SISTEMA DE SOPORTE A LA TOMA DE DECISIONES EN VIVEROS AUTOMATIZADOS UTILIZANDO OLAP Y MINERÍA DE DATOS

JIMENA ADRIANA TIMANÁ PEÑA
RENE VALENCIA VALLEJO

Director: MSc. CARLOS ALBERTO COBOS LOZADA

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE SISTEMAS
POPAYÁN, NOVIEMBRE DE 2007

AGRADECIMIENTOS

A Dios por iluminar nuestro camino a cada instante de nuestras vidas y por habernos permitido culminar con éxito nuestras carreras.

A nuestros docentes, al programa de Ingeniería de Sistemas y a la Universidad del Cauca que nos forjaron como personas y profesionales.

Al ingeniero Carlos Alberto Cobos Lozada, nuestro mentor y amigo por su colaboración, dedicación y consejos que nos permitieron alcanzar los objetivos de este proyecto.

A Miguel Corchuelo, Juan Pablo Paz, Román Ospina y Diego Bravo, por su orientación en las diferentes etapas del proyecto.

A nuestros compañeros y amigos de la universidad con los que compartimos tantos momentos.

Nuestros agradecimientos más profundos a nuestros padres y familias, los cuales nos brindaron su apoyo incondicional y amor infinito.

Muchas gracias.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

A lo largo de este documento se desarrollan las temáticas y los conceptos teóricos relevantes en el desarrollo del proyecto. Esta sección presenta una visión general del trabajo desarrollado y una ubicación contextual en general. El documento se encuentra organizado de la siguiente manera:

PARTE 1 – CONCEPTOS DEL ÁMBITO DE APLICACIÓN

Inicialmente, se Incluyen varios conceptos que permitirán, en el transcurso del documento comprender la temática expuesta.

Descripción del problema. Esta sección brinda una visión general del proyecto, la problemática que originó su desarrollo, la justificación y la viabilidad del mismo.

Contribución a la solución. Presenta los objetivos propuestos al iniciar el proyecto y los resultados obtenidos a la finalización del mismo.

PARTE 2 – OLTP

Marco teórico. Recopila la base conceptual necesaria para la elaboración de un sistema de Procesamiento Transaccional en Línea (OLTP), incluyendo sus características y funcionalidades.

Metodología de desarrollo del sistema OLTP Web. Presenta y describe de forma general, la metodología de desarrollo utilizada, incluyendo los artefactos generados en cada una de sus fases.

Problemas y soluciones. Compendia los problemas presentados en la elaboración del sistema OLTP Web de principio a fin, junto con las soluciones planteadas para cada inconveniente presentado.

PARTE 3 – BODEGA DE DATOS

Marco teórico. Recopila la base conceptual necesaria para la elaboración de una Bodega de Datos, incluyendo sus características y su arquitectura general.

Metodología de desarrollo de la Bodega de Datos. Presenta y describe detalladamente la metodología de desarrollo utilizada para la construcción de la Bodega de Datos, a saber:

- Definición y planeación del proyecto: En esta sección se muestra la definición y el alcance del proyecto, igualmente se expresan los requerimientos definidos.

- Modelado dimensional: Muestra la matriz bus creada para la bodega, además de los modelos dimensionales generados, identificando minuciosamente a cada uno de sus componentes: tablas de hechos y dimensiones.

- Implementación de la bodega: Se muestra la implementación de los modelos físicos generados a partir del modelado dimensional.

Diseño y desarrollo del proceso de ETL (Extracción, Transformación, Carga del inglés Load. En esta sección se muestra el diseño y la implementación de los paquetes que estarán encargados de los procesos de ETL a los datos que salen del OLTP y van hacia la Bodega de Datos.

Problemas y soluciones. Compendia los problemas presentados en la elaboración de la Bodega de Datos, de principio a fin, junto con las soluciones planteadas para cada inconveniente presentado.

PARTE 4 – OLAP

Marco teórico. Recopila la base conceptual necesaria para comprender un sistema de Procesamiento Analítico en Línea (OLAP), incluyendo su clasificación y características.

Proceso de selección de la herramienta OLAP. Detalla los criterios y etapas que se tuvieron en cuenta para la selección de una herramienta OLAP.

Integración de la herramienta OLAP. Describe de forma general el proceso de integración de la herramienta OLAP seleccionada con el proyecto.

PARTE 5 – MINERÍA DE DATOS

Marco teórico. Recopila la base conceptual necesaria para la elaboración de un prototipo de herramienta para descubrimiento de conocimiento, incluyendo su definición, técnicas y aplicaciones, entre otras.

Metodología de desarrollo de la herramienta de minería de datos. Presenta y describe detalladamente la metodología de desarrollo utilizada.

Problemas y soluciones. Compendia los problemas presentados en la elaboración de un prototipo de herramienta para descubrimiento de conocimiento, de principio a fin, junto con las soluciones planteadas para cada inconveniente presentado.

PARTE 6 – CONCLUSIONES

Conclusiones, recomendaciones y trabajo futuro: Expone las conclusiones que se generaron después de la culminación del proyecto de grado e igualmente, las recomendaciones pertinentes para trabajos posteriores relacionados con la continuidad del proyecto.

Referencias bibliográficas. Bibliografía utilizada para el desarrollo del trabajo.

PARTE 1 – CONCEPTOS DEL ÁMBITO DE APLICACIÓN

1 GLOSARIO

Cultivo: Conjunto de labores, operaciones y cuidados que se efectúan para que el suelo dé mayores y mejores cosechas.

Fuente semillera: Grupo de árboles de la misma especie o grupo de especies donde predominan individuos fenotipicamente de conformación aceptable o deseable en cuanto a forma, vigor y sanidad, el cual es manejado técnicamente para aumentar y sostener la producción de semilla en calidad y cantidad.

Germinación: Es el proceso en cual la semilla en estado de vida latente entra en actividad y origina una nueva planta.

Invernadero: Estructura totalmente cerrada cubierta por materiales transparentes como vidrio, plástico o fibra de vidrio, dentro de la cual se generan condiciones artificiales de temperatura, humedad y luminosidad para obtener producciones elevadas antes de su época normal o en climas que no le son propios, mejor control de plagas y enfermedades, entre otras.

Pregerminación: La pregerminación es la acción de dejar desarrollar los brotes en la semilla, algún tiempo antes de la germinación.

Repique: Es el proceso de sacar las plántulas de la cámara de germinación y ponerlas en las bolsas o platabandas.

Semilla: La semilla o pepita es la estructura mediante la que realizan la propagación las plantas. Una semilla contiene un embrión del que puede desarrollarse una nueva planta bajo condiciones apropiadas.

Sustrato: Un sustrato es todo material sólido, de síntesis o residual, mineral u orgánico, que, colocado en un contenedor, en forma pura o en mezcla, permite el anclaje del sistema radicular de la planta, desempeñando, por tanto, un papel de soporte para la planta.

Viveros: Espacio dedicado al cultivo de una o más especies vegetales. Puede estar conformado por uno o más invernaderos.

2 DESCRIPCIÓN DEL PROBLEMA

2.1 PLANTEAMIENTO DEL PROBLEMA

A finales de los años 80 y principios de los 90´s, la tecnificación del campo en los grandes países industrializados ha tenido un gran auge. Maquinaria de ultima generación que involucra sistemas de posicionamiento global, monitoreo de cultivos a través de sensores foto sensibles, Imágenes satelitales o fotografías aéreas para el control de crecimiento en cultivos, entre otros, son el común denominador en estas potencias industriales; sin embargo, en los países tercermundistas, incluido Colombia, la aplicación de conocimientos empíricos y la utilización de herramientas rústicas y manuales es todavía notoria. Es difícil encontrar en las pequeñas y medianas zonas de siembra del área rural, maquinaria, sistemas informáticos (a bajo precio y en lengua nativa), etc., que brinden información y soporte referente a las labores desarrolladas del campo, que involucran en muchas ocasiones, factores que no se tienen en cuenta para lograr una producción de alimentos en óptimas condiciones [1][2].

Factores como el contenido de humedad del suelo y la temperatura del aire pueden tener un efecto enorme sobre el rendimiento de los cultivos. La habilidad de registrar datos de cultivo, suelo y ambiente de forma regular durante el crecimiento de los cultivos, puede aportar información crítica cuando un productor comienza a responder preguntas concernientes al rendimiento de los cultivos y a las relaciones causa efecto.

La recolección, procesamiento, análisis y validación son pasos necesarios para que los datos puedan transformarse en información útil en la toma de decisiones en el campo. Es por eso que se vio la necesidad de brindar herramientas hardware y software generadas en asocio con los programas de Ingeniería Física, Forestal y Sistemas de la Universidad del Cauca, con el fin de que el productor pueda empezar a responder preguntas sobre que pasó en una situación productiva en particular y predecir que pasará para un conjunto de circunstancias dadas, por lo que las decisiones de manejo puedan ser hechas y probadas inclusive antes que se siembre el cultivo.

2.2 JUSTIFICACIÓN

Este proyecto se soporta en tecnologías ampliamente difundidas en el área de gestión de negocios en organizaciones a nivel mundial como lo son la Minería de Datos y el Procesamiento Analítico en Línea (OLAP); dichas tecnologías han tenido un excelente desempeño y aceptación dentro de las organizaciones, al minimizar los tiempos de análisis de datos, entregando información precisa y oportuna para el apoyo a la toma de decisiones estratégicas [3][4][5].

El aporte innovativo de este proyecto consiste en adaptar y aplicar éstas tecnologías para el análisis de información generada dentro de los viveros automatizados, facilitando la labor de productores e investigadores en la identificación de patrones y técnicas de germinación y cultivo más exitosas. A partir de estos resultados los productores de pequeños y medianos cultivos podrán contar con semillas y plántulas que garantizan a futuro, una producción arbórea viable bajo unas condiciones ambientales predeterminadas. A nivel académico, hay dos aspectos muy importantes del proyecto que cabe destacar. Primero, continuar con la investigación en nuevos campos de aplicación para los Sistemas de Soporte a la toma de Decisiones y todas sus tecnologías subyacentes, área en que la Universidad del Cauca es pionera en la región. Una referencia es el proyecto de grado [6] realizado por los ingenieros Acosta y Muñoz con el uso de Sistemas de Soporte a la toma de Decisiones en ambientes educativos virtuales.

Por otra parte, a pesar que este proyecto es formulado por estudiantes del programa de Ingeniería de Sistemas, la naturaleza del mismo hace necesario un desarrollo interdisciplinario con otras

áreas del conocimiento. Por primera vez en el programa de Ingeniería de Sistemas un proyecto de grado tiene como uno de sus pilares el trabajo interdisciplinario con docentes y estudiantes de los programas de Ingeniería Física e Ingeniería Forestal de la Universidad del Cauca, lo que constituye un gran desafío y a su vez una oportunidad para fortalecer la formación profesional. Es preciso aclarar que el software obtenido en el marco de ésta iniciativa, será probado con el roble; sin embargo, su diseño y desarrollo estará proyectado para ser usado en cualquier tipo de cultivo (sin que se pretenda en este proyecto y por limitaciones de tiempo, demostrar su uso con otros cultivos).

2.3 ANTECEDENTES

2.3.1 INVERNADEROS AUTOMATIZADOS

Control y Monitoreo de Variables Ambientales usando SCADA y PLC: (aplicativo invernadero Universidad de Pamplona - Colombia) [7]. Se desarrolló un sistema SCADA para el control y monitoreo de variables ambientales dentro de un invernadero. El sistema puede controlar la temperatura y humedad dentro del invernadero, y según la programación que se le definida al autómata, él actuará. Por ejemplo si la temperatura sobrepasa el máximo, se activa un ventilador y un extractor para hacer circular aire en el interior, en el caso que la temperatura baje demasiado y pase del mínimo, serían encendidas las resistencias de calefacción; al igual el sistema de riego se activará en caso que la humedad del suelo sea mínima o según se haya programado. Todo el proceso es controlado mediante un PLC que sirve de interfaz con el computador que tiene instalado un software de gestión (OLTP). Esta aplicación tiene tres funciones principales: monitoreo, control y simulación. El panel de monitoreo brinda al operario una visualización completa del comportamiento del sistema en tiempo real. El panel de control como su nombre lo indica se encarga del control de los procesos de regulación del ambiente dentro del invernadero por medio de dispositivos como ventiladores, extractores, resistencias de calefacción, etc. Por ultimo el panel de simulación permite ver en forma animada como transcurren los procesos en tiempo real.

Automatización de Invernaderos Mediante Sistemas de Control Distribuidos Industriales: (Universidad Politécnica de Valencia) [8]. Se desarrolló una aplicación software que se encarga de monitorear y controlar el funcionamiento global del invernadero, gestionar las alarmas, generar registros históricos, etc. Además, la aplicación permite la administración remota vía Internet. El sistema de control distribuido está constituido por 4 subsistemas:

- El bus de campo, es la plataforma física de comunicación entre el autómata programable (PLC) y la periferia descentralizada.
- La periferia descentralizada, se encarga de tomar la información de los sensores (de temperatura, humedad relativa, etc.), digitalizarla y mandarla por la red al PLC. Por otra parte recoge las órdenes de éste y las ejecuta sobre los sistemas de actuación (riego, calefacción, etc.).
- El PLC actúa como unidad de control. A partir de la información de todos los sensores y de la lógica de control, determina las actuaciones a realizar.
- El PC ejecuta las tareas del interfaz de usuario (monitorización, establecimiento de consignas, alarmas, configuración, etc.), trasladando la información del PLC al usuario y viceversa, además del registro de la evolución de las variables en una base de datos accesible a través de Open DataBase Connectivity (ODBC).

Cabe mencionar dos antecedentes adicionales: Laboratorio Remoto para el Control de una Maqueta de Invernadero [9] y Sistema de Control de Humedad, Temperatura y Riego para Invernaderos Industriales [10], que reafirman el estado actual de los procesos de automatización

de invernaderos. El aporte diferenciador de este proyecto, es el desarrollo de una herramienta software para complementar los beneficios de la agricultura protegida y la automatización, con un servicio de soporte a la toma de decisiones en base a datos recolectados dentro del invernadero.

2.3.2 SISTEMAS DE SOPORTE A LA TOMA DE DECISIONES

The National Agricultural Decision Support System: (NADSS) [11] es un proyecto de la Universidad de Nebraska en los Estados Unidos que consiste en una colección de herramientas de soporte a las decisiones basadas en la Web que buscan ayudar a los productores agrícolas norteamericanos en el análisis y la mitigación efectiva de los efectos de las sequías. El NADSS es alimentado con índices de sequías, indicadores climáticos y archivos históricos que son analizados para estimar sequías, frecuencias de inundaciones, duración e intensidad. Las herramientas automáticamente generan mapas y tablas que ayudan a ilustrar el peligro de las sequías en la infraestructura agrícola.

Prairie Agriculture Research Initiative Decision Support System: (PARI DSS) [12] es un sistema de soporte a la decisión que se está desarrollando por el Agriculture and Agri-Food Canada en conjunto con productores, gobierno, universidades e industria. El PARI Decision Support System consiste en un aplicación llamada Farm Smart 2000 que provee tres diferentes niveles de soporte a la decisión de acuerdo a las necesidades del productor. El sistema podrá apoyar la conservación de la producción mediante información sobre la rotación de la misma, la variedad de selección del producto, fertilidad, control de pestes y la selección o modificación de maquinaria. Este sistema se refleja como un sistema de conservación de producción, el cual asegura la sostenibilidad del recurso del suelo, y a su vez, de la agricultura.

Sistema de soporte de decisiones para la producción agrícola de los Valles Cordilleranos Patagónicos: [13] Es un proyecto del Instituto Nacional de Tecnología Agropecuaria (INTA) de Argentina, iniciado a mediados del año 2003 y programado para tres años, fue planteado con el objetivo de disponer de un instrumento que favorezca la toma de decisiones calificadas para el desarrollo agrícola sustentable en los valles cordilleranos patagónicos. Las actividades se llevan a cabo en diferentes valles cordilleranos de la Patagonia Argentina, ubicados en las provincias de Neuquén, Río Negro, Chubut y Santa Cruz. Se busca obtener un producto que muestre en forma organizada y detallada toda la información disponible, y la que se obtendrá, en lo que refiere a zonas cultivadas y zonas más aptas, tipo de cultivos, superficie por tipo de cultivo, volúmenes de la producción primaria, volumen de materia industrializada y los productos obtenidos, costos de producción, canales de comercialización y perspectivas de los distintos productos.

Estos tres proyectos son ejemplos de la aplicación actual de los DSS en la agricultura. Sin embargo son sistemas que por su gran envergadura no responden a la pregunta ¿Cuál es la mejor técnica para cultivar? (que en éste proyecto es el problema que se pretender analizar con herramientas DSS) sino ¿cuando y qué cultivar?. Además, su orientación es hacia la agricultura tradicional de cielo abierto, y no sobre invernadero.

2.4 VIABILIDAD DEL PROYECTO

El proyecto contó con recursos económicos, tecnológicos, de infraestructura y con la asesoría multidisciplinaria calificada para la realización exitosa del mismo, gracias al apoyo brindado por los grupos GTI (Grupo de I+D en Tecnologías de la Información), SEPA (Seminario Permanente de Formación Avanzada del Doctorado de Educación), I+D en Ingeniería Física y TULL (Grupo de Investigaciones para el Desarrollo Rural) de la Universidad del Cauca y al convenio Academic Alliance con Microsoft.

3 CONTRIBUCIÓN A LA SOLUCIÓN

3.1 OBJETIVOS DEL PROYECTO

A continuación se muestran los objetivos del proyecto, conforme fueron aprobados por el Comité de Investigación de la FIET en el documento de anteproyecto

3.1.1 OBJETIVO GENERAL

Desarrollar un Sistema de Soporte a la Toma de Decisiones para los agricultores e investigadores en viveros automatizados, utilizando Bodegas de Datos, técnicas de análisis multidimensional y Minería de Datos para la identificación de patrones y técnicas eficientes de germinación y cultivo del roble.

3.1.2 OBJETIVOS ESPECÍFICOS

- Desarrollar un sistema OLTP Web para la gestión del almacenamiento de datos provenientes del hardware de monitoreo instalado en un invernadero, a través del uso del Proceso Unificado como metodología de desarrollo, una arquitectura multinivel y patrones de software.

- Construir un Sistema de Soporte a la toma de Decisiones a partir de los siguientes componentes:

    1. Implementación de una Bodega de Datos que brinde el almacenamiento y los servicios de consulta al DSS soportada en el motor de base de datos Microsoft SQL Server 2005.

    2. Selección e integración de una herramienta OLAP para el análisis multidimensional y la generación de reportes derivados del modelo de la Bodega de Datos.

    3. Desarrollo de un prototipo de herramienta para el descubrimiento de patrones y tendencias potencialmente útiles dentro del conjunto de datos almacenados en la Bodega de Datos con el uso de técnicas de Minería de Datos soportadas en la plataforma Microsoft Analysis Services.

3.2 RESULTADOS OBTENIDOS

- Sistema OLTP Web, para gestión del almacenamiento de los datos operacionales recolectados en invernaderos automatizados: Código fuente e instaladores.

- Sistema de Soporte a la Toma de Decisiones en viveros automatizados utilizando OLAP y Minería de Datos: Código fuente e instaladores.

- Monografía del trabajo de grado: Este documento que describe el proceso seguido en el desarrollo del proyecto, los problemas que se presentaron, las respectivas soluciones, los aportes más sobresalientes, las conclusiones y recomendaciones para desarrollos futuros. Además, en los anexos se encuentra en forma detallada los casos de uso, dimensiones, tablas de hechos, entre otros.

- Ayuda en línea: Este hiper-documento describe la forma de utilizar el Sistema Web OLTP y el sistema de soporte a la toma de decisiones.

- Artículo: “Sistema OLTP Web utilizado en Cultivos Protegidos”, a publicar en Enlace Informático. Revista de Ciencia y Tecnología del Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Sexta Edición, Diciembre de 2007. ISSN: 1692-374X. http://enlaceinformatico.unicauca.edu.co/ (En proceso de evaluación).

- Artículo “Sistema de Soporte a la Toma de Decisiones para Procesos de Germinación y Cultivo en Invernaderos”. Revista Biotecnología en el Sector Agropecuario y Agroindustrial, Facultad de Ciencias Agropecuarias, Universidad del Cauca. Cuarta Edición, Diciembre de 2007. ISSN: 1909-9959. http://www.unicauca.edu.co/biotecnologia/ (En proceso de evaluación).

- El proyecto representará a la Universidad del Cauca en el mes de abril de 2008 en la eliminatoria nacional del concurso Imagine Cup organizado por Microsoft por un cupo a la final mundial a desarrollarse en Francia en el mes de junio.

PARTE 2 – OLTP

4 MARCO TEÓRICO

4.1 DEFINICIÓN

El término Procesamiento Transaccional en Línea ó OLTP (On-Line Transactional Processing) es utilizado para referirse a sistemas de información que usan bases de datos relacionales para el almacenamiento diario y continuo de multiples operaciones de un negocio en particular por parte de muchos usuarios [14][7].

Este tipo de sistemas frecuentemente maneja grandes volúmenes de datos a un nivel muy detallado; por ejemplo, los movimientos de facturas de ventas en las cajas de un supermercado. Esos movimientos se realizan en transacciones atómicas muy simples y breves, que actualizan datos en tablas de una base de datos relacional. Estas tablas están optimizadas para poder soportar muchas operaciones de cambios en los datos en lapsos muy breves. Las tablas disponen de filas y columnas, por lo que se define como un modelo basado en 2 dimensiones [15].

4.2 CARACTERÍSTICAS

A nivel de organización de datos:

- Datos organizados substancialmente por aplicación
- Focalizado en encontrar requerimientos de aplicaciones específicas para tareas específicas

A nivel de integración de datos:

- Generalmente no integrados.
- Cada tema del negocio, puede tener información en diversos sistemas.
- Diferentes sistemas contienen diferentes tipos de datos.

A nivel de acceso y manipulación de datos por parte de usuarios finales:

- Los usuarios del sistema, son los encargados de ingresar nuevos datos, abren y cierran registros, corrigen datos antiguos, eliminar, modificar, etc.
- Se ejecutan muchas veces las mismas acciones.

A nivel de administradores:

- Manipulación de datos registro a registro.
- Transacciones y/o rutinas de validación a nivel de registro.

A nivel de transacción:

- Se manejan cientos de transacciones por día.
- Si la transacción ha sido exitosa, se asegura la consistencia de ese único pedazo de datos.

A nivel del tiempo:

- Los datos operacionales son altamente volátiles, cambian en la medida en que opera la empresa y sus sistemas computacionales reflejan la operación.
- Puede haber cambios en las base de datos mientras se está consultando en ella.

5 METODOLOGÍA DE DESARROLLO DEL SISTEMA OLTP WEB

5.1 DESCRIPCIÓN GENERAL DE LA METODOLOGÍA

El enfoque metodológico del sistema OLTP Web esta guiado por una instanciación del Proceso Unificado, estructurado en ciclos de desarrollo iterativos e incrementales, teniendo en cuenta las siguientes fases [16]:

5.1.1 PLANEACIÓN Y ELABORACIÓN

El objetivo de esta fase consiste en obtener los requerimientos necesarios para el desarrollo del sistema, así como el estudio de las diversas opciones para conseguir su desarrollo. Durante esta fase se obtuvieron los siguientes artefactos de manera preliminar: Casos de uso de alto nivel, Modelo conceptual, Diagrama de secuencia, Glosario, Arquitectura.

5.1.2 CONSTRUCCIÓN

A partir de las actividades y productos obtenidos en la planeación y elaboración se da inicio a esta fase, cuyo propósito consiste en obtener una versión operativa del sistema en las siguientes etapas:

- Análisis: En esta fase se obtiene una concepción clara de los requisitos a desarrollar en cada ciclo, se desarrolla la descripción de los casos de uso de alto nivel y el modelo conceptual específico para cada ciclo.

- Diseño: Teniendo en cuenta el análisis del sistema se procede a generar una solución lógica del prototipo software que finalmente será implementada, para lo cual se diseñan los casos de uso reales, los diagramas de clases de diseño, de secuencia, de paquetes y de despliegue.

- Implementación: Se implementan los componentes lógicos obtenidos en la etapa de diseño.

- Pruebas: Se definen un conjunto de valores de prueba y se realizan formalmente las pruebas de caja blanca y caja negra con el fin de verificar el resultado de la implementación generada en esta fase.

5.1.2.1 CICLOS DE DESARROLLO

Los ciclos de desarrollo permiten dividir la funcionalidad completa del sistema en tareas más pequeñas que facilitan la labor de construcción del sistema cumpliendo con cada una de las etapas mencionadas anteriormente.

- Ciclo 1 – Base de datos: En este ciclo se modeló e implementó sobre Microsoft SQL Server 2005, la base de datos relacional que proporcionaría la persistencia al sistema OLTP Web. Además se definieron los tipos de usuarios que existirían y los privilegios que tendrían.

- Ciclo 2 – Acceso a datos: En este ciclo se integró el Data Access Application Block del Enterprise Library 3.1 como capa de acceso a datos del sistema OLTP Web.

- Ciclo 3 – Servicios generales: En este ciclo se desarrolló y/o integró los servicios de criptografía (Cryptography Application Block - Enterprise Library 3.1) y envió de correo electrónico.

- Ciclo 4 – Excepciones: En este ciclo se desarrolló e integró una arquitectura para el manejo de la excepciones a nivel de la base de datos, la lógica de servicios y la lógica del negocio por medio del Exception Handling Application Block y el Logging Application Block del Enterprise Library 3.1.

- Ciclo 5 – Sistema OLTP Web: En este ciclo se desarrollaron y/o integraron las funcionalidades básicas de autenticación y autorización, gestión de usuarios, gestión de ubicaciones geográficas, gestión de variables ambientales, gestión de sustratos básicos y compuestos, actualización de datos personales, actualización de contraseña, estructura de navegación por tipo de usuario y la apariencia del sitio (por medio de hojas de estilo en cascada o CSS[1]).

- Ciclo 6 – Trabajo de campo: En este ciclo se desarrolló el componente de Trabajo de Campo que incluye todas las funcionalidades necesarias para gestionar los datos sobre las Fuentes Semilleras, Arboles Productores y Recolecciones.

- Ciclo 7 – Pregerminación: En este ciclo se desarrolló el componente de Pregerminación que incluye todas las funcionalidades necesarias para gestionar los datos sobre el proceso de Pregerminación y los Tratamientos Pre-germinativos.

- Ciclo 8 – Germinación: En este ciclo se desarrolló el componente de Germinación que incluye todas las funcionalidades necesarias para gestionar los datos sobre el proceso de Germinación y las Mediciones Automáticas y Manuales de las variables ambientales en el proceso de Germinación.

- Ciclo 9 – Repique: En este ciclo se desarrolló el componente de Repique que incluye todas las funcionalidades necesarias para gestionar los datos sobre el proceso de Repique, los Tratamientos de Repique, las Plagas y las Mediciones Automáticas y Manuales de las variables ambientales en el proceso de Repique.

5.2 DEFINICIÓN DE REQUERIMIENTOS

En la etapa inicial del proyecto se llevo a cabo la fase de recolección de requerimientos en la que se identificó y documentó las necesidades y expectativas del cliente, para tal efecto se realizaron alrededor de 10 reuniones con miembros de los grupos SEPA (Seminario Permanente de Formación Avanzada del Doctorado de Educación), I+D en Ingeniería Física y TULL (Grupo de Investigaciones para el Desarrollo Rural) y GEA (Grupo de Estudios Ambientales).
Durante estas reuniones algunas veces se presentaron prototipos funcionales a medida que se iba avanzando en los ciclos de desarrollo, con el objetivo de identificar y corregir a tiempo inconformidades con los requerimientos y así evitar que se presentaran sorpresas al momento de hacer la entrega del sistema OLTP Web.

5.2.1 PANORAMA GENERAL

Este proyecto tuvo como objeto crear un sistema OLTP Web para productores e investigadores que lo utilizarán como una herramienta para la gestión de la información en los procesos de germinación y cultivo en un vivero automatizado.

[1] CSS - Cascading Style Sheets

5.2.2 USUARIOS

Las funcionalidades del sistema OLTP Web a las que un usuario tiene acceso dependerán del rol del usuario para tal motivo los usuarios deben pasar por un proceso de autenticación y autorización para ingresar.

Los roles definidos para el sistema OLTP Web son: Administrador, Operario e Investigador y serán explicados más adelante en este apartado.

5.2.3 METAS

La implementación del sistema OLTP Web, tiene como meta principal permitir a productores e investigadores en viveros automatizados, gestionar rápida y eficientemente los datos recolectados manualmente o automáticamente (desde el hardware de monitoreo) durante los procesos de recolección, pregerminación, germinación y repique.

