# SISTEMA DE SOPORTE A LA TOMA DE DECISIONES PARA EL PROYECTO DE INVESTIGACIÓN ANÁLISIS MULTIFRACTAL DEL GENOMA HUMANO

**UNIVERSIDAD DEL CAUCA**

**Alba Viviana Camayo Otero**

**Adrian Fernando Martinez Molina**

**UNIVERSIDAD DEL CAUCA**
**FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**
**PROGRAMA INGENIERÍA DE SISTEMAS**
**Grupo de Investigación GTI I + D**
**Grupo de Investigación y Desarrollo en Tecnologías de la Información**
**Popayán**
**2.009**

# SISTEMA DE SOPORTE A LA TOMA DE DECISIONES PARA EL PROYECTO DE INVESTIGACIÓN ANÁLISIS MULTIFRACTAL DEL GENOMA HUMANO

**UNIVERSIDAD DEL CAUCA**

**Alba Viviana Camayo Otero**

**Adrian Fernando Martinez Molina**

Trabajo de investigación para optar al título de Ingenieros de Sistemas

Director:
MSc. Martha Eliana Mendoza

**UNIVERSIDAD DEL CAUCA**
**FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**
**PROGRAMA INGENIERÍA DE SISTEMAS**
**Grupo de Investigación GTI I + D**
**Grupo de Investigación y Desarrollo en Tecnologías de la Información**
**Popayán**
**2.009**

# AGRADECIMIENTOS

Gracias…

A Dios por permitirnos cumplir esta meta y a María Auxiliadora por su protección e intersección.

A nuestros padres: Carmen Lucia Otero, Ángel Miro Camayo, Belén Molina, Servio Martinez y a nuestras familias por su apoyo incondicional, sin su patrocinio y paciencia nada de esto hubiera sido posible.

A nuestra directora, la Magister Martha Eliana Mendoza por su apoyo a lo largo de todo el desarrollo de este proyecto, por su visión y organización, sin esta no hubiéramos alcanzado la meta propuesta.

Al Ingeniero Ember Ubeimar Martinez, por sus empecinadas ideas, sin ellas no nos hubiéramos enfrentado a tantos desafíos que hicieron de esta experiencia, algo de nunca olvidar.

Al grupo BIMAC, por su apoyo y sus grandes ideas, las cuales son gestoras de más conocimiento, conocimiento que contribuye con el bienestar de la humanidad, en especial a la Profesora Patricia Vélez, gracias por acogernos en su casa (Laboratorio).

A los chicos del laboratorio de Bioinformatica, Yohn Jairo Acosta, John Jaime Delgado, Carlos Tellez, Adrian Rodriguez, por los buenos momentos que nos hicieron pasar, por los trasnochos compartidos y por tantos momentos vividos que hacen que de este proyecto una experiencia inolvidable.

A nuestros amigos quienes nos apoyaron a lo largo de la carrera y que nos dieron palabras de ánimo y aliento en todo momento, especialmente a los Joscos, a Rodrigo, Boris y Cuellar, a Diego Ruiz, a Diego Luna, John Betancourt, Hector Alarcon, Jimmy Certuche, Ruben Orozco, Jorge Galindez, Jarvein Mauricio Rivera, Mauricio Hurtado, Gustavo Aponzá, Andres Felipe Manzano, Maria Virginia Lucena, Norma Rivera y Oscar Rendón. Gracias totales incluyendo todos los que se nos olvidan.

Alba Camayo y Adrian Martinez

# TABLA DE CONTENIDO

## LISTA DE FIGURAS

# LISTA DE TABLAS

# RESUMEN

El objetivo de este proyecto de grado es construir una herramienta para el soporte a la toma de decisiones para la investigación Análisis Multifractal del Genoma Humano y generar un conjunto de recomendaciones para la construcción de sistemas que posean características similares a las expuestas por el problema planteado.

La herramienta construida tiene dos componentes, el primero es la bodega de datos en la cual se encuentran almacenados todos los datos necesarios para generar los reportes que los usuarios finales requieran y el segundo es la herramienta OLAP la cual le proporciona al usuario final una mejor visualización de las consultas requeridas.

# CAPÍTULO I. INTRODUCCIÓN

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

El proyecto de investigación llamado “Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información” (Análisis Multifractal Del Genoma Humano), es desarrollado por el Grupo de Biología Molecular, Ambiental y Cáncer (BIMAC)[1] y el Grupo de I+D en Tecnologías de la Información (GTI), en asocio con la Universidad del Valle, Universidad de Cantabria (Santander, España) y Triesta Sciences (India) Pvt. Ltd. (Bangalore, India/Menlo Park, California). Esta investigación fue financiada por COLCIENCIAS y la Universidad del Cauca.

La investigación “Análisis Multifractal del Genoma Humano” aplica el análisis multifractal con el objeto de cuantificar la variación en la información genética y generar una clasificación del genoma humano con significado biológico y potencial uso tecnológico, explicando así la irregularidad topológica en la distribución de secuencias codificantes y no codificantes, a lo largo de los cromosomas [25].

Para llevar a cabo las tareas propias de la investigación, el equipo de trabajo accede al GenBank [10], el cual es un repositorio de datos público en que se encuentran almacenadas secuencias de nucleótidos y proteínas, las cuales son el insumo de esta investigación.

Los investigadores del proyecto análisis multifractal del genoma humano, acceden al FTP del GENBANK y descargan las secuencias en un formato de texto plano, estos registros contienen toda la información relacionada con un cromosoma. Posterior a la descarga, los investigadores realizan todos los análisis requeridos por la investigación y generan resultados, que son almacenados en archivos de texto, que posteriormente serán utilizados para realizar análisis estadísticos, los cuales tienen por objeto comprobar la hipótesis concebida para la realización de la investigación.

Generar una clasificación del genoma humano y cuantificar la variación en la información genética, requiere para fines de esta investigación realizar comparaciones genómicas; es decir realizar comparaciones entre las características de un conjunto de genes, tales como longitud, número de unidades de información, función, etc. y algunos otros datos obtenidos a través de la manipulación de las secuencias; como ejemplo, de estos datos se encuentra la dimensión fractal, la regresión lineal y el promedio de genes existentes en un rango determinado por los investigadores sobre la secuencia de un cromosoma dado [25].

[1] http://bimac.unicauca.edu.co/

La búsqueda de información relevante para la investigación mediante la interpretación de los datos resultantes, representa un desafío debido a la ausencia de almacenamiento estructurado que permita centralizar e integrar la gran cantidad de archivos planos y los datos que estos contienen. Además existen otras limitantes en el manejo de la información resultante de la investigación, tal como lo son: dificultad para realizar análisis estadísticos, detección de posibles errores de consistencia en los campos de los archivos provenientes del GenBank, encontrar relaciones o asociaciones entre las variables involucradas y la ausencia de una herramienta para representar gráficamente los datos resultantes.

Para contribuir en la integración, organización y análisis de los datos obtenidos en la investigación "Análisis Multifractal del Genoma Humano", este trabajo de grado plantea como una alternativa de solución, la construcción de un sistema de soporte a la toma de decisiones (DSS) soportado por una bodega de datos (DW) que permite extraer los datos de los archivos planos generados por los procesos de la investigación, integrarlos, depurarlos e involucrarlos dentro de un mismo repositorio, y una herramienta OLAP para manipular y analizar los datos existentes, presentando relaciones y estadísticas entre los datos, de una forma sencilla para los investigadores.

**JUSTIFICACIÓN**

El grupo BIMAC y GTI, junto a otros grupos e instituciones ganaron la convocatoria realizada por COLCIENCIAS, para la creación del "**Centro de Excelencia** en **Metagenómica y Bioinformática**", bajo la propuesta llamada "**Conformación de una plataforma en metagenómica y bioinformática para la caracterización y el aprovechamiento de recursos genéticos de ambientes extremos**", el desarrollo de este proyecto representa un aporte para el Centro de Excelencia en meta genómica y bioinformática, debido a que con la implantación de estas tecnologías se pretende generar un soporte tecnológico, que les permita a los investigadores del Centro de Excelencia consultar información relevante a la investigación "Análisis Multifractal del Genoma Humano" de una forma integrada y con una visualización de los resultados estructurada.

Debido al crecimiento de los archivos que contienen los resultados de la investigación y al no tener un repositorio que permita integrar los datos resultantes de la investigación "Análisis Multifractal del genoma humano", este proyecto plantea una alternativa de solución con la construcción de un DW, el cual permite recolectar los datos resultantes de la investigación, integrarlos, depurarlos e involucrarlos dentro de un mismo repositorio. Además se construyo un prototipo de herramienta OLAP la cual permite manipular y analizar los datos existentes, presentando relaciones y estadísticas entre los datos, de una forma sencilla para los investigadores.

Este proyecto representa una solución innovadora en el área de aplicación (BioInformática), porque no se encontró un DSS que solucione el problema planteado en este Trabajo de grado, teniendo en cuenta que la investigación presenta características de análisis de información muy particulares. Por lo tanto con la realización de un DSS en el área de la bioinformática se abre un nuevo campo de aplicación en el uso de estos sistemas generando una experiencia en el desarrollo de este tipo de aplicaciones para esta área en particular, y permite realizar un aporte académico a la temática de DW, con la generación de recomendaciones para el diseño y construcción de un DW en el campo de la bioinformática, que tengan en cuenta características similares a las presentadas en la investigación "Análisis Multifractal del Genoma Humano", buscando de esta manera contribuir en el área de Aplicaciones de un DW.

En este proyecto se construyo una solución de un DSS basado en tecnologías de DW y una herramienta OLAP, las cuales permiten a los usuarios hacer análisis dimensionales de los datos almacenados y de esta manera contrastar estos análisis con los resultados obtenidos en su investigación, de igual manera la solución permitirá que los investigadores publiquen su investigación y permitan que otros investigadores realicen sus propias consultas y saquen sus propias conclusiones.

Este proyecto de grado fue posible realizarlo gracias a que se dispone de toda la infraestructura tecnológica necesaria para su diseño e implementación, así mismo, se contó con las metodologías, técnicas y teorías propuestas por Ralph Kimball [19] para el desarrollo de Bodegas de Datos y con el apoyo de la Universidad del Cauca que ofrece la infraestructura y personal altamente calificado para guiar el desarrollo del proyecto. De la misma manera se cuenta con una fuente de datos que da soporte al DSS.

**OBJETIVOS**

**Objetivo General**

- Desarrollar un sistema de soporte a la toma de decisiones estratégicas (DSS) basado en tecnologías de DW y OLAP, que centralice los datos provenientes de las diferentes fuentes de la investigación "Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información", permitiendo el análisis que de esta información deben realizar los investigadores.

**Objetivos Específicos**

- Diseñar y construir un DW que centralice los datos provenientes de la investigación "Análisis Multifractal del Genoma Humano", permitiendo almacenar información relacionada con los genes, unidades de información y proteínas de dicha investigación.

- Construir un prototipo de herramienta OLAP que utilice herramientas y/o componentes, que permita representar de forma gráfica y en grillas los datos obtenidos de la investigación "Análisis Multifractal del Genoma Humano" para su respectivo análisis.

- Generar recomendaciones para el diseño y construcción de un DW en el campo de la bioinformática, que aporten a la construcción de aplicaciones que tengan características similares a la investigación "Análisis Multifractal del Genoma Humano", buscando de esta manera contribuir en el área de Aplicaciones de un DW.

**ESTRUCTURA DEL DOCUMENTO**

El documento se compone de las siguientes secciones:

**CAPÍTULO I. INTRODUCCIÓN:** Permite tener una visión general del proyecto, el origen y la justificación de la idea a realizar.

**CAPÍTULO II. MARCO CONCEPTUAL:** Contiene los conceptos teóricos necesarios utilizados para el desarrollo del proyecto, información fundamental de conceptos como DSS, DW y OLAP.

**CAPITULO III - DESCRIPCIÓN DEL PROCESO DE DESARROLLO DEL DSS:** Se hace una descripción de todo el ciclo de vida usado para el desarrollo del sistema de DSS, desde su concepción inicial a hasta su implementación y despliegue.

**CAPITULO IV - DESCRIPCIÓN DEL PROCESO DE DESARROLLO DE LA HERRAMIENTA OLAP:** Se describe el proceso de desarrollo que se utilizo para la construcción de la herramienta OLAP y se presentan los artefactos que se obtuvieron en cada una de las fases de la metodología de trabajo.

**CAPITULO V. DESCRIPCIÓN DE PROBLEMAS Y LA SOLUCIÓN ENCONTRADA:** En este capítulo se describen los problemas que presentó la realización del proyecto y de la misma manera se describe las estrategias de solución encontradas para estos.

**CAPITULO VI. RECOMENDACIONES PARA EL DISEÑO DE DATA WAREHOUSE:** Contiene la descripción y/o adaptación de diferentes conceptos teóricos utilizados en el diseño e implementación de la DW en el campo de la bioinformática.

**CAPÍTULO VI: CONCLUSIONES Y TRABAJOS FUTUROS:** En este capítulo se presentan las conclusiones del proyecto y las recomendaciones para futuros trabajos en el área.

**CAPÍTULO VII: BIBLIOGRAFÍA:** En este capítulo se presenta la bibliografía empleada para la realización del proyecto

# CAPÍTULO II. MARCO CONCEPTUAL

## *2.1. Sistemas de Información (SI) [9]*

Un sistema de información puede definirse como un conjunto de procesos y componentes interrelacionados que tienen como objetivo **recopilar, elaborar y distribuir** información relativa a las operaciones y actividades de control y dirección, apoyando actividades de acuerdo a las estrategias del negocio.

En un sistema de información todos los elementos pertenecientes a él, pueden ser clasificados en cualquiera de las siguientes categorías: personas, datos, actividades o técnicas de trabajo y recursos materiales en general (Ej. recursos informáticos y de comunicación). Dichos elementos interactúan entre sí, con el objetivo de procesar los datos y la información, para posteriormente distribuirla adecuadamente en la organización teniendo en cuenta los objetivos de esta.

La Figura 1 ilustra todos los elementos de un sistema de información y las relaciones entre estos.

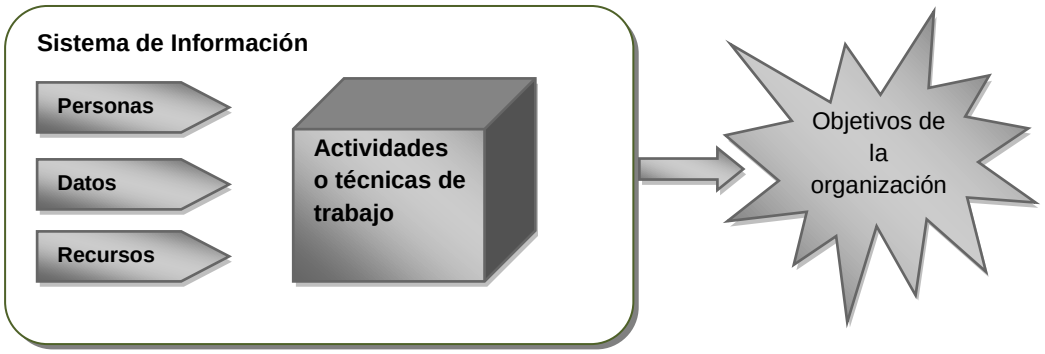



**Figura 1. Elementos de un sistema de información**

**Clasificación de los sistemas de información** [4]
Existen dos tipos de clasificación de los sistemas de información, de acuerdo a la función que desempeñan o según el entorno de aplicación.

1. Según la función que desempeñan:

    a. Sistemas transaccionales (OLTP): Sistemas usados para el procesamiento de datos (almacenamiento, clasificación, cálculos, etc.), de las actividades transaccionales de una empresa.

    b. Sistemas de Información Gerencial (MIS): Construidos para satisfacer las necesidades empresariales en términos generales.

    c. Sistemas de soporte a decisiones (DSS): Construidos para apoyar la toma de una decisión específica, a una unidad de la organización específica, con un problema específico.

    d. Sistemas de información para oficinas (OAS): Estos sistemas están orientados a dar solución al trabajo diario de los administrativos de una empresa u organización (envió de correspondencia, generación de reportes, etc.)

2. Según el entorno de aplicación: [4]

    a. Entorno transaccional: Una transacción es una operación atómica por medio de la cual se creo o modifica los datos. En los entornos transaccionales los datos son capturados, manipulados y modificados, al igual que son usados para la preparación de informes, lo importante en este entorno es saber que datos son modificados y como, al finalizar la transacción.

    b. Entorno decisional: En este entorno se encuentran los sistemas de apoyo a la toma de decisiones, teniendo en cuenta que en una organización son tomadas decisiones en todos los niveles, por lo tanto los sistemas pertenecientes a este entorno en caso de ser aplicados, deben adaptarse a cualquier nivel de la organización.

## *2.2. Sistemas de Soporte a la Toma de Decisiones (DSS)*

Existen múltiples enfoques en lo que respecta a la toma de decisiones y se encuentra un amplio espectro en los ámbitos en los cuales se hace necesario tomar decisiones, por lo tanto el concepto de **un sistema de soporte a la toma de decisiones (DSS** Decision support system – por sus siglas en inglés**),** es un concepto también muy amplio. Un DSS puede adoptar diversas formas. En términos generales se puede decir que un DSS, es un sistema de información, el cual es utilizado como apoyo, más que para la automatización del proceso de toma de decisiones de una organización. Este apoyo se ve reflejado en la recopilación de información, la cual mostrada de una manera estructurada, que apoya a un individuo, equipo o grupo de personas en el proceso de la toma de decisiones. En la práctica, las referencias a este tipo de sistemas suelen ser referencias a aplicaciones informáticas que realizan dicha función de apoyo [22].

Para definir un DSS se integrarán varias definiciones dadas por diferentes autores, con lo cual se dirá que un DSS puede ser catalogado como un sistema interactivo que apoyan a los encargados de tomar decisiones, utilizando datos y modelos para resolver problemas no estructurados o semi estructurados [22],[13],[21].

Los DSS son sistemas que típicamente se ubican en el área de la inteligencia de negocios, sin embargo, en los últimos años se ha observado que la arquitectura y el manejo de los datos de estas tecnologías pueden ser adoptados en disciplinas ajenas a la inteligencia de negocios.

El DSS es un sistema que transforma los datos que tiene una organización en información, con el objetivo de tomar decisiones estratégicas basadas en un análisis dimensional, es decir, considerando unas variables en relación con otras y no de forma independiente entre sí, permitiendo enfocar el análisis desde distintos puntos de vista [19].

El DSS traduce los requerimientos de los usuarios en las correspondientes sentencias de consulta para el DW (núcleo central de los DSS) y es el que interpreta los resultados devueltos para mostrarlos según lo solicitado por el usuario por medio de herramientas OLAP que permiten navegar a través de los datos almacenados en el DW y analizarlos dinámicamente desde una perspectiva multidimensional [19].

### 2.2.1.Bodegas de Datos (DW- Data Warehouse)

El DW es una colección de datos integrados, orientado a temas, que dan soporte a las funcionalidades del DSS, donde cada unidad de dato es relevante en algún momento en el tiempo [19].
La información almacenada en el DW es histórica por naturaleza, y es el resultado de transformar, mejorar la calidad e integrar, datos provenientes de bases de datos fuentes pertenecientes a la organización en la cual se implanta el sistema. Estos datos históricos se analizan para tomar

decisiones empresariales a diferentes niveles, desde la planeación estratégica, a la evaluación del rendimiento de una unidad determinada de la organización [19].

El objetivo principal de un DW es apoyar la toma de decisiones. Estos sistemas junto a las "Herramientas OLAP", permite a los analistas, gerentes y ejecutivos sintetizar información sobre la empresa a través de comparaciones, visiones personalizadas, análisis histórico y proyecciones de datos en varios escenarios [19] [4].

**Características.**
Cuatro características son importantes para definir un DW [27]:

- *Orientado a temas:* orientado a temas implica que se seleccionan áreas importantes del negocio, como por ejemplo el área de "Finanzas" en una organización y alrededor de esta se construyen los modelos, buscando responder a una amplia gama de preguntas que los usuarios tengan sobre el área .

- *Integrado:* Porque los datos que contiene, a pesar de que vienen de diferentes fuentes de datos operacionales, están en un formato consistente.

- *No volátil:* Significa que los datos que contiene un DW no son regularmente accedidos o manipulados, el propósito del DW es permitir que se analice lo que ha ocurrido.

- *De tiempo variante:* A diferencia de los sistemas de Procesamiento de Transacciones en Línea (OLTP) donde su funcionalidad no depende del cambio sobre los datos en el tiempo (datos históricos), un DW basa su funcionalidad en la recolección de datos en horizontes de tiempo muy grandes para realizar análisis o pronósticos a través de acontecimientos pasados en la organización.

**Elementos del Data Warehouse (DW) [19].**
La Figura 2 muestra gráficamente los elementos que constituyen un DW y las relaciones existentes entre estos

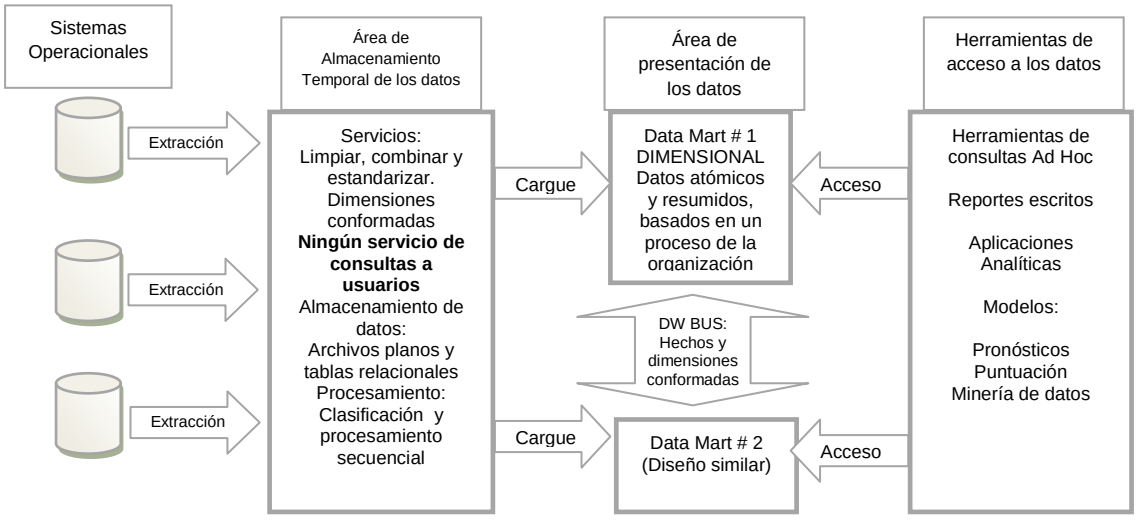



**Figura 2. Elementos básicos de una Bodega de Datos[2].**

[2] Traducción de la figura Basic elements of the data warehouse. [19].

Los elementos básicos que conforman un DW son los siguientes [19]:

**Sistemas Operacionales Fuentes**
Un sistema de registros operacional tiene como función capturar las transacciones de una organización. Esta captura de las transacciones es almacenada en diferentes tipos de repositorios, de acuerdo a las necesidades y criterios de una organización. Dentro de los tipos o formas de almacenamiento, pueden encontrarse los OLTP (OnLine Transaction Processing), archivos planos, tablas de Excel o archivos físicos.

**Área de Almacenamiento Temporal de los Datos (Zona de los datos)**
Parte del DW que es referida a la extracción, limpieza, creación de relaciones y carga de datos desde los sistemas fuente. Esta área se encuentra fuera de los límites de acceso de los usuarios, esto significa que no soporta servicios de consultas y reportes. Las herramientas de limpieza utilizan esta área para resolver problemas de transformación de datos y limpieza, es decir, es un área que se utiliza como paso intermedio para el almacenamiento de los datos antes de ser colocados en el servidor de presentación.

**Área de presentación de los datos**
Máquina física objetivo sobre la cual los datos de la bodega de datos son organizados y almacenados para consultas directas por los usuarios finales, reportes escritos y otras aplicaciones. Si el servidor de presentación es basado sobre tecnología de procesamiento analítico en línea no relacional, entonces las tablas serán organizadas como esquemas estrellas, contrario a esto si el servidor de presentación es basado en una base de datos relacional, entonces las tablas serán organizadas relacionalmente pero aun pueden ser reconocidas las dimensiones sobre estos datos.

**Herramientas de Acceso a los datos**
Un DW puede tener varias formas de acceso o varias formas de visualizar los datos contenidos en la bodega, los cuales pueden clasificarse de la siguiente manera:

- Consultas e informes simples (Consultas Ad Hoc, reportes escritos)
- Cubos OLAP (Aplicaciones analíticas)
- Vistas para minería de datos y Sistemas de previsión empresarial; predicción mediante estudio de series temporales (Modelos).

**Modelo Dimensional [19]**
Disciplina específica para el modelado de datos en el área de los DW, alternativa al modelado de entidad relación (E/R). Un modelo dimensional contiene la misma información que un modelo E/R pero los modelos dimensionales contienen paquetes de datos diseñados en un formato simétrico cuyo objetivo es que sea comprensible a los usuarios, y muestre un gran desempeño en la consulta y resistencia al cambio.

Cada modelo dimensional representa un data mart, el cual es un subconjunto del DW. Cada data mart se construye a partir de dimensiones conformadas y hechos conformados. Estas conforman la arquitectura Bus de las Bodegas de datos, sin dimensiones conformadas, ni hechos conformados, una bodega de datos es sistema aislado.

El modelo dimensional se representa por medio de un esquema estrella, el cual presenta de forma gráfica las tablas de hechos y las dimensiones que conforman un data mart, cada dimensión lleva el nombre y los atributos que la componen, de igual manera la tabla de hechos tiene un nombre y además contiene las llaves foráneas de las dimensiones a las que se encuentra asociada y las medidas de esta. La Figura 3 presentara la estructura gráfica de un esquema estrella.

