



Universidad de La Frontera

Facultad de Ingeniería, Ciencias y Administración

Departamento de Ingeniería de Sistemas.

**“Desarrollo de Piloto de Almacén de Datos
para Análisis de Costos por Carrera anuales
de La Universidad de La Frontera”**

Profesor Guía: Srta. Alicia Josefa Castro Parra

CARLOS SEBASTIÁN HENRÍQUEZ LARA

2010

II

H519 d

2010

C-1



Universidad de La Frontera

Facultad de Ingeniería, Ciencias y Administración

Departamento de Ingeniería de Sistemas.

“Desarrollo de Piloto de Almacén de Datos para Análisis de Costos por Carrera anuales de La Universidad de La Frontera”

**TRABAJO PARA OPTAR AL TITULO
DE INGENIERO INFORMÁTICO**

87222

Profesor Guía: Srta. Alicia Josefa Castro Parra



CARLOS SEBASTIÁN HENRÍQUEZ LARA

2010

**DESARROLLO DE PILOTO DE ALMACÉN DE DATOS PARA ANÁLISIS DE
COSTOS POR CARRERA ANUALES DE LA UNIVERSIDAD DE LA FRONTERA**

CARLOS SEBASTIÁN HENRÍQUEZ LARA

COMISIÓN EXAMINADORA



ALICIA JOSEFA CASTRO PARRA

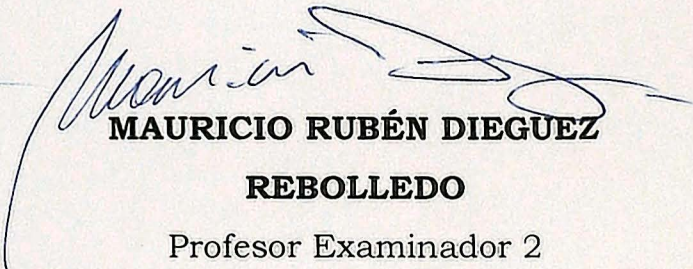
Profesor guía



RUTH ELISABETH NOVOA

TROQUIÁN

Profesor Examinador 1



MAURICIO RUBÉN DIEGUEZ

REBOLLEDO

Profesor Examinador 2

Nota Trabajo Escrito

: 7.0

Nota Examen

: 6.8

Nota Final

: 7.0

Agenda de la semana

Este fin de semana y agosto es un tiempo muy bueno para ir a la playa y disfrutar de la vida. Hay muchas cosas que hacer y ver, y es importante que lo aprovechemos. En este momento, estamos viviendo una época de cambios y es necesario que nos adaptemos a ellos. La vida es un viaje y debemos disfrutar de cada momento. Hay muchas cosas que hacer y ver, y es importante que lo aprovechemos. En este momento, estamos viviendo una época de cambios y es necesario que nos adaptemos a ellos. La vida es un viaje y debemos disfrutar de cada momento.

*"El éxito más hermoso no es aquel que llega como un gato casero domado,
sino aquel que hay que aprender a domar y forzar como un caballo salvaje."*

(Peter Ustinov)

Agradecimientos.

Este fruto de esfuerzo y logros se los dedico y agradezco a todas las personas que estuvieron en todos los momentos, sobre todo en los momentos difíciles, a mi Madre Laura que luchó para que pudiera estudiar y poner todo mi esfuerzo, a mi amor Camila, a mi abuela Elena, a mi hermano Ricardo, a mi padre, a mis amigos y compañeros que me apoyaron. A mi profesora guía que estuvo siempre para apoyarme en este trabajo y que me orientó para obtener los mejores resultados, y a los profesores que se interesaron y me brindaron ayuda cuando la necesité. Muchas gracias por apoyarme para lograr terminar esta etapa con éxito.

Resumen

En este trabajo se plantea como objetivo implementar un almacén de datos piloto para análisis de costos por carrera de la Universidad de La Frontera utilizando tecnología Business Intelligence. Este objetivo se realizó a través de la recopilación información de algún sistema planteado por otras universidades, investigando las herramientas, generando distintas alternativas de modelos de bases de datos, investigando las herramientas de Pentaho OpenSource para trabajar con Business Intelligence, aplicar el conocimiento tanto teórico como práctico de estas herramientas para realizar procesos de extracción, transformación y carga de datos (ETL), consultas multidimensionales, y análisis de estas consultas.

La finalidad de este trabajo es presentar un sistema de análisis de costos por carrera, debido a que en la actualidad no existe un sistema informático en la Universidad que permita determinar los costos asociados a impartir una carrera, orientado a usuarios de niveles de toma de decisiones, como directivos a nivel de la Universidad y de Facultades.

Existe interés institucional en instalar un sistema de este tipo en la Universidad, que permita analizar en detalle los costos asociados a impartir cada una de las carreras. La implementación de un almacén de datos permitiría hacer consultas específicas de manera rápida y eficiente, debido a las múltiples posibilidades de presentación de la información en forma dinámica que las herramientas de inteligencia de negocios ofrecen.

Durante aproximadamente cuatro meses se trabajó en el cumplimiento de estos objetivos para llevar a cabo la implementación del almacén de datos. Se investigó respecto a las herramientas y tecnologías necesarias para la ejecución, se diseñaron propuestas de almacenes de datos que, a través del tiempo y de los avances fueron siendo depuradas, para llegar a un modelo consensuado con la Dirección de Informática para realizar el poblamiento del

almacén y realizar las consultas OLAP para analizar los costos y así considerar exitoso el trabajo de la implementación.

La estructura del almacén de datos no tiene que normalizarse, lo cual permite mayor rapidez en el proceso de consultas OLAP.

En la estructura del modelo de datos, lo principal a considerar como dimensiones son el tiempo, las carreras, los clasificadores de costos y los ítemes de costos, algunos de éstos implican más niveles de detalle como por ejemplo las carreras, ya que pertenecen a unidades de distintos tipos de programas; para las medidas se consideran el costo, el número de alumnos por carrera en un año determinado y, como alternativa a implementar y que como prueba resultó exitoso, agregar como medida el costo por alumno.

El resultado de la implementación se analiza con la plataforma Pentaho BI Server, en el cual se selecciona un esquema con el cubo OLAP, que arroja las distintas dimensiones y sus niveles para cada medida (costo, nro. de alumnos y costo por alumno).

Los objetivos se cumplen de manera exitosa implementando el almacén de datos con consultas multidimensionales que permiten analizar a fondo los costos por carrera anuales de la Universidad de La Frontera.

Se concluye que para la implementación del diseño del modelo de datos, éste debe tener una estructura que abarque todos los datos necesarios para poder crear consultas multidimensionales que entreguen la información requerida por el usuario; además, para las consultas deben procesarse grandes cantidades de datos de manera rápida, por lo que las consultas OLAP son las indicadas para esto ya que permiten hacer análisis de grandes cantidades de datos de manera rápida.

Con Business Intelligence, las compañías, empresas, instituciones u organizaciones pueden analizar tendencias que representan oportunidades nuevas e importantes y así anticipar problemas potenciales y hacer ajustes antes de que se conviertan en un problema.

ÍNDICE

1.	INTRODUCCIÓN.....	2
1.2.	Objetivos Generales y Específicos.	3
1.2.1.	Objetivo General.	3
1.2.2.	Objetivos Específicos.	3
2.	FUNDAMENTOS TEÓRICOS.....	5
2.1.	Introducción a la ayuda a la toma de decisiones.....	5
2.2.	Almacenes de datos.	6
2.2.1.	Data Warehouse.....	6
2.2.2.	Data Mart.....	7
2.2.3.	Componentes de los almacenes de datos.....	8
2.3.	OLTP - On-Line Transactional Processing.....	11
2.4.	OLAP. Online Analytical Processing.	11
2.4.1.	Modelo Multidimensional de datos.	12
2.5.	Esquemas de los almacenes de datos.....	13
2.5.1.	Esquema en estrella.	13
2.6.	Business Intelligence.....	15
2.6.1.	Características de Business Intelligence.....	16
2.6.2.	Arquitectura de una solución de Business Intelligence.....	17
3.	ANTECEDENTES DE LA INSTITUCIÓN.....	21
3.1.	Sistemas de Costos por Carrera.....	21
3.2.	Análisis de distintos Sistemas de Costos por Carrera implementados en la Universidad de La Frontera.....	21
3.2.1.	Sistema Costos por Carrera año 1997.....	21
3.2.2.	Trabajo de Título referente a los Costos por Carrera.....	24
3.3.	Cuadro Comparativo de Estudios de Costos por Carrera.	30
4.	DESCRIPCIÓN DE ACTIVIDADES REALIZADAS.....	33
4.1.	Tecnologías Utilizadas.	33
4.1.1.	JAVA.	33
4.1.2.	Pentaho Open Source Business Intelligence.....	33
4.1.3.	Pentaho BI-Server.....	34
4.1.4.	Pentaho Administrator Console.....	35

4.1.5.	Pentaho Data Integration (KETTLE).	36
4.1.6.	Mondrian Schema-Workbench.	39
4.1.7.	MySQL.	41
4.1.8.	Sybase.	42
4.1.9.	Sybase Power Designer.	42
4.1.10.	Linux.	43
4.2.	Algoritmos de obtención de Costos por Carrera.	44
4.3.	Fuentes de Datos de Origen.	46
4.4.	Modelo del Almacén de Datos.	47
4.5.	Instalación de Pentaho BI Server y Pentaho PAC.	48
4.6.	Creación de usuario para Pentaho y creación de la conexión con el Almacén de Datos.	50
4.7.	Creación de primeras consultas Multidimensionales (CUBOS).	52
4.8.	Análisis de Cubo para modelo de prueba de costos por carrera.	58
4.9.	Modelo de Datos de Traspaso para preparación de Proceso ETL	64
4.10.	Proceso de ETL para el Poblamiento del Almacén de Datos.	64
4.11.	Procedimientos almacenados para poblar el almacén de datos.	69
4.12.	Modelo de Datos en su tercera versión.	70
4.13.	Consultas Multidimensionales del nuevo almacén.	73
5.	RESULTADOS Y DISCUSIÓN.	76
5.1.	Visualización y análisis de Cubos en la Plataforma de Pentaho.	76
5.2.	Resultados con respecto a los objetivos propuestos.	79
6.	CONCLUSIONES.	82
	BIBLIOGRAFÍA.	86
	Bibliografía consultada:	86
	ANEXO A.	90
	ANEXO B.	93
	ANEXO C.	97

Índice de Figuras

Figura 2. 1 Arquitectura básica de los almacenes datos (Adaptado de (Silberschatz, Korth, Sudarshan, 2006).....8

Figura 2. 2 Esquema en estrella de un almacén de datos (Adaptado de Silberschatz et ál, 2006)14

Figura 2. 3 Arquitectura de una solución de Business Intelligence19

Figura 4. 1 Pentaho BI Server ejecutado.....34

Figura 4. 2 Pentaho BI analizando una consulta multidimensional.35

Figura 4. 3 Pentaho Administration Console en la vista para crear usuarios. .36

Figura 4. 4. Pentaho Data Integration.37

Figura 4. 5. Herramienta Spoon realizando proceso de ETL.....38

Figura 4. 6. Herramienta Mondrian Schema Workbench.39

Figura 4. 7. Herramienta Mondrian Schema Workbench creando una nueva consulta multidimensional.....40

Figura 4. 8. Herramienta Mondrian Schema Workbench conectándose a Pentaho BI server para publicar un cubo.40

Figura 4. 9. Motor MySQL usando su herramienta MySQL Query Browser para el desarrollo de almacenes de datos de prueba.41

Figura 4. 10. Motor Sybase utilizando la herramienta SQL Advantage para la conexión remota a un servidor y trabajar con un almacén de datos.42

Figura 4. 11. Herramienta Sybase Power Designer generando un modelo de almacén de datos, para generar el código SQL.....43

Figura 4. 12. Primer modelo de datos de prueba del almacén de datos piloto para análisis de costos por carrera.....48

Figura 4. 13. Ejecución de la Plataforma de BI de Pentaho en Linux.49

Figura 4. 14. Ejecución de la Consola de Administración de Pentaho en Linux.50

Figura 4. 15. Creación de un usuario para entrar a la plataforma de pentaho BI Server.51

Figura 4. 16. Creación de la conexión a la base de datos del almacén de datos local.52

Figura 4. 17. Conexión a la base de datos local con Mondrian Schema Workbench para creación de consulta multidimensional.....53

Figura 4. 18. Creación de un nuevo esquema para generar un cubo.53

Figura 4. 19. Creando y configurando las dimensiones del almacén de datos.54

Figura 4. 20. Creando y configurando los niveles de la jerarquía en una dimensión, para este caso Carreras.....55

Figura 4. 21. Creación de las dimensiones con sus respectivas jerarquías y niveles para un cubo de costos por carrera.	56
Figura 4. 22. Creación de la medida “Costos” para la consulta multidimensional.	57
Figura 4. 23. Consulta multidimensional finalizada con dimensiones y medidas.	57
Figura 4. 24. Publicación del Cubo a la plataforma de BI de Pentaho.	58
Figura 4. 25. Selección del cubo del almacén de datos de prueba local.	59
Figura 4. 26. Visualización del cubo con dimensiones y medidas.	59
Figura 4. 27. Visualización del cubo con dimensiones y medidas.	60
Figura 4. 28. Modelo de datos del almacén de datos piloto para análisis de costos por carrera en su segunda versión.....	61
Figura 4. 29. Tablas de hechos del modelo del almacén de datos (centros del esquema en estrella).....	63
Figura 4. 30. Configuración de tabla de entrada para proceso de ETL de carreras.	65
Figura 4. 31. Vista Previa de una consulta a una tabla de origen en KETTLE.	66
Figura 4. 32. Configuración de la tabla de destino en el proceso de ETL para carreras.	67
Figura 4. 33. Visualización del diseño del proceso de ETL para las carreras... ..	68
Figura 4. 34. Resultados de la ejecución del proceso ETL para las carreras.. ..	69
Figura 4. 35. Tercera versión del Almacén de Datos.	71
Figura 4. 36. Esquema multidimensional del almacén de datos.	74

Figura 5. 1. Selección de cubos en Pentaho BI Server	76
Figura 5. 2. Cubo de costos por carrera	77
Figura 5. 3. Análisis de Costos de una carrera en particular.	78
Figura 5. 4. Análisis de Costos por alumno.	79

Figura C. 1. Parámetros de conexión a la base de datos en Mondrian Schema-Workbench.....	98
Figura C. 2. Creación de un nuevo esquema y agregación de un cubo.....	99
Figura C. 3. Creación de una nueva dimensión para un cubo.	100
Figura C. 4. Edición de atributos de una jerarquía.	101
Figura C. 5. Esquema multidimensional finalizado con sus dimensiones y medidas.	102
Figura C. 6. Datos de conexión para publicar esquema.	102
Figura C. 7. Publicación de Esquema.	103

Índice de Tablas

Tabla 3. 1. Cuadro Comparativo de Estudios de Costos por Carreras.31

1. INTRODUCCIÓN

En la Universidad de La Habana se ha desarrollado un sistema de información que permite determinar los costos de la carrera. Los diversos sistemas de información de la administración universitaria, como el presupuesto o el sistema de control de gastos, no están interrelacionados para calcular los costos de la carrera.

En el año 1997 se realizó un estudio para determinar los costos de la carrera de la Universidad de La Habana.

En el año 2003, como resultado de un estudio de control de una instancia superior, se determinó que los costos por carrera no estaban calculados correctamente. Se decidió implementar un sistema de información que permitiera calcular los costos de la carrera de la Universidad de La Habana.

CAPÍTULO I

INTRODUCCIÓN

Las Universidades de la Habana y de La Piedad, en el marco de la cooperación académica, han acordado desarrollar un sistema de información que permita determinar los costos de la carrera.

Una de las principales razones para la implementación de este sistema es la necesidad de determinar los costos de la carrera de la Universidad de La Habana.

En la Universidad de La Habana se ha desarrollado un sistema de información que permite determinar los costos de la carrera. Los diversos sistemas de información de la administración universitaria, como el presupuesto o el sistema de control de gastos, no están interrelacionados para calcular los costos de la carrera.

En el año 1997 se realizó un estudio para determinar los costos de la carrera de la Universidad de La Habana. En el año 2003, como resultado de un estudio de control de una instancia superior, se determinó que los costos por carrera no estaban calculados correctamente.

1. INTRODUCCIÓN

En la Universidad de La Frontera no existe en la actualidad un sistema informático que permita determinar los costos asociados a impartir una carrera. Los diversos sistemas informáticos que posee la institución administran información asociada a currículo, cobranza, contabilidad, presupuesto o remuneraciones, pero no se ha definido un método único y consensuado para calcular los costos por carrera.

En el año 1997 se realizó un primer estudio piloto, que no volvió a ser aplicado ni depurado.

En el año 2008, como parte de un proyecto Mecesus de Gestión, se contrató una asistencia técnica para proporcionar una asesoría de la forma de calcular los costos por carrera. Como resultado de dicha asesoría se concluyó que era recomendable ajustar el modelo propuesto a la realidad de la UFRO y a su mayor disponibilidad de información, y que la mejor opción era también implementar un procedimiento automatizado para la generación de resultados.

1.1 Descripción general del problema.

Las Universidades requieren evaluar permanentemente diversos parámetros para gestionar las carreras que ofrecen.

Uno de los parámetros es el costo que significa para la Universidad, por año, el ofrecer una carrera. La Universidad de La Frontera (UFRO) no posee un sistema estandarizado para medir el costo por carrera.

En la Universidad de La Frontera se encuentra al término de su ejecución el Convenio de Desempeño "UFRO Evoluciona", que compromete el logro de resultados notables en tres áreas de gestión institucional: procesos de formación, producción científica tecnológica y gestión. En esta última área, se compromete entre otros la gestión de la cobranza, la que se verá apoyada por el presente trabajo de título en relación a la gestión de la información.

Existe interés institucional en instalar un sistema de este tipo en la Universidad, que permita analizar en detalle los costos asociados a impartir cada una de las carreras.

La implementación de un Data Mart, o almacén de datos departamental, permitiría hacer consultas específicas de manera rápida y eficiente, debido a las múltiples posibilidades de presentación de la información en forma dinámica que las herramientas de inteligencia de negocios ofrecen.

1.2. Objetivos Generales y Específicos.

1.2.1. Objetivo General.

- Implementar un almacén de datos piloto para análisis de costos por carrera anuales en la Universidad de La Frontera

1.2.2. Objetivos Específicos.

- Recopilar información de sistemas para determinar costos por carreras anuales implementados en universidades Chilenas.
- Diseñar la estructura del almacén de datos de costos por carrera.
- Implementar consultas OLAP al almacén de datos usando Pentaho BI.

2.1. Introducción a la toma de decisiones

Según el *McGraw-Hill Database* (2007), la toma de decisiones en el contexto de una organización implica el uso general de todos los aspectos de los datos, por lo que se requiere un sistema que permita almacenar los datos que constituyen dichos sistemas. El sistema consta de datos operativos y de diversas unidades de negocio, como la información histórica y financiera, estadística, contable, entre otros.

La tendencia hacia el uso de grandes volúmenes de datos se complementa con un creciente interés en la explotación de estos datos. Muchas organizaciones de las compañías de análisis de datos han hecho que los sistemas de SQL tradicionales respondan a las necesidades de análisis de datos.

CAPÍTULO II

FUNDAMENTOS TEÓRICOS

La creciente demanda de datos requiere una clase de consultas más complejas que las que se pueden hacer con SQL tradicional. Se requiere un soporte flexible para las consultas de agregación, funciones estadísticas y operaciones de unión y unión. Las aplicaciones de análisis de datos se centran en el análisis de datos en línea (Online Analytic Processing, OAP). Los sistemas soportan un estilo de consulta en el que resulta más fácil navegar en los datos como arrays multidimensionales y están diseñados para ser utilizados por herramientas para los usuarios finales, como el tipo de cálculo, además de por los programadores de consultas de los sistemas de datos.

En segundo lugar, muchos gestores de bases de datos (SGBD) soportan consultas de agregación y funciones estadísticas, pero están diseñados para soportar también consultas de tipo OLAP de manera eficiente. Estos sistemas se pueden usar para OLAP. Los sistemas optimizados para las aplicaciones de OLAP se conocen como sistemas de OLAP. Muchos fabricantes de SGBD están tratando de mejorar sus productos en este sentido y, con el tiempo, es probable que haya una convergencia entre los sistemas OLAP especializados y los SGBD tradicionales para soportar consultas OLAP.

2. FUNDAMENTOS TEÓRICOS

2.1. Introducción a la ayuda a la toma de decisiones

Según Ramakrishnan y Gehrke (2007), la toma de decisiones en el contexto de una organización exige una visión general de todos los aspectos de ésta, por lo cual muchas organizaciones han creado almacenes de datos que contienen datos obtenidos de distintas bases de datos operativas y de diversas unidades de negocio junto con la información histórica y resumida, considerándose como fuentes externas de datos.

La tendencia hacia el almacenamiento de datos se complementa con un creciente énfasis en herramientas de análisis potentes. Muchas características de las consultas de ayuda a la toma de decisiones hacen que los sistemas de SQL tradicionales resulten inadecuados.

Se dispone de tres amplias clases de herramientas de análisis (Ramakrishnan y Gehrke, 2007):

En primer lugar, algunos sistemas soportan una clase de consultas estilizadas que suelen implicar operadores de agrupación y ofrecen un soporte excelente para las condiciones booleanas complejas, funciones estadísticas y características para el análisis de series temporales. Las aplicaciones dominadas por estas consultas se denominan Procesamiento Analítico en Línea (OnLine Analytic Processing, OLAP). Estos sistemas soportan un estilo de consulta en el que resulta más conveniente pensar en los datos como arrays multidimensionales y están disponibles para ser utilizados por herramientas para los usuarios finales como las hojas de cálculo, además de por los lenguajes de consulta de las bases de datos.

En segundo lugar, algunos Sistemas gestores de bases de Datos (SGBD) soportan consultas tradicionales al estilo SQL pero están diseñados para soportar también consultas OLAP de manera eficiente. Estos sistemas se pueden considerar como SGBD relacionales optimizados para las aplicaciones de ayuda a la toma de decisiones. Muchos fabricantes de SGBD están mejorando actualmente sus productos en este sentido y, con el tiempo, es probable que disminuya la diferencia entre los sistemas OLAP especializados y los SGBD relacionales mejorados para soportar consultas OLAP.

Warehouse normalmente contienen los datos de todas las divisiones, departamentos o secciones de una institución, organización, empresa, etc. (Silberschatz et ál, 2006).

2.2.2. Data Mart.

Un Data Mart es una base de datos departamental, especializada en el almacenamiento de los datos de un área específica de negocio. Son almacenes de datos que guardan datos históricos que ayudan en la toma de decisiones. Se caracterizan por ser de una estructura óptima de datos para analizar la información detalladamente. Un Data Mart puede ser poblado desde los datos de un Data Warehouse, o bien desde distintas fuentes de origen de información (Sinnexus, N/D); sin embargo, los Data Mart, se especializan en cada área, departamento o sección de una empresa. Es como un Data Warehouse pero más pequeño, enfocado a un área en específico.

Las cargas de trabajo por lo general incluyen consultas bastante complejas, por lo que los tiempos de las respuestas deben ser rápidos. Como la cantidad de datos que se manejan es mucho mayor en relación a los sistemas de bases de datos relacionales, y las consultas son complejas, se debe tener un modelo de datos menos restringido por lo que tradicionalmente no se realiza una normalización que evite inconsistencia y duplicidad de los datos. Entonces de esta manera el diseño del modelo de datos permite que las consultas complejas se puedan responder rápidamente. Los datos se pueden encontrar almacenados en bases de datos operativas de una organización, pudiendo quedar clasificadas en distintas secciones, como ventas, recursos humanos, etc. (Ramakrishnan y Gehrke, 2007).