5.2.4 FUNCIONES DEL SISTEMA

Las funciones del sistema son las acciones que debe hacer el sistema; dichas funciones son identificadas y listadas en grupos cohesivos y lógicos, deben establecerse prioridades entre ellas e identificar aquellas que pasan inadvertidas pero que consumen tiempo y otros recursos [16].

Las funciones se categorizan en:

CATEGORÍA	SIGNIFICADO
Evidente	Debe realizarse y el usuario debería saber que se ha realizado.
Oculta	Debe realizarse, pero no es visible para los usuarios. Las funciones ocultas a menudo se omiten erróneamente durante el proceso de obtención de los requerimientos.
Superflua	Es opcional. Su inserción no implica costos significativos ni repercute en otras funciones.

Tabla 1. Categorías de las funciones del sistema

A continuación se listan las principales funciones básicas del sistema y se define su categoría:

REFERENCIA	FUNCIÓN	CATEGORÍA
R.1	Solicitar el login y contraseña del usuario para entrar al sistema.	Evidente
R.2	Conectarse al Sistema Gestor de Base de Datos.	Oculta
R.3	Verificar las restricciones y privilegios del usuario.	Oculta
R.4	Mostrar mensajes de excepciones durante fallos ocurridos de manera inesperada.	Evidente
R.5	Ingresar información referente a una recolección realizada en una fuente semillera.	Evidente
R.5.1	Ingresar información asociada a una fuente semillera.	Evidente
R.5.2	Ingresar información asociada a un árbol productor.	Evidente
R.6	Crear grupos pre-germinativos a partir de una recolección.	Evidente
R.7	Crear un tratamiento pre-germinativo.	Evidente

REFERENCIA	FUNCIÓN	CATEGORÍA
R.8	Consultar los diferentes tratamientos pre-germinativos aplicados a los grupos pre-germinativos.	Evidente
R.9	Crear grupos germinativos a partir de grupos pre-germinativos creados con anterioridad.	Evidente
R.10	Consultar los grupos germinativos existentes.	Evidente
R.11	Consultar los sustratos asociados a los grupos germinativos.	Evidente
R.12	Mostrar gráficamente el porcentaje utilizado en cada grupo de sustrato.	Superflua
R.13	Crear grupos de repique a partir de grupos germinativos creados con anterioridad.	Evidente
R.14	Consultar los grupos de repique existentes.	Evidente
R.15	Consultar los tratamientos de repique asociados a los grupos de repique.	Evidente
R.16	Mostrar gráficamente el comportamiento de las variables ambientales a través del tiempo de un grupo germinativo.	Superflua
R.17	Mostrar gráficamente el comportamiento de las variables ambientales a través del tiempo de un grupo de repique.	Superflua
R.18	Asignar sensores disponibles a grupos germinativos.	Evidente
R.19	Asignar sensores disponibles a grupos de repique.	Evidente
R.10	Ingresar información referente a plagas.	Evidente
R.21	Consultar las plagas que han atacado a un grupo de repique en un periodo de tiempo determinado.	Evidente

Tabla 2. Funciones básicas del sistema OLTP Web

5.2.5 ATRIBUTOS DEL SISTEMA

Los atributos del sistema son las características inherentes del sistema y tienen un conjunto de detalles, los cuales, tienden a ser valores discretos o simbólicos. A continuación se presentan algunos de ellos:

ATRIBUTO	DETALLES Y RESTRICCIONES DE FRONTERA
Tiempo de respuesta	Cuando se consulte cualquier información del sistema, la información referente a ésta aparecerá en un tiempo máximo de 1 segundo.
Metáfora de interfaz	• Navegación fácil, con el uso del teclado y mouse. • Las diferentes consultas serán presentadas a través de varios controles gráficos: grillas descriptivas, cajas de texto, gráficas dinámicas. • Interfaz minimalista con estructura clara y básica, uso de colores tenues, amplio manejo de vínculos o links.
Tolerancia a fallos	El sistema reacciona a los fallos internos atrapándolos por medio de una arquitectura de excepciones transversal a cada una de las capas de la implementación, además registra los fallos en una bitácora de errores para su posterior análisis.

ATRIBUTO	DETALLES Y RESTRICCIONES DE FRONTERA
Facilidad de uso	Menús claros y en la misma secuencia de aparición del proceso a seguir, ayudarán a usuarios inexpertos en el uso de la herramienta, a guiarse y navegar con facilidad.
Plataformas	Microsoft Windows XP Professional, .Net Framework 2.0, Microsoft SQL Server Standard Edition, Enterprise Library 3.1 - May 2007.

Tabla 3. Atributos del sistema OLTP Web

5.3 ANÁLISIS

5.3.1 CASOS DE USO

La Figura 1, presenta un diagrama general de casos de uso para los usuarios del sistema OLTP Web.

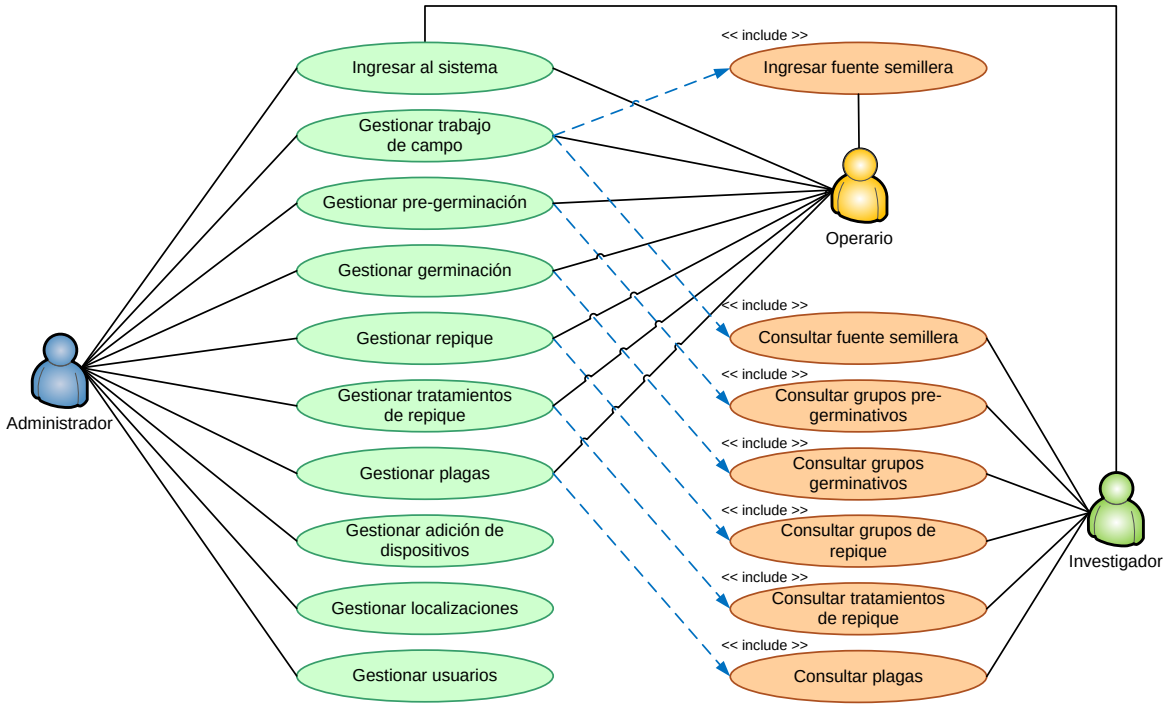


Figura 1. Diagrama general de casos de uso para el Sistema OLTP Web

A continuación se describe el caso de uso “Ingresar Fuente Semillera”, el cual hace parte del caso de uso general “Gestionar Trabajo de Campo”. La descripción de los casos de uso restantes, se encuentra en la sección de Anexos.

5.3.2 CASO DE USO INGRESAR FUENTE SEMILLERA

CASO DE USO: INGRESAR FUENTE SEMILLERA	
Actores: Administrador, Operario.	
Propósito: Ingresar la información referente a una recolección realizada en una fuente semillera.	
Resumen: El administrador o el operario ingresan los datos de la fuente semillera, recolectados en una localidad específica.	
Tipo: Primario	
Referencias Cruzadas: R.5, R.5.1	
CURSO NORMAL DE LOS EVENTOS	
Acción del actor	**Respuesta del sistema**
1. Este caso de uso comienza cuando el usuario (administrador o el operario) da un nombre a una fuente semillera.	
3. El usuario selecciona un departamento.	2. El sistema muestra el nombre de los departamentos.
5. El usuario selecciona un municipio.	4. El sistema muestra el nombre de los municipios.
7. El usuario selecciona un corregimiento.	6. El sistema muestra el nombre de los corregimientos.
9. El usuario selecciona una verada.	8. El sistema muestra el nombre de las veradas.
10. El usuario ingresa las coordenadas planas referentes a la latitud y longitud donde está ubicada la fuente semillera.	
11. El usuario ingresa la altura sobre el nivel del mar de la fuente semillera.	
13. El usuario selecciona el tipo de fuente semillera.	12. El sistema muestra los tipos de fuente semillera que existen.
14. El usuario ingresa las características y observaciones relacionadas a la fuente semillera.	
15. El usuario termina la transacción.	16. El sistema registra la fuente semillera e informa el éxito de la inserción.
CURSO ALTERNO	
Acción del actor	**Respuesta del sistema**
	3, 5, 7, 9, 13. No se puede cargar la información que solicita el usuario. Se muestra un mensaje de error. 16. El cliente no ingresó la información obligatoria. Se informa al usuario para que registre los datos faltantes e intente nuevamente o para que cancela la transacción.
TRAZABILIDAD	
Artefactos anteriores	**Artefactos posteriores**
Artefactos del Análisis: • Diagrama de Casos de Uso	*Artefactos del Análisis:* • Diagrama de Secuencia del Sistema

	• Modelo Conceptual Preliminar *Artefactos del Diseño:* • Caso de Uso Real

Tabla 4. Caso de uso Ingresar Fuente Semillera

5.3.3 MODELO CONCEPTUAL PRELIMINAR

Un paso esencial en el proceso de análisis, es descomponer el ámbito del problema en conceptos u objetos individuales. El modelo conceptual representado a través de diagramas de estructura estática, contribuye a esclarecer la terminología del proyecto y a identificar las relaciones encontradas entre sí [16]. Ver Figura 2.

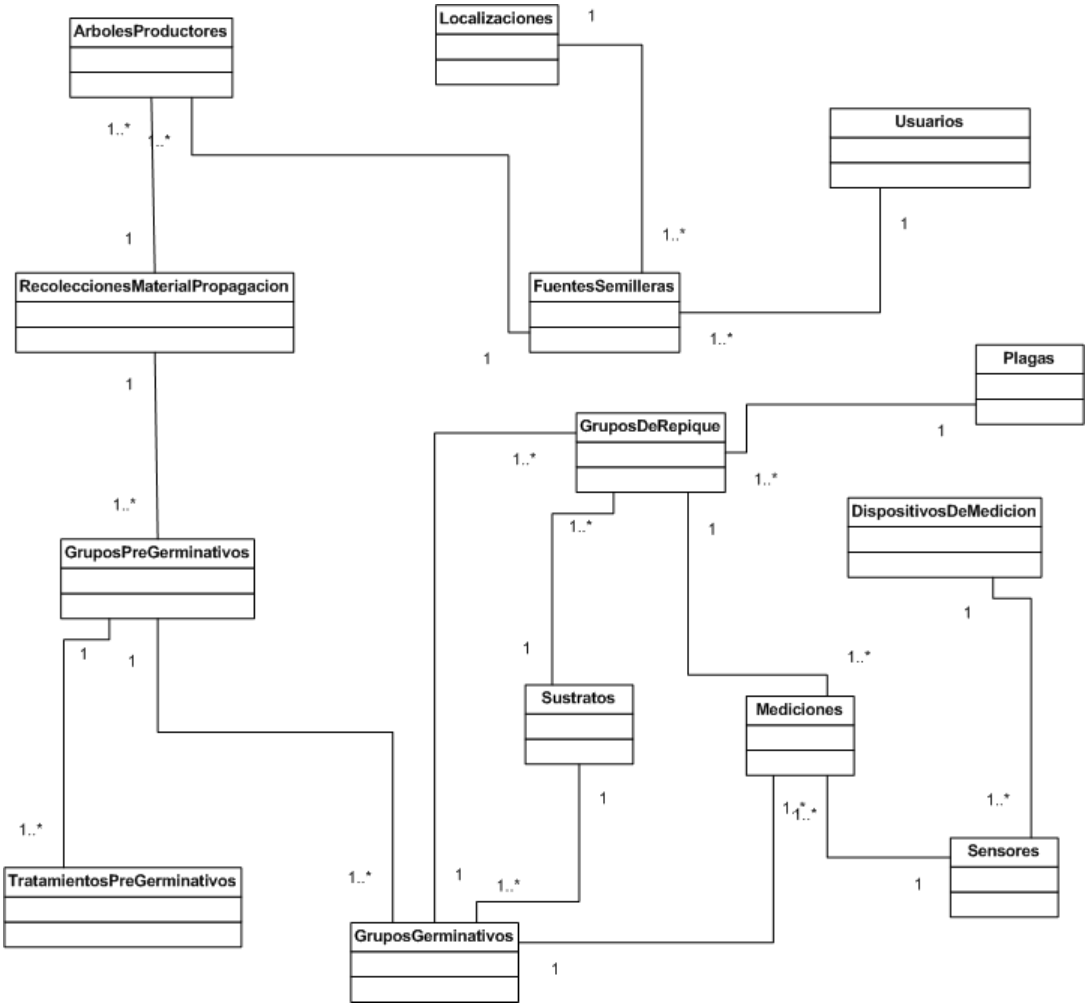


Figura 2. Modelo conceptual preliminar del sistema OLTP Web

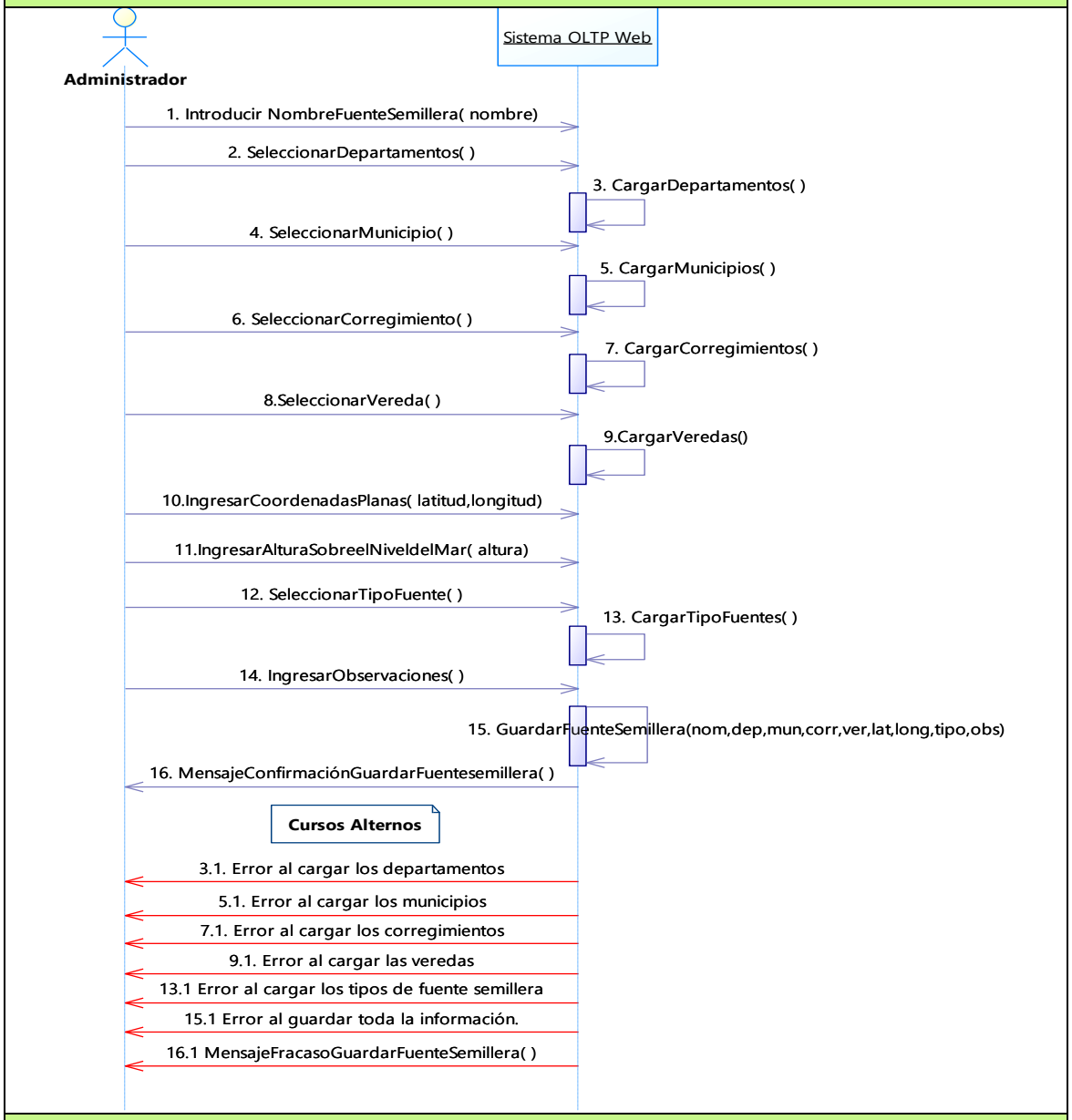
CONCEPTO	DESCRIPCIÓN
ArbolesProductores	Entidad que se encarga de gestionar la información de los árboles productores de semillas.
DispositivosDeMedicion	Entidad que se encarga de gestionar la información de los dispositivos de medición de las variables ambientales.
FuentesSemilleras	Entidad que se encarga de gestionar la información de las fuentes semilleras.
GruposDeRepique	Entidad que se encarga de gestionar la información de los grupos de repique.
GruposGerminativos	Entidad que se encarga de gestionar la información de los grupos germinativos.
GruposPreGerminativos	Entidad que se encarga de gestionar la información de los grupos pregerminativos.
Localizaciones	Entidad que se encarga de gestionar las localizaciones geográficas.
Mediciones	Entidad que se encarga de gestionar las mediciones obtenidas de forma manual o automática.
Plagas	Entidad que se encarga de gestionar la información de las plagas que atacan los grupos de repique.
RecoleccionesMaterialPropagacion	Entidad que se encarga de gestionar la información de las recolecciones de semillas para cada árbol productor.
Sensores	Entidad que se encarga de gestionar la información de los sensores disponibles.
Sustratos	Entidad que se encarga de gestionar la información de la composición de los sustratos en uso.
TratamientosPreGerminativos	Entidad que se encarga de gestionar la información de los tratamientos de los grupos pregerminativos.
Usuarios	Entidad que se encarga de gestionar la información de los usuarios del sistema.

Tabla 5. Descripción del modelo conceptual del sistema OLTP Web

5.3.4 DIAGRAMAS DE SECUENCIA DEL SISTEMA

El siguiente diagrama de secuencia describe, en el curso particular de los eventos del caso de uso "Ingresar Fuente Semillera", los sucesos generados por actores externos y los eventos del sistema.

DIAGRAMA DE SECUENCIA INGRESAR FUENTE SEMILLERA



TRAZABILIDAD

Artefactos anteriores	Artefactos posteriores
Artefactos del Análisis: Caso de uso Ingresar Fuente Semillera	

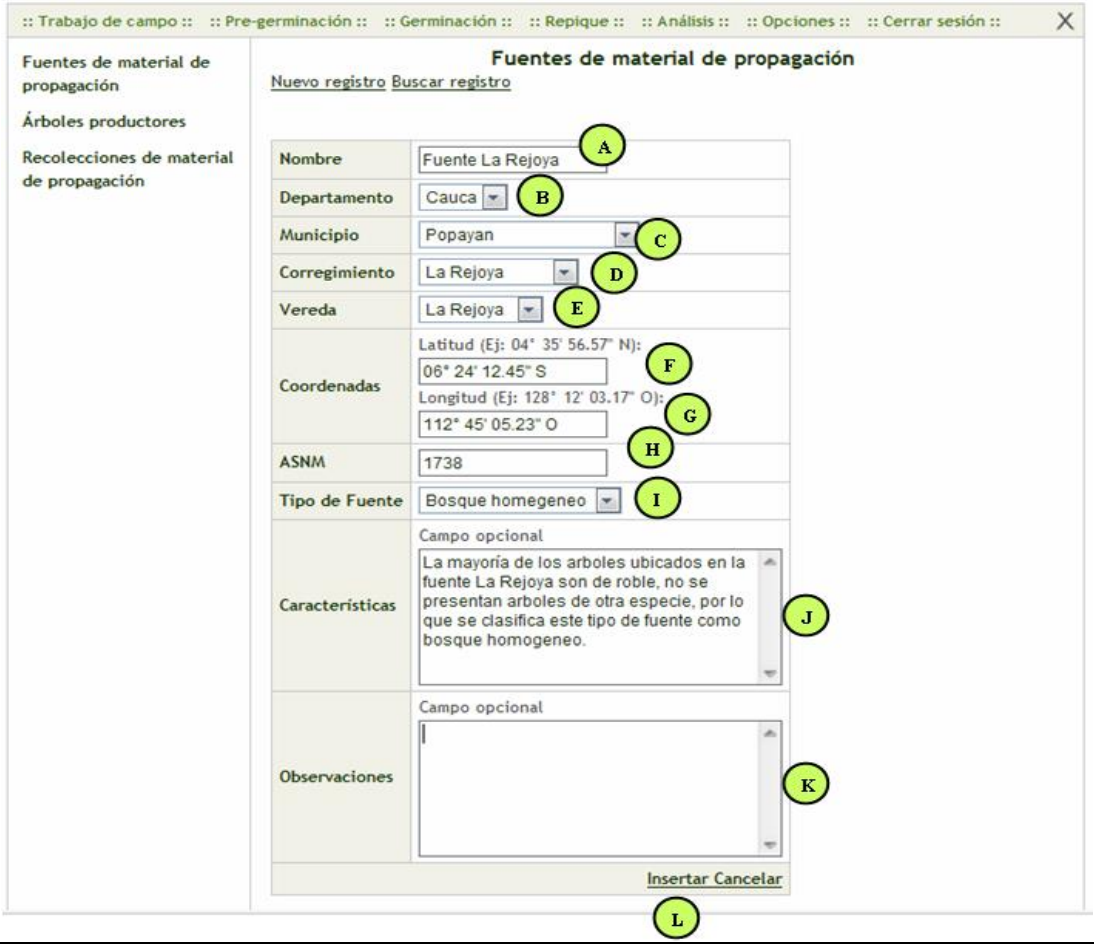
Tabla 6. Diagrama de secuencia caso de uso Ingresar Fuente Semillera

5.4 DISEÑO

5.4.1 CASO DE USO REAL INGRESAR FUENTE SEMILLERA

El siguiente caso de uso real, describe el diseño concreto del caso de uso esencial extendido "Ingresar Fuente Semillera", explicado en el apartado anterior en la etapa de análisis, a partir de una tecnología particular de entrada y salida. La descripción de los casos de uso reales restantes, se encuentra en la sección de Anexos.

CASO DE USO REAL: INGRESAR FUENTE SEMILLERA



CURSO NORMAL DE LOS EVENTOS

Acción del actor	Respuesta del sistema
1. El usuario (administrador o el operario) digita el nombre de la fuente semillera [A].	
	2. El sistema carga el nombre de los departamentos [B].
3. El usuario selecciona un departamento.	
	4. El sistema carga el nombre de los municipios [C].

5. El usuario selecciona un municipio.	
	6. El sistema carga el nombre de los corregimientos [D].
7. El usuario selecciona un corregimiento.	
	8. El sistema muestra el nombre de las veradas [E].
9. El usuario selecciona una verada.	
10. El usuario digita las coordenada planas referentes a la latitud y longitud donde está ubicada la fuente semillera [F] y [G].	
11. El usuario digita la altura sobre el nivel del mar en la que se encuentra la fuente semillera [H].	
	12. El sistema carga los tipos de fuente semillera que existen [I].
13. El usuario selecciona el tipo de fuente semillera.	
14. El usuario ingresa las características y observaciones relacionadas a la fuente semillera [J] y [K].	
15. El usuario da clic en la opción Insertar para poder almacenar la información digitada anteriormente [L].	16. El sistema registra la fuente semillera e informa el éxito de la inserción.

Tabla 7. Caso de uso real ingresar Fuente semillera

5.4.2 DIAGRAMA DE CLASES

El diagrama de clases expuesto en la Figura 3, describe gráficamente las especificaciones, métodos y atributos de las clases de software de la aplicación.

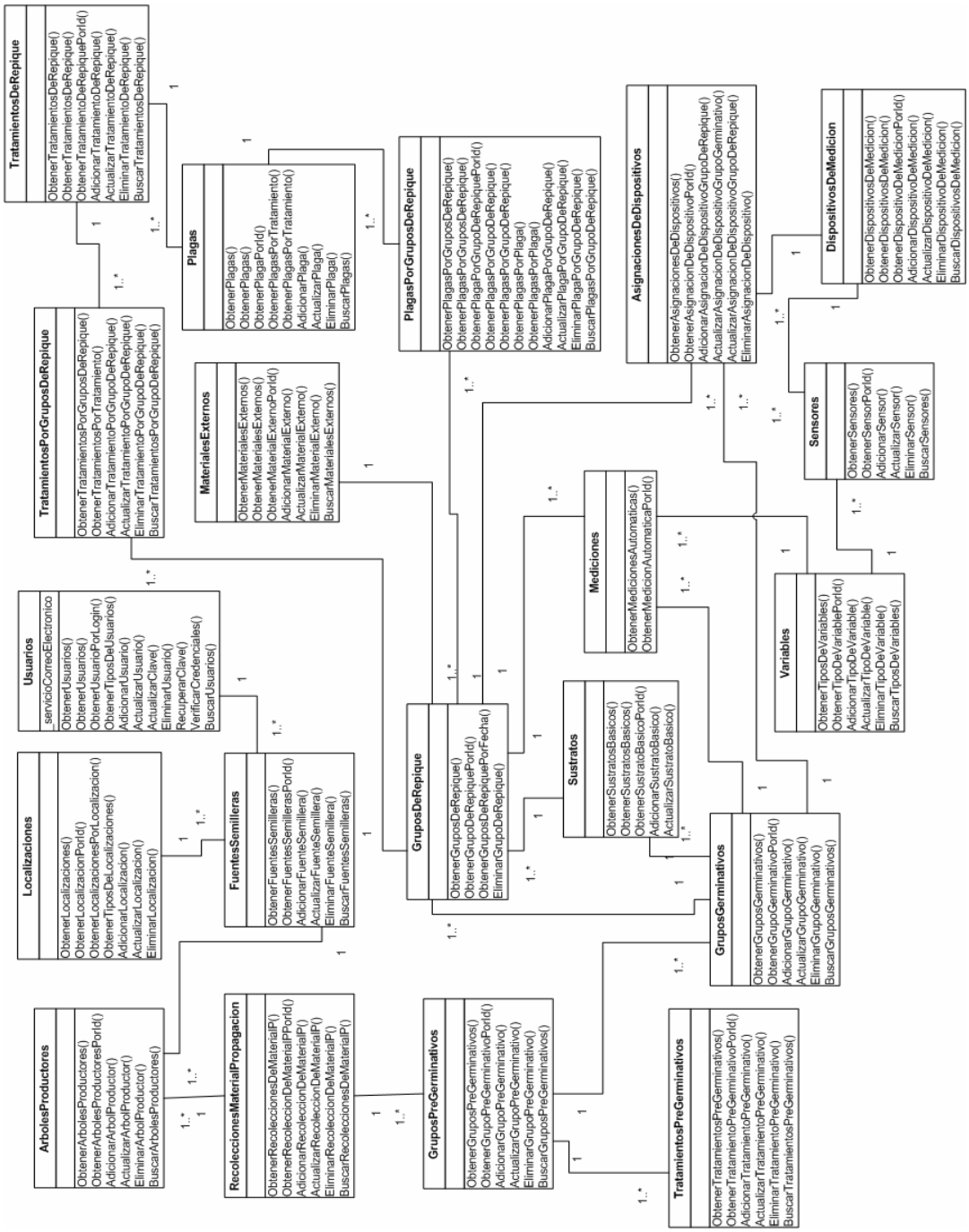


Figura 3. Diagrama de clases del sistema OLTP Web

5.4.2.1 DESCRIPCIÓN DE LAS CLASES

CLASE	FUNCIÓN
ArbolesProductores	Provee los servicios para la gestión de los arboles productores.
AsignacionesDeDispositivos	Provee los servicios para la gestión de las asignaciones de los dispositivos de medición a los grupos germinativos y de repique.
DispositivosDeMedicion	Provee los servicios para la gestión de los dispositivos de medición.
FuentesSemilleras	Provee los servicios para la gestión de las fuentes semilleras.
GruposDeRepique	Provee los servicios para la gestión de los grupos de repique.
GruposGerminativos	Provee los servicios para la gestión de los grupos germinativos.
GruposPreGerminativos	Provee los servicios para la gestión de los grupos pre-germinativos.
Localizaciones	Provee los servicios para la gestión de las localizaciones geográficas y los tipos de localizaciones.
MaterialesExternos	Provee los servicios para la gestión de los materiales externos para los grupos de repique.
Mediciones	Provee los servicios para la gestión de las mediciones manuales y automáticas de los grupos germinativos y de repique.
Plagas	Provee los servicios para la gestión de las plagas.
PlagasPorGruposDeRepique	Provee los servicios para la gestión de las plagas para cada grupo de repique.
RecoleccionesMaterialPropagacion	Provee los servicios para la gestión de las recolecciones de semillas.
Sensores	Provee los servicios para la gestión de los sensores de los dispositivos de medición.
Sustratos	Provee los servicios para la gestión de los sustratos en uso, porcentaje de composición y sustratos básicos.
TratamientosDeRepique	Provee los servicios para la gestión de los tratamientos de repique.
TratamientosPorGruposDeRepique	Provee los servicios para la gestión de los tratamientos de repique para cada grupo de repique.
TratamientosPreGerminativos	Provee los servicios para la gestión de los tratamientos pre-germinativos.
Usuarios	Provee los servicios para la gestión de los usuarios del sistema.
Variables	Provee los servicios para la gestión de las variables ambientales.

