risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 09/12/2017
Aceitação/Acceptance: 06/02/2018

Una Metodología para Procesos Data WareHousing Basada en la Experiencia

Wilson Castillo-Rojas [1], Fernando Medina Quispe [2], Francisco Fariña Molina [2]

wilson.castillo@uda.cl, femedina@unap.cl, franciscofarina@unap.cl

[1] Universidad de Atacama, Facultad de Ingeniería / DIICC, 1530000, Copiapó, Chile.

[2] Universidad Arturo Prat, Facultad de Ingeniería y Arquitectura, 1100000, Iquique, Chile.

DOI: 10.17013/risti.26.83–103

Resumen: El artículo presenta una nueva metodología para procesos data warehousing, que integra diversos enfoques, técnicas y metodologías, tales como: especificación de requisitos de información, modelamiento relacional, modelo de desarrollo combinado a partir de las propuestas de Kimball y Hefesto, un proceso aumentado de extracción-transformación y carga que incorpora explícitamente una fase de validación de indicadores, y finalmente visualizaciones integradas e interactivas para el análisis multidimensional de los indicadores obtenidos. La metodología propuesta no sólo se basa en los aspectos teóricos descritos en el artículo, sino que además en la experiencia lograda por parte del equipo investigador, en el desarrollo de diversos proyectos de data warehousing, principalmente orientados a la generación de indicadores de productividad académica y de gestión de una universidad, lo que corresponde al caso de estudio de aplicación de la metodología que se describe. Los resultados de éxito en diferentes proyectos donde ha sido utilizada esta metodología avalan su eficacia.

Palabras-clave: Inteligencia de negocios, almacén de datos, data warehousing, datamart, ETL, OLAP.

A Methodology for Data WareHousing Processes Based on Experience

Abstract: The article presents a new methodology for data warehousing processes, which integrates different approaches, techniques and methodologies, such as: specification of information requirements, relational modeling, combined development model based on the Kimball and Hefesto proposals, an increased process of extraction-transformation and load that explicitly incorporates a validation phase of indicators, and finally integrated and interactive visualizations for the multidimensional analysis of the obtained indicators. The proposed methodology is not only based on the theoretical aspects described in the article, but also on the experience gained by the research team in the development of various data warehousing projects, mainly oriented to the generation of indicators of academic productivity and management of a university, which corresponds to the case study application of the methodology described. The results of success in different projects where this methodology has been used guarantee its effectiveness.

***Keywords**:* Business intelligence, data warehouse, data warehousing, datamart, ETL, OLAP.

1. Introducción

Actualmente las organizaciones utilizan la información y el conocimiento para apoyar la toma de sus decisiones estratégicas, y de este modo lograr sus metas y mejorar sus procesos. Lo anterior, en un contexto actual de alta competitividad y requerimiento de certificación de procesos, globalización de mercados, etc. En el ámbito de educación superior, las universidades en Chile se encuentran en procesos de alta exigencia para asegurar la calidad en su gestión administrativa y académica, a través de procesos de acreditación institucional donde estas instituciones deben rendir cuentas de toda su actividad académica.

Uno de los desafíos que enfrentan hoy las organizaciones, es el aumento de datos, lo que ha generado dos grandes problemas; el primero, identificar los datos relevantes para dar seguimiento a su estrategia organizacional, y lograr que se cumplan los planes con las metas establecidas. Y el segundo problema, la capacidad para administrar esta gran cantidad de datos.

Para lo anterior, las organizaciones requieren no solamente sistemas de información para dar soporte tecnológico a sus procesos operacionales, sino que requieren de soluciones tecnológicas que les proporcionen, por un lado, indicadores que les permitan medir el desempeño de sus procesos de gestión en general y, por otro lado, patrones o reglas que les permitan describir o predecir comportamientos en sus movimientos o transacciones de sus procesos de negocios. Todo esto, para apoyar la toma de decisión estratégica de una manera mejor preparada (Illescas, Sanchez & Canziani, 2015), (Prieto & Piattini, (2015).

Lo descrito se puede lograr con el uso de tecnologías y herramientas de análisis de datos masivos, que permiten la administración y generación de conocimiento utilizando data histórica de la organización. Entre estas tecnologías se encuentran; herramientas ETL, almacenes de datos o Data Warehouse (DW), herramientas OLAP, minería de datos y otras. Y que son parte de lo que se conoce como Inteligencia de Negocios (IN o BI; acrónimo del inglés Business Intelligence).

Una definición formal de IN, es que corresponde al conjunto de estrategias y herramientas tecnológicas enfocadas a la administración y creación de conocimiento, a través del análisis de datos existentes en una organización (Back, 2002) (Golfarelli, Rizzi & Cella, 2004). Si bien un DW corresponde al repositorio de datos central de un sistema de IN, su diseño, desarrollo e implementación son parte de lo que se denomina Proceso de Data Warehousing (PDW).

Estas tecnologías son muy demandadas actualmente por las organizaciones, que teniendo satisfechas las prestaciones técnicas de sus sistemas informáticos para apoyar los procesos operacionales, requieren de nuevas prestaciones que les permitan mantenerse competitivas en un entorno cambiante y globalizado, siendo los DW una de las tecnologías IN con más desarrollo de proyectos en Chile (Palocsay, Markham & Markham, 2010).

En las instituciones de educación superior, también ha ido en aumento la cantidad de universidades, que utilizan la tecnología de IN para obtener indicadores de su gestión, y con esto enfrentar de mejor manera un proceso de acreditación. Para esto, requieren contar con: unidades administrativas de análisis institucional, recurso humano preparado en tecnología de IN, herramientas de IN, y proyectos de desarrollo de sistemas orientados hacia la toma de decisiones estratégicas.

Bajo este contexto, el artículo presenta y describe una metodología para el desarrollo de PDW que se basa principalmente en la experiencia de desarrollo de varios proyectos llevados a cabo en la Universidad Arturo Prat de Chile (UNAP). Proyectos orientados a obtener indicadores de productividad académica y de gestión de la universidad. No obstante, para elaborar y formalizar esta nueva metodología, se toman como base conceptual dos modelos de desarrollo de PDW, por un lado, el modelo clásico de desarrollo para un DW de (Kimball & Ross, 2002), combinado con la actual metodología Hefesto de (Bernabeu, 2010), de ambas se extraen las fases comunes y se refunden.

Adicionalmente, la metodología propuesta integra diversos enfoques, técnicas y metodologías, tales como: especificación de requisitos de información utilizada en ingeniería de software, proceso aumentado de ETL con una fase de validación de indicadores (ETL+V; Extraction, Transformation, Loading, and Validation), y visualizaciones integradas e interactivas para el análisis multidimensional de los indicadores, basado en el concepto de tablero de control (dashboard en inglés).

