UNIVERSIDAD DE ATACAMA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERIA INFORMATICA Y CIENCIAS DE LA COMPUTACIÓN

DISEÑO E IMPLEMENTACIÓN DE UNA BASE DE DATOS MULTIDIMENSIONAL PARA EL ANÁLISIS DE DATOS EDUCACIONALES.

PROFESOR GUÍA:

CAROLINA ZAMBRANO

GABRIEL FRANCISCO POBLETE CUADRA

2011

Agradecimientos.

En primer lugar, quiero agradecer a mi familia que me han apoyado a lo largos de estos 6 años de carrera y quienes siempre han tenido una palabra de aliento para que yo siguiera adelante. También quiero agradecer a amigos y cercanos quienes siempre me apoyaron y me dieron ánimos en los momentos más difíciles.

Sin embargo, quiero hacer una mención especial a una persona quien siempre creyó en mí. Esta persona es mi hermano Sebastián que me dijo que yo debía estudiar lo que a mí me gustaba y que sabía que no lo defraudaría. Sin ese acto de fe hace casi 6 años atrás yo no estaría aquí hoy escribiendo estos agradecimientos y es gracias a ti que este trabajo de título existe.

A todos muchas gracias.

Resumen.

En este trabajo de título se muestra el diseño e implementación de una Base de Datos Multidimensional para una muestra de datos Educacionales de PISA, un estudio realizado por OECD, con el objetivo de aportar en el análisis para la toma de decisiones educacionales.

Los resultados de la implementación muestran que Chile es el país con los mejores puntajes promedios de Latinoamérica, y que junto con Brasil logran las mejoras mas significativas entre los años 2000 y 2009. También se muestra una relación directa entre el nivel socioeconómico de los alumnos y su puntaje alcanzado, esto es a medida que aumenta el nivel socioeconómico del alumno también lo hace su puntaje promedio. Una tendencia que existe en todos los países de Latinoamérica.

Finalmente para el diseño conceptual de la base de datos multidimensional se ha utilizado el modelo CMDM de Fernando Carpani y para el diseño lógico un esquema estrella. La implementación de la base de datos multidimensional se ha desarrollado en la herramienta SQL SERVER 2008 R2.

Índice de Contenidos.

Índice de Figuras.

Figura Página

Índice de Tablas.

Tabla Página

Índice de Gráficos.

Gráfico Página

Capítulo 1 INTRODUCCIÓN.

En el primer capitulo de este trabajo de titulación se establecen los objetivos generales y específicos del mismo. También se plantea una justificación del trabajo que permite entender el trasfondo y la importancia de este trabajo de titulación.

1.1 Objetivos.

A continuación se establecen los objetivos generales y específicos del trabajo de titulación, que enmarcaran el alcance del trabajo:

Objetivos Generales:

- Diseñar e implementar una Base de Datos Multidimensional para el análisis de datos Educacionales.

Objetivos Específicos:

- Analizar los datos Educacionales de PISA para definir los indicadores que se diseñaran e implementaran en la base de datos multidimensional como medidas.
- Diseñar conceptualmente el esquema multidimensional con las medidas que permitirán hacer el análisis.
- Diseñar el esquema lógico multidimensional para la implementación en la herramienta SQL SERVER.
- Implementar la Base de Datos Multidimensional según el diseño realizado en las etapas anteriores y hacer consultas.
- Generar un conjunto de reportes.

1.2 Justificación.

Hoy en día las organizaciones necesitan de información para la toma de decisiones. Cada vez los sistemas de información operacionales generan más datos que se deben tratar con alguna tecnología. Es por ello que como respuesta a esta necesidad surge un área llamada inteligencia de negocios, que reúne el conjunto de metodologías, aplicaciones y tecnologías que permiten entre otras reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada, para su explotación directa o para su análisis. Cabe destacar que el proceso de inteligencia de negocios completo, no es propio de un ingeniero informático, pues su perfil es estudiar el negocio para aplicar tecnologías que apoyen la estrategia del negocio pero no es experto en todas las áreas de negocio. Por ello es importante considerar dentro del proceso de inteligencia de negocios un experto en el negocio. Dentro del proceso de inteligencia de negocios encontramos un área que es la de Data Warehouse y tecnologías OLAP.

Por otro lado, en el contexto de bases de datos podemos indicar que estas son el núcleo principal de un sistema de información, por lo cual, dependiendo del tipo de sistema de información la base de datos también cambia respecto a los tipos de datos que necesitará estructurar, y también al modelo de datos que soporte el esquema de base de datos específico al área de desarrollo. En este contexto para un sistema de apoyo al análisis se cuenta con bases de datos multidimensionales que son la base de los sistemas de Data Warehousing y la tecnología OLAP.

Un Data Warehouse tiene como componente una base de datos diseñada específicamente para favorecer el análisis de los datos, orientada a un determinado ámbito (empresa, organización, etc.), que ayuda a la toma de decisiones en la organización en la que se utiliza. Un Data Mart es un subconjunto de un Data Warehouse usualmente orientado a un área específica de la organización (ventas, operaciones, etc.). Una base de datos multidimensional está compuesta de dimensiones y medidas, donde las dimensiones son los focos de análisis que se estructuran en tablas dimensionales y las medidas son los

valores que se estructuran en tablas de hecho. Las Bases de Datos Multidimensionales son la base de estas herramientas OLAP debido a que las estructuras de estas se basan en tablas con un campo para las dimensiones y otro para las medidas.

A nivel organizacional son pocas las instituciones que emprenden proyectos de análisis de datos que tengan algún impacto en la comunidad. Es por ello que en este trabajo de título se diseñará e implementará una Base de Datos Multidimensional para una muestra de datos Educacionales de PISA, que es un estudio realizado por OECD dirigido a estudiantes de 15 años y que evalúa si los estudiantes tienen la capacidad de reproducir lo que han aprendido en la escuela para resolver problemas de la vida real, con el objetivo de aportar en el análisis para la toma de decisiones educacionales.

PISA, por sus siglas en inglés, significa P*rogramme for International Student Assessment*, traducido como el Programa para la Evaluación Internacional de los Estudiantes. Es un estudio comparativo de evaluación de los resultados de los sistemas educativos. El propósito de PISA es conocer el nivel de habilidades necesarias que han adquirido los estudiantes para participar plenamente en la sociedad, centrándose en dominios claves como Lectura, Ciencias y Matemáticas. Mide si los estudiantes tienen la capacidad de reproducir lo que han aprendido, de transferir sus conocimientos y aplicarlos en nuevos contextos académicos y no académicos, de identificar si son capaces de analizar, razonar y comunicar sus ideas efectivamente, y si tienen la capacidad de seguir aprendiendo durante toda la vida. Para PISA, esos dominios están definidos como competencia (literacy) científica, lectora o matemática. Bajo esta perspectiva de competencias, PISA se interesa en el repertorio de conocimientos y habilidades adquirido tanto en las escuelas como fuera de ellas.

El estudio se realiza cada tres años, y en cada ciclo se enfatiza uno de los tres dominios de evaluación y los otros son evaluados con menor profundidad. En el 2000 el principal dominio fue Lectura, en el 2003 Matemáticas, en el 2006 Ciencias y en el 2009 se regresa a Lectura, y así sucesivamente.

Entonces para el desarrollo de esta memoria de título se comenzará con el análisis de los datos para definir qué tipo de indicadores de gestión son factibles de diseñar e implementar en la base de datos multidimensional. Cabe destacar que estos indicadores deben ser resueltos desde los datos para no proponer indicadores que sean imposibles de obtener porque no están los datos disponibles. Además, estos indicadores se volverán a validar cuando se realice el diseño conceptual de la base de datos multidimensional. Cabe destacar que desde el punto de vista de ingeniería de datos también es válido definir los indicadores de gestión en la etapa de modelamiento conceptual.

Para la implementación se usará el lenguaje MDX y se obtendrán medidas que permitirán de forma objetiva analizar tendencias en el tiempo según las dimensiones de estudio. Por lo cual en esta memoria de título se estudiará el modelo multidimensional y las consultas al modelo multidimensional como proceso de diseño e implementación.

1.3 Organización del Documento.

En el capítulo 2 se presenta y define el concepto de Inteligencia de Negocios (Business Intelligence) así como también las tecnologías y metodologías que caben bajo dicho concepto, para luego introducir las nociones de Data Warehouse, Data Mart y Bases de Datos Multidimensionales. Luego se presenta una revisión sobre los modelos conceptuales y lógicos para bases multidimensionales. Por último el Capítulo 2 termina con una introducción a los lenguajes SQL y MDX.

En el capítulo 3 se presenta todo el diseño e implementación de la bases de datos multidimensionales, pasando por la construcción de los indicadores, del esquema conceptual y lógico, para finalizar con el desarrollo de las consultas.

Finalmente, en el capítulo 4 se establecen las conclusiones pertinentes así como también la posibilidad de todos aquellos trabajos futuros.

Capítulo 2 MARCO TEÓRICO.

2.1 Business Intelligence.

También conocida como Inteligencia Empresarial o por sus siglas en ingles BI (Business Intelligence, Inteligencia de Negocios), se define como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, alertas, etc.) o para su análisis y conversión en conocimiento, dando así soporte a las personas para la toma de decisiones sobre el negocio. Sin embargo, como se vera a continuación, diferentes autores y organizaciones, que han aportado al desarrollo de BI, han definido BI de diferentes formas y perspectivas:

- "Se refiere a las tecnologías, aplicaciones y prácticas para la recolección, integración, análisis y presentación de la información empresarial y a veces a la información en sí misma" (Luhn, 1958).[1]
- "Es el conjunto de conceptos y métodos para mejorar la toma de decisiones en los negocios, utilizando sistemas de apoyo basado en los hechos" (Dresner, 1989).[2]
- "El propósito de la inteligencia de negocios es convertir grandes volúmenes de datos en valor para los usuarios finales" (Oracle, 2011).[3]
- "Business Intelligence es una disciplina de desarrollo de la información que es concluyente, basado en hechos y acciones concretas. Business Intelligence ofrece a las compañías la capacidad de descubrir y utilizar la

[1] Luhn, H. (1958). A Business Intelligence System. IBM Research.
[2] Dresner, H. (1989). Concepto de BI.
[3] Oracle. (2011). Oracle. Obtenido de www.oracle.com.

información que ya posee, y convertirla en conocimiento que repercute directamente en el rendimiento corporativo" (IBM, 2011).[4]

- "BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un Data Warehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios. Las áreas incluyen clientes, proveedores, productos, servicios y competidores" (Gartner, 2011).[5]

- "Conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa" (Wikipedia, 2011).[6]

Como se puede apreciar la mayoría de las definiciones coinciden en que BI es el conjunto de metodologías, aplicaciones y/o tecnologías que permiten la transformación de los datos, información en nuevo conocimiento, sin embargo el enfoque cambia cuando los autores y organizaciones tratan de definir su uso o propósito.

2.1.1 Historia de BI.

A continuación se hace un recorrido sobre los distintos hitos históricos que nos han llevado hasta la situación actual de BI:

- **1958**: H.P. Luhn escribió un artículo llamado "A Business Intelligence System" donde describe las características que debe tener un sistema de este tipo. Sin embargo la mayor diferencia entre lo que escribió Luhn y lo que se entiende actualmente por BI está en que Luhn no pensaba en

[4] IBM. (2011). IBM. Obtenido de www.ibm.com.
[5] Gartner, G. (2011). Gartner Group. Obtenido de www.gartner.com.
[6] Wikipedia. (2011). Wikipedia. Obtenido de www.wikipedia.org

información estructurada en bases de datos, sino que se refiere siempre a documentos textuales en un sentido más genérico.

- **1969:** Creación del concepto de base de datos (Codd).
- **1970's:** Desarrollo de las primeras bases de datos y las primeras aplicaciones empresariales (SAP, JD Edwards, Siebel, PeopleSoft). Estas aplicaciones permitieron realizar "data entry" en los sistemas, aumentando la información disponible, pero no fueron capaces de ofrecer un acceso rápido y fácil a dicha información.
- **1980s:** Creación del concepto Datawarehouse (Ralph Kimball, Bill Inmon), y aparición de los primeros sistemas de reporting. A pesar de todo, seguía siendo complicado y funcionalmente pobre. Existían relativamente potentes sistemas de bases de datos pero no había aplicaciones que facilitasen su explotación.
- **1989:** Se reintroduce y populariza el término BI (Howard Dresner).
- **1990s:** BI 1.0. Proliferación de múltiples aplicaciones BI. Estos proveedores resultaban caros, pero facilitaron el acceso a la información, y en cierto modo agravaron el problema que pretendían resolver.
- **2000s:** BI 2.0. Consolidación de las aplicaciones BI en unas pocas plataformas Business Intelligence (Oracle, SAP, IBM, Microsoft). A parte de la información estructurada, se empieza a considerar otro tipo de información y documentos no estructurados.

2.2 Herramientas de BI.

Las herramientas de software de BI son usadas para acceder a los datos de los negocios y proporcionar reportes, análisis, visualizaciones y alertas a los usuarios. La gran mayoría de las herramientas de BI son usadas por usuarios finales para acceder, analizar y reportar contra los datos que más frecuentemente

residen en Data Warehouse, Data Marts y almacenes de datos operacionales. Los desarrolladores de aplicaciones usan plataformas de BI para desarrollar y desplegar aplicaciones (las cuales no son consideradas herramientas de BI).

Actualmente el mercado de herramientas de BI se encuentra constituido de dos subsegmentos: EBIS (Enterprise Business Intelligence Suite, Suites de Inteligencia de Negocios Empresal) y plataformas de BI. La mayoría de las herramientas de BI, son BI empresarial.

2.2.1 EBIS vs Plataformas.

Tiedrich (Tiedrich, 2003)[7], menciona que las plataformas de BI son ambientes de desarrollo de aplicaciones, comúnmente ofrecen un lenguaje de codificación como Visual Basic y otros lenguajes para la creación de aplicaciones personalizadas.

Las plataformas de BI se usan cuando hay una necesidad de analizar grandes cantidades de información con muchos cálculos (por ejemplo, rentabilidad de un producto) o para crear aplicaciones amigables para usuarios específicos.

En cambio las herramientas de BI empresarial, contienen una funcionalidad estándar. Una vez que una o más fuente de datos son mapeadas por las herramientas de BI empresarial, la funcionalidad recién comienza. A pesar de que algunas herramientas contienen algunas facilidades de codificación, crear aplicaciones a la medida resulta demasiado complejo.

Según lo dicho por Tiedrich (Tiedrich, 2003)[8], consultor de Gartner, las EBIS contienen ventajas y desventajas que se muestran en la tabla 2.2:

[7] Tiedrich, A. H. (2003). Business Intelligence Tools: Perspective. Gartner Research.
[8] Tiedrich, A. H. (2003). Business Intelligence Tools: Perspective. Gartner Research.

Tabla 2.1 - Ventajas y desventajas de los EBIS [Fuente: Gartner Dataquest, 2003]

Ventajas	Desventajas
Implementación sencilla.	Funcionalidad menos analítica.
Funcionalidad estándar.	Poca facilidad de personalización.

Los EBIS son usualmente utilizados cuando hay muchos usuarios de diversos niveles de habilidad técnica, cada uno con requerimientos de reportes y vistas que son menos analíticos (por ejemplo, reportes administrativos o análisis de variantes simples).

2.2.2 Clasificación de las Herramientas BI.

De acuerdo a su nivel de complejidad se pueden clasificar las herramientas de BI en (Microstrategy, 2004)[9]:

- Reporte empresarial: Los reportes escritos son usados para generar reportes estáticos altamente formateados destinados para ampliar su distribución con mucha gente.
- Cubos de análisis: Los cubos basados en herramientas de BI son usados para proveer capacidades analíticas a los administradores de negocios.
- Vistas Ad Hoc Query y análisis: Herramientas OLAP (On-Line Analytical Processing, Proceso de Análisis en Línea) relacionales son usadas para permitir a los expertos visualizar la base de datos y ver cualquier respuesta y convertirla en información transaccional de bajo nivel.
- Datamining: Son herramientas usadas para desempeñar modelado predictivo o para descubrir la relación causa efecto entre dos métricas.

[9] Microstrategy. (2004). Information Week. Obtenido de http://www.informationweek.com/whitepaper/.

- Entrega de reportes y alertas: Los motores de distribución de reportes son usados para enviar reportes completos o avisos a un gran número de usuarios, dichos reportes se basan en suscripciones, calendarios, etc.

La tabla 2.3 muestra las tecnologías que son usadas para BI las cuales entran dentro de los cinco niveles mencionados anteriormente (Lokken, 2001)[10]:

Tabla 2.2 - Tecnologías de BI [Fuente: Lokken, B., Tool Box, 2001]

Tecnologías de BI
Data Warehouses
Data Mart
Data Mining
Scoreboards
Dashboards
Transformación de datos y herramientas de limpieza
Herramientas de reportes y vistas
Herramientas de análisis y exploración
Herramientas de visualización de datos
Herramientas de predicción y modelamiento
Sistemas de alertas y notificaciones

2.2.3 El proceso de BI.

La figura 2.1 muestra el proceso de Inteligencia de Negocios y sus componentes. Los componentes son:

[10] Lokken, B. (2001). Tool Box. Obtenido de http://businessintelligence.ittoolbox.com/.

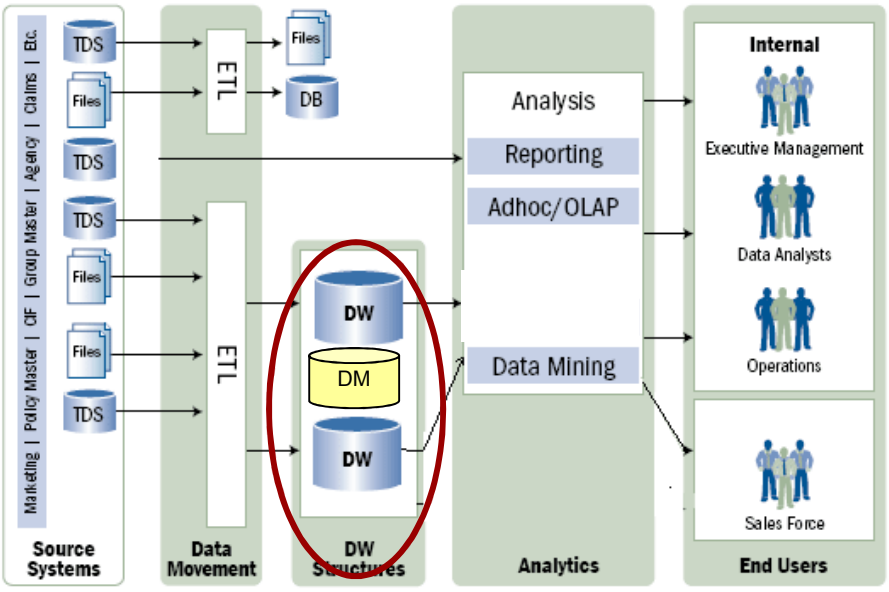

Figura 2.1 - El proceso de Business Intelligence [Fuente: Carolina Zambrano M, Introducción a Data Warehouse, 2010]

Fuentes de información.

Las fuentes de información a las que podemos acceder son:

- Básicamente, de los sistemas operacionales o transaccionales, que incluyen aplicaciones desarrolladas a medida, ERP (Enterprise Resource Planning, Sistemas de Planificación de Recursos), etc.
- Sistemas de información departamentales: previsiones, presupuestos, hojas de cálculo, etc.
- Fuentes de información externa, en algunos casos comprada a terceros, como por ejemplo estudios de mercado. Las fuentes de información externas son fundamentales para enriquecer la información que tenemos de nuestros clientes. En algunos casos es interesante incorporar información referente, por ejemplo, a población, número de habitantes, etc.

