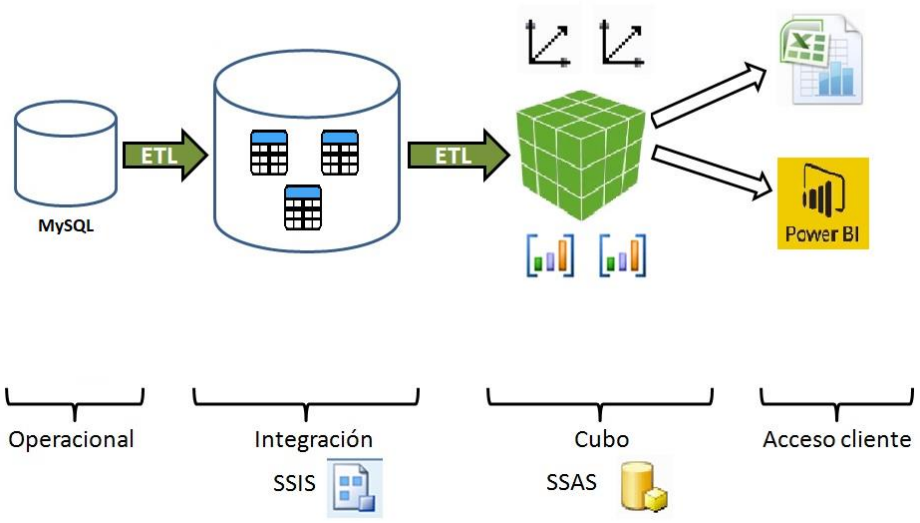


Imagen 28 Fases del Proceso de Transformación.

Fuente: Elaboración Propia.

4.2.4 Topología de servidor

En principio, dada la baja complejidad y costos se elige una topología single-servidor (Microsoft, 2017).

Imagen 29. Arquitectura Single-Servidor.

Fuente: (Microsoft, 2017)

En la arquitectura de BI de single servidor los componentes del servidor son:

Motor de base de datos de SQL Server: Este componente se usa para guardar el Data Warehouse y la base de datos intermedia (Staging Area). Además, el Agente SQL Server se puede utilizar para automatizar la ejecución del paquete SSIS y otras operaciones mediante la creación de trabajos y programas que se almacenan en la base de datos msdb .

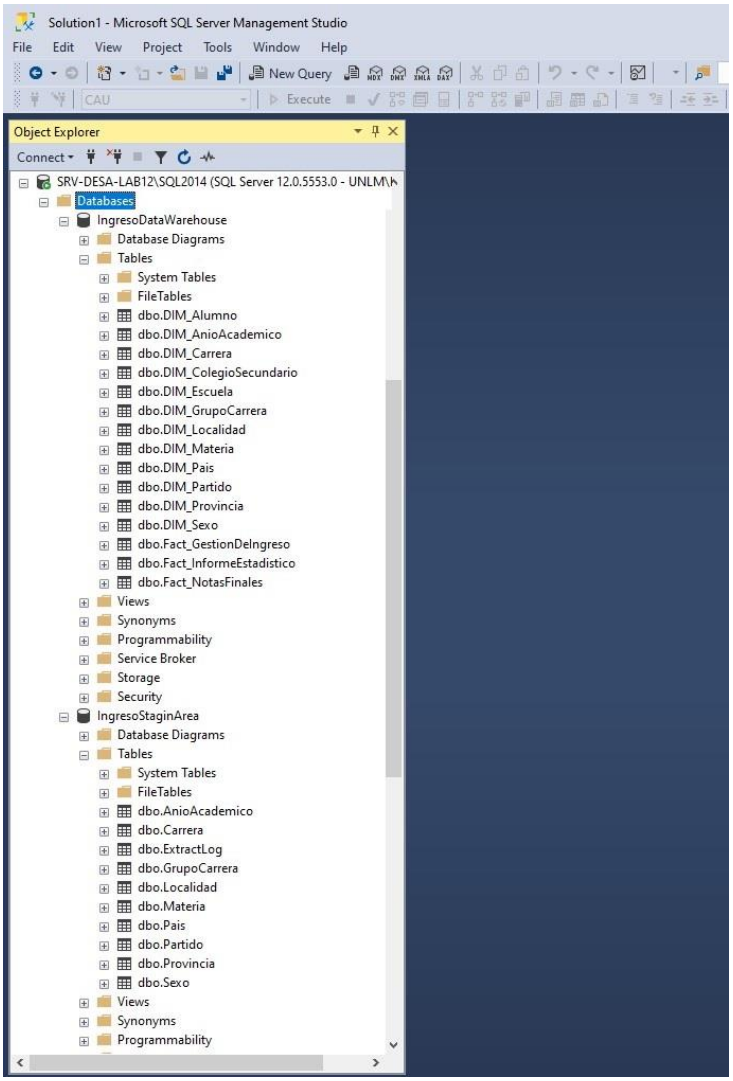


Imagen 30 Motor SQL Server 2014 conectado desde SQL Server Management Studio.

Fuente: Elaboración Propia.

Servicios de integración de SQL Server: Este componente se utiliza para ejecutar paquetes que encapsulan las tareas ETL y los flujos de datos para extraer datos de los sistemas de origen a la base de datos provisional y luego cargarlos en el almacén de datos.

En la ***Imagen 31*** se muestra la ventana de un proyecto *SSIS donde* las transformaciones se estructuran en forma de paquetes (*packages*). Un paquete es el equivalente a un programa ejecutable. En la parte izquierda de la ventana se muestran los elementos que se pueden añadir a un paquete (se adaptan según la parte que estemos definiendo), en la parte principal (parte superior central) se estructuran y relacionan los elementos seleccionados, la parte inferior central está dedicada a gestionar las conexiones de datos para acceder a las fuentes o destinos de los datos y, en la parte derecha están, entre otros componentes, el *Explorador de soluciones*, y la ventana *Propiedades* donde se pueden configurar las propiedades de los elementos que se definen.

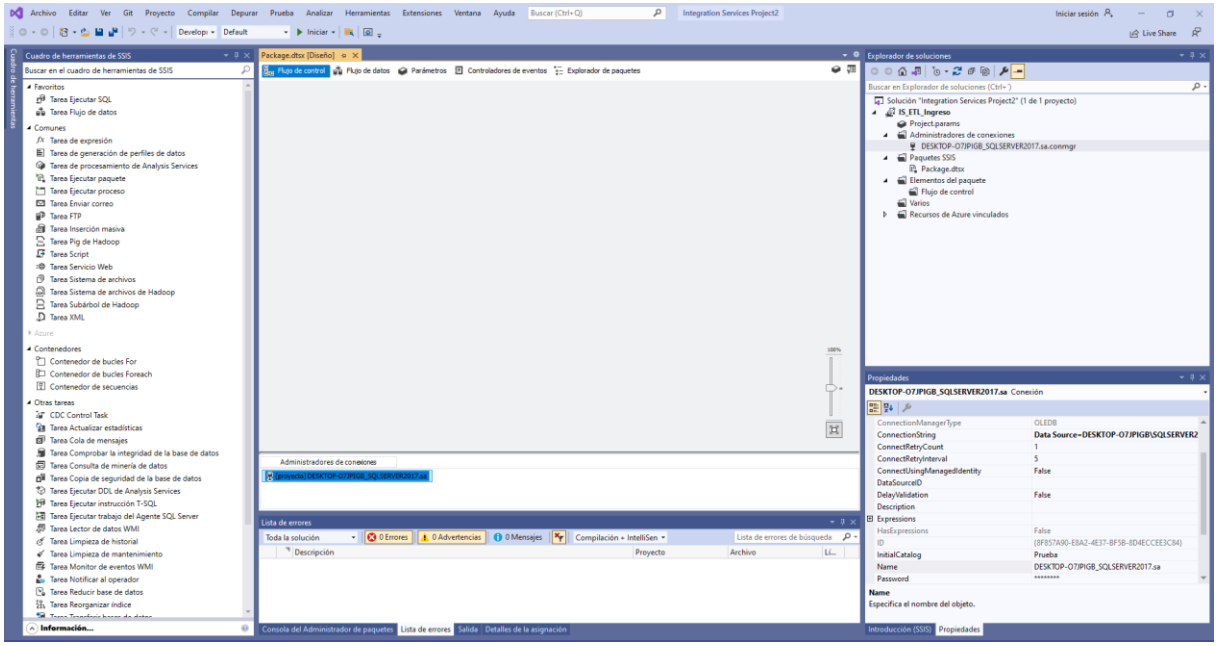


Imagen 31 Inicio del proyecto Integration Service en Visual Studio NET.

Fuente: Elaboración Propia.

Servicios de análisis de SQL Server: Este componente se utiliza para proporcionar modelos de datos analíticos y funcionalidad de minería de datos.

En la ***Imagen 32*** vemos un proyecto de inicio de Servicio de Análisis (Analisis Services) en Visual Studio NET. En la parte derecha de la ventana se encuentran las ventanas Explorador de soluciones, donde definiremos los componentes del proyecto, y Propiedades, donde se muestran las propiedades del elemento seleccionado. En la parte central está el área de trabajo, donde se mostrará el contenido de los elementos del proyecto con los que estemos trabajando.

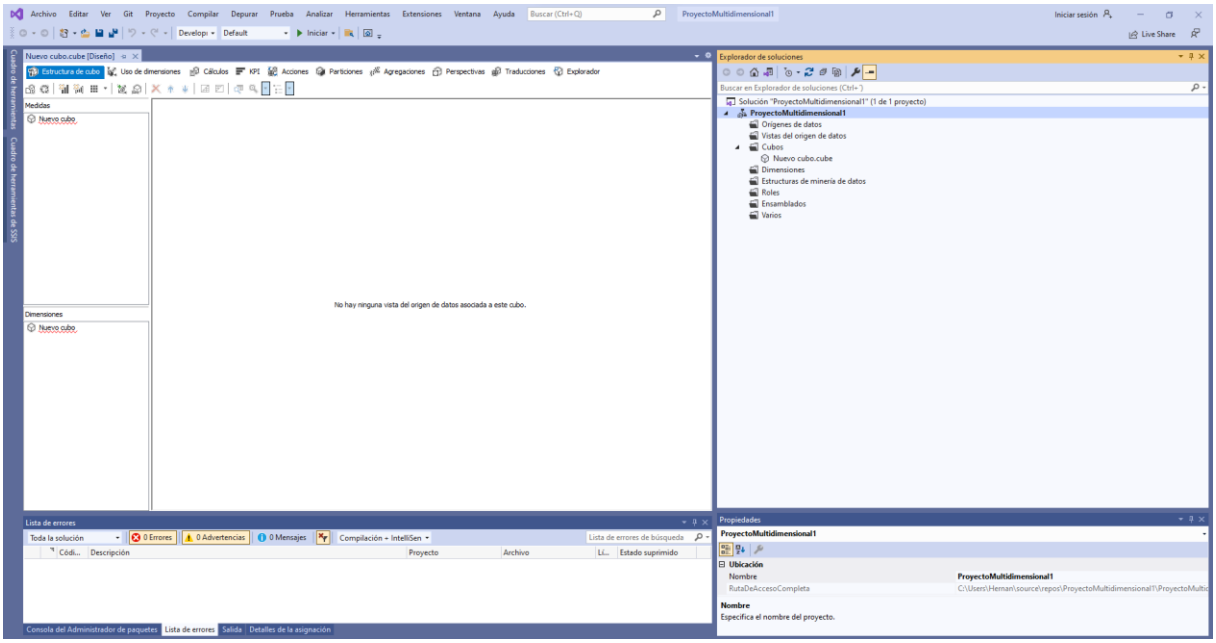


Imagen 32 Proyecto de Servicios de Análisis (Analysis Services) en Visual Studio NET.

Fuente: Elaboración Propia.

4.2.4 Requerimientos de memoria y procesador

Procesador:

La fórmula específica para aplicar es:

((Tamaño medio de la consulta en MB ÷ MCR) x Usuarios simultáneos) ÷ Tiempo de respuesta objetivo (Microsoft, 2017).

Supongamos que el MCR[7] del núcleo de la CPU que vamos a utilizar es de 200 megabytes por segundo (Mbps). Si se espera que una consulta promedio devuelva 18.000 MB, el número anticipado de usuarios simultáneos es 5 y cada consulta debe responder en 60 segundos, el cálculo para encontrar el número de núcleos que se requieren es:

[7] Representa un ancho de banda de E/S máximo estimado para el servidor, la CPU y la carga de trabajo.

((18000 ÷ 200) x 5) ÷ 60 = 7.5

Se necesitan mínimo 8 núcleos

Memoria: vamos a considerar un mínimo de 4 GB por núcleo. En nuestro caso serían 32GB necesarios.

4.2.5 Estimación de datos de hechos iniciales

Utilizamos una estimación conservadora, de 100 bytes por fila y un almacén de datos de 200.000 filas de hechos, cada una de 100 bytes de longitud, tendrá un volumen de datos de hechos inicial de aproximadamente 20MB (Microsoft, 2017) .

4.2.6 Crecimiento de datos

Los datos de hecho en nuestro almacén de datos representan inscriptos, estimamos en cada inscripción 50.000 y se realizan 2 instancias anuales, tomamos nuevamente la longitud de la fila de hechos de 100 bytes de datos. Esto equivale a una tasa de crecimiento de datos de 10MB por año (Microsoft, 2017).

4.3 Requerimientos del negocio

Los requerimientos especifican qué es lo que el sistema debe hacer (sus funciones) y sus propiedades esenciales y deseables. La captura de los requerimientos tiene como objetivo principal la comprensión de lo que los clientes y los usuarios esperan que haga el sistema. Un requerimiento expresa el propósito del sistema sin considerar como se va a implantar. En otras palabras, los requerimientos identifican el porqué del sistema, mientras que el diseño establece el cómo del sistema. La captura y el análisis de los requerimientos del sistema es una de las fases más importantes para que el proyecto tenga éxito (Norris & Rigby, 1994).

Para el proceso de recolección de **requerimientos del negocio** se realizaron entrevistas al Secretario Académico. Además de ello se recopilo información, bases de datos transaccionales, reportes históricos y estadísticas del sistema actualmente utilizado. También se entrevistó a los encargados del sistema "Gestión de Ingreso" que pertenecen al área de Sistemas Académicos.

Conforme a la información recolectada podemos definir los siguientes requerimientos:

1. Promedio de inscriptos hombres/ mujeres por materia, carrera y año académico.
2. Cantidad de inscriptos hombres/ mujeres por materia, carrera y año académico.
3. Cantidad de inscriptos por materia, carrera y año académico.
4. Cantidad de inscriptos: primario / secundario incompleto / secundario completo por materia, carrera y año académico. Excepcionalidad por articulo N° 7 de la Ley de Educación Superior.
5. Cantidad de inscriptos con terciario por materia, carrera y año académico.
6. Cantidad de inscriptos Universitario incompleto/ Universitario completo por materia, carrera y año académico.
7. Cantidad de inscriptos que trabajan/no trabajan por materia, carrera y año académico.
8. Cantidad de inscritos que estudiaron en escuelas privadas o públicas.
9. Cantidad de inscriptos argentinos/extranjeros por materia, carrera y año académico.
10. Cantidad promocionados por materia, carrera y año académico.
11. Cantidad de ausentes por materia, carrera y año académico.
12. Cantidad desaprobados por materia, carrera y año académico.
13. Cantidad aprobados por materia, carrera y año académico.

4.4 Modelo dimensional

La creación del modelo dimensional es muy importante ya que es la base de como crearemos el modelo de DW, la metodología de Kimball lo establece ampliamente.

El proceso iterativo consta de 4 pasos:

a) Elegir el proceso de negocio

b) Establecer el nivel de granularidad

c) Elegir las dimensiones

d) Identificar las tablas de hechos y medidas

En el trabajo realizado a partir de los requerimientos diferenciamos dos modelos (cubos) que los llamamos Informe estadístico y Notas Finales.

4.4.1 Informe estadístico

Medida

Promedio de hombres

Promedio de mujeres

Cantidad de hombres

Cantidad de mujeres

Cantidad de inscriptos

Cantidad primario

Cantidad secundario incompleto

Cantidad secundario completo

Cantidad terciario

Cantidad Universitario incompleto

Cantidad Universitario completo

Cantidad Trabaja

Cantidad No Trabaja

Cantidad Escuela privada

Cantidad Escuela publica

Dimensión

Grupo de Carrera

Carrera

Año académico

4.4.2 Notas finales

Medida

Cantidad de promocionados

Cantidad Ausentes

Cantidad desaprobados

Cantidad aprobados

Dimensiones

Grupo de Carrera

Carrera

Materia

Año académico

Realizamos los diagramas lógicos, esquema estrella de cada modelo propuesto, donde tenemos por un lado la tabla de hechos que contiene los datos para el análisis y las dimensiones que almacenan detalles acerca de los hechos.