La estructura del artículo es la siguiente: la sección 2 provee un marco teórico sobre conceptos y tecnologías asociadas al trabajo. La sección 3 describe la metodología propuesta. La sección 4 presenta la aplicación de esta metodología en el desarrollo de un PDW para la generación de indicadores de productividad académica de una universidad. En la sección 5 se presentan las conclusiones del trabajo. Finalmente, en la sección 6 se listan las referencias bibliográficas.

2. Marco Teórico

En esta era conocida como la sociedad del conocimiento, una de las características tecnológicas es la utilización y consolidación de las bases de datos, para el registro de transacciones a través de Internet, lo cual representa actualmente, la herramienta principal que proporciona información en una organización. Sin embargo, con el paso de los años esta información tiende a crecer y generar grandes volúmenes de datos, por lo que surge la necesidad de cómo manejarla y qué hacer con ella. Es aquí donde aparece el término IN, que surge como solución para analizar y explotar áreas específicas de información, generando nuevas perspectivas y conocimiento con el fin de apoyar la toma de decisiones.

Una de las primeras definiciones de IN se conoce a partir de 1958 con Hans Petter Luhn, investigador de IBM que lo define como: *"La habilidad de aprehender las relaciones de hechos presentados de forma que guíen las acciones hacia una meta deseada"* (Conesa & Curto, 2010), para luego en la década de los 60 dar origen a los Sistemas de Soporte de Decisiones (SSD). Luego evolucionan en Sistemas de Información Ejecutivos (SIE), que son sistemas más robustos capaces de generar reportes consolidados. Si bien estos

sistemas cumplen su labor en la administración de la información, aún son insuficientes, ya que estas herramientas no cuentan adecuadamente, con visualización y explotación de la información.

El término IN se acuña formalmente en el año 1989 por Howard Dresden, analista de Gartner que lo define como: *"Conceptos y métodos para mejorar las decisiones de negocio mediante el uso de sistemas de soporte basado en hechos"* (Venter, 2005). IN actualmente se encuentra presente en las organizaciones más prestigiosas del mundo, ya que sirve como respaldo y soporte a la toma de decisiones de la parte estratégica de una organización, mediante el análisis de gran cantidad de datos en forma rápida y sencilla, que son procesados según reglas y criterios del negocio. IN es un concepto que se aplica de forma transversal a cualquier tipo de negocio, además de integrar información de diferentes procesos, en distintos periodos, generando un análisis completo de la situación actual de la organización, con el fin de progresar en el rendimiento interno de sus procesos.

Desde un punto de vista pragmático, IN es el conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento a través del análisis de datos existentes en una organización. Esto permite a la organización, contar con información privilegiada para responder a problemas internos del negocio como: optimización de costos, rentabilidad tanto de clientes, obtener perfiles de clientes, mejorar procesos de producción entre otras, todo lo anterior se puede responder de forma rápida y eficiente, con un buen proceso e implementación de la IN, lo cual brinda a la organización una ventaja competitiva y estratégica en el mercado (Mazón, Trujillo & Lechtemborger, 2007).

Una de las tecnologías centrales de IN son los DW y DataMart (DM). Según la clásica y más referenciada definición de Inmon, un DW: "*es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza*" (Inmon, 2005).

Un DW corresponde al almacén de datos corporativo que abarca todas las áreas o procesos de la organización. El origen de esta información proviene de diferentes fuentes de datos, ya sean de sistemas operacionales, plantillas de cálculos o archivos de textos planos. Estas fuentes de datos se integran para darle un formato homogéneo y consistente a los datos. Además, un DW se compone de bases de datos multidimensionales denominadas DM, que se enfocan al análisis de un área de negocio en particular, y tiene como objeto proporcionar indicadores o KPI.

Existen 2 enfoques respecto a los conceptos DW y DM (Kimball & Ross, 2002):

- Bill Inmon propone la idea que los DM se sirven de un DW para extraer información que esta almacenada de manera estructurada en un modelo relacional, a esto se le conoce como enfoque top-down.
- Ralph Kimball, plantea la idea de un enfoque dimensional para el diseño de un DW, y afirma que la unión de todos los DM de una organización constituye el DW corporativo, a lo cual se le conoce como el enfoque bottom-up.

El esquema convencional de un sistema de IN lo componen las siguientes etapas (Vaisman & Zimányi, 2014):

i. Especificación de requisitos estratégicos; que corresponde a la definición del proyecto IN, la naturaleza del negocio y sus propósitos, así como la especificación de indicadores claves de desempeño (Key Performance Indicator ó KPI), que se requieren medir en los procesos de la organización.
ii. El proceso ETL; tiene como función la integración de los datos provenientes desde las distintas fuentes heterogéneas, ya sean de sistemas transaccionales, archivo de textos, planillas de cálculos, etc. La integración consiste en la extracción, transformación, cálculos preliminares de KPI, limpieza y homogenización de datos, y carga de datos en el DW. Para esto, se pueden utilizar herramientas ETL, lenguajes de programación y/o lenguaje de consultas de base de datos relacional (Structured Query Language ó SQL).
iii. El DW; corresponde al repositorio de datos, cuyo diseño conceptual está basado en un Modelo Multidimensional (MM). Esto es, considera a la base de datos como un conjunto de hechos u objeto de análisis, y dimensiones que son los puntos de vistas desde los que se pueden analizar estos hechos. Las informaciones relevantes sobre los hechos se representan por un conjunto de indicadores o medidas (valores numéricos) y que corresponde a los KPI a calcular. La información descriptiva de cada dimensión se presenta por un conjunto de atributos alfanuméricos. La tecnología utilizada por lo general son sistemas gestores de base de datos relacionales.
iv. Explotación del DW; a través de herramientas de análisis de datos tales como: OLAP (On-Line Analytical Processing), data mining, generadores de reportes y gráficos estadísticos, dashboard, y en general todo tipo de SSD.

Como se señala en la sección anterior, un PDW corresponde al proceso que permite el diseño, implementación y explotación de un DW. Por lo general, un PDW incluye las estrategias y tecnologías descritas en i), ii), iii), y respecto a la explotación del DW, las que tienen relación al procesamiento analítico en línea (OLAP).

3. Metodología Propuesta

La metodología elaborada y propuesta en este trabajo, recoge la experiencia obtenida por el desarrollo de diferentes proyectos de PDW, orientados a generar indicadores o KPI que miden el desempeño y productividad, tanto académica como de gestión, de una universidad chilena. Desde el punto de vista conceptual, esta metodología integra diversos enfoques, técnicas y metodologías.

La elaboración de esta metodología, se basa en el modelo clásico de desarrollo para un DW de (Kimball & Ross, 2002) combinado con la actual metodología Hefesto de (Bernabeu, 2010), de ambas se extraen las fases comunes y se refunden. La metodología de Kimball se basa en lo que denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). Se consideran de esta metodología las siguientes etapas: definición de requerimientos del negocio, modelado dimensional, diseño físico, y diseño e implementación del subsistema ETL. Por otro lado, la metodología Hefesto consta de 4 etapas: análisis de requerimientos, análisis de los OLTP, modelo lógico del DW, e

integración de datos. Estas etapas son consideradas, pero refundiéndolas y dando un orden de desarrollo distinto.

