SISTEMA DE APOYO A LA TOMA DE DECISIONES PARA EL REPOSITORIO DIGITAL DE OBJETOS DE APRENDIZAJE SPAR 1.0

DIEGO FERNANDO BAYONA VALVERDE

ALEXANDER CALVACHE FERNÁNDEZ

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE SISTEMAS
GRUPO DE I+D EN TECNOLOGÍAS DE LA INFORMACIÓN
POPAYÁN
Enero de 2008

SISTEMA DE APOYO A LA TOMA DE DECISIONES PARA EL REPOSITORIO DIGITAL DE OBJETOS DE APRENDIZAJE SPAR 1.0

Trabajo de grado para optar el titulo de Ingenieros de Sistemas

Diego Fernando Bayona Valverde

Alexander Calvache Fernández

Director:

MSc. Martha Eliana Mendoza

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE SISTEMAS
GRUPO DE I+D EN TECNOLOGÍAS DE LA INFORMACIÓN
POPAYÁN
Enero de 2008

AGRADECIMIENTOS

Al Ser Supremo, a la Vida, al Conocimiento y a la Sabiduría Universal.

A los seres vivos con los que compartimos experiencias que nos permiten tomar conciencia de lo que somos y hacia dónde vamos.

A la mano extendida y generosa de aquellos que nos aman, que nos regalan sus sueños y sus esperanzas.

A la Calidez de los amigos y compañeros que con su invaluable apoyo y colaboración nos brindan confianza y fortaleza para alcanzar nuestros anhelos.

Y aquellas personas que compartieron sus conocimientos y sembraron en nuestro ser las aspiraciones de compartir los nuestros.

Gracias…

Diego Fernando Bayona Valverde, Alexander Calvache Fernández

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

En esta sección se presenta la definición del problema, justificaciones del desarrollo del proyecto y los objetivos general y específicos que pretende dar solución a la problemática planteada, finalmente se hace una descripción de la estructura de este documento.

DEFINICIÓN DEL PROBLEMA

La educación es un factor primordial, estratégico, prioritario y condición esencial para el desarrollo social y económico de cualquier conglomerado humano. Además, es un derecho universal, un deber del Estado y de la sociedad, y un instrumento esencial en la construcción de sociedades autónomas, justas y democráticas. De su cobertura y calidad dependen las posibilidades que tiene un país de incrementar su desarrollo social, económico y cultural, además de poder enfrentar las exigencias competitivas a nivel internacional.[1]

En la actualidad Colombia está desarrollando estrategias como La Agenda de Conectividad, Computadores para Educar, Compartel, Centros Regionales de Educación Superior (CERES) [1], para crear programas educativos a distancia y virtuales que contribuyan a dar respuesta a las necesidades de cobertura y calidad de la educación que requiere el país para alcanzar mejores condiciones de desarrollo social y económico y mejorar la calidad de vida de la población. Estas estrategias se desarrollan en alianza con instituciones de educación superior que posibilitan el uso compartido de recursos humanos, tecnológicos, de infraestructura y conectividad.

Con el firme propósito de apoyar las estrategias educativas del gobierno, la Universidad del Cauca ha desarrollado varios proyectos educativos entre los cuales se encuentra un Repositorio Digital de Objetos de Aprendizaje SPAR 1.0 [2], que da la posibilidad de almacenar y compartir recursos educativos empleados para apoyar experiencias de enseñanza y aprendizaje, ofreciendo una alternativa para potenciar el uso y aprovechamiento de las Tecnologías de la Información y la Comunicación(TIC), permitiendo descentralizar, mejorar la oferta y lograr la equidad de la educación.

Teniendo en cuenta que SPAR 1.0 ya se encuentra en funcionamiento e inicialmente lo están utilizando docentes del Departamento de Ingeniería de Sistemas, se presenta una oportunidad de desarrollar un Sistema a la Toma de Decisiones (DSS), para analizar la información, marcar tendencias, señalar problemas, fortalecer servicios y tomar decisiones estratégicas que contribuyan a mejorar la calidad de los servicios y recursos educativos del Repositorio Digital, logrando un impacto favorable en la comunidad universitaria de la región. En la medida en que los programas presenciales o de educación en línea, utilicen el Repositorio Digital para el desarrollo de las asignaturas se verán los resultados esperados.

Poder contar con DSS para medir el grado de satisfacción de los usuarios con respecto a los recursos educativos ofrecidos y los servicios prestados por el Repositorio Digital, brindará la posibilidad de establecer estrategias y descubrir información útil para

determinar intereses temáticos, características de los usuarios y el nivel de demanda y oferta de los recursos educativos entre otros.

JUSTIFICACIÓN

Justificación Tecnológica

Para la realización de este proyecto se utilizaron técnicas y teorías de *Data Warehouse* (DW), procesamiento analítico en línea (OLAP) y minería de datos, que han dado buenos resultados en el área de gestión de negocios, donde el análisis de información es de gran importancia para el mundo competitivo. Este proyecto aplicó estas metodologías y tecnologías de inteligencia de negocios para almacenar y analizar la información generada por el Repositorio de Objetos Digitales (SPAR) logrando aprovechar al máximo las posibilidades de interacción, acceso e incorporación de contenidos educativos y permitiendo crear un marco de trabajo que ayude a estudiar e investigar como algunas técnicas de minería de datos, *Data Warehouse* y Procesamiento Analítico en Línea, pueden ser aplicadas para descubrir posibles tendencias, patrones y comportamientos de interés de los usuarios para brindar sugerencias, recomendaciones y personalización, basados en el análisis del comportamiento de navegación durante una sesión o en distintas sesiones de un mismo usuario, o en la información brindada por ellos directamente.

El desarrollo de este proyecto representa para la Universidad del Cauca y para la región en general un adelanto tecnológico innovador, que aplica Inteligencia de Negocios a un Repositorio Digital por medio de un DW, una herramienta OLAP y Minería de Datos, ofreciendo la posibilidad de fortalecer la educación en nuestro contexto regional y nacional.

Justificación Académica

La realización del proyecto da la posibilidad a docentes y estudiantes de investigar y profundizar en la aplicación de la minería de datos, OLAP y DW en la educación en línea, buscando generar nuevas soluciones que ayuden a consolidar las ventajas de la educación virtual dentro del nuevo contexto educativo. De la misma manera, contar con un sistema para la toma de decisiones fomenta el uso y el aprovechamiento de los sistemas educativos virtuales, porque brinda sugerencias, recomendaciones y personalización a los usuarios, permitiendo satisfacer de una mejor manera las necesidades académicas de consulta y publicación, igualmente ayuda a proponer estrategias para mejorar la calidad de los servicios al almacenar y analizar información de los recursos educativos y las características de los usuarios.

Justificación Social

En la actualidad el gobierno nacional está desarrollando programas de educación superior en poblaciones apartadas de las ciudades y municipios que cuenten con Universidades[1],

[1] Programas Ministerio de Educación Nacional: Agenda de Conectividad, Computadores para Educar, Compartel, Centros Regionales de Educación Superior (CERES).

la Universidad del Cauca participa de este proceso de descentralización de la educación a través de programas de pregrado que se están llevando a poblaciones del Cauca y a resguardos indígenas a través de educación a distancia y educación virtual. El Repositorio Digital es una alternativa que contribuye con estos programas de educación, ofreciendo acceso a contenidos y servicios educativos, por tal razón, fue importante el desarrollo de un sistema para la toma de decisiones que contribuya a mejorar y fortalecer la calidad del material educativo del Repositorio, de tal forma que se permita analizar la información, marcar tendencias, señalar problemas, fortalecer servicios, con el propósito de realizar seguimiento y evaluación de los recursos y servicios educativos del repositorio para mejorar el impacto sobre los procesos educativos que se llevan a cabo en las comunidades.

OBJETIVOS

OBJETIVO GENERAL

Desarrollar un sistema de soporte a la toma de decisiones estratégicas (DSS) basado en tecnologías de DW, OLAP y Minería de Datos, que permita almacenar, analizar y descubrir información útil y confiable para la gestión de los servicios y recursos educativos con que cuenta el Repositorio Digital de Objetos de Aprendizaje (SPAR).

OBJETIVOS ESPECÍFICOS

- Diseñar y construir un DW que permita almacenar información relacionada con la consulta, publicación y el grado de satisfacción de los usuarios con respecto a los recursos educativos, además descubrir tendencias temáticas y características de navegabilidad de los usuarios a través del repositorio.

- Construir un prototipo de herramienta OLAP que utilice servicios de SQL Server 2005, herramientas y/o componentes, que permitan analizar la información contenida en el DW. Esta aplicación tendrá las siguientes funcionalidades:

  - Acceder el servidor de análisis de SQL Server 2005, permitiendo al usuario crear y administrar cubos OLAP que ofrezcan respuesta a las necesidades de análisis de datos sobre el Repositorio Digital.
  - Consultar datos en cubos y crear cubos locales que permitan el análisis de datos off-line (desconectado del servidor de análisis de SQL Server 2005)

- Definir los criterios para la selección de algoritmos de Minería de Datos, de los proporcionados por Microsoft SQL Server 2005, de forma que se adapten a las necesidades y al contexto del proyecto.

- Diseñar y construir una herramienta de minería de datos que permita descubrir información de interés como patrones y reglas dentro de los datos almacenados en el DW. El desarrollo de este módulo implica:

  - Desarrollar un módulo de Minería de Datos, que utilice un algoritmo de los proporcionados por Microsoft SQL Server 2005, para la elección del algoritmo se tendrán en cuenta algunas de las necesidades de información del repositorio.
  - Permitir la realización de análisis de datos sobre el DW utilizando el algoritmo de minería de datos seleccionado.
  - Integrar el módulo de Minería de Datos en la aplicación de soporte a la toma de decisiones desarrollada.

- Integrar el Sistema de Soporte a la Toma de Decisiones al Repositorio Digital de Objetos de Aprendizaje SPAR 1.0.

El documento se compone de las siguientes secciones:

Capitulo 1 - Marco Teórico: Se describen las bases conceptuales sobre DW, OLAP y Minería de Datos que se utilizaron para el desarrollo del proyecto de grado.

Capitulo 2 - Descripción del proceso de desarrollo del DSS: Se hace una descripción de todo el ciclo de vida usado para el desarrollo del sistema de DSS, desde su concepción inicial a hasta su implementación y despliegue.

Capitulo 3 - Descripción del proceso de desarrollo de la herramienta OLAP: Se describe el proceso de desarrollo que se utilizó para la construcción de la herramienta OLAP y se presentan los artefactos que se obtuvieron en cada una de las fases de la metodología de trabajo.

Capitulo 4 - Descripción del proceso de desarrollo del módulo de minería de datos*:* Se hace una descripción de cada una de las fases del proceso de minería de datos y de sus correspondientes tareas que permitieron el desarrollo del módulo de minería de datos

Capitulo 5 - Descripción del proceso de desarrollo de la herramienta de administración de objetos de análisis: Se describe el proceso de desarrollo de la herramienta de administración (SPARAMO) y se presentan los artefactos que se obtuvieron en cada una de las fases de la metodología de trabajo

Conclusiones, recomendaciones y trabajo futuro: En esta sección se muestran las conclusiones obtenidas del desarrollo de todo el proyecto, se hacen recomendaciones y descripciones sobre trabajos futuros.

Referencias bibliográficas: En esta sección se presenta la bibliografía consultada para la realización de este proyecto.

CAPITULO I:

1. MARCO TEÓRICO

En este capítulo se describen las bases conceptuales sobre DW, OLAP y Minería de Datos que se utilizaron para el desarrollo del proyecto de grado.

1.1 Sistemas de Soporte para la Toma de Decisiones (DSS Decision Support Systems)

El DSS es un sistema de inteligencia de negocios que transforma los datos que tiene una organización en información, con el objetivo de tomar decisiones estratégicas mejores y más rápidas basadas en un análisis dimensional, es decir, considerando unas variables en relación con otras y no de forma independiente, permitiendo enfocar el análisis desde distintos puntos de vista [3].

Para los propósitos del presente proyecto se hace referencia a los DSS como un sistema que se basa en una bodega de datos (DW) y que crea una base de datos multidimensional, la cual provee los mecanismos para navegar y profundizar en los datos almacenados en el DW por medio de herramientas que permiten el procesamiento analítico (OLAP) de la información desde una perspectiva multidimensional [4].

Los DSS pueden incluir técnicas de minería de datos que constituye un paso más en el análisis de los datos para apoyar la toma de decisiones, permitiendo realizar un análisis más avanzado que involucra el descubrimiento de patrones para identificar tendencias y establecer relaciones en los datos [4]. La Figura 1 muestra la arquitectura general de un DSS.

La Figura 1 muestra la arquitectura y el flujo de un DSS basado en un DW, en donde se hace la extracción de datos de distintas fuentes de datos relacionales que posteriormente son transformados, conformados y limpiados por medio de un proceso de ETL (Proceso de extracción, transformación y carga de datos) para finalmente ser cargados en un DW, el cual se compone de subconjuntos lógicos y físicos de información que representa áreas o proceso de negocio denominados Data Marts, el flujo continua hacia la construcción de una bodega de datos OLAP que almacena datos en cubos multidimensionales que producen información analítica valiosa que es extraída por herramientas OLAP. Como muestra la Figura 1, las técnicas de minería de datos pueden ser aplicadas sobre los datos almacenados en fuentes de datos multidimensionales para descubrir tendencias y comportamientos en los datos, apoyando la toma de decisiones estratégicas.

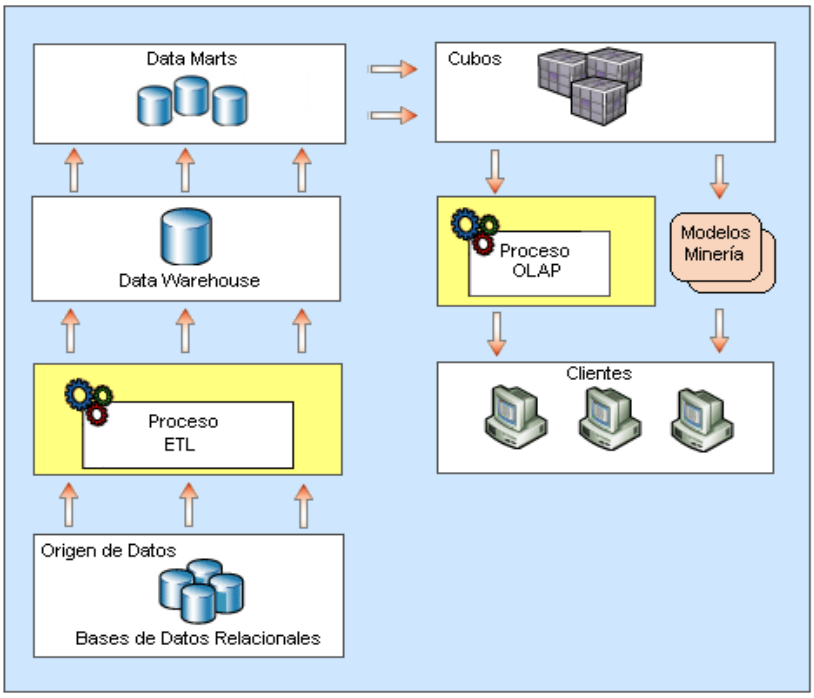


Figura 1: Arquitectura de un DSS (Adoptada de [3])

Aunque los límites y el concepto de un DSS no han sido precisados completamente, actualmente muchos analistas de la industria describen un DSS como un sinónimo de inteligencia de negocios y muchos otros hacen referencia a inteligencia de negocios como un concepto separado del DW.

Para este proyecto se considera un DSS como un sinónimo de *Inteligencia de Negocios* que incluye al DW, debido a que en la Metodología Kimball [7] el DW incluye todo el sistema desde la extracción de datos desde los sistemas fuentes hasta el software y las aplicaciones de los usuarios finales. Kimball usa la frase "Sistema de *Data Warehouse/ Business Intelligence (DW/BI)*" para referirse a la totalidad del sistema y para evitar confusión entre estos términos de esta manera se adopta el concepto de DW/BI por ser el más apropiado para el presente proyecto [5].

1.2 Conceptos Básicos del DW

1.2.1 Bodegas de Datos (DW)

Para la definición del DW se toma la definición de Bill Inmon, "El DW es una colección de datos integrados, orientados a temas, que dan soporte a las funcionalidades del DSS, donde cada unidad de dato es relevante en algún momento en el tiempo [6].

La información almacenada en el DW es histórica y se obtiene de un proceso de transformación, limpieza e integración de datos provenientes de bases de datos fuentes transaccionales. Estos datos históricos son estructurados para consultas y análisis permitiendo tomar decisiones empresariales a diferentes niveles [5].

1.2.2 Modelado Dimensional

Es una metodología para el modelado lógico de datos, un modelo dimensional contiene una estructura simétrica propia que permite rendimiento de consultas, facilidad de uso, entendimiento y flexibilidad a los cambios. El modelo dimensional está compuesto por una tabla central llamada Tabla de Hechos y sus dimensiones asociadas por lo que recibe el nombre de Esquema Estrella [5].

1.2.3 Data Mart

El Data Mart es un subconjunto completo de todo el DW, debe estar basado en datos lo más granulares posibles que puedan ser extraídos desde una sistema fuente operacional y que deben ser representados por un modelo dimensional. El Data Mart es generalmente construido en función de un único proceso del negocio, permitiendo responder a necesidades específicas de un área o departamento específico de la organización [7].

1.2.4 Arquitectura Bus

Esta arquitectura se comporta como la columna vertebral del DW, que permite conectar todos los Data Marts por medio de sus hechos y dimensiones conformados (comunes). Esta arquitectura permite presentar el DW como un todo coherente y no como múltiples islas separadas e independientes [7].

1.2.5 Indicadores o Medidas

Son las variables o métricas que ayudarán a medir el desempeño del negocio. Generalmente son numéricas y aditivas y son almacenadas en las tablas de hechos [7].

1.2.6 Tabla de Hechos

Es la tabla central dentro de un esquema estrella (modelo dimensional) que contiene las medidas asociadas con eventos que ocurren en un proceso de negocio específico. Estos eventos usualmente tienen medidas numéricas que cuantifican la magnitud del evento tal como la cantidad de una orden, cantidad de venta o duración de una llamada. Además de las medidas, la tabla de hechos está compuesta por un conjunto de llaves foráneas que son extraídas de las tablas de dimensiones con las cuales está relacionada [7].

1.2.7 Tipos de Tabla de Hechos

- **Tabla de Hechos de Transacción:** Rastrea o hace seguimiento a cada transacción que ocurre en un punto del tiempo, cuando el evento de la transacción ha ocurrido [5].

- **Tabla de Hechos Factless:** Son tablas de hecho que se utilizan para describir y registrar eventos en un DW dónde no hay ninguna medida numérica natural asociada con el evento. También se usan para garantizar un respaldo de información a otras tablas de hechos permitiendo responder preguntas de los eventos que no ocurrieron [7].

- **Tabla de Hechos de Vista Periódica:** muestra el desempeño del negocio al final de intervalos de tiempo específicos donde la granularidad es de una fila por periodo de tiempo, se utiliza en situaciones donde los cálculos de la historia de las transacciones es muy difícil, también agregan muchos de los hechos a través del periodo de tiempo, provee a los usuarios un modo rápido de obtener totales [5].

- **Tabla de Hechos de Vista Acumulativa:** Son constantemente actualizadas en el tiempo, su diseño generalmente incluye muchas llaves foráneas de la dimensión fecha, para capturar las fechas cuando un ítem en cuestión pasa a través de cada proceso del negocio o puntos en la cadena de valor [5].

1.2.8 Tabla de Dimensión

Son la base del modelo dimensional, describen los objetos del negocio, tales como empleado, cliente, producto, servicio, suscriptor, entre otras. “Las dimensiones son los sustantivos del DW, los procesos del negocio (hechos) son los verbos o acciones del negocio en los cuales los sustantivos (dimensiones) participan. Cada dimensión se enlaza a todos los procesos de negocio (Data Marts) donde ella participa” [5]. La tabla de dimensión está compuesta de una llave primaria y de columnas de atributos descriptivas.

1.2.9 Granularidad

El nivel de detalle contenido en la tabla de hechos es llamado granularidad. La declaración de la granularidad de una tabla de hecho es el segundo de los cuatro pasos clave en el diseño de un modelo dimensional [8].

1.2.10 Dimensiones Conformadas

Son aquellas dimensiones que permiten la integración de los Data Marts, puesto que son dimensiones que significan lo mismo en cada posible tabla de hechos a la que está enlazada. Generalmente esto significa que una dimensión conformada es idénticamente la misma dimensión en cada data mart [8].

1.2.11 Dimensiones que Cambian Lentamente

Son aquellas dimensiones que tienen atributos que cambian con el tiempo. Como por ejemplo el departamento al que pertenece un empleado. Para hacer rastreo a los cambios en los atributos se utilizan técnicas de seguimiento como se describen a continuación [5]:

- **Tipo 1:** El tipo 1 de dimensión que cambia lentamente sobrescribe el valor con un nuevo valor, se usa si no es necesario hacer un seguimiento histórico.

- **Tipo 2:** Si se necesita hacer un seguimiento histórico. Cuando ocurre un cambio, el proceso ETL crea una nueva fila en la dimensión para capturar el nuevo valor. Este nuevo valor tiene efecto de aquí en adelante, el atributo previo es marcado para mostrar que tuvo efecto hasta antes del cambio.

- **Tipo 3:** Este Tipo de técnica de seguimiento conserva columnas separadas para el viejo y el nuevo valor del atributo. Es poco común que se aplique esta técnica por que involucra cambiar las tablas físicas y no es muy escalable.

1.2.12 Dimensiones Degeneradas o de Hechos

Son dimensiones que no tienen atributos descriptivos propios. No se manejan como dimensiones separadas, si no como un atributo en la tabla de hechos [5].

1.2.13 Tabla de Subdimensión

Es una tabla que se obtiene de la división de un conjunto de atributos que hacen parte de otra dimensión [7].

1.2.14 Relaciones

Como su nombre lo indica son las relaciones existentes entre dimensiones y tablas de hechos o entre dimensiones diferentes o la misma dimensión [5].

1.2.15 Tipos de Relaciones [5]

- **Relaciones Muchos a Muchos o Multivaluadas**
    - **Entre Tabla de Hechos y Dimensión:** Ocurre cuando múltiples valores de una dimensión pueden ser asignados a una transacción de una tabla de hechos (una fila). Ejemplo: cuando múltiples vendedores pueden ser asignados a una venta. La solución es crear una tabla puente que agrupe los vendedores en distintos grupos.
    - **Entre Dimensiones:** Existen dimensiones que no son totalmente independientes de otra. Ejemplo: cuentas de banco y clientes. Es difícil combinarlas en una sola dimensión a causa de su relación de muchos a muchos. Una solución es crear una tabla puente entre la tabla de hechos y la dimensión, la otra solución más efectiva es crear una tabla puente entre las dos dimensiones.
    - **Relación Normal o Regular:** La relación estándar entre una dimensión y una tabla de hechos es de uno a muchos. Significa que una fila de la dimensión se une a muchas filas en la tabla de hechos, pero una fila en la tabla de hechos se une a solo una fila en la tabla dimensión.
    - **Relación de Hechos:** Es la relación existente entre una tabla de hechos y una dimensión de hechos o degenerada.

1.2.16 Llaves sustitutas

Son llaves independientes del sistema transaccional, con un único valor, por lo general un entero, asignado a cada fila de la dimensión, esta llave se convierte en la llave primaria de la dimensión y la foránea de la tabla de hechos a la que está relacionada la dimensión [5].

1.2.17 Jerarquías

Las jerarquías son principalmente modos estándar para agrupar datos dentro de una dimensión. Comenzando desde un nivel alto de detalle hasta un nivel más bajo (drill down). Las jerarquías más comunes son las organizacionales, jerarquías geográficas basadas en una localización física, jerarquías de productos correspondientes a categorías y subcategorias, jerarquías de responsabilidades, jerarquías temporales, entre otras [5].

1.2.18 Sistema de Extracción, Transformación y Carga (ETL)

El sistema ETL es la base fundamental del DW. El proceso ETL involucra los procesos de Extracción, Transformación y Carga de datos. El proceso de extracción es aquel que permite obtener los datos desde los sistemas fuente. Los procesos de transformación se utilizan para limpiar, convertir y conformar datos para asegurar su calidad y su consistencia. Los procesos de carga de datos son los procesos requeridos para poblar el modelo físico del DW, la calidad de los datos es un factor determinante en la construcción del DW [8].

1.2.19 Perfilado de Datos

El perfilado de datos es un paso que debe empezar antes de la construcción del sistema ETL, en donde se debe realizar un análisis de los orígenes de datos, en este paso debe realizar un plan de limpieza y corrección de los datos. El objetivo de hacer el perfilado es obtener la información necesaria para realizar el mapeo de datos desde las fuentes a sus respectivos destinos [5].

1.2.20 Sistema de Auditoría

Sistema que hace parte del DW y que está estrechamente enlazado con el proceso de ETL, tiene como fin hacer un seguimiento completo de los datos que son cargados dentro del DW, permitiendo responder preguntas como: ¿De dónde vienen los datos?, ¿Qué cálculos o transformaciones fueron realizas?, ¿Fue el procesamiento de carga exitosas? ¿Qué tantas filas fueron cargadas en cada tabla?, entre otras [5].

1.2.21 Procesamiento Analítico en Línea (OLAP)

El termino OLAP hace referencia a un conjunto de tecnologías, principios, actividades, procesos, herramientas especializadas para crear, mantener, analizar y elaborar informes que permiten la toma de decisiones de una organización [6]. Dependiendo de la tecnología usada, OLAP puede ser dividido en dos tipos diferentes: [6]

1.2.21.1 ROLAP (Relational OLAP)

Se caracteriza porque la manipulación de los datos se hace sobre una base de datos relacional a la cual se le da un tratamiento multidimensional.

Características:

- Pueden soportar muchos datos.
- Pueden soportar uniones dinámicas de datos.

- Es capaz de soportar procesamientos de actualización para propósitos generales.
- Tiene muy bajo desempeño de consultas.
- No es una capa orientada al usuario.
- Limitado por el lenguaje de consulta SQL, que es un lenguaje desarrollado primariamente para consultas transaccionales y no para consultas analíticas.

1.2.21.2 MOLAP (Multidimensional OLAP)

Se caracteriza porque la manipulación de los datos se hace sobre un sistema de base de datos multidimensional.

Características

- Alto desempeño de consultas
- Puede ser optimizado para accesos rápidos sobre datos.
- Si los patrones de acceso a datos son conocidos, entonces la estructura de datos puede ser optimizada, permitiendo mejorar el tiempo de respuesta de las consultas al tener preparadas las respuestas antes de que se planteen las preguntas.
- Provee un modelo de datos intuitivo y multidimensional que facilitan la selección, recorrido y exploración de los de múltiples formas, este modelo datos generalmente es almacenado dentro de una estructura multidimensional que recibe el nombre de Cubo.
- Provee un lenguaje analítico de consulta que tiene un completo entendimiento de la metadata multidimensional, lo que proporciona la capacidad de explorar las complejas relaciones existentes entre los datos
- Puede tomar mucho tiempo de carga de datos.

1.2.22 Consultas Predefinidas

Son las consultas que se realizan sobre reportes estáticos en donde se presenta información que ha sido procesada con anterioridad y en el mejor de los casos se configuran parámetros para que el usuario pueda realizar algunas consultas personalizadas con respecto a unos valores preestablecidos [5].

1.2.23 Consultas Dinámicas

Son consultas que se realizan sobre herramientas que permiten establecer nuevos parámetros dinámicos, el usuario cuenta con una interfaz que le permite realizar consultas complejas con una amplia gama de posibilidades que permiten obtener y analizar información de acuerdo a sus necesidades [5].

1.3 Conceptos básicos de Minería de Datos

1.3.1 Descubrimiento del Conocimiento

El proceso de Knowledge Discovery in Databases (KDD) se define como “Proceso no trivial de identificar patrones validos, novedosos, potencialmente útiles y comprensibles en los datos almacenados en grandes bases de datos” [9] este proceso de transformación de

datos en conocimiento incluye las siguientes etapas: **Selección de Datos, Transformación y Limpieza de Datos, Minería de Datos, Interpretación y Evaluación**. Etapas que son ejecutadas *iterativa e interactivamente*. La interactividad del proceso se refiere a que el usuario debe entender y participar de todo el proceso porque es él quien plantea los requerimientos para determinar los alcances del proyecto. La iteratividad hace referencia a que es un proceso cíclico en el que las tareas se ejecutan repetidamente para producir incrementos que mejoran y potencian los resultados [9]. La Figura 2. Vista del Proceso de Descubrimiento de Conocimiento de Bases de Datos muestra el proceso de KDD:

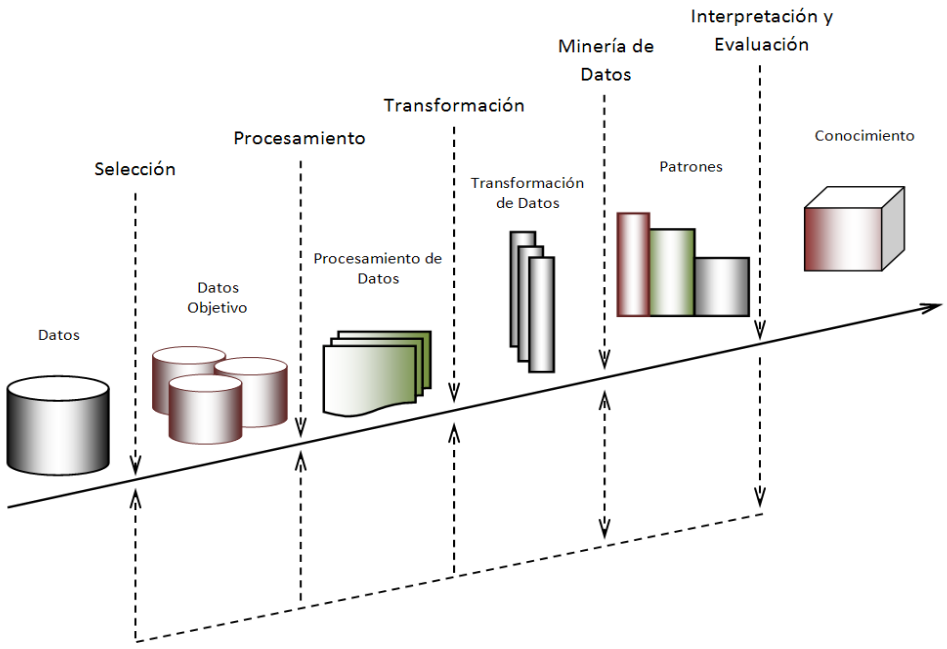


Figura 2. Vista del Proceso de Descubrimiento de Conocimiento de Bases de Datos (Adoptado de [3])

1.3.2 Minería de Datos

La minería de datos puede definirse como un proceso de exploración que tiene como fin encontrar patrones o relaciones que existen en grandes cantidades de datos, permitiendo predecir posibles tendencias y comportamientos que ofrece la posibilidad de tomar decisiones estratégicas. La minería de datos es actualmente aplicada en diferentes áreas como son: la radio, la medicina, procesos industriales de control y especialmente en el área de la inteligencia de negocios, permitiendo a las corporaciones mejorar su mercadeo, incrementar sus ventas y brindar mejor soporte a través de un mejor entendimiento de sus clientes [10]. Además como muestra el presente proyecto puede ser aplicada en la educación virtual, campo donde comienza a brindar potenciales beneficios.

La minería de datos es una de las fases del Proceso de KDD aunque finalmente haya adquirido el significado de todo el proceso en lugar de la fase de extracción de conocimiento. Un proyecto de minería de datos hace uso de diferentes tecnologías y

técnicas que son el resultado de un largo proceso de investigación y desarrollo, y son aplicadas para la solución de diversos problemas como: típicos de agrupamiento automático, clasificación, asociación de atributos, detección de patrones secuenciales, predicciones, estimaciones, detección de anomalías [10].

La minería de datos puede ser abordada desde dos perspectivas diferentes [5]:

La primera permite comprender aspectos relevantes de los datos. Dentro de la Inteligencia de negocios podría ser una comprensión de quienes son sus clientes y cuáles son sus comportamientos. Este acercamiento se llama "exploratorio o minería de datos no dirigida" dónde la meta es encontrar algo interesante.

El segundo acercamiento es llamado "minería de datos dirigida", donde generalmente los modelos generados son aplicados dentro del mismo proceso transaccional para identificar oportunidades o predecir problemas al tiempo que están ocurriendo, permitiendo brindar respuestas adecuadas en una base de tiempo real.

1.3.2.1 Terminología básica de Minería de Datos

A continuación se definen algunos términos que son útiles para la comprensión del módulo de minería de datos del proyecto. Estas definiciones son adoptadas del libro "*The Microsoft Data Warehouse Toolkit*" [5].

- **Algoritmo**

Técnica programática usada para identificar las relaciones o patrones en los datos.

- **El Modelo**

La definición de la relación identificada por el algoritmo que generalmente toma la forma de un conjunto de reglas, un árbol de decisión, un sistema de ecuaciones, o un conjunto de asociaciones.

- **El Escenario**

La colección de atributos y relaciones (variables) que son asociadas con un objeto individual, normalmente un cliente. El escenario también es también conocido como una observación.

- **Conjunto de Escenarios**

Un grupo de escenarios que comparten los mismos atributos. Un conjunto de escenarios se puede pensar como una tabla con una fila por cada objeto (como cliente), es posible tener un conjunto de escenarios anidado cuando una fila en la tabla padre, como "cliente", se une a las múltiples filas en la tabla anidada, como "compras". Conjunto de escenarios también es conocido como conjunto de observaciones.

- **Variable(s) dependiente (columna de predicción)**

La variable que usa el algoritmo para predecir o clasificar. Un algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos, utiliza los valores o estados de estas columnas, para predecir los estados de una columna que se designa como predicción.

- **Variable(s) independiente (columna de entrada)**

La variable con información descriptiva usada para construir el modelo. El algoritmo crea un modelo que usa combinaciones de variables independientes para definir una agrupación o predecir la variable dependiente.

- **Las variables discretas o continuas**

Las columnas numéricas que contienen los valores continuos o discretos. Una columna en la tabla de Empleado llamada Sueldo que contiene los valores del sueldo reales es una variable continua. Pero es posible agregar una columna a la tabla, que contienen los enteros para representan el rango de sueldo, así: (1 = “0 a $25,000”; 2 = “entre $25,000 y $50,000”; y así sucesivamente). Ésta es una columna numérica discreta. Las variables Discretas también son conocidas como categóricas.

- **La estructura de minería**

Un término de minería de datos de Microsoft usado como un nombre para la definición del conjunto de escenarios en Analysis Services. La estructura de minería es esencialmente una capa de metadata encima de una vista de origen de datos que incluye banderas adicionales de minería de datos y propiedades de columna, como el campo que identifica una columna como entrada, de predicción, ambos, o ignorar. Una estructura de minería puede ser usada como la base para múltiples modelos de minería.

- **El modelo de minería**

La aplicación específica de un algoritmo a una estructura particular de minería. Se puede construir varios modelos de minería con parámetros diferentes o algoritmos diferentes partiendo de la misma estructura de minería.

1.3.3 Tareas básicas de negocio

Muchos problemas de interés intelectual, económico y comercial pueden expresarse en términos de tareas de negocios, a continuación se listan y se describen siete tareas de negocio básicas que son atendidas por las técnicas de minería de datos [10] [5]: Clasificación, Estimación, Predicción, Asociación o Afinidad de Grupos, Clustering, Descripción y Perfilado y Detección de Anomalías.

Las tres primeras tareas son ejemplos de minería de datos dirigida, dónde la meta es encontrar el valor de una variable objetivo en particular. La afinidad de grupos y clustering son ejemplos de minería de datos no dirigidas donde la meta es descubrir una estructura

en los datos sin considerar una variable objetivo en particular. El perfilado y la detección de anomalía son tareas descriptivas que pueden ser dirigidas o no dirigidas.

1.3.3.1 Clasificación

La clasificación consiste en analizar las características, atributos y comportamientos de un objeto para asignarlo a un grupo de un conjunto de grupos que han sido predefinidos con anterioridad. Los algoritmos de clasificación se caracterizan porque se usan un número limitado de variables de entrada (discretas) que son evaluadas y que los conjunto de clases o grupos están predefinidos y son opciones discretas como por ejemplo: alto, medio, bajo; si, no; Plata, Oro, Platino dependiendo del contexto [10].