Tabla 8. Descripción de las clases del sistema OLTP Web

5.4.3 ARQUITECTURA DE LA APLICACIÓN

La arquitectura empleada en el diseño e implementación del sistema es una variante de la arquitectura presentada por Microsoft Patterns & Practices [17] para el desarrollo de aplicaciones en .NET. Los elementos que componen la arquitectura son los que se presentan en la Figura 4.

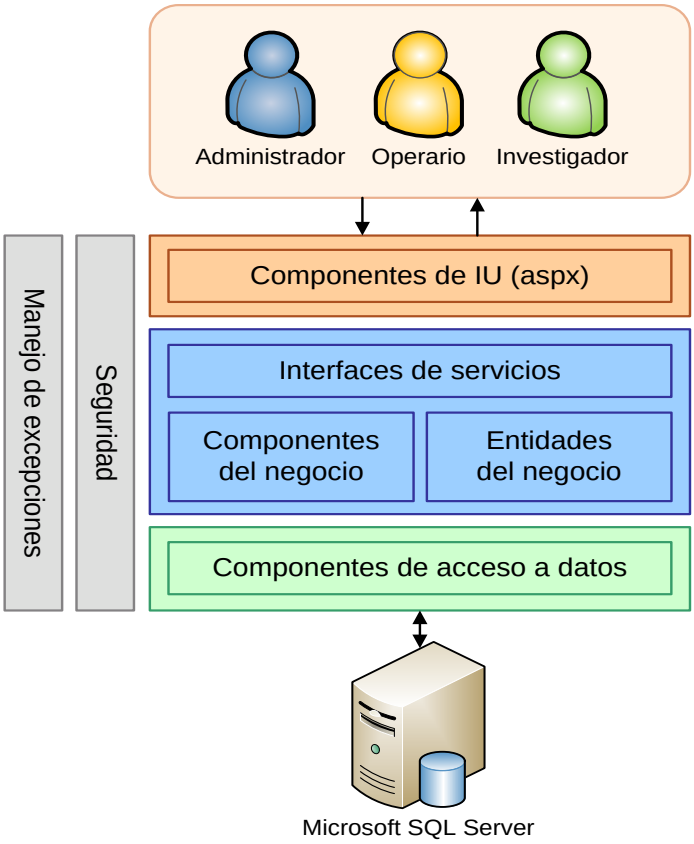


Figura 4. Arquitectura del sistema OLTP Web

A continuación se describe cada uno de los componentes principales de la arquitectura del sistema OLTP Web:

Clientes:

- El Administrador*:* Encargado de la gestión completa del sistema OLTP Web, es responsable de la gestión de usuarios, administración de los dispositivos de medición, de las operaciones de inserción, consulta, eliminación y actualización de la información referente a las etapas del proceso que incluyen la recolección, pregerminación, germinación y repique, entre otras funciones.

- El Operario: personaje activo en el proceso de gestión del OLTP, pero con menos privilegios que el administrador. Se encarga de registrar en el sistema la información recolectada a través de todo el proceso: desde el trabajo de campo hasta la etapa de repique, entre otras funciones.

- El Investigador: Realiza consultas sobre la información que ha ingresado previamente el operario. Al tener poco privilegios sobre la manipulación del sistema, sus acciones son más orientadas al análisis de la información y al control en el registro de los datos.

Componentes de interfaz de usuario (IU): o capa de presentación; es un proyecto ASP.NET independiente. Recopila la lógica de interfaz que permite la visualización, captura y validación de los datos. Esta lógica de interfaz se encapsuló dentro de controles de usuario (ascx). El uso de controles de usuario permitió la reutilización de funcionalidad entre distintas páginas, un mantenimiento de código centralizado y una separación eficiente del contenido y la presentación.

Capa de lógica del negocio: se diseñó como una librería de clases portable, previendo la creación de futuras interfaces a otros entornos distintos al Web sin necesidad de ajustar la lógica del negocio. La capa de lógica del negocio se divide en tres sub-capas: Interfaz de servicios, Componentes del negocio y Entidades del negocio.

- Interfaz de servicios: Exponen como servicios, la lógica del negocio a los clientes de la aplicación, en este caso la aplicación Web desarrollada en ASP.NET.

- Componentes del negocio: Agrupan la verdadera lógica del negocio de la aplicación y operan sobre las entidades del negocio.

- Entidades del negocio: Conjunto de clases orientadas a objetos que representan entidades reales de los proceso de pregerminación, germinación y repique. Creadas en forma automática con el software DataTierGenerator, basado en el modelo de base de datos relacional usado para la persistencia de la información.

Componente de acceso a datos: Como capa de acceso a datos se usó Enterprise Library Data Access Application Block. Este bloque de aplicación simplifica las tareas de desarrollo porque implementa con las mejores prácticas las funcionalidades más comunes para el acceso a datos, tales como gestión de conexiones, ejecución de consultas y procedimientos, transacciones, etc.

A nivel de la base de datos, todas las operaciones CRUD (Create, Retrieve, Update y Delete/Crear, Obtener, Actualizar y Borrar registros) se realizan por medio de procedimientos almacenados debido a dos factores claves: capacidad de mantenimiento y seguridad. Concentrar el código T-SQL (Transact-SQL, Lenguaje de programación de Microsoft SQL Server) fuera de la capa de lógica del negocio permite que se puedan realizar cambios y mejoras en la base de datos de manera transparente a la aplicación y en menor tiempo de lo que tomaría hacerlo directamente en el código de la aplicación.

A nivel de seguridad los procedimientos permiten controlar con precisión el acceso de los usuarios a los datos almacenados garantizado la integridad de la información. En el proyecto, los perfiles o roles de usuario del sistema OLTP Web son a su vez usuarios reales de la base de datos con permisos limitados a la ejecución de ciertos procedimientos y con restricción a la ejecución directa de consultas y acceso a las tablas.

Para el manejo de excepciones, el uso del Enterprise Library Exception Handling Application Block permitió crear una estrategia centralizada de procesamiento de excepciones transversal a toda la aplicación. Por ejemplo, las excepciones que ocurren en la base de datos o en la capa de acceso a datos son atrapadas, registradas en una bitácora y reemplazadas por excepciones del sistema con mensajes más claros y expresivos a los usuarios y que no comprometen la seguridad de la aplicación.

5.4.4 MODELO FÍSICO DE LA BASE DE DATOS

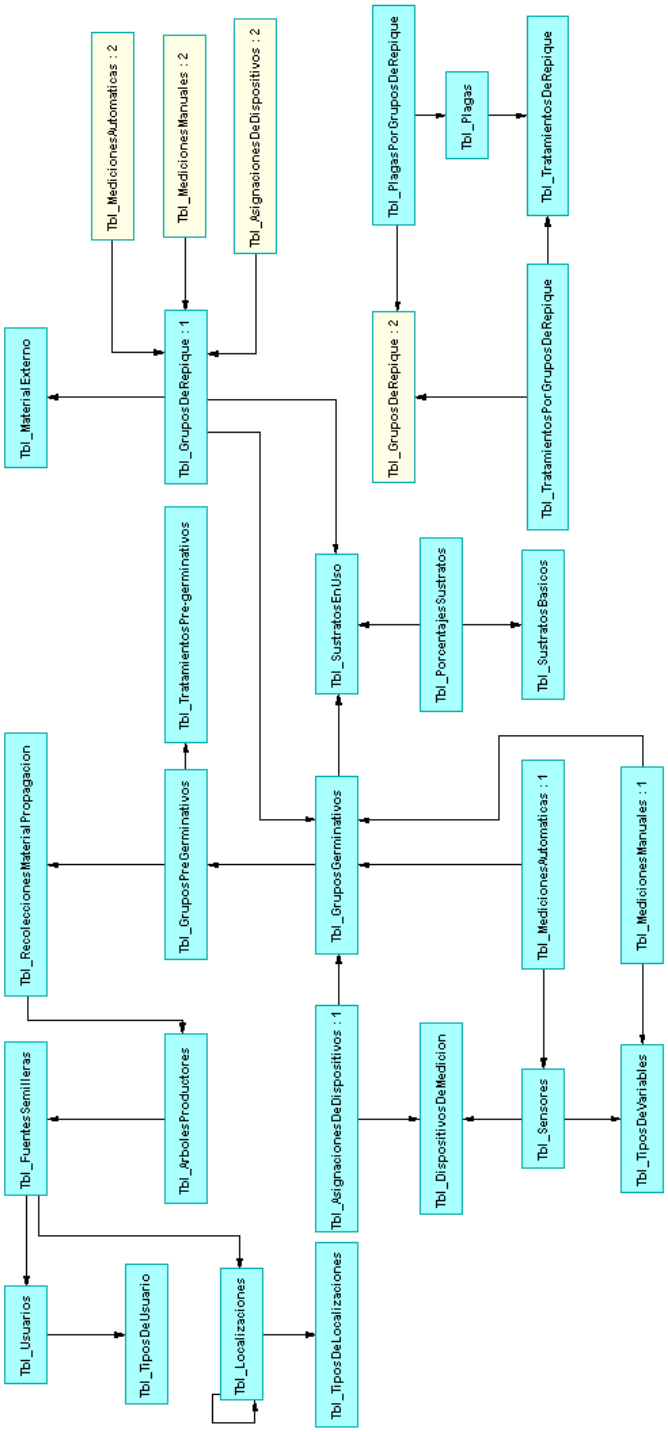


Figura 5. Modelo físico de la base de datos relacional del sistema OLTP Web

Nota: La versión detallada y ampliada del modelo físico se encuentra en la sección de anexos.

5.4.4.1 DESCRIPCIÓN DE LAS TABLAS

TABLA	FUNCIÓN
Usuarios	Contiene la información referente a los usuarios del sistema OLTP Web.
TipoDeUsuario	Contiene la información referente a los tipos de usuario o roles del sistema OLTP Web.
Localizaciones	Contiene la información referente a las localizaciones geográficas en las que se encuentra las fuentes semilleras.
TiposDeLocalizaciones	Contiene la información sobre los tipos de localizaciones geográficas, como lo son departamento, municipio, corregimiento y vereda.
FuentesSemilleras	Contiene la información referente a las fuentes semilleras.
ArbolesProductores	Contiene la información referente a los arboles productores de cada una de las fuentes semilleras.
RecoleccionesMaterialPropagacion	Contiene la información referente a las recolecciones de semillas de cada uno de los arboles productores.
GruposPreGerminativos	Contiene la información referente a los grupos pre-germinativos de cada una de las recolecciones.
TratamientosPre-germinativos	Contiene la información referente a los tratamientos pre-germinativos aplicados a cada grupo pre-germinativo.
GruposGerminativos	Contiene la información referente a los grupos germinativos de cada uno de los grupos de repique.
SustratosBasicos	Contiene la información referente a los sustratos básicos.
PorcentajesSustratos	Contiene la información referente a los porcentajes de sustratos básicos que componente un sustrato en uso.
SustratosEnUso	Contiene la información referente a los sustratos en uso.
MaterialExterno	Contiene la información referente a los materiales externos.
GruposDeRepique	Contiene la información referente a los grupos de repiques de cada uno de los grupos germinativos.
TratamientosDeRepique	Contiene la información referente a los tratamientos de repique.
TratamientosPorGruposDeRepique	Contiene la información referente a los tratamientos de repique aplicados a cada uno de los grupos de repique.
Plagas	Contiene la información referente a las plagas.
PlagasPorGruposDeRepique	Contiene la información referente a las plagas que atacaron a cada uno de los grupos de repique.
TiposDeVariables	Contiene la información referente a los tipos de variables ambientales que pueden ser medidos por los sensores.
Sensores	Contiene la información referente a los sensores de cada uno de los dispositivos de medición.

TABLA	FUNCIÓN
DispositivosDeMedicion	Contiene la información referente a los dispositivos de medición.
AsignacionesDeDispositivos	Contiene la información referente a las asignaciones de los dispositivos de medición a cada uno de los grupos germinativos y/o de repique.
MedicionesAutomaticas	Contiene la información referente a las mediciones obtenidas por los sensores para cada uno de los grupos germinativos y/o de repique.
MedicionesManuales	Contiene la información referente a las mediciones manuales para cada uno de los grupos germinativos y/o de repique.

Tabla 9. Descripción de las tablas de la base de datos del sistema OLTP Web

6 PROBLEMAS Y SOLUCIONES

El principal inconveniente en el desarrollo del Sistema OLTP Web fue la disponibilidad de tiempo para reuniones por parte de los miembros de los grupos de investigación involucrados en esta iniciativa. El objetivo de estas reuniones consistía inicialmente en identificar los requerimientos y posteriormente analizar conjuntamente si los componentes desarrollados cumplían con esos requerimientos, por esas razones postergar o cancelar una reunión significaba un atraso de una o varias semanas en el desarrollo del sistema.

La presentación de prototipos funcionales en las reuniones fue muy importante, primero porque motivo y en cierta forma comprometió aun más a los miembros de los grupos de investigación a colaborar en el desarrollo del proyecto. Los prototipos ayudaron a aclarar dudas sobre algunos requerimientos y a identificar errores y falencias. El problema de esta metodología fue que los prototipos ya tenían toda la funcionalidad implementada en cada una de la capas de la arquitectura, por ejemplo, cuando se identificaba un error, la corrección de este error implicaba un arduo trabajo por que se necesita hacer modificaciones en cada una de la capas. Lo recomendable hubiese sido desarrollar los prototipos con su funcionalidad encapsulada a nivel de componentes de Interfaz de Usuario lo que permite realizar modificaciones directamente. Una vez el prototipo es aprobado, la funcionalidad se extrae del componente de Interfaz de Usuario y se distribuye y/o desarrolla a través de las capas definidas de la arquitectura. Es decir, usar una herramienta basada en la filosofía RAD (Rapid Applications Development, Desarrollo Rápido de Aplicaciones) y luego con los requerimientos bien definidos, realizar el sistema con la arquitectura y patrones adecuados

Ante la necesidad de disminuir el tiempo de desarrollo, el uso de Enterprise Library y sus diferentes application blocks en la construcción del sistema OLTP permitió disminuir notablemente y en unos casos obviar el desarrollo de varios componentes (componentes de acceso a datos, manejo de excepciones y registro de errores) que hubiesen requerido un tiempo importante de desarrollo y prueba.

PARTE 3 – BODEGA DE DATOS

7 MARCO TEÓRICO

7.1 DEFINICIÓN

El concepto de Bodegas de Datos o Data Warehouse surgió de la necesidad de las organizaciones de utilizar los datos históricos para el planeamiento y toma de decisiones. Para tal objetivo era indispensable la consulta de grandes volúmenes de datos almacenados en sus sistemas transaccionales afectando negativamente el rendimiento de las demás transacciones concurrentes. Fue entonces cuando se decidió separar los datos usados para reportes y toma de decisiones de los sistemas transaccionales y diseñar e implementar las Bodegas de Datos para el almacenamiento de estos datos [18]. W. H. Inmon [19] define una Bodega de Datos como 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.

7.2 CARACTERÍSTICAS DE LAS BODEGAS DE DATOS

Las principales características de una Bodega de Datos son [19]: Orientada a la información relevante de la organización, integrada, no volátil y variante en el tiempo. A continuación se explican cada una de ellas:

7.2.1 ORIENTADO A LA INFORMACIÓN RELEVANTE DE LA ORGANIZACIÓN

Dentro de una Bodega de Datos la información se organiza según las áreas de interés para la organización. Por ejemplo, una compañía de seguros organizaría sus datos por cliente, premios, y demandas, en lugar de por diferentes productos (automóviles, vida, etc.). Los datos organizados de esta forma contienen solo la información necesaria para los procesos de soporte a la toma de decisiones.

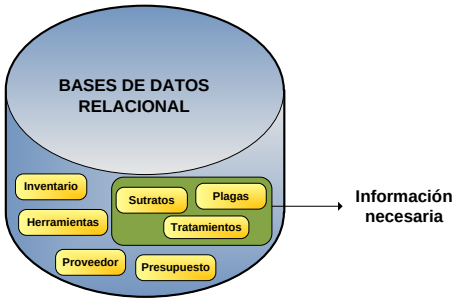


Figura 6. Orientado a la información relevante de la organización (adaptado de [26])

7.2.2 INTEGRADA

Integra los datos provenientes de diferentes fuentes internas o externas a la organización, estas fuentes van desde simples archivos planos hasta sistemas transaccionales. La integración se logra por medio de procesos de normalización de los datos que permiten un formato consistente. Un ejemplo, en la Figura 7, es la representación del sexo de una persona, en una fuente de datos se puede representar con las letras M y F, mientras que en otra fuente se puede notar con las letras H y M, para evitar estas inconsistencias los datos deben filtrarse según un estándar, para el anterior ejemplo se podría definir el estándar de representación con los números 0 y 1.

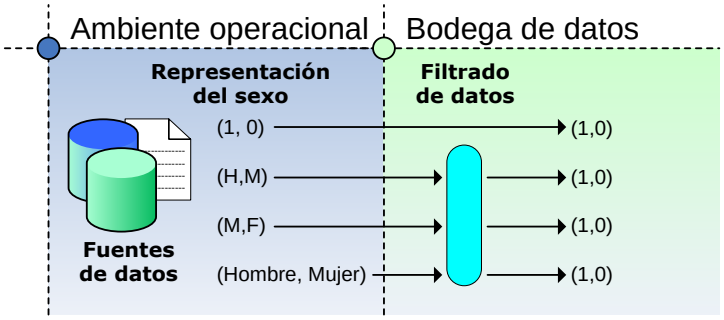


Figura 7. Integración de datos desde sistemas transaccionales hacia la Bodega de Datos

7.2.3 NO VOLÁTIL

Una Bodega de Datos regularmente se alimenta con nuevos datos pero no es muy frecuente que los datos ya almacenados, sean actualizados o modificados.

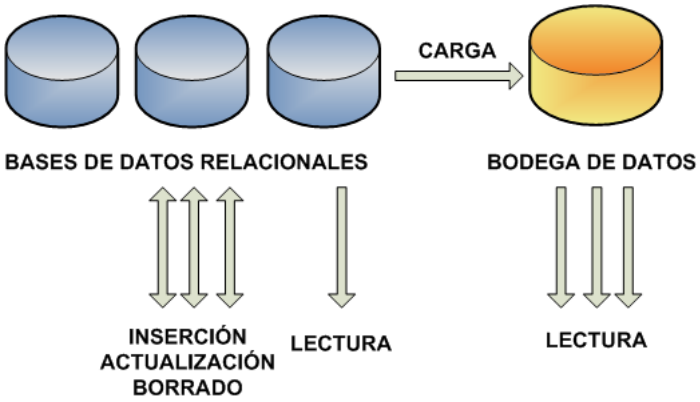


Figura 8. Característica Bodega de Datos – no volátil (adaptado de [26])

7.2.4 VARIANTE EN EL TIEMPO

Los datos en las bodegas de datos son por lo general datos históricos, de poco o ningún uso en el procesamiento operacional, son relativos a un periodo de tiempo y deben ser integrados periódicamente [19].

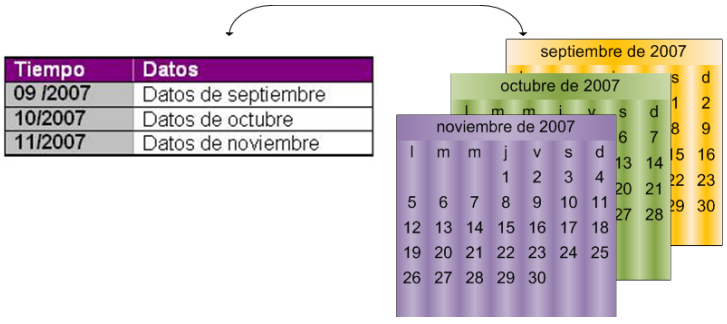


Figura 9. Característica Bodega de Datos – variante en el tiempo (adaptado de [26])

7.3 ARQUITECTURA GENERAL

La arquitectura general de una Bodega de Datos se presenta en la Figura 10 e incluye:

- Datos operacionales. Son los datos de origen para el componente de almacenamiento físico de la Bodega de Datos. Contiene datos primitivos que son permanentemente actualizados, usados por los sistemas operacionales para realizar operaciones transaccionales. Estos datos pueden provenir de fuentes internas y externas.
- Extracción de Datos. Selección sistemática de datos operacionales usados para poblar el componente de almacenamiento físico de la Bodega de Datos.
- Transformación de datos. Procesos para sumarizar y realizar otros cambios en los datos operacionales para cumplir principalmente con los objetivos de orientación a temas e integración. Este proceso incluye corrección de errores, resolución de problemas de dominio, borrado de campos que no son de interés, generación de claves, agregación de información, etc.
- Carga de Datos. Inserción sistemática de datos en el componente de almacenamiento físico de la Bodega de Datos.
- Bodega de Datos. Almacenamiento físico de datos, donde se almacenan datos estratégicos, tácticos y operativos.
- Data Marts. Conjuntos de información de la Bodega de Datos para un departamento o un área especifica.
- Herramientas de Acceso a datos. Herramientas que proveen acceso a los datos de clientes o usuarios finales, con la finalidad de navegar a través de los datos contenidos en la Bodega de Datos.

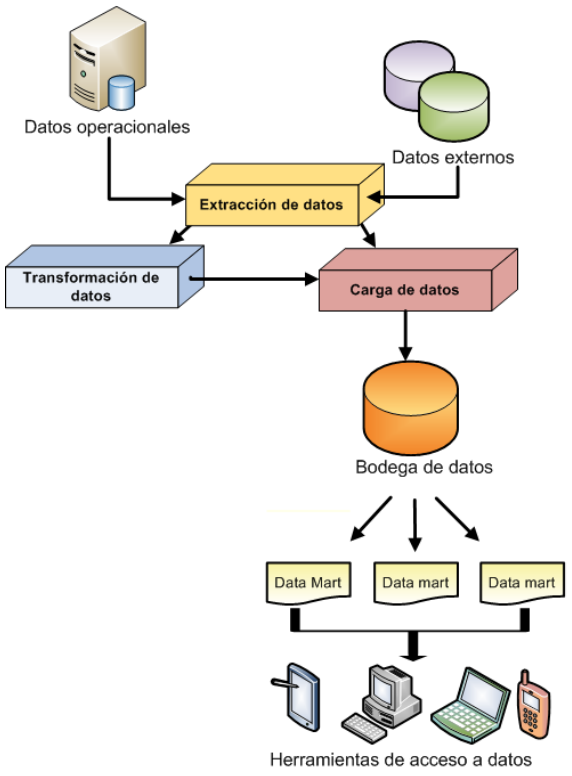


Figura 10. Arquitectura general de una Bodega de Datos

7.4 EL PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS (ETL)

La migración de los datos desde las fuentes operacionales a la Bodega de Datos requiere de procesos para extraer, transformar y cargar los datos, actividad que se conoce como ETL. Estos procesos se originan en la necesidad de reformatear, conciliar y limpiar los datos de origen [18].

La mayoría de los datos de origen son los datos operacionales actuales, aunque parte de ellos pueden ser datos históricos archivados. Si los requerimientos de datos incluyen algunos años de historia es necesario desarrollar tres conjuntos de programas ETL [18]:

Carga Inicial. Se asemeja mucho al proceso de migración de sistemas que se da en las organizaciones cuando pasan, por ejemplo, de sus viejos sistemas operacionales a un producto ERP (Enterprise Resource Planning), y consiste en la extracción, transformación, y carga de los datos existentes en el sistema viejo al nuevo.

Carga Histórica. Este proceso es una extensión de la carga inicial, pero la conversión aquí es un poco diferente porque los datos históricos son datos estáticos. A diferencia de los datos operacionales, los datos estáticos ya se archivaron en dispositivos de almacenamiento offline. Es común que con el transcurso del tiempo se eliminen elementos de datos que ya no sirven, se agreguen nuevos, se modifiquen los tipos de ciertos datos o los formatos de los registros, lo que implica que los datos históricos no necesariamente se puedan sincronizar con los datos operacionales. Por lo tanto los programas de conversión escritos para la carga inicial quizá no sean aplicables a la carga de datos históricos, ya que normalmente requieren cambios.

Carga Incremental. Una vez que la Bodega de Datos está cargada con los datos iníciales e históricos, hay que desarrollar otro proceso para la carga incremental, que se ejecutara mensual, semanal o diariamente. Existen dos formas de diseñar la carga incremental:

- Extraer todos los registros: se extraen todos los registros operacionales, independientemente de los valores que hayan cambiado desde la última carga realizada. En general esta opción no es viable debido al volumen de los datos.

- Extraer Deltas solamente: solo se extraen registros nuevos o registros que contengan valores que cambiaron desde la última carga realizada. Diseñar procesos ETL para extracciones delta es más fácil cuando las fuentes consisten en bases de datos relacionales y se cuenta con una columna timestamp[2] para determinar los deltas.

La mayor parte del trabajo ETL ocurre durante la transformación de los datos, porque es donde se requieren la integración y limpieza de datos. Entre los problemas más habituales en esta etapa se encuentran claves primarias inconsistentes, valores inconsistentes, datos con diferentes formatos, valores erróneos, sinónimos y homónimos.

El paso final en el proceso ETL es la carga de los datos en a la Bodega de Datos. Esta etapa se puede aplicar de dos maneras: insertar filas nuevas en las tablas mediante código escrito a medida, o hacer una carga masiva usando alguna herramienta de importación del DBMS. Este último enfoque es el más eficiente y el más usado en la mayoría de las organizaciones.

[2] Tipo de datos que expone números binarios únicos generados automáticamente en una base de datos. timestamp suele utilizarse como mecanismo para marcar la versión de las filas de la tabla.

8 METODOLOGÍA DE DESARROLLO DE LA BODEGA DATOS

La metodología utilizada para el desarrollo de la Bodega de Datos se basó en la propuesta por Ralph Kimball [14] la cual se ilustra en la Figura 11.

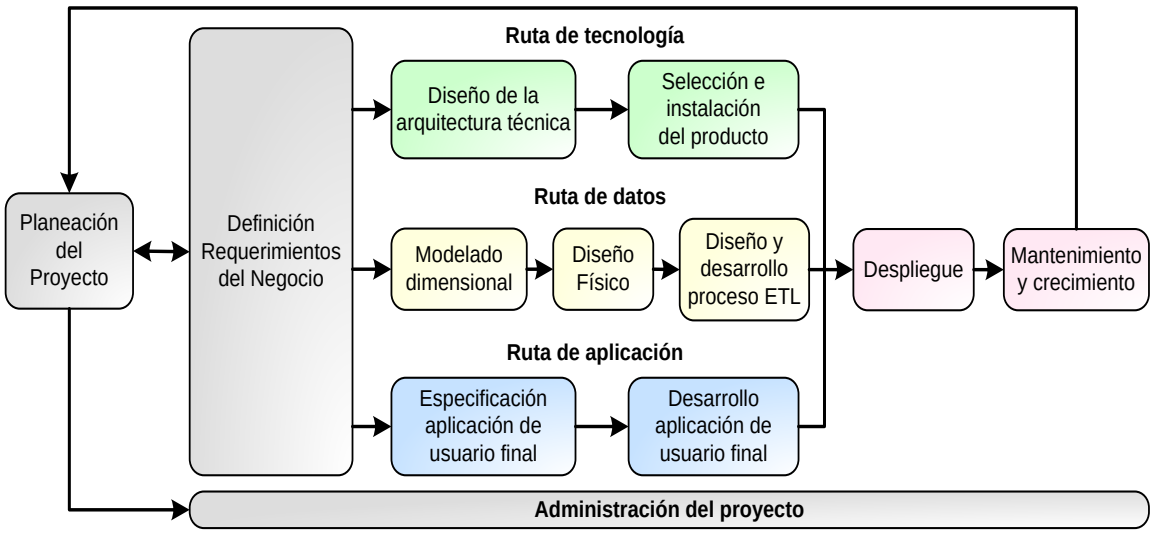


Figura 11. Ciclo de vida para la construcción de una Bodega de Datos según Ralph Kimball

*Planeación del proyecto***.** Permite establecer la definición y el alcance del proyecto de la Bodega de Datos, el cual incluye la valoración y justificación del negocio.