Como se puede observar en la Figura 1, se funden las primeras etapas de ambas metodologías, en una sola etapa inicial llamada Análisis en la nueva metodología, agregando a ésta la segunda etapa de Hefesto (análisis de los OLTP). Luego, se combinan las etapas modelado dimensional y diseño físico de Kimball con la tercera etapa de Hefesto (modelo lógico del DW), lo que da a lugar la segunda etapa denominada Diseño MM. En la tercera etapa denominada Proceso ETL+V, se integra la etapa diseño e implementación del subsistema ETL de Kimball, con la etapa integración de datos de Hefesto. La cuarta etapa Procesamiento Analítico, corresponde a una fusión entre lo que propone Kimball en su ciclo de vida de un DW y las de varios autores, que consideran a OLAP, como una opción de herramienta, entre otras, para la explotación del almacén y no como parte de su método de desarrollo.

Cada una de estas etapas tienen definidas sus entradas y salidas, representadas a través de flechas con línea segmentada. A su vez, estas etapas tienen un sentido bidireccional de funcionamiento, que permite retroceder a la anterior o avanzar a la siguiente, sin perder el sentido del PDW. También, se establece la relación y participación de los usuarios, con flecha de línea continua y en sentido bidireccional, tanto en la entrada al proceso, como en la salida.

Se destacan en esta metodología, elementos claves señalados de la Figura 1 con un círculo de color rojo: plantilla ad-hoc para la especificación de los KPI, repositorio temporal como área de staging, el código SQL de validación de los KPI, y las visualizaciones gráficas integradas a través de tableros de control o dashboard. Lo anterior, según las conclusiones del equipo de desarrollo del DW y los usuarios finales, favorecieron la efectividad y éxito del desarrollo de los proyectos en que se ha utilizado esta metodología. Se explica a continuación, cada una de las etapas de la metodología propuesta en este trabajo.

3.1. Etapa Análisis.

Esta etapa inicial se vincula fuertemente con los usuarios estratégicos. Consta de dos actividades; análisis de los requerimientos de KPI, y análisis de los sistemas operacionales. Para la primera actividad, la entrada corresponde a una plantilla ad-hoc diseñada para la especificación de los requisitos de información estratégica o KPI. La estructura de esta plantilla, es considerada como uno de los elementos claves en el PDW, y se diseña tomando como referencia las plantillas y patrones lingüísticos de (Duran, 2000), que están enfocados para el proceso de educción de requisitos de información en un proyecto de software. Esta plantilla, permite definir y entender claramente, por parte de todos los usuarios, los indicadores obtenidos, tanto como interfaz o herramienta, en la educción inicial de los requisitos de KPI, su análisis, así como en la validación de cada uno. La salida de esta actividad, es el conjunto de KPI especificados a través de la plantilla diseñada.

La segunda actividad de esta etapa, corresponde al análisis de los sistemas operacionales, cuyas entradas la componen las fuentes de datos de la organización, ya sean internas o externas, más toda la información acerca de los datos que se dispongan, tales como

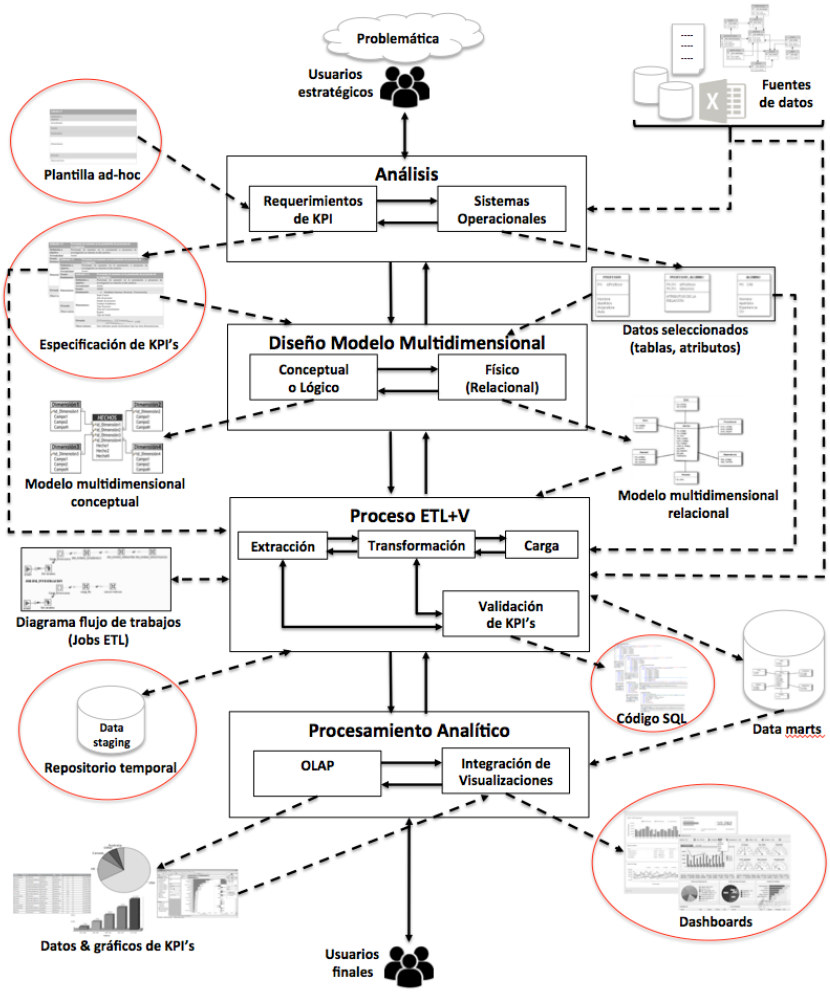

Figura 1 – Metodología propuesta para PDW (elaboración propia)

modelo dato-lógico, meta-datos, etc. Estas fuentes de datos, deben ser analizadas para contrastar los KPI especificados, y entender con qué datos poder realizar sus cálculos, o en caso de no contar con ellos poder obtenerlos. Tiene como salida esta actividad, el meta-datos con la selección de tablas, atributos y datos, que son necesarios para lograr calcular los KPI requeridos.

3.2. Etapa Diseño del MM.

Las salidas de la etapa de análisis, son las entradas en esta etapa. Y en ella, se encuentran dos actividades; el diseño conceptual y el diseño físico (relacional) del MM, cuyas salidas, corresponden a los modelos bajo el mismo nombre, conceptual y físico respectivamente. Para esto, es necesario analizar los KPI especificados junto al meta-datos de las fuentes

de origen, para poder realizar el diseño conceptual. La parte tecnológica de esta etapa la constituye las bases de datos.