1.3.3.2 Estimación

La estimación es un tipo de clasificación, se diferencia en que los grupos o clases a donde se asigna un objeto son valores continuos y no discretos. El proceso es esencialmente el mismo: Un conjunto de atributos se usa para determinar una relación. La mayoría de los algoritmos de estimación están basados en las técnicas de análisis de regresión. Por esta razón esta tarea de negocio recibe también el nombre de regresión [5].

1.3.3.3 Predicción

La predicción también es un tipo de clasificación, se diferencia de las dos tareas anteriores en que los objetos o registros son clasificados de acuerdo a valores o comportamientos que pueden ser validos en el futuro. Los algoritmos de predicción buscan determinar una clase o estimar con tanta precisión como sea posible un valor antes de que este sea conocido. Las variables de entrada existen u ocurren antes de la variable predictiva [5].

1.3.3.4 Asociación o Afinidad de Grupos

La Asociación busca las correlaciones entre los elementos en un grupo de conjuntos. La asociación es llamada también análisis de cesta de mercado. Un problema típico de negocio que involucra asociación es analizar una tabla de ventas e identificar los productos que a menudo se venden juntos. Esto permite determinar estrategias de mercadeo como hacer recomendaciones a los clientes, planificar el arreglo de los productos en las tiendas o en un catalogo a fin de que los productos que se han comprado juntos se encuentren juntos. La asociación es generalmente usada para identificar conjuntos de elementos relacionados y reglas para propósitos de ventas cruzadas. En términos de asociación, cada conjunto de atributos/valores es considerado un elemento. La tarea de asociación tiene dos metas: Encontrar conjuntos de elementos frecuentes y encontrar reglas de asociación.

La mayoría de los algoritmos de asociación encuentran conjuntos de elementos que ocurren con frecuencia, escaneando los conjuntos de datos muchas veces. El límite de frecuencia (soporte) es definido por el usuario antes de procesar el modelo. Por ejemplo para un sistema de E-Learning un soporte del 2% significa que el modelo analiza solo elementos que aparecen al menos en el 2% de las consultas de los usuarios. Un conjunto

de elementos frecuente puede verse así: {Recurso Educativo = “Introducción a UML”, Recurso Educativo: “Programación estructurada”, Recurso Educativo: “Programación modular”}. Cada conjunto de elementos tiene un tamaño, el cual es el número de elementos que este contiene. El tamaño de este particular conjunto de elementos es de tres (3).

Aparte de identificar conjuntos de elementos frecuentes basados en el soporte, la mayoría de algoritmos de asociación también encuentran reglas. Una regla de asociación tiene la siguiente forma: A, B → C con una probabilidad, donde A, B, C son los conjuntos de elementos frecuentes.

Esta probabilidad también es conocida como la *confianza* en la literatura de Minería de Datos. La probabilidad es un valor límite que el usuario necesita especificar antes de entrenar un modelo de asociación. Por ejemplo, revisemos la siguiente regla:
Recurso Educativo = “Introducción a UML”, Recurso Educativo: “Programación estructurada”, → Recurso Educativo: “Programación modular” con una probabilidad del 80%; esto quiere decir que si un usuario consulta el recurso “Introducción a UML” y el recurso “Programación Estructurada” hay un 80% de probabilidad de que él consulte el recurso educativo “Programación Modular”. La Figura 3 muestra los patrones de asociación de los recursos, cada nodo en la figura representa un recurso, cada flecha representa la relación. La dirección de la flecha representa la relación de la predicción, Por ejemplo, la flecha de “Dos y DDos” a “Tipos Abstractos de Datos” indica que quienes consultan “Dos y DDos” podrían consultar también “Tipos Abstractos de Datos” [11].

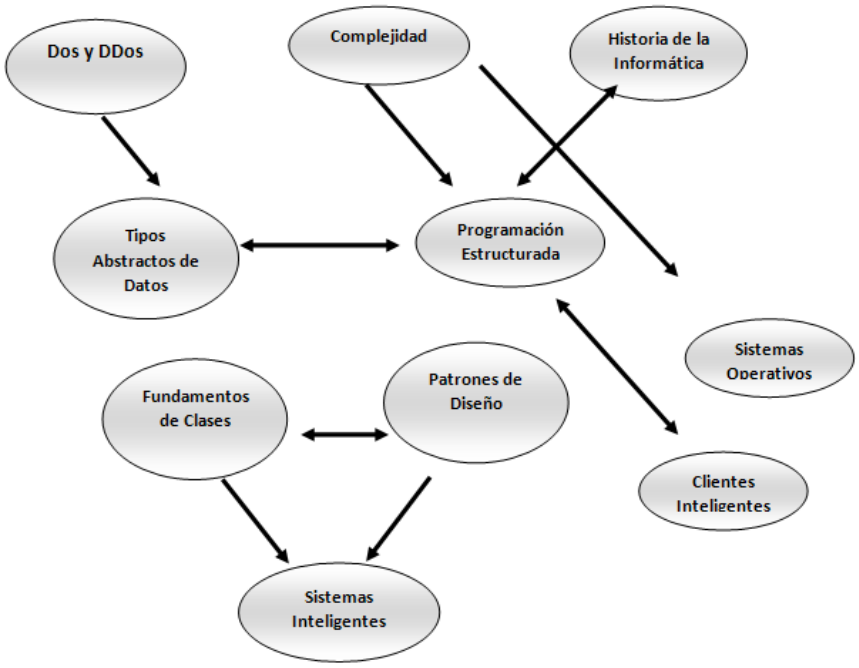


Figura 3. Asociación de Recursos de Aprendizaje. (Adoptada de [11])

1.3.3.5 Clustering

La tarea de Clustering también es llamada Segmentación. Esta tarea también es un tipo de clasificación, pero a diferencia de las otras, los grupos de asignación no se predeterminan si no que son generados e identificados por el algoritmo de minería de datos, al cual le corresponde examinar los objetos o registros y agruparlos de tal forma que se crean conjuntos únicos y homogéneos. Una vez el modelo de clustering ha sido entrenado, puede ser usado para clasificar los nuevos registros de acuerdo a sus atributos, características o comportamientos. Esto ayuda a menudo a agrupar clientes basados en sus patrones de compra y demografía, y posteriormente entonces ejecutar los modelos de predicción separadamente en cada grupo [5]

1.3.3.6 Descripción y Perfilado

Esta tarea de negocio se caracteriza porque se aplican varias técnicas de minería de datos para obtener un mejor entendimiento de las complejidades de los datos. Es decir, tiene como objetivo buscar explicaciones para incrementar el entendimiento de determinados comportamientos que han sido encontrados. La descripción y perfilado también pueden ser usados como una extensión a las tareas de perfilado de datos que se realizan en las etapas de ETL, se puede usar la minería de datos para identificar anomalías de errores de datos específicos y patrones más amplios de problemas de datos que no serían obvios a simple vista [5].

1.3.3.7 Detección de Anomalía

Esta tarea de negocio se caracteriza porque busca a partir del uso de varias técnicas de minería de datos identificar registros que se desvían de la normalidad significativamente. La detección de la anomalía involucra algunas iteraciones extras en el proceso de minería de datos. Para lograr el éxito en los algoritmos usados se debe establecer el entrenamiento a favor de los eventos excepcionales [5].

CAPITULO II:

2. DESCRIPCIÓN PROCESO DE DESARROLLO DEL DSS

2.1 Metodología para el desarrollo DW/BI

En la actualidad existen varias metodologías para el desarrollo de DW/BI, una de las más conocidas es la propuesta por Ralph Kimball, metodología ampliamente aplicada en este ámbito, y con bases conceptuales bien definidas.

Para el desarrollo del DW/BI se tomó como base el ciclo de vida dimensional propuesto por Ralph Kimball [8], metodología que ilustra las diferentes etapas por las que debe pasar todo proceso de construcción de un DW/BI. En la Figura 4, puede observarse el ciclo de vida dimensional con sus diferentes caminos y etapas requeridas para el diseño, desarrollo e implementación del DW/BI, en esta metodología sobresalen tres rutas, la ruta superior que se centra en la tecnología, la ruta central que se centra en los datos y la ruta inferior que se centra en las aplicaciones de usuario.

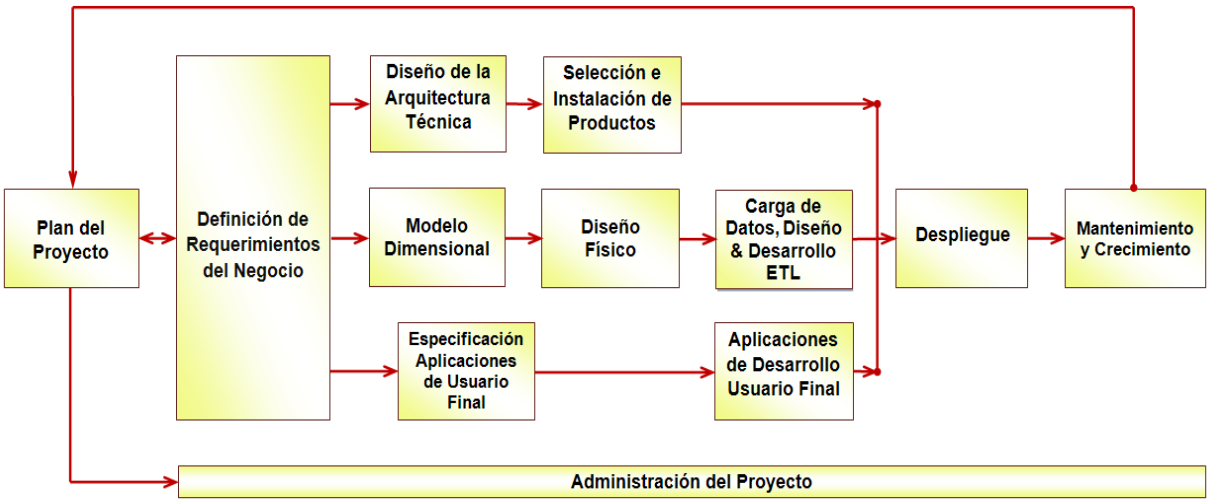


Figura 4: Ciclo de Vida Dimensional (Adoptado de [8])

El ciclo de vida dimensional se caracteriza por ser cíclico, concentrarse en la identificación de los requerimientos del negocio y por hacer desarrollos basados en incrementos que dan soporte a procesos comerciales específicos que tienen alto impacto analítico en el negocio.

A continuación se presentan cada una de las etapas haciendo primero una breve descripción de su objetivo general y luego mostrando los respectivos resultados obtenidos en cada una de ellas.

2.1.1 PLANEACIÓN DEL PROYECTO:

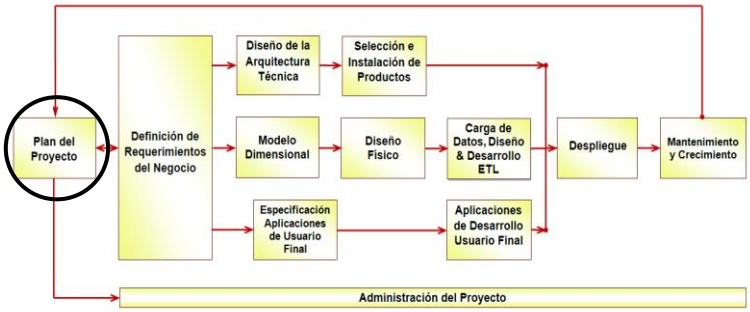


Figura 5: Ciclo de vida dimensional, etapa de planeación del proyecto (Adoptada de [8]))

La planeación busca identificar el escenario del proyecto para saber de dónde surge la necesidad del DW y el alcance del proyecto del DW/BI, incluyendo el impacto y evaluaciones de factibilidad. La planificación del proyecto es dependiente de los requerimientos del negocio porque estos determinan el alcance del proyecto y definen los recursos necesarios. Para la realización de esta tarea de alto nivel, se especifica las actividades y tareas necesarias, se define y se determina el alcance y la justificación del proyecto [8].

2.1.1.1 Definición del proyecto

Esta etapa se centra en el entendimiento del negocio, para saber de donde surge la necesidad del DW/BI, es así como el equipo de desarrollo del proyecto SPAR, considera importante que el repositorio además de brindar el acceso a los recursos educativos que se publiquen en él, permita conocer información que ayude a mejorar la calidad de los servicios y recursos educativos ofrecidos por el repositorio, contribuyendo a satisfacer en gran medida las necesidades de sus usuarios. Por lo tanto, se propuso a los administrativos de SPAR desarrollar un proyecto de inteligencia de negocios que permita obtener la información necesaria para tomar decisiones estratégicas que brindará soporte a estas nuevas necesidades del repositorio.

Seguidamente se identificó que para desarrollar un proyecto de estas características se contaba con el sistema transaccional que contiene los datos que dan soporte al análisis de información, con el apoyo suficiente dentro del equipo administrativo del repositorio SPAR y con los recursos técnicos necesarios para el desarrollo de dicho proyecto. Por consiguiente el primer paso por parte del equipo de desarrollo fue conocer el repositorio tanto a nivel técnico como funcional y el contexto educativo sobre el cual está enmarcado, para poder identificar unos requerimientos generales y un alcance inicial del DW/BI.

Esta preparación o capacitación que realiza el equipo, involucra mayor entendimiento del negocio y del sistema de información, se necesita identificar quiénes son las personas más importantes para realizar una serie de entrevistas y reuniones: Directores y Responsables de la Organización, Administradores, Expertos del Sistema y usuarios. Por

medio de todo el proceso de preparación, entrevistas y reuniones se identifican los requerimientos que posteriormente se clasifican en temas o necesidades analíticas [8].

Con el análisis de cada tema analítico se identifican los procesos de negocio que dan soporte a cada uno de ellos, para posteriormente ser priorizados y dar comienzo con la construcción del sistema de DW/BI.

A continuación se describe la tarea de alto nivel de recolección de requerimientos que permite por medio de los resultados obtenidos en cada una de sus etapas, establecer y detallar el plan y alcance del proyecto.

2.1.2 DEFINICIÓN DE REQUERIMIENTOS DEL NEGOCIO:

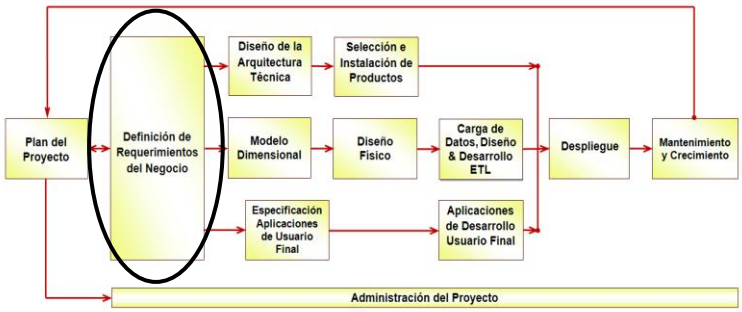


Figura 6: Ciclo de vida dimensional, etapa de Definición de Requerimientos del Negocio (Adoptada de [8])

Los requerimientos del negocio son el centro del "Universo del DW", estos conducen el alcance del proyecto y determinan qué datos deben estar disponibles en el DW/BI, como deben ser organizados y qué tan constante deben ser actualizados [8]. Esta etapa se centra en la recolección y análisis de requerimientos, los cuales son clasificados en necesidades analíticas (conocidas también como temas analíticos [8]) para la posterior priorización de las áreas de implementación, lo cual determina el alcance del proyecto. Este paso ayuda a identificar y a entender las necesidades analíticas del negocio e involucra entender los problemas técnicos, volúmenes de datos, limpieza y movimiento de datos, acceso de usuarios, entre otros. Los requerimientos del negocio guían al equipo de desarrollo para tomar las mejores opciones de estrategia, diseño y arquitectura técnica y establecen la base para continuar el ciclo de desarrollo sobre las tres rutas paralelas: Tecnología, Datos y Aplicaciones [8].

Para identificar los requerimientos del Repositorio de Acceso Público Basado en SCORM (SPAR 1.0) se siguieron algunos pasos como: Conocimiento detallado del contexto del negocio y del sistema de información, entrevistas y reuniones con los administradores, responsables y desarrolladores de SPAR 1.0.

Durante las reuniones se identificaron varias necesidades analíticas que fueron agrupadas en los llamados "Temas Analíticos" que posteriormente se clasifican en determinados procesos de negocio que son identificados y que darán inicio al proceso de

implementación del DW/BI de acuerdo a su priorización y nivel de impacto en el negocio. Para lograr obtener los requerimientos del negocio, se realizaron los siguientes pasos:

PASO 1. Se identificó el personal administrativo y técnico del repositorio SPAR 1.0 y equipo de desarrollo del sistema del DW/BI. La Tabla 1 muestra el personal con sus respectivos roles.

Participantes	**Propósitos/Roles en el modelado de procesos**
Ejecutivos del Negocio	Msc. Carlos Cobos Msc. Erwin Meza
Comité de Dirección	Msc. Martha Mendoza
Modelador de Datos	Alexander Calvache y Diego Bayona
Analista de Negocios	Alexander Calvache y Diego Bayona
Auxiliar de Datos	Alexander Calvache y Diego Bayona
Desarrollador del Sistema Fuente	Ingenieros Iván Giraldo y Jesús Muñoz
DBA	Ingenieros Iván Giraldo y Jesús Muñoz Msc. Carlos Cobos
Diseñador ETL	Alexander Calvache y Diego Bayona
Desarrollador ETL	Alexander Calvache y Diego Bayona
Diseñador Herramienta OLAP	Alexander Calvache y Diego Bayona
Desarrollador Herramienta OLAP	Alexander Calvache y Diego Bayona
Diseñador Herramientas de minería de datos	Alexander Calvache y Diego Bayona
Desarrollador Herramientas de minería de datos	Alexander Calvache y Diego Bayona

Tabla 1: Personal administrativo, Técnico y de Desarrollo con sus respectivos roles.

PASO 2: Después de haber identificado roles y participantes del negocio (Repositorio SPAR 1.0), se formularón una serie de preguntas y se construyeron cuestionarios de entrevistas que posteriormente fueron aplicadas al personal administrativo y técnico de SPAR 1.0. Los formularios de entrevistas pueden encontrarse en el **Anexo 1.**

PASO 3: Se hizo un resumen de los resultados obtenidos con las entrevistas que permiten agrupar requerimientos similares en temas analíticos comunes. Seguidamente se identificarón los procesos de negocio que dan soporte a cada uno de ellos. La Tabla 2 muestra temas analíticos y el proceso de negocio asociado a ellos, en este caso el proceso de negocio es: Gestión, Oferta y Demanda de Contenidos. Los demás tablas de temas analíticos y procesos de negocio asociados se encuentran en el **Anexo 2.**

Tema Analítico	Solicitud de Análisis	Proceso de Negocio que lo soporta
Planeación de Ofertas y Demandas de Contenidos	• Análisis de objetos publicados • Análisis de objetos consultados • Análisis históricos de publicaciones • Análisis históricos de consultas • Previsión de Consultas • Previsión de Publicaciones	Gestión, Oferta y Demanda de Contenidos
Desempeño de Ofertas y Demandas	• Publicaciones por lugar geográfico • Publicaciones por usuarios • Demandas por Lugar Geográfico • Demandas por usuario	Gestión, Oferta y Demanda de Contenidos
Reportes de Consultas y Publicaciones	• Reportes para publicadores. • Reportes para los usuarios de consultas. • Reportes por lugares geográficos • Reportes Históricos	Gestión, Oferta y Demanda de Contenidos
Ofertas especiales de contenidos	• Contenidos con mas ranking basados en consultas y descargas	Gestión, Oferta y Demanda de Contenidos
Listado de Contendidos	• Listado de Contenidos no vistos y/o no descargados • Identificar y promocionar los contenidos • Listados de Contenidos educativos Inactivos	Gestión, Oferta y Demanda de Contenidos

Tabla 2: Temas Analíticos y Proceso de Negocio asociado

PASO 4: El siguiente paso consiste en la construcción de la **Matriz Bus,** que es un mapa de datos de la organización, donde las filas de la matriz identifican los procesos de negocio de la organización y las columnas representan las entidades u objetos que participan en estos procesos (que posteriormente se convierten en dimensiones) [5]. La Figura 7 muestra la Matriz Bus con los procesos de negocio que son relevantes para SPAR y sus entidades u objetos relacionados.

	Fecha	Tiempo	Calendario Nacional	Objetos de Aprendizaje	Usuario	Demografía	Preguntas	Respuestas	Encuesta	Tipo Transacción	Taxonomía	País IP	Sesión	Pagina
Evaluación de Contenidos	*	*	*	*	*	*	*	*	*	*	*	*		
Oferta, Demanda Contenidos	*	*	*	*	*	*				*	*	*		
Sesiones de Usuario	*	*	*	*	*	*				*		*	*	*
Usabilidad Repositorio	*	*			*									*
Búsqueda de Contenidos	*	*		*	*						*		*	*

Figura 7: Matriz Bus de Procesos de Negocio de SPAR 1.0

Los procesos de negocio identificados que dan soporte a los temas analíticos son: Evaluación de Contenidos, Oferta y Demanda de Contenidos, Sesiones de Usuario, Usabilidad del Repositorio y Búsquedas de Contenidos.

PASO 5: A continuación se hizo una priorización de los procesos de negocio basada en su impacto y viabilidad. Para llevar a cabo la priorización se realizaron reuniones conjuntas del equipo de desarrollo del sistema DW/BI y el equipo administrativo y técnico de SPAR 1.0, de esta manera se pudo establecer si existía un correcto entendimiento de los requerimientos y de los procesos de negocio encontrados, para posteriormente hacer su priorización. La Figura 8 representa los resultados obtenidos en la priorización de procesos de negocio.

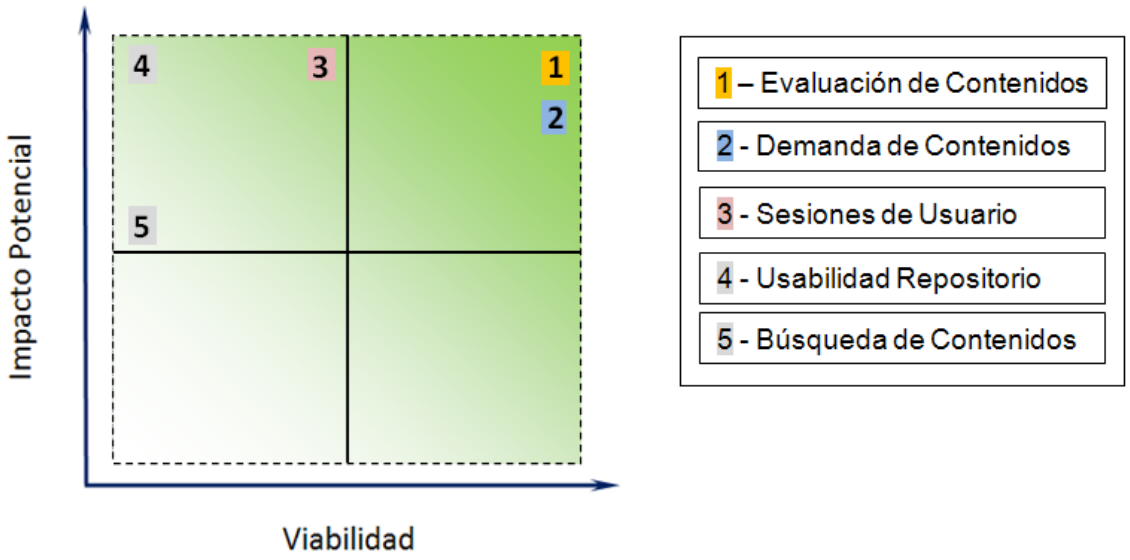


Figura 8: Gráfico de Impacto Vs Viabilidad para los procesos de negocio de SPAR 1.0

A continuación se hace una descripción de cada uno de los procesos de negocio con respecto a su impacto y viabilidad:

Proceso de Negocio 1 (Evaluación de contenidos): Tiene alta viabilidad porque el repositorio digital SPAR 1.0 permite evaluar los recursos de aprendizaje ofrecidos a los usuarios de tal manera que el sistema fuente dispone de esta información. Tiene alto impacto porque permite determinar estrategias para mejorar los recursos en cuanto a calidad de contenidos, presentación y nivel de satisfacción de los usuarios, permitiendo consolidar la calidad del repositorio y sus recursos educativos.

Proceso de Negocio 2 (Oferta, Demanda y Gestión de contenidos): Tiene alta viabilidad porque el repositorio digital SPAR 1.0 almacena completa información sobre los diferentes recursos publicados, consultados y modificados. Tiene alto impacto por que le permite al repositorio establecer estrategias para impulsar las consultas y publicaciones de recursos educativos.

Proceso de Negocio 3 (Sesiones de usuario): Tiene viabilidad media porque el repositorio digital SPAR 1.0 no almacena completa información con respecto a las sesiones de cada usuario. Tiene alto Impacto porque es relevante y favorable para el repositorio obtener información sobre el comportamiento de los usuarios en las sesiones realizadas.

Proceso de Negocio 4: (Usabilidad del Repositorio) Tiene viabilidad baja porque el sistema fuente del repositorio digital SPAR 1.0 no almacena información que permita determinar facilidad de uso del repositorio, facilidad de navegación, facilidad de encontrar objetos, entre otros. Tiene alto Impacto porque es determinante para el repositorio que los usuarios pueden hacer una navegación intuitiva, fácil, entendible y sencilla.

Proceso de Negocio 5: (Búsqueda de Contenidos) Tiene viabilidad baja porque en el sistema fuente del repositorio digital SPAR 1.0 no se almacena información referente a las búsquedas realizadas por los usuarios. Tiene impacto medio por que actualmente no es una necesidad para el repositorio construir un buscador de recursos educativos más preciso.

Luego de obtener los procesos priorizados, se empieza por abordar y entender en más detalle los procesos con mayores prioridades en los cuales se va a trabajar. Kimball dice al respecto: “Los procesos de negocio son las unidades coherentes de trabajo para el sistema de DW/BI, por tanto los de más alta prioridad se convierten en el enfoque inicial del proyecto”[8].

Se aplica nuevamente reuniones, entrevistas y documentación con más detalle para lograr un mayor entendimiento de los procesos de negocio más relevantes. Para este proyecto se decidió que se va a trabajar sobre los tres primeros procesos de negocio de la matriz bus, los cuales son: Evaluación de contenidos; Gestión, Oferta y Demanda de Contenidos y Sesiones de Usuario, los cuales tienen el más alto impacto y viabilidad como lo muestra la Figura 8.

PASO 6: Después de haber obtenido mayor entendimiento de los procesos de negocio que darán origen el desarrollo del sistema de DW/BI, se puede establecer un alcance del proyecto. El alcance del sistema de DW/BI para el repositorio digital SPAR 1.0 queda enmarcado en la construcción de tres Data Mart, que son:

Evaluación de Contenidos: El objetivo de este Data Mart es producir información relevante que permita satisfacer necesidades de consulta analíticas en cuanto a los niveles de satisfacción de los usuarios con respecto a la calidad de contenidos, calidad de presentación y satisfacción en general ofrecida por los recursos educativos publicados en el repositorio digital. Este Data Mart debe permitir análisis de información desde múltiples perspectivas dimensionales, que involucran obtener y cruzar atributos de los usuarios, de los recursos de aprendizaje, de las fechas y tiempos del día, de las áreas temáticas, preguntas y respuestas. Esta Data Mart debe permitir responder a inquietudes como: ¿Cuáles son los objetos mejor calificados?, ¿Cuáles objetos tiene baja calificación en su contenido, presentación o satisfacción en general producida a los usuarios?, ¿Cuál ha sido el promedio de calificación que han tenido los recursos a través del tiempo? ¿Cómo ha sido la aceptación de los recursos en determinados usuarios y países?, ente otros.

Gestión, Oferta y Demanda de Contenidos: La implementación de este Data Mart tiene como objetivo producir información analítica con respecto a la cantidad de consultas, publicaciones y modificaciones hechas sobre los recursos educativos. Este Data Mart debe producir diferentes perspectivas dimensionales de la información permitiendo hacer consultas que cruzan atributos de los usuarios, de los recursos de aprendizaje, de los tipos de acceso a los recursos, de las fechas y tiempos del día, de las áreas temáticas, entre otros. Este Data Mart debe permitir responder a inquietudes como: ¿Cuál es la cantidad de transacciones de los usuarios, cuales son los países que más consultan?,

¿Cuáles son los recursos de aprendizaje más consultados?, ¿Cuáles son las clasificaciones temáticas más buscadas?, ¿Cuáles son los horarios y fechas de consulta más habituales?, ¿Cuáles son los usuarios que mas hacen uso del repositorio?, ¿Cuáles son sus intereses y qué tipo de usuarios son (administrador, LMS, anónimo, registrado)?, ¿Cuáles son los recursos poco consultados?, entre muchas otras.

Sesiones de Usuario: El objetivo de esta Data Mart es producir información analítica relacionada con el comportamiento que tiene los usuarios en las sesiones realizadas. Igualmente este Data Mart debe permitir cruce dimensional entre los atributos de usuario, pagina, fecha, sesión, localización, país, entre otras. Permitiendo responder a inquietudes con respecto al tiempo promedio de duración de las sesiones, éxito de las sesiones con respecto a si se realizo algún tipo de transacción (consulta, descarga, publicación, modificación de metadata, etc.), cantidad de páginas consultadas, formas de inicio y de finalización de sesión en el repositorio SPAR, entre otras.

PASO 7: Después de establecer el alcance del proyecto se define un plan detallado que permita conducir el desarrollo del sistema de DW/BI. El plan del proyecto donde se describen tiempos, actores, responsabilidades, puede ser encontrado en el **Anexo 3.**

2.1.3 MODELADO DIMENSIONAL:

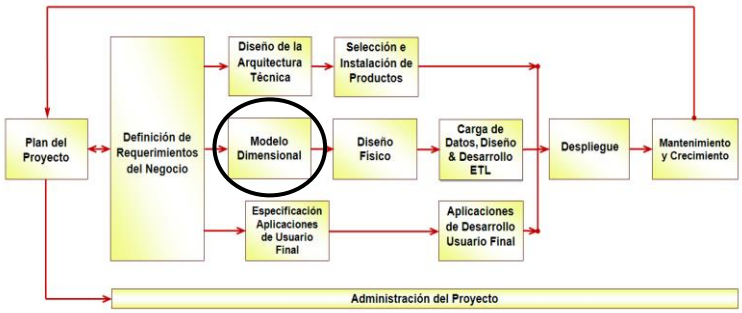


Figura 9: Ciclo de vida dimensional, etapa de Modelado Dimensional. (Adaptada de [8])

En esta etapa se definen los modelos dimensionales que dan soporte a los requerimientos analíticos identificados en la etapa de recolección de requerimientos. Este proceso comienza tomando como base la matriz bus que representa los procesos claves del negocio y su dimensionalidad. Posteriormente se comienza a desarrollar los modelos dimensionales de alto nivel para cada Data Mart, con base en un análisis detallado de los sistemas fuentes y en el entendimiento de los requerimientos del negocio adquirido con anterioridad. El proceso de desarrollo de modelos dimensionales de alto nivel contiene los siguientes pasos: Primero: Se selecciona el Data Mart; Segundo: se define la granularidad; Tercero: Se eligen las dimensiones; Cuarto: Se hace la elección de los hechos, medidas o indicadores. Por último se hace una validación y actualización de los modelos dando origen a los modelos dimensionales finales [8]. El proceso en detalle se ilustra en la Figura 10.

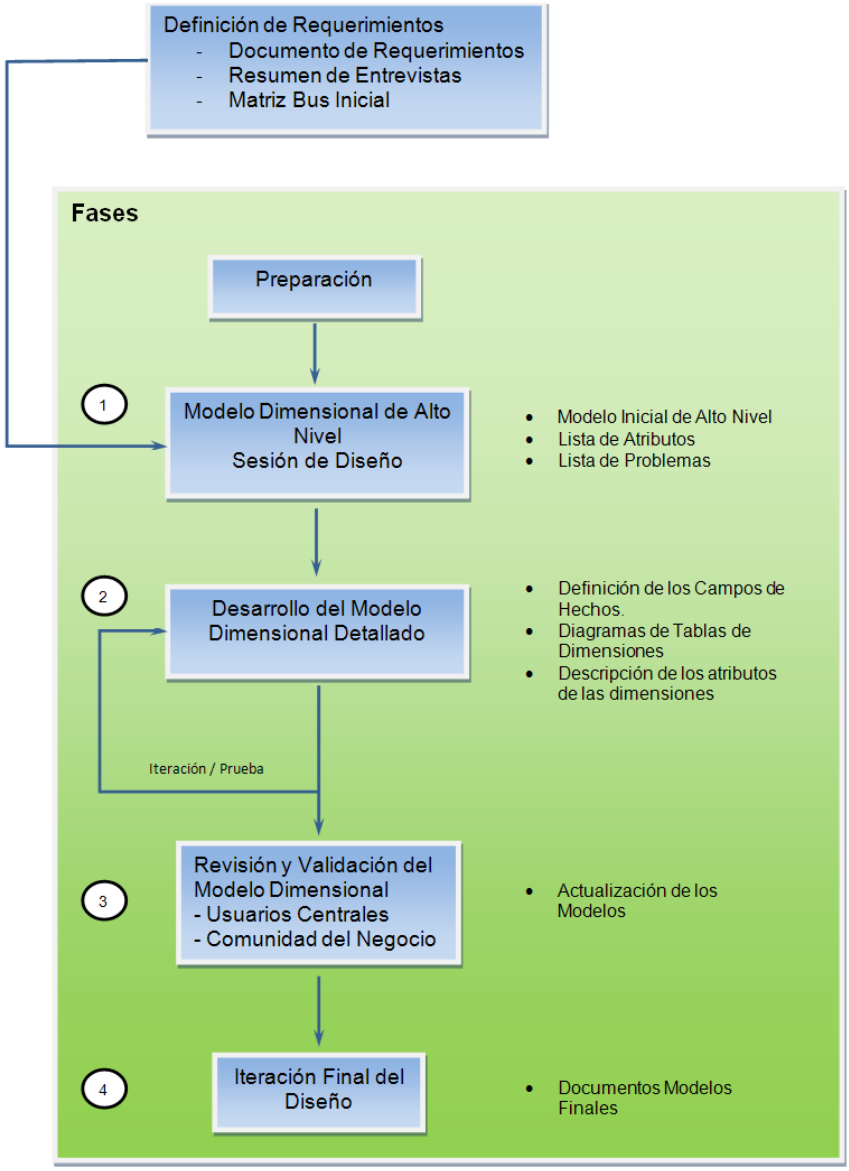


Figura 10: Diagrama de flujo del proceso de modelado dimensional (Adoptado de [5])

El proceso de modelado comienza con el desarrollo del modelo inicial de alto nivel y una lista de atributos, seguido de una fase de desarrollo de modelado dimensional detallado donde se obtienen los campos para la tabla de hechos, la descripción de los atributos y el diagrama de las tablas de dimensiones, luego se continua con revisiones y validaciones del modelo dimensional hasta llegar al modelo dimensional definitivo.

2.1.3.1 Modelado Inicial de Alto Nivel:

Para crear los modelos iníciales de SPARDW se siguieron cuatro pasos [8]:

1. Escoger el Proceso de Negocio.

2. Definir la granularidad: Definir el nivel de detalle o granularidad para el proceso de negocio seleccionado, generalmente es una granularidad atómica que consiste de una fila en la tabla de hechos por una fila en el sistema fuente transaccional (evento).

3. Escoger las Dimensiones: Escoger las dimensiones con base en los objetos asociados a cada proceso de negocio de la matriz bus inicial, escoger las dimensiones puede involucrar redefinir la granularidad.

4. Identificar los Hechos o medidas: identificar hechos o medidas de desempeño generados por el proceso de negocio. Los hechos usualmente se enlazan directamente a la declaración de la granularidad.

Para comenzar con el proceso nos remitimos a la matriz bus construida en la etapa de recolección de requerimientos (Figura 7) y comenzamos la construcción con cada uno de los tres procesos de negocios más relevantes.

