UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE POSGRADO EN ASTROFÍSICA
INSTITUTO DE ASTRONOMÍA

UNA HERRAMIENTA PARA EL MANEJO/ANÁLISIS DE BASES DE DATOS ASTRONÓMICAS, EN LA ERA DEL TSUNAMI DIGITAL: APLICACIÓN AL ANÁLISIS DE BULBOS EN LAS GALAXIAS DEL UNIVERSO LOCAL.

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN CIENCIAS (ASTRONOMÍA)

PRESENTA:
JULIETA RUT SALAZAR CONTRERAS.

TUTORES:
DR. JOSÉ OCTAVIO VALENZUELA TIJERINO.
INSTITUTO DE ASTRONOMÍA
DR. HÉCTOR MANUEL HERNÁNDEZ TOLEDO.
IINSTITUTO DE ASTRONOMÍA

MÉXICO, D. F. MAYO 2014.

Universidad Nacional Autónoma de México

Instituto de Astronomía

Una Herramienta para el manejo/análisis de bases de datos astronómicas, en la era del Tsunami digital: Aplicación al análisis de bulbos en las galaxias del Universo Local.

T E S I S

QUE PARA OBTENER EL GRADO DE:

Maestra en Ciencias (Astronomía)

PRESENTA:

Julieta Rut Salazar Contreras

DIRECTORES DE TESIS:
Dr. José Octavio Valenzuela Tijerino y
Dr. Héctor Manuel Hernández Toledo.

2014

A mi mamá.

"No busco, encuentro."

PABLO PICASSO.

Agradecimientos

En primer lugar quiero agradecer a mis dos directores de tesis, Octavio Valenzuela y Héctor Toledo, por todo el tiempo que dedicaron a este trabajo, sus enseñanzas, su paciencia, las conversaciones, la invaluable ayuda que en todo momento me han proporcionado y sin duda, por la gran cantidad de cosas que me han ayudado a aprender, no sólo astronómicas.

Agradezco profundamente la lectura, el tiempo, la paciencia y cada una de las sugerencias que me ha hecho Christophe Morisset, pues en definitiva enriqueció con su labor – de manera notable – este trabajo.

Agradezco a Héctor Velázquez quien no sólo ha leído cuidadosamente como sinodal mi trabajo de tesis y hecho sugerencias valiosas, sino que siempre ha tenido palabras alentadoras y amables hacia mi persona y mi trabajo.

Agradezco a Liliana Hernández quién dedicó parte de su tiempo a leer mi trabajo. Por sus sugerencias invaluables que sólo resultaron en la ampliación de este trabajo.

Agradezco a Ivanio Puerari por todas sus atenciones, su apoyo, por leer mi tesis y desde luego, por sus sugerencias a mi trabajo.

A las personas que durante este trayecto me ayudaron de manera invaluable: Dr. Vladimir Ávila Reese, Dr. Luis Aguilar Chiu, Dr. Luis Felipe Rodríguez, Dr. Divakara Mayya Yalia y Dr. Alejandro González.

Agradezco a la Universidad Nacional Autónoma de México, porque además de ser mi Alma Máter, sin duda se acerca a la concepción del paraíso en la Tierra.

Agradezco al Instituto de Astronomía por albergarme como estudiante y por las infinitas facilidades que me brindó para la realización de esta tesis.

Mi eterno agradecimiento a Claude Thions por permitirme incursionar y adentrar en el mundo de la docencia, por sus enseñanzas, amistad, confianza y tiempo.

Sin duda, debo agradecer al apoyo más grande e incondicional que siempre he recibido del mejor ejemplo de humanidad que he tenido en la vida: mi mami. Sabes que eres mi motor y que te amo.

Agradezco las enseñanzas que forjó en mí mi padre, cuya ausencia se nota cada día.

Mauricio: me gustaría tener las palabras adecuadas y perfectas para hacerte saber lo importante que has sido en ésto, lo agradecida que estoy contigo y lo mucho, mucho que te quiero.

Le doy las gracias de todo corazón y con mucho cariño a mi hermano Armando, quien prefiere que le digan Arturo.

A mi Tía Queta que es un ser humano maravilloso y que está conmigo en las tan trilladas buenas y malas. Y ya que hablo de familia, tendría que mencionar a las que permití que me adoptaran y que adopté como tal: Olmo González, Mauricio Garza, Yaxk'in Coronado y Antonio Vázquez – con sus respectivas familias – que considero mías también, Patricia Guiza, Yareli Quintana, Hugo Ramírez, Jorge Avella y Marianela Noriega. Gracias por llenar y estar presentes en mi vida de la manera en que lo hacen.

Gracias Yaxk'in Coronado, Jesús Zendejas, Jonnathan Reyes, Marisol Mayen, Santi Roca, Antonio Vázquez...

A todas las personas que sin duda hicieron que en la vida hubiera sonrisas y astronomía: Alma González, Patricia Lara, Eva Martínez, Miguel Monroy, Francisco J. Hernández, Nahiely Fajardo, Omar Anguiano, Tula Bernal, Luc Jament, Simon Verley, Bernardo Cervantes, Sandro Mendoza, Moisés Magaña, Paco Guillén, Mari Loly Martínez, Mónica Sánchez, Zeús Valtierra, Claudio Toledo, Alejandro González, Enrique Anzures, Celia Fierro, Fátima Robles,

Jorge Gacía Rojas, Alex Farah, Roberto Figuera, Juan Abonza, Nancy Ávila, Mónica Lozada, Bertha Vázquez, Alfredo Díaz. Yo sé – y ustedes saben – que seguramente sí me faltan estrellas, de esas que iluminaron mi firmamento, pero seguramente sonreirán con las omisiones probables y recordarán la terrible mala memoria que tengo.

Aún así – con mi mala memoria – gracias a todos por estar.

"Parte de mi amor a la vida
se lo debo a mi amor
a la astronomía".

Rut Salazar.
Mayo del 2014.
México, D.F.

Índice

Índice de figuras

Índice de tablas

Resumen

El avance de las tecnologías de información y comunicación han puesto al alcance de cualquier usuario de internet una cantidad enorme y creciente de información. A este crecimiento se le conoce como: el *Tsunami Digital* [Zhang & Chen, 2010]. La astronomía actual, debido al incremento exponencial en la cantidad, complejidad y calidad de los datos astronómicos provenientes de los grandes catastros observaciones actuales como lo son GAIA[1] [Jordan, 2008], SDSS[2] [York et al., 2000], LSST[3] [Axelrod, 2006], etc. o de las simulaciones numéricas por computadora de procesos físicos que por su naturaleza no-lineal no pueden ser estudiados de manera analítica (como lo es la formación de estructura en el Universo, la formación estelar, etc.) se enfrenta ante un *Tsunami Digital* [Zhang & Chen, 2010].

El crecimiento de datos astronómicos requiere de nuevos métodos para su análisis y organización. En astronomía se ha hecho imperativo analizar grandes bases de datos que permitan detectar con significancia estadística, efectos y sistematicidades no encontradas en estudios previos. Esto se puede extender a algoritmos que traten de manera simultánea más de un gran conjunto de datos. La demanda de herramientas y de recursos computacionales para realizar análisis de datos crece rápidamente.

Tratando de encarar los diversos problemas que se presentarán para la astronomía observacional, debido a la gran información que brindarán los grandes telescopios, en un futuro muy cercano, en esta tesis se presenta una herramienta computacional que se desarrolló para manipular de manera eficiente estas grandes bases de datos – como parte de este trabajo de tesis –.

Dicha herramienta utiliza el principio de **reciclaje de software** [Reifer,

[1] http://sci.esa.int/gaia/
[2] http://www.sdss.org/
[3] http://www.lsst.org/lsst/

1997], que se refiere al comportamiento y a las técnicas, que garantizan que una parte o la totalidad de un programa informático existente, se pueda emplear en la construcción de otro programa. De esta forma se aprovecha trabajo preexistente, se economiza tiempo, y se reduce la redundancia.

Dicha herramienta está basada en un gestor de bases de datos "Open Source"[4] común, llamado **MySQL** que utiliza el lenguaje **MySQL** (Lenguaje de Consulta Estructurado o Structured Query Language por sus siglas en inglés), que permite manipular bases de datos. Así pues, utilizando herramientas computacionales ya existentes, como el gestor de bases de datos, se implementa un conjunto de herramientas bajo el nombre de **"PICASSO"**, que son, distintos módulos de programas cuyo lenguaje nativo es **MySQL**, **C** y **Python**. Estos últimos dos lenguaje se utilizan para poder proporcionar una estructura jerárquica a las tablas que conforman las diferentes bases de datos astronómicas. Con dicha jerarquía, tanto el anlisis, la búsqueda así como la manipulacin de datos en grandes bases se vuelve no sólo más eficiente sino que nos permite utilizar una cantidad menor de memoria RAM en la computadora, que con las herramientas existentes actuales.

Para poner a prueba la herramienta computacional desarrollada para esta tesis, se trabajó con distintos catálogos astronómicos. Haciendo énfasis en dos distintos catálogos observacionales públicos provenientes del Sloan Digital Sky Survey[5] (SDSS)[6][York et al., 2000], en particular, provenientes del DR7 [Abazajian et al., 2009].

El primer catálogo de galaxias que se utiliza en este trabajo de tesis es el de Simard et al. [2011]. Dicho catálogo consta de 1,123,718 galaxias para las cuales Simard et al. [2011] presenta tres distintas descomposiciones Bulbo al Total $(\boldsymbol{B/T})$. Utiliza un modelo de Sérsic puro, uno con bulbo y disco con $n_b = 4$ y uno con el parámero de Sérsic libre (n_{free}). En particular, sólo se trabaja con el catálogo que tiene una descomposición Bulbo al total $(\boldsymbol{B/T})$ con el parámetro de Sérsic libre (n_{free}).

[4]Open Source: es el término con el que se conoce al software distribuido y desarrollado libremente.

[5]http://www.sdss.org

[6]SDSS: Sloan Digital Sky Survey Es un proyecto de inspección del espacio mediante imágenes en el espectro visible y de corrimiento al rojo, realizada en un telescopio de 2.5 metros situado en el observatorio Apache Point de Nuevo México.

El segundo catálogo es el publicado por Tojeiro et al. [2009], que contiene información de masas estelares, formación estelar, historias de metalicidad y el contenido de polvo para un poco más de 800,000 galaxias. Para poder obtener dichos parámetros se utiliza **VESPA** - VErsatile SPectral Analysis [Tojeiro et al., 2007], que es un código que utiliza el rango espectral completo de una galaxia para construir de manera robusta información sobre la formación estelar y las historias de metalicidad de espectros galácticos.

Al relacionar dichos catálogos, se pueden comparar los resultados con distintos estudios estadísticos sobre la dependencia que tienen distintos valores para galaxias del cociente Bulbo al total $(\boldsymbol{B/T})$ contra sus masas estelares (M_*). Estos estudios están hechos con el fin de profundizar en el entendimiento de los procesos físicos que intervienen en la formación y evolución de las galaxias, pues la descomposición de galaxias se usa para estimar distintos parámetros estructurales de las componentes galácticas [Gadotti, 2008].

1
Introducción.

Los actuales telescopios y las simulaciones que se hacen por computadora están creando grandes volúmenes de datos que requieren tanto métodos de análisis como herramientas – no existentes hasta el momento – para poder realizar dichos análisis de manera sistematizada y organizada.

Hoy en día los volúmenes de datos científicos se están duplicando aproximadamente cada año [Bell et al., 2007], este proceso también está presente en los datos astronómicos, en donde además con cada uno de los distintos y nuevos instrumentos – con los que se obtiene información astronómica – se tiene una precisión mayor a los instrumentos anteriores, lo que se puede traducir en que el aumento exponencial no sólo es en la cantidad de datos sino además, es en la calidad de ellos. A este incremento – conocido como *El Tsunami Digital* – es a lo que se enfrenta un astrónomo en la actualidad.

En los últimos años las bases de datos han ayudado al quehacer astronómico. Hemos visto una explosión de distintos servicios como datos recopilados o incluso páginas web dedicadas enteramente a la difusión de datos astronómicos.

Prácticamente todos los observatorios han desarrollado sistemas que permiten la fácil distribución de su información en internet. En estos tipos de páginas

es posible encontrar una descripción técnica detallada de los instrumentos, la asignación del tiempo de observación, publicaciones, catálogos, además de que se cuenta con archivos y bases de datos astronómicas junto con las herramientas necesarias para su propia calibración y análisis. Ejemplos de estos tipos de páginas que cuentan con información del objeto son: Sloan Digital Sky Survey (SDSS)[1] [2], the Infrared Processing and Analysis Center (IPAC)[3] [4], The Database for physics of galaxies (HyperLeda)[5] [6], The SIMBAD astronomical database (SIMBAD)[7] [8], etc. Los archivos contenidos en estas páginas son herramientas fundamentales para la astronomía actual, ya que hay fenómenos variables que requieren el uso de registros durante largos intervalos de tiempo, los nuevos instrumentos requieren estar calibrados a partir de objetos de referencia observados con otros instrumentos, es decir, correlacionando archivos de datos astronómicos anteriores con los nuevos [Gray et al., 2007]. Los nuevos catálogos – provenientes de los nuevos telescopios – tendrán grandes cantidades de información y además, se requerirá correlacionarlos con distintos catálogos igual de grandes o más. Es evidente que tanto usar las bases de datos como a la informática, se vuelve una herramienta fundamental en la astronomía.

La tecnología que se utiliza para hacer los telescopios así como la instrumentación de los mismos, ha mejorado sustancialmente en las últimas décadas. Aún así la información que se tiene de ciertos fenómenos sigue siendo limitada. Por ejemplo: cómo se forma una estrella, cómo se forman los planetas, cómo se forman las galaxias. etc. Por ello la astronomía actual se tiene que apoyar en simulaciones por computadora, utilizadas para modelar el comportamiento de los sistemas naturales complejos, que van desde la interacción de partículas subatómicas a la evolución del Universo. Ayudando a predecir el comportamiento de fenómenos astronómicos para los cuales las observaciones contemporáneas no son suficientes.

[1]SDSS: Es un proyecto de inspección del espacio (que cubre el cielo en el hemisferio norte) mediante imágenes en el espectro visible y de corrimiento al rojo, realizada con un telescopio de 2.5 metros situado en el observatorio Apache Point de Nuevo México.

[2]http://www.sdss.org/

[3]IPAC: Infrared Processing and Analysis Center, cuenta con una gran colección de archivos de datos astronómicos, o de misiones planetarias, con un énfasis especial en el infrarrojo submilimétrico.

[4]http://www.ipac.caltech.edu

[5]HyperLeda Es una base de datos con información física de las galaxias.

[6]http://leda.univ-lyon1.fr

[7]SIMBAD es una base de datos de objetos astronómicos fuera del sistema solar.

[8]http://simbad.u-strasbg.fr/simbad

Actualmente existen diferentes proyectos de gran envergadura, que a partir de simulaciones numéricas, vinculan el comportamiento y entendimiento de diferentes fenómenos que ocurren en el Universo. Al igual, que su contraparte observacional, este tipo de conjuntos de datos son cada vez más grandes y por ende se vuelven cada vez más difícil de gestionar usando el software convencional.

Los avances en la computación de alto rendimiento están teniendo un impacto transformador, debido a los avances en el poder de cómputo y mayor capacidad, que permite entre otras tareas ejecutar simulaciones en una escala sin precedentes y cada vez más complejas. Aún así, la capacidad para analizar los datos resultantes sigue siendo limitada.

Una simulación puede llegar a estar en uso durante 10 o más años y es utilizada por un gran número de investigadores. Las aplicaciones de análisis de datos, no son utilizadas conjuntamente por todos los investigadores que utilizan la misma simulación, de manera que no se ha concentrado el esfuerzo en desarrollar aplicaciones rápidas que traten con los datos de manera colosal, así como tampoco cada aplicación que se ha desarrollado es altamente escalable, motivo por el cual el análisis de datos está frenando el avance científico.

En 2008 la Universidad de Washington corrió una simulación N-cuerpos para estudiar la formación y evolución de estructuras a gran escala del Universo [Kwon et al., 2010]. Se usaron 7.5 millones de horas en CPU y se generaron 50 Terabytes de información de datos con un extra de 25 Terabytes de información procesada. Las simulaciones como Millennium [Lemson & Virgo Consortium, 2006][9] y Aquarius [Springel et al., 2008][10] produjeron más de 30 Terabytes de datos cada una. Sin embargo, simulaciones más recientes como la simulación cosmológica Bolshoi [Riebe et al., 2011][11] produjo 75 TB de información y la simulación Millennium-II (MS-II) [Boylan-Kolchin et al., 2009][12] cerca de 100 TB de información.

Con todo estos datos se ha podido dar respuestas a varias preguntas que

[9] http://www.mpa-garching.mpg.de/galform/virgo/millennium/
[10] http://www.mpa-garching.mpg.de/aquarius/
[11] http://hipacc.ucsc.edu/Bolshoi/
[12] http://www.mpa-garching.mpg.de/galform/millennium-II/

hace unas décadas eran impensables de determinar, debido a la captura no sistematizada de los datos; pero ahora con los grandes proyectos como el **SDSS** o con las simulaciones por computadora se cuenta con muestras de datos astronómicos homogéneos y se puede hablar de significancia estadística. Sin embargo, hay muchas preguntas que siguen abiertas tales como ¿cuándo y cómo se formó la Galaxia?, ¿cuándo se formaron las estrellas en ésta?, ¿cómo se distribuye la materia oscura?, etcétera. Para contestarlas se requiere de un censo representativo del contenido de la Galaxia, cuantificar la estructura espacial a partir de las distancias, conocer las velocidades espaciales para determinar el campo gravitatorio y las órbitas estelares, así como caracterizar la composición química y la edad. ¿cuándo y cómo se formaron las Galaxias? ¿cuándo se formaron sus estrellas? ¿cómo afecta el medio ambiente a la formación y evolución de las galaxias?.

Hay una gran necesidad de dar respuesta a preguntas como las mencionadas anteriormente, que forman parte de los temas actuales de investigación en cosmología, evolución estelar o clasificación de objetos; por lo que se deben adoptar nuevas metodologías de análisis estadístico, que sin duda requieren un acceso masivo a grandes cantidades de observaciones o datos obtenidos a partir de las simualciones, y son sólo posibles si existen los archivos. Los volúmenes de datos están creciendo y se deben crear recursos para poder manipular, acceder, modificar, etc. dichos archivos.

El método tradicional de convertir los datos en conocimiento, consistía en un análisis e interpretación realizada manualmente. Esta forma de actuar, sin duda es lenta y cara. Además, con el actual crecimiento exponencial en volumen y complejidad de datos – tanto observacionales como teóricos – que se van generando en todo el mundo [Szalay et al., 2000], se tiene como desafío crear una herramienta actualizada, que permita a la comunidad astronómica manipular y analizar las cantidades de datos que ahora se pueden adquirir a niveles antes no imaginados.

Se ha impulsado la evolución de los sistemas de bases de datos por dos lados: uno debido al usuario pues ha demandado una serie de capacidades que se han ido incorporando en los sistemas de bases de datos. Por el otro lado, la tecnología, per se, ha hecho posible que algunas facilidades se conviertan en realidad.

La astronomía debe utilizar a la informática como una herramienta, al

menos, desde tres distintas perspectivas, las cuales son:

1. Que sirva al astrónomo como herramienta tecnológica, esto es, que permita agilizar los diferentes procedimientos de adquisición de datos.

2. Que sirva al astrónomo como soporte para la gestión y organización de la información.

3. Como metodología para el diseño de distintas aplicaciones, que sean capaces de gestionar los datos y extraer conocimiento útil a partir de la información.

1.1. Minería de Datos.

Gracias a los avances tecnológicos, es posible almacenar grandes volúmenes de datos. Entre las ventajas de contar con amplias bases de datos siempre está latente la posibilidad de descubrir información de interés, así como adquirir conocimiento mediante el análisis de dichos datos. Sin embargo, el volumen mismo de las bases de datos es con frecuencia una limitante para poder hacer análisis manuales, además con el crecimientos de las bases de datos es impensable seguir trabajando de esta manera.

La necesidad de analizar datos, así como la extracción de conocimiento no implícito en los mismos de forma automática, derivó en el nacimiento de una nueva disciplina: la **Minería de Datos** o en inglés data mining, donde los datos pasan de ser el producto generado por los diferentes procesos inherentes a la actividad desarrollada, a ser la materia prima, puesto que de estas cantidades de datos se extrae conocimiento útil.

La **Minería de Datos** es un proceso cuyo propósito radica –justamente– en descubrir, extraer y almacenar información relevante de las bases de datos. Se ha desarrollado para facilitar el manejo, extraer y almacenar información de diferentes y amplias bases de datos, así como poder analizar datos a través de programas de búsqueda e identificación de patrones, encontrar relaciones globales, diferentes tendencias, desviaciones o incluso entender algunos indicadores que aparentemente podrían parecer caóticos.

Por medio de la minería de datos, se cuenta con mecanismos útiles para la extracción de información, a partir de series extensas de datos. Lo que ha he-

cho que la minería de datos, sea una de las técnicas más utilizadas actualmente para analizar la información proveniente de las bases de datos. La minería de datos está fundamentada en varias disciplinas como la estadística, la visualización de datos, la computación paralela y distribuida.

1.2. Gestor de Bases de Datos.

Hoy en día existen diferentes sistemas gestores de bases de datos (**DBMS**), que son conjuntos de procesos que permiten el almacenamiento, modificación y extracción de la información en una base de datos, además de proporcionar herramientas para añadir, borrar, modificar y analizar los datos. Los usuarios pueden acceder a la información, usando herramientas específicas de interrogación y de generación de informes. Los **DBMS** también proporcionan métodos para mantener la integridad de los datos, para administrar el acceso de usuarios a los datos y para recuperar la información si el sistema se corrompe.

Un **DBMS** permite presentar la información de la base de datos en variados formatos. También puede incluir un módulo gráfico que permita presentar la información con gráficos y tablas. Hay muchos tipos de **DBMS** distintos, según manejen los datos y muchos tamaños distintos, según funcionen sobre computadoras personales y con poca memoria hasta sistemas que funcionan en grandes computadores (mainframes[13]) con sistemas de almacenamiento especiales. Generalmente, se accede a los datos mediante lenguajes de interrogación. Un **DBMS** permite controlar el acceso a los datos, asegurar su integridad, gestionar el acceso concurrente a ellos, recuperar los datos tras un fallo del sistema y hacer copias de seguridad.

1.3. PICASSO.

Las bases de datos astronómicas siguen creciendo, por esto se pensó en crear una herramienta que permita almacenar, analizar y manipular bases de datos

[13]Se le llama mainframe a una computadora que es capaz de realizar el procesamiento de datos complejos, se utilizan como sistemas centrales. Se caracterizan por una alta velocidad de ejecución de tareas individuales y una arquitectura diseñada para permitir el equilibrio de beneficios y un mayor nivel de seguridad. Un solo mainframe pueden reemplazar cientos de pequeños servidores físicos.

astronómicas. A esa herramienta que se generó, como producto de esta tesis, se le llamó **PICASSO**, en homenaje a **Pablo Picasso** el pintor.

PICASSO utiliza las herramientas existentes actuales para manipulación y gestión de bases de datos. A los datos de dichas bases, se les proporciona una jerarquía, que permite manipular dichas bases de datos y hacer búsquedas de datos, de manera más eficiente. Así pues, **PICASSO** es un conjunto de herramientas para el análisis de grandes bases de datos astronómicos.

Su cuerpo está dividido por tres grandes bloques

1. El **gestionador de Base de Datos**, que es el software dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que utilizaremos.

2. El conjunto de módulos que nos permite hacer el análisis de las bases de datos. El modulo más importante de éste bloque es el que nos **permite darle jerarquía a una tabla lineal**.
   Los diversos módulos están escritos en un lenguaje de acceso a bases de datos que permite especificar diversos tipos de operaciones y efectuar consultas, con el fin de recuperar – de una forma sencilla – la información de interés, así como también hacer cambios sobre ella.

3. El tercer bloque de **PICASSO** es una **herramienta gráfica interactiva** que permite hacer análisis y manipulación de datos tabulares.

Sin embargo **PICASSO** no es solamente estos tres grandes bloques, puesto que tiene la posibilidad de ser compatible y leer scripts en diferentes lenguajes como **C** o **python**, lo que permite incorporar un conjunto de diversas tareas, para ser tratadas en las bases de datos.

PICASSO tiene diversos módulos que fueron desarrollados para crear una solución a un par de problemas astronómicos específicos así como problemas computacionales concretos, pensando en las grandes bases de datos que se presentarán en un futuro casi inmediato, es decir que sea altamente escalable. En este trabajo de tesis, se presenta un ejemplo concreto de la potencia que tiene dicha herramienta computacional.

1.4. Estructura de la Tesis.

A continuación, se hablará de la estructura que tiene este trabajo de tesis.

En el **capítulo 2** se amplían algunos de los conceptos más importante sobre las bases de datos, la minería de las bases de datos y la importancia de la jerarquía en las bases de datos.
Se habla sobre el concepto de gestor de base de datos.
En la exposición de este capítulo, se enfatiza la importancia que tiene las diferentes estructuras de datos, en particular la estructura de árboles, ya que, justamente es con este tipo de estructura el que permite realizar búsquedas o sustituciones de elementos, de manera altamente eficiente en las bases de datos de tamaños muy grandes.

También se presenta el conjunto de rutinas que integran a la herramienta **PICASSO**, así como algunos ejemplos con bases de datos sintetícas de tamaños variados (desde miles hasta doscientos millones de datos con diferentes atributos).

En el **capítulo 3** se presentan algunos de los conceptos astronómicos utilizados en esta tesis y se discuten las propiedades generales que caracterizan a la población de galaxias, tales como la morfología, la distribución de brillos superficiales y colores.

En el **capítulo 4** se introducen los catálogos astronómicos que fueron utilizados para poner a prueba nuestra herramienta computacional: **PICASSO**. Los catálogos astronómicos utilizados, son catálogos de galaxias del Universo Local, que contienen información espectroscópica, fotmétrica así como información de la masa estelar al día de hoy de las galaxias.

En el **capítulo 5** se presentan y se discuten los resultados astronómicos, que se obtienen al poner a prueba nuestra herramienta computacional.

Como se menciona en el capítulo 4, se trabaja con diferentes catálogos astronómicos. Se plantean diferentes estrategias de cortes, para dichos catálogos, los cuales fueron aplicados a la muestra de galaxias que tienen infomación tanto fotométrica como espectroscópica para galaxias del Universo Local, así como la masa estelar calculada a partir de la Spectral Energy Distribution por sus siglas en inglés **SED** y que, además, tienen una extensión y brillo central bien defi-

nido por nuestros propios cortes para evitar licitantes provenientes del seeing, es decir, de la baja resolución espacial, del promedio del brillo de las galaxias, esto, con el fin de buscar resultados comparables en robustez a los estudios de galaxias del Universo Local publicados por Fisher & Drory [2011], Gadotti [2008] y Berg et al. [2014]. Dichos resultados son comparados también con predicciones teóricas de modelos semi-empíricos de de evolución de galaxias combinados con modelos semi-analíticos de crecimiento de bulbo [Zavala et al., 2012a].

Finalmente, en el **capítulo 6** se discuten los resultados –tanto astronómicos como computacionales – que se obtuvieron como resultado de este trabajo.
Se elaboran conclusiones generales, sobre algunas propiedades de los bulbos de las galaxias locales, haciendo referencia a otros estudios recientes y a nuestro trabajo a futuro con este tipo de estudios.

También, se discute, sobre las potencialidades a la vista de sondeos planeados como lo son, por ejemplo, el Large Synoptic Survey Telescope (LSST)[14] para el cuál se planea que en 10 años de operación obtenga cerca de 150,000 Terabytes de información. Es decir, la astronomía contemporánea, necesita herramientas específicas para lidiar con el **Tsunami Digital**.

La tesis, también consta de diferentes **apéndices**, que permitirán al lector familiarizarse con algunos de los conceptos computacionales que se utilizan a lo largo de ésta y, que aunque antes de utilizarlos se definieron, permite una lectura más fluida, también se presentan una serie de tablas, para que el lector se pueda familiarizar con las distintas cantidades y unidades de medición en cómputo.

[14] http://www.lsst.org/

2
Bases y Minería de datos

2.1. Introducción.

En el 2000 se puso en operación el Sloan Digital Sky Survey (**SDSS**)[1] [York et al., 2000], que en ese momento y desde su concepción fue el más ambicioso catastro en la historia de la astronomía observacional. En el momento en el que inició sus operaciones, hubo una explosión de datos astronómicos nunca antes vista. El **SDSS** es un proyecto de inspección del espacio que cubre un cuarto del cielo – del hemisferio norte – mediante espectros e imágenes en la región visible del espectro electromagnético, realizada con un telescopio de 2.5 metros de diámetro situado a 2,788 metros sobre el nivel del mar en el observatorio Apache Point situado en Nuevo México.

El **SDSS** ha obtenido imágenes y ha confeccionado un mapa cósmico tridimensional que contiene más de un 1,000,000 de galaxias y más de 200,000 espectros de cuásares, logró medir las posiciones y brillos absolutos de cientos de millones de objetos celestes, es un catálogo que en su última publicación contaba con más de 930,000,000 objetos y 668,054 espectros de estrellas [Ahn et al., 2012].

[1] http://www.SDSS.org

El **SDSS** ha ampliado desde el 2000 el número de inspecciones que se planeaba. En 2008 aparte de continuar con la labor original (**Sloan Legacy Survey**[2]) se amplió a:

- **SEGUE** (Sloan Extension for Galactic Understanding and Exploration, por sus siglas en inglés): Inspeccionó una región del cielo de 3,500 grados cuadrados. Obtuvo el espectro de 240,000 estrellas (con un promedio de velocidad radial típica de 10 km/s), para crear un mapa tridimensional detallado de la Vía Láctea. Sus datos revelaron la edad, composición y distribución de fase espacial de las estrellas dentro de varios componentes galácticos, proporcionando información crucial para la comprensión de la estructura, formación y evolución de nuestra Galaxia.

- **The Sloan Supernova Survey** fue una campaña que duró 3 meses, en la que se observó en repetidas ocasiones una región de 300 grados cuadrados, con el objeto de descubrir Supernovas tipo **Ia**, al final, se detectaron un poco más de 300.

En la actualidad, el **SDSS** está en una tercera fase del proyecto, que abarcará hasta finales del 2014, y que tiene como objetivo profundizar en el conocimiento de:

- Energía oscura y parámetros cosmológicos.

- Estructura, dinámica y evolución química de la Vía Láctea.

- Estudios de sistemas planetarios.

Además del **SDSS**, que hasta la fase 2 (DR7)[3] [Abazajian et al., 2009] había recaudadado un poco más de 60 TB de información, entre distintos catálogos (algunos de ellos en formato SQL – que están en línea– así como también imágentes en formato FITS, imágenes en formato jpeg, etc.), también se tienen proyectos como 2dF Galaxy Redshift Survey **2dFGRS**[4] [Colless et al., 2001], 6dF Galaxy Survey **6dFGS**[5] [Jones et al., 2004], Visible and Infrared

[2]La exploración cubre más de 7,500 grados cuadrados de la Región Galáctica sur, con datos de casi 2 millones de objetos y espectros de más de 800,000 galaxias y 100,000 quásaress. Esta información de la posición y la distancia de los objetos permitió investigar por primera vez la estructura a gran escala del Universo con sus vacíos y filamentos.

[3]http://www.**SDSS**2.org/dr7/

[4]http://www2.aao.gov.au/2dFGRS/

[5]http://www.aao.gov.au/6dFGS/

Survey Telescope for Astronomy **VISTA**[6] [McPherson et al., 2006], en el infrarrojo UKIRT Infrared Deep Sky Survey **UKIDSS**[7] [Dye et al., 2006], el Large Synoptic Survey Telescope **(LSST)**[8] [Tyson, 2002] etc., que redituarán en cantidades de datos impensables – de obtener y de utilizar – hasta hace unos años.

Con todos los telescopios activos en el 2003 el contenido de bases de datos astronómicas era de cientos de terabytes. En el 2005 se estimaba que 1 TB era la tasa de recolección de datos astronómicos diaria. En el 2009, tan sólo el **SDSS** tenía más de 70 TB con información de millones de objetos astronómicos, entre los cuales hay cúmulos de galaxias, galaxias, supernovas, nebulosas, cúmulos de estrellas, estrellas, asteroides, etc. [Ivezic et al., 2010]. Si sumamos toda la información que han recabado estos telescopios, más la información de los proyectos que están por venir y cuyos datos alcanzarán un nivel de detalle aún mayor, como **GAIA**[9] [Jordan, 2008], **Pan-STARRS** [Kaiser et al., 2002] que en 10 años se calcula que obtenga 60 PB (400 veces más grande que el **SDSS**), el Large Synoptic Survey Telescope **LSST**[10] [Axelrod, 2006] que producirá del orden de cientos de megabytes (MB) por segundo, es decir 20 TB por noche; se espera que en una semana adquiera, en cantidad, la misma que todo el tiempo que lleva funcionando el **SDSS** (es decir, desde el 2000). Con todos estos telescopios, obtenemos en los próximos años, decenas de petabytes en información astronómica, que deberá ser almacenada en diferentes bases de datos. Podemos deducir que trabajar y analizar la información astronómica proveniente de telescopios será un gran reto.