*Definición de requerimientos del negocio***.** Consiste en identificar los factores claves que direccionan las actividades del negocio para determinar efectivamente los requerimientos y traducirlos en consideraciones de diseño.

Modelado dimensional. Esta etapa y la anterior generan el modelo dimensional. Este modelo identifica la granularidad de las tablas de hechos, las dimensiones asociadas, los atributos, las jerarquías y los hechos. Este conjunto de actividades termina con el desarrollo de un plan inicial de agregación.

*Diseño físico***.** Definir las estructuras físicas necesarias para soportar el diseño lógico. Dentro de las primeras actividades que se deben realizar, se encuentra definir un estándar de nombrado y configurar todo el ambiente de la Bodega de Datos.

*Diseño y desarrollo del proceso ETL***.** Esta actividad cuenta con tres pasos importantes: la extracción, la transformación y la carga de datos. El proceso de extracción siempre señala aspectos relacionados con la pureza de los datos que se encuentran en las fuentes de datos operacionales, en este caso en el OLTP.

*Diseño de la arquitectura técnica***.** Las bodegas de datos requieren la integración de numerosas tecnologías. El diseño de la arquitectura técnica establece una visión general de la arquitectura y requiere considerar factores como los requerimientos del negocio, el ambiente técnico actual y las directrices técnicas.

*Selección del producto e instalación***.** Se selecciona la plataforma hardware, el sistema gestor de base de datos, la herramienta ETL y las herramientas de acceso.

Especificación de la aplicación de usuario final. Definir el grupo de usuarios finales de la aplicación.

Desarrollo de la aplicación de usuario final. Se configuran los metadatos de la herramienta y se construyen los reportes especificados.

Despliegue. El despliegue involucra poner la tecnología, los datos y las aplicaciones de usuario final accesibles desde el escritorio de los usuarios.

Administración del proyecto. La administración del proyecto garantiza que las actividades de ciclo de vida dimensional permanezcan sincronizadas y encaminadas. Las actividades de la administración del proyecto se realizan durante todo el ciclo de vida. Estas actividades se enfocan en el monitoreo del estado del proyecto y los controles sobre cambios.

Mantenimiento y Crecimiento. El trabajo de mantenimiento surge después del despliegue inicial de la bodega de datos. Es importante seguir enfocándose en los usuarios del negocio para poder suministrarles soporte y educación acordes a sus necesidades. Después de realizar una priorización de las nuevas demandas de los usuarios, se regresa al inicio del ciclo de desarrollo y se construye sobre lo que ya esta establecido en el ambiente de la bodega enfocándose en los nuevos requerimientos. Esta etapa fue omitida en el desarrollo de la Bodega de Datos de este proyecto por limitaciones de tiempo.

8.1 DEFINICIÓN DEL PROYECTO

Antes de iniciar un proyecto de Bodega de Datos, se identifica el escenario apropiado en el que se encuentra la organización, asegurándose de encontrar demanda; y en caso de no contar con un sponsor o patrocinador del proyecto junto con usuarios deseosos, es mejor posponer el proyecto [14].

La Tabla 10, presenta los escenarios más comunes, en los que puede encontrarse la organización:

ESCENARIO	DESCRIPCIÓN
Demanda de un solo fanático de la empresa	Un ejecutivo de la empresa tiene una visión sobre cómo conseguir mejor acceso a la información principal para tomar la mejor decisión.
Mucha demanda	Un considerable número de ejecutivos de la empresa, manifiestan la necesidad de adquirir mejor información de la que se lleva con los procesos actuales. Este escenario resulta un poco más complejo que el anterior, debido a que se necesita priorizar los requerimientos antes de seguir.
En búsqueda de la demanda	Demanda de un solo jefe de los Sistemas de Información de la empresa. No quiere ser el único sin una Bodega de Datos.

Tabla 10. Escenarios de la organización

Basado en lo anterior, el proyecto se enmarcó en el escenario de *"Mucha demanda"*, ya que varios miembros adscritos al programa de Ingeniería Forestal de la Universidad del Cauca, manifestaron

la necesidad de obtener de una forma más efectiva, la información necesaria para los procesos que realizan, además de tener la posibilidad de efectuar análisis de datos para tomar mejores decisiones.

8.1.1 ANÁLISIS DE REQUERIMIENTOS DE ALTO NIVEL DEL NEGOCIO

Por lo general, los requerimientos de los usuarios afectan cada decisión tomada durante toda la puesta en funcionamiento de la Bodega de Datos y a través de ellos, se determinan que datos deben estar disponibles, como deben ser organizados, desplegados y cuan a menudo serán actualizados [14].

Para un mayor entendimiento de los requerimientos, se recomienda empezar a hablar en términos coloquiales con los usuarios finales y patrocinadores del negocio. Después se sugiere establecer un diálogo un poco más técnico con el personal responsable de los sistemas de información. Finalmente se podrán llevar a cabo reuniones con los DBA´s[3] o expertos operacionales del sistema fuente.

Durante varias reuniones acordadas con los implicados del proyecto, se logró, por medio de sesiones grupales facilitadoras y entrevistas personales e informales, identificar los temas y requerimientos generales más relevantes a desarrollar en la Bodega de Datos.

TEMAS	REQUERIMIENTOS GENERALES
Recolecciones	Poder identificar la fuente semillera y el árbol de mayor producción de semilla.
Pregerminaciones	Obtener los mejores tratamientos pre- germinativos para tratar las semillas.
Germinaciones	Identificar los mejores sustratos de siembra para un grupo de semillas.
Repique	Identificar el mejor sustrato para las plántulas.
Tratamientos de repique	Identificar los mejores tratamientos de repique aplicados a las plántulas.
Plagas	Poder identificar las plagas que atacan las plántulas y su número de incidencia en el tiempo.

Tabla 11. Temas y requerimientos generales del negocio

Durante esta etapa, solo se pudo hacer reuniones con los usuarios del negocio; no se realizó entrevistas a DBA´s ni al personal responsable de los sistemas de información debido a que no existían en el contexto de la organización.

8.1.2 PRIORIZACIÓN DE LOS REQUERIMIENTOS DEL NEGOCIO

Esta técnica es apropiada, cuando se evalúan los temas del negocio a implementar en la Bodega de Datos. Se busca priorizarlos, dependiendo del impacto potencial que tengan sobre el negocio y su viabilidad, en función de disponibilidad de datos, recursos, experiencia, facilidad de desarrollo y despliegue [14].

[3] Siglas en inglés para Data Base Administrator o administrador de la base de datos.

Dicha priorización se diagrama en un cuadrante de análisis, tratando de empezar la implementación, con la selección de los temas encontrados en el cuadrante de alto impacto en el negocio y alta viabilidad (cuadrante D). Por lo general los temas ubicados en el cuadrante A se descartan.

En la Figura 12, se muestra la priorización de los temas del negocio identificados en el proyecto.

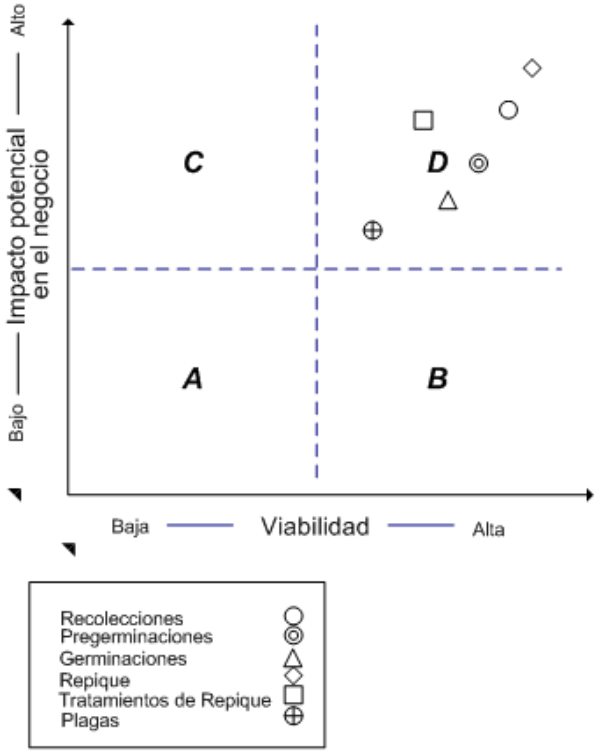


Figura 12. Cuadrante de análisis para la priorización de los temas del negocio

8.2 PLANEACIÓN DEL PROYECTO

Durante la planeación del proyecto, se describen consideraciones claves para obtener en conjunto, el personal del proyecto y el plan del proyecto.

8.2.1 IDENTIDAD DEL PROYECTO

"Bodega de Datos para viveros automatizados", nombre establecido con el propósito de crear identidad con el proyecto; acrónimo GREENDSS_DW.

8.2.2 PERSONAL EN EL PROYECTO

Durante el ciclo de vida del proyecto, son requeridas varias personas con habilidades y destrezas, que a través de roles establecidos, cumplen responsabilidades y desarrollan tareas específicas [14].

Los siguientes roles fueron identificados en el proyecto:

- Patrocinador del proyecto: es el propietario del proyecto y a menudo tiene responsabilidad financiera en él. Este rol fue desempeñado por la Universidad del Cauca.

- Director del proyecto: responsable de la dirección del proyecto. Coordina y monitorea el progreso de tareas y actividades. Tiene un claro conocimiento de los requerimientos del negocio. Este rol fue desempeñado por el magíster Carlos Alberto Cobos Lozada.

- Analista del negocio: responsable de encabezar las actividades de definición de requerimientos y representación de dichos requerimientos en el modelado dimensional. este rol fue realizado por el director y desarrolladores del proyecto Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

- Modelador de datos: responsable de llevar a cabo el análisis detallado de datos y desarrollar el modelo dimensional. Rol asignado a Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

- Administrador de la Bodega de Datos: encargado de trasladar los modelos dimensionales a estructuras de tablas físicas. Asegura la integridad de los datos, la disponibilidad de la bodega y el desempeño de la misma. Los encargados de este rol fueron Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

- Diseñador y desarrollador del proceso de ETL: responsable del diseño y desarrollo del proceso de extracción, transformación y carga de los datos provenientes del OLTP a la Bodega de Datos. Roles ejecutados por Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

- Desarrollador de aplicaciones de usuario final: crea y mantiene actualizado la aplicación de usuario final. Se encarga de la integración de la herramienta OLAP a la Bodega de Datos. Roles ejecutados por Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

- Educador de la Bodega de Datos: educa en el contexto de la Bodega de Datos a usuarios finales. Roles ejecutados por Jimena Adriana Timaná Peña y Rene Valencia Vallejo.

8.2.3 PLAN DEL PROYECTO

El plan del proyecto expuesto en la Tabla 13, lista todas las tareas involucradas en el diseño, el desarrollo y despliegue de la Bodega de Datos, además de los roles y responsabilidades involucrados en cada tarea. El nivel de responsabilidad del rol en cada tarea, está designado en la Tabla 12.

LEYENDA
Responsable principal de la tarea = ●
Involucrado en la tarea = ○
Provee entradas a la tarea = ◗
Informado del resultado de la tarea = ❑
Involucrado opcional en la tarea = ▲

Tabla 12. Nivel de responsabilidad del rol en cada tarea

	Patrocinador del proyecto	Director del proyecto	Analista del negocio	Modelador de datos	DBA	Diseñador y desarrollador ETL	Desarrollador aplicaciones usuario final	Educador de la Bodega de Datos
ADMINISTRACIÓN DEL PROYECTO Y REQUERIMIENTOS								
Definición del proyecto								
1. Identificar escenario de la organización.			●					
2. Realizar las entrevistas programadas.	◗	○	●					
3. Análisis de los resultados obtenidos en las entrevistas.	❑	●	●					
4. Identificación y priorización de los temas de interés para la Bodega de Datos.		○	●					
Planeación del proyecto								
1. Establecer la identidad del proyecto.			●					
2. Identificar los roles del proyecto.		○	●					
3. Definir el plan del proyecto.		○	●					
DISEÑO DE DATOS								
Modelo dimensional								
1. Construcción de la matriz bus.		❑	○					
2. Diseñar los Data Marts.		○	●	●				
3. Crear el diagrama general de la Bodega de Datos		○	●	●				
4. Crear los diagramas de las tablas de hecho.		❑	○	●				
5. Identificar los detalles de las tablas de hecho.		❑	○	●				
6. Identificar los detalles de las tablas de dimensión.		❑	○	●				
IMPLEMENTACIÓN DE LA BODEGA DE DATOS								
1. Creación de objetos dimensionales y tablas físicas.		❑	◗	◗	●	○		
2. Implementación de los Data Mart.		❑	◗	◗	●	○		
3. Instalar la base de datos.		❑	◗	◗	●	○		
DISEÑO E IMPLEMENTACIÓN DEL PROCESO DE ETL								
1. Diseño del proceso de extracción de datos.		▲	◗	◗	◗	●		

2. Diseño del proceso de transformación de datos.		▲	○	○	◗	●		
3. Diseño del proceso de carga de datos.		▲	○	○	◗	●		
4. Creación del paquete maestro del proceso de ETL.		❐	○	○		●		
5. Ejecución del paquete maestro del proceso de ETL.		❐	○	○	○	●		
DESARROLLO DE LA APLICACIÓN DEL USUARIO FINAL								
1. Selección de la herramienta OLAP.		○					●	
2. Integración de la Bodega de Datos y la herramienta OLAP.		❐	○	○		○	●	
3. Creación y gestión de consultas analíticas.		❐	○	○			●	
4. Creación de reportes estáticos.		❐	○	○			●	
5. Elaborar documentación.		❐	▲	▲		▲	●	
DESPLIEGUE								
1. Desarrollar la estrategia de educación inicial para el usuario.	▲	▲	○	▲	▲	▲	●	●
2. Definir el plan de entrega de los productos.	❐	●	●	●	●	●	●	●

Tabla 13. Plan del proyecto

8.3 MODELADO DIMENSIONAL

Gran parte del proceso de gestión del modelado dimensional, está en establecer una buena comunicación entre las personas involucradas en el proyecto. Según Ralph. Kimball, la visualización es esencial durante dicho trabajo de comunicación y recomienda la elaboración de cuatro herramientas gráficas para facilitar el modelado [14]. A continuación se describen las herramientas obtenidas, a saber: matriz bus, diagrama de la tabla de hechos, detalle de la tabla de hechos y detalle de las dimensiones.

8.3.1 MATRIZ BUS

Es usada como ayuda de representación durante las reuniones con diseñadores, administradores y usuarios finales, con el propósito de brindar una vista de los alcances que llegará a tener la Bodega de Datos.

Para su elaboración, es esencial identificar los Data Marts o subconjuntos especializados de información de la bodega de datos y las dimensiones conformadas (dimensiones que tienen el mismo contexto, en cada una de las tablas de hechos a la que se encuentren vinculadas) [14]

Data Marts identificados:

- Recolecciones: además de obtener y almacenar información referente al registro de las actividades que se hacen en las zonas productoras, este Data Mart permite identificar la fuente semillera y el árbol de mayor producción de semillas.

- Pregerminaciones: además de llevar un registro y control sobre los grupos pregerminativos conformados a partir de las recolecciones de semillas, este Data Mart tiene como objetivo,

identificar los mejores tratamientos pre- germinativos para tratar las semillas e identificar el grupo pregerminativo con mayor porcentaje de semillas viables (semillas que sobreviven después de un tratamiento específico).

- Germinaciones: a través de este Data Mart, se puede Identificar los mejores sustratos de siembra para un grupo de semillas y conocer la cantidad de semillas germinadas dependiendo de la posición y sistema de siembra utilizado.

- Repiques: con este Data Mart, además de conocer el número de plántulas que ingresaron y salieron viables del proceso de repique, se puede identificar al grupo de repique con el mejor porcentaje de viabilidad de plántulas.

- Tratamientos de repique: a través de este Data Mart se pueden Identificar los tratamientos de repique aplicados a las plántulas con su respectiva duración.

- Plagas: este Data Mart permite identificar las plagas que atacan las plántulas y su número de incidencia en el tiempo.

Dimensiones conformadas identificadas:

- Recolecciones: almacena información descriptiva de la recolección de semilla.

- Fuentes semilleras: almacena información descriptiva del sitio donde se presenta las recolecciones de semilla y de los árboles productores.

- Grupos pregerminativos: almacena información descriptiva sobre los grupos de semillas que se encuentran en un proceso de pregerminación.

- Grupos germinativos: almacena información descriptiva sobre los grupos de semillas que se encuentran en un proceso de germinación.

- Grupos de repique: almacena información descriptiva sobre los grupos de plántulas que se encuentran en procesos de repique.

- Sustratos en uso: almacena información descriptiva de los sustratos que se encuentran en uso en los procesos de germinación y repique.

- Tratamientos pregerminativos: almacena información descriptiva de los tratamientos que se deben aplicar a los grupos de pregerminativos.

- Tratamientos de repique: almacena información descriptiva de los tratamientos que se deben aplicar a los grupos de repique.

- Plagas: almacena información descriptiva sobre las plagas que afectan a los grupos de repique.

- Fecha: almacena las fechas de los diferentes eventos que acontecen.

- Hora: almacena el tiempo de los diferentes eventos que acontecen.

- Geografía: subdimensión. almacena información descriptiva de las localizaciones geográficas en las cuales se ubican las fuentes semilleras.

La matriz bus fuerza la identificación de todos los Data Marts que posiblemente se construirán y todas las dimensiones implicadas con esos Data Marts. La Tabla 14, establece la relación encontrada entre las áreas del negocio o Data Marts (filas) y las dimensiones (columnas). La intersección es marcada cuando una dimensión existe para un Data Mart.

	Recolecciones	Fuentes semilleras	Grupos pre-germinativos	Grupos germinativos	Grupos de repique	Sustratos en uso	Tratamientos pregerminativos	Tratamientos de repique	Plagas	Fecha	Hora
Recolecciones	✓	✓								✓	✓
Pre-germinaciones	✓	✓	✓			✓	✓			✓	✓
Germinaciones	✓	✓	✓	✓		✓				✓	✓
Repiques	✓	✓	✓	✓	✓					✓	✓
Tratamientos de repique	✓	✓	✓	✓	✓			✓		✓	✓
Plagas	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓

Tabla 14. Matriz bus (Data Marts vs Dimensiones)

8.3.2 DIAGRAMAS DE LAS TABLAS DE HECHOS

Es la visión general de todas las dimensiones que han sido identificadas para el negocio. Cada diagrama se identifica con un nombre, establece la granularidad del hecho, muestra y conecta todas las dimensiones asociadas. En este proyecto, se ha realizado un diagrama de tabla de hechos por cada Data Mart identificado.

La Figura 13 describe a alto nivel el Data Mart Recolecciones, el cual responderá preguntas relacionadas con las recolecciones de semillas, la fecha y hora de la recolección, la fuente semillera y la localización geográfica.

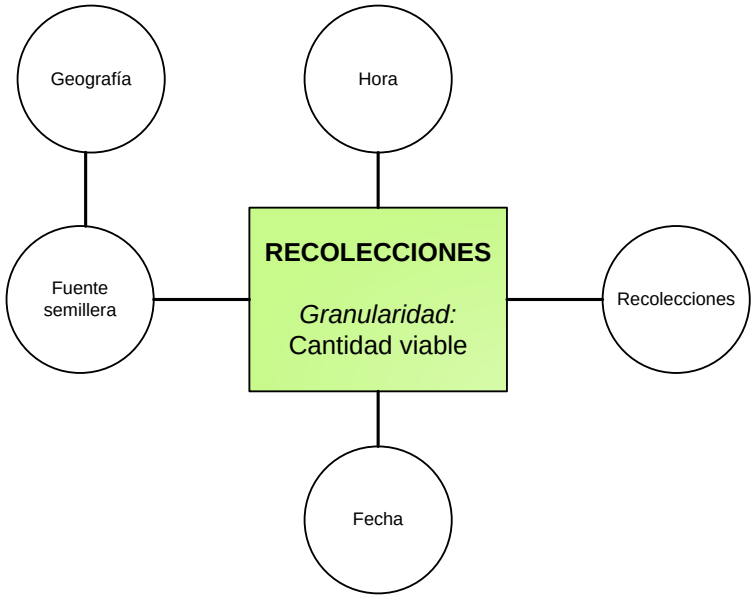


Figura 13. Diagrama tabla de hechos – Data Mart Recolecciones

La Figura 14 describe a alto nivel el Data Mart Pregerminaciones, el cual responderá preguntas relacionadas con los grupos Pregerminativos, los tratamientos aplicados, la duración del proceso de pregerminación y su trazabilidad.

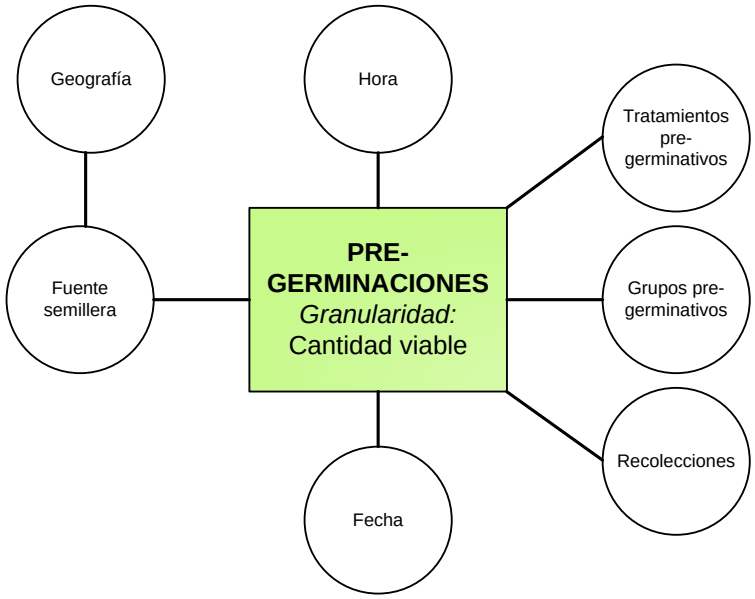


Figura 14. Diagrama tabla de hechos – Data Mart Pregerminaciones

La Figura 15 describe a alto nivel el Data Mart Germinaciones, el cual responderá preguntas relacionadas con los grupos Germinativos, los sustratos utilizados, la duración del proceso de germinación y su trazabilidad.

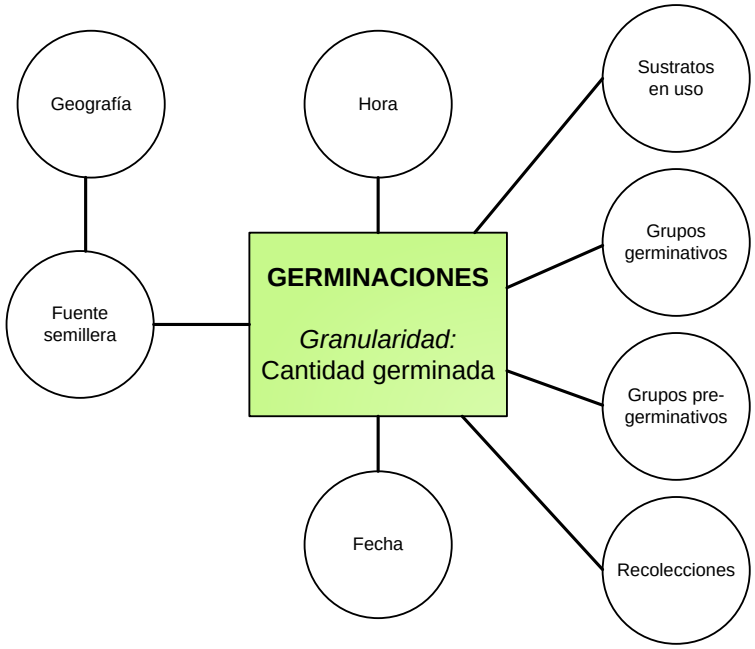


Figura 15. Diagrama tabla de hechos – Data Mart Germinaciones

La Figura 16 describe a alto nivel el Data Mart Repiques, el cual responderá preguntas relacionadas con los grupos de repique, los sustratos utilizados, la duración del proceso de repique y su trazabilidad.

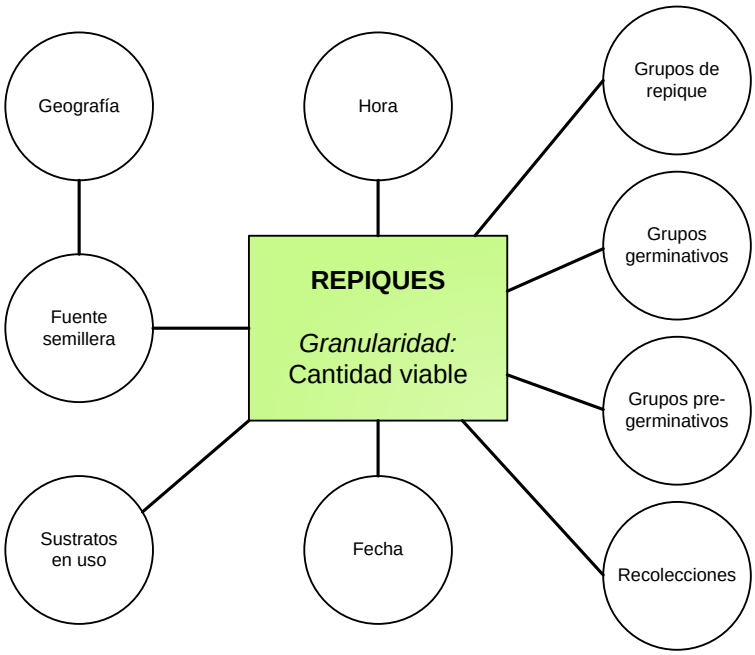


Figura 16. Diagrama tabla de hechos – Data Mart Repiques

La Figura 17 describe a alto nivel el Data Mart Tratamientos de Repique, el cual responderá preguntas relacionadas con los tratamientos aplicados a los grupos de repique, la duración del tratamiento y la trazabilidad del grupo de repique.

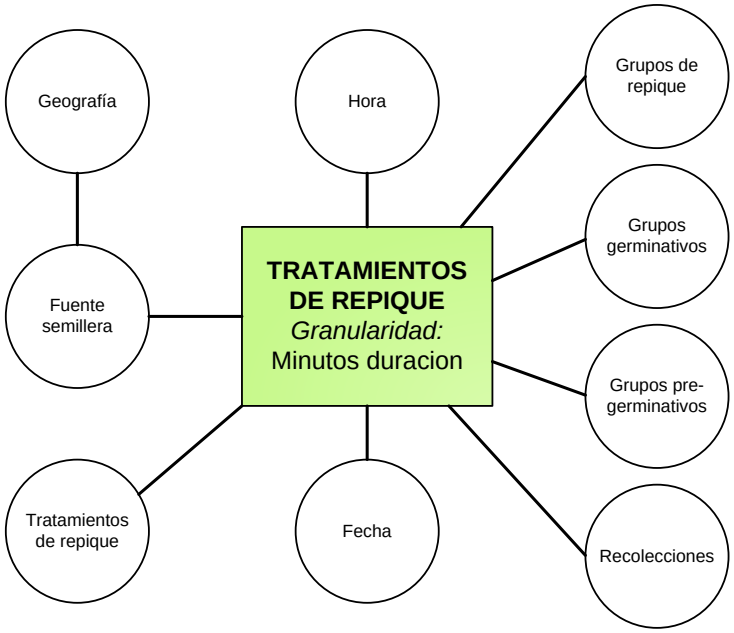


Figura 17. Diagrama tabla de hechos – Data Mart Tratamientos de repique

La Figura 18 describe a alto nivel el Data Mart Ataques Plagas, el cual responderá preguntas relacionadas con los ataques de plagas a los grupos de repique, los tratamientos aplicados, la duración del ataque y la trazabilidad del grupo de repique.