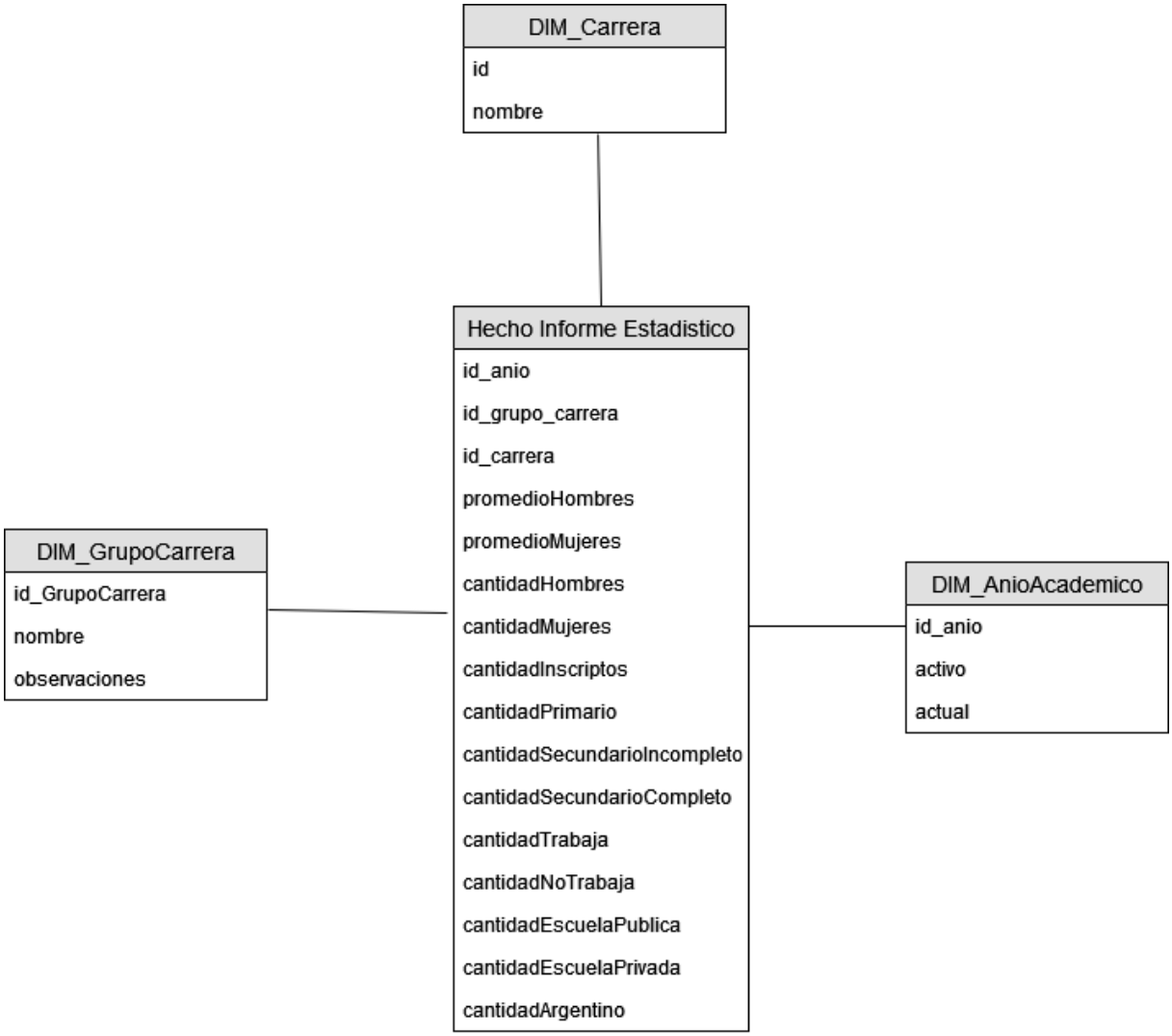


Imagen 33 Esquema estrella del cubo Informe Estadístico.

Fuente: Elaboración Propia.

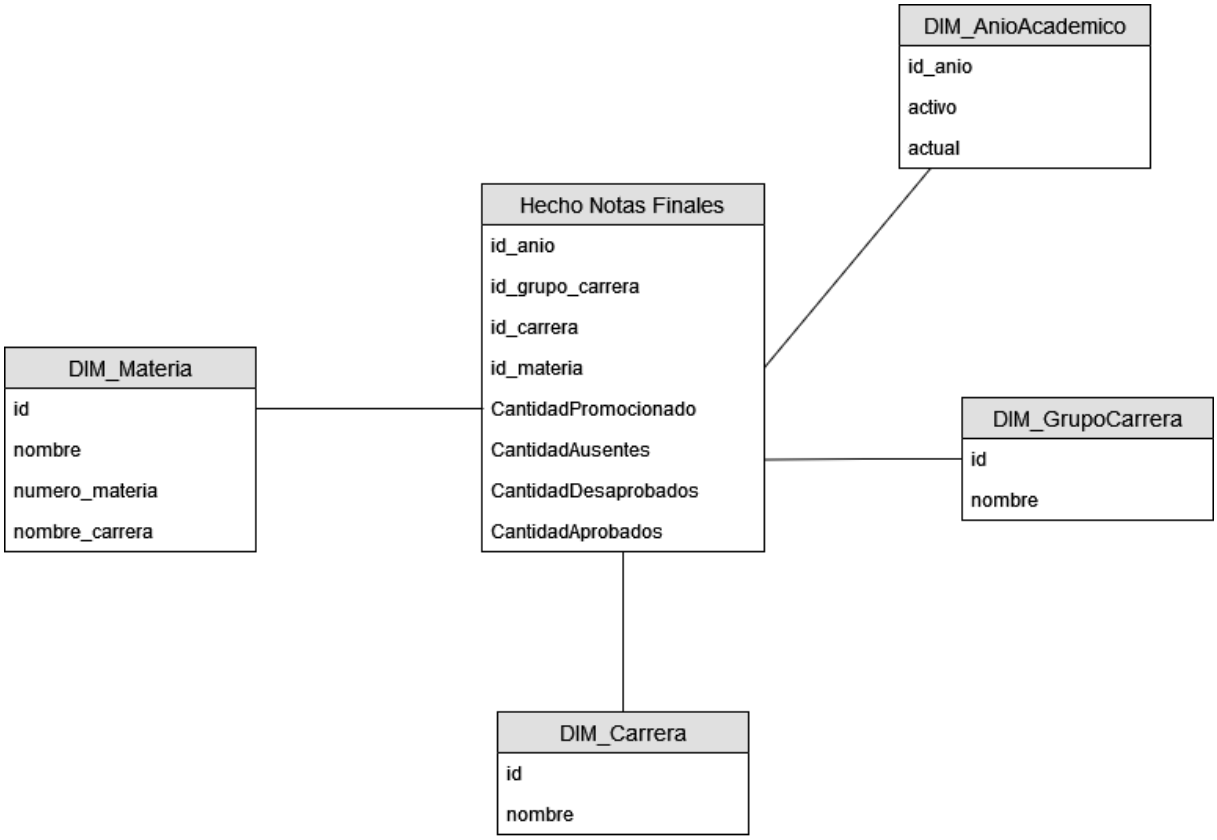


Imagen 34 Esquema estrella del cubo Notas Finales.

Fuente: Elaboración Propia.

4.5 Diseño Físico

En esta fase de la metodología se realiza la implementación del modelo dimensional construido con anterioridad al modelo físico. Preparamos las estructuras necesarias para soportar el diseño lógico propuesto. Utilizamos Microsoft SQL Server Management Studio creando las tablas con sus respectivos campos y relaciones como se muestra en la ***Imagen 35***.

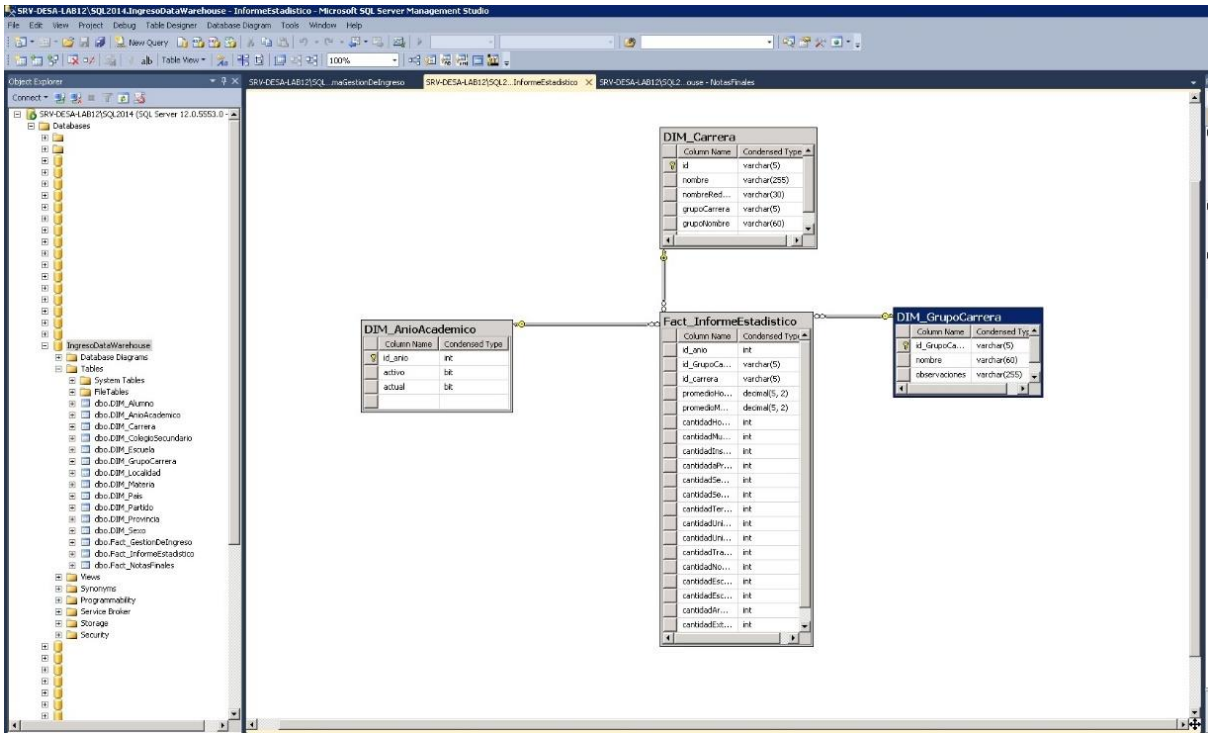


Imagen 35 Diseño físico. Representación del esquema Informe Estadístico en SQL Server[8].

Fuente: Elaboración Propia.

[8] Para scripts de Informe Estadístico ver anexo

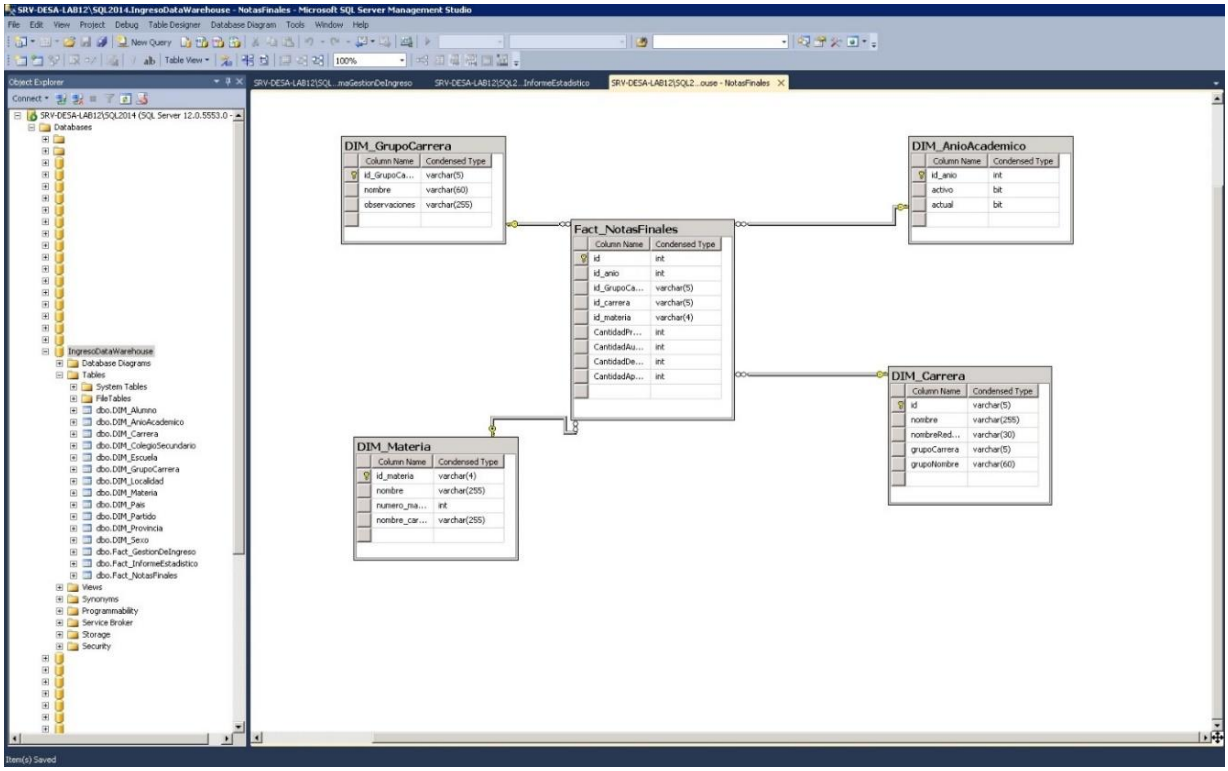


Imagen 36 Diseño físico. Representación del esquema Notas Finales en SQL Server[9].

Fuente: Elaboración Propia.

4.6 Diseño e Implementación del Subsistema ETL

En esta fase de define el diseño e implementación del subsistema de ETL el cual consiste en identificar y recopilar la información inicial de los datos fuentes, de acuerdo con la lógica de negocio y a los requerimientos diseñados anteriormente. Clasificar la información, para posteriormente cargar la información procesada en el modelo diseñado.

Luego de haber creado la Base de Datos y haber definido las dimensiones y hechos del modelo dimensional se procede con la etapa ETL. Para realizar este proceso se empleará Microsoft Visual Studio 2019 con la Herramienta Integration Services de Business Intelligence, el cual facilita la Extracción, Transformación y Carga de los datos de ingresantes que se encuentran en una base de datos de MySQL a la base de datos en Microsoft SQL Server. Vemos en la ***Imagen 37*** la representación gráfica.

[9] Para scripts de Notas Finales ver anexo.

Imagen 37 Diagrama del Proceso ETL.

Fuente: Elaboración Propia.

Vista del entorno Visual Studio con las herramientas utilizadas para realizar el ETL del cubo Gestión de Ingreso.

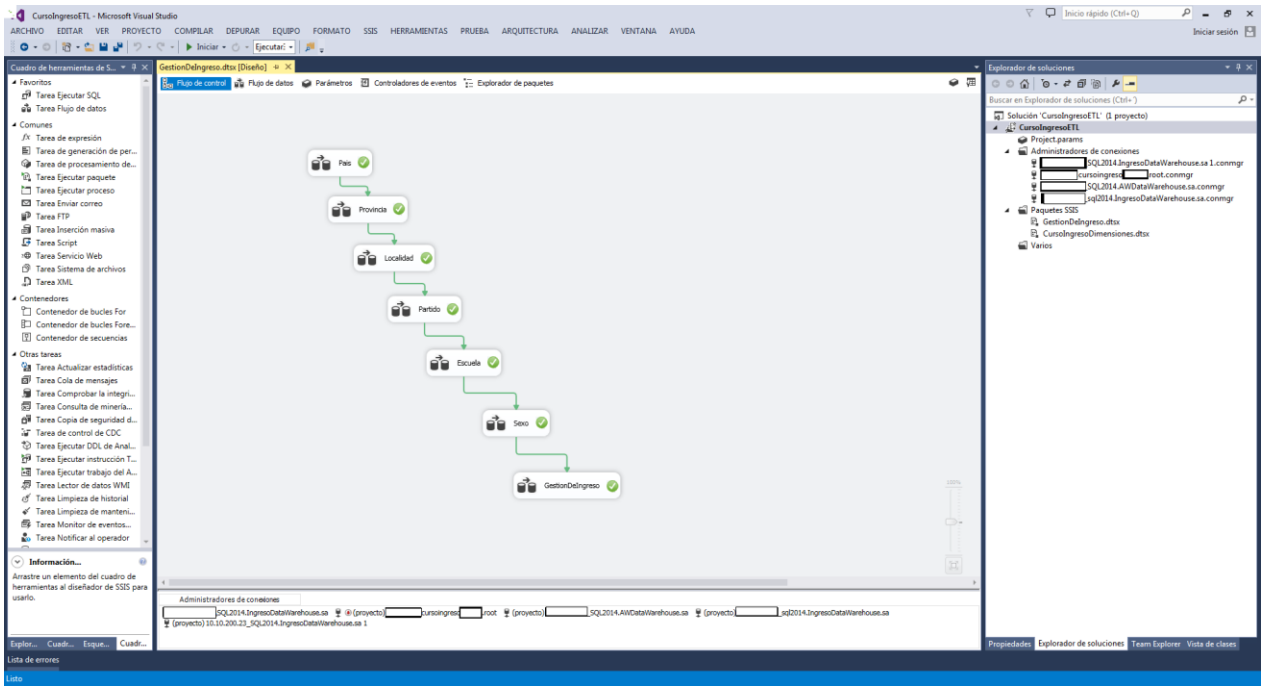


Imagen 38 Flujos de datos para alimentar el cubo Gestión de Ingreso.

Fuente: Elaboración Propia.

Vista del entorno Visual Studio con las herramientas utilizadas para realizar el ETL del cubo Notas Finales.

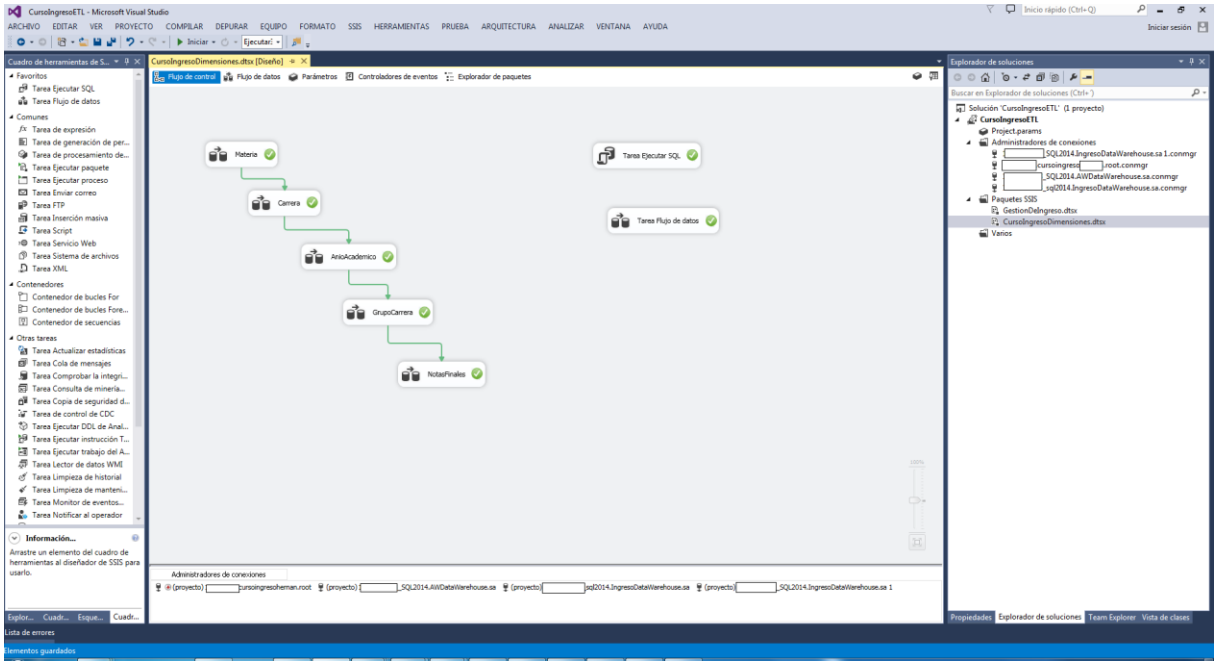


Imagen 39 Flujos de datos para alimentar el cubo Notas Finales.

