DATAWAREHOUSING Y SU APLICACIÓN COMO APOYO A LA TOMA DE DECISIONES EN LA ESPOL

Juan Carlos Bustamante[1], José Francisco Rodríguez[2], Fabricio Echeverría[3]

[1] Ingeniero en Computación, 2005, FIEC-ESPOL; email: jbustama@espol.edu.ec
[2] Ingeniero en Computación, 2005, FIEC-ESPOL; email: jrodrig@espol.edu.ec
[3] Director de Tesis, Ingeniero en Computación, Escuela Superior Politécnica del Litoral, 1998, Postgrado Ecuador, Escuela Superior Politécnica del Litoral, 2005. Profesor FIEC-ESPOL desde 1998, email: pechever@espol.edu.ec

RESUMEN

El presente artículo presenta información sobre el diseño e implementación de un Datawarehouse y su aplicación como apoyo a la toma de decisiones en la ESPOL. El sistema incluye componentes de ubicación de fuentes de datos, extracción, limpieza, montado, sumarización y consultas de cubos de información. Los diversos módulos están diseñados para poder ser ejecutados de manera independiente; los componentes de administración y control como una aplicación bajo Windows cliente-servidor y el componente de reportes y consultas por intermedio de un navegador Web. Vale mencionar que esta implementación corresponde a una plataforma de información que está en capacidad de ser accesada a través de herramientas de terceros o propietarias.

Se presenta una metodología de implementación del datawarehouse que abarca aspectos tales como identificación de etapas, determinación de roles y responsabilidades y tiempos para culminación de tareas. Contiene las características del diseño e implementación del datawarehouse identificando aspectos tales como definición de la arquitectura, necesidades de información, análisis del área objetivo, estudio de las fuentes de datos, diseño de las transformaciones, base de datos física, acceso de usuarios finales y la selección de las herramientas para el desarrollo del proyecto.

Palabras claves: *Datawarehouse, tecnología, inteligencia comercial, OLAP, soporte de decisión.*

SUMMARY

The project to be exposed, designs and it implements a Datawarehouse and its application like support to the taking of decisions in the ESPOL. The system includes components of location of sources of data, extraction, cleaning, mounted, summarization and consultations of cubes of information. The diverse modules are designed to be able to be executed in an independent way; the administration components and control like an application under Windows client-server and the component of reports and consultations through a navigator web. It is worth to mention that this implementation corresponds to a platform of information that is in capacity of being accessed through tools of third or owners.

A methodology of implementation of the Datawarehouse is presented that embraces such aspects as identification of stages, determination of lists and responsibilities and times for culmination of tasks. It contains the characteristics of the design and implementation of the Datawarehouse identifying such aspects as definition of the architecture, necessities of information, analysis of the area objective, study of the sources of data, design of the transformations, physical database, access of end users and the selection of the tools for the development of the project.

I. INTRODUCCION

Datawarehousing es el centro de la arquitectura para los sistemas de información desde la década de los '90. Soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico, informático sobre una amplia perspectiva de tiempo.

Un Datawarehouse o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales [1]. Se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales o transaccionales. La data extraída es transformada para eliminar inconsistencias y resumir si es necesario y luego, cargadas en el datawarehouse. El proceso de transformar, crear el detalle de tiempo variante, resumir y combinar los extractos de datos, ayudan a crear el ambiente para el acceso a la información. Este nuevo enfoque ayuda a las personas individuales (analistas del negocio, usuarios, investigadores), a efectuar su toma de decisiones con más responsabilidad y con fundamento histórico de respaldo.

Se puede caracterizar un datawarehouse haciendo un contraste de cómo los datos de un negocio almacenados en un datawarehouse difieren de los datos operacionales usados por las aplicaciones de producción [2].

Tabla 1: Comparativo entre tecnologías de Almacenamiento

Base de Datos Operacional	Datawarehouse
Datos Operacionales	Datos del negocio para Información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

El objetivo de este proyecto es proveer a la ESPOL innovación de la Tecnología de Información dentro de un ambiente de Datawarehousing que le permita hacer un uso más óptimo de la información, como un ingrediente clave para un proceso de toma de decisiones más efectivo, con soporte al procesamiento analítico e informático sobre una amplia perspectiva de tiempo provisto por una sólida plataforma a partir de datos históricos.