La información que cargamos normalmente es estructurada, es decir, aquella que se puede almacenar en tablas: en la mayoría de los casos es

información numérica. Cada vez más, la tecnología nos permite trabajar con información no estructurada, y se espera que este tipo de información sea cada vez más importante.

En relación a ello, una encuesta ha indicado que el 60% de los directores de Sistemas de Información y de los de Tecnología considera que la información semi-estructurada es crítica para mejorar las operaciones y para la creación de nuevas oportunidades de negocio (Blumberg, 2003)[11].

Proceso extracción, transformación y carga.

Proceso que permite mover datos desde múltiples fuentes, reformatear, limpiar, y cargar en otra base de datos, Data Mart o Data Warehouse, para utilizar en su análisis, o en otro sistema operacional con el fin de apoyar un proceso de negocio.

Esta parte del proceso de construcción es costosa y consume una parte significativa de todo el proceso, por ello requiere recursos, estrategia, habilidades especializadas y tecnologías. El proceso ETL (Extraction Transformation and Load, Extracción Transformación y Carga) se divide en 5 subprocesos:

- Extracción: Este proceso recupera los datos físicamente de las distintas fuentes de información. En este momento disponemos de los datos en bruto.
- Limpieza: Este proceso recupera los datos en bruto y comprueba su calidad, elimina los duplicados y, cuando es posible, corrige los valores erróneos y completa los valores vacíos, es decir, se transforman los datos -siempre que sea posible- para reducir los errores de carga. En este momento disponemos de datos limpios y de alta calidad.

[11] Blumberg, R. (Septiembre de 2003). Information Management. Obtenido de http://www.information-management.com

- Transformación: Este proceso recupera los datos limpios y de alta calidad y los estructura y sumariza en los distintos modelos de análisis. El resultado de este proceso es a obtención de datos limpios, consistentes, sumarizados y útiles.

- Integración: Este proceso valida que los datos que cargamos en el Data Warehouse son consistentes con las definiciones y formatos del Data Warehouse; los integra en los distintos modelos de las distintas áreas de negocio que hemos definido en el mismo. Estos procesos pueden ser complejos.

- Actualización: Este proceso es el que nos permite añadir los nuevos datos al Data Warehouse.

En la figura 2.2 se muestran los principales problemas que se pueden encontrar al acceder a los datos para extraerlos: básicamente se refieren a que provienen de distintas fuentes, bases de datos, plataformas tecnológicas, protocolos de comunicaciones, juegos de caracteres, y tipos de datos.

Figura 2.2 - Problemas del proceso ETL [Fuente: Josep Lluís Cano, Business Intelligence: Competir con Información, 2007]

OLAP

Existen distintas tecnologías que nos permiten analizar la información que reside en un Data Warehouse, pero la más extendida es el OLAP.

Los usuarios necesitan analizar información a distintos niveles de agregación y sobre múltiples dimensiones: Por ejemplo, ventas de productos por zona de ventas, por tiempo, por clientes o tipo de cliente y por región geográfica. Los usuarios pueden hacer este análisis al máximo nivel de agregación o al máximo nivel de detalle. OLAP provee de estas funcionalidades y algunas más, con la flexibilidad necesaria para descubrir las relaciones y las tendencias que otras herramientas menos flexibles no pueden aportar. A estos tipos de análisis se les llama multidimensionales, porque facilitan el análisis de un hecho desde distintas perspectivas o dimensiones.

Esta es la forma natural que se aplica para analizar la información por parte de los tomadores de decisiones, ya que los modelos de negocio normalmente son multidimensionales. La visualización de la información es independiente respecto de cómo se haya almacenado.

El OLAP Council sumarizó las 12 reglas de Codd en lo que ellos llamaban el concepto FASMI (Fast, Analysis, Shared and Multidimensional, Rapido, Analisis, Compartido y Multidimensional)que los productos OLAP deben cumplir (Council, 2005)[12].

El concepto FASMI proviene de las siglas de las iniciales en inglés:

- FAST (Rápido): Debe ser rápido, se necesita lanzar consultas y ver los resultados inmediatamente.
- ANALYSIS (Análisis): Debe soportar la lógica de negocio y análisis estadísticos que sean necesarios para los usuarios.
- SHARED (Compartido): Tiene que manejar múltiples actualizaciones de forma segura y rápida.
- MULTIDIMENSIONAL (Multidimensional): Tiene que proveer de una visión conceptual de la información a través de distintas dimensiones.

[12] Council, O. (2005). FASMI.

La representación gráfica del OLAP son los cubos y se muestra en la figura 2.3:

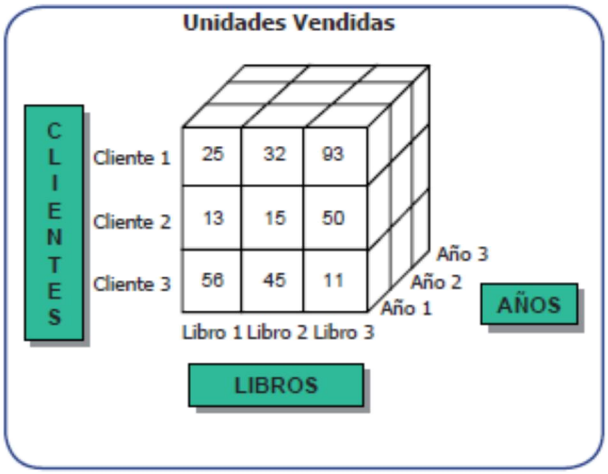

Figura 2.3 - Ejemplo de Cubo [Fuente: Josep Lluís Cano, Business Intelligence: Competir con Información, 2007]

En el cubo se tienen las unidades vendidas de cada uno de los libros, para los distintos clientes y en los distintos años. Este es el concepto de multidimensionalidad. Disponemos de las unidades vendidas de cada uno de los libros para cada uno de los clientes y en cada uno de los años: el contenido de un cubo individual son las ventas de un libro a un cliente en un año. Los contenidos de cada uno de los cubos individuales del cubo recogen lo que se llaman “hechos” (en el ejemplo las unidades vendidas). En la actualidad, las soluciones OLAP permiten que cada una de los cubos individuales pueda contener más de un hecho.

Data Mining

Es el proceso de analizar de manera "semi-automática" grandes bases de datos para buscar patrones útiles. Similar al descubrimiento de conocimiento en inteligencia artificial, la minería de datos encuentra reglas y patrones.

La tecnología Data Mining trata con volúmenes de datos almacenados principalmente en disco. Es semi-automático porque requiere de intervención

manual, un pre proceso (qué patrón buscar) y un post proceso (encontrar nuevos patrones novedosos).

2.3 Data Warehouse

Cuando se necesite la información esta requerirá en un mismo entorno para facilitar su análisis. Normalmente, en los sistemas transaccionales no tenemos preparada la información para ser analizada, sólo tenemos la información de las transacciones actuales, pero no la de los periodos anteriores o la de las previsiones. Si quiere estudiar la evolución de las ventas se necesitara saber:

- Ventas actuales.
- Ventas del/os periodo/s anterior/es.
- Presupuesto de ventas del ejercicio.

Si se sigue con la tendencia actual, incluso quizá sea preciso las ventas estimadas hasta el final del período. Inicialmente, se puede sentir la tentación de recuperar esa información, introducirla en una hoja de cálculo y a partir de ahí comenzar el análisis. Obviamente este no es el camino correcto.

La aparición de los Data Warehouse o almacenes de datos son la respuesta a las necesidades de los usuarios que necesitan información consistente, integrada, histórica y preparada para ser analizada para poder tomar decisiones.

Al recuperar la información de los distintos sistemas, tanto transaccionales como departamentales o externos, y almacenándolos en un entorno integrado de información diseñado por los usuarios, el Data Warehouse permite analizar la información contextualmente y relacionada dentro de la organización.

Hay muchas definiciones de Data Warehouse, entre las cuales se tiene:

- “Un datawarehouse es una colección de información creada para soportar las aplicaciones de toma de decisiones” (Watson, 2011).[13]

[13] Watson, H. J. (2011). Terry College of Business. Obtenido de www.terry.uga.edu/~hwatson/dw_tutorial.ppt.

- “El Data Warehouse es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión” (Inmon, 1992).[14]

- "Una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis" (Kimball, 1996).[15]

Por lo tanto se puede concluir, que un Data Warehouse, es el proceso de extraer y filtrar datos de las operaciones comunes de las empresas, procedentes de los distintos subsistemas operacionales, para transformarlos, integrarlos, sumariarlos y almacenarlos en un depósito o almacén de datos, para poder acceder a ellos cada vez que se necesite mediante mecanismos flexibles para el usuario.

2.3.1 Características de un Data Warehouse

Siguiendo con la definición de Inmon, un Data Warehouse se caracteriza por ser:

- **Integrado**: El aspecto más importante en los Data Warehouses es que la información encontrada siempre está integrada. La información debe ser transformada en medidas comunes, códigos comunes y formatos comunes para que pueda ser útil. La integración permite a las organizaciones implementar la estandarización de sus definiciones. La integración de los datos se muestra de muchas maneras: en conversiones de nombres consistentes, en la medida uniforme de las variables, en la conversión de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples, entre otros. Los datos almacenados en el Data Warehouse, deben integrarse en una estructura consistente, por lo que las

[14] Inmon, B. (1992). Building the datawarehouse (1ª edición).
[15] Kimball, R. (1996). The datawarehouse Toolkit.

inconsistencias existentes en los diversos sistemas operacionales, deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

Como se muestra en la figura 2.4 un Data Warehouse integra datos recogidos de diferentes sistemas operacionales de la organización y/o fuentes externas.

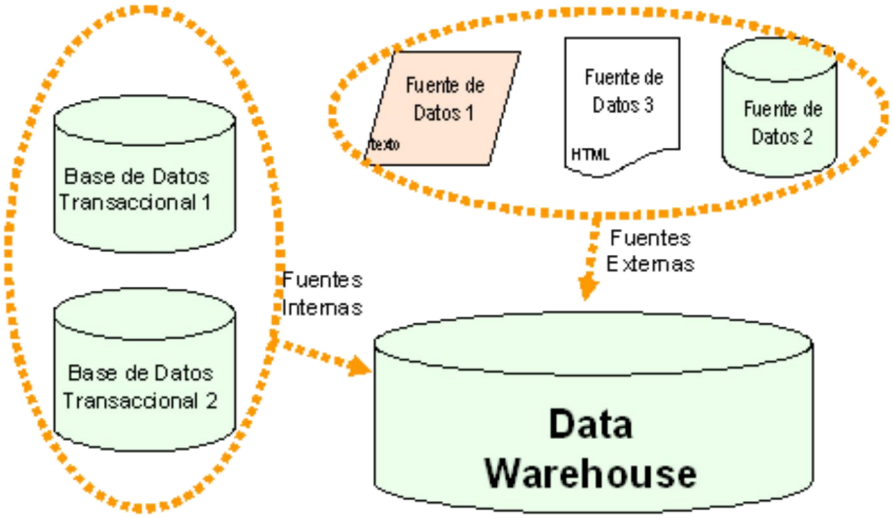

Figura 2.4 - Data Warehouse Integrado [Fuente: Jorge Castro, Luis Urrutia, Seminario de Titulo, 2004]

- **Temáticos:** Significa que cada parte del Data Warehouse está construida para resolver un problema de negocio, que ha sido definido por los encargados de tomar decisiones. Por ejemplo: Entender los hábitos de compra de nuestros clientes, analizar la calidad de nuestros productos, analizarla productividad de una línea de fabricación, etc. Para poder analizar un problema de negocio necesitamos información que proviene de distintos sistemas y la organizamos entorno a áreas: ventas, clientes, elementos de transporte, etc. Provee a los tomadores de decisiones de una visión completa y concisa sobre una problemática de negocio, obviando toda aquella información que no necesitan para la toma de decisiones. Por ejemplo, como se aprecia en la figura 2.5, todos los datos sobre

vendedores pueden ser consolidados en una única tabla del Data Warehouse. Así, las peticiones de información sobre vendedores serán más fáciles de responder dado a que toda la información está en el mismo lugar.

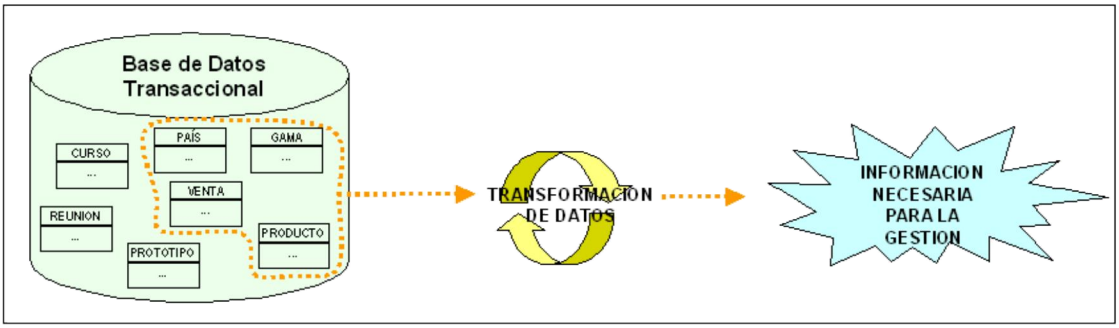

Figura 2.5 - Data Warehouse Temático [Fuente: Jorge Castro, Luis Urrutia, Seminario de Titulo, 2004]

- **No Volátil**: La información es útil solo cuando es estable. Los datos operacionales cambian sobre una base momento a momento. La perspectiva más grande, esencial para el análisis y la toma de decisiones, requiere una base de datos estable.

En la figura 2.6, se puede ver que la actualización (insertar, actualizar, borrar), que se hace en el ambiente operacional. Pero la manipulación básica de los datos que ocurre en el Data Warehouse son: la carga inicial de los datos y el acceso a los mismos. Por lo tanto, la información de un Data Warehouse existe para ser leída, y no modificada.

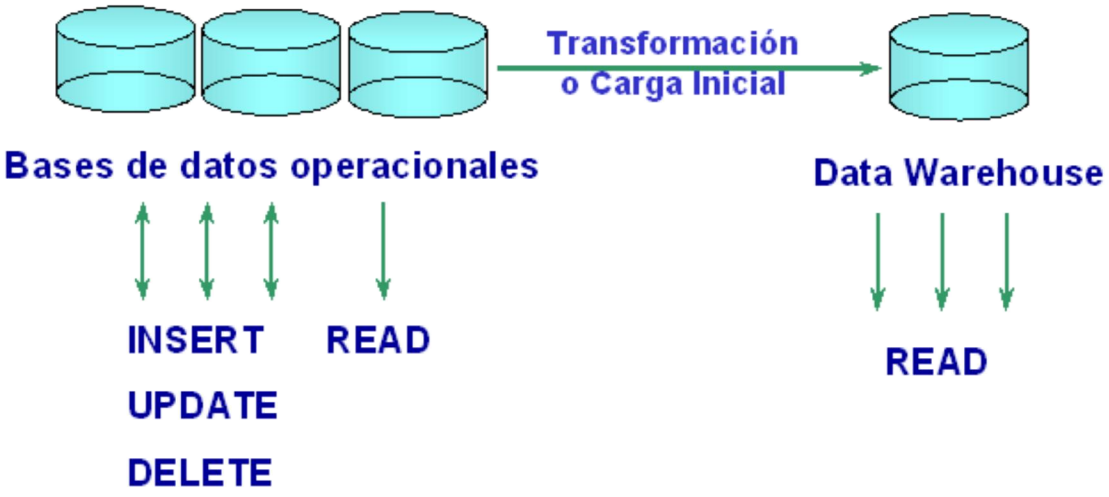

Figura 2.6 - Data Warehouse No Volátil [Fuente: Jorge Castro, Luis Urrutia, Seminario de Titulo, 2004]

- **Histórico:** Significa que se mantiene la información histórica y se almacena referida a determinadas unidades de tiempo, tales como horas, días, semanas, meses, trimestres o años. Ello nos permitirá analizar, por ejemplo, la evolución y la tendencia de las ventas en los periodos que queramos.

2.3.2 Procesos de un Data Warehouse

- **Extracción**: obtención de información de las distintas fuentes tanto internas como externas.
- **Elaboración**: filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- **Carga**: organización y actualización de los datos y los metadatos en la base de datos.
- **Explotación**: extracción y análisis de la información en los distintos niveles de agrupación.

Desde el punto de vista del usuario, el único proceso visible es la explotación del almacén de datos, aunque el éxito del Data Warehouse radica en

los tres procesos iníciales que alimentan la información del mismo y suponen el mayor porcentaje de esfuerzo (en torno a un 80%) a la hora de desarrollar el almacén.

2.3.3 Diferencias entre Sistema Tradicional y Data Warehouse

En cuanto a los requerimientos de diseño y características de implementación, los sistemas tradicionales OLTP (Online Transactional Processing, Proceso Transaccional en Linea) y los Data Warehouse poseen muchas diferencias, las que se pueden resumir en la Tabla 2. 4.

Tabla 2.3 - Diferencia entre un sistema a tradicional y un Data Warehouse [Fuente: SAS INSTITUTE EUROPEAN OFFICE, 2003]

SISTEMA TRADICIONAL	DATA WAREHOUSE
Predomina la actualización	Predomina la consulta
La actividad más importante es del tipo operativo (día a día)	La actividad mas importante es el análisis y la decisión estratégica
Predomina el proceso puntual	Predomina el proceso masivo
Mayor importancia a la estabilidad	Mayor importancia al dinamismo
Datos en general desagregados	Datos en distintos niveles de detalle y agregación
Importancia del dato actual	Importancia del dato histórico
Importancia del tiempo de respuesta de la transacción instantánea	Importancia de la respuesta masiva
Estructura relacional	Visión multidimensional
Usuarios de perfiles medios o bajos	Usuarios de perfiles altos
Explotación de la información relacionada con la operativa de cada aplicación	Explotación de toda la información interna y externa del negocio

2.3.4 Factores que inciden en la construcción de un Data Warehouse

Los factores que se deben tener en cuenta cuando estamos evaluando una alternativa tecnológica para la construcción de un Data Warehouse son (Strange, 2001)[16]:

- Tamaño del Data Warehouse: Es el volumen de datos que contiene el Data Warehouse.
- Complejidad de los esquemas de datos: Si el modelo de datos es complejo, puede dificultar la optimización y el rendimiento de las consultas.
- Número de usuarios concurrente: Éste es un factor determinante. Si distintos usuarios pueden lanzar consultas concurrentes (a la vez), el Data Warehouse debe gestionar sus recursos para poder dar respuesta a las distintas consultas.
- Complejidad de las consultas: Si las consultas necesitan acceder a un número elevado de tablas y los cálculos a realizar son complejos, podemos poner en dificultades al motor de la base de datos del Data Warehouse.

2.3.5 Ventajas y Desventajas del Data Warehouse

Los Data Warehouse convierten los datos operacionales de una organización en una herramienta competitiva, que permite a los usuarios finales (gerentes de alto nivel, o nivel medio) examinar los datos de modo estratégico, realizar análisis y detección de tendencias, seguimiento de medidas críticas, producir informes con mayor rapidez, un acceso más fácil, más flexible y más intuitivo a la información que se necesite en cada momento.

Frecuentemente, datos que son difíciles de interpretar, desde varias fuentes, se convierten en información lista para el usuario final. Estas ventajas competitivas para el negocio, se pueden traducir en:

[16] Strange, K. (2001). The Challenges of Implementing a datawarehouse to Achieve Business Agility.

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares.

Aun con los beneficios que se incorporan a la organización, junto con la implantación de un Data Warehouse, también se pueden presentar ciertos problemas, que según José Hernández Orallo[17], son:

- Infravaloración del esfuerzo necesario para su diseño y creación.
- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos.
- Incremento continuo de los requisitos de los usuarios.
- Privacidad de los datos.
- Alto coste.
- Complejo desarrollo.