El panorama que se nos presenta en la actualidad y a un futuro cercano (menos de 10 años), nos indica que los datos recolectados crecen – se espera que en algunos años más, la tasa de recolección sea de de casi 100 terabytes diarios [Bell et al., 2007] – por esto se debe pensar en nuevos métodos de análisis, organización de datos y manejo eficiente para el quehacer científico en estas escalas (cientos de petabytes).

De la necesidad de manipular los datos y extraer información de manera eficiente (y escalable) surge **PICASSO**. **PICASSO** tiene diversos módulos que

[6] http://www.vista.ac.uk/
[7] http://www.ukidss.org/
[8] http://www.lsst.org/lsst/
[9] http://www.rssd.esa.int/
[10] http://www.lsst.org/

fueron desarrollados para crear una solución a un par de problemas astronómicos específicos, así como problemas computacionales concretos, pensando en las grandes bases de datos que se presentarán en un futuro casi inmediato, es decir, que sea altamente escalable. Así pues, **PICASSO** es una herramienta que desarrollé y que consiste en utilizar diferentes herramientas que ya se conocen actualmente y se utilizan para gestionar y manipular bases de datos, como **MySQL**, **TopCat**, etc. y a los datos contenidos en las tablas de las bases, se les proporciona una jerarquía a partir de diferentes lenguajes, como **SQL**, **C** o **python**. Lo que permite manipular dichas bases de datos y hacer búsquedas de datos de manera más eficiente, es decir, que permite crear una nueva herramienta a partir de **la reutilización de código**, con la que se pueda manejar de manera eficiente las bases de datos tanto actuales, como las que se programa que se tendrán para la siguiente década. Así pues, **PICASSO** es un conjunto de herramientas para el análisis de grandes bases de datos astronómicos que se basa en la reutilización de código.

La herramienta **PICASSO** se ha creado pensando en múltiples usos y en diferentes tópicos astronómicos. **PICASSO** permite correlacionar de manera eficiente bases de datos con propiedades de objetos astronómicos generadas en grandes catastros observacionales o en catastros sintéticos. Dicha herramienta puede ampliarse tanto como el usuario o diferentes usuarios necesiten. A lo largo de este trabajo de tesis se hablará de las posibilidades que tiene **PICASSO**.

Actualmente, las bases de datos son uno de los sistemas de gestión de la información más extendido, pues mejora notablemente la manera de interactuar con los datos, es decir almacenarlos, organizarlos, manipularlos y analizarlos.

Para poder explicar con detalle cómo funciona **PICASSO** de manera más extendida se presenta una introducción a

- Reutilización de código
- Bases de datos
- Minería de bases de datos
- Sistemas de gestión de bases de datos,
- Jerarquización de las tablas contenidas en las bases de datos.

2.2. Reutilización de código.

La reutilización de código se refiere al comportamiento y a las técnicas que garantizan que una parte o la totalidad de un programa informático existente se pueda emplear en la construcción de otro programa. De esta forma se aprovecha el trabajo anterior, se economiza tiempo, y se reduce la redundancia. La manera más fácil de reutilizar código es copiarlo total o parcialmente desde el programa antiguo al programa en desarrollo. Este proceso se conoce como abstracción. La abstracción puede verse claramente en las bibliotecas de software, en las que se agrupan varias operaciones comunes a cierto dominio para facilitar el desarrollo de programas nuevos. Hay bibliotecas para convertir información entre diferentes formatos conocidos, acceder a dispositivos de almacenamiento externos, proporcionar una interfaz con otros programas. Para que el código existente se pueda reutilizar, debe definir alguna forma de comunicación o interfaz. Esto se puede dar por llamadas a una subrutina, a un objeto, o a una clase.

En astronomía, la reutilización de código, se vuelve bastante valiosa, debido a que la preparación del astrónomo no está basada en las ciencias de la computación. Existen varios programas que son base de distintas aplicaciones que se pueden utilizar en astronomía, en particular, utilizar programas que gestionen las bases de datos, aunado a programas o scripts que nos permitan darle una jerarquía, a los datos de las tablas contenidos en distintas bases de datos, para así poder manipularlas de manera mucho más eficiente y logrando escalabilidad.

2.2.1. Introducción.

Hoy en día se busca la producción de sistemas software de calidad de forma ágil, eficiente y sistemática [Müller et al., 1993]. Se han dirigido muchos esfuerzos de investigación y desarrollo en la búsqueda de técnicas, metodologías, herramientas y procedimientos que de alguna manera permitan mejorar el desarrollo de los sistemas.

La **reutilización de software** es una excelente manera de ahorrar costos y esfuerzos de desarrollo, motivo por el cual, ha surgido recientemente un interés por crear y utilizar modelos de procesos para desarrollo de software que contemplen la reutilización. El concepto de reutilización, no es nuevo, sin embargo, recientemente ha emergido, como elemento central el concepto de **componente de software reutilizable** [McClure, 2001]. Las características

de los componentes de software reutilizables han obligado a crear nuevos modelos de desarrollo de software basados en la reutilización [Davis & Bersoff, 1991]. Estos procesos buscan la solución a un problema. Esta solución no es única, por lo que los métodos deben ir evolucionando y adaptándose a las nuevas y diferentes necesidades de los diferentes tipos de usuarios y los nuevos conceptos de la Ingeniería de Software [McClure, 2001].

2.2.2. Conceptos básicos de reutilización.

La reutilización es uno de los conceptos más simples y antiguos en la programación, y es algo que a menudo no es muy utilizado. En astronomía este concepto, tiene un punto a favor extra a parte de que permite ahorrar tiempo y dinero, puesto que el astrónomo no es un experto en cómputo, pero el astrónomo actual y con mayor énfasis cada vez, necesita de las herramientas de última generación para encarar, a lo que llamamos, el **tsunami digital**.

2.2.3. Beneficios de la reutilización.

2.2.3.1. Tiempo y costo.

La reutilización de software es una excelente manera de ahorrar costos y esfuerzos de desarrollo [Shioji et al., 2012]. De forma ideal, el tiempo de desarrollo es reducido debido a que los componentes reutilizables relevantes pueden ser aplicables al proyecto dado en un marco de tiempo menor que re-desarrollar desde cero Haefliger et al. [2008]. Como el tiempo y costo de desarrollo tienen una correlación positiva, reducir los tiempos de desarrollo trae un ahorro en los costos del proyecto. Esto puede llevar a proyectos y productos más baratos, y más utilidades a la organización.

El tiempo, en el que se desarrolla desde cero una herramienta computacional, puede ser mayor al tiempo en el que la herramienta sea necesaria. Un panorama real es aquel en el que podría tardar más tener la herramienta habilitada para trabajar con los datos que los datos mismos.

2.2.3.2. Calidad.

Cualquier optimización, refactorización y pruebas hechas en componentes reutilizables ya han sido completados. Todas las lecciones aprendidas al producir el componente están implícitamente incluidas dentro de esto y son au-

tomáticamente llevadas al siguiente proyecto. Consecuentemente, los proyectos desarrollados son de una mayor calidad [Han, 2011].

La reutilización efectiva puede reducir la densidad de defectos de 5 a 10 veces [Shioji et al., 2012]. Es más, debido a los ahorros en tiempo, los desarrolladores pueden invertir el mismo en el software nuevo que necesita ser producido, incrementando potencialmente su calidad o cualidades, adecuándolas a las necesidades actuales.

2.2.3.3. Estandarización.

Los componentes dentro de un dominio dado (área de aplicación), requieren cierta clase de estandarización para hacerlos compatibles con otros componentes [Davis & Bersoff, 1991]. Esto puede llevar a hacer estándares de interfaces de componentes, así como de código y documentación. Estas entidades animan a mejores prácticas de desarrollo [Müller et al., 1993].

2.2.4. Utilizando los principios de reutilización de software, se crea: PICASSO.

La **reutilización de software** resulta una excelente manera de ahorrar costos y esfuerzos de desarrollo.

Usando el principio de reutilización de software se crea **PICASSO**. **PICASSO** está pensado de la siguiente forma:

- Se utiliza un gestor de bases de datos **MySQL**.
- Se diseñan e implementan diferentes algoritmos en diversos lenguajes **SQL**, **C** y **Python**, con los cuales se le puede proporcionar una estructura de jerarquía a los datos de las tablas que conforman las bases de datos, lo que permite que el análisis y manejo de las bases de datos sea mucho más eficiente e incluso puede ser escalable, que si sólo se usara **MySQL**.
- Se utiliza un visualizador estándar llamado **TopCat**, el cual también permite gestionar tablas, pero para una cantidad grande de datos (tablas con más de 300 millones de renglones y unas pocas decenas de columnas – donde cada renglón representa a un objeto astronómico – colapsa, sin embargo, si se trabaja con las submuestras generadas a partir de lo descrito anteriormente **TopCat** resulta muy eficiente.

2.3. Bases de Datos.

El término de *Bases de Datos* aparece por primera vez en los años sesentas, dentro del marco de un congreso de análisis computacional [Yao, 1985]. Desde ese entonces y hasta ahora, las *bases de datos* son uno de los sistemas de gestión de la información más utilizados y sus avances, tanto en el campo teórico como en la realización de programas, ha ido variando, puesto que se ha ido avanzando y mejorando en términos no sólo de su funcionalidad, sino también en términos de su utilidad a razón de las exigencias actuales.

2.3.1. Definición de Bases de Datos.

Una *base de datos* es una colección o depósito de datos organizados y estructurados, según un determinado modelo de información que refleja, no sólo los datos en sí mismos, sino también las relaciones que existen entre ellos [King & McLeod, 1985]. Una base de datos se diseña con un propósito específico y debe ser organizada con una lógica coherente [Yao, 1985]. Los datos podrán ser compartidos por distintos usuarios y aplicaciones, pero deben conservar su integridad y seguridad al margen de las interacciones. La definición y descripción de los datos ha de ser única para maximizar la independencia en su utilización [King & McLeod, 1985].

Los datos son independientes de los programas que los usan, la definición y las relaciones entre los datos se almacenan junto a éstos y se puede acceder a los datos de diversas formas [Date, 2004].

Debido a la importancia que tienen las interrelaciones entre los datos para poder realizar un diseño de su estructura, es necesario que la base de datos sea capaz de almacenar correctamente todas estas relaciones, así como los atributos de los datos, entidades y restricciones al modelo [Teorey et al., 1986]. El modelo, se debe de definir previamente al realizar el diseño y posteriormente, al introducir los datos. Ésto, es una de las diferencias fundamentales entre las bases de datos y los ficheros[11] utilizados hasta los años sesentas.

[11] Los archivos también denominados ficheros (*file*) son una colección de información (datos relacionados entre sí), se encuentra almacenada como una unidad.

2.3.2. Ventajas de las Bases de Datos.

Son muchas las ventajas que las bases de datos y los sistemas gestores de ellas ofrecen. A continuación se va a enfatizar y profundizar en algunas de estas ventajas.

- Obtener más información de la misma cantidad de datos.- La base de datos facilita al usuario, obtener la información de manera ordenada y así contar con la mayor información, debido a la facilidad que provee esta estructura para analizar datos.

- Compartir los Datos.- Varios usuarios pueden compartir datos si están autorizados. Esto implica, que si un dato cambia, todos los usuarios que pueden acceder a ese dato, verán inmediatamente el cambio efectuado.

- Redundancia controlada.- Debido al sistema tradicional de archivos independientes, al compartirlo con más de un usuario, los datos se podrían duplicar constantemente, lo cuál, crearía un problema de sincronización cuando se actualiza un dato en un archivo.

- Consistencia.- Al controlarse la redundancia, cuando se actualiza un dato, todos los usuarios autorizados de la Base de Datos pueden ver dicho cambio, independientemente de que estén trabajando en distintos sistemas.

- Integridad.- La base de datos tiene la capacidad de validar ciertas condiciones cuando los usuarios aceptan datos o rechazar entradas.

- Seguridad - Se tiene control de los datos. La Base de Datos provee mecanismos que le permiten crear niveles de seguridad para distintos tipos de usuarios. Desde no poder modificar un archivo hasta poder cambiarlo enteramente.

- Flexibilidad y rapidez al obtener datos.- El usuario puede fácilmente obtener información de la Base de Datos con tan solo escribir breves oraciones, dicha flexibilidad, es aún mayor al darle jerarquía a la base de datos – de esto, se hablará con mayor detalle en la sección *2.6*.

- Mejora el mantenimiento de los programas.- Debido a que los datos son independientes de los programas, si ocurre un cambio en la estructura de una tabla (archivo), el código no se afecta.

- Independencia de los Datos.- Los datos pueden modificarse pero, no se tiene que modificar los programas.

Sin embargo, en todos los sistemas de gestión de información existen inconvenientes, en particular el tamaño; al proveer todas las ventajas anteriomente nombradas, el sistema gestor de base de datos requiere de mucho espacio en disco duro y también, requiere de mucha memoria principal (RAM)[12].

2.4. Sistema Gestor de Bases de Datos (DBMS).

Un Sistema Gestor de Base de Datos (DataBase Management System **DBMS** por sus siglas en inglés), es el software que permite a los usuarios introducir, organizar, procesar, describir, administrar y recuperar la información almacenada en las bases de datos. Consiste de una base de datos y un conjunto de aplicaciones (programas) para tener acceso a ellos.

En estos sistemas gestores se proporciona al usuario un conjunto coordinado tanto de programas, procedimientos y lenguajes que permiten a los distintos usuarios realizar sus tareas habituales con los datos, garantizando además la seguridad de los mismos.

El objetivo primordial de un **DBMS** es crear un ambiente en el que sea posible almacenar y recuperar información de forma eficiente y conveniente. Su éxito reside en mantener la seguridad e integridad de los datos.

2.4.1. Características Generales de los DBMS.

Algunas de las características comunes que tienen los distintos gestionadores de Bases de Datos son:

- Manipular los datos siguiendo exclusivamente las órdenes de los usuarios.
- Acceso a las bases de datos de forma simultánea por varios usuarios y/o varias aplicaciones.
- Permitir definir diferentes usuarios, así como las restricciones de acceso para cada uno de ellos.

[12]RAM: Random Access Memory, por sus siglas en inglés.

- Seguridad de la base de datos y de los datos mismos, en forma de permisos y privilegios, es decir, determinados usuarios tendrán permiso para consulta o modificación de determinadas tablas. Esto permite compartir datos, sin que peligre la integridad de la base de datos o protegiendo determinados contenidos.

2.4.2. Arquitectura de los sistemas gestores de Bases de Datos.

En el años de1975, el comité ANSI-SPARC (American National Standard Institute - Standards Planning and Requirements Committee, por sus siglas en inglés) propuso una arquitectura de tres diferentes niveles para los **DBMS** cuyo objetivo principal radicaba en separar los programas de aplicación de la Base de Datos física [Han, 2011].

Se definen tres distintos niveles de abstracción, los cuales son:

- **Nivel interno** o también conocido como nivel físico. Es el nivel más cercano al almacenamiento físico. Tal y como cualquier dato se almacena en la computadora. Describe la estructura física de la DB mediante un esquema interno. Este esquema se especifica con un modelo físico y describe los detalles de cómo se almacenan físicamente los datos: los archivos que contienen la información, su organización, los métodos de acceso a los registros, los tipos de registros, la longitud, los campos que los componen, etcétera [Han, 2011].

- **Nivel externo** o también conocido como el nivel de visión. Este nivel es el más cercano a los usuarios. Donde se describen varios esquemas externos o vistas de usuarios. Cada esquema describe la parte de la DB que interesa a un grupo de usuarios [Han, 2011].

- **Nivel conceptual**. Este nivel describe la estructura de toda la DB para un grupo de usuarios mediante un esquema conceptual. Este esquema describe las entidades, atributos, relaciones, operaciones de los usuarios y restricciones, ocultando los detalles de las estructuras físicas de almacenamiento. Representa la información contenida en la DB [Han, 2011].

  En la Figura 2.1 se representa de manera esquemática los tres diferentes niveles de abstracción que se presenta en la arquitectura ANSI.

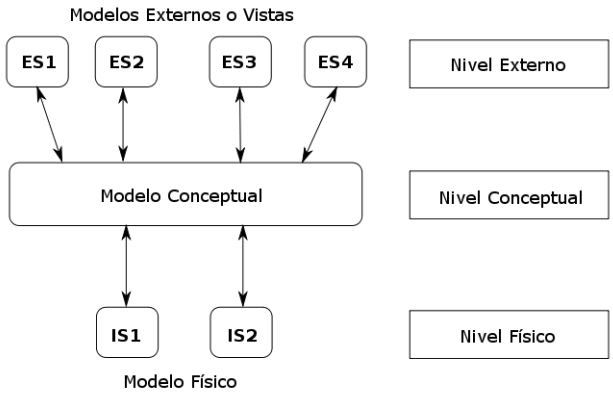

Figura 2.1: Niveles de abstracción en la arquitectura ANSI.

Esta arquitectura, describe los datos en tres distintos niveles de abstracción. Se debe tomar en cuenta que, en realidad los únicos datos que existen, están a nivel físico, almacenados en discos u otro tipo de dispositivo. Los **DBMS** basados en este tipo de arquitectura, permiten que cada grupo de usuarios pueda hacer referencia a su propio esquema externo, mientras que el **DBMS** debe ser capaz de transformar cualquier petición del usuario (que cumpla adecuadamente al esquema externo) en una petición expresada en términos de esquema conceptual, para así, finalmente ser una petición expresada en lo que se llama esquema interno y que se procesará sobre la DB almacenada.

El proceso de transformar peticiones y resultados de un nivel a otro se denomina **correspondencia** o **transformación**. El **DBMS** tiene que ser capaz de interpretar una solicitud de datos y realiza cada uno de los siguientes pasos:

- Un usuario solicita datos (crea una consulta).
- El **DBMS** verifica el esquema externo para ese usuario.
- El **DBMS** acepta el esquema externo para ese usuario.
- El **DBMS** transforma la solicitud al esquema conceptual.
- El **DBMS** verifica el esquema conceptual.
- El **DBMS** acepta el esquema conceptual.

- El **DBMS** transforma la solicitud al esquema físico (o interno).
- El **DBMS** selecciona la o las tablas implicadas en la consulta.
- El **DBMS** ejecuta la consulta.
- El **DBMS** transforma del esquema interno al esquema conceptual
- El **DBMS** transforma del esquema conceptual al esquema externo.
- Finalmente, el usuario es capaz de ver/interactuar con los datos solicitados.

Para una DB específica sólo hay un esquema interno y uno conceptual, pero puede haber varios esquemas externos definidos para uno o para varios usuarios.

Se ha elegido a **MySQL** como el gestor de bases de datos en esta tesis.

2.4.3. MySQL.

MySQL[13] es un sistema gestor de bases de datos (**DBMS**) muy conocido, ampliamente usado y que ha ganado un lugar importante, como uno de los sistemas gestores más utilizados por su simplicidad y rendimiento. Está disponible para múltiples plataformas [Date., 2006].

2.4.3.1. Caracterśticas generales del gestor de Bases de Datos MySQL.

MySQL tiene una serie de características positivas que se listan a continuación:

- Está desarrollado en **C/C++**.
- Se distribuyen ejecutables para cerca de diecinueve plataformas diferentes.
- La API se encuentra disponible en **C**, **C++**, **Eiffel**, **Java**, **Perl**, **PHP**, **Python**, **Ruby** y **TCL**.
- Está optimizado para equipos de múltiples procesadores.
- Es muy destacable su velocidad de respuesta.

[13] http://dev.mysql.com/

- Se puede utilizar como cliente-servidor o incrustado en aplicaciones.
- Cuenta con un variado conjunto de tipos de datos.
- Soporta múltiples métodos de almacenamiento de las tablas, con prestaciones y rendimiento diferentes para poder optimizar el **DBMS** a cada caso concreto.
- Su administración está basada en usuarios y privilegios, que permiten o no modificar las distintas tablas.
- Se tiene constancia de casos en los que maneja cientocincuenta millones de registros, sesenta mil tablas y cinco millones de columnas. Este es el límite hasta ahora conocido, sin embargo, en los foros de desarrolladores de **MySQL** hay personas que dicen que se está ya trabajando para que este límite crezca en las siguientes versiones de **MySQL**.
- Es altamente confiable en cuanto a estabilidad se refiere.

Otra de las características que tiene **MySQL** es el tipo de esquema de tabla que **MySQL** soporta, además que permite la incorporación de distintos tipos de esquemas definidos por el usuario. El esquema se define al inicio, pero incluso se puede modificar posteriormente.

El modelo relacional para la gestión de una base de datos es el modelo que más se utiliza en la actualidad para modelar problemas reales y administrar datos de manera dinámica [Silberschatz et al., 2008].

2.4.4. Base de datos relacional.

Este tipo de tablas fue definido por Edgar Frank Codd a principios de los años setentas [Yao, 1985].

En una base de datos relacional, los datos se muestran en forma de tablas y relaciones. En las bases de Codd, se tiene definidos los siguientes objetivos para este modelo relacional:

- Independencia física. La forma en que se almacenan los datos, no debe influir en su manipulación lógica.
- Independencia lógica. Las aplicaciones que utilizan la base de datos no deben ser modificadas, al modificarse elementos de la base de datos.

- Uniformidad. Las estructuras lógicas siempre tienen una única forma conceptual (en este caso: las tablas)

- Sencillez.

Las **tablas** se representan gráficamente como una estructura rectangular formada por **filas** y **columnas**.

Cada una de las **columnas** almacena información sobre una propiedad determinada de la **tabla** (se le llama también campo).

Cada **fila** a su vez, posee una ocurrencia o ejemplar de la instancia o relación representada por la tabla (a las filas se las llama también registro).

Atributo 1	**Atributo 2**	**Atributo 3**	...	**Atributo n**	
valor 1,1	valor 1,2	valor 1,3	...	valor 1,n	← Tupla 1
valor 2,1	valor 2,2	valor 2,3	...	valor 2,n	
...	...	...	...	...	
valor m,1	valor m,2	valor m,3	...	valor m,n	← Tupla m

Cuadro 2.1: Tabla relacional

En el cuadro 2.1, se presenta una tabla relacional. Dicha tabla tiene los valores: valor 1,1, valor 1,2, valor 1,3, ..., valor 1,n; los cuáles forman la tupla 1, también valor m,1, valor m,2, valor m,3, ..., valor m,n, que forman la tupla m.

2.4.4.1. Terminología Relacional.

De manera muy intuitiva en la sección anterior, se definieron ya algunos conceptos básicos sobre la estructura relacional. A contiuación se presentan algunas definiciones de algunos términos que se tienen que tener en cuenta, al crear una tabla relacional

- **Tupla**. Cada fila de la tabla (es decir, cada uno de los ejemplares que la tabla representa).

- **Atributo**. Es cada columna de la tabla.

- **Grado**. Número de atributos de la tabla.
- **Cardinalidad**. Número de tuplas de una tabla.
- **Dominio**. Conjunto válido de valores representables por un atributo.

2.4.4.2. Las 12 reglas de Codd.

Como ya se mencionó anteriormente, las tablas de tipo relacional, fueron definidas por Edgar Frank Codd [Yao, 1985], quién, preocupado por los productos que decían ser sistemas gestores de bases de datos relacionales **(RDBMS)** sin serlo, publica las **12 reglas que debe cumplir todo DBMS para ser considerada relacional** [Codd, 1979].

A continuación se presentan las 12 reglas:

1. **Información**. Toda la información de la base de datos debe estar representada explícitamente en el esquema lógico. Es decir, todos los datos deben estar en las tablas.

2. **Acceso garantizado**. Todo dato debe ser accesible, sabiendo el valor de su clave[14] y el nombre de la columna o atributo que contiene el dato.

3. **Tratamiento sistemático de los valores nulos**. El **DBMS** debe permitir el tratamiento adecuado para estos valores, antes de este tipo de tablas, históricamente, se presentaron demasiados problemas al tratar con este tipo de valores.

4. **Catálogo en línea basado en el modelo relacional**. Los metadatos deben de ser accesibles usando un esquema relacional.

5. **Sublenguaje de datos completo**. Al menos debe de existir un lenguaje que permita el manejo completo de la base de datos. Este lenguaje, por lo tanto, debe permitir realizar cualquier operación definida.

6. **Actualización de vistas**. El **DBMS** debe encargarse de que las vistas muestren – siempre – la última información.

[14] Las claves se refieren a un atributo o un conjunto de atributos que permiten identificar unívocamente un registro.

7. **Inserciones, modificaciones y eliminaciones de dato nivel**. Cualquier operación de modificación, debe actuar sobre conjuntos de filas, nunca deben actuar registro a registro.

8. **Independencia física**. Los datos deben de ser accesibles desde la lógica de la base de datos, aún cuando se modifique el almacenamiento.

9. **Independencia lógica**. Los programas, no deben verse afectados por cambios en las tablas.

10. **Independencia de integridad**. Las reglas de integridad, deben almacenarse en la base de datos (en el diccionario de datos), no en los programas de aplicación.

11. **Independencia de la distribución**. El sublenguaje de datos, debe permitir que sus instrucciones funcionen igualmente en una base de datos distribuida[15] que en una que no lo es.

12. **No subversión**. Si el DBMS posee un lenguaje que permite el recorrido registro a registro, éste no puede utilizarse para incumplir las reglas relacionales.

En este trabajo de tesis, todas las tablas, se manejaron como tablas contenidas en bases de datos de tipo relacional, como ya se dijo se utiliza **MySQL** como sistema gestor de bases de datos. Hasta ahora, no se ha hablado del lenguaje que se utiliza en este gestor, el cuál es **SQL** Structured Query Language por sus sigas en inglés. A continuación se hablara de él.

2.4.5. Lenguaje de gestión de Bases de Datos.

En 1980 **SQL** (Structured Query Language) se hace el lenguaje estándar para las bases de datos. **SQL** es un lenguaje estándar internacional, comúnmente aceptado por los fabricantes de sistemas gestores de bases de datos [Samuel & Pedersen, 2004].

El **SQL** trabaja con estructura **cliente/servidor** sobre una red de computadoras. El **cliente** es el que inicia la consulta; el **servidor** es que atiende esa

[15]Una Base de Datos Distribuida (**DDB**) es una colección de datos distribuidos en diferentes nodos de una red de computadoras. Cada sitio de la red es autónomo, puede ejecutar aplicaciones locales y al menos una aplicación global, lo cual requiere el acceso a datos, ubicados en varios sitios, usando un subsistema de comunicación [Ceri & Pelagatti, 1984].

consulta.

El **cliente** utiliza toda su capacidad de proceso para trabajar; se limita a solicitar datos al **servidor**, sin depender para nada más del exterior. Estas peticiones y las respuestas son transferencias de textos que cada **cliente** se encarga de sacar por pantalla, presentar en informes tabulados, imprimir, guardar, etc., dejando el servidor libre [Chen et al., 2001].

El **SQL** permite:

- Definir una base de datos mediante tablas.
- Almacenar información en tablas.
- Hacer consultas en la base de datos.
- Seleccionar la información que sea necesaria de la base de datos.
- Borrar, insertar y modificar datos en las tablas.
- Realizar cambios en la información y estructura de los datos.
- Combinar y calcular datos para conseguir la información necesaria.

SQL es el lenguaje de comunicación entre el programa cliente y programa servidor.

SQL se puede emplear para:

- Crear y modificar la estructura de una tabla de datos.
- Seleccionar información de una tabla.
- Añadir datos a una tabla.
- Introducir información en una tabla.
- Realizar consultas entre tablas con campos comunes.

2.4.6. ¿Porqué utilizamos como gestionador de Bases de Datos a MySQL?.

MySQL, como ya se mencionó es un gestor de base de datos sencillo de usar, rápido, gratuito: e incluso, cualquier persona puede descargarlo en internet.

Las características no mencionadas anteriormente, que tiene **MySQL** son:

- Es una base de datos relacional. Es decir, es un conjunto de datos que están almacenados en tablas, entre las cuales se establecen relaciones matemáticas y lógicas para manejar, acceder y gestionar los datos, para esto último se usa el lenguaje estándar de programación **SQL** (Structured Query Language).
- Es una base de datos muy rápida, segura y fácil de usar. Gracias a la colaboración de muchos usuarios, la base de datos se ha ido mejorando y optimizándose en velocidad.

Cuenta con las características típicas y genéricas de las bases mencionadas anteriormente además de:

- Controlar la concurrencia y las operaciones asociadas a la recuperación de los fallos.
- Potencia: **SQL** es un lenguaje muy potente para consulta de bases de datos.
- Portabilidad: **SQL** es también un lenguaje estandarizado, de modo que las consultas hechas usando **SQL** son fácilmente portables a otros sistemas y plataformas. Esto, unido al uso de **C/C++**, **python**, etc. proporciona una portabilidad enorme.

Vale la pena, resaltar las ventajas adicionales:

- **Escalabilidad**: es posible manipular bases de datos enormes, del orden de seis mil tablas y alrededor de cincuenta millones de registros, y hasta 32 índices por tabla.
- **MySQL** está escrito en **C** y **C++** y probado con multitud de compiladores

- Dispone de **APIs** para muchas plataformas diferentes. **API** (del inglés **A**pplication **P**rogramming **I**nterface) es una serie de servicios o funciones que el Sistema Operativo ofrece al programador.
- Conectividad: es decir, permite conexiones entre diferentes máquinas con distintos sistemas operativos.
- Permite manejar multitud de tipos para columnas.
- Permite manejar registros de longitud fija o variable.

La versión de **MySQL** que se ha utilizado durante la redacción de este trabajo de tesis es la 5.6.17 la última versión estable en ese momento, y que además se encuentra disponible para diversos sistemas operativos, aunque no tendría que habrá ningún problema en ejecutarl los programas tal y como están estructurados en versiones anteriores, hasta la 4.17.

2.5. Exploración de datos y estudios estadísticos.

Entre los análisis estadísticos típicos que se pueden derivar de tener un gran conjunto de datos, podemos contar la creación de muestras uniformes, ensamblado de submuestras relevantes, descartar datos erróneos, etc. En el caso de conjuntos masivos de datos, podemos explorarlos de forma manual o automática, sin embargo, para la cantidad de datos que se preveé que habrá – debido a los diferentes telescopios – hacerlo manualmente cada vez será más complicado.

El concepto de **estructura** en informática se utiliza con mayor frecuencia para referirse a las estructuras de datos – así se hará a lo largo de este trabajo de tesis – sin embargo, se aplica también a los lenguajes de programación y en general a las diferentes aplicaciones que se utilizan.

2.6. Estructura de datos.

Se considera una estructura de datos como un conjunto de variables, quizá de tipos distintos, que se relacionan entre sí y que se pueden operar como un to-

do, esto implica un conjunto de celdas[16] en las que se puede almacenar los datos.

El componente básico de la estructura de datos es la celda, las estructuras de datos se implementan a través de los lenguajes y son un modelo que caracteriza y permite almacenar y utilizar una determinada organización de datos [Frakes & Baeza-Yates, 1992].

Las estructuras de datos son fundamentales para el manejo de información y el desarrollo de sistemas. Las diferentes maneras como se relacionan los datos originan conformaciones que son estructuras de datos de mayor complejidad.

2.6.1. Relaciones entre elementos.

Las relaciones entre los elementos de una estructura establecen la conexión física o lógica entre los distintos elementos que componen la estructura. Los principales tipos de relación son:

- De precedencia: permite representar la secuencia u orden de los elementos.
- De equivalencia: con esta relación es posible equiparar o igualar elementos.
- De jerarquía: permite indicar niveles de importancia entre los elementos.
- De pertenencia: determina que algunos elementos están incluidos en otros elementos, conformando de esta manera estructuras más complejas.
- De adyacencia: permite representar la igualdad de importancia entre los elementos de la estructura.

Es importante destacar que una estructura puede estar constituida por un conjunto de estructuras elementales o básicas, las cuáles son válidas entre estructuras.

2.6.2. Funciones básicas para estructuras.

Para disponer de una estructura de datos se necesita una serie de algoritmos que ejecuten las tareas fundamentales, los cuales reciben el nombre de funciones básicas y se enumeran a continuación:

[16]Celda: se utiliza para indicar el espacio de memoria que un equipo informático destina para almacenar un dato.

- Funciones constructoras: Se encargan de crear la estructura; definiendo las características, la delimitación, las relaciones, y asignando el espacio correspondiente, dejando la estructura a disposición del desarrollador para que proceda a colocar la información.