A continuación se resumen los beneficios que un Datawarehouse puede aportar:

- Poner tanta información de interés general y comercial como sea posible en manos de tantos usuarios diferentes como sea posible.
- Mejorar el tiempo de espera que se consumen en la generación de los reportes habituales.
- Proporcionar una herramienta que permita tomar las decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilitar la aplicación de técnicas estadísticas de análisis y modelamiento para encontrar relaciones en los datos del almacén; obteniendo un valor agregado para el negocio de dicha información.
- Proporcionar la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplificar dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.

II. CONTENIDO

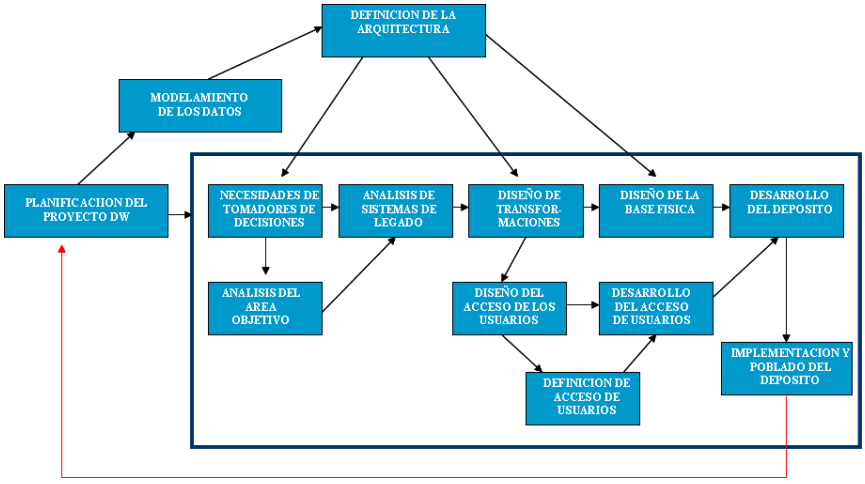
Alcance

El alcance de un datawarehouse puede ser tan amplio como toda la información estratégica de la empresa desde su inicio, es por esto que el proyecto contempla la cobertura de una de las áreas de mayor impacto en la Institución como lo es la Académica. Dentro de este ámbito se crearán modelos de datos iniciales (cubos de información) que intenten resolver las necesidades de información más críticas y consultadas actualmente.

Metodología de Desarrollo

La figura 1 refleja la metodología genérica [3], que es la seleccionada para la implementación del Datawarehouse en la ESPOL. Las etapas presentadas son iterativas; cada etapa provee más detalle así como puede ser repetitiva. Cada cuadro representa una etapa significante que debería suceder durante un proyecto de tecnología de Datawarehouse.

Figura 1: Metodología Genérica para el Desarrollo de un Datawarehouse



En esencia, la planificación y desarrollo de un proyecto de Datawarehouse requiere lo siguiente [4]:

- Los proyectos de Datawarehouse deben ser iterativos.
- Las necesidades del negocio deben ser constantemente reflejadas.
- Herramientas específicas deben habilitar al equipo de desarrollo en la identificación del progreso y enfoque de cada etapa mientras se añade al “gran conjunto”.

Implementación de los modelos del negocio

“Un modelo de negocio es un sistema que define cómo una empresa debe planificar, construir y emplear sus recursos para ofrecer a sus clientes un mejor beneficio o valor agregado superior” [7].

En la implementación del modelo de negocio de la ESPOL se abarcan los elementos desarrollados en la tabla 2 a continuación mostrada.

Tabla 2: Modelo de Negocio del Datawarehouse

COMPONENTE	CARACTERISTICAS DEL MODELO
Valor del Cliente	El desarrollo del proyecto permitirá dar a los usuarios un trato diferenciado por cuanto contarán con apoyo en la toma de decisiones.
Alcance	Los usuarios del datawarehouse (Directores, Analistas, Investigadores) serán los beneficiarios del valor agregado que se obtendrá.
Costos	Dado que la implementación ha sido realizada in-house, con plataforma establecida y herramientas de libre uso, el costo del proyecto es mínimo.
Beneficios	Vanguardia tecnológica. Soporte a la toma de decisiones. Calidad de información. Satisfacción de los usuarios del sistema
Actividades relacionadas	Plataforma para herramientas de inteligencia de negocios. Soporte a sistemas de información ejecutivos. Bases para nuevos proyectos o tópicos de investigación como Datamining
Implementación	Desarrollo por parte del equipo de TI de ESPOL. Empleo de componente de hardware, software y comunicaciones propios.
Capacidades	Soporte al flujo transaccional del negocio. Posibilidad de crecimiento iterativo en cobertura. Facilidad de distribución
Vigencia	Independencia de herramientas comerciales brinda escalabilidad y flexibilidad de crecimiento. Conocimiento propio de la implementación garantiza seguridad. Ampliación de cobertura se traduce en usabilidad.