[17] Jose Hernadez Orallo. (2003). Obtenido de http://users.dsic.upv.es/~jorallo/cursoDWDM/dwdm-I.pdf.

2.3.6 Data Mart

El trabajo de construir un Data Warehouse corporativo puede generar inflexibilidades, o ser costoso y requerir plazos de tiempo que las organizaciones no están dispuestos a aceptar. En parte, estas razones originaron la aparición de los Data Mart. Los Data Mart están dirigidos a una comunidad de usuarios dentro de la organización, que puede estar formada por los miembros de un departamento, o por los usuarios de un determinado nivel organizativo, o por un grupo de trabajo multidisciplinar con objetivos comunes.

Los Data Mart almacenan información de un número limitado de áreas; por ejemplo, pueden ser de marketing y ventas o de producción. Normalmente se definen para responder a usos muy concretos. Normalmente, los Data Mart son más pequeños que los Data Warehouses. Tienen menos cantidad de información, menos modelos de negocio y son utilizados por un número inferior de usuarios.

Los Data Mart pueden ser independientes o dependientes. En las figuras 2.7 y 2.8 se muestra un ejemplo de Data Mart dependiente e independiente respectivamente. Los primeros son alimentados directamente de los orígenes de información, mientras que los segundos se alimentan desde el Data Warehouse corporativo. Los Data Mart independientes pueden perpetuar el problema de los "silos de información" y en su evolución pueden llegar a generar inconsistencias con otros Data Mart.

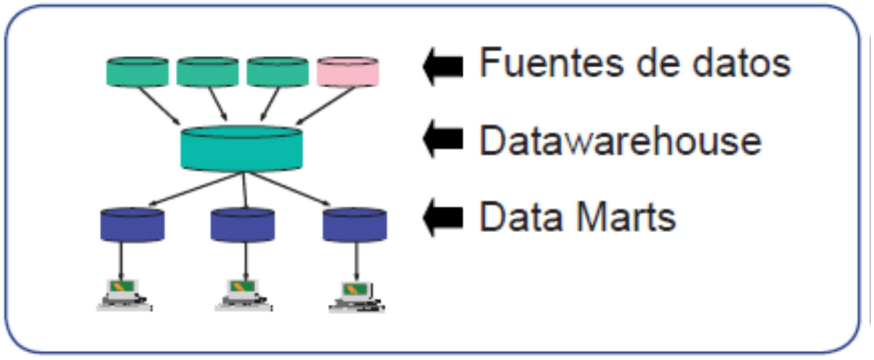

Figura 2.7 - Data Mart Dependiente [Fuente: Josep Lluís Cano, Business Intelligence: Competir con Información, 2007]

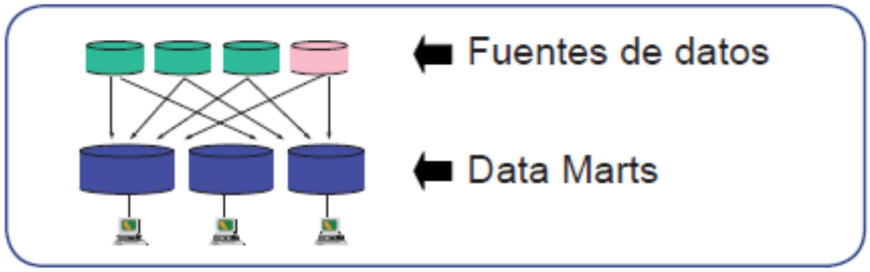

Figura 2.8 - Data Mart Independiente [Fuente: Josep Lluís Cano, Business Intelligence: Competir con Información, 2007]

Para la construcción de un Data Warehouse se han definido dos estrategias básicas:

- La defendida por W.H. Inmon, que propone definir un Data Warehouse corporativo y a partir de él ir construyendo los modelos de análisis para los distintos niveles y departamentos de la organización; es decir, una estrategia de arriba abajo, desde la estrategia a lo más operativo.

- La defendida por R. Kimball es la de construir distintos Data Marts que cubran las distintas necesidades de la organización, sin la necesidad de construir un Data Warehouse.

Con la estrategia de definir un Data Warehouse corporativo, el Data Warehouses desarrollado en fases y cada una de las mismas debe ser diseñada para generar valor para el negocio. Se construye un Data Warehouse corporativo, del que se cuelga un Data Mart dependiente con una parte de la información del Data Warehouse. En fases posteriores se van desarrollando Data Marts usando subconjuntos del Data Warehouse. Igual que los proyectos complejos, es caro, necesita mucho tiempo y es propenso al fracaso. Cuando tenemos éxito conseguimos un Data Warehouse integrado y escalable.

Si optamos por la estrategia más común, la de construir distintos Data Marts, el proyecto comienza con un Data Mart único al que posteriormente se irán añadiendo otros Data Marts que cubrirán otras áreas de negocio. Normalmente no requiere de grandes inversiones y es fácil de implementar, aunque con lleva

algunos riesgos; de entre ellos, cabe destacar fundamentalmente dos: puede perpetuar la existencia del problema de “silos de información” y posponer la toma de decisiones que conciernen a la definición de criterios y modelos de negocio. Si seguimos esta estrategia debemos tener claro el plan de acción, es decir, qué áreas cubriremos y la integración de los distintos modelos. Esta estrategia se utiliza a veces como un paso previo al desarrollo de un Data Warehouse corporativo.

Las dos aproximaciones abogan por construir una arquitectura robusta que se adapte fácilmente a los cambios de las necesidades de negocio y que nos proporcione una sola versión de la verdad.

2.3.7 Bases de Datos Multidimensionales

Son bases de datos ideadas para desarrollar aplicaciones muy concretas, como creación de Cubos OLAP.

Una MDB (Multidimensional Data Base, Base de Datos Multidimensional) es un repositorio de datos que proporciona un entorno integrado para consultas de soporte a las decisiones que requieren de agregaciones, y de enormes cantidades de datos históricos.

Su estructura básica se puede representar con el simple diagrama de entidad-relación como se muestra en la figura 2.9, en el que todas las entidades D representan las dimensiones de la MDB, mientras que la entidad de conexión F es la tabla de hechos.

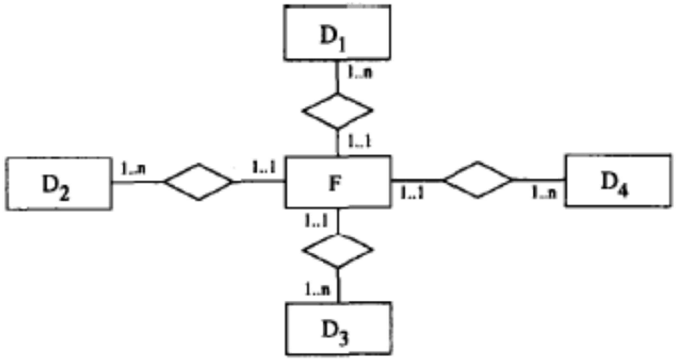

Figura 2.9 - Estructura de una base de datos multidimensional

Cada tabla de dimensiones “D” contiene toda la información que es específica solo a la propia dimensión, mientras que la tabla “F” de hechos se correlaciona con todas las dimensiones y contiene información sobre los atributos de interés.

En una base de datos multidimensional, la información se representa como matrices multidimensionales, cuadros de múltiples entradas o funciones de varias variables sobre conjuntos finitos. Cada una de estas matrices se denomina Cubo. El esquema de un cubo queda determinado dando a conocer sus ejes con sus respectivas estructuras y la estructura de los datos que se presentan en cada celda de la matriz. Se asume que los datos en todas las celdas son uniformes, es decir, todas las posiciones de la matriz tienen datos con igual estructura. Una instancia de un cubo, queda determinada por un conjunto de datos para cada eje y un conjunto de datos para la matriz.

Bases de Datos Multidimensionales Vs Bases de Datos Relacionales

Bases de Datos Relacionales

El modelo de base de datos relacional usa una estructura de dos dimensiones de filas y columnas para almacenar datos. Las tablas pueden estar unidas por valores comunes.

El acceso a los datos desde bases de datos relacional puede requerir joins demasiados complejos de muchas tablas y es claramente un elemento no trivial para usuarios finales sin entrenamiento.

Para obtener la información deseada de los datos, se requieren consultas complejas las cuales pueden tomar demasiado tiempo para devolver los resultados.

Bases de Datos Multidimensionales

Mejora la presentación y navegación de los datos. El análisis de datos y toma de decisiones es mucho más fácil a través de las base de datos multidimensionales comparado con bases de datos relacionales.

2.4. Modelado Multidimensional: Conceptos fundamentales.

2.4.1 Introducción

Actualmente podemos abordar el estudio del modelado multidimensional desde dos perspectivas. Una correspondiente a la parte estructural permite representar los aspectos estáticos del análisis multidimensional, y la parte dinámica permite aplicar ciertas operaciones para manipular los datos, comúnmente conocidas como operaciones OLAP.

La mayoría de los modelos multidimensionales están constituidos principalmente por un conjunto hechos y un conjunto de dimensiones. Los hechos, generalmente con relaciones “muchos a muchos” con las dimensiones, representan el objeto de análisis mientras que las dimensiones son las diferentes perspectivas a utilizar para analizar los hechos.

Dentro de la mayoría de los modelos multidimensionales podemos encontrar una dimensión común a todos ellos, ésta es la dimensión tiempo. Esto es debido a que es muy común llevar a cabo análisis de los datos históricos conforme su evolución en el tiempo.

Generalmente, las dimensiones se estructuran en jerarquías de agregación. A cada nivel de una jerarquía se le llama nivel de agregación o simplemente nivel. De esta forma, se puede considerar que toda dimensión siempre tiene por lo menos una jerarquía con un único nivel. En la figura 2.10 se muestra una instancia de una dimensión con una jerarquía de agregación en donde los vendedores se agrupan en ciudades y las ciudades en regiones.

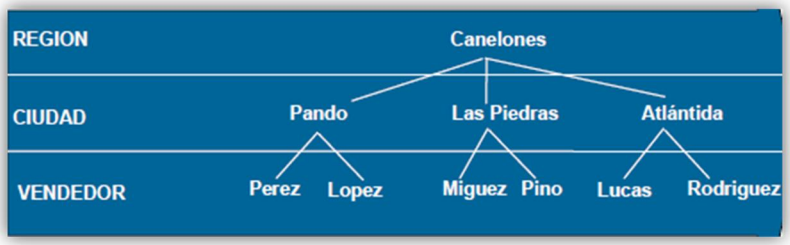

Figura 2.10 - Instancia de una Dimensión [Fuente: Fernando Carpani, Tesis de Maestría, 2000]

Estas jerarquías han de estar dirigidas y no formar ciclos, es decir, un nivel A se clasifica con uno B y no al contrario; y no puede existir alguna clasificación que parta de un nivel A y nos conduzca al mismo nivel A.

Finalmente, un nivel de una dimensión se puede clasificar en más de un atributo distinto para dar origen a lo que se denomina jerarquía múltiple. Así si dos o más jerarquías simples se unen en un nivel común, tenemos lo que se denomina jerarquías de caminos alternativos o múltiples que se pueden representar mediante Grafos Aciclicos Dirigidos; ya que estos permiten expresar distintas jerarquías (caminos) preservando las propiedades de dirigidas y no cíclicas.

Por ejemplo en la figura 2.11, la dimensión tiempo puede tener una jerarquía formada por *fecha →mes →trimestre → año*; además de otra jerarquía formada por *fecha →semana →año*.

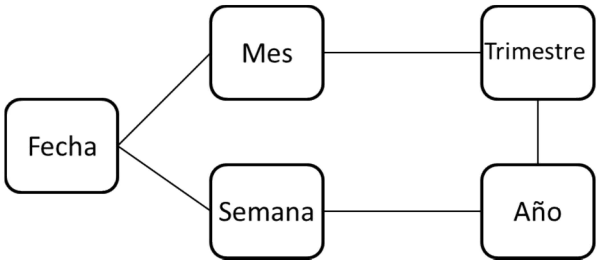

Figura 2.11 - Dimensión de Jerarquía Múltiple [Fuente: Carlos Neil, Tesis de Doctorado, 2010]

Por último en cuanto a las jerarquías de clasificación son los conceptos de jerarquías estrictas y completas. Una jerarquía es estricta cuando los elementos de un nivel inferior de una jerarquía pertenecen únicamente a un elemento del nivel de jerarquía superior. Por otro lado, una jerarquía es completa cuando todos los elementos de un nivel inferior pertenecen únicamente a un elemento del nivel

superior y que un elemento del nivel superior consiste únicamente de esos miembros.

Por ejemplo en la figura 2.12, la jerarquía producto →tipo de producto →departamento →región puede considerarse una jerarquía no-estricta si un tipo de producto pertenece a más de un departamento.

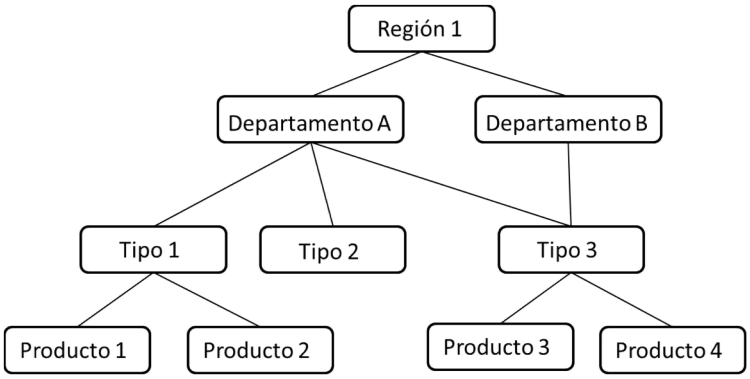

Figura 2.12 - Dimensión con Jerarquía No-Estricta [Fuente: Carlos Neil, Tesis de Doctorado, 2010]

El objetivo final de definir jerarquías de clasificación es poder utilizar sus niveles para agrupar los datos que se desean analizar y obtener así datos resumidos o agregados.

Una vez definido el modelo multidimensional desde la perspectiva estructural, a continuación, se presenta desde su perspectiva dinámica, que se compone de dos partes:

- Los requisitos iniciales definidos para analizar los datos representados en la parte estructural y,
- Las operaciones OLAP.

En primer lugar, los usuarios solicitan requisitos iniciales para analizar los hechos con respecto a las dimensiones definidas. Una vez que el requisito ha sido resuelto y los datos devueltos, el usuario comienza un proceso interactivo en función de estos datos devueltos a través de las operaciones OLAP.

Al conjunto de operaciones que permiten al usuario interactuar con los datos devueltos por un requisito se les denomina operaciones OLAP. Las operaciones OLAP son las siguientes:

- Roll-Up: permite aumentar el nivel de agregación de los datos a lo largo de los niveles de jerarquía definidos en las dimensiones.
- Drill-Down: permite disminuir el nivel de agregación de los datos a lo largo de los niveles de jerarquía definidos en las dimensiones.
- Slice: permite restringir los valores de una o más dimensiones a un valor o rango de valores.
- Dice: permite establecer restricciones en los datos del hecho.

2.4.2 Modelado Multidimensional Conceptual

En los últimos años han aparecido una serie de propuestas para modelar conceptualmente una base de datos. La propuesta que ofrece mayor expresividad y términos semánticos de dimensión y hecho necesarios para modelar esta aplicacion es la que se describe a continuación:

- CMDM: un modelo conceptual para la especificación de BDM (Carpani, 2000)[18].

2.4.2.1 Modelo CMDM

Para realizar el modelado multidimensional, el modelo CMDM (Conceptual MultiDimensional Model, Modelo Conceptual Multidimensional) presenta tres estructuras básicas: niveles, dimensiones y relaciones multidimensionales.

Los niveles representan un conjunto de objetos que son del mismo tipo. Para representar un nivel el modelo utiliza un rectángulo que contiene el nombre y

[18] Carpani, F. (2000). CMDM: un modelo conceptual para la especificacion de BDM.

la estructura del tipo de ese nivel. Los niveles se organizan en jerarquías y cada jerarquía está compuesta por uno o varios niveles. En cada jerarquía se tiene una relación 1-n entre objetos de nivel superior e inferior.

Las dimensiones están determinadas por una jerarquía de niveles. En el modelo una dimensión se representa por un rectángulo dentro del cual aparece un nombre para la dimensión y un grafo dirigido en donde los nodos son los niveles que participan de esa dimensión.

Una relación dimensional representa un conjunto de cubos, tomado del conjunto de todos los cubos que se pueden construir a partir de los niveles de un conjunto dado de dimensiones.

Se asume que en cada uno de los cubos que pertenecen a la instancia de la relación dimensional, debe aparecer al menos un nivel de cada una de las dimensiones que participan en la relación.

Notar que, en CMDM, los cubos son elementos de las instancias de las relaciones dimensionales y no hay una estructura que los represente directamente.

Por lo tanto, el esquema de una relación dimensional está dado por un grafo en forma de estrella. El nodo central es de forma oval y tiene el nombre de la relación dimensional y los nodos “satélite” son rectangulares y tienen el nombre de cada una de las dimensiones que participan de la relación. En la figura 2.13 se muestra un ejemplo del modelo CMDM.

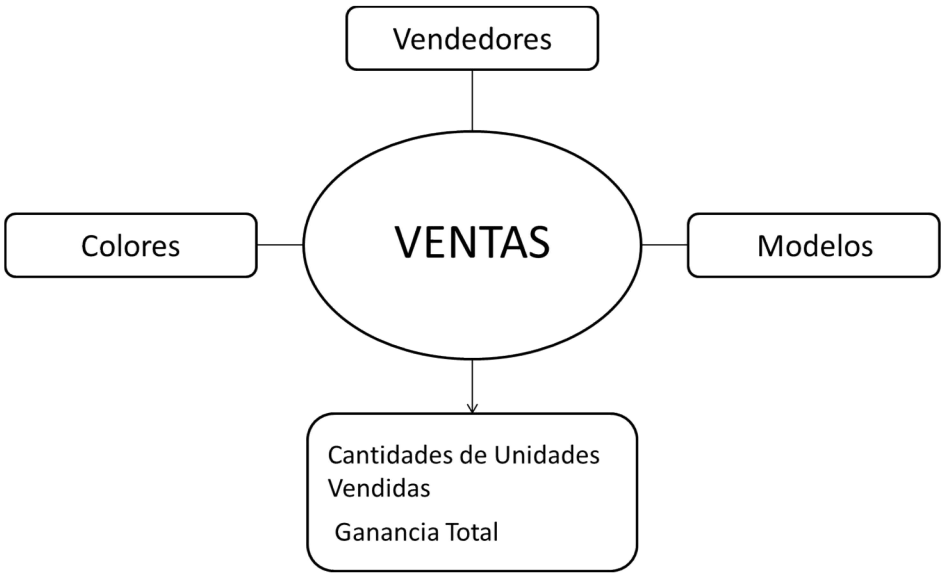

Figura 2.13 - Ejemplo de esquema multidimensional modelado con CMDM [Fuente: Fernando Carpani, Tesis de Maestria, 2000]

2.4.3 Modelos Multidimensionales Lógicos

En el ámbito de bases de datos diversos métodos han sido propuestos para la generación de un modelo lógico de un Data Warehouse. Hoy en día en los servidores ROLAP (Relational On-Line Analytic Processing, Procesamiento Analítico Relacional en Linea) los esquemas más utilizados son los esquema estrella y copo de nieve. A continuación se presentan cada uno de estos esquemas.