- Funciones para acceso: Estas facilitan la llegada de un elemento perteneciente a la estructura; la función puede ser simple o bastante compleja, dependiendo del tipo de estructura. Las formas de acceso se pueden considerar directas cuando a partir de un parámetro o una dirección se trata de encontrar el valor correspondiente, y son inversas en el caso contrario.

- Funciones destructoras: Se encargan de devolver al sistema los recursos asignados a la estructura de datos.

2.6.3. Operaciones.

Sobre una estructura de datos se puede efectuar diferentes tipos de operaciones, entre las más importantes están las siguientes:

- Navegar por la estructura o recorrido todos los elementos de la estructura. Esta es una operación básica y garantiza que se puede recuperar la información almacenada.

- Búsqueda, permite determinar si un elemento se encuentra o no en la estructura.

- Consulta de la información, permite obtener información de uno o más elementos de la estructura.

- Copia parcial o total, es aquella operación mediante la cual se puede obtener total o parcialmente una estructura con características similares a la original.

- Prueba, permite determinar si uno o varios elementos cumplen determinadas condiciones.

- Modificación, permite variar parcial o totalmente el contenido de la información de los elementos de la estructura.

- Inserción, es la operación mediante la cual se incluye un nuevo elemento en la estructura.

- Eliminación, es la operación que permite suprimir elementos de la estructura.

2.6.4. Estructura de datos con base en la relación entre los elementos.

Para poder establecer una clasificación bastante básica, se necesita tener en cuenta cuál es el tipo de relación que existe entre los elementos que conforman las estructuras de datos.
Por ejemplo, una primera abstracción consiste en considerar que los elementos son datos – de los que no nos interesa su conformación – y es que en realidad, pueden ser estructuras de gran complejidad, lo importante es que pueden ser consideradas como datos simples y en consecuencia pueden ser tratados como si fueran elementos. De esta manera se puede obtener la siguiente clasificación:

- **Estructura de tipo lista**: Es la que surge al observar la manera como se relacionan los elementos y se puede concluir que una forma de relación es la secuencia u orden de la cual deriva la estructura de lista.

- **Estructura de tipo árbol**: La estructura de tipo lista es muy utilizada, sobre todo cuando los elementos tienen la misma categoría; sin embargo cuando entre éstos elementos existen diferentes niveles, surgen relaciones del tipo jerárquico que se plasman en una estructura jerárquica o de árbol.

- **Estructura de tipo red**: Las relaciones que existen entre los elementos en muchos casos son complejas y las dos estructuras anteriores son insuficientes; por ello se requiere de una estructura más abierta, en la que los elementos puedan relacionarse sin tener las restricciones de secuencia o jerarquía, sino básicamente de adyacencia.

Cuando se está haciendo una búsqueda en una tabla, hacer una consulta en una tabla con cualquier tipo de estructura, es algo que se vuelve trivial utilizando un sistema gestor de bases de datos, pero es importante el tipo de estructura que dicha tabla tiene, pues de esto dependerá la rapidez y eficacia de la búsqueda [Samuel & Pedersen, 2004].

Existen dos diferentes formas de acelerar las consultas que se hacen. La primera es afinando nuestro servidor para que responda lo mejor posible, añadir

memoria y esperar el tiempo que sea necesario, pero queda claro que con la cantidad de información que pronto habrá disponible en astronomía, esto resultará inviable. La otra forma de hacer una búsqueda acelerada en una base de datos es haciendo uso de los índices [Celko, 2010].

Los índices son usados para encontrar rápidamente los registros que tengan un determinado valor en alguna de sus columnas. Sin un índice, **MySQL** tiene que iniciar con el primer registro y leer a través de toda la tabla para encontrar los registros relevantes. Aún en tablas pequeñas, de unos 1,000 registros, es por lo menos 100 veces más rápido leer los datos usando un índice, que haciendo una lectura secuencial [DuBois, 2003]. Cuando **MySQL** trata de responder una consulta, examina una variedad de estadísticas acerca de nuestros datos y decide como buscar los datos que deseamos de la manera más rápida [Celko, 2010]. Sin embargo, como se acaba de mencionar, cuando en una tabla no existen índices en los cuales pueda auxiliarse **MySQL** para resolver una consulta, se tendrán que leer todos los registros de la tabla de manera secuencial. Esto es comúnmente llamado un escaneo completo de una tabla, y es muchas veces algo que se debe evitar.

Se debe evitar el escaneo completos de tablas por las siguientes razones:

- Sobrecarga de CPU. El proceso de leer y revisar cada uno de los registros en una tabla es insignificante cuando se tienen pocos datos, pero puede convertirse en un problema a medida que va aumentando la cantidad de registros en nuestra tabla. Existe una relación la cuál es proporcional entre el número de registros que tiene una tabla y la cantidad de tiempo que le toma a **MySQL** revisarla completamente.
- Concurrencia. Mientras **MySQL** está leyendo los datos de una tabla (actualizándolos o eliminándolos), éste la bloquea, de tal manera que nadie más puede escribir en ella (es posible leerla).
- Sobrecarga de disco. En una tabla muy grande, un escaneo completo consume una gran cantidad de entrada/salida en el disco. Esto puede alentar de manera siginificativa el servidor de bases de datos.

En resumen, lo mejor es tratar de que los escaneos completos de tablas sean mínimos – especialmente si nuestra aplicación necesita escalabilidad en

tamaño, número de usuarios, o ambos [Beaulieu, 2005].

2.6.5. Tablas.

Una tabla en una base de datos es bastante similar – en apariencia – a una hoja de cálculo típica, pues los datos se encuentran almacenados en filas y columnas. Una consecuencia a esto es, la relativa facilidad de importar una hoja de cálculo a algún gestor de bases de datos, como una base de datos. La principal diferencia entre almacenar los datos en una hoja de cálculo y hacerlo en una base de datos es la forma en la que se organizan los datos.

Para lograr la máxima flexibilidad para una base de datos, la información tiene que estar organizada en tablas, para que no haya redundancias. Por ejemplo, si se almacena información sobre una galaxia, cada galaxia se debe insertar una sola vez en una tabla que se configurará para contener únicamente datos de las galaxias.

Cada fila de una tabla se denomina registro. En los registros es donde se almacena cada información individual. Cada registro consta de campos (al menos uno). Los campos corresponden a las columnas de la tabla.

2.6.6. Árboles.

El árbol es una estructura de datos muy importante en informática y en ciencias de la computación. Se trata de estructuras no lineales a diferencia de las listas, listas encadenadas, tablas o algún cierto tipo de arreglos, que se considera lineal.

Puede representarse como un conjunto de nodos enlazados entre sí por medio de ramas. La información contenida en un nodo puede ser de cualquier tipo, ya sea simple o una estructura de datos compleja.

Los árboles se utilizan para representar fórmulas algebraicas, organizar objetos en orden de tal forma que las búsquedas sean muy eficientes y en algunas aplicaciones diversas como inteligencia artificial o algoritmos de cifrado. También se utilizan en el diseño de compiladores, procesamiento de texto y algoritmos de búsqueda [Celko, 2010].

2.6.7. Árboles generales y terminología.

De manera intuitiva, el concepto de árbol implica una estructura en la que los datos se organizan de modo que los elementos de información están relacionados entre sí a través de ramas.

Formalmente: un árbol es una estructura de base de datos que cumple una de estas dos condiciones:

- Es una estructura vacía, o
- Es un nodo de tipo base que tiene de 0 a N subárboles disjuntos entre sí.

El nodo base, que debe ser único, es conocido como raíz y se ha establecido, por convención, el representarlo de manera gráfica, en la parte superior.
En particular, un **árbol** es un conjunto de uno o más **nodos** tales que: hay un nodo especial llamado **raíz**, que debe ser único y e ha establecido, por convención, el representarlo de manera gráfica, en la parte superior.

Los nodos restantes se dividen en $n \geq 0$ conjuntos disjuntos, tal que cada uno de estos conjuntos es un árbol y se los conoce como **subárboles** [Celko, 2004].

Se hablará sobre algunos términos utilizados en la descripción de los atributos de un árbol. un **nodo** que tiene **subárboles** se conoce como **padre** de ellos, y los **nodos** sucesores se llaman **hijos**.

Los hijos de un nodo se denominan **descendientes** así como el padre y los abuelos se conocen como **ascendientes**.

Los nodos del mismo padre suelen llamarse **hermanos** y los nodos que no tienen descendientes se conocen como **hojas**.

Se define **nivel de un nodo** a la distancia que ese nodo tiene al nodo raíz, consecuente la raíz tiene **nivel** igual **cero**.

Un **camino** es una secuencia de nodos en los que cada nodo es **adyacente al siguiente**. Cada nodo del árbol puede ser alcanzado siguiendo un único camino que comienza en el nodo raíz.

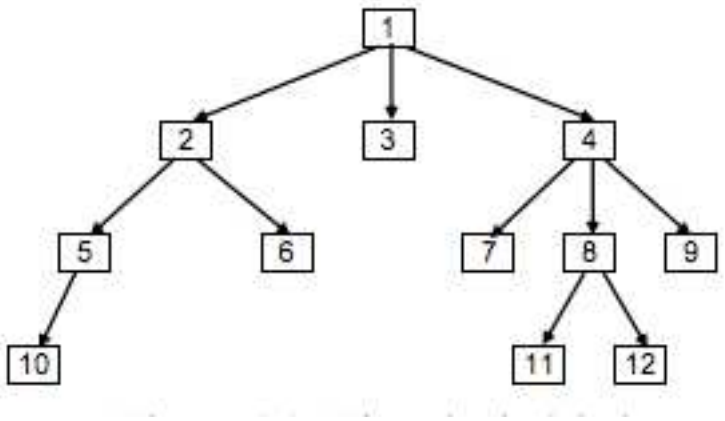

Figura 2.2: Ejemplo de un árbol, cuya numeración de nodos se seleccionó de manera arbitraria.

La **altura** o **profundidad** de un árbol es el nivel de la hoja del camino más largo desde la raíz más uno. Por definición la altura de un árbol vació es cero.

Un árbol se divide en **subárboles**, que es cualquier estructura conectada por debajo del nodo raíz. Cada **nodo de un árbol** es la raíz de un subárbol que se define por el nodo y todos sus descendientes.

La figura 2.2. muestra un ejemplo de una estructura en árbol (se ha puesto una numeración de los nodos de manera arbitraria). Se entiende por topología de un árbol a su representación geométrica.

En un **árbol** se representa una **relación jerárquica** a partir del nodo raíz en sentido vertical descendente, definiendo niveles. Por ejemplo el nivel del nodo raíz es: 1.

Desde la raíz se puede llegar a cualquier nodo, avanzando por las distintas ramas y atravesando los sucesivos niveles, estableciendo así un **camino**. Por ejemplo, en la figura 2.2 el nodo 10 está a nivel 4 y para llegar a él, se puede seguir un camino (o subcamino) de nodos, es decir pasar por el nodo raíz así como por los nodos 2, 5 y 10.

Un nodo es **antecesor** de otro cuando ambos forman parte de un camino, pero el primero se encuentra en un nivel superior (es decir, en la numeración arbitraria designada en el ejemplo mostrado en la figura 2.2 tiene una numeración más baja) al del segundo (el cual tiene una numeración más alta). En el ejemplo anterior el nodo 5 es antecesor del 10. Por el contrario, el nodo 10 es **descendiente** del nodo 5.

La relación entre dos nodos separados de forma inmediata por una rama se denomina **padre/hijo**.

En el ejemplo de la figura 2.2, el nodo 5 es **hijo** del nodo 2 y recíprocamente, el nodo 2 es **padre** del nodo 5.

En un árbol, un padre puede tener varios hijos pero un hijo sólo puede tener un único padre.

Se denomina **grado** al número de hijos que tienen un nodo. Por ejemplo, en la figura 2.2 el nodo 2 tiene grado 2 y el nodo 6 tiene grado 0.

En esta analogía de árbol se dice que un nodo es **hoja** cuando no tiene descendientes (grado 0).

2.6.8. La estructura en árbol.

Al utilizar una estructura de árbol, se tienen las siguientes ventajas:

- Relaciones jerárquicas entre los datos de una colección.
- Búsqueda en tiempos sublineales, lo que no se puede conseguir al usar una representación lineal.

La altura del árbol de la figura 2.2 es 4 (la alcanzan sus nodos 10, 11 y 12). El número de nodos del nivel más poblado en el ejemplo es 5 para el nivel 3. E mayor de los grados de los nodos, en el ejemplo es 3 (para dos de los nodos 1 y 4).

Se dice que un árbol es completo sólo cuando todos sus nodos (excepto las hojas) tienen el mismo grado y los diferentes niveles están poblados por completo. En un árbol completo de grado g, la amplitud del nivel n es $A_n ==$

g^{n-1}, y el número total de nodos del árbol es:

$$N_n = \frac{g^n - 1}{g - 1} \tag{2.1}$$

- Una trayectoria del nodo n_i al nodo n_k, es una secuencia de nodos desde n_i hasta n_k, tal que n_i es el padre de n_{i+1}
- Existe un solo enlace entre un padre y cada uno de sus hijos.
- El largo de una trayectoria es el número de enlaces en la trayectoria.
  - Una trayectoria de k nodos tiene largo $k - 1$.
- La altura de un nodo es el largo de la trayectoria más larga de ese nodo a una hoja. La profundidad de un nodo es el largo de la trayectoria desde la raíz a ese nodo.
  - La profundidad del árbol es la profundidad de la hoja mas profunda.
  - Nodos a una misma profundidad están al mismo nivel.

2.6.9. Árbol Binario.

Los árboles binarios son aquellos que tienen grado 2. Esto significa que cada nodo puede tener dos, uno o ningún hijo, pero nunca más de dos subárboles. En particular, al tratarse de dos hijos - como máximo -, cada uno de ellos puede identificarse como **hijo izquierdo** o **hijo derecho**, y esos nodos pueden ser la raíz de subárboles (binarios) de manera recursiva.

Un árbol binario está formado por un nodo raíz, un subárbol izquierdo **I** y uno derecho **D**.

En cualquier nivel **n**, un árbol binario puede contener de 1 a 2^n nodos. El número de nodos por nivel contribuye a la densidad del árbol.

2.6.10. Árbol binario completo.

Un **árbol binario completo** de profundidad n es un árbol en el que para cada nivel desde el 0 al $n - 1$ tiene un conjunto lleno de nodos, y cada uno de los nodos hoja a nivel n ocupan las posiciones más a la izquierda del árbol.

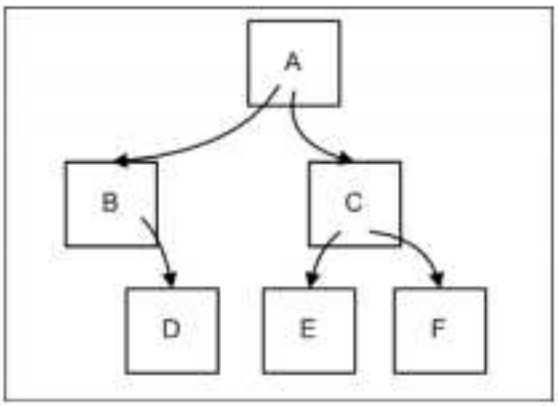

Figura 2.3: Ejemplo de un árbol binario.

Un **árbol binario completo** que contiene $2n$ nodos a nivel n es un **árbol lleno**. Un **árbol lleno** es un árbol binario que tiene el máximo número de entradas para su altura. Esto sucede cuando el último nivel está lleno.

En la figura 2.4 se muestra un esquema de un árbol binario completo, mientras que en 2.5 se muestra de manera esquemática un árbol binario lleno.

2.6.11. Árboles binarios de búsqueda y sus aplicaciones.

Una de las aplicaciones más importantes que tienen los árboles de búsqueda binarios, es la de organizar la información de manera **jerárquica**, para así acelerar los procesos de búsqueda, inserción y borrado. Se logra una mejora por ejemplo, sobre las búsquedas lineales [Celko, 2004].

Para efectuar las búsquedas, se utiliza una clave que se llama **la clave de búsqueda**. Dicha clave, en un árbol binario de nodos, está relacionada con los nodos y debe cumplir con la propiedad de que el subárbol a la izquierda de cada nodo – si existe – debe contener sólo nodos con claves menores o iguales al padre y el subárbol a la derecha – si existe – debe contener sólo nodos con claves mayores o iguales al padre.

La **clave de búsqueda** se extrae de la propia información, directamente o mediante transformaciones adecuadas. Para poder clasificar la información sin ambigüedad, es necesario establecer entre las claves de búsqueda un conjunto

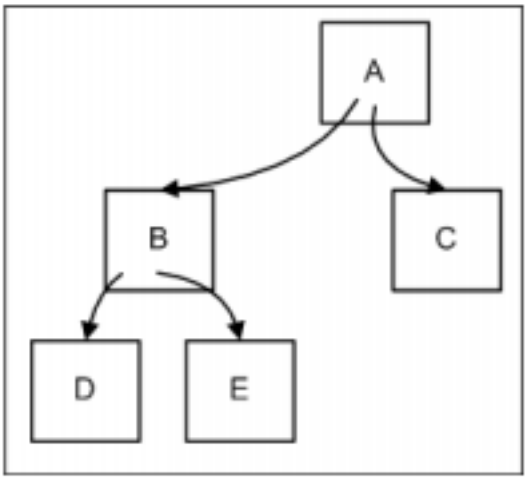

Figura 2.4: Árbol binario completo.

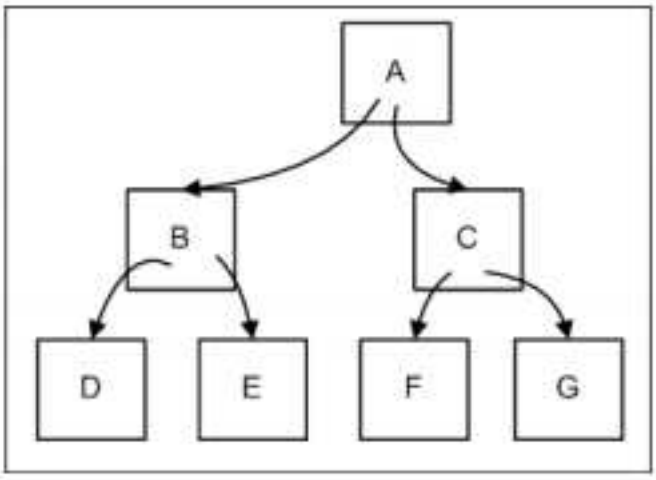

Figura 2.5: Árbol binario lleno.

de condiciones mutuamente excluyentes, tal que una y sólo una de ellas sea cierta.

Un **árbol de búsqueda** se define como un árbol en el que, para cada nodo, las claves de los subárboles hijos satisfacen una y sólo una condición de un conjunto de n condiciones mutuamente excluyentes.

Si $n = 2$, se tendrá un árbol de búsqueda binario; si $n = 3$, se tendrá un árbol de búsqueda ternario; etc. Así pues, un árbol binario de búsqueda (ABB) es un árbol binario, en el que para cada nodo se definen dos condiciones mutuamente excluyentes, de forma que las claves de los nodos del subárbol izquierdo cumplen una de ellas, y las del subárbol derecho la otra. Habitualmente estas condiciones determinan una relación de orden, de manera que el recorrido en orden simétrico del árbol produce una secuencia ordenada de nodos.

2.7. Análisis de Algoritmos de búsqueda.

Gran parte de la motivación para el diseño de árboles se debe a la compatibilidad que presentan con algoritmos eficaces de búsqueda [Celko, 2004].

2.7.1. Los costos en tiempo y en espacio.

Una de las características básicas que debe tener un algoritmo es que éste sea correcto, esto quiere decir, que produzca el resultado deseado en un tiempo finito. Adicionalmente puede interesarnos, desde luego, que sea claro, que esté bien estructurado, que sea fácil de usar, que sea fácil de implementar y que sea **eficiente**.

Entendemos por **eficiencia** de un algoritmo la cantidad de recursos de cómputo que requiere; es decir, cuál es su tiempo de ejecución y qué cantidad de memoria utiliza [Celko, 2010].

A la **cantidad de tiempo** que requiere la ejecución de un cierto algoritmo se le suele llamar **costo en tiempo** mientras que **costo en espacio** se le llama a la cantidad de memoria que requiere.

Conviene buscar algoritmos correctos que mantengan tan bajo como sea posible el consumo de recursos en el sistema, es decir, que sean lo más eficientes posible. Cabe hacer notar que el **concepto de eficiencia de un algoritmo** es un concepto relativo, esto quiere decir que ante dos algoritmos correctos que resuelven el mismo problema, uno es más eficiente que otro si el primero consume menos recursos. Así, el concepto de eficiencia y en consecuencia el concepto de costo nos permitirá comparar distintos algoritmos entre ellos [Cormen et al., 2001].

Definición 1 *Dado un algoritmo A cuyo conjunto de entradas es $\mathcal{A}$, su eficiencia o costo (en tiempo, en espacio, en número de operaciones de entrada/salida, etc.) es una función T tal que:*

$$T: \quad \mathcal{A} \longrightarrow \mathbb{R}^+$$

$$\alpha \longmapsto T(\alpha)$$

2.7.2. Costo en los casos: mejor, promedio y peor.

Utilizando la definición 1 que se dio anteriormente, caracterizar a la función T puede ser complicado. Por lo que se definen tres diferentes funciones las cuáles dependen exclusivamente del tamaño que tengan las entradas, describiendo de manera resumida las caractersticas de la función T.

Definición 2 *Sea $\mathcal{A}_n$ el conjunto de las entradas de tamaño n y T_n: $\mathcal{A}_n \longrightarrow \mathbb{R}$ la función T restringida a $\mathcal{A}_n$. Los costos en el mejor, promedio y peor caso se definen como sigue:*

***Costo en el mejor caso**: $T_{mejor}(n) = mín\ \{T_n(\alpha) | \alpha \in \mathcal{A}_n\}$.*

***Costo en el caso promedio**: $T_{prom}(n) = \sum_{\alpha \in \mathcal{A}_n} Pr(\alpha) T_n(\alpha)$, donde $Pr(\alpha)$ es la probabilidad de ocurrencia de la entrada α.*

***Costo en el peor caso**: $T_{peor}(n) = máx\ \{T_n(\alpha) | \alpha \in \mathcal{A}_n\}$.*

En general, en ningún caso se excede el costo del peor caso.

Una característica que tiene el costo de un algoritmo – en cualquiera de los tres casos descritos anteriormente – es su **tasa de crecimiento**. La tasa de

crecimiento de una función marca una diferencia importante con las funciones que tengan un tasa de crecimiento distinta [Cormen et al., 2001].

2.7.3. Notación Asintótica.

Se hablará de tres distintas funciones que son: M N y O.

Definición 3 *Dada una función* f*:* $\mathbb{N} \longrightarrow \mathbb{R}^+$ *la clase* $M(f)$ *se define por:* $M(f) = \{\ g$*:* $\mathbb{N} \longrightarrow \mathbb{R}^+ |\ \exists\ c \in \mathbb{R}^+, \exists\ n_0 \in \mathbb{N}$*:* $\forall\ n \geq n_0,\ g(n) \leq cf(n)\}$.

Intuitivamente, la definición anterior refleja el hecho de que el crecimiento asintótico de las funciones g es, a lo mucho proporcional al de la función f.

Se puede decir – informalmente – que la tasa de crecimiento de la función f es una cota superior para las tasas de crecimiento de las funciones g.

2.7.3.1. Propiedades básicas de la notación M.

- Si $\lim_{n\to\infty} \frac{g(n)}{f(n)} < \infty$ entonces $g = M(f)$.
- $\forall f$: $\mathbb{N} \longrightarrow \mathbb{R}^+$, $f = M(f)$ (reflexividad).
- Si $f = M(g)$ y $g = M(h)$ entonces $f = M(h)$ (transitividad).
- $\forall\ c > 0$, $M(f) = M(cf)$

La úlltima propiedad justifica la preferencia por omitir factores constantes. En el caso de la función log se omite la base porque $log_c(x) = \frac{log_b(x)}{log_b(c)}$.

Definición 4 *Dada una función* f*:* $\mathbb{N} \longrightarrow \mathbb{R}^+$ *la clase* $N(f)$ *se define por:* $N(f) = \{\ g$*:* $\mathbb{N} \mid \mathbb{R}^+ \longrightarrow \exists\ c \in \mathbb{R}^+, \exists\ n_0 \in \mathbb{N}$*:* $\forall n \geq n_0,\ g(n) \geq cf(n)\}$.

Esta definición muestra que: el crecimiento asintótico de las funciones g es más rápido que el de la función f. Así que la tasa de crecimiento de la función f puede verse como una cota inferior para las tasas de crecimiento de las funciones g.

2.7.3.2. Propiedades básicas de la notación N.

- Si $\lim_{n\to\infty} \frac{g(n)}{f(n)} > 0$ entonces $g = N(f)$.
- $\forall f$: $\mathbb{N} \longrightarrow \mathbb{R}^+$, $f = N(f)$ (reflexividad).
- Si $f = N(g)$ y $g = N(h)$ entonces $f = N(h)$ (transitividad).
- Si $f = M(g)$ entonces $g = N(f)$ y viceversa.

Definición 5 *Dada una función* f*:* $\mathbb{N} \longrightarrow \mathbb{R}^+$ *la clase* $O(f) = M(f) \bigcap N(f)$

Con esta definición, de manera intuitiva se refleja el hecho de que el crecimiento asintótico de las funciones g es similar al de la función f.

2.7.3.3. Propiedades básicas de la notación O.

- $\forall f$: $\mathbb{N} \longrightarrow \mathbb{R}^+$, $f = O(f)$ (reflexividad).
- Si $f = O(g)$ y $g = O(h)$ entonces $f = O(h)$ (transitividad).
- $f = O(g) \Longleftrightarrow g = O(f)$ (simetría).
- Si $\lim_{n\to\infty} \frac{g(n)}{f(n)} = c$, $0 < c < \infty$, entonces, $g = O(f)$.

2.7.3.4. Otras propiedades de las notaciones asintóticas.

- Para cualesquiera constantes positivas $\alpha < \beta$, si f es una función creciente entonces $M(f^\alpha) \subset M(f^\beta)$.
- Para cualesquiera constantes a y b mayores que cero, si f es una función creciente entonces $M(log(f^a) \subset M(f^b)$.
- Para cualquier constante $c > 0$, si f es una función creciente, $M(f) \subset M(c^f)$.

2.7.3.5. Reglas útiles.

Regla de las sumas $O(f) + O(g) = O(f + g) = O(\text{máx}\{f,g\})$.

Regla del producto $O(f)O(g)$=$O(fg)$.

2.7.4. Costo de los algoritmos iterativos.

Existen algunas reglas que facilitan el cálculo o incluso el análisis del costo de los algoritmos iterativos en el peor de los casos.

- El costo de una operación elemental es $O(1)$.
- Cuando el costo de un fragmento F_1 es f_1 y el de un fragmento es F_2 es f_2 entonces el costo en el peor de los casos del fragmento:
  F_1;F_2
  es:
  $f_1 + f_2$

2.7.5. Costo de los algoritmos recursivos.

El costo que tiene $T(n)$ (en cualquiera de los tres casos: peor, medio o mejor) de un algoritmo recursivo, satisface una ecuación recurrente. Esto quiere decir que $T(n)$ dependerá del valor de T para valores más pequeños de n.
Con frecuencia la recurrencia adopta una de las siguientes formas :

- $T(n) = aT(n-c) + g(n)$, con a y c constantes y tales que $a > 0$ y $c > 1$.
- $T(n) = AT(n/b) + g(n)$, con a y b constantes tales que $a > 0$ y $b > 1$.

La primera recurrencia se le conoce como **recurrencia sustractora** y ésta, corresponde a algoritmos que tienen una parte que no es recursiva con un costo $g(n)$ y hace a llamadas recursivas con problemas de tamaño $n - c$.

La segunda recurrencia presentada, se conoce como **recurrencia divisora**. Y corresponde a algoritmos que tienen una parte que no es recursiva con costo $g(n)$ y hacen a llamadas recursivas con subproblemas de tamaño aproximado n/b.

2.7.6. Teoremas.

A continuación se citarán algunos teoremas, los cuáles proporcionan un mecanismo que permite calcular el costo de algoritmos recursivos que se expresen mediante recurrencias sustractoras o divisoras.

Teorema 1 *Sea $T(n)$ el costo (en cualquiera de los tres casos: peor, medio y mejor) de un algoritmo recursivo que satisface la recurrencia:*

$$T(n) = \begin{cases} f(n) & si\ 0 \leq n < n_0 \\ aT(n-c) + g(n) & si\ n \geq n_0 \end{cases}$$

Donde n_0 es un número natural, $c \geq 1$ es una constante, $f(n)$ es una función arbitraria y $g(n) = O(n^k)$ para una constante $k \geq 0$, entonces:

$$T(n) = \begin{cases} O(n^k) & si\ a < 1 \\ O(n^{k+1}) & si\ a = 1 \\ O(n^{n/c}) & si\ a > 1 \end{cases}$$

Teorema 2 *Sea $T(n)$ el costo (en los tres casos, es decir peor, medio o mejor) de un algoritmo recursivo que satisface la recurrencia:*

$$T(n) = \begin{cases} f(n) & si\ 0 \leq n < n_0 \\ aT(n/b) + g(n) & si\ n \geq n_0 \end{cases}$$

donde n_0 es un número natural $b > 1$ es una constante, $f(n)$ es una función arbitraria y $g(n) = O(n^k)$ para una constante $k \geq 0$. Sea $\alpha = log_b(a)$. Entonces:

$$T(n) = \begin{cases} O(n^k) & si\ \alpha < k \\ O(n^k log\ n) & si\ \alpha = k \\ O(n^\alpha) & si\ \alpha > k \end{cases}$$

2.7.7. Búsquedas en una lista ordenada.

Para éste tipo de búsqueda no existe forma más eficaz de hacerlo que iniciar por la primera entrada y analizar cada una de ellas hasta encontrar un elemento coincidente o hasta llegar al final de la lista.

Si la lista está ordenada de forma ascendente, se podrá reconocer una ausencia tan pronto como se encuentre una clave mayor que la que se busca.

La eficacia que tiene este método está determinada por la posición que tiene el elemento buscado. Si sólo se busca un elemento y la lista contiene n elementos, habría que esperar a recorrer máximo n elementos para encontrar la

entrada. Si la lista contiene $2n$ elementos, de forma intuitiva habría que esperar el doble que si contuviese n elementos, y así conforme la lista crezca.

Para hacer un análisis del tiempo de ejecución de un algoritmo, se inicia dividiendo todo el cálculo en los diferentes pasos que tarden el mismo tiempo cada vez que se ejecuten [Cormen et al., 2001]. Se debe tomar en cuenta si hay o no *bucles* que se puedan ejecutar un número variable de veces, así como si existen diferentes condiciones que puedan afectar de alguna manera el desarrollo del algoritmo. Por lo que el análisis debería identificarlos para que se pueda calcular el número medio y el máximo (peor) número de repeticiones que se efectuarán.

El resultado del costo de tiempo de ejecución de este análisis, suele ser una suma de términos. Cada término consta del número de veces que se repite una constante que representa el tiempo estimado de cada subtarea. Así la suma de la búsqueda lineal de una lista enlazada sencilla puede expresarse como

$$C_s + C_c * k \tag{2.2}$$

Donde:

- C_s = Costo constante de definir una búsqueda. Por ejemplo: analizar argumentos.

- C_c = Costo de comprobar un elemento de la lista.

- $k = 1$, en el mejor de los casos, esto es, si el elemento que se busca es el primero de la lista, $n/2$ en el caso habitual donde n = el número de elementos de la lista; n en el peor de los casos.

En realidad, no es nada fácil calcular "a priori" el tiempo preciso que algún programa se va a tardar en ejecutar las subtareas. Puesto que esto en gran medida va a depender del CPU, la configuración del equipo, la eficacia del compilador, etc. Sin embargo, el recuento de repeticiones sí puede ser muy preciso y además se suele expresar como la función de una variable que representa algún aspecto de la envergadura del problema [Cormen et al., 2001].

En el caso de buscar en una lista enlazada simple, la cantidad clave es la longitud de la lista en la que estamos buscando.

El objetivo se centra en prever qué pasará con el tiempo de ejecución a medida que crece la envergadura del propio problema.

En una lista enlazada, el término C_s no nos importa demasiado, ya que no cambia a medida que la lista crece. Lo que sí es relevante es que podemos esperar un crecimiento lineal del tiempo de ejecución directamente proporcional a la envergadura del problema (la longitud de la lista). Si nuestro algoritmo tuviese dos o más bucles dependientes del número de elementos, la fórmula para el peor de los casos sería una suma de los términos dependientes de la variable n.