Implementación de las interfaces

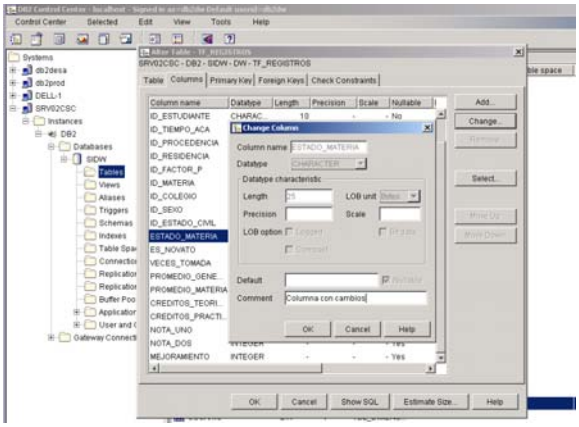
La Interfaz de Usuario, en adelante IU, de un programa es un conjunto de elementos hardware y software de una computadora que presentan información al usuario y le permiten interactuar con la información y con el computadora [5]. En la implementación de las interfaces del Datawarehouse de la ESPOL vale diferenciarlas de acuerdo a la naturaleza de las mismas, en función del componente al que pertenecen, esto es, se dará una muestra de las interfaces de administración, procesamiento, control, verificación, seguridades y consulta del proyecto [8].

a) Interfaz de Administración

La administración comprende tareas de manejo y control de la base de datos a través del DBMS y de las operaciones propias del datawarehouse con el EDW-Security.

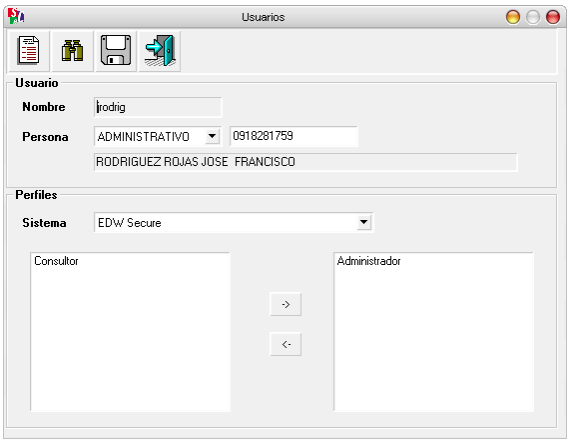
La herramienta que administra la base de datos física (SIDW) es el Control Center de DB2. Permite realizar operaciones como crear bases de datos, columnas, disparadores, procedimientos almacenados y gestionar parámetros de la configuración del motor. Una vista de esta herramienta es mostrada en la siguiente figura 2 que representa un cambio en una de las columnas.

Figura 2: Herramienta de administración de la base de datos del EDW



La herramienta EDW-Security permite administrar usuarios, perfiles, acceso a reportes o consultas tratadas como operaciones, generación masiva de reportes con el procesamiento respectivo de extracción, sumarizado y presentación en formato XML para su consulta a través del EDW-Consult [8]. La siguiente vista ilustra la interfaz de administración (creación, modificación y eliminación de usuarios del EDW. A este componente se denomina EDW-Security.

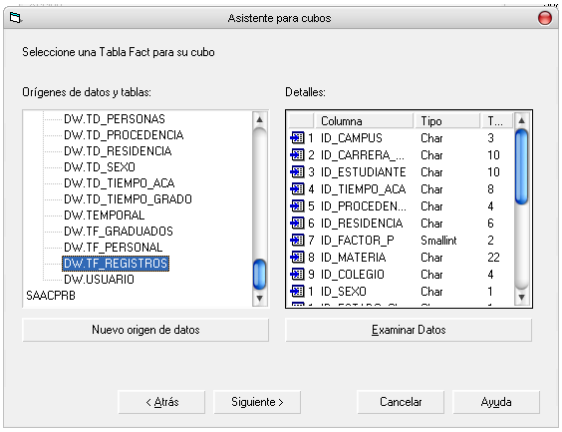
Figura 3: Administración de Usuarios a través del EDW-Security



La gestión de accesos a los diferentes niveles de información se realiza a través del componente antes descrito, pero en la modalidad de Sistemas, Perfiles y Operaciones.