1. **Proceso de Negocio o Data Mart:**
   Evaluación de Objetos de Aprendizaje

   Alcance: El objetivo de este Data Mart es producir información relevante que permita satisfacer necesidades de consulta analíticas en cuanto a los niveles de satisfacción de los usuarios con respecto a la calidad de contenidos, calidad de presentación y satisfacción en general ofrecida por los recursos educativos publicados en el repositorio digital. Este Data Mart debe permite análisis de información desde múltiples perspectivas dimensionales, que involucran obtener y cruzar atributos de los usuarios, de los recursos de aprendizaje, de las fechas y tiempos del día, de las áreas temáticas, preguntas y respuestas. Esta Data Mart debe permitir responder a inquietudes como: ¿Cuáles son los objetos mejor calificados?, ¿Cuáles objetos tiene baja calificación en su contenido, presentación o satisfacción en general producida a los usuarios?, ¿Cuál ha sido el promedio de calificación que han tenido los recursos a través del tiempo? ¿Cómo ha sido la aceptación de los recursos en determinados usuarios y países?, ente otros.

2. **Granularidad:**
   Cada registro de la tabla de hechos representa la calificación que el usuario hace a un recurso educativo, por cada pregunta respondida de la evaluación (tres preguntas).

3. **Dimensiones**

- **Dimensión Fecha** (Conformada)
- **Dimensión Objetos de Aprendizaje** (Conformada)
- **Dimensión Usuario** (Conformada)
- **Dimensión Tiempo del día** (Conformada)
- **Dimensión Preguntas**
- **Dimensión Respuestas**

- **Dimensión Calendario Nacional** (Conformada)
- **Dimensión País IP** (Conformada)

Subdimensiones
- **Subdimensión Localización**
- **Subdimensión Taxonomía**
- **Subdimensión Puente Taxonomía**

4. **Hechos:**
   En este caso se registra el evento de calificar un recurso educativo donde la medida es la calificación cuantitativa y cualitativa y el promedio de calificaciones cuantitativas por recurso**.** Se rastrea las repuestas seleccionadas por el usuario para determinar o categorizar los objetos de contenido de acuerdo a los valores de calificación.

   El modelo de alto nivel construido es mostrado en la Figura 11.

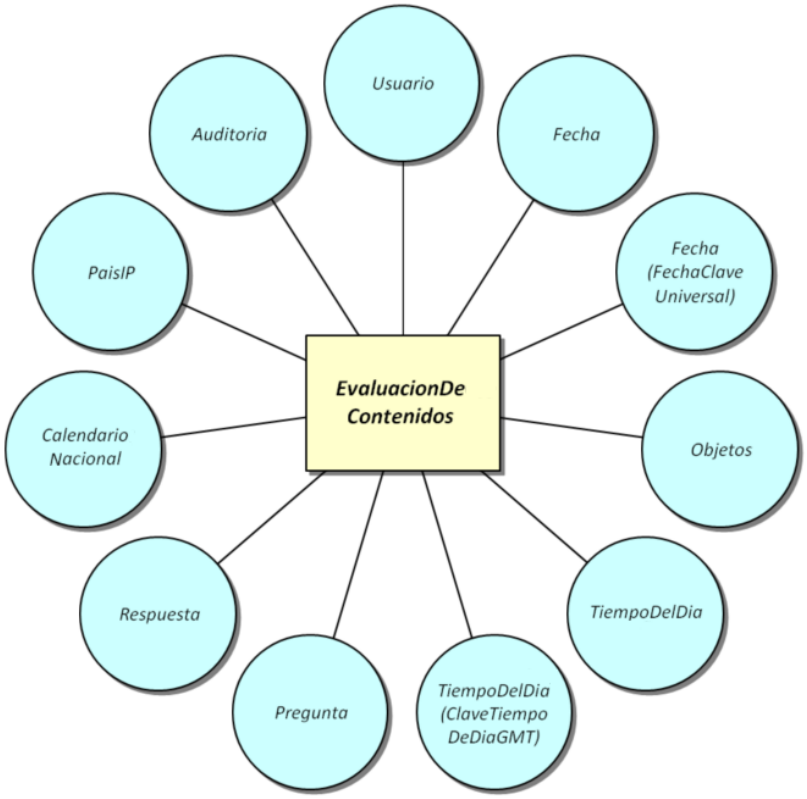


Figura 11: Modelo Inicial de Alto Nivel para el Proceso de Evaluación de contenidos

1. **Proceso de Negocio o Data Mart:**

   Gestión, Oferta, Demanda de Contenidos
   Alcance: La implementación de este Data Mart tiene como objetivo producir información analítica con respecto a la cantidad de consultas, publicaciones y modificaciones hechas sobre los recursos educativos. Este Data Mart debe

producir diferentes perspectivas dimensionales de la información permitiendo hacer consultas que cruzan atributos de los usuarios, de los recursos de aprendizaje, de los tipos de acceso a los recursos, de las fechas y tiempos del día, de las áreas temáticas, entre otros. Este Data Mart debe permitir responder a inquietudes como: ¿Cuales la cantidad de transacciones de los usuarios, cuales son las países que más consultan?, ¿Cuales son los recursos de aprendizaje más consultados?, ¿Cuáles son las clasificaciones temáticas más buscadas?, Cuales son los horarios y fechas de consulta más habituales?, ¿Cuáles son los usuarios que más hacen uso del repositorio, cuáles son sus intereses y qué tipo de usuarios son (administrador, LMS, anónimo, registrado)?, ¿Cuáles son los recursos poco consultados?, entre muchas otras.

2. **Granularidad**
   Tabla de hechos Factless (Sin medidas) en la que cada registro representa una transacción (publicación, consulta, descarga, cambio de recurso o edición de metadata) que el usuario realiza en el repositorio.

3. **Dimensiones:** Este modelo incluye las siguientes dimensiones:

- **Dimensión Fecha** (Conformada)
- **Dimensión Objetos de Aprendizaje** (Conformada)
- **Dimensión usuario** (Conformada)
- **Dimensión Tiempo del día** (Conformada)
- **Dimensión Tipo de Transacción**
- **Dimensión Calendario Nacional** (Conformada)
- **Dimensión País IP** (Conformada)

Subdimensiones
- **Subdimensión Localización**
- **Subdimensión Taxonomía**
- **Subdimensión Puente Taxonomía**

4. **HECHOS:**
   En este caso se registran los eventos (Factless) que pueden ser consultas o publicaciones. Se hace seguimiento de la consulta, publicación o modificación que realiza un usuario sobre el repositorio. **No contiene hechos (medidas).**

El modelo de alto nivel construido es mostrado en la Figura 12.

Figura 12: Modelo Inicial de Alto Nivel para el Proceso de Gestión, Oferta y Demanda de Contenidos

1. **Proceso de Negocio o Data Mart: Sesiones de Usuario**
   Alcance: El objetivo de esta Data Mart es producir información analítica relacionada con el comportamiento que tienen los usuarios en las sesiones realizadas. Igualmente este Data Mart debe permitir cruce dimensional entre los atributos de usuario, pagina, fecha, sesión, localización, país, entre otras. Permitiendo responder a inquietudes con respecto al tiempo promedio de duración de las sesiones, éxito de las sesiones con respecto a si se realizó algún tipo de transacción (consulta, descarga, publicación, modificación de metadata, etc.), cantidad de páginas consultadas, formas de inicio y de finalización de sesión en el repositorio SPAR, entre otras.

2. **Granularidad** Se inserta una fila por cada sesión de usuario completada.

3. **Dimensiones:**

- **Dimensión Fecha** (Conformada)
- **Dimensión Objetos de Aprendizaje** (Conformada)
- **Dimensión Usuario** (Conformada)
- **Dimensión Tiempo del día** (Conformada)
- **Dimensión Pagina**
- **Dimensión Sesión**

- **Calendario Nacional** (Conformada)
- **Dimensión País IP** (Conformada)

Subdimensiones:

- **Subdimensión Localización**

4. **Hechos:**
   Se mide la cantidad de páginas visitadas por sesión, el éxito o no éxito de sesión, determinado por la realización o no de transacciones, tiempo de duración de sesión, salida normal o salida abrupta de la sesión.

El modelo de alto nivel construido es mostrado en la Figura 13.

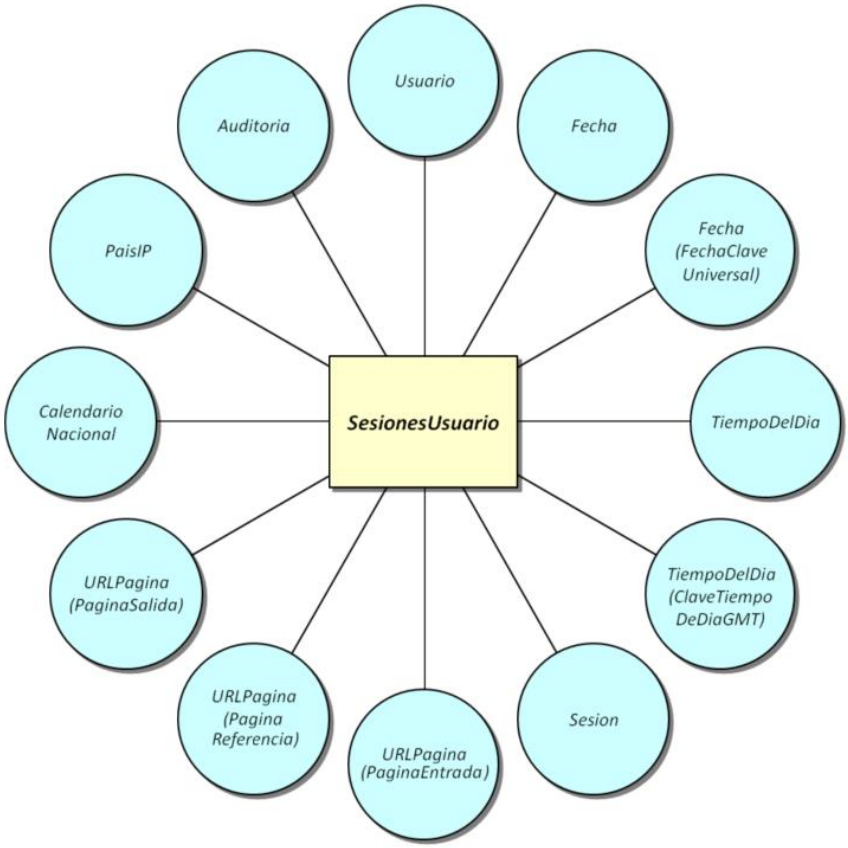


Figura 13: Modelo Inicial de Alto Nivel para el Proceso de Sesiones de usuario.

Este proceso de desarrollo finaliza con la identificación de una lista de atributos de todas las dimensiones y una lista de problemas identificados. Estos entregables se encuentran en el **Anexo 4.**

2.1.3.2 Desarrollo Detallado del Modelado Dimensional.

En esta fase se examina en detalle tabla por tabla, profundizando en las descripciones, las fuentes, relaciones, tipos de datos, tipos de atributos SCD (atributos que cambian

lentamente), problemas de calidad en los datos, transformaciones requeridas para poblar los modelos [5].

La descripción detallada de estos atributos de las dimensiones y de las tablas de hechos para cada proceso de negocio de SPAR se encuentra en el **Anexo 5.**

2.1.3.3 Revisión y Validación.

Involucra revisar los modelos dimensionales obtenidos en reuniones conjuntas: expertos técnicos, administrativos del negocio y equipo de desarrollo del sistema de DW/BI. De la misma manera se identifican algunas necesidades analíticas extras que deberían soportar los modelos, lo que implica modificar la estructura de los mismos [8]**.**

Algunas descripciones de las revisiones y modificaciones hechas de presentan a continuación:

En el modelo de gestión, oferta y demanda de contenidos se realizó una revisión sobre los atributos demográficos de los usuarios, se decidió que la dimensión usuario no debería contener los datos demográficos debido a que muchos usuarios comparten datos demográficos de país y departamento. Por lo tanto, se creó una subdimensión de datos demográficos llamada Localización relacionada con la Dimensión Usuario. Esto contribuye con el ahorro de espacio y obtiene ventajas analíticas al tener una tabla de datos demográficos separada para realizar cruces dimensionales.

En otra revisión se sugirió que el modelo permitiera almacenar las transacciones realizadas los días festivos en distintos países. Por tal razón, se decidió crear una dimensión de Calendario Nacional con información de los festivos para distintos países en el mundo, independientemente de la dimensión primaria de fecha que contiene atributos genéricos sobre la fecha, sin importar el país. Calendario Nacional se debe relacionar directamente con la tabla de hechos para saber si las transacciones se realizaron en días festivos y también se relaciona a la dimensión fecha con calendario nacional para saber qué fechas son festivos en distintos países.

Para la dimensión fecha también se agrega una funcionalidad multinacional en donde se relaciona la dimensión fecha con la tabla de hechos a través de dos relaciones, a esto se le llama roles de dimensión, esto con el fin de manejar fechas locales y fechas universales, dado que el repositorio es de acceso multinacional y la fecha en que ocurren las transacciones difiere entre países por las zonas horarias GMT.

Lo mismo sucede con la dimensión tiempo del día donde es necesario que se maneje dos roles, un rol para hora local y el otro rol para la hora universal GMT.

Otro cambio significativo identificado como resultado de las reuniones fue la identificación de un caso particular de relación entra tablas: se presenta cuando se busca relacionar la clasificación temática de los objetos con las tabla de hecho del modelo dimensional de Gestión, Oferta y Demanda de Contenidos y el modelo de Evaluación de Contenidos, se puede identificar que existe una relación muchos a muchos entre los objetos y sus

clasificaciones temáticas, esto se puede visualizar en el sistema transaccional donde existen tres tablas: la tabla taxonomía que tiene una relación reflexiva, la tabla objetos y la tabla puente entre objetos y taxonomía. Para este tipo de casos se han identificado dos alternativas en donde una de ellas es crear una tabla puente entre una dimensión y una tabla de hechos, la otra alternativa consiste en crear una tabla puente entre las dos dimensiones. La opción adoptada fue introducir una tabla intermediaria entre las dimensiones Objetos y Taxonomía. Se decidió adoptar esta opción porque este tipo de relación ya existe en el sistema fuente, y por que permite responder a determinadas preguntas de interés.

En la dimensión página se identificó la necesidad de manejar tres roles, uno para página de entrada, otro para página de referencia y otro para página de salida.

Finalmente, después de haber realizado las últimas revisiones y validaciones se construyen los modelos dimensionales finales. Cada modelo dimensional representa un proceso de negocio o Data Mart y la unión de estos representa el completo *Data Warehouse.*

A continuación se presentan los tres modelos dimensionales finales que representan el DW del repositorio digital SPAR 1.0. Estos tres modelos dimensionales dan origen a la construcción del diseño físico del DW y a su posterior proceso de extracción, transformación y carga.

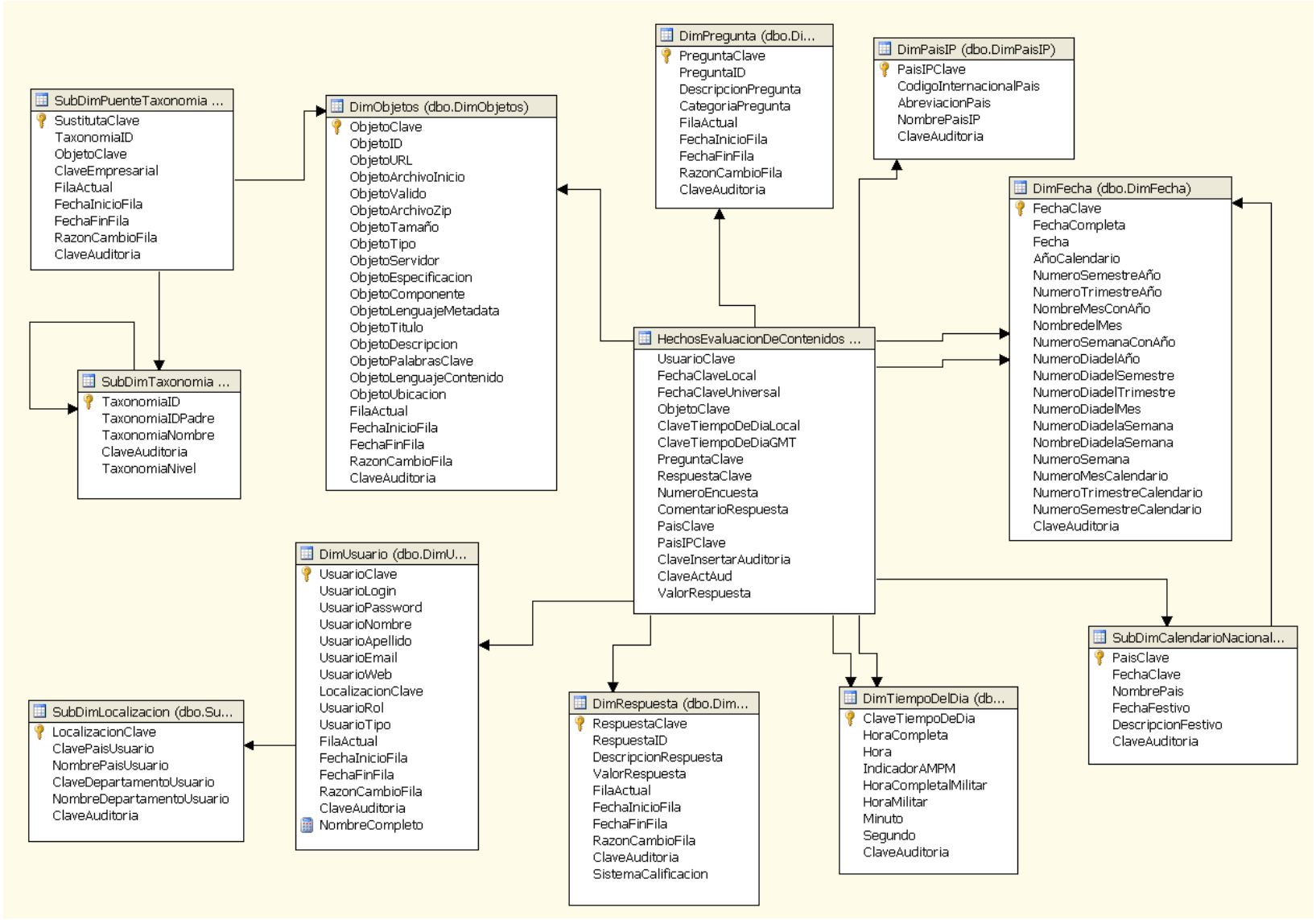


Figura 14: Modelo Dimensional de Evaluación de Contenidos.

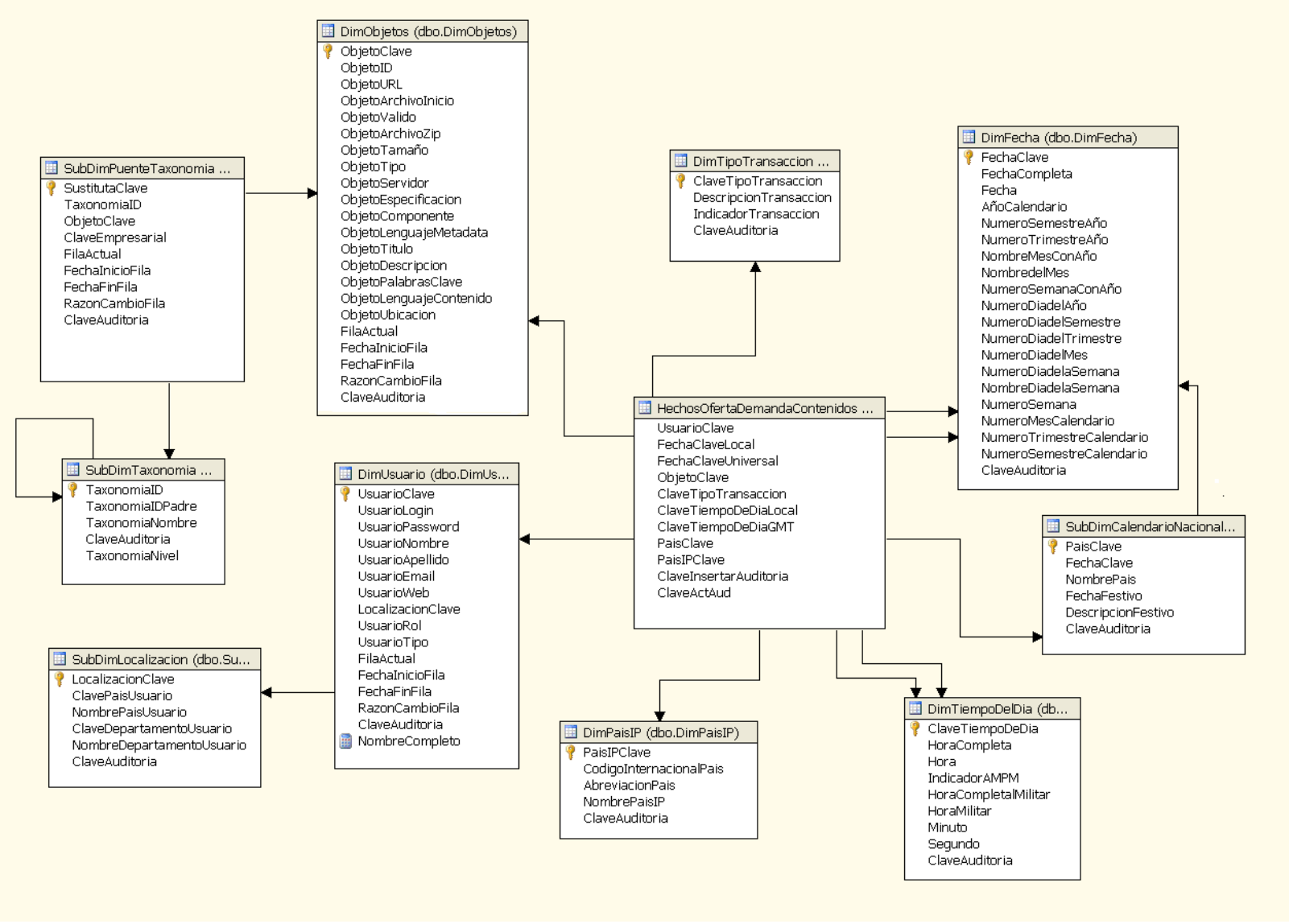


Figura 15: Modelo Dimensional de Gestión, Oferta y Demanda de Contenidos.

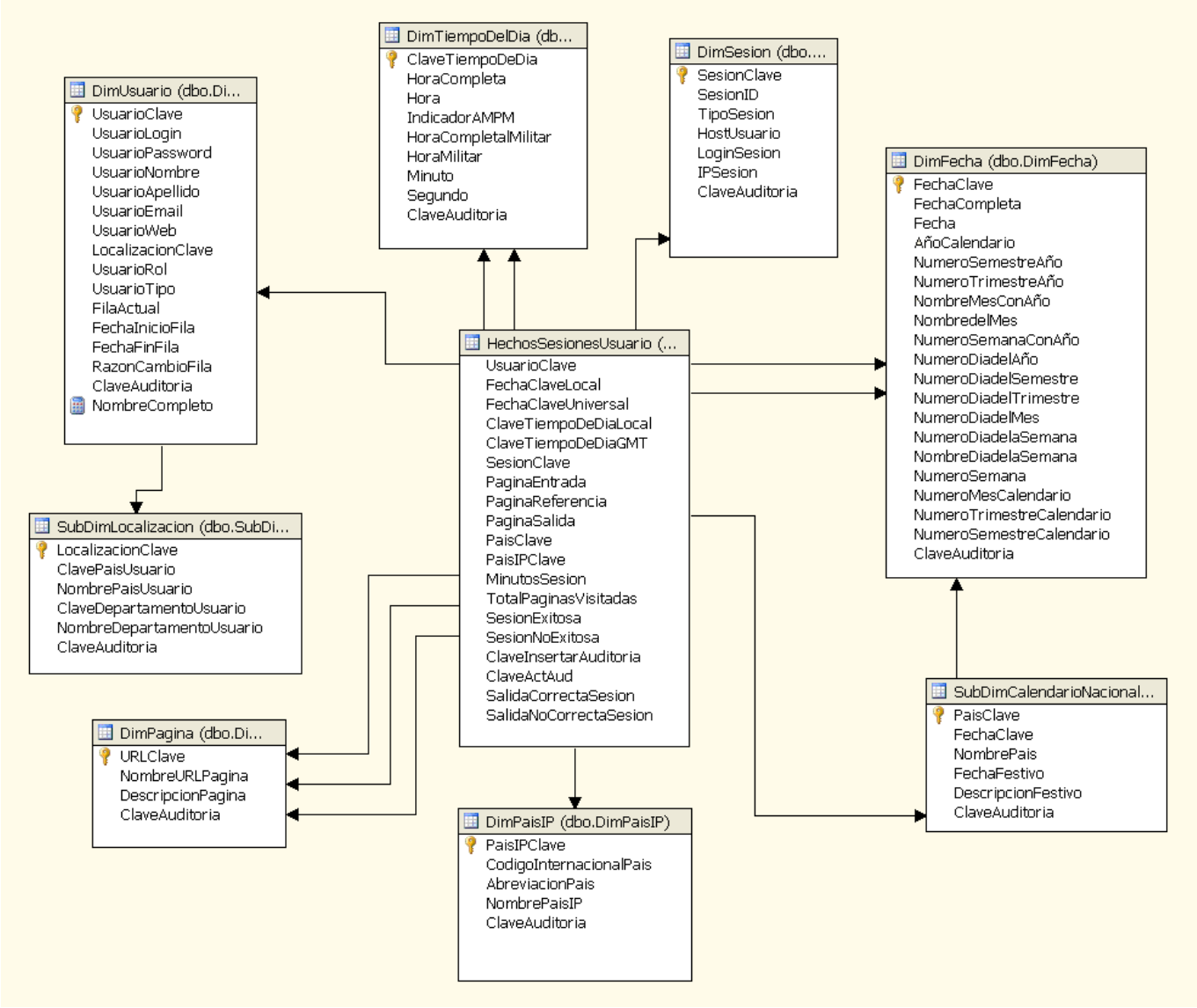


Figura 16: Modelo Dimensional de Sesiones de Usuario.

2.1.4 DISEÑO FÍSICO DEL DATA WAREHOUSE:

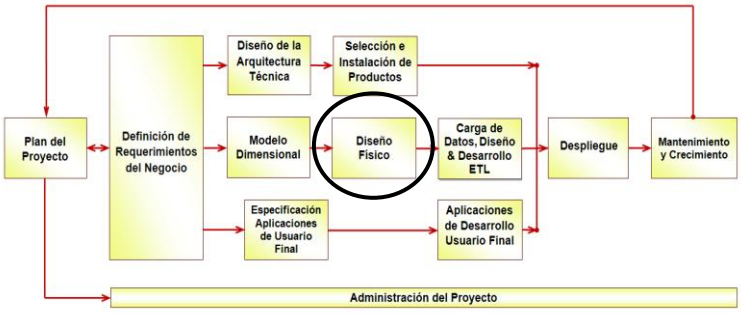


Figura 17: Ciclo de vida dimensional, etapa de Diseño físico. (Adoptada de [8])

La arquitectura de diseño físico con respecto al hardware para el sistema de DW/BI de SPAR 1.0, es una configuración conocida como **All-in-One** (Todo en Uno), que consiste en tener todos los componentes del servidor corriendo en una sola maquina [5]. Los componentes son: RDBMS, Integration Services, Reporting Services, Analysis Services con OLAP y Data Mining. La Figura 18 muestra este tipo de configuración.

Figura 18: Configuración All-in-One para un Sistema de Inteligencia de Negocios. (Adoptada de [5])

El diseño físico de SPARDW se realizó en SQL Server 2005, donde se crearon las tablas, columnas, llaves primarias, foráneas, sustitutas y demás atributos.

2.1.5 DISEÑO DEL SISTEMA ETL.

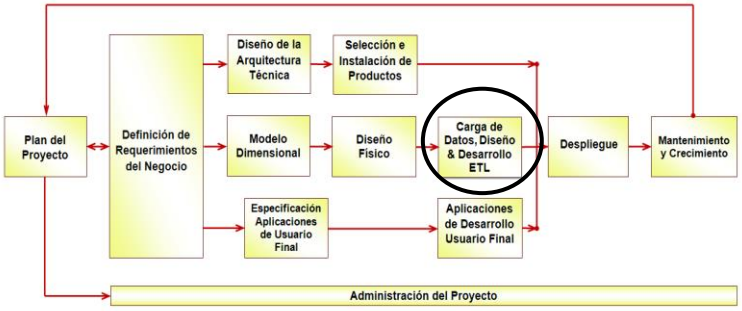


Figura 19: Ciclo de vida dimensional, etapa de ETL (Adoptada de [8]).

Las principales etapas de esta fase son: la Extracción, la Transformación y la Carga de datos. El proceso de extracción se utiliza para obtener los datos desde las diferentes fuentes de datos, el proceso de transformación se utiliza para la limpieza, conformación y conversión de los datos y el proceso de carga de datos es el que se encarga de poblar los datos en el DW [8]. Otro sistema que se diseña y se construye en esta etapa es el sistema de auditoría, que tiene como fin hacer un completo seguimiento de los datos que son cargados dentro del DW.

El Diseño del Sistema ETL, es uno de los procesos más importantes en la construcción de un sistema de DW/BI. El primer paso en el desarrollo del sistema ETL es empezar con los Diagramas de Mapas de Alto Nivel de ETL, para la construcción de estos diagramas debe identificarse con anterioridad cuáles son las fuentes de datos desde donde se obtendrán los mismos, cuáles serán las posibles transformaciones y limpieza de datos y finalmente el destino [5], que en nuestro caso es el DW relacional.

La principal fuente de datos identificada para el sistema de DW/BI de SPAR es el sistema de datos transaccional del repositorio SPAR 1.0, las otras fuentes de datos son archivos de Excel.

Después de identificar las fuentes de datos, se inicia el proceso de construcción de los diagramas Alto Nivel de ETL. Se construyen diagramas para cada Dimensión y cada Tabla de Hechos del DW relacional, para las cuales se identifican sus respectivos orígenes, transformaciones y destinos. Es necesario conocer la secuencia de carga de tablas, debido a la existencia de dependencias entre algunas dimensiones, y en el caso particular la tabla de hechos que no puede ser poblada antes de poblar sus tablas dimensiones asociadas. La Figura 20 muestra el diagrama de Alto Nivel de ETL de la dimensión Usuario, Localización y Taxonomía. Los otros diagramas Alto Nivel de ETL del sistema DW/BI de SPAR se presentan en el **Anexo 6.**

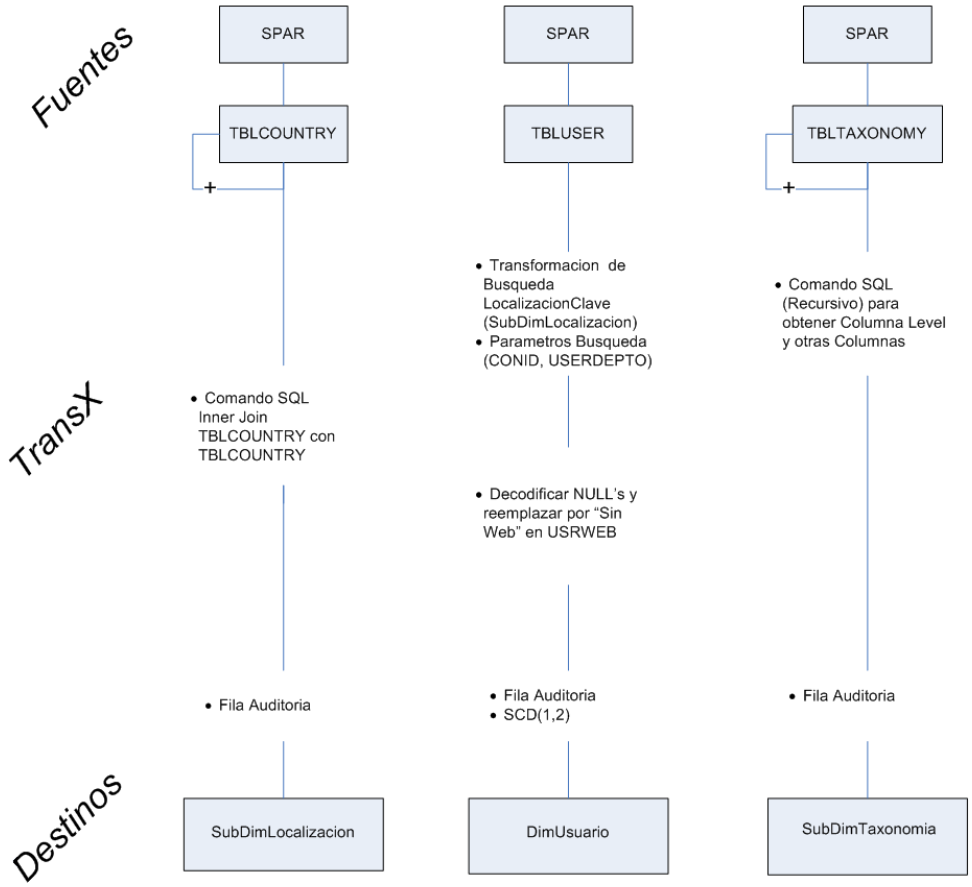


Figura 20: Diagrama de Alto Nivel de ETL de la dimensión Usuario, Localización y Taxonomía.

Para la construcción del sistema ETL de SPARDW se utilizó la herramienta Microsoft SQL Server 2005 Integration Services (SSIS), ésta es una plataforma para crear soluciones de integración de datos de alto rendimiento, incluyendo paquetes de extracción, transformación y carga (ETL) de datos. Integration Services incluye herramientas gráficas y asistentes para crear y depurar paquetes; tareas para realizar funciones de flujo de datos, como operaciones FTP, ejecución de instrucciones SQL y envío de mensajes de correo electrónico; orígenes y destinos de datos para extraer y cargar datos; transformaciones para borrar, agregar, mezclar y copiar datos, además de otras funciones que permiten una serie de cambios sobre los datos [5].

Para la implementación del proceso ETL se hace uso de los paquetes de Integration Services, los paquetes son archivos XML que tiene la extensión .dtsx. Un paquete contiene una o más tareas que permiten hacer manipulaciones sobre los datos.

Siguiendo recomendaciones de buenas técnicas de diseño y construcción hechas por Ralph Kimball [5], se creó un paquete hijo para la carga de cada tabla (Tabla Dimensión/Tabla de Hechos), y paquetes maestros que contienen tareas de ejecución de paquetes que permiten que la secuencia de carga de tablas sea adecuada. Estas técnicas de construcción de paquetes incrementan la flexibilidad y el entendimiento del sistema.

El proceso que se siguió para crear el sistema ETL, es el siguiente:

Comenzó con la creación de un proyecto de Integration Services, donde se crean una serie de paquetes de acuerdo a la dependencia de dimensiones, para cada paquete se establecen los administradores de conexión para datos de origen y destino, a continuación se agrega una tarea de flujo de datos en el paquete. La tarea de flujo de datos mueve datos entre orígenes y destinos, y proporciona la funcionalidad para transformar, limpiar y modificar los datos a medida que se mueven. En la tarea de flujo de datos se lleva a cabo la mayor parte del proceso de extracción, transformación y carga. Integration Services permite la manipulación de dimensiones SCD (Dimensiones de Variación Lenta), para hacer seguimiento de cambios en este tipo de dimensiones, para lo cual se crearon columnas de administración en las dimensiones indicadas. Estas columnas son adicionales para rastrear el rango de datos para los cuales la fila de la dimensión es válida, y se adicionan solamente a dimensiones que tiene atributos que cambian lentamente de tipo 2 [5]).