Hoy en día estamos familiarizados con todo tipo de búsquedas: en un texto plano, por ejemplo podemos buscar exactamente la palabra o el conjunto de caracteres que deseamos. Mientras más información se tenga en una base de datos, más lento se vuelve el proceso de búsqueda.

Por ejemplo, se puede realizar una búsqueda en1 MB de información en un segundo; con el mismo comando se puede buscar en 1 GB de información tardando un minuto, en 1 TB de información tardarías dos días. Este tipo de búsqueda es totalmente ineficiente para grandes cantidades de información, ya que para buscar en 1 PB de información se tardaría al rededor de tres años.
Así que en algún punto se necesitan los índices para limitar la búsqueda.

En particular, un índice le permite a **MySQL** determinar si un valor dado coincide con cualquier fila en una tabla [Beaulieu, 2005]. Cuando indexamos una columna en particular, **MySQL** almacena información extra acerca de los valores en la columna indexada, es decir, **MySQL** almacena todas las claves de los índices, creando una estructura de datos no lineal, la cuál permite a **MySQL** encontrar los índices rápidamente.

Cuando **MySQL** encuentre que hay un índice en una columna, lo usará en vez de hacer un escaneo completo de la tabla. Esto reduce de manera imporante los tiempos de CPU y las operaciones de entrada/salida en disco, también mejora la concurrencia porque **MySQL** bloqueará la tabla únicamente para obtener las filas que necesite (en base a lo que encontró utilizando el índice). Cuando tenemos grandes cantidades de datos en nuestras tablas, la mejora en la obtención de los datos puede llegar a ser muy significativa [Celko, 2010].

2.7.8. Eficacia en una búsqueda binaria.

Se ha demostrado ya que la búsqueda en estructuras lineales no es lo más eficiente. Ahora se hablará de la eficiencia que tienen las búsquedas en los árboles binarios.
El método de búsqueda binaria es eficiente para búsquedas en arreglos de gran

tamaño, el único inconveniente es que requiere que el arreglo de datos este ordenado [Cormen et al., 2001]. Consiste en ubicar el elemento de la posición central del arreglo, si el elemento de la posición central coincide con la clave de búsqueda, finaliza la búsqueda porque ya se encontró. Si no coincide, se determina si elemento se encuentra en la mitad superior o inferior del arreglo y a continuación se repite el proceso utilizando el elemento central de la sublista. El algoritmo es como sigue: Las operaciones básicas del árbol binario de búsqueda deberían requerir un tiempo $O(h)$, donde h es la altura del árbol. Pero, se deduce que la altura de un árbol binario equilibrado es, aproximadamente $log_2(n)$, donde n es el número de elementos si el árbol permanece equilibrado. Se puede demostrar que, si las claves se insertan aleatoriamente en un árbol binario de búsqueda, esta condición se cumplirá y que el árbol permanecerá lo suficientemente equilibrado para que la hora de búsqueda y de inserción sea aproximadamente $O(log\ n)$.

2.7.9. Operaciones básicas: Búsqueda, Inserción y Borrado en árboles.

La complejidad de estas operaciones está en $O(h)$, siendo h la altura del árbol; en ABB equilibrados, $h = log_2(n)$, siendo n el número de nodos. La operación buscar k-ésimo(t,i) devuelve el nodo con la i-ésima clave más pequeña. Suponiendo que todos los elementos del Árbol Binario tienen claves distintas, esta operación puede realizarse en tiempo $t = O(logn)$, sin que el costo de las demás operaciones se vea afectado, sin que el costo de las demás operaciones se vea afectado.

Como ya se ha dicho anteriormente, con un gestor de bases de datos (DBSM) –en particular, **MySQL** – y un lenguaje como es **SQL**, se tiene como primer resultado tablas de datos que tienen una naturaleza lineal o unidimensional, llamada anteriormente como estructura de tipo lista, esto ocurre aún que entre los datos existan relaciones. En los tipos abstractos de datos lineales, existen exactamente un elemento previo y otro siguiente (excepto para el primero y el último, si es que los hay). En las estructuras no lineales, como conjuntos o árboles, este tipo de secuencialidad no existe, aunque en los árboles existe una estructura jerárquica, de manera que un elemento tiene un sólo predecesor, pero varios sucesores.

Una exploración amplia a los tipos de datos que se utilizan en la astronomía (aunque no sólo en astronomía), nos lleva a situaciones en que las representa-

ciones lineales son inadecuadas pues crecerán las bases de datos y con esto los tiempos de búsqueda de manera lineal.

Un árbol impone una estructura jerárquica sobre una colección de objetos. Es una de las estructuras más utilizadas en computación. Siempre que se quiera representar información jerarquizada.

2.8. PICASSO.

El origen del concepto de **PICASSO** surge con la intención de dar respuesta al enorme problema de operar adecuadamente la información masiva producida tanto por los grandes catastros observacionales, como por los datos obtenidos a partir de las simulaciones cosmológicas. Y es que, de manera natural surge una demanda de nuevos métodos, herramientas y recursos computacionales eficientes y escalables.

En **PICASSO** se utiliza el principio de **reciclaje de software** [Reifer, 1997], que se refiere al comportamiento y a las técnicas, que garantizan que una parte o la totalidad de un programa informático existente, se pueda emplear en la construcción de otro programa. De esta forma se aprovecha trabajo preexistente, se economiza tiempo, y se reduce la redundancia.

PICASSO, utiliza un gestor común de bases de datos llamado **MySQL** – que permiten trabajar hoy en día con bases de datos de tamaños moderados – y además, se diseñan e implementan diferentes algoritmos en distintos lenguajes, como **SQL** –principalmente–, **C** y **Python**, con los cuales, se le puede **proporcionar una estructura de jerarquía a los datos** contenidos en las tablas que conforman las bases de datos.

Ya que los datos en las tablas, contienen una jerarquía, se puede trabajar, manipular, hacer búsquedas y un análisis y manejo de las información en diferentes tablas, con distinto tipo de información, mucho más eficiente. Con dicha jerarquía, es muy fácil que los programas hechos para **PICASSO**, puedan escalar el tiempo de búsqueda con bases de datos cada vez más grandes, que, por ejemplo, si sólo se usara **MySQL** y datos contenidos en las tablas sin ningún tipo de jerarquía. Además, se utiliza un visualizador estándar que es compatible con el formato **SQL**, llamado **TopCat**,

Sin embargo, **PICASSO** no es el único esfuerzo que se ha hecho, hoy en

día existe Big Data y Hadoop.

Hadoop[17] es una de las implementaciones de código abierto de MapReduce[18]. Hadoop es un proyecto administrado por Apache Software Foundation. Es un Framework para el desarrollo de aplicaciones de procesamiento paralelo que utiliza MapReduce. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos

Big data es el término que designa un crecimiento, disponibilidad y uso exponenciales de la información estructurada y desestructurada. Existe mucha literatura en torno a la tendencia sobre big data y a cómo este puede ser la base para la innovación, la diferenciación y el crecimiento. La parte más positiva de esto, es que se esté generando un esfuerzo conjunto para lidiar con el volumen, la variedad e incluso la velocidad cada vez mayor de la información.

2.9. Reflexiones hacia el futuro.

Si miramos el panorama de proyectos observacionales, el problema del manejo de datos toma una adecuada dimensión. Tanto los futuros satélites astronómicos previstos para los próximos años, como los futuros telescopios proveerán un sin fín de datos observacionales. Además claro, los grandes relevamientos digitales, proyectos en curso como: 6dF GRS, Dark Energy Survey, Southern Sky Survey y Pan-STARRS elevarán la tasa de adquisición de datos a varios PB por noche de observación.

Cabe preguntarse si resultará práctico el almacenamiento de todos los datos adquiridos o pasaremos a esquemas donde las técnicas de software extraerán la información importante de las imágenes que será anualmente almacenada.

Uno de los proyectos más ambiciosos en desarrollo es el Large Synoptic Telescope a ser instalado en Chile con una tasa de 20 TB por noche, lo que implica una tasa anual de 7.3 PB. Los datos anteriores indican que las bases de datos en astronomía continuarán resultando sumamente útiles. La implementacion de bibliotecas que extiendan la funcionalidad de **SQL** es necesaria y la posibilidad de integrar búsquedas con **SQL** en software estadísticos que permitirá una mejor interacción con los grandes repositorios de datos existentes y

[17] http://hadoop.apache.org/

[18] MapReduce es un modelo de programación diseñado por Google, escrita en Java.

futuros.

Hoy se pueden comparar y estudiar, por ejemplo:

1. Catálogos de galaxias simuladas vs. galaxias observados de catastros como **SDSS** [York et al., 2000] y 2dF [Cole et al., 2005].

2. La morfología global (distribución de galaxias) observadas con la distribución de halos [Kirscher et al., 2009], [Moster et al., 2013].

3. Modelos físicos apropiados que midan estadísticamente el acumulamiento de galaxias [Cameron et al., 2009].

3
Propiedades Generales de Galaxias

En esta sección se presentan varios de las conceptos astronómicos que serán utilizados a lo largo de esta tesis.

3.1. Magnitud y Luminosidad.

La información del Universo se puede obtener gracias al estudio detallado de la luz emitida por las estrellas, las galaxias, las nubes de gas y el polvo.

El primer catálogo de estrellas fue hecho por Hiparco, quien usó dos criterios: el primero fue la posición de cada estrella y el segundo su brillo aparente. Para lo cual creó una escala numérica donde la estrella más brillante era de primera magnitud y la estrella más débil que podía observar a ojo era de sexta magnitud, es decir, la escala de Hipparco asignaba números más pequeños a las estrellas más brillantes.

La escala de Hipparco se calibró para que una estrella de primera magnitud fuese 100 veces más brillante que una de sexta. La diferencia de una magnitud a otra es de $100^{1/5} = 2{,}51$. Hoy en día se utilizan distintos detectores de luz para medir la magnitud aparente de un objeto. La escala de Hiparco fue extendida

en ambas direcciones.

El **brillo de una estrella** está medido en términos de su flujo de radiación F. El **flujo de radiación** es la cantidad total de luz emitida en todas las longitudes de onda que pasan a través de una unidad de área perpendicular a la dirección del flujo por unidad de tiempo.

El **flujo de radiación** de un objeto dependerá de su luminosidad (energía emitida por segundo) intrínseca y su distancia al observador.

$$F = \frac{L}{4\pi r^2} \tag{3.1}$$

Recordando que la diferencia de 5 magnitudes entre la magnitud aparente de dos estrellas corresponde a una razón de 100 podremos escribir el cociente de flujos de radiación como:

$$\frac{F_2}{F_1} = 100^{(m_1 - m_2)/5} \tag{3.2}$$

donde m es la magnitud aparente del objeto.

Usando la distancia de 10 párstecs, se obtiene:

$$100^{(m_1 - m_2)/5} = \frac{F_{10}}{F} = \left(\frac{d}{10pc}\right)^2 \tag{3.3}$$

Usando la ecuación 3.1 podemos expresar la razón entre las luminosidades como:

$$\frac{L_2}{L_1} = 100^{(m_1 - m_2)/5} \tag{3.4}$$

También podemos reescribir la ecuación 3.4 en términos de magnitudes absolutas:

$$\frac{L_2}{L_1} = 100^{(M_1 - M_2)/5} \tag{3.5}$$

Si deseamos normalizar esta ecuación respecto al Sol, tendremos una relación entre la **magnitud absoluta de una estrella** y su **luminosidad**:

$$M = M_{Sol} - 2{,}5 log_{10} \frac{L}{L_{Sol}} \tag{3.6}$$

Casi la totalidad de los objetos del firmamento con excepción del Sol, la Luna, y algunos objetos como cometas, son observados como objetos puntuales a simple vista. Con el uso del telescopio podemos llegar a resolver volumen en objetos muy cercanos (como los planetas), muy brillantes (como las nebulosas), o de gran tamaño (como las galaxias).

Por lo regular se emplean medidas de flujo (intensidad) puntuales, asignándole así una magnitud a los objetos celestes dependiendo de su brillo puntual. En estos casos especiales en los que el objeto posee un volumen aparente no se puede hablar del brillo de un objeto para cualquier punto, ya que la intensidad de brillo varia dependiendo de la región del objeto que observemos.

3.2. Las Galaxias.

Las **galaxias** son objetos formados básicamente por acumulaciones de miles de millones de estrellas, polvo interestelar, gas, planetas y materia oscura, se encuentran ligados gravitacionalmente y orbitan alrededor de un centro común.

3.2.1. Clasificación Morfológica.

Hubble [1926] utilizando la resolución del telescopio de Monte Wilson desarrolló una clasificación para las galaxias, teniendo en cuenta el aspecto visual (la forma) que éstas tenían. La clasificación de Hubble [1926] fue la primera que se hizo, así que es natural que con el tiempo aparecieran en la literatura otros criterios de clasificación basados en sus mismos parámetros o inclusive

en otros. Algunas de las clasificaciones basadas en la apariencia óptica de las galaxias son las de Holmberg [1958] y la de de Vaucouleurs [1959]. Sin embargo, estas clasificaciones son sólo extensiones del esquema de Hubble el cual hasta nuestros días sigue siendo ampliamente usado.

3.2.1.1. Sistema de Clasificación de Hubble.

La secuencia de Hubble o esquema de Hubble mostrada en la figura 3.1, agrupa distintos tipos de galaxias en **elípticas**, **lenticulares** (a las cuales también se les conoce como galaxias de disco), **espirales** e **irregulares**.

A continuación se detallan algunas de las principales características de los distintos tipos morfológicos de la secuencia o esquema de Hubble:

- **Galaxias Elípticas**: Tienen forma elíptica, sin ninguna otra estructura notable superficialmente. Están caracterizadas por su elipticidad $\epsilon = 1 - b/a$, donde a es el semieje mayor y b es el semieje menor. Esta elipticidad se relaciona con el tipo elíptico, de manera que aquellas con apariencia esférica (cuya elipticidad es *0*) tendrán una clasificación *E0*, mientras que las más achatadas (o elongadas) tendrán elipticidades hasta llegar a *7*, es decir: *E7*.

- **Galaxias Lenticulares**: Este tipo de galaxias presenta características comunes a las galaxias elípticas y a las galaxias espirales: poseen una componente esferoidal y una componente de disco, pero carecen de brazos espirales. No presentan rasgos ni detalles, a este tipo de galaxias se les denota como *S0*. Fueron introducidas en el esquema por el mismo Hubble [1936].

- **Galaxias Espirales**: Poseen un bulbo esferoidal, más un disco estelar, el cual puede poseer un patrón espiral (brazo) definido o poco definido, estos brazos nacen de la zona central, la cuál algunas veces presenta la forma de una barra, de allí el nombre de espirales barradas. Las galaxias espirales son designadas como *Sa*, *Sb* o *Sc* cuando no hay una barra presente, en caso contrario se designan como *SBa*, *SBb* y *SBc*, y ambas clasificaciones (para galaxias espirales con o sin barra) están basadas en la apertura de sus brazos espirales.

- **Galaxias Irregulares**: Tal y como su nombre lo indica este tipo de galaxias no tiene una forma definida y son galaxias con un bajo brillo

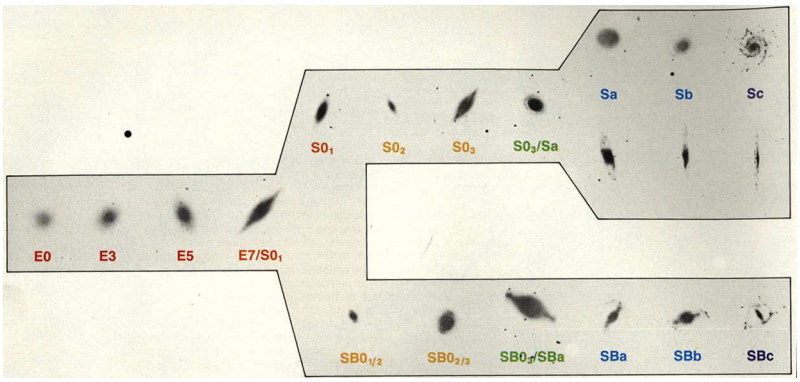

Figura 3.1: Secuencia de Hubble

superficial. Su estructura irregular se debe a un predominio de velocidades de turbulencia relativas a las rotacionales. Tienden a ser más pequeñas que las galaxias elípticas y que las galaxias espirales. Se les clasifica como *Irr I* e *Irr II*.

3.2.1.2. Esquema de Clasificación de De Vaucouleurs.

Otro de los esquemas de clasificación de galaxias es el de de Vaucouleurs [1959]. Que en realidad es un esquema modificado del de Hubble [1926], se le introdujeron tipos tardíos de galaxias espirales, *Sd* y *Sm*, también entre las *Sc* y las *Irr I*, se propuso una subdivisión para las espirales en tipos intermedios, por ejemplo *S0/a*, *Sap* y *Scd*, y se incluyó información sobre la estructura interna y externa del anillo en las espirales. Las galaxias *S0* también fueron separadas en distintas clases en función de la absorción por polvo en sus discos o, en el caso de las *SB0* por la importancia de su barra. Se presenta, en la figura 3.2 el esquema de Clasificación de De Vaucouleurs. En este sistema se utilizan tres ejes para clasificar a galaxias: en el eje principal se encuentra la secuencia: $E - E^{+} - S0^{-} - S0^{0} - S0^{+} - Sa - Sb - Sc - Sd - Sm - Im$. Donde la "*m*" significa magallánica refiriéndose a las nubes de Magallanes, el "-" significa temprano (suave) y el "+" significa tardío (irregular).

Un segundo eje indica si la galaxia contiene barra *(SB)*, no tiene *(SA)* o es muy débil *(SAB)*.

Por último, hay un eje que describe objetos que muestran anillos *(r)*, o que son puramente espirales *(s)* o que cuentan con características intermedias *(rs)*.

El eje principal de este sistema puede ser representado por un parámetro T. Este parámetro está definido en la tabla 3.1. que muestra la relación que existe entre la clasificación de Hubble y el valor T. Los valores de T están fuertemente correlacionados con los colores.

Hubble	E	E/S0	S0	S0/a	Sa	Sa-b	Sb	Sb-c	sc	Sc-Irr	Irr
T	-5	-3	-2	0	1	2	3	4	6	8	10

Cuadro 3.1: Relación existente entre la clasificación de Hubble y el valor T.

3.2.2. Otros sistemas de clasificación.

Existen otros sistemas de clasificación, los cuáles son usados con menor frecuencia.

Por mencionar algunos tenemos el sistema de clasificación de **Yerkes** o **Morgan** [Morgan, 1958], basado en la morfología y la concentración central de luz de las galaxias.
Este sistema unidimensional sigue la secuencia *a-f-g-k*.

Donde los objetos de tipo *a* tienen una débil concentración de luz hasta *k* que son objetos que cuentan con la más alta concentración de luz.

Dicha clasificación muestra que galaxias con alta concentración de luz, tienen una población estelar más vieja y galaxias con baja concentración de luz, cuentan con una población estelar dominante, más joven.

Otra clasificación es la de van den Bergh [van den Bergh, 1960a], [van den Bergh, 1960b]. Es una clasificación para galaxias espirales, dependientes de su luminosidad, y adicionando a la secuencia de Hubble las siguientes clasificaciones:

- **Luminosidad clase** *I* para súper gigantes,

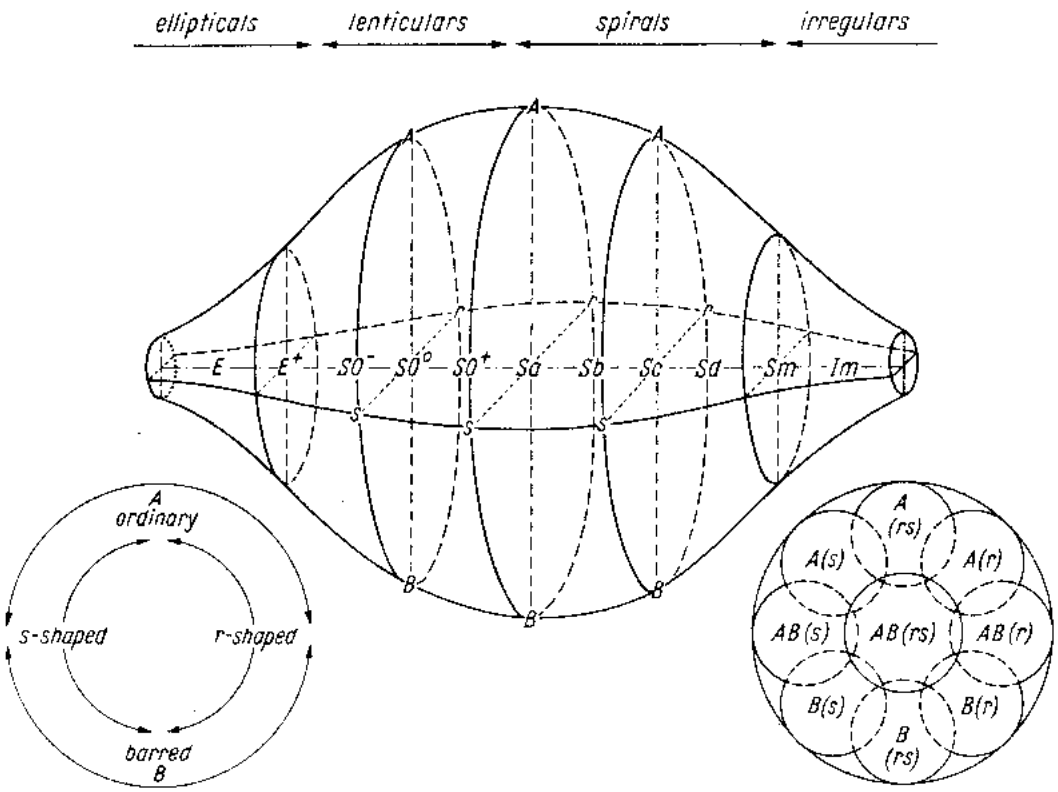

Figura 3.2: Diagrama de los componentes barra y anillo del sistema de clasificación de De Vaucouleurs.

- **luminosidad clase** II para gigantes brillantes,
- **luminosidad clase** III para gigantes,
- **luminosidad clase** IV para subgigantes y
- **luminosidad clase** V para enanas.

Para automatizar los esquemas de clasificación los cuales están basados en cantidades morfológicas, se hace uso de la fotometría de superficie, que es usada para estudiar objetos extendidos (no puntuales) como las galaxias.

En la fotometría de galaxias se define el término **brillo superficial**, como el flujo por unidad de ángulo sólido. Teniendo así, unidades de $[mag/arcsec^2]$. Por lo tanto a primer orden, el brillo superficial es independiente de la distancia al objeto. Algunos objetos astronómicos presentan mayor brillo en su centro y conforme uno se aleja, la luminosidad decrece, hasta llegar a cero en el infinito. Este rango de decreción (representado por funciones matemáticas) es conocido como **perfil de brillo superficial**.

3.3. Propiedades integradas de las galaxias.

3.3.1. Perfiles de Luminosidad.

Una de las herramientas más importantes para estudiar a las galaxias es la **fotometría superficial**.

Las componentes bulbo y disco pueden ser distinguidas por la dependencia de su brillo superficial.

Existen diferentes funciones analíticas que describen el brillo superficial de las galaxias dependiendo del tipo de objeto y el tipo de estructura que se pretende modelar. Mientras que las galaxias elípticas o las componentes esferoidales de las galaxias de disco han sido originalmente ajustadas mediante una única componente, una gran cantidad de galaxias parecen bien descritas por un perfil de dos componentes principales: un bulbo que domina la zona central y un disco que domina las regiones más externas.

El brillo superficial de los discos de galaxias se pueden modelar generalmente bien hacia las regiones externas, mediante una función exponencial del tipo

$$I(r) = I_0 e^{-\left(\frac{r}{r_s}\right)} \tag{3.7}$$

Freeman [1970], donde $I(r)$ e I_0 son la intensidad a la distancia(r), y r_s es el **radio de escala** o la distancia a la que la intensidad de perfil cae e^{-1} de su valor en el centro. Los valores típicos para la longitud de escala del disco r_0, son del orden de 2 a 5 kpc para galaxias espirales normales.

Los perfiles de luminosidad de las galaxias elípticas y de los bulbos de las galaxias espirales pueden ser representados por una función hallada de forma empírica por de Vaucouleurs [1959] , también conocida como la ley $r^{1/4}$, dada por:

$$I(r) = I_e e^{[-7{,}67[\left(\frac{r}{r_e}\right)^{1/4} - 1]]} \tag{3.8}$$

Donde r es la distancia al centro de la galaxia a lo largo de su semieje mayor, r_e es el **radio efectivo** definido como el radio que encierra la mitad de la luminosidad total de la galaxia, y por último I_e la **intensidad** que corresponde al **radio efectivo**.
Algunos de los valores típicos para el radio efectivo r_e de la componente esferoidal de galaxias espirales normales van desde 0,5 hasta 4 kpc, y es mayor en los tipos tempranos.

La descomposición del perfil de luminosidad en una ley $r^{1/4}$ o en una ley exponencial no implica necesariamente un bulbo o disco desde el punto de vista dinámico. La descomposición está regida por dispersión de velocidades para los esferoides y velocidad de rotación para los discos.

Con la llegada de la fotometría CCD, se ha mostrado que para la mayor parte de las galaxias elípticas y esferoidales, la ley $r^{1/4}$ es sólo una aproximación de primer orden que en numerosas ocasiones sólo ajusta bien en un intervalo limitado de perfil de brillo.

En tiempos recientes, se han introducido ciertas dudas sobre su universalidad, y sobre su fiabilidad para determinar parámetros efectivos precisos en galaxias de tipo tempranas [Trujillo & Aguerri, 2004], ya que éstas resultan ser demasiado dependientes de la porción de luz ajustada. Este hecho así como los errores provenientes de las incertidumbres en la sustracción del cielo de fondo, son la razón principal de las discrepancias en los valores de los parámetros ajustados.

Debido a esto, se ha generalizado el uso de una ley de potencias en la cual el exponente es también un parámetro libre. Se conoce como La Ley de Sérsic [Sérsic, 1963], ya que la misma describe satisfactoriamente el comportamiento de los perfiles de luminosidad de un gran porcentaje de galaxias e involucra 3 parámetros a diferencia de los 4 necesarios para caracterizar un bulbo más un disco. Esta ley tiene la siguiente expresión:

$$I(r) = I_e exp\left(-\left(\frac{r}{r_s}\right)^n\right) \tag{3.9}$$

donde I_e es la intensidad central, r_e es la longitud de escala, definida como el radio donde la intensidad ha caído en un factor e^{-1} de la intensidad central y n es el parámetro de forma, o el índice de la ley de potencia, llamado **índice de Sérsic**, el cual toma el valor $n = 1$ para galaxias tipo disco y $n = 0.25$ para las galaxias esferoidales.

La luminosidad de cada componente puede calcularse a través de integrar el perfil de luminosidad en una área circular de radio r:

$$L = \int_o^\infty I(r) 2r\pi \, dr \tag{3.10}$$

Reemplazando las ecuaciones anteriores en 3.10 se obtiene:

$$L_{bulbo} = 7{,}2\pi I_e r_e^2 \tag{3.11}$$

$$L_{disco} = 2\pi I_0 r_0^2 \tag{3.12}$$

De manera análoga la luminosidad obtenida con el perfil de Sérsic está dada por:

$$L_s = \frac{2\pi I_s r_s^2 \Gamma\left(\frac{r}{n}\right)}{n} \tag{3.13}$$

donde $\Gamma\left(\frac{r}{n}\right)$ es la función gamma.

El uso de la función de Sérsic no sólo ajusta el perfil de luz de una mejor manera que el de de Vacouleurs, sino que también se han encontrado correlaciones entre los parámetros de Sérsic y las propiedades galácticas, tales como su tamaño y su magnitud absoluta.

Un método para ajustar los datos con el perfil de Sérsic es tomar el logaritmo natural a la función ya dada:

$$ln\left[I\left(r\right)\right] = \left[ln\left(I_e\right) + b_n\right] + \left[\frac{-b_n}{r_e^{1/n}}\right] r_e^{1/n} \tag{3.14}$$

Al definir $x = r$ y $y = ln\left[I\left(r\right)\right]$ la cuación se transforma a $y = A - Bx^C$. De aquí se obtendrá r_e dado por:

$$r_e = \left(\frac{b_n}{B}\right)^n \tag{3.15}$$

la intensidad efectiva I_e

$$I_e = e^{(A-b_n)} \tag{3.16}$$

y por último el índice de Sérsic

$$n = 1/C \tag{3.17}$$

El cociente entre la luminosidad del bulbo y del disco se denomina relación bulbo a disco

$$\frac{B}{D} = 3{,}61 \left(\frac{r_e}{r_0}\right)^2 \frac{I_e}{I_0} \tag{3.18}$$

Estudios realizados principalmente en galaxias de campo, encuentran que esta relación correlaciona con el tipo morfológico de Hubble ([Kent, 1985], [Schechter & Dressler, 1987], [Andredakis et al., 1995]) y con el parámetro de forma n de la ley de Sérsic ([Andredakis et al., 1995], [Graham, 2001]). En esta tesis utilizaremos el cociente bulbo/total como un indicador morfológico.

3.4. Colores de las Galaxias.

Se hicieron estudios sobre la dependencia del color de las galaxias con la morfología [Holmberg, 1958], [Roberts & Haynes, 1994].

Lo que se encuentra es que las galaxias de tipo esferoidal son generalmente rojas, mientras que las galaxias tipo disco e irregulares, muestran colores azules. Los colores de las galaxias proporcionan información sobre sus historias de formación estelar, reflejando la población estelar dominante. En particular, el color $(B - V)$ mide la actividad de formación estelar actual y pasada de una galaxia [Roberts & Haynes, 1994]. Las poblaciones estelares jóvenes, tienen emisión en el ultravioleta y colores ópticos muy azules.

A medida que la población estelar envejece, sus colores ópticos se enrojecen.

Cuando todos los colores son considerados juntos, la distribución de los colores de las galaxias puede ser aproxiamada por la suma de dos gaussianas normales, es decir, una función bimodal ([Strateva et al., 2001], [Blanton et al., 2003], [Kauffmann et al., 2003]). El valor medio y la varianza de estas dos distribuciones son funciones de la luminosidad o de la masa estelar ([Bernardi et al., 2003], [Blanton et al., 2003], [Hogg et al., 2004]).

La explicación más natural para la distribución bimodal de los colores de las galaxias, es que la misma representa dos poblaciones de estrellas diferentes. No obstante, tanto Driver et al. [2006] como Allen et al. [2006] proponen que

la bimodalidad refleja las dos componentes de las galaxias (bulbo y disco), más que dos poblaciones diferentes de galaxias.

Baldry et al. [2004] y Balogh et al. [2004] han propuesto un escenario donde la fusión de las galaxias es la responsable de la bimodalidad. La población roja se forma a través de fusiones de galaxias, mientras que la población azul, forma estrellas a una tasa determinada por procesos internos. En este escenario, la naturaleza gaussiana de la distribución bimodal de los colores se conserva.

3.5. Masas estelares.

Para poder medir las masas de las galaxias se utilizan distintas técnicas basadas en distintos parámetros como: tamaños, movimientos relativos e incluso utilizando la hipótesis de que el sistema en estudio se encuentra ligado gravitacionalmente.

Algunas de las técnicas utilizadas para poder medir las masas de las galaxias son imprecisas, como por ejemplo se calcula a partir de las poblaciones estelares, pero debido a la falta de conocimiento en el límite de masas pequeñas de la Función inicial de Masa (IMF)[1]. Casi siempre se utiliza la razón masa-luminosidad para las galaxias, en lugar de solamente las masas. Tanto la masa como la luminosidad están bien correlacionadas dentro de una misma clase morfológica, y la relación M/L media es una de las claves en la determinación de la densidad promedio de materia en el Universo.

Para calcular las masas o la relación M/L de las galaxias, debe especificarse el valor de la constante de Hubble [Poveda, 1958], [Poveda, 1961], ya que el valor de la luminosidad varía como el cuadrado de la distancia. También se indica el sistema de magnitudes empleado para determinar la luminosidad, ya que distintos sistemas miden luminosidades a diferentes radios y además, las galaxias tienen un rango amplio de colores.

Las masas de las galaxias espirales se determinan a partir de sus **curvas de rotación**. Estas galaxias son circularmente simétricas, de tal modo que las **velocidades radiales aparentes** se pueden corregir por inclinación.

[1]La función inicial de masas (IMF) es la distribución diferencial de estrellas en función de sus masas en regiones de formación estelar

De esta manera, la velocidad de las partes más externas sirve para estimar la masa total a distancias más próximas del centro galáctico.