Fuente: Elaboración Propia.

4.7 Implementación

En esta fase del ciclo de vida, propuesto por Kimball ***Imagen 40***, se realiza la implementación de todo el modelo, lo cual implica la convergencia del diseño lógico, diseño físico y la visualización de la solución hacia los usuarios de negocio. Aunque sin duda se realizaron pruebas durante las tareas de desarrollo del DW, se deben realizar pruebas de sistema de extremo a extremo, calidad de los datos, procesamiento de operaciones, rendimiento y pruebas de usabilidad. Además de evaluar críticamente la preparación de los entregables del DW, también se debe complementar con educación y soporte para la implementación. Debido a que la comunidad de usuarios debe adoptar el sistema DW para que se considere exitoso, la educación es fundamental.

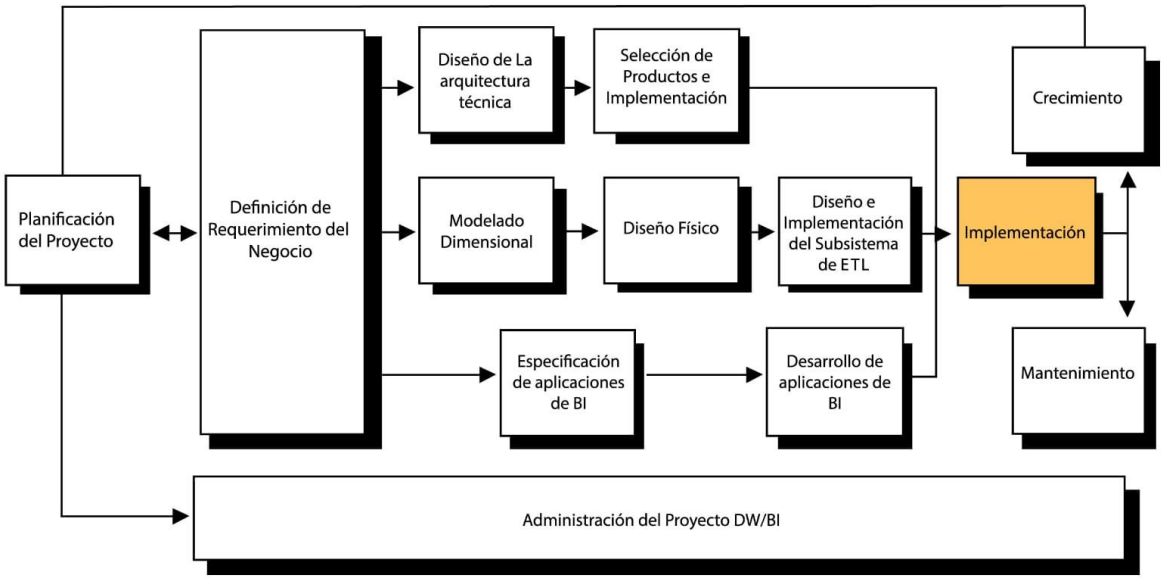


Imagen 40 Metodoligia de Kimball en la fase de Implementación.

Fuente: Traducción de (Kimball & Ross, 2013)

La implementación comprende de varias piezas que debemos realizar, pero podemos agrupar estas tareas en la Infraestructura tecnológica y la implementación de la solución con los usuarios.

Para la infraestructura de la solución necesitamos poner en producción el DW. Debemos disponer del hardware y software necesarios donde quedara finalmente implementado. Además, debemos dejar funcionando las aplicaciones que permiten explorar nuestra solución, en nuestro caso Microsoft Excel y Power BI con sus correspondientes accesos a los usuarios interesados.

La implementación de la solución con los usuarios consiste en asegurarnos que los usuarios hagan un uso productivo de la solución. Entendemos que esta es la parte más complicada ya que la mayoría de los usuarios particularmente de este tipo de soluciones, son personas con muy poco tiempo que no son fáciles de capacitar, así como para redefinir sus nuevas formas de trabajo, como podría ser acceder a la información directamente en lugar de pedirla. Trabajar con altos niveles directivos de la organización hacen al proyecto más vulnerable. Desde la parte informática de los desarrolladores de la solución, podemos caer en la trampa de enamorarnos

demasiado de la tecnología y los datos en lugar de centrarnos en los requisitos y objetivos de la organización.

En el capítulo siguiente veremos este punto del trabajo realizado en detalle aplicado a nuestra solución de DW.

Capítulo 5 Validación

5.1 Validación de la solución

La metodología para la evaluación del proyecto consiste en verificar los requerimientos relevados en la Secretaría Académica y mostrar los resultados obtenidos a partir de las necesidades de información que precisaban. La herramienta que vamos a utilizar para ello es Microsoft Excel que nos permite seleccionar diferentes combinaciones de datos para obtener los resultados que necesitamos con cierto grado de detalle y la posibilidad de analizar los datos desde distintos puntos de vista. La validación de los datos con el usuario nos garantiza que al conocer los datos de su negocio puede contrastar los resultados obtenidos con la realidad. Además, obtenemos los datos del Data Warehouse y los informes del sistema operacional "Gestión de ingreso" para demostrar que los datos son válidos.

Debemos tener en cuenta que necesitamos brindar una capacitación para que puedan realizar una buena evaluación del proyecto. Es decir, comprendan los conceptos que van a utilizar y puedan aprovechar el máximo potencial que le puede brindar el Data Warehouse. Es de suma importancia la capacitación para lograr la aceptación del usuario y cumplir con sus expectativas.

Es de suma importancia en este punto realizar:

a. Pruebas de aceptación de usuario: para asegurar que los datos que se proporcionan al usuario final cumplen con sus expectativas y además verificar que las herramientas que se ponen a su disposición son las adecuadas.

b. Pruebas de validación de datos: esta prueba es mediante el uso de una herramienta de consulta ad hoc (Excel) que permita recuperar datos en un formato similar a los informes operativos existentes. Cuando se detecta la existencia de un vínculo entre el Data Warehouse y el informe operacional, se demuestra que los datos son válidos. Esta prueba ha de ser llevada a cabo por un representante del negocio, ya que este perfil es quien mejor conoce los datos y puede validarlos con mayores garantías de éxito.

Veamos como el usuario final accede a los cubos realizados. Primero debemos configurar el Excel para conectarnos a la base de datos que se encuentra en SQL Server. Utilizamos las credenciales de nuestro usuario de dominio. A su vez los usuarios que tienen acceso son asignados en SQL Server para poder acceder. Luego nos aparece el Data Warehouse generado y lo seleccionamos ***Imagen 41***.

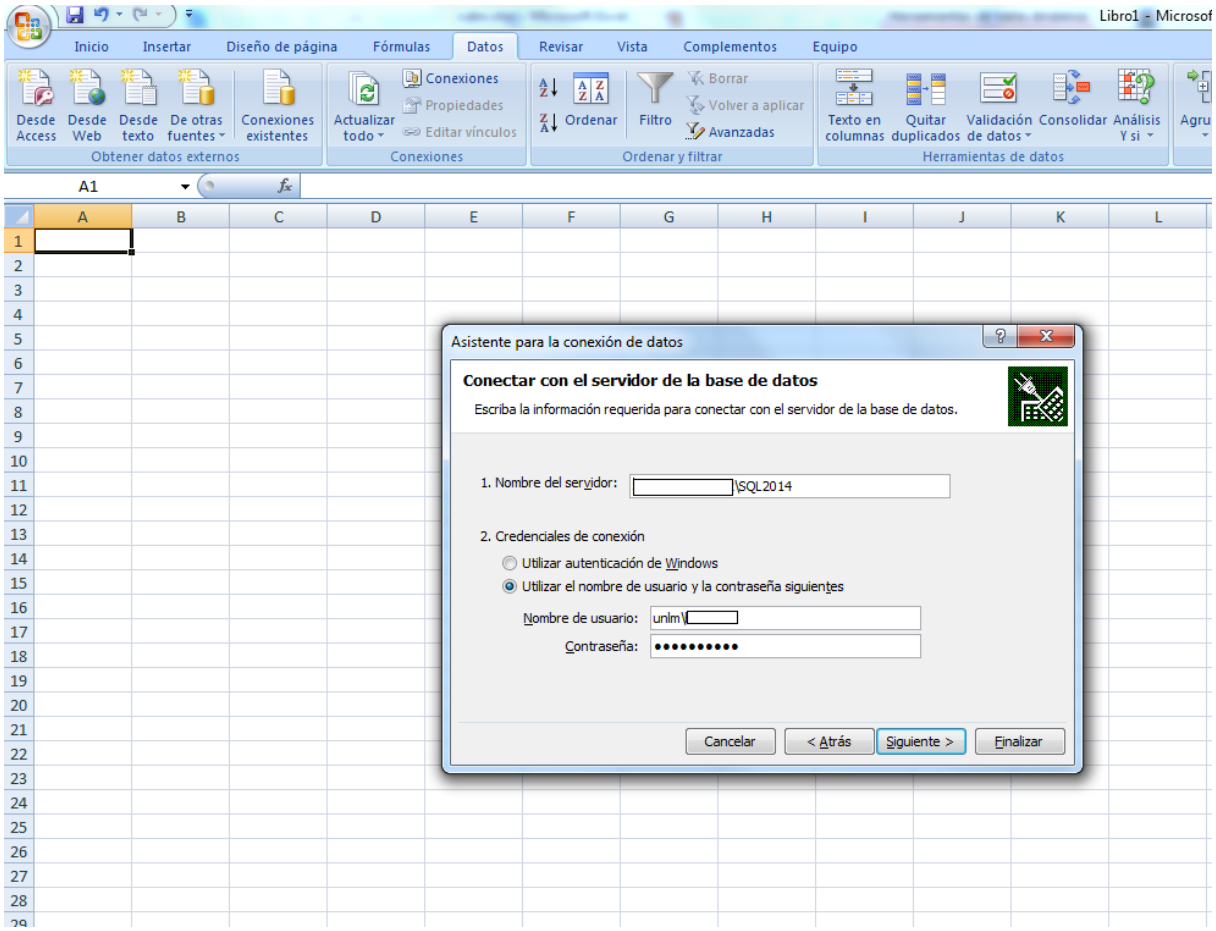


Imagen 41 Conexión a la base de datos SQL Server donde tenemos el cubo.

Fuente: Elaboración Propia.

Para conectarnos al cubo, desde el cliente, debe tener usuario de dominio y los permisos necesarios que otorgamos como administradores para conectarse a la base de datos.

Una vez conectados nos muestra las distintas bases, en nuestro caso como tipo cubo ***Imagen 42***.

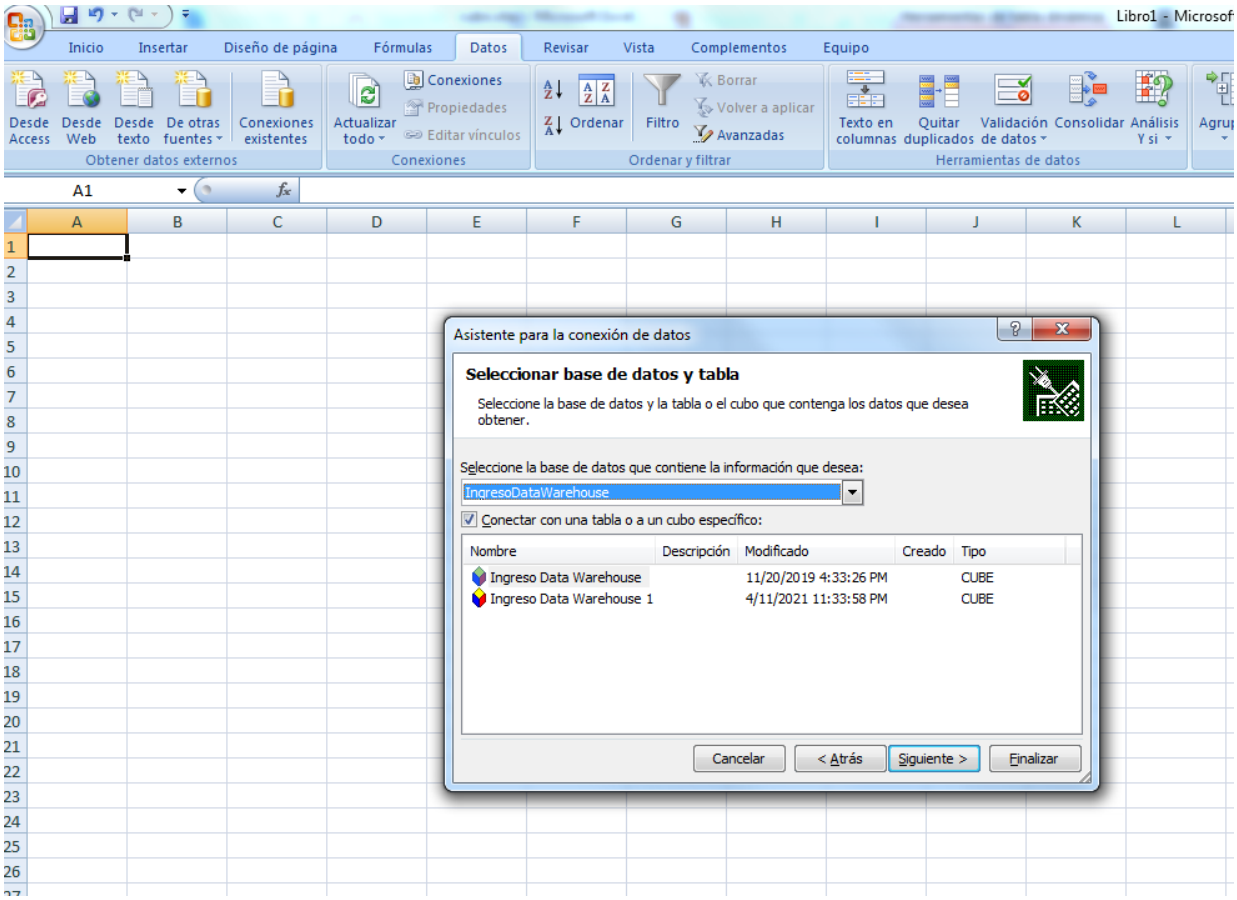
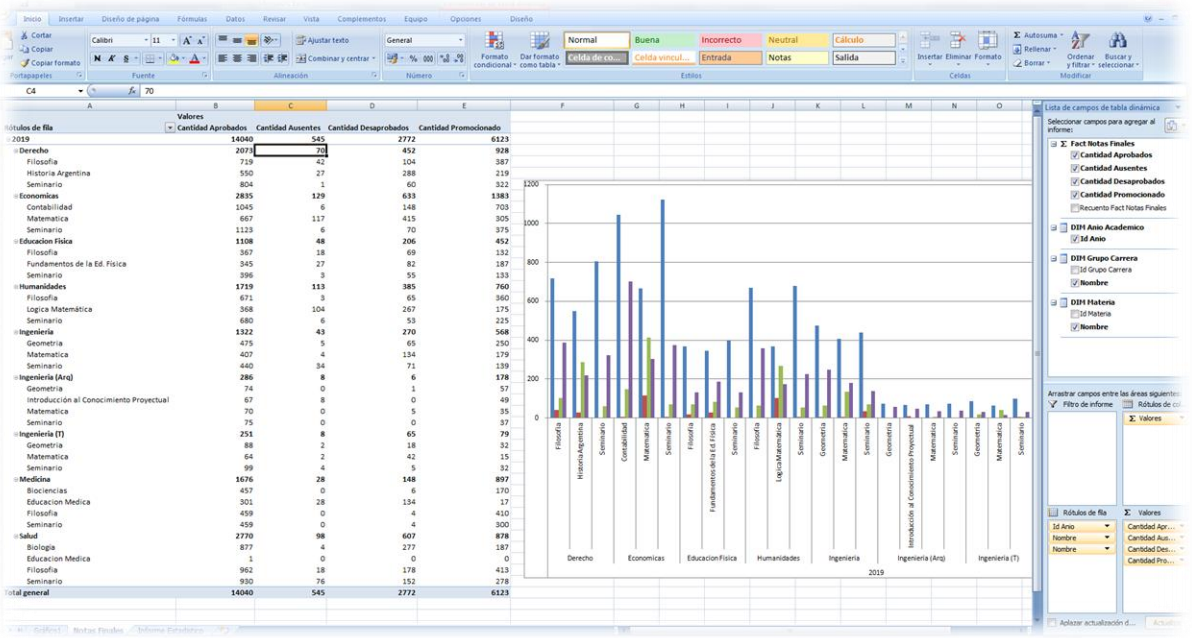


Imagen 42 Vista del Data Warehouse generado a seleccionar.

Al realizar la conexión y seleccionar el cubo nos muestra en el margen derecho del archivo Excel las dimensiones y hechos que definimos y podemos seleccionar. Excel además nos permite agregar gráficos seleccionando la tabla con los resultados para tener una visión como muestra la ***Imagen 43***.