Para el desarrollo del sistema de ETL de SPARDW se crearon un total 19 paquetes hijos y 3 paquetes maestros. La Figura 21 muestra un paquete (DimObjetos.dtsx) y su flujo de tareas que hacen la extracción, transformación y carga de datos en la Dimensión Objetos. A continuación se explica el flujo de tareas de este paquete:

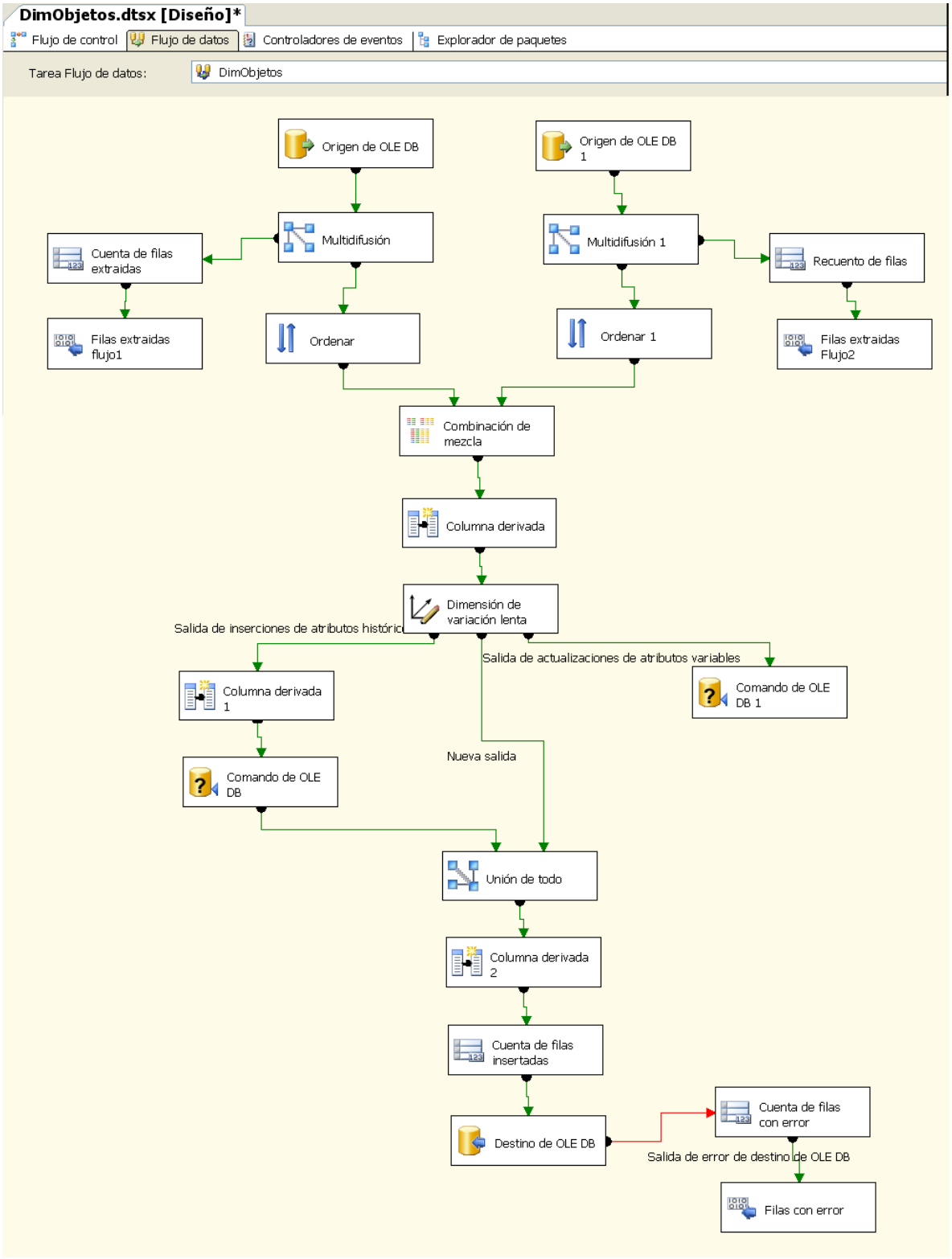


Figura 21: Paquete que extrae, transforma y carga datos en la Dimensión Objetos.

1. Este paquete comienza con la extracción de datos, para lo cual utiliza dos tareas de extracción, que extraen datos de la base de datos relacional del repositorio SPAR 1.0, la extracción de datos en este caso se hace por medio de sentencias SQL.

2. Estos datos pasan hacia transformaciones de tipo multidifusión que hacen una copia de los datos para mandarlos por flujos diferentes, por un lado el flujo continua hacia una transformación de conteo de filas con el fin de hacer seguimiento de la cantidad de filas extraídas y por el otro lado el flujo de datos continua para recibir las transformaciones necesarias.

3. Las transformaciones de ordenamiento, ordenan los datos para que pueden fluir hacia la transformación de combinación de mezcla.

4. La siguiente transformación llamada combinación de mezcla recibe dos flujos de datos que contienen diferentes atributos de información de los objetos de aprendizaje, por ejemplo por un flujo de datos viene datos con atributos como ObjetoID, ObjetoURL, etc., y por el otro flujo de datos viene datos con atributos como ObjetoID ,ObjetoLenguajeMetadata, ObjetoTitulo, etc., lo que hace la transformación de mezcla es combinar los datos por medio de un atributo común (ObjetoID) de los dos conjuntos de datos, creando filas únicas que contienen todos los atributos para cada objeto.

5. EL siguiente paso es llevar el flujo de salida desde la transformación de mezcla a una transformación de columna derivada para adicionar al flujo de datos el valor de una clave de auditoría.

6. El flujo continúa con una serie de transformaciones de dimensiones de variación lenta, la cual permite identificar si hay cambios en los datos de columnas a las cuales se les hace un seguimiento de cambios y poder tratar estos datos adecuadamente.

7. Antes de cargar los datos en el destino se hace un conteo de filas que se van a insertar, esto lo hace la transformación de conteo (cuenta de filas insertadas).

8. Por último el flujo entra a una transformación de destino que carga los datos en una tabla relacional (Dimensión Objetos) de la bodega de datos relacional.

El resto de paquetes con sus respectivos flujos y tareas pueden ser encontrados en el **Anexo 7.**

Otro aspecto importante dentro del proceso ETL es la construcción del sistema de auditoría, usado para el desarrollo de este proyecto es basado en el sistema propuesto por Ralph Kimball [5]. Este sistema se compone de tres tablas, con las cuales se busca hacer un seguimiento de los paquetes y los procesos que cargan los datos en las dimensiones y tablas de hechos. Estas tablas brindan información como: cantidad de filas extraídas, insertadas y con error, entre otros [5]. La dimensión de auditoría en SPARDW tiene dos roles, el primero se utiliza para llevar el registro de las filas extraídas y el segundo rol para llevar el registro de las filas actualizadas. Todas las dimensiones y tablas de hechos del

DW de SPAR 1.0 tienen una relación con la dimensión auditoria. La Figura 22 muestra el sistema de auditoría desarrollado para SPARDW.

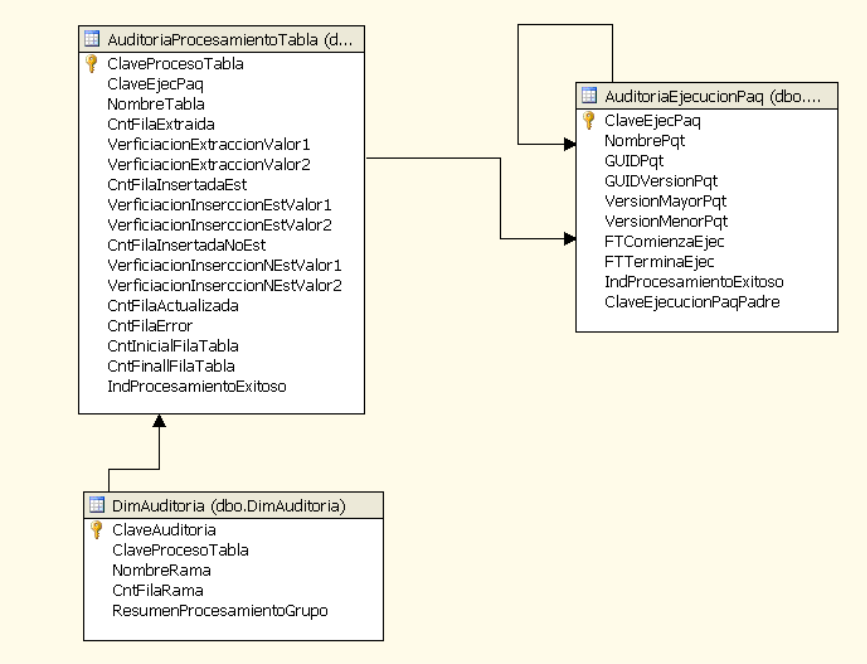


Figura 22: Sistema de Auditoria De SPARDW.

Después de haber finalizado el proceso de ETL que permitió tener el DW relacional de SPAR con datos limpios y transformados se continúa con el desarrollo del DW multidimensional.

2.1.6 CONJUNTO DE HERRAMIENTAS.

En esta etapa se describe la arquitectura técnica y la selección de productos para el sistema DW/BI de acuerdo a la ruta superior del ciclo de vida dimensional.

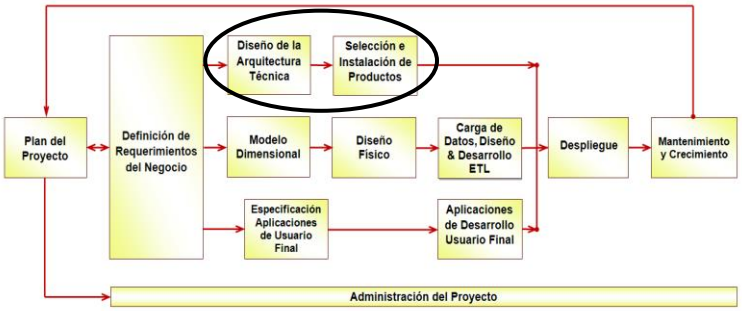


Figura 23: Ruta de la Tecnología del ciclo de vida dimensional (Adaptada de [8]).

2.1.6.1 DISEÑO DE LA ARQUITECTURA TÉCNICA

"Los ambientes del DW/BI requieren la integración de varias tecnologías. Se debe tener en cuenta algunos aspectos como son: los requerimientos del negocio, los ambientes técnicos actuales y lineamientos técnicos de planeación para poder establecer el diseño de la arquitectura técnica [8][4]". En esta etapa se especifica las herramientas y tecnologías que se requieren para responder a la pregunta cómo se debe hacer. El siguiente grafico muestra la arquitectura técnica de un sistema de DW/BI.

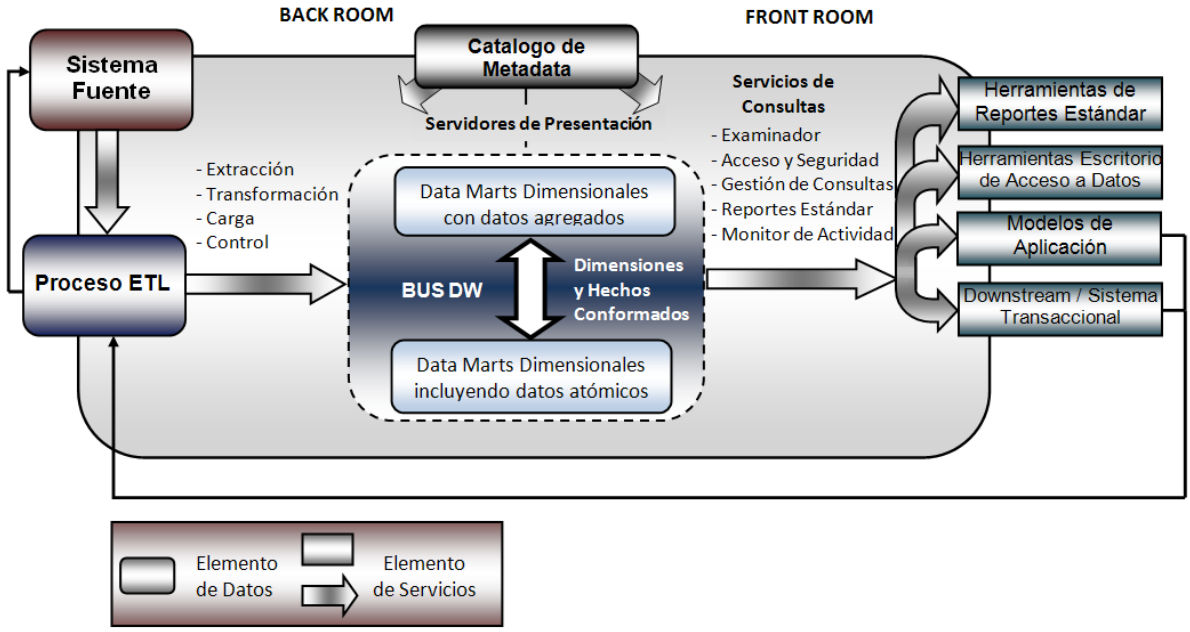


Figura 24: Arquitectura Técnica de un Sistema de DW/BI (Adoptado [8])

Como muestra la Figura 24: Arquitectura Técnica de un Sistema de DW/BI (Adoptado [8]), la arquitectura se divide en dos partes, la parte interna conocida como back room y la parte externa conocida como front room. En el back room se describe los orígenes de datos, el flujo de datos, consulta de servicios, metadata, automatización de procesos entre otros. En el front room se describe las herramientas estáticas y dinámicas para acceder y consultar los datos del DW/BI [8].

Para poder establecer una arquitectura técnica se debe tener en cuenta requerimientos de hardware y software. Para el desarrollo del sistema de DW/BI de SPAR 1.0 se cuenta con un servidor de alto desempeño y de gran volumen de almacenamiento , que brinda soporte para instalar nuevos productos software que requieren una alta capacidad de máquina, como son los productos que requiere la construcción y el mantenimiento de un sistema de DW/BI.

Para determinar la arquitectura técnica de SPARDW se necesita identificar un conjunto de herramientas con funciones específicas para las partes interna y externa de la arquitectura. Con relación a la *parte interna*, se tiene:

- El sistema fuente transaccional del repositorio SPAR 1.0 en un motor de base de datos de SQL Server.

- Para el flujo de datos es necesario una o varias herramientas que realicen extracción desde el sistema fuente, transformación, limpieza y carga de datos.

- Para la construcción y mantenimiento del DW relacional se necesita un motor de base de datos.

- Para la construcción y mantenimiento del DW multidimensional se necesita un motor de administración de bases de datos multidimensionales OLAP.

Con relación a la *parte externa* o cara pública se debe determinar dos tipos de herramientas:

- La herramienta de reportes estándares.

- La herramienta de reportes dinámicos (ad hoc).

2.1.6.2 SELECCIÓN E INSTALACIÓN DEL PRODUCTO

Utilizando el diseño de la arquitectura técnica como marco, se evalúan y seleccionan los componentes específicos de la arquitectura como la plataforma hardware, el motor de base de datos, la herramienta de ETL, herramientas de acceso y otros. Una vez evaluados y seleccionados los productos, se procede con la instalación y prueba de los mismos en un ambiente integrado de DW [8]. La finalización de esta etapa permite comenzar con el proceso de construcción física del DW relacional y multidimensional de SPAR 1.0 y el proceso de construcción de la herramienta OLAP.

Para la selección de productos se tiene en cuenta que este proyecto se encuentra enmarcado dentro del proyecto I+D titulado “Scorm Public-Access Repository” cofinanciado por Microsoft Research, por lo que se dispone de la infraestructura tecnológica proporcionada por Microsoft para el diseño e implementación y puesta en producción de todo el sistema de DW/BI de SPAR 1.0, por tal razón, se escogió SQL Server 2005 Enterprise Edition como motor de base de datos relacional y multidimensional junto con sus respectivos ambientes de desarrollo y de administrzación integrados, esto son: Business Intelligence Development Studio (Herramienta de diseño y construcción) y SQL Server Management Studio (Herramienta primaria de administración). Estas herramientas proveen de todas las funcionalidades requeridas para desarrollar completamente un sistema de DW/BI.

Otro conjunto de herramientas son diseñadas para los usuarios analíticos del negocio, estas incluyen Microsoft Office, Excel, Componentes Web de office y servicios de Share Point. Office y Share point proveen herramientas que se pueden usar para construir aplicaciones de usuario final para acceder al DW relacional y/o multidimensional.

De acuerdo a la arquitectura técnica se instalan las siguientes herramientas de desarrollo y administración de SQL Server 2005:

SQL Server Management Studio se usa para administrar el sistema de DW/BI. A través de esta herramienta se puede conectar y administrar los servidores de base de datos relacionales, de Analysis Services (Bases de Datos Multidimensionales), Integration Services (Servicios de ETL) y Reporting Services (Servidor de Reportes OLAP). También permite realizar consultas de tipo MDX, DMX y XMLA para bases de datos de análisis dimensional como por ejemplo cubos y modelos de minería de datos [5].

Business Intelligence Development Studio BI Studio se utiliza para desarrollar y diseñar sistemas de BI, esta herramienta proporciona interfaces de usuario que facilitan el diseño, desarrollo e implementación de bases de datos de Analysis Services, modelos de minería de datos, paquetes de Integration Services y reportes de Reporting Services [5].

Analysis Services (AS): Es un motor OLAP que permite almacenar, gestionar y consultar bases de datos OLAP (multidimensionales), diseñadas y construidas para usos de Inteligencia de Negocios. AS tiene las siguientes características [5]:

- **Metadata orientada al usuario**: La estructura de las bases de datos de AS permite definir nombre de los objetos (dimensiones, hechos, jerarquías) orientados al usuario, que le permite navegar fácilmente sobre los datos.

- **Análisis Complejos almacenados en la BD:** Las bases de datos de AS almacenan información sobre los cálculos, desde cálculos muy simples hasta muy complejos, como miembros calculados, indicadores de desempeño, cálculos inteligentes, perspectivas específicas, medidas semiaditivas, traducciones y otros que permiten construir una base de datos OLAP con mejores características.

- **Riqueza del lenguaje Analítico:** SQL es un lenguaje orientado a las consultas, no es un lenguaje analítico, las herramientas de consulta para Analysis Services generan MDX (Expresiones Multidimensionales) en vez de SQL. La fortaleza más grande del MDX es el entendimiento de la metadata dimensional: Hechos, dimensiones, atributos, jerarquías, miembros, etc.

- **Desempeño de Consulta:** El motor OLAP ha sido diseñado como un servidor de alto desempeño. AS tiene características de almacenamiento y gestión de tablas resúmenes o agregadas que mejoran significativamente el desempeño de las consultas.

- **Minería de Datos:** Provee las herramientas y funcionalidades necesarias para desarrollar modelos de minería de Datos.

Integration Services (IS): Es una herramienta de diseño y desarrollo de paquetes que son usados para hacer extracción, transformación y carga del DW relacional. Sin embargo IS es

más que una herramienta de ETL, esta puede ser usada para gestionar y mantener bases de datos e integrar datos entre aplicaciones o realizar cálculos complejos en tiempo real.

Reporting Services: Es una herramienta de reportes que permite el almacenamiento, la gestión y consulta de reportes predefinidos. Igualmente provee a los usuarios la posibilidad de construir informes ad hoc (personalizados) simples, por medio de una herramienta llamada el "Report Builder" (constructor de Reportes).

A continuación se describe la arquitectura de un sistema de DW/BI que usa herramientas Microsoft para su diseño, construcción y mantenimiento. La Figura 25 muestra la arquitectura de un Sistema Microsoft de DW/BI.

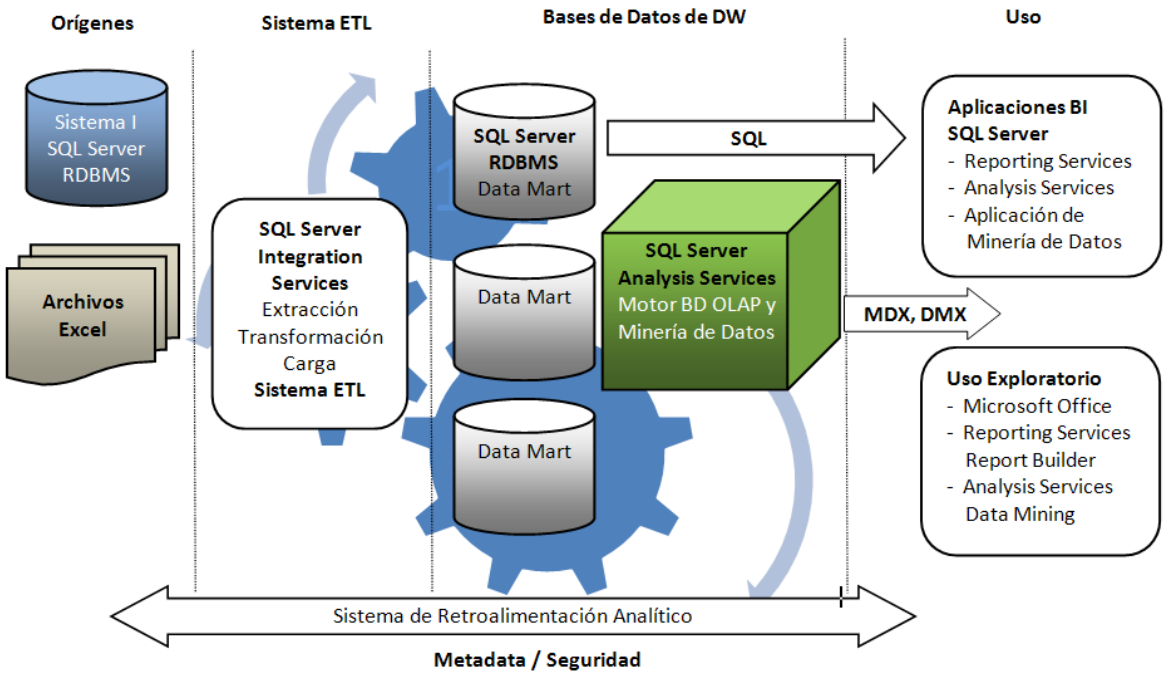


Figura 25: Arquitectura de un Sistema Microsoft de DW/BI. (Adaptada de [5])

Los cuatro componentes de esta arquitectura son: sistemas de datos fuente, sistema ETL, Bodegas de datos relacional y multidimensional, y aplicaciones de Inteligencia de Negocios. El flujo es el siguiente: Los modelos dimensionales se traduce en la construcción de en una bodega de datos relacional para lo cual se usa el motor de Base de datos relacional de SQL Server 2005, luego se desarrolla un sistema ETL donde se construyen paquetes de Integration Services que permiten extracciones, transformaciones y cargas del DW relacional, a continuación se realiza la construcción de una base de datos OLAP a partir del DW relacional, la cual se administra y se almacena usando el motor OLAP de Analysis Services, luego se utilizan herramientas OLAP que permiten la construcción y consulta de reportes predefinidos al igual que consultas dinámicas para acceder a la información. Las herramientas OLAP de Microsoft son Reporting Services, Report Builder, Componentes de Office Web, Microsoft Office Excel, entre otros. Como muestra la Figura 25, el acceso a la

información puede hacerse desde el DW relacional o desde la base de datos multidimensional, pero se recomienda que se haga desde una capa superior orientada al usuario, en este caso sería la base de datos de AS, que además de ser orientada al usuario provee un lenguaje de consultas (MDX) que es más eficiente para realizar análisis multidimensionales que el lenguaje de consultas SQL usado para acceder a bases de datos relacionales.

Para la selección de herramientas front room (usuarios finales) de la Arquitectura Técnica es necesario considerar algunos aspectos:

Para poder tomar una decisión en cuanto a construir o hacer uso de las herramientas OLAP que ofrece Microsoft dentro del área de Inteligencia de Negocios, se identificaron un conjunto básico de funcionalidades que debería soportar la aplicación OLAP basadas en las propuestas por Ralph Kimball [5]. Considerando que el principal objetivo de las herramientas OLAP es proporcionar a los usuarios la información analítica que ellos necesitan un una manera adecuada, útil, entendible, manejable, flexible y permitirles tomar decisiones comerciales apropiadas, se construye la siguiente tabla que lista un conjunto de requerimientos y sus implicaciones funcionales que debe soportar la herramienta OLAP de SPAR.

REQUERIMIENTOS COMERCIALES	IMPLICACIONES FUNCIONALES
Crear Reportes	Variedad de formatos de presentación (tablas, gráficos, matrices, etc.). Una herramienta poderosas, rápido, facial para construir reportes.
Encontrar Reportes	Marco de Navegación Metadata Búsqueda
Visualizar Reportes	Acceso por medio de una gran variedad de formas. Como por ejemplo Navegadores ó auto email.
Recibir Resultados en la forma más apropiada	Salida de resultados en una variedad de tipos de archivos.
Cambiar Reportes de acuerdo a las necesidades	Parámetros Drill Down/ Atributos Adicionales Enlaces
Sistema solido y confiable.	Desempeño Escalabilidad Gestión
Consultas Dinámicas Personalizadas (AD HOC)	Arrastrar y Colocar
Consultas off-line	Arrastrar y Colocar

Figura 26: Requerimientos de usuario e implicaciones funcionales. (Adaptada de [5])

Comparación de Herramientas OLAP:

Basados en esta lista de requerimientos funcionales se hizo un estudio comparativo de las funcionalidades ofrecidas por Microsoft y las distintas herramientas para consultas OLAP más reconocidas en el mercado. Esta tabla comparativa se muestra en el **Anexo 11.** Esta tabla muestra que varias herramientas del mercado cumplen con la gran mayoría de funcionalidades requeridas para el diseño, construcción y despliegue de un DW/BI, sin embargo estas herramientas requieren de la compra de licencias. En el caso de las herramientas Microsoft se cuenta con las licencias necesarias para llevar a cabo el desarrollo del sistema del DW/BI de SPAR 1.0, como se explico con anterioridad.

2.1.7 DISEÑO DE LA BASE DE DATOS MULTIDIMENSIONAL

Luego de la construcción y carga del DW relacional, el siguiente paso se centra en la construcción de la base de datos multidimensional en la herramienta de Analysis Services de Microsoft Business Intelligence. Esta parte del proyecto se enmarca dentro del camino de datos del ciclo de vida dimensional. Este camino involucra el modelado dimensional, diseño físico, y proceso de ETL para poblar la base de datos.

La construcción de la base de datos multidimensional en Analysis Services comprende algunos pasos principales como son [5]:

- Preparar el diseño y ambiente de desarrollo
- Crear una Vista de Origen de Datos
- Crear y refinar las dimensiones
- Crear y Editar el cubo
- Procesar el cubo
- Crear cálculos
- Realizar iteraciones

El proceso de construcción de la base de datos multidimensional se describe a continuación.

El proceso que se siguió para el diseño de la base de datos multidimensional fue:

1. Crear un proyecto de Analysis Services con el nombre de SPARAS (Bodega de Datos de SPAR en Analysis Services).

2. Crear una conexión de origen de datos a la base de datos SPARDW (DW relacional).

3. Crear una vista de origen de datos, la cual es una capa que contiene metadatos de las tablas sobre las cuales se va a definir la base de datos multidimensional. Al crear la vista de origen de datos se eligen las tablas de dimensiones y tablas de hechos correspondientes para el Data Mart de Evaluación de Contenidos, de Gestión, Oferta y Demanda de Contenidos y Sesiones de Usuario. Las vistas permiten modificar los nombres de los objetos (Tablas de Dimensiones, Tablas de Hechos, Atributos, Medidas) para que sean más entendibles para el usuario final. Las modificaciones hechas en la vista no afectan el origen de datos relacional.

4. Crear el cubo multidimensional, el cual se basa en la vista de origen de datos para hacer la selección de las tablas de dimensiones y tablas de hechos correspondientes (Figura 27), las tablas de dimensión taxonomía y dimensión puente taxonomía no se seleccionan porque existe una relación muchos a muchos entre la dimensión taxonomía y la dimensión objetos para lo cual es necesario seguir otro proceso para adicionar estas tablas al cubo. Posteriormente se seleccionan las medidas que estarán en el cubo en donde cada medida pertenece a un grupo de medida, cada grupo de medida es equivalente a una Tabla de Hechos, para finalizar se le da un nombre al cubo (Cubo SPARDW). En este momento se ha definido el cubo y algunos objetos multidimensionales (Dimensiones y hechos).

5. Agregar las dimensiones de taxonomía y puente taxonomía por separado y para crear una relación de muchos a muchos entre estas dos dimensiones, esto se explicará más adelante.

6. Se personalizan las dimensiones, por medio de una interfaz de diseño en Análisis Services, donde aparecen un conjunto de pestañas llamadas Estructura de dimensiones, Traducciones y Examinador. La pestaña de estructura de dimensiones es muy importante porque es aquí donde se agregan, se quitan y se editan sus atributos, se editan las relaciones entre ellos y las propiedades de los mismos, además permite la creación de jerarquías y niveles, por ejemplo para la Dimensión de Tiempo Del Día se construye un jerarquía que involucra los siguientes niveles de atributos: Hora, Minuto, Segundo. De la misma manera se crearon jerarquías para otras dimensiones. La Figura 28 muestra la pestaña de estructura de dimensión con las jerarquías creadas para la Dimensión Tiempo del Día.

   Cuando se edita una nueva dimensión, se siguen los siguientes pasos si es necesario [5]:

   - Editar los nombres de la dimensión y cada atributo dentro de la dimensión.
   - Editar las propiedades de la dimensión.
   - Editar las propiedades de cada atributo.
   - Crear las relaciones de atributo y editar sus propiedades.
   - Crear las jerarquías. Editar las propiedades de las jerarquías y las propiedades de cada nivel.
   - Si es necesario, definir las traducciones de la dimensión.
   - Procesar la dimensión para mirar los datos de la dimensión.
   - Iterar

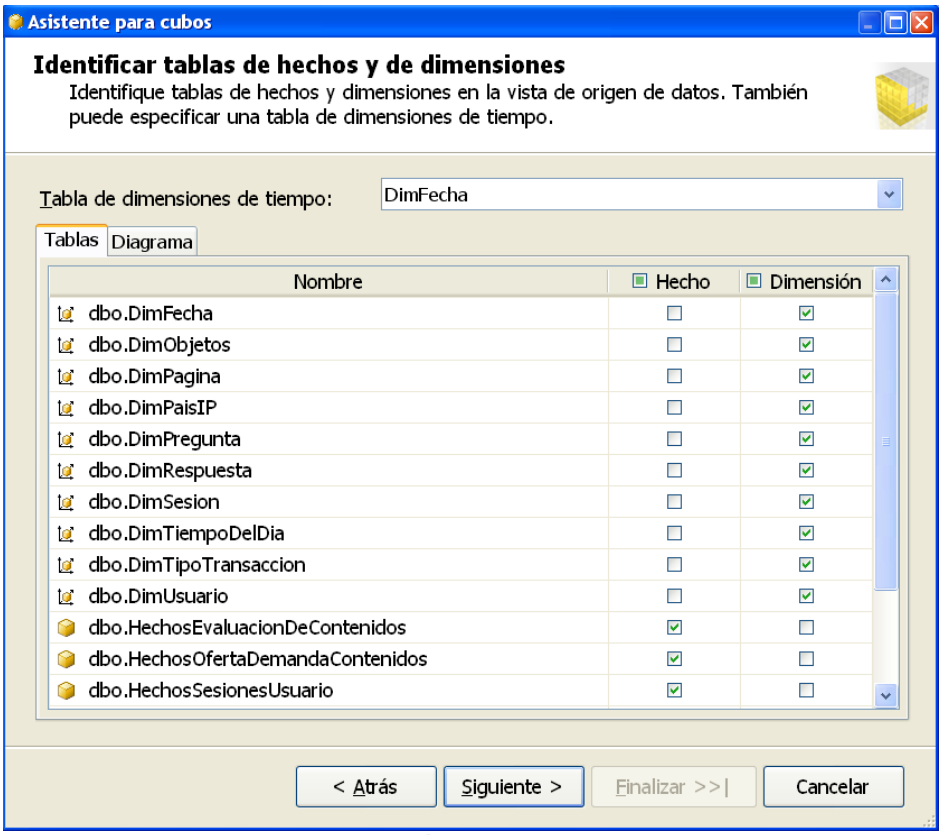


Figura 27: Ventana del asistente de creación de cubos que permite seleccionar las tablas de hechos y de dimensiones asociadas para la creación del cubo OLAP.

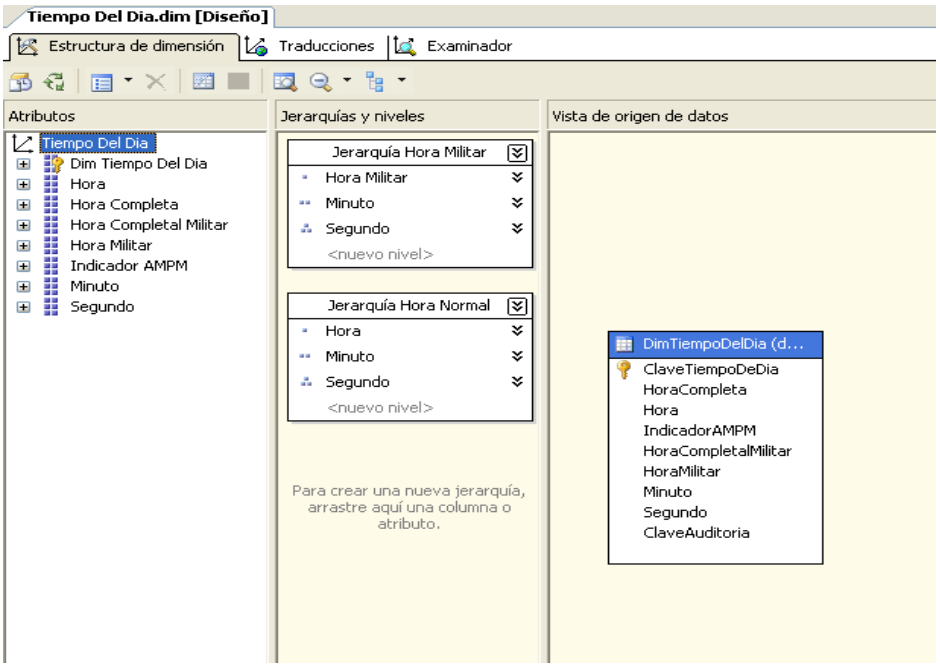


Figura 28: Pestaña de Estructura de Dimensión de la Dimensión Tiempo del Día y su jerarquías.

Después de crear el cubo, se puede observar su interfaz de diseño dentro del Business Intelligence Development Studio, que contiene pestañas como: Estructura de Cubo, Uso de Dimensiones, Perspectivas, Traducciones, Examinador, etc. como lo muestra la Figura 29. Se procederá a explicar de forma general lo que se puede hacer en las pestañas que se nombraron anteriormente.

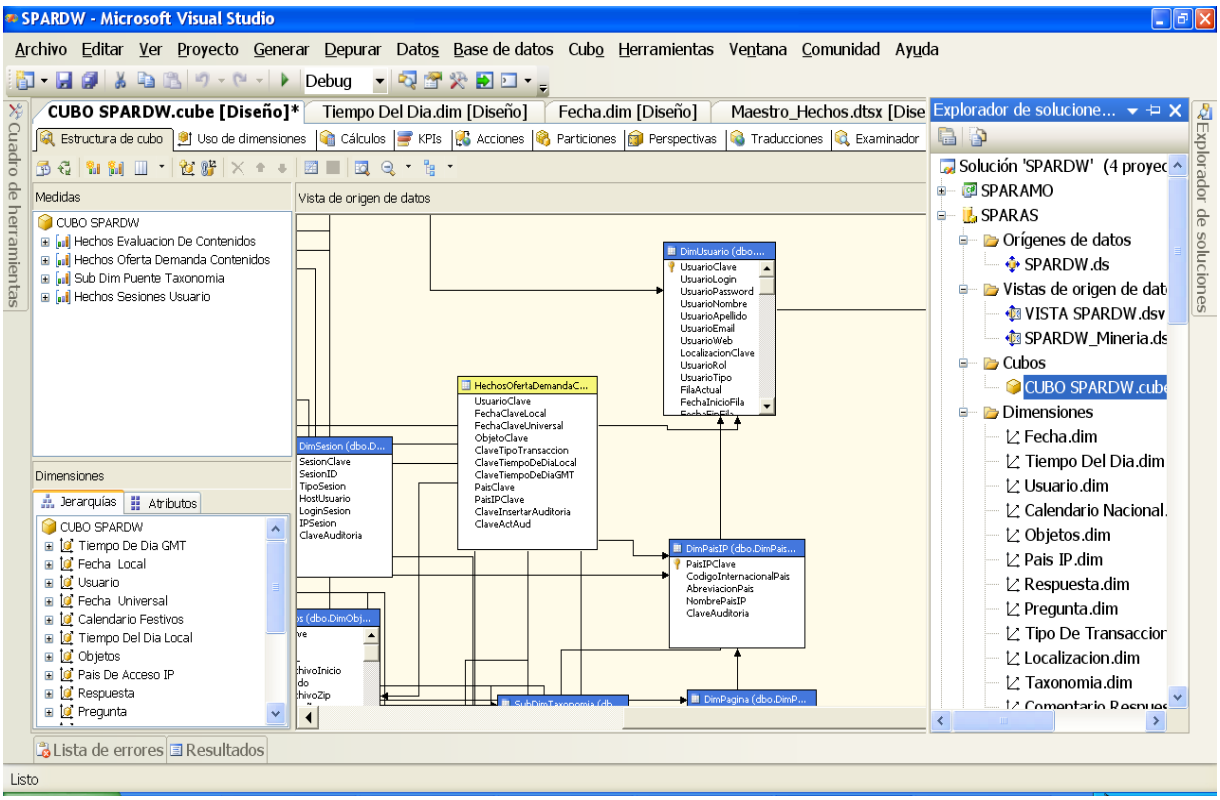


Figura 29: Interfaz de diseño de cubos multidimensionales.