Si la distribución de masas fuese esférica, el problema se reduciría al caso de una órbita circular de radio R al rededor de una masa puntual M:

$$\frac{1}{2}mV^2 = \frac{GmM}{R} \tag{3.19}$$

En donde m es la masa de la partícula (estrella) y V su velocidad orbital. De la ecuación anterior, despejando la masa puntual M:

$$M = \frac{1}{2}\frac{V^2R}{G} \tag{3.20}$$

Dado que la distribución real de materia es aplanada, se debe aplicar una pequeña corrección. Si la distribución de masa de una galaxia espiral disminuye con el radio al igual que la luz en el óptico, las curvas de rotación serían keplerianas, con una caída de velocidad $V\alpha R^{-1/2}$. Sin embargo, las observaciones a 21 cm, que se extienden muy por encima de los diámetros ópticos de las galaxias, muestran una curva de rotación plana. Este hecho implica que las masas de las espirales aumentan linealmente con el radio, lo que a su vez significa que la relación M/L en las regiones externas de las espirales aumenta exponencialmente. Las masas de las galaxias elípticas se determinan a partir de sus velocidades de dispersión y de medidas de sus radios característicos. Para un sistema en equilibrio, se tiene por el **Teorema del Virial**:

$$W + 2T = 0 \tag{3.21}$$

Donde T es la energía cinética del sistema y W la energía potencial gravitatoria. Para un sistema esférico y con la hipótesis de que la dispersión de la velocidad radial es una medida de las velocidades de las estrellas respecto a su centro de masas, se tiene:

$$M = \int_o^R \frac{M(r)}{r}\, dM \tag{3.22}$$

Si la galaxia se puede ajustar bien por la ley de de Vaucouleurs, se tiene que la energía potencial gravitatoria está dada por:

$$W = -0{,}33\frac{G}{M^2 r_e} \tag{3.23}$$

3.6. Conclusión

Queda claro que aún existen diversos parámetros de las galaxias –incluso en el Universo Local – que no se han podido estudiar a detalle, ya que la propia resolución de los telescopios actuales no nos permite llegar más lejos, como por ejemplo, el caso de los bulbos. Sin embargo, se espera que con los nuevos telescopios se puedan hacer estudios, con estadísticas significativas, que nos permitirán hablar con mayor certeza de los modelos de evolución y formación de las galaxias, y para ello es preciso tener muestras completas de objetos.

En este trabajo de tesis se pone a prueba la herramienta computacional desarrollada **PICASSO**, la cual tiene un gran potencial – incluso con las bases de datos actuales – y sobre todo, con la bases de datos que se presentarán en los siguientes 10 años. Se obtiene una muestra bastante grande, comparadas con las que hay en la literatura actualmente, de descomposición Bulbo/Disco contra la Masa estelar, que se aplicará para hacer un estudio formal de las propiedades de los bulbos de las galaxias en el Universo local.

4

Catastros Observacionales: La muestra.

4.1. Introducción.

El imparable auge que representa la cantidad de datos astronómicos así como el acceso a éstos, es un fenómeno que se está viviendo no sólo en astronomía, sino en diversas ciencias, en los dispositivos electrónicos, aplicaciones, individuos e incluso organizaciones o instituciones. A este auge, se le conoce como **Tsunami Digital**, y en particular, está provocando una serie de cambios en la forma en que se hace la astronomía actual y la forma en la que se hará en la siguiente década.

Gracias a la disponibilidad de grandes conjuntos de datos, al desarrollo de herramientas computacionales para un análisis más eficiente, ahora ya es posible explorar y analizar muestras de objetos astronómicos que antes eran impensables, en particular se pueden hacer estudios en muestras de galaxias de forma totalmente cuantitativa [Simard et al., 2011].

Gracias a dichas mediciones cuantitativas, en los últimos años se ha tenido un avance significativo en el conocimiento de las propiedades de galaxias [Peng et al., 2002], [Simard et al., 2002], [de Souza et al., 2004], [Lotz et al., 2004], [Conselice, 2006], [Pignatelli et al., 2006]. Este tipo de estudios proveen

un marco adecuado para hacer comparaciones entre teoría y observaciones [Simard et al., 2011].

En este trabajo de tesis, se utiliza la herramienta computacional llamada **PICASSO** (ver Capítulo 2), desarrollada para manipular, manejar y analizar grandes bases de datos astronómicas.

Con **PICASSO**, a los elementos de las tablas que conforman las diversas bases de datos, se les asigna una jerarquía, por medio de diferentes índices. Los cuales se seleccionan para crear una tabla cuya jerarquía sea como la de un árbol binario. Esto, permite hacer las búsquedad bajo el tiempo, costo y eficiencia que tienen las búsquedas en los árboles binarios.

Se crean distintos subniveles ya dentro del árbol binario, como ya se explicó (ver capítulo 2), lo que permite explorar con detalle e incluso correlacionar distintas propiedades de objetos astronómicos que se encuentran en distintas tablas o en distintas bases de datos.

Se pueden hacer búsquedas eficientes y no sólo de los objetos cercanos a una galaxia, o encontrar todos los objetos dentro de una misma región, o encontrar la misma galaxia en diversos catálogos, sino también se pueden hacer búsquedas entre todas las galaxias, de uno o varios catálogos, que cumplan con ciertas propiedades. Para que estas búsquedas sean lo más eficientes posibles, se debe tener una claridad de cómo se va a usar el árbol y los tipos de búsqueda que se van a hacer, de preferencia, cuando se genera el árbol y se le asigna una jerarquía, es decir, cuando se decide la cantidad de subniveles que éste debe tener, qué propiedades, de las galaxias, en este caso, son las más adecuadas para ligarlas al árbol, dependiendo del problema al que se le trate de dar solución. Se pueden crear en una misma tabla distintas jerarquías, o para distintas características crear distintos árboles, desde luego la manera más eficiente, en espacio, es crear un único árbol que se comporte de la misma manera para todas las tablas de las diferentes bases de datos, sin embargo, no es el único esquema.

Utilizando **PICASSO** *se obtiene un catálogo* – de los más grandes en la literatura – *de la relación que existe entre la fracción Bulbo al total contra la masa estelar* estimada para galaxias provenientes del Sloan Digital Sky Survey. La obtención de dicho catálogo se describe a continuación.

4.2. La muestra.

En este capítulo nos enfocaremos en la correlación de dos distintos catálogos observacionales de galaxias. También se hicieron distintas pruebas y correlaciones con catálogos más grandes, sobre todo para las comparaciones de tiempos de búsqueda.

En este trabajo de tesis se utiliza **PICASSO** para correlacionar galaxias contenidas en diferentes catálogos públicos de galaxias.

Los catálogos de galaxias observacionales utilizados en este trabajo de tesis son públicos y ambos provienen del Sloan Digital Sky Survey, Data Release Seven (SDSS DR7) [Abazajian et al., 2009] y más concretamente, originados por "The Legacy Survey"[1] del SDSS.

A continuación, se hablará de cada uno de los dos catálogos.

4.2.1. Catálogo de descomposición Bulbo+Disco para 2.20 millones de galaxias: Catálogo fotométrico.

El primero de los catálogos que se utiliza en este trabajo de tesis es el publicado por Simard et al. [2011], al cuál le llamamos: el catálogo **S11_entero**.

El **catálogo S11_entero** es, hasta ahora el catálogo más grande – en la literatura – que contiene información de la relación que existe entre la fracción Bulbo al Total **(B/T)**, en las bandas g y r. Las diferentes mediciones fotométricas fueron obtenidas al hacer un ajuste paramétrico a las galaxias. Se aplicaron ajustes bidimensionales de brillo superficial a las galaxias, usando un programa llamado **GIM2D** [Simard et al., 2002]. Cada uno de los elementos de la muestra se analizó haciendo uso de las funciones de Sérsic, exponencial, y la función **Sky** (cielo) del propio programa.

GIM2D es un algoritmo que se utiliza para ajustar modelos bidimensionales axisimétricos de galaxias directamente a las imágenes [Simard et al., 2002].

[1] El SDSS Legacy Survey proporciona un mapa uniforme y bien calibrado en u, g, r, i y z para más de $7,500$ grados cuadrados en el Norte y tres rayas en el Sur (740 grados cuadrados).

Las funciones de los modelos son: Sérsic/de Vaucouler, exponencial, PSF, así como la opción **Sky** o cielo, el cual hace referencia a la oscuridad del cielo y al número de cuentas que éste provoca en la lectura del CCD [Simard et al., 2002]. Además de **GIM2D** existen otras rutinas que ayudan a modelar perfiles galácticos como son **Ellipse** [Jedrzejewski, 1987], **GALFIT** [Peng et al., 2002] y **BUDDA** [de Souza et al., 2004].

S11_entero contiene información de parámetros estructurales (como tamaño del bulbo, tamaño del disco, ángulo de posición, índice de Sérsic, etc.), colores, así como tres distintas descomposiciónes del cociente de las escalas características de la componente del bulbo contra el total (**Total = Bulbo + Disco**) de las galaxias **B/T** para **2,195,875** galaxias. A dicha muestra se le aplican dos criterios principales, a la muestra que se obtiene con dichos criterios, se le llama **S11_limitada** o sólo **S11**. Dichos criterios son:

- Buscar objetos con $14 \leqslant m_{petro,r,corr} \leqslant 18$, donde $m_{petro,r,corr}$ es la magnitud de Petrosian en la banda r corregida por extinción galáctica, de acuerdo a los valores de extinción dados en la misma base de datos del SDSS.
- Que fuesen objetos extendidos, es decir que tengan un tipo morfológico en las bases de datos del SDSS tal que: $obj\_type = 3$, lo que se traduce a que sean *galaxias*.

Sobre este último criterio, en la clasificación morfológica del SDSS, se proporciona una descripción detallada que caracteriza la forma y la morfología de un objeto (criterio basado sólo en las imágenes), y así provee un separador muy simple para estrellas o galaxias a partir de la fotometría, el número tres, corresponde a la clase: *galaxias*.

Después de aplicar dichos criterios a la muestra, que como ya se mencionó contenía más de dos millones de galaxias, Simard et al. [2011] aplica distintos algoritmos ya estandarizados para estar seguros de eliminar los casos espurios y que las galaxias se encuentren tanto en **SEGUE** como en **The Legacy Area**.

Al final, **S11** contiene: $1,123,718$ galaxias con distinta información en las bandas fotométricas g y r.

Después de hacer esta selección Simard et al. [2011] utilizó tres diferentes modelos de ajuste para las galaxias:

1. Un perfil de Sérsic puro.
2. Un perfil de dos componentes: el cual utiliza un modelo de bulbo-disco fijo, para el cual utiliza un bulbo de Vacouleurs [de Vaucouleurs, 1948] y un disco exponencial, es decir, fijando la $n_B = 4$.
3. Un perfil de dos componentes: el cual utiliza un modelo de bulbo-disco libre o Sérsic libre, (es decir, n_B= *libre*) y un perfil de disco exponencial.

Estos tres distintos catálogos de galaxias ofrecen la posibilidad de hacer comparaciones y una amplia gama de estudios, tanto observacionales como teóricos, sin embargo – para los fines de esta tesis – sólo **se utiliza la base de datos generada a partir de aplicar modelos de bulbo-disco con indice de Sérsic libre** (el tercero de la lista anterior).

Se presenta la tabla 4.1 en la que se muestra el número de objetos de la muestra **S11** con el que se trabaja dentro de esta tesis.

Criterio de selección	Galaxias que permanecen (Removidas)
Simard et al. [2011] o **S11** entero	$2,195,875$
S11	$1,123,718$ $(1,072,157)$

Cuadro 4.1: Número de elementos en la muestra fotométrica **S11** sin y con limitaciones, ambas provenientes del SDSS.

4.2.2. Catálogo con información de Masas Estelares.

La segunda base de datos que se utiliza en este trabajo de tesis, es la publicada por Tojeiro et al. [2009]. Es un catálogo conformado por $683,113$ galaxias, contiene información de **masas estelares** al día de hoy, formación estelar, registro del enriquecimiento químico - metalicidades -, así como el contenido de polvo para los espectros de galaxias, todos estos datos fueron obtenidos a partir de utilizar galaxias del Sloan Digital Sky Survey.
A esta base de datos se le conoce como **VESPA** [Tojeiro et al., 2009], que es un repositorio de los datos obtenidos al utilizar el código **VESPA** (**VE**rsatile **SP**ectral **A**nalysis)[2], el cual está descrito a detalle en el artículo de Tojeiro

[2]http://www-wfau.roe.ac.uk/vespa/

et al. [2007].

En la página oficial de **VESPA**, se encuentran tablas con diferente tipo de información. Las cuales se describirán a continuación, pues fueron utilizadas en este trabajo de tesis.

4.2.2.1. Primera tabla de VESPA: *GalProp*.

La primera tabla de **VESPA** es *GalProp*. Contiene información de las galaxias como: un identificador único de **VESPA**, que se obtiene a partir de algunas propiedades descritas en las tablas del **SDSS**, la masa estelar M_*, el error para dicha masa estelar $\sigma(M_*)$ y más información del tipo espectroscópico.

Para los fines de este trabajo, sólo se utiliza la información descrita en el cuadro 4.2 de esta tesis.

Campo	Unidades	Descripción
Índice		Indicador único, construido con información del SDSS
Masa estelar	$M_\odot$	M_* Ecuación 4.10
Error de masa estelar	$M_\odot$	$\sigma(M_*)$ Ecuación 4.11

Cuadro 4.2: Se presentan los datos que se utilizaron de la tabla GalProp de Tojeiro et al. [2009], con información relevante acerca de las masas estelares para galaxias, en unidades de masas solares [$M_\odot$], así como un identificador único.

4.2.2.2. Masas y fracción de masas.

Lo que hace el programa **VESPA** [Tojeiro et al., 2007] para calcular las masas, es resolver el siguiente problema:

$$F_\lambda = \int_0^1 f_{dust}(\tau_\lambda, t)\psi(t)S_\lambda(t, Z)\, dt \tag{4.1}$$

Donde F_λ es el sistema de reposo observada de una galaxia, $\psi(t)$ es la SFR es la razón de formación estelar en términos de masas solares $M_\odot$, por unidad

de tiempo y $S_\lambda(t, z)$ es la luminosidad por unidad de onda de una población estelar individual (**S**ingle **s**tellar **p**opulation o SSP por sus siglas en inglés) con edad t y metalicidad Z, por unidad de masa. La dependencia de la metalicidad con la edad no tiene restricciones, convirtiéndolo en un problema no lineal.

VESPA recupera la fracción de formación estelar, que se transforma en masas absolutas.

Si λ está en el sistema observado, entonces se puede relacionar el flujo de una galaxia con la luminosidad emitida por medio de:

$$F_\lambda = \frac{L_{\lambda/(1+z)}}{4\pi D_L^2(1+z)} \tag{4.2}$$

En términos prácticos, se puede reescribir la ecuación 4.1, de manera simplificada:

$$F_j = \sum_\alpha x_\alpha S_{\alpha j} \tag{4.3}$$

Donde F_j se ha desplazado del sistema en reposo de la fuente y x_α es la fracción de formación estelar formadas en Δt_α.

La luminosidad de una galaxia, escrita en términos de los modelos de Tojeiro et al. [2009], es simplemente $L_j = \sum_\alpha u_\alpha S_{\alpha j}$, donde u_α es la masa estelar formada en α.

De 4.2:

$$F_j = \frac{\sum_\alpha u_\alpha G_{\alpha j}}{4\pi D_L^2(1+z)} \tag{4.4}$$

Para cualquier bin de edad α:

$$u_\alpha = x_\alpha 4\pi D_L^2(1+z) \tag{4.5}$$

Por lo que se puede distinguir entre la masa estelar formada en una galaxia y la masa estelar que queda en una galaxia al día de hoy:

$$M(t) = \int_0^1 \psi(t')[1 - R(t - t')]\, dt' \tag{4.6}$$

Donde $R(t - t')$ Es la fracción de masa estelar perdida en el tiempo t, por una población estelar de edad t' y $\psi(t')$ es la SFR en un tiempo t'.

En términos prácticos, se calcula como:

$$m_\alpha = u_\alpha R_\alpha \tag{4.7}$$

$$M^u_{*,fiber} = \sum_\alpha u_\alpha \tag{4.8}$$

$$M^u_{*,fiber} = \sum_\alpha m_\alpha \tag{4.9}$$

Donde R_α que es la fracción de reciclado, está dada por los modelos, para cada una de las metalicidades. R_α es típicamente del orden de 0,5 para las poblaciones de más edad, mientras que en los *bines* más jóvenes, la pérdida de masa es mucho menos significativa con R_α entre 0,7 y 0,9, dependiendo de la anchura del bin [Tojeiro et al., 2009].

Por último, se corrige por el hecho de que la fibra tiene una abertura de 3 segundos de arco, que significa que típicamente se observa la totalidad de la galaxia. Utilizamos la fibra observado y magnitudes de Petrosian en la banda z para aumentar la escala de la masa estelar como:

$$M_* = M_{*,fiber} x 10^{0,4(Z_p - f z_p)} \tag{4.10}$$

Donde z_p y Fz_p son la magnitud de Petrosian y la magnitud de fibra en la banda z, respectivamente.

4.2.2.3. Estimación de errores.

Se utiliza la matriz de covarianza completa para estimar estadísticamente el error en la masa estelar total M_*:

$$\sigma^2(M_*) = \sigma^2(M_{*,fiber})\beta^2 + M^2_{*,fiber}\sigma^2(\beta) \tag{4.11}$$

4.2.2.4. Segunda tabla de VESPA: *lookupTable*.

La siguiente tabla que se presenta en la página de **VESPA** se llama *lookupTable*, la cual contiene información espectroscópica obtenida de la tabla *SpecObj*[3] del **SDSS** para las galaxias de la tabla *GalProp*.

Los datos de la tabla *lookupTable* que se utilizaron para este trabajo de tesis, se presentan en el Cuadro 4.3.

Campo	Descripción
Índice	Indicador único, construido con información del SDSS.
SpecObjID	Es un identificador para la parte espectroscópica.

Cuadro 4.3: Se presentan los datos que se utilizaron de la tabla *lookupTable* de Tojeiro et al. [2009] con información del identificador espectroscópico del SDSS.

El *SpecObjID* es un identificador único del **SDSS**, todo los objetos observados en el cielo con el SDSS, tiene su código de identifiación.

[3]SpecObj es una tabla del Sloan Digital Sky Survey con información espectroscópica, la información que contiene dicha tabla se puede consultar en http://skyserver.sdss3.org/dr8/en/help/browser/browser.aspn=Specobj&t=U

4.2.3. Correlacionando las tablas de VESPA.

La tabla *GalProp* contiene información para 683, 113 galaxias, mientras que la tabla *lookupTable* contiene información para 781, 788 galaxias. Una de las primera búsquedas que se hicieron en este trabajo de tesis, fue una simple correlación entre dos tablas, las cuales contenían una columna, en donde los datos que coincidían deberían tener, lo que se llama *coincidencia de valor exacto*.

Para buscar en las dos tablas la coincidencia entre el índice único de **VESPA**, se necesita – como se dijo en el Capítulo 2–, un lenguaje que permita una comunicaión con la base de datos, éste es *Structured Query Language* (**SQL** por sus siglas en inglés) o *Lenguaje de Consultas Estructuradas*.

Para enviar consultas a la base de datos, primero debemos saber qué archivos contiene la base de datos. En realidad, dichos archivos son distintas *tablas* de datos – con columnas y filas –, por lo que se debe saber qué se busca y en dónde puede estar (en qué tablas se encuentran los datos). Así, tanto de la tabla *GalProp* como de la tabla *lookupTable*, se requiere saber el índice único de **VESPA** para tener el identificador espectroscópio del **SDSS**, junto con la información de las masas estelares, para esto se hace una simple consulta en **MySQL**:

$SELECT$
$lookupTable.SpecID, GalProp.index, GalProp.M\_stellar, GalProp.errM\_stellar$
$FROM$
$lookupTable, GalProp$
$WHERE$
$GalProp.index = lookupTable.index$

De este primer emparejamiento se obtuvieron 666, 266 galaxias.

A partir de esta búsqueda, ninguna otra fue hecha de manera directa. En particular, éste es el ejemplo más sencillo de emparejamiento: *coincidencia exacta*, pues se tienen dos tablas con una cantidad de sólo algunos cientos de miles de objetos (en este caso galaxias). Motivo por el cuál no es realmente necesario utilizar toda la infraestructura de construir un árbol, para poder comparar dos tablas, ya que se busca eficiencia no sólo de tiempo, sino de espacio. Sin embargo, debe quedar claro que se pueden hacer búsquedas cada vez más complicadas, y se trata de que éstas sean eficientes, pues el tipo de

búsqueda variará y dependiendo del problema, es la medida que se debe tomar. Teniendo siempre presente que los catálogos (cantidad de datos) serán cada vez más grandes [Simard et al., 2011].

Correlacionar como se ha hecho entre las tablas de **VESPA**, tiene un tiempo de búsqueda al menos proporcional a n, donde n es el número de objetos que hay en las tablas, desde luego, al utilizar un árbol binario, el tiempo de búsqueda se reduciría.

Para que las búsquedas sean lo más eficientes posibles se debe tener claridad de cómo se va a usar el árbol y los tipos de búsqueda que se van a hacer, de preferencia cuando se genera el árbol y se le asigna una jerarquía, es decir, cuando se decide la cantidad de subniveles que éste debe tener y qué propiedades de las galaxias – en este caso –, son las más adecuadas para ligarlas al árbol, pero debe quedar claro, que esto dependerá del problema al que se le trate de dar solución y sobre todo de la cantidad de objetos astronómicos que se tenga en un catálogo.

Es posible crear para una misma tabla distintas jerarquías, es decir, utilizar varios subniveles para distintos valores o propiedades: crear distintos árboles. Desde luego, la manera más eficiente – en espacio – es crear un único árbol con la cantidad mínima necesaria de subíndices (operaciones) y que la búsqueda se haga con la cantidad menor de objetos e incluso, cada árbol creado, tenga las mismas propiedades para todas las tablas de las diferentes bases de datos. Sin embargo, no es el único esquema a seguir.

En la tabla 4.4 se presenta el número de galaxias que se utilizan en las tablas de Tojeiro et al. [2009], que cuentan con información del identificador espectroscópico del SDSS, así como con las masas estelares para las galaxias, a esta muestra de $666,266$ galaxias a lo largo de este trabajo se le llama **VE09**.

4.3. Restricciones a la muestra para este trabajo.

Se utilizan los dos catálogos: el de Simard et al. [2011], **S11** que cuenta con $1,123,718$ y el de Tojeiro et al. [2009] **VE09** que cuenta con $666,266$ galaxias.

Criterio de selección	Galaxias que permanecen (Removidas)
GalProp	$683,113$
lookupTable	$781,788$
VESPA_correlación o **VE09**	$666,266$ $(16,847)$

Cuadro 4.4: Se presentan los datos del número de galaxias, que se utilizan en las tablas de Tojeiro et al. [2009], que cuentan con información del identificador espectroscópico del SDSS, así como con las masas estelares para las galaxias.

A continuación, en el cuadro 4.5 se presenta una tabla que contiene los números de galaxias para cada una de las tablas, la que tiene información espectroscópica y la que tiene información fotométrica.

Base de datos con información del tipo:	Galaxias que permanecen (Removidas)
Fotométrica **S11**	$1,123,718$
Espectroscópica **VE09**	$666,266$

Cuadro 4.5: Se presenta el número de galaxias que tiene cada una de las dos muestras con las que se va a trabajar, la de Simard et al. [2011] **S11** o fotométrica y la de Tojeiro et al. [2009], **VE09** o espectroscópica.

4.3.1. Definición de la muestra jerárquica: Correlación fotométrica y espectroscópica.

Ambos catálogos (Simard et al. [2011] y Tojeiro et al. [2009]) tienen un identificador proveniente del **SDSS** y aunque ambos fueron obtenidos del Data Release Seven (SDSS, $DR7$) [Abazajian et al., 2009], en particular de *The Legacy Area*, cada uno de ellos fue obtenido de una tabla central diferente.

Como ya se ha dicho anteriormente, los datos del **SLOAN** son de libre acceso a través de su servidor *web*, el *Skyserver*[4]. Tanto las imágenes, los espectros y las tablas con información de los objetos están disponibles para ser descargados, siempre y cuando el uso de dicha información no tenga fines comerciales.

[4]2http://cas.sdss.org/

El **SDSS** posee millones de datos sobre muchos objetos diferentes pero se necesita de alguna manera, poder comunicarse con el *lugar* en donde se encuentran esos datos para que al hacer una búsqueda, éste entregue todos los datos que solicitamos, es decir, saber la información que tiene cada tabla dentro de una base de datos. Para poder hacer una búsqueda en el *Skyserver* de **SDSS** se utiliza un lenguaje que nos pueda comunicar con la base de datos. Dicho lenguaje, es **SQL**.

Simard et al. [2011] utilizó la tabla *PhotoObj* mientras que Tojeiro et al. [2009] utilizó la tabla *SpecObj* del **SDSS**. El identificador único de *PhotoObj* es conocido como $objID$, mientras que, el identificador de *SpecObj* es el $SpecObjID$.

También existe un identificador llamado $bestObjId$ dentro de la tabla *SpecObj*, el cuál, entrega el código de identificación del objeto (correspondiente al $ObjID$), si es que se tiene información espectroscópica y fotométrica ya comparada de manera previa. Cabe resaltar que *NO* todos los objetos de la tabla *SpecObj* están dentro de la tabla *PhotoObj*, ni todos los objetos que están tanto en *PhotoObj* como en *SpecObj*, tienen el $bestObjId$. Aún así, para una primera correspondencia entre las galaxias de los catálogos **VE09** y **S11**, se buscó en el *Skyserver* los objetos que tuvieran el $bestObjId$ dentro de *SpecObj*, sin embargo, el resultado obtenido fue de casi $400,000$ galaxias. Motivo por el cuál, se obtuvieron para cada galaxia contenida en **VE09** y en **S11**, su ascensión recta (RA) y su declinación (Dec).

Después de explorar gráficamente el espacio de parámetros cubierto por estos dos catálogos – tanto el fotométrico como el espectroscópico – para entender mejor los criterios de selección, así como los posibles sesgos presentes en las muestras, se establecieron varios y distintos criterios de corte en el espacio de parámetros y así, como resultado obtener una muestra de galaxias, la cuál fuera adecuada para los objetivos del estudio que se hace en este trabajo de tesis.

4.3.2. Primer Criterio: Galaxias existentes en los catálogos ($S11$) y ($VE09$).

El primer criterio de corte consistió en correlacionar las dos muestras descritas en las secciones anteriores. Es decir, se hizo una correlación entre las

bases de datos de Simard et al. [2011] ($S11$) con el modelo **Bulbo al Total** (B/T) para Sérsic libre y la base de datos de **VESPA** ($VE09$).

Para esto, a partir de las coordenadas de cada galaxia, en particular de la declinación, se creó un modelo jerárquico de árbol binario, que a continuación se describe.

4.3.2.1. Creación del Árbol binario: Utilizando la herramienta PICASSO.

Dependiendo de la complejidad que tenga el problema a resolver, se puede utilizar un árbol directamente creado en **MySQL**, sin embargo si el árbol tiene una estructura más compleja, lo más sencillo, es crearlo con algún lenguaje de programación como C, $C++$, $Python$, etc.

Entre las cualidades disponibles que tiene **MySQL** están las **APIs** (Application Programming Interface por sus siglas en inglés, o Interfaz de Programación de Aplicaciones). Se puede trabajar con programas en diferentes lenguajes como C, $C++$, $Python$, etc.

A partir de las declinación (Dec), se van a construir los diferentes subniveles del árbol. En este caso, sólo **se selecciona crear un árbol binario** para la tabla de Simard et al. [2011], debido a que si el número de subniveles que contiene el árbol es suficiente, no se necesita emplear la estructura de árbol binario para ambas coordenadas (RA, Dec). Refiriéndose a *suficente* si la cantidad de objetos contenidos en cada hijo es pequeña en comparación al número n de objetos contenidos en la tabla principal.

Así que en realidad, en este caso no se construyen dos diferentes árboles, sino solamente uno.

Así pues, la forma en que se construye el árbol – al hacer la búsqueda sobre la coordenada Dec – para la muestra **S11** que contiene $1,112,606$ galaxias, cumple con los siguientes pasos, de manera ordenada:

1. La raíz del árbol representa un espacio de búsqueda en el intervalo [a,b].

2. Si la muestra contenida en el intervalo es mayor a 20 objetos astronómicos, entonces se crean dos nuevos árboles de búsqueda binarios que serán

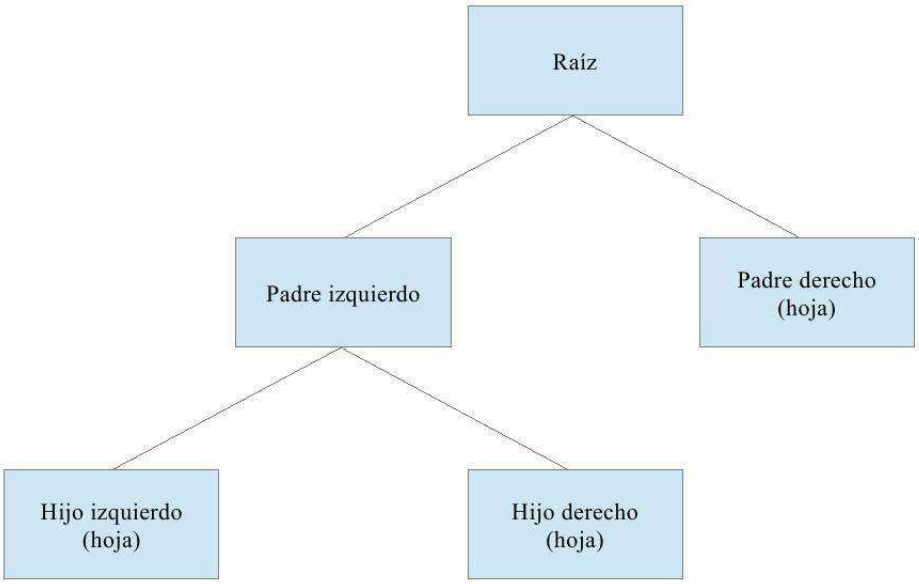

Figura 4.1: Representación esquemática del árbol binario producido con la herramienta **PICASSO.**

hijos del nodo raíz, de tal forma que el hijo izquierdo corresponde al espacio de búsqueda $[a, (a+b)/2]$ y el hijo derecho al espacio de búsqueda $[(a+b)/2, b]$.

3. Si el elemento buscado tiene $a \leq Dec \leq (a+b)/2$ y el hijo izquierdo tiene más de 20 elementos, entonces se repite el paso 2, usando como árbol al hijo izquierdo, de otro modo se revisa si $(a+b)/2 \leq Dec \leq b$ y de ser necesario se repite el paso 2 con el hijo derecho como árbol.

4. En la última iteración se tiene una muestra de a lo más 20 elementos astronómicos, y con esta cantidad de objetos, ya se puede hacer una búsqueda secuencial en tiempo constante.

Así para la muestra **S11**, con un nivel 15 de refinamientos, se llega a muestras que contienen al rededor de 20 elementos, es decir, de 20 galaxias. Como se ilustra en las figuras 4.1 y 4.2.