Rótulos de fila	Valores: Cantidad Aprobados	Cantidad Ausentes	Cantidad Desaprobados	Cantidad Promocionado
2019	14040	545	2772	6123
Derecho	2071	70	452	928
Filosofia	719	42	104	387
Historia Argentina	550	27	288	219
Seminario	804	1	60	322
Economicas	2835	129	633	1383
Contabilidad	1045	6	148	703
Matematica	667	117	415	305
Seminario	1123	6	70	375
Educacion Fisica	1108	48	206	452
Filosofia	367	18	69	132
Fundamentos de la Ed. Fisica	345	27	82	187
Seminario	396	3	55	133
Humanidades	1719	113	385	760
Filosofia	671	3	65	360
Logica Matemática	368	104	267	175
Seminario	680	6	53	225
Ingenieria	1322	43	270	568
Geometria	475	5	65	250
Matematica	407	4	134	179
Seminario	440	34	71	139
Ingenieria (Arq)	286	8	6	178
Geometria	74	0	1	57
Introducción al Conocimiento Proyectual	67	8	0	49
Matematica	70	0	5	35
Seminario	75	0	0	37
Ingenieria (T)	251	8	65	79
Geometria	88	2	18	32
Matematica	64	2	42	15
Seminario	99	4	5	32
Medicina	1676	28	148	897
Biociencias	457	0	6	170
Educacion Medica	301	28	134	17
Filosofia	459	0	4	410
Seminario	459	0	4	300
Salud	2770	98	607	878
Biologia	877	4	277	187
Educacion Medica	1	0	0	0
Filosofia	962	18	178	413
Seminario	930	76	152	278
Total general	14040	545	2772	6123

Imagen 43 Vista del cubo Notas Finales desde Excel.

Fuente: Elaboración Propia.

Vista del cubo con cantidad de inscriptos aprobados, desaprobados, ausentes y promocionados, del año 2019 y agrupado por departamento académico con sus respectivas materias para el ingreso ***Imagen 44***.

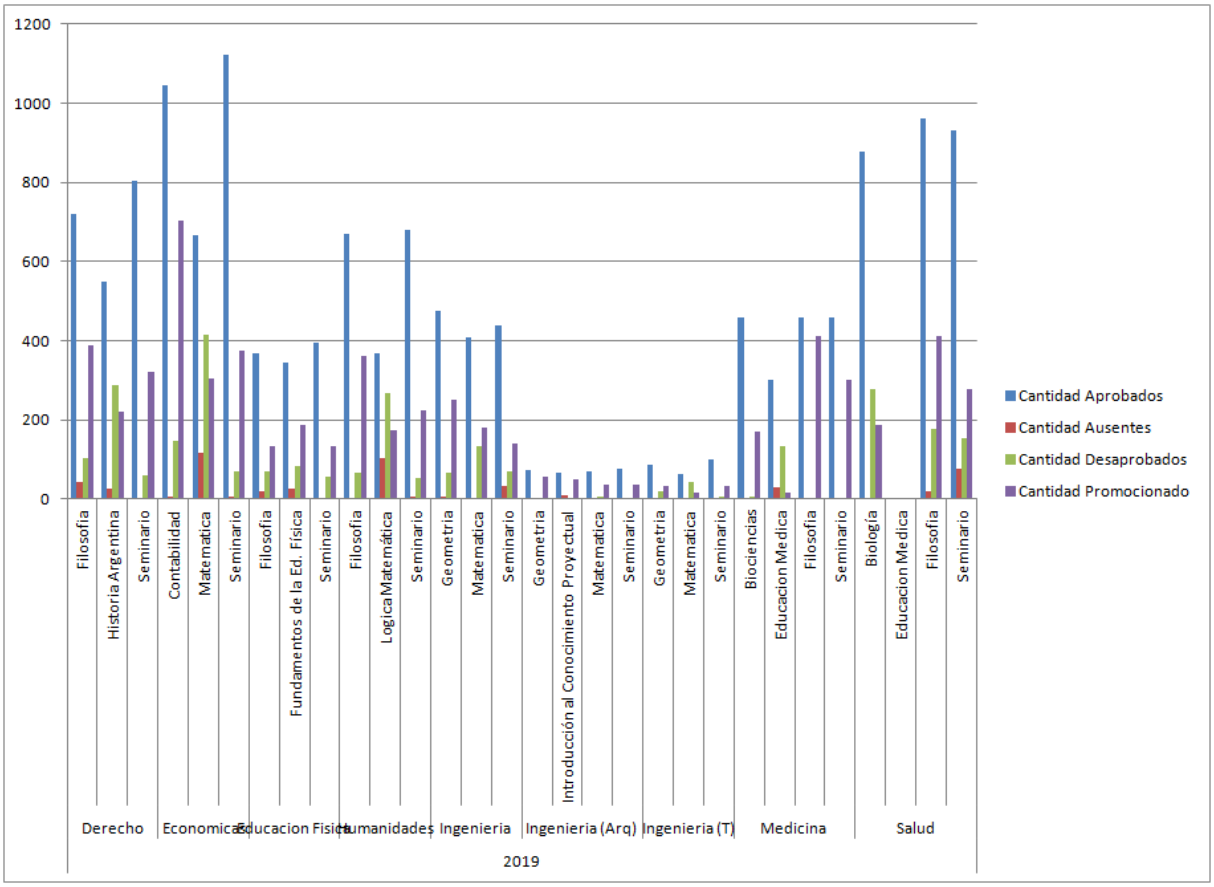
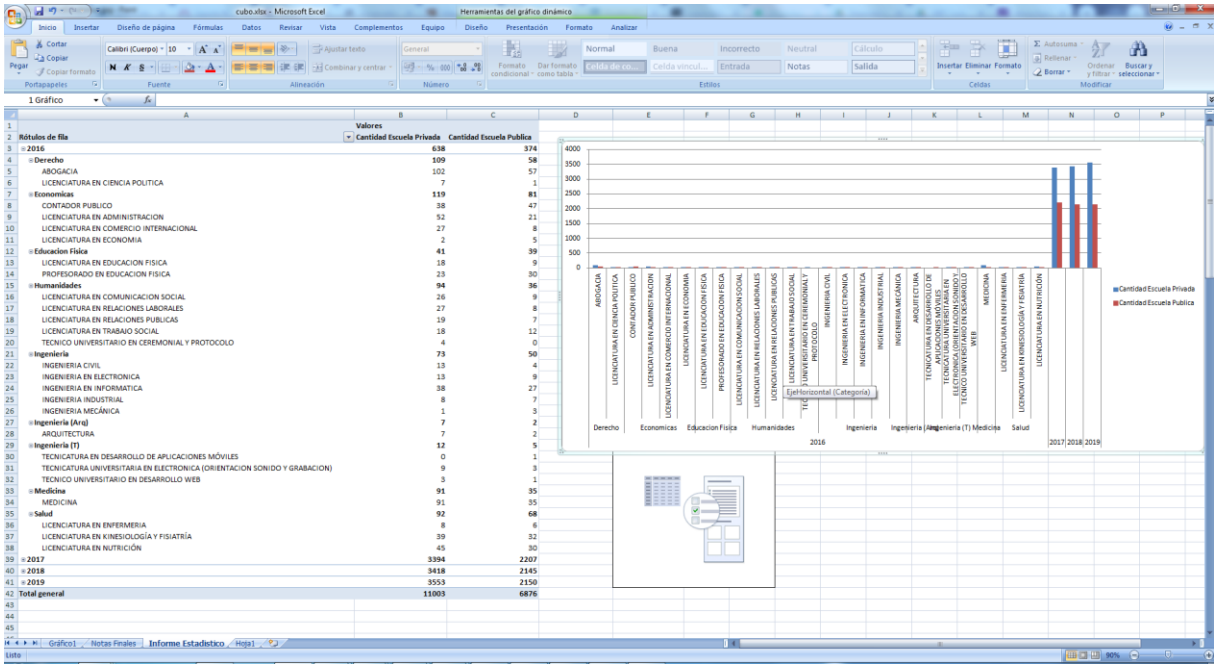


Imagen 44 Gráfico del cubo Notas Finales.

Fuente: Elaboración Propia.

Vista del cubo agrupado por año, carrera y materia con cantidad de inscriptos que provienen de escuela pública y privada. Se puede observar que ponemos visualizar el total por departamento académico y como está conformado ese total con las materias correspondientes ***Imagen 45***.



Rótulos de fila	Cantidad Escuela Privada	Cantidad Escuela Publica
2016	638	374
Derecho	109	58
ABOGACIA	102	57
LICENCIATURA EN CIENCIA POLITICA	7	1
Economicas	119	81
CONTADOR PUBLICO	38	47
LICENCIATURA EN ADMINISTRACION	52	21
LICENCIATURA EN COMERCIO INTERNACIONAL	27	8
LICENCIATURA EN ECONOMIA	2	5
Educacion Fisica	41	39
LICENCIATURA EN EDUCACION FISICA	18	9
PROFESORADO EN EDUCACION FISICA	23	30
Humanidades	94	36
LICENCIATURA EN COMUNICACION SOCIAL	26	9
LICENCIATURA EN RELACIONES LABORALES	27	8
LICENCIATURA EN RELACIONES PUBLICAS	19	7
LICENCIATURA EN TRABAJO SOCIAL	18	12
TECNICO UNIVERSITARIO EN CEREMONIAL Y PROTOCOLO	4	0
Ingeniería	73	50
INGENIERIA CIVIL	13	4
INGENIERIA EN ELECTRONICA	13	9
INGENIERIA EN INFORMATICA	38	27
INGENIERIA INDUSTRIAL	8	7
INGENIERIA MECÁNICA	1	3
Ingeniería (Arq)	7	2
ARQUITECTURA	7	2
Ingeniería (T)	12	5
TECNICATURA EN DESARROLLO DE APLICACIONES MÓVILES	0	1
TECNICATURA UNIVERSITARIA EN ELECTRONICA (ORIENTACION SONIDO Y GRABACION)	9	3
TECNICO UNIVERSITARIO EN DESARROLLO WEB	3	1
Medicina	91	35
MEDICINA	91	35
Salud	92	68
LICENCIATURA EN ENFERMERIA	8	6
LICENCIATURA EN KINESIOLOGÍA Y FISIATRÍA	39	32
LICENCIATURA EN NUTRICIÓN	45	30
2017	3394	2207
2018	3418	2145
2019	3553	2150
Total general	11003	6876

Imagen 45 Vista del cubo Informe Estadístico.

Fuente: Elaboración Propia.

Vista grafica de inscriptos por año, departamento, materia de hombres y mujeres. En este caso vemos un lapso desde 2016 a 2019 con las cantidades por departamento académico. De acuerdo con la necesidad de análisis podríamos visualizar las cantidades por materias de cada departamento académico ***Imagen 46***.

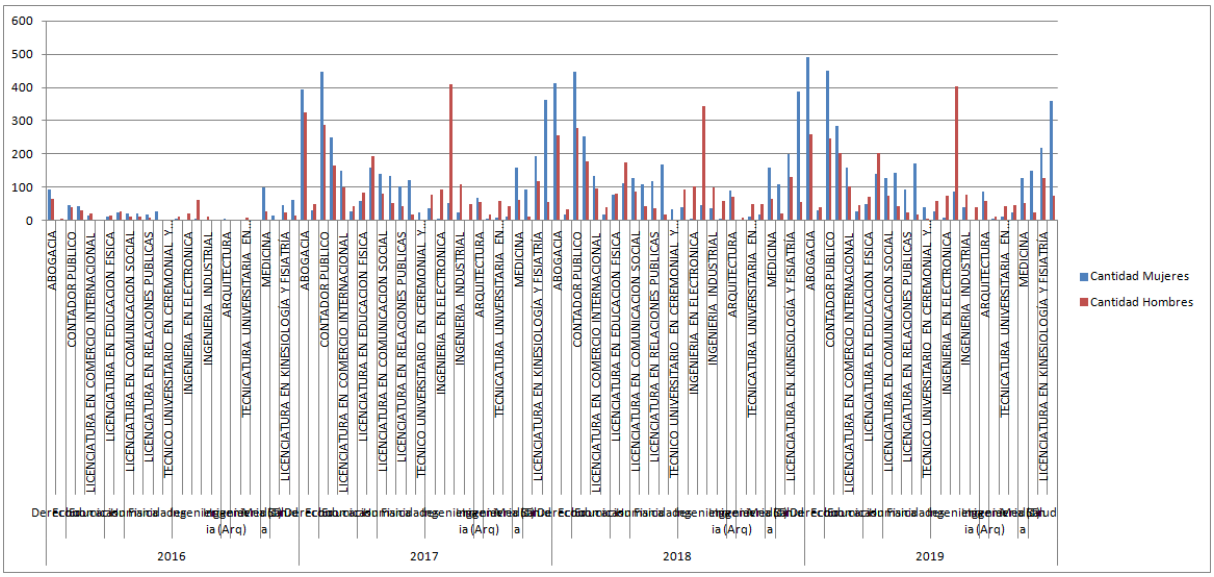


Imagen 46 Gráfico cubo Informe Estadístico.

Fuente: Elaboración Propia.

En el transcurso del trabajo, si bien se planifico la visualización de los cubos a través de Microsoft Excel, se pudo obtener acceso a la herramienta de visualización Microsoft Power BI que además de brindarnos un entorno enriquecido nos da la posibilidad de mostrar la información en una página web con los indicadores que se quieran exponer de forma pública.

Se muestran las imágenes de conexiones que se realizan desde Power BI Desktop con SQL Server y las publicaciones que se realizan en la web en un espacio común de Power BI donde podemos administrar los accesos a los distintos usuarios. Por último, se utiliza una réplica de la página de la universidad para mostrar cómo se visualizaría un cubo de forma pública en nuestra web institucional.

De igual forma que hicimos la conexión con Excel necesitamos conectarnos a SQL Server con la misma instancia y base de datos para comenzar a trabajar. Como vemos en la ***Imagen 47*** Power BI nos brinda distintas posibilidades de conexión y seleccionamos SQL Server para obtener nuestros datos.

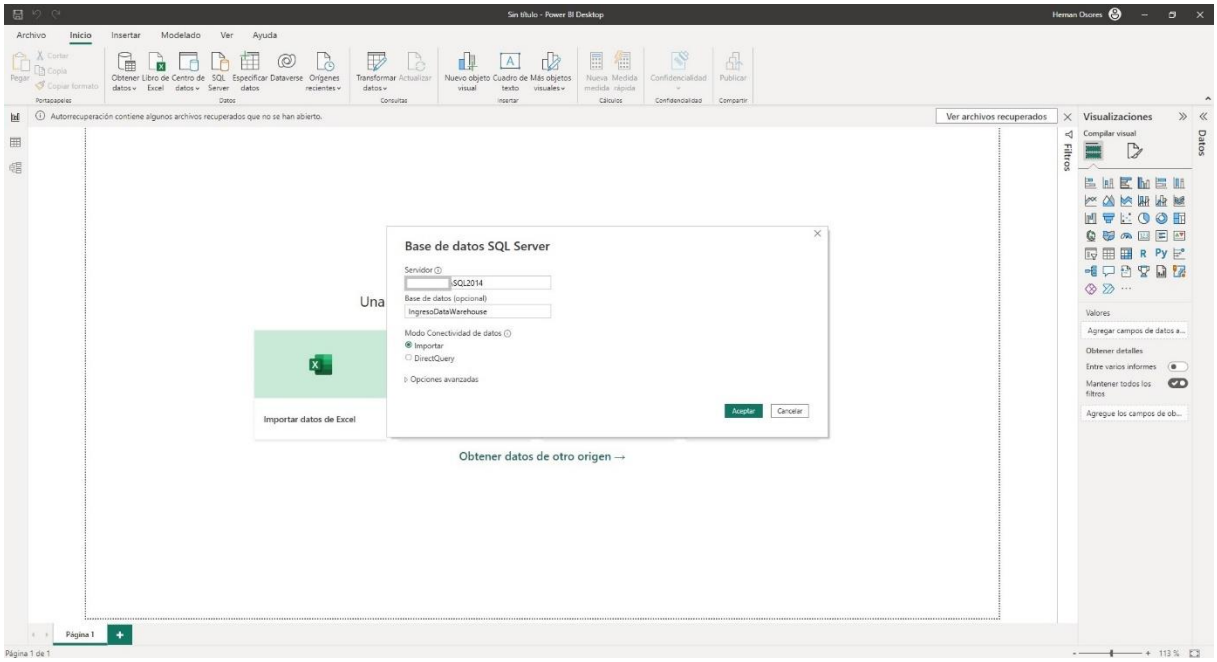


Imagen 47 Conexión con SQL Server para obtener los datos.

Fuente: Elaboración Propia.

Una vez conectado nos muestra las dimensiones y hechos que tenemos disponibles y debemos seleccionar los que vamos a utilizar en este análisis particular ***Imagen 48***.

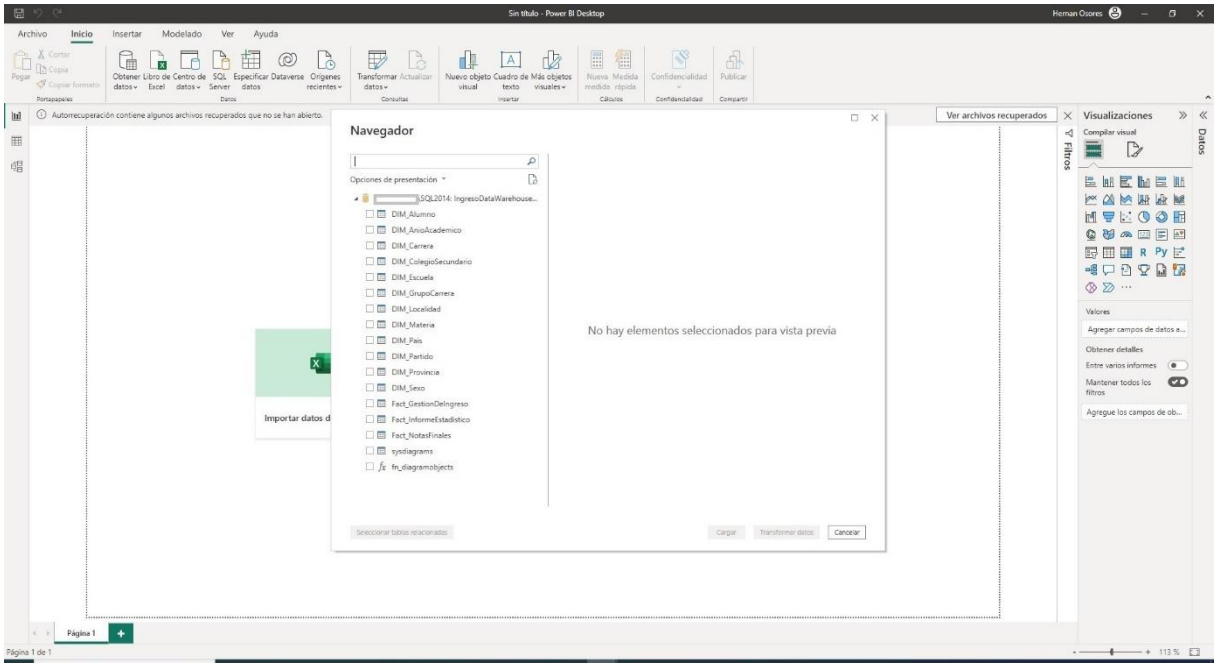


Imagen 48 Selección de las dimensiones y hechos que se van a utilizar.