Los datos se extraen de las bases de datos operativas y de fuentes externas, se limpian para minimizar errores, se completa la información que falta en caso que sea posible, y se transforman para corregir incoherencias semánticas, esto se realiza debido a que si los datos se encuentran en distintas bases de datos operativas, existe una gran probabilidad de que se encuentren fallas para el almacén de datos, como pueden ser tablas con los mismos nombres, o campos de nombres iguales, entre otros problemas, por lo que la transformación de éstos permite corregir problemas que pueden causar errores al momento de crear el almacén de datos.

Los datos, una vez limpios y transformados, se cargan en el almacén, a través de una vista relacional de las tablas de las fuentes de datos (bases de datos operativas y fuentes externas). Estas vistas se materializan y se guardan en el almacén.

Debido al gran volumen de datos, la tarea en general de carga de datos, puede demorar un tiempo considerable.

Una vez cargados los datos en el almacén, se deben tomar medidas para asegurar que se actualicen periódicamente los datos.

2.2.3. Componentes de los almacenes de datos

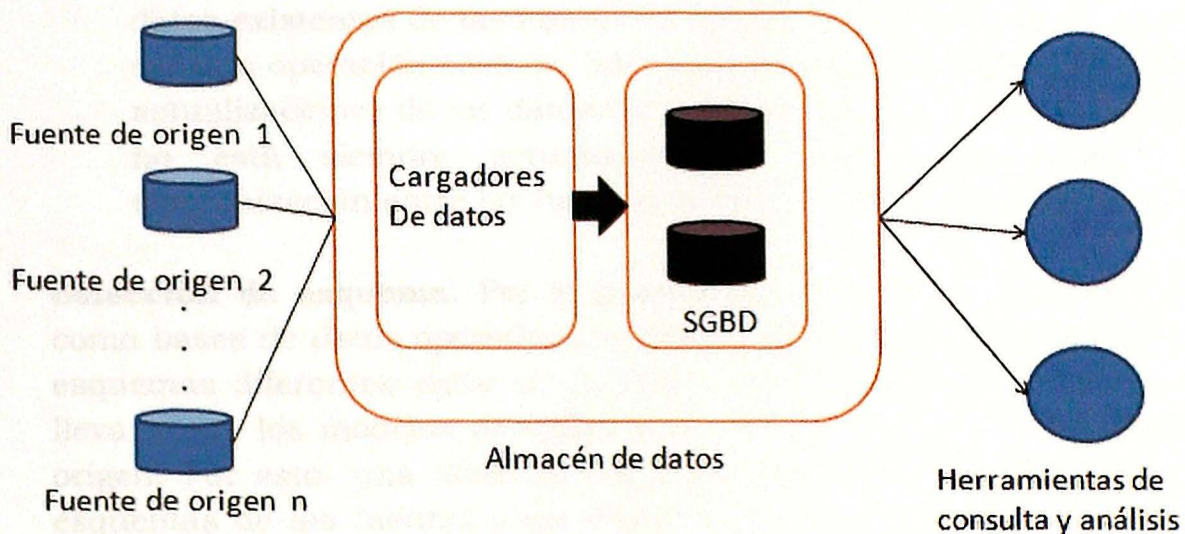


Figura 2. 1 Arquitectura básica de los almacenes datos (Adaptado de (Silberschatz, Korth, Sudarshan, 2006)

La Figura 2.1. muestra la arquitectura básica de los almacenes de datos en donde los orígenes de los datos, de distintas fuentes, se cargan para quedar guardados en los SGBD del almacén de datos, es aquí donde se guardan los datos históricos que son consultados por herramientas de consultas y análisis y permiten ayudar a la toma de decisiones con la información generada.

Según Silberschatz et ál (2006) para crear un almacén de datos se necesita resolver los siguientes problemas:

- **Momento y modo de la recogida de datos.** Existen dos arquitecturas para la obtención de datos, estas son:
 - Arquitectura dirigida por los orígenes: en la cual las fuentes de origen transmiten los datos hacia el almacén de manera continua o de manera periódica.
 - Arquitectura dirigida por el destino: en la cual el almacén de datos realiza periódicamente solicitudes o peticiones de datos nuevos a las fuentes de origen. Debe existir una sincronización entre las fuentes de datos de origen, ya sean de bases de datos operativas o desde fuentes externas. Esta sincronización es para actualizar los datos existentes de las fuentes de origen en el almacén de datos; al ser una operación costosa, hay veces en que no siempre se realizan actualizaciones de los datos del almacén, por lo que la información no está siempre actualizada; por esto debe existir una sincronización entre las fuentes de origen y el almacén de datos.
- **Selección de esquema.** Por lo general las fuentes de datos de origen como bases de datos operativas y fuentes externas tienen estructuras y esquemas diferentes entre sí, ya que cada uno es independiente; esto lleva a que los modelos de datos sean diferentes entre cada fuente de origen. Por esto, una labor de los almacenes de datos es integrar los esquemas de las fuentes y los datos de éstas, realizando un esquema que integre los datos de las diversas fuentes. Este proceso muestra que los datos que contiene el almacén de datos no son una copia exacta de los datos de origen de las fuentes de datos.
- **Transformación y limpieza de los datos.** Se debe realizar la labor de corregir los datos, la cual se denomina limpieza de datos. Las fuentes de origen frecuentemente entregan datos con inconsistencias de carácter menor que se pueden corregir. Por ejemplo, nombres de calles, direcciones, nombres de ciudades, códigos de productos, códigos postales incorrectos, etc.

Puede ocurrir que datos obtenidos de distintas fuentes de datos se repitan existiendo duplicación de datos, por lo que se tiene que

realizar una operación de eliminación de estos datos por lo general conocida como “*desduplicación*” (Silberschatz et ál, 2006).

Los datos, además de pasar por una etapa de limpieza para la corrección, se pueden transformar, como por ejemplo cambiar unidades de medida, o incluso generar un esquema o modelo diferente que reúna los datos relacionales de varios orígenes, de manera de que sean compatibles entre ellos para evitar errores y que estén integrados para manejar bien el almacén de datos. Los almacenes de datos regularmente se acompañan de herramientas gráficas que dan soporte a la transformación de datos.

- **Propagación de las actualizaciones.** Las actualizaciones de las relaciones establecidas en las fuentes de origen de datos deben extenderse a los almacenes de datos. En caso de que las relaciones del almacén de datos sean exactas a la de los orígenes de datos, se realiza una propagación directa; de no ser así, el problema de la propagación (extensión de actualización de relaciones al almacén de datos) pasa a ser un problema de mantenimiento de vistas que se debe trabajar.
- **Resúmenes de los datos.** Los datos brutos que se generan por el sistema transaccional de datos al almacén de datos pueden ser demasiado grandes como para realizar este proceso en línea. Por consiguiente, se pueden responder consultas manteniendo datos resumidos obtenidos por agregación de relaciones en vez de mantener las relaciones completas. Un ejemplo puede ser, en vez de almacenar los datos de cada venta de productos, almacenar las ventas totales de los productos por nombre de artículo y categoría.

Los distintos pasos que se deben realizar para la obtención de datos para el almacenamiento en el almacén de datos se denominan: ETL (Extract, Transform, Load) o también ETC (Extracción, Transformación, Carga).

2.3. OLTP - On-Line Transactional Processing

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones (Sinnexus, N/D). Una transacción genera un proceso atómico (que debe ser validado con un *commit*, o invalidado con un *rollback*), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura (por ejemplo, la enorme cantidad de transacciones que tienen que soportar las BD de bancos o hipermercados diariamente).
- Los datos se estructuran según el nivel de aplicación (programa de gestión a medida, ERP o CRM implantado, sistema de información departamental, etc.).
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad y la existencia de islas de datos).
- El historial de datos suele limitarse a los datos actuales o recientes.

2.4. OLAP. Online Analytical Processing.

Online Analytical Processing (Procesamiento Analítico en Línea) es un sistema interactivo que permite a los analistas ver diferentes resúmenes de los datos multidimensionales (Silberschatz et ál, 2006). Consiste en consultas a estructuras multidimensionales que contienen datos resumidos de grandes bases de datos o sistemas transaccionales.

La razón de usar OLAP para las consultas es la velocidad de respuesta. Una base de datos relacional almacena entidades en tablas que han sido normalizadas. Esta estructura es buena en un sistema OLTP pero para consultas complejas y de muchos campos de diferentes tablas es relativamente lenta. Un modelo mejor para búsquedas es una base de datos multidimensional. La principal característica que potencia a OLAP, es que es lo más rápido a la hora de hacer consultas, en contraposición con OLTP que es la mejor opción para la operación normal del sistema, que incluye inserciones, actualizaciones y eliminaciones de datos.

Existen algunas clasificaciones entre las implementaciones OLAP. La clasificación está hecha sobre la base de en qué tipo de motor son almacenados los datos (Sinnexus, N/D):

- **ROLAP** es una implementación OLAP que almacena los datos en un motor relacional. Típicamente, los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas. La arquitectura está compuesta por un servidor de banco de datos relacional y el motor OLAP se encuentra en un servidor dedicado. La principal ventaja de esa arquitectura es que permite el análisis de una enorme cantidad de datos.
- **MOLAP** es una implementación OLAP que almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta. Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis.
- **HOLAP** (Hybrid OLAP) almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.

Existen sistemas OLAP que se implementan como sistemas de tipo cliente-servidor. La parte del servidor contiene la base de datos relacional y los cubos de datos MOLAP. Y los sistemas clientes obtienen las vistas de los datos, a través de la comunicación con el servidor (Silberschatz et ál, 2006).

2.4.1. Modelo Multidimensional de datos.

El modelo multidimensional de datos se enfoca en un conjunto de medidas numéricas y cada medida depende de un conjunto de dimensiones.

Son adecuados para resumir y organizar datos, un ejemplo se consideran las hojas de cálculo; son enfocados para trabajar sobre datos de tipo numérico. Es más fácil de visualizar y entender que el modelo entidad-relación. Este modelo multidimensional puede realizarse cuando existen atributos de medida y atributos de dimensión.

2.5. Esquemas de los almacenes de datos

Los almacenes de datos normalmente tienen esquemas diseñados para los análisis de los datos, y usan herramientas como OLAP. Por consiguiente, los datos suelen ser datos multidimensionales, en los cuales se encuentran atributos de dimensión y atributos de medida. Las tablas que almacenan datos multidimensionales se llaman **tablas de hechos**, que por lo general son de gran tamaño. Los atributos de dimensiones suelen ser identificadores para minimizar los requisitos de almacenamiento, estos atributos son claves foráneas de otras tablas llamadas **tablas de dimensiones** (Silberschatz et ál, 2006).

2.5.1. Esquema en estrella.

La figura 2.2 muestra un ejemplo de un esquema utilizado como modelo de datos de un almacén de datos. Explica la estructura de cómo serán almacenados los datos de una tienda de artículos (Silberschatz et ál, 2006).

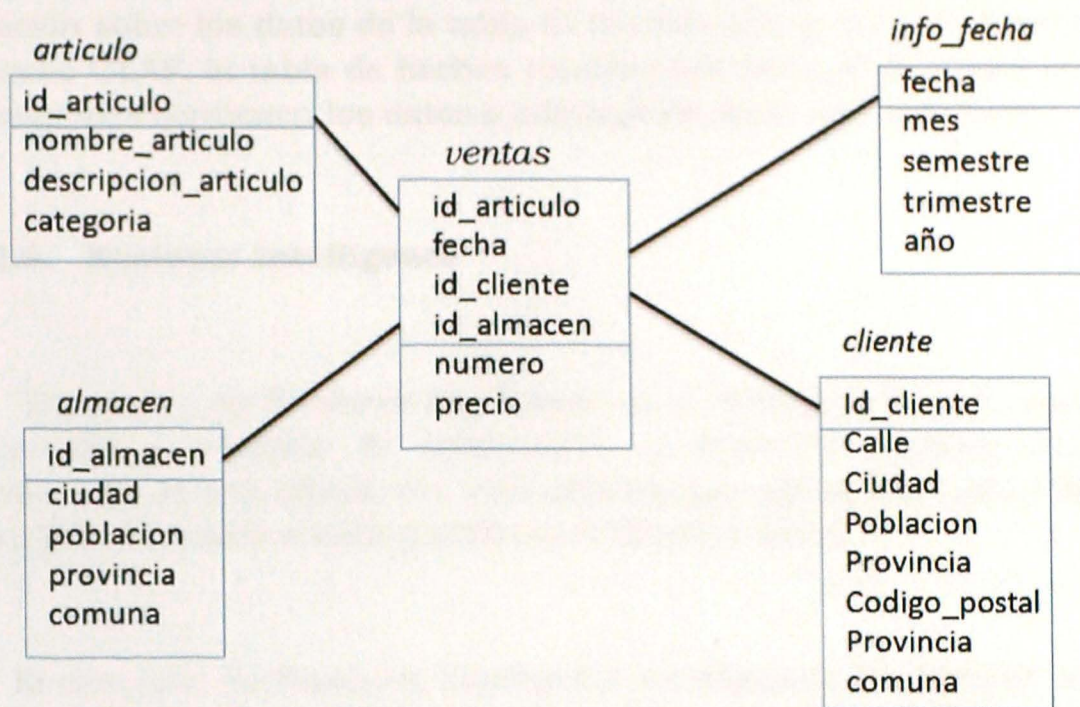


Figura 2. 2 Esquema en estrella de un almacén de datos (Adaptado de Silberschatz et ál, 2006)

La tabla que se ubica en el centro es la tabla de hechos, la cual tiene como clave compuesta las claves foráneas de las demás tablas (“almacen”, “cliente”, “articulo”, “fecha”) que son las tablas de dimensiones. Se le llama esquema estrella por formar una figura de estrella en la relación de la tabla de hechos con las tablas de dimensión.

El esquema estrella es una estructura lógica que tiene una tabla de hechos que contiene datos factuales en el centro, rodeadas por tablas de dimensión que contienen datos de referencia (pueden estar desnormalizados) (Connolly y Begg, 2005).

En la construcción de cubos OLAP, las tablas de dimensiones son elementos que tienen atributos que se utilizan para restringir y agrupar los datos que se almacenan en una tabla de hechos, cuando se realizan consultas sobre dichos datos en un almacén de datos.

Estos datos sobre dimensiones son parámetros de los que dependen otros datos, que serán objeto de análisis y que están contenidos en la tabla de hechos. Las tablas de dimensiones ayudan a realizar ese análisis aportando

información sobre los datos de la tabla de hechos, por lo que puede decirse que en un cubo OLAP, la tabla de hechos contiene los datos de interés y las tablas de dimensiones contienen los datos e información sobre estos hechos.

2.6. Business Intelligence

“El corazón de Business Intelligence es la habilidad de una organización para acceder y analizar la información, y entonces explotar su ventaja competitiva. En la era digital, las capacidades que ofrece Business Intelligence será la diferencia entre el éxito y el fracaso” (Medina 2005).

El concepto de Business Intelligence (Inteligencia de Negocio) se conoce como el conjunto de metodologías, tecnologías y aplicaciones que permiten reunir, depurar y transformar datos de sistemas transaccionales e información desestructurada en información estructurada, para su explotación directa o bien para su análisis y conversión del conocimiento, generar escenarios, pronósticos, facilitando y dando soporte a la toma de decisiones sobre el negocio (Sinnexus, N/D). En términos simples, Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

La solución de Business Intelligence completa, según Sinnexus (N/D), permite:

Observar ¿qué está ocurriendo?

Comprender ¿por qué ocurre?

Predecir ¿qué ocurriría?

Colaborar ¿qué debería hacer el equipo?

Decidir ¿qué camino se debe seguir?

Business Intelligence trabaja como un factor estratégico para una empresa, institución u organización, generando una potencial ventaja, la cual es proporcionar información privilegiada para la toma de decisiones de negocio, ofertas de productos, control financiero, planificación de la producción entre otros (Sinnexus, N/D).

2.6.1. Características de Business Intelligence

La solución de Business Intelligence ayuda a implementar estrategias para el éxito a largo plazo ayudando a la toma de decisiones, al proporcionar funcionalidades que complementen la mejor comprensión de la información y su pronto análisis para la generación de conocimiento.

Las características que se consideran en Business Intelligence son las siguientes (Kehr, 2010):

Análisis de información: Business Intelligence cuenta con características de almacenamiento de datos, herramientas de reporte y análisis, modelos de mejores prácticas, aplicaciones de análisis de negocios y recursos administrativos.

Seguridad: Business Intelligence cuenta con la actualización constante de la información y permite conocer el patrón de comportamiento de un usuario.

Toma de decisiones: Las funciones de la solución permiten a los empleados de todo nivel organizacional tomar decisiones plenamente informadas.

Enfocadas en el valor: Con Business Intelligence, se lograrán menores costos totales y un rápido rendimiento sobre la inversión.

Integradas: Business Intelligence ofrece una integración homogénea con todos los demás componentes que integran la solución, brindando de esta forma confiabilidad al usuario en la información entregada.

Ampliables: Esta herramienta es extensible y adaptable a las cambiantes necesidades de los negocios. La solución incluye herramientas e interfaces que suplementan y expanden los contenidos de negocios existentes o que se integran con herramientas de análisis y reporte desarrolladas por terceros.

Personalizadas: Se pueden personalizar los contenidos y las maneras como acceden a los mismos, manteniéndose ajustados a las políticas de seguridad fijadas por las organizaciones.

Probadas: Business Intelligence ofrece conocimientos y análisis probados. La solución está probada por miles de instalaciones productivas a nivel mundial, dando seguridad en las ventajas empresariales que BI ofrece.

2.6.2. Arquitectura de una solución de Business Intelligence

Una solución de Business Intelligence parte de los sistemas de origen de una organización (bases de datos operativas, fuentes externas como ERP, archivos de texto, etc.), sobre los que suele ser necesario aplicar una transformación estructural para optimizar su proceso analítico (Sinnexus, N/D).

Debido a esto, se realiza la fase de extracción, transformación y carga (ETL) de datos. Esta etapa suele apoyarse en un almacén intermedio, llamado Operational Data Store ODS que actúa como pasarela entre los sistemas fuente y los sistemas destino (generalmente un Data Warehouse), y cuyo principal objetivo consiste en evitar la saturación de los servidores funcionales de la organización.

La información resultante, ya unificada, depurada y consolidada, se almacena en un Data Warehouse que puede servir como base para la construcción de distintos Data Mart departamentales. Estos Data Mart se caracterizan por poseer la estructura óptima para el análisis de los datos de un área en específico de la organización o empresa, ya sea mediante bases de datos transaccionales (OLTP) o mediante bases de datos analíticas (OLAP).

Los datos que se encuentran almacenados en el Data Warehouse o en cada Data Mart se explotan utilizando herramientas de análisis, reportes, etc. para generar el producto de Business Intelligence que apoya la toma de decisiones mostrando pronósticos, distintos escenarios, amenazas, fortalezas, tendencias, etc. Algunos productos de Business Intelligence son: Cuadros de Mando Integrales (CMI), Sistemas de Soporte a la Decisión (DSS), Sistemas de Información Ejecutiva (EIS).

A continuación la figura 2.3 muestra la arquitectura de una solución de Business Intelligence en la cual se muestra el origen de los datos, pasando por cada etapa que son proceso de ETL, almacenamiento en el Data Warehouse, posible almacenamiento en los Data Marts para optimizar y trabajar en un área en particular, explotación de la información a través de distintas herramientas de análisis, hasta la solución del producto Business Intelligence (BI) donde finalmente se genera conocimiento.

Figura 2. 3 Arquitectura de una solución de Business Intelligence

3. ANTECEDENTES DE LA INSTITUCIÓN

3.1. Sistema de Control por Carrera

Un sistema de control de un sistema físico es aquel que permite la regulación de las variables de salida de un sistema a partir de las variables de entrada de un sistema de control (Khalid, 1997).

El Sistema de Control por Carrera es un sistema de control que permite la regulación de las variables de salida de un sistema a partir de las variables de entrada de un sistema de control (Khalid, 1997).

Existen diversos tipos de sistemas de control por carrera, los cuales se clasifican en función de la estructura del sistema de control y de la forma en que se realiza la regulación (Khalid, 1997).

CAPÍTULO III

ANTECEDENTES GENERALES

3. ANTECEDENTES DE LA INSTITUCIÓN.

3.1. Sistemas de Costos por Carrera.

Un sistema de costos es un sistema ideado para buscar una forma de prorrateo de los distintos costos asociados a la obtención de un bien o servicio, de manera de obtener como resultado un valor unitario de dicho bien o servicio (UFRO, 1997).

El Sistema de Costos por Carrera, por tanto, tiene por objetivo obtener el Costo por Alumno de cada una de las carreras impartidas en pregrado en una universidad, en este caso en particular, la Universidad de La Frontera.

Existen diversas formas de enfrentar dicho problema, pero dadas las características del tipo de servicio que la Universidad ofrece, el que en términos generales es la dictación de Carreras Universitarias, que se ofrecen a lo largo de todo un año, se ha optado por la determinación de Costos por Carrera anuales (UFRO, 1997).

3.2. Análisis de distintos Sistemas de Costos por Carrera implementados en la Universidad de La Frontera.

En la Universidad de La Frontera se ha tratado de implementar tres sistemas para calcular los costos de las Carreras que imparten, ya sea en su totalidad o de algunas Carreras.

3.2.1. Sistema Costos por Carrera año 1997

Este fue el primer estudio realizado en la Universidad de La Frontera para el análisis de costos por carrera estudiando y planteando los algoritmos de cálculo de costos (UFRO, 1997).

Los datos que contiene el sistema de costos planteado están abocados a determinar los costos de la Docencia de Pregrado. Para la correcta

determinación de los costos por carrera, se aisló la componente de Docencia de los demás quehaceres normales de la Universidad, como son Investigación y Extensión, por lo tanto se dividieron los costos asociados al normal desarrollo de las actividades Universitarias en:

- a) Investigación.
- b) Extensión
- c) Docencia, que a su vez se dividió en :
 - i) Docencia de Pregrado.
 - ii) Docencia de Perfeccionamiento y Postgrado.

Los datos que contiene el sistema de costo planteado, están orientados a determinar los costos de la Docencia de Pregrado.

En este trabajo se analizaron y definieron los elementos componentes del costo por carrera, los distintos elementos definidos fueron:

- Costo Directo de Docencia.
- Costo Indirecto de Docencia
- Costo Administración Central de la Facultad
- Costo Administración Central.

Para cada una de las componentes del Costo por Carrera, fue preciso determinar los ítemes de costo considerados:

- Remuneraciones de Directivos, Académicos, Administrativos y Honorarios.
- Gastos en Bienes y Servicios Corrientes, Viáticos, Transferencias
- Depreciación de Activos Fijos.

Tipos de costos que componen el Costo por Carrera.

(i) Costo Directo de Docencia:

Está Compuesto por las remuneraciones pagadas a los académicos de la Universidad, los cuales pertenecen a alguno de los diferentes departamentos académicos.

Los costos directos de docencia se calcularon por Departamento académico. Para cada departamento el costo directo de docencia se obtuvo como la suma de las remuneraciones de sus académicos multiplicadas por un

factor de prorratio, que indicaba el porcentaje de horas contratadas, dedicadas a docencia de pregrado.

(ii) Costo Indirecto de Docencia:

En este componente de Costos se consideró gastos en viáticos, honorarios, bienes y servicios, y transferencias de los respectivos departamentos académicos.