El siguiente código es una representación resumida de la construcción del árbol de búsqueda binaria y la búsqueda. Esta implementación se puede lograr

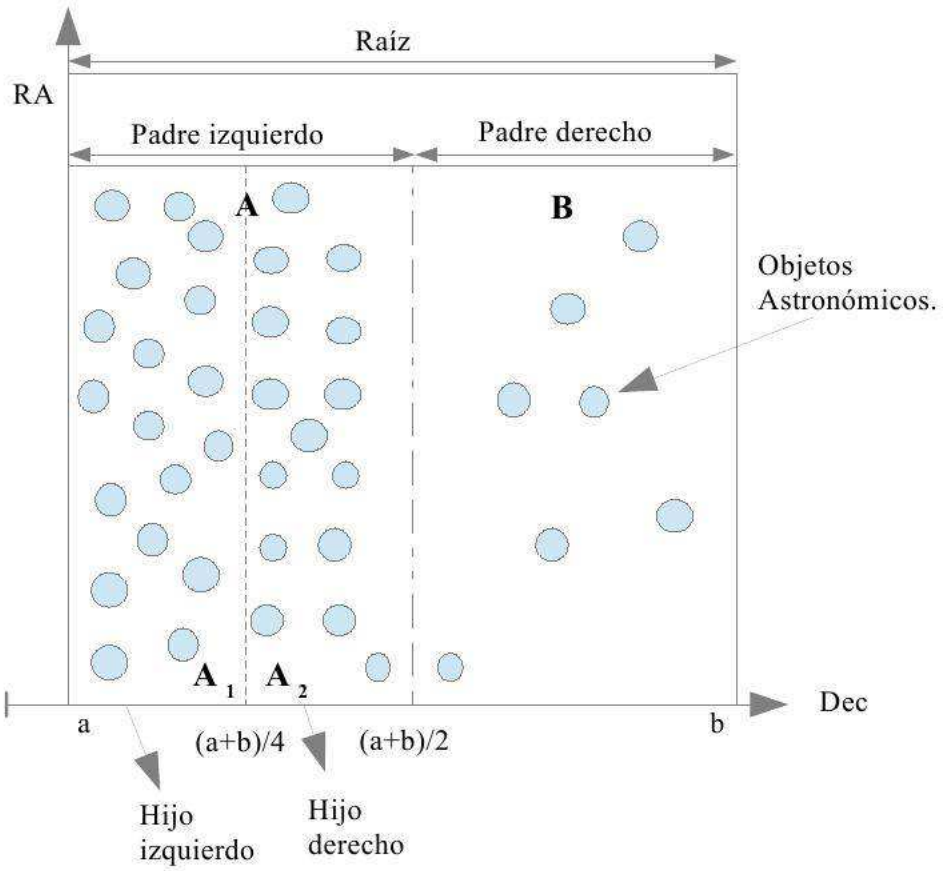

Figura 4.2: Creación del árbol binario con un espacio de búsqueda en el intervalo *raíz* [a,b]. Dado que la muestra contenida en el intervalo [a,b] es mayor a 20 objetos astronómicos, entonces se crean dos nuevos árboles o celdas **A** y **B** que son hijos del nodo raíz, por simplicidad, a cada uno de los hijos de la raíz se les llamó *padre derecho* y *padre izquierdo*. De tal forma, que el padre izquierdo corresponde al espacio de búsqueda $[a, (a+b)/2]$ y el padre derecho al espacio de búsqueda $[(a+b)/2, b]$. En particular, el padre derecho tiene menos de 20 elementos, por lo que ya no se hace ninguna división (el padre derecho, es también una hoja del árbol), sin embargo el padre izquierdo, entre $a \leq Dec \leq (a+b)/2$ tiene más de 20 elementos, entonces en este caso si se vuelve a hacer una partición. Dicha partición crea un **Hijo Izquierdo** o A_1 y un **Hijo Derecho** o A_2. Dado que tanto **Hijo Izquierdo** como **Hijo Derecho** contienen menos de 20 objetos, la anterior fue la última iteración y partición, ya que, con esta cantidad de objetos, se puede hacer una búsqueda secuencial en tiempo constante, y tanto A_1 como A_2 son hojas dentro de éste árbol.

fácilmente en $C++$ usando la API para la interacción con **MySQL** en las funciones espacio(a, b) y elem(a, b) al realizar consultas sobre el número de elementos con la condición de estar en ese intervalo:

```
    busquedaArbol(int n,int t,obj s,int a,int b) {
1           if(n > t) {
2                   return espacio(a, b);
3           } else {
4                   int mid = (a + b)/2;
5                   if(s.dec < mid) {
6                           return busquedaArbol(elem(a, mid), t, s, a, mid);
7                   } else {
8                           return busquedaArbol(elem(mid, b), t, s, mid, b);
9                   }
10          }
.   }
```

Donde n es el número de elementos de la muestra en el que se está buscando, t es el número máximo de elementos deseados en la muestra final, s es el elemento buscado ($s.dec$ representa la coordenada Dec), a y b son los límites del intervalo donde se realiza la búsqueda, espacio(a, b) indica que el espacio final está limitado por a y b, elem(a, b) es el conteo de la cantidad de elementos dentro de los límites a y b.

La construcción de un sólo árbol de búsqueda binaria con los suficientes refinamientos para la coordenada Dec permite tener muestras pequeñas en la misma búsqueda secuencial, para buscar Dec, también se puede encontrar RA. Es decir, no se tiene ya que hacer un árbol de búsqueda binario para la coordenada RA. Y cuando se hace la comparación entre **S11** y **VE09** ambas coordenadas se toman en cuenta.

Sin embargo, con la implementación del algoritmo mostrado, sólo se busca representar la forma en que el árbol de búsqueda binaria se ha implementado. Se trabajó con éste algoritmo asignando índices a los subárboles, de la manera en que se asignan índices con $MySQL$, sin embargo como se muestra en los resultados, asignar un índice por galaxia tiene un costo elevado en memoria y para galaxias cercanas el tiempo de búsqueda no se mejora.

En las ejecuciones realizadas se asignó un índice por cada dos iteraciones, teniendo en cuenta la propiedad de que en un árbol binario completo, el nivel $i+1$ tiene el doble de elementos que el nivel i, es decir que el nivel anterior. La raíz del árbol se encuentra en el nivel 1. Y con esto, se ahorra memoria en índices, pues en lugar de poner un índice por cada galaxia, se pone un índice por zona.

Con esta primera correlación se obtuvieron $656,411$ galaxias.

A continuación, en el cuadro 4.6 se presenta el número de galaxias que se obtiene.

Muestra	Galaxias que permanecen (Removidas)
S11 S11	$1,123,718$
VE09	$666,266$
Correlación S11-VE09	$656,411$ $(467,307)$

Cuadro 4.6: Se presenta el número de galaxias obtenidas al correlacionar tanto la muestra de Simard et al. [2011] **S11** o fotométrica y la de Tojeiro et al. [2009], **VE09** o espectroscópica.

4.3.3. Corrimiento al rojo máximo.

La siguiente submuestra que se genera tiene como parámero principal el corrimiento al rojo máximo (o redshfit). De manera que $z \leq 0{,}05$, donde z es el corrimiento al rojo de la tabla de Simard et al. [2011].

Desde la perspectiva de la medición de la distancia, en esta región: la distancia es independiente (a aproximadamente 10 %) del modelo cosmológico o el método de medición. Se puede utilizar:

$$cz = H_0 d \tag{4.12}$$

Donde $H_0 = 70 - 75 kms^{-1} Mpc^{-1}$ para calcular distancias extragalácticas en este rango. Con esta submuestra se eliminan todas las galaxias cuya magnitud aparente es muy baja.

Haciendo dicho corte se obtienen: 86,625 galaxias.

A continuación, en el cuadro 4.7 se presenta el número de galaxias que se obtienen con la restricción $z \leq 0{,}05$.

Muestra	Galaxias que permanecen (Removidas)
Correlación S11-VE09	$656,411$
$z \leq 0{,}05$	$86,625$ $(569,786)$

Cuadro 4.7: Se presenta el número de galaxias que se obtienen cuando se restringe a la muestra, de manera que: $z \leq 0{,}05$.

4.3.4. Galaxias que se encuentren de Frente.

La siguiente restricción que se aplica tiene la finalidad de obtener galaxias que se encuentren de frente, debido a que se pretende poder hacer una descomposición Bulbo-Total. Para determinar el ángulo se hicieron distintas pruebas, que nos permitieron llegar a la conclusión de que mientras más grande fuera el ángulo, más galaxias incluiríamos, y en particular, el criterio es bastante flexible debido a que si el disco de una galaxia tiene un ángulo de inclinación de incluso 55° aún se puede observar el bulbo y es posible aplicar a este tipo de galaxias una descomposición $Bulbo/Total$.

Después de aplicar esta restricción se obtienen $39,780$ galaxias.

A continuación, se presenta en el cuadro 4.8 el número de galaxias que se obtienen con la restricción $i \leq 55°$.

Muestra	Galaxias que permanecen (Removidas)
$z \leq 0{,}05$	$86,625$
$i \leq 55°$	$39,780$ $(46,845)$

Cuadro 4.8: Se presenta el número de galaxias que se obtienen cuando se restringe el ángulo de inclinación, de manera que $i \leq 55°$.

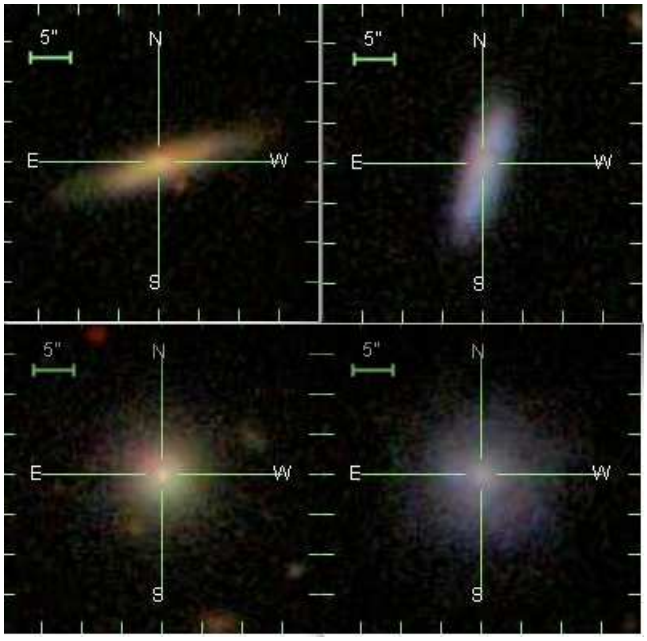

Figura 4.3: Las dos galaxias de la parte superior, son galaxias cuyos ángulos son $i \geq 55°$, es decir, no cumplen con la restricción descrita en esta sección, mientras que las galaxias de la parte inferior de la figura, son galaxias que cumplen con la restricción de $i \leq 55°$.

En la figura 4.3 se presentan 4 diferentes galaxias. Las de la parte superior del cuadro, son galaxias cuyos ángulos cumplen con $i \geq 55°$, ya que la primera galaxia tiene un $i = 79{,}18°$ y la segunda $i = 84{,}84°$.

Las galaxias de la parte inferior de la figura 4.3, son galaxias que cumplen con la restricción planteada en esta sección, es decir: $i \leq 55°$. La primera galaxia que se encuentra en la parte inferior, tiene un ángulo tal que: $i = 7{,}31°$ y la segunda galaxia inferior tiene $i = 1{,}95°$.

Además de hacer distintas pruebas, de la Figura 4.3, se puede ver que para las galaxias que se encuentran en la parte inferior sí se puede aplicar una descomposición $Bulbo/Total$, mientras que para las galaxias que se encuentran en la parte superior de la figura 4.3, es difícil distinguir hacia los centros, motivo por el cual, aplicar una descomposición $Bulbo/Total$ añadiría errores a los modelos, resultando así poco confiables.

4.3.5. Muestra limitada por Volumen.

El siguiente criterio de corte utilizado fue para limitar la muestra por Volumen. En un catálogo que se obtiene a partir de observaciones se tiene una magnitud aparente límite, es decir, el número promedio de objetos por unidad de volumen decrece con la distancia. Esto trae como consecuencia, que para distancias lejanas sólo se observan los objetos más brillantes, para eliminar cualquier posible sesgo, se construye una muestra limitadas por volumen.

Para limitar la muestra por volumen, se utiliza la magnitud absoluta de las galaxias, cantidad proporcionada por Simard et al. [2011] en sus propias tablas.

En la Figura 4.4 se muestra la relación entre la *magnitud absoluta en la banda r* M_r y el *corrimiento al rojo* (z). Se elige la muestra encerrada en rojo, pues es la que contiene mayor número de objetos.

Al hacer este corte y seleccionar la fracción encerrada en rojo de la Figura 4.4 se obtienen $38,893$ galaxias.

A continuación, se presenta en el cuadro 4.9 el número de galaxias que se obtiene al limitar la muestra por volumen.

Muestra	Galaxias que permanecen (Removidas)
$i \leq 55°$	$39,780$
Muestra limitada por volumen	$38,893$ (987)

Cuadro 4.9: Se presenta el número de galaxias que se obtienen cuando se limita la muestra por volumen.

Se seleccionaron distintas submuestras a partir de éste punto, que nos permite ver cómo se comporta la gráfica de Masa estelar con Bulbo-Total ($B/B+D$). Entre los distintos criterios se modificaron los límites del brillo, el índice de Sérsic, el radio efectivo, para una muestra, así como se utilizaron dos criterios extras que Simard et al. [2011] hizo de manera simultánea para restringir tanto el nivel adecuado de señal a ruido como buena resolución espacial.

La muestra se obtuvo a partir de el catálogo VESPA [Tojeiro et al., 2009], el cual, como ya se discutió, contiene información para las masa estelares M_* de

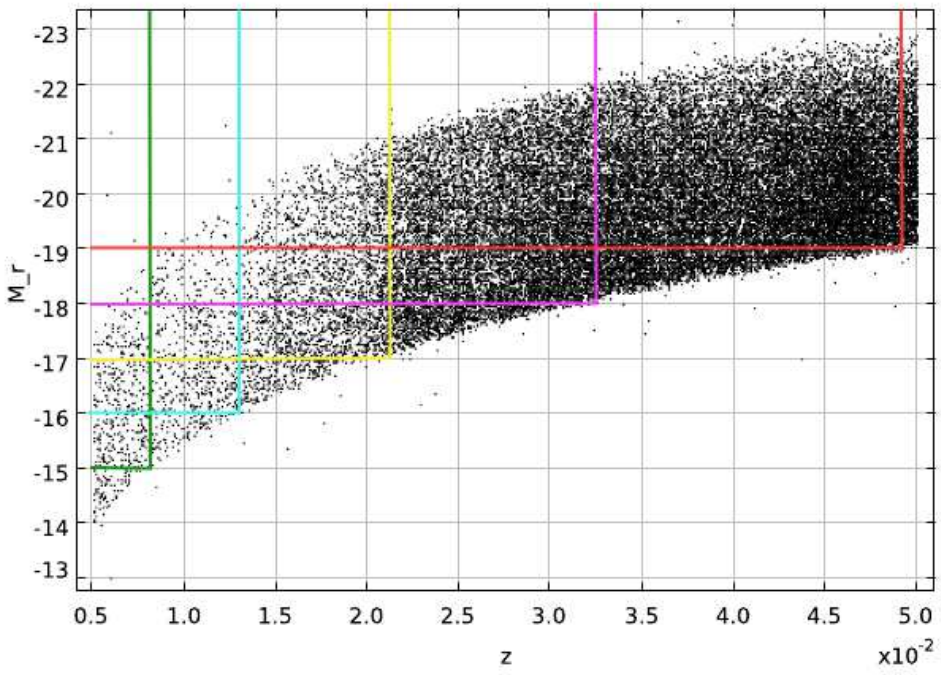

Figura 4.4: Magnitud Absoluta en la banda r (M_r) en función del corrimiento al rojo (z). Se presentan distintas submuestras limitadas por volumen, sin embargo, la que se utilizó en esta tesis es la encerrada en rojo pues contiene un mayor número de objetos.

una fracción de galaxias, en base a la distribución de energía espectral (SED, por sus siglas en inglés). Aunque Simard et al. [2011] asegura que su muestra es completa, al correlacionar tanto la muestra **S11** como **V09** se modifica la representatividad de la muestra original de Simard et al. [2011], por lo cual, el volumen de la muestra en el espacio de corrimiento al rojo es modificado, así como los límites en magnitud para recuperar la completez.

4.3.6. Por brillo superficial.

El siguiente corte que se hace es para obtener galaxias brillantes, así que se seleccionaron las que tuvieran brillos superficiales promedio de entre -19 y 22,5 magnitudes por segundo de arco cuadrado.

Al al aplicar este criterio a la muestra se obtienen $38,545$ galaxias.

A continuación, se presenta en el cuadro 4.10 el número de galaxias que se obtiene si se aplica el criterio $-22{,}5 \leq I_e \leq -19$.

Muestra	Galaxias que permanecen (Removidas)
Muestra limitada por volumen	$38,893$
$-22{,}5 \leq I_e \leq -19$.	$38,545$ (348)

Cuadro 4.10: Se presenta el número de galaxias que se obtienen cuando se limita la muestra por brillo superficial.

4.3.7. Índice de Sérsic

Se hicieron distintas submuestras teniendo en cuenta el índice de Sérsic, ya que en el artículo de Simard et al. [2011] se dan valores de este índice hasta 8. Al final, se optó por no hacer ningún corte en este caso.

4.3.8. Radio efectivo del semi-eje mayor del bulbo.

Este corte se decidió para poder tener imágenes de galaxias con una buena resolución angular, la cuál debe ser mayor a dos segundos de arco, esto al considerar el tamaño en pixeles de la imagen y el contraste con el *seeing*. Es decir, considerando la señal a ruido de la imagen, ya que si el tamaño de el radio efectivo del semi-eje mayor del bulbo es menor que la señal a ruido, las

conclusiones estarán afectadas ya que en realidad no se podrá discernir sobe la estructura interna de la galaxia.

De manera que $r_{eff_{disk}} > 2\ arcsec.$

Al al aplicar este criterio a la muestra se obtienen $38,396$ galaxias.

A continuación, en el cuadro 4.11 se presenta el número de galaxias que se obtiene si se aplica el criterio $r_{eff_{disk}} > 2arcsec.$

Muestra	Galaxias que permanecen (Removidas)
$-22{,}5 \leq I_e \leq -19.$	$38,545$
$r_{eff_{disk}} > 2arcsec.$	$38,396$ (149)

Cuadro 4.11: Se presenta el número de galaxias que se obtienen cuando se limita la muestra por $r_{eff_{disk}} > 2arcsec.$

En la figura 4.5 se presentan tres distintas galaxias con diferente tamaños en su radio efectivo.

4.4. Submuestra

Con todos estos criterios y restricciones, la muestra depurada queda con $38,396$ galaxias.

4.5. El Problema Astronómico

En el paradigma **CDM**, se espera que las galaxias que experimentaron un fusión mayor, al momento en que su masa era ya una fracción importante de la masa al dia de hoy, tengan una componente de bulbo significativa con un cociente **B/T** grande y un indice de Sérsic **n** también grande.

Dependiendo de la historia de estos mergers o fusiones, así como de la fracción de galaxias espirales que satisfacen este criterio, se podría tener una fracción grande/pequña de galaxias, que al día de hoy, presenten un cociente

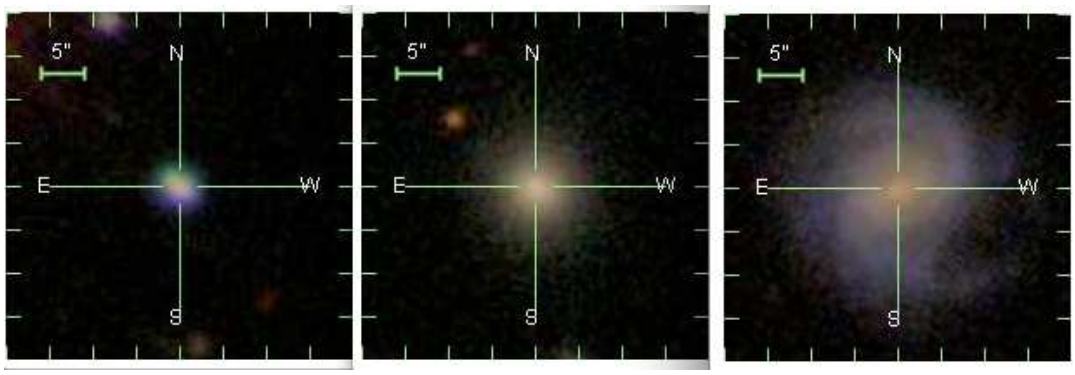

Figura 4.5: Mosaico de galaxias con distintos tamaños en su radio efectivo. La primera imagen, con los criterios descritos ($r_{eff_{disk}} > 2arcsec$) fue eliminada de la muestra final. No se puede hacer ningún tipo de descomposición a una galaxia que es más pequeña que la señal a ruido o el seeing, en las demás galaxias del mosaico se puede hacer una descomposición B/T.

B/T bajo. Sin embargo, recientemente se ha encontrado evidencia de una población significativa de galaxias con cociente **B/T** bajo o bien prácticamente sin bulbo en el Universo Local, especialmente en galaxias de masa baja o de tipo tardío. Barazza et al. [2007] analiza una muestra de varios miles de galaxias tardías del Sloan Digital Sky Survey **(SDSS)** encontrando que entre un (15-20) % de tales galaxias hasta $z < 0{,}03$ no presentan componente de bulbo.

La evidencia también sugiere que aún en galaxias de alta masa y tipos tempranos pueden existir bulbos con cocientes **B/T** bajos e índices de Sérsic **n** bajos.

Laurikainen et al. [2007] encuentra galaxias barradas y no-barradas de tipo temprano con cocientes **B/T** promedio entre 0.25 y 0.35, mientras que los cocientes **B/T** en galaxias de tipos más tardíos eran típicamente menores que 0,2 y con índices de Sérsic $n < 2.5$ a lo largo del esquema de Hubble.

Esta estadística emergente sobre la fracción de galaxias sin bulbo, es decir **B/T** = 0, y de galaxias con cocientes **B/T** bajos va en aumento y es necesario trabajo más detallado del que hay en la actualidad, para explorar la distribución y las propiedades de los bulbos tanto en galaxias de alta masa como de baja masa.

En este trabajo de tesis se aprovecha el recientemente auge de publicaciones en esta dirección por varios autores. Se aplica la herramienta computacional **PICASSO**, desarrollada en este trabajo, para el análisis de bases de datos; para explorar este problema como una primera aproximación.

Al inicio de este capítulo, ya se ha descrito un primer intento para obtener una muestra con una serie de restricciones observacionales, sobre las propiedades de los bulbos en galaxias del Universo Local de alta y baja masa.

De manera genérica los bulbos se dividen comúnmente en varios grupos: bulbos clásicos, bulbos en forma de caja/cacahuate y pseudobulbos o bulbos en forma de disco. Se cree que los bulbos clásicos se forman en procesos de fusiones mayores ($M1/M2 < 1/4$) y relajación violenta. Están asociados con valores del índice de Sérsic relativamente altos en el intervalo $2{,}5 < n < 6$ ([Hopkins, 2009]; [Springel & Hernquist, 2005]). Los bulbos en forma de caja/cacahuate se cree que son el resultado de resonancias verticales e inestabilidades de "buckling" en galaxias barradas y que además son vistos a altas inclinaciones ([Combes & Sanders, 1981]; [Bureau & Athanassoula, 2005]; [Athanassoula, 2006]; [Martinez-Valpuesta et al., 2006]). Los pseudobulbos o bulbos con forma de disco se cree que se forman como resultado de movimientos del gas hacia el *kiloparsec* central, con formación estelar subsecuente y formando paulatinamente una componente estelar con un cociente **B/T** alto, relativamente compacto ([Kormendy, 1993]; [Jogee, 1999]; [Kormendy & Kennicutt, 2004]; [Kormendy & Fisher, 2005]).

Los pseudobulbos tienden a tener índices de Sérsic $n < 2{,}5$ ([Kormendy & Fisher, 2005]; [Fisher & Drory, 2008]). Una posibilidad para la formación de los pseudobulbos es la idea de la evolución secular [Kormendy & Kennicutt, 2004] donde una barra estelar o una estructura globalmente oval en una galaxia no-interactuante y puede llevar flujos de gas hacia el kiloparsec central a través de choques o torcas gravitacionales.

Alternativamente se podrían generar estos flujos de gas por procesos no-seculares como las interacciones gravitacionales y las fusiones menores. En estos casos los flujos de gas podrían ser generados por estructuras no-axisimétricas estimuladas por las interacciones como las barras (e.g., [Quinn et al., 1993]; [Hernquist & Mihos, 1995]), y a través de torcas de marea producidas por la

galaxia compañera. La formación estelar central subsecuente podría formar una componente estelar compacta con un cociente **B/T** alto, como un pseudobulbo.

En esta tesis no se hace ninguna hipótesis sobre el origen de los distintos tipos de bulbos y, simplemente se hace referencia a ellos de acuerdo al valor del índice de Sérsic n y al valor del cociente **B/T**. Consideraremos bulbos con índices de Sérsic altos ($n > 4$), intermedios ($2 < n < 4$) y bajos ($n < 2$), así como cocientes **B/T** altos y bajos.

Las componentes estructurales de las galaxias tales como el bulbo y el disco se pueden obtener a través de la descomposición **2D** de la distribución de luz, tomando en cuenta la **PSF**.

Simard et al. [2011] llevó a a cabo una descomposición de este tipo para $1,123,718$ galaxias provenientes del catálogo fotométrico del **SDSS**. Sin embargo, otros trabajos han demostrado que es importante incluir la componente de barra en la descomposición de la distribucion de luz (e.g., [Laurikainen et al., 2005], [Laurikainen et al., 2007]) de las galaxias barradas, ya que de no hacerlo se pueden obtener valores del cociente **B/T** artificialmente inflados y propiedades de los bulbos que pueden estar sesgadas.

Los trabajos estadísticos recientes sobre la fracción de barras en el Universo Local ([Marinova et al., 2007]; [Menéndez-Delmestre et al., 2007]; [Méndez-Abreu et al., 2008], [Aguerri et al., 2009]; [Hernandez-Toledo et al. 2014 en preparación]) indican que esta fracción podría llegar hasta un 60 %, por lo que la inclusión de las barras es algo verdaderamente importante, que no se debería omitir en la descoposición **2D**. Algunos estudios recientes, han empezado a tomar en cuenta a las barras en la descomposición **2D**. Tal es el caso de Laurikainen et al. [2007]; Reese et al. [2007]; Gadotti [2008], Gadotti & Kauffmann [2009] y [Weinzirl et al., 2009].

En esta tesis se han revisado estos trabajos sobre descomposición bulbo/disco/barra y se ha complementado sus resultados con nuestra propia descomposición bulbo/disco/barra **2D** para una pequeña submuestra de (25 galaxias) del catálogo de descomposición bulbo/disco publicado por Simard et al. [2011], es decir e **S11**, para lo cuál se utiliza la rutina *GALFIT* [Peng et al., 2002]; [Peng, 2010]). Se han seleccionado al azar galaxias barradas y no barradas en el filtro r del Sloan con inclinaciones bajas y llevado a cabo:

- La descomposición bulbo/total para tratar de reproducir los resultados de la descomposición **2D** publicados por Simard et al. [2011]
- La descoposición bulbo/disco/barra para intentar predecir de manera gruesa cuales son los efectos de la no inclusion de la barra en la estimacion de los cocientes **B/T** e índices de Sérsic en las galaxias que presentan barra.

4.5.1. Preparación de las Imágenes y *GALFIT*

En cada bin de magnitud absoluta M_r y corrimiento al rojo z de nuestra muestra, se seleccionó un pequeño subconjunto de galaxias en la banda r teniendo cuidado de cubrir las regiones externas de cada galaxia durante el ajuste, así como regiones suficientemente lejanas para estimar el cielo.

Se llevó a cabo un examen rápido de las estrellas en el campo de cada galaxia para seleccionar las que tuvieran un alto cociente **S/N** para generar la imagen de **PSF**.

Se deja que *GALFIT* calcule de manera interna el mapa de ruido por pixel proporcionando información en el encabezado de las imágenes, como la ganancia, ruido de lectura, tiempo de exposición y el numero de imágenes combinadas.

Antes de ejecutar *GALFIT* se suavizaron las imágenes agrupando pixeles en formato 4 x 4, lo que permite garantizar, que el ajuste no se atorará en aquellas regiones que se encuentren muy saturadas y de esta manera convergiera.

GALFIT requiere de un "guess" razonable por cada componente que se ajusta. Así pues, se utiliza un algoritmo tipo $Levenberg-Marquardt\ algorithm$ (LMA) downhill gradient, para determinar la χ^2_{min} basado en los parámetros de entrada.

Para ajustar las componentes estructurales de las galaxias seleccionadas se utilizó un método iterativo invocando tres veces a *GALFIT*; para así ajustar una componente; dos componentes y tres componentes.

4.6. Conclusiones.

Se logró obtener una muestra completa de galaxias, a las que se le puede aplicar fácilmente distintos algoritmos que permiten la reproducción de los resultados obtenidos por Simard et al. [2011], así como se pueden aplicar otra serie de algorítimos para descomponer y tomar en cuenta la subestructura interna de las galaxias, encontrando así variaciones a los resultados propuestos por Simard et al. [2011], pero comparables a estudios como Weinzirl et al. [2009], Laurikainen et al. [2005] y Laurikainen et al. [2007]. Dichos resultados y comparaciones se muestran en el siguiente capítulo de este trabajo de tesis.

5

Discusión y Resultado.

5.1. Introducción

En este trabajo de tesis, se obtienen dos tipos de resultados. Uno, que es una herramienta computacional creada para manejar grandes bases de datos: **PICASSO**, y otro que se obtiene al aplicar dicha herramienta a problemas astronómicos concretos, en particular, se obtiene un catálogo – de los más grandes en la literatura – de la relación que existe entre la fracción Bulbo al total contra la masa estelar estimada para galaxias provenientes del Sloan Digital Sky Survey.

Al inicio de este capítulo, se presenta una discusión enfocada en los resultados astronómicos obtenidos y al final de este capítulo, se presentan algunos de los resultados computacionales obtenidos al utilizar **PICASSO** en bases de datos grandes ($\sim 200,000,000$ de objetos).

5.2. Parte I: Discusión y Resultados Astronómicos.

Los bulbos son estructuras muy comunes en las galaxias del Universo Local, por ejemplo, la Vía Látcea, Andrómeda, etc. tienen este tipo de estructuras en la parte central, relativamente pequeñas. Sin embargo, su estudio es aún incipiente.

Existen galaxias en las cuales la estructura esferoidal o bulbo es dominante, por lo que caracterizar cuáles son las propiedades más comunes en diferentes tipos de galaxias y en diferentes ambientes, es un elemento necesario para poder restringir cualquier teoría que pretenda explicar la formación y evolución de los bulbos.

Actualmente, existen tres diferentes estudios observacionales acerca de bulbos en galaxias cercanas ([Fisher & Drory, 2008], [Gadotti, 2008] y [Mendel et al., 2014]), de los cuales hablaremos más adelante con detalle, sin embargo, es importante hacer notar que los primeros que se mencionan, son estudios que utilizan galaxias muy cercanas en las cuales un análisis a detalle es abordable. Sin embargo, el tamaño de las muestras de ambos estudios es del orden de 70 galaxias en uno y algunos cientos de galaxias en el otro. Lo cual, implica que aunque se intenten correcciones para acercarse a una representatividad dentro de la población de galaxias, las fluctuaciones estadísticas (simplemente por el número de galaxias), resultan ser muy importantes.

Recientemente, con el sondeo que hizo el Sloan Digital Sky Survey (SDSS), se han abierto diversas posibilidades de analizar muestras mucho más grandes, más representativas y en particular de manera automatizada [Simard et al., 2011].

En particular, Simard et al. [2011] realizó un estudio automatizado, en el cual se realiza la descomposición Bulbo al total **B/T** en dos dimensiones para una muestra limitada por volumen para más de dos millones de galaxias **S11_entero**. En este caso la significancia estadístca no parece representar un grave problema. Sin embargo, no todas las galaxias en dicho estudio son candidatos adecuados para realizar un análisis de descomposición **B/T**. Para este estudio Simard et al. [2011], no se proporcionan otras cantidades físicas de las galaxias como lo son: masa estelar, actividad central, masa de gas, medio am-

biente, etc.

El estudio realizado en este trabajo de tesis, pretende cubrir esta necesidad mencionada, construyendo, diferentes submuestras que cumplan con una serie de criterios, para garantizar la validez de este análisis bulbo-al-total **B/T** y, que a partir de la correlación de la muestra de Simard et al. [2011], con VESPA [Tojeiro et al., 2009], y otras como ALFALFA [Kent et al., 2005] y UNAM-KIAS [Vázquez-Mata et al., 2010], etc., se pretenden buscar posibles correlaciones de las propiedades de los bulbos con los parámetros antes mencionados.

En este trabajo de tesis se aplican una serie de herramientas computacionales (Ver Capítulo 2) que fueron desarrolladas para el manejo y análisis de grandes bases de datos, que permiten explorar y correlacionar con detalle diferentes catálogos, en particular, como ya se ha mencionado, se utilizan dos catálogos observacionales de galaxias públicos, ambos provienentes del Sloan Digital Sky Survey Data Release Seven (SDSS, DR7) [Abazajian et al., 2009] y más concretamente, ambos provenientes de "The Legacy Survey" del SDSS [York et al., 2000].

El primer catálogo que se utiliza en este trabajo de tesis es el publicado por Simard et al. [2011].

A esta muestra, se le aplicaron distintos algoritmos – bastante estandarizados y bien descritos por [Simard et al., 2011] – para la eliminación de los casos espurios y objetos que no fueran galaxias, también se aseguraron de que las galaxias estuvieran tanto en SEGUE como en the Legacy Area y al final, obtuvieron $1,123,718$ galaxias con distinta información en las bandas fotométricas g y r. Después de haber hecho esta selección Simard et al. [2011] utilizó tres diferentes modelos de ajuste para las galaxias. El primero es un modelo de Sérsic puro, el segundo, un modelo de bulbo-disco fijo, en el cual utiliza $n_{B_{disk}} = 4$ y por último un modelo de disco-bulbo libre o Sérsic libre, (es decir, n_B libre). Para los fines de esta tesis, utilizamos la base de datos generada a partir de aplicar modelos de bulbo-disco con indice de Sérsic libre.