Fuente: Elaboración Propia.

En el margen derecho "Datos" podemos observar que hemos seleccionado las dimensiones Año Académico, Carrera y Grupo de Carrera que agrupa las carreras por departamento académico. En lo que respecta a hechos vemos que seleccionamos Notas finales. En la parte de "Visualizaciones" podemos elegir el grafico que deseamos para visualizar la información de análisis. Por último, la ***Imagen 49*** representa la cantidad de aprobados, cantidad de desaprobados, cantidad de promocionados y cantidad de ausentes del año 2019 para los departamentos de Derecho e Ingeniería.

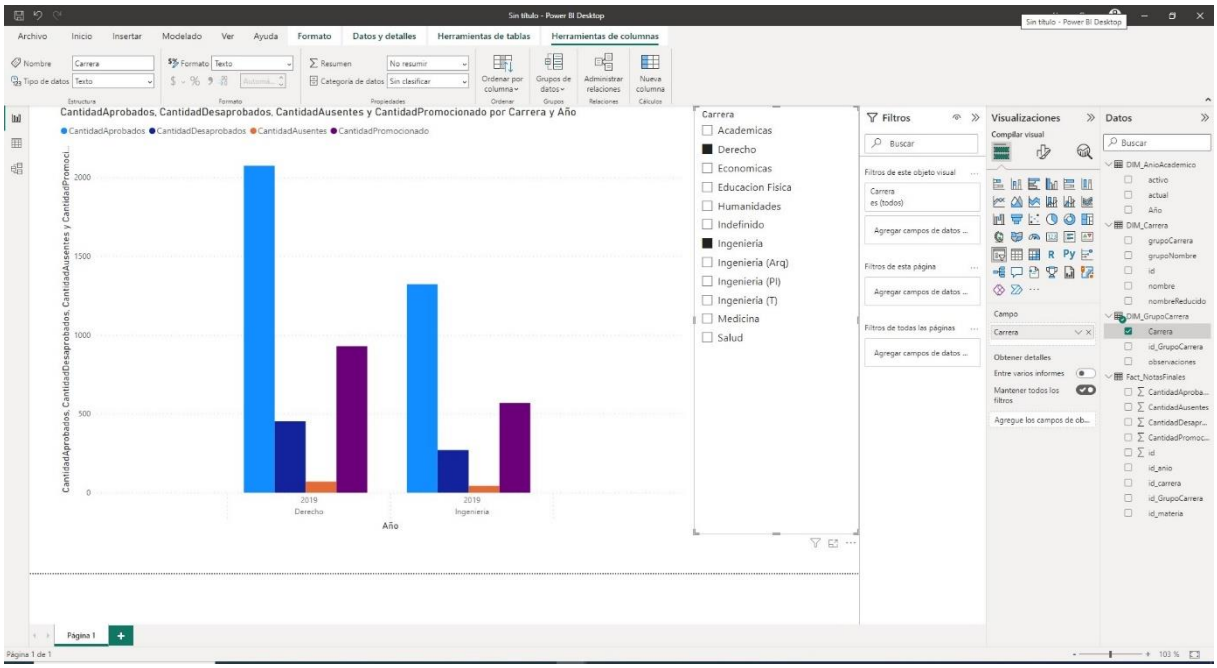


Imagen 49 Vista del cubo Notas Finales en Power BI.

Fuente: Elaboración Propia.

En la ***Imagen 50*** seleccionamos para el analisis los años 2017 y 2018, los departamentos de Ingenieria y Humanidades para mostrar, cantidades de aspirantes de la escuela publica y privada . Ademas, cantidades de hombres y mujeres.

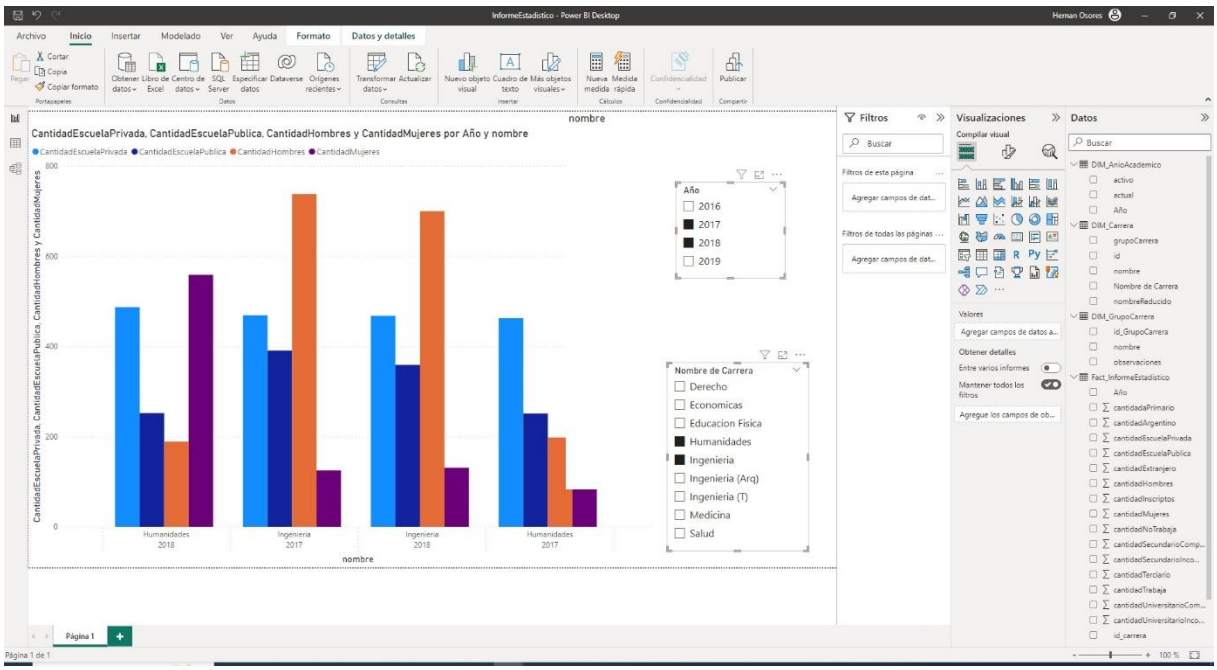


Imagen 50 Vista del cubo Informe Estadístico en Power BI.

Fuente: Elaboración Propia.

Power BI permite hacer publicaciones dentro de un espacio de trabajo donde podemos compartir, como muestra la ***Imagen 51***, el informe estadístico. En este caso el usuario podrá hacer el análisis por año y carrera con las cantidades que muestra la imagen.

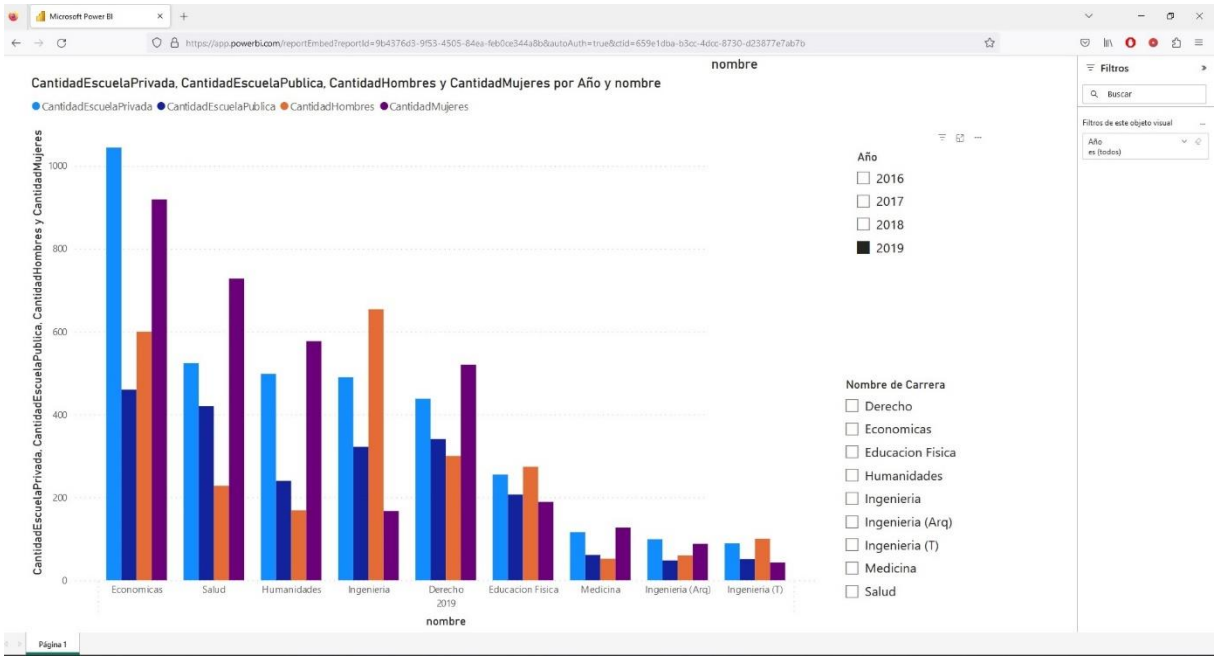


Imagen 51 Vista de una publicación en el espacio de Power BI.

Fuente: Elaboración Propia.

Hasta el momento hemos administrado la visualización por usuarios y espacios de trabajo. Pero podemos tener la necesidad de mostrar determinada información en un espacio público como podría ser la web institucional. Power BI nos brinda esta posibilidad como muestra la ***Imagen 52***.

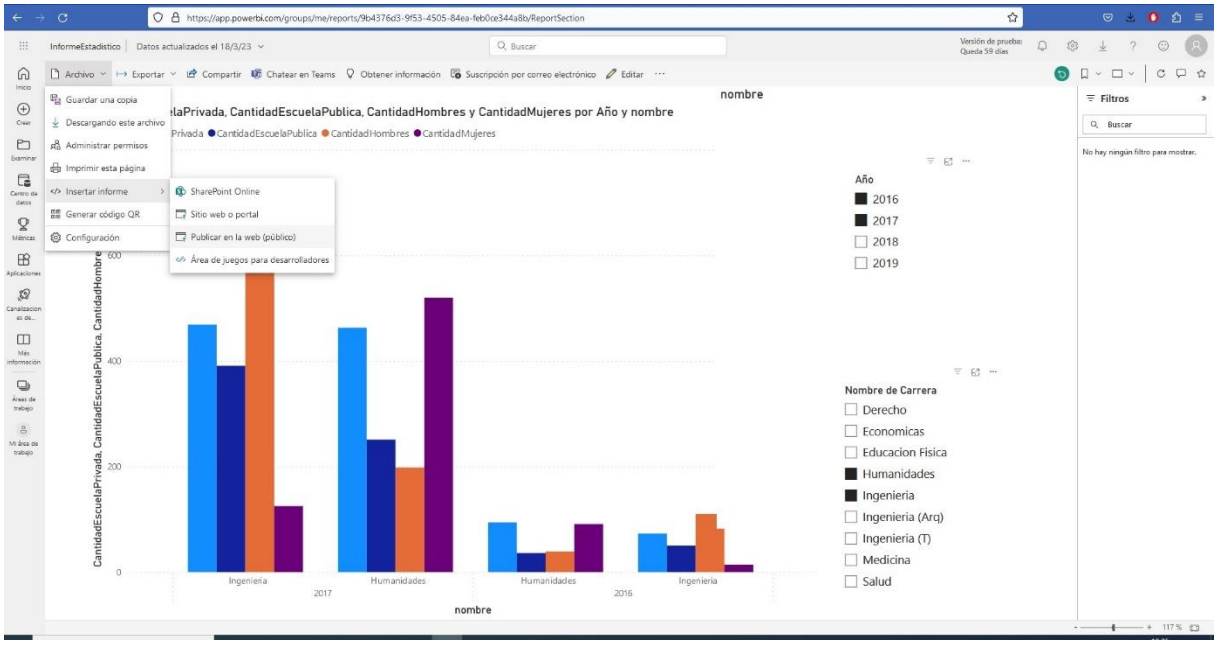


Imagen 52 Como publicar el cubo Informe Estadístico en una página Web. Paso 1.

Fuente: Elaboración Propia.

En la ***Imagen 53*** vemos que cuando seleccionamos en la imagen anterior "Publicar en la web (publico)" nos genera un código HTML que nos va a servir para insertar el informe en nuestra página web.

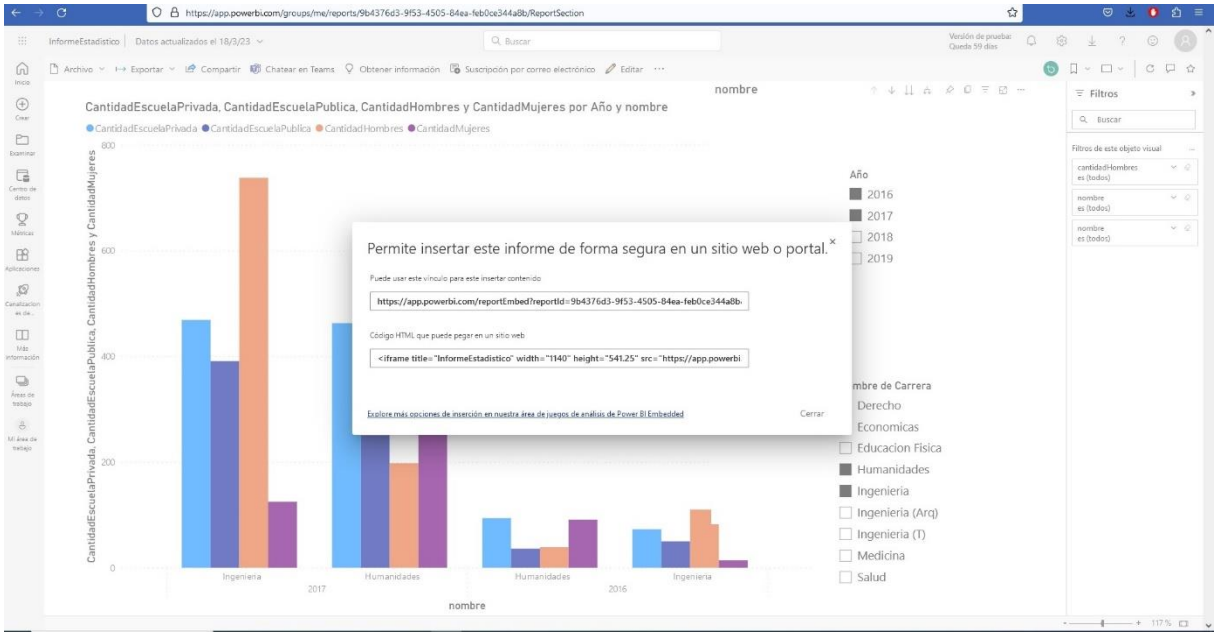


Imagen 53 Como publicar el cubo Informe Estadístico en una página Web. Paso2.

Fuente: Elaboración Propia.

En la ***Imagen 54*** vemos, desde el editor de código fuente Visual Studio Code[10], una parte de la página web donde hemos insertado el código HTML generado por Power BI.

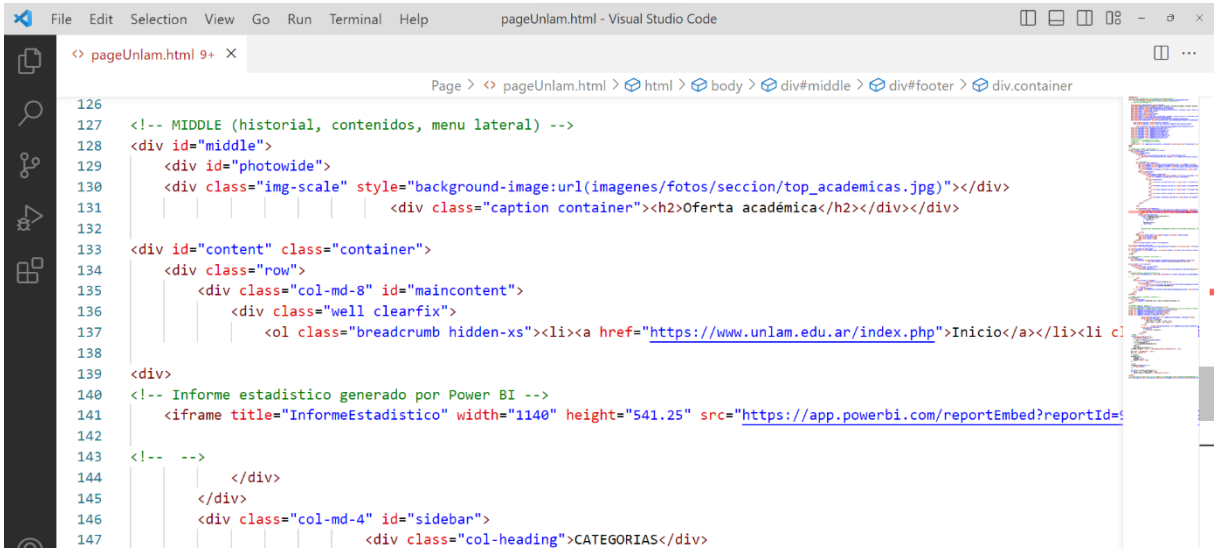


Imagen 54 Código fuente HTML de la página institucional.

Fuente: Elaboración Propia.

La ***Imagen 55*** nos muestra la pagina institucional resultante luego de insertar el codigo HTML con el informe estadistico. En este caso las personas que visualicen el informe estadistico

[10] Visual Studio Code es un editor de código fuente desarrollado por Microsoft para Windows, Linux y macOS.

podran ver las cantidades en el grafico de acuerdo a los años y departamentos academicos seleccionados.