Estructura del cubo

La actividad principal en la edición de la estructura del cubo consiste en la edición de las propiedades de los grupos de medida, medidas y dimensiones del cubo.

Otro aspecto importante es el juego de roles de las dimensiones, en el proyecto se presentó este caso con las dimensiones fecha y tiempo del día, en donde cada una de estas dimensiones son tablas en la base de datos relacional, por lo tanto, se les denomina dimensiones de base de datos, pero cuando tienen más de un rol, se generan dimensiones de cubo por cada rol existente entre la dimensión y la tabla de hechos. Lo mismo sucede con la dimensión página y su relación con el grupo de medida Sesiones de Usuario, en el DW relacional existe una sola tabla de página, pero al tener tres roles (tres relaciones) con la tabla de hechos, se generan tres dimensiones de cubo en la base de datos multidimensional.

Uso de dimensiones

Se utiliza para verificar los tipos de relaciones de las dimensiones con los grupos de medidas y configurar las relaciones faltantes. En los grupos de medidas del proyecto se tienen relaciones de tipo normal con la mayoría de las dimensiones, excepto las siguientes:

- Subdimensión Localización, para la cual se define una relación de referencia, que consiste en relacionar la subdimension Localización con los grupos de medida (tablas de hechos) correspondientes a través de la dimensión usuario.

- Subdimensión Taxonomía, para la cual se define un "grupo de medida intermedio" que para este caso es la dimensión puente taxonomía, este grupo de medida intermedio permite enlazar la dimensión taxonomía con la dimensión objetos y por lo tanto con el grupo de medida (tabla de hechos) correspondientes. Otra de las características de la subdimensión taxonomía es su relación reflexiva, por lo que se debe crear una relación padre hijo para poder visualizar el árbol taxonómico de esta dimensión.

- Un caso especial se presenta con la medida comentario de respuesta del grupo de medida de Evaluación de Contenidos, esta medida es de tipo no aditiva y por tanto no pudo ser configurada en AS, para darle solución se decidió crear una dimensión de hechos o degenerada que si son soportadas por AS, esta dimensión está compuesta de un solo atributo (comentario respuesta) y se maneja en el cubo SPARDW como otra dimensión relacionada con el grupo de medida Evaluación de Contenidos. La Figura 30 muestra la pestaña uso de dimensiones donde pueden visualizarse las dimensiones y el tipo de relaciones con los grupos de medidas para SPARAS.

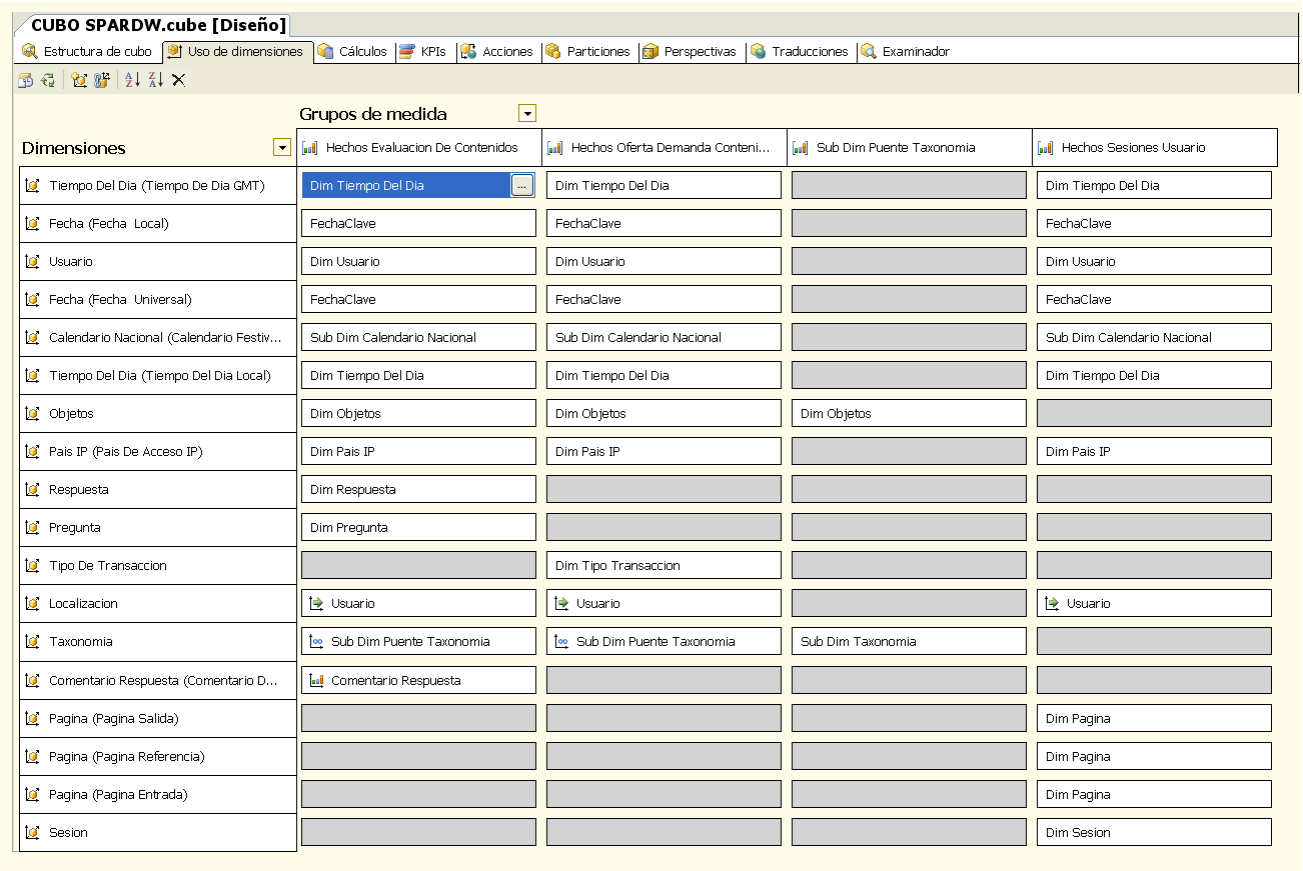


Figura 30: Pestaña Uso de Dimensiones, muestra los tipos de relaciones entre dimensiones y los grupos de medidas para SPARAS.

Perspectivas

Permite la creación de subconjuntos del cubo. Para el cubo de SPARDW se crearon tres perspectivas, una perspectiva por cada Data Mart.

Por último se procedió a procesar el cubo, cuando éste se ha completado en su totalidad y sin errores, se procede a examinar los datos a través de cruces dimensionales para resolver posibles inconsistencias. Cuando el cubo esta correctamente generado, se puede realizar sobre él consultas OLAP a través de una herramienta para Análisis de Datos Multidimensionales.

2.1.8 DESPLIEGUE DEL DW

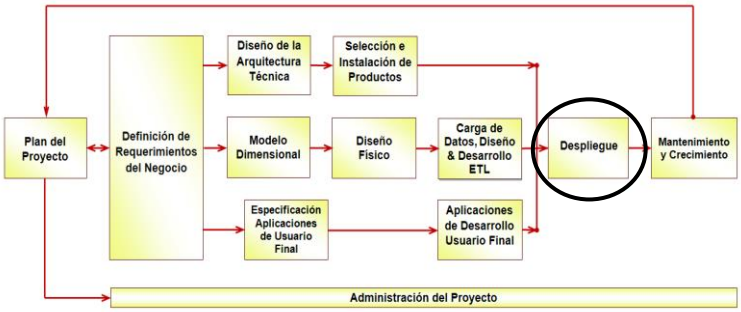


Figura 31: Ciclo de vida dimensional, etapa de Despliegue

El despliegue es donde concurren las tres rutas, la ruta de tecnología, la ruta de datos y la ruta de aplicaciones de usuarios finales. En esta fase se realizan varias actividades, las cuales deben realizarse antes de poner en producción el producto para garantizar al usuario un correcto funcionamiento, entre estas actividades están: configuración de hardware, configuraciones de red, conexiones a base de datos, verificaciones de seguridad, instalación de software, documentación, manuales y capacitación a los usuario, entre otros. Una de las partes más importantes dentro de estas actividades es la capacitación al usuario final, donde se le debe presentar e instruir sobre el contenido, las aplicaciones y herramientas de acceso del DW/BI [8].

Para el despliegue del sistema de DW/BI de SPARDW, fue necesario verificar que los productos estuvieran correctamente instalados, verificar el funcionamiento del servidor en cuanto a hardware como las conexiones de bases de datos con el repositorio SPAR 1.0, el correcto funcionamiento del sitio, del servidor de reportes, del servidor ETL, del servidor de análisis, el acceso a los datos y hacer la entrega de manuales y documentación necesaria para el manejo de SPARDW, también se realizaron capacitaciones a los usuarios administradores.

2.1.9 MANTENIMIENTO Y CRECIMIENTO

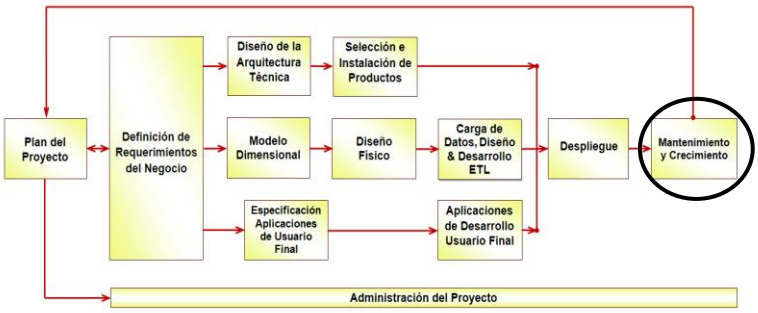


Figura 32: Ciclo de vida dimensional, etapa de Mantenimiento y Crecimiento

Es un proceso con etapas bien definidas, con comienzo y fin, pero de naturaleza espiral que acompaña la evolución de la organización durante toda su historia. Es importante establecer las prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer. Se debe tener en cuenta una serie de puntos para mantener el DW exitosamente, entre ellos se destacan: el continuo soporte y la constante capacitación a usuarios de negocios, el manejo de la infraestructura (monitoreo de base de datos, tráfico, etc.), afinar el rendimiento sobre las consultas, mantenimiento de la metadata y procesos ETL [8]. Estas dos etapas están fuera del alcance de este proyecto.

2.1.10 ADMINISTRACIÓN DEL PROYECTO

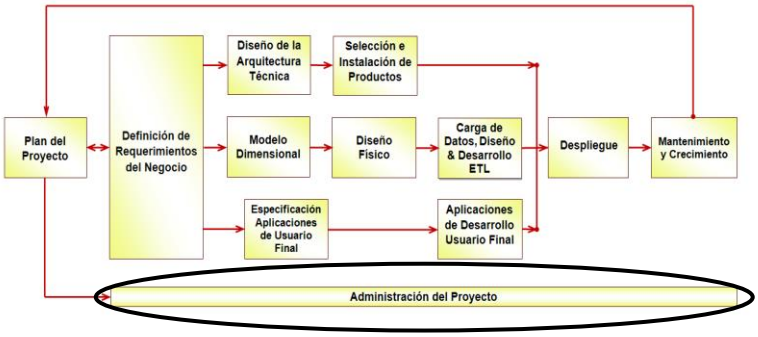


Figura 33: Ciclo de vida dimensional, etapa de Administración del Proyecto.

La etapa de administración del proyecto es la encargada de monitorear, controlar y administrar todas las actividades del proyecto que se realizan en el Ciclo de Vida Dimensional. Las diferentes actividades se centran en verificar el estado del proyecto y su evolución con el tiempo [8]. La administración del presente proyecto se fue realizando a lo largo de desarrollo de todo el sistema de DW/BI, al contrastar los resultados que se estaban obteniendo con el plan proyecto establecido en las primeras etapas de construcción. Para esto fue indispensable el rol que jugó el director del proyecto.

CAPITULO III:

3. DESCRIPCIÓN DEL PROTOTIPO DE LA HERRAMIENTA OLAP

En este capítulo se describe el proceso de desarrollo que se utilizó para la construcción de la herramienta OLAP. De la misma manera, se presentan los artefactos que se obtuvieron en cada una de las fases de la metodología de trabajo.

El proceso de desarrollo de las aplicaciones de Business Intelligence (BI) hace parte de uno de los tres caminos del Ciclo de Vida Dimensional, metodología escogida para el desarrollo completo del DW/BI. Esta metodología divide el proceso de desarrollo de las aplicaciones de BI en dos etapas: La Especificación y el Desarrollo. La Figura 34 muestra las etapas del Ciclo de Vida Dimensional que hacen parte del desarrollo de las aplicaciones de BI.

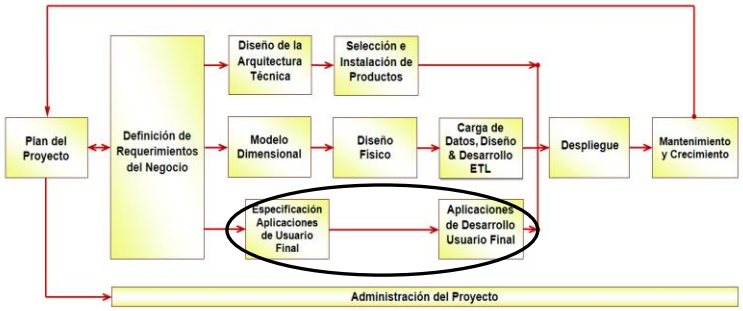


Figura 34: Ciclo de Vida Dimensiona (Adoptada [8])

La metodología usada para el desarrollo de la herramienta OLAP consiste en combinar estas dos fases del Ciclo de Vida Dimensional con el Proceso Unificado Racional de Desarrollo de Software (RUP). Permitiendo que el proceso de desarrollo sea iterativo e incremental, dirigido por la realización de casos de uso y centrado en la arquitectura. De esta forma, se estableció una iteración base con sus diferentes fases, pero con la revisión continua de las fases de esta iteración de modo que se logrará un desarrollo iterativo e incremental.

El proceso Unificado de Desarrollo se divide en ciclos y cada ciclo se divide en cuatro fases que se integrarán a las dos etapas del Ciclo de Vida Dimensional de la siguiente manera: La fase de Preparación Inicial y Preparación Detallada hacen parte de la etapa de Especificación y las dos fases finales Construcción y Transición hacen parte de la Etapa de Desarrollo. La descripción de todo el proceso de desarrollo del proyecto se presenta a continuación.

3.1 ETAPA DE ESPECIFICACIÓN DE LAS APLICACIONES

3.1.1 FASE DE PREPARACIÓN INICIAL

Esta fase incluye la concepción inicial, los objetivos para el ciclo de vida, la investigación de alternativas, la planificación y el alcance del proyecto.

3.1.1.1 Definición de Requerimientos y Concepción Inicial

La definición de los requerimientos hecha en las primeras etapas del Ciclo de Vida Dimensional determina el alcance de las herramientas de BI. Los requerimientos iníciales del proyecto fueron identificados como temas analíticos y clasificados dentro de los procesos o áreas del negocio que posteriormente son priorizadas para determinar el alcance del proyecto.

En el gráfico de impacto Vs Viabilidad del Capítulo II (Figura 8) se identificaron las siguientes áreas de negocio como las de más alta prioridad y de mayor viabilidad: Calidad de contenidos, Gestión Oferta y Demanda de Contenidos, Perfiles de Usuario (Sesiones de Usuario). Los Temas analíticos identificados y clasificados en las primeras etapas permitieron dividir la funcionalidad completa de la herramienta OLAP en dos módulos: ***Reportes Estándares y Aplicación Analítica.***

Los reportes estándar deben ser relativamente simples con formato y parámetros predefinidos. Estos reportes deben proporcionar para el usuario común de SPAR y para el administrador un conjunto central de información de lo que está pasando en cada proceso del negocio en particular. (Evaluación, Gestión Oferta y Demanda, Sesiones de Usuario). La mayoría de las necesidades analíticas identificadas con las entrevistas y reuniones serán presentadas como reportes estándares por parte del modulo de Reportes OLAP de SPAR.

La aplicación analítica permite hacer análisis más complejos que los arrojados por los informes estándar, como consultas dinámicas y personalizadas en las que cualquiera de los usuarios de SPAR puede establecer sus parámetros, filtros y dimensionalidad. El Modulo OLAP de la Aplicación analítica permite realizar consultas sobre cualquiera de los tres modelos dimensionales de la Bodega de Datos: Evaluación de Contenidos, Gestión oferta y Demanda de Contenidos y Sesiones de Usuario. Estos análisis pueden realizarse directamente sobre el cubo multidimensional alojado en el servidor de Analysis Services o puede hacerse sobre archivos XML que representan perspectivas del cubo, que pueden ser descargados por los usuarios, permitiendo hacer análisis off-line (Desconectados del Servidor de Analysis Services). La aplicación analítica también incluye el modulo de minería de datos de recomendaciones de Recursos de Aprendizaje además del modulo de consultas dinámicas (AD HOC). En esta sección se hace referencia sólo al modulo de Consultas Dinámicas y Reportes Estándares, el proceso de desarrollo del modulo de minería de datos es describe en el Capitulo 4.

Entre los artefactos que se obtuvieron durante esta fase se encuentran, los diagramas de casos de uso, los casos de uso de alto nivel, modelo conceptual preliminar, diagramas de

secuencia, una lista de requerimientos funcionales que debe soportar las aplicación OLAP del proyecto, que sirven como criterios para determinar si se debe desarrollar o se debe integrar las herramientas actuales que hacen parte de la plataforma OLAP de Microsoft. Una tabla comparativa entre las funcionalidades ofrecidas por las distintas herramientas OLAP más reconocidas en el mercado. Para el caso particular de los reportes estándar se define la platilla estándar de reportes y lista de reportes candidatos. La descripción de los artefactos mencionados se presenta a continuación:

3.1.1.2 Diagrama de Casos de Uso:

Un diagrama de *casos de uso* describe un conjunto de actividades realizadas por determinados actores sobre un sistema.[12] En la Figura 35 se presenta el diagrama de casos de uso general, en el cual se presentan los dos tipos de usuario que interactúan con cualquiera de los tres módulos OLAP (Herramienta de Reportes, Componente de Consultas Dinámicas – Cubo Remoto, Componente de Consultas Dinámicas –Cubo Local)

Cada uno de los tres módulos OLAP presenta un diagrama de casos de uso detallado, en el cual se visualizan los diferentes casos de uso implementados para que cada módulo cumpla con su funcionalidad específica, estos casos de uso pueden ser encontrados en el **Anexo 8.**

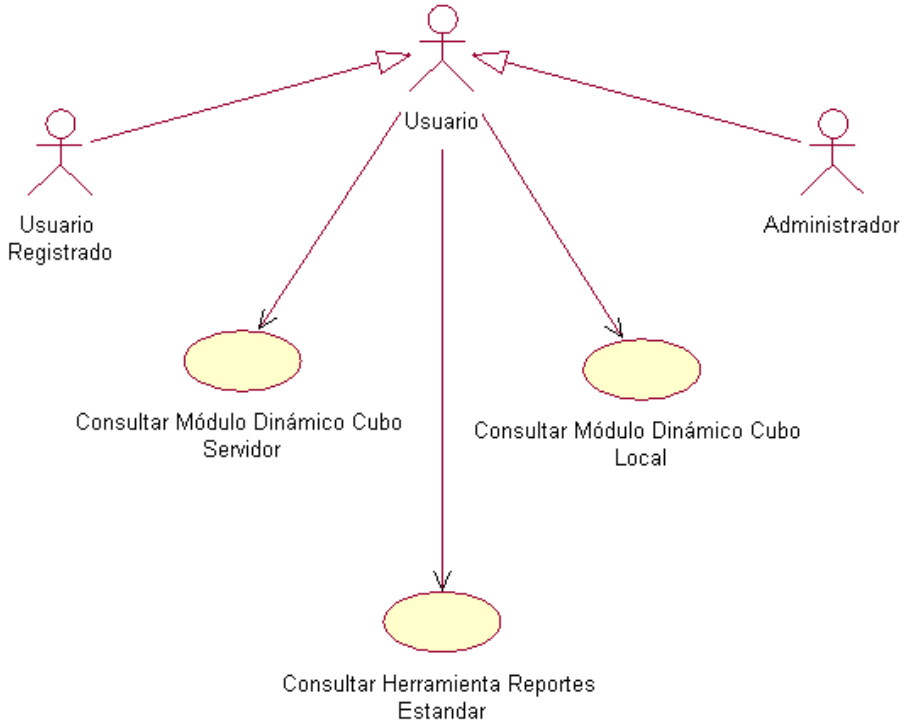


Figura 35: Diagrama general de casos de uso para la herramienta OLAP.

3.1.1.3 Casos de Uso de Alto Nivel:

“El caso de uso de alto nivel es un documento narrativo que describe la secuencia de eventos de un actor (agente externo) que utiliza un sistema para completar un proceso” [12]

Los casos de uso de alto nivel para los tres módulos OLAP de SPAR se describen en el **Anexo 9.**

3.1.1.4 Modelo Conceptual Preliminar:

"Un modelo conceptual es una representación de conceptos en un dominio del problema" [12]. El modelo conceptual preliminar del proyecto se compone de los siguientes conceptos y se presenta en la Figura 36.

- **Interfaz SPAR:** Este concepto hace referencia al mecanismo de comunicación entre los usuarios y el repositorio digital. Para este caso en concreto son las tres herramientas software Search-SCORM, Author-SCORM y Admin-SCORM que hacen parte del Repositorio Digital de Objetos de Aprendizaje SPAR.

- **Herramienta Reportes:** Este concepto hace referencia a la herramienta donde se publican y se consultan los reportes estándares OLAP.

- **Herramienta Dinámica C_Local:** Este concepto hace referencia a la herramienta que permite hacer análisis dinámicos (ad hoc) sobre el cubo OLAP alojado en el servidor de Analysis Services.

- **Herramienta Dinámica C_Servidor:** Este concepto hace referencia a la herramienta que permite hacer análisis dinámicos (ad hoc) sobre los archivos de cubo local almacenados en el equipo del usuario.

- **Catálogo Reportes:** Este concepto hace referencia a la lista de reportes estándar que han sido publicados en el servidor de reportes.

- **Cubo:** Este concepto hace referencia a un cubo definido en la base de datos multidimensional que contiene todos los procesos del negocio (tres Data Mart).

- **Medida:** Este concepto hace referencia a una medida definida para un grupo de medidas del cubo multidimensional.

- **Grupo de Medidas:** Este concepto hace referencia a un grupo de medidas que hace parte del cubo y que está relacionado a las dimensiones.

- **Dimensión:** Este concepto hace referencia a una dimensión definida en la base de datos multidimensional.

- **Jerarquía:** Este concepto hace referencia a una jerarquía definida en una dimensión de la base de datos multidimensional.
- **Atributo:** Este concepto hace referencia a una columna definida en una dimensión de la base de datos multidimensional.

- **Reporte:** Es un reporte almacenado en el servidor de reportes.

- **Perspectiva:** Es una perspectiva definida en el cubo multidimensional, simplifica la vista del cubo.

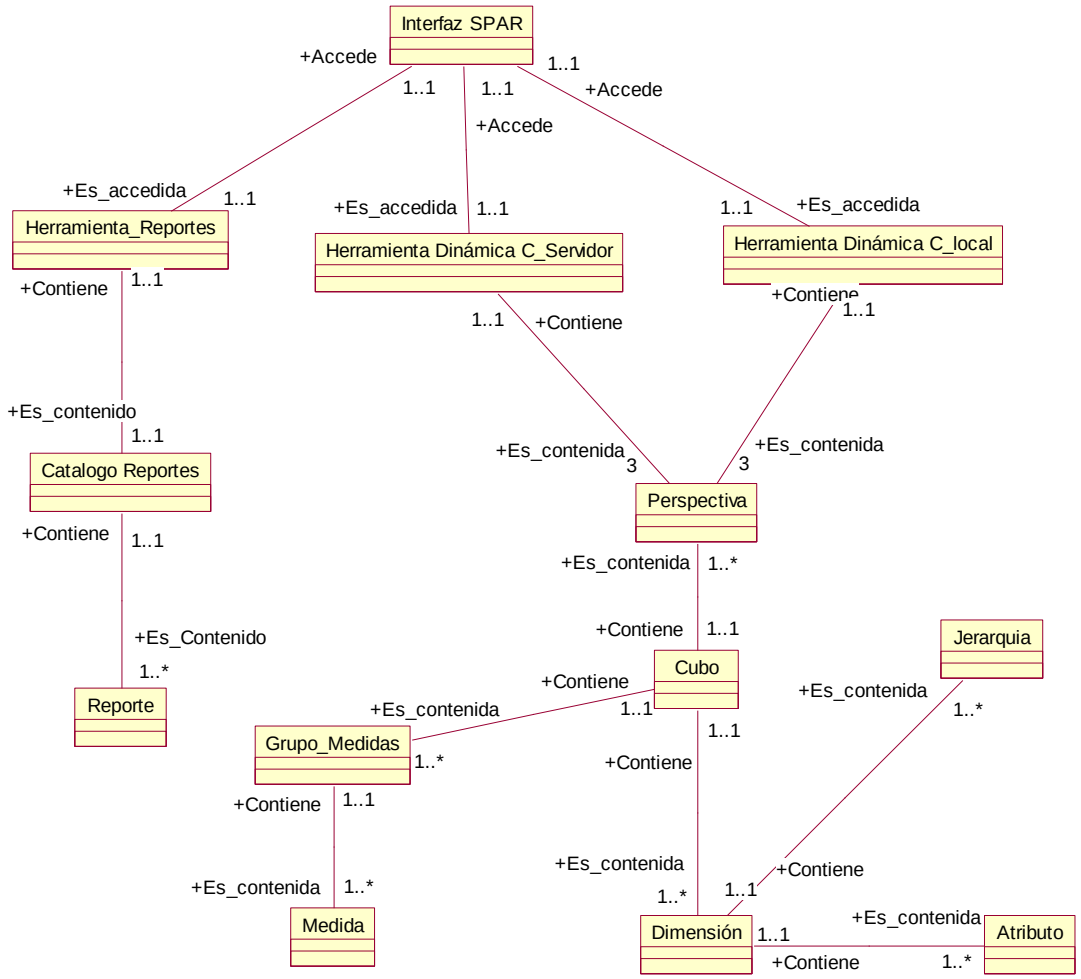


Figura 36: Modelo Conceptual Preliminar

3.1.1.5 Diagramas de Secuencia:

"El diagrama de la secuencia de un sistema muestra gráficamente los eventos que fluyen de los actores al sistema" [12]. Los diagramas de secuencia para las herramientas se describen en el **Anexo 10.**

3.1.1.6 Lista de Funcionalidades Requeridas:

Para poder tomar una decisión en cuanto a construir o hacer uso de las herramientas que ofrece Microsoft dentro del área de Inteligencia de Negocios, se identificaron un conjunto básico de funcionalidades que debería soportar nuestra aplicación OLAP, basadas en la lista de funcionalidades propuestas en el libro "*The Microsoft Data Warehouse Toolkit*" [5], para poder satisfacer su principal objetivo, que es proporcionar a los usuarios la información analítica que ellos necesitan un una manera adecuada, útil, entendible, manejable, flexible y permitiéndoles tomar apropiadas decisiones comerciales. La Tabla 3 lista un conjunto de requerimientos y sus implicaciones funcionales que debe soportar la herramienta OLAP de SPAR 1.0.

REQUERIMIENTOS COMERCIALES	IMPLICACIONES FUNCIONALES
Crear Reportes	Variedad de formatos de presentación (tablas, gráficos, matrices, etc.). Una herramienta poderosas, rápida, fácil para construir reportes.
Encontrar Reportes	Marco de Navegación Metadata Búsqueda
Visualizar Reportes	Acceso por medio de una gran variedad de formas. Como por ejemplo Navegadores ó auto email.
Recibir Resultados en la forma más apropiada	Salida de resultados en una variedad de tipos de archivos.
Cambiar Reportes de acuerdo a las necesidades	Parámetros Drill Down/ Atributos Adicionales Enlaces
Sistema sólido y confiable.	Desempeño Escalabilidad Gestión
Consultas Dinámicas Personalizadas (AD HOC)	Arrastrar y Colocar
Consultas off-line	Arrastrar y Colocar

Tabla 3: Lista de requerimientos y sus implicaciones funcionales. (Adaptada [5])

3.1.1.7 Comparación de Herramientas OLAP:

Basados en esta lista de requerimientos funcionales se hizo un estudio comparativo de las funcionalidades ofrecidas por Microsoft y las distintas herramientas OLAP más reconocidas en el mercado. Esta tabla comparativa se muestra en el **Anexo 11.**

3.1.1.8 Plantilla Estándar de Reportes:

Se crea una plantilla para identificar los elementos estándar que aparecerán en cada reporte, incluyendo sus marcos y estilos. Es útil definir la plantilla estándar antes de empezar a definir el listado de los informes individuales, esta plantilla da un contexto para definir los informes.

Los siguientes elementos conforman la plantilla de reportes [5]:

- El nombre del Informe:
- El título del Informe
- Cuerpo del Informe:
    1. Justificación de los datos:
    2. La precisión de los datos:
    3. Formato de encabezado de columna y fila
    4. El fondo relleno y colores.
    5. Formatear totales o dividir en filas de subtotal.
    6. El encabezado y pie de página: Los elementos siguientes deben encontrarse en alguna parte en el encabezado o pie de página.
        1. El nombre del Informe.
        2. Los parámetros usados.
        3. Categoría de navegación.
        4. Las notas del Informe.
        5. La enumeración de la página.
        6. Hora y fecha de ejecución del reporte.
        7. El origen de los datos:
        8. La instrucción de confidencialidad.
        9. La referencia del DW/BI: el nombre y logotipo.
- Nombre del archivo del Reporte

La Figura 37 muestra la plantilla estándar definida para la creación de informes del proyecto:

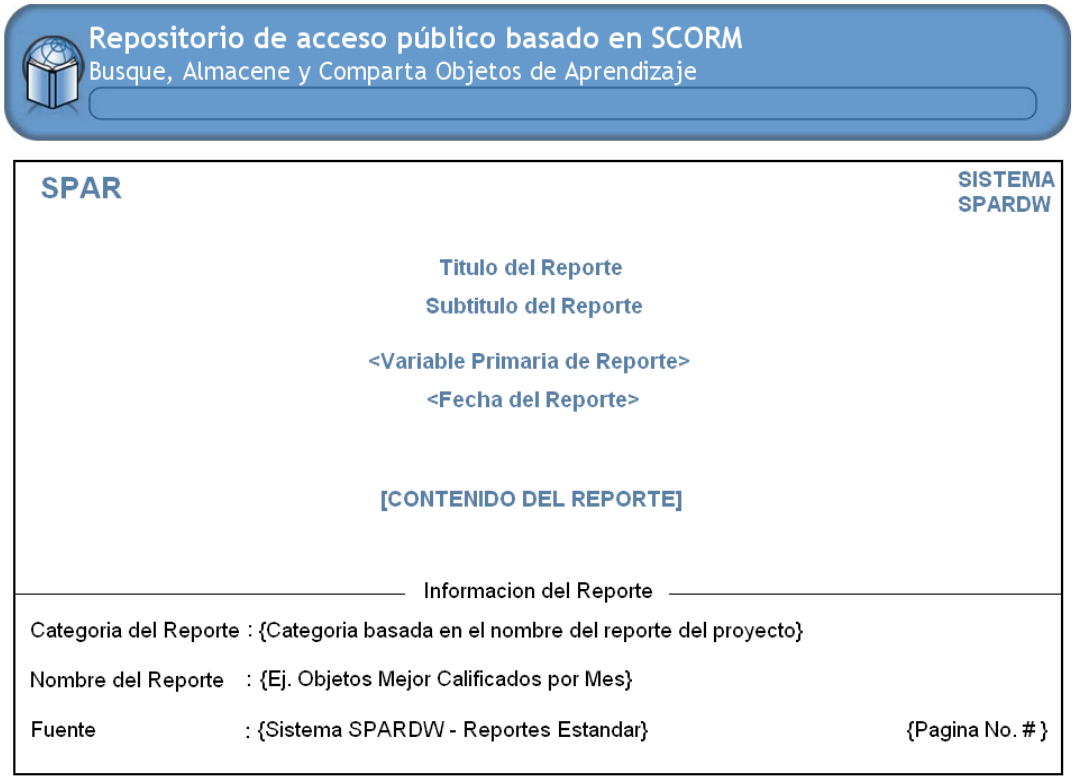


Figura 37: Plantilla Estándar de Reportes para SPAR 1.0

3.1.1.9 Lista de Reportes Candidatos:

Para crear una lista de reportes candidatos a ser construidos y publicados, se hace una revisión de los necesidades analíticas identificadas en la etapa de recolección de requerimientos del ciclo de vida dimensional y se crea una tabla que describe cada reporte, un nivel de esfuerzo de construcción, una categoría a la cual pertenece, un usuario final para quien va dirigido, un valor de impacto en el negocio, una descripción del tipo de reporte que es (Grafico, Tabla, Matriz, combinación), entre otros. La tabla de reportes candidatos se puede observar en el **Anexo 12.**

3.1.2 FASE DE PREPARACIÓN DETALLADA

En esta fase se planteo la arquitectura para el ciclo de vida del proyecto. En esta fase se realiza la captura de la mayor parte de los requerimientos funcionales, acumulando la información necesaria para hacer la construcción de la herramienta.

Entre los artefactos que se obtuvieron durante esta fase se encuentran, *la arquitectura de la aplicación de BI, los casos de uso en formato expandido y los casos reales de uso* para los tres módulos OLAP. La descripción de los artefactos mencionados anteriormente se presenta a continuación:

3.1.2.1 Arquitectura Preliminar de la Herramienta OLAP:

La herramienta OLAP está dividida en dos grandes módulos: Un módulo que corresponde a la herramienta de reportes estándar y un segundo módulo que corresponde a la aplicación analítica, que a su vez se divide en un módulo dinámico de consultas sobre el cubo que es almacenado en el servidor y un modulo dinámico de consultas sobre archivos de cubo que son almacenados en el equipo del usuario. Estos módulos se integran directamente a la arquitectura del Repositorio Digital SPAR 1.0 permitiendo una completa integración de los dos sistemas.

La arquitectura usada para la construcción de la herramienta OLAP es una arquitectura cliente servidor que puede observarse en la Figura 38. A continuación se describe cada uno de los componentes de la arquitectura y como ellos van interactuando para finalmente convertirse en la herramienta OLAP del Repositorio Digital SPAR 1.0.