Además se incluyó la depreciación de los activos fijos asignados a cada uno de estos departamentos. Los honorarios fueron considerados como parte de los costos aquí definidos, debido a la dificultad que se presentaba al analizar si dichos honorarios fueron con el objetivo de contratar académicos para dictar alguna asignatura o bien fueron incurridos para otros fines no directamente asociados con la docencia.

Al igual que para el costo directo, se consideró factores de prorratio de dedicación a docencia de pregrado por parte de los departamentos académicos.

(iii) Costo Administración Central de Facultades:

Se consideró aquellas unidades que corresponden a la Administración Central de las Facultades: Decanato, Secretaría Académica, Secretaría de Pregrado, etc.

Se incluyeron los gastos en remuneraciones del personal directivo y administrativo dependiente de las facultades, los honorarios, viáticos, gastos en bienes y servicios, transferencias y depreciación de los activos fijos asociados a cada una de las unidades de Administración de las Facultades.

(iv) Costo de Administración Central:

Aquí se consideraron todos los costos en que incurre la Administración Central de la Universidad, incluida también la depreciación de sus activos fijos y ponderados por los respectivos porcentajes de docencia que pueden ser asociados a cada una de las unidades.

Se clasificaron tres grandes grupos los que corresponde a las unidades que corresponden a la administración Central de la Universidad:

- **Administración Central:** Este grupo contiene a las unidades de la Administración Central Universitaria.

- **Gastos Centralizados:** En este grupo se encuentran unidades y centros de costos que habitualmente son considerados como Administración Central, pero cuyo objetivo es centralizar gastos, tales como: agua, electricidad, teléfonos, etc., correspondientes a toda la universidad.
- **Otras Unidades de la Administración Central:** En esta clasificación se agruparon las unidades correspondientes a institutos independientes y proyectos, los cuales no estaban asociados directamente a ninguna facultad. Aquí también se incluyeron los proyectos de construcciones, los que, por tratarse de inversiones, no gravan las estructuras de los costos por carrera.

Cada unidad que compone la Administración Central Universitaria fue analizada por separado, para determinar, en función de sus objetivos, en cuál de los tres grupos sería clasificada, y el porcentaje de participación en docencia de pregrado, docencia de perfeccionamiento, investigación y extensión, de manera de poder asignar, de acuerdo a esos porcentajes, los costos que cada una de esas unidades impone a la Docencia de Pregrado en la determinación de los Costos por Carrera.

3.2.2. Trabajo de Título referente a los Costos por Carrera.

Este Trabajo de Título de Aravena (2002) se consideró como caso de estudio, teniendo en cuenta la magnitud que implicaba al considerar sólo un departamento. Su desarrollo se basó en el estudio UFRO (1997).

El objetivo de este estudio también se enfocó principalmente en el área de Docencia de Pregrado, aislándola de las demás actividades que realiza la Universidad, las cuales se clasificaron en Investigación y Extensión, también Docencia de Perfeccionamiento y Postgrado.

Este trabajo se basó en calcular los costos por Carrera que se obtienen al final de un periodo anual, a partir de los costos directos e indirectos, para las Carreras de Ingeniería Civil Industrial, Ingeniería Agroindustria, Ingeniería en Alimentos e Ingeniería Ambiental. Estas tres carreras estaban adscritas al departamento de Ingeniería Química. Cabe señalar que en este estudio el análisis se realizó como si se tratase de una sola carrera.

Este estudio también identifica y propone utilizar los siguientes elementos componentes del costo por carrera:

- Costo Directo de Docencia.
- Costo Indirecto de Docencia.
- Costo Indirecto de Administración de Facultades.
- Costo Indirecto de Administración Central Universitaria.

Para cada componente se incluyeron ciertos ítemes:

(i) Costo directo de Docencia:

- Remuneraciones de los académicos de los distintos departamentos que prestan el servicio de docencia a las carreras analizadas.
- Total de créditos de docencia anuales impartidos por cada departamento que presta servicio de docencia las carreras en estudio.
- Total de créditos de docencia anuales impartidos hacia las carreras foco de análisis.
- Fracción de dedicación a la docencia de pregrado de cada uno de los departamentos que prestan el servicio de docencia a las carreras analizadas.

(ii) Costo Indirecto de Docencia:

- Remuneraciones del personal no académico de los distintos departamentos que prestaban el servicio de docencia a las carreras objeto de costo.
- Gastos en viáticos, honorarios, bienes y servicios corrientes, transferencias y compromisos pendientes a años anteriores, de todos los departamentos que prestaban el servicio de docencia a las carreras estudiadas.
- Total de créditos de docencia anuales impartidos por cada departamento que prestaba servicio de docencia a las carreras analizadas.
- Total de créditos de docencia anuales impartidos hacia las carreras.
- Fracción de dedicación a la docencia de pregrado de cada uno de los departamentos que prestaban servicio de docencia a las carreras analizadas.

(iii) Costo Indirecto de Administración de Facultades.

Aquí se incluyeron estos ítemes:

- Remuneraciones de directivos, personal académico y no académico de las distintas unidades consideradas parte de la administración de facultades.
- Gastos en viáticos, honorarios, bienes y servicios corrientes, transferencias y compromisos pendientes de años anteriores, de las unidades consideradas como parte de la administración de las facultades.
- Total de créditos impartidos por cada facultad a las carreras objeto de costo durante el año considerado base histórica.
- Total de créditos impartidos por cada facultad durante el año considerado como base histórica.
- Fracción de dedicación a la docencia de pregrado de cada una de las facultades analizadas.

(iv) Costo Indirecto de Administración Central Universitaria:

- Remuneraciones de directivos, personal académico y no académico de las distintas unidades consideradas parte de la administración central Universitaria.
- Gastos en viáticos, honorarios, bienes y servicios corrientes, transferencias y compromisos pendientes de años anteriores, de las unidades que se consideraron parte de la administración Universitaria.
- Depreciación de edificios de la Universidad.
- Fracción de dedicación a la docencia de pregrado de la Universidad.
- Matrícula total de alumnos de la Universidad durante el año considerado como base histórica de cálculo de costos.
- Matrícula total de alumnos de las carreras analizadas durante el año considerado como base histórica para el cálculo de costos.

Para este estudio se hicieron algunas consideraciones generales:

- El año considerado como base histórica para el cálculo del costo de las carreras fue el año 2000.
- Las tres carreras en estudio se costearon como si fuese una sola carrera.
- El estudio se realizó bajo dos escenarios, uno sin considerar la planificación académica, esto quiere decir, suponiendo que la Universidad sólo realiza labores de docencia de pregrado, y el otro escenario fue considerando la planificación académica.

- Los datos necesarios para el cálculo de los costos, fueron extraídos de las distintas bases de datos de la Universidad.

En el estudio se identificaron cuáles eran los distintos ingresos de las carreras objeto de estudio. Entre estos ingresos se encuentran:

- Arancel anual de las carreras.
- Crédito solidario.
- Becas (J. Gómez M., Mineduc, etc).
- Aporte Fiscal Indirecto.
- Arancel de Matrícula (Arancel de derechos básicos).
- Crédito UFRO.
- Factor de recuperación de crédito UFRO, que para estos efectos se consideró de un 20% anual. Esta cifra surgida a raíz del comportamiento histórico.

Con estos datos, en el estudio se generaron las fórmulas de cálculo necesarias para estimar los costos de las carreras.

3.2.3. Estudio Costos por Carrera año 2008

En el estudio llamado “Diseño/desarrollo de un modelo para el estudio de costos por carrera Universidad de La Frontera” (UFRO, 2008) y resultados de diseño (UFRO, 2010), se propuso un modelo que consideró estudiar los costos anuales de operación de un total de 12 carreras, obteniéndose indicadores tales como costo por carrera, costos promedio por alumno carrera, y resultados operaciones anuales por carrera.

Para la identificación de los costos se utilizaron cuatro corrientes de recursos, agrupadas en los siguientes servicios:

(i) Servicios Docentes (costos directos):

Están conformados por la utilización de recursos docentes, agrupados en facultades y departamentos académicos.

(ii) Servicio de Administración Docente:

Estos costos están conformados por la utilización de recursos de administración docente de nivel central, tales como Vicerrectoría Académica, Registro Curricular, Admisión, Biblioteca, etc.

(iii) Servicio de Administración Central:

Conformado por la utilización de recursos de gestión directiva y apoyo logístico de nivel central, tales como Rectoría, Vicerrectoría de Administración y Finanzas y unidades de la Administración Central.

(iv) Servicios Estudiantiles:

Conformado por la utilización de recursos de servicios estudiantiles, tales como la Dirección de Desarrollo Estudiantil (DDE) y unidades dependientes.

Todos estos componentes son identificados, ordenados y analizados por medio de la ejecución del presupuesto anual de operaciones de la institución.

El costo total de una carrera para un año presupuestario ejecutado se definió como:

$$\text{CTC} = \text{CTAnualServDoc} + \text{CTAnualServAdmDoc} + \text{CTAnualServAdminCent} + \text{CTAnualServEstu}.$$

Considerando como:

CTC: Costo total de una carrera.

CTAnualServDoc: Costo total anual de servicios docentes de una carrera.

CTAnualServAdmDoc: Costo total anual del servicio de administración docente de una carrera.

CTAnualServAdminCent: Costo total anual del servicio de administración central de una carrera.

CTAnualServEstu: Costo total anual de servicios estudiantiles de una carrera.

Como fuentes de financiamiento el modelo consideró que una carrera obtiene cinco fuentes de ingresos:

- Pago Aranceles de matrícula de la carrera.
- Crédito fiscal de la carrera.
- Becas arancel de matrícula de carrera.
- Aporte fiscal directo por alumno carrera.
- Aporte fiscal indirecto por carrera.

Por lo que los ingresos totales son la suma de estas cinco entradas.

Se realizó un procedimiento con los pasos de cómo aplicar este modelo:

- a. Clasificación y agrupación de centro de costos:** El primer procedimiento consiste en clasificar y agrupar los centros de costos presupuestarios que entregan servicios directos e indirectos a las carreras de la institución: Servicios Docentes, Servicios de Administración Docente, Servicios de Administración Central, Servicios Estudiantiles.
- b. Cálculo ponderador de distribución de Costo por Servicio:** El segundo procedimiento consistió en calcular los ponderadores de distribución de costos de servicios directos e indirectos para cada carrera costeadada, éstos se determinaron sobre la base de los alumnos-asignatura atendidos por departamento / facultad.
- c. Distribución de Costos por Carrera:** El tercer procedimiento consistió en distribuir entre todas las carreras los costos directos que son los servicios docentes, y los costos indirectos que son los servicios de: administración docente, administración central y servicios estudiantiles.

Distribución de Costos de Servicios Directos: Se distribuyen considerando los ponderadores de distribución de costo departamental que resulta de la sumatoria de todos los alumnos-asignatura atendidos por cada departamento académico de la institución.

Distribución de Costos de Servicios Indirectos: Se distribuyeron considerando los ponderadores de distribución de costo institucional que

resulta de la sumatoria de todos los alumnos-asignatura atendidos por todos los departamentos académicos de la institución.

- d. Cálculos de Costos por carrera: El cuarto procedimiento consiste en agrupar los ingresos efectivos por carrera; esto se obtiene directamente del registro de ingreso por carrera, al cual se le suma la distribución del aporte fiscal directo y el aporte fiscal indirecto.
- e. Cálculo de variables de resultados: El modelo en estudio entrega tres variables como resultados para su análisis.
 - Costo Carrera: Valorización económica de todos los servicios directos e indirectos (recursos) que utiliza una carrera de pregrado en el periodo de un año, en la formación de profesionales. Se obtiene por sumatoria de todos los costos ponderados por Servicio/Carrera.
 - Costo Carrera por Alumno: Costo promedio anual que incurre una carrera de pregrado en la formación profesional de un alumno. Esta se obtiene dividiendo el costo total carrera por la matrícula de alumnos regulares de la carrera.
 - Resultado Operacional: Comparación de ingresos y egresos efectivos en que incurre la operación de una carrera de pregrado en el periodo de un año. Este resultado se obtiene de restar a los ingresos totales los costos totales de la carrera.

3.3. Cuadro Comparativo de Estudios de Costos por Carrera.

De los tres estudios realizados para intentar calcular los costos por carrera de la Universidad de La Frontera se pueden encontrar algunas similitudes y diferencias que pueden influir en el enfoque y medida en que se pueden realizar los cálculos.

Tabla 3. 1. Cuadro Comparativo de Estudios de Costos por Carreras.

Criterio	Estudio año 1997	Estudio Trabajo de Título 2002	Estudio año 2008
Cantidad de Carreras en estudio	Se abarcan todas las carreras impartidas vigentes en el año 1997	Se analizan tres carreras (que tienen que ver directamente con el departamento de Ingeniería Química) en el año 2000, considerándolas como una sola carrera	Se realiza el estudio sobre una selección de 12 carreras de las cuatro facultades.
Componentes de Costos que se identifican	Se enfoca en estas cuatro componentes para identificar los costos: Costo Directo de Docencia. Costo Indirecto de Docencia. Costo de Administración Central de Facultad. Costo de Administración Central.	Se enfoca en estas cuatro componentes para identificar los costos (se basa en el estudio de 1997): Costo directo de Docencia. Costo Indirecto de Docencia. Costo Indirecto de Administración de Facultades. Costo Indirecto de Administración Central Universitaria.	Los cuatro componentes de costos identificados son: Servicios Docentes (costos directos). Servicio de Administración Docente. Servicio de Administración Central. Servicios Estudiantiles.

CAPÍTULO IV

DESCRIPCIÓN DE ACTIVIDADES REALIZADAS

4. DESCRIPCIÓN DE ACTIVIDADES REALIZADAS.

4.1. Tecnologías Utilizadas.

Se utilizaron diferentes herramientas, metodologías, tecnologías para llevar de manera correcta el desarrollo del almacén de datos piloto para calcular los costos por carrera de la Universidad de La Frontera.

4.1.1. JAVA.

Para utilizar Pentaho y todas sus herramientas se instaló Java Development Kit para correr las aplicaciones de esta herramienta libre para trabajar con Business Intelligence. Para trabajar localmente con datos de prueba se utilizó la versión de Java "1.6.0_22".

4.1.2. Pentaho Open Source Business Intelligence.

El proyecto Pentaho BI es una iniciativa de una comunidad open source que provee a las organizaciones las mejores soluciones para las necesidades de inteligencia de negocio. La herramienta Pentaho Business Intelligence (BI) es de código abierto, la cual está desarrollada para las necesidades de Análisis de los Datos lo cual ayuda a la toma de decisiones, permite analizar consultas multidimensionales de bases de datos (normalmente almacenes de datos) para ver tendencias, posibles problemas, oportunidades, amenazas, proyecciones, entre otros, logrando que a través de diferentes herramientas que contiene Pentaho se puedan implementar soluciones inteligentes a las empresas (Calderón, N/D).

4.1.3. Pentaho BI-Server.

Pentaho BI-Server es la plataforma para ver análisis y generar reportes con una herramienta que contiene por defecto, esta plataforma ha sido desarrollada en JAVA.

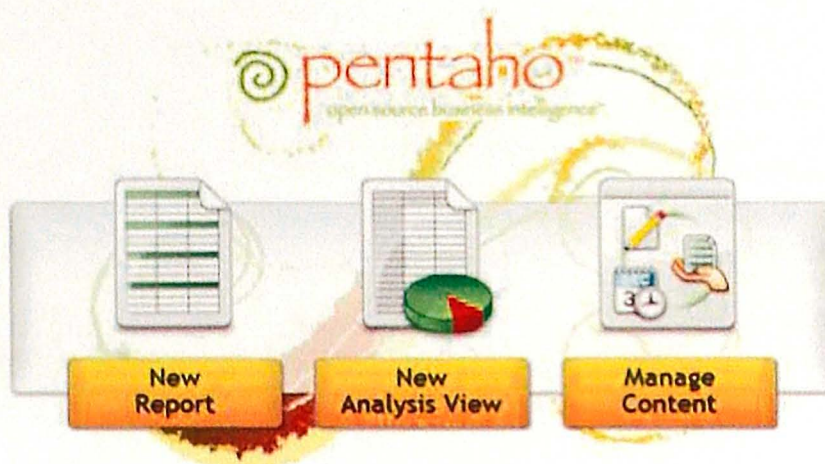


Figura 4. 1 Pentaho BI Server ejecutado.

Su principal objetivo es poder elegir distintos cubos (consultas multidimensionales cargadas anteriormente) para poder verlos, analizarlos, filtrar columnas, realizar consultas multidimensionales para el cubo en análisis, realizar una vista general detallada de todos los registros del cubo, ocultar datos de dimensiones que tengan valor cero en sus medidas respectivas, girar para cambiar de posición valores entre medidas y dimensiones, añadir distintos tipos de gráficos, ordenar campos de las consultas, exportar a .xls o .pdf, entre otras funcionalidades más.

Programa	Grupo	Año	Valor	Contador
Ingenieria Comercial	Agrupacion	Anio	18.343.100	1
Ingenieria Informatica	Agrupacion	Anio	263.057.200	1
Ingenieria Matematica	Agrupacion	Anio	752.995.678	9
	Administracion Central de Facultad	Anio	577.066.000	4
	Remuneraciones	Anio	577.066.000	4
		2007	135.723.000	1
		2008	139.723.000	1
		2009	148.820.000	1
		2010	152.800.000	1
	Costo Indirecto de docencia	Anio	163.584.000	4
		2007	35.626.000	1
		2008	39.026.000	1
		2009	42.206.000	1
		2010	46.726.000	1
	Remuneraciones	Anio	163.584.000	4
Ingenieria en Construccion	Agrupacion	Anio	294.453.100	1

Terminado

Figura 4. 2 Pentaho BI analizando una consulta multidimensional.

4.1.4. Pentaho Administrator Console.

Esta herramienta permite crear usuarios para ingresar a la plataforma de Pentaho, asignándoles distintos roles.

Además permite realizar conexiones a distintas bases de datos de diversos motores, dependiendo de la base de datos que requiera para generar los cubos, para poder publicar cubos en la plataforma Pentaho. Para ver los cubos en la plataforma Pentaho es indispensable crear aquí la conexión con la base de datos (en este caso con el almacén de datos).

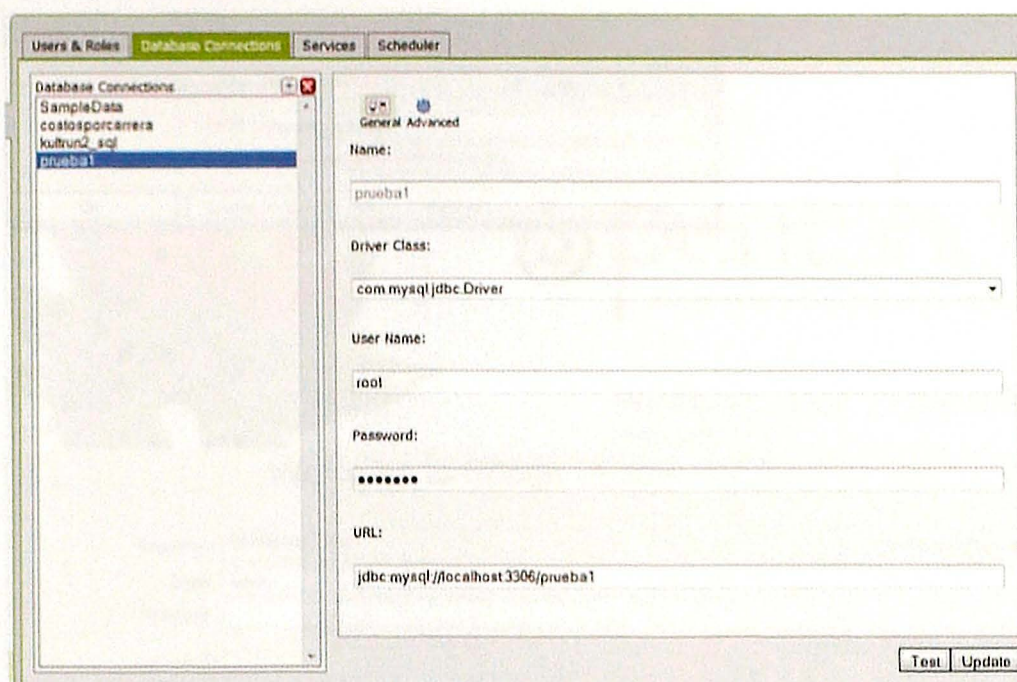


Figura 4. 3 Pentaho Administration Console en la vista para crear usuarios.

4.1.5. Pentaho Data Integration (KETTLE).

Pentaho Data Integration (PDI) es un motor de un conjunto de herramientas para realizar procesos de Extracción, Transformación y Carga, mejor conocidos como ETL. PDI no sólo sirve como una herramienta ETL, también es utilizado para otros fines tales como la migración de datos entre las aplicaciones o bases de datos, exportar datos desde bases de datos de archivos planos, limpieza de datos, y mucho más. PDI tiene una interfaz intuitiva, y sus capacidades para ETL son potentes (Roldán, 2010).



Figura 4. 4. Pentaho Data Integration.

Esta herramienta permite el desarrollo de los procesos ETL (Extraction, Transformation, Load) o bien Extracción, Transformación y Carga. Este proceso se realiza para poblar la base de datos, o por lo general, almacenes de datos, debido a que en éstos los registros que se deben insertar son muchos (ya que son datos mayoritariamente históricos), por esto la solución es extraer los datos de bases de datos funcionales u otro tipo de origen de datos, realizar una transformación si es necesario para evitar inconsistencias o que se ingresen datos erróneos, o que no correspondan, y así luego poder cargarlos en el almacén de datos o base de datos respectiva.

PDI cuenta con un conjunto de herramientas, cada una de ellas con un objetivo específico (Salinas, 2008):

SPOON: Es una herramienta gráfica que permite el diseño de las transformaciones y trabajos. Spoon incluye opciones para pre-visualizar y testear los elementos (procesos ETL) desarrollados. Es la principal herramienta de trabajo de PDI y con la que se construyen y validan los procesos ETL.

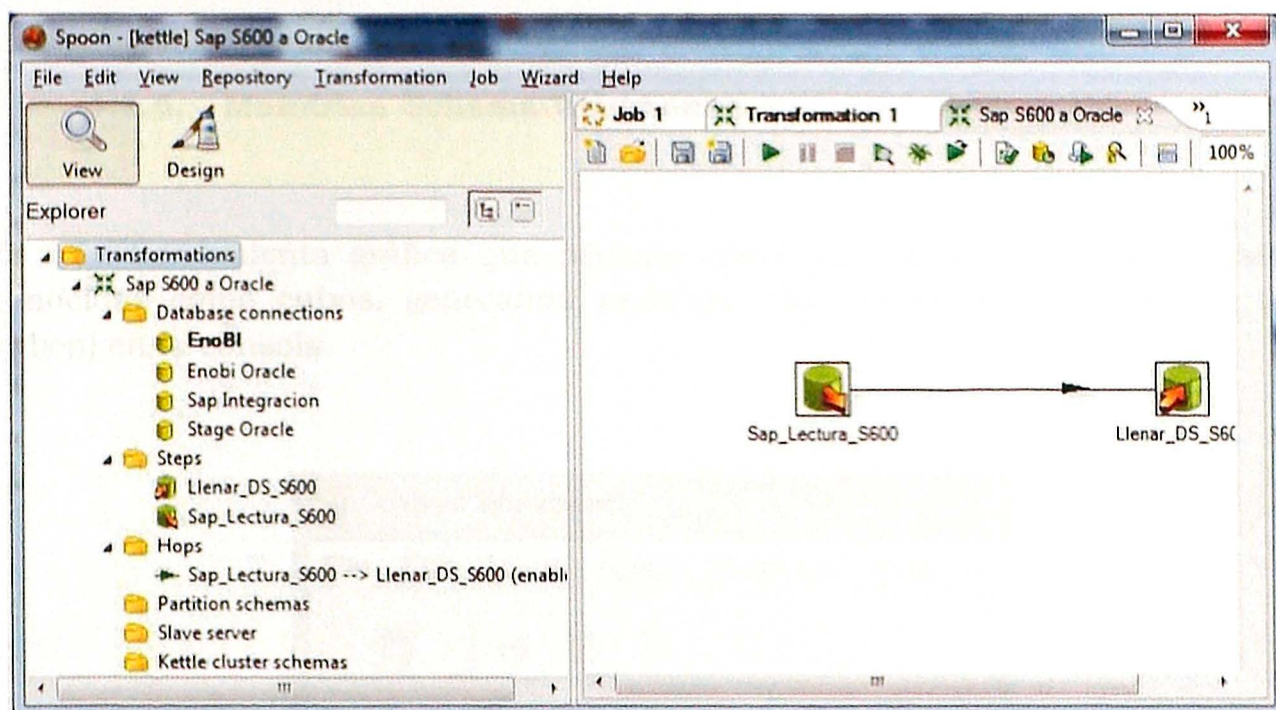


Figura 4. 5. Herramienta Spoon realizando proceso de ETL.