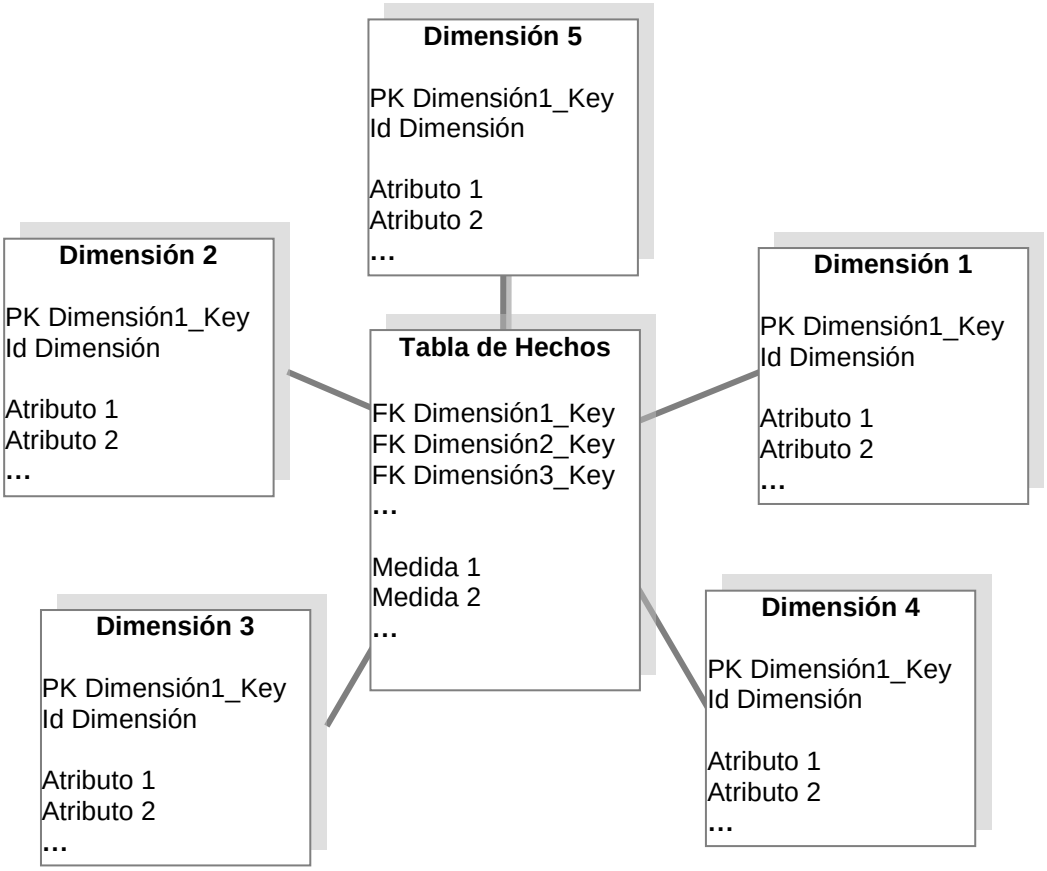



**Figura 3. Esquema Estrella**

Un modelo dimensional tiene los siguientes componentes:

*Dimensiones:*
Una dimensión en un esquema estrella representa una única entidad del negocio, como cliente o producto; son entidades independientes que proveen el contexto para los datos numéricos almacenados en las tablas de hechos, están compuestas por atributos textuales que son descriptivos y que están altamente correlacionados entre sí, ayudan a mostrar múltiples perspectivas de los datos permitiendo realizar análisis y cruces de información que permiten responder preguntas del negocio.

*Atributos:*
Los atributos son una agrupación de elementos dentro de una dimensión. Representan categorías o clases de elementos que tienen el mismo nivel lógico dentro de una dimensión donde todos los elementos de un atributo se relacionan con otros atributos de la dimensión de la misma forma. La finalidad de los atributos es ver la información de cada dimensión a diferentes niveles de detalle y agrupar los datos para ser analizados.

*Jerarquías:*
Representadas por un ordenamiento lógico dentro de la dimensión, se encuentran formadas por los diferentes tipos de relaciones entre los atributos de una misma dimensión. Dentro del contexto de

navegación del modelo dimensional, se puede decir que las diferentes jerarquías definen los caminos para la *"navegación"* sobre los datos.

*Tabla de hechos:*
Una tabla de hechos es una gran tabla dentro del DW que almacena medidas de negocio. La tabla de hechos usualmente contiene hechos y llaves foráneas de las tablas de dimensiones. La tabla de hechos representa datos, usualmente numéricos y aditivos, que pueden ser analizados y examinados.

*Indicadores o Medidas*
Son las variables o métricas que ayudarán a medir el desempeño de un área del negocio modelada.

### 2.2.2.Procesamiento Analítico en Línea (Online Analytical Processing – Olap)

Se define OLAP como el proceso interactivo de crear, mantener, analizar y realizar informes sobre datos, proporciona muchas ventajas a los usuarios que realizan análisis, como por ejemplo [27]:

- Provee un modelo de datos intuitivo y multidimensional que facilitan la selección, recorrido y exploración de los datos.
- Provee un lenguaje analítico de consulta que proporciona la capacidad de explorar las complejas relaciones existentes entre los datos empresariales.
- Ofrece un pre cálculo de los datos consultados con más frecuencia que permite una rápida respuesta a las consultas.

Este proceso permite manipular los datos por dimensiones, lo cual proporciona a los usuarios un fácil acceso a los datos con el objetivo de dar soporte a la toma de decisiones. Esta técnica de análisis, no ejecuta múltiples consultas, en lugar de esto, los datos son estructurados de manera que los usuarios pueden acceder a estos de una manera rápida y fácil, resolviendo las preguntas típicamente formuladas por la organización.

Las herramientas OLAP son capaces de analizar grandes volúmenes de datos, resolver consultas complejas y realizar análisis desde diferentes perspectivas; también permite tener vistas reformateadas y calculadas sin correr el riesgo de dañar los datos, permitiendo que varios usuarios puedan tener la información sin la necesidad de generar copias.

**Conceptos y componentes**

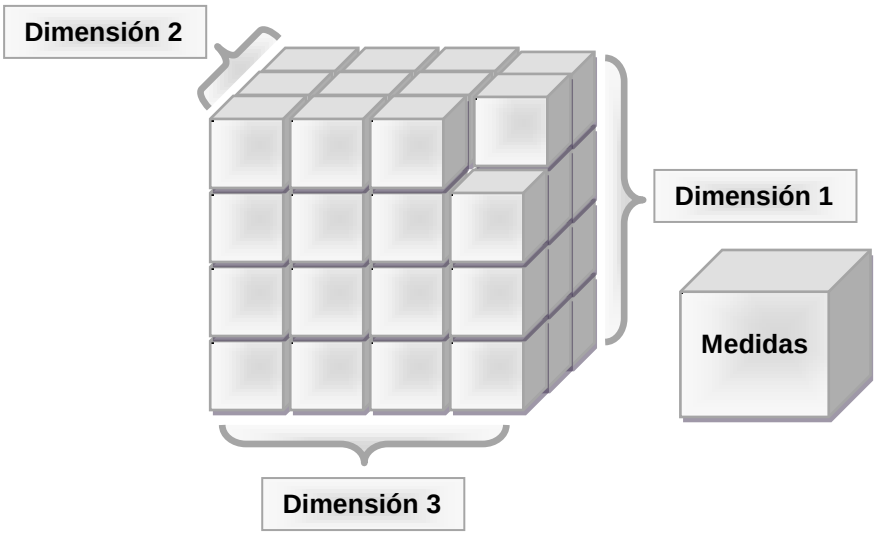



**Figura 4. Cubo tridimensional OLAP**

La Figura 4 muestra la representación gráfica de un cubo OLAP y les componentes los cuales serán explicados a continuación.

**Cubo:** Se denomina cubo a los arreglos o matrices multidimensionales en los cuales es almacenada la información, el cubo contiene información importante para el usuario, la cual es organizada dentro de los cubos en dimensiones y medidas, en una estructura dimensional que es capaz de responder a las preguntas que tienen los usuarios.

**Medida:** Valor que toma una variable de análisis. Las medidas son datos cuantificables, o indicadores de desempeño los cuales son usados para determinar el éxito de un proceso de la organización, las medidas orientan las respuestas a preguntas relacionadas con asuntos de la organización. Se pueden tener medidas regulares, las cuales se encuentran almacenadas en la bodega porque es información que proporciona la fuente, pero también se pueden tener medidas que no son proporcionadas por la fuente, sino calculadas generalmente a partir de las medidas regulares.

**Dimensión:** Los atributos de texto descriptivos, son organizados en dimensiones. Es necesario utilizar criterios de diseño para saber que atributos van a ser almacenados en la bodega y cuales serán descartados.

**Nivel:** Las dimensiones están construidas por niveles, los cuales representan las jerarquías establecidas por la organización y por los modelos de datos que esta usa, los niveles inferiores proveen información más detallada, mientras que los niveles superiores proveen información con menor detalle, esto es conocido como granularidad de la información, a menor grano mas detalle y viceversa.

**Operaciones típicas en una herramienta OLAP**

La información recibida por la herramienta, debe estar estructura y organizada de tal manera que se puedan realizar las siguientes operaciones.

- Drill down y Roll up (profundizar y escalar), permiten visualizar la información a diferentes niveles de análisis, drill down permite ver la información de forma mas detallada, va de lo

general a lo particular, por su parte la operación roll up, va de lo particular a lo general, permitiéndole al usuario ver la información condensada, permitiéndole navegar por los niveles de la información.

- Intercambio de filas por columnas, realizar permutaciones entre dimensiones.
- Respuestas interactivas desde perspectivas diferentes.
- Combinación de varias fuentes contenidas en el DW, para obtener respuestas interactivas.
- Realiza de cálculos de dificultad moderada, tales como sumas, multiplicaciones, divisiones, porcentajes, restas, etc.
- Slice y Dice (cortar - rotar), estas operaciones, permiten navegar por un cubo ya visualizado. Slice le permite al usuario enfocarse en una perspectiva específica y la operación Dice, rota el cubo para presentar otra perspectiva al usuario.

**Tipos de Sistemas OLAP**
Tradicionalmente, los sistemas OLAP se clasifican según las siguientes categorías:

**ROLAP**
Implementación OLAP que almacena los datos en un motor relacional. En esta categoría se puede observar que los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas. Los esquemas más comunes sobre los que se trabaja son estrella ó copo de nieve, aunque es posible trabajar sobre cualquier base de datos relacional. La arquitectura está compuesta por un servidor de banco de datos relacional y el motor OLAP se encuentra en un servidor dedicado. La principal ventaja de esta arquitectura es que permite el análisis de una enorme cantidad de datos.

**MOLAP**
Esta implementación OLAP almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado. Estos valores pre calculados o agregaciones son la base de las ganancias de desempeño de este sistema. Algunos sistemas utilizan técnicas de compresión de datos para disminuir el espacio de almacenamiento en disco debido a los valores pre calculados.

**HOLAP (Hybrid OLAP)**
Almacena algunos datos en un motor relacional y otros en una base de datos multidimensional.

## *2.3. Biología Molecular*

Es una ciencia cuyo objetivo fundamental es la comprensión de todos aquellos procesos celulares, que contribuyen a que la información genética se transmita eficientemente de unos seres a otros, y se exprese en los nuevos individuos [5].

La historia de la biología molecular está marcada por una serie de hitos que contribuyeron decisivamente en su desarrollo: la historia de esta ciencia inicia con Mendel cuando en el año 1866 pública sus experimentos los cuales conducen a los principios de segregación y clasificación independiente de los genes. En 1869 el científico suizo Frederick Miescher descubre en el núcleo de las células una sustancia de carácter ácido a la que llamó nucleina. En los años 20, el químico alemán Robert Feulgen, utilizando una tinción específica, descubrió que el ácido desoxirribonucleico (DNA – por sus siglas en inglés) estaba situado en los cromosomas. En 1944 Avery, McCleod y McCarty comprueban que el DNA es el portador de la información genética. En 1953 Watson y Crick revelan la estructura del DNA como una doble hélice complementaria que

recuerda la estructura de una escalera de caracol. Posteriormente y de forma exponencial se realizan descubrimientos relacionados con las enzimas de restricción (polimerasas) que conducirán a lo que se conoce como tecnología del DNA recombinante [5].

**Nucleótido:** Subunidad del DNA O ácido ribonucleico (RNA – por sus siglas en inglés) compuesto por una base nitrogenada, en el DNA de Adenina, Guanina, Timina y Citosina; en el RNA de Adenina Guanina, Uracilo y Citosina; Millones de nucleótidos se ligan para formar una molécula de DNA o ARN [12].

**Gen:** Unidad física y funcional fundamental de almacenamiento de información y unidad de herencia de los seres vivos. Desde el punto de vista molecular, un gen es una secuencia ordenada de nucleótidos en la molécula de DNA en una posición particular dentro de un cromosoma, un gen contiene la información necesaria para la síntesis de una macromolécula con función celular específica. Puede codificar productos funcionales específicos como: una proteína o una molécula del RNA [12].

**Acido Desoxirribonucleico (DNA):** Molécula que codifica la información genética. El DNA es una molécula de doble hélice que se mantiene unida por débiles enlaces entre pares base de nucleótidos. Los cuatro nucleótidos en el DNA contienen las bases Adenina (A), Guanina (G), Citosina (C) y Timina (T). En la naturaleza, los pares base se forman únicamente entre A – T y entre G – C, así la secuencia de bases de una hélice puede ser deducida mediante sus parejas [12].

**Cromosoma:** Estructura genética de auto replicación de las células que contiene el DNA celular que lleva en su secuencia de nucleótidos los arreglos lineares de genes. En procariotas, el DNA del cromosoma es circular, y el genoma entero es cargado sobre un cromosoma, en Eucariotas el genoma consiste de un numero de cromosomas cuyo DNA es asociado con diferentes tipos de proteínas [12].

**Secuencia de DNA:** Orden relativo de los pares base, ya sea en un fragmento de DNA, un Gen, un cromosoma o en un genoma completo [12].

**Proteína:** Es una larga molécula compuesta de una o mas cadenas de aminoácidos en un orden especifico; el orden es determinado mediante la secuencia de las bases de nucleótidos en el gen que codifican la proteína. Las proteínas son necesarias ya sea para la estructura, la función y regulación de las células del cuerpo, tejidos y órganos, cada proteína tiene funciones únicas. Algunos ejemplos son las hormonas, las enzimas y los anticuerpos [12].

**Genómica:** Consiste en el estudio de genomas y se divide en:

Genómica estructural es la rama de la genómica orientada a la caracterización y localización de las secuencias que conforman el DNA de los genes, la genómica funcional consiste en la recolección sistemática de información sobre la función de los genes, mediante la aplicación de aproximaciones experimentales globales que evalúen la función de los genes haciendo uso de la información y elementos de la genómica estructural. Se caracteriza por la combinación de metodologías experimentales a gran escala con estudios computacionales de los resultados.

La genómica funcional tiene como objeto disminuir la brecha existente entre el conocimiento de las secuencias de un gen y su función, para de esta manera desvelar el comportamiento de los sistemas biológicos. Se trata de expandir el alcance de la investigación biológica desde el estudio de genes individuales al estudio de todos los genes de una célula al mismo tiempo en un momento determinado [11].

**Proteómica:** El proteoma se puede definir como el conjunto de las proteínas expresadas por un genoma. La PROTEOMICA es el estudio de proteomas, Configura una disciplina fundamental de la era post-genómica que trata de descubrir la constelación de proteínas que otorgan a las células su estructura y función. Distintas tecnologías permiten obtener y comparar "instantáneas" de las proteínas que se están expresando en un momento determinado en una célula (robótica, electroforesis 2D, espectrometría de masas, chips, bioinformática) [11].

**Metagenómica:** La metagenómica consiste en el aislamiento de genes expresables y de productos implicados en rutas metabólicas. Dicho aislamiento se consigue mediante secuenciación de bibliotecas de DNA derivadas directamente de muestras del medio ambiente sin cultivar [24].

## *2.4. Proyecto Genoma Humano*

En el año de 1990 se dio inicio al proyecto genoma humano, un esfuerzo coordinado por el departamento de energía de los Estados Unidos y los institutos nacionales de salud de Japón, Francia Alemania, entre otros. Este proyecto culmino 13 años más tarde [26].
Las principales metas del Proyecto Genoma Humano eran [26]:

- Identificar los aproximadamente 20.000 o 25.000 genes en el Genoma Humano.
- Determinar la secuencia de 3 billones de pares base que componen el Genoma Humano.
- Almacenar la información en Bases de Datos.
- Mejorar las herramientas para el análisis de los Datos.
- Transferir tecnologías al sector privado
- Manejar los problemas éticos, legales y sociales (ELSI) que pudieran acarrear la realización del proyecto Genoma Humano.

**Bases de Datos Biológicas**

Para suplir la necesidad de almacenar las secuencias y anotaciones generados por el proyecto y poder realizar análisis posteriores sobre estos datos, se crearon bases de datos tales como: el GenBank[3], la Base de Datos de Secuencias de Nucleótidos del Laboratorio Europeo de Biología Molecular[4] y el DNA Databank de Japón[5], etc.
Una base de datos biológica en una gran cantidad de datos persistentes organizados, usualmente asociados con software diseñado para actualizar y consultar, además componentes para la recuperación de datos almacenados en el sistema. Una simple base de datos podría ser un solo archivo que contenga muchos registros, cada uno de los cuales incluye el mismo sistema de información. Por ejemplo, un expediente asociado a una base de datos de la secuencia de nucleótidos contiene información típica tal como nombre del contacto, la secuencia de entrada con una descripción del tipo de molécula, el nombre científico del organismo de la fuente de el cual fue aislado, y a menudo, las citaciones de la literatura asociadas a la secuencia [Just07].

Los beneficios de tener los datos almacenados en una base de datos [Just07]:

- Fácil acceso a la información.
- Un método de extraer solo la información necesaria para responder una pregunta biológica especifica.

Muchas de las bases de datos Biológicas están enlazadas para permitirle al usuario acceder a la información de una manera más integrada, por ejemplo la Base de datos de Proteínas está ligada a una base de datos Taxonómica, esto le permite a los investigadores encontrar información taxonómica de las especies de las cuales es derivada la misma proteína [Just07].

[3] Base de datos de secuencias genéticas publica, http://www.ncbi.nlm.nih.gov/Genbank
[4] Institución Europea de investigación en biología molecular, http://www.embl.org/aboutus/news/press/2007/08jan07
[5] Base de datos internacional de secuencias de nucleótidos, http://www.ddbj.nig.ac.jp/

El paso de la era genómica a la post genómica trae consigo la construcción de bases de datos con información biológica, nace así una comunidad de información biológica. Las bases de datos biologías que en los últimos años han presentado exponencial tienen como objetivo recopilar y permitir el libre acceso a la comunidad científica; Inicialmente el desarrollo de estas no fue orientado para consultas vía Web, pero la comunidad científica no tardo en aprovechar los beneficios que esta proporciona, HTML, JAVA, JAVASCRIPT, CGI, PERL y muchas otras tantas tecnologías las cuales facilitaron y se convirtieron en herramientas de uso común, en la comunidad de biólogos.

Ante los grandes volúmenes de datos la necesidad primaria es clara, se debe, almacenar, presentar, comparar y analizar. La introducción de técnicas para secuenciamiento de proteínas en los años 70`s, impulsa el incremento masivo de secuencias, un grupo de investigación en la Universidad de Georgetown, se dio a la tarea de recolectar las secuencias y ponerlas a disposición de la comunidad en general en un libro que titularon ¨Atlas of protein sequence and structure¨ (Dayhoff et al 1965). En 1973 se dispuso por vez primera de un sistema para consulta en medio magnético, y no fue sino hasta el año de 1981 cuando se tuvo un sistema para consulta en línea. En la actualidad acceder a los bancos de datos que contienen información relativa tanto a DNA como a proteínas es sumamente sencillo.

De los medios magnéticos de almacenamiento de distribución manual se paso al HTML y después se salto a la implementación de motores de búsqueda, a la integración de sistemas de búsqueda y consulta con herramientas analíticas y comparativas, haciendo que estos desarrollos crecieran exponencialmente lo cual hizo necesario crear catálogos indexados de recursos para tener una guía de acceso a estas tecnologías.

Con el advenimiento de las técnicas de DNA recombinante en los años 70`s un grupo de científicos en el laboratorio de los Alamos empezó a archivar secuencias de ácidos nucleídos. En 1982 en MALIGNS adscrito a NIH se creo el GenBank y se hizo accesible por todos los medios de ese momento. Hacia mediados de los ochenta se contempló la necesidad de mapear el genoma humano; es entonces cuando el Congreso de los Estados Unidos, en respuesta a la necesidad tanto de almacenamiento como de intercambio de información, establece el NCBI como una división de la librería nacional de medicina (NLM). Después de haber tenido publicaciones escritas, y electrónicas en diferentes medios el GenBank está disponible para consulta por parte de cualquier persona a través de la red Internet.

Sin duda alguna en lo relativo a grandes bancos de datos en biología molecular ha experimentado un gran avance en los últimos años, se ha mantenido actual frente a los cambios frecuentes en cuanto a herramientas informáticas se refiere. Sin embargo existe redundancia en las bases de datos existentes, secuencias de proteínas y familias de genes o versiones de genes homólogos encontrados en diferentes organismos. Se tienen casos incluso de secuencias idénticas bajo distintos números o claves de acceso (identificadores en los bancos de datos), las variaciones se dan en cuanto al tejido estudiado, o el organismo del cual provienen las secuencias. El uso de bases de datos que presentan redundancia nos deja posibles fuentes de error. Si el conjunto de datos contiene información acerca de secuencias de ácidos nucleídos o de aminoácidos altamente relacionados el análisis estadístico de esas secuencias o contra esas secuencias va a presentar un sesgo importante hacia la clase de datos blanco. Se hace entonces necesario el evitar usar secuencias muy cercanas, muy relacionadas, hay sobre este punto que decir algo importante, la definición de un alto grado de relación se hace sobre el problema mismo, no existen consideraciones generales a este respecto.

El desarrollo en áreas específicas de la bioinformática es de difícil seguimiento, los recursos ofrecidos son a menudo olvidados y no son mantenidos, debido a que no están a cargo de grandes organizaciones, ni de importantes centros de investigación, lo que hace que los enlaces no sean revisados día a día como seria lo ideal.

El GenBank es una colección pública de secuencias tanto de proteínas como de ácidos nucleídos con soporte bibliográfico (referencias tomadas de la literatura reportada) y notación biológica (especie y origen). La base de datos del GenBank crece de una manera exponencial, este crecimiento es debido a la forma misma en que la base se actualiza. Son los mismos autores quienes se encargan de mantener la base al día, pero además de remisiones de autores, el GenBank se nutre también de las otras bases de datos existentes actualizando interactivamente sus ficheros. Las secuencias son procesadas una vez remitidas, y desde ese momento pueden ser localizadas usando una herramienta de búsqueda basada en una clave taxonómica desarrollada por el NCBI en colaboración con el EMBL y el DDBJ.

Con el objeto de establecer un identificador único para cada entrada en el GenBank el NCBI asigna a cada secuencia un término llamado **gi**. Un nuevo identificador **gi** es asignado a cada secuencia después de que esta ha sido actualizada de alguna manera, esta llave única aparece en el campo ACCESSION de la entrada, justo antes del número de entrada (ACCESSION #). El número de entrada a diferencia del identificador **gi** no varía cada vez que la entrada es modificada, se mantiene invariable aún cuando las anotaciones correspondientes a las secuencias cambian.

**Campos que componen una entrada del GenBank:**

Ver Anexo M, donde se encuentra un ejemplo de un registro del GenBank.

**Locus:** contiene el identificador único de la secuencia en la base de datos, el número de bases y la fecha de entrada de la secuencia

**ACCESSION:** se encuentra el numero gi, después de los dos puntos se encuentra el número aleatorio asignado a la secuencia, el ACCESSION #.

**DEFINITION:** da una descripción corta de la secuencia, incluye el nombre del organismo de origen.

**KEYWORDS:** Lista términos o palabras que facilitan el indexamiento de la secuencia en las posibles búsquedas sistemáticas.

**SOURCE:** se lista el origen biológico de la secuencia.

**REFERENCE:** cubre AUTHORS, TITLE, JOURNAL, y MEDLINE (número de la referencia en el MEDLINE).

**BASE COUNT:** da la información acerca de la composición de la secuencia, el número de A, T, C, G.

El último que se encuentra en el registro, corresponde a la secuencia de un cromosoma en sí, la secuencia se encuentra estructurada en regiones codificantes y no codificantes, las regiones codificantes llamadas Gen, se expresan en unidades de información, llamadas EXONES e INTRONES. Cada región codificante debe iniciar en exón y terminar en exón, lo anterior puede verse gráficamente en la Figura 5.

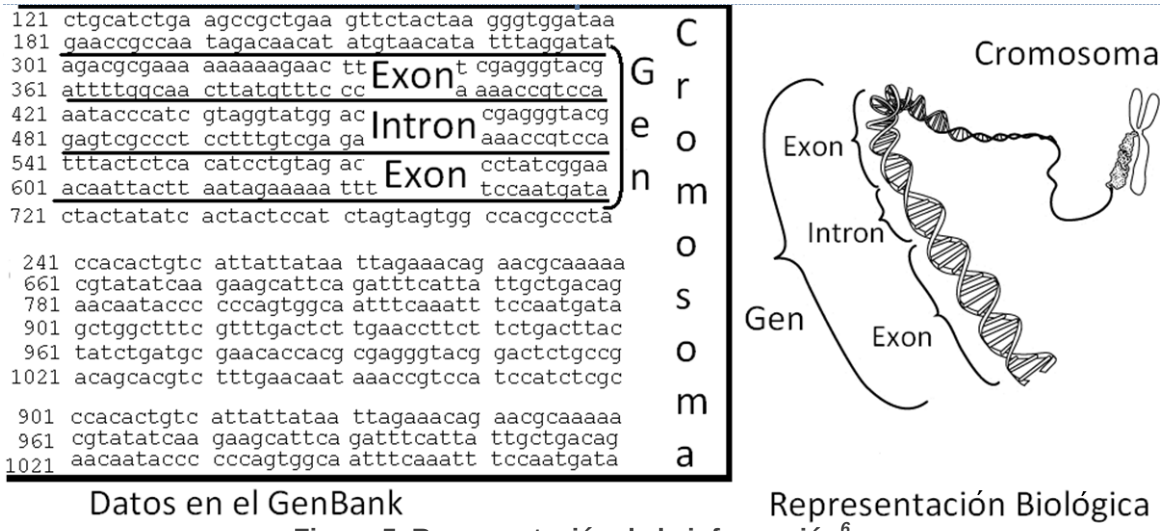



**Figura 5. Representación de la información**[6]

**Artefactos de las bases de datos biológicas [1], [15]**

Las bases de datos biológicas, presentan una tasa de crecimiento exponencial, debido al incremento de proyectos de secuenciación genómica y proteómica, lo cual dificulta la labor de revisión de los datos, recordando que como cualquier base de datos, las bases de datos biológicas o repositorios públicos que almacenan información genómica y proteómica, esta sujeto a problemas en la calidad de los datos contenidos, tales como: datos correctos, uniformes, completos, no redundantes, entre otras cuestiones a tener en cuenta. Existen dos análisis realizados por investigadores bioinformáticos, en los cuales se han clasificado algunos de los artefactos, como son llamados los errores o particularidades encontrados en los datos almacenados en los repositorios públicos, los cuales dificultan la labor de encontrar información por parte de los investigadores. El objetivo de estos estudios es dar a conocer a la comunidad científica que trabaja con estos repositorios, los posibles artefactos que se pueden encontrar y evaluar que tan críticos pueden llegar a ser [1], [15].