Esquema Estrella

El esquema estrella consiste en una o más tablas centrales denominadas tablas de hechos rodeadas por una serie de tablas de dimensiones que forman una especie de “estrella”. Cada tabla de hecho corresponde con cada hecho definido en el modelo conceptual así como cada tabla de dimensión se corresponde con cada dimensión definida. La tabla de hechos representa una relación “muchos a muchos” entre todas las tablas de dimensiones que relaciona. Sin embargo, representa una relación “muchos a uno” con cada tabla de

dimensión por separado. Por lo tanto, la clave primaria de la tabla de hechos está compuesta por las claves de las tablas de dimensiones con las que se relaciona.

En algunas ocasiones puede suceder que la clave primaria compuesta descrita anteriormente no sea suficiente para identificar a las instancias de la tabla de hecho. En estos casos se introduce una componente más en la clave primaria de la tabla de hechos según el dominio que se modele para identificar dichas instancias.

Una consideración hecha por algunos autores referente a las claves primarias de las tablas de dimensiones es el hecho que las tablas de dimensiones tengan claves generadas por el sistema y que en principio sus claves primarias originales sean un campo más en la tabla.

Por otro lado el esquema estrella utiliza la desnormalización para definir las tablas de hechos y de dimensiones por dos razones fundamentales; la primera, porque es mucho más intuitivo para el análisis multidimensional al estar muy próximo al proceso cognitivo seguido al llevar a cabo este tipo de análisis: hechos y dimensiones, y la segunda, porque al existir un número mínimo de relaciones entre tablas, la recuperación de los datos es más rápida, más aun debido al gran volumen de datos manejados por las aplicaciones OLAP.

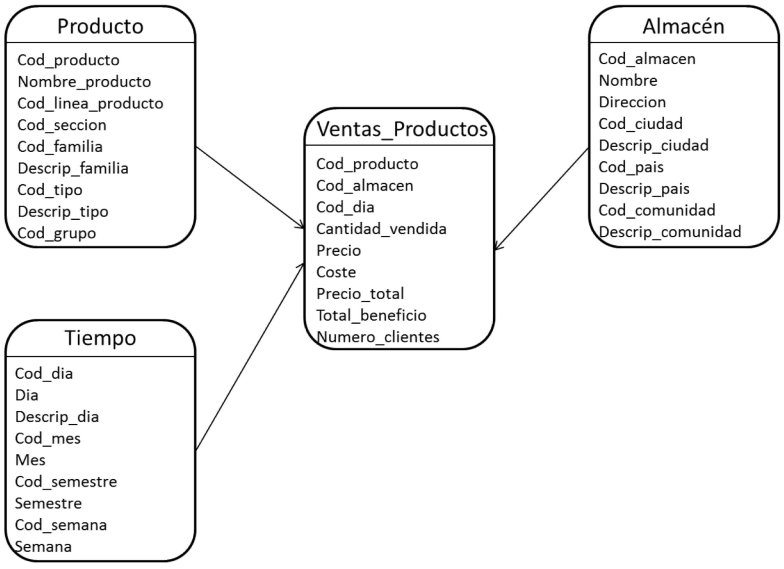

Figura 2.14 - Ejemplo de Esquema Estrella [Fuente: Juan Carlos Trujillo, Tesis de Doctorado, 2001]

Por otro lado, se puede observar en la figura 2.15 que en el esquema estrella no se soporta explícitamente la definición de jerarquías de clasificación de los elementos de dimensión. Sin embargo, las jerarquías son expresadas en la implementación de dicho esquema a través de los denominados atributos de nivel definidos en cada una de las dimensiones. Este atributo contendrá números arbitrarios que identificaran a cada nivel de la jerarquía. Por lo tanto, cada registro con el mismo valor para esta columna serán los registros que pertenezcan al mismo nivel de jerarquía.

Esquema Copo de Nieve

Otra alternativa al esquema estrella para disminuir el tamaño de las tablas de dimensiones y no utilizar el atributo de nivel es utilizar el esquema copo de nieve. Por ello, suele aplicarse cuando un gran número de atributos caracterizan a los niveles más altos de las jerarquías y se tiene conocimiento de que un gran número de requisitos requerirán de datos agregados.

En este esquema se normalizan las tablas de las dimensiones por sus niveles de jerarquía. Ahora las tablas de dimensiones sólo contendrán los atributos que caracterice a los niveles más bajo de las jerarquías y las claves ajenas de las nuevas tablas que forman los niveles de las jerarquías. Cada nivel de jerarquía se representa como una tabla que contendrá solo los atributos que caracterizan a dicho nivel más una clave ajena para relacionar dicha tabla con la tabla que representa el siguiente nivel de jerarquía.

En la figura 2.16 se muestra un ejemplo de esquema copo de nieve parcial ya que únicamente se ha normalizado la dimensión almacén. Se puede apreciar como la tabla de dimensión almacén contiene todas las claves ajenas de las nuevas tablas que representan los niveles de jerarquía y los atributos que describen a los almacenes únicamente. Se puede observar como la tabla ciudad contiene los atributos que describen a las ciudades además de la clave ajena que apunta a la tabla comunidad, ya que esta representa el siguiente nivel de jerarquía.

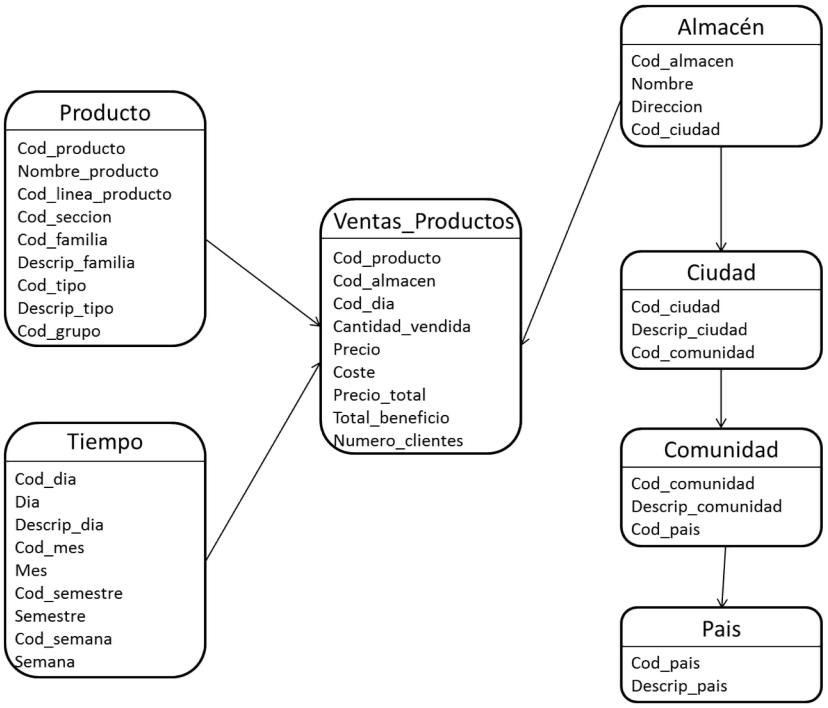

Figura 2.15 - Ejemplo de Esquema Copo de Nieve [Fuente: Juan Carlos Trujillo, Tesis de Doctorado, 2001]

Es conveniente añadir que en la práctica se puede decidir aplicar esquema de copos de nieve solo a algunas dimensiones, mientras otras dimensiones se pueden mantener totalmente denormalizadas.

Para finalizar, en la tabla 2.6 se muestra un resumen de las ventajas y desventajas del esquema estrella y el esquema copos de nieve.

Tabla 2.4 - Resumen Esquema Estrella y Copo de Nieve [Fuente: Elaboración propia]

Modelo	Ventajas	Desventajas
Modelo Estrella	Fácil de entender. Reduce el número de uniones físicas.	La necesidad de manejar el atributo nivel para datos agregados. Las dimensiones tienen un tamaño enorme.
Modelo Copo De Nieve	No es necesario definir el atributo nivel.	Aumenta la complejidad de mantener la meta

	Fácil para definir jerarquías	información debido al aumento en el número de tablas.

2.4.4 Lenguaje de Consulta

Una vez construido el Data Warehouse la siguiente etapa es obtener el nuevo conocimiento oculto en el cubo, para ello existen dos formas de extraer la información. Una de ellas es utilizar el estándar SQL (Structured Query Languaje, Lenguaje de Consulta Estructurado) y la otra consiste en utilizar un lenguaje enriquecido como lo es MDX (Multidimensional eXpressions, Expresiones Multidimensionales).

SQL

SQL es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones sobre las mismas. Sin embargo SQL ha sido extendido para poder realizar consultas multidimensionales sobre cubos OLAP. Una de sus características distintivas es el manejo del álgebra y el cálculo relacional, que permite realizar consultas con el fin de recuperar, de una forma sencilla, información de interés de una base de datos, así como también hacer cambios sobre la misma.

Los orígenes de SQL están ligados a los de las bases de datos relacionales. En 1970 E. F. Codd propone el modelo relacional y asociado a éste un sublenguaje de acceso a los datos basado en el cálculo de predicados. SQL es un lenguaje de acceso a bases de datos que explota la flexibilidad y potencia de los sistemas relacionales permitiendo gran variedad de operaciones sobre los mismos. Es un lenguaje declarativo de “alto nivel” o “no procedimental” orientado al manejo de conjuntos de registros. Es declarativo por cuanto especifica qué es lo que se quiere y no como conseguirlo.

SQL permite definir la estructura de los datos, recuperar y manipular datos, administrar y controlar el acceso a los datos, compartir datos de forma concurrente

y asegurar su integridad. Además, los comandos del lenguaje pueden ser ejecutados interactivamente o bien formando parte de un programa de aplicación, donde las sentencias SQL están insertas dentro del código fuente y mezcladas con las sentencias del código.

4.1.1 Implementación de Consultas Multidimensionales en SQL

Vamos a empezar a construir nuestra primera consulta en SQL. La base de las consultas SQL tiene la siguiente estructura:

SELECT nivel-ID$_1$,…, nivel-ID$_n$, **Funcion**(H.medida) **FROM** Hecho H, Dim$_1$ d$_1$,…, Dim$_n$ d$_n$ **WHERE** H.pk$_1$ = D$_1$.ID **AND**,…, H.pk$_n$ = D$_n$.ID **AND** d$_1$.attr = valor1 **AND GROUP BY** nivel-ID$_1$,…, nivel-ID$_n$ **HAVING Funcion**(H.medida) oper valor2 **ORDER BY** nivel-ID$_1$,…, nivel-ID$_n$

La cláusula FROM contiene la tabla hecho y las tablas de dimensiones, estas tablas están vinculadas mediante la cláusula WHERE que también define las condiciones sobre las columnas de las tablas de dimensión. La cláusula GROUP BY muestra los identificadores de los niveles sobre los cuales agrupamos los datos, esos mismos identificadores aparecen en la cláusula SELECT, previos a la función de agregado. La cláusula HAVING establece restricciones sobre los valores de las medidas. Por último, la cláusula ORDER BY explicita el orden de salida de la consulta.

Sin embargo a pesar de que SQL se ha extendido de muchas formas, incluyendo la adición de soporte para consultas temporales, añadiendo CUBE y ROLLUP etc, SQL enfrenta algunas desventajas al momento de realizar consultas multidimensionales. Algunas de ellas son:

- Muchos cálculos básicos OLAP tienden a ser complejos.
- Otros cálculos son imposibles.
- Siendo un lenguaje que no está optimizado para OLAP, algunas consultas multidimensionales resultan demasiado lentas.

MDX

MDX es un lenguaje de consulta creado especialmente para Bases de datos OLAP. También es un lenguaje de cálculo, con una sintaxis similar a las fórmulas de hoja de cálculo. El lenguaje de expresiones multidimensionales (MDX) proporciona una sintaxis especializada para consultar y manipular los datos multidimensionales almacenados en cubos OLAP.

Si bien es posible traducir algunas de estas sentencias a SQL, generalmente requieren la síntesis de torpes expresiones SQL incluso para las expresiones MDX más sencillas.

MDX se introdujo por primera vez como parte de la especificación OLE DB para OLAP en 1997 por Microsoft. Fue inventado por el grupo de ingenieros de SQL Server incluyendo Mosha Pasumansky. La especificación fue seguida rápidamente por la liberación comercial de Microsoft OLAP Services 7.0 en 1998 y más tarde por Microsoft Analysis Services.

Aunque no es un estándar abierto, sino más bien una especificación de propiedad de Microsoft, MDX ha sido adoptada por una amplia mayoría de los proveedores de OLAP y se ha convertido en el estándar de facto para los sistemas OLAP.

MDX es una extensión de SQL, por lo que su estructura es similar a la de SQL, y en ciertos casos las mismas palabras claves tienen funciones similares, sin embargo existen diferencias.

4.2.1 Implementación de Consultas Multidimensionales con MDX

Se explicara con ejemplo MDX usando un cubo de análisis de ventas para una empresa hipotética que vende accesorios de computadores alrededor del mundo. Este cubo contiene dimensiones típicas tales como tiempo, geográfica, productos, clientes, etc.

Se empezara a construir la primera consulta MDX. La base de las consultas MDX tiene la siguiente estructura:

SELECT $axis_1$ ON COLUMNS, $axis_2$ ON ROWS FROM *cube*

Se va a comparar esto con una instrucción SQL:

```
SELECT column1, column2, …, columnn FROM table
```

Se pueden encontrar algunas de las palabras clave de MDX familiares. De hecho, SELECT y FROM sirven exactamente del mismo modo que en SQL. La palabra clave SELECT establece lo que se quiere ejecutar en una consulta y obtener resultados de datos.

La palabra clave FROM especifica la fuente de los datos. En el caso de SQL recibimos los datos de la tabla, y en caso de MDX que recibimos los datos del cubo. Sin embargo, hay algo diferente entre las dos consultas anteriores. SQL básicamente nos da vista relacional de los datos, que siempre es de dos dimensiones. El resultado de la consulta SQL es una tabla que tiene dos dimensiones: filas y columnas. Y estas dimensiones no son simétricas. Las filas son todas de la misma estructura, y esta estructura se define por las columnas.

En MDX, se puede especificar cualquier número de dimensiones para formar el resultado de una consulta. Se nombran ejes (axis) para evitar la confusión con las dimensiones del cubo. Se puede tener cero, uno, dos, tres o cualquier otro número de ejes. Y todos estos ejes son perfectamente simétricos.

No hay un significado especial para ROWS y COLUMNS que no sean para fines de visualización en MDX.

En MDX, se tiene que definir la estructura de cada eje. Por tanto, en el ejemplo los axis 1 será la definición de los ejes que aparecen en posición horizontal, y axis2 será la definición de los ejes que aparecen en vertical. Axis es una colección de miembros de dimensión, o más generalmente tuplas.

Hay muchas maneras de definir un eje de MDX. Se comenzara con una forma simple. Si se quiere dar a conocer en el eje todos los miembros de cierta dimensión. La sintaxis para este tipo de definición seria:

<dimension name>.MEMBERS

Y del mismo modo, si queremos ver a todos los miembros pertenecientes a la dimensión de cierto nivel de esta dimensión, la sintaxis sería:

<dimension name>.<level name>.MEMBERS

Ahora se construirá la primera consulta MDX con toda la información que se ha aprendido hasta ahora. Se utilizara el cubo de ejemplo. En nuestra primera consulta MDX se muestran las ventas en diferentes continentes durante los años. Por lo tanto la consulta se verá así

```
SELECT
  Years.MEMBERS ON COLUMNS,
  Regions.Continent.MEMBERS ON ROWS
FROM Sales
```

El resultado de la siguiente consulta se puede visualizar en la tabla 2.7:

Tabla 2.5 - Resultados Consulta [Fuente: Mosha Pasumansky, MDX for Everyone, 1998]

Continent	1994	1995	1996	1997
N. America	120,000	200,000	400,000	600,000
S. America	-	10,000	30,000	70,000
Europe	55,000	95,000	160,000	310,000
Asia	30,000	80,000	220,000	200,000

Se tratara de analizar los resultados de esta consulta. Se tiene una tabla con dos ejes. El eje horizontal (es decir, columns) contiene los miembros de la dimensión ‘Years’, y el eje vertical (es decir, rows) contiene los miembros del nivel ‘Continent’ de la dimensión ‘Regions’.

Todos estos son especificados explícitamente en la consulta. Sin embargo, los resultados obtenidos no hacen referencia a los datos cuantitativos. Puesto que, en la nuestra consulta MDX no se específica la medida que se quiere utilizar, la medida inferida se obtiene por defecto. En el ejemplo, la medida por defecto es 'Ventas' así, la interpretación del número de la esquina superior izquierda de la tabla es "en el año 1994, las ventas en América del Norte fueron 120.000".

Las dimensiones o niveles a menudo contienen miles de miembros y por lo tanto, a veces, no se necesita ver los a todos ellos. Algunos miembros pueden ser irrelevantes para el análisis actual. Por ejemplo, si se analiza la consulta, se

observa que se produce información acerca de todos los años. Por lo tanto, si sólo quiere ver las ventas por continentes en 1996 y 1997 se puede especificar una lista de los miembros como una definición de eje. Por ejemplo:

$\{ dim.member_1, dim.member_2, \ldots , dim.member_n\}$

Se reescribirá la consulta MDX para explicar esta sintaxis y representar sólo años 1996 y 1997:

```
SELECT {Years.[1996], Years.[1997] } ON COLUMNS,
Regions.Continent.MEMBERS ON ROWS FROM Sales
```

Se puede notar que los nombres de los miembros de 1996 y 1997 están entre corchetes. Esta cita se hace para evitar la confusión del analizador MDX sobre la naturaleza de 1996 y 1997. Sin los corchetes, el analizador de MDX puede decidir que se trata solo de números en lugar de nombres de miembros. En general, entre corchetes, se permite tener nombres de miembros que contengan cualquier tipo de símbolos, como espacios, puntos, etc.

Operaciones Multidimensionales

Una característica distintiva de las herramientas OLAP es su capacidad de agrupación de medidas considerando una o más dimensiones, la visión multidimensional permite analizar a los distintos niveles de jerarquías en cada dimensión de una manera lógica.

Aunque no hay consenso sobre un conjunto mínimo de operaciones sobre el modelo multidimensional, la mayoría de los trabajos de investigación presentan dos grandes grupos de consultas:

- **Drilling**: estas operaciones permiten navegar a través de los distintos niveles de jerarquías de las dimensiones, con el objetivo de analizar una medida con mayor o menor nivel de detalle. Cuando la consulta navega de

mayor a menor nivel de detalle en los datos, la operación se denomina *ROLL-UP*, la navegación inversa es *DRILL-DOWN***.**

- **Selections*:*** estas operaciones permiten al usuario trabajar con un subconjunto de los datos disponibles. *SLICE* especifica restricciones en los datos de las dimensiones y *DICE* establece restricciones en los datos del hecho.

Las estructuras multidimensionales son consideradas, en general, en forma aislada, sin embargo cuando son utilizadas para realizar consultas puede ser necesario navegar de una estructura a otra (drill-across). Esto significa que se puede analizar un cubo desde un punto de vista y querer luego ver datos, en otra estructura, desde otro punto de vista; para que esto sea posible, las estructuras multidimensionales necesitan tener puntos de equivalencias, esto es, dimensiones en común.

Un aspecto importante a considerar en el diseño de un Data Warehouse está referido a la aplicación de la operación “suma” en los atributos de hecho. La violación de esta propiedad puede conducir a conclusiones y decisiones erróneas.

Conocer cuáles son los atributos no sumarizables se torna imprescindible, ya que esta operación no puede aplicarse a ellos (pero sí otras, por ejemplo promedios, máximos o mínimos) debido a que las reglas semánticas pueden no ser bien conocidas y el resultado, si bien puede ser establecido (es decir, se puede obtener un valor), éste no será correcto. Por ejemplo, el valor cantidad de ítems vendidos por mes en un supermercado es un atributo sumarizable, esto es, se pueden sumar esos valores y determinar, mensualmente, cuantos ítems se vendieron. Por otro lado, el atributo cantidad de ítems que permanecen sin vender en la góndola del supermercado no es sumarizable, ya que esa suma puede repetir ítems de meses anteriores y, por lo tanto, el valor resultante carece de sentido práctico.