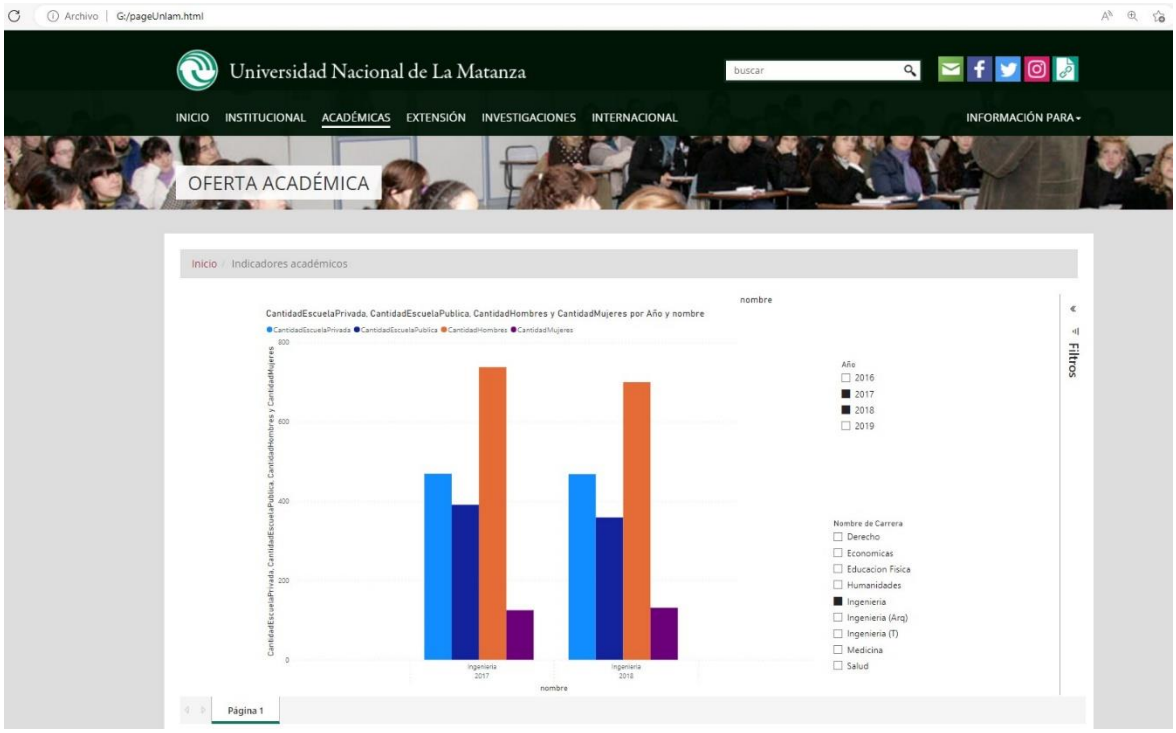


Imagen 55 Ejemplo de la página institucional con una publicación del cubo Informe Estadístico.

Fuente: Elaboración Propia.

Capítulo 6 Conclusión y futuros trabajos

6.1 Conclusión

A partir de los objetivos planteados en este trabajo podemos obtener las siguientes conclusiones:

- Se manifestaron las referencias teóricas más importantes para mejorar la toma de decisiones, con los conceptos para administrar la historia de los datos en un sistema analítico.

- Se analizaron las características de los Data Warehouse y sus diferencias con los sistemas transaccionales. Además, distintas metodologías de acuerdo con sus autores.

- Con la obtención de los requerimientos identificamos el fenómeno que se deseaba medir. A partir de la recolección de los datos provenientes de los sistemas operacionales, sistema de Ingresantes, cumplimos con el objetivo de transformar estos datos en información para dar paso al componente analítico que debe tener nuestro Data Warehouse. Este no cumple sólo una misión informativa también desempeñan una tarea evaluativa. Donde se espera que su lectura señale si determinado curso de los acontecimientos en el ingreso constituye una mejora o un deterioro. Podemos medir aspectos esenciales como notas finales de acuerdo con su departamento académico, carrera, materia, año lectivo. Además de informes estadísticos de aspirantes que provienen de escuelas públicas, sexo, trabajo etc. por departamento académico, carrera y año. El resultado de este trabajo permite utilizar en tiempo y forma la información para que los actores de los niveles estratégicos puedan tomar decisiones en nuestra institución.

- De acuerdo con los conceptos estudiados se optó por la metodología que podía adaptarse a nuestra solución, desarrollando un Data Warehouse que da respuesta a las necesidades de la Secretaría Académica. Con esta implementación se logró, el acceso a los datos de forma fácil y rápida a través de un Excel y Power BI, dando independencia al personal técnico para la obtención de nuevos reportes y facilitar el proceso de comparación, proyección a futuro y muestra de indicadores. Este trabajo permite lograr una visión más

completa e integral de la universidad, entender los eventos en forma sistemática, para así redefinir estrategias.

Con dichos objetivos, estoy convencido que el presente trabajo sirve como referencia y complemento de los trabajos realizados en esta casa, para brindar un marco teórico y de aplicación que pueda significar de ayuda para los docentes y alumnos que quieran abordar el tema tratado.

Los indicadores no son sólo medidas estadísticas que informan, sino que también, permiten construir nuevas visiones y expectativas. Confío en que este es el puntapié inicial para seguir desarrollando soluciones que nos permitan trabajar en conjunto y avanzar en una mejor atención hacia los ingresantes y alumnos y, porque no, mejorar las gestiones locales, provinciales y nacionales con los resultados obtenidos.

6.2 Futuros trabajos

Considero que hemos progresado al incentivar la utilización de este tipo de tecnologías para la toma de decisiones, tan importantes y útiles que nos permitan seguir avanzando hacia un modelo de excelencia universitaria. El desafío es continuar estimulando el uso de almacenes de datos para la toma de decisiones y, como consecuencia de esta necesidad, crear un área específica para su desarrollo que permita brindar soluciones de esta naturaleza.

Se podrían desarrollar nuevos cubos para mejorar los análisis más detallados con respecto a los ingresantes, por ejemplo, incluir información del contexto socio económico. Estas nuevas demandas redundaran en una mejora del sistema operacional de Ingresantes porque debe tener la capacidad de registrar estos nuevos datos.

Capítulo 7 Referencias bibliográficas

Ballard, C., Farrell, D. M., Gupta, A., Mazuela, C., & Vohnik, S. (March 2006). *Dimensional Modeling: In a Business Intelligence Environment* (Vol. First Edition). ibm.com/redbooks.

Bedell, J. (1997). *Data Warehousing, Data Modeling and Design.* MicroStrategy.

Bernabeu, D. (13 de Mayo de 2009). *dataprix*. Obtenido de https://www.dataprix.com/es/data-warehousing-y-metodologia-hefesto/24-cualidades

Bigatti, C., & Grasso, M. (2008). *Data Warehouse.* Obtenido de http://www.edutecne.utn.edu.ar/sist-gestion-II/Apunte%20BI.pdf

Castro, H. M. (22 de agosto de 2014). *Data Warehouse y Sistemas de Soporte a la Decisión Un Enfoque Práctico.* Obtenido de https://www.slideserve.com/afia/data-warehouse-y-sistemas-de-soporte-a-la-decisi-n

Chinkes, E. (2008). *Business Intelligence para mejores decisiones de negocio.* Ciudad Autonoma de Buenos Aires: EDICON.

Elmasri, R., & Navathe, S. (2007). *Fundamentos de Sistemas de Bases de Datos.* Madrid: PEARSON EDUCACION S.A.

Inmon, W. H. (2005). *Building the Data Warehouse* (Fourth ed.). Wiley Publishing, Inc.

Kimball, R. (1996). *The Data Warehouse Toolkit: practical techniques for building dimensional data warehouses.* John Wiley & Sons, Inc. New York, NY, USA ©1996.

Kimball, R. (1999). When A Slowly Changing Dimension Speeds Up. *2*(11).

Kimball, R. (2008). Slowly Changing Dimensions. *18*(9).

Kimball, R., & Caserta, J. (2008). *The Datawarehouse ETL Toolkit.* Wiley.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit.* John Wiley and Sons, Inc.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling* (Third ed.). John Wiley & Sons.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). *The Data Warehouse Lifecycle Toolkit* (Second Edition ed.). Indianapolis: Wiley Publishing, Inc.

Microsoft. (2017). *Implementing a SQL Data Warehouse - 20767B.* Microsoft Corporation.

MicroStrategy LATAM South. (2006). *Teoría sobre Business Intelligence.* (3ra, Ed.)

Norris, M., & Rigby, P. (1994). *Ingeniería de software explicada.* Megabyte – Grupo Noriega Editores.

Oracle, C. (s.f.). *Oracle Cloud Infrastructure (OCI).* Obtenido de https://www.oracle.com/ar/database/what-is-oltp/

Perez Marquez, M. (2011). *MICROSOFT SQL SERVER 2008 R2 MOTOR DE BASE DE DATOS Y ADMINISTRACION* . MADRID: RC LIBROS.

Powell, G. (2006). *Beginning Database Design.* Indianapolis, Indiana: Wiley Publishing, Inc.

Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).

Ross, M. (2000). Combining SCD Techniques.

Ross, M. (Marzo de 2004). The Kimball bus architecture and the Corporate Information Factory: What are the fundamental differences? *Kimball Group*.

Ross, M. (2013). Slowly Changing Dimension Types 0, 4, 5, 6 and 7.

Anexo 1 Protección de los Datos Personales

Ley 25.326

Disposiciones Generales. Principios generales relativos a la protección de datos. Derechos de los titulares de datos. Usuarios y responsables de archivos, registros y bancos de datos. Control. Sanciones. Acción de protección de los datos personales.

Sancionada: Octubre 4 de 2000.

Promulgada Parcialmente: Octubre 30 de 2000.

El Senado y Cámara de Diputados de la Nación Argentina reunidos en Congreso, etc. sancionan con fuerza de Ley:

Ley de Protección de los Datos Personales

Capítulo I

Disposiciones Generales

ARTICULO 1° — (Objeto).
La presente ley tiene por objeto la protección integral de los datos personales asentados en archivos, registros, bancos de datos, u otros medios técnicos de tratamiento de datos, sean estos públicos, o privados destinados a dar informes, para garantizar el derecho al honor y a la intimidad de las personas, así como también el acceso a la información que sobre las mismas se registre, de conformidad a lo establecido en el artículo 43, párrafo tercero de la Constitución Nacional.

Las disposiciones de la presente ley también serán aplicables, en cuanto resulte pertinente, a los datos relativos a personas de existencia ideal.

En ningún caso se podrán afectar la base de datos ni las fuentes de información periodísticas.

ARTICULO 2° — (Definiciones).

A los fines de la presente ley se entiende por:

— Datos personales: Información de cualquier tipo referida a personas físicas o de existencia ideal determinadas o determinables.

— Datos sensibles: Datos personales que revelan origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical e información referente a la salud o a la vida sexual.

— Archivo, registro, base o banco de datos: Indistintamente, designan al conjunto organizado de datos personales que sean objeto de tratamiento o procesamiento, electrónico o no, cualquiera que fuere la modalidad de su formación, almacenamiento, organización o acceso.

— Tratamiento de datos: Operaciones y procedimientos sistemáticos, electrónicos o no, que permitan la recolección, conservación, ordenación, almacenamiento, modificación, relacionamiento, evaluación, bloqueo, destrucción, y en general el procesamiento de datos personales, así como también su cesión a terceros a través de comunicaciones, consultas, interconexiones o transferencias.

— Responsable de archivo, registro, base o banco de datos: Persona física o de existencia ideal pública o privada, que es titular de un archivo, registro, base o banco de datos.

— Datos informatizados: Los datos personales sometidos al tratamiento o procesamiento electrónico o automatizado.

— Titular de los datos: Toda persona física o persona de existencia ideal con domicilio legal o delegaciones o sucursales en el país, cuyos datos sean objeto del tratamiento al que se refiere la presente ley.

— Usuario de datos: Toda persona, pública o privada que realice a su arbitrio el tratamiento de datos, ya sea en archivos, registros o bancos de datos propios o a través de conexión con los mismos.

— Disociación de datos: Todo tratamiento de datos personales de manera que la información obtenida no pueda asociarse a persona determinada o determinable.

Capítulo II

Principios generales relativos a la protección de datos

ARTICULO 3° — (Archivos de datos – Licitud).

La formación de archivos de datos será lícita cuando se encuentren debidamente inscriptos, observando en su operación los principios que establece la presente ley y las reglamentaciones que se dicten en su consecuencia.

Los archivos de datos no pueden tener finalidades contrarias a las leyes o a la moral pública.

ARTICULO 4° — (Calidad de los datos).

1. Los datos personales que se recojan a los efectos de su tratamiento deben ser ciertos, adecuados, pertinentes y no excesivos en relación al ámbito y finalidad para los que se hubieren obtenido.
2. La recolección de datos no puede hacerse por medios desleales, fraudulentos o en forma contraria a las disposiciones de la presente ley.
3. Los datos objeto de tratamiento no pueden ser utilizados para finalidades distintas o incompatibles con aquellas que motivaron su obtención.
4. Los datos deben ser exactos y actualizarse en el caso de que ello fuere necesario.
5. Los datos total o parcialmente inexactos, o que sean incompletos, deben ser suprimidos y sustituidos, o en su caso completados, por el responsable del archivo o base de datos cuando se tenga conocimiento de la inexactitud o carácter incompleto de la información de que se trate, sin perjuicio de los derechos del titular establecidos en el artículo 16 de la presente ley.
6. Los datos deben ser almacenados de modo que permitan el ejercicio del derecho de acceso de su titular.
7. Los datos deben ser destruidos cuando hayan dejado de ser necesarios o pertinentes a los fines para los cuales hubiesen sido recolectados.

ARTICULO 5° — (Consentimiento).

1. El tratamiento de datos personales es ilícito cuando el titular no hubiere prestado su consentimiento libre, expreso e informado, el que deberá constar por escrito, o por otro medio que permita se le equipare, de acuerdo a las circunstancias.

El referido consentimiento prestado con otras declaraciones, deberá figurar en forma expresa y destacada, previa notificación al requerido de datos, de la información descrita en el artículo 6° de la presente ley.

2. No será necesario el consentimiento cuando:

a) Los datos se obtengan de fuentes de acceso público irrestricto;

b) Se recaben para el ejercicio de funciones propias de los poderes del Estado o en virtud de una obligación legal;

c) Se trate de listados cuyos datos se limiten a nombre, documento nacional de identidad, identificación tributaria o previsional, ocupación, fecha de nacimiento y domicilio;

d) Deriven de una relación contractual, científica o profesional del titular de los datos, y resulten necesarios para su desarrollo o cumplimiento;

e) Se trate de las operaciones que realicen las entidades financieras y de las informaciones que reciban de sus clientes conforme las disposiciones del artículo 39 de la Ley 21.526.

ARTICULO 6° — (Información).

Cuando se recaben datos personales se deberá informar previamente a sus titulares en forma expresa y clara:

a) La finalidad para la que serán tratados y quiénes pueden ser sus destinatarios o clase de destinatarios;

b) La existencia del archivo, registro, banco de datos, electrónico o de cualquier otro tipo, de que se trate y la identidad y domicilio de su responsable;

c) El carácter obligatorio o facultativo de las respuestas al cuestionario que se le proponga, en especial en cuanto a los datos referidos en el artículo siguiente;

d) Las consecuencias de proporcionar los datos, de la negativa a hacerlo o de la inexactitud de los mismos;

e) La posibilidad del interesado de ejercer los derechos de acceso, rectificación y supresión de los datos.

ARTICULO 7° — (Categoría de datos).

1. Ninguna persona puede ser obligada a proporcionar datos sensibles.

2. Los datos sensibles sólo pueden ser recolectados y objeto de tratamiento cuando medien razones de interés general autorizadas por ley. También podrán ser tratados con finalidades estadísticas o científicas cuando no puedan ser identificados sus titulares.

3. Queda prohibida la formación de archivos, bancos o registros que almacenen información que directa o indirectamente revele datos sensibles. Sin perjuicio de ello, la Iglesia Católica, las asociaciones religiosas y las organizaciones políticas y sindicales podrán llevar un registro de sus miembros.

4. Los datos relativos a antecedentes penales o contravencionales sólo pueden ser objeto de tratamiento por parte de las autoridades públicas competentes, en el marco de las leyes y reglamentaciones respectivas.

ARTICULO 8° — (Datos relativos a la salud).

Los establecimientos sanitarios públicos o privados y los profesionales vinculados a las ciencias de la salud pueden recolectar y tratar los datos personales relativos a la salud física o mental de los pacientes que acudan a los mismos o que estén o hubieren estado bajo tratamiento de aquéllos, respetando los principios del secreto profesional.

ARTICULO 9° — (Seguridad de los datos).

1. El responsable o usuario del archivo de datos debe adoptar las medidas técnicas y organizativas que resulten necesarias para garantizar la seguridad y confidencialidad de los datos personales, de modo de evitar su adulteración, pérdida, consulta o tratamiento no autorizado, y que permitan detectar desviaciones, intencionales o no, de información, ya sea que los riesgos provengan de la acción humana o del medio técnico utilizado.
2. Queda prohibido registrar datos personales en archivos, registros o bancos que no reúnan condiciones técnicas de integridad y seguridad.

ARTICULO 10. — (Deber de confidencialidad).

1. El responsable y las personas que intervengan en cualquier fase del tratamiento de datos personales están obligados al secreto profesional respecto de los mismos. Tal obligación subsistirá aun después de finalizada su relación con el titular del archivo de datos.
2. El obligado podrá ser relevado del deber de secreto por resolución judicial y cuando medien razones fundadas relativas a la seguridad pública, la defensa nacional o la salud pública.

ARTICULO 11. — (Cesión).

1. Los datos personales objeto de tratamiento sólo pueden ser cedidos para el cumplimiento de los fines directamente relacionados con el interés legítimo del cedente y del cesionario y con el previo consentimiento del titular de los datos, al que se le debe informar sobre la finalidad de la cesión e identificar al cesionario o los elementos que permitan hacerlo.
2. El consentimiento para la cesión es revocable.
3. El consentimiento no es exigido cuando:

a) Así lo disponga una ley;

b) En los supuestos previstos en el artículo 5° inciso 2;

c) Se realice entre dependencias de los órganos del Estado en forma directa, en la medida del cumplimiento de sus respectivas competencias;

d) Se trate de datos personales relativos a la salud, y sea necesario por razones de salud pública, de emergencia o para la realización de estudios epidemiológicos, en tanto se preserve la identidad de los titulares de los datos mediante mecanismos de disociación adecuados;

e) Se hubiera aplicado un procedimiento de disociación de la información, de modo que los titulares de los datos sean inidentificables.