Se recomienda utilizar una herramienta de la tecnología CASE (Computer Aided Software Engineering), donde se puede utilizar una de las notaciones gráficas existentes para diagramar el MM, ya sea entidad-relación, UML (Unified Modeling Language), u otra. Para generar el diseño físico del MM, en este caso relacional ya que se utiliza el enfoque de Kimball, basta con utilizar la opción en la herramienta CASE para generarlo, indicando algunos parámetros como: el motor de base de datos, nombre del servidor, y otros datos de conexión.

3.3. Etapa Proceso ETL+V.

La tercera etapa atañe al clásico proceso ETL de extracción, transformación y carga, aumentado con una cuarta actividad incorporada explícitamente, y denominada validación de KPI. Se trata de la etapa más técnica y que mayor tiempo demanda, la de mayor complejidad, y la que más entradas y salidas tiene. Las entradas a esta etapa son; la especificación de los KPI, el MM relacional, la descripción (meta-datos) de los datos seleccionados, y las fuentes de datos originales. Sus salidas corresponden a; diagrama de flujo de trabajo del proceso ETL (generada a través de una herramienta o software para esto, tal como Kettle de Pentaho u otras, o simplemente con el uso de lenguaje SQL), el repositorio temporal o área de staging, los DM cargados con los KPI, y el código SQL de validación.

La etapa ETL+V, cuya variante es la más destacada en este trabajo, tiene como función realizar toda la extracción y cálculo preliminar de los datos desde las fuentes de datos originales, y que necesarios para el cálculo de los KPI, para luego ser depositados en el repositorio temporal relacional. En este repositorio, se realizan todas las transformaciones, cálculos, agregaciones y validaciones de calidad de los datos. También se debe realizar la actividad de carga de los datos hacia los DM con los KPI ya calculados.

Sin embargo, antes de la carga es necesario verificar la validez de los KPI. Para esto, la actividad de validación, permite realizar la revisión de estos en forma eficiente, utilizando el lenguaje de consulta para base de datos relacionales (SQL), y así poder realizar consultas, cálculos y validaciones necesarias. Se trata de verificar que los KPI obtenidos sean consistentes, con los datos integrados en el repositorio temporal, las fuentes de datos originales, y el DM.

3.4. Etapa Procesamiento Analítico.

La última etapa corresponde a la explotación y análisis de la información contenida en los DM, y para esto se requiere una herramienta de la tecnología OLAP, con la cual generar el análisis multidimensional de los datos, y las visualizaciones gráficas de los KPI. Adicionalmente, se considera como segunda actividad la integración de visualizaciones de los distintos objetos de análisis y KPI, utilizando el concepto de tablero de control o dashboard (Marcus, 2006), y que permiten proporcionar a los usuarios finales, una mirada integral de los indicadores provistos por los DM, a través de interfaces gráficas y con reportabilidad interactiva e intuitiva.

4. Caso de Uso: PDW de Productividad Académica

Se describe a continuación, la aplicación de la metodología propuesta para un caso de uso en particular, y que corresponde a un PDW cuyo objetivo es generar indicadores de productividad académica de una universidad chilena.

4.1. Descripción de la Problemática.

En su plan estratégico institucional, la UNAP declara como eje estratégico la "*Gestión Moderna y competitiva*", que apunta a administrar de manera eficiente los procesos de gestión institucional tanto en recursos humanos, infraestructura y sistemas de información. Para esto, se requiere el fortalecimiento de los sistemas de información a nivel de reportes de indicadores, para la toma de decisiones internas en vicerrectorías, sedes, facultades y centros docentes de la institución.

La Unidad de Análisis Institucional (UAI) de la UNAP, es la encargada de generar y administrar los DM de la institución, con el fin de poder entregar informes tanto cualitativos como cuantitativos del estado actual de la gestión de las diferentes áreas de la institución. Para esta ocasión, la UAI plantea la necesidad del desarrollo de un PDW cuyo objetivo es obtener indicadores sobre la productividad académica. Esto, con el propósito de dar respuesta a requerimientos de información tanto de entidades internas y externas a la Universidad. Además, se busca definir y establecer criterios para la obtención de estos indicadores.

Por lo anterior, se realiza un análisis de los procesos de internos involucrados en la productividad académica de todos los funcionarios del estamento académico de la institución. Se plantean tres aspectos fundamentales dentro de sus actividades que son: docencia, investigación y vinculación con el medio. Para este trabajo, se llega a acuerdo en abordar en las dos primeras labores, ya que son las más requeridas en el trabajo diario de la universidad.

En cuanto a la actividad de docencia, los indicadores buscan conocer como es la distribución del personal académico en las distintas institutos, sedes, facultades y carreras de la universidad, como también la cantidad de horas semanales que dedican a sus labores docentes. También es relevante conocer la cantidad de jornadas completas equivalentes que los docentes representan en la institución, ya que es un indicador que se compara con la cantidad de alumnos matriculados, con el fin de establecer la relación que existen entre ellos.

En lo que concierne a la actividad de investigación, los indicadores deben permitir establecer la producción científica de la universidad, a través de medidas tales como la cantidad de publicaciones y proyectos de investigación realizadas por los académicos de las distintas institutos, sedes, facultades y carreras. Además, se buscan determinar porcentajes y tasas de aumento entre los periodos estudiados. Otro aspecto importante es, conocer el porcentaje de docentes con grado de doctor que realizan las labores de investigación, ya que con este indicador se pueden tomar medidas preventivas o de incentivo, para aumentar la productividad científica, y de ser necesario, contratar nuevos académicos con grado de doctor.

4.2. Etapa Análisis: Requerimientos de KPI.

La toma de requerimientos se lleva a cabo a través de técnicas de educción que son utilizadas en un proceso de software, con la salvedad que el tipo de usuario corresponde al nivel estratégico de la institución. Por lo anterior, se realizan un conjunto de reuniones ejecutivas con las dos vicerrectorías (académica y de investigación) en conjunto con la UAI, que están directamente ligadas a la problemática, con el fin de tomar las inquietudes y necesidades sobre la productividad académica de estas entidades tanto en sus actividades docentes y de investigación. Y de este modo, lograr definir y especificar los KPI requeridos.

También, se consideran relevante en esta etapa, los requerimientos externos provistos por instituciones de gobierno y otras, tales como: Consejo Nacional de Acreditación (CNA), Sistema de Información de Educación Superior (SIES), Consejo de Rectores de Universidad Chilenas (CRUCH), y el Consejo Nacional de Educación (CNED). Por esto, se realiza también un análisis de todos los informes entregados por estas instituciones, y de sus instructivos de solicitud de información. Con ambas visiones, interna y externa, se consolidan los requerimientos de productividad académica, de docencia y de investigación, utilizando la plantilla de requerimientos ad-hoc diseñada para esta actividad. Como muestra se presenta en la Figura 2, la especificación de un KPI de investigación, y uno de docencia en la Figura 3, de un total de catorce especificados (7 de cada uno).