Para iniciar la presentación de la problemática, se exhibirá la estructura de los registros del GENBANK, para posteriormente clasificar los artefactos de acuerdo a dicha estructura, la cual se encuentran de esta manera:

**Cabecera:** Información general del registro.

**Característica:** Descripción de lo estructural, funcional y otras propiedades físico – químicas de la secuencia y las regiones de interés.

**Secuencia:** secuencia de nucleótidos o proteínas.

[6] Basada en la Figura encontrada en http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

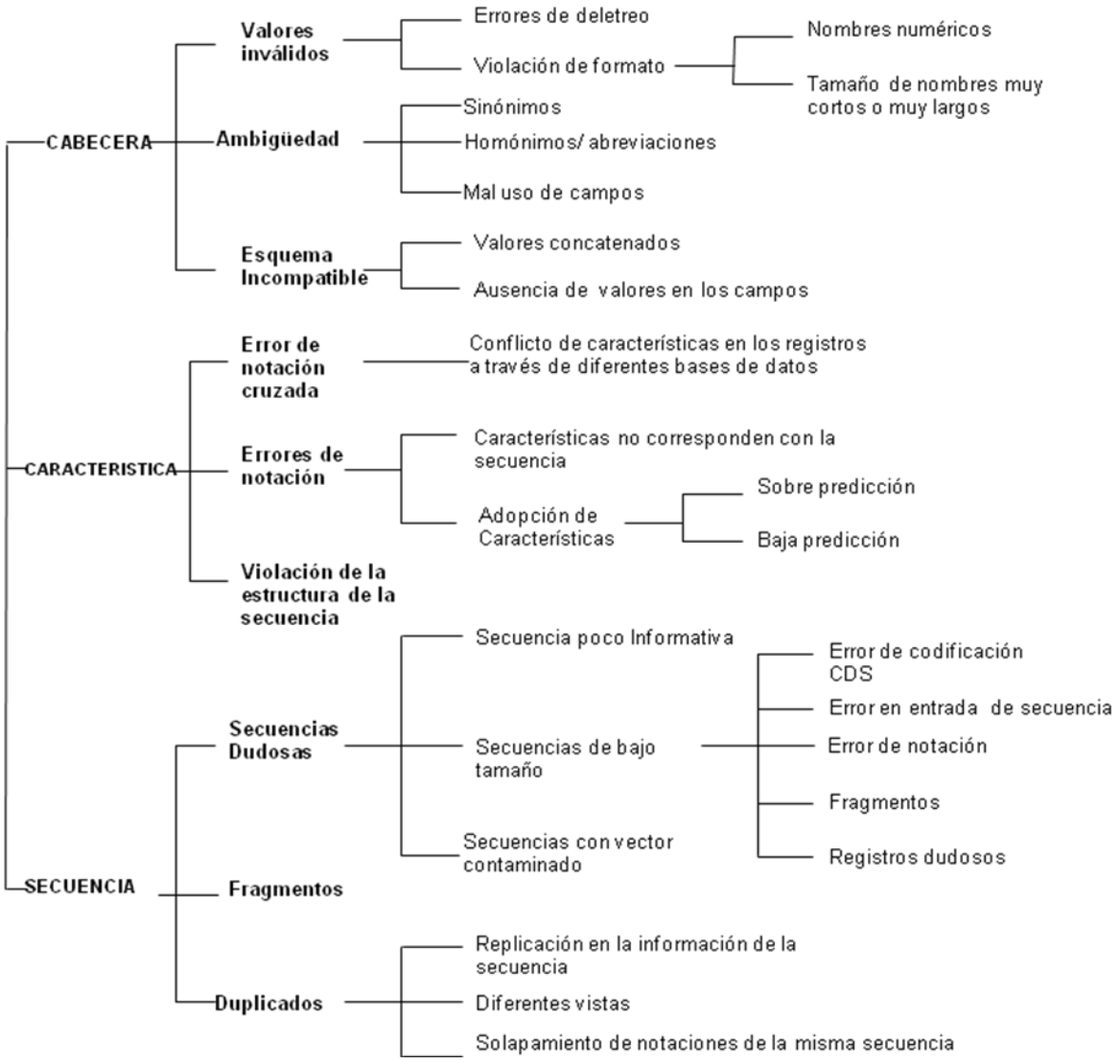



**Figura 6. Esquema de clasificación de los artefactos**[7]

Artefactos en la cabecera

1. Valores Inválidos

   a. Errores de deletreo: Este tipo de error no es crítico, es un error tipográfico, introducidos por los autores en el momento de cargar los datos, por ejemplo "inmmunoglobulin", es muchas veces escrita como "inmmunoglobin" o "inmmunogloblin"

   b. Violación de formato: se mencionó anteriormente que la cabecera del registro proporcionaba la información general, este error no es un error crítico, lo que se puede

[7] Traducción de la figura contenida en la presentación "una clasificación de los artefactos biológicos" [15].

encontrar es campos muy grandes o por el contrario, muy pequeños, los cuales no contribuyen al entendimiento de los registros.

2. Ambigüedad

   a. Sinónimos: Diferentes nombres para nombrar a una misma secuencia, este error hace que se carezca de identificadores únicos por cada una de las secuencias registradas.

   b. Homónimos/ Abreviaturas: Diferentes secuencias pueden tener el mismo nombre, uno de los factores por lo que esto ocurre es la existencia de secuencias que son comunes en diferentes organismos, a lo cual se denomina homónimo, pero también se encuentra otro factor y es el de las abreviaturas en el cual la abreviatura de una secuencia puede ser igual para muchas otras, por ejemplo una secuencia puede ser denominada como GK, puede referirse a GLYCEROL KINASES, GLUTAMITE KINASES o GUANYLATE KINASES, entre otras.

La presencia de homónimos y abreviaturas causa problemas en la identificación de secuencias y la búsqueda de palabras claves, igual que como ocurre cuando se tienen sinónimos.

   c. Mal uso de los campos: este error tampoco se considera crítico, se presenta cuando los investigadores que proporcionan las secuencias incluyen valores en los campos que no corresponden propiamente a dicho campo.

3. Esquema incompatible

   a. Valores concatenados: Debido a la incompatibilidad de campo dentro del esquema, si el campo tiene una granularidad más fina que la presentada por el esquema el dato es concatenado con otro campo en el momento de ser transformado.

   b. Ausencia de valores en los campos: es una falla en el esquema de mapeo, los campos fuentes no son tomados en cuenta en el esquema de transformación de los datos y estos pueden ser mapeados incorrectamente y enviados a campos que no corresponden.

Artefactos en las características.

1. Error de notación cruzada.

   a. Conflicto de características en los registros a través de diferentes bases de datos: Existen múltiples registros en las diferentes bases de datos públicas pertenecientes a una misma secuencia de proteína o nucleótido, los cuales contienen inconsistencias o contradicciones en las notaciones de las características, tales como: errores en la entrada de los datos, ausencia de notaciones en las funciones de la secuencia, diferentes interpretaciones por parte de los investigadores e inferencia de las características o transferencia notaciones basadas en las coincidencias con otras secuencias las cuales no propiamente tienen un alto grado de compatibilidad.

2. Errores de notación.

   a. Adopción de Características: con la notación funcional se busca encontrar el mayor nivel de concordancia entre una secuencia y otra para de esta manera extrapolar las características de la una hacia la otra, pero incluso a veces el nivel más alto de concordancia entre las secuencias no asegura que las secuencias sean bastante símiles y que por lo tanto compartan funciones similares.

3. Violación de la estructura de la secuencia: se encuentran entidades con características ilógicas las cuales no corresponden a las restricciones lógicas de la estructura de un gen.

Artefactos en la secuencia.

1. Secuencias dudosas

   a. Secuencia poco informativa: Existen secuencias que contienen abundantes residuos desconocidos “X” o nucleótidos desconocidos “N” lo cual hace que estas secuencias sean poco informativas.

   b. Secuencias muy pequeñas: son secuencias que tienen un tamaño menor respecto a las otras secuencias almacenas y que puede presentarse por entradas parciales, incompletas o con mala notación.

   c. Secuencias de vector contaminado: una secuencia contaminada contiene una o más segmentos de secuencias ajenos, introducidos en varios pasos de procesos de clonación o durante la recombinación.

2. Fragmentos: Fragmentos de secuencias en diferentes registros: Existe gran redundancia en las secuencias debido a registros que contienen fragmentos o superposición de secuencias con secuencias más completas en otros registros.

3. Duplicados: las secuencias duplicadas pueden ser generadas por una misma secuencia que ha sido subida varias veces por el mismo o por diferentes grupos de notaciones dentro de la base de datos de secuencias o por actualización cruzada.

Ver Anexo N donde se encuentra la explicación sináptica de la problemática.

## *2.5. Bioinformática*

Los análisis realizados sobre las bases de datos Biológicas buscan brindar un significado a todos los datos almacenados desde el punto médico o biológico, dichos análisis se han dividido en distintas áreas de interés tales como la predicción de la estructura de las proteínas, los cambios en la estabilidad inducidos por mutaciones, el estudio de las interacciones de las proteínas con moléculas más pequeñas, la evolución natural basándose en los cambios en la información genética, en lugar de utilizar los cambios morfológicos, mucho más difíciles de medir, el análisis de la expresión y de la regulación de los genes y de las proteínas, para entender los mecanismos que regulan los procesos celulares. Desde la comparación de esas dinámicas en su evolución normal y bajo condiciones patológicas se pueden estudiar los cambios que se verifican en las células cancerígenas, y vislumbrar nuevas líneas de combate contra la enfermedad [Just07].

La complejidad de estos análisis requirió de la interacción de varias disciplinas, surgiendo de esta manera la bioinformática un nuevo campo de investigación, en la cual se fusionan la biología, la informática y la tecnología en una sola. Dentro de la bioinformática, existen tres subdisciplinas [Just07]:

1. El desarrollo de nuevos algoritmos y estadísticas para establecer relaciones entre miembros de grandes grupos de datos.

2. El análisis y la interpretación de varios tipos de datos incluyendo secuencias de nucleótidos y aminoácidos, dominios proteómicos y estructuras de proteínas.

3. El desarrollo y la implementación de herramientas que permitan acceso y manejo eficientes de diferentes tipos de información.

## *2.6. Investigación "Análisis Multifractal del Genoma Humano"*

**Análisis Multifractal**
Conjunto formado por una jerarquía de subconjuntos (variedades), cada uno de ellos de carácter fractal (variedades fractales), considerando que un fractal es definido como objeto semi geométrico cuya estructura básica, fragmentada o irregular, se repite a diferentes escalas. Por lo general, se considera que el multifractal es una variedad topológica, generalmente métrica.[25], [28].

Cualquier reunión de conjuntos fractales por sí sola no puede considerarse un multifractal; para ello es necesario que estén coordinados de cierta manera. Como norma general, se exige que el espectro de singularidad sea una curva convexa. El objetivo es garantizar que el conjunto, y cada una de sus partes sea invariante bajo transformaciones de cambio de escala [25].

El interés por los multifractales nace del estudio de las propiedades de los fluidos turbulentos con alto Número de Reynolds. Éstos son los llamados fluidos en régimen de Turbulencia Completamente Desarrollada. En esos caso, la elevada turbulencia del fluido hace que su estructura abandone todas las simetrías afines propias del régimen laminar. A cualquier escala a la cual se analice el fluido se encontrará que el los grados de libertad no resueltos no son pequeñas variaciones o fluctuaciones sobre el régimen de mayor escala, sino que tienen amplitudes considerables, hasta el punto de que la dirección de la corriente está complemente indeterminada aunque se conozca la dirección en la gran escala [25].

**Planteamiento del Problema de la Investigación "Análisis Multifractal del Genoma Humano"**
"La complejidad del genoma humano quedó de manifiesto tras la secuenciación y anotación de sus 3 Gpb lo cual reveló una intrincada irregularidad topológica en la distribución de secuencias codificantes y no codificantes a lo largo de los cromosomas y cuyo significado biológico aun dista de entenderse por completo. A fin de contribuir a explicar dicha irregularidad se propone aplicar el análisis multifractal con el objeto de: 1) Cuantificar la variación en la información genética y 2) Generar una clasificación del genoma humano con significado biológico y potencial uso tecnológico. Para alcanzar estos objetivos se ha diseñado una estrategia que combina genómica, bioinformática, matemáticas, ciencia no-lineal y análisis filogenético a fin de responder a las necesidades propuestas y con el valor agregado de proveer a los investigadores en el campo de herramientas de biotecnología de la información suplementarias para el análisis de la organización de los genomas eucariotes y de la salud humana" [25].

**Líneas de investigación:** 1) Desarrollo de abordajes genómicos, microbianos y humanos. 2) Bioinformática. 3) Desarrollo de software. 4) Minería de Datos 5) Aprendizaje de Máquina [25].

**Objetivos:** Cuantificar la variación del genoma humano y generar una clasificación con sentido biológico que contribuya a explicar la irregularidad cifrada. Producir una solución tecnológica: la clasificación en sí misma [25].

**Justificación:** Contribuir al análisis de la estructura y función del gen y secuencias intergénicas a fin de ganar un mejor entendimiento de la organización de genes y genomas. Asignar nuevos genes potencialmente útiles [25].

# CAPITULO III. DESCRIPCIÓN DEL PROCESO DEL DESARROLLO DEL DSS

## *3.1. Metodología para el Desarrollo DW*

La realización de este proyecto esta guiada por la metodología propuesta por Ralph Kimball. Esta metodología es una de las más conocidas, empleadas y con bases conceptuales bien definidas [19].

El desarrollo del DW, fue guiado por ciclo de vida dimensional propuesto por Ralph Kimball [19], en el ciclo de vida dimensional se pueden apreciar las diferentes etapas que deben ser llevados a cabo en la construcción de un DW. El marco presentado por Ralph Kimball ilustra las diferentes etapas por las que debe pasar todo proceso para la construcción de un DW. Ver Figura 7.

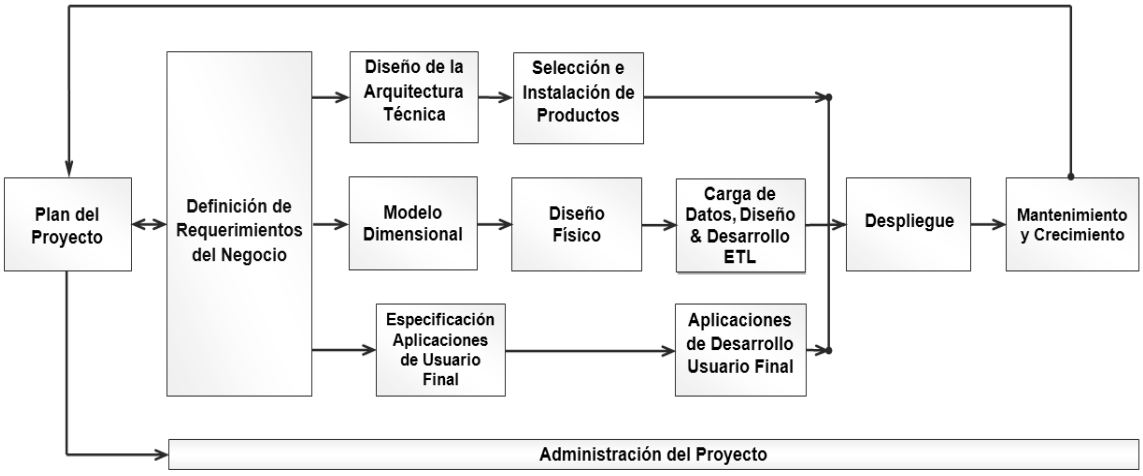



**Figura 7. Ciclo de Vida Dimensional**[8]

La Figura 7 muestra una vista general del mapa de ruta de un proyecto [19] de DW, además presenta la secuencialidad de tareas de alto nivel requeridas para el diseño, desarrollo e implementación del DW.

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

Se presentara a continuación cada una de las etapas haciendo primero una breve descripción de su objetivo general y luego mostrando los respectivos resultados obtenidos en cada una de ellas.

### 3.1.1. Planeación Del Proyecto

El Objetivo de esta etapa fue identificar el escenario del proyecto para saber de dónde surgió la necesidad del DW y construir el alcance del proyecto, se incluye el impacto y la evaluación de

[8] Traducción de la figura The Business Dimensional Lifecycle diagram. [19].

la factibilidad. Esta etapa es dependiente de los requerimientos porque estos determinan el alcance del proyecto y los recursos que se requieren. Para esta tarea de alto nivel fueron necesarias las especificaciones de las actividades y las tareas necesarias, para posteriormente definir el alcance y la justificación del proyecto [19].

**Definición del Proyecto**

Para el desarrollo del sistema para la toma de decisiones para la investigación análisis multifractal del genoma humano, se identificó un alto interés por parte del grupo BIMAC, quienes son los responsables de la investigación y a los cuales les interesa que la información resultante de la investigación sea presentada de una manera entendible y fácil de manipular por los usuarios a los que posiblemente les interese el tema. Se debe tener en cuenta que un escenario en el cual son varias personas las que manifiestan el interés por el sistema, es un escenario algo complejo debido a la necesidad de priorizar los requisitos antes de proceder, siendo a veces una tarea complicada en la cual es difícil complacer a todos los interesados.

Se debe destacar que la realización de este proyecto involucra un amplio estudio del entorno en el cual está enfocada la aplicación, debido a que no es el entorno convencional hacia el cual están enfocados este tipo de sistemas, los cuales nacen como respuesta a necesidades empresariales y que para que sea aplicado en un entorno biológico, se deben realizar ajustes conceptuales, terminológicos y técnicos.

En este proyecto se construyo una solución de un DSS basado en tecnologías de DW y una herramienta OLAP, las cuales permiten a los usuarios hacer análisis dimensionales de los datos almacenados y de esta manera contrastar estos análisis con los resultados obtenidos en su investigación, de igual manera la solución permitirá que los investigadores publiquen su investigación y permitan que otros investigadores realicen sus propias consultas y saquen sus propias conclusiones.

**Evaluación del ambiente organizacional para la ejecución del proyecto**

En las tareas propuestas por Ralph Kimball, en su libro del ciclo de vida dimensional, se encuentran algunos factores que permiten evaluar la disposición que una organización tiene para el desarrollo de un DW y de esta manera establecer las bases para asegurar el éxito de un proyecto DW; sin embargo, estos factores no aplican ampliamente en el área en la que este proyecto fue desarrollado, no obstante se analizaron estos factores y se adaptaron al área de aplicación, como se mostrara a continuación.

Para llevar a cabo esta actividad se hizo uso del artefacto que propone Kimball en su literatura, denominado test de Litmus, el cual puede encontrarse en el Anexo E, en el documento de anexos de este proyecto, a continuación se presentaran las conclusiones a las que se llegaron después de analizar los resultados obtenidos en el test.

**Influencia del sponsor en la organización**: El proyecto se desarrollara como una alternativa para la presentación de los resultados de la investigación análisis multifractal del genoma humano, que está siendo desarrollada por los grupos BIMAC y GTI de la Universidad del Cauca. Dentro del equipo de investigadores se encuentra el Ingeniero Ember Ubeimar Martínez Flor, quien se denominará el sponsor, considerando que es la persona interesada dentro del grupo de investigadores en explorar este tipo de soluciones.

Se considera que el sponsor es de gran influencia dentro del grupo de investigadores porque hace parte fundamental del equipo, maneja la información y tiene acceso a los recursos del grupo. Además se puede decir que es una persona a la que se tiene acceso frecuentemente, porque trabaja en el mismo espacio en el cual está ubicado el equipo de trabajo.

**Motivación del proyecto:** La investigación análisis multifractal realizada por el grupo BIMAC de la universidad del Cauca ha generado en los últimos años un gran volumen de información relativa al genoma humano, la investigación busca encontrar patrones dentro del genoma humano y esperan que a la culminación de la investigación la información pueda ser manipulada por usuarios a quienes les interese dicha información y que puedan realizar diferentes análisis sobre los datos presentados.
La bodega de datos es una alternativa de solución, con la cual se pretende generar un sistema en el que se encuentren los datos integrados

**Relación entre equipos de la organización (Área Biológica – Área tecnologías de la información):** El grupo de investigadores, es un grupo muy bien conformado en el cual cada equipo que lo conforma, tienen definido el rol que desempeña cada uno el grupo. Además la información es su baluarte más preciado por lo cual es fundamental para ellos el desarrollo de aplicaciones que contribuyan con sus análisis.

**Presencia de cultura analítica:** El grupo BIMAC, es un grupo dedicado a realizar estudios relacionados con el entorno biológico, el análisis es una labor que se desempeña a diario.

**Factibilidad:** Debido a que el proyecto se encuentra enmarcado dentro de un trabajo de grado de la universidad del Cauca, se cuenta con los recursos que proporciona la Universidad y el grupo de investigaciones, además se cuenta con disponibilidad de los datos, los cuales son proporcionados por el grupo de investigación BIMAC.

En conclusión se puede observar que se cuenta con una alta disponibilidad para llevar a cabo el proyecto, teniendo en cuenta que los factores evaluados en el test que dieron como altamente disponibles equivalen a un porcentaje de más del 80%.

**Personas relacionadas con el proyecto.**
Grupo BIMAC: Los integrantes del grupo BIMAC, contribuirán en el proyecto en cuanto al entendimiento de la información proporcionada para el desarrollo de la bodega de datos, además de proporcionar el espacio y los recursos necesarios para llevar a cabo las actividades programadas en el proyecto.

Sponsor: Ingeniero Ember Ubeimar Martinez quien también pertenece al grupo BIMAC, es la conexión entre el área de aplicación del proyecto y el equipo de desarrollo.

Gerente del proyecto:Magister Martha Eliana es la directora del proyecto, es la persona experta en el tema de las bodegas de datos encargada de asesorar y supervisar todas las actividades programadas.

Líder del proyecto por parte del grupo BIMAC: Patricia Vélez, directora del grupo BIMAC.

Analistas del sistema: Alba Viviana Camayo Otero y Adrian Fernando Martinez, Estudiantes de decimo semestre de ingeniería de Sistemas.

Modeladores de datos: Alba Viviana Camayo Otero y Adrian Fernando Martinez, Estudiantes de decimo semestre de ingeniería de Sistemas.

Administradores de la bodega de datos: Alba Viviana Camayo Otero y Adrian Fernando Martinez, Estudiantes de X semestre de ingeniería de Sistemas.

Diseñadores del sistema temporal de almacenamiento de los datos: Alba Viviana Camayo Otero y Adrian Fernando Martinez, Estudiantes de decimo semestre de ingeniería de Sistemas.

Desarrolladores de las aplicaciones de usuario final: Alba Viviana Camayo Otero y Adrian Fernando Martinez, Estudiantes de X semestre de ingeniería de Sistemas.
**Desarrollo del plan del proyecto.**
El desarrollo de este proyecto se dividió en dos fases las cuales se explicarán a continuación.

- **FASE 1. Metodología de Desarrollo.**
  El desarrollo total del proyecto involucro la utilización de una metodología, para la construcción del DW y la herramienta OLAP.

- **FASE 2. Generar recomendaciones para el diseño y construcción de un DW en el campo de la bioinformática.**
  Esta etapa comprendió la recopilación de casos especiales de diseño que se pudiesen presentar durante la creación del DW en el campo de la Bioinformática, para generar recomendaciones de diseño y construcción de un DW en esta área de aplicación.

**Plan de Comunicaciones**
El plan de comunicaciones buscaba establecer comunicaciones robustas entre los diferentes actores involucrados en el proceso de desarrollo de una bodega de datos.

El plan de comunicaciones contenía la descripción en forma general de los tipos de mensajes, el contenido, el formato y la frecuencia de dichos mensajes utilizados para poder comunicarse con los diferentes actores del proyecto.

Se establecieron diferentes estrategias de comunicación de acuerdo a las personas con las que se establecería comunicación, a continuación se presentaran los actores implicados y las estrategias de comunicación a ser utilizadas.

**Equipo del proyecto:** El equipo del proyecto se reúne semanalmente con la directora del proyecto para revisar el estado del proyecto, revisar si las actividades han sido llevadas a satisfacción por los responsables y se asignan nuevas actividades.

**Grupo de investigadores (BIMAC):** se planearon reuniones con el equipo de investigadores que para presentar avances y retroalimentar el proceso.

### 3.1.2. Definición De Requerimientos Del Negocio

Para realizar la captura de los requerimientos se tuvo en cuenta las recomendaciones dadas en el libro de Ralph Kimball y el formato del QFD para la toma de requerimientos.

Ralph Kimball expone en su libro que los usuarios y sus requerimientos causan un gran impacto en las decisiones que han de ser tomadas a largo del proyecto, los requerimientos son el centro del universo de las bodegas de datos, la captura de requerimientos permitió determinar cuáles eran los datos que debían encontrarse disponibles en la bodega de datos, como debían estar organizados y que tan a menudo esos datos debían ser actualizados; también fue importante determinar quienes accederían a la bodega de datos, como se deseaba que los datos fuesen presentados y finalmente se planeó el desarrollo, mantenimiento y crecimiento de la bodega teniendo en cuenta las opiniones del cliente. Se inicio formulando respuestas a preguntas relacionadas con todo el ciclo de vida y que permitiera saber que tanto conocimiento se tenía acerca del área en que se planteaba la solución [19].

Es necesario destacar que la aplicación se desarrollo en un ambiente nuevo y con algunas particularidades a tener en cuenta; por tal razón, inicialmente, se realizaron algunos prototipos con los cuales se buscaba explorar y entender un poco más el ambiente de desarrollo del proyecto para posteriormente retomar la metodología que se propuso llevar en el desarrollo del proyecto y

por medio de la captura de requisitos retroalimentar el proceso y confirmar los supuestos con los cuales se construyeron dichos prototipos.

Para realizar esta actividad y retomando lo planteado por Ralph Kimball, las respuestas a las preguntas planteadas parten de un conocimiento previo en el cual se ha explorado el ambiente por medio de los prototipos creados.

Para guiar el proceso de captura de los requerimientos se hizo uso de una plantilla guiada por el concepto del QFD para la captura de requerimientos; el QFD[29] (Despliegue de la función de calidad), es definido por el DR. Yoji Akao, uno de sus desarrolladores; como un método para desarrollar un diseño de calidad dirigido a satisfacer al cliente al traducir sus demandas en metas de diseño y puntos importantes de aseguramiento de la calidad para usarse en toda la fase de producción.