4. El cesionario quedará sujeto a las mismas obligaciones legales y reglamentarias del cedente y éste responderá solidaria y conjuntamente por la observancia de las mismas ante el organismo de control y el titular de los datos de que se trate.

ARTICULO 12. — (Transferencia internacional).

1. Es prohibida la transferencia de datos personales de cualquier tipo con países u organismos internacionales o supranacionales, que no proporcionen niveles de protección adecuados.

2. La prohibición no regirá en los siguientes supuestos:

a) Colaboración judicial internacional;

b) Intercambio de datos de carácter médico, cuando así lo exija el tratamiento del afectado, o una investigación epidemiológica, en tanto se realice en los términos del inciso e) del artículo anterior;

c) Transferencias bancarias o bursátiles, en lo relativo a las transacciones respectivas y conforme la legislación que les resulte aplicable;

d) Cuando la transferencia se hubiera acordado en el marco de tratados internacionales en los cuales la República Argentina sea parte;

e) Cuando la transferencia tenga por objeto la cooperación internacional entre organismos de inteligencia para la lucha contra el crimen organizado, el terrorismo y el narcotráfico.

Capítulo III

Derechos de los titulares de datos

ARTICULO 13. — (Derecho de Información).

Toda persona puede solicitar información al organismo de control relativa a la existencia de archivos, registros, bases o bancos de datos personales, sus finalidades y la identidad de sus responsables.

El registro que se lleve al efecto será de consulta pública y gratuita.

ARTICULO 14. — (Derecho de acceso).

1. El titular de los datos, previa acreditación de su identidad, tiene derecho a solicitar y obtener información de sus datos personales incluidos en los bancos de datos públicos, o privados destinados a proveer informes.
2. El responsable o usuario debe proporcionar la información solicitada dentro de los diez días corridos de haber sido intimado fehacientemente.

   Vencido el plazo sin que se satisfaga el pedido, o si evacuado el informe, éste se estimara insuficiente, quedará expedita la acción de protección de los datos personales o de hábeas data prevista en esta ley.
3. El derecho de acceso a que se refiere este artículo sólo puede ser ejercido en forma gratuita a intervalos no inferiores a seis meses, salvo que se acredite un interés legítimo al efecto.
4. El ejercicio del derecho al cual se refiere este artículo en el caso de datos de personas fallecidas les corresponderá a sus sucesores universales.

ARTICULO 15. — (Contenido de la información).

1. La información debe ser suministrada en forma clara, exenta de codificaciones y en su caso acompañada de una explicación, en lenguaje accesible al conocimiento medio de la población, de los términos que se utilicen.
2. La información debe ser amplia y versar sobre la totalidad del registro perteneciente al titular, aun cuando el requerimiento sólo comprenda un aspecto de los datos personales. En ningún caso el informe podrá revelar datos pertenecientes a terceros, aun cuando se vinculen con el interesado.
3. La información, a opción del titular, podrá suministrarse por escrito, por medios electrónicos, telefónicos, de imagen, u otro idóneo a tal fin.

ARTICULO 16. — (Derecho de rectificación, actualización o supresión).

1. Toda persona tiene derecho a que sean rectificados, actualizados y, cuando corresponda, suprimidos o sometidos a confidencialidad los datos personales de los que sea titular, que estén incluidos en un banco de datos.

2. El responsable o usuario del banco de datos, debe proceder a la rectificación, supresión o actualización de los datos personales del afectado, realizando las operaciones necesarias a tal fin en el plazo máximo de cinco días hábiles de recibido el reclamo del titular de los datos o advertido el error o falsedad.

3. El incumplimiento de esta obligación dentro del término acordado en el inciso precedente, habilitará al interesado a promover sin más la acción de protección de los datos personales o de hábeas data prevista en la presente ley.

4. En el supuesto de cesión, o transferencia de datos, el responsable o usuario del banco de datos debe notificar la rectificación o supresión al cesionario dentro del quinto día hábil de efectuado el tratamiento del dato.

5. La supresión no procede cuando pudiese causar perjuicios a derechos o intereses legítimos de terceros, o cuando existiera una obligación legal de conservar los datos.

6. Durante el proceso de verificación y rectificación del error o falsedad de la información que se trate, el responsable o usuario del banco de datos deberá o bien bloquear el archivo, o consignar al proveer información relativa al mismo la circunstancia de que se encuentra sometida a revisión.

7. Los datos personales deben ser conservados durante los plazos previstos en las disposiciones aplicables o en su caso, en las contractuales entre el responsable o usuario del banco de datos y el titular de los datos.

ARTICULO 17. — (Excepciones).

1. Los responsables o usuarios de bancos de datos públicos pueden, mediante decisión fundada, denegar el acceso, rectificación o la supresión en función de la protección de la defensa de la Nación, del orden y la seguridad públicos, o de la protección de los derechos e intereses de terceros.

2. La información sobre datos personales también puede ser denegada por los responsables o usuarios de bancos de datos públicos, cuando de tal modo se pudieran obstaculizar actuaciones judiciales o administrativas en curso vinculadas a la investigación sobre el cumplimiento de obligaciones tributarias o previsionales, el desarrollo de funciones de control de la salud y del medio ambiente, la investigación de delitos penales y la verificación de infracciones administrativas. La resolución que así lo disponga debe ser fundada y notificada al afectado.

3. Sin perjuicio de lo establecido en los incisos anteriores, se deberá brindar acceso a los registros en cuestión en la oportunidad en que el afectado tenga que ejercer su derecho de defensa.

ARTICULO 18. — (Comisiones legislativas).

Las Comisiones de Defensa Nacional y la Comisión Bicameral de Fiscalización de los Órganos y Actividades de Seguridad Interior e Inteligencia del Congreso de la Nación y la Comisión de Seguridad Interior de la Cámara de Diputados de la Nación, o las que las sustituyan, tendrán acceso a los archivos o bancos de datos referidos en el artículo 23 inciso 2 por razones fundadas y en aquellos aspectos que constituyan materia de competencia de tales Comisiones.

ARTICULO 19. — (Gratuidad).

La rectificación, actualización o supresión de datos personales inexactos o incompletos que obren en registros públicos o privados se efectuará sin cargo alguno para el interesado.

ARTICULO 20. — (Impugnación de valoraciones personales).

1. Las decisiones judiciales o los actos administrativos que impliquen apreciación o valoración de conductas humanas, no podrán tener como único fundamento el resultado del tratamiento informatizado de datos personales que suministren una definición del perfil o personalidad del interesado.

2. Los actos que resulten contrarios a la disposición precedente serán insanablemente nulos.

Capítulo IV

Usuarios y responsables de archivos, registros y bancos de datos

ARTICULO 21. — (Registro de archivos de datos. Inscripción).

1. Todo archivo, registro, base o banco de datos público, y privado destinado a proporcionar informes debe inscribirse en el Registro que al efecto habilite el organismo de control.

2. El registro de archivos de datos debe comprender como mínimo la siguiente información:

a) Nombre y domicilio del responsable;

b) Características y finalidad del archivo;

c) Naturaleza de los datos personales contenidos en cada archivo;

d) Forma de recolección y actualización de datos;

e) Destino de los datos y personas físicas o de existencia ideal a las que pueden ser transmitidos;

f) Modo de interrelacionar la información registrada;

g) Medios utilizados para garantizar la seguridad de los datos, debiendo detallar la categoría de personas con acceso al tratamiento de la información;

h) Tiempo de conservación de los datos;

i) Forma y condiciones en que las personas pueden acceder a los datos referidos a ellas y los procedimientos a realizar para la rectificación o actualización de los datos.

3) Ningún usuario de datos podrá poseer datos personales de naturaleza distinta a los declarados en el registro.

El incumplimiento de estos requisitos dará lugar a las sanciones administrativas previstas en el capítulo VI de la presente ley.

ARTICULO 22. — (Archivos, registros o bancos de datos públicos).

1. Las normas sobre creación, modificación o supresión de archivos, registros o bancos de datos pertenecientes a organismos públicos deben hacerse por medio de disposición general publicada en el Boletín Oficial de la Nación o diario oficial.

2. Las disposiciones respectivas, deben indicar:

a) Características y finalidad del archivo;

b) Personas respecto de las cuales se pretenda obtener datos y el carácter facultativo u obligatorio de su suministro por parte de aquéllas;

c) Procedimiento de obtención y actualización de los datos;

d) Estructura básica del archivo, informatizado o no, y la descripción de la naturaleza de los datos personales que contendrán;

e) Las cesiones, transferencias o interconexiones previstas;

f) Órganos responsables del archivo, precisando dependencia jerárquica en su caso;

g) Las oficinas ante las que se pudiesen efectuar las reclamaciones en ejercicio de los derechos de acceso, rectificación o supresión.

3. En las disposiciones que se dicten para la supresión de los registros informatizados se establecerá el destino de los mismos o las medidas que se adopten para su destrucción.

ARTICULO 23. — (Supuestos especiales).

1. Quedarán sujetos al régimen de la presente ley, los datos personales que, por haberse almacenado para fines administrativos, deban ser objeto de registro permanente en los bancos de datos de las fuerzas armadas, fuerzas de seguridad, organismos policiales o de inteligencia; y aquellos sobre antecedentes personales que proporcionen dichos bancos de datos a las autoridades administrativas o judiciales que los requieran en virtud de disposiciones legales.

2. El tratamiento de datos personales con fines de defensa nacional o seguridad pública por parte de las fuerzas armadas, fuerzas de seguridad, organismos policiales o inteligencia, sin consentimiento de los afectados, queda limitado a aquellos supuestos y categoría de datos que resulten necesarios para el estricto cumplimiento de las misiones legalmente asignadas a aquéllos para la defensa nacional, la seguridad pública o para la represión de los delitos. Los archivos, en tales casos, deberán ser específicos y establecidos al efecto, debiendo clasificarse por categorías, en función de su grado de fiabilidad.

3. Los datos personales registrados con fines policiales se cancelarán cuando no sean necesarios para las averiguaciones que motivaron su almacenamiento.

ARTICULO 24. — (Archivos, registros o bancos de datos privados).

Los particulares que formen archivos, registros o bancos de datos que no sean para un uso exclusivamente personal deberán registrarse conforme lo previsto en el artículo 21.

ARTICULO 25. — (Prestación de servicios informatizados de datos personales).

1. Cuando por cuenta de terceros se presten servicios de tratamiento de datos personales, éstos no podrán aplicarse o utilizarse con un fin distinto al que figure en el contrato de servicios, ni cederlos a otras personas, ni aun para su conservación.

2. Una vez cumplida la prestación contractual los datos personales tratados deberán ser destruidos, salvo que medie autorización expresa de aquel por cuenta de quien se prestan tales servicios cuando razonablemente se presuma la posibilidad de ulteriores encargos, en cuyo caso se podrá almacenar con las debidas condiciones de seguridad por un período de hasta dos años.

ARTICULO 26. — (Prestación de servicios de información crediticia).

1. En la prestación de servicios de información crediticia sólo pueden tratarse datos personales de carácter patrimonial relativos a la solvencia económica y al crédito, obtenidos de fuentes accesibles al público o procedentes de informaciones facilitadas por el interesado o con su consentimiento.

2. Pueden tratarse igualmente datos personales relativos al cumplimiento o incumplimiento de obligaciones de contenido patrimonial, facilitados por el acreedor o por quien actúe por su cuenta o interés.

3. A solicitud del titular de los datos, el responsable o usuario del banco de datos, le comunicará las informaciones, evaluaciones y apreciaciones que sobre el mismo hayan sido comunicadas durante los últimos seis meses y el nombre y domicilio del cesionario en el supuesto de tratarse de datos obtenidos por cesión.

4. Sólo se podrán archivar, registrar o ceder los datos personales que sean significativos para evaluar la solvencia económico-financiera de los afectados durante los últimos cinco años. Dicho plazo se reducirá a dos años cuando el deudor cancele o de otro modo extinga la obligación, debiéndose hace constar dicho hecho.

5. La prestación de servicios de información crediticia no requerirá el previo consentimiento del titular de los datos a los efectos de su cesión, ni la ulterior comunicación de ésta, cuando estén relacionados con el giro de las actividades comerciales o crediticias de los cesionarios.

ARTICULO 27. — (Archivos, registros o bancos de datos con fines de publicidad).

1. En la recopilación de domicilios, reparto de documentos, publicidad o venta directa y otras actividades análogas, se podrán tratar datos que sean aptos para establecer perfiles determinados con fines promocionales, comerciales o publicitarios; o permitan establecer hábitos de consumo, cuando éstos figuren en documentos accesibles al público o hayan sido facilitados por los propios titulares u obtenidos con su consentimiento.

2. En los supuestos contemplados en el presente artículo, el titular de los datos podrá ejercer el derecho de acceso sin cargo alguno.

3. El titular podrá en cualquier momento solicitar el retiro o bloqueo de su nombre de los bancos de datos a los que se refiere el presente artículo.

ARTICULO 28. — (Archivos, registros o bancos de datos relativos a encuestas).

1. Las normas de la presente ley no se aplicarán a las encuestas de opinión, mediciones y estadísticas relevadas conforme a Ley 17.622, trabajos de prospección de mercados, investigaciones científicas o médicas y actividades análogas, en la medida que los datos recogidos no puedan atribuirse a una persona determinada o determinable.

2. Si en el proceso de recolección de datos no resultara posible mantener el anonimato, se deberá utilizar una técnica de disociación, de modo que no permita identificar a persona alguna.

Capítulo V

Control

ARTICULO 29. — (Órgano de Control).

1. El órgano de control deberá realizar todas las acciones necesarias para el cumplimiento de los objetivos y demás disposiciones de la presente ley. A tales efectos tendrá las siguientes funciones y atribuciones:

a) Asistir y asesorar a las personas que lo requieran acerca de los alcances de la presente y de los medios legales de que disponen para la defensa de los derechos que ésta garantiza;

b) Dictar las normas y reglamentaciones que se deben observar en el desarrollo de las actividades comprendidas por esta ley;

c) Realizar un censo de archivos, registros o bancos de datos alcanzados por la ley y mantener el registro permanente de los mismos;

d) Controlar la observancia de las normas sobre integridad y seguridad de datos por parte de los archivos, registros o bancos de datos. A tal efecto podrá solicitar autorización judicial para acceder a locales, equipos, o programas de tratamiento de datos a fin de verificar infracciones al cumplimiento de la presente ley;

e) Solicitar información a las entidades públicas y privadas, las que deberán proporcionar los antecedentes, documentos, programas u otros elementos relativos al tratamiento de los datos personales que se le requieran. En estos casos, la autoridad deberá garantizar la seguridad y confidencialidad de la información y elementos suministrados;

f) Imponer las sanciones administrativas que en su caso correspondan por violación a las normas de la presente ley y de las reglamentaciones que se dicten en su consecuencia;

g) Constituirse en querellante en las acciones penales que se promovieran por violaciones a la presente ley;

h) Controlar el cumplimiento de los requisitos y garantías que deben reunir los archivos o bancos de datos privados destinados a suministrar informes, para obtener la correspondiente inscripción en el Registro creado por esta ley.

2. El órgano de control gozará de autonomía funcional y actuará como órgano descentralizado en el ámbito del Ministerio de Justicia y Derechos Humanos de la Nación.

3. El órgano de control será dirigido y administrado por un Director designado por el término de cuatro (4) años, por el Poder Ejecutivo con acuerdo del Senado de la Nación, debiendo ser seleccionado entre personas con antecedentes en la materia.

El Director tendrá dedicación exclusiva en su función, encontrándose alcanzado por las incompatibilidades fijadas por ley para los funcionarios públicos y podrá ser removido por el Poder Ejecutivo por mal desempeño de sus funciones.

ARTICULO 30. — (Códigos de conducta).

1. Las asociaciones o entidades representativas de responsables o usuarios de bancos de datos de titularidad privada podrán elaborar códigos de conducta de práctica profesional, que

establezcan normas para el tratamiento de datos personales que tiendan a asegurar y mejorar las condiciones de operación de los sistemas de información en función de los principios establecidos en la presente ley.

2. Dichos códigos deberán ser inscriptos en el registro que al efecto lleve el organismo de control, quien podrá denegar la inscripción cuando considere que no se ajustan a las disposiciones legales y reglamentarias sobre la materia.

Capítulo VI

Sanciones

ARTICULO 31. — (Sanciones administrativas).

1. Sin perjuicio de las responsabilidades administrativas que correspondan en los casos de responsables o usuarios de bancos de datos públicos; de la responsabilidad por daños y perjuicios derivados de la inobservancia de la presente ley, y de las sanciones penales que correspondan, el organismo de control podrá aplicar las sanciones de apercibimiento, suspensión, multa de mil pesos ($ 1.000.-) a cien mil pesos ($ 100.000.-), clausura o cancelación del archivo, registro o banco de datos.

2. La reglamentación determinará las condiciones y procedimientos para la aplicación de las sanciones previstas, las que deberán graduarse en relación a la gravedad y extensión de la violación y de los perjuicios derivados de la infracción, garantizando el principio del debido proceso.

ARTICULO 32. — (Sanciones penales).

1. Incorpórase como artículo 117 bis del Código Penal, el siguiente:

"1°. Será reprimido con la pena de prisión de un mes a dos años el que insertará o hiciera insertar a sabiendas datos falsos en un archivo de datos personales.

2°. La pena será de seis meses a tres años, al que proporcionará a un tercero a sabiendas información falsa contenida en un archivo de datos personales .

3°. La escala penal se aumentará en la mitad del mínimo y del máximo, cuando del hecho se derive perjuicio a alguna persona.

4°. Cuando el autor o responsable del ilícito sea funcionario público en ejercicio de sus funciones, se le aplicará la accesoria de inhabilitación para el desempeño de cargos públicos por el doble del tiempo que el de la condena" .

2. Incorpórase como artículo 157 bis del Código Penal el siguiente:

"Será reprimido con la pena de prisión de un mes a dos años el que:

1°. A sabiendas e ilegítimamente, o violando sistemas de confidencialidad y seguridad de datos, accediere, de cualquier forma, a un banco de datos personales;

2°. Revelare a otro información registrada en un banco de datos personales cuyo secreto

estuviere obligado a preservar por disposición de una ley.