Indicador 11	Tasa anual de Crecimiento en la adjudicación de proyectos de investigación
Definición u objetivo	Representa el valor de aumento o disminución de la relación que existe entre los proyectos adjudicado de un periodo a otro.
Periodicidad	Anual.
Fuente	• Entidades Externas: - • Entidades Internas: Rectoría, Vicerrectorías
Destinatario	VRIIP
Dimensiones	Sede-Centro Año de proceso Estado de proyecto Unidad Académica Tipo Proyecto Área del conocimiento. Región Tipo de fondo
Fórmula	$\frac{\sum N°Proyectos_{Año\ x} - \sum N°\ Proyectos_{Año\ x-1}}{\sum N°\ Proyectos_{Año\ x-1}}$
Observaciones	Si el valor es positivo significa que aumento de uno año a otro, pero si es negativo implica una disminución en el indicador.

Figura 2 – Indicador: Tasa anual de crecimiento en proyectos de investigación

Se puede observar en ambas figuras, la detallada especificación del indicador que incluye, no sólo su descripción y objetivo, sino que además su periodicidad, las dimensiones involucradas, las fórmulas sobre la cual se calculan los KPI. Toda esta información es acordada y validad por los usuarios estratégicos, y la plantilla sirve como interfaz de comunicación en todo el proceso de educción de requerimientos.

Indicador 5	Nº Jornada Completa Equivalente
Definición u objetivo	Mide la equivalencia en jornada completa de las horas realizadas por el total de académicos contratados.
Periodicidad	Anual, Semestral, Mensual
Fuente	Unidad de Programación y Registro Académico, RR.HH
Destinatario	• Entidades Externas: SIES, CNA, CRUCH, Agencias Acreditadoras, CNED, proyectos MECESUP. • Entidades Internas: Rectoría, Vicerrectorías, DFT, Direcciones, Facultades, Carreras.
Dimensiones	Sede Año de proceso – Semestre - Mes Grado Académico Nacionalidad Sexo Unidad Académica Carreras Sede Nivel Formación Tipo Contrato Tipo Jornada
Fórmula	$$JCE = \frac{\sum TotalJC + TotalJM + TotalJH}{44}$$ Donde: $$TotalJC = (\sum_{i}^{n} DocentesJC_i) * 44$$ $$TotalJM = (\sum_{i}^{n} DocentesJM_i) * 22$$ $$TotalJH = \sum Cant.Horas_i * NumDocentesJH_i$$ JC=Jornada Completa JM=Jornada Media JH=Jornada Hora
Observaciones	Para el caso del TotalJH el índice i representa la cantidad de horas a multiplicar por el número de docentes que realizan dicha cantidad de horas.

Figura 3 – Indicador: Número de jornadas completas equivalentes

4.3. Etapa Análisis: Sistemas Operacionales.

Una vez definidos los indicadores, se prosigue con la identificación de las fuentes de datos que son necesarias para su cálculo. Para esta actividad, es necesario reunirse con la Unidad de Informática y Comunicaciones (UNICO) de la institución, en conjunto con la UAI, con el objeto de identificar los diferentes modelos de datos involucrados y seleccionar las fuentes de datos. Luego de varias revisiones y análisis a los sistemas de información que maneja la UNAP, se filtran los que se van a utilizar, dando como resultado los que se listan en la Tabla 1.

Sistema	Descripción de la fuente de datos
ICON	Sistema contable y financiero.
SIPER	Sistema de Personal.
SICDO	Sistema Curricular y Docente.
GEDO	Sistema de Gestión de Documentos Online.
VRIIP	Sistema que almacena proyectos de investigación.

Tabla 1 – Selección de fuentes de datos a utilizar

Se obtiene de esta actividad; el listado de los sistemas que contienen los datos a utilizar, el meta-datos de estos, su modelo dato-lógico, y se seleccionan las tablas y atributos específicos que se requieren para el cálculo de los KPI.

4.4. Etapa Diseño: MM Conceptual.

Debido a que la UNAP cuenta con algunos DM previamente desarrollados, es necesario conocer las dimensiones que se encuentran presentes en estos, con el objeto de reutilizar la información en la elaboración de los nuevos DM, y así evitar la redundancia de datos junto con la duplicidad de dimensiones a nivel institucional. Luego de analizar los requerimientos y modelos existentes, se elaboran dos MM, para cumplir con los indicadores a obtener.

Uno de estos modelos enfocado a los indicadores de docencia, y el otro de investigación. El primer modelo es del tipo esquema "Copo de Nieve" (Snowflake), donde se hace un trabajo de jerarquización de algunas dimensiones, y reutilización de otras. Por otro lado, el DM de investigación es un esquema estrella, y la gran mayoría de sus dimensiones son creadas en este trabajo. Las dimensiones utilizadas en este PDW, se puede ver en la Tabla 2.

En cuanto a los indicadores de ambos MM, se presentan en las Figuras 4 a) y b), en sus respectivas tablas hechos con los KPI especificados para el esquema copo de nieve de docencia, y estrella de investigación.

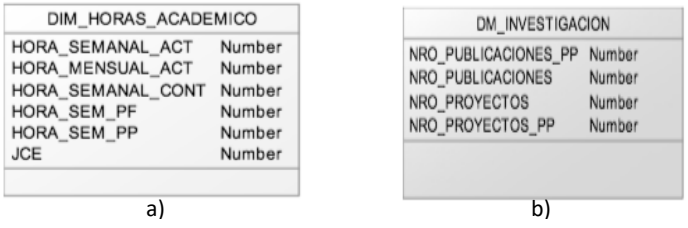

Figura 4 – Tablas de hechos: a) docencia, b) investigación

Dimensión	Descripción
Tiempo	Meses, semestres y años.
Académico	Nombre, fecha nac., fecha ingreso, sexo y nacionalidad.
Jerarquía	identifica la jerarquía docente de cada académico, estas son: “Profesor Asociado”, “Profesor Titular”, “Profesor Asistente”, “Instructor”, “Sin Jerarquía”.
Forma Contractual	Tipos de contratos: Jornada Completa, Media Jornada o Horas.
Grado Académico	Doctorado, Magister, Profesional, Licenciado, Técnico y Sin título.
Tipo de Actividad	“Docencia” o “Investigación”.
Carrera	Actividad del docente asociado a una carrera, cuenta con 3 jerarquías distintas: unidad académica, sede, y nivel de formación.
Tipo de Formación	Tipo de asignatura: Formación General o Formación Profesional.
Fuente de Financiamiento	Internos (UNAP), o externos (públicos o privados).
Área del Conocimiento	Clasificación de un programa académico: Agropecuaria y Ciencias del Mar, Administración y Comercio, Arte y Arquitectura, Ciencias Naturales y Matemáticas, Ciencias Sociales, Derecho, Humanidades, Educación, Tecnología y Salud.
Región	Nombre de regiones
Tipo de Proyecto	El tipo de proyecto categoriza en investigación o prestación de servicios a las postulaciones que realiza un docente a un proyecto en particular.
Estado del Proyecto	Estado del proyecto: adjudicado, en postulación, decretado, o finalizado.
Formato de Publicación	Artículo científico, capítulo de libro, revisión de articulo (review).
Tipo de Indexación	ISI, SCOPUS y SCIELO.
Posición de la Publicación	Que tiene la institución en un artículo científico.