Por lo tanto, en el análisis, es útil la agregación de instancias de hecho a diferentes niveles de abstracción (roll up), por lo que la mayoría de los atributos de hecho deberían ser aditivos, esto significa que el operador suma puede ser usado

para agrupar valores de atributos a lo largo de todas las jerarquías. Un atributo se denomina semi-aditivo si no es aditivo en una o más dimensiones y no-aditivo, si no lo es en ninguna dimensión

A continuación se analizaran las operaciones más importantes descritas anteriormente. Para ejemplificar las distintas operaciones multidimensionales se utilizará como base el siguiente sistema de información. “La empresa fabrica y centraliza sus operaciones comerciales, pero sus clientes están distribuidos geográficamente en todo el país. Realiza ventas de productos a partir de pedidos de clientes que están ubicados en distintas provincias, registra las transacciones realizadas, las fechas y cantidades vendidas”.

La figura 2.17 ofrece una primera aproximación del sistema de ejemplo. En el mismo estableceremos los siguientes niveles de jerarquía (no explicitadas en el gráfico). En la dimensión producto; “producto -> tipo de producto -> departamento -> región”. En la dimensión fecha; “fecha -> mes -> trimestre -> año”. En la dimensión cliente; “cliente -> categoría”.

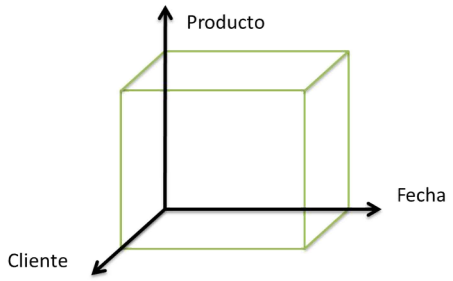

Figura 2.16 - Cubo multidimensional [Fuente: Creación propia]

Operación de Roll-Up

La operación roll-up, permite reducir el nivel de detalle en el que se consultan los datos, realizando operaciones de agregación (sumas, promedios, etc.) a través de los niveles de jerarquías de las dimensiones. Calcula las medidas en función de agrupamientos, es decir, realiza el re-cálculo de la medida de acuerdo a los ajustes de escala.

Por ejemplo, en la Tabla 2.8 se observa una representación tabular de las ventas por producto, cliente y fecha, donde agregaremos una columna que explicita el nivel de jerarquía “tipo de producto”. De acuerdo a los valores de la tabla, las ventas totales por producto son: (azúcar, 60), (leche, 10) y (yogurt, 12).

Tabla 2.6 - Representación tabular de las ventas [Fuente: Carlos Neil, Tesis de Doctorado, 2003]

Tipo de Producto	Producto	Cliente	Fecha	Cantidad
No perecedero	Azúcar	Fernández	19-02-2008	20
No perecedero	Azúcar	Rodríguez	19-02-2008	40
No perecedero	Arroz	Fernández	20-02-2008	30
Lácteos	Leche	Pérez	20-02-2008	10
Lácteos	Yogurt	Pérez	23-12-2008	12

Si ahora agrupamos por “tipo de producto”, las ventas pueden expresarse, mediante la operación roll-up, como (productos no perecederos, 90) y (productos lácteos, 22). La tabla 2.9 muestra el resultado de la operación Roll-Up sobre los datos:

Tabla 2.7 - Representación de las ventas después de Roll-Up [Fuente: Carlos Neil, Tesis de Doctorado, 2003]

Tipo de Producto	Cantidad
No perecedero	90
Lácteos	22

Operación de Drill-Down

Esta operación es la inversa de la operación roll-up, es decir permite aumentar el nivel de detalle al que se consultan los datos, al ir a un nivel más bajo dentro de la jerarquía.

Siguiendo el ejemplo anterior, si tuviéramos las ventas por "tipo de producto" y quisiéramos desagregarlas por producto, el resultado ahora sería: (azúcar, 60), (arroz, 30), (leche, 10) y (yogurt, 12). La tabla 2.10 muestra el resultado de la operación Drill-Down sobre los datos:

Tabla 2.8 - Representación de las ventas después de Drill-Down [Fuente: Carlos Neil, Tesis de Doctorado, 2003]

Producto	Cantidad
Azúcar	60
Arroz	30
Leche	10
Yogurt	12

Operaciones de Slice y Dice

Estas operaciones permiten reducir el conjunto de datos consultados por medio de operaciones de proyección y selección de datos (filtrado), basándose en los atributos de las dimensiones y hechos. La operación slice, consiste en restringir los valores de una o más dimensiones a un valor o rango de valores.

Por ejemplo, en la tabla 2.8, si fijamos la dimensión cliente en un valor (por ejemplo, Fernandez), reducimos el número de dimensiones consideradas; la representación resultante se muestra en la tabla 2.11.

Tabla 2.9 - Representación de las venteas después de Slice [Fuente: Carlos Neil, Tesis de Doctorado, 2003]

Producto	Fecha	Cantidad
Azúcar	19-02-2008	20

Arroz	20-02-2008	30

La operación dice, establece restricciones en los datos del hecho. Siguiendo con el ejemplo, si ahora restringiéramos la consulta para los productos cuyas ventas fueron superiores a 15 unidades, la representación resultante se muestra en la tabla 2.12.

Tabla 2.10 - Representación de las ventas después de Dice [Fuente: Carlos Neil, Tesis de Doctorado, 2003]

Producto	Cliente	Fecha	Cantidad
Azúcar	Fernández	19-02-2008	20
Azúcar	Rodríguez	19-02-2008	40
Arroz	Fernández	20-02-2008	30

2.5 Datos de PISA

El siguiente apartado tiene como objetivo comprender la naturaleza de los datos para poder definir los indicadores dentro del modelo conceptual multidimensional. Este proceso podemos definirlo en dos etapas como se muestra a continuación:

- Compresión y recopilación de los datos de PISA.
- Proceso de análisis de los datos PISA.

2.5.1 Compresión y recopilación de los datos de PISA.

PISA (Programme for International Student Assessment, Programa para la Evaluación Internacional de los Estudiantes) es un estudio comparativo de evaluación de los resultados de los sistemas educativos de los países miembros y colaboradores de la OECD (Organisation for Economic Co-operation and Development, Organización para la Cooperación y el Desarrollo Económico).

Como se ha explicado anteriormente en el primer capítulo de este trabajo de título, PISA es un estudio realizado por OECD dirigido a estudiantes de 15 años que evalúa si los estudiantes tienen la capacidad de reproducir lo que han aprendido en la escuela para resolver problemas de la vida real.

El estudio se realiza cada tres años, y en cada ciclo se enfatiza uno de los tres dominios de evaluación (lenguaje, matemáticas y ciencias) y los otros son evaluados con menor profundidad.

Dentro de los países sudamericanos que han participado desde un comienzo en el estudio de PISA se encuentran Chile, Argentina y Brasil, destacando la adhesión de Chile como miembro número 31 de la OECD desde el 11/01/10.

2.5.2 Proceso de análisis de los datos.

El proceso de análisis de los datos comprende como primera parte del trabajo la recolección de los distintos manuales de las bases de datos del estudio de PISA. Estos documentos se encuentran en la página oficial de la OECD[19] y se encuentran bajo el nombre de:

- Manual for the PISA 2000 Database.
- Manual for the PISA 2003 Database.
- Manual for the PISA 2006 Database.
- Manual for the PISA 2009 Database.

Cada uno de estos documentos proporciona toda la información relacionada con las bases de datos para cada una de las versiones del estudio, dentro de lo que destacamos:

[19] Documentos obtenidos de www.pisa.oecd.org/

- Estructura general del estudio: En este apartado del documento se realiza un resumen general de la estructura del estudio.
- Información disponible del estudio: Este apartado del documento describe todos aquellos archivos disponibles para la descarga desde la página oficial de la OECD, los cuales se encuentran divididos en seis categorías:
    - Cuestionarios: Cuestionario del Estudiante, Cuestionario del Colegio, Cuestionario de Competencias, Cuestionario IT.
    - Codebooks: Información relacionada con cada una de las variables y todos sus respectivos valores válidos para esa variable.
    - Archivos de Control SAS: Archivos de compilación en formato SAS.
    - Archivos de Control SPSS: Archivos de compilación en formato SPSS.
    - Archivos de Datos: Datos relacionados con las respuestas de los alumnos para cada uno de los cuestionarios.
    - Compendia.
- Estructura de los cuestionarios: En esta sección se establece la nomenclatura utilizada en el estudio.
- Estimadores de rendimiento: En este apartado se introduce el concepto de valores plausibles para calcular estimaciones a nivel de población y su diferencia con respecto a estimadores puntuales para determinar el rendimiento de los estudiantes.
- Índices derivados de los cuestionarios: En esta sección del documento se hace una breve descripción de cada uno de los índices derivados de los cuestionarios de los estudiantes.

2.5.2.1 Estructura Base de datos PISA.

La segunda parte del análisis de los datos comprendió el análisis de la estructura de los mismos. Cada versión del estudio de PISA consta de una tabla

de 400 a 450 atributos por alumno. Los atributos se dividen en cuatro categorías principalmente:

- Atributos de identificación.
- Atributos de cuestionarios.
- Atributos de índices calculados.
- Atributos de estimadores de rendimiento.

La cantidad de atributos por cada estudio varía debido a que ciertas preguntas se agregan o se eliminan de los cuestionarios de los alumnos implicando que ciertos índices dejen de calcularse.

Atributos de identificación.

Los atributos de identificación permiten individualizar de manera única a un alumno dentro del estudio. La identificación de un alumno del estudio consiste básicamente de tres atributos, que juntos forman de manera única un identificador para cada alumno:

- El atributo de identificación del país etiquetado como COUNTRY. Los códigos de los países usados en PISA son los códigos de los países establecidos en la ISO 3166.
- El atributo de identificación del colegio etiquetado como SCHOOLID.
- El atributo de identificación del alumno etiquetado como STIDSTD.

Atributos de cuestionarios.

Estos atributos contienen las respuestas de los alumnos a cada una de las preguntas de los distintos cuestionarios. Los nombres que son usados para identificar estos atributos en la base de datos internacional están directamente

relacionados con la versión internacional de los cuestionarios. Cada nombre de atributo consiste de 7 caracteres, como se muestra a continuación:

Por ejemplo para la codificación ST ■■ Q ■■ se tiene:

- Los dos primeros caracteres se refieren al tipo de cuestionario. ST para el cuestionario del estudiante. IT para el cuestionario de familiaridad computacional. Y CC para el cuestionario de competencias interdisciplinarias.
- El tercer y cuarto carácter se refieren al número de la pregunta como aparece en la versión internacional del cuestionario. Por ejemplo, ST01 se refiere a la primera pregunta del cuestionario del estudiante relacionada con la fecha de nacimiento.
- El sexto y séptimo carácter se refieren al ítem de la pregunta. Por ejemplo, ST01Q01 se refiere al día de nacimiento, ST01Q02 se refiere al mes de nacimiento y ST01Q03 se refiere al año de nacimiento.

Atributos de índices calculados.

Estos atributos son índices calculados a partir de las respuestas de los alumnos en los cuestionarios. La nomenclatura que utilizan estos atributos para ser identificados dentro del estudio es solamente una abreviación del significado del índice. Por ejemplo, HISEI corresponde a Highest International Socio-Economic Index.

Atributos de estimadores de rendimiento.

El estudio proporciona un estimador de rendimiento conocido como valores plausibles que permiten determinar estadísticas de rendimiento a nivel de población.

Los valores plausibles fueron desarrollados para el análisis de los datos de la NAEP de 1983-1984 (Evaluación Nacional del Progreso Educativo), por Mislevy, Sheehan, Beaton y Johnson, basado en el trabajo de Rubin en múltiples imputaciones. Los valores plausibles se utilizaron en todos los estudios posteriores NAEP, TIMSS y posteriormente PISA.

La manera más simple de describir los valores plausibles es decir que los valores plausibles son una representación de la gama de capacidades que pueden suponerse razonablemente de un alumno. En lugar de estimar directamente la capacidad θ de un alumno, se estima una distribución de probabilidad para θ. Es decir, en lugar de obtener una estimación puntual para θ de un alumno, un abanico de valores posibles para la magnitud θ de un alumno, con una probabilidad asociada para cada uno es estimado. Los valores plausibles son valores aleatorios de esta distribución de θ para un alumno.

A continuación se muestra un ejemplo de la prueba de ciencias del año 2000 para ayudar a ilustrar el concepto de valores plausibles.

Suponga que un comité de la ciudad decide imponer un nuevo impuesto a los inmuebles para aumentar la recaudación de la ciudad. El nuevo impuesto será proporcional a la longitud del salón de la casa familiar. Los inspectores visitan todas las casas de la ciudad midiendo la longitud de los salones de las casas. Se les dan huinchas de medir y son instruidos de registrar la longitud en términos de enteros. Por ejemplo 1 metro, 2 metros, 3 metros, y así sucesivamente.

Los resultados se muestran en la figura 2.17. Cerca del 3% de los salones de las casas tienen una longitud de 4 metros, algo más del 16% de los salones reportan una longitud de 9 metros y así sucesivamente.

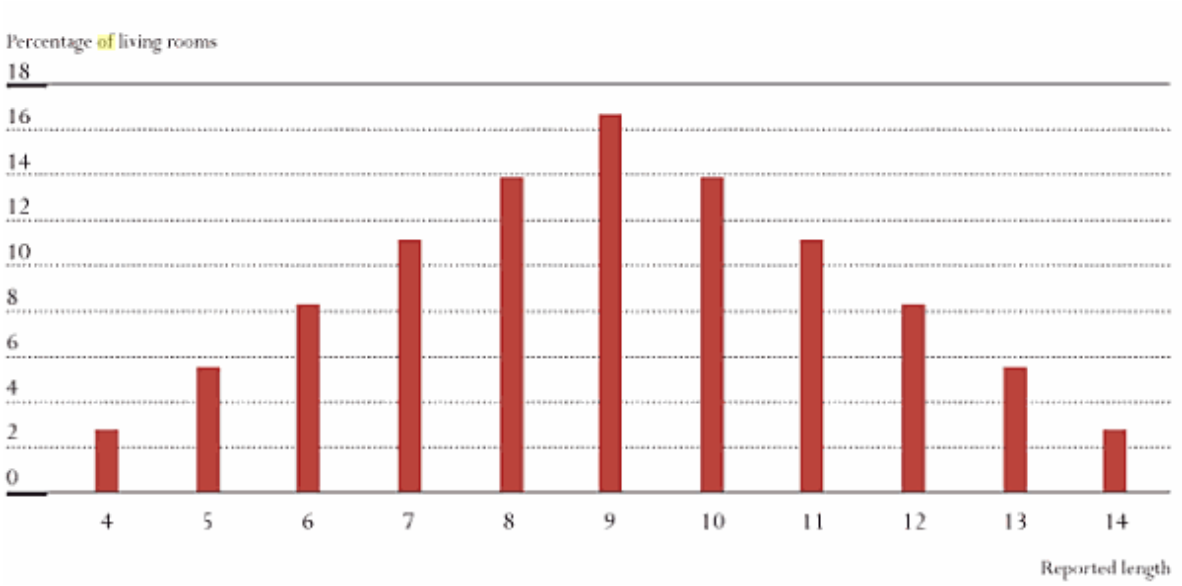

Figura 2.17 - Distribución de los salones reportados [Fuente: PISA Data Analysis Manual: SPSS and SAS, Second Edition]

Por supuesto, la realidad es bastante diferente, debido a que la longitud es una variable continua. Con una variable continua, las observaciones pueden tomar cualquier valor entre el máximo y el mínimo. Por otro lado, con una variable discontinua, las observaciones pueden solamente tomar un número predefinido de valores. La figura 2.18 muestra la distribución de la longitud de los salones anotada.

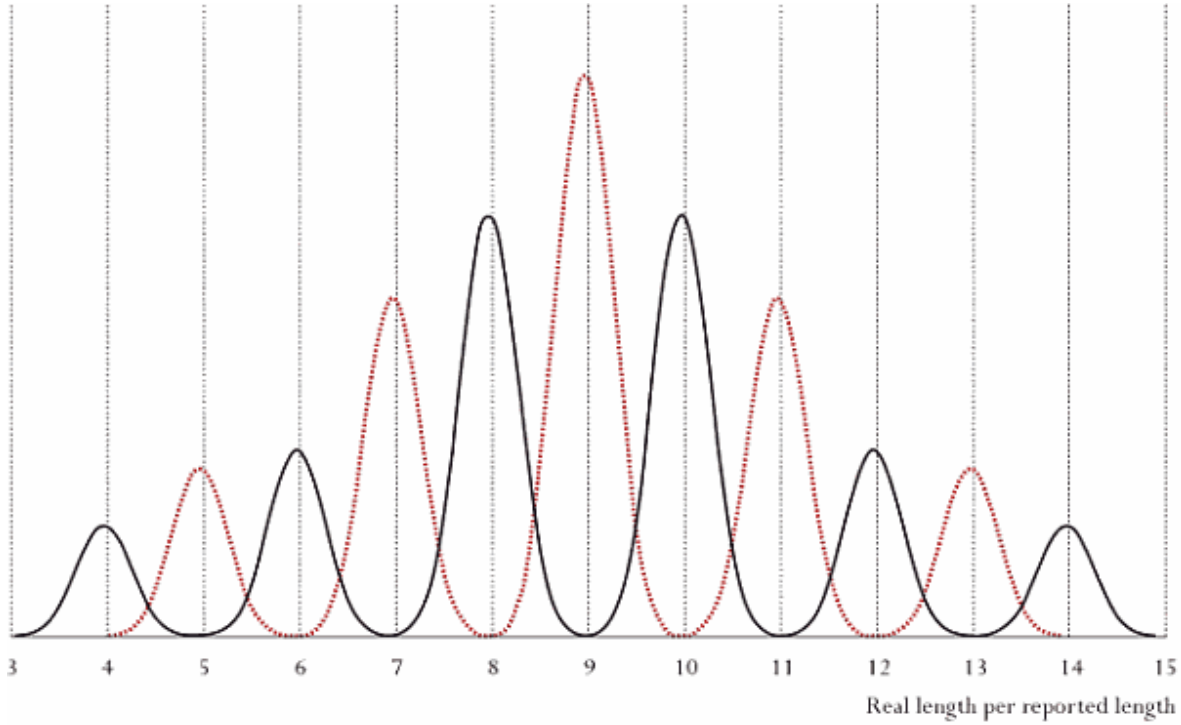

Figura 2.18 - Distribución real de la longitud reportada [Fuente: PISA Data Analysis Manual: SPSS and SAS, Second Edition]

Todos los salones con una longitud registrada de 5 metros no son exactamente 5 metros de largo. En promedio, son 5 metros de largo, pero su longitud varia alrededor de la media. La diferencia entre la longitud anotada y la real se debe al proceso de redondeo y a errores de medición. Por ejemplo un inspector puede anotar incorrectamente 5 metros de longitud para un salón en particular, cuando en realidad mide 4.15 metros. Si el error de redondeo fuera la única fuente del error entonces la longitud registrada seria 4 metros. Sin embargo la segunda fuente de error, el error al medir, explica el solapamiento de la distribución.