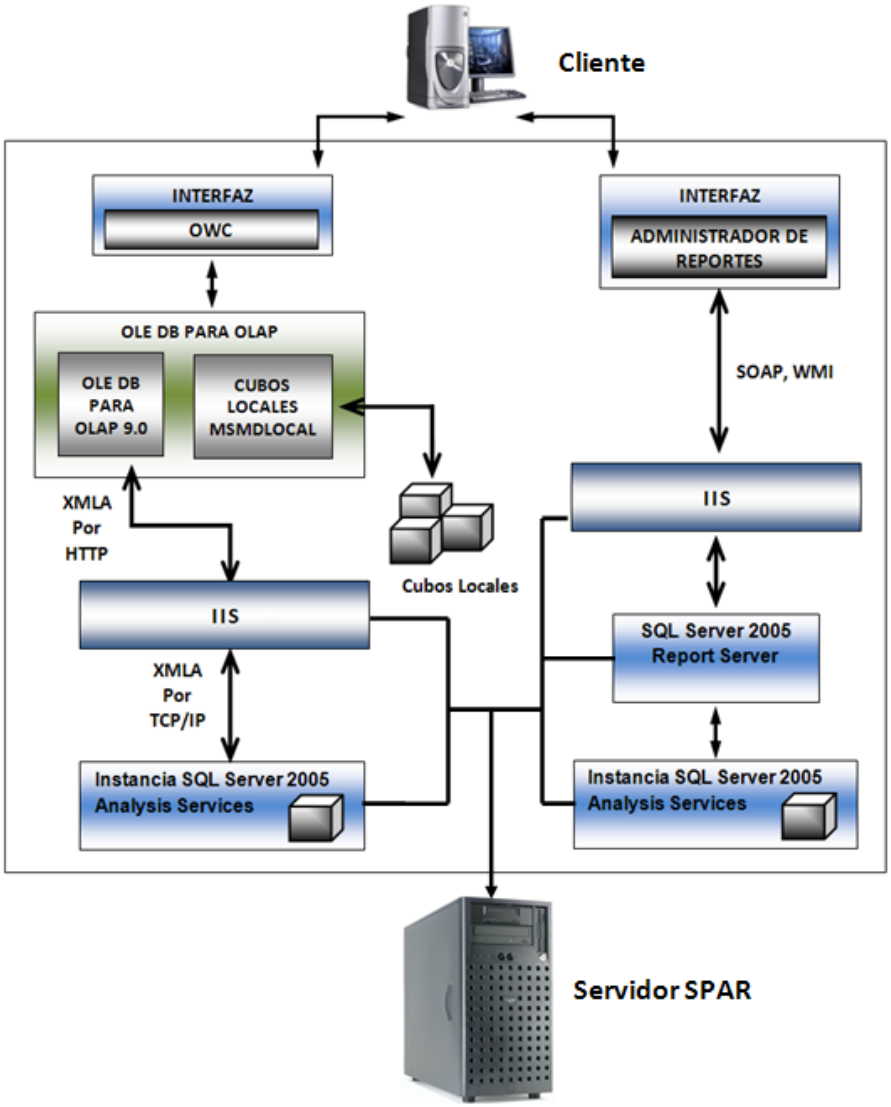


Figura 38: Arquitectura de la Herramienta OLAP de SPAR 1.0

La Figura 38 muestra la arquitectura de la herramienta dividida en dos flujos principales, un primer flujo que se origina en la interfaz que corresponde a la aplicación analítica (flujo de la izquierda en la Figura 38, el cual se describe a continuación, junto con sus componentes:

- **Interfaz Aplicación Analítica (OWC):**

La interfaz de la aplicación analítica enlazada al repositorio SPAR, integra Componentes Web de Office (OWC), que incluye una interfaz de usuario y funcionalidad semejante a la de Excel. Los OWC integrados a la aplicación analítica de SPAR fueron: **Tabla dinámica de Office y Grafico de Office** (La descripción de estos componentes se encuentran en el Capítulo II, sección: Conjunto de Herramientas). Estos dos componentes integrados a la aplicación analítica se configuraron de tal forma que obtiene datos de la base de datos multidimensional de SPAR 1.0 (SPARAS) (MOLAP). La conexión se hace directamente con el servidor de AS por medio de OLE DB para OLAP (ODBO)[2].

Como muestra la Figura 38 los **OWC** usan ODBO para conectarse al cubo SPARDW de la base de datos multidimensional OLAP (SPARAS), para conectarse con el servidor de AS el proveedor ODBO envía y recibe XMLA[3] en paquetes SOAP sobre HTTP a través de servicios de Internet Information Server (IIS). Por medio de este flujo de conexiones se permite los usuarios hacer análisis multidimensionales rápidos, consistentes e interactivos sobre cualquiera de los tres Data Mart construidos en el cubo SPARDW: Gestión, Oferta y Demanda de contenidos, Evaluación de Contenidos y Sesiones de Usuario

Otro punto importante dentro del primer flujo, representa al modulo dinámico de consultas off-line (desconectado del servidor de análisis), los OWC además de hacer una conexión al servidor de AS como se describió anteriormente, puede hacer una conexión a los archivos de cubo local, que son archivos que contienen los datos multidimensionales de los tres Data Mart del cubo SPARDW. Este tipo de conexión agiliza las consultas dimensionales por que los archivos pueden ser almacenados en el equipo local del usuario, permitiéndole hacer consultas off-line, en este caso los OWC usan el mismo protocolo ODBO para interactuar con los datos y mostrarlos en los componentes de tabla dinámica y/o grafico de office integrados en la interfaz.

El segundo flujo (flujo de la derecha en la Figura 38) se origina en la interfaz que corresponde herramienta de reportes estándares, se describe a continuación, junto con sus componentes:

[2] **OLE DB para OLAP (ODBO):** Son un conjunto de objetos e interfaces diseñadas por Microsoft que extienden las funcionalidades de OLE DB para proveer acceso y conectividad a bases de datos multidimensionales.

[3] **XMLA:** Es un protocolo de acceso de objetos (SOAP) basado en el protocolo XML. Es un estándar abierto que fue diseñado para acceder a datos de cualquier fuente multidimensional que reside en la web. XMLA es el mecanismo central de comunicación de Analysis Services.

- **Interfaz de Herramienta de Reportes Estándar:**

Como se observa en la arquitectura, la interfaz que corresponde a los reportes estándar usa una herramienta llamada el **Administrador de Reportes** que se describe en la etapa de selección de productos junto con los **Servicios de Reportes y Servidor de Reportes**.

El Administrador de Reportes fue integrado dentro del repositorio digital SPAR 1.0, y como muestra la figura, esta herramienta usa los protocolos SOAP[4] y WMI[5] para obtener acceso a través de Internet Information Services (IIS) a los informes predefinidos que han sido almacenados en el servidor de reportes (Report Server) de SQL Server 2005, antes de ser almacenados, estos reportes han sido diseñados en etapas descritas anteriormente y elaborados en la etapa de construcción de ciclo de desarrollo de la herramienta OLAP de SPAR. En el caso de la herramienta OLAP de SPAR el servidor de reportes obtiene los datos de una base de datos multidimensional de AS (SPARAS).

3.1.2.2 Casos de Uso Formato Expandido

"Un caso de uso expandido describe un proceso más a fondo que el de alto nivel. La diferencia básica con el caso de uso de alto nivel consiste en que tiene una sesión destinada al curso normal de los eventos, que los describe paso por paso" [12]. A continuación se presenta el caso de uso expandido Consultar Módulo Dinámico – Cubo Servidor, los demás casos de uso expandido para los tres módulos OLAP se describen en el **Anexo 13.**

Caso de Uso Expandido Consultar Módulo Dinámico – Cubo Servidor

Caso de uso:	Consultar Módulo Dinámico – Cubo Servidor
Actores:	Usuario Registrado, Administrador
Propósito:	Hacer consultas Dinámicas personalizadas (ad hoc) sobre el cubo almacenado en el servidor de análisis.
Resumen:	Cualquier de los dos usuarios tiene la posibilidad de ingresar dentro del modulo Dinámico OLAP y realizar consultas personalizadas sobre cualquiera de los tres modelos Dimensionales de la Bodega de Datos
Tipo:	Primario.
Curso normal de eventos	
Acción de los Actores	**Respuesta del sistema**
1. Este caso de uso se inicia cuando un visitante quiere hacer consultas personalizadas sobre el cubo multidimensional.	**2.** El sistema presenta la herramienta OLAP encargada de hacer consultas dinámicas.

[4] **SOAP:** Es un protocolo estándar creado por Microsoft, IBM y otros, está actualmente bajo el auspicio de la W3C y define cómo dos objetos en diferentes procesos pueden comunicarse por medio de intercambio de datos XML. SOAP es uno de los protocolos utilizados en los servicios Web.

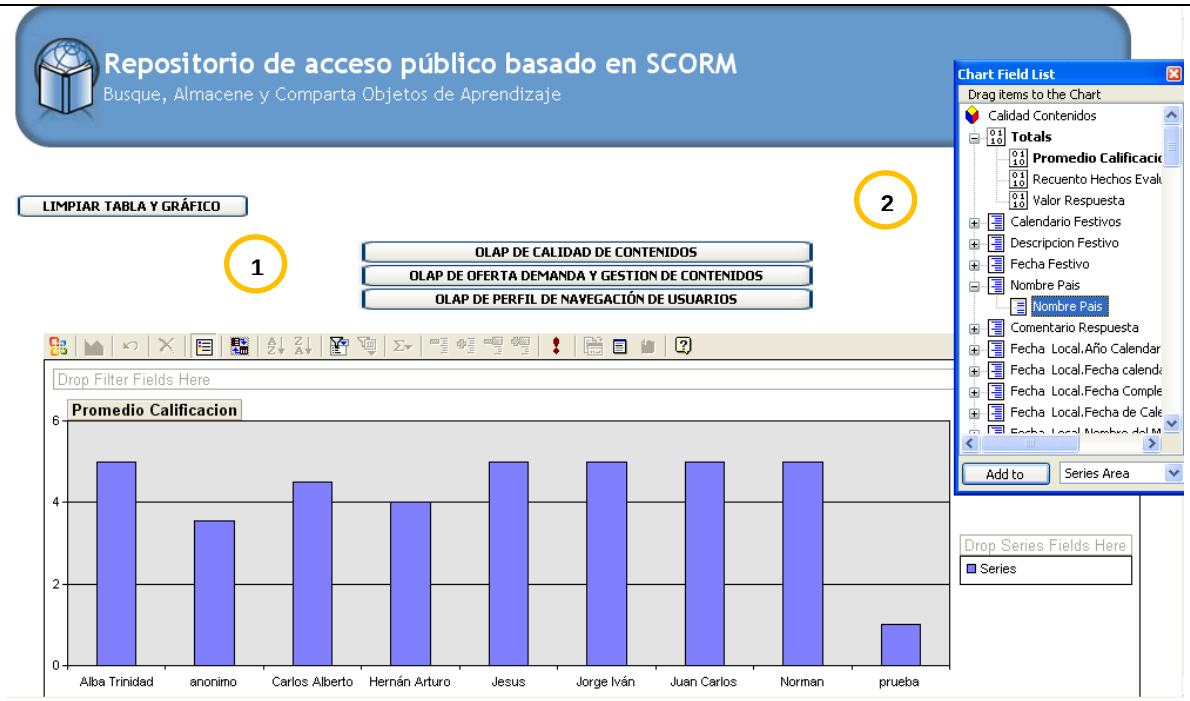
[5] **WMI:** Es la infraestructura para la gestión de datos y operaciones sobre sistema operativos Windows, es la implementación de Microsoft de Web-Based Enterprise Management (WBEM), una iniciativa del sector que pretende establecer normas estándar para tener acceso y compartir la información de administración a través de una red empresarial.

3. El usuario identifica cual perspectiva del cubo consultar y hace la solicitud. **5.** El usuario genera consultas dinámicas arrastrando y colocando dimensiones y medidas sobre los visores de tabla o gráficos.	**4.** El sistema obtiene los datos multidimensionales (Dimensiones y Medidas) de determinada perspectiva del cubo y los presenta al usuario.

3.1.2.3 Casos de Uso Reales:

"Un caso real de uso describe el diseño concreto del caso de uso a partir de una tecnología particular de entrada y salida, así como de su implementación global" [12]. A continuación se presentan el caso de uso real general de la herramienta OLAP; los demás casos de uso reales se pueden encontrar en el **Anexo 14.**

CASO DE USO REAL CONSULTAR MÓDULO DINÁMICO CUBO SERVIDOR:



Curso normal de los eventos	
Acción del Actor	**Respuesta del Sistema**
1. El usuario accede al sitio Web http://spar.unicauca.edu.co/spar/default.aspx	

2. El usuario se autentica para poder hacer uso de los servicios OLAP.	
3. El usuario entra a la sección OLAP.	
4. El usuario entra al Módulo Dinámico Cubo Servidor.	**5.** El sistema muestra vínculos a cualquiera de las tres posibles perspectivas del cubo correspondiente a cada área del negocio **[1].**
6. El usuario selecciona un área del negocio en particular para hacer consultas personalizadas de acuerdo a sus necesidades.	**7.** El sistema carga las dimensiones y hechos correspondientes a esa área del negocio (Data Mart/Modelo Dimensional) **[2].**

3.2 ETAPA DE DESARROLLO DE LAS APLICACIONES

3.2.1 FASE DE CONSTRUCCIÓN

Esta fase involucra la creación de la Herramienta OLAP, para lo cual se definieron dos ciclos de vida iterativos con el fin de cumplir con los requerimientos funcionales del sistema. Cada ciclo de desarrollo se enfoca en la construcción de cada uno de los dos módulos en los cuales se divide la funcionalidad completa de la herramienta OLAP: Reportes Estándares y Aplicación Analítica. A continuación se describen los dos ciclos.

Ciclo 1: Herramienta de Reportes Estándar:
En este ciclo se realizó la construcción de cada uno de los reportes estándar basados en la lista de reportes candidatos obtenida en la fase de preparación inicial. De la misma manera los reportes son publicados en el Servidor de Reportes de Microsoft SQL Server 2005 (Reporting Services) para posteriormente ser accedidos por medio del portal de navegación (Report Manager) que hace parte de Servidor de Reportes de Microsoft y que es integrado dentro del Repositorio Digital SPAR 1.0. La Lista de Reportes Finales con su respectiva descripción puede ser encontrada en el **Anexo 15.** La Figura 39 ilustra un ejemplo de uno de los reportes estándar que muestra el total de calificaciones y el promedio de calificaciones para determinados objetos, en determinado año.

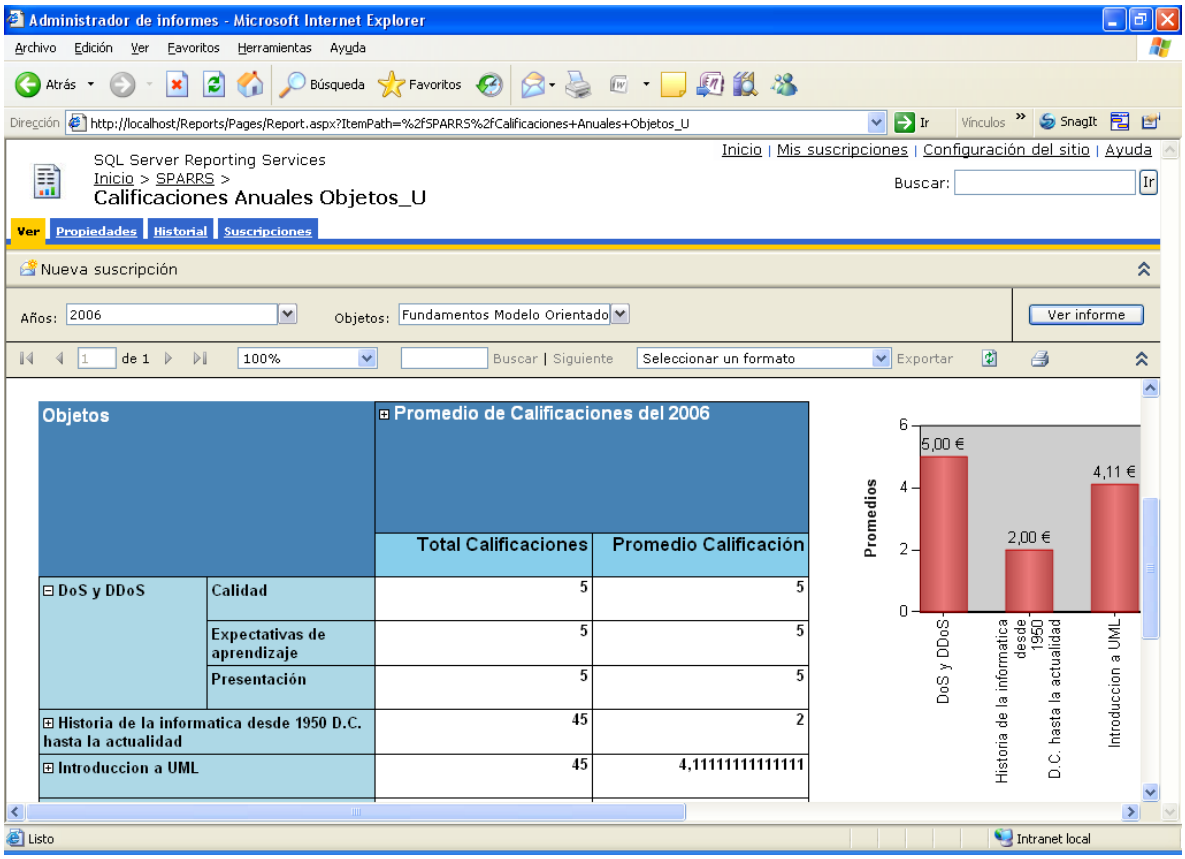


Figura 39: Reporte Estándar de Cantidad de Calificaciones y Promedios de Calificaciones para determinados objetos de aprendizaje.

Ciclo 2: Aplicación Analítica: Módulos Dinámicos

En este ciclo basado en la determinación de usar Office Web Componentes (determinación que se tomo en el Capítulo II, etapa de Selección de Productos) se realizaron las siguientes tareas:

- Conexión entre los dos controles usados: Microsoft Office Chart 11.0 y Microsoft PivotTable 11.0, para eso se construye un script de Visual Basic que se integra en el código HTML de la interfaz OLAP del Módulo Dinámico.

- Enlazar los controles al origen de datos, que en este caso es el cubo multidimensional de la bodega de datos almacenado en el Servidor de Análisis de SQL Server 2005, para eso se construye un script de Visual Basic que se integra en el código HTML de la interfaz OLAP del Módulo Dinámico.

- Integrar los Componentes al Repositorio Digital de Objetos de Aprendizaje SPAR 1.0

- Crear los cubos locales con base en el cubo multidimensional almacenado en el Servidor de Análisis de SQL Server 2005. Para esto se crean tres consultas MDX que

se ejecutan sobre el servidor de Análisis de SQL Server 2005 para poder generar los archivos de los cubos locales.

- Para el Módulo Dinámico de Cubo Local se configuran los componentes para que permitan la lectura de los archivos de cubo local, para eso se construye un script de Visual Basic que se integra en el código HTML de la interfaz OLAP del Módulo Dinámico.

Figura 40 muestra un ejemplo de una consulta realizada con el módulo dinámico de consulta que muestra los resultados en los componentes web de office.

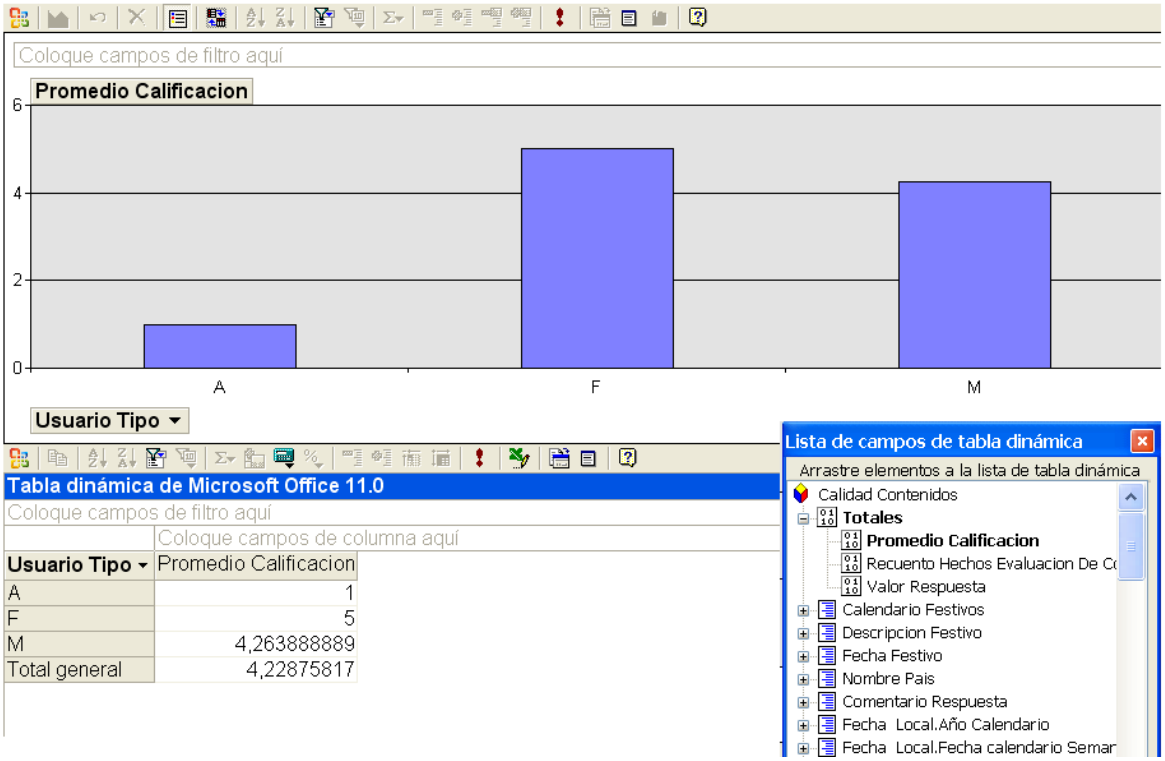


Figura 40: Consulta personalizada hecha sobre el Modulo Dinámico de consultas

3.2.2 FASE DE TRANSICIÓN:

En esta fase se entregó el producto después de haber realizado las pruebas de aceptación por un grupo especial de usuarios y efectuado los ajustes y correcciones requeridas. En esta fase se hace un despliegue al servidor de producción poniendo a disposición de los usuarios finales los reportes y los dos módulos dinámicos de consultas, desde este momento los usuarios pueden comenzar a satisfacer sus necesidades analíticas sobre los diferentes áreas de negocio del Repositorio Digital (Evaluación de Contenidos, Gestión, Oferta y Demanda de Contenidos y Sesiones de Usuarios).

CAPITULO IV

4. DESCRIPCIÓN DEL MÓDULO DE MINERÍA DE DATOS

El módulo de minería de datos de este proyecto tiene como objetivo hacer recomendaciones de los recursos de aprendizaje específicos que podrían ser particularmente útiles a un usuario dado, a tal grado, que la lista de recomendaciones personalizada es más atractiva que una lista aleatoria de recomendaciones, el visitante tiene más probabilidades de hacer una consulta o descarga. La lista de recomendaciones hechas está basada en una serie de asociaciones encontradas en los datos por el algoritmo de minería, que reflejan los intereses de aprendizaje de los usuarios.

La Metodología usada para el desarrollo de este módulo de minería está basada en el acercamiento propuesto por el Grupo Kimball en su libro "*The Microsoft Data Warehouse Toolkit*" [5], acercamiento que se soporta en tres orígenes, como son: El proceso de Minería de Datos Llamado "Ciclo Virtuoso de Minería de Datos" propuesto por Michael Berry y Gordon Linoff [10], El Proceso Industrial de Estándar Cruzado para Minería de datos (CRISP) [9] y el acercamiento presentado por Zhao Hui Tang y Jamie MacLennan, miembros claves del equipo de desarrollo de Microsoft SQL Server 2005 en su libro "*Data Mining with SQL Server 2005"* [11]. El proceso se divide en tres fases principales: *Fase del Negocio, Fase de Minería de Datos, Fase de operaciones,* y varias áreas de tareas dentro de estas fases, como muestra la Figura 41.

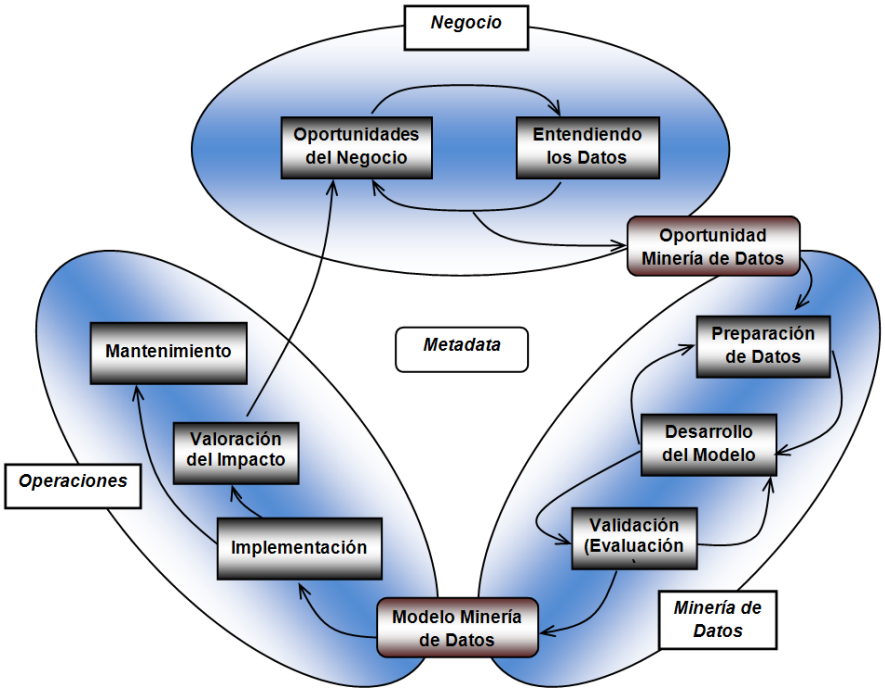


Figura 41: El proceso de Minería de Datos (Adoptada de [5])

La figura muestra que el proceso de minería de datos es un proceso iterativo como todos los procesos en el sistema de DW/BI. Los puntos de iteración más comunes se reflejan por las flechas que apuntan hacia atrás, de la misma manera existen iteraciones adicionales entre las fases principales del proceso que no son representadas en la figura pero que igualmente son muy comunes.

A continuación se hace una descripción de cada una de las fases y de sus correspondientes áreas de tareas que permiten el desarrollo del presente módulo de minería de datos.

4.1 FASE DEL NEGOCIO:

El principal objetivo en esta fase es identificar las oportunidades comerciales que pueden dar origen al desarrollo de un proyecto de minería de datos, estas oportunidades se priorizan dependiendo de su impacto y viabilidad, la cual está determinada por la existencia y disponibilidad de los datos, tal como muestran la conexión existente entre las dos tareas: Oportunidades de Negocio y Entendimiento de los datos mostrados en la Figura 41.

4.1.2 Oportunidades de Negocio:

Dentro de esta tarea se identifican y priorizan las oportunidades de negocio, se establece el objetivo del proyecto de minería de datos, se determina su impacto en el negocio y su dificultad de implementación.

La idea de incorporar minería de datos en SPAR se presentó con el planteamiento inicial de este proyecto como una oportunidad para mejorar los servicios ofrecidos por el repositorio digital. A lo largo de muchos encuentros con los directivos de SPAR se fueron identificando un rango de oportunidades para la realización de un proyecto de minería de datos sobre SPAR. Entre las que se pueden nombrar: aumentar la cantidad de usuarios en el repositorio, mejorar la utilidad del repositorio, aumentar descargas, publicaciones y consultas de objetos, conservar a los usuarios y convertir visitantes en usuarios de los servicios del Repositorio Digital.

Después de algunas discusiones con el equipo administrativo de SPAR se estableció que algunas de las necesidades actuales del Repositorio se podrían satisfacer con la posibilidad de influir en el comportamiento transaccional (descargas y consultas) de los usuarios haciendo recomendaciones pertinentes de recursos educativos de interés particular a cada uno de ellos. Por tanto se decidió hacer recomendaciones en determinadas secciones del sitio web de SPAR, manteniendo una lista de cuatro recursos educativos recomendados a los intereses de cada visitante.

Esto implicó la modificación de dos interfaces del sitio web del Repositorio: descargar.aspx y consultar.aspx, además la creación de varios servicios web que permitieran hacer consultas directas al modelo de minería de datos almacenado en el servidor de análisis de Microsoft (AS) y retornan los resultados para posteriormente ser visualizados como recomendaciones en las dos interfaces del Repositorio SPAR 1.0 en tiempo real que ocurre la interacción del usuario con el sitio web.

A continuación se procedió a analizar la disponibilidad de datos necesarios para continuar con el desarrollo del proyecto.

4.1.3 La Comprensión de los Datos

Esta tarea involucra hacer una exploración que permita determinar si los datos están limpios y disponibles para dar soporte a las oportunidades comerciales de mayor prioridad identificadas anteriormente.

El sistema de DW/BI tiene todo el historial de transacciones realizadas por cada usuario a nivel de recurso educativo individual, esta información está almacenada en el DW del repositorio y hace parte del Data Mart de Gestión, Oferta y Demanda de Contenidos, por lo tanto la información puede ser encontrada en la Tabla de Hechos Transacciones de Usuario y en la Dimensión Usuario y la Dimensión Objetos. De la misma manera, se identificó que el sistema dispone también de cierta información demográfica del usuario almacenada en la Dimensión Usuario. Por lo tanto se determinó que existían datos disponibles para crear un modelo útil para hacer las recomendaciones de los recursos de aprendizaje. A causa de que no se tienen muchos de los datos demográficos de los usuarios almacenados en el sistema transaccional de SPAR 1.0, el modelo estaría basado en los recursos de interés como los indicados por los recursos consultados o descargados con anterioridad.

Teniendo en cuenta lo anterior, se definió como *el objetivo global* mejorar la utilidad del repositorio, satisfaciendo en gran medida los intereses temáticos de los usuarios. De la misma manera, aumentar la probabilidad de incrementar usuarios y de conservar a los actuales. *La estrategia* es hacer recomendaciones de recursos educativos que tienen una alta probabilidad de interesar a cada visitante del repositorio SPAR. Esta estrategia se soporta sobre un modelo de minería de datos basado en las consultas y descargas de recursos de aprendizaje.

4.2 LA FASE DE MINERÍA DE DATOS

Esta sección continúa con la preparación de los datos, el desarrollo del modelo de minería y su validación, estas tres tareas se retroalimentan entre sí para lograr un desarrollo iterativo.

4.2.1 La Preparación de los datos

Esta tarea involucra la limpieza y transformación de los datos para crear el conjunto de escenarios de minería de datos que dan soporte al modelo. El origen de datos para el modelo de minería de datos es el DW de SPAR 1.0, porque este dispone de datos limpios y transformados, lo que facilita y optimiza este proceso.

El objetivo es relacionar la información de los usuarios con la información de los recursos educativos consultados o descargados, lo que implica la creación de un conjunto de escenarios del usuario y un conjunto de escenarios anidado de transacciones. Por tanto, después de identificar el origen de datos se hace una exploración que involucra consultas y reportes que examinan detalladamente las tablas de origen. Con esta exploración se decidió sacar los datos del escenario de usuarios de la tabla “DimUsuario” y sacar los datos del

escenario anidado de transacciones de la tabla "HechosOfertaDemandaContenidos" y de la tabla DimObjetos. Cada fila en el conjunto de escenarios de transacciones tiene la clave del usuario y el nombre recurso educativo. Cada fila del conjunto de escenarios del usuario tiene relación uno a muchos con el conjunto de escenarios de transacciones.

Para poder obtener los datos para el conjunto de escenarios de usuarios y el conjunto de escenarios anidado de transacciones se construyó un paquete de Integration Services (IS) que obtiene determinados datos de los usuarios de la dimensión usuario y los almacena en una tabla llamada EscDatosUsuario y al mismo tiempo obtiene los datos de las transacciones de los clientes y los almacena en una tabla llamada EscDatosTransacciones. La Figura 42 muestra el flujo de datos que permite la creación del conjunto de escenarios de usuarios.

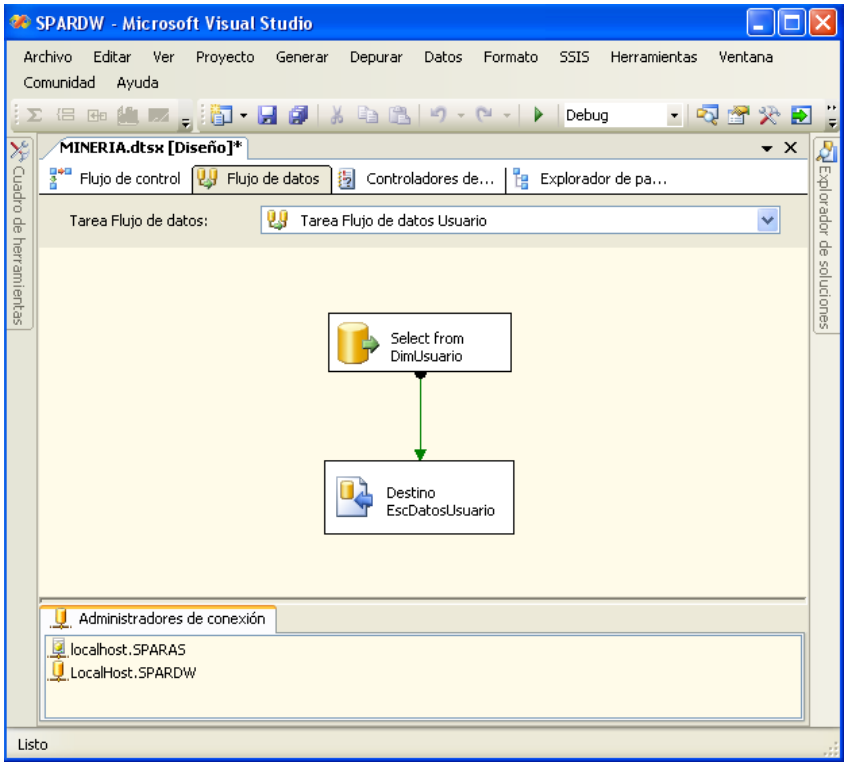


Figura 42: Flujo de datos de un paquete de Integration Services que crea el escenario de usuarios

La Figura 43 muestra una vista de los dos conjuntos de escenarios que dan soporte a la construcción del modelo de minería de recomendaciones de objetos de aprendizaje.

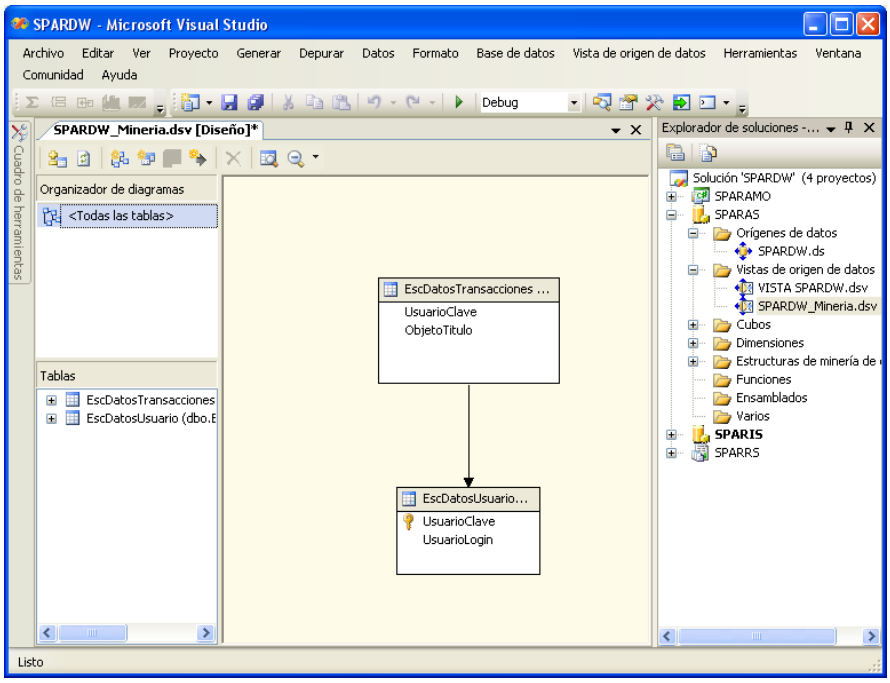


Figura 43: Conjunto de escenarios de usuario y conjunto de escenarios anidado de transacciones

4.2.2 Desarrollo del Modelo

El proceso de desarrollo del modelo involucra una serie de pasos que se realizan directamente sobre la herramienta de diseño y construcción de bases de datos multidimensionales de Microsoft el ***Business Intelligence Development Studio.***

Antes de comenzar con el proceso de desarrollo del modelo de minería de datos, es necesario hacer la selección del algoritmo de los proporcionados por Microsoft que permitan satisfacer el objetivo global del proyecto de minería de datos, que es: mejorar la utilidad del repositorio, satisfaciendo en gran medida los intereses temáticos de los usuarios.

La estrategia identificada en las primeras etapas del proceso de desarrollo para satisfacer este objetivo, consiste en hacer recomendaciones de recursos educativos que tienen una alta probabilidad de interesar a cada visitante del repositorio SPAR. Lo que reduce el conjunto de posibilidades de selección del algoritmo de minería de datos, a aquellos que permiten hacer recomendaciones basadas en conjuntos. Del conjunto de algoritmos ofrecidos por Microsoft solo dos algoritmos satisfacen esta tarea de negocio, estos son:

- Algoritmo de Reglas De Asociación de Microsoft (Microsoft Association Rules)
- Algoritmo de Arboles de Decisión (Microsoft Decision Trees).