Tabla 2 – Dimensiones a utilizar en el MM

En la Tabla 3 se describen cada uno de los indicadores considerados en el diseño conceptual del MM, para estas dos tablas de hechos.

Indicador	Descripción (hecho docencia)
HORA_SEMANAL_ACT	Cantidad de horas semanales que imparte un docente.
HORA_MENSUAL_ACT	Total de horas mensuales de un docente en una actividad.
HORA_SEM_PF	Cantidad de horas semanales que un docente realiza.
HORA_SEM_PP	Total de horas semanales que imparte un docente.
JCE	Cantidad de Jornada Completas Equivalentes.
Indicador	**Descripción (hecho Investigación)**
NRO_PUBLICACIONES_PP	Cantidad de publicaciones proporcionales.
NRO_PUBLICACIONES	Cantidad total de publicaciones que realiza el docente.
NRO_PROYECTOS	Cantidad total de proyectos en el que participa un docente.
NRO_PROYECTOS_PP	Número de proyectos proporcionales que realiza el docente.

Tabla 3 – Descripción de indicadores

Una vez definidas las componentes (hechos y dimensiones), se diagraman los diseños conceptuales de ambos MM en una herramienta CASE, conectándolos a través de las dimensiones comunes, y el nivel de las jerarquías que sea necesario. En este trabajo se utiliza la herramienta CASE PowerDesigner.

4.5. Etapa Diseño: Modelo Relacional del MM.

Una vez terminado el diseño conceptual, se genera el modelo relacional a través de la herramienta CASE, generando el código script en lenguaje de definición de datos (o DDL: Data Definition Language), indicando el motor de base de datos a utilizar en la implementación. Se presenta como anexo al final del artículo, el MM en la Figura 12, esto debido al tamaño de la imagen.

4.6. Etapa Proceso ETL+V: Extracción.

Para este proceso, se utiliza la herramienta Pentaho's Data Integration (PDI), que permite modelar el proceso ETL con diagramas de flujos de transformaciones, que consisten en un conjunto de pasos fijos encapsulados en flujos integrados denominados trabajos (job), que pueden contener sentencias o script de ejecución.

En la actividad de extracción, se utiliza el meta-datos, donde está registrada la información de las fuentes de datos seleccionadas, en conjunto con las especificaciones de los KPI, y se proceden a extraer los datos y se van dejando de manera agregada, en caso de ser necesario, directamente en el repositorio temporal. Se genera un flujo de trabajo que realiza las transformaciones necesarias para esta parte del proceso, y a modo de ejemplo se presenta en la Figura 5 un job que extrae información de los académicos.

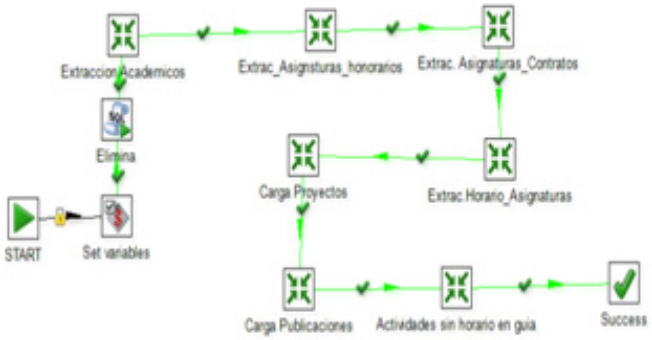

Figura 5 – Flujo de trabajo para la extracción de datos seleccionados

4.7. Etapa Proceso ETL+V: Transformación.

Al diseñar dos DM, se opta por elaborar un flujo de trabajos para cada uno, lo cual permite una depuración más sencilla. En la Figura 6 a) se presenta el flujo de trabajo para el DM que calcula los indicadores de docencia. En la Figura 6 b), se presenta el flujo de trabajo para calcular los indicadores asociados a la actividad de investigación. Cabe señalar, que sólo se presentan los flujos de trabajos resumidos para ambos DM, y

el detalle de cada transformación se deja fuera por limitación de números de páginas para el artículo

a) b)

Figura 6 – Flujo de trabajo para el cálculo de indicadores de: a) docencia b) investigación

4.8.Etapa Proceso ETL+V: Carga.

La carga de las tablas hechos es la parte más compleja de esta etapa, puesto que se debe realizar la comprobación de los datos cargados o cambio de algún nodo de la transformación. Antes de la carga, se deben verificar si los cálculos obtenidos son válidos en relación a los datos en los sistemas operacionales. Para el DM de docencia, la transformación que presenta la Figura 7, calcula las métricas a nivel mensual, y luego las carga a la tabla hechos de docencia.

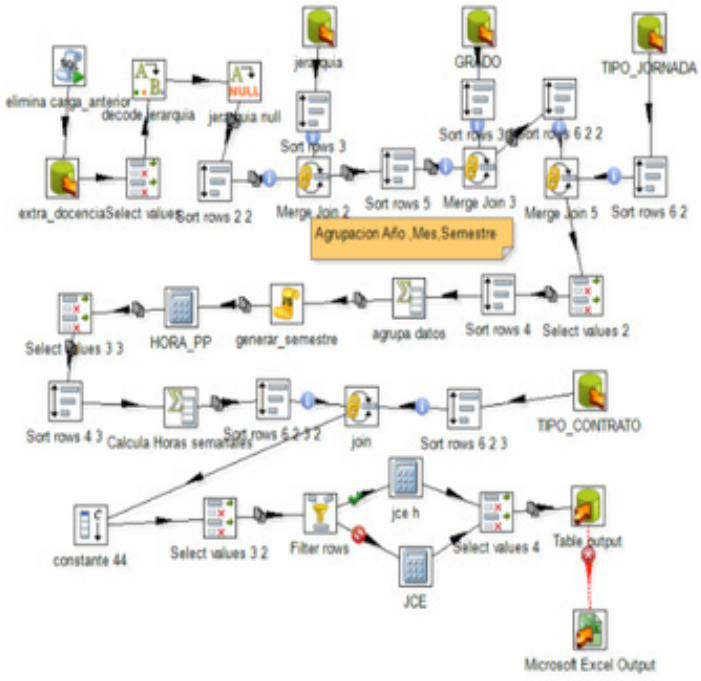

Figura 7 – Flujo de trabajo para la carga de tabla hechos de docencia

Primero se realiza una limpieza de la carga anterior de los datos si es que es necesario, para que en el nodo “extra_docencia” de la Figura 7, ejecute la consulta que extrae los datos de las asignaturas impartidas a una carrera y la cantidad de horas involucradas. En los pasos siguientes se hacen cruces de información para extraer los identificadores únicos de las dimensiones que correspondan.