El ingreso, modificación y eliminación de elementos del datawarehouse como tablas de hechos y dimensiones son realizados por medio del componente EDW-OLAP, controlador del cubo de datos, a través de la siguiente pantalla:

Figura 4: Vista del componente EDW-OLAP



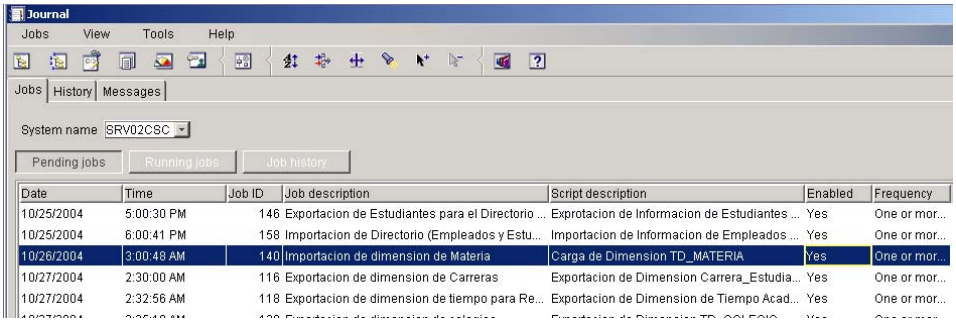
Esta herramienta, el EDW-OLAP, permite conectarse con cualquier repositorio de datos a través de una sentencia UDL, otorgándole a la aplicación soporte multiplataforma.

b) Interfaz de procesamiento y programación

El procesamiento y ejecución de las tareas está a cargo del Journal Center de DB2, una herramienta que administra los trabajos y permite monitorear su planificación, ejecución e historia de manera natural.

En la figura 5 muestra el conjunto de trabajos programados para su ejecución con detalles tales como fecha y hora de ejecución, descripción, estado y demás atributos.

Figura 5: Programador de tareas de limpieza, exportación y cargado de datos para el EDW



Date	Time	Job ID	Job description	Script description	Enabled	Frequency
10/25/2004	5:00:30 PM	146	Exportacion de Estudiantes para el Directorio ...	Exprotacion de Informacion de Estudiantes ...	Yes	One or mor...
10/25/2004	6:00:41 PM	158	Importacion de Directorio (Empleados y Estu...	Importacion de Informacion de Empleados ...	Yes	One or mor...
10/26/2004	3:00:48 AM	140	Importacion de dimension de Materia	Carga de Dimension TD_MATERIA	Yes	One or mor...
10/27/2004	2:30:00 AM	116	Exportacion de dimension de Carreras	Exportacion de Dimension Carrera_Estudia...	Yes	One or mor...
10/27/2004	2:32:56 AM	118	Exportacion de dimension de tiempo para Re...	Exportacion de Dimension de Tiempo Acad...	Yes	One or mor...

La programación de ejecución de procesos de importación y exportación de datos se realiza a través de un scheduler. En esta herramienta se determina la frecuencia, acciones, inicio y fin de cada tarea, que representa la ejecución de un archivo bien sea de exportación o importación.

c) Interfaz de consultas

En este campo entran tanto el acceso web que tiene el datawarehouse a través del EDW-Consult, así como a través de herramientas de terceros.

Una vista del sitio web creado para las consultas predefinidas del e-datawarehouse se muestra en la figura siguiente:

Figura 6: Interfaz de consulta del EDW

Desde esta herramienta es posible realizar las cuatro principales operaciones de lo que un modelo conceptual de datawarehouse [1], esto es:

Pivoting: Se rota el cubo para ver una cara en particular. Por Ej. : analizar registros referidas a los diferentes campus.