Se hizo uso de la metodología del *Blitz QFD,* esta permitió alinear los recursos existentes con las verdaderas necesidades del cliente, es una herramienta muy práctica que no requiere de *software,* ni de herramientas específicas para ofrecer resultados, en el Anexo A, Anexo C y Anexo D. Se podrá encontrar una tabla en la cual se muestra la definición de las entrevistas, los siete pasos sugeridos por la metodología seguida para la captura de requerimientos y el resultado de la actividad.

A continuación se presentará el resultado de la actividad para la toma de requerimientos, mostrando todas las consultas, las cuales se dividen en dos grupos como se verá a continuación.

**Tabla 1.Definición de las consultas**

<table>
<tr><th></th><th>Filtros</th><th>Información visualizada después del filtrado</th><th>Resultados Gráficos de las consultas</th></tr>
<tr><td rowspan="4">Primer Escenario</td><td>Selecciona un organismo</td><td>Número de unidades de información existentes en un gen</td><td>Graficar la relación entre el número de genes y los rangos definidos</td></tr>
<tr><td>Selecciona un cromosoma</td><td>Promedio de unidades de información que se encuentran en cualquiera de los rangos de análisis.</td><td>Graficar la relación entre el promedio de unidades de información y los rangos definidos.</td></tr>
<tr><td>Determina un intervalo de análisis</td><td rowspan="2">Promedio de las longitudes de los genes que se encuentran en cualquiera de los rangos de análisis.</td><td>Graficar la relación entre el promedio de longitudes de los genes y los rangos definidos.</td></tr>
<tr><td>Determina el numero de partes en que será dividido el intervalo (se denominara rangos de análisis)</td><td>Graficar relaciones entre las variables de análisis</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td rowspan="2">Segundo Escenario</td><td>Selecciona un cromosoma de análisis</td><td>Longitud de cada unidad de información.</td><td rowspan="2"></td></tr>
<tr><td>El usuario determina el número de unidades de información que desea analizar</td><td>Orden de acuerdo a la longitud de las unidades de información.</td></tr>
</table>

Como resultado del análisis de los requerimientos se tuvo que para el desarrollo del proyecto se requería construir dos Data marts, el primero que se denomino análisis fractal y el segundo análisis de las unidades de información.

Para poder comprender el entorno en que se desarrollo la aplicación se realizo un análisis de la información manejada por el grupo BIMAC, como complemento a la captura de requisitos y en pro de tener claridad en el momento de construir la Matriz Bus (ver Tabla 2), para mayor detalle de la actividad ver Anexo B.

**Matriz de Data Marts**
La matriz de data marts muestra la relación entre los posibles data marts y las dimensiones. A continuación se muestra una breve descripción de cada uno de los data mart y las dimensiones.

**Tabla 2. Matriz Bus**

| Data Marts | Dimensiones | | | | |
|---|---|---|---|---|---|
| | Cromosoma | Gen | Estructura | Tipo | Band |
| Análisis fractal | X | X | X | | X |
| Análisis unidades de información | X | X | X | X | |

**Descripción del Data Mart Análisis Fractal**

El objetivo de este datamart consistía en escoger un intervalo, dividirlo en un número de partes finito y dentro de estas particiones clasificar el valor D de cada Estructura, si el valor de D estaba dentro de los límites superior e inferior de la partición. Y así poder consultar dentro de cada una de estas particiones las longitudes, número de unidades de información, simetría de los genes en la partición, también porcentajes por número genes y calcular el número de D por familia de genes.

**Descripción del Data Mart Análisis de Unidades de Información**

Este data mart tiene como objeto permitir al investigador seleccionar el número todas las estructuras que tuvieran el mismo número de unidades de información, poder visualizar a que cromosomas, genes ó a que familia de genes pertenecen esas estructuras y finalmente poder organizar en orden ascendente o descendente las longitudes de las unidades las de información (Exón o Intrón).

### 3.1.3. Modelado Dimensional

En esta etapa fueron definidos los modelos dimensionales guiados por los requerimientos obtenidos en la fase anterior. Para la realización del modelado dimensional se tuvo en cuenta que el objetivo de este es el análisis de la información, la cual se puede visualizar en los reportes generados por la investigación, por lo cual al realizar el modelado se debió tener en cuenta dichos reportes, los cuales fueron establecidos en la captura de requerimientos. (Ver **¡Error! No se encuentra el origen de la referencia.**), de igual manera se debe tener en cuenta la matriz bus construida, la cual representa los procesos claves de la investigación y la dimensionalidad de estos.

Definidos los procesos de la investigación, se desarrollaron de los modelos dimensionales de alto nivel para cada uno de los data mart, dichos modelos se construyeron siguiendo el orden que se presenta en la Figura 8, primero se selecciono un data mart; segundo fue definida la granularidad; tercero se seleccionaron las dimensiones; cuarto se seleccionaron las medidas y por último se realizo una validación y actualización de los modelos dando origen a los modelos finales [19].

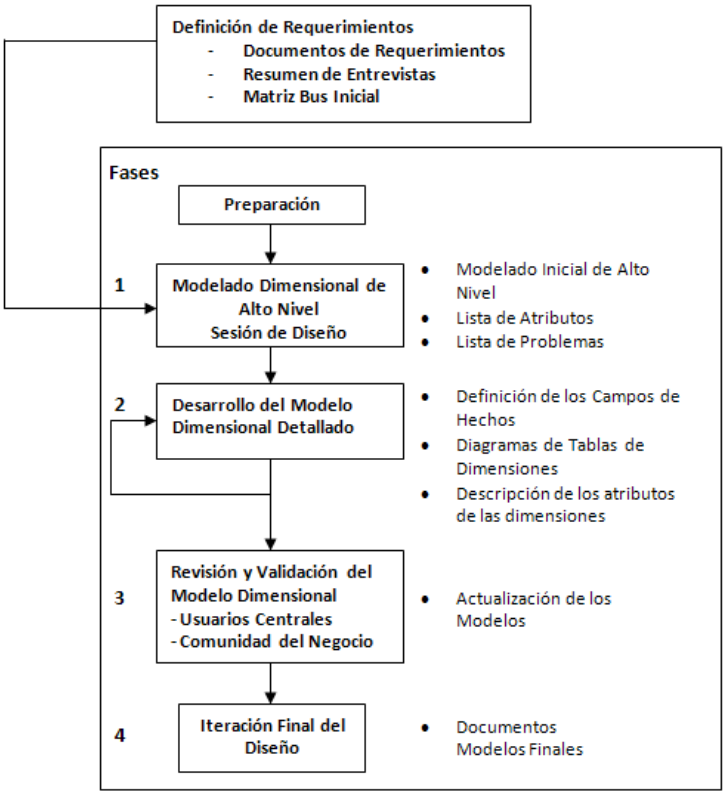



**Figura 8. Diagrama de flujo del proceso de modelado dimensional [14]**

El inicio del proceso de modelado es la creación de los modelados iniciales de alto nivel y una lista de atributos, que permitirán seguir con el segundo paso que es el desarrollo detallado del modelado dimensional, para posteriormente ser validados, la descripción de los data marts iniciales puede verse en el Anexo F.

**Modelado Inicial de Alto Nivel**

Para crear los modelos iníciales del DW para la investigación análisis multifractal del genoma humano se siguieron los cuatro pasos definidos para la construcción de los modelos [19]:

1. Seleccionar el Proceso de la investigación.
2. Definición de la granularidad: Definición del nivel de detalle o granularidad para el proceso de investigación, generalmente es una granularidad atómica que consiste de una fila en la tabla de hechos por una fila en el sistema fuente transaccional.
3. Selección de las Dimensiones: Selección de las dimensiones basándose en los objetos asociados a cada proceso de negocio de la matriz bus inicial, la selección de las dimensiones puede involucrar en ocasiones redefinir la granularidad.
4. Identificar las medidas: identificar las medidas de desempeño generados por el proceso de negocio. Los hechos usualmente se enlazan directamente a la declaración de la granularidad.

Para dar inicio a la creación de los modelos se debe revisar la matriz bus construida en la etapa de recolección de los requerimientos (Ver Tabla 2)

1. **Data Mart Análisis Fractal:**
   El objetivo de este data mart consistía en escoger un intervalo, dividirlo en un número de partes finito y dentro de estas particiones clasificar el valor D de cada Estructura, si el valor

de D estaba dentro de los límites superior e inferior de la partición. Y así poder consultar dentro de cada una de estas particiones las longitudes, número de unidades de información, simetría de los genes en la partición, también porcentajes por número genes y calcular el número de D por familia de genes.

2. **Granularidad:**
   Un aspecto teórico importante a resaltar es el hecho que un gen pudo haber sido secuenciado por múltiples métodos, lo cual da como resultado, tener el mismo gen expresado de formas posiblemente distintas, esto será denominado como estructura de un gen.

   Una vez realizada la aclaración teórica, se tiene que la granularidad definida para este data mart, es por estructura, que es la granularidad más baja que se puede tener en la tabla de hechos, la cual según la clasificación de Ralph Kimball, asigna a la tabla de hechos una granularidad transaccional.

3. **Dimensiones**

- **Dimensión Cromosoma** (Compartida por los dos Datamarts)
- **Dimensión Gen** (Compartida por los dos Datamarts)
- **Dimensión Estructura** (Compartida por los dos Datamarts)
- **Dimensión Band**

4. **Medidas:**
   Se requiere para esta tabla de hechos clasificar D, R2, #UI, simetría, size, %gen# UI`s, %gen_size, %gen_ simetría, #D's X Gen y #D´s X familia, de acuerdo a los intervalos y rangos especificados.

El modelo de alto nivel construido es mostrado en la Figura 99.

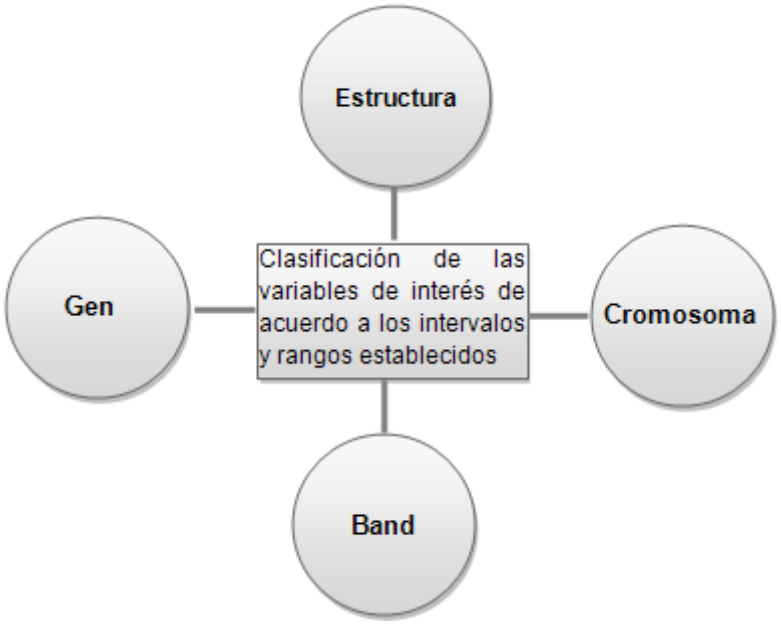



**Figura 9. Modelo Inicial de Alto Nivel para el análisis fractal**

1. **Data Mart Análisis Unidades de Información:**
   Este data mart tiene como objeto permitir al investigador seleccionar el número todas las estructuras que tuvieran el mismo número de unidades de información, poder visualizar a que cromosomas, genes ó a que familia de genes pertenecen esas estructuras y finalmente poder organizar en orden ascendente o descendente las longitudes de las unidades las de información (Exón o Intrón).

2. **Granularidad:**
   La granularidad de la tabla de hechos análisis de las unidades de información, está dada por unidad de información, es la granularidad más baja que se puede obtener en la tabla, la cual según la clasificación de Ralph Kimball, asigna a la tabla de hechos una granularidad transaccional.

3. **Dimensiones**

- **Dimensión Cromosoma** (Compartida por los dos Datamarts)
- **Dimensión Gen** (Compartida por los dos Datamarts)
- **Dimensión Estructura** (Compartida por los dos Datamarts)
- **Dimensión Tipo**

4. **Medidas:**
   Análisis de las unidades de información de acuerdo a las medidas, orden_estruct, orden_tipo, descrip_orden_tipo, pos_inicial, pos-final, longitud, Nombre del Campo, orden_estruct, orden_tipo, descrip_orden_tipo.

El modelo de alto nivel construido es mostrado en la Figura 100

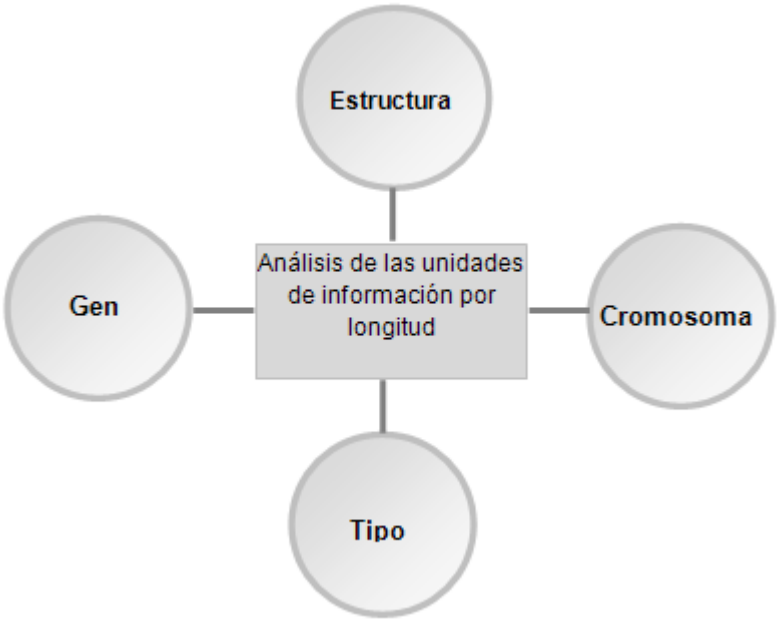



**Figura 10. Modelo inicial de alto nivel para el análisis de las unidades de información**

**Desarrollo Detallado del Modelado Dimensional**

**Descripción del Data Mart Análisis de Unidades de Información**

El objetivo de este data mart consistía en permitir al investigador seleccionar el número todas las estructuras que tuvieran el mismo número de unidades de información, poder visualizar a que cromosomas, genes ó a que familia de genes pertenecen esas estructuras y finalmente poder organizar en orden ascendente o descendente las longitudes de las unidades las de información (Exón o Intrón).

**Descripción de las Dimensiones**

- **Dimensión Cromosoma:** Contienen los cromosomas de los organismos, un identificador único para cada cromosoma, una descripción y un identificador del organismo al que pertenecen.
- **Dimensión Gen:** Contiene los genes, con un identificador, la banda donde se ubica el gen, la familia, subfamilia y descripción.
- **Dimensión Estructura:** Contiene las estructuras de cada gen, un identificador de la estructura y su función molecular.
- **Dimensión Tipo:** Contiene las unidades de información de cada estructura, un identificador único para cada unidad, junto con una descripción para saber si son Exones o Intrones.
- **Dimensión Band:** Contiene los posibles intervalos y particiones donde se puede clasificar el valor D.

**Definición detallada de las Dimensiones:** La definición de las dimensiones se realizó de acuerdo a los requerimientos obtenidos y a la granularidad definida. A continuación se presentan cada una de las dimensiones que pertenecen a este data mart, se debe recordar que los data marts comparten algunas dimensiones, lo cual también será especificado.

**Dimensión Gen – DimGen (Dimensión compartida):** Contiene los genes de cada cromosoma, con un identificador, la banda, la familia, subfamilia y descripción.

**Tabla 3. Dimensión Gen**

| Nombre del Campo | Tipo – Tamaño | Descripción | Llave primaria |
|---|---|---|---|
| Gen_Key | Int | | SI |
| Gen_Id | Varchar(20) | Identificador del Gen | NO |
| Descripción | Varchar(300) | Descripción de las funciones del gen. | NO |
| Familia | Varchar(50) | Familia a la que pertenece el Gen. | NO |
| Subfamilia | Varchar(50) | Subfamilia a la que pertenece el Gen. | NO |

| | | | |
|---|---|---|---|
| Inicio | Int | Posición Inicial del Gen en la secuencia. | NO |
| Final | Int | Posición Final del Gen en la secuencia. | NO |
| Band | Varchar(1) | Hélice del ADN donde está ubicado el gen. | NO |

**Dimensión Cromosoma – DimChr (Dimensión compartida):** Provee información relevante del cromosoma

**Tabla 4. Dimensión Cromosoma**

| Nombre del Campo | Tipo – Tamaño | Descripción | Llave primaria |
|---|---|---|---|
| Chr_Key | Int | | SI |
| Chr_Id | Varchar(5) | Identificador del cromosoma | NO |
| Organismo | Varchar(6) | Organismo al que pertenecen los cromosomas | NO |

**Dimensión Estructura – DimEstructura (Dimensión compartida):** Contiene información acerca de una estructura específica.

**Tabla 5. Dimensión Estructura**

| Nombre del Campo | Tipo – Tamaño | Descripción | Llave primaria |
|---|---|---|---|
| Estruct_Key | Int | | SI |
| Estruct_Id | Varchar(20) | Identificador de la estructura | NO |
| UI`s | Varchar(4266) | Arreglo de las longitudes de los exones e Intrones en la estructura separadas por coma. | NO |
| Num_unid_info | Int | Número de unidades Exones + Intrones. | NO |
| Molecular_funcion | Varchar(1500) | Función molecular de la estructura. | NO |

**Dimensión Band (DimBand):** Esta dimensión fue construida con el objetivo de poder suplir la necesidad de los investigadores de realizar análisis de los datos por intervalos y rangos de análisis.

**Tabla 6. Dimensión Band**

| Nombre del Campo | Tipo – Tamaño | Descripción | Llave primaria |
|---|---|---|---|
| Band_Key | Int | | SI |
| intervalo_1_unid | Varchar(30) | Intervalos de una unidad desde 1 hasta 50 | NO |
| intervalo_2_unid | Varchar(30) | Intervalos de dos unidades desde 1 hasta 50 | NO |
| intervalo_3_unid | Varchar(30) | Intervalos de diez | NO |

| | | | |
|---|---|---|---|
| | | unidades desde 1 hasta 50 | |
| band_lower_value | Float | Valor mínimo del rango | NO |
| band_upper_value | Float | Valor máximo del rango | NO |
| lower_limit_string | Varchar(39) | Valor mínimo del rango convertido a cadena para mostrar al usuario | NO |
| upper_limit_string | Varchar(39) | Valor máximo del rango convertido a cadena para mostrar al usuario | NO |
| band_div10 | Int | Orden para dividir en 10 partes cada unidad | NO |
| band_div20 | Int | Orden para dividir en 20 partes cada unidad | NO |
| limits_div10 | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en 10 partes cada unidad | NO |
| limits_div20 | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en 20 partes cada unidad | NO |
| 10_limite_bajo | Varchar(39) | Cadena de texto con el límite inferior, para dividir en 10 partes cada unidad | NO |
| 20_limite_bajo | Varchar(39) | Cadena de texto con el límite inferior, para dividir en 20 partes cada unidad | NO |
| 10_limite_alto | Varchar(39) | Cadena de texto con el límite superior, para dividir en 10 partes cada unidad | NO |
| 20_limite_alto | Varchar(39) | Cadena de texto con el límite inferior, para dividir en 20 partes cada unidad | NO |
| limits_div10_2_unids | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en 10 partes cada dos unidades | NO |
| limits_div20_2_unids | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en 20 partes cada dos unidades | NO |
| limits_div10_10_unids | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en | NO |

| | | | |
|---|---|---|---|
| | | 10 partes cada diez unidades | |
| limits_div20_10_unids | Varchar(39) | Cadena de texto con el Orden, límite inferior y superior para dividir en 20 partes cada diez unidades | NO |
| 10X2U_limite_bajo | Varchar(39) | Cadena de texto con el límite superior, orden para dividir en 10 partes cada dos unidades | NO |
| 20X2U_limite_bajo | Varchar(39) | Cadena de texto con el límite superior, orden para dividir en 20 partes cada dos unidades | NO |
| 10X10U_limite_alto | Varchar(39) | Cadena de texto con el límite superior, orden para dividir en 10 partes cada diez unidades | NO |
| 20X10U_limite_alto | Varchar(39) | Cadena de texto con el límite superior, para dividir en 20 partes cada diez unidades | NO |
| 10_parts_2_unids | Varchar(3) | Orden para dividir en 10 partes cada dos unidades | NO |
| 20_parts_2_unids | Varchar(3) | Orden para dividir en 20 partes cada dos unidades | NO |
| 10_parts_10_unids | Varchar(3) | Orden para dividir en 10 partes cada diez unidades | NO |
| 20_parts_10_unids | Varchar(3) | Orden para dividir en 20 partes cada diez unidades | NO |

**Definición de las Medidas:** Las medidas son transacciones individuales que cumplen con la granularidad definida para la tabla de hechos.

**Tabla 7. Medidas de la tabla de hechos Análisis fractal**

| Nombre del Campo | Tipo - Tamaño | Descripción | Tipo de medida (Regular - Calculada) | Método de Agregación |
|---|---|---|---|---|
| D | Numeric(24, 12) | Valor calculado y asignado al Gen por los investigadores del grupo BIMAC | NO | Avg |

| | | | | |
|---|---|---|---|---|
| R2 | Numeric(24, 12) | Valor calculado y asignado al Gen por los investigadores del grupo BIMAC | NO | Avg |
| #UI | Int | Número de unidades de información por Gen | NO | Avg |
| simetría | Numeric(25, 11) | Valor calculado y asignado al Gen por los investigadores del grupo BIMAC | NO | Avg |
| size | Numeric(7) | Longitud del Gen | NO | Avg |
| %gen# UI`s | Numeric(7) | Porcentaje de unidades de información por Gen(es) | NO | Avg |
| %gen_size | Numeric(7) | Porcentaje de logitud del gen(es) | NO | Avg |
| %gen_ simetría | Numeric(25) | Porcentaje de simetría del gen(es) | NO | Avg |
| #D's X Gen | Numeric(7) | Número de D por gen dependiendo de un R2 seleccionado | SI | Avg |
| #D´s X familia | Numeric(7) | Número de D por familia de genes dependiendo de un R2 seleccionado | SI | Avg |

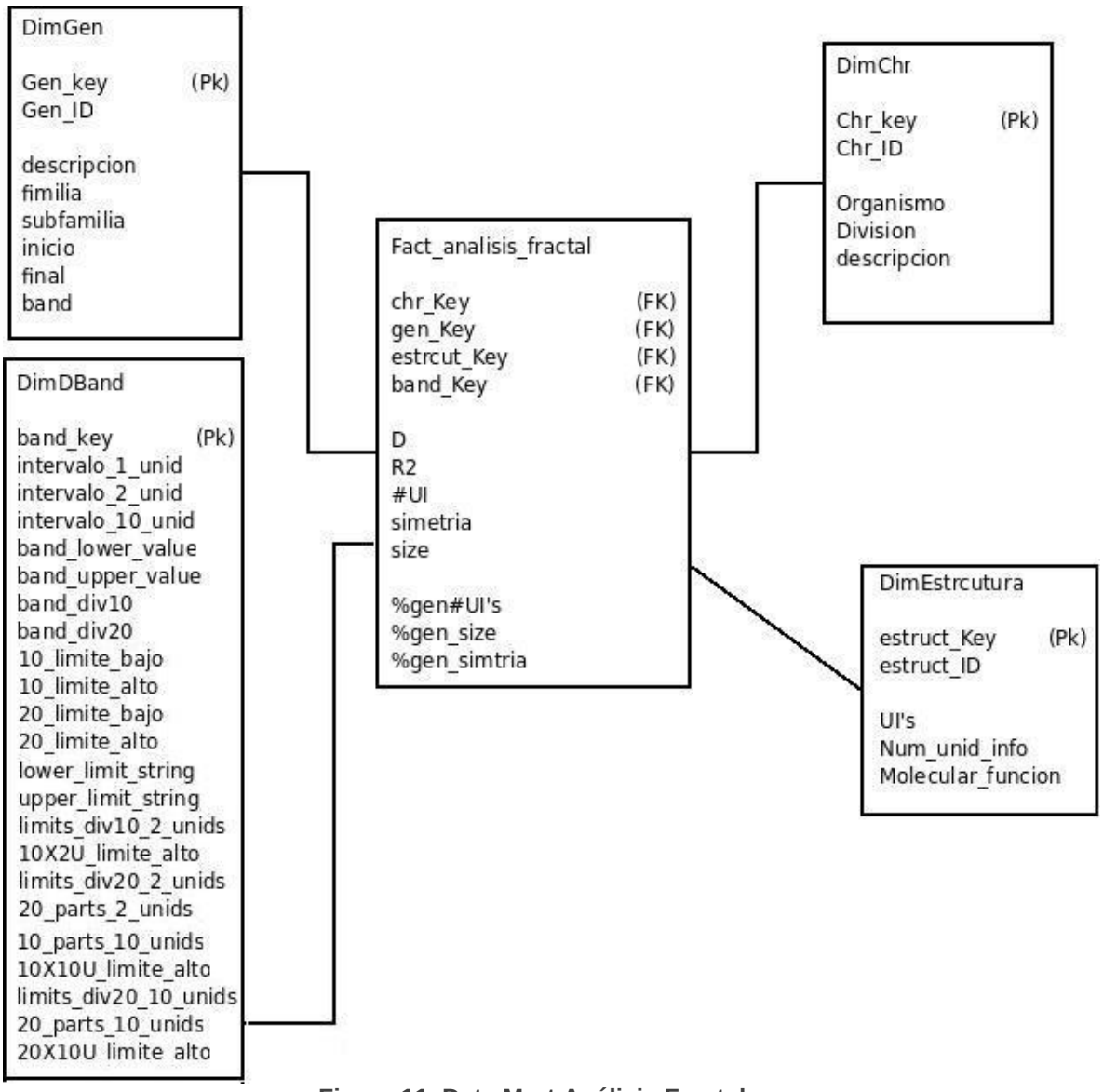



**Figura 11. Data Mart Análisis Fractal**

**Data Mart Análisis de Unidades de Información**

**Descripción del Data Mart Análisis de Unidades de Información**

El objetivo de este data mart consistía en permitir al investigador seleccionar el número todas las estructuras que tuvieran el mismo número de unidades de información, poder visualizar a que cromosomas, genes ó a que familia de genes pertenecen esas estructuras y finalmente poder organizar en orden ascendente o descendente las longitudes de las unidades las de información (Exón o Intrón).

**Dimensión Tipo – DimTipo:** Contiene información relevante de cada unidad de información perteneciente a una estructura.