El nodo agrupa las dimensiones para realizar el cálculo total de horas semanales de la actividad a una carrera en particular. Luego en el siguiente nodo, se suman todas las horas semanales calculadas con el fin de evitar duplicación de información, y no romper las reglas de integridad al momento de cargar los datos. Posterior a este paso, se genera un filtro para separar los docentes según su tipo de jornada, con el fin de poder calcular cuál es su aporte a las JCE de su actividad. Finalmente, se procede a cargar la información a la tabla de hechos de docencia. Algo similar se realiza para el DM de investigación.

4.9.Etapa Proceso ETL+V: Validación de KPI.

Finalmente, se validan que los indicadores obtenidos sean iguales o semejantes a los que se puedan obtener desde los sistemas operacionales. Para esto, se utiliza como método de prueba, calcular cada indicador mediante una consulta SQL a la base de datos desde los sistemas transaccionales, comparándolo con la consulta que se genera con las tablas del MM. En la Figura 8, se muestra la validación de un indicador de un total de 14. En particular, para este ejemplo se puede corroborar que el KPI está bien calculado, ya que su valor corresponde a los datos de origen.

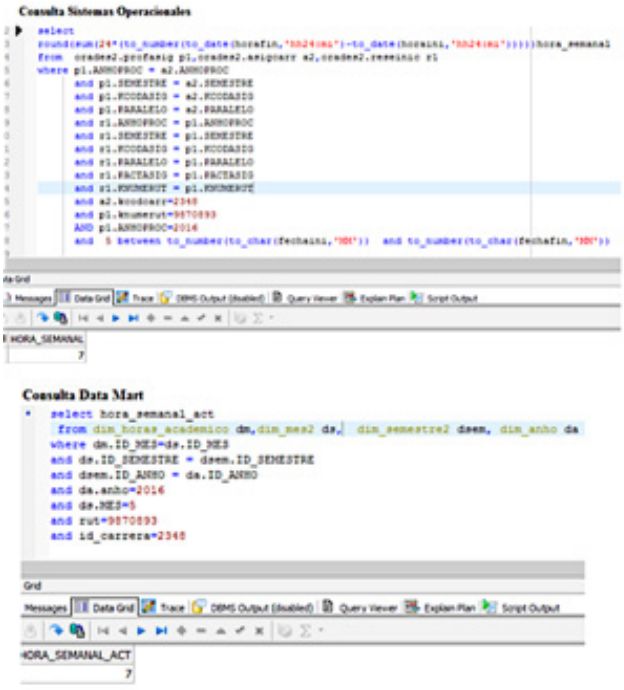

Figura 8 – Código SQL para validación de un KPI

4.10. Etapa Procesamiento Analítico: OLAP.

Se presentan los resultados del DM mediante la explotación y visualización de los indicadores a través de la herramienta OLAP Qlikview. De esta forma se da respuesta a los requerimientos definidos y planteados en un comienzo del proyecto. Estos resultados son expuestos con gráficas, las que permiten visualizar la información y poder realizar el análisis que corresponda. A modo de muestra, se presenta en la Figura 9, una gráfica de un indicador de docencia. En total son 14 gráficas, una por cada KPI.

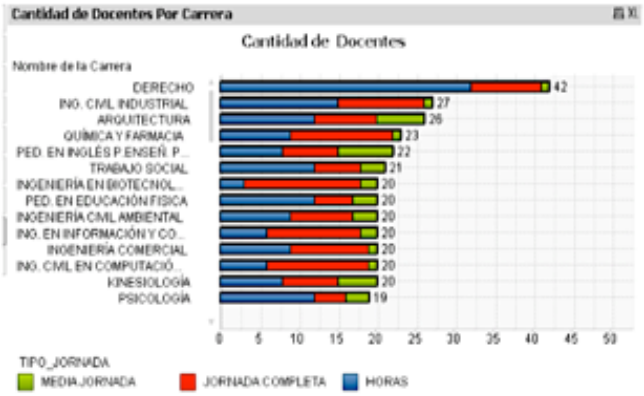

Figura 9 – Gráfica: cantidad de profesores por carrera

4.11.Etapa Procesamiento Analítico: Dashboard.

Una vez generadas las gráficas de los indicadores, se pueden integrar en un formato tipo tablero de control (dashboard), que permite a los usuarios finales, realizar un análisis multidimensional con distintos controles y mecanismos de interacción. El dashboard que muestra la Figura 10, presenta una consolidación de los indicadores relacionados a docencia en la cual, en los extremos tanto izquierdo como derecho, se encuentran las dimensiones manejadas por el modelo, y que pueden ser seleccionadas para cambiar los gráficos presentados al interior del panel. Los datos resaltados en color verde significan las selecciones actuales, y los de color gris que esos valores no están relacionados a la selección actual.

En la Figura 11, se muestran indicadores de proyectos de investigación en un dashboard, que se pueden visualizar por diferentes dimensiones. Los gráficos se encuentran agrupados en un elemento contenedor en la cual se muestra la pestaña sobre la dimensión en la que se muestra el indicador con algún gráfico asociado.

5. Conclusiones

Para el desarrollo de un PDW, toma gran relevancia la comunicación efectiva con las unidades con las que se realizan los trabajos en general. En particular, en el trabajo de educción fue clave la plantilla de especificación de requisitos de los KPI, ya que actúo como interfaz entre los usuarios estratégicos y el equipo del DW.

Figura 10 – Tablero de control de indicadores de docencia

Figura 11 – Tablero de control indicadores de proyectos de investigación

Durante el proceso de desarrollo de un DM, muchas veces se tiene que volver a un paso anterior para corregir o ajustar detalle del proceso en sí, es decir, a prueba y error lo que se ve reflejado mayormente en el proceso ETL+V.

Un factor importante es el análisis y comprensión, que se debe realizar sobre las estructuras de datos y modelos de los sistemas operacionales.

El desarrollo de un repositorio temporal, ayuda a dar repuesta eficiente a requerimientos operacionales durante la elaboración de un DM.

La validación de los KPI fue otro elemento clave en el proceso, ya que permitió generar confianza en el equipo, y las unidades de la institución sobre los resultados obtenidos a través del PDW.

Las visualizaciones gráficas integradas a través de los dashboard fueron muy bien valoradas por parte de los usuarios estratégicos.

Los resultados generaron un gran impacto en los usuarios, ya que en algunos casos los indicadores no se encontraban sobre la media de la institución, por el contrario, existían casos que los resultados presentados eran bastante buenos y aceptables, sin embargo, esta presentación de los resultados, implica que se tome cierta atención a ello, con el fin de apoyar la toma de decisiones desde la perspectiva estratégica y así´ poder mantener y mejorar los indicadores en general.

Finalmente, la metodología elaborada y utilizada en este PDW, fue muy valorada en general por la institución, así como en el equipo de trabajo.

