TECNOLÓGICO DE ESTUDIOS SUPERIOR DE TIANGUISTENCO

Ingeniería en Sistemas Computacionales

PROTOCOLO DE TESIS
DE LICENCIATURA

Nombre del alumno:	"Iber Alberto Conde Sánchez"
Tutor(a) de tesis:	Mtra. "Karol Baca López"
Tutor(a) de tesis:	Dr. "Jesús Espinal Enríquez"
Título de la tesis:	"*ESTRUCTURA DE DATOS EN MAPEO DE DATOS CLÍCOS EN CÁNCER DE MAMA DE Genomic Data Commons Data Portal.*"

Abstract

El cáncer de mama es la neoplasia más frecuente y con mayor tasa de mortalidad entre las mujeres en México. En este trabajo se revisan datos clínicos públicos de Cáncer de Mama del portal Genomic Data Commons (GDC) en el que es posible compartir información sobre estudios genómicos sobre el cáncer en apoyo de la medicina de precisión. Esto consiste en una construcción de una aplicación para la presentación de datos clínicos, bioespecíficos y pequeños volúmenes de datos moleculares, con herramientas específicas y seleccionadas para mayor uso de los formatos descargados del GDC como Gestor de Base de Datos (MongoDB) y lenguajes de programación como Java (NetBeans). Se hace la gestión de Base de Datos para la conexión con el lenguaje de programación y así obtener nuestra propio Software para facilitar el uso de Datos al Investigador.

Key words: Mapeo de Datos Clínicos, Cáncer de Mama , GDC, MongoDB, Java.

Índice

1 Introducción

El tema que se propone de tesis surge de la necesidad de facilitar y optimizar flujos de trabajo para estudios genómicos de cáncer. Y en particular, del cáncer de mama que es la primera causa de muerte por neoplasia maligna en la mujer. La incidencia aumenta con la edad; sin embargo, la relación entre la edad y la supervivencia de las pacientes con cáncer de mama no está bien definida. Se observa que las mujeres jóvenes con cáncer de mama tienen patrones biológicos de comportamiento más agresivos.

En el Instituto Nacional de Medicina Genómica, el Mapeo de Datos Clínicos involucra manejo de información y Datos Personales de un paciente, que se integra dentro de todo tipo de establecimiento para la atención médica, ya sea público, social o privado, el cual, consta de documentos escritos, gráficos, imagen lógicos, electrónicos, y de cualquier otra índole, en los cuales, el personal de salud deberá hacer los registros, anotaciones, en su caso, constancias y certificaciones correspondientes a su intervención en la atención médica del paciente, con apego a las disposiciones jurídicas aplicables.

Se utilizan en gran medida para documentar diagnósticos y procedimientos. Son más de 150 los sistemas de codificación conocidos y sus modificaciones clínicas. Esto requiere que el investigador tenga la facilidad de obtener información de los Datos Clínicos de Cáncer de Mama (disponibles en TCGA-BRCA), que se encuentran en el portal: Genomic Data Commons (GDC, por sus siglas en inglés) del NCI (National Cancer Institute) [**?**]. Ya que los investigadores del Área de Ciencias Genómicas que trabajan con investigaciones que requieran datos clínicos, usan archivos de secuencia sin procesar, normalmente almacenados como BAM o FASTQ, constituyen la mayor parte de los datos.

El tamaño de un solo archivo puede variar mucho, dependiendo del análisis específico; Sin embargo, algunos de los archivos completos del genoma BAM en el Atlas del genoma del cáncer (TCGA) alcanzan tamaños de 200-300 GB. En tales casos, una llamada de descarga de datos en alto rendimiento es esencial. Para esto las recomendaciones del sistema para utilizar la herramienta de transferencia de datos GDC son especificadas para tener los procesos con mayor fluidez en tráfico de datos, rendimiento en específico que debe contar para su descarga y almacenamiento de datos.

Con tal fin, se propone implementar un Software que transforme archivos con diferentes formatos del GDC Portal, en un modo para que el usuario tenga la facilidad de manejo de datos. Así, las consultas sean más eficientes para el usuario en cuanto a tiempo y esfuerzo.

1.1 Plantiamiento Del Problema

El cáncer de mama ha tenido un desarrollo ascendente en las últimas décadas, así como el incremento del número de muertes por esta. Esta situación se vincula con los estilos de vida, así como con transiciones demográficas y epidemiológicas, procesos que han favorecido el aumento en la esperanza de vida de la población y por ende el desarrollo de enfermedades crónico-degenerativas [1,2]. El cáncer se encuentra entre los padecimientos que destacan en el panorama epidemiológico mundial, cuya tasa de incidencia y mortalidad varía en relación con el área geográfica y las condiciones de vida [3,4].