Figura 7: Muestra de Pivoting del EDW-Consult

Rep. Registros por Campus, Carrera, Año y Sectores

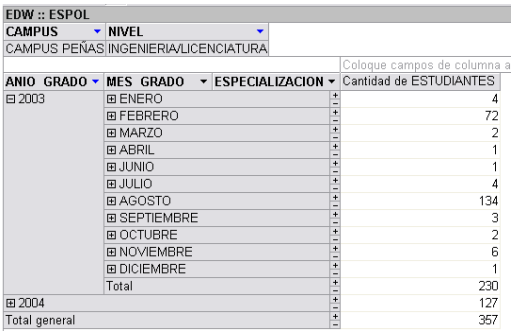
EDW :: ESPOL

Coloque campos de filtro aquí

	Coloque campos de columna aquí
CAMPUS	Cantidad de ESTUDIANTES
CAMPUS DAULE	1543
CAMPUS PEÑAS	65112
CAMPUS PLAYAS	97
CAMPUS PROSPERINA	236516
CAMPUS SAMBORONDÓN	576
CAMPUS SANTA ELENA	5722
Total general	309566

Slicing Dicing: Se selecciona algún subconjunto del cubo. Analizar el cubo de graduados restringiéndolo para algunos campus, carreras, años y niveles.

Figura 8: Muestra de Slicing Dicing del EDW-Consult



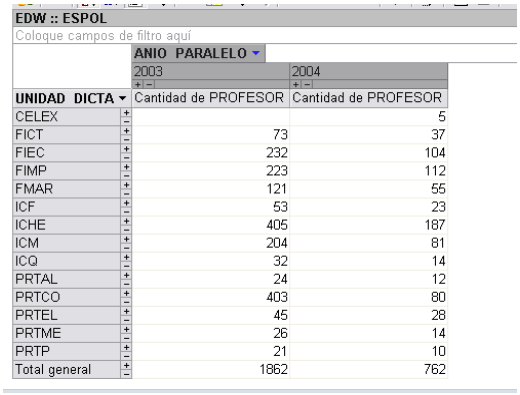
EDW :: ESPOL

CAMPUS	NIVEL
CAMPUS PEÑAS	INGENIERIA/LICENCIATURA

ANIO_GRADO	MES_GRADO	ESPECIALIZACION	Cantidad de ESTUDIANTES
2003	ENERO		4
	FEBRERO		72
	MARZO		2
	ABRIL		1
	JUNIO		1
	JULIO		4
	AGOSTO		134
	SEPTIEMBRE		3
	OCTUBRE		2
	NOVIEMBRE		6
	DICIEMBRE		1
	Total		230
2004			127
Total general			357

Roll Up: Se agrupa por alguna dimensión determinada. Por Ej. : Analizar los profesores de una Unidad por años.

Figura 9: Muestra de Roll-up del EDW-Consult



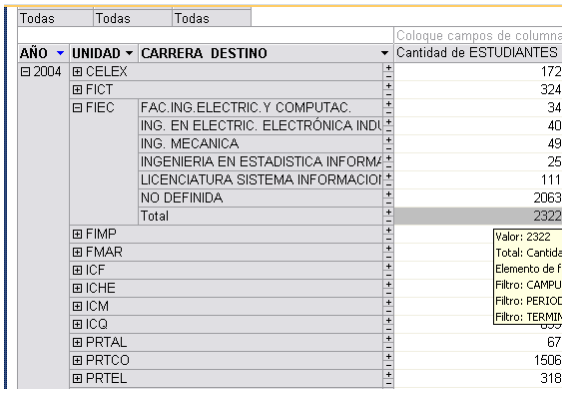
EDW :: ESPOL

Coloque campos de filtro aquí

	ANIO_PARALELO	
	2003	2004
UNIDAD_DICTA	Cantidad de PROFESOR	Cantidad de PROFESOR
CELEX		5
FICT	73	37
FIEC	232	104
FIMP	223	112
FMAR	121	55
ICF	53	23
ICHE	405	187
ICM	204	81
ICQ	32	14
PRTAL	24	12
PRTCO	403	80
PRTEL	45	28
PRTME	26	14
PRTP	21	10
Total general	1862	762

Drill Down: Operación inversa: muestra información detallada de cada agrupamiento. Por Ej. Analizar los registros por Unidad que dicta la materia y la carrera a la que fue destinada.