En cuanto a trabajo futuro, si bien la metodología tiene definidas sus etapas y funciones en cada una de estas, así como tiene determinadas sus entradas y salidas en cada una de sus etapas, le falta establecer una formalización para que pueda responder a un proceso repetible. En esta línea, el trabajo a seguir es documentar detalladamente todos los aspectos y métodos de esta nueva metodología. Así como también aplicarlas en nuevos PDW, y realizar un análisis comparativo de los resultados con y sin la metodología propuesta.

Referencias

Rocha, Á. (2012). Framework for a Global Quality Evaluation of a Website. *Online Information Review, 36*(3), 374-382. doi:10.1108/14684521211241404

Antunes, A. A. (2004). Sistemas XYZ. In Sousa A. J. (Ed.), *Tecnologias Internet*. Lisboa: Editora Xxxpto.

Back, T. (2002). Adaptive business intelligence based on evolution strategies: some application examples of self-adaptive software. Information Sciences, 148(1-4), 131–121.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond Data Warehousing: What's Next in Business Intelligence. In: Proceedings of the *7th ACM international workshop on data warehousing and OLAP*, (pp. 1-6). Washington DC, USA: ACM.

Palocsay, S., Markham, I., & Markham, S. (2010). Utilizing and teaching data tools in Excel for exploratory analysis. *Journal of Business Research*, 63 (2), 191-206.

Kimball, R., & Ross, M. (2002). *The Data WarehouseToolkit - The Complete Guide to Dimensional Modeling* (2nd edition). Hoboken, NJ, USA: Jhon Wiley and Sons, Inc.

Bernabeu, R. (2010). HEFESTO: *Metodología para la Construcción de un Data Warehouse*. Córdob, Argentina.

Conesa, J., & Curto, J. (2010). *Introducción al Bussiness Intelligence* [en línea]. Barcelona: Editorial UOC, ISBN 978-84-9788-886-8.

Venter, M. (2005). Business Intelligence Initiatives: Failures Versus Success. *Revista Interdisciplinary Journal* [en línea], 4 (1).

Mazón, J., Trujillo, J., & Lechtenborger, J. (2007). Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms. *Data & Knowledge Engineering*, 63(3) 725–751.

Inmon, W. (2005). Building the Data Warehouse (3rd edition). Hoboken, NJ, USA: Jhon Wiley and Sons, Inc.

Vaisman, A., & Zimányi, E. (2014). Data Warehouse Systems Design and Implementation. *Springer Series: Data-Centric Systems and Applications*, (XXVI), 603.

Duran, A. (2000). *Un Entorno Metodológico de Ingeniería de Requisitos para Sistemas de Información*, Tesis doctoral Universidad de Sevilla.

Marcus, A. (2006). Dashboards in Your Future. *The Art of Prototyping*, 13 (1), 48–60.

Illescas, G., Sanchez, M., & Canziani, G. (2015). Métodos de Pronóstico por Indicadores dentro de la Gestión del Conocimiento Organizacional. *RISTI - Revista ibérica de Sistemas y Tecnologías de Información*, (E3), 29 – 41. DOI: 10.17013/risti.e3. 29–41.

Prieto, A., & Piattini, M. (2015). Propuesta de marco de mejora continua de gobierno TI en entidades financieras. *RISTI - Revista ibérica de Sistemas y Tecnologías de Información* (15), 51 – 67. DOI: 10.17013/risti.15.51-67.

Anexo

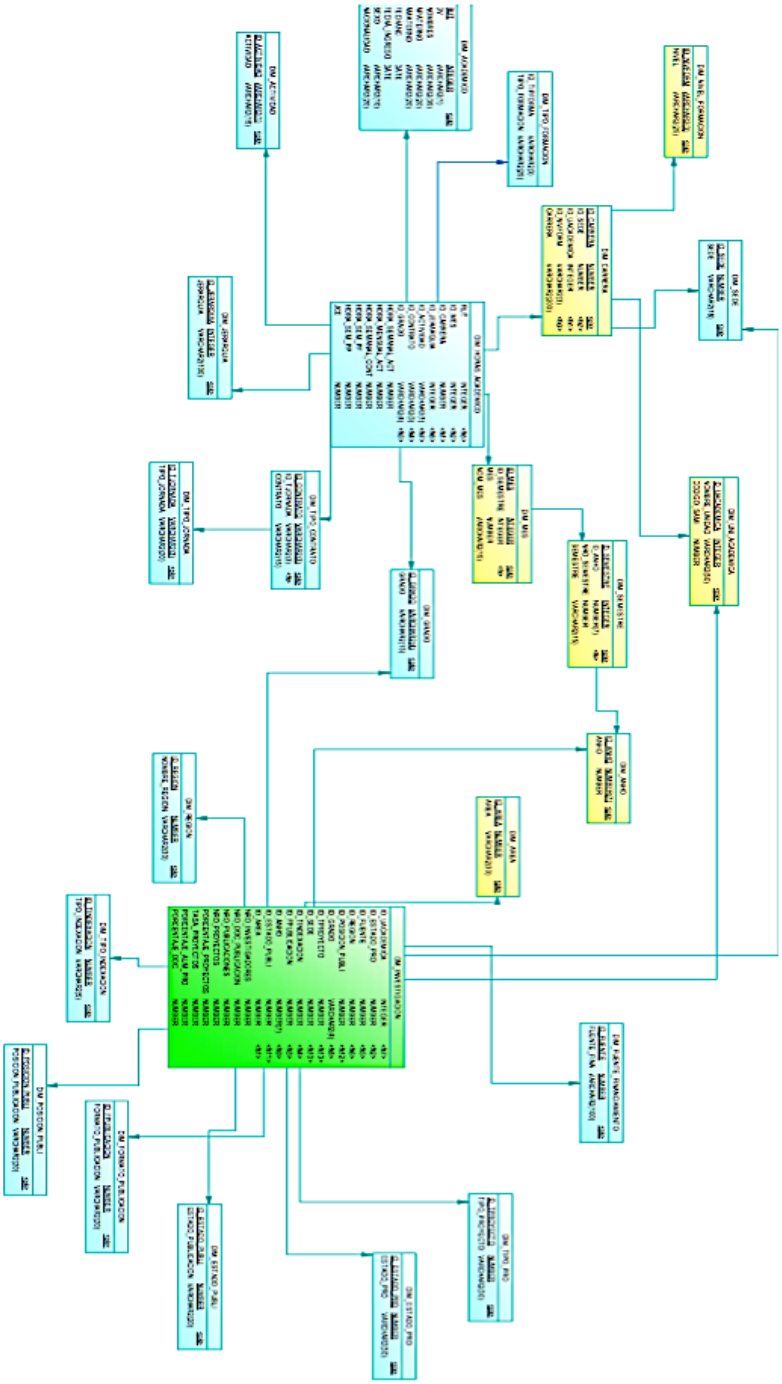

Figure 12 – MM relacional de los DM