PAN: Es una herramienta que permite la ejecución de las transformaciones diseñadas con Spoon (como un archivo o desde el repositorio creado con Spoon). Pan tiene la ventaja de que permite preparar la ejecución desde la línea de comandos (consola de comandos) a través de scripts. Un motor de transformación de datos.

KITCHEN: Es una herramienta similar a PAN, pero ésta permite ejecutar trabajos "Job", funcionalidad que se realiza en Spoon donde se desarrolla todo el proceso ETL y queda guardado como un trabajo).

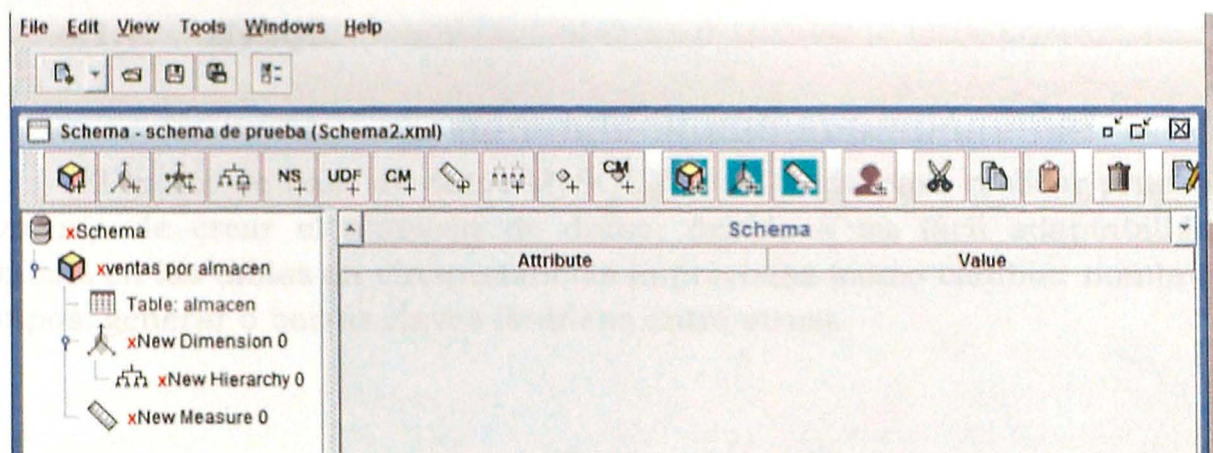


Figura 4. 7. Herramienta Mondrian Schema Workbench creando una nueva consulta multidimensional.

Los cubos diseñados con sus respectivas dimensiones y medidas se deben publicar (subir el cubo como .xml) a Pentaho BI server, a través de la conexión que se establece en la consola de administración PAC (herramienta de Pentaho).

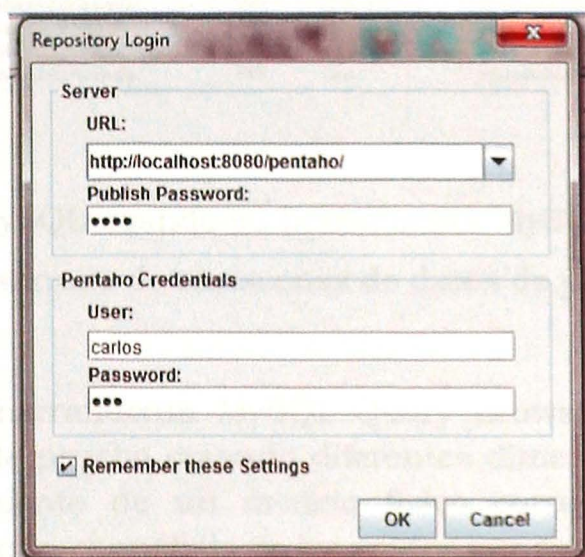
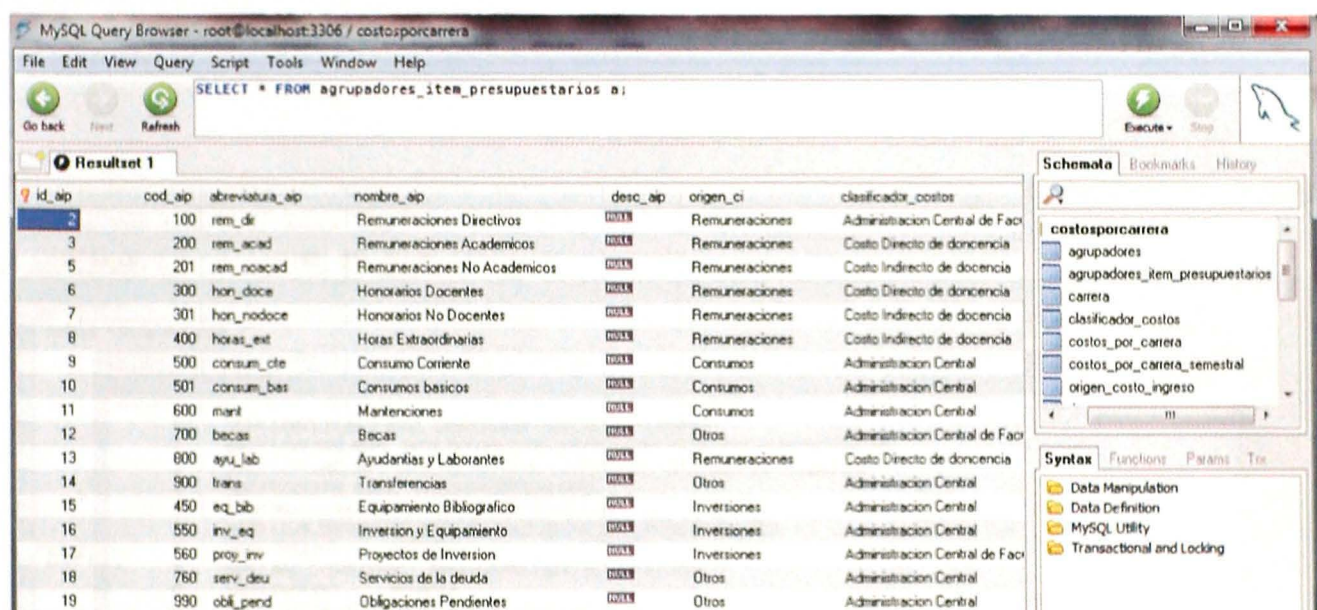


Figura 4. 8. Herramienta Mondrian Schema Workbench conectándose a Pentaho BI server para publicar un cubo.

4.1.7. MySQL.

El motor de base de datos MySQL fue utilizado para realizar pruebas al momento de crear el almacén de datos, debido a su fácil adaptabilidad a cambios en las tablas en circunstancias imprevistas (como cambiar nombres de campos, generar o borrar claves foráneas entre otros).



The screenshot shows the MySQL Query Browser interface. The query bar contains the command: `SELECT * FROM agrupadores_item_presupuestarios a;`. The results are displayed in a table with the following columns: `id_aip`, `cod_aip`, `abreviatura_aip`, `nombre_aip`, `desc_aip`, `origen_ci`, and `clasificador_costos`. The table contains 19 rows of data, including items like 'Remuneraciones Directivos', 'Remuneraciones Academicos', 'Honorarios Docentes', etc. On the right side, the 'Schemata' panel shows the database structure, including tables like 'agrupadores', 'agrupadores_item_presupuestarios', 'clasificador_costos', etc. The 'Syntax' panel at the bottom right lists various SQL functions and utilities.

id_aip	cod_aip	abreviatura_aip	nombre_aip	desc_aip	origen_ci	clasificador_costos
2	100	rem_dir	Remuneraciones Directivos	REMUN	Remuneraciones	Administración Central de Fac
3	200	rem_acad	Remuneraciones Academicos	REMUN	Remuneraciones	Costo Directo de docencia
5	201	rem_noacad	Remuneraciones No Academicos	REMUN	Remuneraciones	Costo Indirecto de docencia
6	300	hon_doce	Honorarios Docentes	REMUN	Remuneraciones	Costo Directo de docencia
7	301	hon_nodoce	Honorarios No Docentes	REMUN	Remuneraciones	Costo Indirecto de docencia
8	400	horas_ext	Horas Extraordinarias	REMUN	Remuneraciones	Costo Indirecto de docencia
9	500	consum_cte	Consumo Corriente	CONSUM	Consumos	Administración Central
10	501	consum_bas	Consumos basicos	CONSUM	Consumos	Administración Central
11	600	mant	Mantenciones	CONSUM	Consumos	Administración Central
12	700	becas	Becas	OTROS	Otros	Administración Central de Fac
13	800	ayud_lab	Ayudantías y Laborantes	REMUN	Remuneraciones	Costo Directo de docencia
14	900	trans	Transferencias	OTROS	Otros	Administración Central
15	450	eq_bib	Equipamiento Bibliografico	INVER	Inversiones	Administración Central
16	550	inv_eq	Inversion y Equipamiento	INVER	Inversiones	Administración Central
17	560	proy_inv	Proyectos de Inversion	INVER	Inversiones	Administración Central de Fac
18	760	serv_deu	Servicios de la deuda	OTROS	Otros	Administración Central
19	990	oblig_pend	Obligaciones Pendientes	OTROS	Otros	Administración Central

Figura 4. 9. Motor MySQL usando su herramienta MySQL Query Browser para el desarrollo de almacenes de datos de prueba.

Utilizando la herramienta MySQL Query Browser se desarrollaron los almacenes de datos de prueba creando diferentes dimensiones y medidas para revisar el funcionamiento de un modelo físico correcto y adecuado a las necesidades que requería el análisis de los costos por carrera anuales.

4.1.8. Sybase.

Para el desarrollo del almacén de datos definitivo y el poblamiento de éste se utiliza el motor de base de datos Sybase el cual es el utilizado en la Dirección de Informática de la Universidad de La Frontera.

Para trabajar con el almacén de datos se utiliza la herramienta SQL Advantage, la cual permite conectarse a los servidores de manera remota para realizar los procesos pertinentes al poblamiento, generar cambios y analizar dificultades y posibles errores.

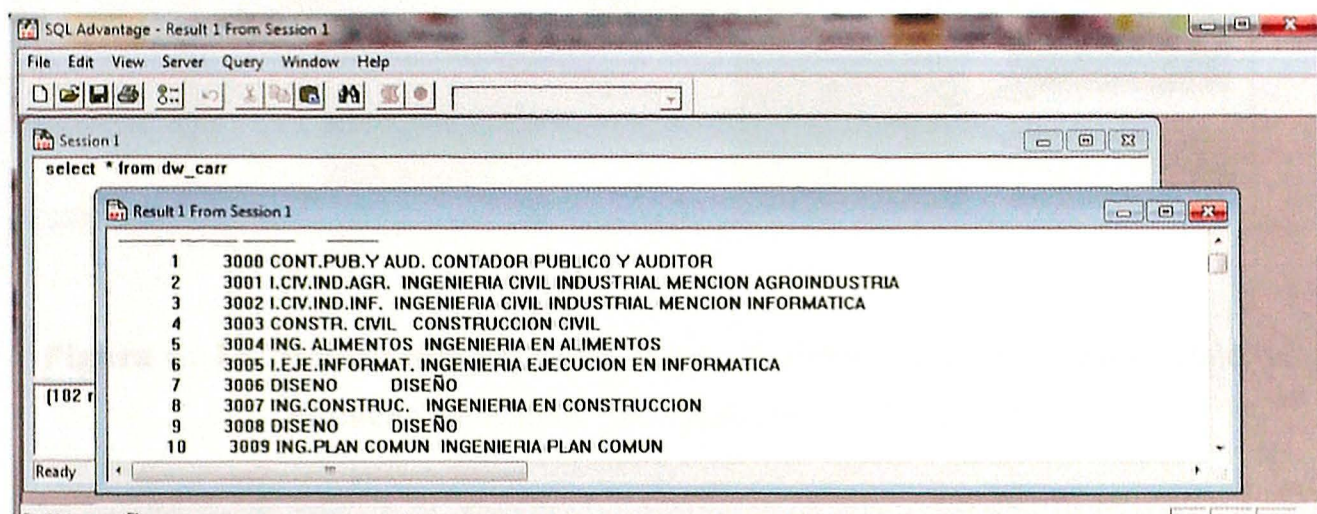


Figura 4. 10. Motor Sybase utilizando la herramienta SQL Advantage para la conexión remota a un servidor y trabajar con un almacén de datos.

4.1.9. Sybase Power Designer.

Herramienta de Sybase que permite diseñar modelos de datos relacionales, físicos, ya sea para bases de datos relaciones, transaccional, almacenes de datos. Desarrollando el modelo de datos se puede crear directamente la base de datos a través de herramientas, conectándose a

servidores, generar archivos XML, generar código SQL para personalizarlo y escribir directamente las sentencias para crear las tablas; entre muchas otras posibilidades que ofrece Sybase Power Designer.

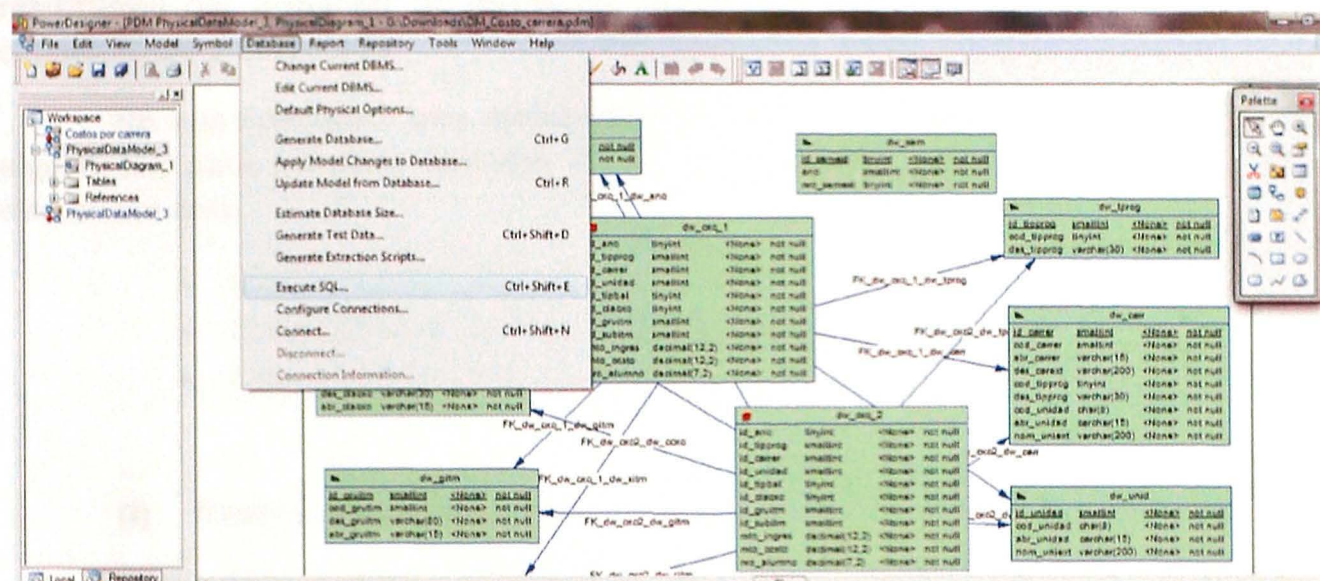


Figura 4. 11. Herramienta Sybase Power Designer generando un modelo de almacén de datos, para generar el código SQL.

4.1.10. Linux.

Durante el proceso de investigación, desarrollo de almacenes de datos de prueba, implementación de las distintas herramientas de Pentaho y sus configuraciones respectivas, se utilizó sistemas operativos Linux en sus distribuciones OpenSUSE 11.3 y Ubuntu 10.04.

Debido a que los servidores de la Dirección de Informática utilizan sistemas operativos Linux, el investigar y realizar la instalación de todas las herramientas de manera local debía ser en los mismos entornos para poder detectar errores y darles solución en caso de generarse en los servidores.

4.2. Algoritmos de obtención de Costos por Carrera.

Para la obtención de los costos por carreras y todos los datos necesarios para poblar el almacén de datos, se plantearon algoritmos de obtención de datos. Estos algoritmos cumplieron un papel fundamental para que los resultados del almacén de datos en cuanto al contenido de sus datos sean consistentes y correctos.

Se consideraron tres elementos de costos en donde se plantearon los algoritmos para calcular factores (f) de prorrateo para estos elementos. Estos elementos son:

- Costos Administrativos.
- Costo Directo de Docencia (por departamento)
- Costo Indirecto de Docencia.

(i) Costos Administrativos.

Para los costos administrativos se consideraron los siguientes algoritmos:

a. Centro de Costo de Carreras o grupos de Carreras.

Se calcula un factor (f_{GC}) que se planteó como:

$$(f_{GC}) = \frac{nro_al_carr}{nro_tot_al_gru_carr}$$

Siendo nro_al_carr el número de alumnos de una carrera, y $nro_tot_al_gru_carr$ el número total de alumnos de un grupo de carreras.

b. Centro de Costo de Facultad.

Se calcula el factor (f_F) como:

$$(f_F) = \frac{nro_al_carr}{nro_al_fac}$$

En donde nro_al_carr es el número de alumnos de una carrera; y nro_al_fac el número de alumnos de la Facultad a la cual la carrera corresponde.

c. Centro de Costo Global.

Se calcula el factor (f_U) como:

$$(f_U) = \frac{nro_al_car}{nro_alu_ufro}$$

Siendo nro_al_carr el número de alumnos de una carrera; y nro_alu_ufro el número de alumnos de la Universidad de La Frontera.

(ii) Costo Directo de Docencia.

Se calculó el costo directo por departamento que imparte clases a las carreras de la Universidad de La Frontera, el costo directo por departamento determinado por:

$$\sum Ctos_Asig f(horas)$$

Siendo Cto_Asig los costos de una asignatura de un departamento. El factor (f_{CA}) de una carrera_i con una asignatura_j está determinado por:

$$(f_{CA}) = \frac{nro_alu_carrerai_asignaturaj}{nro_alu_asignaturaj}$$

Siendo $nro_alu_carrerai_asignaturaj$ la cantidad de alumnos de una carrera_i con una asignatura_j. Y $nro_alu_asignaturaj$ el número total de alumnos de la asignatura_j.

(iii) Costo Indirecto de Docencia.

Se calculó un factor (f_{ID}) determinado por:

$$(f_{ID}) = \frac{nro_alu_atend_carr}{nro_total_alu_atend}$$

Siendo `nr_alu_atend_carr` el número de alumnos de una carrera atendidos por un departamento. Y `nro_total_alu_atend` el número total de alumnos atendidos por ese departamento.

4.3. Fuentes de Datos de Origen.

Para poblar tanto la base de datos relacional de paso (ODS) que se crea para los procesos ETL como el almacén de datos, se obtienen los datos necesarios de distintas fuentes de origen las cuales son:

(i) Sistema Presupuestario.

Del Sistema Presupuestario de la Universidad de La Frontera se extraen:

- Centros de Costos.
- Ítemes de Costos.
- Gastos del año (compromisos).

(ii) Sistema Contable.

De este sistema se obtienen:

- Depreciaciones por año calculadas sobre los bienes de la Universidad.

(iii) Sistema Académico.

Del Sistema Académico de la Universidad se obtiene:

- Carreras (Pregrado) y Programas (Postgrado, etc.).
- Unidades
- Tipos de Programa.
- Alumnos por carrera y año.
- Alumnos por asignatura y carrera por año.
- Asignaturas dictadas por año.

Todos estos datos se obtienen desde fuentes operativas de datos, sistemas en funcionamiento de la Universidad, pero también se planteó la obtención de otros datos para poblar el ODS que es la base de datos relacional normalizada de paso o intermedia entre las fuentes de datos de origen y el almacén de datos.

- Origen de los Costos por Carrera.

Además se mapeó los ítemes para relacionarlos con los clasificadores de costos y los agrupadores de ítemes. Por último se planteó la relación de los clasificadores de costos con centros de costos y las carreras.

4.4. Modelo del Almacén de Datos.

Una vez realizada la investigación de Pentaho y todas sus herramientas para utilizarlas en el almacén de datos y así implementar Business Intelligence, se diseñó el modelo de la base de datos del almacén de datos.

Se creó un modelo inicial para analizar su estructura y la manera de poder acceder a los datos, la obtención de estos, el detalle de éstos y la velocidad de respuesta utilizando consultas OLAP.

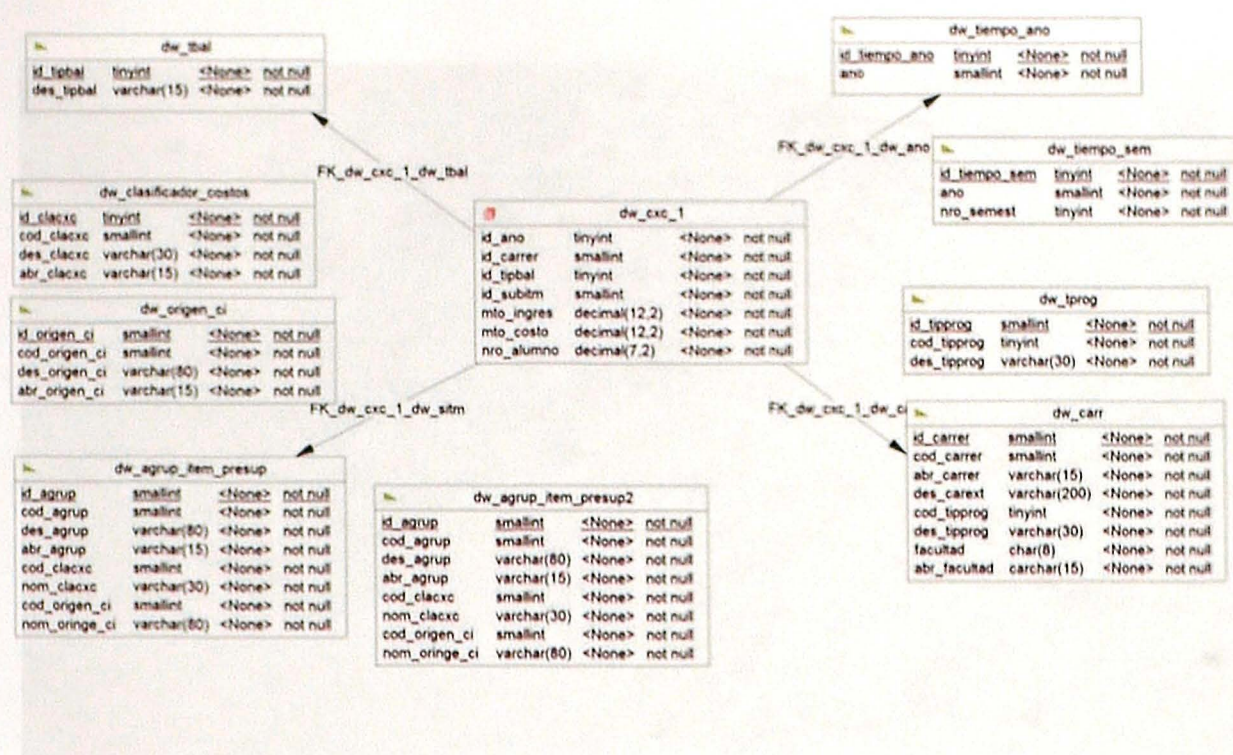


Figura 4. 12. Primer modelo de datos de prueba del almacén de datos piloto para análisis de costos por carrera.