La otra base de datos que utilizamos, es la publicada por [Tojeiro et al., 2009], que es un catálogo con información sobre masas estelares al día de hoy, formación estelar, registro del enriquecimiento químico - metalicidades - así como el contenido de polvo para los espectros de galaxias obtenidos del Sloan

Digital Sky Survey. Esta base de datos conocida como VESPA [Tojeiro et al., 2009], es un repositorio de los datos obtenidos al utilizar el código **VESPA** (VErsatile SPectral Analysis) el cual está descrito en detalle en Tojeiro et al. [2007].

En este trabajo, se presentan distintas submuestras que fueron obtenidas a partir de la unión de éstos dos catálogos observacionales. (***Simard+VESPA***). A partir de esto, logramos presentar una de las muestra más grande hasta el momento que contiene información Bulbo al Total $\mathbf{B/T}$ contra masas estelares $\mathbf{M}_*$. En particular, con masas más pequeñas que en cualquier muestra que hasta hoy haya hecho estudios comparables a este. Así pues, en este capítulo se discuten los resultados que se obtuvieron al correlacionar la razón de Bulbo al total $\mathbf{B/T}$ contra masa estelar ($\mathbf{M}_*$).con distintos trabajos, como el de [Gadotti, 2008], quién realizó una descomposición bulbo/barra/disco usando imágenes en las bandas g, r e i para una muestra representativa del Sloan Digital Sky Survey con $1,000$ galaxias. [Gadotti, 2008] para obtener su muestra, tomó todos los objetos espestroscópicamente clasificados como galaxias en el SDSS Data Release 2 (DR2) aplicando los siguientes criterios:

- Galaxias que tuvieran un corrimiento al rojo en el rango $0{,}02 \leqslant z \leqslant 0{,}07$ Esto proporciona una muestra con un número estadísticamente significativo de las galaxias, cuyas imágenes tienen una resolución espacial física relativamente comparables.
- Dado que las galaxias enanas no eran objeto de su estudio, excluyó, todas las galaxias con masas estelares por debajo de 10^{10} masas solares. Las masas estelares las obtuvo de [Kauffmann et al., 2003]
- Teniendo la muestra con las dos restricciones anteriores, [Gadotti, 2008] obtuvo una muestra limitada por volumen, es decir una muestra que incluye todas las galaxias más masivas que 10^{10} masas solares en el volumen definido por el corte del corrimiento al rojo.

A partir de ahora, nos referiremos a esta muestra como la muestra **G08**.

La siguiente muestra con la que comparamos nuestros resultados es la presentada por [Fisher & Drory, 2011], la cual consta al principio de 320 galaxias, provenientes del Spitzer Space Telescope (**SST**) (3,6 micras) y del Telescopio

Espacial Hubble (**HST**), es una muestra limitada por volumen, de galaxias que no estén de canto, cumple con: $i < 80°$ dentro de una esfera de 11 Mpc (en el Universo Local), provenientes de Kennicutt et al. [2008] completo para galaxias espirales con B = 15 mag (correspondientes a $M_B = -15, 2$). Requirió también una Latitud Galáctica $\mid b \mid < 20°$. Sin embargo en su trabajo discute distintas muestras dependientes de la fotometría y su morfología, pero como esta información es poco confiable, introduce distintos sesgos y terminó con una muestra de 78 galaxias.

La muestra de Mendel et al. [2014] es un catálogo de bulbo/disco/masas estelares para un total de $\sim 660,000$ galaxias, todas del catálogo textbfS11.

A continuación se presenta la gráfica 5.1 Bulbo al total **B/T** contra Masa estelar $\mathbf{M}_*$, presentada por el artículo de Zavala et al. [2012a], en ella se presentan las muestras de Gadotti [2008] **G08**, Fisher & Drory [2011] así como predicciones teóricas sobre la formación de galaxias, y la formación de bulbos a partir de dos simulaciones la de Milenio I y Milenio II.

Lo puntos rojos representan la muestra observacional de Gadotti [2008] **G08**, mientras que los cuadrados azules, representan la muestra observacional de Fisher & Drory [2011]. La mediana de la distribución de las muestras de galaxias observacionales se muestran con líneas sólidas.

Se compa nuestros resultados sobre la gráfica 5.1. Los presentamos en la gráfica 5.2 con puntos verdes sobre la imagen, en esta primera gráfica no hay ningún tipo de selección o corte. Presentamos los datos obtenidos al hacer un match entre Simard et al. [2011] y Tojeiro et al. [2009].

Es deseable, que la muestra con la información fotométrica y espectroscópica tenga la mayor cantidad de elementos posibles, pero que estos sean independientes de sesgos y errores sitemáticos para poder poner a prueba tanto las observaciones echas como las conclusiones de estudios teóricos que se basan en utilizar como restricción la función de masa estelar de galaxias a diferentes épocas utilizando una predicción teórica de la función de masa de halos.

La muestra graficada es la **Correlación S11-V09**, es decir, es una muestra que no contiene ningún tipo de selección, por lo que al hacer esto, aunque es una muestra comparable a la de Mendel et al. [2014], se está introduciendo

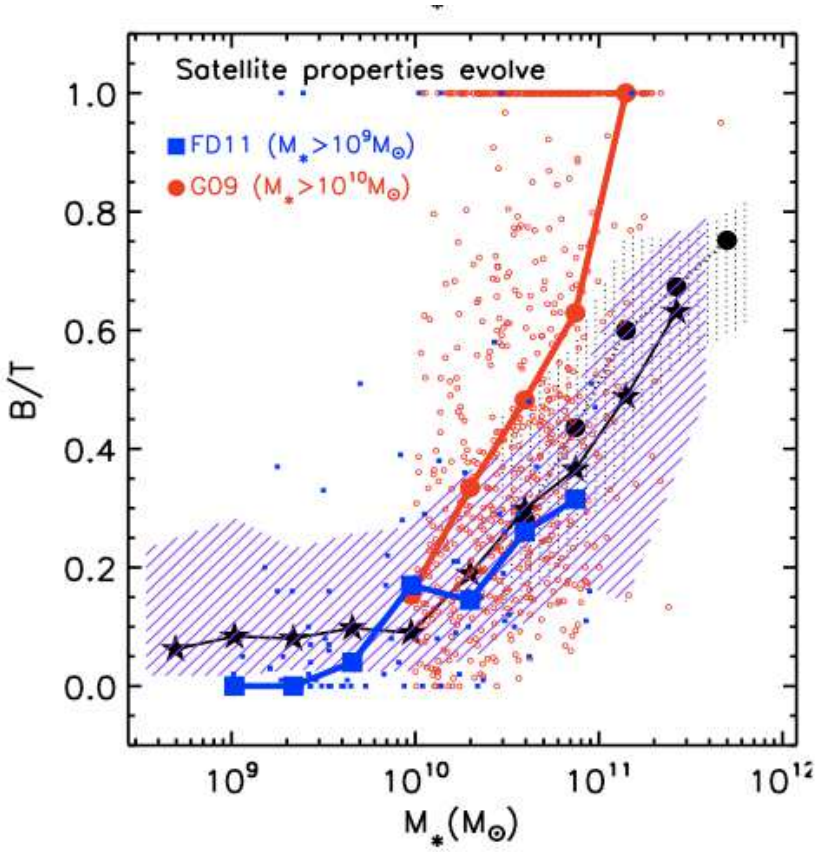

Figura 5.1: Razón entre el Cociente Bulbo al Total $\mathbf{B/T}$ con la Masa estelar $\mathbf{M_*}$. En esta gráfica se presenta en rojo las observaciones del artículo de Gadotti [2008] que consta de 1000 galaxias, en azul se presentan las de Fisher & Drory [2011] que son casi 100 y la parte sombreada – líneas azules – es la predicción teórica de Zavala et al. [2012a].

muchos sesgos, no es una muestra completa por volumen y pueden encontrarse galaxias que tengan cualquier inclinación, lo cuál, en realidad, dificultaría aplicar el algoritmo para hacer una descomposición Bulbo-Disco **B/T**.

Para esta muestra se tiene un $< V/V_{max} > = 0{,}32$.

La necesidad de crear una muestra completa con los dos catálogos observacionales Simard et al. [2011] **S11** y Tojeiro et al. [2009] **VE09**, da como resultado, distintas selecciones entre las galaxias obteniendo así, dos muestras, la primera es la que se describe en el *Capítulo 4*, **Muestra I**, pero, hubo a necesidad de crear otra muestra **Muestra II**.

Se trabajó con galaxias elegidas al azar para aplicar distintos algoritmos y reproducir las mediciones de Simard et al. [2011], y aunque se pensó que se había sido lo suficientemente estricto en las condiciones para que la muestra que se obtuviera, la de $38,396$ (**Muestra I**) descrita enteramente su obtención en el capítulo 4 de este trabajo de tesis, se pudo constatar, cuandoo se trabajaron con galaxias de manera individual, que no se había sido tan estricto en la selección, pues aún había varias galaxias que en realidad, al observarse su imagen (en http://cas.sdss.org/), resultaban difícil o imposible para hacerles la descomposición **B/T**, motivo por el cual se crea una segunda muestra, en la que se aplicaron criterios más estrictos a favor de obtener una mejor resolución espacial.

En la primera muestra: **Muestra I**, se hacen todos los cortes necesarios para poder tener una buena resolución en las imágenes, tener a las galaxias cercanas, tener una señal a ruido adecuada, galaxias con un rango de brillo amplio así como una muestra de galaxias limitada por volumen. Es decir: los cortes descritos con detalle en el capítulo 4:

- Galaxias que tuvieran un corrimiento al rojo en el rango $0{,}05 \leqslant z$ Esto proporciona una muestra con un número estadísticamente significativo de galaxias, cuyas imágenes tienen una resolución espacial física relativamente comparable entre ellas.
- Galaxias con un rango de brillo entre: $19 \leq s \leq 23$
- Galaxias con un bulbo cuyo radio efectivo fuese mayor que el seeing medio: $R_s \geq 2''$

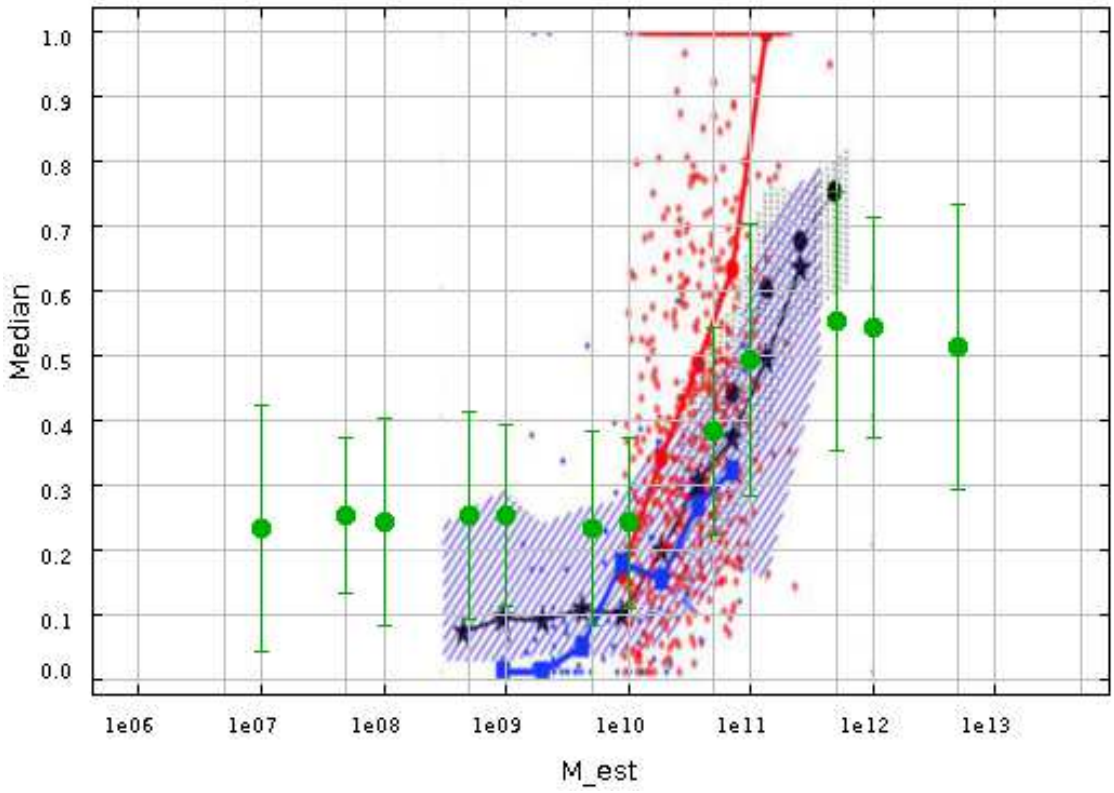

Figura 5.2: Razón entre el Cociente Bulbo al Total **B/T** con la Masa estelar $\mathbf{M}_*$. En esta gráfica se presenta en rojo las observaciones del artículo de Gadotti [2008], muestra llamada **G08**, la cual consta de ∼1000 galaxias, en azul se presentan las de Fisher & Drory [2011] que son del orden de 70 galaxias, la parte sombreada – líneas azules – es la predicción teórica de Zavala et al. [2012a]. Además, en esta gráfica se presenta la muestra entera de Simard et al. [2011] **S11_entera** con la muestra de Tojeiro et al. [2009] **V09** en verde, NO se ha hecho ningún tipo de tratamiento, por lo que consta de 656, 421 galaxias, esta muestra no es una muestra entera.

- Galaxias de frente, es decir que tuvieran un disco con un ángulo de inclinación (i) tal que: $i \leq 55°$
- Muestra limitada por volumen

Para dicha muestra se obtiene un $< V/V_{max} > = 0{,}52$ y es la gráfica 5.3.

Poniendo criterios de selcción con mayor restricción, sobre todo en cuando a resolución angular, se obtiene la **Muestra II** que contiene 35, 348 galaxias, y que se presente en la gráfica 5.4. Su $< V/V_{max} >$ es de 0,62.

En la gráfica 5.5 se presenta una comparación entre los resultados obtenidos con una muestra completa (en verde) con 38, 396 galaxias y los resultados que obtuvo Mendel et al. [2014], la muestra de Mendel et al. [2014] tiene 660, 000 galaxias, de las que con los criterios de selección que se describen en este trabajo fueron eliminadas, pues se ha demostrado que es difícil aplicar una descomposición **B/T**.

Se han presentado ya las distintas sub-muestras que se obtuvieron con los diferentes cortes que se describieron anteriormente. Estas muestras, con las distintas restricciones que se tomaron en cuenta como es que tengan buena resolución espacial, completez por corte de magnitud, tomando sólo galaxias con redshift pequeño, permite obtener la relación que existe entre la masa estelar $\mathbf{M}_*$ y la descomposición Bulbo-total **B/T** para un número de galaxias representativamente mucho mayor que las muestras de estudios previos hasta ahora con la más grande la de [Gadotti, 2008] que tiene cerca de 1,000 galaxias, que era hasta ahora la muestra más grande por número, pero no estadísticamente significativa.

5.3. El Impacto de las Barras en la descomposición 2D.

Resumimos brevemente los efectos de incluir el ajuste de la barra en la descomposición **2D** en las galaxias barradas seleccionadas.

En una descomposición bulbo-disco-barra de una galaxia barrada, hay una redistribución de la luz en tres componentes. Se encuentra que hay una tendencia a disminuir la luminosidad del bulbo así como una redistribución de

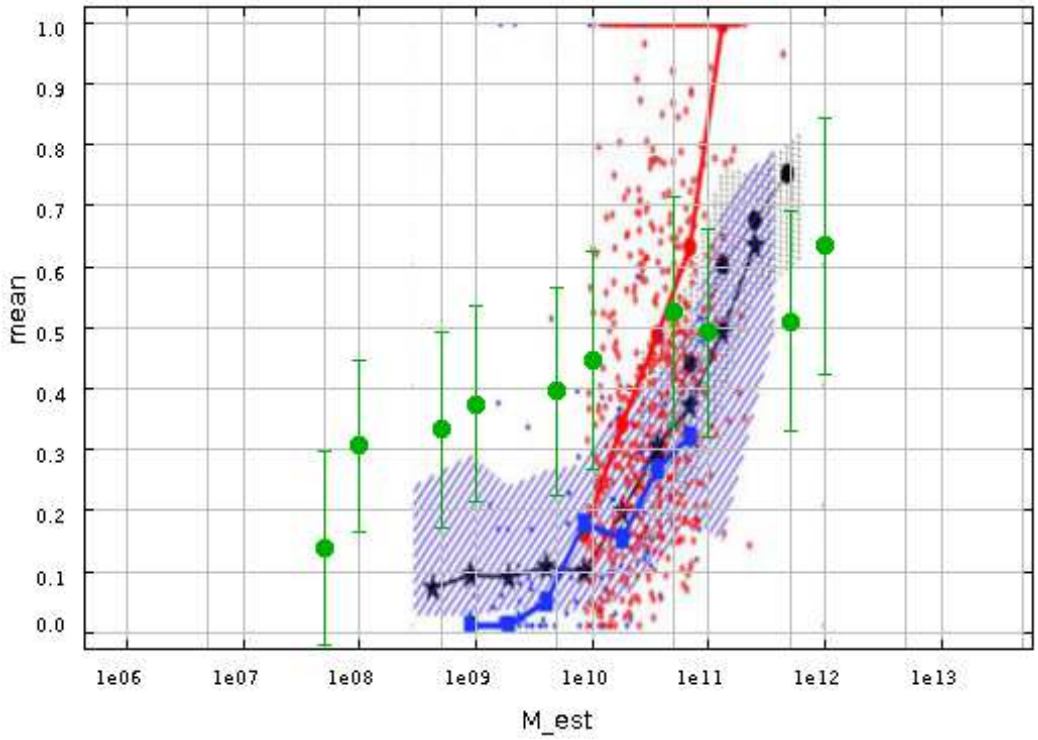

Figura 5.3: Razón entre el Cociente Bulbo al Total **B/T** con la Masa estelar $\mathbf{M}_*$. En esta gráfica se presenta en rojo las observaciones del artículo de Gadotti [2008], muestra llamada **G08**, la cual consta de ∼1000 galaxias, en azul se presentan las de Fisher & Drory [2011] que son del orden de 70 galaxias, la parte sombreada – líneas azules – es la predicción teórica de Zavala et al. [2012a]. Además, en esta gráfica se presenta la muestra con todos los cortes descritos en el Cap. 4 en verde, que consta de 38, 396 galaxias, llamado en la tesis: **Muestra I**.

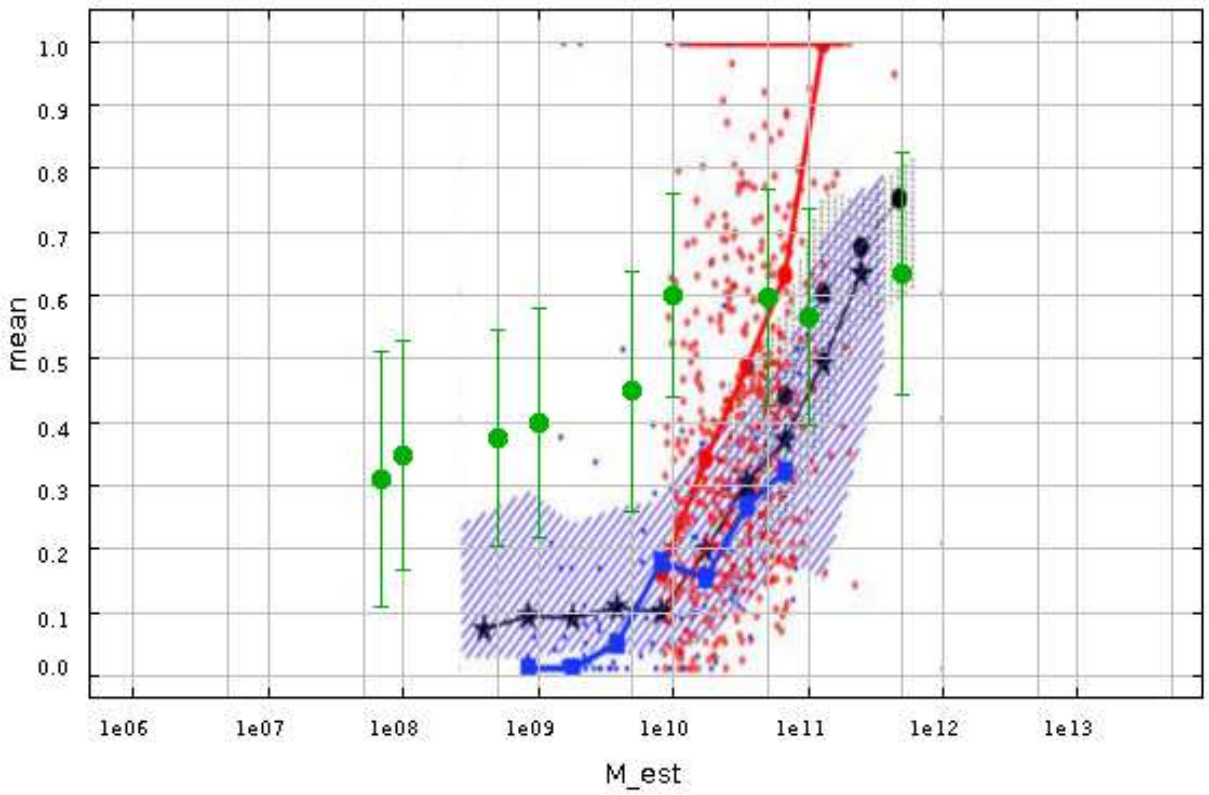

Figura 5.4: Razón entre el Cociente Bulbo al Total **B/T** con la Masa estelar $\mathbf{M}_*$. En esta gráfica se presenta en rojo las observaciones del artículo de Gadotti [2008], muestra llamada **G08**, la cual consta de ∼1000 galaxias, en azul se presentan las de Fisher & Drory [2011] que son del orden de 70 galaxias, la parte sombreada – líneas azules – es la predicción teórica de Zavala et al. [2012a]. Además, en esta gráfica se presenta la muestra en verde, que consta de 35, 348 galaxias, llamado en la tesis: **Muestra II**.

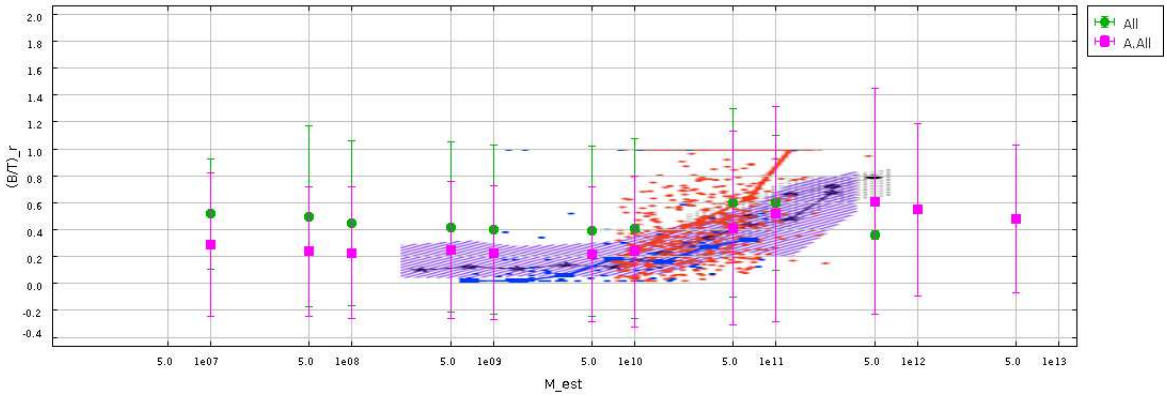

Figura 5.5: Razón entre el Cociente Bulbo al Total $\mathbf{B/T}$ con la Masa estelar $\mathbf{M_*}$. En esta gráfica se presenta en rojo las observaciones del artículo de Gadotti [2008], muestra llamada **G08**, la cual consta de ∼1000 galaxias, en azul se presentan las de Fisher & Drory [2011] que son del orden de 70 galaxias, la parte sombreada – líneas azules – es la predicción teórica de Zavala et al. [2012a]. Además, en esta gráfica se presenta la muestra con todos los cortes descritos en el Cap. 4 en verde, y se grfica en rosa la muestra de Mendel et al. [2014].

luz en el disco. La inclusión de la barra reduce la cantidad fracciónal de luz en el bulbo y el cociente **B/T** comparada con una descomposición bulbo-disco **B/D**. Esta reducción corresponde en promedio a factores del orden 2,5 con los cambios más notorios en galaxias más fuertemente barradas.

El radio de escala del disco es un parámetro robusto que cambia poco al incluir el ajuste de la barra. Sin embargo si el ajuste del disco es pobre, la inclusión de la barra puede causar cambios en los parámetros ajustados del disco.

Los resultados en esta tesis indican que se podría *sobre-estimar* el cociente **B/T** y el índice de Sérsic hasta en un 25 % si no se considera la barra, lo cual es consistente con los resultados de Gadotti [2008] y Weinzirl et al. [2009].

5.4. Conclusiones

Con esto, se pretende ilustrar y dar un ejemplo realista de la importancia actual que tiene una herramienta computacional como la que se desarrolló en esta tesis, **PICASSO**, la cual nos permite trabajar, correlacionar y manipular distintas bases de datos astronómicas.

5.5. Parte II: Discusión y Resultados Astronómicos.

Otra gran vertiente que se desarrolla en este trabajo de tesis es el crear una herramienta eficiente, que permita trabajar con bases de datos grandes, hoy en día la mayoría de bases de datos astronómicas no tienen más que unos cuantos cientos de miles de millones, pero como ya se ha discutido, este no es el panorama que se prevé para la siguiente década [Gray et al., 2007].

A partir de herramientas computacionales ya existentes, como lo son el gestor de datos **MySQL**, el visualizador **TopCat**, además de lenguajes de programación como Python y C++ se elaboraron rutinas que permiten integrar distintos tipos de informaciones generada, que nos permiten explotar las capacidades de dichas herramientas. **PICASSO** nos permite manipular grandes bases, a las que se les atribuye una jerarquía, que en este caso se atribuye

en forma de árbol binario y árbol binario completo. Con dicha jerarquía se logra correlacionar diferentes catálogos, encontrar galaxias con distintas propiedades, etc. de manera eficiente.

De manera general, se decide cómo implementar el árbol a partir de las propiedades mismas de la tabla y del problema que se planea resolver, por ejemplo, en el problema que se presentó anteriormente, al buscar las galaxias a partir de la coordenada *Dec*, se buscaron crear secciones que contuvieran una cantidad de al rededor de 20 objetos astronómicos, galaxias, ahora se pone a prueba la búsqueda de objetos dentro de tablas más grandes. Los resultados se muestran en la tabla 5.1. En particular, la primera búsqueda, la de 1290000, si es el tiempo que tomó hacer la búsqueda entre la tabla **S11** y **V09**.

Todos los árboles que se construyeron (5), los de la tabla 5.1, se construyeron de manera análoga al presentado anteriormente.

Se buscó que en cada hoja cuando menos hubiera del orden de 50 objetos y se probó haciendo cortes geométricamente iguales, es decir creando árboles binarios completos, aunque, en muchos casos si uno excede el número de subniveles por mucho, resulta que en algunas hojas ya no hay ningún objeto, es decir, se crea una construcción en forma de árbol binario completo, pero en realidad, sigues usando un árbol binario.

En particular, la herramienta desarrollada en este trabajo de tesis es altamente escalable.

Se utiliza una computadora de 8 GB de memoria RAM.

Se logró reducir los tiempos de búsqueda en comparación a otras estructuras como se muestra en la figura 5.6 y en la tabla 5.1.

No. de elementos	**TopCat**	**MySQL**	**MySQL** index	**MySQL** hierarc.: **PICASSO**
1 290 000	4 min	9 min	4 min	1 min
20 000 000	12 min	26 min	13.5 min	3 min
50 000 000	47 min	53 min	28 min	6 min
100 000 000	CRASH	260 min	39 min	8 min
200 000 000	CRASH	530 min	90 min	9 min

Cuadro 5.1: Tiempos entre distintos tipos de tablas

Con esta tabla, se puede observar que el tiempo si se reduce logariítmicamente cuando se utilizan los árboles binarios.

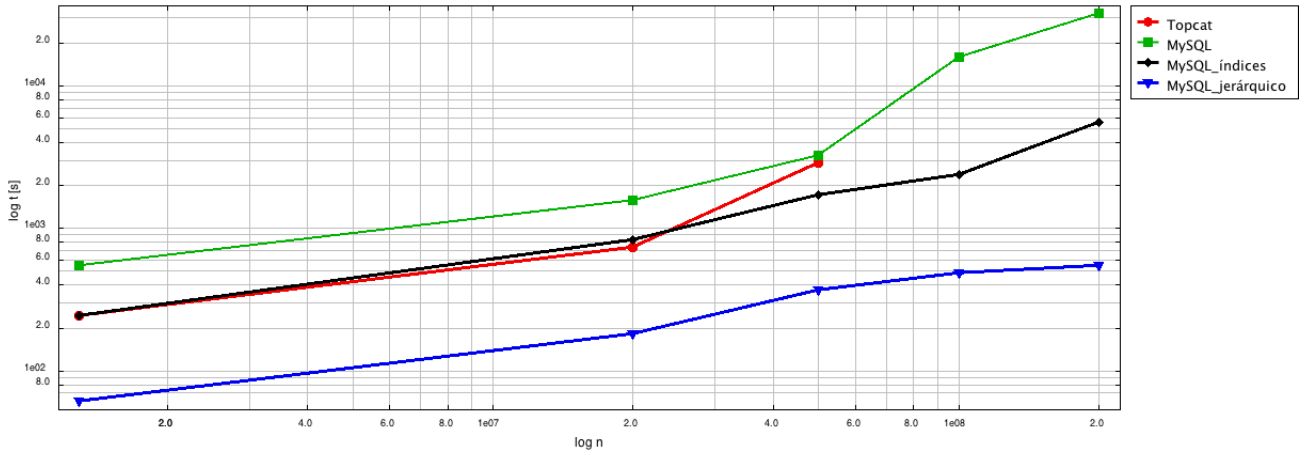

Figura 5.6: Comparación entre tiempos de ejecución para tablas con diferentes estructuras de datos. Línea roja es para un gestor llamado TopCat el cuál, tiene un límite y ya no funciona para más de 100,000,000 de objetos, el resto de líneas son construidas a partir de utilizar MySQL, la línea verde es cuando no se le da ningún tipo de estructura a los datos, la línea negra es cuando se le asigna un índice a cada uno de los datos y la línea azul es cuando se le asigna una esteuctura de árbol a los diferentes datos de la tabla.

Se ha desarrollado una herramienta que permite trabajar de manera eficiente con bases de datos actuales, pero que además es una herramienta que tiene alta escalabilidad y podrá ser útil aún con las diferentes bases de datos astronómicas que se prevén en los siguientes 10 años.

6
Conclusiones

En los últimos años ha existido un gran crecimiento en las capacidades de generar, administrar, trabajar y colectar datos [Bell et al., 2007].

La información disponible – de cualquier índole – y en particular, información astronómica proveniente, ya sea de telescopios o simulaciones, está teniendo un crecimiento, con el que resulta difícil lidiar. El crecimiento de información astronómica, se calculaba –hasta ahora– como un crecimiento que duplicaba la cantidad de información que se tenía anualmente [Gray et al., 2007], [Golden et al., 2013], dicho crecimiento es mayor que el que se predice con La ley de Moore [Moore, 1975].

Es importante almacenar toda la información recabada por los telescopios y por las simulaciones, sin embargo, almacenar dicha información, es sólo el principio; se tiene que saber qué hacer con dicha información, cómo analizarla y sobre todo, tener un acceso, que resulte eficiente y que se pueda trabajar y analizar dicha información.

La información astronómica que se tiene hoy en día, provenientes de diferentes telescopios, ha permitido generar diversos escenarios, pronósticar y confrontar distintas teorías. Lo que se puede traducir, en un mejor entendi-

miento del Universo [Simard et al., 2011].

En esta última década hemos presenciado la inversión de enormes esfuerzos enfocados a la generación de catastros tanto en el universo local (p.ej., el SDSS) como cosmológicos (incluyendo estudios del fondo cósmico en microondas, las oscilaciones bariónicas acústicas, la estructura a gran escala y la búsqueda a altos corrimientos al rojo de supernovas, galaxias y cuásares).