En las investigaciones que realizan las personas en distintos sectores de investigación requieren Datos Clínicos y suelen recurrir al portal Genomic Data Commons (GDC, por sus siglas en inglés) del NCI (Instituto Nacional de Cáncer) en el que está unificado la información de numerosos estudios y permite compartir datos entre los estudios genómicos sobre el cáncer en apoyo de la medicina de precisión.

El investigador requiere tener la facilidad de obtener la información de los Datos Clínicos en Cáncer de Mama (TCGA-BRCA), que proporciona a la comunidad de investigación sobre el cáncer un repositorio unificado de datos. Se proporcionan herramientas para guiar las presentaciones de datos de los investigadores e instituciones. Apoya varios programas del genoma del Cáncer en el centro del NCI para la genómica del cáncer (CCG) y en nuestra especificación Cáncer de Mama (TCGA-BRCA), incluyendo el atlas del genoma del cáncer (TCGA). También proporciona herramientas para guiar la presentación de datos, incluyendo el Portal de Presentación de Datos de GDC, una herramienta basada en la web para la presentación de datos clínicos, bio-específicos y pequeños volúmenes de datos moleculares, así como la herramienta de Transferencia de Datos de GDC. También está disponible una API GDC segura para las presentaciones de datos por lotes.

A todo esto en lo particular para el investigador que trabaje en este portal, ésta información tiene una problemática peculiar aparte de las recomendaciones que el portal ofrece, en relación a la visualización de los datos, en el caso de los usuarios hay métodos complejos con diferentes aplicaciones, aparte de tener en específico el manejo de los lenguajes de programación, teniendo a consideración la especificación de trabajo y contando con las especificaciones requeridas, es complejo llegar y especificar los datos requeridos en el cual se basa el Mapeo de Datos Clínicos en Cáncer de Mama, para esto; las consultas de los usuarios de datos determinados que necesite específicamente, no reúne las cualidades fáciles de uso y comparativa para seleccionar el muestreo clínico, en una visualización compresible al usuario que necesite, a todos los factores a evaluar un dato requerido. En el cual los datos que requieren los usuarios por lo regular, son diagnósticos y procedimientos, que están en un sistema codificado que clasifican y muestran un historial clínico de una población asignada y encontrada en Mapeo de Datos Clínicos en Cáncer de Mama de especialización que se trabaja en el área de Bioinformática Computacional o en otra área del Instituto que lo requiera.

1.2 Justificación

En este proyecto se busca desarrollar un modelo de Software, de mejora para ser aplicado en la Biología Computacional y otras áreas que lo requieran, a fin de mejorar la capacidad de obtención de datos y metodológica-mente el proceso actual que existe en el portal GDC reducir el tiempo y costo del usuario. El desarrollo permitirá:

- No descargar datos cada vez que requiera hacer una consulta.
- Agilizar el tiempo de búsqueda de datos.
- La agilidad de visualizar datos específicos sin ningún lenguaje de programación u otro tipo de aplicaciones.

- No tener que aplicar códigos para visualizar y/o descarga de datos, o en otro caso pérdida de la información.

1.3 Delimitación

En el año actual los investigadores tienen un problema para la visualización de forma rápida y segura de ver Mapeo de Datos Clínicos de Cáncer. Las descargas en la Web a veces pueden tardar dependiendo la velocidad del internet que esté a su disposición. Lo que no parece un problema importante pero en ocasiones tiende a detenerse la descarga o tiene fallos en los servidores de GDC y el tiempo que tarden en cargar los datos y en el lenguaje de programación que estén utilizando. La aplicación también tiene sus limitaciones, porque la Base de Datos se va a alojar en el servidor del INMEGEN y no podrá mostrarse en la WAM, Así, también la integridad de los datos se resguarda de un mal uso. La aplicación se limita al tratado de Cáncer de Mama, en el software tiene la posibilidad de ampliarse hacia los demás tipos de cáncer en el cual el uso se podrá hacer más frecuente en los investigadores. Los recursos son suficientes para operar la aplicación ya que el INMEGEN cuenta ya con la estructura requerida para el desarrollo de una sistematización.

1.4 Hipótesis

- La conexión de la Base de Datos en MongoDB a Java (NetBeans) para la habilitación de una aplicación gráfica de visualización y obtención de datos para el usuario permitirá un mejor rendimiento en el tratamiento y mapeo de la información de datos clínico de mama.