Este modelo se utilizó para crear una base de datos en MySQL para probar su estructura y ver si funcionaba para analizar los costos por carrera de la Universidad de La Frontera. En base a este modelo se fueron generando distintas propuestas, logrando así que este modelo sea más robusto.

4.5. Instalación de Pentaho BI Server y Pentaho PAC.

La instalación de la plataforma Pentaho BI Server y de la herramienta Pentaho Administration Console (PAC) se realizó de forma local para ir realizando pruebas y ver posibles errores que se podían generar al momento de instalar y configurar estas herramientas. A medida que se instalaron de manera correcta las herramientas, se fueron instalando las mismas en un servidor de desarrollo de la Dirección de Informática de la Universidad de La Frontera.


```
carlos@linux-eobz:/opt/pentaho> cd biserver-ce/
carlos@linux-eobz:/opt/pentaho/biserver-ce> ./start-pentaho.sh
/opt/pentaho/biserver-ce
DEBUG: Using JAVA_HOME
DEBUG: _PENTAH0_JAVA_HOME=/usr/lib/jvm/jdk1.6.0_21
DEBUG: _PENTAH0_JAVA=/usr/lib/jvm/jdk1.6.0_21/bin/java
/opt/pentaho/biserver-ce/data
DEBUG: Using JAVA_HOME
DEBUG: _PENTAH0_JAVA_HOME=/usr/lib/jvm/jdk1.6.0_21
DEBUG: _PENTAH0_JAVA=/usr/lib/jvm/jdk1.6.0_21/bin/java
classpath is :./lib/hsqldb-1.8.0.jar
Using CATALINA_BASE: /opt/pentaho/biserver-ce/tomcat
Using CATALINA_HOME: /opt/pentaho/biserver-ce/tomcat
Using CATALINA_TMPDIR: /opt/pentaho/biserver-ce/tomcat/temp
Using JRE_HOME: /usr/lib/jvm/jdk1.6.0_21
carlos@linux-eobz:/opt/pentaho/biserver-ce> [Server@13582d]: [Thread(main,5,main)]:
[Server@13582d]: [Thread(main,5,main)]: checkRunning(false) exited
[Server@13582d]: Startup sequence initiated from main() method
[Server@13582d]: Loaded properties from [/opt/pentaho/biserver-ce/data/server.properties]
[Server@13582d]: Initiating startup sequence...
[Server@13582d]: Server socket opened successfully in 137 ms.
```

Figura 4. 13. Ejecución de la Plataforma de BI de Pentaho en Linux.

Al ejecutar el script “start-pentaho.sh” el sistema configura las variables de entorno y prepara la plataforma web de Business Intelligence.

Similar es el proceso que se realiza para ejecutar la consola de administración; se ejecuta otro script de otra ruta, cuyo directorio es donde se encuentra la consola de administración para crear cuentas de usuario y conexiones a bases de datos.


```

carlos@linux-eobz ~$ cd /opt/pentaho/administration-console/
carlos@linux-eobz /opt/pentaho/administration-console$ ./start-pac.sh
/opt/pentaho/administration-console
DEBUG: Using JAVA_HOME
DEBUG: _PENTAHO_JAVA_HOME=/usr/lib/jvm/jdk1.6.0_21
DEBUG: _PENTAHO_JAVA=/usr/lib/jvm/jdk1.6.0_21/bin/java
2010-12-30 10:47:30.806 INFO Logging to STDERR via org.mortbay.log StdErrLog
30-dic-2010 10:47:31 org.pentaho.pac.server.JettyServer startServer
INFO: Console is starting
2010-12-30 10:47:31.574 INFO jetty-6.1.2
2010-12-30 10:47:31.897 INFO Started SocketConnector @ 0.0.0.0:8099
30-dic-2010 10:47:31 org.pentaho.pac.server.JettyServer startServer
INFO: Console is now started. It can be accessed using http://linux-eobz.site:8099 or http://127.0.0.2:8099

```

Figura 4. 14. Ejecución de la Consola de Administración de Pentaho en Linux.

Para acceder a la plataforma de BI de Pentaho, en el navegador se tiene que ir a la ruta (URL del servidor:8080) por defecto. Como en una primera instancia se trabaja con Pentaho de forma local, la URL para acceder es:

`http://localhost:8080/pentaho`

Y para acceder a la consola de administración de Pentaho para la creación de cuentas y conexiones a bases de datos la ruta es similar pero cambiando el puerto que por defecto es 8099. Por lo tanto en la ruta local se accedió a:

`http://localhost:8099`

Con las plataformas corriendo exitosamente se realizó el proceso de crear un usuario para acceder al BI Server y crear la conexión con su configuración para el almacén de datos.

4.6. Creación de usuario para Pentaho y creación de la conexión con el Almacén de Datos.

Con la instalación finalizada de Pentaho BI server y el Pentaho Administrator Console (PAC), se crearon cuentas de usuarios con los permisos respectivos para el acceso a la plataforma de Business Intelligence. Por otra parte, en el PAC se creó la conexión a la base de datos que en un comienzo fue local para datos de prueba en una primera instancia con el primer almacén de datos diseñado.

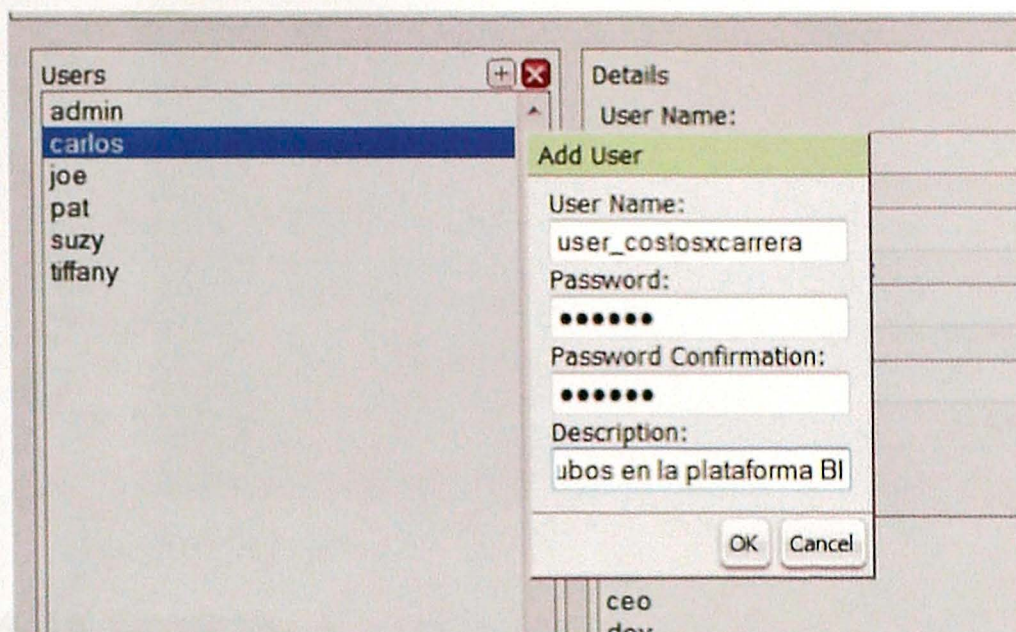


Figura 4. 15. Creación de un usuario para entrar a la plataforma de pentaho BI Server.

Por defecto se encuentran algunos usuarios con los cuales se puede acceder a la plataforma de de Business Intelligence de Pentaho. Una vez creado el usuario, se creó la conexión a la base de datos local de prueba en MySQL, la cual sirve para posteriormente publicar en la plataforma las consultas multidimensionales creadas en Mondrian Schema Workbench.

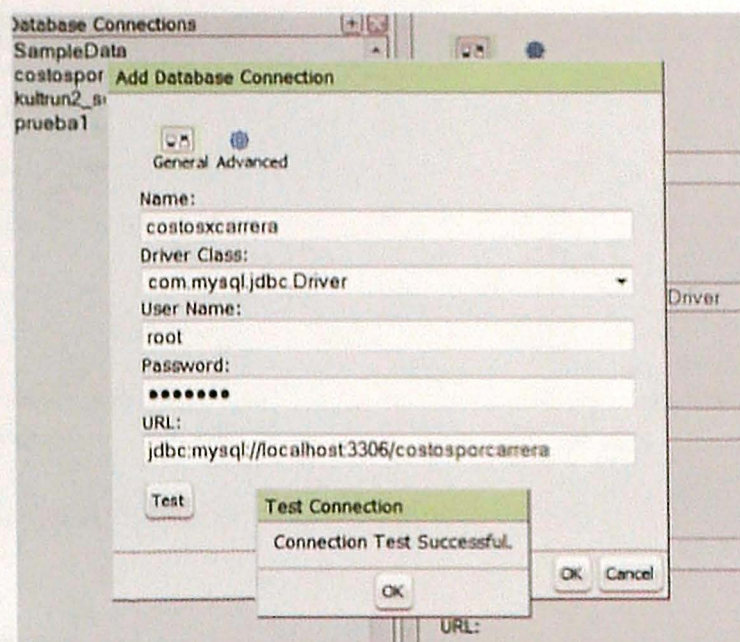


Figura 4. 16. Creación de la conexión a la base de datos del almacén de datos local.

4.7. Creación de primeras consultas Multidimensionales (CUBOS).

Con la herramienta de Mondrian Schema-Workbench habilitada, se creó la primera consulta multidimensional de análisis para el almacén de datos de prueba creado en base al primer modelo de datos.

En "Tools"-> "Connection", se abre el formulario para ingresar los datos de conexión (Nombre de la clase del Driver, URL de la conexión, nombre de usuario de la base de datos y su contraseña).

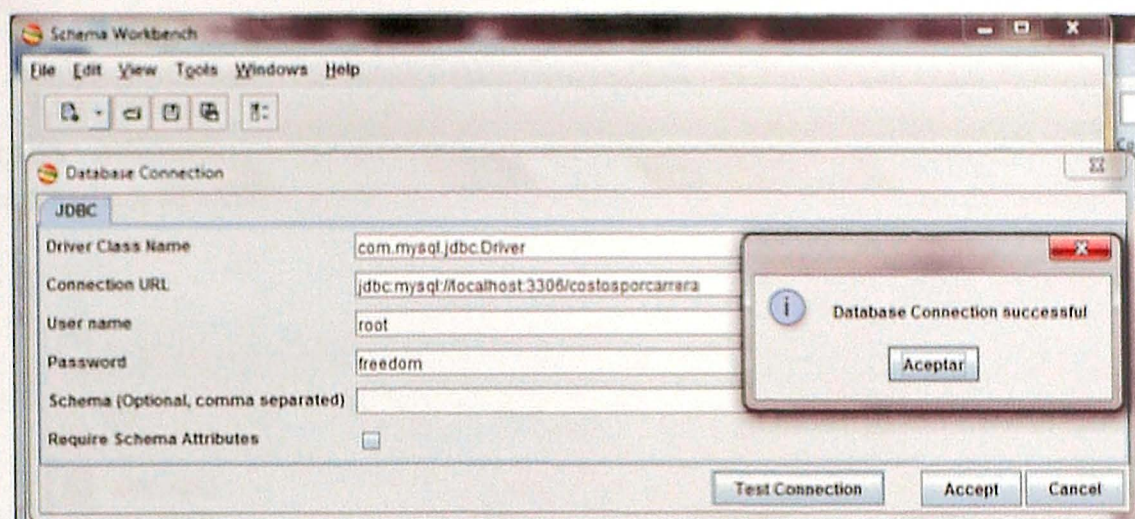


Figura 4. 17. Conexión a la base de datos local con Mondrian Schema Workbench para creación de consulta multidimensional.

Con la conexión exitosa a la base de datos se crea la consulta multidimensional, en donde se crea un nuevo esquema, para luego crear un cubo. En este cubo se pueden crear las dimensiones del almacén de datos en donde se define las distintas jerarquías con sus respectivos niveles. Además se deben ingresar las medidas que se necesitan para armar el cubo.

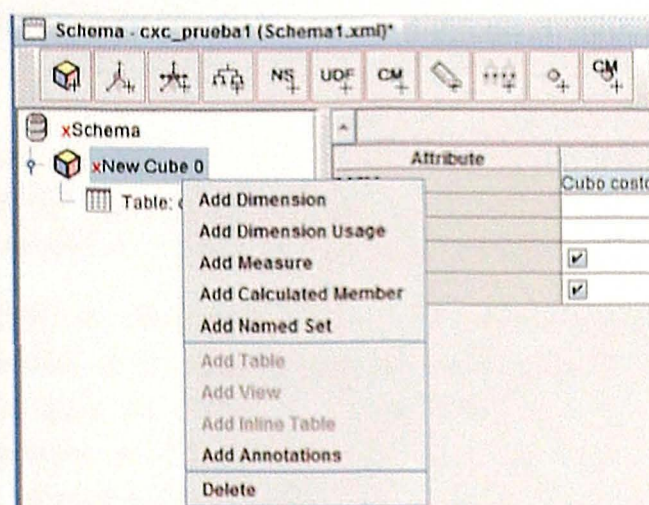


Figura 4. 18. Creación de un nuevo esquema para generar un cubo.

Una vez que se crea la nueva dimensión (de Carreras), se debe crear una jerarquía mínima, y para ésta, también se debe asignar una tabla que corresponde a la tabla de dimensión.

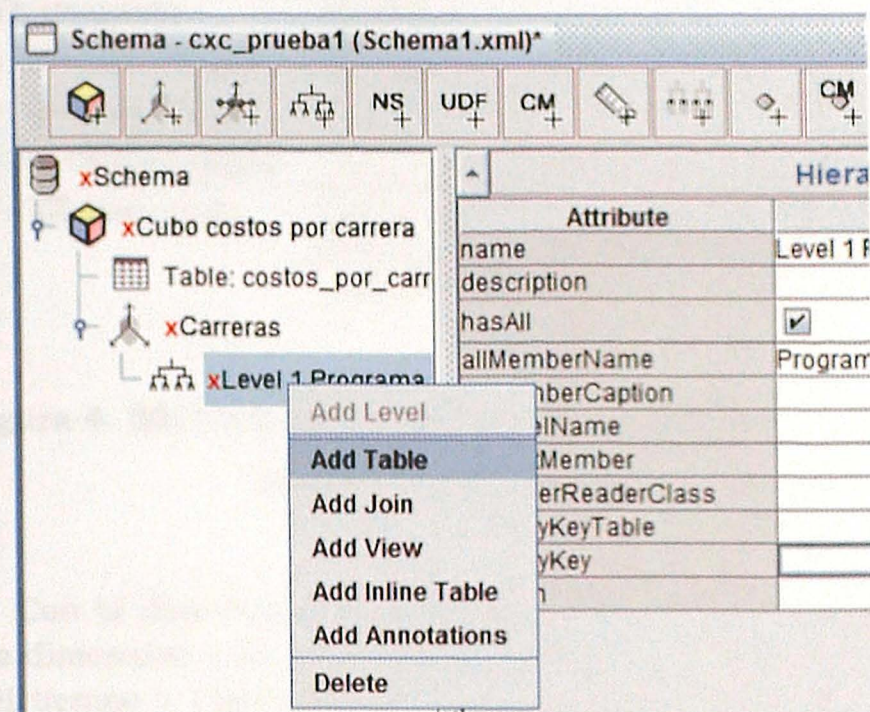


Figura 4. 19. Creando y configurando las dimensiones del almacén de datos.

Para este caso la tabla que se le asigna es la tabla "carreras" la cual contiene el identificador, un código, el nombre, la facultad a la cual pertenece y el tipo de programa al que corresponde (pregrado, postgrado, etc.).

Primero se crea la dimensión que se llama Carreras, luego se crea la jerarquía, aquí es donde se le asigna la tabla que corresponde a su dimensión, luego para cada jerarquía se crean los diferentes niveles que serán los que se desglosan en la consulta multidimensional por ejemplo: año, mes, día, si el tiempo fuese una dimensión donde se necesitara más detalle, para la dimensión de los ítemes de costos, los niveles de la jerarquía son: Programa, Facultad, Carrera.

Level for 'Programas' Hierarchy	
Attribute	Value
name	Level 2 Carrera
description	
table	
column	carrera->nombre_carrera - VARCHAR(30)
nameColumn	carrera->abreviatura_carrera - VARCHAR(5)
parentColumn	carrera->cod_carrera - INTEGER(10)
nullParentValue	carrera->desc_carrera -
ordinalColumn	carrera->facultad - VARCHAR(50)
type	carrera->id_carrera - INTEGER(10)
uniqueMembers	carrera->nombre_carrera - VARCHAR(30)
levelType	carrera->programa - VARCHAR(30)
hideMemberIf	
approxRowCount	

Figura 4. 20. Creando y configurando los niveles de la jerarquía en una dimensión, para este caso Carreras.

Con la dimensión de carreras configurada, se realizó el mismo proceso para la dimensión tiempo y para la dimensión de Ítemes. Siendo los niveles año para el tiempo y Clasificador de costos, Origen de costo y además ítem como niveles de la dimensión de los ítemes presupuestarios.

Con esto, el proceso de definir las dimensiones, con la creación de éstas y la configuración para que se hagan bien las consultas multidimensionales del almacén de datos y así mostrar datos en el cubo quedó finalizado, dando paso a la creación de las medidas.

Costos por carrera

Los niveles de la dimensión de carreras se configuraron en la dimensión de carreras. Para ello se usó el siguiente código SQL:

Mientras que para la dimensión de tiempo se configuraron los niveles año, semestre y trimestre. Para ello se usó el siguiente código SQL:

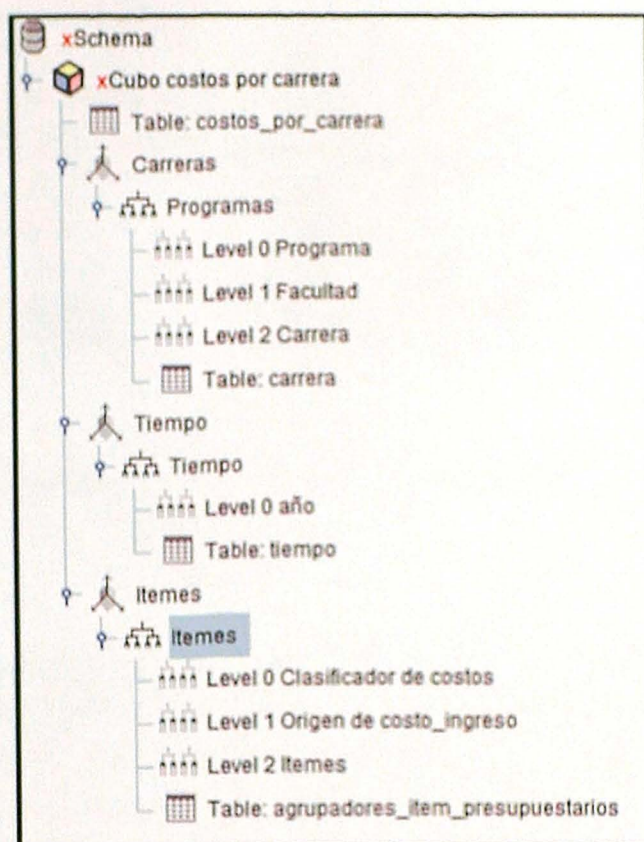


Figura 4. 21. Creación de las dimensiones con sus respectivas jerarquías y niveles para un cubo de costos por carrera.

En este modelo de datos de prueba, la tabla de hechos “costosporcarrera”, tiene como atributos de medida los campos: “ingreso”, “costo” y “nro_alumnos”. Se crearon nuevas medidas en el cubo “Cubo de costos por carrera”.

Los niveles de la jerarquía son fundamentales para el desglose de los registros en la consulta multidimensional que se carga en la plataforma de Pentaho una vez finalizado y publicado el cubo.

Mientras más niveles, más se sabe el detalle de los registros; dependiendo de la variedad de estos niveles, se pueden tener grandes desgloses, lo que ayuda a la toma de decisiones a un nivel más específico gracias a los detalles de cada rama de una jerarquía.

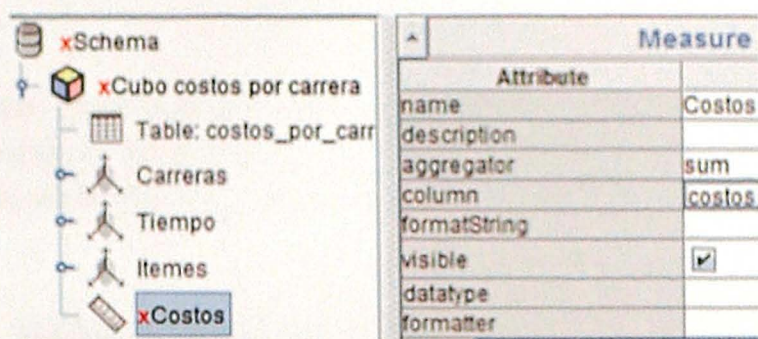


Figura 4. 22. Creación de la medida “Costos” para la consulta multidimensional.

Se Agregó la medida “Costos” la cual lleva el agregador “sum” para sumar los valores de los diferentes registros de la consulta multidimensional correspondiente a los gastos de los distintos ítemes, de cada carrera por cada año. También se agrega le medida “ingreso” de la misma manera. Lo mismo se realizó para la medida “nro_alumnos” con la diferencia que el agregador es “avg” (promedio), ya que va a contar solamente los números de alumnos para las distintas carreras para los distintos años registrados en el almacén de datos.

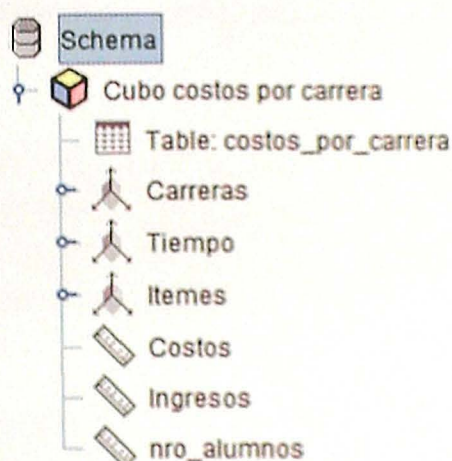


Figura 4. 23. Consulta multidimensional finalizada con dimensiones y medidas.

El siguiente paso fue publicar el cubo en la plataforma de Pentaho BI Server, con los datos de la conexión y los datos del usuario creado en la consola de administración de Pentaho.