El siguiente paso consiste en hacer un estudio de las características de cada uno de estos algoritmos que se muestran a continuación:

4.2.2.1 Reglas de Asociación de Microsoft (Microsoft Association Rules)

Este algoritmo da soporte a la tarea de negocio descrita como asociación, afinidad de grupos o análisis de cesta de mercado. Microsoft Association Rules es uno de los algoritmos proporcionados por Microsoft SQL Server 2005 Analysis Services (SSAS), es útil para hacer recomendaciones basadas en la identificación de un conjunto de correlaciones existentes en los datos almacenados en las bases de datos de una organización. En el caso de cesta de compra el algoritmo recomienda productos a los clientes basado en productos que ya han adquirido o tienen interés [10].

Microsoft Association Rules es un algoritmo que pertenece a la familia de algoritmos "A Priori", que son algoritmos muy comunes y eficientes para encontrar conjuntos frecuentes de elementos con atributos comunes. El algoritmo de Reglas de Asociación de Microsoft trabaja de la siguiente manera: El primer paso es recorrer y explorar un conjunto de datos para hallar los elementos que aparecen juntos en un escenario (evento) tal como una consulta de un recurso educativo, a continuación, crea conjuntos de elementos asociados que aparecen juntos, como mínimo, un determinado número de veces dentro del conjunto de datos inicial. Este valor mínimo de ocurrencias de los conjuntos de elementos asociados, es un valor que se predefine como parámetro para el algoritmo y recibe el nombre de ***Soporte Mínimo***, valor que permite generar reglas de asociación para los conjuntos en el que el número de ocurrencias sea mayor o igual al soporte mínimo establecido. El segundo paso consiste en generar las reglas de asociación a partir de los conjuntos de elementos frecuentes. Estas reglas describen cómo estos elementos se agrupan dentro de los conjuntos y pueden utilizarse por ejemplo, para predecir las probables compras de un cliente en el futuro. Este paso de generación de reglas usa menos tiempo porque requiere menos cantidad de iteraciones que el primer paso
[11]. La Figura 44 ilustra los dos pasos de procesamiento del algoritmo de Reglas de Asociación de Microsoft.

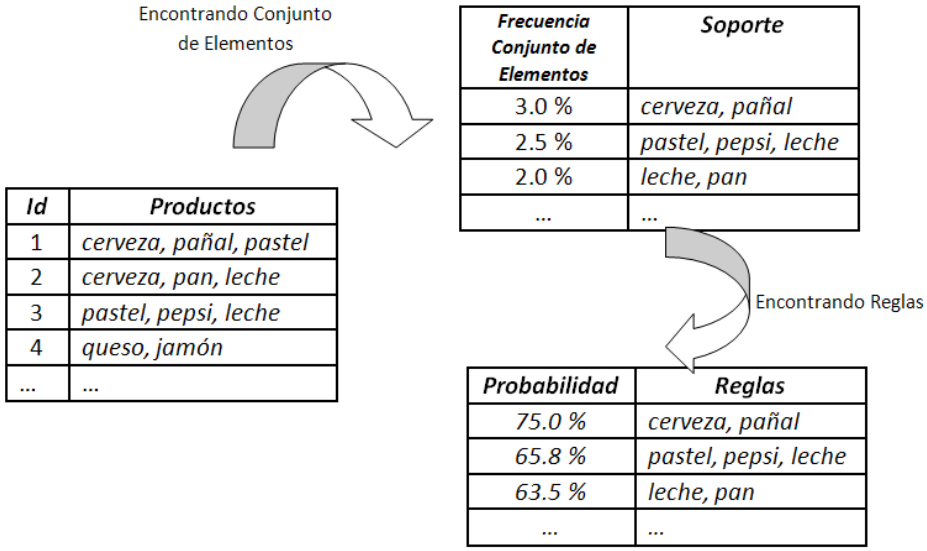


Figura 44: El Proceso de dos pasos del Algoritmo de Asociación (*Adoptada [11]*)

Conceptos Básicos del algoritmo de asociación de Microsoft:

A continuación se describe algunos conceptos que permiten comprender los principios básicos del algoritmo de asociación de Microsoft [11]:

- **Conjunto de Elementos (Itemset):**

Son grupos de elementos que componen los modelos de asociación. Cada conjunto de elementos tiene un tamaño, el cual es el número de elementos contenidos en el conjunto. Por ejemplo, el tamaño del conjunto de elementos de {arroz, aceite, azúcar} es 3.
Conjunto de elementos Frecuentes: son conjuntos de elementos que son comunes y populares dentro de un conjunto de datos (conjunto de escenarios). El límite de popularidad u ocurrencia para un conjunto de elementos se define como el soporte como se describió anteriormente y que será explicado en detalle a continuación.

- **Soporte:**

El soporte se define como el número de ocurrencias de un conjunto de elementos dentro de un conjunto de datos. El soporte de un conjunto de elementos {A, B} se compone del total de transacciones que contienen A y B. El *Soporte Mínimo (Minimum_Support)* es un parámetro límite que se necesita definir antes de procesar el algoritmo de asociación. Este valor significa que se está interesado solo en el conjunto de elementos que cumplen al menos con el soporte mínimo dentro de un conjunto de datos. Para el algoritmo de Reglas de Asociación de Microsoft este parámetro puede especificarse como un porcentaje o como un valor absoluto.

- **Probabilidad (Confianza):**

La probabilidad se define como el grado de posibilidad de que se produzca un suceso. En este caso la probabilidad es una propiedad de una regla de asociación que ha sido generada a partir de los conjuntos descubiertos de elementos asociados. Una regla de asociación tiene la forma de A→B donde A es el antecedente y B el consecuente. La probabilidad de una regla A->B es calculada usando el soporte del conjunto de elementos {A, B} divido por el soporte de {A}. Esta probabilidad también es llamada ***Confianza.*** La probabilidad se define así:

$$\text{Probabilidad } (A \rightarrow B) = \text{Probabilidad } (B|A) = \text{Soporte } (A, B) / \text{Soporte } (A)$$

Probabilidad Mínima (Minimum_Probability): La Probabilidad Mínima es un parámetro que se necesita especificar antes de procesar el algoritmo. Especifica la probabilidad mínima de que se cumpla una regla. Por ejemplo, si se establece este valor en 0.5, se determina que no se genera ninguna regla con menos del 50% de probabilidad.

- **Importancia:**

La importancia es una propiedad que puede aplicarse a conjunto de elementos o reglas.

La importancia para el conjunto de elementos se define usando la siguiente fórmula:

Importancia ({A, B}) = Probabilidad (A, B) / (Probabilidad (A)* Probabilidad (B))

Si la importancia = 1, A y B son elementos independientes. En el ejemplo de cesta de mercado esto significa que la compra del producto A y la compra del producto B son dos eventos independientes. Si la importancia <1, A y B están negativamente correlacionados. Esto significa que si un cliente compra A, no es muy probable que él también compre B. Si la importancia> 1, A y B están positivamente correlacionados. Esto significa que si un cliente compra A, muy probablemente él también compra B.
La importancia aplicada a las reglas se calcula usando la siguiente fórmula:

Importancia (*A* → *B*) = log (*p* (*B*|*A*) /*p* (*B*|not *A*))

Si la importancia = 0 significa que no hay ninguna asociación entre A y B. Una valor positivo de importancia significa que la probabilidad de B sube cuando A ocurre. Un valor negativo de importancia significa que la probabilidad de B baja cuando A ocurre.

4.2.2.2 Árboles de Decisión (Decision Trees)

Los árboles de decisión son usados para las tareas de negocio de clasificación, estimación, predicción y análisis de asociación. La idea principal del algoritmo es generar una serie de divisiones (nodos) basado en las relaciones existentes entre las variables de entrada y la variable objetivo (variable o atributo de predicción). Para la generación del árbol, el algoritmo identifica las columnas de entrada que se correlacionan de forma significativa con la columna de predicción y crea el nodo correspondiente por cada columna de influencia. Por ejemplo en un escenario para predecir la consulta de recursos educativos, si nueve de diez usuarios jóvenes consultan un recurso educativo, pero solo lo hacen dos de diez usuarios de edad mayor, el algoritmo infiere que le edad es un elemento relevante en la consulta de recursos educativos. La forma en que el algoritmo determina la influencia de una columna de entrada varía en función de si se predice una columna continua o una columna discreta. Por esta razón, el algoritmo de arboles de decisión de Microsoft es un algoritmo híbrido, que soporta clasificación y regresión, es decir, puede hacer modelados de predicción de atributos discretos y continuos. Para atributos continuos, el algoritmo usa la regresión lineal para determinar donde se divide un árbol de decisión. Para atributos discretos el algoritmo usa los valores, o estados de las columnas de entrada para predecir los estados de la columna de predicción. Este algoritmo también tiene una característica especial y es que puede ser aplicado para análisis de asociación, como se explica en el siguiente ítem.
Los árboles de decisión ofrecen muchas ventajas sobre otros algoritmos de minería de datos, por ejemplo los árboles de decisión son fáciles de construir y de interpretar. Cada camino de la raíz a una hoja forma una regla [11].

Usando árboles de decisión para análisis de asociación:

Cuando se usa el algoritmo de arboles de decisión para análisis de asociación con el fin de hacer recomendaciones, el algoritmo construye un bosque de arboles, es decir, el algoritmo construye un árbol para cada atributo de recomendación. Por ejemplo un modelo de recomendaciones de productos de canasta familiar genera un árbol por cada producto que aparece en el conjunto de escenarios usado para el entrenamiento del modelo, en este caso por cada producto que aparece en la tabla compras. Esto implica que un algoritmo de árboles de decisión para análisis de asociación puede utilizar mucho tiempo y recursos cuando hay muchos elementos en los conjuntos de datos de entrenamiento, pero puede igualmente producir resultados interesantes si el número de elementos es limitado. De la misma manera existen otras limitantes de este algoritmo cuando es usado para tareas de asociación y es que el número máximo de arboles es de 255. Otro problema es que este algoritmo no regresa un conjunto de elementos y reglas como lo hace un algoritmo de asociación común, el usuario tiene que entender las relaciones usando un visor de contenidos [11].

4.2.2.3 Criterios de Selección del Algoritmo de Minería de Datos.

Después de conocer las características y modos de funcionamiento de cada uno de los dos algoritmos, se definieron unos criterios o consideraciones que hay que tener en cuenta para hacer la selección correcta, la definición de estos criterios se baso en algunos estudios hechos en la Universidad de Granada [13]

- **Recomendaciones basadas en conjuntos:** El algoritmo debe permitir hacer recomendaciones o sugerencias de recursos de aprendizaje a los usuarios, basados en recursos que ellos ya han consultado o descargado, sin importar el orden en que ocurrieron estos eventos.

- **Rendimiento Computacional:** El algoritmo debe usar la menor cantidad de recursos computacionales para realizar sus procesamientos y poder generar la lista de recomendaciones.

- **Tiempo de Ejecución:** El algoritmo debe producir resultados basado en las asociaciones encontradas en el menor tiempo posible.

- **Despliegue de Resultados:** El algoritmo debe permitir al usuario conocer las asociaciones encontradas entre los elementos e identificar fácilmente el conjunto de reglas generadas por el algoritmo.

- **Menor cantidad de acceso a la fuente de datos:** El algoritmo debe hacer la menor cantidad de accesos a la base de datos para encontrar los conjuntos de elementos asociados.

Tabla Comparativa de Algoritmos

A continuación se presenta una tabla comparativa que permite contrastar los algoritmos con los criterios, para finalmente hacer la selección del algoritmo. La asignación de valores comparativos se hace de acuerdo a la información obtenida que describe las características y modos de funcionamiento de cada uno de los algoritmos descrita anteriormente.

CRITERIOS	Recomendaciones		Rendimiento		Tiempo de Ejecución		Despliegue		Acceso a Datos	
ALGORITMO	Si	No	Mayor	Menor	Mayor	Menor	Si	No	Mayor Cantidad	Menor Cantidad
Reglas de Asociación	**X**		**X**			**X**	**X**			**X**
Árboles de Decisión	**X**			**X**	**X**			**X**	**X**	

Tabla 4: Tablas Comparativa entre el Algoritmo de Arboles de Decisión y Algoritmo de Reglas de Asociación, en contraste con los criterios de selección.

Con base en las tabla anterior se puede observar que el algoritmo de Reglas de Asociación tiene mayor rendimiento que el algoritmo de Árboles de Decisión, dado a que el algoritmo de Árboles de Decisión genera un árbol por cada recomendación, lo que implica el uso de una cantidad mayor de recursos en comparación con el algoritmo de Reglas de Asociación, de la misma manera el algoritmo de Árboles de Decisión gasta una mayor cantidad de tiempo en la construcción de cada árbol de recomendación en el caso de que existan muchos elementos en el conjunto de entrenamiento, además como se describió anteriormente este algoritmo genera un número máximo de 255 árboles, lo que no sería útil en el caso del repositorio que actualmente tiene más de 255 recursos de aprendizaje. Por último, el algoritmo de arboles de decisión no cumple con el criterio de despliegue de resultados que es muy útil para examinar las asociaciones encontradas y las reglas generadas. Después de haber hecho la selección del algoritmo se continúa con los pasos que involucra el desarrollo del modelo de minería de datos, estos pasos se describen a continuación:

Paso 1: Se crea una nueva vista de origen de datos en el proyecto SPARAS (Base de Datos Multidimensional de SPAR) de Analysis Services (AS) llamada SPARDW_Mineria.dsv que incluye las dos tablas creadas en el paquete de Integration Services: EscDatosUsuario y EscDatosTransacciones. El proyecto de AS y la vista de origen de datos se observa en la Figura 43.

Paso 2: Se crea una nueva estructura de minería, este paso involucra:

- Seleccionar el tipo de origen de datos: DW Relacional
- Seleccionar la técnica de minería: Reglas de Asociación de Microsoft

- Especificar cuáles son las tablas de escenarios: EscDatosUsuario y EscDatosTransacciones (Escenario anidado).

- Especificar los datos de entrenamiento, que son las columnas de clave, de entrada y de predicción:

    - Columnas Claves: UsuarioClave (columna clave del escenario de usuarios), Objeto Titulo (columna clave para el escenario anidado de transacciones).

    - Columnas de Entrada: Objeto título

    - Columnas de Predicción: Objeto título (contiene los títulos de los objetos a recomendar)

- Establecer un nombre para la estructura de minería de datos: EstRecObjetos

- Establecer un nombre para el modelo de minería de datos: RecObjetos-RA
- Ajustar los parámetros de procesamiento del algoritmo de reglas de asociación de Microsoft, esto son:
    - ***Minimum_Support* =** 0.03: Esto significa que el porcentaje mínimo de escenarios que deben contener los conjuntos de recursos antes de que se genere una regla es del tres por ciento del total escenarios.
    - ***Minimum_Probability*** = 0.3: Esto significa que no se genera ninguna regla con menos del treinta por ciento de probabilidad.

  Nota: El valor de estos parámetros se establece después de algunas discusiones con el equipo administrativo del proyecto, considerando el número actual de usuarios y transacciones hechas sobre el Repositorio SPAR y las características del algoritmo de reglas de asociación de Microsoft.

- Desplegar y procesar el modelo de minería de datos: En este paso se hace el entrenamiento del modelo, lo que le permite identificar conjuntos de recursos relevantes que aparecen juntos en las transacciones de los clientes y establecer las reglas de asociación que determinan la probabilidad de consulta o descarga de recursos basado en estas asociaciones.

4.2.3 Validación del Modelo

Esta fase involucra hacer una revisión del modelo propuesto, la cual se hace con los usuarios administrativos para examinar sus resultados, rendimiento y evaluar su impacto.

La validación y evaluación del impacto del modelo sólo se puede hacer hasta después de haberlo puesto en producción, para poder confrontarlo con el objetivo original del proyecto, sin embargo, la herramienta de diseño y construcción permitió al equipo administrativo de

SPAR y al equipo de desarrollo examinar sus resultados haciendo una prueba de racionalidad de los conjuntos de recursos asociados encontrados y reglas de asociación generadas por el modelo de recomendaciones.

A continuación se muestran y se describen los tres visores que hacen parte de la herramienta que permitieron evaluar los resultados:

El visor de asociación de Microsoft contiene tres fichas: **Conjuntos de elementos**, **Reglas** y **Red de dependencias**

- Ficha Conjuntos de elementos

La ficha Conjuntos de Elementos muestra tres extractos de información importantes que se relacionan con los conjuntos de elementos que el algoritmo de asociación de Microsoft detecta: *el soporte*, que es el número de transacciones en las que tiene lugar el conjunto de elementos; *el tamaño*, que es el número de elementos incluidos en el conjunto; y la composición real del conjunto de elementos. Dependiendo de cómo se configuren los parámetros del algoritmo, éste puede generar un número elevado de conjuntos de elementos.

Todos los conjuntos de elementos que muestra el visor contienen información sobre las transacciones hechas por los usuarios. Por ejemplo, el conjunto de elementos que contiene el valor 5 en la columna Soporte indica que, de todas las transacciones, 5 personas que consultaron Representación de la información, Fundamentos Modelo Orientado a Objetos, también consultaron DoS y DDoS. La Figura 45 ilustra la apariencia de los conjuntos de elementos encontrados por el algoritmo:

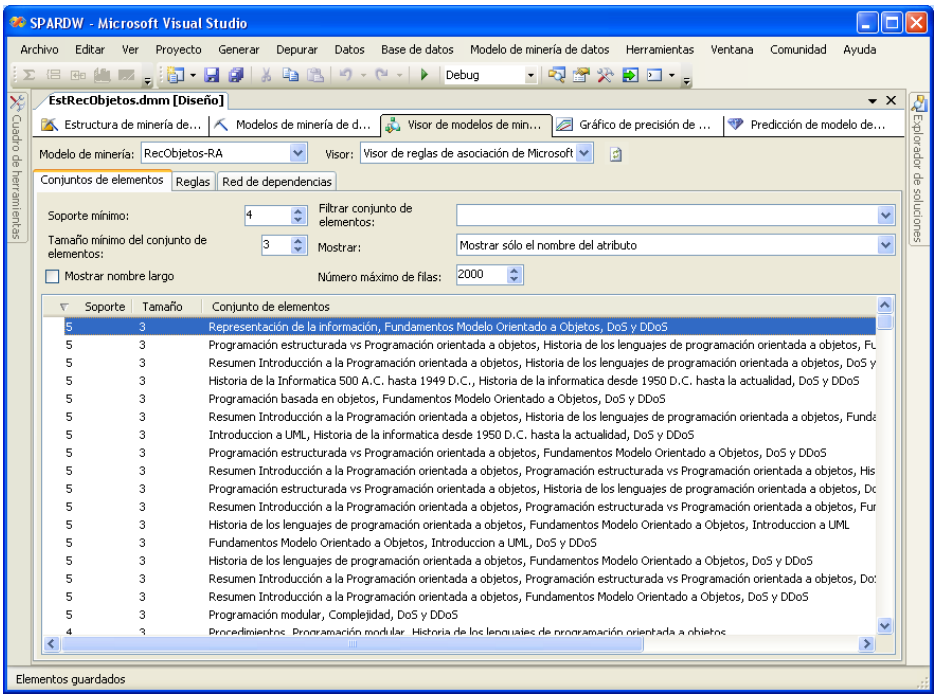


Figura 45: Visor que muestra el conjunto de recursos asociados encontrados por el modelo de minería de datos.

- Ficha Reglas

La ficha Reglas muestra la siguiente información relacionada con las reglas que el algoritmo encuentra. **Probabilidad:** Posibilidad de que se produzca una regla, **Importancia:** Mide la utilidad de una regla; un valor elevado significa que la regla es mejor. **Regla:** Definición de la regla.

Cada regla puede utilizarse para predecir la presencia de un elemento de una transacción en función de la presencia de otros elementos. Una ejemplo de una regla generada por el modelo es: (Fundamentos Modelo Orientado a Objetos = Existing, DoS y DDoS = Existing -> Tipos Abstractos de Datos = Existing), esta regla dice que cuando alguien consulta o descarga “Fundamentos Modelo Orientado a Objetos” y “DoS y DDoS”, hay una probabilidad de 0,5 de que esta persona consulte o descargue “Tipos Abstractos de Datos”. La Figura 46 ilustra la apariencia de los conjuntos de reglas encontradas por el algoritmo:

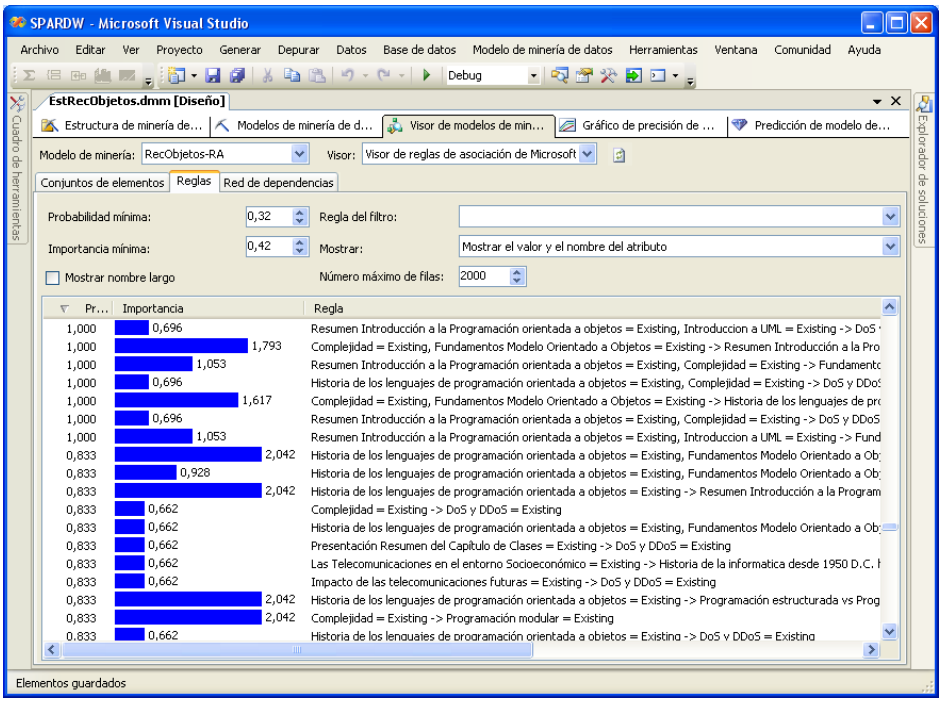


Figura 46: Visor que muestra el conjunto de reglas generadas por el modelo de minería de datos.

- Ficha Red de dependencias

Mediante la ficha Red de dependencias, se puede examinar la interacción entre los diferentes elementos del modelo. Cada nodo en la red representa una de las variables o Recursos de Aprendizaje (ObjetoTitulos) en el modelo de minería. Al seleccionar un nodo, puede utilizar la leyenda de color de la parte inferior de la ficha para establecer los elementos que determinan o son determinados por otros elementos del modelo. La Figura 47 ilustra la apariencia de la Red de Dependencias encontradas por el algoritmo:

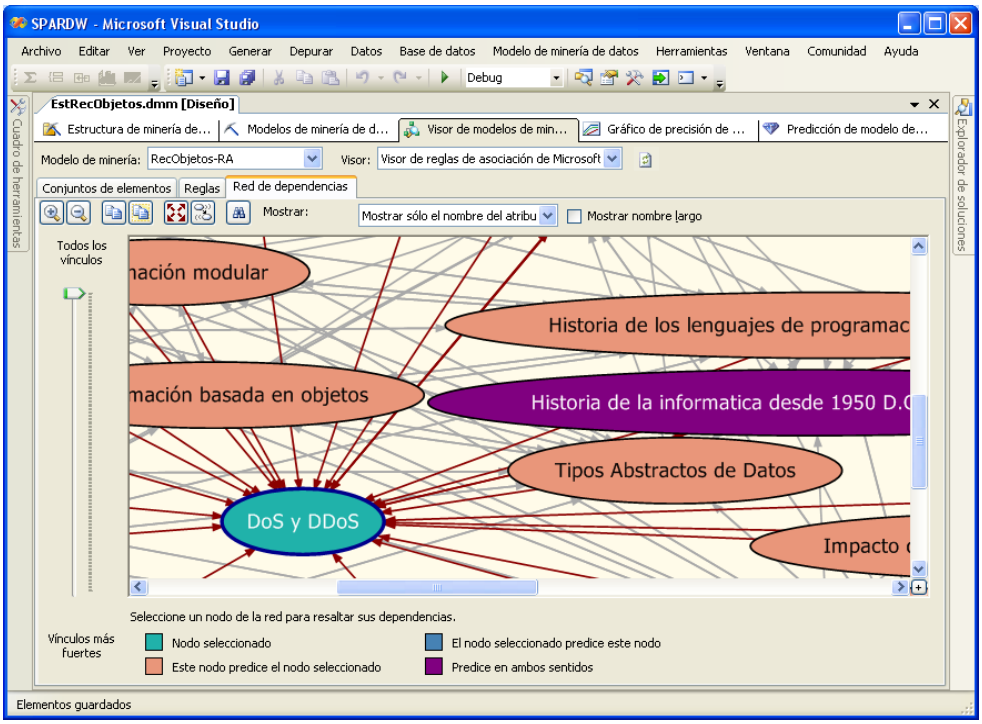


Figura 47: Visor que muestra las relaciones existentes de los recursos encontradas por el modelo.

Después de haber hecho un estudio de los resultados obtenidos, trabajando a través de varias iteraciones ajustando los parámetros del propio algoritmo. El próximo paso en el proceso es la Fase de Operaciones.

4.3 FASE DE OPERACIONES

Esta fase involucra poner el modelo en producción, ver qué impacto tiene y definir su mantenimiento. Esta fase involucra las siguientes tareas:

4.3.1 Implementación:

A continuación se presentan los detalles de implementación del modelo.

- Poner el modelo a disposición del servidor web:

Para poner el modelo a disposición del Repositorio Digital SPAR, se crea dentro del Web Services de SPAR un Método Web que permite obtener los recursos consultados o descargados anteriormente por determinado usuario (el que realiza la transacción). Después de obtener el historial de consultas o descargas se genera una consulta DMX (Data Mining Extension Language) con el Titulo del Objeto actual sobre el cual se está realizando la transacción, junto con los objetos que hacen parte del historial del usuario. Para esto se crean una conexión de tipo ADOMD.NET hacia el servidor de Analysis Services y se envía la consulta DMX, que retorna una lista de 4 objetos recomendados en orden de más alta probabilidad. Un ejemplo de una consulta DMX que genera el Web Services es:

```
SELECT flattened  PREDICT([RecObjetos-RA].[Esc Datos Transacciones],INCLUDE_STATISTICS,4)
AS Recommendation
From [RecObjetos-RA] NATURAL PREDICTION JOIN
(SELECT( SELECT 'Fundamentos Modelo Orientado a Objetos' AS [Objeto Titulo]
UNION  SELECT 'Definición de la Idea de la Investigación' AS [Objeto Titulo]
UNION  SELECT '¿Por qué fallan los equipos?' AS [Objeto Titulo]
UNION  SELECT 'Modelo de Investigación Documental' AS [Objeto Titulo]) as [Datos Transacciones])
AS Consulta
```

En este caso, un usuario que ha comenzado sesión en SPAR realiza una consulta sobre el Recurso educativo “Modelo de Investigación Documental” , el servidor genera la consulta y le muestra una lista de recomendaciones de Titulos de Objetos que pueden ser de su interes.

- Anunciar la Lista de Recomendaciones:

Desde el momento que se carga la ventana de visualización del Recurso Educativo que ha sido de interes del usuario, se presentan los resultados de recomendaciones en la parte superior de la ventana. La Figura 48 muestra como se verian las recomendaciones en SPAR.

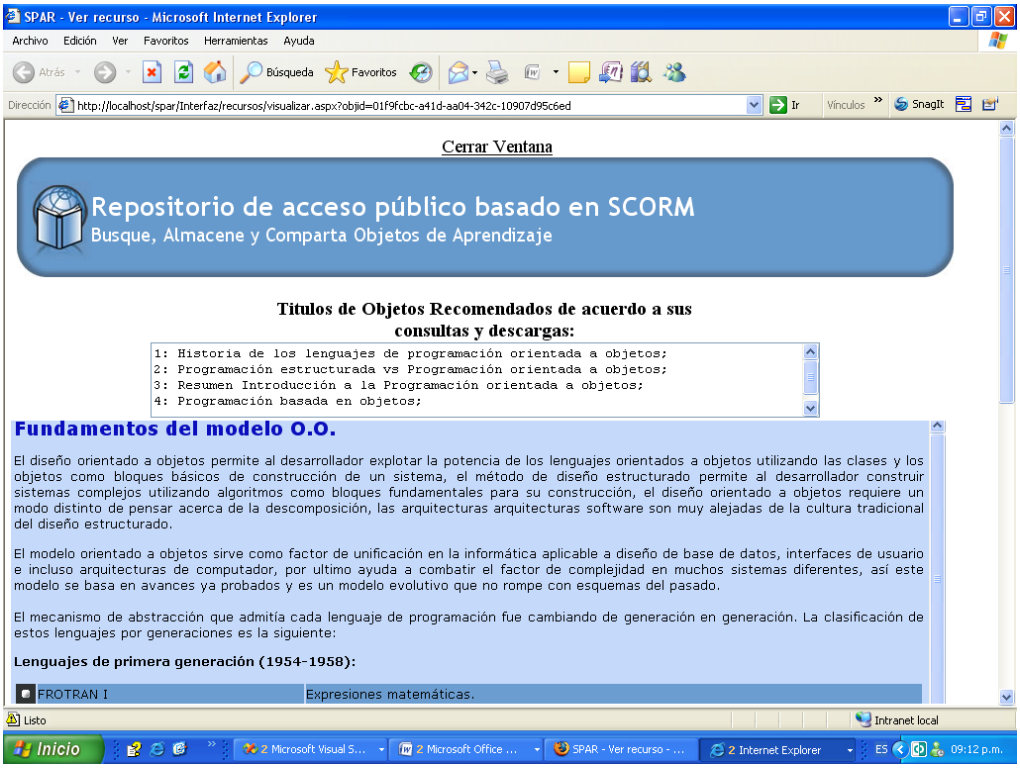


Figura 48: Recomendaciones hechas basadas en determinada consulta

En este ejemplo, para un usuario que consulta “fundamentos del modelo O.O”, las recomendaciones del modelo incluyen “Historia de los lenguajes de Programación orientada a objetos”, “Programación estructurada vs Programación orientada a objetos”, “Programación basada en objetos”, “Resumen Introducción a la Programación orientada a objetos”. Es de notar la fuerte relación taxonómica de los objetos recomendados con el objeto visualizado actualmente.

En el caso de las descargas de Recursos educativos, se anuncia de la misma manera el conjunto de recomendaciones en la parte superior de la ventana donde se hace la descarga directa del recurso actual, como muestra la Figura 49.

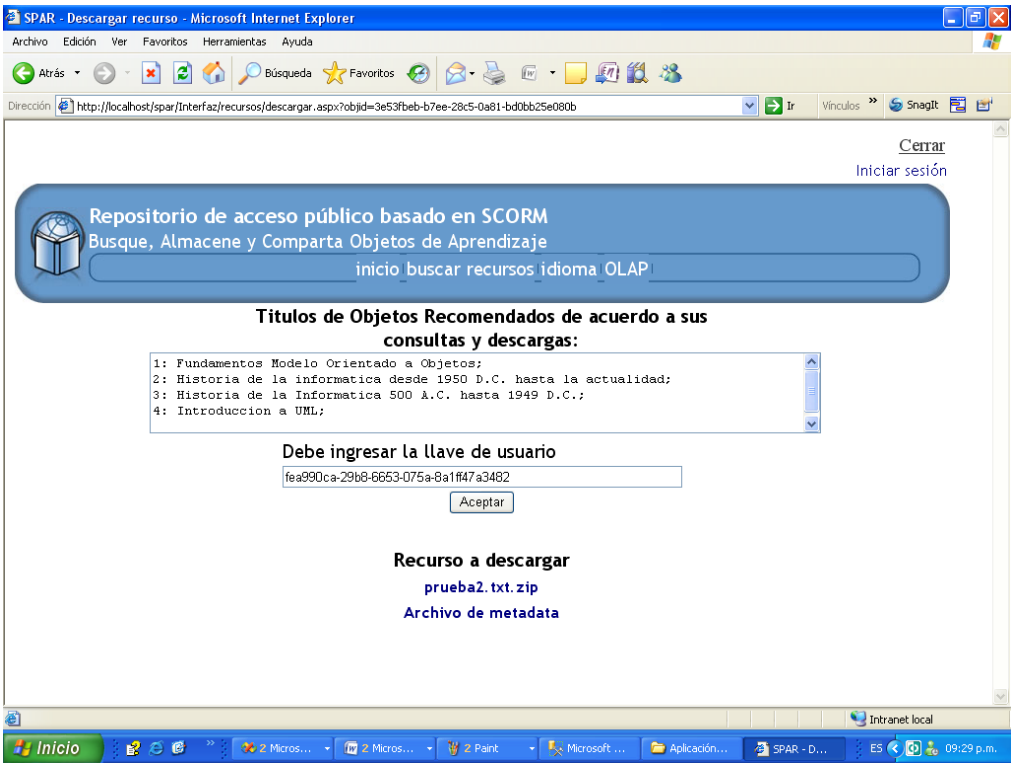


Figura 49: Recomendaciones hechas basadas en determinada descarga.

4.3.2 Evaluación del Impacto:

La evaluación del impacto se hace después de que el modelo tiene determinado tiempo de haber estado en producción, por tanto no es posible determinar el impacto actualmente. Sin embargo, se dejan establecidos los análisis que se harán con este fin: Primero, se examinará el número promedio de usuarios que realizan transacciones de consultas y descargas antes y después de la introducción de la lista de recomendación. Segundo: Se examinará el cambio en el valor promedio de la cantidad de transacciones hechas por los usuarios antes y después de la puesta en producción del modelo de recomendaciones. El aumento de los porcentajes determina un impacto positivo del modelo.

4.3.3 Mantenimiento:

El modelo de minería de datos tiene que ser mantenido periódicamente. El proceso de mantenimiento involucra reentrenamiento del modelo con los datos actualizados de las transacciones hechas por los usuarios. Con el fin de hacer este mantenimiento se configuraron dos tareas dentro del paquete Maestro “CORRER TODO.dtsx” del proyecto de Integration Services “SPARIS”, una tarea llamada “Cargar Escenarios Minería” actualiza el escenario de clientes y el escenario anidado de transacciones haciendo un llamado al

paquete “Mineria.dtsx”, una segunda tarea llamada “Procesar BD SPARAS” hace posteriormente un reprocesamiento del modelo de recomendaciones de minería de datos al reprocesar completamente la base de datos multidimensional del proyecto “SPARAS”. El flujo del paquete Maestro puede observarse en la Figura 50.

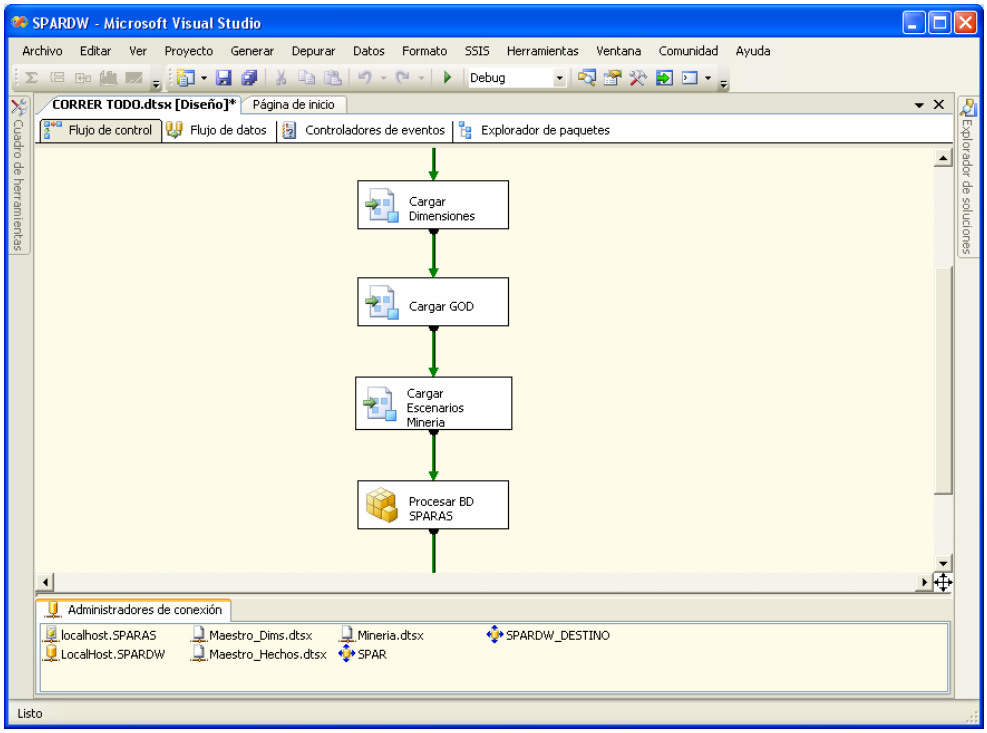


Figura 50: Flujo del paquete Maestro "CORRER TODO.dtsx" que permite el mantenimiento del modelo de recomendaciones.