1.5 Objetovos (General Y Específicos)

1.5.1 Objetivo General

Implementar un Software que trabaje con datos del GDC, en un modo para que el usuario tenga la facilidad de manejo de datos. Así, las consultas de datos sea más eficiente para el usuario.

1.5.2 Objetivos Específicos

1. Buscar la forma de descarga más confiable de TCGA-BRCA que se encuentran en el GDC.
    - Analizar el formato si es factible de descargar .xlsx.
    - Analizar el formato si es factible de descargar .xml.
    - Analizar el formato si es factible de descargar .json.
    - Diferenciar los formatos para su uso.
2. Analizar los Gestores de Base de Datos No Relacional más conveniente para el uso del formato escogido.
3. Analizar los Lenguajes de Programación para realizar la aplicación.
4. Implementar la API para facilitar al usuario el uso de Datos: Mapeo de Datos Clínicos en Cáncer de Mama.

1.6 Aportaciones De La Tesis

Una problemática en cuanto a la elaboración de una estructura de Datos Clínicos para consultar y lograr una buena visualización; buscar la manera de descarga, lenguajes de programación para compilación, cosas cotidianas que hacemos para recabar datos en actividades en el área de trabajo, etc. En nuestro caso con la herramienta de trabajo que utilizan los investigadores que son datos de secuencia sin procesar, que deben leer y expresarlos para un buen entendimiento al trabajar con ellos. En el Mapeo de Datos Clínicos, los datos son tan importantes e indispensables para una investigación y seguimiento de pacientes que se trataron en TCGA-BRCA. El logro de hacer una estructura de datos amigable que se encuentran en el portal Genomic Data Commons (GDC), reducirá el tiempo de descarga como de búsqueda de los datos, ya que al tener el Software en una plataforma podrá observar que al ver los datos ya están curados y se encuentran en mayor parte de su totalidad sin procesar. Así, los datos están en una forma factible para su descarga o bien nada más para consulta o comparación en un ámbito que se pueda laborar con los datos [**?**].

1.7 Organización De La Tesis

Se muertra como paulatinemente se va a ir trabajando las actividades marcadas.

2 Estado Del Arte (Antecedentes)

2.1 Estado Actual Del Cáncer De Mama

El Cáncer de Mama (C.M.) es un problema de salud pública a nivel mundial. Su alta frecuencia, las implicaciones biológicas, el impacto emotivo y económico que acarrea en la paciente y sus familiares, hacen de esta enfermedad uno de los problemas de salud más discutidos a nivel médico-familiar y en la sociedad en la actualidad. El cáncer de mama es un padecimiento crónico, heterogéneo, y con una evolución irregular, tan lenta que permite a un 10% vivir más de 12 años a enfermas inoperables que resisten todo tipo de tratamiento y por otro lado, mujeres con tumores tempranos menores de 1 cm presentan enfermedad diseminada en un 10-20% de los casos [**?**].

El Cáncer de Mama es la neoplasia más frecuente en las mujeres a nivel mundial aunque puede presentarse en hombres, la proporción es de 1 caso por 150 mujeres [**?**]. De acuerdo al informe de la International Agency for Research on Cancer (IARC), en el año 2008, se diagnosticarón 1,380 300 nuevos casos, representando el 23% de los cánceres en las mujeres. El número de casos fue casi igual en los países desarrollados que en las que vivían en países en desarrollo, 692,000 en los primeros y de 691,000 en los segundos. Sin embargo, es de hacer notar que la población en los primeros países se calculó en 1 billón y en los segundos de 6 billones, de acuerdo a cifras del Banco Mundial en el 2006 [**?**]. El riesgo de una mujer mexicana de desarrollar un C.M., durante su vida es de 2.9% comparado con el 4.27% para Latinoamérica y de 7.14% para mujeres de países desarrollados [**?**, **?**].