**Tabla 8. Dimensión Tipo**

| Nombre del Campo | Tipo – Tamaño | Descripción | Llave primaria |
|---|---|---|---|
| Tipo_Key | Int | | SI |
| Tipo_Id | Varchar(10) | Identificador de la unidad de información | NO |
| Descripción | Varchar(10) | Descripción de la unidad de información | NO |

**Dimensión Gen:** Ver Tabla 3

**Dimensión Cromosoma:** Ver Tabla 4

**Dimensión Estructura:** Ver Tabla 5

**Definición de las Medidas:** Las medidas son transacciones individuales que cumplen con la granularidad definida para la tabla de hechos.

**Tabla 9. Medidas de la tabla de hechos análisis de las unidades de información**

| Nombre del Campo | Tipo – Tamaño | Descripción | Tipo de medida (Regular - Calculada) |
|---|---|---|---|
| orden_estruct | Int | | NO |
| orden_tipo | Int | Orden de la unidad de información dentro de la estructura de acuerdo a su tipo. | NO |
| descrip_orden_tipo | Varchar(10) | El orden de la Unidad de información dentro de la estructura junto con una descripción del tipo de unidad de información que es (Exón o Intrón). | NO |
| pos_inicial | Int | Posición inicial de la unidad de información en la secuencia. | NO |
| pos-final | Int | Posición inicial de la unidad de información en la secuencia. | NO |
| longitud | Int | Longitud de la unidad de información. | NO |
| descrip_orden_tipo | Varchar(10) | El orden de la Unidad de información dentro de la estructura junto con una descripción del tipo de unidad de información que es (Exón o Intrón). | NO |

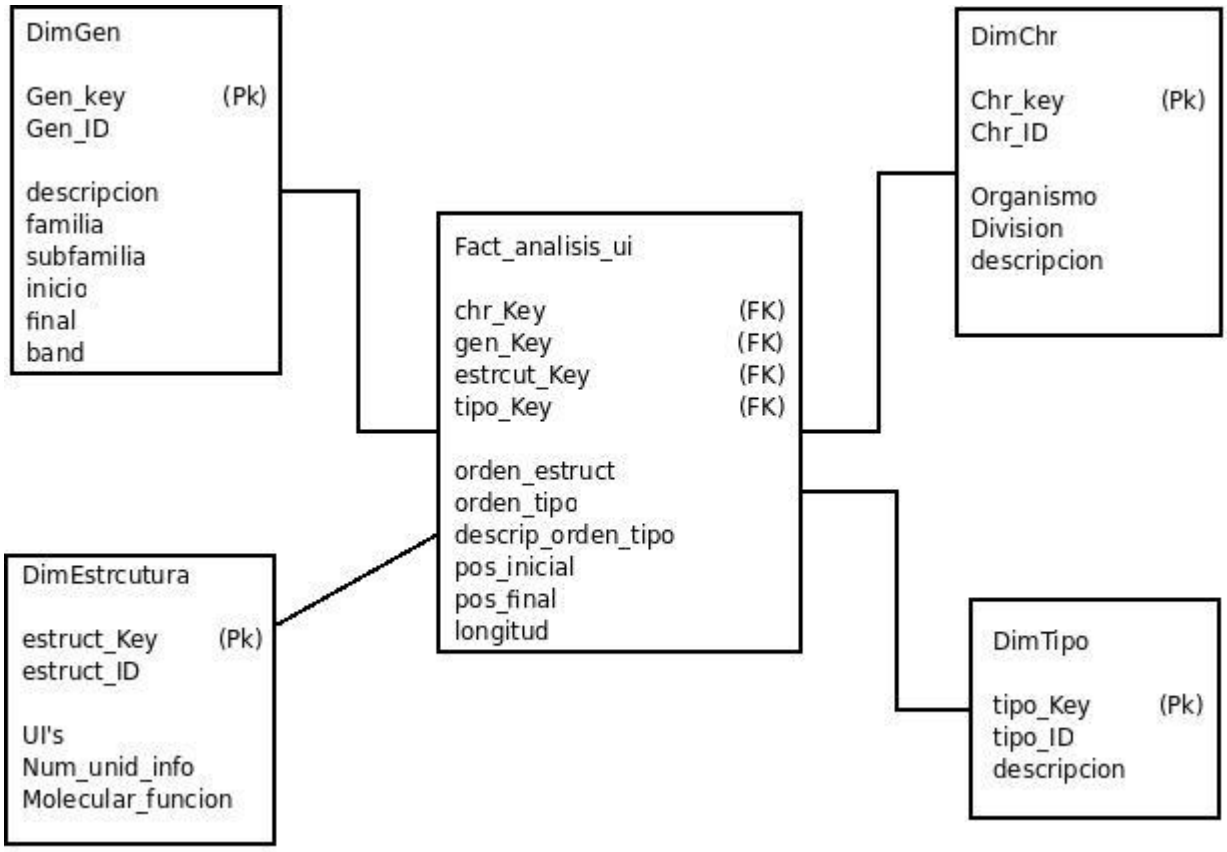
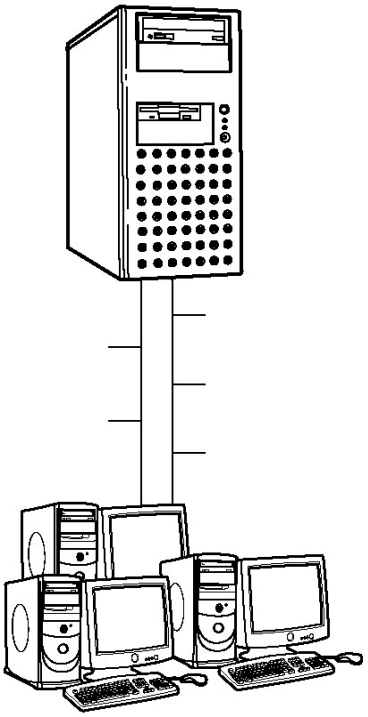



**Figura 12. Data Mart Análisis Unidades de Información**

### 3.1.4.Diseño Físico del Data Warehouse

Se hizo uso la arquitectura "ALL IN ONE", propuesta como una opción de arquitectura física por Ralph Kimball [14], teniendo en cuenta que el grupo de investigación cuenta con un solo servidor, además el tamaño de datos cargados en el data warehouse no sobrepasa las capacidades de almacenamiento y tampoco las capacidades de procesamiento con las que se cuenta en el momento.



**Pentho in Jboss Server**

- MySQL RDBMS
- Kettle ETL-tool
- Mondrian OLAP engine
- Jpivot Reporter
- Open flash charts

- Browser
(IE, Firefox, Chrome)
- Herramientas de análisis

**Figura 13. Arquitectura Física ALL IN ONE [14]**

En la Figura 13 podemos apreciar la Arquitectura Física usada en la construcción de la bodega de Datos.

**Áreas de la arquitectura**

**Tabla 10. Área de Datos**

| Nivel de detalle | Datos (qué) |
|---|---|
| **Auditoría y requerimientos del negocio** | **¿Qué información es necesaria para que la investigación obtenga un mejor análisis de los datos?, ¿Cuáles de estos datos están Disponibles?** |
| | Los datos disponibles, son los necesarios para llevar a cabo el desarrollo de este proyecto, teniendo en cuenta que dichos datos son la fuente con la cual se generan los reportes de la investigación, estos datos han sido descritos en este documento anteriormente, ver modelado dimensional. |
| **Modelos y documentos de** | **¿Cuáles son sus entidades principales? (los hechos y las dimensiones)** |

| | |
|---|---|
| **la Arquitectura** | Las dimensiones principales son: Gen, cromosoma, estructura, tipo<br>Las medidas: todas las medidas descritas en las tablas anteriores. |
| **Modelos detallados y especificaciones** | **¿Cuáles son los elementos individuales, sus definiciones, dominios y reglas de derivación?** |
| | Ver Tabla 3. Dimensión Gen<br>Ver Tabla 4. Dimensión Cromosoma<br>Ver Tabla 5. Dimensión Estructura<br>Ver Tabla 7. Medidas de la tabla de hechos Análisis fractal<br>Ver Tabla 8. Dimensión tipo<br>Ver **¡Error! No se encuentra el origen de la referencia.**. Medidas de la tabla de hechos Análisis de las unidades de información |

**Tabla 11. Área Técnica**

| **Nivel de detalle** | **Técnica (Cómo)** | |
|---|---|---|
| | **Back Room** | **Front Room** |
| **Auditoría y requerimientos del negocio** | **¿Cómo se obtendrán los datos, transformarán y se harán disponibles a los investigadores?** | **¿Cuáles son los mayores retos que se afrontan en el grupo de investigación?** |
| | Los datos son obtenidos de los archivos de texto plano, los cuales son los resultados de la investigación, posteriormente se les realizara el proceso ETL, correspondiente y se presentaran a los usuarios en una herramienta WEB, que les permitirá, navegar y manipular los datos. | Manipular los datos resultantes de la investigación y difundirlos en la comunidad científica. |
| | **¿Cómo se está haciendo esto hoy?** | |
| | Se solicita a un investigador que cree un script en python que genere los resultados con los filtros deseados y se entrega un archivo plano. | |
| **Modelos y documentos de la Arquitectura** | **¿Dónde está almacenada la mayor cantidad de datos y donde debería estar localizada?** | **¿En qué forma necesitan los investigadores obtener la información para que sea útil?** |
| | La mayoría de datos esta dividida en archivos planos individuales, debería estar centralizados y ser de fácil acceso y consulta. | La información se debe encontrar estructurada de acuerdo a los requerimientos planteados por el usuario |
| **Modelos detallados y especificaciones** | **¿Qué normas y productos proveen lo necesario?** | |
| | La metodología de Ralph Kimball y la suite de pentaho. | |

**Tabla 12. Área de la Infraestructura**

| Nivel de detalle | Infraestructura (Dónde) |
|---|---|
| **Auditoría y requerimientos del negocio** | **¿Qué capacidades de hardware y nivel de sistemas necesarios?** |
| | Se requieren sistemas que faciliten las actividades de desarrollo del proyecto, suites de inteligencia de negocios y un servidor con la capacidad necesaria para el almacenamiento de los datos, teniendo en cuenta el crecimiento del sistema.<br>Para tener un ejemplo, se estima que los datos procesados de un genoma, oscilan entre 60 y 70 MB y el servidor debe tener capacidad para almacenar 10 genomas a futuro. |
| **Modelos y documentos de la Arquitectura** | **¿Cuáles son las capacidades específicas con las que se cuenta?** |
| | Servidor con 8 GB en RAM, almacenamiento en disco duro de 200 GB. |

### 3.1.5. Diseño del Sistema ETL

El proceso de ETL de este proyecto no contó con desafíos respecto a la transformación y extracción de los datos, esto debido a que los archivos proporcionados por los investigadores presentan homogeneidad en los nombres y en el formato de los datos debido a que todos provienen de la misma fuente. Se presentaron algunos problemas en el momento del cargue, pero estos problemas están fuera del alcance de las herramientas y serán especificados más adelante. Los Diagramas de cargue de cada una de las dimensiones y tablas de la bodega pueden verse en los anexos de este proyecto. A continuación se presentaran los diagramas de alto nivel para el ETL.


**Figura 14. Diagrama de alto nivel ETL Dimensión Cromosoma**

En la Figura 14 se ilustra el flujo realizado para cargar los datos de la dimensión cromosoma, dichos datos se encontraban en un archivo de texto titulado cromosoma, en el cual se encontraba información tal como el identificador del cromosoma, el organismo al que pertenece el cromosoma y una descripción.


**Figura 15. Diagrama de alto nivel ETL Dimensión Rangos**

En la Figura 15 se ilustra el flujo realizado para cargar los datos de la dimensión Band, dichos datos se encontraban en un archivo de texto titulado Rangos, en el cual se encontraba información tal como el identificador del intervalo, un campo para identificar a que agrupación pertenece un valor, los limites superiores e inferiores de un intervalo y diferentes tipos de notación de los límites.

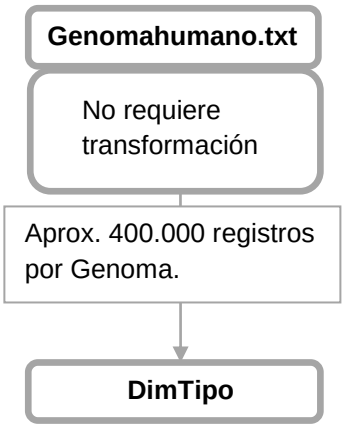



**Figura 16. Diagrama de alto nivel ETL Dimensión Tipo**

En la Figura 16 se ilustra el flujo realizado para cargar los datos de la dimensión Tipo, dichos datos se encontraban en un archivo de texto titulado Genomahumano, en el cual se encontraba información tal como el identificador de la unidad de información y una descripción.

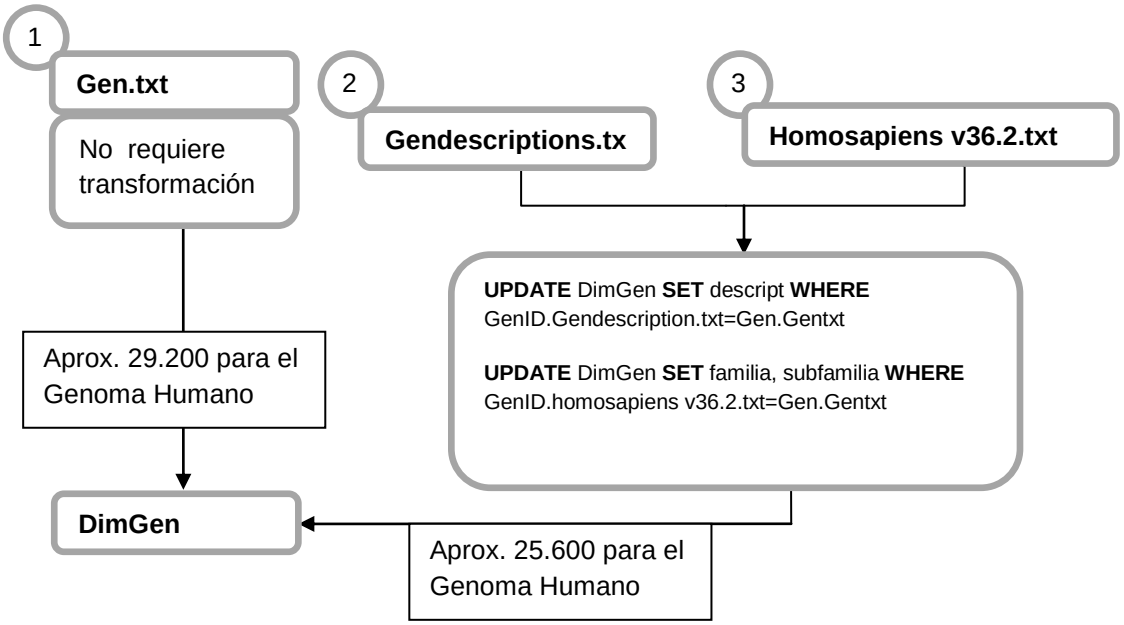



**Figura 17. Diagrama de alto nivel ETL Dimensión Gen**

En la Figura 17 se ilustra el flujo realizado para cargar los datos de la dimensión Gen, para realizar el cargue de esta dimensión fue necesario extraer los datos desde múltiples archivos de texto, tales como Gen, Gendescriptions y Homosapiens. De estos archivos se extrajo la información requerida para poblar la dimensión, tal como, el nombre del gen, la hebra donde se encuentra ubicado el gen (band), la descripción del gen, la familia y la subfamilia a la que pertenece.

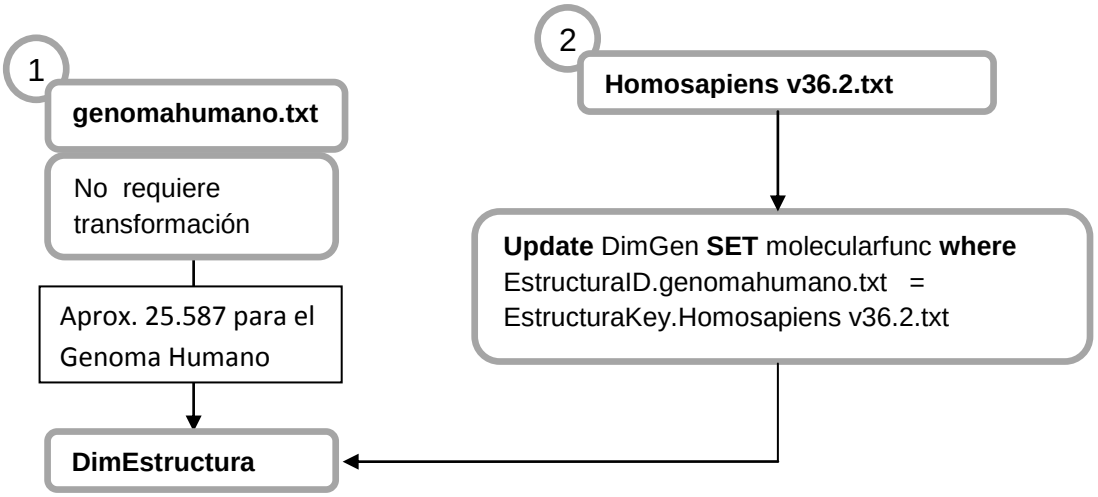



**Figura 18. Diagrama de alto nivel ETL Dimensión Estructura**

En la Figura 18 se ilustra el flujo realizado para cargar los datos de la dimensión Estructura, para realizar el cargue de esta dimensión fue necesario extraer los datos desde múltiples archivos de

texto, tales como: genomahumano y Homosapiens. De estos archivos se extrajo la información requerida para poblar la dimensión, tal como: el identificador de la estructura y la función molecular.

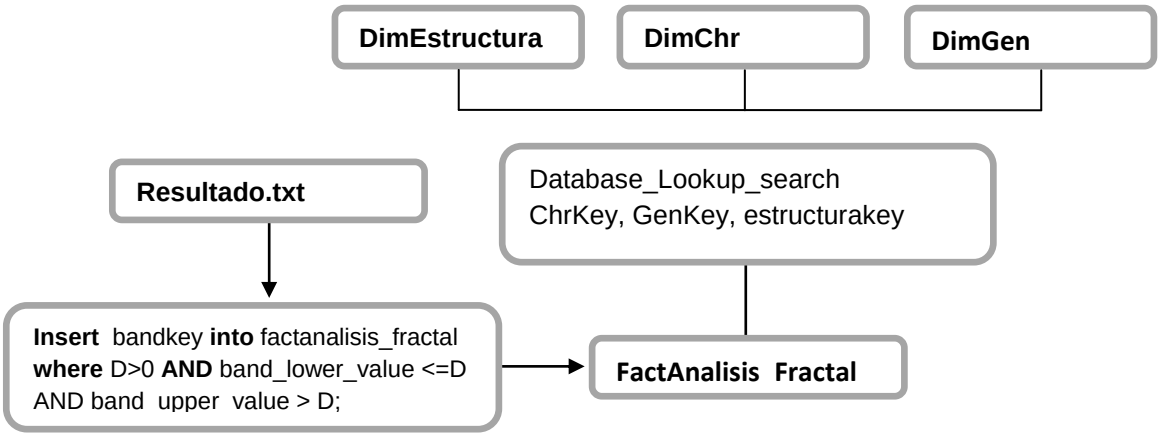



**Figura 19. Diagrama de alto nivel ETL Tabla de Hechos Análisis Fractal**

En la Figura 19 se ilustra el flujo realizado para cargar los datos de la tabla de hechos análisis fractal, para realizar el cargue de esta tabla es necesario asociar las dimensiones a la tabla y extraer las medidas desde el archivo de texto, titulado resultado.

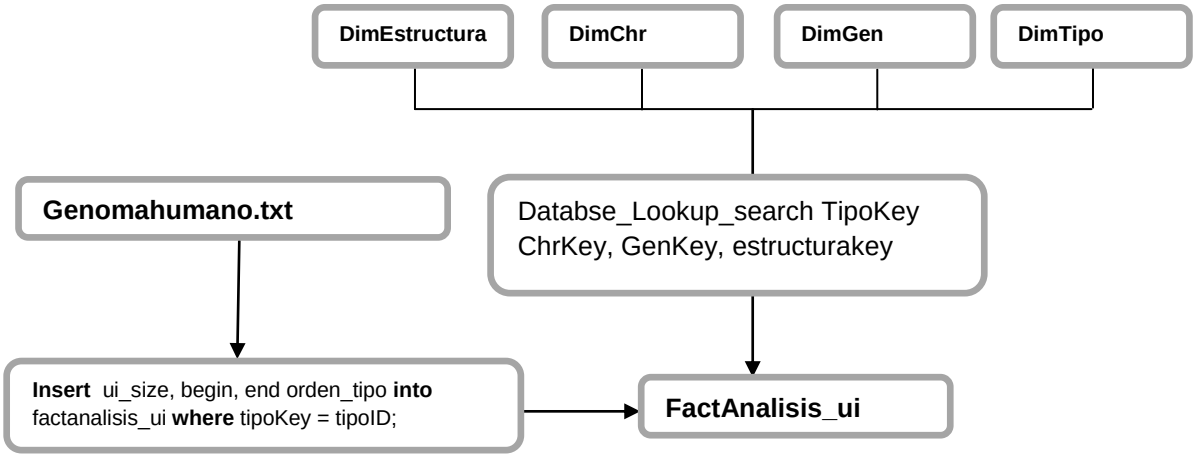



**Figura 20. Diagrama de alto nivel ETL Tabla de Hechos Unidades de información**

En la Figura 20 se ilustra el flujo realizado para cargar los datos de la tabla de hechos unidades de información, para realizar el cargue de esta tabla es necesario asociar las dimensiones a la tabla y extraer las medidas desde el archivo de texto, Genomahumano.

En el Anexo H, se encuentran los paquetes de carga ETL, guiados por los diagramas de alto nivel.

### 3.1.6.Conjunto de Herramientas

**Diseño de la Arquitectura Técnica**

La arquitectura técnica del proyecto está orientada bajo la arquitectura de la suite Pentaho. En la Figura 21 se observan todos los componentes del DSS.

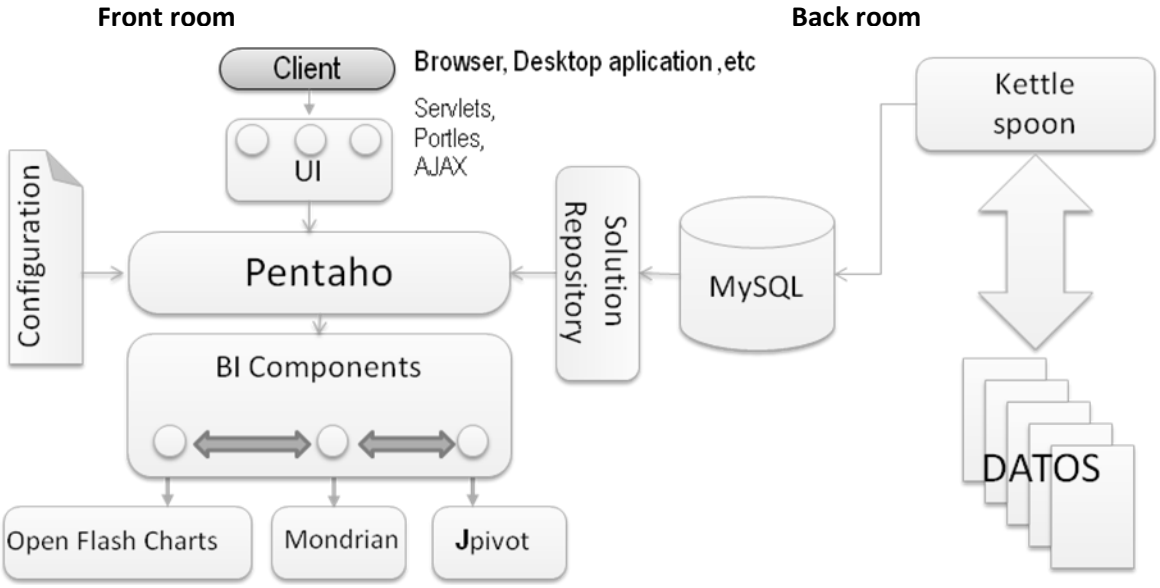



**Figura 21. Arquitectura Técnica**

Como muestra la Figura 21, la arquitectura se divide en dos partes, la parte interna conocida como back room y la parte externa conocida como front room. En el back room se describen los orígenes de datos, el flujo de datos, consulta de servicios, metadata, automatización de procesos entre otros. En el front room se describe las herramientas estáticas y dinámicas para acceder y consultar los datos del DW [14]

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

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

- Los datos pertenecientes a la investigación, requeridos para poblar el DW se encuentra en archivos planos.

- Para poblar el DW, de acuerdo a los flujos diseñados es necesario una herramienta de ETL. Para este proyecto se utilizó **KETTLE** con la cual se realizó la extracción desde el sistema fuente, la transformación, la limpieza y el cargue de datos.

- Para la construcción y mantenimiento del DW relacional se necesita un motor de base de datos, en el caso de este proyecto se hizo uso del motor de base de datos **MySQL.**

- Para la construcción y mantenimiento del DW multidimensional se necesita un motor de administración de bases de datos multidimensionales OLAP, en el caso de este proyecto se hizo uso de la suite **Pentaho**, y de **Mondrian** el cual es el motor multidimensional**.**

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

- Para la creación de los reportes dinámicos (ad hoc) se hizo uso de la herramienta **Jpivot**.

- Para generar los gráficos requeridos se utilizo **Open flash charts**.

Para mayor detalle de las arquitecturas de los productos utilizados, ver Anexo L.

**Funcionamiento de la arquitectura**

Una vez los datos son cargados en el DW mediante la herramienta de ETL Kettle - Spoon, un repositorio en MySQL.

Pentaho posee un motor de soluciones que crea un ambiente de ejecución, el cual es responsable de cargar los recursos, ejecutar los actions sequences y auditar, se puede pensar en él como una especie de maquina virtual BI (Bussiness intelligence).

Lo que Pentaho ejecuta es una colección de componentes BI, componentes para graficar, reportar, imprimir, OLAP, mail… etc, en este caso se usa el motor OLAP **mondrian**, herramienta de reportes dinámicos (ad hoc), **Jpivot** y para graficar **Open flash charts**. Para mayor detalle de los artefactos implementados ver Anexo I.

**Selección e Instalación del Producto**

La selección del producto para el desarrollo de este proyecto tuvo un criterio determinante por parte de los usuarios, quienes sugirieron que el desarrollo fuese en herramientas libres, se realizaron algunos prototipos en software propietario, pero las herramientas fueron descartadas y se le dio prioridad a la petición del usuario.
Los manuales de instalación y configuración del producto se encuentran en los anexos de este documento.