El paquete Mineria.dtsx está compuesto por dos tareas de flujos de datos. La primera tarea de flujo de datos selecciona los datos de usuario y los almacena en la tabla EscDatosUsuario. La segunda tarea hace el mismo proceso para almacenar los datos en EscDatosTransacciones. La Figura 51 ilustra el proceso:

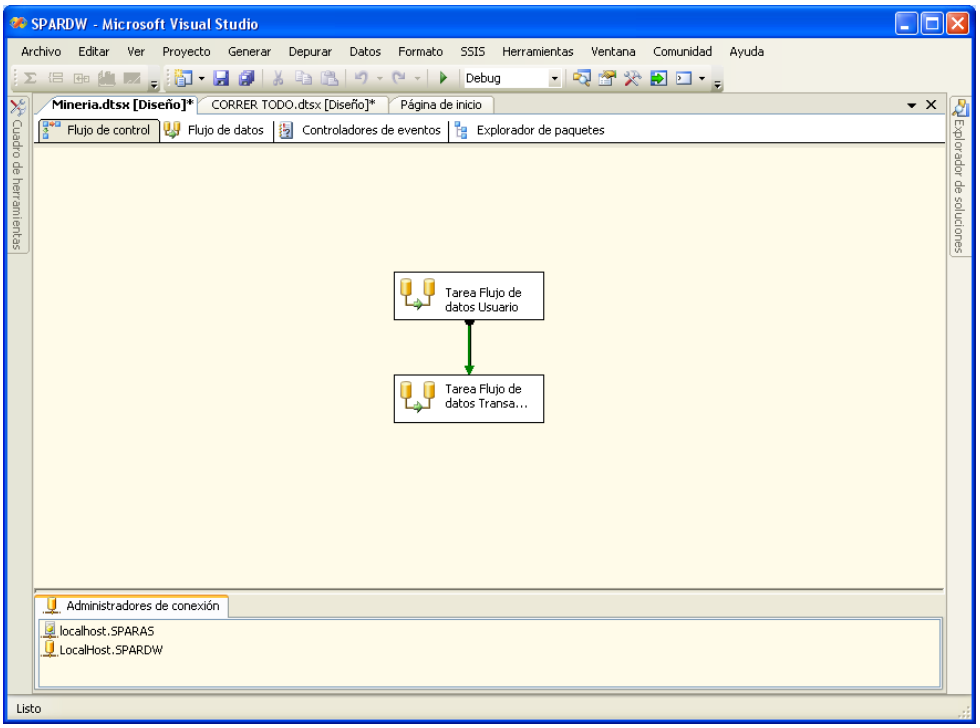


Figura 51: Flujo del paquete "Mineria.dtsx" que carga los datos dentro del escenario de usuarios y el escenario anidado de transacciones.

El paquete maestro "CORRER TODO.dtsx" es el encargado de hacer la carga y actualización continua de las dimensiones, las tablas de hechos, los escenarios de minería de datos, y de hacer el reprocesamiento periódico de la base de datos multidimensional "SPARAS" permitiendo actualizar los datos multidimensionales del cubo y el modelo de minería de datos. El llamado a este paquete se hace a través de una tarea automatizada creada en el SQL Agent (Herramienta de SQL Server 2005) que se ejecuta en determinados periodos de tiempo. En este caso el servidor ejecutará la tarea todos los lunes a las 12:00 am, permitiendo mantener la bodega de datos relacional y multidimensional así como el modelo de minería de datos actualizados.

CAPITULO V

5. DESCRIPCIÓN DE LA HERRAMIENTA DE ADMINISTRACIÓN

Este capítulo describe el proceso de desarrollo de la herramienta de administración SPARAMO.

La construcción de esta herramienta tiene un fin académico, que es brindar conocimiento de cómo una aplicación cliente puede hacer uso de una serie de capacidades y servicios administrativos disponibles en el servidor de base de datos multidimensional de Microsoft Analysis Services (AS). La construcción de esta herramienta permitió conocer a fondo cómo se puede hacer una conexión y trabajar directamente sobre objetos (cubos, dimensiones, perspectivas, entre otros) en una instancia de AS, desde su creación hasta su manipulación. El desarrollo de esta herramienta brindó una idea de cómo trabaja el ***Business Intelligence Development Studio*** (BIDS), que es la principal herramienta de diseño y construcción de bases de datos multidimensionales proporcionada por Microsoft, mediante el BIDS se construyo el cubo, las dimensiones, jerarquías, perspectivas, vistas, cadenas de conexión y todos los objetos necesarios para la creación del DW multidimensional de SPAR, que posteriormente es usado por la herramienta OLAP para satisfacer necesidades de consultas analíticas sobre el repositorio.

SPARAMO es una herramienta realizada en C# que muestra el uso de Objetos de Administración de Análisis (AMO), que son un conjunto de librerías que proporcionan un modelo de objetos de .NET Framework que las aplicaciones cliente pueden utilizar para administrar una instancia de AS. La herramienta SPARAMO permite realizar una serie de funcionalidades básicas sobre AS, como la creación y eliminación de bases de datos (DW) multidimensionales basados en un DW relacional, además permite la creación de cubos locales basados en cubos existentes en AS, estos cubos locales son archivos que contienen datos multidimensionales que pueden ser utilizados por los usuarios para hacer consultas dinámicas desde una herramienta OLAP desconectado del servidor de AS. Para la creación de cubos locales la herramienta hace uso de ADOMD.Net, que son un conjunto de clases de .NET Framework que se utilizan para obtener acceso a los objetos y los datos de AS y trabajar con ellos.

Para la construcción de la herramienta se utilizó el Proceso Unificado Racional de desarrollo de software (RUP). De la misma forma que para el desarrollo de la herramienta OLAP, se estableció una iteración base en la cual se hacen revisiones continuas de cada fase para lograr un desarrollo iterativo e incremental.

5.1 Fase de Preparación Inicial

Esta fase incluye la concepción inicial, la planificación y alcance del proyecto de desarrollo de la herramienta. Los artefactos que se obtuvieron durante esta fase son: Los diagramas de casos de uso, los casos de uso de alto nivel, modelo conceptual preliminar y diagramas de secuencia, estos artefactos pueden ser encontrados en el **Anexo 16.**

5.2 Fase de Preparación Detallada

El objetivo principal de esta fase fue plantear la arquitectura para el ciclo de vida del proyecto. En esta fase se realiza la captura de la mayor parte de los requerimientos funcionales, acumulando la información necesaria para realizar la construcción.

Los artefactos que se obtuvieron durante esta fase son: La arquitectura de la aplicación, los casos de uso en formato expandido y los casos reales de uso para los tres módulos.

Arquitectura de la Aplicación:

La arquitectura de la herramienta de administración tiene una arquitectura de tres capas que permiten separar los datos de la aplicación, la interfaz del usuario y la lógica de control en distintos componentes. Este diseño modular permite mayor flexibilidad, facilita la reusabilidad y la escalabilidad de la aplicación. La Figura 52 muestra la arquitectura preliminar de la aplicación:

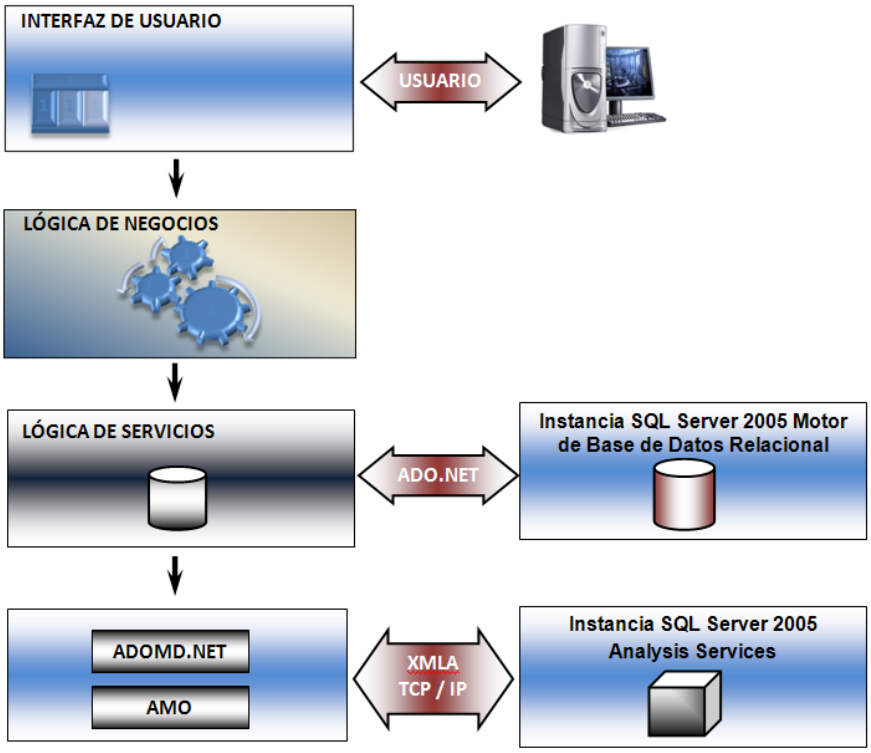


Figura 52: Arquitectura de SPARAMO

- **Usuarios:**
    - **Administrador:** Este actor tiene completo conocimiento del dominio del problema por tanto tiene la posibilidad de realizar las tres funciones administrativas ofrecidas por la herramienta que son: La creación de nuevas bases de datos multidimensionales con características de dimensionalidad determinadas por el usuario administrador, eliminación de bases de datos multidimensionales que no están siendo útiles y la creación de cubo locales con determinadas características dimensionales que permitan satisfacer necesidades analíticas de otros usuarios.

- **Las tres capas son:**

    - **Capa Interfaz de Usuario (SPARAdmin):** Capa de interacción con los usuarios, permite hacer uso de las distintas funcionalidades de la herramienta por medio de solicitudes directas que posteriormente será redirigidas a la capa lógica del negocio.

    - **Capa Lógica del Negocio (Lógica):** Gestiona e interpreta los eventos que realiza el usuario, e invoca acciones en la lógica de servicios para posteriormente retornar resultados a la capa de presentación. Esta capa funciona como el controlador lógico de la aplicación que da soporte al dominio del problema.

    - **Capa Lógica de Servicios (Obtener Tablas):** Capa que interactúa directamente con el servidor de base de datos relacional y el servidor de base de datos multidimensional gestionando el acceso y la administración de datos.

- **Librerías y Clases:**

    - **ADOMD.NET:** Son un conjunto de clases de .Net Framework que han sido creadas para que las aplicaciones cliente pueden comunicarse con AS, permitiendo el acceso a los datos multidimensionales por medio de consultas MDX (Lenguaje de consultas para bases de datos multidimensionales) [14].

    - **AMO:** Son un conjunto de clases de .Net Framework que permite a las aplicaciones cliente administrar objetos de una instancia de AS. Este conjunto de librerías permiten crear, eliminar, modificar objetos tales como dimensiones, cubos, estructuras de minería y bases de datos de Análisis Services. Con AMO no pueden consultarse datos multidimensionales, para hacer consultas de datos se usa ADOMD.NET [15].

    - **ADO.NET:** Son un conjuntos de clases de .Net Framework que exponen servicios de acceso a datos relacionales, XML y de aplicaciones [16].

- **Protocolos de Comunicación:**

  - **XMLA:** Es un protocolo de acceso de objetos (SOAP) basado en el protocolo XML. Es un estándar abierto que fue diseñado para acceder a datos de cualquier fuente multidimensional que reside en la web. XMLA surgió como iniciativa de Microsoft y rápidamente ha sido adoptado por un gran número de fabricantes de bases de datos multidimensionales (Hyperion, SAS, Mondrian, entre otros.), convirtiéndose actualmente en el único estándar universal para acceso a datos multidimensionales. XMLA es el mecanismo central de comunicación de AS, por esta razón AMO y ADOMD.NET toman comandos desde una aplicación cliente y los convierten en mensajes XMLA para ser enviados a una instancia del AS.

Casos de Uso de Formato Expandido:

A continuación se muestra el caso de uso en formato expandido de crear una base de datos multidimensional, los demás casos de uso de formato expandido se describen en el **Anexo 16.**

Caso de uso en formato expandido: Crear Base de Datos Multidimensional

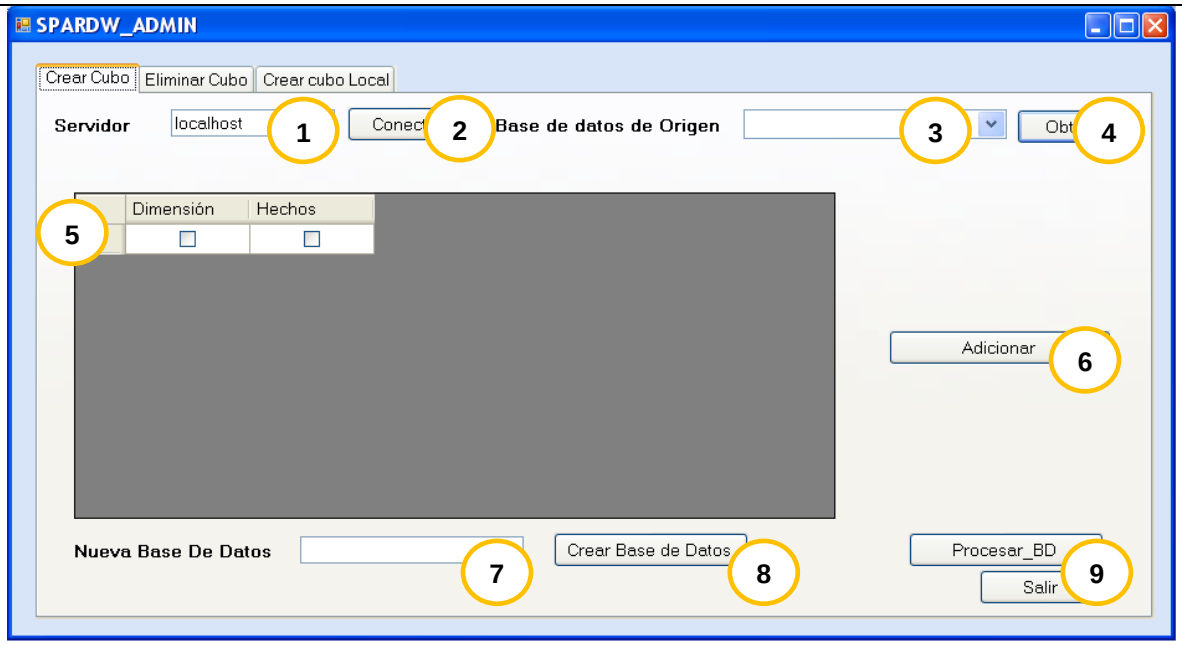
Caso de uso:	Crear Base de Datos Multidimensional.
Actores:	Administrador
Propósito:	Crear la Base de Datos multidimensional a partir de un DW relacional.
Resumen:	El administrador ingresa dentro del modulo de creación de la base de datos multidimensional de la herramienta y realiza la secuencia de pasos que involucra la creación. Estos pasos son: • Conectarse al servidor relacional. • Seleccionar el DW relacional. • Seleccionar las tablas de dimensiones y tablas de hechos. • Crear la BD multidimensional. • Procesar la BD multidimensional.
Tipo:	Primario.

Curso normal de eventos	
Acción de los Actores	**Respuesta del sistema**
1. Este caso de uso se inicia cuando un usuario desea crear una base de datos multidimensional.	**2.** El Sistema muestra una interfaz con tabuladores

3. El usuario selecciona el tabulador Crear	
4. El usuario establece el nombre del servidor de base de datos relacional con el cual se va a conectar.	**5**. El Sistema muestra el catalogo de los nombres de las bases de datos relacionales.
6. El Usuario selecciona la base de datos relacional del DW y solicita obtener las tablas de hechos y dimensiones	**7**. El Sistema obtiene y muestra las tablas de hechos y dimensiones.
8. El usuario selecciona las tablas de hechos y dimensiones y solicita adicionarlas a la base de datos multidimensional a crear.	**9**. El Sistema muestra mensaje de éxito de adición.
10. El usuario digita el nombre de la base de datos multidimensional y solicita crearla.	**11**. El Sistema crea la base de datos multidimensional en el servidor.
12. El usuario solicita procesar la base de datos multidimensional.	**13**. El Sistema procesa la base de datos multidimensional y obtiene los datos.

Casos Reales de Uso: El caso real de uso de Crea la Base de Datos Multidimensional se muestra a continuación. Los demás casos reales de uso se describen en el **Anexo 16.**

Caso de Uso Real Crear Base de Datos Multidimensional



Curso normal de los eventos	
Acción del Actor	**Respuesta del Sistema**
1. El usuario accede a la aplicación entrando en el tabulador de crear cubo.	2. El sistema muestra la interfaz de creación de base de datos multidimensional
3. El usuario digita el nombre del servidor **[1]** y solicita conectarse con el servidor **[2].**	4. El sistema se conecta con el servidor de acuerdo al nombre del servidor introducido y obtiene el catalogo de las bases de datos relacionales **[3].**
5. El usuario selecciona la base de datos relacional del DW **[3].**	
6. El usuario solicita obtener el catalogo de tablas de la base de datos del DW. **[4]**	7. El sistema obtiene el catalogo de tablas de la base de datos relacional del DW. **[5]**
8. El usuario selecciona las tablas de dimensiones y de hechos **[5]** y solicita adicionarlas a la base de datos multidimensional **[6]**	9. El sistema guarda las tablas de dimensiones y de hechos seleccionadas. **[6]**
10. El usuario escribe el nombre de la base de datos multidimensional que desea crear **[7]** y solicita crearla. **[8]**	11. El sistema crea la base de datos multidimensional con el nombre que el usuario introdujo. **[8]**
12. El usuario solicita procesar la base de datos multidimensional. **[9]**	13. El sistema procesa la base de datos multidimensional obteniendo los datos. **[9]**

5.3 Fase de Construcción:

Esta fase involucra la creación de la herramienta de administración, para lo cual se definieron tres ciclos de vida iterativos, cada uno de los cuales desarrolla una funcionalidad específica de la herramienta.

Ciclo 1: Creación de una base de datos multidimensional de AS

En este ciclo se desarrollo la funcionalidad completa de la herramienta que permite la creación de una bodega de datos multidimensional a partir de una bodega de datos relacional. Esta funcionalidad permite desde la creación de una cadena de conexión a la fuente de datos relacional hasta la selección de tablas de dimensiones y tablas de hechos, construcción de la bodega de datos multidimensional y su posterior procesamiento. Esta funcionalidad de la herramienta permite la construcción de una bodega de datos multidimensional que contiene todos los tipos de relaciones estándares entre tablas de dimensiones y las tablas de hechos, tales como: relaciones uno a muchos, relaciones referenciadas (subdimensiones) y relaciones muchos a muchos entre dimensiones.

Ciclo 2: Eliminación de una base de datos multidimensional de AS

En este ciclo se desarrollo la funcionalidad que permite la conexión, selección y eliminación de determinada base de datos multidimensional alojada en el servidor de AS.

Ciclo 3: Creación de cubos locales

En este ciclo se desarrollo la funcionalidad que permite la creación de cubos locales. La herramienta permite establecer una conexión al servidor de análisis para obtener el catálogo de las bases de datos multidimensionales, seleccionar la base de datos de AS y el cubo multidimensional sobre el cual se va a construir y llenar un cubo local, seleccionar las dimensiones de cubo y los grupos de medidas de interés y posteriormente hacer la creación y el llenado del cubo local. Los cubos locales son archivos que contienen datos multidimensionales que mejoran el desempeño de consultas analíticas al permitir a los usuarios hacer consultas multidimensionales off-line, usando como fuente un archivo que se almacena en su equipo local evitando acceder a un servidor análisis remoto.

La Figura 53 muestra un ejemplo de la herramienta de administración en la creación de un cubo local.

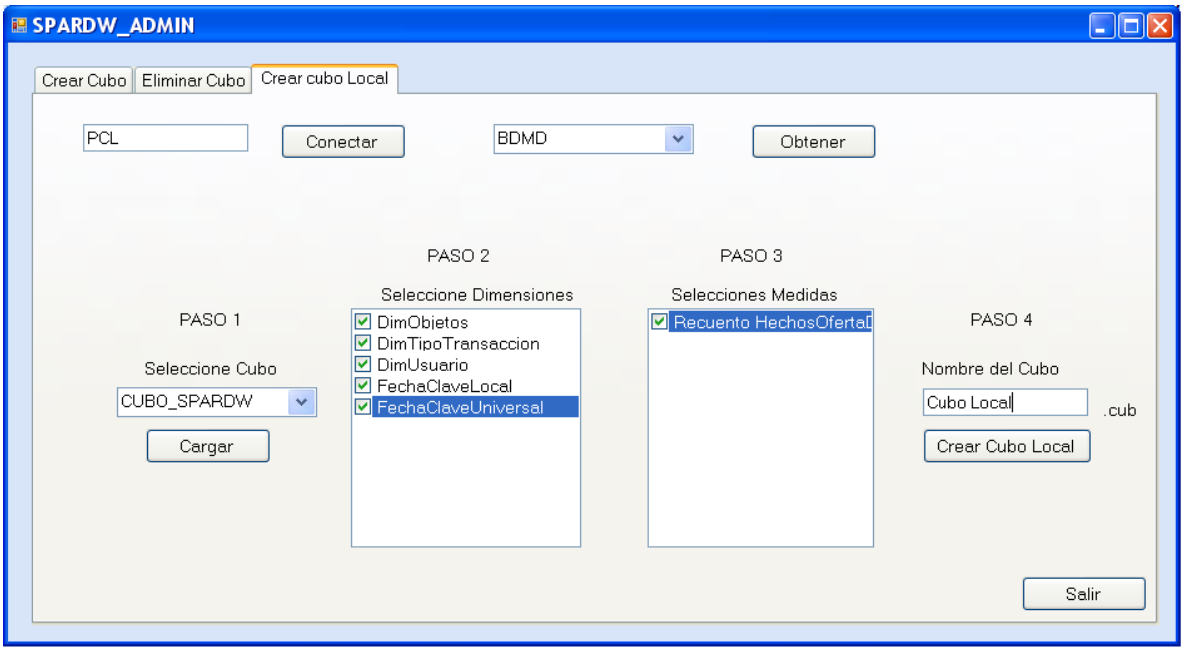


Figura 53: Creación de cubos locales en SPARAMO

5.4 Fase de Transición

En esta fase se hicieron ajustes y correcciones finales requeridas y se hace entrega de una herramienta de administración completamente funcional que tiene como fin ilustrar el uso de Objetos de administración de análisis (AMO) y el conjunto de clases ADOMD.NET. Esta herramienta permite estudiar el comportamiento de herramientas que administran y obtienen acceso a los datos de una instancia de AS y puede servir como soporte para el desarrollo de aplicaciones cliente personalizadas de construcción y de consulta de bases de datos multidimensionales bajo plataformas Microsoft.

CONCLUSIONES Y RECOMENDACIONES Y TRABAJO FUTURO

De la experiencia obtenida con el desarrollo del proyecto se puede decir que:

- La creación de un módulo de minería de datos basado en el algoritmo de reglas de asociación ofrece ventajas con respecto a otros algoritmos, porque este produce resultados que causan un impacto favorable en el negocio sin la necesidad de que el usuario final tenga que realizar interpretaciones complejas.

- Complementar un sistema de Bodegas con Técnicas de Minería de Datos permite optimizar la obtención de resultados de un DSS, porque las técnicas de minería ayudan a realizar análisis más profundos de los datos que no necesariamente son dependientes de las capacidades que tengan los usuarios para hacer consultas y generar hipótesis, como es el caso del modelo de minería de recomendaciones construido para el repositorio digital SPAR 1.0.

- Las metodologías y técnicas de inteligencia de negocios además de dar buenos resultados en el área de gestión de negocios, pueden ser aplicadas dentro del contexto de la educación en línea, permitiendo obtener buenos resultados.

- La arquitectura bus propuesta por Ralph Kimball facilita el proceso de construcción de sistemas DSS, por que permite hacer la construcción basado en incrementos o subconjuntos, a los cuales se le definen metas especificas que soportan a determinados áreas del negocio, estos subconjuntos posteriormente son integrados para producir un sistema completo que da soporte a la toma de decisiones estratégicas.

- El ciclo de vida dimensional propuesto por Ralph Kimball es una metodología que satisface completamente las necesidades de diseño, construcción e implementación de sistemas DSS. Porque involucra todas los procesos y tareas requeridos para realizar desde la extracción de los datos, al desarrollo y despliegue de aplicaciones.

Del desarrollo del proyecto se obtuvo:

1) Una bodega de datos relacional (SPARDW) que da soporte a tres áreas de negocio (Data Mart) que son:

- La Gestión, la oferta y la demanda de contenidos educativos: Este Data Mart construido produce información analítica con respecto a la cantidad de consultas, publicaciones y modificaciones hechas por los usuarios sobre los recursos educativos. Esta Data Mart produce diferentes perspectivas dimensionales de la información al permitir hacer consultas que cruzan atributos de los usuarios, de los

recursos de aprendizaje, de los tipos de acceso a los recursos, de las fechas y tiempos del día, de las áreas temáticas, entre otros. Este Data Mart permite responder a inquietudes como: ¿Cuál es la cantidad de transacciones de los usuarios, cuales son las países que más consultan?, ¿Cuales son los recursos de aprendizaje más consultados? Cuáles son las clasificaciones temáticas más buscadas?, Cuales son los horarios y fechas de consulta más habituales?, ¿ Cuáles son los usuarios que mas hacen uso del repositorio, cuáles son sus intereses y qué tipo de usuarios son (administrador, LMS, anónimo, registrado)?, ¿ Cuáles son los recursos poco consultados?, entre muchas más preguntas que pueden ser de importantes y que son resueltas perfectamente por el sistema de DW/BI.

- Evaluación de Contenidos: Este Data Mart produce información en cuanto a los niveles de satisfacción de los usuarios, con respecto a la calidad de contenidos de los recursos, calidad de presentación de los recursos y satisfacción en general de los usuarios con respecto a estos. De la misma manera que el anterior data mart este permite analizar la información desde múltiples perspectivas dimensionales, que involucran obtener y cruzar atributos de los usuarios, de los recursos de aprendizaje, de las fechas y tiempos del día, de las áreas temáticas, preguntas y respuestas. Este data mart permite responder a inquietudes como: ¿Cuáles son los objetos mejor calificados?, ¿Cuáles objetos tiene baja calificación en su contenido o presentación o satisfacción en general que ofrece a los usuarios?, ¿cuál ha sido el promedio de calificación que han tenido los recursos a través del tiempo? ¿Cómo ha sido la aceptación de los recursos en determinados usuarios y países?

- Sesiones de Usuario: Este Data Mart produce información analítica relacionada con el comportamiento que tiene los usuarios en las sesiones realizadas. Igualmente este data mart permite cruce dimensional entre sus dimensiones de usuario, pagina, fecha, sesión, localización, país, entre otras. Permitiendo responder a inquietudes con respecto al tiempo promedio de duración de las sesiones, éxito de las sesiones con respecto a si se realizo algún tipo de transacción (consulta, descarga, publicación, modificación de metadata, etc.), cantidad de páginas consultadas, formas de inicio y de finalización de sesión en el repositorio SPAR.

2) Una base de datos multidimensional (SPARAS), construida con base en la bodega de datos relacional, es el sistema fuente sobre el cual la herramienta OLAP obtiene los datos para producir información útil para la toma de decisiones. La base de datos multidimensional contiene un cubo multidimensional que optimiza las consultas analítica al proveer tablas agregadas (resúmenes) que son construidas en función de consultas comunes que ya se han hecho, agilizando de esta manera la entrega de resultados, de la misma manera el cubo multidimensional por ser una capa de nivel superior a la bodega de datos relacional posee metatadata orientada al usuario que aumentan el entendimiento de los objetos analíticos (dimensiones, atributos, jerarquías, medidas), además provee de un lenguaje de consultas dimensionales (DMX) desarrollado por Microsoft exclusivamente para análisis multidimensional.

3) Una herramienta OLAP que está dividida en dos grandes módulos: Un módulo que corresponde a la herramienta de reportes estándares y un segundo módulo que corresponde a la aplicación analítica, que a su vez se divide en un módulo dinámico de consultas sobre el cubo que es almacenado en el servidor y un modulo dinámico de consultas sobre archivos de cubo que son almacenados en el equipo del usuario. Estos módulos se adhieren directamente a la arquitectura del Repositorio Digital SPAR 1.0 permitiendo una completa integración de los dos sistemas. Cada uno de los módulos integra componentes y servicios ofrecidos por las plataformas Microsoft entre los que se encuentra: Componentes Web Office (OWC) y Servicios de Reportes (Reporting Services), los OWC permiten hacer análisis personalizados sobre el cubo de la base de datos multidimensional y sobre archivos de cubos de acuerdo a las necesidades de los usuarios, los datos son mostrados en tablas, matrices y gráficos. Los Servicios de Reportes permitieron integrar al repositorio una aplicación web que permite la administración y el acceso a informes (predefinidos) a través de la interfaz de análisis de reportes estándares de SPAR. (Más detalles ver Capitulo II, sección Arquitectura de la Herramienta OLAP)

4) Un módulo de minería de datos que tiene como objetivo mejorar la utilidad del repositorio, satisfaciendo en gran medida los intereses temáticos de los usuarios. Este módulo de minería de datos hace recomendaciones de recursos educativos que tienen una alta probabilidad de interesar a cada visitante del repositorio SPAR. Para hacer las recomendaciones se uso el algoritmo de Reglas de Asociación de Microsoft que es uno de los algoritmos proporcionados por Microsoft SQL Server 2005 Analysis Services (SSAS), el cual identifica un conjunto de correlaciones existentes en los datos basado en las consultas y/o descargas que hacen los usuarios. El modelo de minería construido se puso a disposición del Repositorio Digital SPAR permitiendo mostrar automáticamente desde el momento que un usuario descarga o consulta determinado recurso educativo una lista de 4 objetos recomendados en orden de más alta probabilidad. (Para una descripción más detallada remitirse al Capítulo IV)

5) Una Herramienta de Administración de Objetos de Análisis (SPARAMO) que tiene como objetivo brindar conocimiento de cómo una aplicación cliente puede hacer uso de una serie de capacidades y servicios administrativos disponibles en el servidor de base de datos multidimensional de Microsoft Analysis Services (AS). La construcción de esta herramienta permitió conocer a fondo cómo se puede hacer una conexión y trabajar directamente sobre objetos (cubos, dimensiones, perspectivas, entre otros) en una instancia de AS. Este proceso de administración de una instancia de AS ocurre a muy bajo nivel y no es conocido por los usuarios que hacen uso de herramientas de diseño y construcción de objetos de análisis que existen en el mercado, puesto que estas son herramientas de alto nivel en las que el usuario diseña, crea e implementa mediante el uso de asistentes. La herramienta SPARAMO permite realizar una serie de funcionalidades básicas sobre AS, como la creación y eliminación de bases de datos (DW) multidimensionales basados en un DW

relacional, además permite la creación de cubos locales basados en cubos existentes en AS.

6) Se logró publicar un artículo en una revista nacional:

BAYONA, Diego, CALVACHE, Alexander, MENDOZA y Martha. Sistema de Apoyo a la Toma de Decisiones para el Repositorio Digital de Objetos de Aprendizaje SPAR 1.0 Enlace Informático. Revista de Ciencia y Tecnología Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca, Cuarta Edición, Junio 2006. ISSN: 1692-374X. http://enlaceinformatico.unicauca.edu.co/**.**

Como recomendaciones se plantea:

- Para el desarrollo de sistemas de soporte a la toma de decisiones, se recomienda hacer un análisis detallado de los sistemas de información para evaluar la disponibilidad de los datos y poderlos contrastar con los requerimientos de los usuarios, de tal modo que se permita satisfacer ampliamente las necesidades analíticas de los usuarios.

- Se recomienda hacer un estudio detallado de las teorías, técnicas, metodologías y tecnologías disponibles, antes de desarrollar un proyecto de estas características con el fin de lograr mejores resultados en cada uno de los procesos de desarrollo.

- Antes de comenzar con el desarrollo de un sistema DSS, es necesario entender los procesos de negocio, con el fin de hacer una buena captura de requerimientos y poder guiar correctamente el desarrollo del DSS.

- El proceso de desarrollo de un DSS debe comenzar por dividir la solución en varios procesos de negocio priorizados, ejecutando primero aquellos que tengan mayor valor para el negocio y mayor viabilidad para llevar a cabo la construcción del sistema.

Como trabajos futuros se propone:

- La implementación del Data Mart de Usabilidad del Repositorio, que tiene como objetivo producir información con respecto a la facilidad de uso del repositorio, facilidad de navegación, accesibilidad, facilidad de encontrar objetos, entre otros.

- La implementación del Data Mart de Búsquedas del Repositorio que tiene como objetivo producir información con respecto al tipo de búsquedas que realizan los usuarios, tipos de temáticas, títulos, palabras claves, entre otros. Permitiendo posibles aplicaciones de Minería de Texto para mejorar los resultados de búsqueda de tal forma que los resultados sean más precisos a las necesidades del usuario.

- Investigaciones en minería de datos que permitan la implementación de otras técnicas de minería para aumentar la cantidad de usuarios en el repositorio, mejorar la utilidad del repositorio, aumentar descargas, publicaciones y consultas de objetos, conservar los usuarios, entre otros.

REFERENCIAS BIBLIOGRÁFICAS:

[1]. Artículo "Plan Sectorial 2002-2006" Ministerio de Educación Colombia. http://www.mineducacion.gov.co/1621/propertyvalue-30966.html

[2]. J. A. Muñoz, J. I. Giraldo, SPAR 1.0 - Repositorio Digital de acceso público basado en IMS DRI 1.0 con soporte a múltiples especificaciones y estándares de meta datos. Proyecto de Grado, Universidad del Cauca, Popayán Colombia 2006. http://spar.unicauca.edu.co/spar/default.aspx

[3]. Wang John, Encyclopedia of Data Warehousing and Mining, Idea Group Inc. 2006.

[4]. Teoría sobre Business Intelligence" Concurso MicroStrategy Experiencia Business Intelligence 2da. EdiciónAño 2005 http://www.microstrategy.com.pe/ExperienciaBI2/teoriadw.pdf

[5]. The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset. 2006 by Wiley Publishing, Inc., Indianápolis, Indiana

[6]. W. H. Inmon. Building the Data Warehouse. Third Edition. Wiley Computer Publishing. 2002.

[7]. Kimball, R. (1998). The Data Warehouse Toolkit: Practical techniques for building dimensional Data Warehouse. Wiley Computer Publishing

[8]. Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, The data Warehouse Lifecycle Toolkit, Wiley computer Publishing. 1998.

[9]. Fayyad U., Piatestky-Shapiro G., Smyth P. Discovery and Data Mining. AAAI Press/The MIT Press 1996.

[10]. Michael J. A. Berry and Gordon S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management Second Edition, Wiley, 2004.

[11]. Zhao Hui Tang, Jamie MacLennan. Data Mining with SQL Server 2005, Wiley Publishing, Inc 2005.

[12]. LARMAN, Craig. UML y Patrones: Introducción al Análisis y Diseño Orienta a Objetos. Ed. Prentice Hall. Mexico, 1999.

[13]. Alejandro Amat Bedmar - "Ingeniería de Conocimiento Minería de Datos Empresariales" – M.S./E.T.S Ingeniería Informática de la Universidad de Granada. 2005.

[14]. Microsoft TechNet:
http://technet.microsoft.com/es-es/library/ms171022.aspx

[15]. MSDN, Introducing AMO Concepts
http://msdn2.microsoft.com/en-us/library/ms345089.aspx

[16]. MSDN, ADO.NET:
http://msdn2.microsoft.com/es-es/library/e80y5yhx(VS.80).aspx