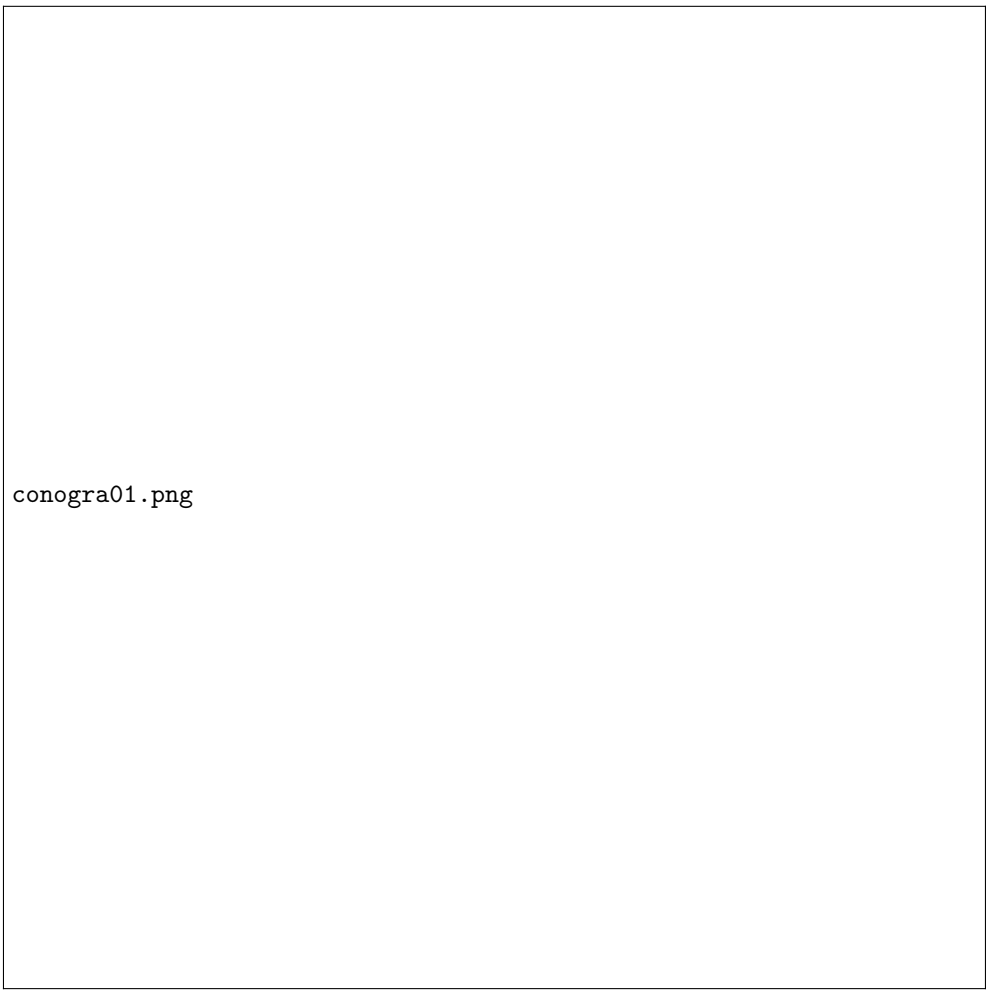

Figure 1: Cronograma de actividades.

2.2 Integración Semántica Y Estandarización De Datos Clínicos Basada En Arquetipos

La información de Mapeo de Datos Clínicos es un documento médico-legal que surge del paciente, donde se recoge la información necesaria para la correcta atención y muestra historia clínica que construye un documento principal en un sistema de información sanitario, imprescindible en su vertiente asistencial, administrativa, y además constituye el registro completo de la atención prestada al paciente durante su enfermedad, de lo que se deriva su trascendencia como documento legal. Además de los datos clínicos que tengan relación con la situación actual del paciente, incorpora los datos de sus antecedentes personales y familiares, sus hábitos y todo aquello vinculado con su salud biopsicosocial. También incluye el proceso evolutivo, tratamiento y recuperación. La historia clínica no se limita a ser una narración o exposición de hechos simplemente, sino que incluye en una sección aparte los juicios, documentos, procedimientos, informaciones y consentimiento informado. El consentimiento informado del paciente, que se origina en el principio de autonomía, es un documento donde el paciente deja registrado y firmado su reconocimiento y aceptación sobre su situación de salud y/o enfermedad y participa en la toma de decisiones del profesional de la salud. [**?**]

3 Marco Teórico

Se pretende desarrollar un software que pueda ser aplicado como una herramienta útil para la investigación en Biología Informática o en otras especializaciones que lo necesiten. Es necesario tener en cuenta que, en todo desarrollo de sistemas de software es de suma importancia definir una metodología. Esta permite a los desarrolladores seguir alguna especificaciń en cada una de las etapas del desarrollo del sistema, desde los requerimientos iniciales hasta las pruebas finales, que haga que el software sea coherente y además formal. Abordaremos los conceptos computacionales tomados en cuenta durante todo el proceso de elaboración del software de este proyecto. Los conceptos que a continuación trataremos son:

- La ingeniería de software.
- Metodología Orientada a Objetos.
- Arquitectura cliente-servidor.
- Modelo no relacional (NoSQL).