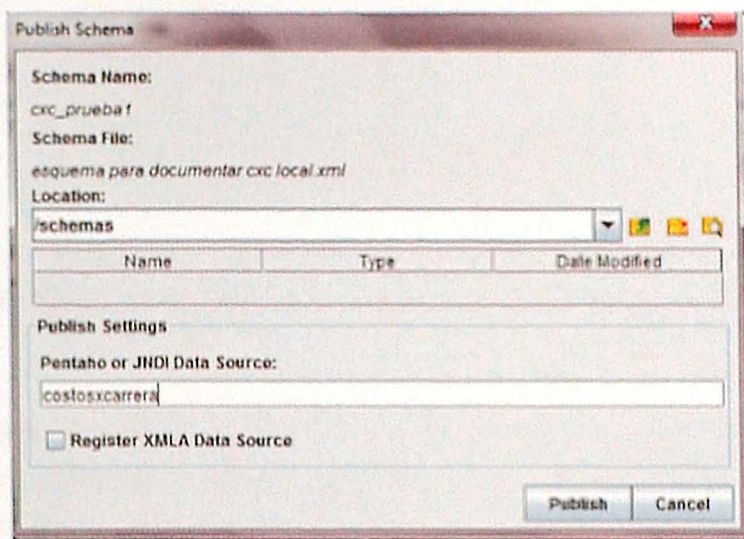


Figura 4. 24. Publicación del Cubo a la plataforma de BI de Pentaho.

4.8. Análisis de Cubo para modelo de prueba de costos por carrera.

Con el análisis publicado, se accedió a la plataforma de BI de Pentaho para visualizar el cubo y analizar los datos y observar si los datos obtenidos son claros y para ver qué otros datos se pueden obtener, o modificar las tablas de dimensiones agregando campos o creando otras tablas para poder entregar más detalles en las consultas.

Primero se cargó el esquema diseñado con su respectivo cubo:

El esquema de nombre `cxc_prueba1` es el esquema realizado con Mondrian Schema Workbench en donde se creó el cubo. El cubo de nombre “Cubo costos por carrera” está diseñado para mostrar las consultas multidimensionales desde el almacén de datos local de prueba.

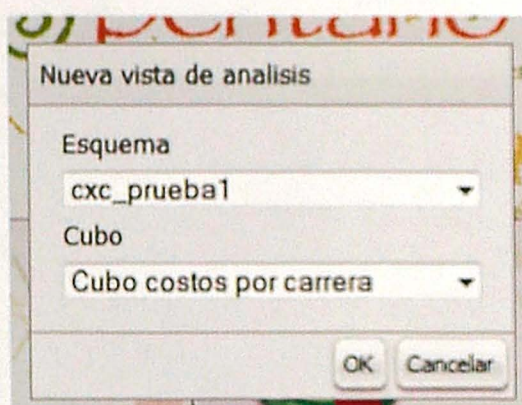


Figura 4. 25. Selección del cubo del almacén de datos de prueba local.

Con la vista de análisis de la consulta multidimensional, se cargan los registros correspondientes a todas las medidas y las dimensiones, las cuales se van desglosando para ver en detalle los valores correspondientes para los distintos años y programas por los distintos ítemes de costos.

Nueva vista de ana..			Medidas		
Programas	Tiempo	Ítemes	• Costos	• Ingresos	• nro_alumnos
Programa	Año	Ítemes	1.623.302.178	0	2.226
		Administracion Central de Facultad	1.447.372.500	0	1.480
		Costo Indirecto de docencia	163.584.000	0	566
Pregrado.	Año	Ítemes	1.623.302.178	0	2.226

Figura 4. 26. Visualización del cubo con dimensiones y medidas.

Nueva vista de ana...

Ingenieria Informatica	2004	Itemes	18.343.100	0	204
	Año	Itemes	263.057.200	0	240
	2004	Itemes	263.057.200	0	240
		Administracion Central de Facultad	263.057.200	0	240
Ingenieria Matematica		Remuneraciones	263.057.200	0	240
		Remuneraciones Directivos	263.057.200	0	240
	Año	Itemes	752.995.678	0	1.312
	2007	Itemes	171.349.000	0	240
	2008	Itemes	178.749.000	0	246
	2009	Itemes	191.026.000	0	286
		Administracion Central de Facultad	148.820.000	0	143
		Remuneraciones	148.820.000	0	143
		Remuneraciones Directivos	148.820.000	0	143
		Costo Indirecto de docencia	42.206.000	0	143
		Remuneraciones	42.206.000	0	143
		Remuneraciones No Academicos	42.206.000	0	143
	2010	Itemes	211.871.678	0	540

Figura 4. 27. Visualización del cubo con dimensiones y medidas.

En la Figura 4.27. se puede visualizar el cubo más desglosado para ver más detalles (los distintos niveles) de las distintas dimensiones y los valores de las medidas asociados a éstas. Se puede apreciar también en la figura, que se presenta el desglose de las remuneraciones de directivos y no académicos con diferentes tipos de origen de costos como son Administración Central de Facultad como para Costo indirecto de docencia, para el año 2009 en la Carrera de Ingeniería Matemáticas que cuenta con 143 alumnos. Estos datos de prueba se ingresaron sin tener aún una referencia aproximada de los valores reales.

Con el primer prototipo de prueba del almacén de datos de tipo local, se procedió a depurar el modelo de datos para dejarlo más robusto, generando más conexiones entre tablas de hechos con las tablas de dimensiones, creando otras tablas, entre otros.

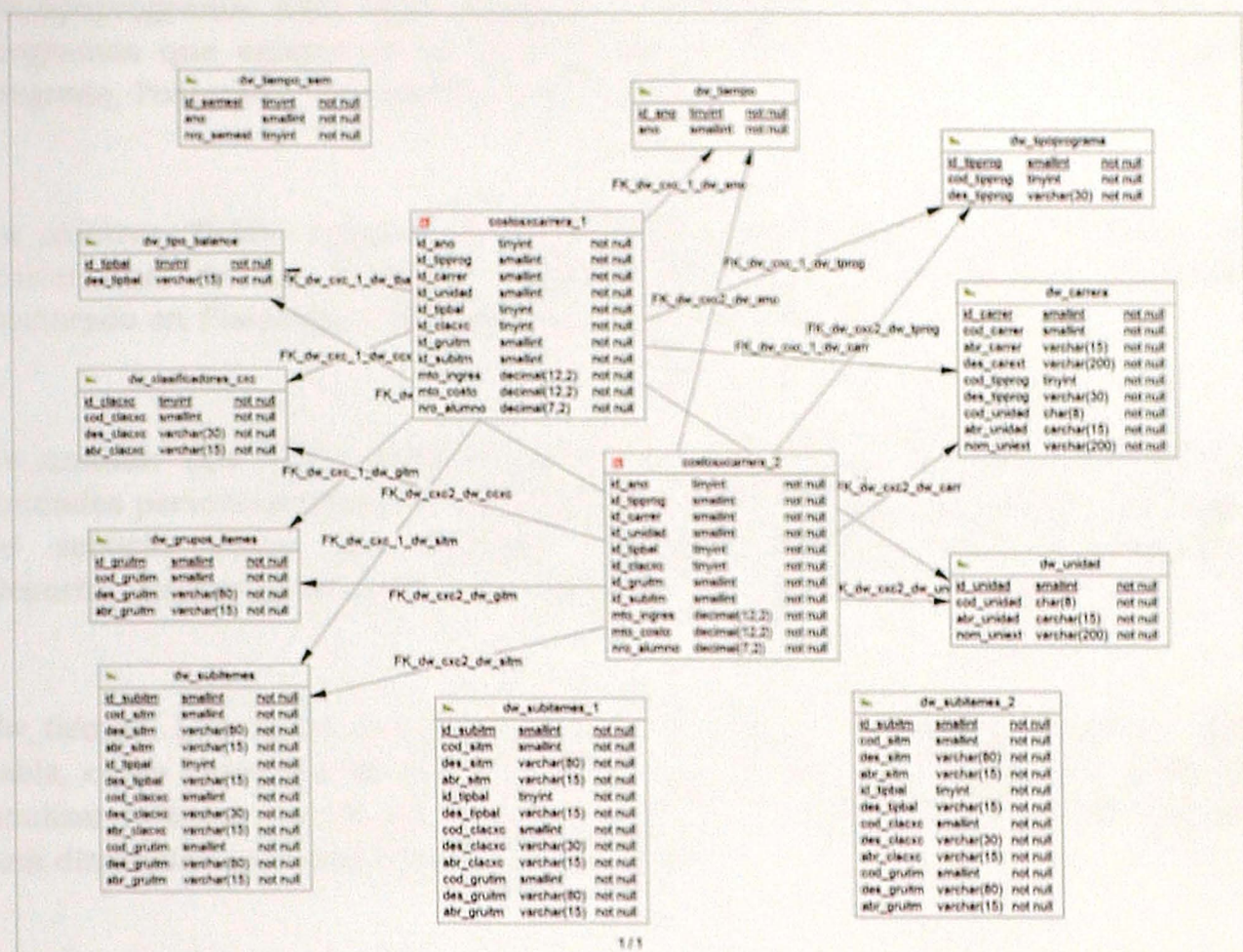


Figura 4. 28. Modelo de datos del almacén de datos piloto para análisis de costos por carrera en su segunda versión.

En la figura 4.28. se observa el modelo anterior, con cambios en su estructura. Este nuevo modelo estrella considera dos tablas de hechos y once tablas de dimensiones, sin embargo, para este piloto se están considerando ocho tablas conectadas a las tablas de hechos. En este esquema las “Puntas” de la estrella son las tablas de dimensiones que son atributos relativos a las variables, también se puede considerar como perspectivas de análisis, o bien la información complementaria a cada uno de los registros de la tabla de hechos:

dw_tipoprograma: Esta tabla contiene datos relacionados a todos los tipos de programas que existen en la Universidad de La Frontera, como por ejemplo Pregrado, Postgrado, Formación Continua, etc.

dw_carrera: Tabla relativa a las carreras o programas que imparte la Universidad de La Frontera. Ejemplo: Ingeniería Informática, Psicología, Doctorado en Fisiología y Biología Molecular Vegetal, etc.

dw_unidad: Los datos almacenados en esta tabla se refieren a todas las unidades pertenecientes a la Universidad de La Frontera que prestan algún tipo de servicio como por ejemplo: Rectoría, Dirección de Investigación, Departamentos Académicos, Direcciones de Carreras, etc.

dw_tiempo: Los datos que contiene son los distintos años transcurridos. Esta tabla, como todas las demás, cumple un rol importante, ya que se si se desea analizar información de los costos a través de los distintos años se debe tener una dimensión que corresponda al tiempo (año).

dw_clasificadores_cxc: Tabla que contiene los clasificadores propios de los costos por carrera, por ejemplo: Costo Directo de Docencia, Costo de Administración Central, Costo de Administración de la Facultad, etc.

dw_grupos_items: Aquí se almacenan los distintos grupos de ítems para costos por carrera, a partir del plan de cuentas presupuestarias.

dw_subitemes: Similar a dw_grupos_items, esta tabla contiene los subítemes de los costos por carrera, a partir del plan de cuentas presupuestarias de la Universidad.

Todas estas tablas están conectadas a dos tablas de hechos que forman el centro del esquema estrella en el modelo del almacén de datos. Estas tablas de hechos son: "costoporcarrera_1" y "costoporcarrera_2", las cuales obtienen el Desarrollo de Piloto de Almacén de Datos para Análisis de Costos por Carrera anuales de la Universidad de La Frontera

valor del identificador o clave primaria de cada tabla de dimensión y además contienen los valores numéricos asociados a cada tabla (medidas). Para este caso los campos correspondientes a costos, ingresos y número de alumnos por carrera son las medidas (valores) de las distintas dimensiones.

costosxcarrera_1			costosxcarrera_2		
id_ano	tinyint	not null	id_ano	tinyint	not null
id_tipprog	smallint	not null	id_tipprog	smallint	not null
id_carrer	smallint	not null	id_carrer	smallint	not null
id_unidad	smallint	not null	id_unidad	smallint	not null
id_tipbal	tinyint	not null	id_tipbal	tinyint	not null
id_clacxc	tinyint	not null	id_clacxc	tinyint	not null
id_gruitm	smallint	not null	id_gruitm	smallint	not null
id_subitm	smallint	not null	id_subitm	smallint	not null
mto_ingres	decimal(12,2)	not null	mto_ingres	decimal(12,2)	not null
mto_costo	decimal(12,2)	not null	mto_costo	decimal(12,2)	not null
nro_alumno	decimal(7,2)	not null	nro_alumno	decimal(7,2)	not null

Figura 4. 29. Tablas de hechos del modelo del almacén de datos (centros del esquema en estrella).

En este modelo se tienen en consideración los siguientes casos:

- Las tablas de hechos consideran: La primera considera gastos en inversiones y la segunda considera la depreciación en lugar de la inversión.
- Se podría crear una tabla de hechos que separe la medida de ingresos de las tablas de hechos existentes, de tal manera de dejar los ingresos independientes de los costos.

4.9. Modelo de Datos de Traspaso para preparación de Proceso ETL

Ahora, con la segunda versión del modelo del almacén de datos, se diseñó un modelo de datos relacional, cuya base de datos almacena datos de traspaso extraídos de otras fuentes de datos, como lo son aquellos que maneja la Dirección de Finanzas de la Universidad de La Frontera.

Esta base de datos de traspaso, almacena datos correspondientes a los saldos, y los relacionados a los ítemes de costos.

Estos datos son necesarios para hacer el prorrateo de costos de los diferentes ítemes en los distintos tipos de programa que ofrece la Universidad.

4.10. Proceso de ETL para el Poblamiento del Almacén de Datos.

Con la herramienta Pentaho Data Integration (PDI o Kettle), se realizó el proceso de ETL para extraer los datos de fuentes de datos externas (ajenas al almacén de datos) cuya información permite el análisis de los costos por carrera.

Con la herramienta Kettle disponible, se creó un nuevo trabajo (job) en el cual se agregarán entradas de tablas las cuales son una forma de obtener datos de fuentes que se extraen directamente de tablas de bases de datos provenientes de algún servidor, en este caso de los servidores de la Universidad de La Frontera, cuyas bases de datos contienen los datos de: los diferentes programas de la Universidad, los diferentes tipos de unidades que existen, las carreras que imparte, el tiempo que se considera en la universidad como año para considerar los costos, los distintos clasificadores de costos y grupos de ítemes de costos y los diferentes tipos de ingresos que percibe la Universidad de La Frontera.

Primero se realizó el proceso de ETL para las dimensiones que tienen que ver con carrera, los tipos de programas y unidades.

Configurando el proceso de entrada para las consultas a las fuentes de datos funcionales, como se ve en la siguiente figura se van a obtener los datos correspondientes a las carreras y diferentes tipos de unidades y programas. Debido a que la tabla con datos de las carreras del almacén de datos no se encuentra normalizada para efectos de rapidez en las consultas, debe llevar

asociados los datos referentes a los programas y unidades directamente como campos de la tabla.

El proceso para configurar la tabla con datos de entrada consiste en crear una conexión con la base de datos de origen. Al ser exitosa esta conexión, se elige la tabla de la cual se quieren consultar los datos que serán transformados, si es necesario, para luego cargarlos en la tabla de destino.

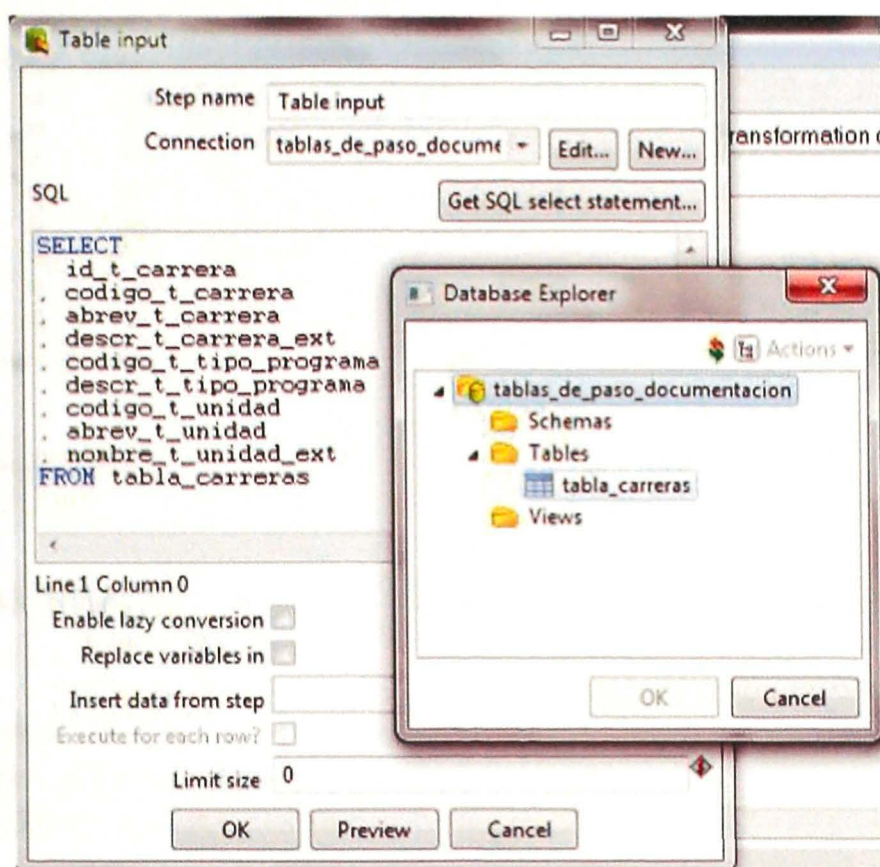


Figura 4. 30. Configuración de tabla de entrada para proceso de ETL de carreras.

Configurando de manera correcta los datos de entrada, se pueden observar los datos desde una vista previa para saber si los resultados que se obtienen son los que se requieren, y detectar a tiempo algún error en la selección de la tabla.

Al seleccionar la tabla, se puede editar la consulta para obtener los registros que se requieren, para el caso de necesitar menos campos de los que existen en la tabla de origen.

O también para hacer filtros de datos si se requiere, por ejemplo, obtener carreras de pregrado que pertenezcan a la Facultad de Ingeniería, Ciencias y Administración.

#	id_t...	codig...	abrev_t_carrera	descr_t_carrera_ext	cod...	descr_t_tipo_progra
63	63	3081	I. MECANICA	INGENIERIA MECANICA	10	Carreras de Pregrad
64	64	3082	ING. AMBIENTAL	INGENIERIA AMBIENTAL	10	Carreras de Pregrad
65	65	3083	LIC.FIS.APLIC.	LICENCIATURA EN FISICA APLICADA	10	Carreras de Pregrad
66	66	3084	ING. MATEMA...	INGENIERIA MATEMATICA	10	Carreras de Pregrad
67	67	3085	ING. COMERCIAL	INGENIERIA COMERCIAL	10	Carreras de Pregrad
68	68	3086	I. INFORMATICA	INGENIERIA INFORMATICA	10	Carreras de Pregrad
69	69	3087	I. ELECTRICA	INGENIERIA ELECTRICA	10	Carreras de Pregrad
70	70	3088	I. ELECTRONICA	INGENIERIA ELECTRONICA	10	Carreras de Pregrad

Figura 4. 31. Vista Previa de una consulta a una tabla de origen en KETTLE.

De forma similar se agregó a este trabajo (job de KETTLE) una salida a través de una tabla, la cual se también se configuró creando una conexión a una base de datos, que fue el destino de los datos extraídos desde el origen.

Aquí se cargaron los datos, para el poblamiento del almacén, en la tabla de destino se almacenaron todos los datos de las carreras correspondientes a los distintos tipos de programas, y también las distintas unidades de la Universidad.

Se tuvo que configurar el proceso de carga creando la conexión a la base de datos a un motor Sybase, agregando el Driver correspondiente de Java para éste, dándole un nombre a la conexión, agregando la URL del servidor con el nombre del base de datos de destino, incorporando el nombre de usuario del almacén y su contraseña correspondiente.

Luego se escogió la tabla de destino, para este caso la tabla de carreras, al igual que la de origen. Dentro de las opciones de configuración se puede ejecutar sentencias para borrar los campos de la tabla y crear nuevos campos con los mismos nombres que los de la tabla de origen, también especificar la tabla que contiene los identificadores (clave primaria) para cargar datos sin sobrescribir en otros datos o para evitar duplicidad de identificadores al momento de cargar los datos con los existentes.

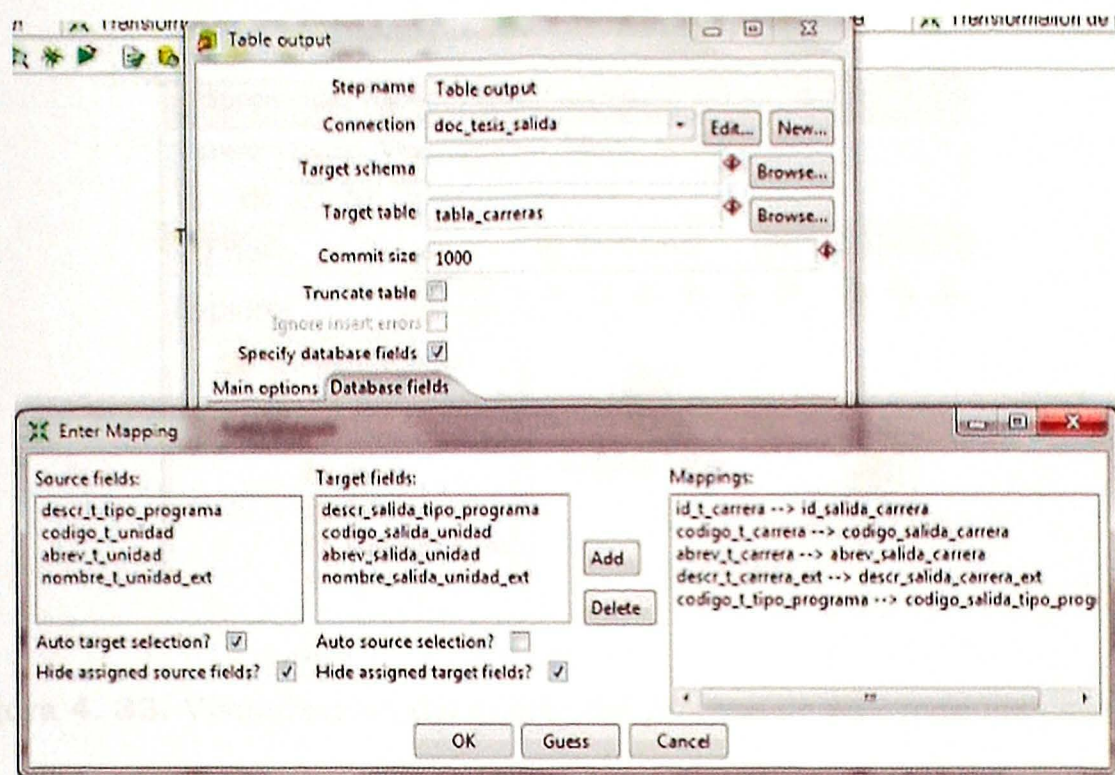


Figura 4. 32. Configuración de la tabla de destino en el proceso de ETL para carreras.

Dentro de otras opciones, se encuentra definir el tamaño de la transacción que es la cantidad de registros máximo que se van a transportar y cargar en la tabla de destino, vaciar la tabla antes de realizar la carga o ignorar errores de inserción. También se puede repartir la información en más de una tabla o particionar la información por periodos de meses o días. Luego se

realiza el mapeo de los campos de origen (Source fields) hacia los campos de destino (Target fields).

KETTLE tiene la ventaja de hacer un mapeo inteligente en donde reconoce los campos de nombres similares, tanto de origen como destino, para luego realizar el mapeo de manera casi automática.

Con la configuración exitosa para la tabla de entrada y de salida, se realizó el proceso de transporte y carga.