A continuación se presentan dos tablas comparativas de las herramientas libres existentes en el mercado para desarrollar aplicaciones BI.

**Herramientas de Software libre disponibles al momento de selección**

**Tabla 13. Comparación herramientas suites BI**

| Nombre | Análisis | Gráficas | Almacenamiento de datos | Facilidad de implementación de una solución sin experiencia en la herramienta. | Soporte |
|---|---|---|---|---|---|
| **JasperSoft 2.0** | JasperAnalysis | Ireport | **Compatible con las bases de datos más comunes.** | Alta | Bajo |
| **Pentaho 1.7** | Jpivot | JfreeReport Ireport | | Alta | Bajo |

| | | | | | |
|---|---|---|---|---|---|
| | | Bird<br>Open flash Charts | | | |
| **Bissgress Green Plum** | - | - | PostgreSQL | desconocida | ninguno |

La Tabla 13 muestra los criterios de comparación de las suites libres de BI existentes, se seleccionó la suite de Pentaho 1.7, teniendo en cuenta que ofrece más posibilidades en el manejo gráfico el cual era un elemento muy importante para el proyecto y cuenta con una amplia variedad para la conexión a múltiples bases de datos.

**Herramientas OLAP**

**Tabla 14. Comparación herramientas OLAP**

| **Nombre** | **Incluida en la herramienta** | **Basado en** | **Aplicación Web** | **Funcionalidades estándar(Drill, down, roll up, slice and dice, graficar)** | **Servidor de mdx** |
|---|---|---|---|---|---|
| **Jpivot** | Pentaho 1.7 | Javascript | Si | Todas | Mondrian |
| **WebPalo** | Independiente | GWT | Si | No grafica | Mondrian |
| **Jasperanalisis** | Jasper 2.0 | Jpivot | Si | Todas | Mondrian |
| **freeanalysis** | Independiente | Eclispe | No | Todas | Mondrian |
| **rubik** | Independiente | Eclipse | No | Todas | Mondrian |
| **HaloGen** | Independiente | Jpivot - Gwt | Si | Todas | Mondrian |

En la Tabla 14 se observan los criterios de comparación que se tuvieron en cuenta para la selección de la herramienta OLAP, se escogió la herramienta jpivot teniendo en cuenta que viene incluida en la suite seleccionada para construir la solución y además cuenta con las funcionalidades requeridas para el desarrollo del proyecto.

### 3.1.7. Diseño De La Base De Datos Multidimensional

Luego de la construcción y carga del DW relacional (ver Anexo J.), el siguiente paso se centra en la construcción de la base de datos multidimensional, la cual es creada con la ayuda de la herramienta Workbench, la cual pertenece a la suite de Pentaho y facilita la construcción del esquema dimensional.

El esquema es creado como un XML, el cual permitirá que el servidor dimensional y el gestor de consultas mdx mondrian interprete de manera dimensional los datos que existen en la base de datos relacional, se debe tener en cuenta seguir ciertos pasos durante la construcción del esquema en workbench:

- Preparar el diseño
- Crear un Origen de Datos
- Crear y refinar las dimensiones, sus atributos y jerarquías.
- Crear y refinar las tablas de hechos.
- Crear y Editar cubos
- Crear y Editar cubos virtuales
- Crear medidas calculadas
- Realizar iteraciones

El proceso que se siguió para la construcción del esquema multidimensional fue:

1. Crear una conexión a la base de datos DwBio
2. Crear un nuevo esquema en workbench (con el nombre Dwbiofinal).
3. Crear las dimensiones compartidas y cada uno de sus atributos con sus jerarquías

    a. Cuando se edita una nueva dimensión, se siguen los siguientes pasos:
        i. Editar los nombres de cada atributo dentro de la dimensión y su origen
        ii. Editar las propiedades de la dimensión.
        iii. Editar las propiedades de cada atributo.
        iv. Crear las jerarquías. Editar las propiedades de las jerarquías y las propiedades de cada nivel.

4. Crear un cubo multidimensional por cada tabla de hechos y asignar sus medidas
5. Crear las dimensiones que no son compartidas y solo pertenecen a un cubo, sus atributos, y jerarquías
6. Crear las medidas calculadas si las hay.

En la Figura 22, se puede observar la herramienta workbench al momento de la creación del esquema, la herramienta permite crear los cubos, con sus respectivas dimensiones, medidas y jerarquías requeridas por el modelo.

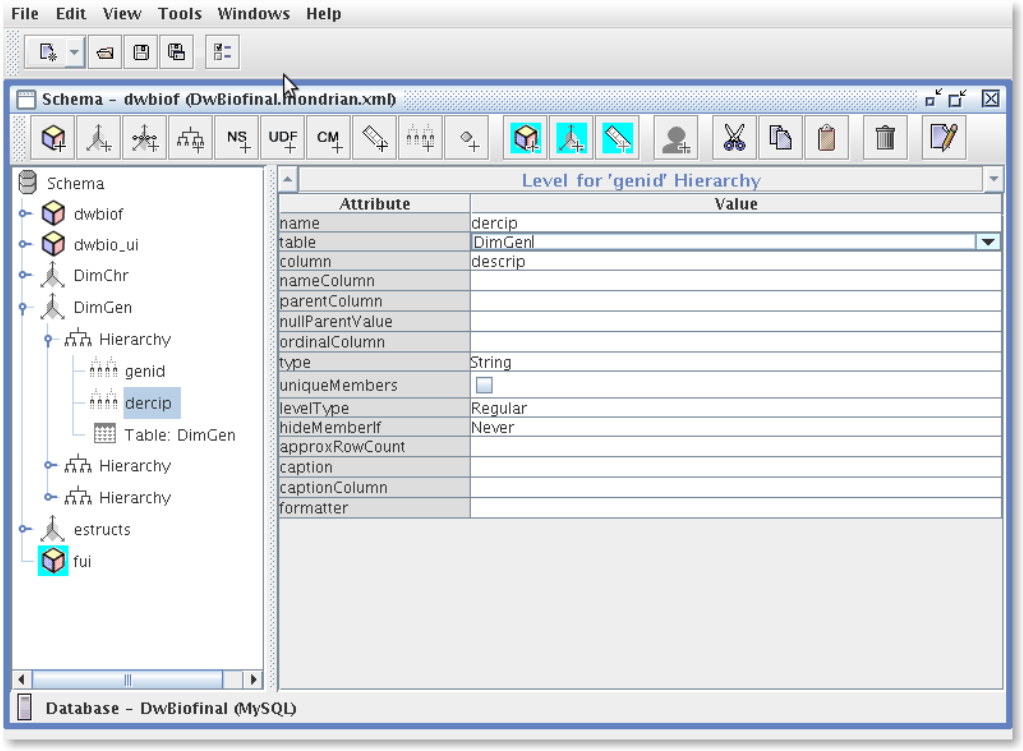



**Figura 22. Creación del esquema dimensional con Workbench**

### 3.1.8. Despliegue del DW.

Esta etapa es el punto de convergencia de las tres rutas, la de la tecnología, la de los datos y la de aplicaciones. En esta etapa se realizan varias actividades, que deben llevarse a cabo antes de poner en producción el DW, garantizando de esta manera el correcto funcionamiento del producto, las actividades que pueden encontrarse en esta etapa son: la configuración de hardware, configuraciones de red, conexiones a bases de datos, chequeos de seguridad, instalación de software, documentación, manuales y capacitaciones a los usuarios, etc. [19].

Para el despliegue del sistema se verificó que los productos estuvieran correctamente instalados, verificar el funcionamiento del servidor en cuanto a hardware como las conexiones de bases de datos, el correcto funcionamiento del sitio, del servidor ETL, del servidor de análisis, el acceso a los datos y hacer la entrega de manuales y documentación necesaria para el manejo de la aplicación, también se realizaron capacitaciones a los usuarios administradores. Finalmente fue puesta en funcionamiento en un servidor Dell Power Edge 2800 de propiedad y uso del grupo de investigación BIMAC.

En el Anexo K. se encuentra un test de verificación de los resultados de las consultas obtenidas por el DSS, en contraste con los resultados obtenidos por los investigadores sin la herramienta.

### 3.1.9. Mantenimiento y crecimiento

Es un proceso con etapas bien definidas, con comienzo y fin, pero de naturaleza espiral que acompaña la evolución de la organización durante toda su historia. Es importante establecer las prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer. Se debe tener en cuenta una serie de puntos para mantener el DW exitosamente, entre ellos se destacan: el continuo soporte y la constante capacitación a usuarios de negocios, el manejo de la infraestructura (monitoreo de base de datos, tráfico, etc.), afinar el rendimiento sobre las consultas, mantenimiento de la metadata y procesos ETL [19]. Estas dos etapas están fuera del alcance de este proyecto.

### 3.1.10. Administración del proyecto

La etapa de administración del proyecto es la encargada de monitorear, controlar y administrar todas las actividades del proyecto que se realizan en el Ciclo de Vida Dimensional. Las diferentes actividades se centran en verificar el estado del proyecto y su evolución con el tiempo [19]. La administración del presente proyecto se fue realizando a lo largo de desarrollo de todo el sistema de DW/BI, al contrastar los resultados que se estaban obteniendo con el plan proyecto establecido en las primeras etapas de construcción. Para esto fue indispensable el rol que jugó el director del proyecto, quien por ser experta en el tema de los DSS, guió el desarrollo del proyecto y contribuyó en la verificación de que la construcción de la solución cumpliera con las fases requeridas para el proyecto. Adicionalmente se llevo de manera general una tabla de seguimiento del desarrollo del proyecto. (Ver Anexo G).

# CAPITULO IV. DESCRIPCIÓN DEL PROTOTIPO DE LA HERRAMIENTA OLAP

Este capítulo describe el proceso de desarrollo usado en la construcción de la herramienta OLAP y los artefactos obtenidos en el proceso.

El desarrollo de las aplicaciones BI (Business Intelligent), forma parte de una de las tres rutas establecidas por el ciclo de vida dimensional, la ruta de aplicación como es llamada esta ruta propone el desarrollo completo de una herramienta de usuario final. La metodología divide el proceso de desarrollo de las aplicaciones en dos etapas: la de la especificación de la herramienta y la del desarrollo de la herramienta.

El desarrollo del prototipo de la herramienta OLAP estuvo guiado por las fases del ciclo de vida dimensional, propuesto por Kimball en lo referente al desarrollo de aplicaciones de usuario final y se utilizo como complemento de las actividades propuestas en estas fases, algunos artefactos del lenguaje unificado de modelado (UML – por sus siglas en inglés), los cuales permitieron una mejor especificación del diseño del prototipo de la herramienta.

**Especificación y desarrollo del prototipo**

Las especificaciones de las aplicaciones de usuario son guiadas por los requerimientos recolectados y que finalmente fueron priorizados como se especifica a continuación.

**Priorizar las Necesidades del Cliente.**

**Tabla 15. Requerimiento priorizados**

| # | Código | Necesidad | Prioridad |
|---|---|---|---|
| 1 | **AF-N1** | El usuario determina en cuantos rangos se dividirá el intervalo de análisis. | **ALTA** |
| | **AF-N1** | El usuario puede determinar el tamaño de los rangos. | **ALTA** |
| 2 | **AF-N5** | Obtener la cantidad de genes que se encuentra en cualquiera de los rangos de análisis. | **ALTA** |
| | **AF-N5** | Obtener el promedio de unidades de información se encuentra en cualquiera de los rangos de análisis. | **ALTA** |
| | **AF-N5** | Obtener el promedio de las longitudes de los genes que se encuentra en cualquiera de los rangos de análisis. | **ALTA** |
| | **AF-N5** | Graficar la relación entre el número de genes y los rangos definidos | **ALTA** |
| | **AF-N5** | Graficar la relación entre el promedio de unidades de información y los rangos definidos. | **ALTA** |
| | **AF-N5** | Graficar la relación entre el promedio de longitudes de los genes y los rangos definidos. | **ALTA** |
| 3 | **AF-N14** | Visualización de la consulta en la herramienta OLAP | **ALTA** |
| 4 | **AUI-N1** | Permitir al usuario determinar el número de unidades de información que desea analizar | **ALTA** |
| | **AUI-N1** | Permitir al usuario escoger el cromosoma que desea analizar | |
| | **AUI-N1** | Visualizar el resultado de la consulta según las especificaciones del usuario | **ALTA** |
| | | | |
| 1 | **AF-N15** | Establecer tareas típicas de la herramienta OLAP. | **MEDIA** |
| 2 | **AF-N16** | Permitir la adición de dimensiones y medidas en la herramienta OLAP | **MEDIA** |

La Tabla 15 hace parte de los artefactos usados para el análisis de los requerimientos del sistema, la estructura de la tabla tiene un identificador para cada uno de los requerimientos y un campo con la priorización de estos. La notación del identificador asignado corresponde a las letras iníciales del data mart al que pertenece el requerimiento, por ejemplo AF – N16 corresponde a el data mart, análisis fractal.

Teniendo en cuenta tabla anterior, se agruparon los requerimientos de acuerdo a los parámetros de filtrado inicial, las agrupaciones dieron como resultado dos escenarios los cuales fueron especificados en la toma de requerimientos (ver **¡Error! No se encuentra el origen de la referencia.**). Los escenarios se convirtieron posteriormente en los data marts, análisis fractal y análisis de las unidades de información.

Para las consultas relacionadas con el datamart análisis fractal se tienen las siguientes especificaciones:

- Permitir a los usuarios definir el escenario deseado de acuerdo a los filtros planteados por el grupo de investigación y los cuales se encuentran especificados en los requerimientos del sistema.
- Desplegar las medidas de la tabla de hechos de acuerdo a los filtros realizados.
- Permitir al usuario construir gráficas de relación de los datos.

Una consulta de este data mart debe visualizarse con las siguientes filas y columnas.

**Tabla 16. Especificación visualización de los datos del data mart Análisis fractal**

| | | %D | #D´s X familia | #D´s XGen | %#UI | %Size | %Simetria |
|---|---|---|---|---|---|---|---|
| **Rango1** | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| **Rango2** | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |
| | Gen | 9.9999 | 99 | 99 | 9.9999 | 9.9999 | 9.9999 |

Para las consultas relacionadas con el datamart de las unidades de información se tienen las siguientes especificaciones:

- Permitir a los usuarios definir el escenario deseado de acuerdo a los filtros planteados por el grupo de investigación y los cuales se encuentran especificados en los requerimientos del sistema.
- Desplegar las medidas de la tabla de hechos de acuerdo a los filtros realizados.

Una consulta de este data mart debe visualizarse con las siguientes filas y columnas.

**Tabla 17. Especificación visualización de los datos del data mart Análisis unidades de información**

| | | Longitud | Descripción- Orden/Tipo |
|---|---|---|---|
| **Gen1** | Estructura | 9.9999 | xxxxxx |
| | Estructura | 9.9999 | xxxxxx |
| | Estructura | 9.9999 | xxxxxx |

| | | | |
|---|---|---|---|
| **Gen2** | Estructura | 9.9999 | xxxxxx |
| | Estructura | 9.9999 | xxxxxx |
| | Estructura | 9.9999 | xxxxxx |

**Casos de uso**

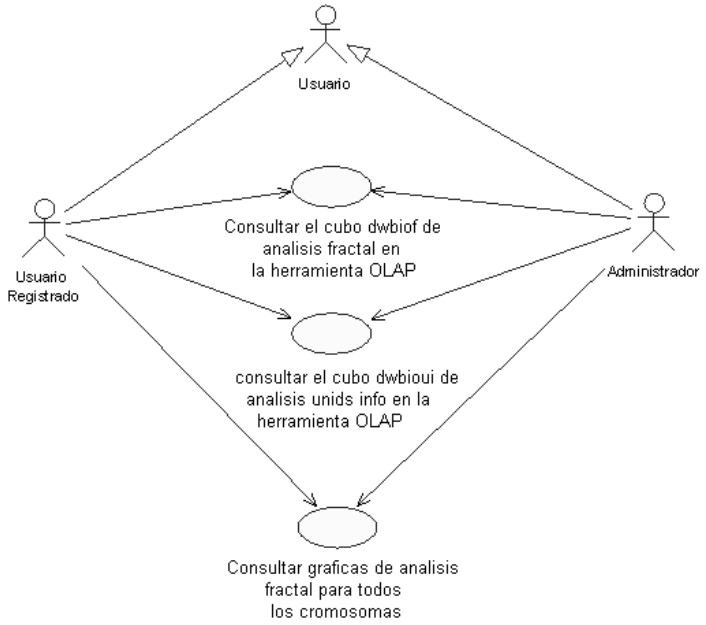



**Figura 23. Diagrama de Caso de Uso Consultar dwbioui**

**Tabla 18. Caso de uso Consulta del cubo dwbioui**

| | |
|---|---|
| **Caso de uso:** | Consultar el cubo dwbioui – análisis fractal en la herramienta OLAP |
| **Actores:** | Usuario Registrado, Administrador |
| **Propósito:** | Hacer uso de los filtros para mostrar los datos deseados y realizar consultas Dinámicas personalizadas (ad hoc) sobre el cubo almacenado en el servidor mondrian. |
| **Resumen:** | Cualquier de los dos usuarios tiene la posibilidad de ingresar dentro del modulo Dinámico OLAP y realizar consultas personalizadas sobre el cubo. |
| **Tipo:** | Primario. |
| **Curso normal de eventos** | |
| **Acción de los Actores** | **Respuesta del sistema** |
| **1.** Este caso de uso se inicia cuando un visitante quiere hacer consultas personalizadas sobre el cubo multidimensional.<br>**3.** El usuario selecciona mediante el filtro de Jpivot el número de unidades de información que desea.<br>**5.** El usuario configura las gráficas dinámicamente seleccionando valores en los filtros. | **2.** El sistema presenta la herramienta OLAP encargada de hacer consultas dinámicas.<br>**4.** Pentaho gestiona los recursos necesarios para que mondrian ejecute la consulta mdx y open flash chart la grafique. |

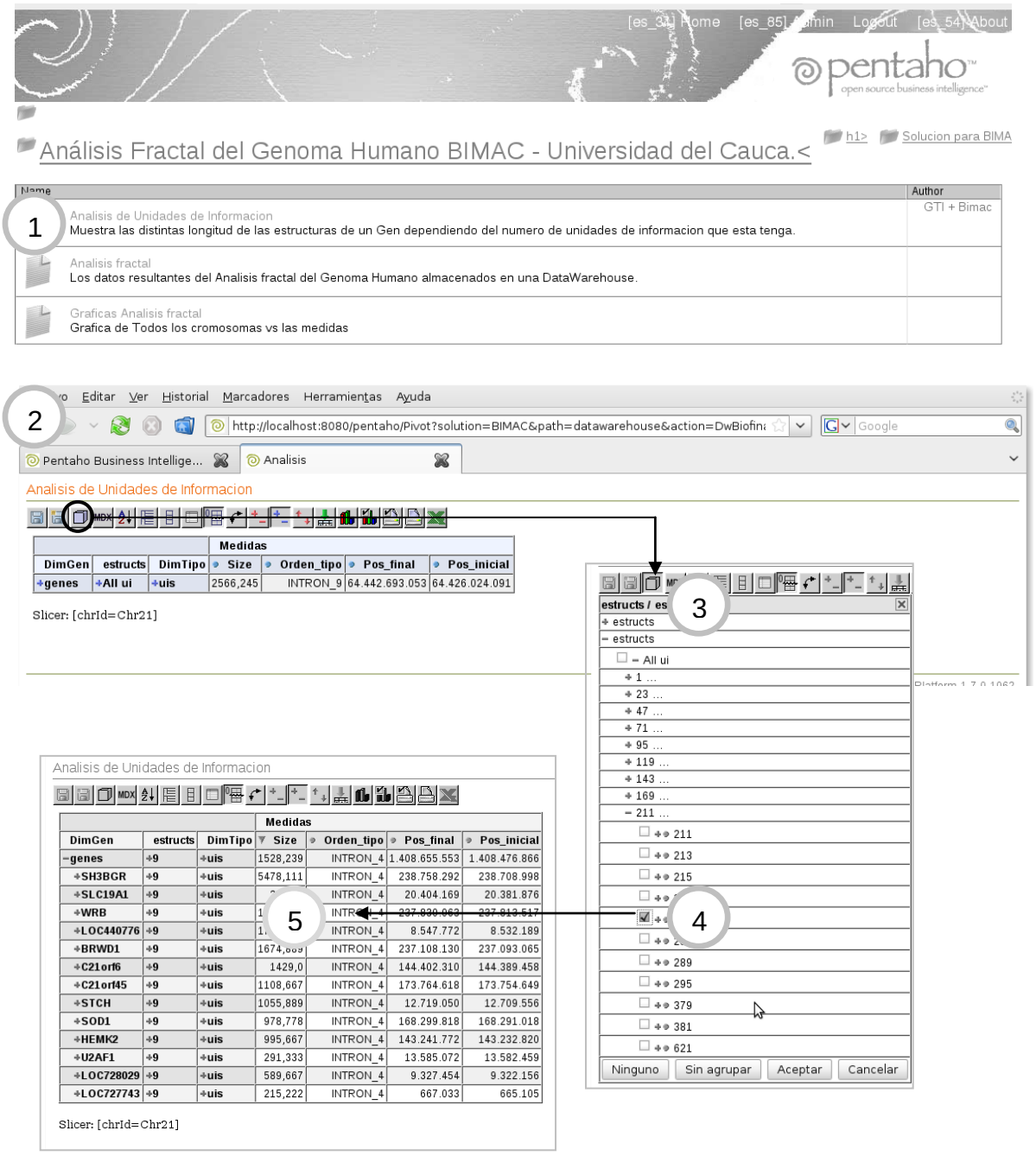



**Figura 24. Caso de Uso Real Consultar dwbioui**

**Tabla 19. Caso de uso Consulta del cubo dwbiof**

| | |
|---|---|
| **Caso de uso:** | Consultar el cubo dwbiof – análisis de unidades de información en la herramienta OLAP |
| **Actores:** | Usuario Registrado, Administrador |

| | |
|---|---|
| **Propósito:** | Hacer uso de Jpivot para seleccionar el número de unidades de información, para mostrar los datos deseados y realizar consultas Dinámicas personalizadas (ad hoc) sobre el cubo con el servidor mondrian. |
| **Resumen:** | Cualquier de los dos usuarios tiene la posibilidad de ingresar dentro del modulo Dinámico OLAP y realizar consultas personalizadas sobre el cubo. |
| **Tipo:** | Primario. |

| **Curso normal de eventos** | |
|---|---|
| **Acción de los Actores** | **Respuesta del sistema** |
| **1.** Este caso de uso se inicia cuando un visitante quiere hacer consultas personalizadas sobre el cubo multidimensional. | **2.** El sistema presenta un conjunto de filtros adicionales y la herramienta OLAP encargada de hacer consultas dinámicas. |
| **3.** El usuario selecciona mediante el conjunto de filtros adicionales a Jpivot los valores q desea visualizar en Jpivot. | **4.** El sistema obtiene los datos multidimensionales (Dimensiones y Medidas) determinadas por los filtros en el cubo. |
| **5.** El usuario genera consultas dinámicas seleccionando nuevos filtros o agregando dimensiones a las filas o columnas de Jpivot. | **7**. Pentaho gestiona los recursos necesarios para que mondrian ejecute la consulta mdx y open flash chart la grafique. |
| **6**. El usuario selecciona una o varias columnas para graficarlas | |