Las cuales darán la pauta sobre medicina del cual se basan los datos clínico, genética de cáncer de mama que padecen las personas, los estándares utilizados tanto para el análisis, almacenamiento, diseño, implementación, pruebas y mantenimiento de la aplicación; la ingeniería examinará la aplicación existente para actualizarla y mejorarla; las bases de datos permitirán el manejo y manipulación de la gran cantidad de datos que existan; MongoDB, NetBeans y JQuery ayudarán en la automatización de ciertas tareas.

3.1 Ingeniería De Software

El término “Ingeniería de Software" fue introducido por primera vez a finales de 1960 en una conferencia destinada a su discusión, la cual fue posteriormente llamada “crisis del software". Esta crisis de software fue el resultado directo de la introducción del hardware de la tercera generación computacional [**?**]. Para tener una idea clara de lo que es la ingeniería de software vamos a definirlo según varios autores:

1. La aplicación de un enfoque sistemático, disciplinado y cuantificable hacia el desarrollo, operación y mantenimiento del software; es decir, la aplicación de ingeniería al software.

2. Es una disciplina o área de la Informática o Ciencias de la Computación, que ofrece métodos y técnicas para desarrollar y mantener software de calidad que resuelven problemas de todo tipo [**?**].

El factor común en estas definiciones es que la ingeniería de software se enfoca a los sistemas computacionales, utilizando los principios de la ingeniería para el desarrollo de estos sistemas, y está compuesta por aspectos técnicos y no técnicos. La ingeniería de Software no es una disciplina que sólo deba aplicarse en proyectos de ciertas áreas, sino que también trata con áreas diversas dentro de las ciencias computacionales, tales como: construcción de compiladores, sistemas operativos, o desarrollos empresariales como es el caso de ésta aplicación de software. La Ingeniería de Software abarca todas las fases del ciclo de vida en el desarrollo de cualquier sistema de información aplicables a áreas tales como investigación científica, medicina, logística, y para este caso en particulares. En un nivel técnico la ingeniería de software empieza con una serie de tareas de modelado que llevan a una especificación completa de los requisitos y a una representación del diseño general del software a construir. Con los años se han propuesto muchos métodos para el modelado del análisis. Sin embargo, ahora dos tendencias dominan el modelado del análisis, el análisis estructurado y el análisis orientado a objetos.

3.2 Metodología Orientada A Objetos

Vivimos en un mundo de objetos. Estos objetos existen el a naturaleza, en entidades y en los productos que usamos. Los objetos pueden ser clasificados, descritos, organizados, combinados, manipulados y creados. Es por esto que se propuso un análisis y desarrollo orientado a objetos, que nos permita aprovechar las características, individualidad y facilidad de manipulación que nos ofrecen los objetos.

Es así que al estar hablando de objetos es importante describir las ideas fundamentales implícitas en la tecnología orientada a objetos incluyen [**?**]:

- *Objetos.* Un objeto es cualquier cosa, real o abstracta, acerca de la cual almacenamos datos y aquellos métodos que los manipulan.

- *Clases.* Una clase es la implementación de un tipo de objeto. Específica la estructura de datos y los métodos operacionales permitidos que se aplican a cada uno de sus objetos.

- *Métodos.* Específica la manera en la cual los datos de un objeto son manipulados. Los métodos en un tipo de objeto hacen solamente referencia a la estructura de datos de ese tipo de objeto. No deben de accesar directamente a la estructura de datos de otro objeto.
- *Peticiones.* Una petición solicita una operaciín específica debe ser invocada usando uno o varios objetos como parámetros.

Una vez que se han mencionado las ideas fundamentales del modelo orientado a objetos, es importante saber que existen tres conceptos importantes que diferencian el enfoque OO de la ingeniería del software convencional:

1. *Encapsulamiento* empaqueta los datos y las operaciones que manejan estos datos en un objeto simple con denominación.
2. *Herencia* permite que los atributos y operaciones de una clase sean heredados por todas las subclases y objetos que se instancian de ella.
3. *Polimorfismo* permite que una cantidad de operaciones diferentes posean el mismo nombre, reduciendo la cantidad de líneas de código necesarias para implementar un sistema y facilita los cambios en caso que se produzcan.

Como sabemos, los objetos están compuestos por atributos los cuales describen un objeto; que en esencia, son los que definen al objeto, a la vez que clarifican lo que se representa con el objeto en el contexto del espacio del problema.

Para poder manipular los atributos de los objetos existen los algoritmos que los procesan, los cuales son llamados operaciones, métodos o servicios y pueden ser vistos como módulos en un sentido convencional. Cada una de las operaciones encapsuladas por un objeto proporciona una representación de uno de los comportamientos del objeto. Las operaciones definen el comportamiento de un objeto y cambian, de alguna manera, los atributos de dicho objeto.