En este ejemplo en particular, las longitudes de los salones están distribuidas normalmente alrededor de la media, que es también la longitud registrada. Si la diferencia entre la longitud y el entero más cercano es pequeña, entonces la probabilidad de no anotar esta longitud con el entero más cercano es muy baja. Por ejemplo es muy poco probable que una longitud de 4.15 metros sea registrada como 5 metros o 3 metros. Sin embargo, mientras la distancia entre la

longitud real y el entero más cercano aumenta, la probabilidad de no registrar esta longitud con el entero más cercano también aumenta. Por ejemplo, es muy probable que una longitud de 4.95 metros sea registrada como 5 metros, por otro lado una longitud de 4.50 metros se anotara tantas veces 4 metros como 5 metros.

Por lo tanto la metodología de los valores plausibles consiste de:

- Calcular matemáticamente las distribuciones alrededor de los valores registrados.
- Asignar a cada valor observado un conjunto de valores aleatorios tomados a partir de las distribuciones posteriores.

Para el ejemplo, un salón de 7.15 metros que fue registrado como 7 metros podría asignársele cualquier valor de la distribución normal alrededor de la longitud registrada de 7 metros. Puede tener 7.45 así como también 6.55 o 6.95 metros. Por lo tanto los valores plausibles no deben ser considerados para estimaciones de nivel individual porque no lo son.

Por ultimo si θ es el estadístico poblacional y θ_i el estadístico de interés calculado sobre un valor plausible, entonces:

$$\theta = \frac{1}{M} * \sum_{i=1}^{M} \theta_I$$

Siendo M el número de valores plausibles.

Resumiendo los valores plausibles nos entregan las siguientes ventajas:

- Algunos estimadores de población están sesgados cuando estimadores puntuales son usados para construir estadísticas a nivel de población.
- Los valores plausibles facilitan el cálculo del error estándar para estimadores de diseños muéstrales complejos.

2.5.2.2 SPSS

La última parte dentro del proceso de análisis de datos contempla examinar los datos e intentar calcular las medidas de interés junto la herramienta SPSS de análisis estadístico. Se calculó el promedio en la prueba de ciencias para el país de Australia en el año 2000, como una forma de explorar la herramienta.

Los pasos para poder calcular el promedio fueron los siguientes:

1. Se cargan los datos de la prueba de ciencias para el año 2000.
2. Luego de cargar los datos, debemos filtrar los casos de manera de calcular solamente la media para el país de Australia. Para ello debemos ir “Data-> Select Cases”.
3. Una vez dentro del menú debemos hacer click en el radio “If condition is satisfied”, y luego hacer click en “If”.
4. Luego se debe elegir en base a que variable filtraremos los datos, para nuestro ejemplo elegiremos la variable COUNTRY con un valor igual a 036 (Codigo de la ISO 3166 para Australia).
5. Una vez creado el subconjunto de datos debemos ir a “Analyze-> Descriptive Stadistics->Descriptive”.
6. Dentro del menú seleccionamos los cinco valores plausibles disponibles para el cálculo de la media. Dentro del menú podemos elegir en “Options” la posibilidad de calcular la desviación estándar, el valor máximo, el valor mínimo, entre otras estadísticas de interés.

El resultado de dicha operación se ilustra en la figura 2.19.

Descriptive Statistics

	N	Mean	Std. Deviation
Plausible value in science	2860	524,9708	98,30587
Plausible value in science	2860	525,0613	97,50829
Plausible value in science	2860	524,3738	97,64468
Plausible value in science	2860	526,7764	97,94183
Plausible value in science	2860	524,8758	98,01555
Valid N (listwise)	2860		

Figura 2.19 - Resultado del análisis mediante SPSS [Fuente: Elaboración propia]

Luego el puntaje promedio en ciencias para Australia el año 2000 está dado por la media de estos cinco valores, que da como resultado 525, 208. La operación debe repetirse para cada una de las pruebas y por cada país si lo que queremos es conocer el puntaje medio de los países para el año 2000 y por tipo de prueba.

Como podemos observar el trabajo para el cálculo de estadísticas a nivel de población es bastante laborioso y a veces algo confuso esto porque para cada nueva estadística debemos realizar una serie de pasos ya antes efectuados. Por otro lado la comparación de una misma estadística resulta compleja debido a que la herramienta no compara estadísticas de forma natural.

Por último la incorporación de nuevas estadísticas de población se complica más a medida que intentamos realizar nuevos cálculos.

Capítulo 3 DESARROLLO E IMPLEMENTACIÓN DE LA SOLUCIÓN.

El proceso de desarrollo de la implementación estará determinado por el desarrollo iterativo de cubos. El desarrollo iterativo tiene como propósito crear cada vez una versión más completa de la implementación. Por ultimo este proceso de desarrollo se completará en tres iteraciones las cuales darán como resultado al final de cada iteración un cubo con un objetivo en particular:

- El primer cubo tendrá como objetivo probar el rendimiento de la herramienta SQL Server 2008 y sus características de diseño de cubos, medidas, medidas calculadas, dimensiones entre otros.
- El segundo cubo tendrá como objetivo aumentar la funcionalidad del cubo, agregando una nueva dimensión y medida. Y analizar la influencia de las características socioeconómicas en el desempeño de los alumnos. Esto a través de la inclusión de una dimensión con estos datos.
- Por último, el tercer cubo tendrá como objetivo determinar la relevancia de la inclusión de la dimensión tiempo dentro del análisis de los datos, así como el nivel de escolaridad alcanzado por los padres.

3.1 Metodología de Diseño.

El diseño de cada cubo estará definido por una metodología en común la cual nos permitirá definir los procedimientos para alcanzar nuestros objetivos. A continuación se describen cada uno de los procesos necesarios para implementar una Base de Datos Multidimensional (cubo), los que según C. Zambrano y D. Rojas son[20]:

[20] C. Zambrano, D. Rojas, "Data Warehouse para analizar el comportamiento académico", XXIV Congreso Chileno de Educación en Ingeniería, 2010.

- Proceso de Modelado Conceptual: Este proceso permite capturar los requerimientos de información necesarios para poder generar los indicadores de gestión. El esquema resultante, que contempla las dimensiones, medidas y relaciones multidimensionales, es independiente del motor utilizado para generar el cubo resultante.
- Proceso de Modelado Lógico y Físico: Este proceso tiene como entrada un esquema conceptual multidimensional y genera un esquema lógico y físico, la principal dificultad de este proceso radica en generar un modelo lógico que satisfaga no solo los requerimientos funcionales de información sino también las restricciones.
- Proceso de ETL: Este proceso considerado uno de los más largos dentro del desarrollo, consta básicamente de extraer los datos desde los sistemas fuentes, transformarlos y posteriormente cargarlos en la Base de Datos Multidimensional.
- Proceso de Análisis ROLAP: Proceso mediante el cual los usuarios exploran la información mediante las distintas operaciones ROLAP.

3.2 Implementación del Cubo.

Este apartado del proceso de desarrollo tiene como objetivo documentar los procesos necesarios para la implementación. El proceso de desarrollo documentado que se muestra a continuación corresponde al proceso de desarrollo para el Cubo Nº3.

Como primera etapa del proceso de desarrollo se diseña el esquema conceptual el cual presenta las dimensiones, medidas y relaciones multidimensionales como se muestra en la figura 3.1. El modelo conceptual utilizado para desarrollar el esquema resultante se basa en el modelo CMDM de Carpani.

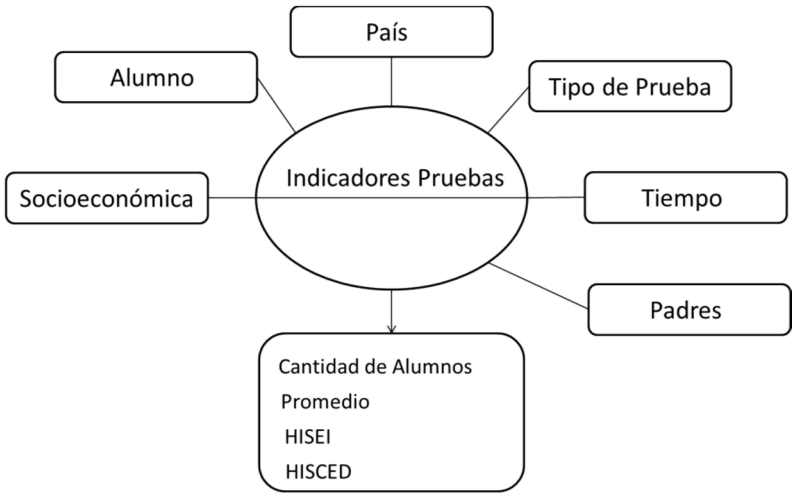

Figura 3.1 - Esquema CMDM de la implementación [Fuente: Elaboración propia]

El esquema posee 6 dimensiones, que serán las diferentes perspectivas desde la cuales analizaremos la información:

- Alumno: Contiene los datos de los alumnos como lo son año de nacimiento, sexo, entre otros datos.
- País: Contempla los países de los cuales provienen los alumnos.
- Tipo de Prueba: Describe las pruebas que rinden los alumnos.
- Socioeconómica: Describe el nivel socioeconómico al cual un alumno pertenece.
- Tiempo: Posee las fechas de las pruebas.
- Padres: Esta dimensión contiene información acerca de los niveles de escolaridad alcanzados por los padres de los alumnos que rinden las pruebas.

Por otro lado las medidas resultantes como cantidad de alumnos y promedio pueden agregarse o desagregarse a través de las dimensiones.

Luego de haber diseñado el esquema conceptual multidimensional para la implementación, la siguiente etapa dentro del proceso de desarrollo consta del diseño del esquema lógico de la implementación. Para ello se ha utilizado el

esquema en estrella de forma de simplificar las uniones entre las dimensiones y la tabla de hecho. Cabe destacar que el esquema, corresponde a un esquema desnormalizado. La figura 3.2 ilustra el esquema multidimensional lógico de la implementación

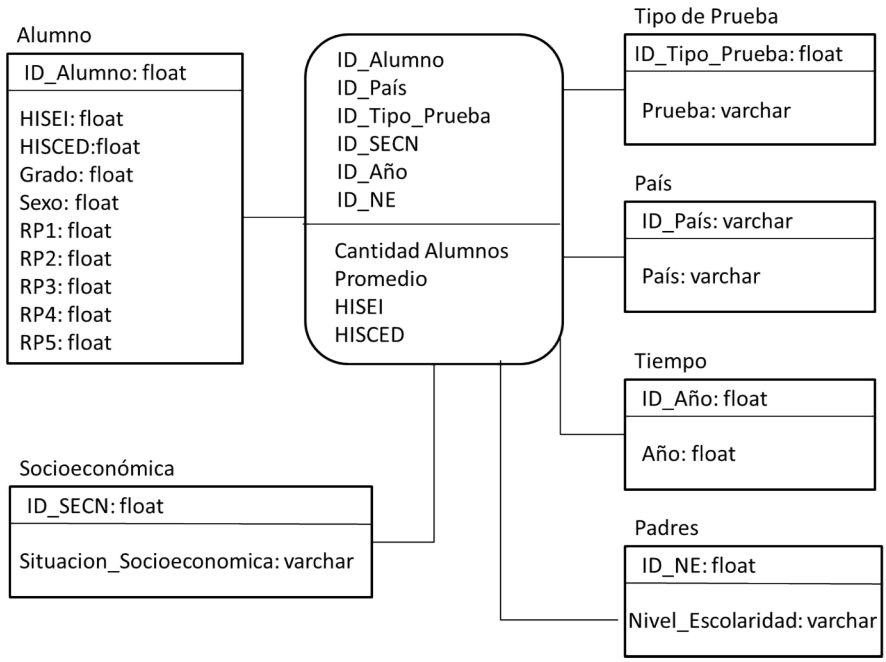

Figura 3.2 - Esquema Lógico Estrella de la implementación [Fuente: Elaboración propia]

Dentro de la etapa del proceso de ETL, la implementación considera los siguientes pasos como podemos ver en la figura 3.3:

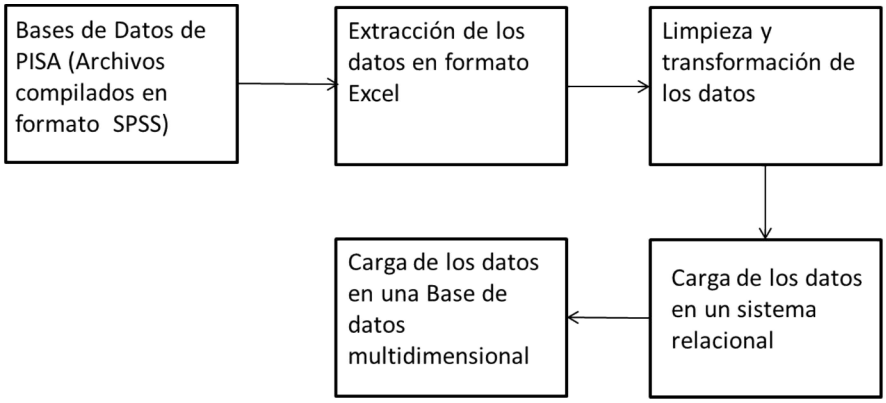

Figura 3.3 - Proceso de ETL [Fuente: Elaboración Propia]

Sin embargo cabe destacar que no existe un sistema transaccional propiamente tal para los datos de PISA, esto porque los datos se encuentran en archivos de texto que deben ser compilados mediante el software SPSS de análisis estadístico que permite analizar y visualizar la información de forma tabulada (1).

Por lo tanto para poder obtener los datos fuentes se realizaron los siguientes pasos:

1. Descarga de los archivos fuentes: Lo primero que debemos hacer es descargar los datos necesarios desde la página de PISA de la OECD desde la cual podemos conseguir los resultados de los estudios desde el año 2000 en adelante. Cada estudio posee dos tipos de archivos, uno que contiene los datos propiamente tal del estudio en formato txt (sin ningún tipo de formato) y otro en formato spss que permite compilar y ordenar los datos del estudio aplicando etiquetas, formato y especificaciones para datos erróneos o perdidos.
2. Modificar archivo de compilación: Una vez descargados los archivos lo siguiente es modificar cada uno de los archivos de compilación. Primero modificamos al principio de cada archivo la ruta en la cual se encuentra nuestro archivo txt con los datos, y luego al final del archivo especificando la ruta del archivo de salida generando un nuevo archivo en formato sav.
3. Compilación de los datos: Por ultimo debemos ejecutar el software estadístico SPSS y compilar el archivo siguiendo los siguientes pasos:

    a. Ir a File->Open->Syntax
    b. Buscar el archivo .spss y abrirlo.
    c. Finalmente en el menú desplegado ejecutar Run.

Los pasos antes descritos permiten solamente visualizar los datos de PISA de forma tabulada y etiquetada con nombres distintivos para cada columna así como también una pequeña descripción de la misma.

Una vez realizados los pasos anteriores, el siguiente paso dentro del proceso ETL consta de la extracción de los datos propiamente tal (2). Primero guardamos los datos de interés en formato Excel. Para ello solo basta ir al menú del software SPSS hacer click en File->Save As. El programa permite guardar los datos en diversos formatos y además seleccionar aquellos datos que son de interés del usuario haciendo click en “variables” del cual se desplegara otro submenú del cual podremos elegir los datos que más nos interesen.

Siguiendo con el proceso ETL, la siguiente etapa contempla la transformación y limpieza de los datos (3). En esta etapa podemos destacar la transformación de los puntajes de las pruebas esto debido a que por defecto el archivo Excel generado por la herramienta establece los decimales con el carácter coma siendo interpretado esto en la base de datos como miles. Por otro lado podemos destacar en la limpieza de los datos aquellos puntajes que aparecían en negativo.

Una vez extraídos los datos fuentes y transformados y limpiados en unidades compatibles se cargan dichos datos en una base de datos relacional diseñada específicamente para el proceso de desarrollo (4). La figura 3.4 muestra la base de datos relacional diseñada para la implementación a través de la herramienta SQL Server Management Studio 2008.

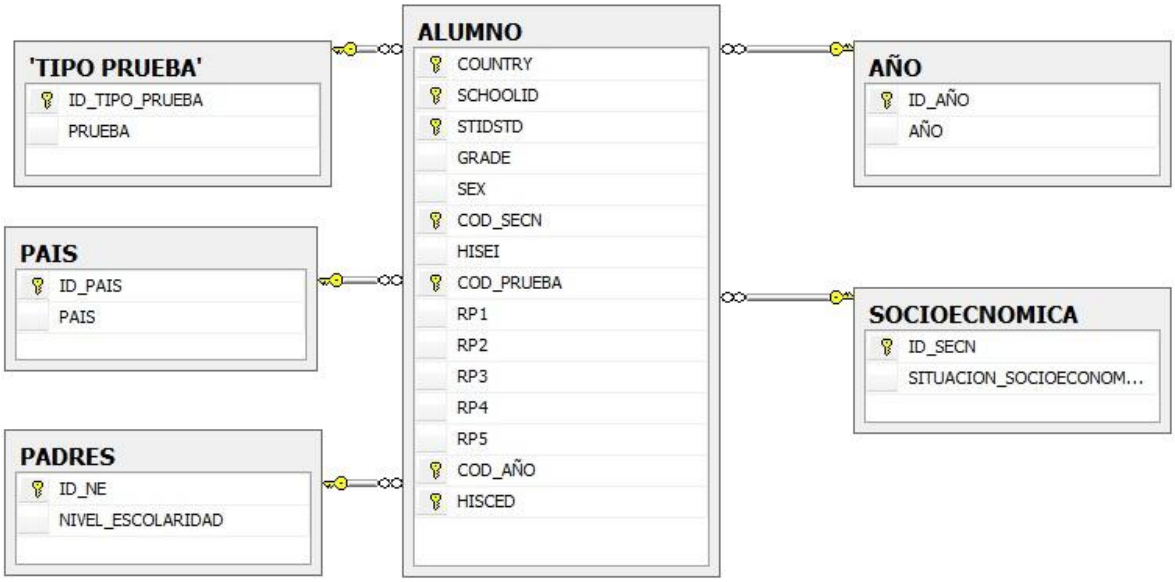

Figura 3.4 - Base de Datos relacional de la implementación [Fuente: Elaboración propia]

Por último los datos son cargados en la base de datos multidimensional diseñada en las etapas anteriores (5).

3.3 Resultados de la Implementación.

La última etapa dentro del proceso de desarrollo consta de la exploración la información mediante las distintas operaciones ROLAP. En esta etapa los usuarios de forma casi intuitiva exploran la información buscando tendencias y patrones que les resultan de interés.

A continuación se muestran una serie de reportes generados a través de la implementación. Los reportes se muestran agrupados por cubos como una forma de mostrar el grado de avance que supone uno con respecto al otro.

3.3.1 Reportes Cubo N°1

El grafico 3.1 muestra los puntajes promedios para Chile para cada una de las pruebas en el año 2000. Como se puede apreciar la prueba de lenguaje es la prueba con el promedio más bajo y ciencias el promedio más alto.

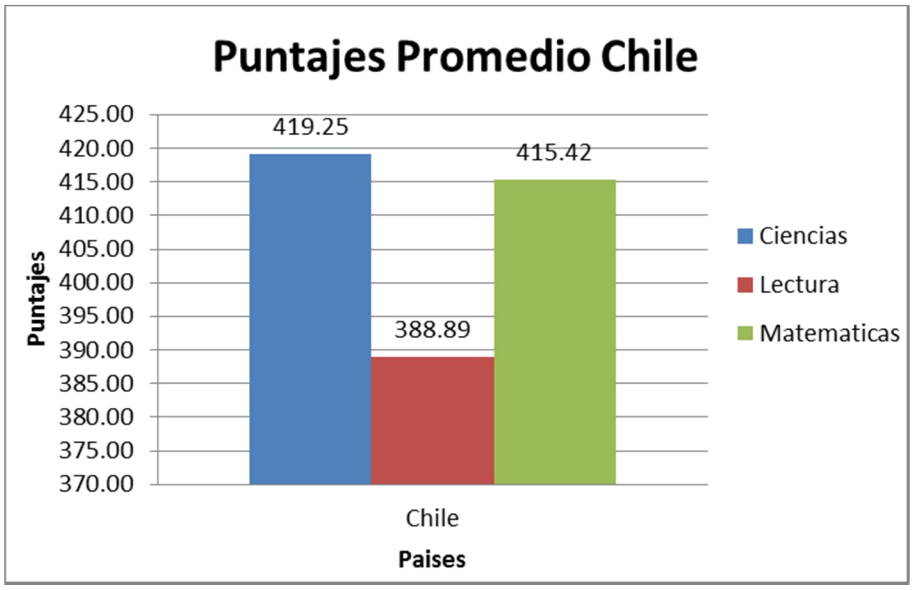

Gráfico 3.1 - Puntajes Promedio Chile [Fuente: Elaboración propia]

El grafico 3.2 muestra los distintos puntajes promedios para los países latinoamericanos participantes del estudio en el año 2000. Como se observa Chile y Argentina lideran los puntajes a nivel latinoamericano sobrepasando la barrera de los 400 puntos. Además existe una diferencia de más de 30 puntos para cada una de las pruebas entre Chile y Brasil.