Cuando el autor sea funcionario público sufrirá, además, pena de inhabilitación especial de uno a cuatro años".

Capítulo VII

Acción de protección de los datos personales

ARTICULO 33. — (Procedencia).

1. La acción de protección de los datos personales o de hábeas data procederá:

a) para tomar conocimiento de los datos personales almacenados en archivos, registros o bancos de datos públicos o privados destinados a proporcionar informes, y de la finalidad de aquéllos;

b) en los casos en que se presuma la falsedad, inexactitud, desactualización de la información de que se trata, o el tratamiento de datos cuyo registro se encuentra prohibido en la presente ley, para exigir su rectificación, supresión, confidencialidad o actualización.

ARTICULO 34. — (Legitimación activa).

La acción de protección de los datos personales o de hábeas data podrá ser ejercida por el afectado, sus tutores o curadores y los sucesores de las personas físicas, sean en línea directa o colateral hasta el segundo grado, por sí o por intermedio de apoderado.

Cuando la acción sea ejercida por personas de existencia ideal, deberá ser interpuesta por sus representantes legales, o apoderados que éstas designen al efecto.

En el proceso podrá intervenir en forma coadyuvante el Defensor del Pueblo.

ARTICULO 35. — (Legitimación pasiva).

La acción procederá respecto de los responsables y usuarios de bancos de datos públicos, y de los privados destinados a proveer informes.

ARTICULO 36. — (Competencia).

Será competente para entender en esta acción el juez del domicilio del actor; el del domicilio del demandado; el del lugar en el que el hecho o acto se exteriorice o pudiera tener efecto, a elección del actor.

Procederá la competencia federal:

a) cuando se interponga en contra de archivos de datos públicos de organismos nacionales, y

b) cuando los archivos de datos se encuentren interconectados en redes interjurisdicciones, nacionales o internacionales.

ARTICULO 37. — (Procedimiento aplicable).

La acción de hábeas data tramitará según las disposiciones de la presente ley y por el procedimiento que corresponde a la acción de amparo común y supletoriamente por las normas del Código Procesal Civil y Comercial de la Nación, en lo atinente al juicio sumarísimo.

ARTICULO 38. — (Requisitos de la demanda).

1. La demanda deberá interponerse por escrito, individualizando con la mayor precisión posible el nombre y domicilio del archivo, registro o banco de datos y, en su caso, el nombre del responsable o usuario del mismo.

En el caso de los archivos, registros o bancos públicos, se procurará establecer el organismo estatal del cual dependen.

2. El accionante deberá alegar las razones por las cuales entiende que en el archivo, registro o banco de datos individualizado obra información referida a su persona; los motivos por los cuales considera que la información que le atañe resulta discriminatoria, falsa o inexacta y justificar que se han cumplido los recaudos que hacen al ejercicio de los derechos que le reconoce la presente ley.

3. El afectado podrá solicitar que mientras dure el procedimiento, el registro o banco de datos asiente que la información cuestionada está sometida a un proceso judicial.

4. El Juez podrá disponer el bloqueo provisional del archivo en lo referente al dato personal motivo del juicio cuando sea manifiesto el carácter discriminatorio, falso o inexacto de la información de que se trate.

5. A los efectos de requerir información al archivo, registro o banco de datos involucrado, el criterio judicial de apreciación de las circunstancias requeridas en los puntos 1 y 2 debe ser amplio.

ARTICULO 39. — (Trámite).

1. Admitida la acción el juez requerirá al archivo, registro o banco de datos la remisión de la información concerniente al accionante. Podrá asimismo solicitar informes sobre el soporte técnico de datos, documentación de base relativa a la recolección y cualquier otro aspecto que resulte conducente a la resolución de la causa que estime procedente.

2. El plazo para contestar el informe no podrá ser mayor de cinco días hábiles, el que podrá ser ampliado prudencialmente por el juez.

ARTICULO 40. — (Confidencialidad de la información).

1. Los registros, archivos o bancos de datos privados no podrán alegar la confidencialidad de la información que se les requiere salvo el caso en que se afecten las fuentes de información periodística.

2. Cuando un archivo, registro o banco de datos público se oponga a la remisión del informe solicitado con invocación de las excepciones al derecho de acceso, rectificación o supresión, autorizadas por la presente ley o por una ley específica; deberá acreditar los extremos que hacen aplicable la excepción legal. En tales casos, el juez podrá tomar conocimiento personal y directo de los datos solicitados asegurando el mantenimiento de su confidencialidad.

ARTICULO 41. — (Contestación del informe).

Al contestar el informe, el archivo, registro o banco de datos deberá expresar las razones por las cuales incluyó la información cuestionada y aquellas por las que no evacuó el pedido efectuado por el interesado, de conformidad a lo establecido en los artículos 13 a 15 de la ley.

ARTICULO 42. — (Ampliación de la demanda).

Contestado el informe, el actor podrá, en el término de tres días, ampliar el objeto de la demanda solicitando la supresión, rectificación, confidencialidad o actualización de sus datos personales, en los casos que resulte procedente a tenor de la presente ley, ofreciendo en el mismo acto la prueba pertinente. De esta presentación se dará traslado al demandado por el término de tres días.

ARTICULO 43. — (Sentencia).

1. Vencido el plazo para la contestación del informe o contestado el mismo, y en el supuesto del artículo 42, luego de contestada la ampliación, y habiendo sido producida en su caso la prueba, el juez dictará sentencia.

2. En el caso de estimarse procedente la acción, se especificará si la información debe ser suprimida, rectificada, actualizada o declarada confidencial, estableciendo un plazo para su cumplimiento.

3. El rechazo de la acción no constituye presunción respecto de la responsabilidad en que hubiera podido incurrir el demandante.

4. En cualquier caso, la sentencia deberá ser comunicada al organismo de control, que deberá llevar un registro al efecto.

ARTICULO 44. — (Ámbito de aplicación).

Las normas de la presente ley contenidas en los Capítulos I, II, III y IV, y artículo 32 son de orden público y de aplicación en lo pertinente en todo el territorio nacional.

Se invita a las provincias a adherir a las normas de esta ley que fueren de aplicación exclusiva en jurisdicción nacional.

La jurisdicción federal regirá respecto de los registros, archivos, bases o bancos de datos interconectados en redes de alcance interjurisdiccional, nacional o internacional.

ARTICULO 45. — El Poder Ejecutivo Nacional deberá reglamentar la presente ley y establecer el organismo de control dentro de los ciento ochenta días de su promulgación.

ARTICULO 46. — (Disposiciones transitorias).

Los archivos, registros, bases o bancos de datos destinados a proporcionar informes, existentes al momento de la sanción de la presente ley, deberán inscribirse en el registro que se habilite

conforme a lo dispuesto en el artículo 21 y adecuarse a lo que dispone el presente régimen dentro del plazo que al efecto establezca la reglamentación.

ARTICULO 47. — Los bancos de datos prestadores de servicios de información crediticia deberán suprimir, o en su caso, omitir asentar, todo dato referido al incumplimiento o mora en el pago de una obligación, si ésta hubiere sido cancelada al momento de la entrada en vigencia de la presente ley.

ARTICULO 48. — Comuníquese al Poder Ejecutivo.

DADA EN LA SALA DE SESIONES DEL CONGRESO ARGENTINO, EN BUENOS AIRES, A LOS CUATRO DIAS DEL MES DE OCTUBRE DEL AÑO DOS MIL.

— REGISTRADO BAJO EL N° 25.326 —

RAFAEL PASCUAL. — JOSE GENOUD. — Guillermo Aramburu. — Mario L. Pontaquarto.

NOTA: Los textos en negrita fueron observados.

Anexo 2 Scripts

Informe Estadístico

```
USE [IngresoDataWarehouse]
GO
/****** Object:  Table [dbo].[DIM_AnioAcademico]    Script Date: 22/07/2023 10:27:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[DIM_AnioAcademico](
	[id_anio] [int] NOT NULL,
	[activo] [bit] NULL,
	[actual] [bit] NULL,
 CONSTRAINT [PK_DIM_AnioAcademico] PRIMARY KEY CLUSTERED
(
	[id_anio] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
/****** Object:  Table [dbo].[DIM_Carrera]    Script Date: 22/07/2023 10:27:29 a.m. ******/
```

```
SET ANSI_NULLS ON

GO

SET QUOTED_IDENTIFIER ON

GO

SET ANSI_PADDING ON

GO

CREATE TABLE [dbo].[DIM_Carrera](

        [id] [varchar](5) NOT NULL,

        [nombre] [varchar](255) NULL,

        [nombreReducido] [varchar](30) NULL,

        [grupoCarrera] [varchar](5) NULL,

        [grupoNombre] [varchar](60) NULL,

 CONSTRAINT [PK_DIM_Carrera] PRIMARY KEY CLUSTERED

(

        [id] ASC

)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]

) ON [PRIMARY]


GO

SET ANSI_PADDING OFF

GO

/****** Object:  Table [dbo].[DIM_GrupoCarrera]    Script Date: 22/07/2023 10:27:29 a.m.
******/

SET ANSI_NULLS ON
```

```
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO



CREATE TABLE [dbo].[DIM_GrupoCarrera](
	[id_GrupoCarrera] [varchar](5) NOT NULL,
	[nombre] [varchar](60) NULL,
	[observaciones] [varchar](255) NULL,
 CONSTRAINT [PK_DIM_GrupoCarrera] PRIMARY KEY CLUSTERED
(
	[id_GrupoCarrera] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]
) ON [PRIMARY]


GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[Fact_InformeEstadistico]    Script Date: 22/07/2023 10:27:29
a.m. ******/
SET ANSI_NULLS ON
```

```
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[Fact_InformeEstadistico](
    [id_anio] [int] NOT NULL,
    [id_GrupoCarrrera] [varchar](5) NOT NULL,
    [id_carrera] [varchar](5) NOT NULL,
    [promedioHombres] [decimal](5, 2) NULL,
    [promedioMujeres] [decimal](5, 2) NULL,
    [cantidadHombres] [int] NULL,
    [cantidadMujeres] [int] NULL,
    [cantidadInscriptos] [int] NULL,
    [cantidadaPrimario] [int] NULL,
    [cantidadSecundarioIncompleto] [int] NULL,
    [cantidadSecundarioCompleto] [int] NULL,
    [cantidadTerciario] [int] NULL,
    [cantidadUniversitarioIncompleto] [int] NULL,
    [cantidadUniversitarioCompleto] [int] NULL,
    [cantidadTrabaja] [int] NULL,
    [cantidadNoTrabaja] [int] NULL,
    [cantidadEscuelaPublica] [int] NULL,
    [cantidadEscuelaPrivada] [int] NULL,
    [cantidadArgentino] [int] NULL,
    [cantidadExtranjero] [int] NULL
```

```
) ON [PRIMARY]


GO

SET ANSI_PADDING OFF

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_AnioAcademico] FOREIGN KEY([id_anio])

REFERENCES [dbo].[DIM_AnioAcademico] ([id_anio])

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico] CHECK CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_AnioAcademico]

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_Carrera] FOREIGN KEY([id_carrera])

REFERENCES [dbo].[DIM_Carrera] ([id])

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico] CHECK CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_Carrera]

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_GrupoCarrera] FOREIGN KEY([id_GrupoCarrrera])

REFERENCES [dbo].[DIM_GrupoCarrera] ([id_GrupoCarrera])

GO

ALTER TABLE [dbo].[Fact_InformeEstadistico] CHECK CONSTRAINT
[FK_Fact_InformeEstadistico_DIM_GrupoCarrera]

GO
```

Notas Finales

```
USE [IngresoDataWarehouse]
GO
/****** Object:  Table [dbo].[DIM_AnioAcademico]    Script Date: 22/07/2023 10:29:27 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[DIM_AnioAcademico](
	[id_anio] [int] NOT NULL,
	[activo] [bit] NULL,
	[actual] [bit] NULL,
 CONSTRAINT [PK_DIM_AnioAcademico] PRIMARY KEY CLUSTERED
(
	[id_anio] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
/****** Object:  Table [dbo].[DIM_Carrera]    Script Date: 22/07/2023 10:29:27 a.m. ******/
SET ANSI_NULLS ON
GO
```

```
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[DIM_Carrera](
	[id] [varchar](5) NOT NULL,
	[nombre] [varchar](255) NULL,
	[nombreReducido] [varchar](30) NULL,
	[grupoCarrera] [varchar](5) NULL,
	[grupoNombre] [varchar](60) NULL,
 CONSTRAINT [PK_DIM_Carrera] PRIMARY KEY CLUSTERED
(
	[id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO

/****** Object:  Table [dbo].[DIM_GrupoCarrera]    Script Date: 22/07/2023 10:29:27 a.m.
******/
SET ANSI_NULLS ON
GO
```

```
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[DIM_GrupoCarrera](
	[id_GrupoCarrera] [varchar](5) NOT NULL,
	[nombre] [varchar](60) NULL,
	[observaciones] [varchar](255) NULL,
 CONSTRAINT [PK_DIM_GrupoCarrera] PRIMARY KEY CLUSTERED
(
	[id_GrupoCarrera] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[DIM_Materia]    Script Date: 22/07/2023 10:29:27 a.m.
******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
```

```
GO
CREATE TABLE [dbo].[DIM_Materia](
	[id_materia] [varchar](4) NOT NULL,
	[nombre] [varchar](255) NULL,
	[numero_materia] [int] NULL,
	[nombre_carrera] [varchar](255) NULL,
 CONSTRAINT [PK_DIM_Materia] PRIMARY KEY CLUSTERED
(
	[id_materia] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[Fact_NotasFinales]    Script Date: 22/07/2023 10:29:27 a.m.
******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[Fact_NotasFinales](
```

```
	[id] [int] IDENTITY(1,1) NOT NULL,
	[id_anio] [int] NULL,
	[id_GrupoCarrera] [varchar](5) NULL,
	[id_carrera] [varchar](5) NULL,
	[id_materia] [varchar](4) NULL,
	[CantidadPromocionado] [int] NULL,
	[CantidadAusentes] [int] NULL,
	[CantidadDesaprobados] [int] NULL,
	[CantidadAprobados] [int] NULL,
 CONSTRAINT [PK_Fact_NotasFinales] PRIMARY KEY CLUSTERED
(
	[id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF,
IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS =
ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
ALTER TABLE [dbo].[Fact_NotasFinales]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_NotasFinales_DIM_AnioAcademico] FOREIGN KEY([id_anio])
REFERENCES [dbo].[DIM_AnioAcademico] ([id_anio])
GO
ALTER TABLE [dbo].[Fact_NotasFinales] CHECK CONSTRAINT
[FK_Fact_NotasFinales_DIM_AnioAcademico]
GO
```

```
ALTER TABLE [dbo].[Fact_NotasFinales]  WITH CHECK ADD  CONSTRAINT
[FK_Fact_NotasFinales_DIM_Carrera] FOREIGN KEY([id_carrera])

REFERENCES [dbo].[DIM_Carrera] ([id])

GO

ALTER TABLE [dbo].[Fact_NotasFinales] CHECK CONSTRAINT
[FK_Fact_NotasFinales_DIM_Carrera]

GO

ALTER TABLE [dbo].[Fact_NotasFinales]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_NotasFinales_DIM_GrupoCarrera] FOREIGN KEY([id_GrupoCarrera])

REFERENCES [dbo].[DIM_GrupoCarrera] ([id_GrupoCarrera])

GO

ALTER TABLE [dbo].[Fact_NotasFinales] CHECK CONSTRAINT
[FK_Fact_NotasFinales_DIM_GrupoCarrera]

GO

ALTER TABLE [dbo].[Fact_NotasFinales]  WITH NOCHECK ADD  CONSTRAINT
[FK_Fact_NotasFinales_DIM_Materia] FOREIGN KEY([id_materia])

REFERENCES [dbo].[DIM_Materia] ([id_materia])

GO

ALTER TABLE [dbo].[Fact_NotasFinales] CHECK CONSTRAINT
[FK_Fact_NotasFinales_DIM_Materia]

GO
```

Staging Area

```
USE [IngresoStagingArea]
GO
/****** Object:  Table [dbo].[AnioAcademico]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[AnioAcademico](
	[id_anio] [int] NOT NULL,
	[activo] [bit] NULL,
	[actual] [bit] NULL,
 CONSTRAINT [PK_AnioAcademico] PRIMARY KEY CLUSTERED
(
	[id_anio] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]


GO
/****** Object:  Table [dbo].[Carrera]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
```

```
CREATE TABLE [dbo].[Carrera](
	[id] [varchar](5) NOT NULL,
	[nombre] [varchar](255) NULL,
	[nombreReducido] [varchar](30) NULL,
	[grupoCarrera] [varchar](5) NULL,
	[grupoNombre] [varchar](60) NULL,
 CONSTRAINT [PK_Carrera] PRIMARY KEY CLUSTERED
(
	[id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[ExtractLog]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[ExtractLog](
	[DataSource] [varchar](20) NULL,
	[LastExtract] [datetime] NULL,
	[LastVersion] [int] NULL
) ON [PRIMARY]
```

```
GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[GrupoCarrera]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[GrupoCarrera](
	[id_GrupoCarrera] [varchar](5) NOT NULL,
	[nombre] [varchar](60) NULL,
	[observaciones] [varchar](255) NULL,
 CONSTRAINT [PK_GrupoCarrera] PRIMARY KEY CLUSTERED
(
	[id_GrupoCarrera] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[Materia]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
```

```
SET QUOTED_IDENTIFIER ON
GO
SET ANSI_PADDING ON
GO
CREATE TABLE [dbo].[Materia](
	[id_materia] [varchar](4) NOT NULL,
	[nombre] [varchar](255) NULL,
	[numero_materia] [int] NULL,
	[nombre_carrera] [varchar](255) NULL,
 CONSTRAINT [PK_Materia] PRIMARY KEY CLUSTERED
(
	[id_materia] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
SET ANSI_PADDING OFF
GO
/****** Object:  Table [dbo].[Sexo]    Script Date: 22/07/2023 11:25:29 a.m. ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[Sexo](
	[id] [int] NOT NULL,
	[descripcion] [nvarchar](50) NULL,
 CONSTRAINT [PK_Sexo] PRIMARY KEY CLUSTERED
```

```
(
        [id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]

GO
```