No sólo se requiere conocer la forma en la que los objetos interactúan entre sí, sino también es necesario saber que el proceso se mueve a través de una espiral evolutiva, que comienza con la comunicación con el usuario. Es aquí donde se define el dominio del problema y se identifican las clases básicas del problema.

Esta es la metodología que se empleará para el desarrollo de la aplicación. El análisis y diseño orientado a objetos tiene dos aspectos. Al primer aspecto le conciernen los tipos de objeto, clases, relaciones entre los objetos y la herencia, y se conoce como el Análisis de Estructura de Objetos (AEO) y Diseño de Estructura de Objetos (DEO). Al otro aspecto le concierne el comportamiento de los objetos y que les pasa con el tiempo, y se conoce como el Análisis del Comportamiento de Objetos (ACO) y Diseño del Comportamiento de Objetos (DCO) [**?**].

3.3 Arquitectura Cliente-Servidor

El término cliente-servidor se refiere a una arquitectura o división lógica de responsabilidades; donde el cliente (parte frontal o aplicaciones para el usuario o

interfaces) es la aplicación que se ejecuta sobre el DBMS, aplicaciones escritas por el usuario y aplicaciones integradas; y el servidor (parte dorsal o servicios de fondo) es el DBMS y soporta la definición, manipulación, seguridad e integridad de los datos entre otros [**?**]. El uso de la arquitectura cliente-servidor brinda ciertas ventajas como son: o El servidor puede ser una máquina construida a la medida y por lo tanto proporcionar un mejor desempeño, o maneja el procesamiento paralelo normal, es decir el procesamiento del servidor y del cliente se están haciendo en paralelo, por lo que el tiempo de respuesta y velocidad real de trasporte mejoran, o varias máquinas cliente pueden acceder a la misma máquina servidor y por lo tanto una sola base de datos puede ser compartida entre varios sistemas clientes distintos.

3.4 Modelo No Relacional (Nosql)

En un mundo de constantes cambios a nivel de sistemas, es necesario volver a pensar acerca de los paradigmas que manejan la industria. Necesitamos ajustar nuestras herramientas a las necesidades reales que tenemos hoy en día con el fin de tener sistemas a la altura de nuestros requerimientos.

Las bases de datos relacionales requieren que los esquemas ser definidos antes de poder añadir datos. Por ejemplo, es posible que desee almacenar datos sobre sus clientes, tales como números de teléfono, nombre y apellido, dirección, ciudad y estado - una base de datos SQL necesita saber qué va a almacenar por adelantado.

Esto encaja mal con los enfoques de desarrollo ágil, ya que cada vez que se complete nuevas características, el esquema de la base de datos a menudo tiene que cambiar. Así que si usted decide, unas pocas iteraciones en el desarrollo, que desea guardar los elementos favoritos de los clientes, además de sus direcciones y teléfonos, tendrá que añadir que la columna de la base de datos, y luego migrar toda la base de datos para el nuevo esquema.

Si la base de datos es grande, este es un proceso muy lento que implica un tiempo muerto significativo. Si se cambian con frecuencia los datos de sus tiendas de aplicaciones porque usted está iterando rápidamente este tiempo de inactividad también puede ser frecuente. Tampoco hay manera, utilizando una base de datos relacional, para abordar eficazmente los datos que es completamente desestructurada o conozca con antelación.

Bases de datos NoSQL están construidos para permitir la introducción de datos sin un esquema predefinido. Eso hace que sea fácil hacer cambios significativos de la aplicación en tiempo real, sin tener que preocuparse por las interrupciones del servicio lo que significa que el desarrollo es más rápido, integración de código es más fiable, y se necesita menos tiempo del administrador de base de datos. Generalmente, los desarrolladores han tenido que añadir el código de la zona de aplicación para aplicar los controles de calidad de datos, tales como ordenar la presencia de campos específicos, tipos de datos o valores permisibles.

Las bases de datos NoSQL más sofisticadas permiten las reglas de validación que deben aplicarse dentro de la base de datos, permitiendo a los usuarios para

hacer cumplir la gobernabilidad a través de los datos, manteniendo al mismo tiempo las ventajas de agilidad de un esquema dinámico. Aun así, hemos llegado a un punto en que seguir usando bases de datos relacionales para todos los casos es simplemente inviable. Existen varios problemas con los RDBMS actuales que pueden suponer una seria limitante para la construcción de aplicaciones. Estos problemas son en gran medida el motivo por el que surgió el movimiento NoSQL [**?**].