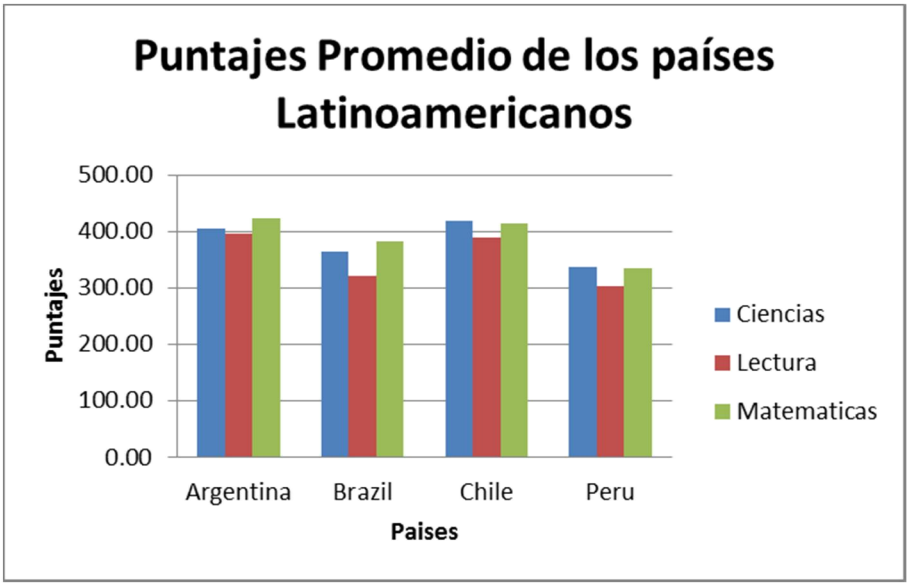

Gráfico 3.2 - Puntajes Promedio Países Latinoamericanos [Fuente: Elaboración propia]

El grafico 3.3 muestra los países con los mejores puntajes promedios del estudio para el año 2000. Destacamos los puntajes alcanzados por Los Países Bajos, Japón y Hong Kong quienes superan los 550 puntos promedio en la prueba de Lenguaje y los 540 puntos en Ciencias. También se observa que los únicos países miembros de la OECD en esta grafica son Canadá, Inglaterra y Australia.

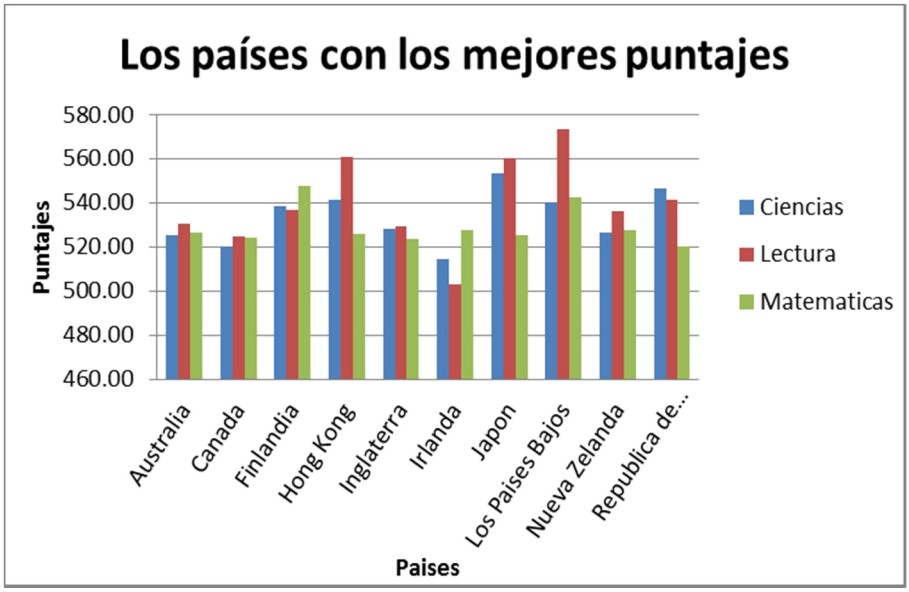

Gráfico 3.3 - Países mejores evaluados [Fuente: Elaboración propia]

3.3.2 Reportes Cubo N°2

El grafico 3.4 muestra los puntajes promedios agrupados por nivel socioeconómico para Chile en el año 2000. La grafica muestra que a medida que el nivel socioeconómico del alumno aumenta también lo hace su rendimiento en la prueba, llegando a existir una diferencia de más de 100 puntos promedio entre los niveles socioeconómicos más bajos y los más altos.

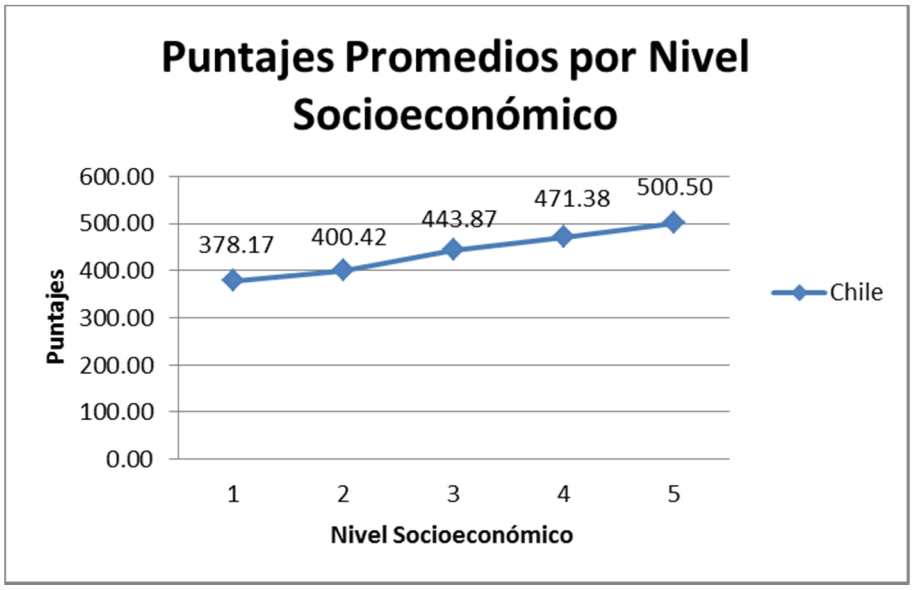

Gráfico 3.4 - Puntajes Promedios por Nivel Socioeconómico Chile [Fuente: Elaboración propia]

El grafico 3.5 muestra los puntajes promedios agrupados por nivel socioeconómico para todos los países latinoamericanos participantes del estudio en el año 2000. Como se observa la tendencia al igual que en el grafico anterior se repite para cada uno de los países participantes. Es decir mientras más alto sea el nivel socioeconómico más alto es el puntaje alcanzado por el alumno.

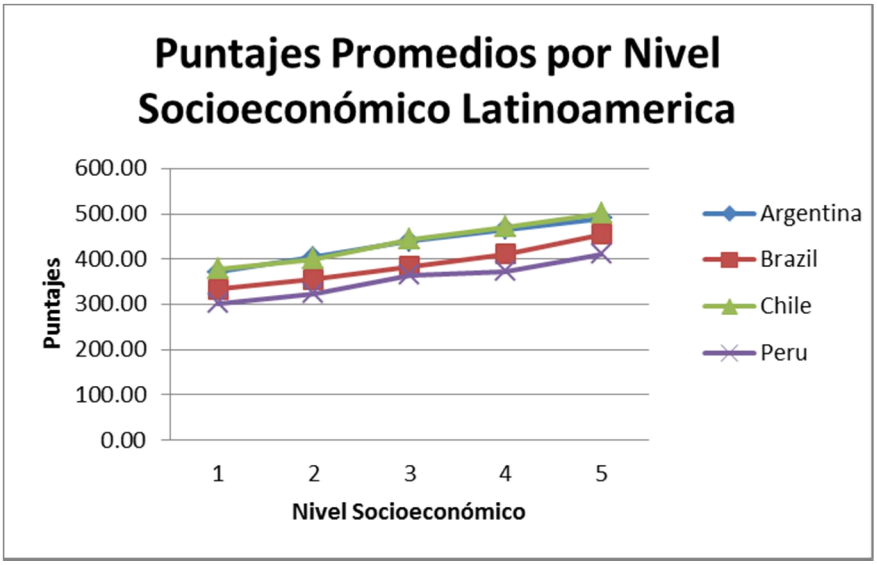

Gráfico 3.5 - Puntajes Promedio por Nivel Socioeconómico Latinoamérica [Fuente: Elaboración propia]

El grafico 3.6 muestra los puntajes promedios por género para las distintas pruebas en el año 2000 en Chile. Las mujeres coloreadas de color azul y los hombres de rojo, muestran solo diferencias significativas en la prueba de matemáticas llegando a una diferencia de 24 puntos.

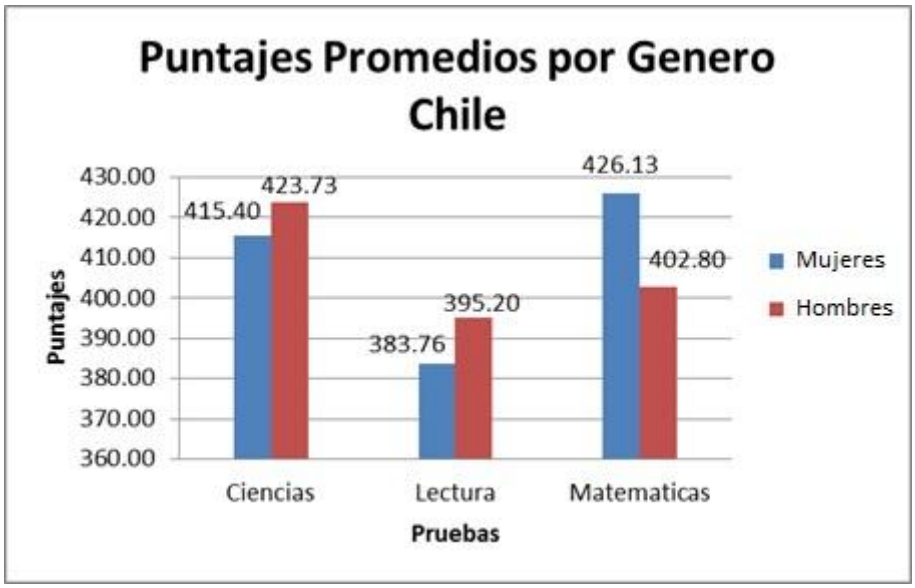

Gráfico 3.6 - Puntajes Promedio por Genero Chile [Fuente: Elaboración propia]

3.3.3 Reportes Cubo N°3.

El grafico 3.7 muestra la evolución de los puntajes promedios para los países participantes de Latinoamérica desde el año 2000 al 2009. La tendencia muestra un crecimiento sostenido en Chile y Brasil logrando una mejora por sobre los 30 puntos cada uno.

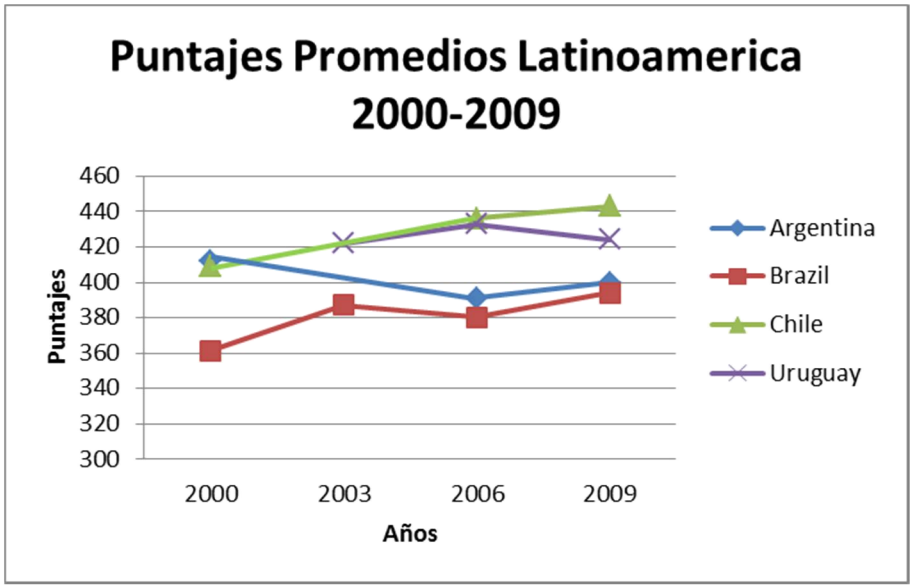

Gráfico 3.7 - Puntajes Promedio Latinoamérica 2000-2009 [Fuente: Elaboración propia]

El grafico 3.8 muestra los puntajes promedios agrupados por nivel socioeconómico en Chile desde el año 2000 al 2009. Como se observa la curva del año 2000 es similar a las del año 2006 y 2009 excepto que estas últimas se encuentras desplazadas más arriba. Esto se debe a que los puntajes promedios para todos los niveles socioeconómicos suben, sin embargo sigue existiendo una diferencia de más de 100 puntos entre los niveles socioeconómicos más bajos y los más altos.

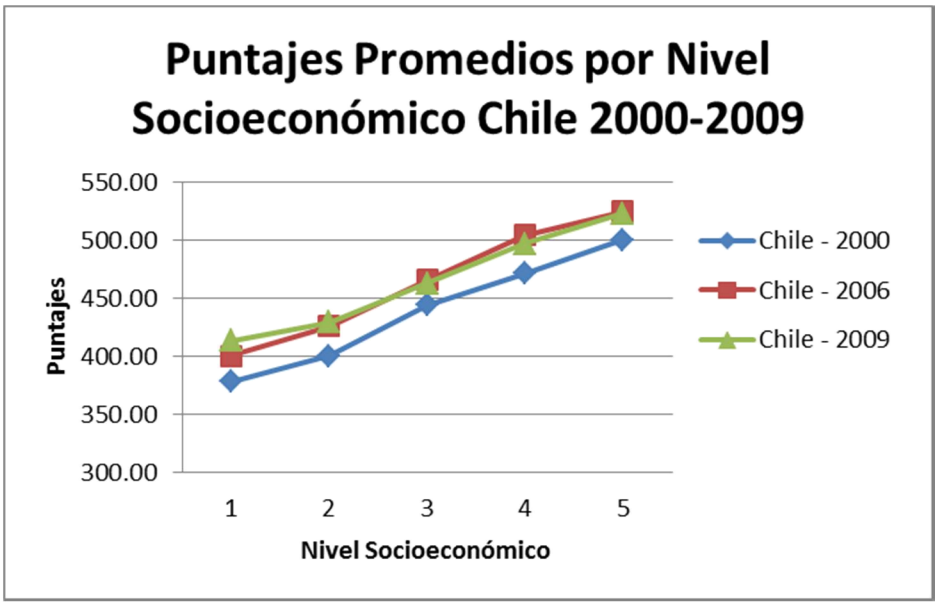

Gráfico 3.8 - Puntajes Promedio por Nivel Socioeconómico Chile 2000-2009 [Fuente: Elaboración propia]

La grafica 3.9 muestra los puntajes promedios agrupados por el nivel de escolaridad alcanzado por los padres de los alumnos. Los niveles son:

a. Nivel 0: Sin Educación.
b. Nivel 1: Primaria.
c. Nivel 2: Primer ciclo de Secundaria.
d. Nivel 3: Secundaria Humanista.
e. Nivel 4: Secundaria Nivel Técnico.
f. Nivel 5: Pregrado.
g. Nivel 6: Postgrado.

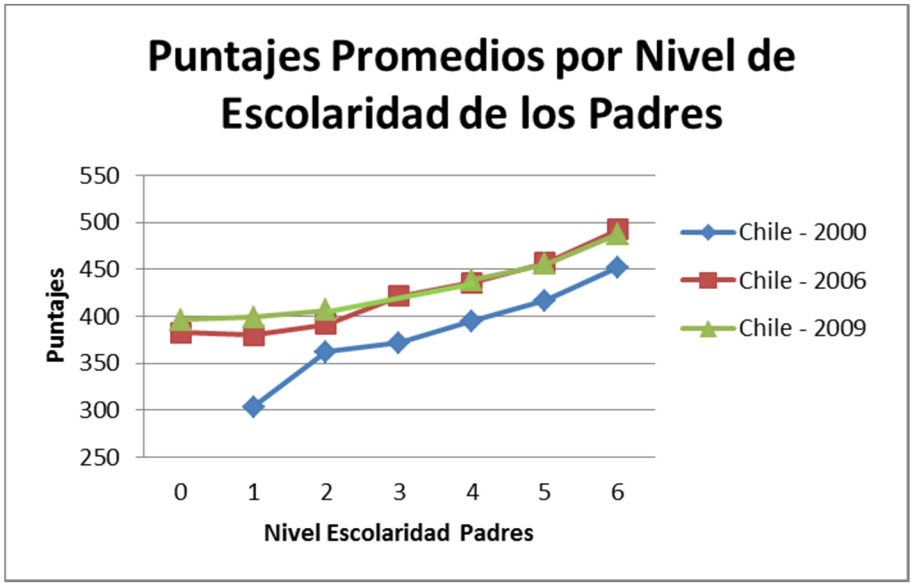

Gráfico 3.9 - Puntajes Promedio por Nivel de Escolaridad de los Padres Chile [Fuente: Elaboración propia]

El grafico 3.10 muestra la evolución de los puntajes promedios de las distintas pruebas desde el año 2000 al 2009 en Chile. Como se observa la prueba de lenguaje ha sido aquella que presenta el mejor progreso con una diferencia de 60 puntos entre el 2000 y el 2009, seguida de ciencias con una mejora de 30 puntos.