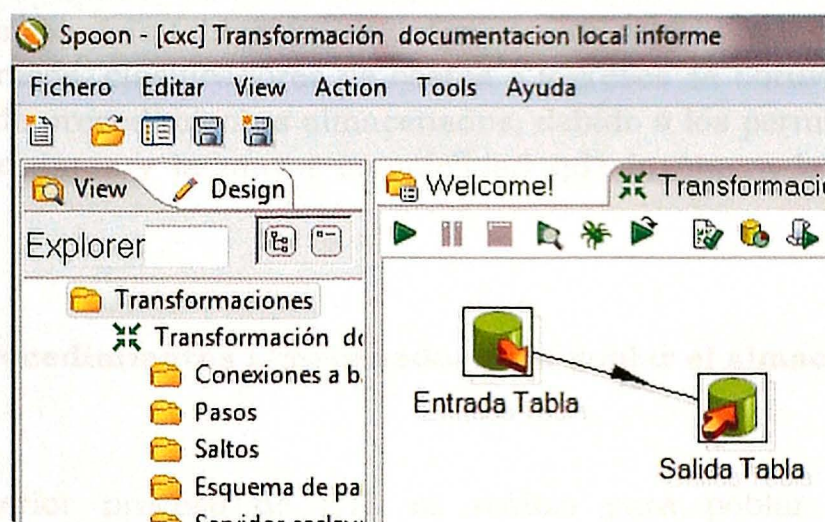


Figura 4. 33. Visualización del diseño del proceso de ETL para las carreras.

Al ejecutar el proceso de ETL se debe habilitar el modo seguro y luego pulsar el botón Play. Al finalizar el proceso, se muestran los resultados de la ejecución: Se despliega en detalle para las tablas de origen y destino los diferentes procesos de escritura, lectura, actualizaciones, la cantidad de datos de entrada y salida, y el estado de cada una, además del tiempo que se demoró en realizar el proceso de extracción, transformación y carga, como se puede apreciar en la **Figura 4. 34**. Cuando existen errores en el traspaso por alguna configuración, toda la fila correspondiente a la tabla donde ocurrió el error se pone de color rojo, permitiendo observar si se realizó la extracción o transformación o carga y observar los errores para solucionarlos.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph

#	Nombre paso	Numero Copia	Leído	Escrito	Entrada	Salida	Actualizado	Rejected	Errores	Activo	Tiempo
1	Entrada Tabla	0	0	182	182	0	0	0	0	Finalizado	0.1s
2	Salida Tabla	0	182	182	0	182	0	0	0	Finalizado	0.2s

Figura 4. 34. Resultados de la ejecución del proceso ETL para las carreras..

El proceso de ETL se realizó para las tablas de carreras, unidades, programas, tiempo. Las tablas cuyos datos son los relacionados con los ítemes de costos, ingresos, clasificadores de costos e ingresos se obtuvieron mediante la realización de procedimientos almacenados, debido a los permisos de accesos a las bases de datos y la mayor complejidad que involucra la generación de estos datos.

4.11. Procedimientos almacenados para poblar el almacén de datos.

El anterior proceso de ETL se realizó para poblar las tablas de dimensiones del almacén de datos piloto. Para la obtención de los demás datos y completar el poblamiento del almacén, se realizaron procedimientos almacenados, estos, utilizan algoritmos que se plantearon que son necesarios para la obtención de los costos por carrera y así poder completar el poblamiento del almacén.

De estos procedimientos, los datos que se obtienen son:

Los datos de las tablas de dimensiones de:

- Clasificadores de Costos.
- Agrupadores de Ítemes de Costos.
- Ítemes de Costos.

También se obtienen los datos de la tabla de hechos de los costos por carrera, en donde se obtienen los valores correspondientes a:

- Costos.
- Número de Alumnos.
- Costos por Alumno.

Para realizar todos estos procedimientos, se diseñó un ODS relacional, que es una base de datos intermedia normalizada, la cual contiene las tablas en donde se almacenaron los datos resultantes de la ejecución de los procedimientos trabajados con los algoritmos de obtención de costos.

Este ODS relacional contiene finalmente los datos de los cuales se realiza la extracción para su posterior transformación (denormalización) y carga en el almacén de datos piloto.

4.12. Modelo de Datos en su tercera versión.

El modelo de datos en su segunda versión contemplaba que en una tabla de hechos se almacenaran registros de costos e ingresos. En la tercera versión y definitiva del modelo de datos piloto, se creó una tercera tabla de hechos en la cual se guardan los datos de ingresos de manera independiente de la tabla y registros de costos, para analizarla sin mezclar datos de costos con ingresos en una vista.

Analizando los requerimientos y cómo estructurar bien los datos para el almacén de datos, se considera que hay centros de costos que representan unidades las cuales prestan servicios ya sea una carrera o a un grupo de carreras, también que prestan servicios a todas las carreras de una facultad. Y además se observa que existen centros de costos que representan unidades que son transversales para todas las carreras de la Universidad de La Frontera. El modelo de almacén de datos en su tercer diseño y definitivo cumple estos requerimientos a cabalidad.

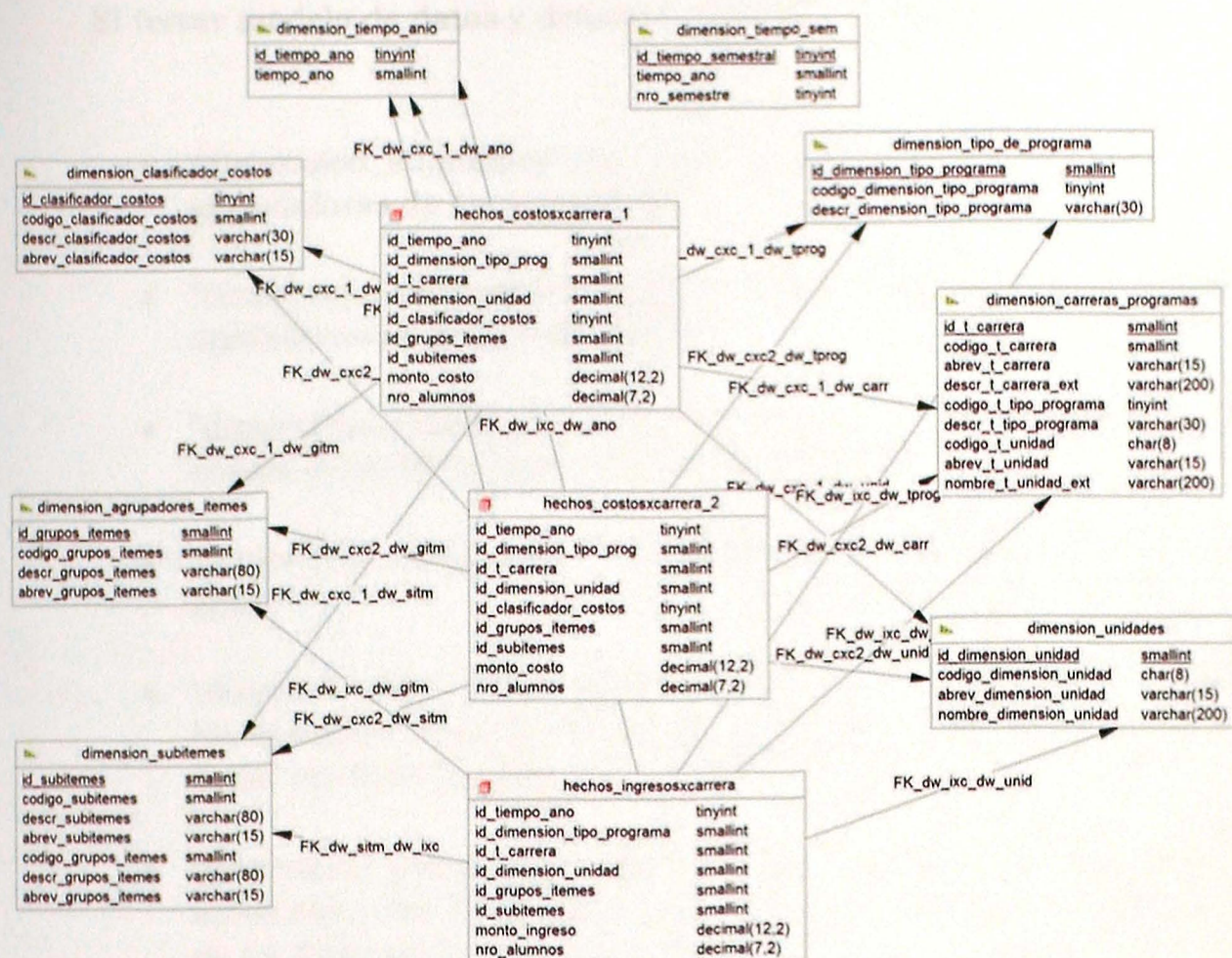


Figura 4. 35. Tercera versión del Almacén de Datos.

Las primeras dos tablas de hechos de la segunda versión del modelo mantienen su función, solamente el atributo de medida de ingreso se elimina. Almacenando en una los datos de costos considerando la inversión en bienes y la otra considera la depreciación de los mismos.

El tercer modelo de datos y definitivo contiene las siguientes tablas:

- “dimension_subitemes”: la cual almacena los subítemes, y los agrupadores de los subítemes.
- “dimension_agrupadores_itemes”: Esta tabla sólo almacena los agrupadores o grupos de ítemes.
- “dimension_clasificador_costos”: Guarda los clasificadores de costos como dimensión en el almacén de datos.
- “dimension_tiempo_ano”: Guarda los datos correspondientes a los años.
- “dimension_tipo_de_programa”: Esta tabla de dimensión contiene los distintos tipos de programa de la Universidad de la Frontera como carreras de pregrado, magíster, doctorado, etc.
- “dimensión_carreras_programas”: En esta tabla de dimensión de almacenan las Carreras y Programas que imparte la Universidad de La Frontera con sus respectivas unidades como Facultades y el tipo de programa al que pertenecen.
- “dimension_unidades”: Aquí se almacenan los datos de las distintas unidades de la Universidad, como las Facultades o Departamentos.

Las tablas de hechos del modelo de datos definitivo son:

- “hechos_costoxcarrera_1”: Esta tabla de hechos contiene las medidas: costos y número de alumnos que son los valores numéricos que se relacionan con las dimensiones del almacén de datos; existe como posible alternativa, agregar una medida de costos por alumno. Esta tabla de hechos considera la inversión como ítem de costo.
- “hechos_costoxcarrera_2”: Esta tabla, al igual que la anterior, contiene las medidas: costos y número de alumnos que son los

valores numéricos que se relacionan con las dimensiones del almacén de datos; también existe como alternativa, agregar una medida de costos por alumno. Esta tabla de hechos considera la depreciación en vez de la inversión como ítem de costo.

- “hechos_ingresosxcarrera”: Esta tabla de hechos contiene las medidas: ingresos, número de alumnos; como posible alternativa a considerar se puede agregar la medida ingresos por alumno.

4.13. Consultas Multidimensionales del nuevo almacén.

Con un modelo de datos definitivo para el almacén de datos Piloto, se crea la consulta multidimensional que abarca todas las dimensiones para las distintas medidas de la tabla de hechos de costos del almacén de datos.

Existen tres tablas de hechos, para distintas propuestas que se consideran para desarrollar a futuro. En este trabajo se realizó el trabajo de poblar, analizar y trabajar con una tabla de hechos correspondiente a los costos en donde se prorratan 100% a pregrado, considerando las inversiones. Este cubo muestra las distintas medidas para las dimensiones ya conocidas, el análisis sirve para observar y analizar los costos por carrera, la cantidad de alumnos por carrera de tal manera de conocer entonces los costos por alumno como otra medida de la tabla de hechos.

Con Mondrian Schema Workbench se creó el nuevo cubo similar al que se creó para las carreras en la primera versión del modelo de datos; este cubo considera las siguientes dimensiones: Carreras (que corresponden a los programas, las unidades a las que corresponden y las carreras), Tiempo (medido en años), los ítemes de costos e ingresos (en donde se encuentran los agrupadores de costos y los ítemes de costo) y como cuarta dimensión se agregan los clasificadores de costos.

La Tabla “hechos_costosxcarrera1” corresponde a la tabla de hechos en donde se considera el centro de costos de inversión. Los atributos de medida de esta tabla son costos y el número de alumnos, como posible medida se considera el costo por alumno.

Las demás tablas corresponden a las tablas de dimensiones que complementan la información que muestra la tabla de hechos en cada registro, de tal manera que por cada medida (costo, número de alumnos o costo por alumno) existe un año considerado, una carrera en una unidad y tipo de programa, un ítem de clasificación en distintos agrupadores de ítemes y también existe una clasificación de costos, cumpliéndose el rol de las dimensiones de facilitar la información para estudio y análisis de los costos por carrera de la Universidad.

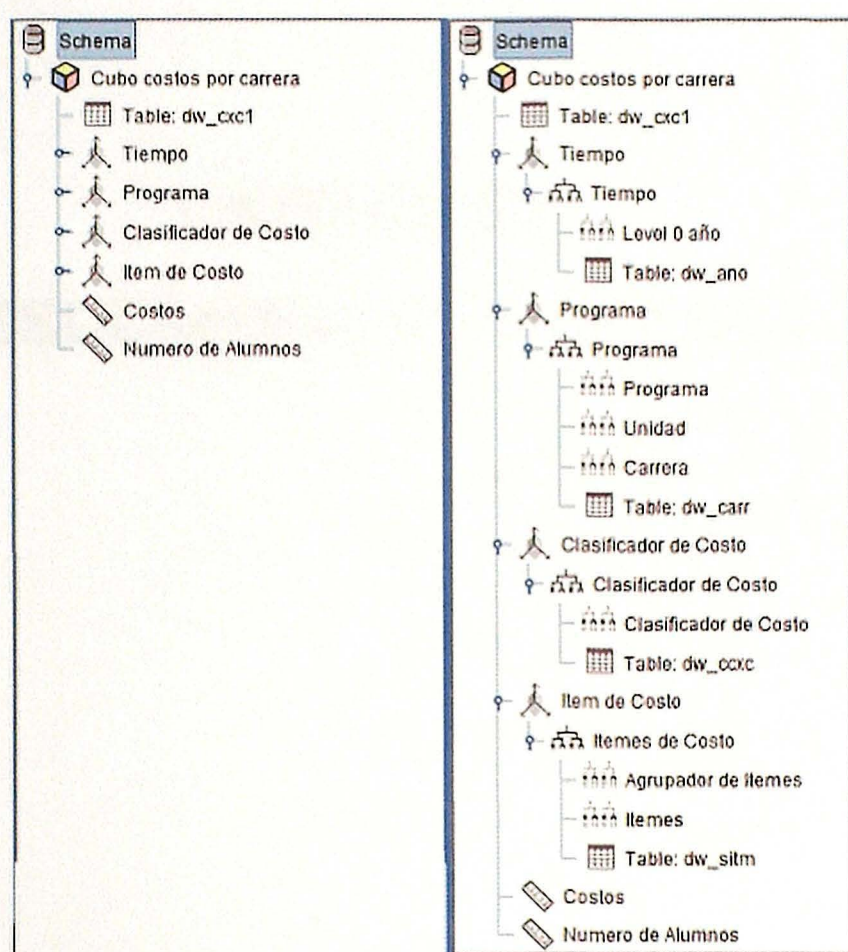


Figura 4. 36. Esquema multidimensional del almacén de datos.

CAPÍTULO V

RESULTADOS Y DISCUSIÓN

5. RESULTADOS Y DISCUSIÓN.

5.1. Visualización y análisis de Cubos en la Plataforma de Pentaho.

Con las consultas multidimensionales finalizadas para el almacén de datos piloto definitivo, se publicaron en la plataforma de Business Intelligence de Pentaho del servidor de desarrollo de la Universidad. Estas consultas o cubos son los que permiten ver todo el trabajo que implica la creación correcta del almacén de datos con su estructura en forma de estrella con tres núcleos (tablas de hechos) y un proceso de ETL que cargue los datos de manera correcta para no tener incoherencias de datos.

Los cubos se publicaron desde la herramienta Mondrian Schema Workbench con la conexión que se creó al servidor de la Universidad a la base de datos en Sybase.

Con el software Pentaho BI Server corriendo en el servidor, se accede a la plataforma a través del navegador.

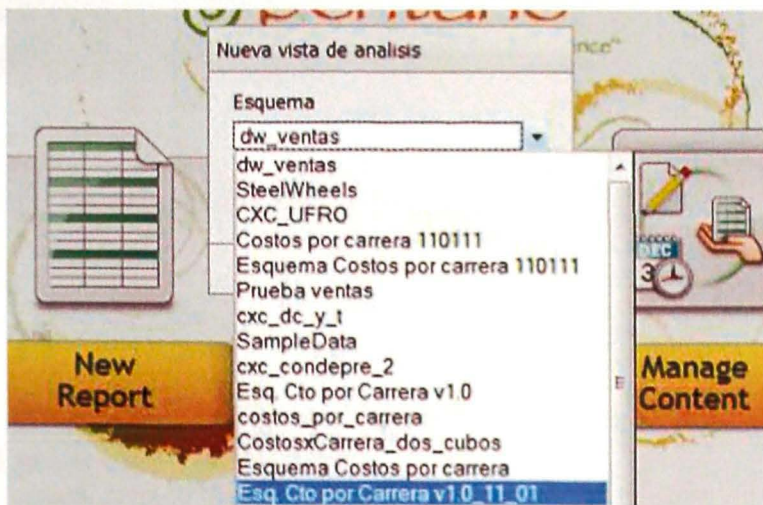


Figura 5. 1. Selección de cubos en Pentaho BI Server

clasificadores de costos y algunos ítemes de costos con los valores de los costos asociados a éstos.

OBST. Y PUERIC.	+	Clas. de Costos	+	Ítemes	407.055.676	169	2.408.613,43
ODONTOLOGIA	-	Clas. de Costos	+	Ítemes	994.588.129	410	2.425.824,69
		Adm. Facultad	-	Ítemes	93.695.231	410	228.524,94
			-	Cons. Corriente	385.749	410	940,85
				VIATICOS NACIONALES	385.749	410	940,85
			-	Hon. Docentes	17.710.254	410	43.195,74
				HONORARIOS DOCENTES	17.710.254	410	43.195,74
			+	Inversión Equip	123.046	410	300,11
			-	Rem. Académicos	14.694.320	410	35.839,8
				REMUNERACIONES ACADEMICOS	14.694.320	410	35.839,8
			+	Rem. Directivos	6.949.045	410	16.948,89
			+	Rem. No Acad.	39.863.522	410	97.228,1
		Admin. Carrera	-	Ítemes	8.662.969	410	21.129,19
			+	Inversión Equip	301.350	410	735
		Admin. Central	+	Ítemes	788.404.269	410	1.922.937,23
		Conv. Desempeño	+	Ítemes	75.284.147	410	183.619,87
		Proy. FDI	+	Ítemes	7.294.250	410	17.790,86
		Proy. Mecesus	+	Ítemes	21.247.263	410	51.822,6

Figura 5. 3. Análisis de Costos de una carrera en particular.

Por cada medida en la consulta multidimensional se debe seleccionar un agregado u operación a la medida, como sumar, contar, promedio, mínimo, máximo, etc. Para este cubo de análisis las medidas se muestran como:

Suma para los costos totales y para los costos por alumno.

Promedio para el número de alumnos por carrera de un año, lo cual tiene la desventaja de que para los niveles superiores como la raíz (primer nivel sin realizar desglose) y para los primeros niveles como unidades sin llegar aún a las carreras, se muestra un número de alumnos mucho menor que no corresponde, debido a que se muestra un promedio del número de alumnos; sin embargo, la ventaja de esto es que para los niveles de detalle, que es donde interesa analizar los costos, la medida mostrándose como promedio de número

de alumnos hace que las tres medidas sean coherentes al momento de analizar y calcular.

ODONTOLOGIA	[-] Clas. de Costos	[+] Ítemes	994.588.129	410	2.425.824,69
	Adm. Facultad	[-] Ítemes	93.695.231	410	228.524,94
		[-] Cons. Corriente	385.749	410	940,85
		VIATICOS NACIONALES	385.749	410	940,85
		[-] Hon. Docentes	17.710.254	410	43.195,74
		HONORARIOS DOCENTES	17.710.254	410	43.195,74
		[+] Inversión Equip	123.046	410	300,11
		[-] Rem. Académicos	14.694.320	410	35.839,8
		REMUNERACIONES ACADEMICOS	14.694.320 /	410 →	35.839,8
		[+] Rem. Directivos	6.949.045	410	16.948,89
		[+] Rem. No Acad.	39.863.522	410	97.228,1

Figura 5. 4. Análisis de Costos por alumno.

En la Figura 5. 4. Se aprecia que asignando el agregador promedio a la medida del número de alumnos se realiza un correcto análisis de los costos por alumno, priorizando el detalle por carrera a que el total de los cálculos de los niveles superiores, donde el número de alumnos no es el correcto debido a que se trata de un promedio.

Con esta consulta multidimensional queda finalizado el trabajo de la implementación del almacén de datos. Su implementación se realizó de manera exitosa, dejando como desarrollo posible a futuro, el prorratio, ETL, poblamiento de las demás tablas de hechos y consultas multidimensionales variadas para todas las tablas de hechos.

5.2. Resultados con respecto a los objetivos propuestos.

Los objetivos planteados en este trabajo se realizaron con éxito, tanto el objetivo general como los objetivos específicos, ya que se implementó el almacén de datos piloto en un servidor de prueba (objetivo general) a través del diseño estructural del almacén de datos quedando finalizado con la tercera versión (Figura 4. 35.). Se crearon consultas OLAP (multidimensionales) para el

análisis con Pentaho (utilizando casi todas sus herramientas), se buscó información de otros sistemas de costos por carreras sin encontrarse fuentes suficientes como para tener referencias al momento de realizar el almacén de datos, y se analizó en detalle los estudios de sistemas de costos por carrera planteados en años anteriores en la Universidad de La Frontera.

Considerando estos puntos, la implementación del almacén de datos con sus consultas para análisis fueron exitosos.

Con la etapa de desarrollo del piloto concluida, comienza la etapa de pruebas del sistema de análisis de costos por carrera en certificación. Continuación a esto, sigue la etapa de validación de usuarios del sistema, lo cual da paso a la etapa de producción y socialización del sistema.

CAPÍTULO VI

CONCLUSIONES

6. CONCLUSIONES.

La información se considera como un activo de vital importancia en organizaciones, entidades, instituciones, empresas, etc. Debido a que con la correcta obtención de los datos, las fases el procesamiento, análisis y presentación pueden generar perspectivas amplias de lo que ocurre para tomar mejores decisiones en comparación a un análisis más deficiente, lo cual implica que la información que una entidad posee debe ser correcta, oportuna y precisa.

Un punto importante a analizar al momento de crear un almacén de datos es investigar, conocer, aprender el contexto y dominio sobre el cual se va a tratar el almacén de datos, los datos a almacenar y manipular son de distintos contextos dentro de instituciones, organizaciones, empresas, etc.

El diseño del modelo de datos debe tener una estructura que abarque todos los datos necesarios para poder crear consultas multidimensionales que entreguen la información requerida por el usuario; además, en las consultas deben procesarse grandes cantidades de datos de manera rápida, por lo que las consultas OLAP son las indicadas para esto puesto que permiten hacer análisis de grandes cantidades de datos de manera eficiente.

Para diseñar el modelo de un almacén de datos no se debe realizar el proceso de normalización, la duplicidad de datos no se considera como una metodología errónea para diseñar modelo de datos, al contrario, al existir duplicidad de datos en las tablas de dimensiones las consultas OLAP se efectúan más rápido obteniendo el acceso a los datos de forma más expedita.

Las distintas herramientas libres de Pentaho permiten apoyar en cada uno de los diversos procesos necesarios para llevar a cabo la implementación de un Data Warehouse o Data Mart. Para ETL, migraciones, consultas multidimensionales, reportes, datamining, y ver y analizar cubos, Pentaho OpenSource es una herramienta factible de utilizar con éxito.