4 Metodología

La metodología se basa través de etapas que gradualmente se cumplirán mismas que se procede a la búsqueda y recopilación de las fuentes de información en multimedia como también el manejo de la plataforma de trabajo que los datos de secuencia en bruto están alojados bajo el esquema que se revisan especialmente sobre el tema de Genomic Data Commons Data Portal. Y las fuentes bibliografías, artículos que también son relevantes para el desarrollo del Software, en la figura 2 se muestra la estructura a trabajar.

Vamos a basarnos en el desarrollo que se fue formando heurísticamente en etapas que se mencionan a continuación:

1. Revisión de bibliografías, artículos y trabajos especiales sobre el tema.
2. Se mencionan las formas de descarga que en el portal de GDC Data Commons y los usuarios pueden elegir a su necesidades la formaa de descarga.

   Nota: El trabajo que se realizá está desarrollado en el sistema operativo GNU/Linux, Si existe algún problema con su sistema operativo revisar los manuales que el soporte ofrece a su sistema.

   (a) Descargar datos usando un archivo de manifiesto.

   Una forma conveniente de descargar varios archivos desde el GDC es usar un archivo de manifiesto generado por el Portal de datos de GDC. Después de generar un archivo de manifiesto, inicie la descarga utilizando la herramienta de transferencia de datos GDC suministrando la opción -m o –manifiesto, seguido de la ubicación y el nombre del archivo de manifiesto. Los usuarios de OS X pueden arrastrar y soltar el archivo de manifiesto en la Terminal para proporcionar su ubicación.

   (b) Descargar datos usando UUID de archivos GDC.

   La herramienta de transferencia de datos GDC también admite la descarga de uno o más archivos individuales utilizando UUID (s) en lugar de un archivo de manifiesto. Para hacer esto, ingrese los UUID (s) después del comando de descarga.

   (c) Descargar datos de acceso controlado.

   Se requiere un token para autenticar el usuario para descargar datos de acceso controlado desde GDC. Los tokens pueden obtenerse en el

Figure 2: Diagrama de flujo de trabajo de software a trabajar.

Portal de datos en GDC. Una vez descargado, el archivo token se puede pasar a la herramienta de transferencia de datos GDC usando la opción -t o –token-file.

(d) Descarga obteniendo un archivo de manifiesto para la descarga de datos en forma manual.

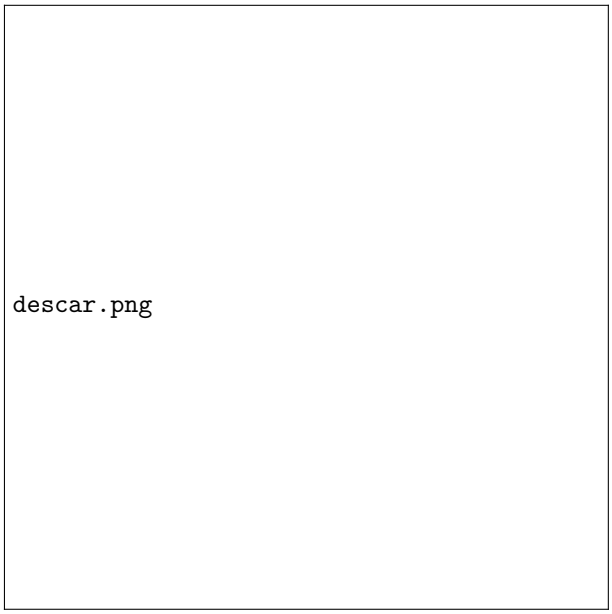

Figure 3: Imagen de descarga de un rachivo de manifiesto en forma manual.

La herramienta de transferencia de datos de GDC admite la descarga de múltiples archivos enumerados en un archivo de manifiesto de GDC. Los archivos manifiestos se pueden generar y descargar directamente desde el Portal de datos de GDC: Primero, seleccione los archivos de datos de interés. Haga clic en el botón Carro en la fila correspondiente al archivo deseado. El botón se pondrá verde para indicar que el archivo ha sido seleccionado.

Una vez que se hayan seleccionado todos los archivos de interés, haga clic en el botón Carro en la esquina superior derecha. Esto abrirá la página del carro, que proporciona una descripción general de todos los archivos seleccionados actualmente. Esta lista de archivos se puede descargar como un archivo de manifiesto haciendo clic en el botón verde Descargar y seleccionando Manifiesto en el menú desplegable.

La ultima forma de descarga fue la que seleccione para trabajar dado que las demás formas son un poco más complejas. Se obtuvierón los datos en formato de .xml ya que la plataforma cuenta con diferentes tipos de archivos, está es por que se obtuvierón en forma general en un solo formato.