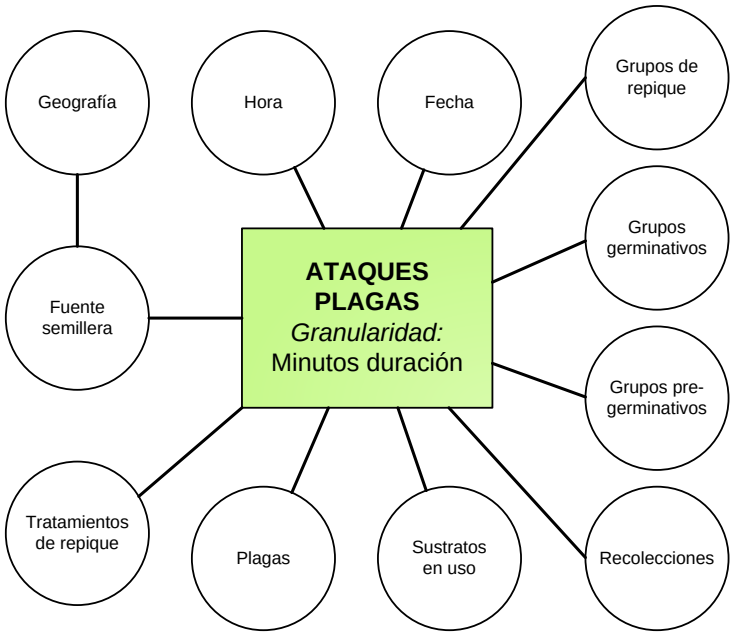


Figura 18. Diagrama tabla de hechos – Data Mart Ataques de plagas

8.3.3 DETALLES DE LAS TABLAS DE HECHOS

Provee la lista completa de los hechos disponibles a través de la tabla de hechos. Muestra llaves foráneas, hechos básicos y hechos derivados. A continuación se muestra los detalles para cada Data Mart desde la Figura 19 hasta la Figura 24.

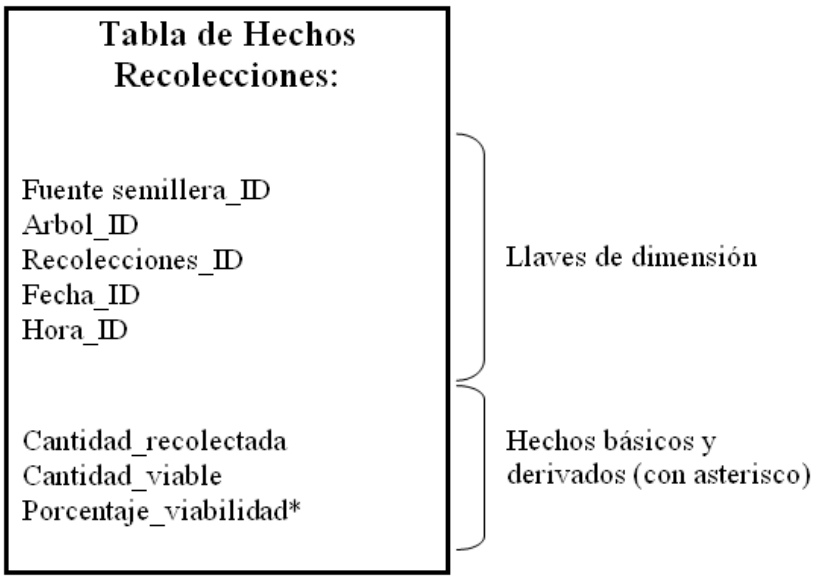


Figura 19. Detalles tabla de hechos – Data Mart Recolecciones

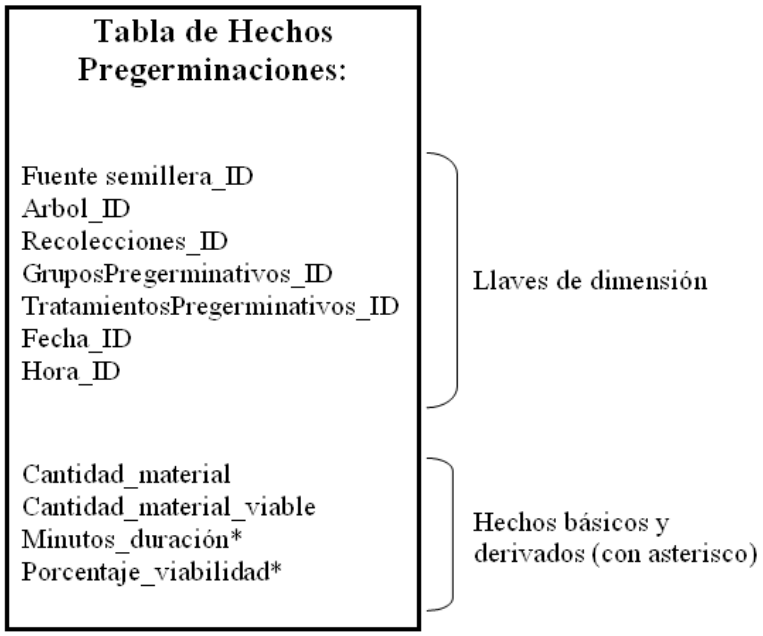


Figura 20. Detalles tabla de hechos – Data Mart Pre-germinaciones

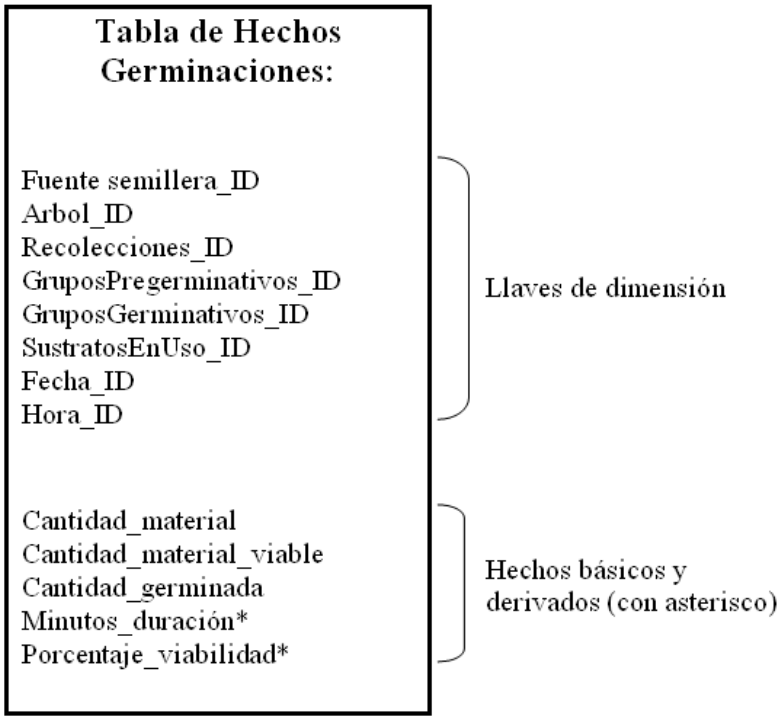


Figura 21. Detalles tabla de hechos – Data Mart Germinaciones

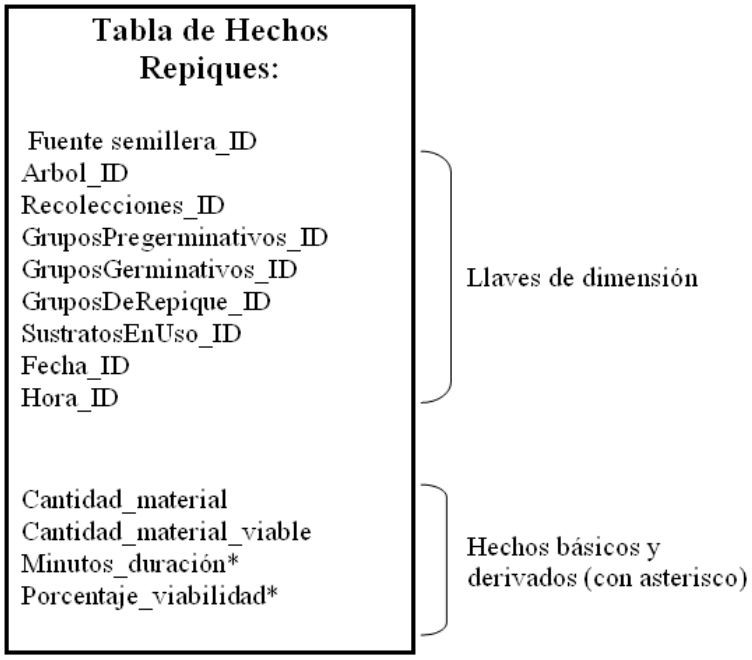


Figura 22. Detalles tabla de hechos – Data Mart Repiques

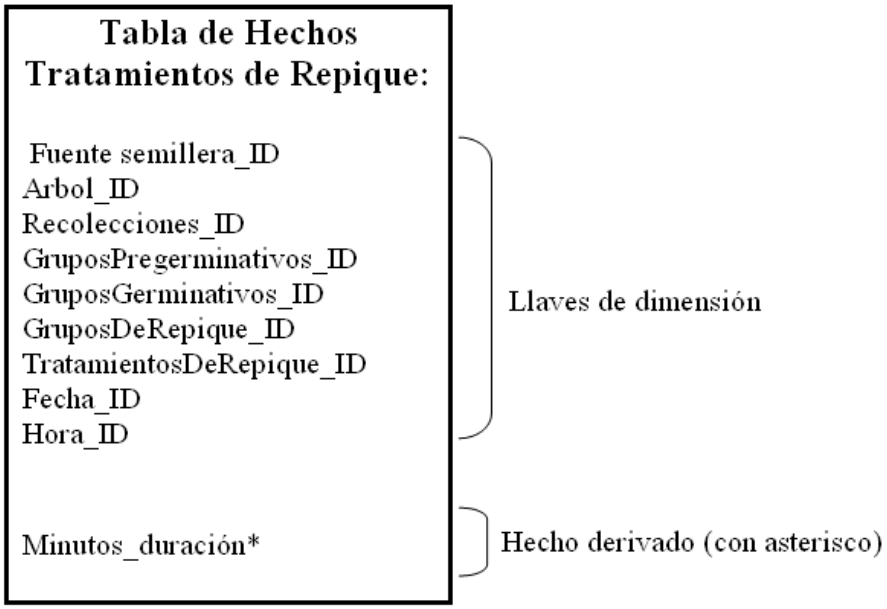


Figura 23. Detalles tabla de hechos – Data Mart Tratamientos de repique

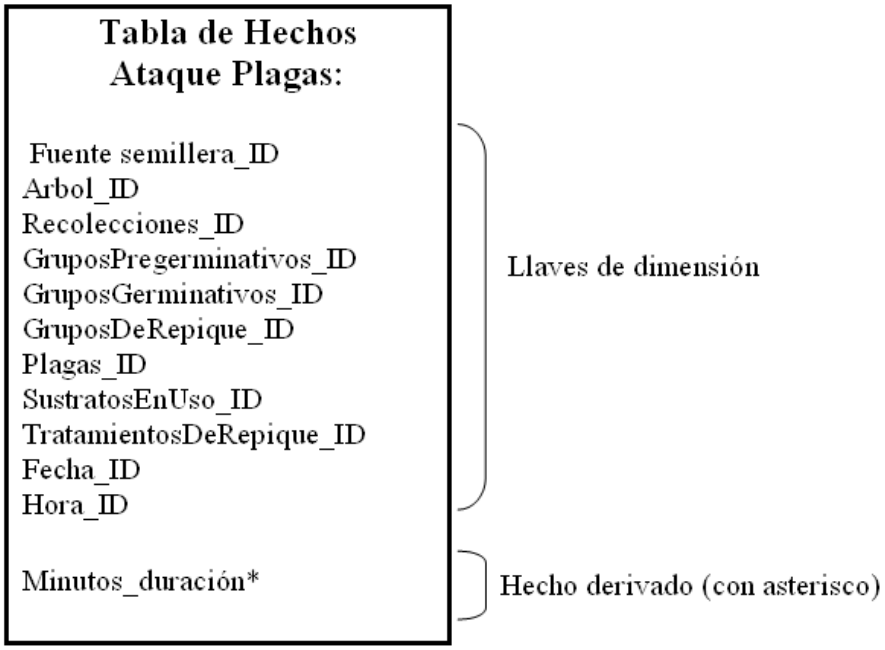


Figura 24. Detalles tabla de hechos – Data Mart Ataques de plagas

8.3.4 DETALLES DE LAS DIMENSIONES

Muestra los atributos individuales para cada dimensión. Incluye: nombre del atributo, descripción, cardinalidad, valores de ejemplo y políticas para dimensiones que cambian lentamente (tipo 1 -

sobrescribe el atributo a modificar, tipo 2 - crea un nuevo registro cuando se detecta un cambio). A continuación se presenta los detalles para cada dimensión identificada.

- Dimensión Fecha

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			20070816
Fecha completa	Indica la fecha completa.	*..*		16/08/2007 12:00:00 a.m.
Año	Representa el año del calendario.	1	Tipo 1	2006, 2007, 2008
Semestre	Representa al semestre de un año.	2	Tipo 1	1, 2
Nombre día de la semana	Nombre del día en la semana.	7	Tipo 1	Lunes, martes
Número día mes	Número día en el mes	31	Tipo 1	1, 2, 3
Nombre mes	Nombre del mes del año	12	Tipo 1	Enero, Junio

Tabla 15. Detalle dimensión Fecha

- Dimensión Fuente Semillera

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Arbol_Id	Clave principal de la dimensión.			1, 2, 3…
Loc_Id	Clave para identificar la subdimensión de localidad.			1, 2, 3…
Nombre	Nombre establecido para una fuente semillera	*..*	Tipo 1	Fuente Clarete
ASNM	Altura sobre el nivel del mar	*..*	Tipo 1	1487 mts.
Coordenadas	Coordenadas latitudinales y longitudinales de ubicación	*..*	Tipo 1	05° 36' 56.57" N 120° 15' 03.19" O
Tipo Fuente	Clasificación en	3	Tipo 1	Bosque

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
	la que se encuentra la fuente semillera.			heterogéneo Bosque homogéneo Árbol aislado
Características de la zona	Información opcional asociada a la zona donde se encuentra la fuente semillera.	*..*	Tipo 1	Zona muy húmeda.
Observaciones	Información opcional sobre la fuente semillera.	*..*	Tipo 1	La fuente semillera es homogenea.
código árbol	Identifica individualmente a cada árbol.	*..*	Tipo 1	PUE-01 CAJ-47
Altura árbol	Altura en metros del árbol.	*..*	Tipo 1	12 mts.
Altura comercial árbol		*..*	Tipo 1	10 mts.
Diámetro de copa árbol		*..*	Tipo 1	2 mts.
Diámetro altura pecho árbol		*..*	Tipo 1	4 mts.
Coordenadas árbol	Coordenadas latitudinales y longitudinales de ubicación del árbol.	*..*	Tipo 1	04° 35' 56.27" N 128° 12' 03.17" O
Tipo material árbol	Clasificación en la que se encuentra el árbol.	2	Tipo 1	Árbol productor Fuente de clones
Observaciones árbol	Información opcional sobre el árbol.	*..*	Tipo 1	No hay observaciones.

Tabla 16. Detalle dimensión Fuente Semillera

- Dimensión Geografía

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Departamento	Nombre	*..*	Tipo 1	Cauca

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
	establecido para un departamento.			
Municipio	Nombre establecido para un municipio.	*..*	Tipo 1	Popayán
Corregimiento	Nombre establecido para un corregimiento.	*..*	Tipo 1	La Rejoya
Vereda	Nombre establecido para una vereda.	*..*	Tipo 1	Villa Nueva

Tabla 17. Detalle dimensión Geografía

- Dimensión Grupos Germinativos

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para un grupo germinativo	*..*	Tipo 1	G1-PG1-1-CAJ-001
Posición de siembra	Clasificación de la posición en la que se puede sembrar una semilla.	3	Tipo 1	No determinada Horizontal vertical
Sistema de siembra	Clasificación del sistema de siembra a utilizar para sembrar una semilla.	3	Tipo 1	Al voleo Individual Lineal
Observaciones	Información opcional sobre el grupo germinativo.	*..*	Tipo 1	No hay observaciones.

Tabla 18. Detalle dimensión Grupos Germinativos

- Dimensión Grupos Pregerminativos

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para un grupo pregerminativo	*..*	Tipo 1	PG1-1-CAJ-001
Semillas por kilogramo	Cantidad de semillas presentes en un kilogramo.	*..*	Tipo 1	300
Diámetro longitudinal	Medida tomada longitudinalmente a la semilla.	*..*	Tipo 1	4
Diámetro transversal	Medida tomada transversalmente a la semilla.	*..*	Tipo 1	5
observaciones	Información opcional sobre el grupo pregerminativo.	*..*	Tipo 1	Semilla muy pequeña

Tabla 19. Detalle dimensión Grupos Pregerminativos

- Dimensión Grupos de Repique

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para un grupo de repique.	*..*	Tipo 1	R1-G1-PG1-1-CAJ-001
Observaciones	Información opcional sobre el grupo de repique.	*..*	Tipo 1	No hay observaciones.

Tabla 20. Detalle dimensión Grupos de Repique

- Dimensión Hora

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Hora completa	Indica la hora completa	1440	Tipo 1	02:07 AM
Hora	Clave para identificar la	12	Tipo 1	1, 2, 3…
Minutos	Nombre establecido para una fuente semillera	60	Tipo 1	0, 1, 2
AMPM	Indicador tiempo del día	2	Tipo 1	A.M. P.M.

Tabla 21. Detalle dimensión Hora

- Dimensión Plagas

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para una plaga	*..*	Tipo 1	Hongo, hormiga
Descripción	Información opcional sobre la plaga.	*..*	Tipo 1	La plaga solo ataca raíces.

Tabla 22. Detalle dimensión Plagas

- Dimensión Recolecciones

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido a una recolección	*..*	Tipo 1	Recolección del 16 de Agosto de 2007 a las 7:23 PM
Porcentaje de pureza	Indica el valor de pureza en un	1001	Tipo 1	80%, 95%, 0.7%, 78.9%

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
	conjunto de semillas.			
Observaciones	Información opcional sobre la recolección.	*..*	Tipo 1	La recolección fue realizada en la mañana.

Tabla 23. Detalle dimensión Recolecciones

- Dimensión Sustratos en Uso

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para una un sustrato.	*..*	Tipo 1	Sustrato Tierra + Arena

Tabla 24. Detalle dimensión Sustratos en Uso

- Dimensión Tratamientos Pregerminativos

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para un tratamiento pregerminativo.	*..*	Tipo 1	Escarificación mecánica. Escarificación con agua caliente. Escarificación con ácido.
Descripción	Información opcional sobre el tratamiento pregerminativo.	*..*	Tipo 1	El tratamiento de Escarificación consiste en sumergir las semilla en ácido por 3 minutos.

Tabla 25. Detalle dimensión Tratamientos Pregerminativos

- Dimensión Tratamientos de Repique

NOMBRE ATRIBUTO	DESCRIPCIÓN	CARDINALIDAD	POLÍTICA DE DIMENSIONES QUE CAMBIAN LENTAMENTE	VALORES DE EJEMPLO
ID	Clave principal de la dimensión.			1, 2, 3…
Nombre	Nombre establecido para un tratamiento de repique.	*..*	Tipo 1	Podar raíz.
Descripción	Información opcional sobre el tratamiento de repique.	*..*	Tipo 1	El tratamiento consiste en podar cuidadosamente la raíz de la plántula.

Tabla 26. Detalle dimensión Tratamientos de Repique

8.4 IMPLEMENTACIÓN DE LA BODEGA DE DATOS

El proceso de implementación de la Bodega de Datos lleva los modelos lógicos generados durante el modelado dimensional, a su representación física; para ello, se tuvo en cuenta la aplicación de ciertos criterios y estándares que facilitan su implementación. A continuación se mencionan algunos:

- Se utilizó el prefijo Dim para el nombrado de las dimensiones y el prefijo Fact para las tablas de hechos.
- Los nombres de las dimensiones son alusivos al proceso o propósito que cumplen. Ej.: DimRecolecciones, DimFecha, DimHora, etc.
- Los nombres utilizados en el modelo lógico son iguales en el modelo físico.
- Para la implementación en SQL Server Analysis Services se creó un único cubo que alberga todos los Data Marts o áreas del negocio de la Bodega de Datos.
- Se crearon perspectivas para cada Data Mart, para proveer un mejor acceso y exploración del cubo.
- Se utilizó la herramienta Sybase Power Designer para crear el modelo físico (ver Anexo B) y general los scripts para la creación de la Bodega de Datos.

A continuación se muestra la implementación realizada en SQL Server Analysis Services para cada Data Mart.

La Figura 25 muestra la implementación del Data Mart Recolecciones que incluye las medidas de Cantidad Recolectada que corresponde a la cantidad de semillas recolectadas en una recolección, Cantidad Viable es la cantidad de semillas viables o buenas obtenidas de las semillas recolectadas y Porcentaje de Viabilidad que hace referencia al porcentaje de éxito de la recolección. El Data Mart responde preguntas relacionadas con las recolecciones de semillas, la fecha y hora de la recolección, la fuente semillera, el árbol productor y la localización geográfica.

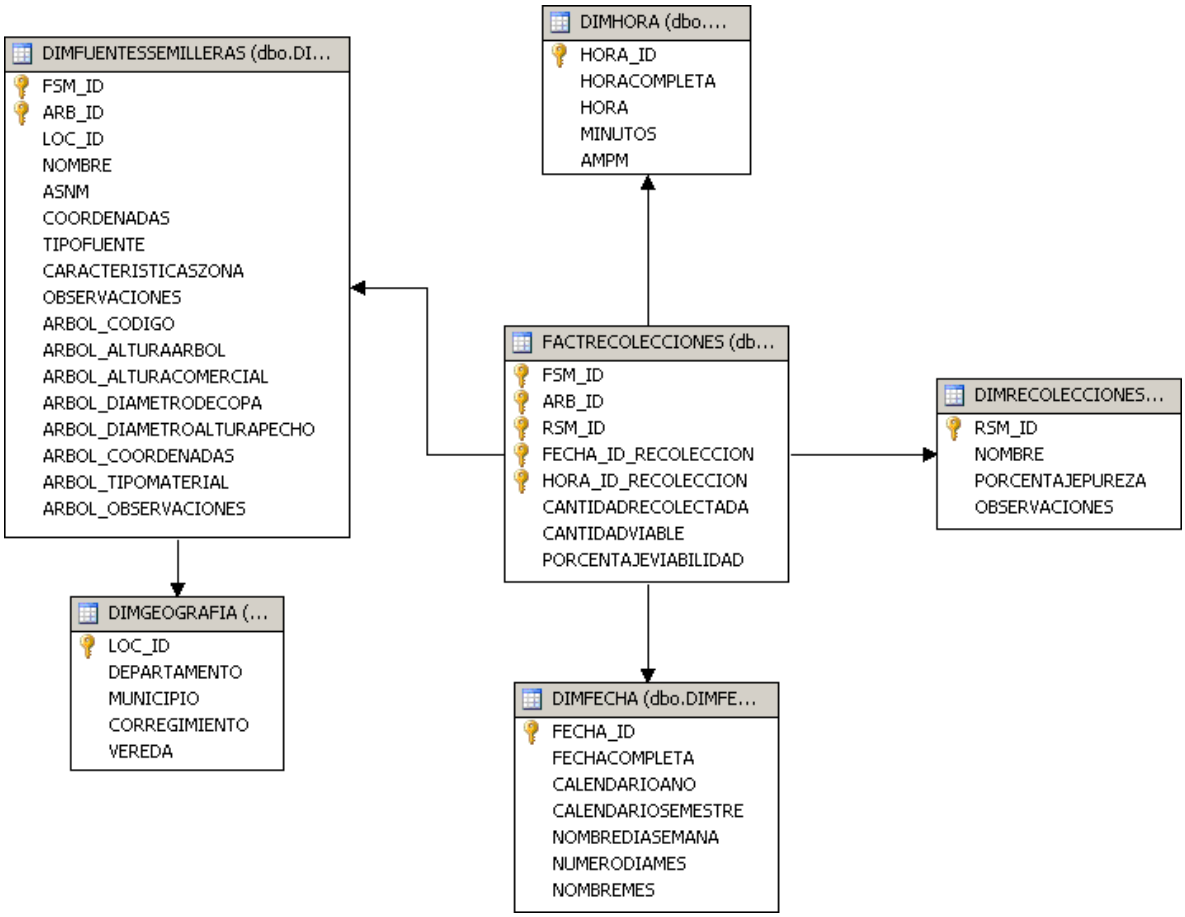


Figura 25. Implementación del Data Mart Recolecciones

La Figura 26 muestra la implementación del Data Mart Pregerminaciones que incluye las medidas de Cantidad de Material que corresponde a la cantidad de semillas que conforman el grupo Pregerminativo, Cantidad Material Viable es la cantidad de semillas viables o buenas obtenidas al final del proceso pregerminativo y Porcentaje de Viabilidad que hace referencia al porcentaje de éxito del grupo Pregerminativo. El Data Mart responde preguntas relacionadas con los grupos Pregerminativos, los tratamientos aplicados, la duración del proceso de pregerminación y su trazabilidad.

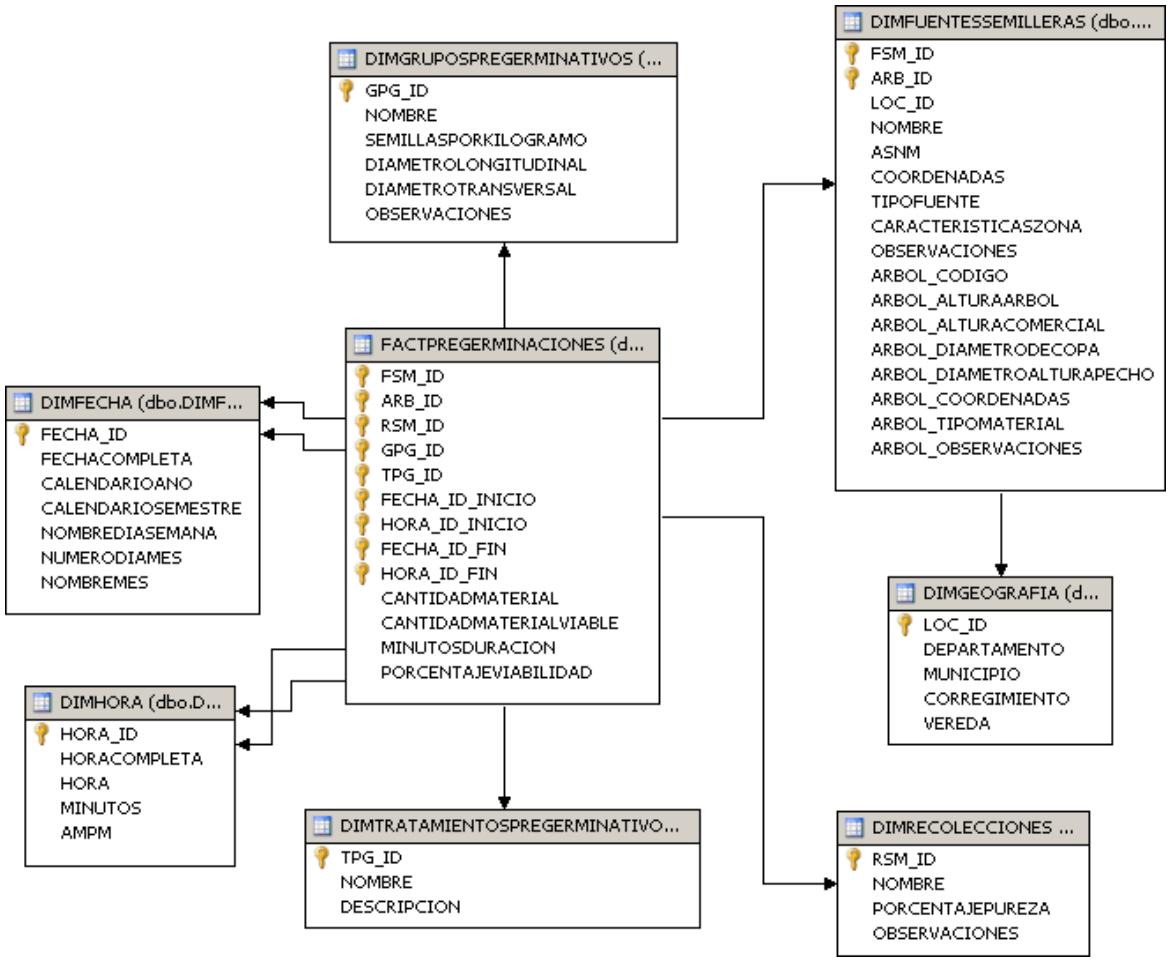


Figura 26. Implementación del Data Mart Pre-Germinaciones

La Figura 27 muestra la implementación del Data Mart Germinaciones que incluye las medidas de Cantidad de Material que corresponde a la cantidad de semillas que conforman el grupo Germinativo, Cantidad Material Viable es la cantidad de semillas viables o buenas obtenidas al final del proceso germinativo, Cantidad Germinada es la cantidad de semillas que germinaron y Porcentaje de Viabilidad que hace referencia al porcentaje de éxito del grupo Germinativo. El Data Mart responde preguntas relacionadas con los grupos Germinativos, los sustratos utilizados, la duración del proceso de germinación y su trazabilidad.