Los algoritmos que se plantean para obtener datos de fuentes de origen son fundamentales para el poblamiento de los almacenes de datos, para esto, al momento de realizarlos, se deben realizar en función de obtener resultados esperados y correctos, ya que si se realizan mal, se pueden obtener datos erróneos, lo cual conlleva a que la información procesada y analizada por un sistema de inteligencia de negocio no sea correcta y se tomen malas decisiones. Por lo tanto, los algoritmos para obtención de datos, si se van a utilizar para poblar una base de datos ODS o para poblar un almacén de datos deben ser correctos, de tal manera que los resultados obtenidos al ser procesados y analizados por un sistema de Business Intelligence ayuden y apoyen a la toma de decisiones de manera correcta.

El proceso de ETL es un procedimiento complejo que debe realizarse teniendo en cuenta diversos factores que pueden traer errores. Debido a esto, se deben conocer los datos y modelos de bases de datos de origen en los cuales se realizan las extracciones. Al realizarse distintos tipos de operaciones y transformaciones hay que conocer los datos tanto a un nivel conceptual (lo que significan, lo que contienen) como de formato (tipo de dato, etc.), esto implica que para conocer los datos y sus fuentes, hay que tener acceso a bases de datos funcionales y externas, lo cual implica tener permisos de acceso y que no cualquier persona puede realizar procesos ETL. En el caso de la Universidad de La Frontera, se trabajó en conjunto con personas que tuvieran acceso a las bases de datos y que conocieran su estructura y contenido, para realizar el ETL. Entonces para realizar ETL, tenemos que tener en cuenta los privilegios del usuario, tiene que haber confidencialidad, confianza, integridad y seguridad al momento de manipular los datos para la extracción, transformación y carga de éstos.

Utilizar la tecnología Business Intelligence para realizar este trabajo es una forma necesaria de dar solución al tema de realizar un sistema de análisis de costos por carrera, debido a que gracias a esta tecnología se puede obtener información estratégica para los altos ejecutivos o cargos de gran responsabilidad e importancia al interior de la Universidad y así tomar decisiones de gran relevancia gracias al apoyo que ofrece la disponibilidad de la información procesada por la tecnología de Business Intelligence.

Con Business Intelligence, las compañías, empresas, instituciones u organizaciones pueden analizar tendencias que representan oportunidades nuevas e importantes y así anticipar problemas potenciales y hacer ajustes antes de que se conviertan en un problema.

BIBLIOGRAFÍA.

BIBLIOGRAFÍA.**Bibliografía consultada:**

- Aravena, H. (2002). Trabajo de título: "Sistemas de Costos para las Carreras de: Ingeniería Civil Industrial M. Agroindustria, Ingeniería en Alimentos e Ingeniería Ambiental". Universidad de La Frontera, Temuco, Chile.
- Connolly T. M.; Begg C. E. (2005). "Sistemas de Bases de Datos, un enfoque práctico para diseño, implementación y gestión", 4ª Edición, Pearson Educación S.A. España.
- Kehr, R., (2010). Trabajo de título: "Implementación del Módulo Business Intelligence para SAP (R3)". Universidad de La Frontera, Temuco, Chile
- Ramakrishnan, R.; Gehrke J. (2007). "Sistemas de Gestión de Bases de Datos", 3ª Edición, Mc Graw-Hill, España.
- Roldán, M. C. (2010). "Pentaho 3.2 Data Integration", Editorial Packt Publishing Ltd., Reino Unido.
- Silberschatz, A.; Korth, H.; Sudarshan S. (2006). "Fundamentos de bases de datos", 5ª Edición, Mc Graw-Hill, España.
- UFRO (1997). Universidad de La Frontera, 1997, Trabajo de Estudio: "Sistema Costos por Carrera". Chile.
- UFRO (2008). Universidad de La Frontera, 2008, Diseño/Desarrollo de un modelo para el estudio de costos por carrera Universidad de La Frontera: "Informe Estudio Costos por Carrera". Chile.
- UFRO (2010). Universidad de La Frontera, 2010, Diseño/Desarrollo de un modelo para el estudio de costos por carrera Universidad de La Frontera: "Entrega Diseño/Manual de procedimiento/plan comunicacional". Chile.

Referencias web consultada:

- Calderón, M. (N/D). "Pentaho Open BI", Recuperado el 4 de Octubre del 2010, desde <http://www.cognus.cl/content/view/271452/Pentaho-Open-BI.html>
- Medina, J. (2005). "Business Intelligence: conceptos y actualidad", Recuperado el 20 de Septiembre del 2010, desde <http://www.gestiopolis.com/recursos5/docs/ger/buconce.htm>
- Salinas, A. (2008). "Introducción a Pentaho", Recuperado el 5 de Octubre del 2010, desde <http://www.gravitar.biz/index.php/bi/introduccion-pentaho-parte-1/>
- Sinnexus (N/D). Empresa Sinnexus, "Business Intelligence", Recuperado el 23 de Septiembre del 2010, desde http://www.sinnexus.com/business_intelligence/index.aspx

ANEXOS

GLOSARIO DE TÉRMINOS

• **Alcaldía de Bogotá:** es el municipio de Bogotá, D.C., el cual es el centro administrativo y político de la ciudad.

• **Bombas de agua:** son dispositivos mecánicos que convierten energía en movimiento para mover el agua desde una fuente subterránea o superficial hacia un punto de uso.

• **Catón Street:** es una zona comercial y residencial ubicada en el centro de Bogotá, D.C.

• **Data River:** es un río que fluye a través de Bogotá, D.C., y es uno de los principales cursos de agua de la ciudad.

• **Desarrollo urbano:** es el proceso de transformación física y social de un territorio, que incluye la construcción de infraestructura, vivienda y servicios.

• **Medio ambiente:** es el conjunto de factores físicos, químicos y biológicos que rodean a los seres vivos y que influyen en su desarrollo.

• **ODS: Objetivos de Desarrollo Sostenible:** son 17 objetivos globales que buscan mejorar la calidad de vida y el bienestar de la humanidad.

• **OLAP: Organización Local de Asesoría y Promoción:** es una organización que trabaja en la promoción y desarrollo de las comunidades locales.

ANEXO A.

GLOSARIO DE TÉRMINOS

- **Almacén de Datos:** base de datos que almacena grandes cantidades de datos integrados para realizar análisis para ayuda a la gestión y a la toma de decisiones.
- **Business Intelligence:** (Inteligencia de Negocio) Son la habilidad y tecnologías para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.
- **Cubos OLAP:** Vista de la información almacenada en una base de datos, con una estructura donde se reconocen las dimensiones y hechos. Se conocen también como consultas multidimensionales.
- **Data Mart:** Es un almacén de datos departamental, esto quiere decir que concentra datos específicamente de un área de la empresa.
- **Dimensiones:** Características de cada proceso de negocio, es el área temática de una empresa, es la información adicional que complementa los datos de las tablas de hechos.
- **Hechos:** Representa una métrica para medir el estado de desarrollo de una organización, es la información numérica procesada de un almacén de datos, por ejemplo: el número de ventas.
- **ODS:** Base de datos relacional normalizada que sirve de paso intermedio entre los datos de las fuentes de origen y el almacén de datos.
- **OLAP:** procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia empresarial (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

- **Pentaho:** Conjunto de herramientas y tecnologías para trabajar con Business Intelligence; existe la versión Enterprise y la versión Community,
- **Prorrateo:** Es la base para la asignación de un importe de manera proporcional; es la repartición o división de forma proporcional.

ANEXO B.

INSTALACIÓN, CONFIGURACIÓN DE PENTAHO BI SERVER

ANEXO B.

En este anexo se explica cómo instalar configurar y levantar la plataforma Pentaho BI Server en Linux (Distribución Ubuntu u OpenSUSE).

- Requisitos:

Como esta guía se instala en las distribuciones señaladas anteriormente, se tiene que tener Ubuntu o sino OpenSUSE. Para Windows el procedimiento es similar con la diferencia de los archivos de comandos a ejecutar (en Linux tienen extensión .sh y en Windows tienen extensión .bat)

Tener Instalado JAVA (jdk).

- Instalación:

(i) Descargar Pentaho BI Server Community Edition (recomendado alguna versión estable) desde:

<http://sourceforge.net/projects/pentaho/files> y descomprimir el archivo en algún directorio, por ejemplo:

`/opt/pentaho`

Al descomprimir el servidor descargado contiene los siguientes directorios:

administrator-console: Es el servicio administrativo para gestionar los usuarios a la plataforma de Pentaho BI Server y para crear las conexiones a bases de datos en donde van a interactuar las consultas multidimensionales publicadas en la plataforma de Pentaho, se conoce como PAC (Pentaho Administration Console).

El otro directorio es **biserver-ce:** Este directorio contiene los archivos de la plataforma Pentaho BI Server.

(ii) Levantar la plataforma Pentaho BI Server:

Para levantar e iniciar la plataforma, se tiene que ejecutar un script en el directorio donde se descomprimió el archivo descargado.

```
/biserver-ce/start-pentaho.sh
```

Para iniciar la consola de administración de Pentaho se tiene que ejecutar el archivo ubicado en /administration-console/start-pac.sh

Para acceder a la plataforma BI, acceder desde el navegador a:
`http://nombre_del_servidor:8080/pentaho.`

Trabajando de forma local sería: `http://localhost:8080/pentaho`

Para acceder a la consola PAC, acceder desde el navegador a:
`http://nombre del servidor:8099.`

Trabajando de forma local sería: `http://localhost:8099` y acceder con los datos por defecto "**user:** *admin*" y "**password:** *password*".

- Configuración:

Pentaho en sus herramientas que se conectan a bases de datos, utiliza conectores de JAVA para comunicarse con las bases de datos. Por lo tanto debe contar con su Driver (.jar) específico para los distintos motores de datos con los que va a trabajar:

Pentaho BI Server: Necesita el driver en la ruta

```
/biserver-ce/tomcat/common/lib
```

Para la consola de administración de Pentaho se requiere el driver en:

```
/administration-consola/jdbc
```


Cambiar la clave de publicación en el servidor Pentaho:

Para publicar consultas multidimensionales y reportes es necesario una clave de publicación que es solicitada por las herramientas de diseño de Pentaho como por ejemplo: Mondrian Schema-Workbench.

El archivo para modificar la clave es `publisher-config.xml` y se encuentra en el directorio: `biserver-ce/pentaho-solutions/system/` Aquí la clave se cambia por la que se desea, en este ejemplo es *"password"*:

```
<?xml versión="1.0"?>
<publisher-config>
<publisher-password>password</publisher-password>
</publisher-config>
```


ANEXO C.

CREACIÓN Y PUBLICACIÓN DE UN ESQUEMA MULTIDIMENSIONAL

ANEXO C.

Este anexo explica los pasos para crear un esquema multidimensional y publicarlo con Mondrian Schema-Workbench, para poder analizar desde la plataforma de Pentaho todos los cubos OLAP (consultas multidimensionales) creadas en este esquema.

- Instalación:

Descargar la herramienta Mondrian Schema-workbench de:

<http://sourceforge.net/projects/mondrian/files/schema%20workbench>

Se recomienda descargar una versión estable, para este caso se utiliza la versión 3.2.1-stable. Después, descomprimir la herramienta en cualquier directorio, por ejemplo:

`/opt/pentaho`

Copiar el driver (.jar) del motor de base de datos con el que se quiere trabajar para conectar la herramienta a una base de datos y pegarlo en el directorio: `/schema-workbench/drivers`.

Para ejecutar la herramienta es necesario tener instalado Java (jdk), luego se tiene que ejecutar el script: `/schema-workbench/workbench.sh`.

- Conexión con la base de datos.

Al abrir la herramienta, ir a la pestaña Tools -> Connections.

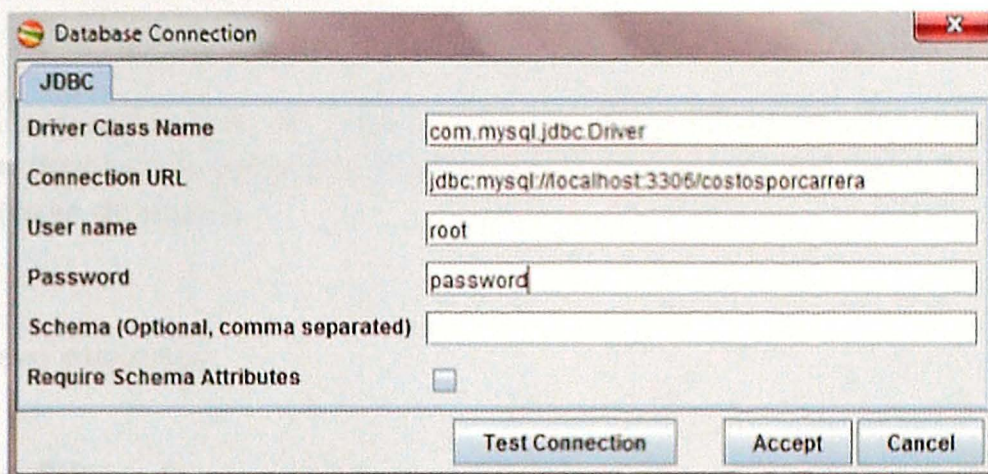


Figura C. 1. Parámetros de conexión a la base de datos en Mondrian Schema-Workbench.

Como se observa en la Figura C.1. Los Parámetros a ingresar son:

- Driver Class Name (nombre de la clase del Driver).
- Connection URL (URL para la conexión a la base) en donde al final de la ruta se agrega /nombre de la base de datos. Para este ejemplo la base se llama "costosporcarrera".
- User name (nombre de usuario) para este caso es root.
- Password (contraseña) es la contraseña del usuario para ingresar tener acceso a la base de datos en este ejemplo la contraseña es "password".
- Schema (opcional).

Para saber si la conexión ha sido exitosa, ir a "Test Connecion" y la herramienta debe arrojar el mensaje "Database Connection Successful". Luego ir al botón Accept.

- Creación del esquema:

El esquema que se crea es un archivo XML con las propiedades y elementos necesarios para que el sistema OLAP pueda trabajar con el almacén de datos (base de datos).

Nuevo esquema:

Ir a **File -> New -> Schema**

Primero definir el atributo **name** en el nuevo esquema. Luego clic derecho sobre el schema y elegir "**add Cube**". Definir el atributo **name** del cubo.

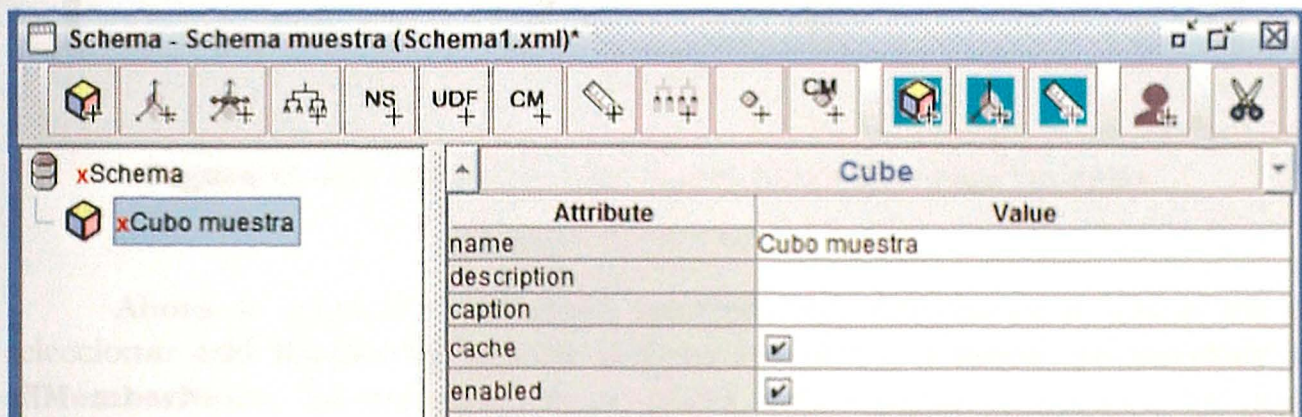


Figura C. 2. Creación de un nuevo esquema y agregación de un cubo.

Ahora a este nuevo cubo se le tiene que asociar una tabla de hechos: clic derecho en el cubo y elegir **add table**, aquí, en **name** seleccionar la tabla de hechos del almacén de datos.

Con el cubo creado y su tabla de hechos agregada, hay que crear las dimensiones, la cantidad de dimensiones dependen de la cantidad de tablas de dimensiones de la base de datos.

En el cubo, clic derecho y seleccionar **add Dimension**, aquí definir los atributos **name** y **foreignKey**; el primero lleva el nombre que se le quiere dar a la dimensión y en el segundo se debe elegir la clave foránea correspondiente de la dimensión, en la figura C.3. se muestra que la clave foránea que se ingresa es **“id_carrera”** que hace referencia a la tabla de dimensión **“carrera”**.

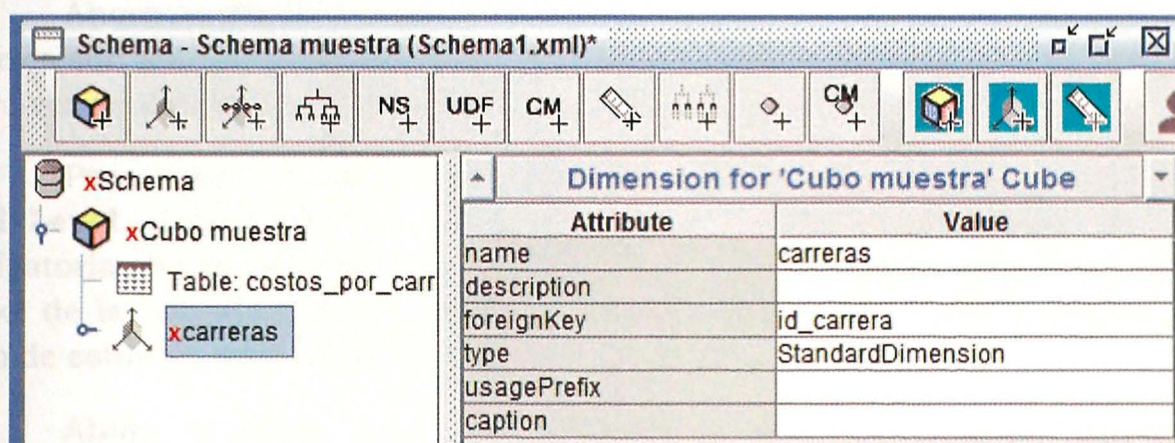


Figura C. 3. Creación de una nueva dimensión para un cubo.

Ahora se crean las jerarquías, para esto clic derecho en la dimensión y seleccionar **add Hierarchy**, aquí se definen los atributos **name**, **primaryKey** y **allMemberName**. De inmediato no se podrá definir algunos atributos, así que ahora en la jerarquía creada clic derecho y elegir **add Table** y seleccionar la tabla que corresponde, para este ejemplo es carreras. Ahora definir los atributos de la jerarquía.

xSchema

- xCubo muestra
 - Table: costos_por_carrera
 - xcarreas
 - xjearquia carreras
 - Table: carrera

Hierarchy for 'carreras' Dimension	
Attribute	Value
name	jearquia carreras
description	
hasAll	☒
allMemberName	Carreras
allMemberCaption	
allLevelName	
defaultMember	
memberReaderClass	
primaryKeyTable	
primaryKey	id carrera
caption	

Figura C. 4. Edición de atributos de una jerarquía.

Ahora se deben agregar los distintos niveles de por cada jerarquía, estos niveles son los que después en la vista del cubo multidimensional se desglosan para ver en detalle los diferentes registros del almacén de datos.

Para crear un nuevo nivel, clic derecho en la jerarquía y seleccionar **add Level**, aquí se deben definir los atributos **name**, y **column** de forma obligatoria. Normalmente se crea más de un nivel debido a que se muestra el árbol de la jerarquía, por ejemplo: Programa->Facultad->Carrera. Siendo cada uno de estos un nivel distinto de una jerarquía.

Ahora, se deben agregar las medidas al cubo, para esto clic derecho en el cubo y elegir **add Measure**. Aquí se definen los atributos **name**, **aggregator** y **column**, en donde **aggregator** debe ser la característica que opere sobre la medida como suma, promedio, distinto, contar, etc. Y la columna debe ser el campo de la tabla de hechos correspondiente a la medida que se quiere seleccionar, para este caso "costos".

Con esto el cubo OLAP o consulta multidimensional de ejemplo queda finalizada para ser publicada en la plataforma de BI de Pentaho y poder visualizarla.

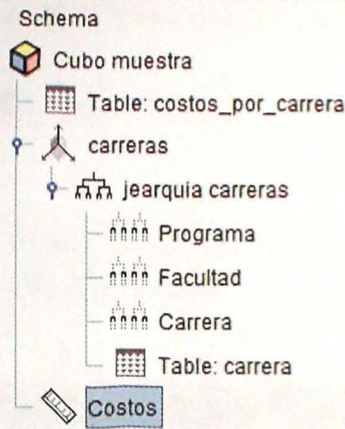


Figura C. 5. Esquema multidimensional finalizado con sus dimensiones y medidas.

- Publicar el esquema:

Ahora ir a **File-> Publish.**

Aquí, se deben completar los parámetros de la conexión para publicar el esquema en la plataforma.

Repository Login

Server

URL:
http://localhost:8080/pentaho/

Publish Password:
....

Pentaho Credentials

User:
carlos

Password:
...

☒ Remember these Settings

OK Cancel

Figura C. 6. Datos de conexión para publicar esquema.

- **URL:** corresponde al servidor donde está instalado Pentaho BI Server.
- **Publish Password:** Es la clave de publicación de esquemas del servidor.
- **User y Password:** Son los datos de un usuario de Pentaho BI Server con privilegios de administrador.

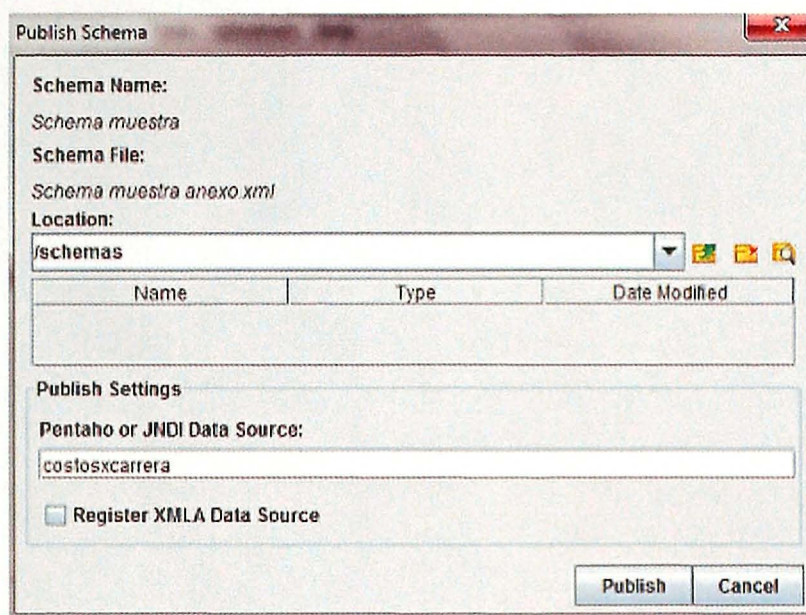


Figura C. 7. Publicación de Esquema.

Por último seleccionar un directorio donde es almacena el esquema, y definir el campo: **Pentaho or JNDI Data Source**, el cual corresponde al nombre del conexión a la base de datos creada en la consola de administración de Pentaho, finalmente seleccionar el botón Publish y el esquema con el cubo multidimensional queda publicado.