Figura 10: Muestra de Drill down del EDW-Consult



Todas	Todas	Todas

AÑO	UNIDAD	CARRERA_DESTINO	Cantidad de ESTUDIANTES
2004	CELEX		172
	FICT		324
	FIEC	FAC.ING.ELECTRIC.Y COMPUTAC.	34
		ING. EN ELECTRIC. ELECTRÓNICA INDU	40
		ING. MECANICA	49
		INGENIERIA EN ESTADISTICA INFORMA	25
		LICENCIATURA SISTEMA INFORMACIO	111
		NO DEFINIDA	2063
		Total	2322
	FIMP		
	FMAR		
	ICF		
	ICHE		
	ICM		
	ICQ		
	PRTAL		67
	PRTCO		1506
	PRTEL		318

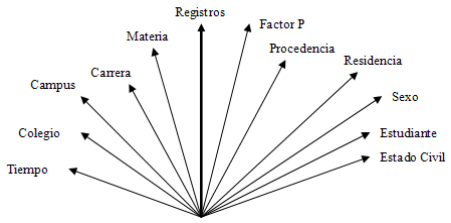
Valor: 2322
Total: Cantida
Elemento de f
Filtro: CAMPU
Filtro: PERIOD
Filtro: TERMIN

Implementación de las dimensiones en la Base de Datos

La implementación de las dimensiones del Datawarehouse se basa en un proceso de diseño con estructura multidimensional. La visualización conceptual de requerimientos no tiene que ser a nivel atómico pero debe representar la vista del mundo del analista de negocios [6].

Consideremos en primer lugar el modelo de consultas de **Registros** con las dimensiones involucradas. Se ilustra en el gráfico a continuación:

Figura 11: Modelo de consultas de Registros



Cada extremo del modelo representa una "cara" o criterio del cubo de datos a través del cual éste puede ser manipulado, es decir, se pueden responder requerimientos tales como:

- ¿Cuántos registrados del sexo femenino hay por campus, colegio de procedencia y año de ingreso?
- Comparación entre cantidades de graduados por carrera, factor p, ciudad de varios periodos de tiempo.
- ¿Cuál será la proyección de crecimiento (o decrecimiento) en ingreso de estudiantes para el próximo ciclo académico de una especialización determinada?
- Lista los 5 principales colegios por ciudad que aportan más estudiantes a la Escuela.

El modelo de consulta de **Graduados** abarca las siguientes dimensiones mostradas en la figura 12, en tanto que el modelo de consulta de Estudiantes abarca las dimensiones mostradas en la figura 13:

Figura 12: Modelo de consultas de Graduados

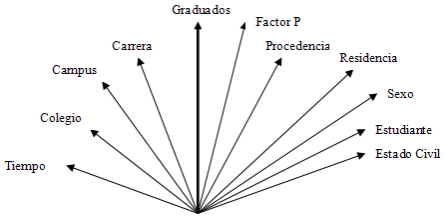
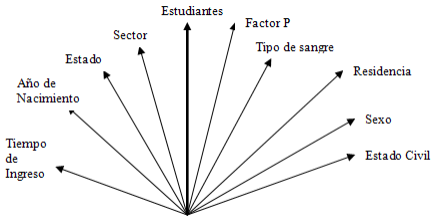
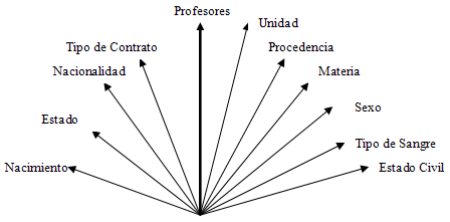


Figura 13: Modelo de consulta de Estudiantes



El modelo de consulta de **Profesores** abarca las siguientes dimensiones mostradas en el gráfico:

Figura 14: Modelo de consultas de Profesores



Estos modelos permiten dar una idea más clara de lo que un diseño multidimensional persigue, que no es otra cosa que ofrecer una vista más entendible y familiar al analista del negocio (usuarios del área escogida).

Cabe mencionar el empleo de dimensiones jerárquicas que incluyen agrupamiento y especialización de un criterio de selección, como es el caso de las dimensiones Procedencia, Residencia, Colegio, Tiempo, entre otras.

Migración de datos

"La migración de datos abarca tanto las fases de diseño de las transformaciones, sumarización, conversión histórica, controles, comprobación de datos, corrida del poblado y verificación auditada de información"[3].

La siguiente tabla lista las principales actividades realizadas en la migración de datos del datawarehouse de la ESPOL [8].