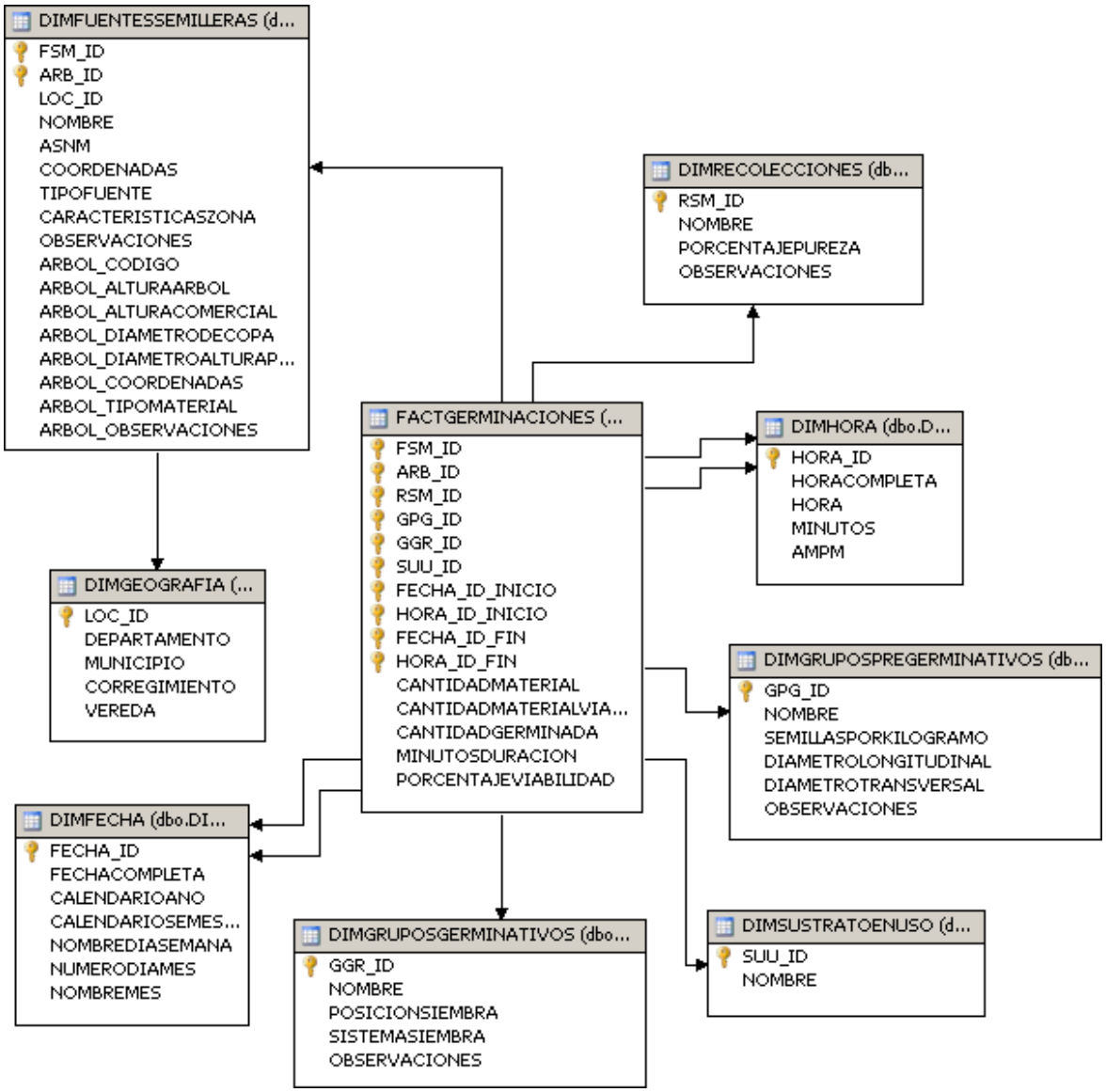


Figura 27. Implementación del Data Mart Germinaciones

La Figura 28 muestra la implementación del Data Mart Repique que incluye las medidas de Cantidad de Material que corresponde a la cantidad de plántulas que conforman el grupo de Repique, Cantidad Material Viable es la cantidad de plántulas viables o buenas obtenidas al final del proceso de repique, Minutos de Duración que indica la duración del proceso de repique en minutos y Porcentaje de Viabilidad que hace referencia al porcentaje de éxito del grupo de Repique. El Data Mart responde preguntas relacionadas con los grupos de repique, los sustratos utilizados, la duración del proceso de repique y su trazabilidad.

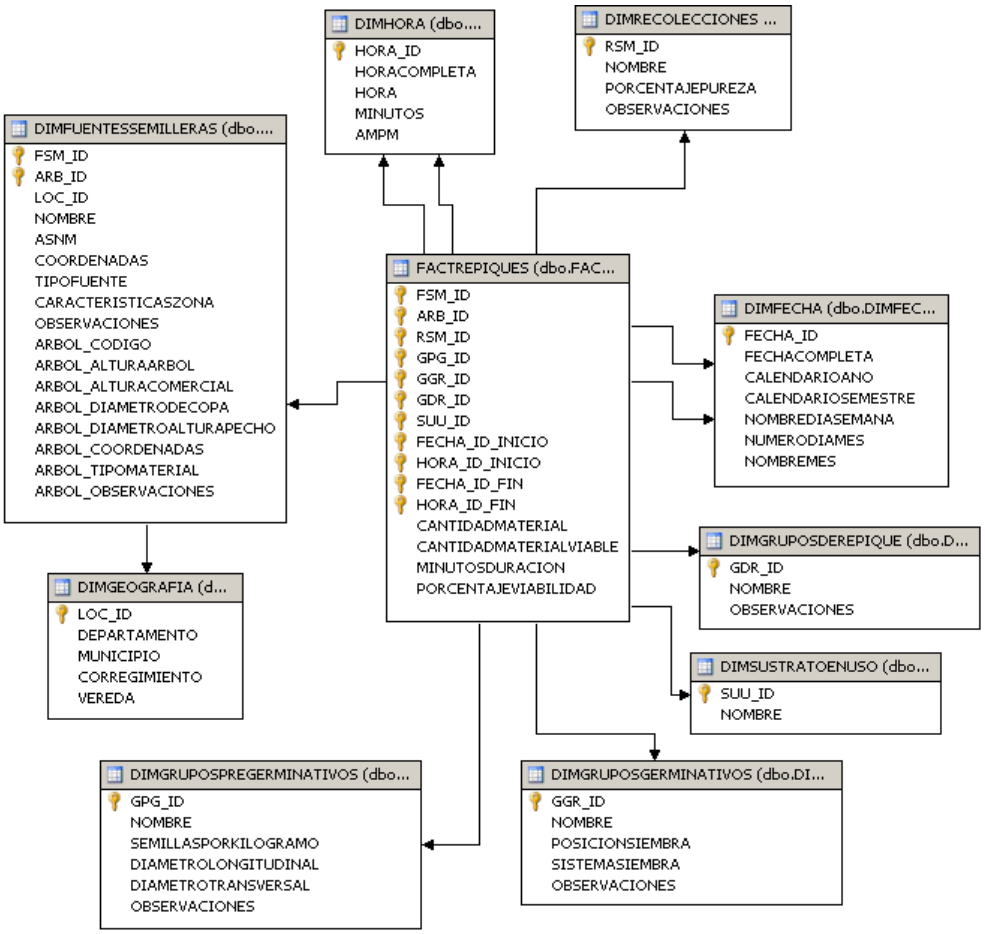


Figura 28. Implementación del Data Mart Repiques

La Figura 29 muestra la implementación del Data Mart Tratamientos de Repique que incluye la medida Minutos de Duración que indica la duración del tratamiento de repique en minutos aplicado a un grupo de Repique. El Data Mart responde preguntas relacionadas con los tratamientos aplicados a los grupos de Repique, la duración del tratamiento y la trazabilidad del grupo de Repique.

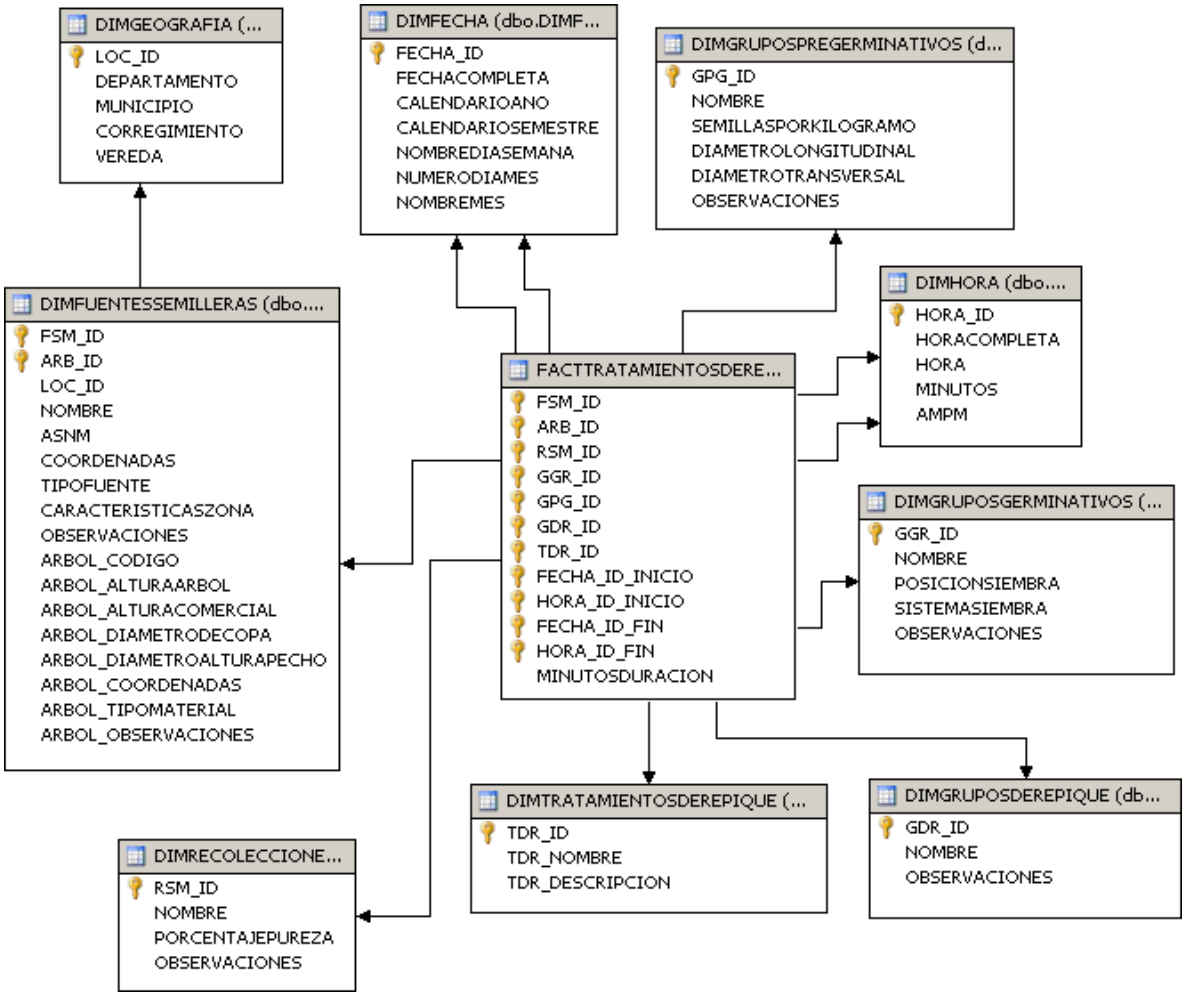


Figura 29. Implementación del Data Mart Tratamientos de Repique

La Figura 30 muestra la implementación del Data Mart Ataques Plagas que incluye la medida Minutos de Duración que indica la duración del ataque de la plaga en minutos a un grupo de Repique. El Data Mart responde preguntas relacionadas con los tratamientos aplicados para una plaga, la duración del ataque y la trazabilidad del grupo de Repique.

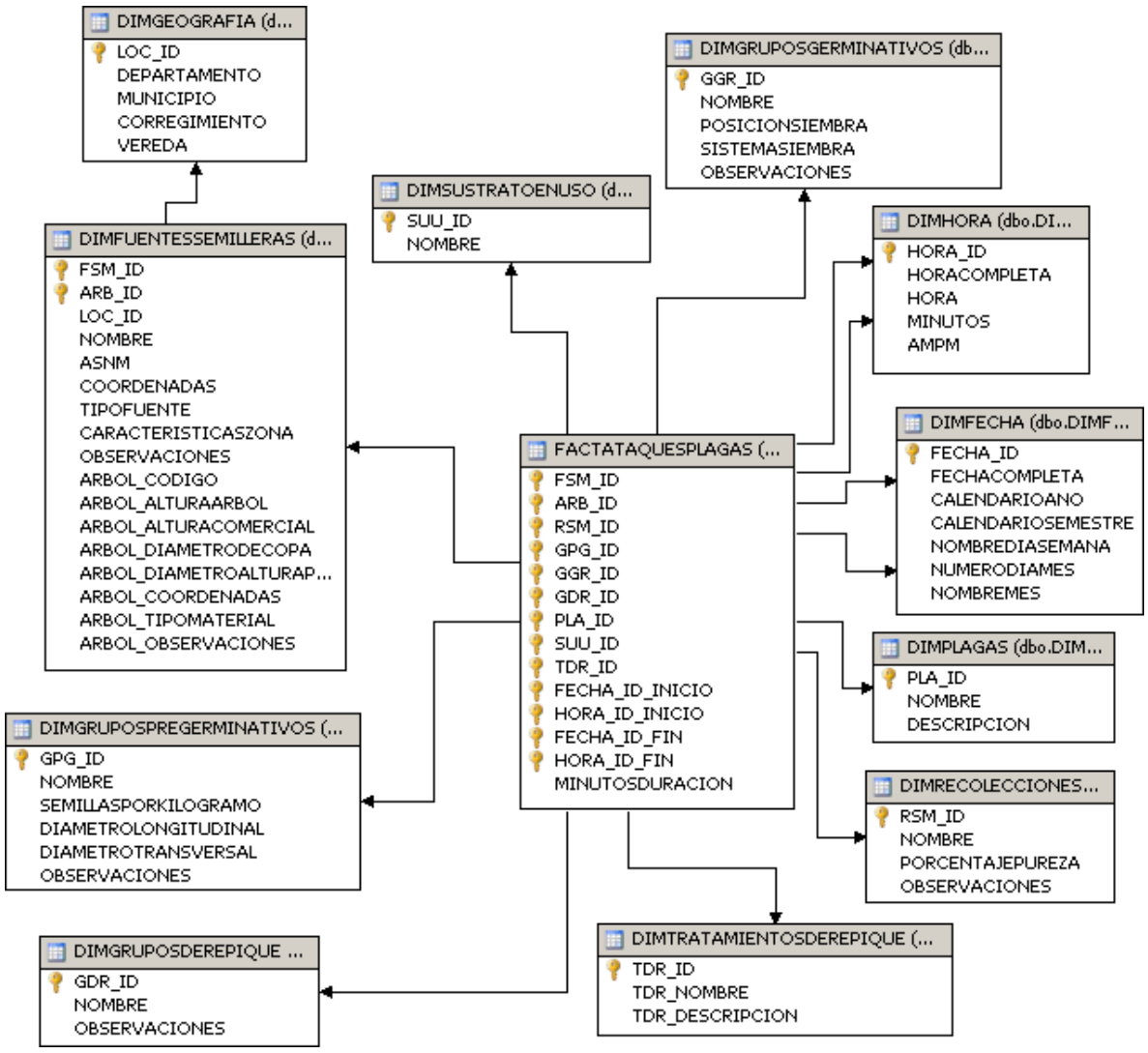


Figura 30. Implementación del Data Mart Ataques de plagas

9 DISEÑO Y DESARROLLO DEL PROCESO DE ETL

El proceso ETL se realizó sobre SQL Server Integration Services (SSIS), que es la plataforma que permite generar soluciones de integración de datos de Microsoft SQL Server 2005. En esta etapa se identificó como única fuente de datos operacionales la base de datos del sistema OLTP Web.

En SSIS, un proyecto o paquete está compuesto básicamente por dos componentes, El flujo de control y el flujo de datos. Los elementos de flujo de control preparan o copian datos, interactúan con otros procesos o implementan ciclos de trabajo. Las restricciones de precedencia ordenan en una secuencia la ejecución de las tareas. Para el proyecto se desarrolló un único flujo de control que abarca todas las tareas para la extracción, transformación y carga de los datos. Este flujo de control permite realizar la carga inicial y las posteriores cargas incrementales de los datos, en la Figura 31 se aprecia una sección del flujo de control correspondiente al Data Mart Pre-germinaciones. El flujo comprende cuatro tareas, las primeras tres se encargan de cargar los datos para las dimensiones y la última lo hace para la tabla de hechos. El orden en que están dispuestas las tareas es el orden en que se ejecutarán, para iniciar su ejecución una tarea deberá esperar que su antecesora finalice con éxito.

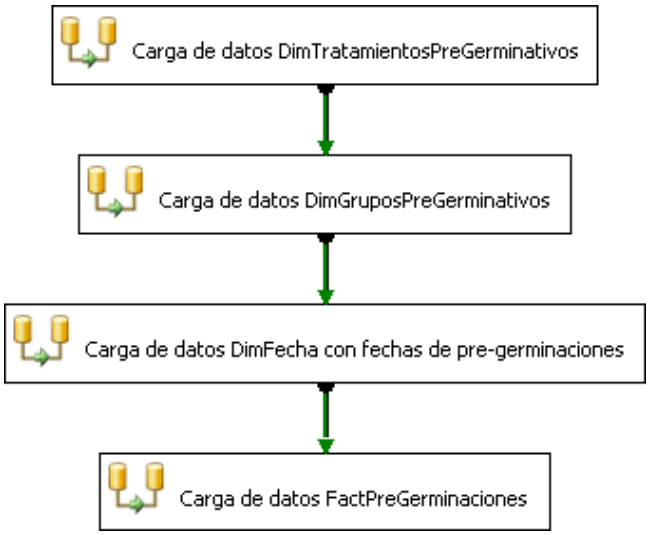


Figura 31. Flujo de control – Data Mart Pre-germinaciones

Cada tarea tiene un flujo de datos asociado, un flujo de datos es una estructura compuesta por orígenes que extraen datos, transformaciones que modifican y agregan datos, destinos que cargan datos y rutas que conectan las salidas y entradas de los componentes del flujo de datos. La tarea "Carga de datos DimTratamientosPreGerminativos" ejecuta el flujo de datos de la Figura 32. Este flujo carga datos para la dimensión Tratamientos Pre-germinativos. La fuente u origen de datos es una vista de la base de datos relacional del sistema OLTP Web (VIS_TRATAMIENTOS-PREGERMINATIVOS).

Una vez los datos son extraidos pasan a un proceso de transformacion, denominado Columna Derivada, que crea nuevos valores de columna aplicando expresiones a las columnas de entrada. Una expresión puede contener cualquier combinación de variables, funciones, operadores y columnas de la entrada de transformación. El resultado puede agregarse como una nueva columna o insertarse en una columna existente como un valor de reemplazo. Este proceso es utilizado para reemplazar los campos vacios o nulos por mensajes mas dicientes o informativos para un usuario.

A continuación el flujo de datos pasan al proceso de Dimension de Variación Lenta, que coordina la actualización e inserción de registros en las tablas de dimensiones, para este caso, la dimensión Tratamientos Pregerminativos.

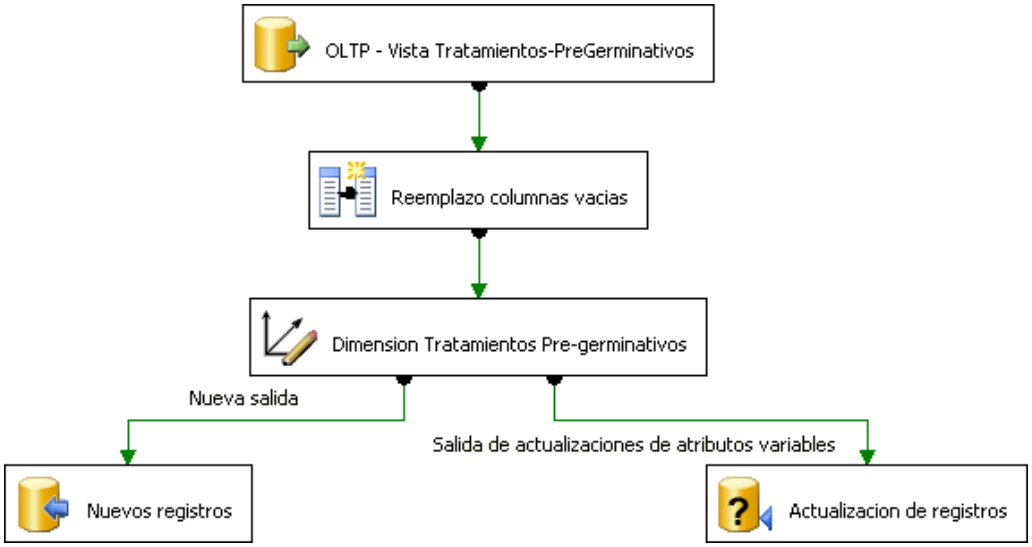


Figura 32. Flujo de datos – Dimensión Tratamientos Pre-germinativos

Procedimientos análogos ocurren para las tareas del flujo de control “Carga de datos DimGruposPreGerminativos”, “Carga de datos DimFecha con fechas de pre-germinaciones” y “Carga de datos FactPreGerminaciones”.

10 PROBLEMAS Y SOLUCIONES

Inicialmente cada Data Mart de la Bodega de Datos respondía preguntas propias de su área del negocio; sin embargo, se identificó la necesidad de contar con información sobre las etapas de los procesos previos de un área del negocio. Esta información permite conocer la trazabilidad de una semilla a lo largo de su ciclo de recolección, pregerminación, germinación y repique. Por ejemplo, en el Data Mart Repique se puede identificar la Fuente Semillera (proceso de recolección) de donde provino una semilla y determinar que Fuente Semillera produce semillas con mayor porcentaje de viabilidad. Para lograr esta trazabilidad se propuso inicialmente crear varias dimensiones que agruparan la información más relevante de los Data Marts anteriores. Por ejemplo, en el Data Mart Germinaciones tendría una dimensión Trazabilidad, que agruparía información de los Data Marts Recolecciones y Pregerminaciones, que son procesos previos a la germinación. La alternativa de adicionar estas dimensiones se descartó por la duplicidad innecesaria de los datos y se optó por relacionar directamente cada tabla hechos con las dimensiones conformadas para obtener la trazabilidad.

Durante la etapa de implementación de la Bodega de Datos surgió el dilema acerca de la usabilidad de la exploración. Si la implementación se realizaba en un solo cubo de Analysis Services, la exploración seria muy compleja para el usuario, porque el cubo estaría compuesto de varios grupos de medida, tablas de hechos y dimensiones de distintas áreas del negocio. A partir de esto, se analizó la posibilidad de crear un cubo de Analysis Services por cada Data Mart, pero esta opción se descartó por implicaciones negativas en el rendimiento. Finalmente se optó por el uso de perspectivas. Las perspectivas son subconjuntos visibles predefinidos de un cubo, en este caso particular cada subconjunto representa un Data Mart y permiten una exploración centrada en las áreas del negocio del interés para el usuario sin perder rendimiento.

El asistente o wizard del Business Intelligence Development Studio para la creación de los cubos de Analysis Services es una herramienta muy didáctica para el aprendizaje inicial, pero no es la mejor opción en un escenario real de desarrollo por los problemas en la identificación de tipos de dimensión, atributos y jerarquías. Lo recomendable es generar los cubos manualmente, adicionando las dimensiones e identificando sus atributos y jerarquías.

PARTE 4 – OLAP

11 MARCO TEÓRICO

11.1 DEFINICIÓN

El término OLAP (On-Line Analytic Processing) fue presentado en un artículo escrito por Arbor Software Corp (en la actualidad conocida como Hyperion Solutions Corporation), y puede ser definido como el proceso interactivo de crear, mantener, analizar y realizar informes sobre datos y es usual añadir que los datos en cuestión, son percibidos y manejados como si estuvieran almacenados en un "arreglo multidimensional" [20] . El análisis multidimensional consiste en combinar distintas áreas de la organización, y así ubicar ciertos tipos de información que revelen el comportamiento del negocio [18].

OLAP funcionalmente esta caracterizado por el análisis multidimensional dinámico de datos de la empresa, soportando actividades analíticas y de navegabilidad de los usuarios finales. Algunas de las características son:

- Aplicación de cálculos y modelado a través de las dimensiones, a través de Jerarquías y/o a través de miembros.
- Análisis de tendencias sobre periodos de tiempo secuenciales.
- Dividir (slicing) subconjuntos para visualizarlos sobre la pantalla.
- Navegar (drill down - drill up) por los niveles mas profundos consolidados.
- Alcance a través de los datos de detalle subyacentes.
- Rotación a las nuevas comparaciones dimensionales de un área determinada.
- Exportar datos en diferentes formatos.
- Generación de reportes.

OLAP esta implementado en un modo multiusuario cliente/servidor y ofrece respuestas rápidas consistentes a una consulta, indiferente del tamaño de la base de datos y la complejidad. OLAP ayuda a los usuarios a sintetizar la información de la empresa a través de comparaciones, vistas personalizadas, así como a través del análisis histórico.

11.2 CLASIFICACIÓN

A continuación se describen las implementaciones tecnológicas OLAP más populares:

11.2.1 MOLAP

MOLAP (Multidimensional OLAP), suministra capacidad de análisis de datos almacenados con un sistema de bases de datos multidimensional (MDBS). Algunas características son [14][23]:

Ventajas:

- Excelente desempeño: Los cubos MOLAP son construidos para realizar una rápida recuperación de datos.

- Puede ejecutar cálculos complejos: Todos los cálculos han sido previamente generados cuando se crea el cubo. De aquí que los cálculos complejos aunque sean dobles, se obtiene un resultado rápido.

Desventajas:

- Limitación en la cantidad de datos que puede manejar: Porque todos los cálculos son ejecutados cuando se construye el cubo, no es posible incluir una gran cantidad de datos dentro del mismo. Esto no quiere decir que los datos en el cubo no sean derivados de una gran cantidad de datos. De hecho esto es posible, pero en este caso solamente información de nivel de resumen será incluida en el cubo.

- Requieren inversión adicional: La tecnología de cubo es usualmente propietaria y aun no es muy común en las organizaciones. Por consiguiente, adoptar la tecnología MOLAP, involucra necesidades adicionales de inversión de recursos humanos y económicos.

11.2.2 ROLAP

ROLAP (Relational OLAP), puede ser definida como un conjunto de aplicaciones e interfaces que le dan a las bases de datos relacionales un tratamiento multidimensional [14], es decir OLAP sobre una base de datos relacional. Algunas características son [23]:

Ventajas:

- Puede manejar gran cantidad de datos: La limitación del tamaño de datos en la tecnología ROLAP depende de los límites de la cantidad de datos que la base de datos pueda manejar. En otras palabras, ROLAP por si mismo no tiene limites de cantidad de datos.

- Puede apoyar funcionalidades inherentes en la base de datos relacional: A menudo, las bases de datos relacionales ya vienen con un contenedor de funcionalidades. Las tecnologías ROLAP, desde que se ubicaron sobre las bases de datos relacionales, pueden apoyar estas funciones.

Desventajas:

- El desempeño puede ser bajo: Porque cada reporte ROLAP es esencialmente una consulta SQL (o múltiples consultas SQL) en la base de datos relacional, la duración de la consulta puede ser largo si el tamaño de la base de datos es grande.

- Limitado para funcionalidades SQL: Porque la tecnología ROLAP principalmente cuenta con la generación de sentencias SQL para consultar la base de datos relacional y las sentencias SQL no cumplen con todas las necesidades, por consiguiente las tecnologías ROLAP esta tradicionalmente limitadas por lo que puede hacerse con SQL.

11.2.3 HOLAP

HOLAP (Hybrid OLAP), es una combinación de los dos anteriores. Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en el relacional [24].

11.3 OLAP EN MICROSOFT ANALYSIS SERVICES 2005

Analysis Services 2005 almacena la información de la metadata de la base de datos en formato XML. Analysis Services 2005 proporciona la opción de almacenar los datos eficientemente en formato propietario del Analysis Services o almacenarlo en la base de datos relacional. Si se elige que los datos sean almacenados en formato propietario conseguirá un mejor desempeño de las consultas que en el caso en que los datos sean recuperados desde la base de datos relacional. Este formato propietario le ayuda al Analysis Services 2005 a recuperar los datos eficientemente y a mejorar el desempeño de las consultas. De acuerdo a donde los datos estén almacenados se tienen los siguientes tipos de almacenamiento: MOLAP, ROLAP y HOLAP [25].