**Figura 25. Caso de Uso Real Consultar dwbiof 1**

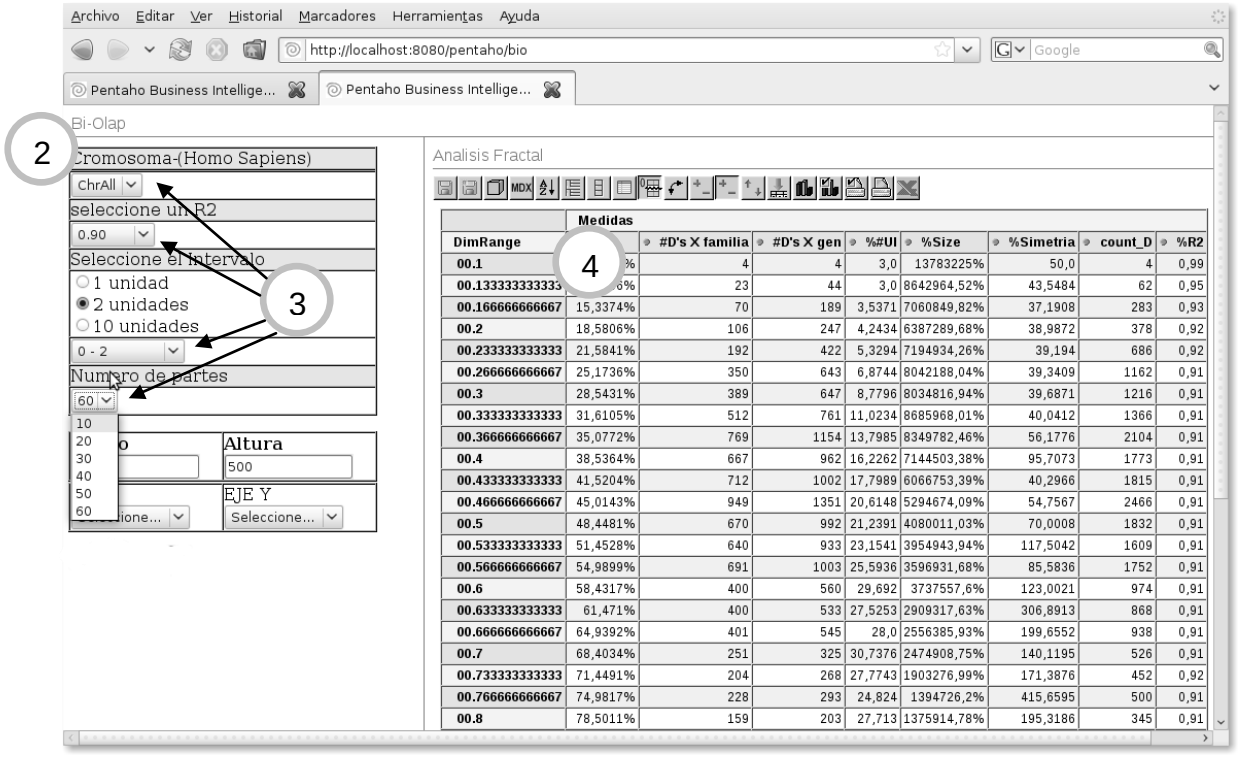



**Figura 26.Caso de Uso Real Consultar dwbiof 2**

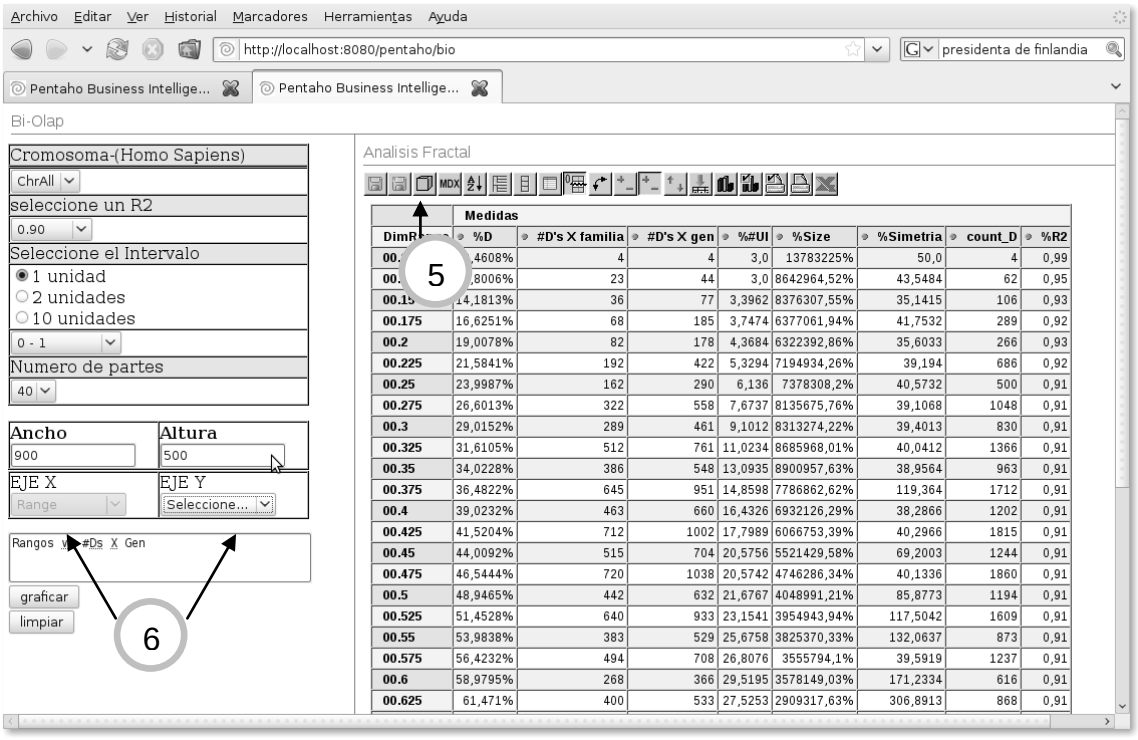



**Figura 27.Caso de Uso Real Consultar dwbiof 3**

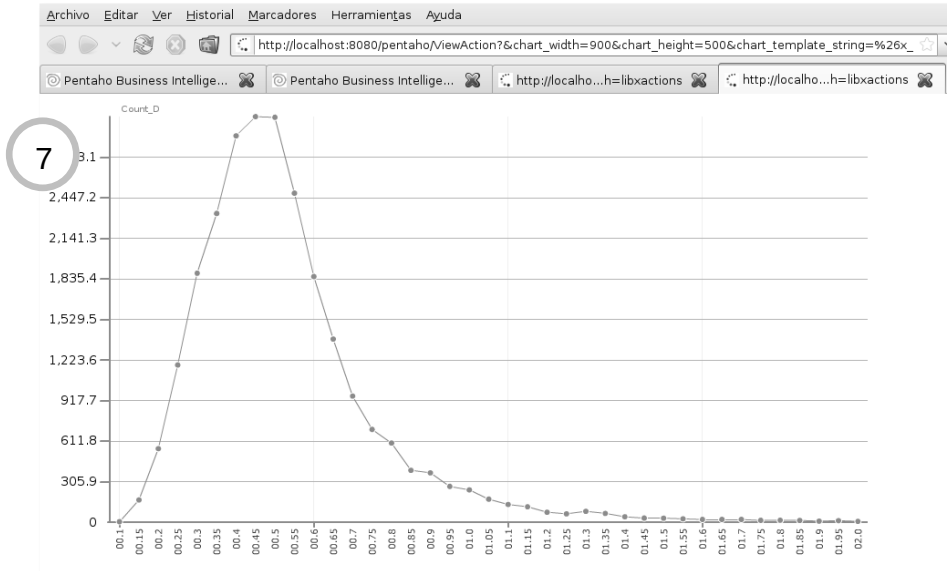



**Figura 28. Caso de Uso Real Consultar dwbiof 4**

En la Figura 28 podemos el resultado de la petición realizada por el usuario, graficar rangos contra cuenta de D.

**Tabla 20. Caso de uso para consultas gráficas**

| | |
|---|---|
| **Caso de uso:** | Consultar gráficas de análisis fractal para todos los cromosomas |
| **Actores:** | Usuario Registrado, Administrador |
| **Propósito:** | Hacer uso de los filtros para mostrar los datos deseados y realizar gráficas sobre el cubo con el servidor mondrian. |
| **Resumen:** | Cualquier de los dos usuarios tiene la posibilidad de ingresar dentro del modulo Dinámico OLAP y realizar consultas personalizadas sobre el cubo. |
| **Tipo:** | Primario. |
| **Curso normal de eventos** | |
| **Acción de los Actores** | **Respuesta del sistema** |
| **1.** Este caso de uso se inicia cuando un visitante quiere realizar gráficas personalizadas sobre el cubo multidimensional.<br><br>**3.** El usuario selecciona mediante el conjunto de filtros adicional de los datos desea.<br><br>**5.** El usuario genera consultas dinámicas seleccionando nuevos filtros o agregando dimensiones y medidas a las filas o columnas de Jpivot. | **2.** El sistema presenta un conjunto de filtros adicionales y la herramienta OLAP encargada de hacer consultas dinámicas.<br><br>**4.** El sistema obtiene los datos multidimensionales (Dimensiones y Medidas) determinadas por los filtros en el cubo. |

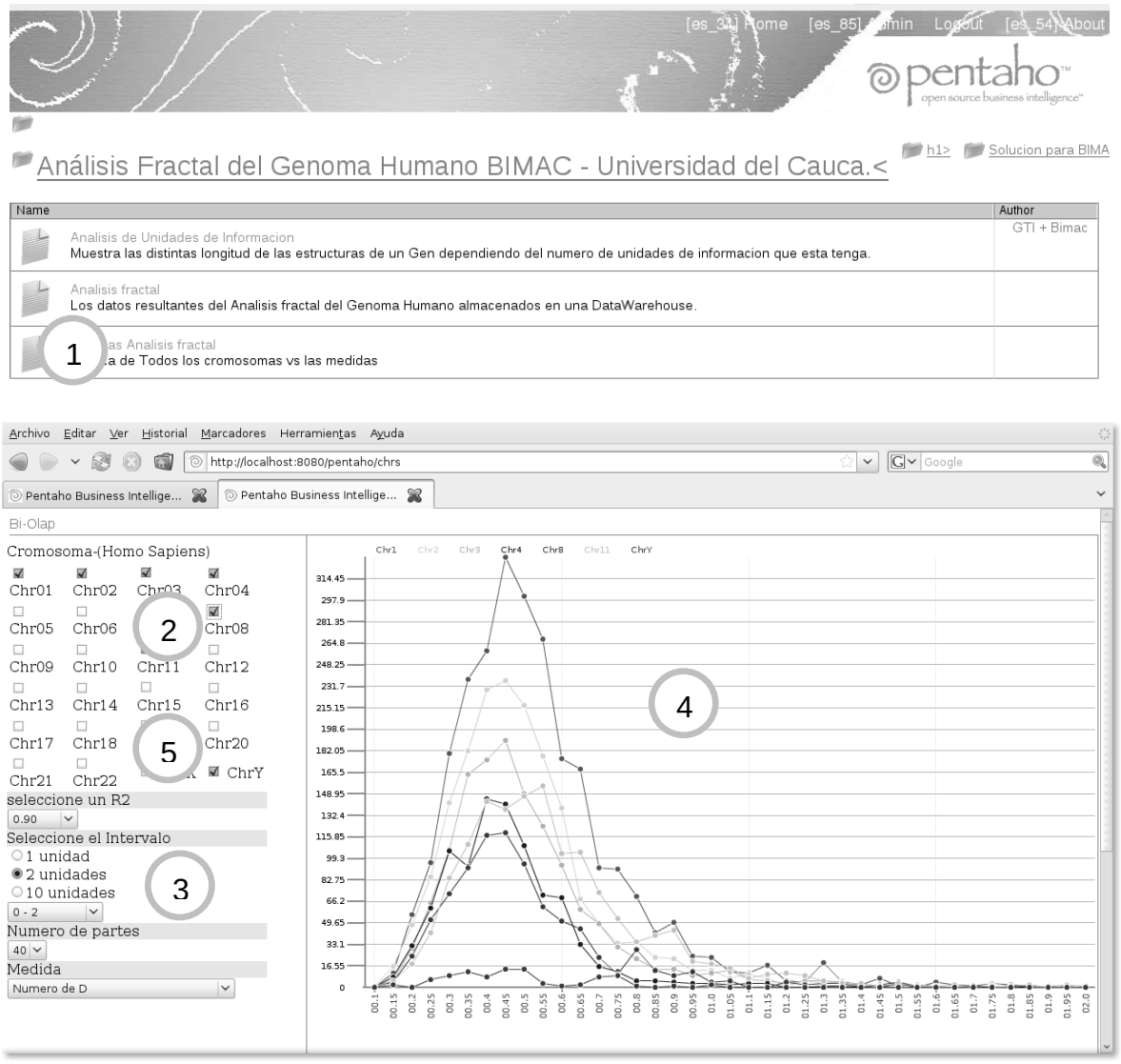



**Figura 29. Caso de Uso Real Consultar gráficas de análisis fractal para todos los cromosomas**

En la Figura 29 se puede observar una alternativa de visualización para el data mart Análisis fractal, donde los investigadores encuentran un conjunto de filtros que les permiten seleccionar un intervalo, dividirlo en un numero de partes de acuerdo a su criterio, dicha selección es graficada contra las diferentes medidas del data mart para todos de los cromosomas simultáneamente, teniendo también la posibilidad de agregar o quitar cromosomas a la grafica.

# CAPITULO V. DESCRIPCIÓN DE PROBLEMAS Y LA SOLUCIÓN ENCONTRADA

En este capítulo se describen los problemas que se presentaron durante la realización del proyecto y de igual manera se describe las estrategias de solución encontradas para estos.

La construcción de la solución presentó algunos problemas en diferentes áreas del proyecto, estos problemas serán desglosados a continuación:

## *5.1. Problemas Presentados en los Procesos ETL*

Los procesos de ETL, involucran extracción, limpieza y cargue de los datos que serán almacenados en el DW. Para este proyecto de grado se encontraron algunas consideraciones propias del proceso aplicado por los investigadores para llevar a cabo sus tareas y las fuentes en las cuales se encuentran los datos utilizados por dicha investigación.

Se abordaran los problemas que se tuvieron en el proceso de ETL dividiéndolo en las dos consideraciones antes descritas:

1. La naturaleza de la fuente de donde se extraen los datos para la investigación.

2. La naturaleza de la investigación y los procesos propios de esta.

Por cada una de las consideraciones se explicarán las decisiones tomadas teniendo en cuenta los análisis realizados con base a la revisión bibliográfica, las experiencias de otros proyectos a fines, de los investigadores del grupo BIMAC y de los desarrolladores de este proyecto.

### 5.1.1. La naturaleza de la fuente de donde se extraen los datos para la investigación.

En el proceso de ETL realizado para este proyecto se encontraron algunos errores asociados a los artefactos expuestos en el marco conceptual, localizados esencialmente en la cabecera y las características de los registros. Los errores que se presentaron en el momento de realizar el cargue de los datos a la bodega de datos fueron los siguientes:

1. La herramienta utilizada para el proceso de ETL, detectó 22 datos erróneos de 29.200 datos, en el proceso de cargue de las descripciones de los genes a la dimensión gen, 2 de estos datos están clasificados como abreviaturas y los 20 restantes se clasificaron como homónimos.

   El error presentado por estos 22 datos, es un error de integridad referencial, el cual se presentó porque el archivo de origen, con el cual se debía realizar la actualización presentaba solo dos columnas, la primera contenía el nombre del gen y la segunda la descripción del gen, por lo tanto solo se contaba con un campo para realizar la comparación necesaria la cual permitiría asignar la descripción a cada gen, pero este campo que es el nombre del gen, puede encontrarse varias veces en la tabla destino, debido a que un gen puede tener el mismo nombre, por eso no se logra obtener un identificador único al cual asignarle la descripción.

   Para darle solución a este problema, se reconstruyó el archivo que contenía las descripciones adicionándole el identificador del gen, el cual es único y realizando la comparación por el identificador.

2. Otro error encontrado en los datos fue la violación de la estructura de la secuencia, que como ya se explicó son entidades que no corresponden a la estructura lógica de una secuencia.

Estos registros fueron eliminados porque son registros dañados.

Por lo tanto, algunos artefactos enunciados pueden pasar desapercibidos, debido a que las herramientas no los detecten como los sinónimos y otros artefactos que no se presentaron en el cargue de los datos actuales.

### 5.1.2. La naturaleza de la investigación y los procesos propios de esta.

Para poblar la bodega de datos no se utilizó como fuente de datos el GENBANK, es decir el GENBANK es el repositorio desde el cual los investigadores obtienen los datos, los procesan de acuerdo a sus necesidades y los resultados son almacenados en archivos planos, los cuales son convierten en la fuente de datos de la bodega de datos.

Inicialmente se estudio la posibilidad de integrar el proceso con el cual se obtienen los resultados de la investigación al proceso de la bodega de datos, pero es algo que no hace parte de la solución que se planteó; ya que la bodega de datos es una alternativa por medio de la cual una organización puede centralizar la información proveniente de los procesos propios de esta. Además al estudiar el proceso que los investigadores realizan sobre los datos del GENBANK, se encontró que por la naturaleza de esta investigación dentro de los procesos se encuentra la aplicación de técnicas de minería de datos, procesos que se salen del alcance del proceso de ETL.

Por lo tanto como se mencionó anteriormente la fuente de datos de la bodega de datos fueron archivos planos que generaban los investigadores. Sin embargo, la fuente de datos (GENBANK) utilizada para generar los datos de la investigación fue de gran importancia en el proyecto. Gracias a esta se pudo comprender la naturaleza de algunos errores que se presentaron en el proceso de creación de la bodega y los cuales fueron explicados en el punto anterior (La naturaleza de la fuente).

Otro aspecto importante es que el grupo de investigadores no realiza una limpieza exhaustiva de los datos, debido a que consideran que el porcentaje de los datos que presentan error o ruido, es insignificante en la producción de los resultados, por ello en el proceso de construcción de la bodega de datos, se presentaron errores en los datos en el momento de poblar la bodega los cuales se explicaron en el punto anterior.

También es necesario considerar que los archivos obtenidos como resultado de la investigación, no fueron pensados como un suministro de un sistema como la bodega de datos, lo cual contribuyó a que se tuvieran algunas dificultades en el proceso de ETL, debido a la falta de campos únicos en algunos de los archivos, que permitieran realizar adecuadamente el cargue de los datos.

## *5.2. Problemas Presentados en el Modelado de la Solución*

### 5.2.1. Dimensión Rango (Tabla de Hechos Análisis Fractal)

Para los investigadores del grupo BIMAC, es relevante generar intervalos y rangos de análisis dentro de los intervalos de acuerdo a una variable establecida, la idea inicial contempló la posibilidad de que los usuarios finales pudieran definir los límites de los intervalos y posteriormente definir la cantidad de particiones, llamadas rangos, en las cuales dividirían dicho intervalo. Esta alternativa se contempló por petición del sponsor del proyecto.

Al querer satisfacer esta necesidad se tuvieron algunas limitantes debido a que a nivel teórico se encuentra parte de la solución a este problema de modelado, pero en la práctica las herramientas no soportan la teoría expuesta. Buscando dar claridad a la problemática se expondran las soluciones que se contemplaron y las razones por las cuales no se logro dar la solución.

- La primera solución que se planteó fue construir la consulta a partir de sentencias MDX, pero no se encontró una sentencia o un conjunto de sentencias que permitieran realizar las actividades necesarias para construir los intervalos y dividirlos en rangos

- La otra alternativa de solución estudiada fue una solución planteada a un problema similar en el libro "THE DATA WAREHOUSE TOOLKIT'', de la autoría de Ralph Kimball, en el libro el señor Kimball, plantea un reporte en el cual el usuario final puede crear unos rangos de valores, tal como en los balances de cuentas, en busca de mayor claridad, se mostrará la tabla de ejemplo expuesta en el libro y el modelo con el cual el señor Kimball pretende dar solución a este tipo de consultas:

| Balance Range | Number of Accounts | Total of Balances |
|---|---|---|
| 0-1,000 | 45,678 | $10,222,543 |
| 1,001-2,000 | 36,788 | $45,777,216 |
| 2,001-5,000 | 11,775 | $31,553,884 |
| 5,001-10,000 | 2,566 | $22,438,287 |
| 10,001 and up | 477 | $8,336,728 |

**Figura 30. Ejemplo reporte por intervalos**

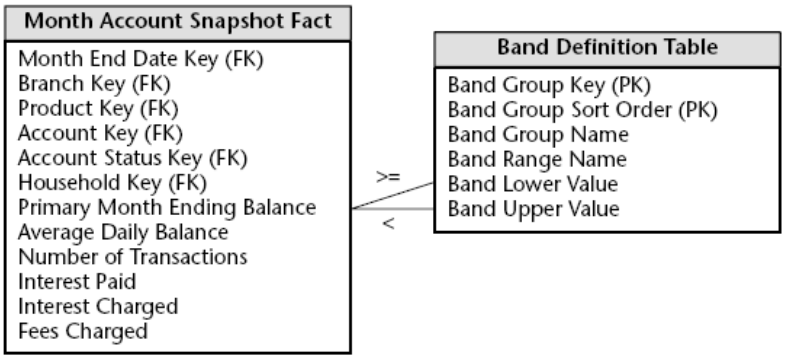



**Figura 31. Definición de la dimensión band**

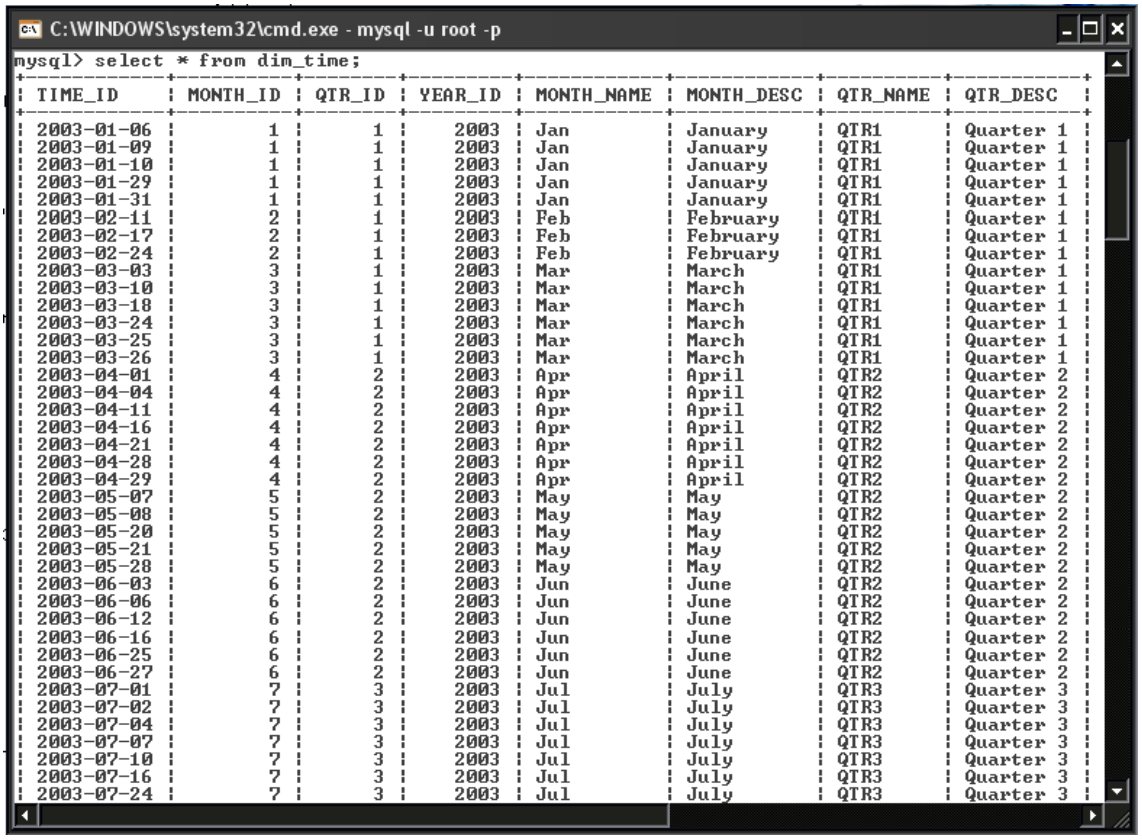
En la solución, se puede ver una particularidad, la tabla que permite definir los rangos, no está asociada a la tabla de hechos, lo cual se puede ver por la ausencia de una llave foránea en la tabla de hechos, lo que involucra realizar un cruce entre la tabla BAND DEFINITION y el hecho PRIMARY MONTH ENDING BALANCE, mediante el uso de un par de cruces de menor que y mayor que.

El autor sugiere que la solución planteada es muy limitada, debido a que el cruce con la tabla BAND DEFINITION es poco convencional y que esta no es una buena base para encontrar una restricción limitante, teniendo como resultado tan solo una agrupación de ***n*** balances. El autor también sugiere que para que esta solución pueda ser aplicada se debería crear un índice sobre el hecho PRIMARY MONTH ENDING BALANCE, encontrando una limitante en

los sistemas de gestión de bases de datos (DBMS), esto puede ser una enorme mejora para estos, ordenando y comprimiendo individualmente cada hecho de una forma eficiente.

No se logró encontrar una solución dinámica posible, la cual pudiera ser implementada en el modelado o en las consultas, por lo cual se analizaron los requerimientos de los usuarios en conjunto con los reportes generados por estos. Este análisis permitió establecer un aspecto desconocido y que posteriormente fue confirmado por los investigadores; los intervalos y rangos de análisis utilizados para la investigación, no son dados al azar, si no que por el contrario han sido seleccionados bajo criterios propios de la investigación, lo que no implica que se puedan considerar otros intervalos y rangos de análisis, pero estos intervalos y rangos definidos son con los que típicamente se trabajan y son los que se desean mostrar a los usuarios, recordando que el objetivo del proyecto es permitir crear una herramienta por medio de la cual los resultados de la investigación sean divulgados.

Teniendo en cuenta esto se inicio la búsqueda de una solución que permitiera presentar la información, considerando los intervalos y rangos planteados por los investigadores. La solución encontrada se planteó a partir de dos casos de modelado presentados por el autor seguido, el primer caso es el reporte por intervalos, expuesto anteriormente y el segundo caso es el de la definición de la tabla física de la dimensión fecha que será explicado a continuación:



C:\WINDOWS\system32\cmd.exe - mysql -u root -p

```
mysql> select * from dim_time;
+------------+----------+--------+---------+------------+------------+----------+-----------+
| TIME_ID    | MONTH_ID | QTR_ID | YEAR_ID | MONTH_NAME | MONTH_DESC | QTR_NAME | QTR_DESC  |
+------------+----------+--------+---------+------------+------------+----------+-----------+
| 2003-01-06 |        1 |      1 |    2003 | Jan        | January    | QTR1     | Quarter 1 |
| 2003-01-09 |        1 |      1 |    2003 | Jan        | January    | QTR1     | Quarter 1 |
| 2003-01-10 |        1 |      1 |    2003 | Jan        | January    | QTR1     | Quarter 1 |
| 2003-01-29 |        1 |      1 |    2003 | Jan        | January    | QTR1     | Quarter 1 |
| 2003-01-31 |        1 |      1 |    2003 | Jan        | January    | QTR1     | Quarter 1 |
| 2003-02-11 |        2 |      1 |    2003 | Feb        | February   | QTR1     | Quarter 1 |
| 2003-02-17 |        2 |      1 |    2003 | Feb        | February   | QTR1     | Quarter 1 |
| 2003-02-24 |        2 |      1 |    2003 | Feb        | February   | QTR1     | Quarter 1 |
| 2003-03-03 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-03-10 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-03-18 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-03-24 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-03-25 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-03-26 |        3 |      1 |    2003 | Mar        | March      | QTR1     | Quarter 1 |
| 2003-04-01 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-04 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-11 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-16 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-21 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-28 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-04-29 |        4 |      2 |    2003 | Apr        | April      | QTR2     | Quarter 2 |
| 2003-05-07 |        5 |      2 |    2003 | May        | May        | QTR2     | Quarter 2 |
| 2003-05-08 |        5 |      2 |    2003 | May        | May        | QTR2     | Quarter 2 |
| 2003-05-20 |        5 |      2 |    2003 | May        | May        | QTR2     | Quarter 2 |
| 2003-05-21 |        5 |      2 |    2003 | May        | May        | QTR2     | Quarter 2 |
| 2003-05-28 |        5 |      2 |    2003 | May        | May        | QTR2     | Quarter 2 |
| 2003-06-03 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-06-06 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-06-12 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-06-16 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-06-25 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-06-27 |        6 |      2 |    2003 | Jun        | June       | QTR2     | Quarter 2 |
| 2003-07-01 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-02 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-04 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-07 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-10 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-16 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
| 2003-07-24 |        7 |      3 |    2003 | Jul        | July       | QTR3     | Quarter 3 |
```

**Figura 32. Ejemplo de tabla física de la dimensión tiempo**

En la Figura 32 se puede apreciar la estructura de la dimensión tiempo, lo que es interesante de esta tabla para este proyecto, es la manera como se definen las diferentes agrupaciones que pueden existir para realizar los análisis, en este caso se puede agrupar la información por meses,

por cuartiles o por años, se puede apreciar que para llevar a cabo estas agrupaciones, la tabla tiene unos campos identificadores, los cuales le permiten agrupar la información según sea requerido, gracias a que asigna a cada fecha un conjunto de identificadores que permiten saber a que mes, que cuartil y que año pertenece una fecha, de tal manera que si alguien desea consultar los meses, lo que debe hacerse es agrupar por el identificador de mes, teniendo en cuenta que todos los registros que contengan un mismo identificador es porque pertenecen al mismo mes.