Tabla 3: Actividades realizadas en la migración de datos

ACTIVIDAD	DESCRIPCIÓN
Diseño de transformaciones	Definición de requerimientos del datawarehouse. Creación de procesos de limpieza, filtrado, conversión y estandarización de datos. Definición de proceso de extracción de datos ya trasformados.
Sumarización	Generación de campos calculados a partir de datos individuales
Conversión histórica	Se abarca la transformación de los períodos académicos que habían sido llevados a lo largo de la historia de la ESPOL con dos componentes que eran año y término, hacia un nuevo esquema en el que se incorpora la entidad del período que representa la duración de un ciclo dependiendo de la naturaleza de la carrera bajo la que este sustentada.
Controles de transformación	Se basan en obtener totales tanto en el datawarehouse como en los sistemas OLTP de los cuales se extrajeron; dicha comprobación no es sino la verificación de estos totales en cada una de la entidades que compongan el conjunto de información
Comprobación de datos	Existe un log que detalla la cantidad de datos tanto exportados (obtenidos de las fuentes de legado) como importados (cargados en el repositorio del datawarehouse).
Corrida del poblado	Se realiza a través de procesos programados según los requerimientos de los usuarios. La programación del poblado es realizada a través de una herramienta propia del DBMS
Verificación de información	Debe ser realizada con usuarios representativos de las áreas involucradas. Requiere verificaciones de totales por rango de tiempo, comprobación de cálculos y demás auditorias de datos.

III. CONCLUSIONES

- Desde comienzos de la era de la computación, las organizaciones han usado los datos desde sus sistemas operacionales para atender sus necesidades de información. Algunas proporcionan acceso directo a la información contenida dentro de las aplicaciones operacionales. Otras, han extraído los datos desde sus bases de datos operacionales para combinarlos de varias formas no estructuradas, en su intento por atender a los usuarios en sus necesidades de información.

- Ambos métodos han evolucionado a través del tiempo y ahora las organizaciones manejan una data no limpia e inconsistente, sobre las cuales, en la mayoría de las veces, se toman decisiones importantes.

- La gestión administrativa reconoce que una manera de elevar su eficiencia está en hacer el mejor uso de los recursos de información que ya existen dentro de la organización. Sin

embargo, a pesar de que esto se viene intentando desde hace muchos años, no se tiene todavía un uso efectivo de los mismos.

- La razón principal es la manera en que han evolucionado las computadoras, basadas en las tecnologías de información y sistemas. La mayoría de las organizaciones hacen lo posible por conseguir buena información, pero el logro de ese objetivo depende fundamentalmente de su arquitectura actual, tanto de hardware como de software.

- Aunque diversas organizaciones y personas individuales logran comprender el enfoque de un Warehouse, la experiencia ha demostrado que existen muchas dificultades potenciales.

- Reunir los elementos de datos apropiados desde diversas fuentes de aplicación en un ambiente integral centralizado, simplifica el problema de acceso a la información y en consecuencia, acelera el proceso de análisis, consultas y el menor tiempo de uso de la información.

- Las aplicaciones para soporte de decisiones basadas en un datawarehousing, pueden hacer más práctica y fácil la explotación de datos para una mayor eficacia del negocio, que no se logra cuando se usan sólo los datos que provienen de las aplicaciones operacionales o transaccionales (que ayudan en la operación de la empresa en sus operaciones cotidianas), en los que la información se obtiene realizando procesos independientes y muchas veces complejos.

- DataWarehousing es un proceso, no un producto. Estos sistemas de información no se pueden comprar, hay que construirlos, pero en su diseño y construcción es necesario buscar estrategias tecnológicas de garantía.

REFERENCIAS

[1]. Ralph Kimball, Ed. John Wiley & Sons. "The Datawarehouse Toolkit", [1996]
[2]. Jill Dyché, Ed. Prentice Hall, "E-data. Transformando datos en información con Datawarehousing", [2000]].
[3]. Bill Inmon, Ed. John Wiley & Sons, Inc, "Building the Datawarehouse",[1998]
[4]. John Ladley, Ed. Prentice Hall, "Practical Advice from the Experts", [1998]
[5]. Poe, Vidette, Klauer, Patricia and Brobst, Stephen, Ed. Prentice Hall, "Building a Datawarehouse for Decision Support, Second Edition" [1998].
[6]. Ramon Barquin and Herb Edelstein, chapter 10, Prentice Hall PTR, "Planning and Designing the Datawarehouse", [1996].
[7]. Francis Stevens George, http://www.krooman.com, "Internet Technology and your Business: An Introduction and Guide"
[8]. Juan C. Bustamante y José F. Rodríguez, "Datawarehousing y su aplicación como Apoyo a la Toma de Decisiones en la Espol" (Tesis, Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, 2005)