MOLAP es el modo de almacenamiento en el cual los datos están almacenados en un formato propietario del Analysis Services. Este es el modo de almacenamiento por defecto. Las ventajas claves de este modo de almacenamiento es la rápida recuperación de datos mientras se realiza el análisis y por consiguiente proporciona un buen desempeño en las consultas y la capacidad de manejar cálculos complejos [25].
Las dos desventajas potenciales del modo MOLAP son: necesidad de almacenamiento para grandes bases de datos e incapacidad para ver datos nuevos que ingresan a la Bodega de Datos [25].

En modo ROLAP las agregaciones o sumarias también están almacenadas en la base de datos relacional. Las consultas contra el Analysis Services son apropiadamente cambiadas a consultas hacia la base de datos relacional para recuperar la correcta sección de datos requerida. La más importante desventaja del modo de almacenamiento ROLAP es la lentitud en las consultas debido a que cada consulta hacia el Analysis Services es traducida en una o más consultas SQL hacia la base de datos relacional [25].

El HOLAP si la consulta hacia el Analysis Services requiere de datos agregados, es recuperada desde los datos almacenados en la instancia del Analysis Services y será más rápida que recuperarlos desde la base de datos relacional. Si la consulta requiere de datos detallados, una consulta adecuada será enviada a la base de datos relacional y estas consultas pueden tomar un poco más de tiempo [25].

12 SELECCIÓN DE LA HERRAMIENTA OLAP

Para la selección de la herramienta OLAP que se integraría al proyecto, se realizó un análisis y en los casos donde fue posible, pruebas con diferentes herramientas OLAP disponibles en el mercado. Se estableció un conjunto de criterios de selección basándose en las necesidades y expectativas de un usuario general de una herramienta de análisis multidimensional, además de los requisitos mínimos para la integración de la herramienta con la bodega de datos y el sistema OLTP Web.

12.1 CRITERIOS DE SELECCIÓN

CÓDIGO	DESCRIPCIÓN
C1	Capacidad de integración con el sistema OLTP Web: Si la herramienta puede integrarse directamente con el sistema OLTP Web (aplicación ASP.NET 2.0) corriendo sobre Microsoft Internet Information Services (IIS).
C2	Compatibilidad con navegadores: Que la herramienta pueda ser usada con distintos navegadores sin la necesidad de adicionar complementos o plugins. En este caso específico, se tomaron como referencia los navegadores Internet Explorer y Mozilla Firefox con los cuales se abarca la mayoría del mercado.
C3	Compatibilidad con SQL Server Analysis Services (SSAS) 2005.
C4	Costo de una licencia para un ambiente de producción.
C5	Drag-and-drop (arrastrar y soltar): Capacidad de arrastrar y soltar con el ratón objetos de una ventana a otra o entre partes de una misma ventana.
C6	Drill down and Drill up: Capacidad para navegar apropiadamente por niveles de las jerarquías de las dimensiones.
C7	Exportar datos a Excel: Capacidad de exportar los datos a Excel.
C8	Exportar datos en formato XML.
C9	Filtering and sorting: Capacidad para filtrar y ordenar los datos desplegados.
C10	Generación de gráficos: Capacidad de generar gráficos estadísticos a partir de los datos analizados.
C11	Generación y carga de reportes: Capacidad para que el usuario genere, almacene y cargue reportes remotamente.
C12	Herramienta independiente: Si la herramienta se adquiere independiente o acoplada a otros productos.
C13	Impresión: Capacidad para imprimir los datos y gráficos directamente desde la herramienta.
C14	Versión de prueba: Si la herramienta OLAP cuenta con una versión de prueba o trial disponible para la descarga y prueba previa a la adquisición.

Tabla 27. Criterios para la selección de una herramienta OLAP

12.2 HERRAMIENTAS OLAP ANALIZADAS

CÓDIGO	DESCRIPCIÓN
BOWI XI	**Producto:** Business Objects Web Intelligence XI **Tipo de componente:** Web **Fabricante:** Business Objects
CFXOE	**Producto:** Chart FX OLAP Extension **Tipo de componente:** ASP.NET **Fabricante:** Software FX

CÓDIGO	DESCRIPCIÓN
	Observaciones: Requiere la instalación de Chart FX for .NET 6.2 o Chart FX for Visual Studio 2005. Tiene problemas de incompatibilidad con navegadores distintos a Microsoft Internet Explorer.
COGNOS	**Producto:** Cognos 8 Business Intelligence – Analysis **Tipo de componente:** Web **Fabricante:** Cognos
CONTOUR	**Producto:** ContourCube **Tipo de componente:** Active X **Fabricante:** Contour Components **Observaciones:** Por ser un control Active X tiene problemas de compatibilidad con los navegadores y con las políticas de seguridad de los equipos clientes.
DUNDAS	**Producto:** Dundas OLAP Services **Tipo de componente:** ASP.NET **Fabricante:** Dundas
DYNAMI	**Producto:** DynamiCube **Tipo de componente:** Active X **Fabricante:** Data Dynamics **Observaciones:** Por ser un control Active X tiene problemas de compatibilidad con los navegadores y con las políticas de seguridad de los equipos clientes. No tiene incorporada la capacidad de generar gráficos.
HYPERION	**Producto:** Hyperion System 9 BI+ Web Analysis **Tipo de componente:** Web **Fabricante:** Hyperion
INSTANT	**Producto:** instantOLAP **Tipo de componente:** Web **Fabricante:** instantOLAP **Observaciones:** Requiere que los navegadores de los clientes tengan instalado un plugin de Java.
PIVOTTABLE	**Producto:** Office Web Components - PivotTable **Tipo de componente:** COM **Fabricante:** Microsoft
JPIVOT	**Producto:** Pentaho - JPivot **Tipo de componente:** JSP **Fabricante:** Proyecto open source **Observaciones:** Debe correr sobre un servidor web Apache Tomcat
RADAR	**Producto:** RadarCube ASP.NET para SSAS **Tipo de componente:** ASP.NET **Fabricante:** Radar-Soft
ASPXPIVOT	**Producto:** ASPxPivotGrid Suite **Tipo de componente:** ASP.NET **Fabricante:** Developer Express

Tabla 28. Herramientas OLAP analizadas

12.3 ANÁLISIS DE LAS HERRAMIENTAS OLAP

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
BOWI XI	--	✓	✓	--	✓	✓	✓	--	✓	✓	✓	--	--	X
CFXOE	✓	X	✓	US$ 999	✓	✓	--	--	✓	✓	X	X	X	✓
COGNOS	--	✓	✓	--	✓	✓	✓	✓	✓	✓	✓	--	✓	X
CONTOUR	✓	X	✓	US$ 864	✓	✓	✓	X	✓	X	✓	✓	✓	✓
DUNDAS	✓	✓	✓	US$ 699	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DYNAMI	✓	X	✓	US$ 599	✓	✓	✓	X	✓	✓	X	✓	✓	✓
HYPERION	--	--	✓	--	✓	✓	--	--	✓	✓	--	--	✓	X
INSTANT	X	X	X	--	X	✓	✓	X	✓	✓	X	✓	✓	✓
PIVOTTABLE	✓	X	✓	US$ 0	✓	✓	✓	X	✓	X	X	✓	✓	✓
JPIVOT	X*	✓	✓	US$ 0	X	✓	X	X	✓	--	X	✓	X	✓
RADAR	✓	✓	✓	US$ 300	✓	✓	✓	✓	✓	X	✓	✓	X	✓
ASPXPIVOT	✓	✓	✓	US$ 179	X	✓	✓	X	✓	X	X	✓	X	✓

✓ Cumple con el criterio X No cumple con el criterio -- No hay suficiente información

Tabla 29. Análisis de las herramientas OLAP

*: La herramienta podría integrarse indirectamente al proyecto realizando un desarrollo independiente (Java) y estableciendo los mecanismos de comunicación adecuados.

12.4 PROCESO DE SELECCIÓN

Etapa I:

Una vez realizado el análisis a las herramientas OLAP, se descartaron aquellas que no contaban con versiones de prueba (C14) y las que no son compatibles con SQL Server Analysis Services 2005 (C3). Posteriormente se descartaron las herramientas que no se pudieran integrar de manera DIRECTA con el sistema OLTP Web sobre Microsoft Internet Information Services (C1). Las herramientas OLAP que aprobaron esta etapa fueron:

- Chart FX OLAP Extension
- ContourCube
- Dundas OLAP Services
- DynamiCube
- Office Web Components – PivotTable
- RadarCube ASP.NET para SSAS
- ASPxPivotGrid Suite

Etapa II:

Esta etapa consistió en seleccionar de las herramientas OLAP pre-seleccionadas en la Etapa I aquellas que cumplieran con el criterio de generación y carga de reportes (C11). Las herramientas OLAP que aprobaron esta etapa fueron:

- ContourCube
- Dundas OLAP Services
- RadarCube ASP.NET para SSAS

Etapa III:

Esta etapa consistió en seleccionar de las herramientas OLAP pre-seleccionadas en la Etapa II aquellas que cumplieran en orden de importancia con los criterios de generación de gráficos (C10) y compatibilidad con navegadores (C2). La herramienta OLAP que aprobó esta etapa fue:

- Dundas OLAP Services

12.5 HERRAMIENTA OLAP SELECCIONADA

Dundas OLAP Services fue la herramienta que aprobó las tres etapas definidas para el proceso de selección de la herramienta OLAP del proyecto. Además cumplió con todos los criterios adicionales a las etapas de selección como lo fueron Drag-and-drop (C5), Drill down and Drill up (C6), exportar datos a Excel (C7), exportar datos en formato XML (C8), Filtering and sorting (C9), herramienta independiente (C12) e impresión (C13).

12.5.1 FUNCIONALIDAD DE DUNDAS OLAP SERVICES

La herramienta permite seleccionar los distintos cubos y perspectivas definidos en la Bodega de Datos y crear dinámicamente consultas multidimensionales arrastrando las medidas y dimensiones a la grilla de visualización de datos. Una vez ahí, se puede navegar a través de los niveles de los datos usando funcionalidades de drill down y drill up. El usuario tiene la posibilidad de imprimir o exportar los resultados a Microsoft Excel o en formato XML (ver Figura 33).

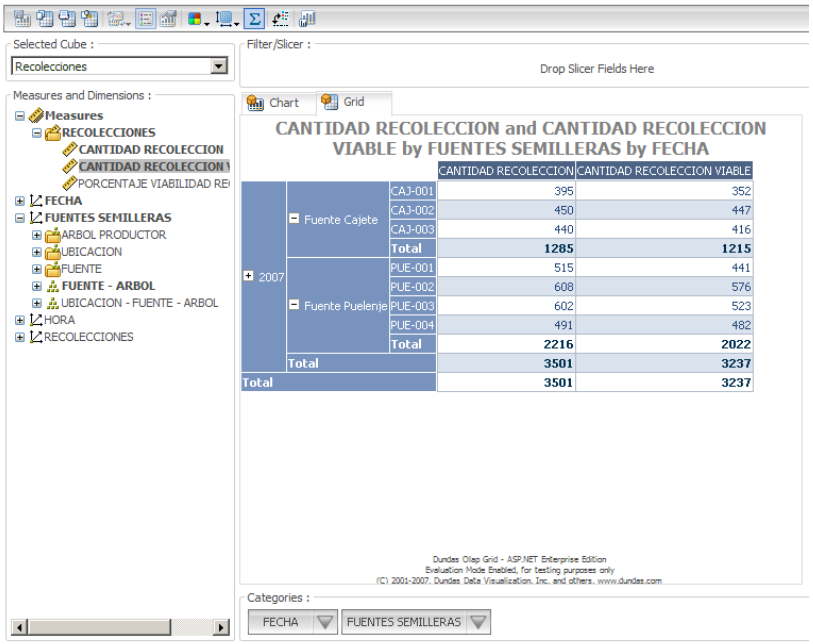


Figura 33. Grilla de visualización de datos - Dundas OLAP Services

La herramienta permite visualizar gráficamente los resultados de las consultas multidimensionales y navegar a través de ellos (ver Figura 34).

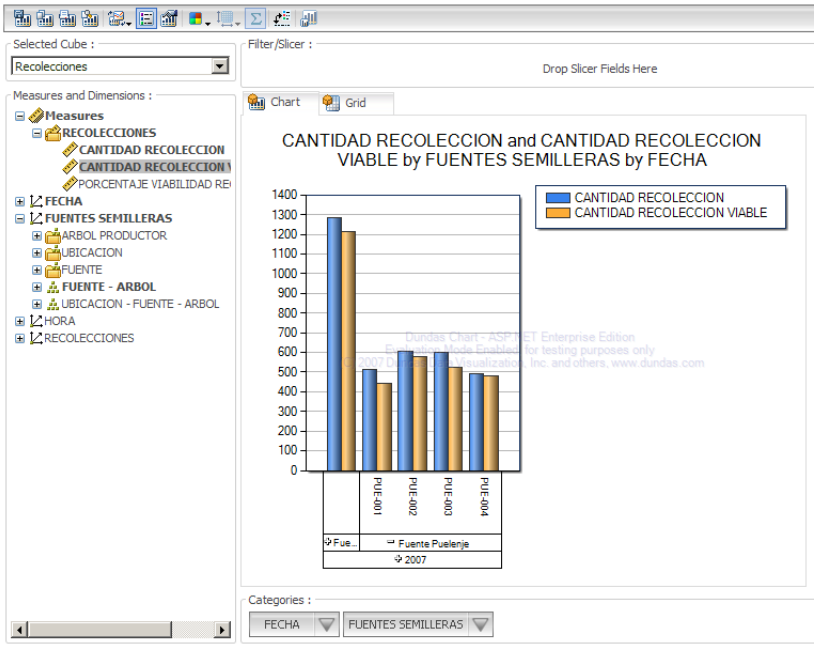


Figura 34. Visualización gráfica de los datos - Dundas OLAP Services

El usuario tiene la flexibilidad de escoger el tipo de gráfico (12 opciones), seleccionar las etiquetas y la distribución de las leyendas, asignar una paleta de colores y determinar las unidades en las que se presentaran los datos, como por ejemplo, porcentaje, moneda, tiempo, cantidad, etc. Ver Figura 35.

Otra de las funcionalidades es la posibilidad de guardar localmente las consultas multidimensionales creadas, de esta forma el usuario crea una sola vez la consulta que le interesa, la guarda y la carga cuando la necesite nuevamente. Adicionalmente se creó un conjunto de reportes predefinidos que permiten consultar y visualizar los eventos y relaciones más comunes en los procesos de germinación y cultivo. Ver Figura 36.

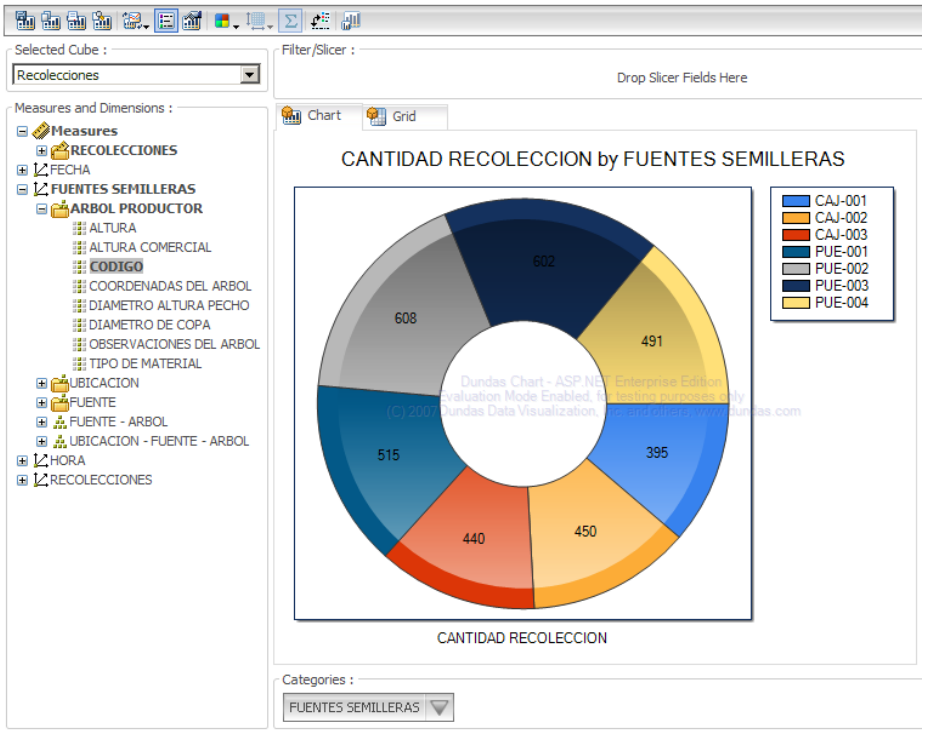


Figura 35. Tipos de gráficos - Dundas OLAP Services

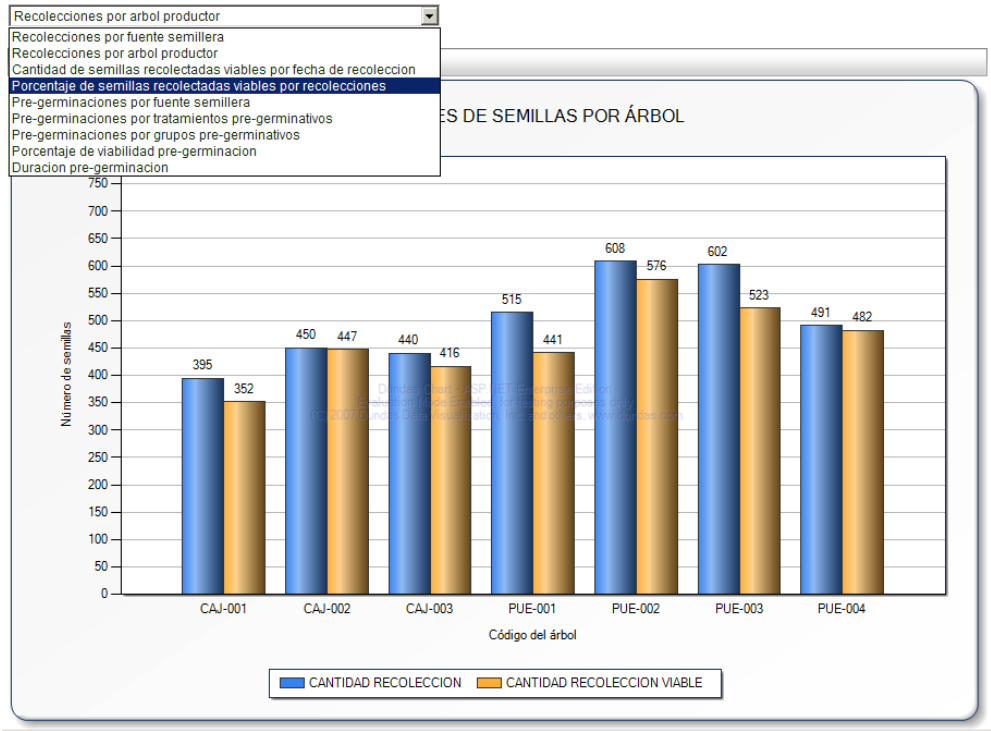


Figura 36. Reportes - Dundas OLAP Services

13 INTEGRACIÓN DE LA HERRAMIENTA OLAP

Una vez finalizado el desarrollo del sistema OLTP Web y seleccionada la herramienta OLAP se decidió integrar ambas soluciones en un nuevo proyecto ASP.NET denominado GREENDSS. La principal razón de esta decisión fue centralizar la administración y ofrecer al usuario final una plataforma unificada de servicios. La integración de Dundas OLAP Services al sistema OLTP Web se realizó en tiempo de desarrollo y sin inconvenientes debido a que Dundas OLAP Services es un control de servidor totalmente compatible con aplicaciones ASP.NET. El acceso a la Bodega de Datos se realiza por una conexión directa e independiente entre Dundas OLAP Services y Microsoft Analysis Services sin necesidad de código de programación adicional. Ver Figura 37 y Figura 38.

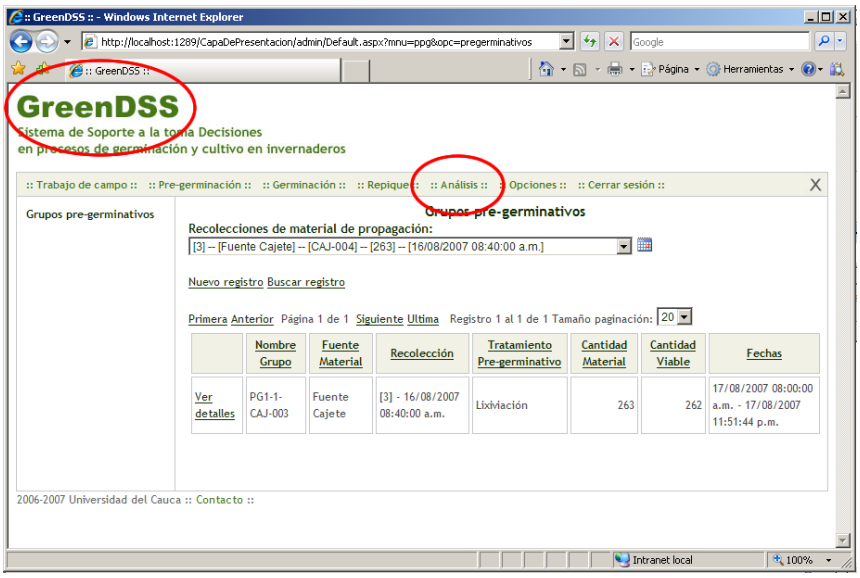


Figura 37. GreenDSS

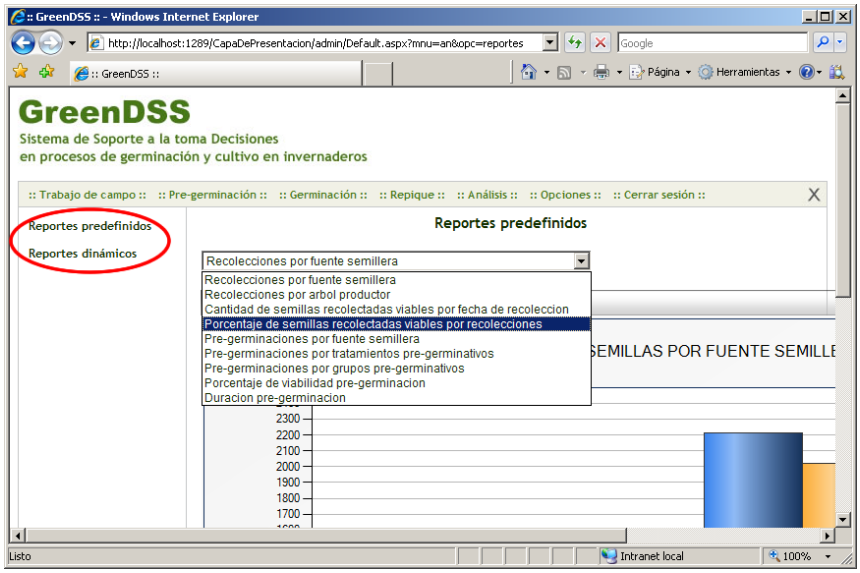


Figura 38. Dundas OLAP Services en GreenDSS

PARTE 5 – MINERÍA DE DATOS

14 MARCO TEÓRICO

14.1 DEFINICIÓN

Las herramientas de Minería de Datos, mediante algoritmos de análisis de datos especializados predicen posibles patrones, tendencias o comportamientos futuros de la organización, permitiendo al experto tomar decisiones en los negocios con base en un conocimiento que estaba inmerso en los datos y que no se había contemplado. Las implementaciones de los algoritmos, están basadas en la preparación y automatización previa de los datos, con el fin asegurar la escalabilidad y rendimiento en grandes volúmenes de datos. Es por esto que el uso de la tecnología de Bodega de Datos es recomendable (aunque no es imprescindible) en este tipo de soluciones [21].

14.2 OLAP VS MINERÍA DE DATOS

Es preciso tener claro que los productos OLAP responden a cuestiones del tipo ¿Qué?, como por ejemplo, ¿qué condiciones ambientales en el vivero hacen que las plantas de roble tengan un mayor crecimiento?, ¿qué técnica en el proceso de germinación obtuvo las mejores semillas?, etc. Mientras que las técnicas de Minería de Datos responden a cuestiones estratégicas del tipo ¿Por qué?, como por ejemplo, ¿por qué las plántulas[4] mueren tempranamente?, ¿por qué las semillas obtenidas no son uniformes?, entre otras posibles preguntas.

14.3 TÉCNICAS Y APLICACIONES

Las técnicas y aplicaciones de la Minería de Datos difieren de acuerdo a la clase de problema que se quiere resolver, a continuación se da una lista de las diferentes técnicas y sus respectivas aplicaciones [21]:

- Análisis de Asociaciones. Esta técnica establece relaciones entre los diferentes atributos de un conjunto de datos para identificar comportamientos; es comúnmente utilizada para entender el patrón de compra de los clientes, mediante el análisis del conjunto histórico de transacciones de compra de los mismos, que se conoce como Market Basket Analisis.

- Análisis basado en Secuencias. Esta técnica se trata de una variante del Análisis de Asociaciones, la cual permite encontrar relaciones que se presentan secuencialmente en el tiempo. En esta situación, no solo son importantes las transacciones, sino también el orden en el cual se presentan. Este tipo de análisis se utiliza para predecir, por ejemplo, cuales serán las próximas compras que una persona hará, con base en los patrones de secuencia de compras encontrados.

- Clustering. Es utilizado para problemas de segmentación. Esta técnica agrupa automáticamente registros de datos que presentan características similares, en segmentos o grupos diferentes. El Clustering es a menudo el primer paso en los análisis de Minería de Datos y sirve por ejemplo, para apoyar segmentación de clientes basándose en sus atributos demográficos; definir los hábitos de compra de múltiples segmentos de población, los segmentos podrían ser comparados para determinar a cuales de ellos enfocar las próximas campañas de mercadeo y publicidad.

- Clasificación. En ésta técnica se emplea un conjunto histórico de datos de ejemplos preclasificados, con el fin de construir un modelo que permita categorizar nuevos registros de

[4] Planta joven, al poco tiempo de brotar de la semilla

información. Las aplicaciones de detección de fraudes, estudio de riesgo de créditos y análisis de deserción de clientes emplean frecuentemente este tipo de análisis.

- Estimación ó Scoring. Se trata de una variante del problema de Clasificación que envuelve la generación de puntajes o "scores" en vez de realizar una simple clasificación binaria o n-aria de los datos. Generalmente se utiliza para asignar a un solicitante de un préstamo un puntaje basado en un conjunto de datos de entrenamiento.

- Predicción. La predicción consiste en crear un modelo matemático que permita predecir un valor numérico a partir de un conjunto de variables de entrada. En estadística, este tipo de análisis es comúnmente conocido como análisis de regresión.

15 METODOLOGÍA PARA MINERÍA DE DATOS

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de Minería de Datos: el modelo de referencia y la guía del usuario. El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de Minería de Datos en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de Minería de Datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase [22].

CRISP-DM estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. En la Figura 39 se puede ver un esquema de las diferentes fases de la metodología [22]:

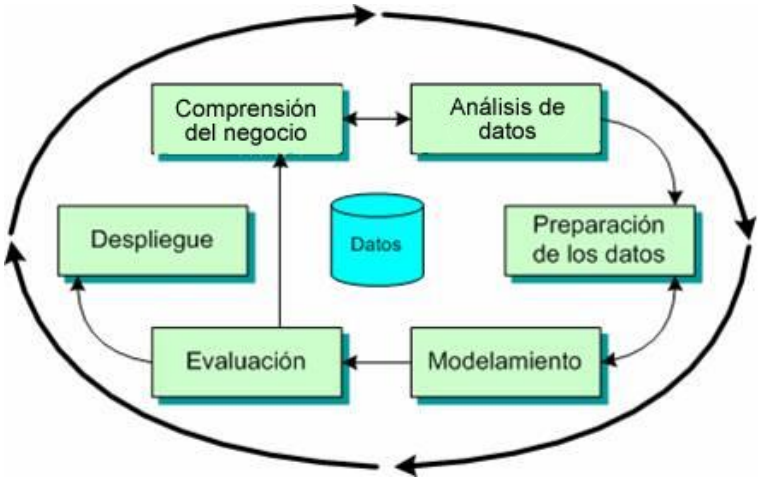


Figura 39. Fases del modelo de proceso CRISP-DM (Adaptado de [22])

Comprensión del negocio. Incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

Análisis de datos. Comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Preparación de los datos. Incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos, se encuentra muy relacionada con la cuarta fase, de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Las fases de preparación y modelado interactúan de forma sistemática.