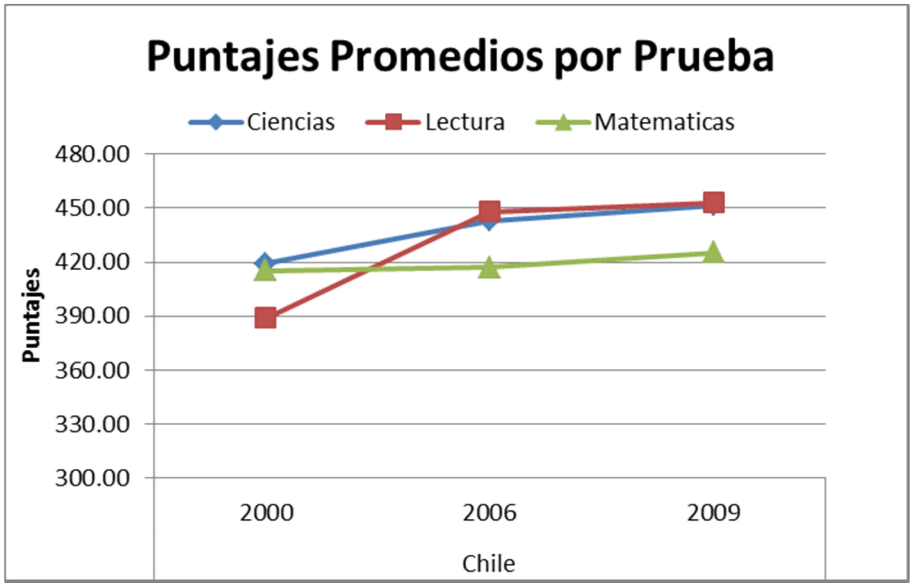

Gráfico 3.10 - Puntajes Promedio por Prueba 2000-2009 Chile [Fuente: Elaboración propia]

El grafico 3.11 muestra la evolucion de los puntajes promedios para la prueba de lenguaje por genero entre los años 2000 y 2009 en Chile. Las mujeres colareadas de color azul y los hombres de rojo. Las mujeres muestran una clara tendencia a la suba en esta prueba, mientras que los hombres mantienen sus resultados entre el 2006 y el 2009.

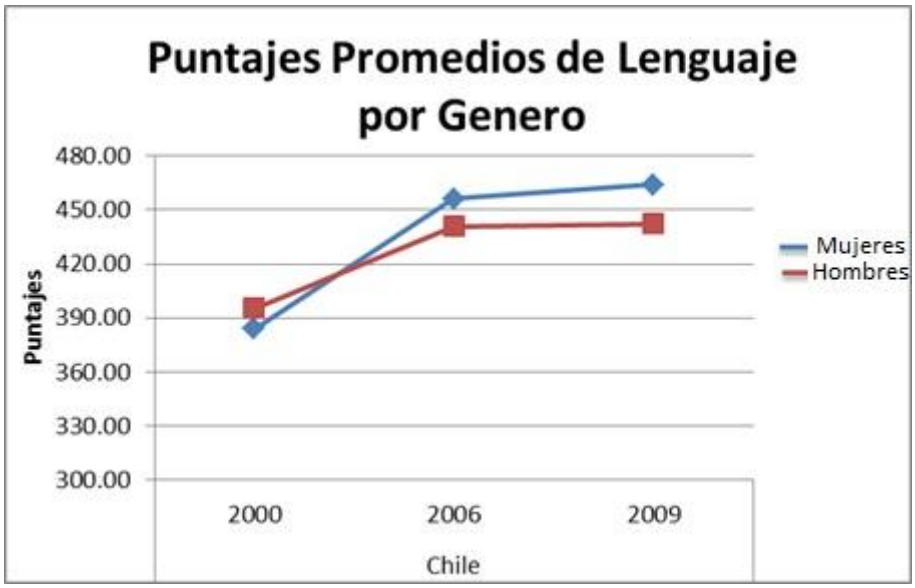

Gráfico 3.11 - Puntajes Promedio Lenguaje por Genero [Elaboración propia]

El grafico 3.12 muestra la evolucion de los puntajes promedios para la prueba de ciencias por genero entre los años 2000 y 2009 en Chile. Las mujeres colareadas de color azul y los hombres de rojo. Tanto mujeres como hombres muestran una tendencia a la suba, sin embargo los hombres en promedio obtienen mejores resultados para esta prueba.

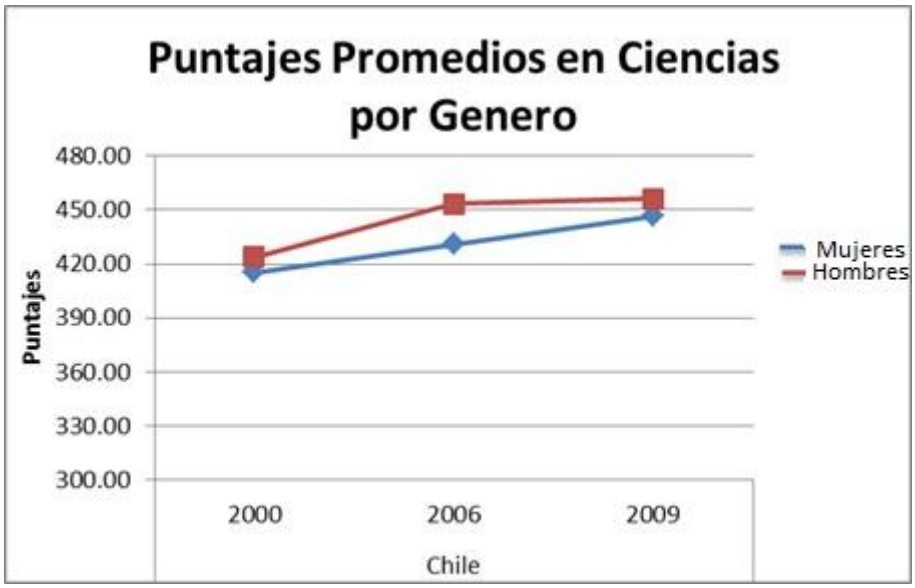

Gráfico 3.12 - Puntajes Promedio Ciencias por Genero Chile [Fuente: Elaboración propia]

El grafico 3.13 muestra la evolucion de los puntajes promedios para la prueba de matematicas por genero entre los años 2000 y 2009 en Chile. Las mujeres colareadas de color azul y los hombres de rojo. En esta prueba los puntajes de los hombres muestra una clara tendencia a la suba, mientras que las mujeres tienen un comportamiento mas anómalo.

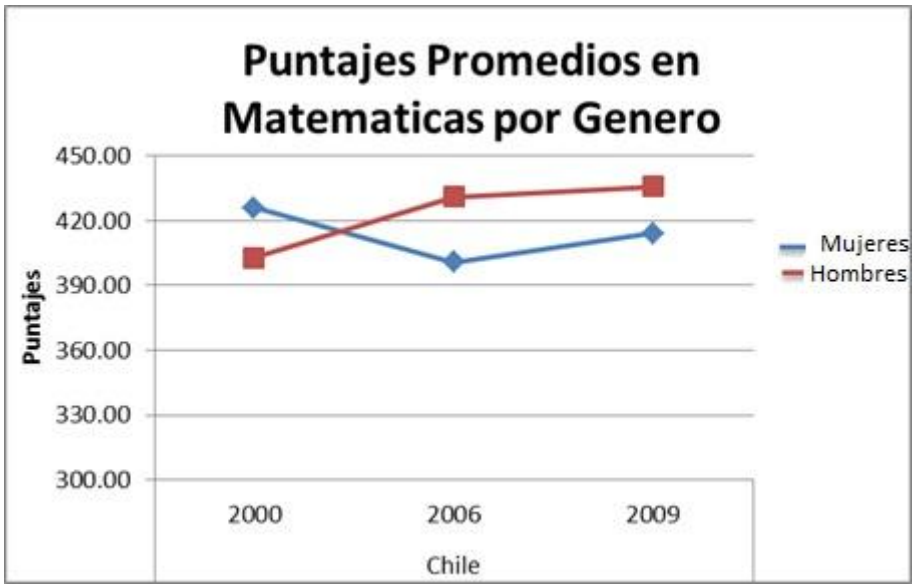

Gráfico 3.13 - Puntajes Promedio Matemáticas por Genero Chile [Fuente: Elaboración propia]

El grafico 3.14 muestra los puntajes promedios de Chile y algunos países miembros de la OECD. Como se observa algunos países miembros de la OECD tienen una clara tendencia a la baja, mientras que otros mantienen niveles de rendimiento. En el 2009 Chile se encuentra a solamente a 30 puntos de los países miembros de la OECD, siendo uno de los países con los mejores progresos académicos.

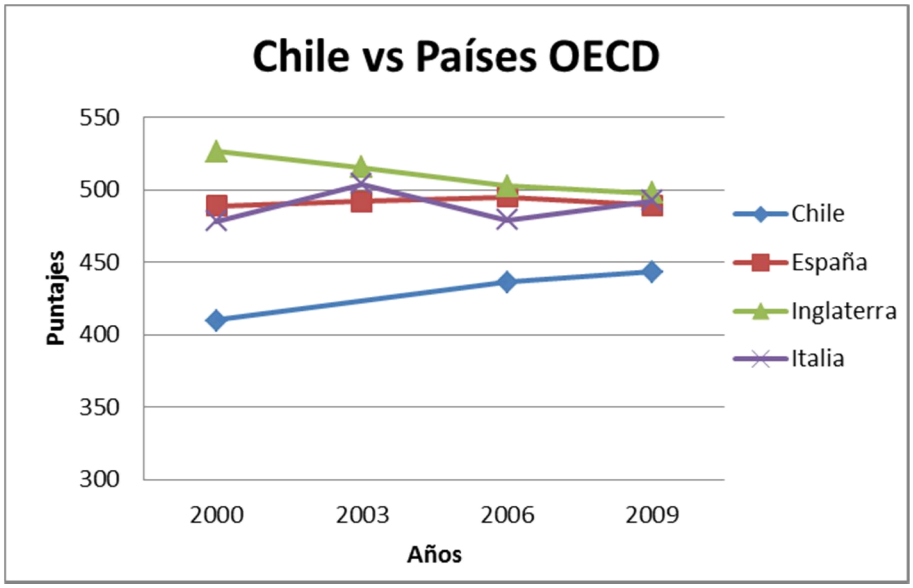

Gráfico 3.14 - Puntaje Promedio Chele vs OECD [Fuente: Elaboración propia]

El grafico 3.15 muestra los índices socioeconómicos promedios para los países participantes de Latinoamérica entre el año 2000 y 2009. Se observa que Chile es el país con los índices promedios más bajos, con una pequeña tendencia a la suba.

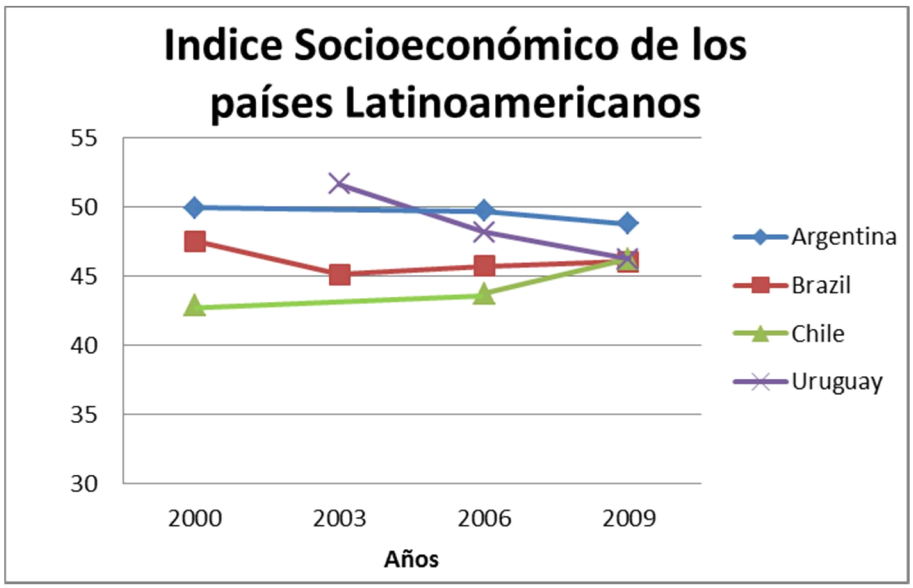

Gráfico 3.15 - Índice Socioeconómico de los países latinoamericanos

El grafico 3.16 muestra los niveles de escolaridad promedios para los países participantes de Latinoamérica entre el año 2000 y 2009. Se observa que Chile es el país con el promedio de escolaridad más alto, seguido de Argentina. Los niveles de escolaridad son:

a. Nivel 0: Sin Educación.
b. Nivel 1: Primaria.
c. Nivel 2: Primer ciclo de Secundaria.
d. Nivel 3: Secundaria Humanista.
e. Nivel 4: Secundaria Nivel Técnico.
f. Nivel 5: Pregrado.
g. Nivel 6: Postgrado.

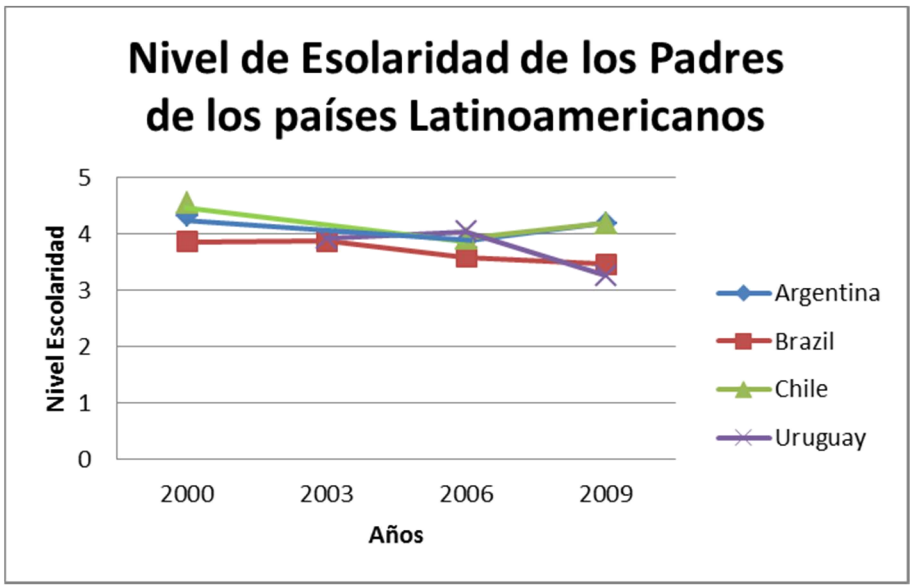

Gráfico 3.16 - Nivel de Escolaridad de los Padres Latinoamérica [Fuente: Elaboración propia]

Capítulo 4 CONCLUSIONES

4.1 Del Trabajo de Titulación

En el presente trabajo de titulación, se ha diseñado e implementado una base de datos multidimensional para datos educacionales con el objetivo de mostrar la utilidad de la técnica y la metodología de bases de datos multidimensionales. En el proceso de Inteligencia de negocios la toma decisiones es un punto clave dentro del que hacer de las organizaciones, esto porque delimita los planes de acción a seguir para cumplir los objetivos de la organización, así como también la utilización de los recursos necesarios para que dichos planes puedan ser ejecutados.

Los resultados de la implementación muestran que Chile es el país con los mejores puntajes promedios de Latinoamérica, y que junto con Brasil logran las mejoras mas significativas entre los años 2000 y 2009. También se muestra una relación directa entre el nivel socioeconómico de los alumnos y su puntaje alcanzado, esto es a medida que aumenta el nivel socioeconómico del alumno también lo hace su puntaje promedio. Una tendencia que existe en todos los países de Latinoamérica. Por otro lado los resultados mostraron que algunos países de desarrollados han paulatinamente disminuido sus puntajes promedios en las diferentes pruebas, a diferencia de Chile que ha significativamente mejorado sus puntajes.

Un punto clave dentro del proceso de desarrollo e implementación fue el análisis de los datos proporcionados por el estudio de PISA, que permitió la generación de aquellas medidas de mayor interés. Cabe destacar que en esta etapa también se descartaron otras posibles medidas de interés por no tener la continuidad necesaria dentro de todas las versiones del estudio, es decir no habían datos.

El desarrollo iterativo permitió entre otras cosas ir ajustando el proceso de desarrollo a las capacidades técnicas operativas a las cuales se estaba sujeto. Por otro lado permitió también generar una retroalimentación de la información que se

iba generando dando pautas, de cuáles eran los tipos de datos que podían aportar más al análisis de la información. Y finalmente ir desarrollando y perfeccionando los esquemas multidimensionales lógico y conceptuales.

Por otro lado la reportabilidad generada supero las expectativas. Eso se debió principalmente a la incorporación de la dimensión tiempo que permitió darle toda una nueva arista o perspectiva a la información disponible.

Finalmente se dan por cumplidos los objetivos planteados al principio de este trabajo de título dado que la implementación se ha desarrollado en su totalidad cumpliendo el objetivo general así como los específicos.

4.2 De la Experiencia Personal

A través del siguiente trabajo el autor se ha dado cuenta del valor de los datos que residen en los sistemas transaccionales. Esto porque los datos son una fuente inexplorada de información de las cuales se pueden generar ventajas competitivas.

Por otro lado, desde la perspectiva de la implementación, trabajar bajo una metodología permitió que el desarrollo de la implementación finalizara dentro de los tiempos establecidos. Además el autor se ha dado cuenta que no es necesario tener todo el conocimiento técnico sobre una tecnología o área especifica de la informática para desarrollar alguna aplicación o software, ya que solo bastan las ganas y la disciplina para poder lograrlo.

Por ultimo el autor ha podido incrementar mas aun los conocimientos en bases de datos, además de aprender una nueva técnica y metodología de desarrollo que actualmente es muy utilizada por las grandes empresas.

4.3 De los Trabajos Futuros

El siguiente paso de este trabajo de titulación es el desarrollo de una versión mas eficiente de la implementación. Esto es porque durante todo el desarrollo de este trabajo los objetivos han estado centrados en aspectos de

diseño y funcionalidad. La eficiencia ha sido un aspecto secundario, y queda ahora como una primera línea de continuación. Por esta razón se considera apropiado desarrollar la implementación con un esquema lógico de copo de nieve.

Referencias Bibliográficas

Libros.

1. Carpani, F. "*CMDM: un modelo conceptual para la especificación de BDM"*, 2000.
2. Inmon, B. "*Building the datawarehouse"*, 1992.
3. Kimball, R. "*The datawarehouse Toolkit"*, 1996.

Sitios Web.

4. Articulo *Información*. Obtenido de http://www.information-management.com. Blumberg, R. Ultima consulta Septiembre de 2011.
5. Articulo Información. Obtenido de http://businessintelligence.ittoolbox.com/. Lokken, B. Ultima consulta Noviembre 2011.
6. Articulo *Información*. Obtenido de http://www.informationweek.com/whitepaper/. Microstrategy. Ultima consulta Septiembre 2011.
7. Articulo *Información*. Obtenido de http://users.dsic.upv.es/~jorallo/cursoDWDM/dwdm-I.pdf. Orallo, J. H. Ultima consulta Noviembre 2011.
8. Articulo *Información*. Obtenido de http://www.terry.uga.edu/~hwatson/dw_tutorial.ppt. Watson, H. J. Ultima consulta Septiembre 2011.
9. Articulo "*Business Intelligence Tools: Perspective*". Obtenido de www.gartner.com. Tiedrich, A. H. Ultima consulta Septiembre 2011.
10. Articulo "*Definición de Business Intelligence*". Obtenido de www.gartner.com. Gartner Group. Ultima consulta Noviembre 2011.
11. Articulo "*Definición de Business Intelligence*". Obtenido de www.gartner.com. Dresner, H. Ultima consulta Octubre 2011.
12. Articulo "*Definición de Business Intelligence*". Obtenido de www.ibm.com. IBM. Ultima consulta Octubre 2011.
13. Articulo "*Definición de Business Intelligence*". Obtenido de www.wikipedia.com. Wikipedia. Ultima consulta Octubre 2011.
14. Articulo "Definición de Business Intelligence". Obtenido de www.oracle.com. Oracle Group. Ultima consulta Octubre 2011.

Artículos y otras referencias.

15. Kamble, A. "*Modelo CGMD: un modelo conceptual multidimensional basado en MER*", 2008
16. Luhn, H. "*A Business Intelligence System*". IBM Research, 1958.
17. Zambrano, Carolina., & Rojas, Dario. "*Data Warehouse para analizar el comportamiento académico*", 2010.
18. Strange, K. "*The Challenges of Implementing a datawarehouse to Achieve Business Agility*", 2001.