Teniendo la explicación de estos dos casos de modelado explicaremos lo requerido por el proyecto y la forma cómo estos dos casos aportan a la solución que se planteó. A continuación se analizará un intervalo de los propuestos por los investigadores:

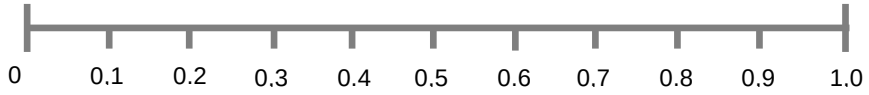



**Figura 33. Intervalos de 0 – 1 divididos en 10**

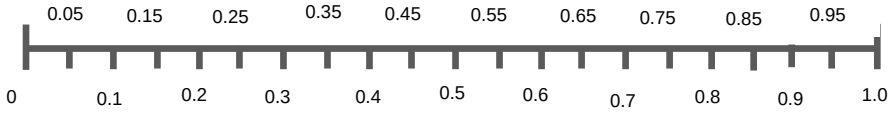



**Figura 34. Intervalos de 0 – 1 divididos en 20**

La Figura 33 y la Figura 34, son una muestra del intervalo y las divisiones que pueden encontrarse en el problema planteado, se puede observar el intervalo con diez particiones contiene al que tiene veinte, igual que un cuartil contiene meses, por lo tanto los intervalos serán definidos de forma similar que como se hace en la tabla física tiempo. Las divisiones pueden ser definidas como la tabla band porque esta nos permite definir el valor inferior y superior para la división más pequeña en cada intervalo.

Unas consideraciones importantes a tener en cuenta son:

- Cuando se trata de la tabla tiempo se sabe que un año tiene 365 días o 366 en caso de ser bisiesto lo que hace que siempre se tenga la misma cantidad de valores para la tabla, lo que hace que esta sea invariante, al igual que las agrupaciones por cuartil y mes. Esto no ocurre cuando se definen intervalos debido a que los intervalos no tienen todos igual cantidad de particiones, por lo tanto las agrupaciones que se podían hacer sobre estos intervalos eran variables.

- Se pensó en los intervalos como años fiscales para poder solucionar la variabilidad antes expuesta.

- Cada intervalo se creó con sus propias divisiones para poder llevar a cabo las agrupaciones necesarias.

## *5.3. Problemas Presentados en la Presentación de los datos*

Se presentaron algunos inconvenientes en el momento de la presentación de los datos, haciendo uso de las herramientas OLAP, teniendo en cuenta que las consultas que solicitaban el grupo de investigadores, eran consultas especializadas, atípicas a las soluciones conocidas en los DW. Los

problemas presentados en la adaptación de la herramienta OLAP, a la solución requerida fueron los siguientes:

### 5.3.1. Problemas Relacionados con los Filtros

Los filtros requeridos por los investigadores para el Data Mart análisis fractal sobrepasaban las capacidades de las herramientas OLAP, analizadas hasta el momento con las cuales se deseaba satisfacer las necesidades de los investigadores, los estándares de funcionalidad típicos de las herramientas OLAP analizadas no contemplan la capacidad de mostrar automáticamente valores que estén por encima o por debajo de un valor definido por el usuario y mucho menos la clasificación en rangos de las medidas, seleccionando un intervalo y el numero de partes en las que se quiere dividir dicho intervalo.

Para dar solución a este problema debió adicionarse funcionalidad a la aplicación creando los filtros necesarios, construyendo una página que contenía las consultas MDX, necesarias para realizar los filtros y que permitía que estos pudieran ser parametrizados.

### 5.3.2. Problemas Relacionados con el Graficador

- Las gráficas solicitadas en el proyecto requerían de seleccionar las columnas que se deseaban graficar y el eje para cada una de las variables, sin embargo, los gráficadores de las herramientas OLAP no tienen esas opciones tan avanzadas, las opciones gráficas se limitan a generar histogramas, tortas, líneas de relación, entre otros gráficos. Todas estas gráficas muestran la relación de dimensiones contra medidas, lo cual no contribuía en la solución a ser planteada debido a que algunas de las solicitudes de la aplicación del proyecto requerían gráficas de relaciones de medidas contra medidas.

  La solución a este problema fue adicionar un graficador ajeno a la herramienta usada y que daba la libertad de seleccionar las columnas que se deseaban graficar y en que ejes irían las variables, esto en conjunto con las consultas MDX parametrizadas en los filtros, dieron solución a las gráficas requeridas.

- Se encontró que el graficador externo usado, sólo reconocía las columnas que tenían medidas, por lo cual no permitía graficar dimensiones contra medidas.

  Lo que se realizó para darle solución al problema fue crear una medida calculada de un atributo de la dimensión de interés y graficarla, teniendo en cuenta de no mostrarla en el visualizador de los datos

- Las dimensiones son típicamente cadenas de caracteres, teniendo en cuenta que lo esperado son descripciones y nombres, en el caso de este proyecto se encontró con que los atributos de las dimensiones eran numéricos y las herramientas OLAP no esperan este tipo de valores, por lo que son convertidos automáticamente en cadenas de caracteres para ser visualizados, el problema es que al realizar la conversión el visor les asignaba un orden alfabético, degenerando el orden de la presentación de los reportes.

  La solución dada consistió en adicionar un 0 a la izquierda de todos los valores, para poder conservar el orden.

# CAPITULO VI. RECOMENDACIONES PARA EL DISEÑO DE DATA WAREHOUSE

Contiene la descripción y/o adaptación de diferentes conceptos teóricos utilizados en el diseño e implementación de la DW en el campo de la bioinformática.

Para dar apertura a este capítulo es indispensable aclarar lo que es una recomendación, para tal fin se tomara una definición propuesta por el Dr. Carlos Sabino, en su libro llamado, Como hacer una tesis [Carlo94], en el cual el autor dice que las recomendaciones "suponen que es posible extraer líneas prácticas de conducta sobre la base del desarrollo analítico que se haya hecho previamente. Para poder establecerlas es preciso que los conocimientos obtenidos en la investigación sean examinados a la luz de ciertas metas o valores que posee el autor y que son, necesariamente, subjetivos. Son por lo tanto siempre relativas al punto de vista adoptado y a los fines que se persiguen en relación al problema tratado." [Carlo94]. Por lo tanto las recomendaciones son el reflejo de la experiencia vivida en el proyecto y que tienen como finalidad contribuir en el esclarecimiento de dudas de aquellos a quienes puedan llegar a tener experiencias similares a las vividas en este proyecto.

## *6.1. Recomendaciones en el diseño del DW.*

La construcción del DW que modelará las consultas propuestas por los investigadores del grupo BIMAC representó un reto de construcción debido a la naturaleza de los datos, por lo cual se presentaran recomendaciones respecto al cargue de datos biológicos y al modelado de datos biológicos.

### 6.1.1. Cargue de datos Biológicos

El cargue de datos que provienen de repositorios biológicos debe ser analizado desde otra perspectiva un poco diferente a la planteada para el área de los negocios esto a razón de que las transformaciones propuestas por las herramientas para el proceso ETL analizadas, no suplen las necesidades en cuanto un proyecto se ve enfrentado a datos como los tratados en este, debido a qué estas herramientas están enfocadas en homogeneizar los datos provenientes de diversas fuentes y no están pensadas para procesar archivos de texto que no tengan un formato en columnas como es el caso de las bases de datos biológicas públicas.

Este proyecto no contempló muchas de las transformaciones propuestas por Kimball para el proceso de ETL, debido a que los datos provenían de una sola fuente (archivos planos generado por los investigadores), y no se presentaban las inconsistencias usuales del cargue de datos de una Bodega de datos como son el manejo de los nombres de los campos o en cuanto al manejo de los tipos de datos; sin embargo las bases de datos biológicas se enfrentan a otra serie de problemas debido a qué presentan inconsistencias referentes al almacenamiento de los datos en las bases de datos públicas, de las cuales los investigadores del grupo BIMAC, extraen los datos para su investigación. Estas bases de datos tienen, se podría decir, ruido en los datos, que los investigadores los denominan como artefactos y se hace alusión a ellos en el capítulo que contempla los problemas que se presentaron en este proyecto, por el momento es importante enfocar el tema en cuáles deberían o podrían ser las alternativas recomendadas que permitan realizar un proceso de cargue exitoso, recordando que es un proceso que se repite en toda la vida del DW y del que depende que el DW, contenga información acertada para los usuarios.

Teniendo en cuenta lo planteado en el párrafo anterior, los autores de este proyecto recomiendan a los grupos de investigación, los cuales son los que extraen los datos directamente de los repositorios públicos realizar la siguiente actividad:

- Realizar una revisión y limpieza de los datos antes de ser usados teniendo en cuenta el Framework de limpieza de datos biológicos provenientes de repositorios públicos, el cual será presentado a continuación y a partir de este se estructurara una posible solución de acuerdo a las necesidades de este proyecto y que puede ser de ayuda a proyectos con características similares. Este Framework fue propuesto en el taller de discusión, acerca de las bases de datos biológicas (Workshop on Database Issues in Biological Databases (DBiBD)) [15]

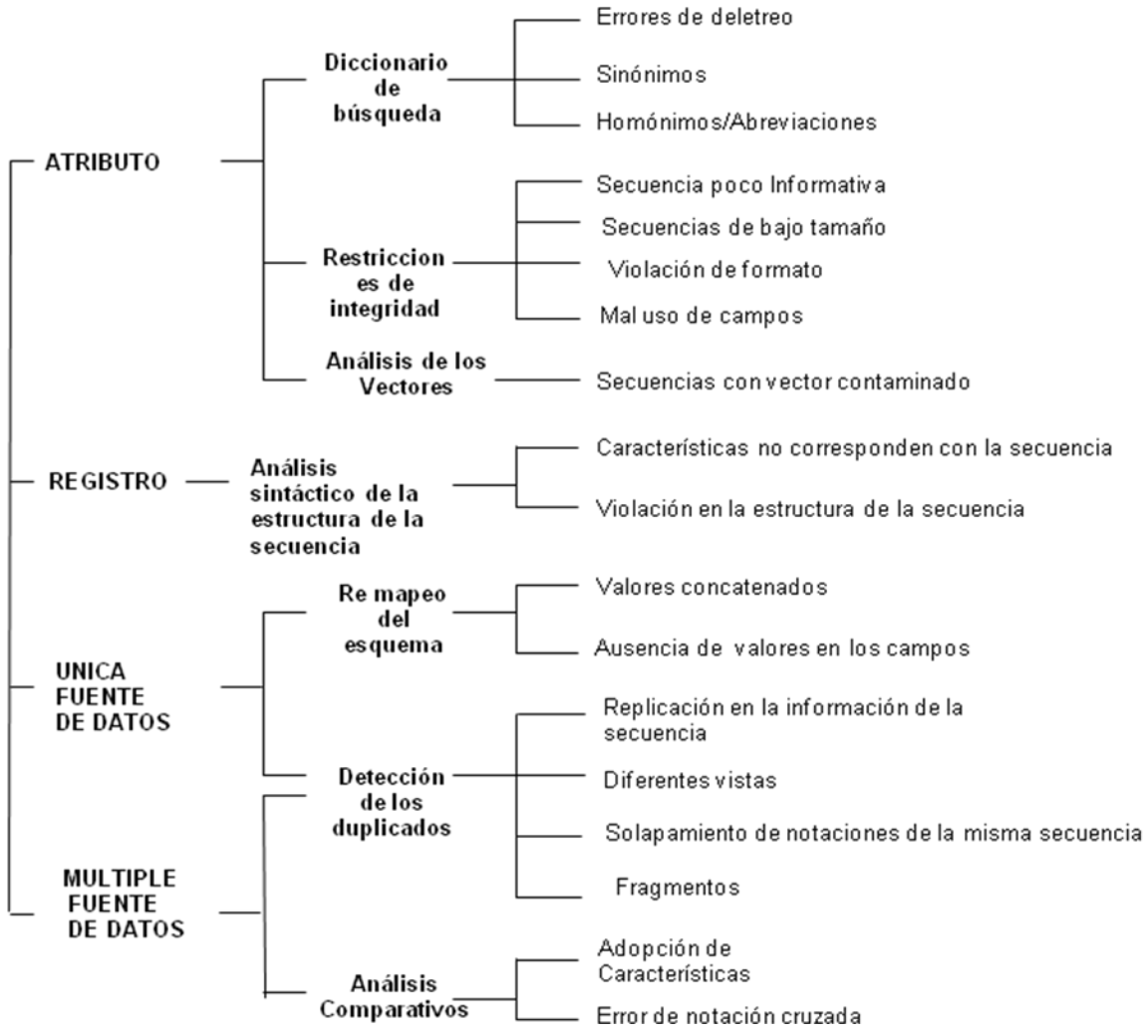



**Figura 35. Framework de limpieza de datos biológicos**

Se destaca que el Framework expuesto agrupa los posibles artefactos que pudiesen presentarse y les proporciona una solución, es importante también comentar que el Framework contempla diferentes posibilidades a la hora de clasificar los artefactos, contemplando una o varias posibles

soluciones dependiendo si el artefacto se encuentra en el atributo, en el registro como tal o si los datos tomados provienen de una o a múltiples fuentes de datos.

Para este proyecto se presentaron dos situaciones, los resultados obtenidos por los investigadores son calculados a partir de los datos provenientes de un único repositorio de datos, el GENBANK, y se encontraron artefactos en los atributos, tales como datos con sinónimos, homónimos o abreviaturas, lo cual dificultó el cargue de los datos. Este aspecto está contemplado en Framework en la parte del "Diccionario de datos" de "ATRIBUTO". Por lo tanto, se recomienda tratar de implementar las acciones propuestas en el Framework, de esta manera los investigadores contarán con información más limpia y más exacta para sus labores investigativas. Sería interesante que se considere la posibilidad de crear una herramienta ETL, la cual considere estos aspectos.

En caso de tener datos resultantes de las investigaciones, que fueron calculados a partir de los datos provenientes de repositorios públicos y que no fueron corregidos de acuerdo a los criterios antes expuestos, los autores recomiendan que por lo menos, los archivos resultantes contengan dos o más campos de comparación, que puedan ser usados en el momento del cargue, teniendo en cuenta que si se dejan campos de comparación únicos, los cuales puedan tener sinónimos, homónimos o abreviaturas, es posible que el proceso de cargue de los datos hacia el DW genere errores. Esta opción fue la que se llevo a cabo para realizar el cargue de los datos de la investigación análisis multifractal del genoma humano, solo en dos casos de cargue no fue posible tener más de dos campos de comparación, por lo cual se debió llevar a cabo una revisión manual, para poder ser cargados la totalidad de los datos.

### 6.1.2. Modelado de Datos Biológicos

**Dimensión Rango**

En este proyecto se encontró una necesidad muy particular por parte de los investigadores y que es posible se encuentre en otras investigaciones, la necesidad de construir intervalos y poder fraccionar estos intervalos en rangos de análisis, permitiéndole al usuario clasificar los datos respecto a una variable de interés. Lo anterior puede verse como un intervalo de intervalos.

Lo que se recomienda en caso de tener un escenario similar es realizar una fusión entre dos conceptos de diseño, uno, el de cómo definir una tabla dinámica para generar intervalos de valores arbitrarios y el otro concepto, la definición de la dimensión tiempo, ambos casos de modelado definidos por Ralph Kimball en su libro, Guía completa para el modelado dimensional [20].

Esta fusión debe realizarse al no encontrarse un caso de modelado que se ajuste a las necesidades del modelo a construir para el proyecto; esta fusión aporta a la solución que se desea y que no puede aplicarse de manera excluyente, es decir, la aplicación individual de cada uno de ellos no logra satisfacer el objetivo de lo que se desea modelar. Por lo tanto, se recomienda en caso de tener un desafío de modelado similar incluir los campos pertenecientes a la tabla band para definir, las divisiones de los intervalos, y generar identificadores y nombres de los identificadores como se generan en la dimensión tiempo para reconocer y agrupar la variable de interés, por intervalos y por divisiones.

Además se recomienda seguir los siguientes pasos para crear la tabla de la dimensión rango:

- Revisar cuales son las agrupaciones y divisiones que se consideran relevantes para la investigación y listarlas, por intervalos de análisis y divisiones requeridas para cada uno de los intervalos.

- Determinar cuál es la división más pequeña que se pueda llegar a tener, o lo que es igual el máximo número de partes en que será dividido el intervalo más pequeño que se pueda tener en los intervalos de análisis. Esto determina el incremento más pequeño que pueda llegar a tenerse.

- Se generan un conjunto de valores de la siguiente manera, se inicia desde el menor valor posible (en el caso de este proyecto es 0), incrementando de acuerdo al valor resultante de la división que se realizó con anterioridad, para obtener así el siguiente valor hasta llegar máximo valor requerido (en el caso de este proyecto es 50).

  En resumen al generar este conjunto de datos lo que obtiene es el intervalo más grande que se pueda llegar a tener dividido en la menor fracción posible de acuerdo a lo requerido. Estos valores son los que agrupados de diferentes formas permiten definir los intervalos y rangos de análisis.

- Teniendo esta información se construyen las columnas de la dimensión teniendo en cuenta definir los valores superior e inferior de cada división y definir los identificadores que permitan realizar las diferentes agrupaciones.

## *6.2. Recomendaciones en la Presentación de los datos*

Las herramientas existentes en el mercado para la presentación de los datos, no soportan los análisis, las consultas y los gráficos requeridos para este tipo de proyectos. Por lo cual las recomendaciones que se generan en este punto tienen que ver con adicionar funcionalidad específica a las herramientas OLAP como se muestra a continuación:

- Las herramientas de presentación deben contemplar la posibilidad de realizar el manejo de valores numéricos en las dimensiones, ya que en estos momentos las herramientas existentes en el mercado sólo permiten cadenas de carácteres. Esto solucionaría los problemas presentados en el caso de tener datos numéricos, dentro de las dimensiones, con un orden específico, ya que las herramientas actualmente toman los valores que se encuentran en las dimensiones como cadenas de caracteres y al momento de mostrarlos, los ordena como un vector de cadenas de caracteres. Una solución inmediata al no contar con la posibilidad planteada en este párrafo es que durante la preparación de los datos a los valores numéricos que se necesite darles un campo equivalente en cadena de caracteres en caso de necesitar estos valores numéricos en un orden de presentación numérico.

- Las herramientas OLAP, deberían considerar ampliar el rango de posibilidades en cuanto a los análisis gráficos, permitiendo cosas tales como: graficar medidas contra medidas y dar flexibilidad y facilidad a los usuarios de poder seleccionar las columnas que desea graficar. Esto permite dar solución a las necesidades gráficas de este tipo de proyectos.

- Otro aspecto importante en este tipo de proyecto son los análisis estadísticos, para lo cual sería deseable encontrar en las herramientas OLAP una mayor cobertura de funciones estadísticas para aplicar sobre los datos de la bodega de datos en el momento de la consulta.

# CAPÍTULO VII: CONCLUSIONES Y TRABAJO FUTURO

- En el modelado dimensional propuesto por Kimball, el cuál guió la construcción del DW, no se contempla la generación de reportes que involucren rangos de análisis, lo cual implicó, una adaptación que permitiera cumplir con lo requerido por el proyecto, explorando nuevas alternativas de aplicación del modelado usado.

- Las herramientas OLAP analizadas para la construcción del prototipo, se encuentran diseñadas para dar solución a problemas en el área de los negocios, limitando el uso que pueda darse a otras áreas de aplicación, sin embargo en este proyecto de grado, se logró adaptar un conjunto de herramientas las cuales dan solución a lo requerido y amplían el campo de acción de estas tecnologías.

- Se deben realizar análisis previos de los datos en pro de mitigar el impacto que pueda llegar a tenerse en caso que los datos tengan inconsistencias o artefactos como es comúnmente llamado el ruido encontrado en las bases de datos biológicas, es importante destacar que para el equipo de investigadores el ruido encontrado en los datos es un porcentaje insignificante, pero para las bodegas de datos las inconsistencias de los datos pueden generar errores que impidan poblar la bodega exitosamente o más aún que estos datos erróneos generen resultados indeseables o no válidos en el momento de las consultas.

- Las Bodegas de datos pueden ser una buena alternativa para el almacenamiento y consulta de los datos de investigaciones biológicas, pero para poder utilizar todas las características de flexibilidad y dinamismo de las consultas, es necesario que las herramientas provean funcionalidad específica para las particularidades de este tipo de investigaciones.

- Implementar soluciones de DW, en ambientes ajenos para los que fue concebida esta tecnología, requiere un compromiso por parte de la organización, es necesario que la organización afronte costos en tiempo, en patrocinio y en la capacitación que implique la aplicación y adaptación de la tecnología a un nuevo campo.

### Trabajos Futuros

- Ampliar el modelo actual añadiendo nuevos atributos y medidas al modelo actual, por ejemplo la inclusión de los atributos C, P y Q en la dimensión cromosoma que permitirá ubicar los genes con sus estructuras en un lugares determinado de los cromosomas y poder observar si existe alguna regularidad entre la posición de los genes, su función y su D.

- Añadir a la dimensión gen un atributo de sobre oncogenes que permita observar con ayuda de la clasificación del D y la anterior tarea si existe alguna regularidad.

- El cargue de más genomas.

- Un PreETL con los datos del grupo y posiblemente un automatización completa del cargue de datos.

- Realizar mejoras o implementar una herramienta OLAP que permita una mayor funcionalidad para trabajar con datos de este tipo.

- La adición de nuevas dimensiones y tablas de hechos que completen el modelo biológico, como Proteina, estructura de la proteína, Alus, repeats, la inclusión de la Secuencia del Gen, Genoma y todas las posibles tablas de hechos que se puedan

llegar a derivar. Claro está que la implementación y el análisis de los datamarts que se puedan generar en estas dimensiones y tablas de hechos, dependerán de los avances que puedan llegarse a presentar en el manejo de cadenas de caracteres en la herramienta OLAP.

# CAPÍTULO VIII: BIBLIOGRAFÍA

En este capítulo se presenta la bibliografía empleada para la realización del proyecto

[1] A. V. Kavitha, V. Brusic, J. L. Koh, BioDArt – Catalogue of Biological Data Artifact Examples, Research Publishing Services, 2006

[2] S. P. Shah, Y. Huang, T. Xu, M. Yuen, J. Ling, F. Ouellette, Atlas – a data warehouse for integrative bioinformatics, Página Web: http://www.biomedcentral.com/1471-2105/6/34, Último Acceso: 12 de Noviembre de 2008.

[3] C. Sabino, Como hacer una tesis. Ed. Panapo, Caracas, 1994.

[4] C. Imhoff, N. Galemmo, J. G. Geiger, Mastering Data Warehouse Desing, Relational and Dimensional Techniques, First Edition, Wiley Publishing 2003.

[5] M. A. Dasi, Biologia molecular, Aspectos Básicos de la Biologia Molecular, Página Web: http://www.hsa.es/org/dmedica/centrales/ap/docs/biomol/index.htm, Último Acceso: 28 de Agosto de 2007.

[6] ELSI Research Program, The Ethical, Legal and Social Implications (ELSI) Research Program, Página Web: http://www.genome.gov/10001618, Último Acceso: 28 de Agosto de 2007.

[7] EMBnet Colombia, Centro de Bioinformática - Instituto de Biotecnología ©2005 Universidad Nacional de Colombia, Página Web: http://www.co.embnet.org/, Último Acceso: 20 de Agosto de 2007.

[8] Ethical, Legal, and Social Issues Research, Human Genome Project information, Página Web:http://www.ornl.gov/sci/techresources/Human_Genome/research/elsi.shtml, Último Acceso: 28 de Agosto de 2007.

[9] F. G. de la Fuente, M. Gil Estallo. Los sistemas de información en la sociedad del conocimiento, ESIC Editorial. 2004

[10] GenBank Overview, Pagina Web: http://www.ncbi.nlm.nih.gov/Genbank/, Último Acceso: 28 de Agosto de 2007.

[11] Genómica funcional, Página Web: http://www.solociencia.com/biologia/bioinformatica-genomica-funcional.htm, Último Acceso: 28 de Agosto de 2007.

[12] Glossary of genetic terms from the DOE Human Genome Program, Página Web: http://www.ornl.gov/sci/techresources/Human_Genome/glossary/, Último Acceso: 28 de Agosto de 2007.

[13] J.H Moore, M.G. Chang. "Design of Decision Support Systems." *Data Base*, Vol.12, Nos.1 and 2, 1980.

[14] J. Mundy, W. Thornthwaite, R. Kimball. The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset, Wiley Publishing, Inc., Indianápolis, Indiana, 2006

[15] J.L.Y. Koh, M. L. Lee, V. Brusic. Workshop on Database Issues in Biological Databases (DBiBD), 2005.

[16] Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources Página Web: http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html, Último Acceso: 28 de Agosto de 2007.

[17] M. Fischer, Q. K. Thai, M. Grieb, J. Pleiss. DWARF – a data warehouse system for analyzing protein families.Institute of Technical Biochemistry, University of Stuttgart Germany. Página Web: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1647292, Último Acceso: 11 de Septiembre de 2007.

[18] P. G. Keen, Decision support systems: an organizational perspective. Reading, Mass, Addison-Wesley, 1978.

[19] R. Kimball, L. Reeves, M. Ross, W. Thornthwaite, The data Warehouse Lifecycle Toolkit, Wiley Publishing, 1998.

[20] R. Kimball, M. Ross, The Data Warehouse Toolkit, the Complete Guide to Dimensional Modeling. Second Edition, 2002

[21] R. H. Sprague, E. D. Carlson. Building effective decision support systems. Englewood Cliffs, N.J, Prentice-Hall, 1982.

[22] S. L. Alter. Decision support systems: current practice and continuing challenges. Reading, Mass, Addison-Wesley Pub, 1980.

[23] T. Kirsten, H. H. Do, E. Rahm, A Data Warehouse for Multidimensional Gene Expression Analysis, published by the Interdisciplinary Centre for Bioinformatics, 2004, Página Web: http://www.izbi.uni-leipzig.de/izbi/Working%20Paper/2004/01_geware.pdf, Último Acceso: 12 de Noviembre de 2008.

[24] Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación Página Web: http://eisc.univalle.edu.co/index.php?option=com_content&task=view&id=113, Último Acceso: 28 de Agosto de 2007.

[25] P. Vélez, Propuesta para la participación en la convocatoria nacional para el concurso de proyectos de investigación programa nacional de biotecnología, "Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información", universidad del cauca, Presentada 2004.

[26] What is the Human Genome Project? Human Genome Project information, Página Web:http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml, Último Acceso: 28 de Agosto de 2007.

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

[28] Benoît Mandelbrot, La Geometría Fractal de la Naturaleza, Tusquets, 1997

[29] Mazur, Glenn, “QFD Black Belt Notes”, Japan Bussiness Consultants, E.U., 2002