Modelamiento. Se seleccionan las técnicas de modelado más apropiadas para el proyecto de Minería de Datos específico. Las técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios: ser apropiada al problema, disponer de datos adecuados, cumplir los requerimientos del problema, tiempo necesario para obtener un modelo y conocimiento de la

técnica. Antes de proceder al modelado de los datos se debe establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez realizadas estas tareas genéricas se procede a la generación y evaluación preliminar del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

Evaluación. Se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la sexta fase.

Despliegue. Normalmente los proyectos de Minería de Datos no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. En la fase de despliegue se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

CRISP-DM consta de cuatro niveles de abstracción, para cada una de las distintas fases, los cuales se encuentran organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. Ver Figura 40.

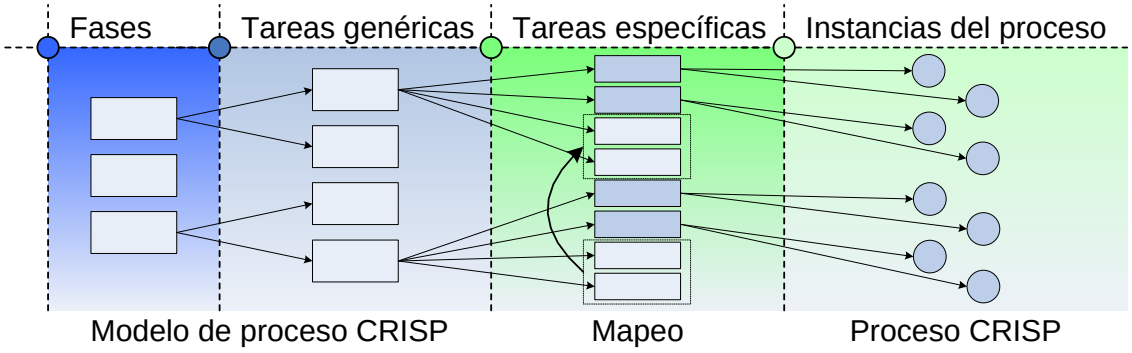


Figura 40. Los cuatro niveles de la Metodología CRISP–DM (Adaptado de [22])

A nivel más general, el proceso está organizado en las seis fases mencionadas anteriormente, estando cada fase a su vez estructurada en varias tareas genéricas de segundo nivel. Las tareas genéricas se proyectan en tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas.

Así, si en el segundo nivel se tiene la tarea general “limpieza de datos”, en el tercer nivel se dicen las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, Instancias del Proceso, recoge el conjunto de acciones, decisiones y resultados que se desarrollaran sobre el proyecto de Minería de Datos específico.

15.1 COMPRENSIÓN DEL NEGOCIO

En nuestro contexto especifico, para garantizar una mayor productividad en los procesos de germinación y repique es importante identificar las condiciones ambientales ideales a las que deben ser sometidas las semillas y plántulas.

El objetivo principal de esta etapa es, primero, determinar si existe una relación entre la temperatura, humedad, acidez y la luminosidad con el porcentaje de viabilidad para los procesos de germinación y de repique; y segundo, encontrar los rangos de variación ideales para dichos procesos.

15.2 ANÁLISIS Y PREPARACIÓN DE DATOS

Se dispondrá como único origen de datos, la Bodega de Datos. La Calidad y limpieza de los datos esta asegurada puesto que han pasado por un proceso previo de ETL. Para poder construir un modelo de Minería de Datos que permita determinar la relación entre las variables ambientales y el porcentaje de viabilidad de los grupos germinativos y de repique se han definido dos vistas minables o vistas de origen de datos que ofrecen una abstracción del origen de datos para los procesos de germinación y repique. Ver Figura 41 y Figura 42.

En la vista minable Germinación, los atributos Temperatura, Humedad, Acidez y Luminosidad, corresponden a los valores promedio obtenidos para estas variables ambientales durante las mediciones automáticas realizadas al grupo Germinativo en el invernadero y son las variables de entrada en el modelo de Minería de Datos. El atributo Clasificación es la variable discreta del modelo y puede tener los valores de clasificación del grupo Germinativo: Excelente, Bueno, Regular, Malo y Deficiente. Los valores de clasificación dependen del rango en que se encuentra el porcentaje de viabilidad promedio del grupo Germinativo (PV_TOTAL). Los atributos restantes de la vista minable permiten desplegar información complementaria al usuario. La estructura de la vista minable Repique es análoga a la de Germinación pero sus datos corresponden al proceso de Repique.

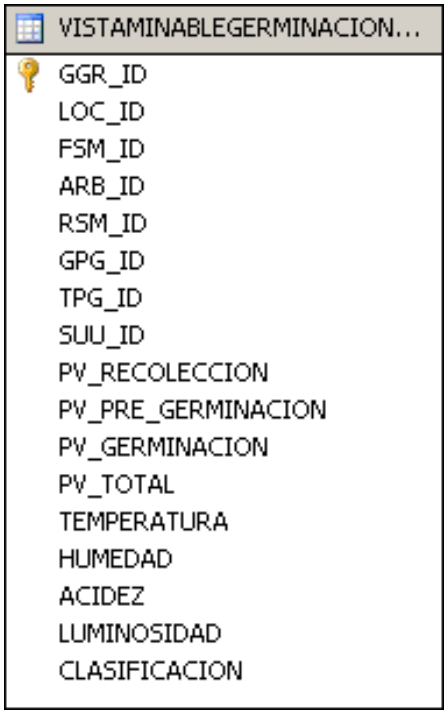


Figura 41. Vista Minable Germinación

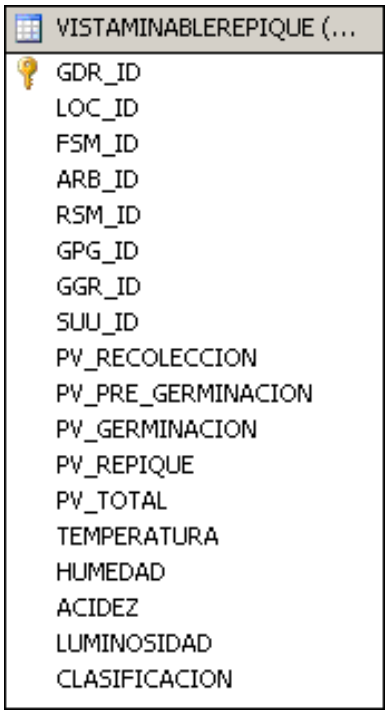


Figura 42. Vista Minable Repique

NOMBRE	DESCRIPCIÓN
GGR_ID	Id del grupo germinativo.
LOC_ID	Id de la localización geográfica.
FSM_ID	Id de la fuente semillera.
ARB_ID	Id del árbol productor.
RSM_ID	Id de la recolección.
GPG_ID	Id del grupo pregerminativo.
TPG_ID	Id del tratamiento pregerminativo.
SUU_ID	Id del sustrato en uso del grupo germinativo.
PV_RECOLECCION	Porcentaje de viabilidad de la recolección.
PV_PRE_GERMINACION	Porcentaje de viabilidad del grupo pregerminativo.
PV_GERMINACION	Porcentaje de viabilidad del grupo germinativo.
PV_TOTAL	Porcentaje de viabilidad promedio.
TEMPERATURA	Temperatura promedio del grupo germinativo.
HUMEDAD	Humedad promedio del grupo germinativo.
ACIDEZ	Acidez promedio del grupo germinativo.
LUMINOSIDAD	Luminosidad promedio del grupo germinativo.
CLASIFICACION	Clasificación del grupo germinativo por su PV_TOTAL.

Tabla 30. Descripción de las columnas de la Vista Minable Germinación

NOMBRE	DESCRIPCIÓN
GDR_ID	Id del grupo repique.
SUU_ID	Id del sustrato en uso del grupo de repique.
PV_RECOLECCION	Porcentaje de viabilidad de la recolección.
PV_PRE_GERMINACION	Porcentaje de viabilidad del grupo pregerminativo.
PV_GERMINACION	Porcentaje de viabilidad del grupo germinativo.
PV_REPIQUE	Porcentaje de viabilidad del grupo de repique.
PV_TOTAL	Porcentaje de viabilidad promedio.
TEMPERATURA	Temperatura promedio del grupo de repique.
HUMEDAD	Humedad promedio del grupo repique.
ACIDEZ	Acidez promedio del grupo repique.
LUMINOSIDAD	Luminosidad promedio del grupo repique.
CLASIFICACION	Clasificación del grupo de repique por su PV_TOTAL

Tabla 31. Descripción de las columnas de la Vista Minable Repique

15.3 MODELAMIENTO

El algoritmo de Minería de Datos es el mecanismo que crea modelos de Minería de Datos. Microsoft SQL Server 2005 Analysis Services proporciona varios algoritmos que se pueden usar en los modelos de Minería de Datos. Estos algoritmos son un subconjunto de todos los algoritmos que pueden utilizarse en Minería de datos. Analysis Services incluye los siguientes tipos de algoritmos [27]:

- Algoritmos de clasificación, que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos. Entre ellos están: algoritmo de árboles de decisión de Microsoft, algoritmo Bayes Naive de Microsoft y algoritmo de Red Neuronal Microsoft.

- Algoritmos de regresión, que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos. Entre ellos están: algoritmo de Árboles de Decisión de Microsoft, algoritmo de Serie Temporal de Microsoft y algoritmo de Red Neuronal Microsoft.

- Algoritmos de segmentación, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares. Entre ellos esta el algoritmo de Clústeres de Microsoft.

- Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden utilizarse en un análisis de la cesta de compra. Entre ellos esta el algoritmo de Asociación de Microsoft.

- Algoritmos de análisis de secuencias, que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas Web. Entre ellos esta el algoritmo de Clústeres de Secuencia de Microsoft.

De todos los algoritmos, el tipo de algoritmo más apropiado para el proyecto es el de clasificación, porque permite realizar predicciones sobre atributos discretos, en este caso la columna clasificación de las vistas minables de germinación y repique.

15.4 EVALUACIÓN

Se evaluaron tres modelos de Minería de Datos correspondientes a los algoritmos de Árboles de Decisión, Bayes Naive y de Red Neuronal de Microsoft con el uso de la herramienta de gráficos de precisión del SQL Server Business Intelligence Development Studio. Específicamente se utilizaron gráficos de elevación que permitieron comparar la precisión de las predicciones de cada modelo a partir de varios conjuntos de datos simulados con un promedio de 3550 registros cada uno. Los resultados obtenidos para un conjunto de datos se muestran en la Figura 43 y la Tabla 32. Resultados de la evaluación

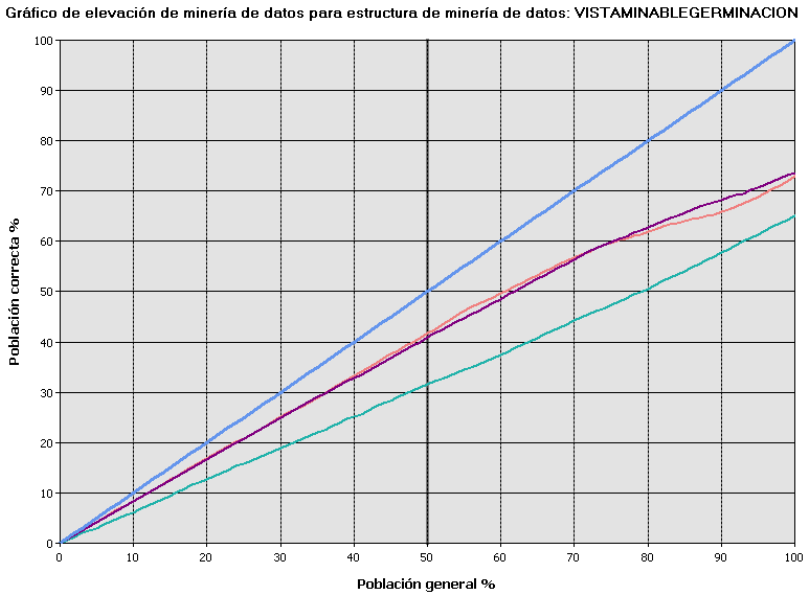


Figura 43. Gráfico de elevación

MODELO	PUNTUACIÓN	POBLACIÓN	PREDICCIÓN
Árboles de Decisión	0,79	41,75%	81,41%
Naive Nayes	0,63	31,56%	64,95%
Red Neuronal	0,63	31,56%	64,51%
Modelo ideal		50,00%	

Tabla 32. Resultados de la evaluación

El modelo de Árboles de Decisión tuvo un porcentaje mayor de predicción frente a los otros dos modelos, lo que significa que es el modelo que más se acerca al modelo ideal, es decir, un modelo teórico que predice el resultado correcto el 100% de las veces.

15.5 DESPLIEGUE

Al proyecto GreenDSS se integró la biblioteca Data Mining Web Controls de Microsoft. Esta biblioteca proporciona una versión ligera de los visores del modelo de Minería de Datos. Data Mining Web Controls permite explorar modelos de Minería de Datos complejos desde la Web. Específicamente se seleccionó el control DMDecisionTreeViewer para el modelo de Árboles de Decisión. Ver Figura 44 y Figura 45.

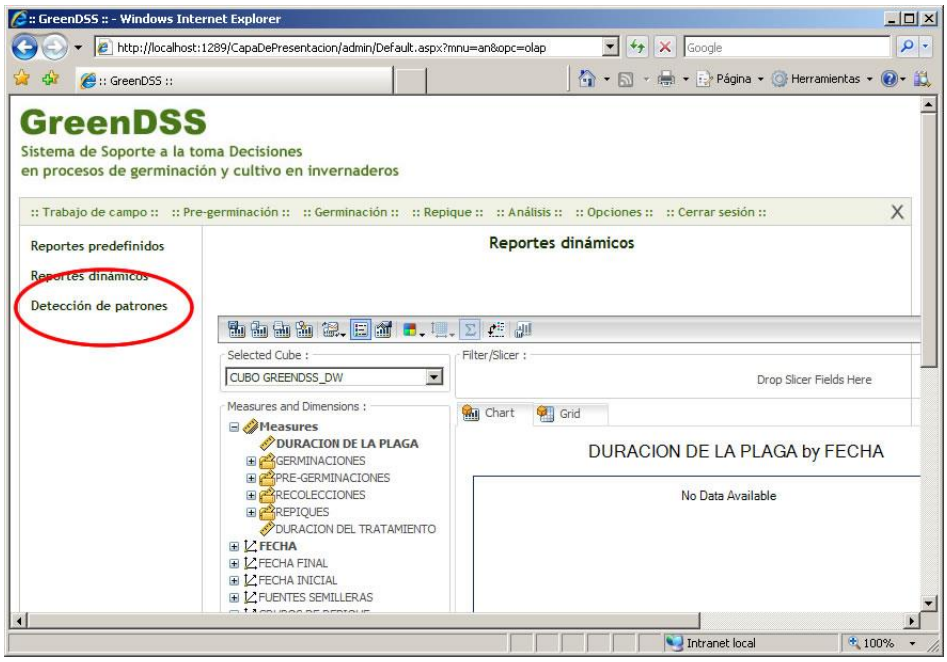


Figura 44. Integración del componente de Minería de Datos

El control DMDecisionTreesViewer permite mostrar el árbol de decisión completo para cada vista minable y realizar las siguientes operaciones:

- Expandir y contraer nodos.
- Seguir las divisiones en árboles de decisión.
- Realizar sombreados en función de la compatibilidad con un estado del atributo de predicción.
- Examinar mediante la información sobre las distribuciones de un nodo determinado.

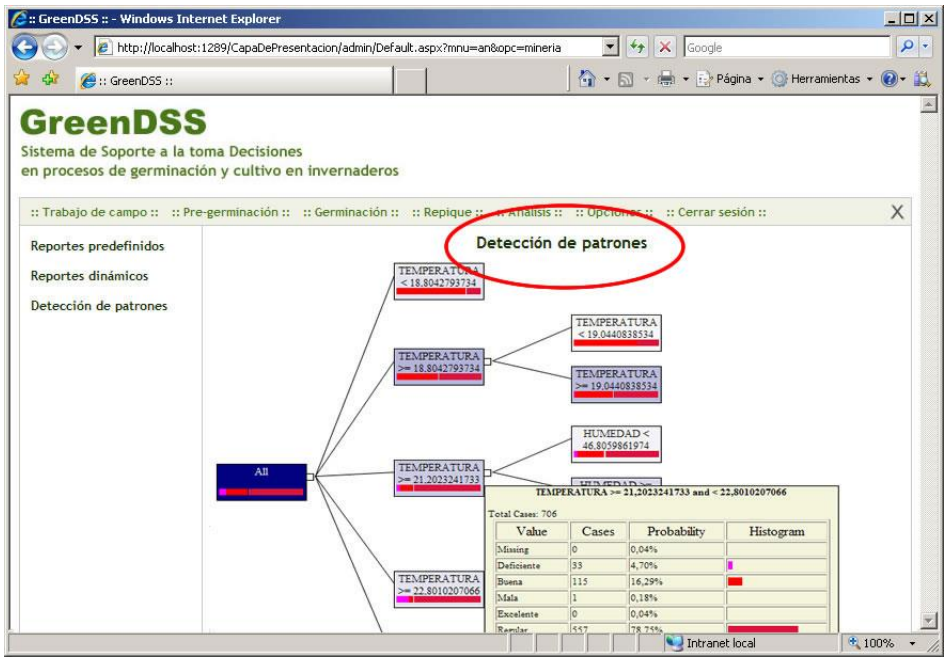


Figura 45. Data Mining Web Controls en GreenDSS

16 PROBLEMAS Y SOLUCIONES

El principal problema de la creación del modelo de Minería de Datos, consistió en no contar con un conjunto de datos reales que permitieran comprobar si el modelo de Minería de Datos seleccionado funciona correctamente en un ambiente de producción. La elección del modelo se realizó con base a datos simulados por una aplicación desarrollada para tal fin. Los datos se generaron tan reales como fue posible y con predisposiciones hacia ciertos patrones que posteriormente sirvieron como criterio de precisión de los modelos evaluados. Por ejemplo, la aplicación se programaba para generar datos con una predisposición a arrojar porcentajes de viabilidad más altos si el rango de temperatura estaba entre los 21° y 23° y la humedad entre el 40% y 50%. Conociendo la tendencia de antemano, se podía determinar que modelo tienen mayor precisión de predicción.

Durante las pruebas iniciales en la fase de evaluación de los modelos de Minería de Datos se presentaron problemas con las predicciones cuando el volumen de datos simulados era muy bajo. A medida que la cantidad de datos aumenta el comportamiento de los modelos mejora.

PARTE 6 – CONCLUSIONES, RECOMENDACIONES Y TRABAJO FUTURO

17 CONCLUSIONES

- Se desarrolló un sistema OLTP Web que gestiona el almacenamiento de los datos recolectados manual o automáticamente en un vivero durante los procesos de recolección, pregerminación, germinación y repique. El sistema OLTP Web se implementó bajo la plataforma Microsoft .NET siguiendo los lineamientos del Proceso Unificado y las especificaciones y recomendaciones propuesta por Microsoft patterns & practices.

- Se desarrolló un Sistema de Soporte a la toma de Decisiones con los siguientes componentes:

  1. Una Bodega de Datos implementada sobre un motor de base de datos Microsoft SQL Server 2005 que cuenta con los modelos dimensionales necesarios para soportar la toma de decisiones en las etapas de recolección, pregerminación, germinación, repique, tratamientos de repique y plagas.
  2. Una herramienta OLAP (Dundas OLAP Services) que se seleccionó e integró al proyecto y que permite realizar análisis multidimensionales dinámicos, visualización gráfica de resultados y la generación y carga de reportes derivados del conjunto de datos almacenados en la Bodega de Datos.
  3. Un prototipo de herramienta para el descubrimiento de patrones y tendencias potencialmente útiles para los procesos de Germinación y Repique a partir del conjunto de datos almacenados en la Bodega de Datos, con el uso del modelo de Minería de Datos de Árboles de Decisión de Microsoft soportado en la plataforma Microsoft Analysis Services 2005.

- Durante el desarrollo del proyecto se tuvo la oportunidad de interactuar con personas de los programas académicos de Ingeniería Forestal e Ingeniería Física de la Universidad Cauca. Esta experiencia fue muy enriquecedora a nivel personal y profesional porque se compartió y adquirió conocimientos desde cada campo disciplinar. Por esta razón es importante impulsar en la Universidad del Cauca la conformación de iniciativas o proyectos interdisciplinarios como apoyo a los procesos de formación profesional.

- A nivel de desarrollo se destaca las ventajas del uso de los application blocks del Enterprise Library para la construcción de aplicaciones sobre la plataforma Microsoft .NET. Su inclusión simplifica los procesos de acceso a datos, administración y manejo de excepciones, configuración y encriptación.

- Fue muy útil seguir las guías para el desarrollo de aplicaciones .NET de Microsoft patterns & practices porque son una gran fuente de ayuda al proporcionar un conjunto de prácticas, lineamientos y recomendaciones ampliamente probados en la industria y que contribuyen a mejorar la calidad de las aplicaciones.

- Dado que el proyecto contemplaba la realización de tres productos software con características distintas, fue acertada la decisión de utilizar una metodología específica para cada producto. Además, el uso de una plataforma única de desarrollo, en este caso Microsoft .NET contribuyó a agilizar el proceso de desarrollo y la integración transparente de los productos obtenidos.

- En la Bodega de Datos, la naturaleza y necesidades del proyecto requerían conocer la trazabilidad de una semilla a lo largo de los procesos de recolección, pregerminación, germinación y repique. En la literatura revisada no se encontró una guía o modelo que permitieran realizar este tipo de análisis. La solución encontrada consistió en relacionar la tabla de hechos con dimensiones que aunque no pertenezcan a su área del negocio específica, ayudan a responder preguntas sobre la trazabilidad.

18 RECOMENDACIONES Y TRABAJO FUTURO

- Realizar la implementación del proyecto en un entorno de producción con usuarios, situaciones y datos reales que permitan evaluar la eficiencia, eficacia y viabilidad de cada uno de los componentes desarrollados y realizar los ajustes convenientes. Además ampliar el área de aplicación a especies no arbóreas. Desafortunadamente, por circunstancias ajenas al proyecto no se contó con el invernadero, dispositivos de medición, material de prueba (semillas y plántulas) y usuarios finales que permitieran realizar pruebas en un entorno real de producción, y se necesito recurrir a pruebas con datos simulados.

- Agregar la capacidad de operar sobre dispositivos móviles en ambientes desconectados a los módulos de Fuentes semilleras, Árboles productores y Recolecciones de material de propagación del sistema OLTP Web, dado que son tareas que comúnmente se realizan en zonas rurales de difícil acceso. El uso de dispositivos móviles garantizaría una mayor integridad de los datos recolectados porque eliminaría su proceso de transcripción, además permitiría realizar corroboraciones directas con los datos almacenados.

- Finalmente, se recomienda seguir profundizando en el estudio de la adopción de las tecnologías de Bodegas de Datos, OLAP, Minería de Datos, y demás involucradas en el soporte a la toma de decisiones hacia nuevos escenarios distintos al administrativo. Por la experiencia adquirida en el desarrollo de este proyecto de grado podemos afirmar que detrás de estas áreas hay grandes oportunidades a nivel investigativo, laborales y de emprendimiento para la Universidad del Cauca, el Departamento de Sistemas y sus egresados.

REFERENCIAS BIBLIOGRÁFICAS

[1]. GÓMEZ, Diego Fernando. Construcción de lo posible, un marco prospectivo para el desarrollo del país. Medellín - Colombia. Centro de Estudios en Economía Sistémica. 2005. ISBN: 958-97719-0-4.
[2]. GÓMEZ, Diego Fernando. Repensando el desarrollo, una aproximación sistémica. Medellín - Colombia. Centro de Estudios en Economía Sistémica. 2005. ISBN: 958-97719-1-2.
[3]. Gartner Group. The Outlook for Business Intelligence and Data Warehousing (Visitado 2006, Noviembre 26). URL: http:// www.gartnerpress.com/reports.
[4]. Computerworld. Business Intelligence at Age 17 (Visitado 2006, Noviembre 26). URL: http://www.computerworld.com/action/article.do?command=view ArticleBasic&articleId=266298.
[5]. Hyperion Solutions Ibérica. Business Intelligence: El Tesoro de Saber Utilizar el Tiempo (Visitado 2006, Noviembre 26). URL: http://www.hyperion.es/downloads/es/Saber_utilizar_el_tiempo_Mayo_05.pdf.
[6]. ACOSTA Mejía, Lisandro., MUÑOZ Córdoba, Jaime Alberto. Sistema de Apoyo para la Toma de Decisiones en Unicauca Virtual utilizando Data Warehouse y OLAP. Colombia, Universidad del Cauca, 2006.
[7]. ROZO Ibañez, Durvvin A. Control y Monitoreo de Variables Ambientales Utilizando PLC y SCADA. Revista Colombiana de Tecnología Avanzada. ISSN: 1692-7257. Volumen 2, 2003.
[8]. RAMOS Fernández, C., HERRERO Durá, J.M., RODRÍGUEZ García, S. Automatización de Invernaderos Mediante Sistemas de Control Distribuidos Industriales. Universidad Politécnica de Valencia. (Visitado 2006, Julio 28) URL: http://ctl-predictivo.upv.es/documentos/congresos/2001/agro01.pdf.
[9]. GUZMÁN, J. L., BERENGUEL, M., RODRÍGUEZ, F. Laboratorio Remoto para el Control de una Maqueta de Invernadero. España, Universidad de Almería.
[10]. TABARES Tovar, Juan Carlos. Sistema de Control de Humedad, Temperatura y Riego para Invernaderos Industriales. Venezuela, Universidad Nacional Experimental Politécnica de la Fuerza Armada Nacional – UNEFA, 2006.
[11]. University of Nebraska-Lincoln. National Agricultural Decision Support System (Visitado 2006, Agosto 10). URL: http://nadss.unl.edu/.
[12]. Prairie Agriculture Research Initiative Decision Support System (Visitado 2006, Agosto 6). URL: http://paridss.usask.ca/index.html.
Instituto Nacional de Tecnología Agropecuaria. Sistema de Soporte de Decisiones para la Producción Agrícola de los Valles Cordilleranos Patagónicos (Visitado 2006, Agosto 10) URL: http://www.inta.gov.ar/bariloche /desarrollo/gesrural/documentos/SSDVC%20proyecto.pdf.
[13]. Instituto Nacional de Tecnología Agropecuaria. Sistema de Soporte de Decisiones para la Producción Agrícola de los Valles Cordilleranos Patagónicos (Visitado 2006, Agosto 10) URL: http://www.inta.gov.ar/bariloche /desarrollo/ gesrural/documentos/SSDVC%20proyecto.pdf.
[14]. KIMBALL, Ralph. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouse. Wiley Computer Publishing, 1998.
[15]. LARRIERA, Gustavo. Analysis Services 2005 para Desarrolladores (Visitado 2006, Agosto 9). URL: http://www.microsoft.com/spanish /msdn/academico.mspx.
[16]. LARMAN, Craig. UML y Patrones: Introducción al Análisis y Diseño Orientado a Objetos. Prentice Hall, 1999.
[17]. Building Distributed Applications. Application Architecture for .NET: Designing Applications and Services. (Visitado 2007, Julio 25). URL: http://msdn2.microsoft.com/en-us/library/ms954595.aspx.
[18]. ZVENGER, Patricia A. Introducción al Soporte de Decisiones, Incorporación de Soluciones OLAP en Entornos Empresariales. Argentina, Universidad Nacional del Sur, 2005.
[19]. INMON, W. H. Building the Data Warehouse. Second Edition. Wiley Computer Publishing, 1996.

[20]. DATE, C. J. Introducción a los Sistemas de Bases de Datos. Séptima edición. Addison Wesley Iberoamericana, 2001.
[21]. Latino BI. Inteligencia de Negocios: Minería de Datos para Apoyar la Toma de Decisiones (Visitado 2006, Septiembre 2). URL: http://www.latino-bi.com/.
[22]. LAROSE, Daniel T. Discovering Knowledge in Data: An Introduction to Data Mining. Wiley Computer Publishing, 2005.
[23]. 1keydata. Información relacionada con los conceptos de OLAP, ROLAP y MOLAP. (Visitado 2007, Noviembre 10). URL: http http://www.1keydata.com/datawarehousing/datawarehouse.
[24]. Data Warehousing Review. Designing OLAP Solutions. (Visitado 2007, Noviembre 10). URL: http://www.dwreview.com/OLAP/OLAP_Comparison.html
[25]. Libros en pantalla de SQL Server 2005. Introducción al Analysis Services 2005. (Visitado 2007, Noviembre 10). URL: http://technet.microsoft.com/es-es/library/ms130214.aspx
[26]. IBARRA, María de los Ángeles. Procesamiento Analítico en Línea (OLAP). Argentina, Universidad Nacional del Nordeste.
[27]. Libros en pantalla de SQL Server 2005. Algoritmos de minería de datos. (Visitado 2007, Octubre 18). URL: http://technet.microsoft.com/es-es/library/ ms175595.aspx