3. Para crear una base de datos del portal https://portal.gdc.cancer.gov/cart, que está los archivos en .xml. Necesitamos convertirlos a .json para facil-

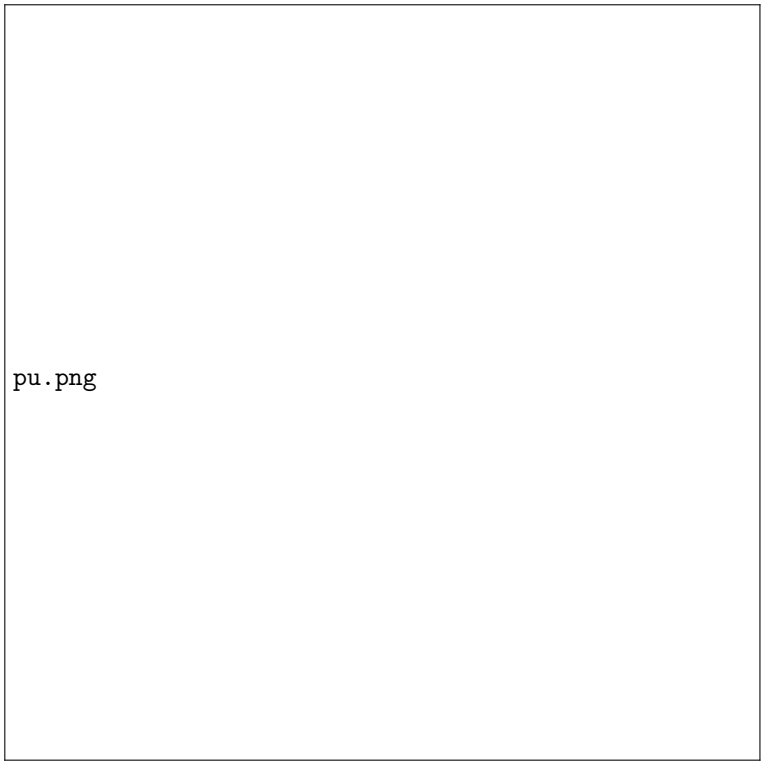

Figure 4: Imagen de lista de archivos a descargar.

itar la carga de datos para esto hacemos una conversión masiva de .xml a json. Para esto necesitamos aplicaciones que esten instaladas para que puedan correr en un mk.
Verificar si ya tenemos instalados los programas requeridos y escribimos:

* apt search xml etree
* apt search python xml

Instalamos.

* pip install https://github.com/hay/xml2json/zipball/master
* pip install –upgrade pip
* sudo apt install python-pip ipython
* sudo apt install 9base

La directorio que estamos utilizando para nuestro proyecto. Utilizamos un editor de textos, el archivo con el nombre de mkfile. Así, que quedaría para la conversión en nuestro directorio de mk.

Nota: Se debe ubicar en la directorio que se tiene los archivos .xml, el directorio bin que dentro debe contener su README.md y su targets, un index.php con su README.md y el mk. Para hacer funcionar esto nos vamos al directorio .bashrc.

Guardamos esto, y nos indica que busqué el mk que existan en nuestro sistema operativo. Como queremos hacer la conversión masivo, también en el directorio que estamos utilizando se crea otra directorio con el nombre de bin/ y ahí creamos un archivo targets. Para que sea utilizado con la aplicación xargs. Ya con esto damos permisos y corremos la aplicación en terminal.

4. Ya obtenido los datos en .json. Ahora se hace la carga de la Base de Datos a nuestro GBD (Gestor de Base de Datos) que en nuestro caso ocupamos MongoDB. En este paso es utilizado desde terminal un código que hace llamar todos los archivos que se encuentran en nuestro directorio para importar a la base de datos ya destinada.

   Si tiene alguna duda para hacer la importación o instalación consultar a la pagina https://docs.mongodb.com/manual/. En este caso la importación se hace masiva para agilizar el tiempo de carga de datos a la base de datos.

5. Ya obtenida la base de datos hacemos unas consultas de ejercicio. Ahora podemos crear nuestra primera conexión a la base de datos MongoDB con Java, nos vamos a conectar con la configuración por defecto que es en localhost y el puerto 27017. Para conectar con la base de datos simplemente utilizamos el método getDatabase(“test”), donde test es el nombre de la base de datos que estamos utilizando como ejemplo de una conexión. Ahora, nada más depender del programador para toda la visualización y los detalles que requiere el usuario y hacer las pruebas para la aplicación.