En astronomía, se han generado diversas bases de datos, provenientes de las observaciones obtenidas de los diferentes telescopios. Estas bases de datos están disponibles, para cualquiera que las solicite, en las diversas páginas web de los proyectos. Sin embargo, se tiene previsto que en los siguientes años, el tamaño, de dichas bases de datos, así como la calidad de información de las mismas, crecerá. Por mencionar un ejemplo: la cantidad de información que se capturó con el SDSS [York et al., 2000] – desde el 2000 hasta el 2010, se obtendrá en menos de una semana, en cuanto el telescopio LSST [Tyson, 2002] empiece a generar información.

A este gran crecimiento exponencial en la cantidad de datos, es a lo que se le conoce como **tsunami digital** [Zhang & Chen, 2010].Y, se tiene que tomar en cuenta, desde ahora, para poder dar respuesta a cómo es que se va a trabajar con dichas cantidades de información. En ese sentido y previendo dichas necesidades, se pensó en crear la herramienta **PICASSO**.

Por un lado, los estudios de las propiedades de las galaxias locales ahora son posibles para muestras que en número antes eran inimaginables. El SDSS ha logrado obtener imágenes para aproximadamente $11,663\ deg^2$ del cielo en 5 bandas y espectros para $929,555$ galaxias en el Universo Local. Ahora contamos con información en forma fotométrica y de parámetros estructurales para unos 360 millones de objetos, y se hace notar, que justamente el análisis de estas enormes muestras de objetos, tanto espectralmente como fotométricamente, son los principales objetivos detrás de la investigación astronómica actual. Sin embargo, en los próximos años, otros catastros a gran escala incluyendo Pan-STARRS ([Kaiser et al., 2002]) y LSST([Tyson, 2002]) incrementarán en orden de magnitud la cantidad de datos accesibles.

Por otro lado, los catastros cosmológicos han permitido un entendimiento considerable de la formación de estructura jerárquica, permitiendo una mejor

estimación de los parámetros cosmológicos y mejorando nuestro entendimiento sobre las condiciones iniciales que gobernaban el crecimiento de estructuras a todas escalas.

Pero estos enormes avances, van de la mano de otros, como la capacidad de analizar volúmenes de datos cada vez mayores, para extraer su valor en tiempo real y revelar comportamientos o tendencias, según las propiedades exploradas, o bien, los enormes avances tanto en métodos de computo y precisión numérica que se requieren para simular, por ejemplo, la formación de estructura.

Estas enormes bases de datos nos proveen un marco de referencia para estudiar las propiedades de las galaxias y sus componentes estructurales. Los métodos de análisis computacional nos proveen de herramientas para llevar a cabo comparaciones entre observaciones y teoría, no sólo a nivel individual, sino también ahora de manera estadística.

En esta tesis se logró desarrollar una herramienta computacional, denominada **PICASSO**, utilizando lo que se conoce en ciencias computacionales como **reutilización de código**. [Boehm, 1981], [Ishihara et al., 2013]. La **reutilización de código**, se refiere al comportamiento y a las técnicas, que garantizan que una parte o la totalidad de un programa informático existente, se pueda emplear en la construcción de otro programa. De esta forma se aprovecha el trabajo anterior, se economiza tiempo, y se reduce la redundancia.

Así pues, utilizando las herramientas computacionales ya existentes, que permiten trabajar hoy en día con bases de datos de tamaños moderados, y a partir de generar diferentes programas en distintos lenguajes como **MySQL**, principalmente, **C++** y **Python**, se le otorga una jerarquía a los datos contenidos en tablas, que forman las bases de datos, lo que nos permite poder trabajar, manipular y hacer búsquedas de distintos objetos astronómicos, en diferentes tablas, con distinto tipo de información, de manera mucho más eficiente que cuando los datos contenidos en las tablas no tienen ningún tipo de jerarquía, además, con dicha jerarquía, es muy fácil que los programas hechos para **PICASSO**, puedan escalar el tiempo de búsqueda en bases de datos cada vez más grandes.

Cuando una base de datos no tiene una estructura jerárquica, el tiempo de búsqueda es proporcional al número de elementos que tiene dicha base de

datos, ésto si sólo se efectuara una única búsqueda, sin embargo, el tiempo de búsqueda, si se realizan más, puede llegar a ser del orden de $O(n^2)$ donde n es el tamaño de la entrada, es decir, el número de objetos que tiene dicha tabla. Si ésta crece, el tiempo de búsqueda también lo hará proporcional al número de entradas que la base tenga.

Por otro lado, al volver los datos de la tabla jerárquica – del tipo de **árbol binario**– por medio de índices, se obtiene una relación logarítmica, entre el número de elementos y el tiempo de búsqueda, es decir: tiene un tiempo en cota superior asintótica[1], con un orden tal que $O(log\ n)$.

Si aumenta el numńero de elementos, pero se tiene una cantidad de niveles adecuados (en los árboles binarios), el tiempo de búsqueda se reduce logarítmicamente, por lo que además, puede ser extendida a diferentes bases de datos que contengan millones de datos.

Utilizando catálogos de descomposición estructural fotométrica (bulbo + total) y de masas estelares basados en las observaciones del SDSS, en esta tesis se ha desarrollado una serie de herramientas de análisis de grandes bases de datos, que se han explotado para explorar la relación entre la masa estelar de las galaxias y sus componentes estructurales (discos y esferoides), mostrando a **PICASSO** como una herramienta poderosa para entender los procesos involucrados en la formación y evolución de las galaxias.

El **estudio observacional de bulbos galácticos**, en este momento es aún incipiente, ya que con las observaciones actuales y utilizando sólo la distribución de brillo superficial a lo largo de la imagen para muchas galaxias, en estos momentos, sus componentes estructurales no se pueden resolver, ya que el tamaño angular con el cual se está observando resulta estar al límite de la resolución, además se sabe que los bulbos son multicomponentes [Perez et al., 2013], [Gargiulo et al., 2012], [Zavala et al., 2012b], así que se necesita tener una resolución espacial adecuada para hacer dichos estudios, se necesita caracterizar el medio ambiente, así como hacer estudios a distintos redshifts. Un trabajo reciente es el de Tasca et al. [2014] en el cual, se estudia la evolución de la fracción de contribución relativa entre el bulbo galáctico y el disco en la banda B, para diferentes épocas, siendo éste un trabajo pionero en el que se realiza una descomposición bulbo a disco de imágenes del Hubble Space

[1]En análisis de algoritmos, se llama cota superior asintótica a una función que sirve como cota superior de otra función cuando el argumento tiende a infinito.

Telescope HST[2] para $3,266$ galaxias con corrimientos al rojo espectroscópico en el rango $0{,}7 \leq z \leq 0{,}9$, encontrando que la fracción de luz en la banda B de los discos decrece aproximadamente un 30 % desde el redshift de 0.9 a 0.

En este trabajo de tesis se presenta un estudio que contiene una gran cantidad de objetos (galaxias) comparados con los que hay hasta el momento en la literatura, con una descomposición bulbo-total, dicha muestra consta de unos cuantos miles de objetos. Se crean diferentes submuestras utilizando distintos criterios de selección, para que se pueda garantizar completez estadística con respecto a la función de luminosidad, modificando el volumen y los cortes en la magnitud límite de la muestra [Hwang & Park, 2009].

Se hace una comparación cuantitativa de las distintas submuestras con las pocas muestras publicadas en la literatura Fisher & Drory [2011], Gadotti [2008] y Mendel et al. [2014].

Como resultado general de este trabajo de tesis se obtiene que: el valor promedio de Bulbo al total ($\mathbf{B/T}$), que representa el cociente entre la luminosidad de la componente del bulbo y la luminosidad total de la galaxia, en galaxias de masa estelar $(M_*) \leq 10^{10} M_{\odot}$ es considerablemente mayor que lo concluido en estudios anteriores y con una menor significancia estadística. Aunque, el volumen ocupado por estas muestras es diferente, por lo que el valor promedio también difiere, pero menos con respecto al valor promedio a galaxias con masas comparables de $10^{11} M_{\odot}$, que es lo concluido por estudios anteriores Fisher & Drory [2011] y Gadotti [2008] y Berg et al. [2014].

Es importante mencionar que hay posiblemente dos errores sistemáticos que en trabajos posteriores debemos estudiar a detalle. El primero es un efecto sistemático introducido al utilizar la base de datos de VESPA [Tojeiro et al., 2009] la cual no proporciona un mapeo uniforme entre magnitudes absolutas y masas estelares, lo que posiblemente esté afectando la completez de nuestra muestra y se deja para un estudio posterior utilizar métodos basados en los colores de las galaxias [Bell & de Jong, 2001] que conectan estos colores con la masa estelar, aunque son menos precisos, la corrección es más uniforme, lo cual nos permitirá estudiar los errores sitemáticos que introdujimos al utiizar VESPA [Tojeiro et al., 2009].

El uso de una herramienta para manejar y correlacionar diferentes bases de datos nos permitió crear lo que hasta ahora es la muestra más grande de

[2]http://www.stsci.edu/hst/

galaxias en el Universo Local con descomposición Bulbo al Total (**B/T**) contra masa estelar (definida con la SED). De manera preliminar, se estudiaron otras correlaciones: con el medio ambiente, actividad nuclear, masa de gas, etc.

Se considera que nuestro estudio permite ilustrar el poder que tendrán este tipo de herramientas computacionales en la era de los grandes sondeos de las galaxias, como por ejemplo el telescopio sinóptico (LSST), GAIA, PAN-STARRS, etc., que en conjunto proveerán mucho mas de 60 TB de información por noche.

6.1. Trabajo a Futuro

Desarrollar una versión lo más amigable posible de la herramienta de búsqueda aplicada a grandes bases de datos, implementada en esta tesis, para su uso colectivo.

Estimación de la masa estelar para toda la muestra con información estructural (bulbo + disco) a partir de los colores observados utilizando métodos de síntesis de poblaciones [Bell & de Jong, 2001]

Reconstrucción de los diagramas M_* vs (**B/T**) con esta nueva estimación de la masa.

Hacer un estudio de la completez estadística de la(s) muestra(s) generada(s) con la herramienta en esta tesis a fin de poder generalizar los resultados obtenidos.

Considerar los efectos de la presencia de las barras y de fuentes puntuales (AGNs y cúmulos nucleares) en la descomposición bulbo + disco y discutir sus consecuencias para el presente análisis.

Explorar nuevos esquemas con arquitecturas distribuidas que sean más fáciles de escalar.

Apéndices

A
Apéndice 1

A.1. Sobre Unidades de medición en cómputo

Una computadora digital moderna se puede definir en gran medida como un conjunto de interruptores electrónicos, los cuales se utilizan para representar y controlar el recorrido de los datos denominados bits (dígitos bits) [Ifrah, 2001].

Bit: Es el elemento más pequeño de información que puede manipular una computadora, su nombre es un acrónimo de *Binary Digit* (o dígito binario). Adquiere un único dígito en el sistema numérico binario, es decir 0 ó 1. Están representados físicamente por un elemento como un único pulso enviado a través de un circuito, o bien como un pequeo punto en un disco magnético capaz de almacenar un 0 o un 1. La representación de información se logra mediante la agrupación de bits pues así se logran formar unidades más grandes de datos que permiten manejar mayor información en los sistemas de las computadoras [Shannon, 1948].

Byte: Se describe como la unidad básica de almacenamiento de información. Agrupa al menos ocho bits, pero, debe quedar claro que su tamaño depende del código de información en el que está definido. Es, el equivalente a un

único carácter, como puede ser una letra, un número o un signo de puntuación, es decir, es lo que se utiliza para poder representar todo tipo de información. [Kozierok, 2005].

El byte representa sólo una pequeña cantidad de información, la cantidad de memoria y de almacenamiento de una máquina suele indicarse en kilobytes (1,024 bytes), en Megabytes (1,048,576 bytes) o en Gigabytes (1,024 Megabytes).

Kilobyte: Es una unidad que equivale a 1024 bytes. Es una unidad común para la capacidad de memoria o almacenamiento de las microcomputadoras .

Megabyte : Es una unidad de medida de cantidad de datos informáticos. Es un múltiplo binario del byte que equivale a un millón de bytes, es decri 1 048576 bytes.

Gigabyte : Es un múltiplo del byte de símbolo gb que se describe como la unidad de medida más utilizada en los discos duros. El cual también es una unidad de almacenamiento. Un gigabyte es con exactitud (1,073,742,824 bytes o mil 1024 megabytes)

Terabyte: Es la unidad de medida de la capacidad de memoria y de dispositivos de almacenamiento informático. Su símbolo es TB y coincide con algo más de un trillón de bytes.

A continuación se presenta el Cuadro A.1 con los términos que son más utilizados en el cómputo para describir el espacio que tiene un disco, o los datos de espacio de almacenamiento, así como la memoria del sistema todos con sus equivalentes específicas en bytes.

A.2. De bytes a Yottabytes

A continuación se muestra una lista comparativa, que permite asociar elementos conocidos para el lector, con sus unidades equivalentes utilizadas en computación, como Kilobytes, Terabytes, etc. La mayoría de los siguientes da-

Cantidad	Equivalencia
8 Bit	1 Byte
1024 Byte	1 Kbyte
1024 KByte	1 MegaByte
1024 MByte	1 GigaByte
1024 GByte	1TeraByte
1024Terabytes	1 Petabyte

Cuadro A.1: Equivalencias para las unidades de medición más comunes en el cómputo

tos fueron tomados de Williams [2006] [1]

- **Byte** equivale a 8 bits.
- 0.1 bytes: Una decisión binaria.
- 1 byte: un solo carácter.
- 10 bytes: una sola palabra.
- **Kilobyte** equivale a 1000 bytes o 10^3 bytes.
- 1 Kilobyte: Una historia muy corta.
- 10 Kilobytes: Una página enciclopédica.
- 15 Kilobytes: Una página web estática.
- 50 Kilobytes: Una página que contenga una imagen comprimida.
- **Megabyte** equivale a 1,000,000 bytes o 10^6 bytes.
- 1 Megabyte: Una pequeña novela.
- 5 Megabytes: Las obras completas de William Shakespeare ó 10 segundos de vídeo con calidad de televisión.
- 10 Megabytes: Medio minuto de sonido de alta fidelidad o una radiografía digital de rayos X.

[1] http://www2.sims.berkeley.edu/research/projects/how-much-info/summary.html

- 50 Megabytes: Una mamografía digital.
- 100 Megabytes: Un libro enciclopédico de dos volúmenes.
- **Gigabyte** equivale a 1,000,000,000 bytes ó 10^9 bytes.
- 1 Gigabyte: una camioneta llena de papel, sólo el sonido de una sinfonía de alta fidelidad o una película con calidad de televisión.
- 20 Gigabytes: Una colección – bastante entera – de las obras de Beethoven.
- 500 Gigabytes: El mayor sitio FTP.
- **Terabyte** equivale a 1,000,000,000,000 bytes ó 10^{12} bytes.
- 1 Terabyte: todas las películas de rayos X en un hospital tecnológico o 50 000 árboles convertidos en papel.
- 2 Terabytes: Una biblioteca de investigación académica.
- 10 Terabytes: La colección impresa de la Biblioteca del Congreso de EE.UU.
- 400 Terabytes: La base de datos de The National Climactic Data Center (NOAA).
- **Petabyte** equilvale a 1,000,000,000,000,000 bytes ó 10^{15} bytes.
- 1 Petabyte: 3 años de datos de EOS (hasta el 2001) donde EOS es el Earh Observing System de la NASA.
- 2 Petabytes: Todas las bibliotecas de investigación académica de Estados Unidos de Norte América.
- 20 Petabytes: Toda la información disponible en la Web.
- 200 Petabytes: Todo el material impreso.
- **Exabyte** equivale a 1,000,000,000,000,000,000 bytes ó 10^{18} bytes.
- 2 Exabytes: Volumen total de la información generada en todo el mundo anualmente.

- 5 Exabytes: Todas las palabras alguna vez pronunciadas por los humanos.
- 161 exabytes de datos fueron creadas en 2006, esto es 3 millones de veces la cantidad de información contenida en todos los libros que se habían escrito.
- 800 exabytes El volumen global de datos alcanzado a finales de 2009.
- **Zettabyte** equivale a 1,000,000,000,000,000,000,000 bytes ó 10^{21} bytes.
- **Yottabyte** equivale a 1,000,000,000,000,000,000,000,000 bytes ó 10^{24} bytes.
- **Xenottabyte** equivale a 1,000,000,000,000,000,000,000,000,000 bytes ó 10^{27} bytes.
- **Shilentnobyte** equivalente a 1,000,000,000,000,000,000,000,000,000,000 bytes ó 10^{30} bytes.
- **Domegemegrottebyte** equivalente a 1,000,000,000,000,000,000,000,000,000,000,000 bytes ó 10^{33} bytes.

A.3. Datos interesantes.

El año pasado excedimos los 1,2 Zettabytes de información generada, y se tiene un pronóstico de crecimiento de $44x$ en la presente década [Mayer-Schnberger & Cukier, 2013].

Facebook contiene aproximadamente $10,000$ millones de fotos, las cuales requieren un Petabyte de almacenamiento.

El gran colisionador de partculas se prevé que producirá 15 Petabytes de datos cada año

Google procesa 20 Petabytes de datos diariamente [Dean & Ghemawat, 2008].

B
Apéndice 2

B.1. Términos técnicos utilizados a lo largo de la tesis y sus abreviaturas

En este apéndice se presentan la definiciones y las abreviaturas – en caso de que tenga – de los términos técnicos utilizados a lo largo de la tesis.

- **DB** Abreviatura de *Data Base* en espaol Base de Datos. Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. Es una colección estructurada de datos.

- **DBMS** *Data Base Management System*, en espaol Sistema gestor de bases de datos. Es el software encargado de administrar, producir y gestionar bases de datos.

- **SQL** Abreviatura de Structured Query Language, en espaol lenguaje de interrogaciones estructuradas. Es el lenguaje estandarizado para pedir información de una base de datos. Fue diseñada en un centro de investigación de IBM en 1974.

C
Apendice 3

En este apéndice se presenta la información de los catálogos observaciones provenientes del SDSS, DR7, que se utilizan en esta tesis. En particular, se describen las carcterístcas que tienen. El primero es el utilizado por Simard et al. [2011], descomposición bulbo + disco con y parámetros estructurales con n_b libre.

ObjID	SDSS Object ID
z	SDSS Redshift (Spectroscopic if available. Photometric otherwise)
SpecClass	SDSS SpecClass value (set to -1 if z is photometric or -2 if no redshift available at all)
Scale	Physical scale in arcsec/kpc at redshift z
V_{max}	Galaxy volume correction in Mpc^3
g_{g2d}	g-band apparent magnitude of GIM2D output B+D model
r_{g2d}	r-band apparent magnitude of GIM2D output B+D model
$g_{g2d,f}$	g-band apparent fiber magnitude of output B+D model
$r_{g2d,f}$	r-band apparent fiber magnitude of output B+D model
Δ(fiber color)	Delta fiber color defined as $(g-r)_{gim2d,fiber} - (g-r)_{SDSS,fiber}$ (set to -99.99 if no SDSS fiber magnitudes available)
$(B/T)_g$	g-band bulge fraction
$(B/T)_r$	r-band bulge fraction
$(B/T)_{g,f}$	g-band fiber bulge fraction
$(B/T)_{r,f}$	r-band fiber bulge fraction
$R_{hl,g}$	g-band galaxy semi-major axis, half-light radius in kiloparsecs
$R_{hl,r}$	r-band galaxy semi-major axis, half-light radius in kiloparsecs
$R_{chl,g}$	g-band galaxy circular half-light radius in kiloparsecs
$R_{chl,r}$	r-band galaxy circular half-light radius in kiloparsecs
R_e	Bulge semi-major effective radius in kiloparcsecs
e	Bulge ellipticity ($e \equiv 1 - b/a$, e = 0 for a circular bulge)
ϕ_b	Bulge position angle in degrees (measured clockwise from the $+y$ axis of SDSS images)
R_d	Exponential disk scale length in kiloparsecs
i	Disk inclination angle in degrees ($i \equiv 0$ for a face-on disk)
ϕ_d	Disk position angle in degrees (measured clockwise from the $+y$ axis of SDSS images)
$(dx)_g$	B+D model center offset from column position given by `colc_g` on SDSS corrected g-band image (arcsec)
$(dy)_g$	B+D model center offset from row position given by `rowc_g` on SDSS corrected g-band image (arcsec)
$(dx)_r$	B+D model center offset from column position given by `colc_r` on SDSS corrected r-band image (arcsec)
$(dy)_r$	B+D model center offset from row position given by `rowc_r` on SDSS corrected r-band image (arcsec)
$S2_g$	g-band image smoothness parameter
$S2_r$	r-band image smoothness parameter
$M_{g,g}$	g-band GIM2D galaxy rest-frame, absolute magnitude
$M_{g,b}$	g-band GIM2D bulge rest-frame, absolute magnitude
$M_{g,d}$	g-band GIM2D disk rest-frame, absolute magnitude
$M_{r,g}$	r-band GIM2D galaxy rest-frame, absolute magnitude
$M_{r,b}$	r-band GIM2D bulge rest-frame, absolute magnitude
$M_{r,d}$	r-band GIM2D disk rest-frame, absolute magnitude
n_b	Bulge Sérsic index
P_{pS}	F-test probability that a B+D model is *not* required compared to a pure Sérsic model
P_{n4}	F-test probability that a free n_b B+D model is *not* required compared to a fixed n_b=4 B+D model

Bibliografía

Abazajian K. N. et al., 2009, ApJS, 182, 543

Aguerri J. A. L., Méndez-Abreu J., Corsini E. M., 2009, Astronomy and Astrophysics, 495, 491

Ahn C. P. et al., 2012, ApJS, 203, 21

Allen P. D., Driver S. P., Graham A. W., Cameron E., Liske J., de Propris R., 2006, MNRAS, 371, 2

Andredakis Y. C., Peletier R. F., Balcells M., 1995, MNRAS, 275, 874

Athanassoula E., 2006, ArXiv Astrophysics e-prints

Axelrod T. S., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astronomical Society of the Pacific Conference Series Vol. 351, Astronomical Data Analysis Software and Systems XV. p. 103

Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681

Balogh M. L., Baldry I. K., Nichol R., Miller C., Bower R., Glazebrook K., 2004, ApJL, 615, L101

Barazza F. D., Jogee S., Marinova I., 2007, in Combes F., Palouš J., eds, IAU Symposium Vol. 235, IAU Symposium. pp 76–76

Beaulieu A., 2005, Learning SQL. O'Reilly Media

Bell E. F., de Jong R. S., 2001, ApJ, 550, 212

Bell G., Gray J., Szalay A., 2007, eprint arXiv:cs/0701165

Berg T. A. M., Simard L., Mendel Trevor J., Ellison S. L., 2014, MNRAS, 440, L66

Bernardi M. et al., 2003, Astronomical Journal, 125, 1882

Blanton M. R. et al., 2003, ApJ, 592, 819

Boehm B. W., 1981, Software Engineering Economics, 1st edn. Prentice Hall PTR, Upper Saddle River, NJ, USA

Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150

Bureau M., Athanassoula E., 2005, ApJ, 626, 159

Cameron E., Driver S. P., Graham A. W., Liske J., 2009, ApJ, 699, 105

Celko J., 2004, Joe Celko's Trees and Hierarchies in SQL for Smarties. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science

Celko J., 2010, Joe Celko's SQL for Smarties: Advanced SQL Programming. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science

Ceri S., Pelagatti G., 1984, IEEE Transactions on Computers, 31, 119

Chen Z., Gehrke J., Korn F., 2001, in In ACM SIGMOD. ACM Press, pp 271–282

Codd E. F., 1979, ACM Transactions on Database Systems, 4, 397

Cole S. et al., 2005, Monthly notices of the Royal Astronomical Society., 362, 505

Colless M. et al., 2001, MNRAS, 328, 1039

Combes F., Sanders R. H., 1981, Astronomy and Astrophysics, 96, 164

Conselice C. J., 2006, MNRAS, 373, 1389

Cormen T. H., Stein C., Rivest R. L., Leiserson C. E., 2001, Introduction to Algorithms, 2nd edn. McGraw-Hill Higher Education

Date C. J., 2004, An Introduction to Database Systems, 8 edn. Pearson Addison-Wesley, Boston, MA

Date. C. J., 2006, An Introduction to Database Systems. Pearson Education

Davis A. M., Bersoff E. H., 1991, Commun. ACM, 34, 104

de Souza R. E., Gadotti D. A., dos Anjos S., 2004, ApJS, 153, 411

de Vaucouleurs G., 1948, Annales d'Astrophysique, 11, 247

de Vaucouleurs G., 1959, Handbuch der Physik, 53, 275

Dean J., Ghemawat S., 2008, Commun. ACM, 51, 107

Driver S. P. et al., 2006, MNRAS, 368, 414

DuBois P., 2003, MySQL: the definitive guide to using, programming, and administering MySQL4. Developer's library, Sams Publishing

Dye S. et al., 2006, MNRAS, 372, 1227

Fisher D. B., Drory N., 2008, in Funes J. G., Corsini E. M., eds, Astronomical Society of the Pacific Conference Series Vol. 396, Formation and Evolution of Galaxy Disks. p. 309

Fisher D. B., Drory N., 2011, The Astrophysical Journal Letters, 733, L47

Frakes W., Baeza-Yates R. A., 1992, Information Retrieval: Data Structures & Algorithms. Prentice-Hall

Freeman K. C., 1970, ApJ, 160, 811

Gadotti D. A., 2008, MNRAS, 384, 420

Gadotti D. A., Kauffmann G., 2009, MNRAS, 399, 621

Gargiulo I. D., Pérez M. J., Cora S. A., Valenzuela O., Avila Reese V., Padilla N. D., Tecce T. E., Ruiz A. N., 2012, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 55, 293

Golden A., Djorgovski S. G., Greally J. M., 2013, ArXiv e-prints

Graham A. W., 2001, Astronomical Journal, 122, 1067

Gray J., Szalay A. S., Budavari T., Lupton R., Nieto-Santisteban M. A., Thakar A., 2007, CoRR, abs/cs/0701172

Haefliger S., von Krogh G., Spaeth S., 2008, Manage. Sci., 54, 180

Han S., 2011, PhD thesis, Cambridge, MA, USA

Hernquist L., Mihos J. C., 1995, ApJ, 448, 41

Hogg D. W. et al., 2004, ApJL, 601, L29

Holmberg E., 1958, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136, 1

Hopkins P. F., 2009, in Jogee S., Marinova I., Hao L., Blanc G. A., eds, Astronomical Society of the Pacific Conference Series Vol. 419, Galaxy Evolution: Emerging Insights and Future Challenges. p. 228

Hubble E. P., 1926, ApJ, 64, 321

Hubble E. P., 1936, Realm of the Nebulae

Hwang H. S., Park C., 2009, The Astrophysical Journal, 700, 791

Ifrah G., 2001, The Universal History of Computing: From the Abacus to the Quantum Computer, 1st edn. John Wiley & Sons, Inc., New York, NY, USA

Ishihara T., Hotta K., Higo Y., Kusumoto S., 2013, in Lmmel R., Oliveto R., Robbes R., eds, WCRE. IEEE, pp 457–461

Ivezic Z. et al., 2010, in American Astronomical Society Meeting Abstracts 215. pp 401–403

Jedrzejewski R. I., 1987, MNRAS, 226, 747

Jogee S., 1999, PhD thesis, California Institute of Technology

Jones D. H. et al., 2004, MNRAS, 355, 747

Jordan S., 2008, Astronomische Nachrichten, 329, 875

Kaiser N. et al., 2002, in Tyson J. A., Wolff S., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4836, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. pp 154–164

Kauffmann G. et al., 2003, MNRAS, 341, 54

Kennicutt Jr. R. C., Lee J. C., Funes José G. S. J., Sakai S., Akiyama S., 2008, ApJS, 178, 247

Kent B. R. et al., 2005, in American Astronomical Society Meeting Abstracts. p. 179

Kent S. M., 1985, ApJS, 59, 115

King R., McLeod D., 1985, in , Principles of Database Design (I). pp 115–150

Kirscher J., Grießhammer H. W., Shukla D., Hofmann H. M., 2009, ArXiv e-prints

Kormendy J., 1993, in Dejonghe H., Habing H. J., eds, IAU Symposium Vol. 153, Galactic Bulges. p. 209

Kormendy J., Fisher D. B., 2005, in Torres-Peimbert S., MacAlpine G., eds, Revista Mexicana de Astronomia y Astrofisica Conference Series Vol. 23, Revista Mexicana de Astronomia y Astrofisica Conference Series. pp 101–108

Kormendy J., Kennicutt Jr. R. C., 2004, Ann. Rev. Ast. & Ast., 42, 603

Kozierok C., 2005, The TCP/IP Guide: A Comprehensive, Illustrated Internet Protocols Reference. No Starch Press, San Francisco, CA, USA

Kwon Y., Nunley D., Gardner J. P., Balazinska M., Howe B., Loebman S., 2010, in Gertz M., Ludscher B., eds, Lecture Notes in Computer Science Vol. 6187, SSDBM. Springer, pp 132–150

Laurikainen E., Salo H., Buta R., Knapen J. H., Speltincx T., Block D. L., 2007, in Combes F., Palouš J., eds, IAU Symposium Vol. 235, IAU Symposium. pp 36–38

Laurikainen E., Salo H., Buta R., Vasylyev S., 2005, VizieR Online Data Catalog, 735, 51251

Lemson G., Virgo Consortium t., 2006, ArXiv Astrophysics e-prints

Lotz J. M., Primack J., Madau P., 2004, Astronomical Journal, 128, 163

McClure C., 2001, Software Reuse: A Standards-Based Guide. Software Engineering Standards Series, Wiley

Marinova I. et al., 2007, in American Astronomical Society Meeting Abstracts. p. 905

Martinez-Valpuesta I., Shlosman I., Heller C., 2006, ApJ, 637, 214

Mayer-Schnberger V., Cukier K., , 2013, Big data a revolution that will transform how we live, work, and think

McPherson A. M. et al., 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.

Mendel J. T., Simard L., Palmer M., Ellison S. L., Patton D. R., 2014, VizieR Online Data Catalog, 221, 3

Méndez-Abreu J., Aguerri J. A. L., Corsini E. M., 2008, in Funes J. G., Corsini E. M., eds, Astronomical Society of the Pacific Conference Series Vol. 396, Formation and Evolution of Galaxy Disks. p. 369

Menéndez-Delmestre K., Sheth K., Schinnerer E., Jarrett T. H., Scoville N. Z., 2007, ApJ, 657, 790

Moore G. E., 1975, 21, 11

Morgan W. W., 1958, Publications of the ASP, 70, 364

Moster B. P., Naab T., White S. D. M., 2013, MNRAS, 428, 3121

Müller H. A., Tilley S. R., Wong K., 1993, in Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Software Engineering - Volume 1. CASCON '93. IBM Press, pp 217–226

Peng C., 2010, in American Astronomical Society Meeting Abstracts #215. pp 229–232

Peng E. W., Ford H. C., Freeman K. C., White R. L., 2002, Astronomical Journal, 124, 3144

Perez J., Valenzuela O., Tissera P. B., Michel-Dansac L., 2013, MNRAS, 436, 259

Pignatelli E., Fasano G., Cassata P., 2006, Astronomy and Astrophysics, 446, 373

Poveda A., 1958, Boletin de los Observatorios Tonantzintla y Tacubaya, 2, 3

Poveda A., 1961, ApJ, 134, 910

Quinn P. J., Hernquist L., Fullagar D. P., 1993, ApJ, 403, 74

Reese A. S., Williams T. B., Sellwood J. A., Barnes E. I., Powell B. A., 2007, Astronomical Journal, 133, 2846

Reifer D. J., 1997, Practical Software Reuse, 1st edn. John Wiley & Sons, Inc., New York, NY, USA

Riebe K. et al., 2011, ArXiv e-prints

Roberts M. S., Haynes M., 1994, in Meylan G., Prugniel P., eds, European Southern Observatory Conference and Workshop Proceedings Vol. 49, European Southern Observatory Conference and Workshop Proceedings. p. 197

Samuel M. L., Pedersen A. U., , 2004, MySQL in a Main Memory Database Context

Schechter P. L., Dressler A., 1987, Astronomical Journal, 94, 563

Sérsic J. L., 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 99

Shannon C. E., 1948, Bell system technical journal, 27

Shioji E., Kawakoya Y., Iwamura M., Hariu T., 2012, in Proceedings of the 28th Annual Computer Security Applications Conference. ACSAC '12. ACM, New York, NY, USA, pp 309–318

Silberschatz A., Korth H., Sudarshan S., Pérez F., 2008, Fundamentos de bases de datos. McGraw-Hill

Simard L., Trevor Mendel J., Patton D. R., Ellison S. L., McConnachie A. W., 2011, VizieR Online Data Catalog, 219, 60011

Simard L. et al., 2002, ApJS, 142, 1

Springel V., Hernquist L., 2005, ApJL, 622, L9

Springel V. et al., 2008, MNRAS, 391, 1685

Strateva I. et al., 2001, Astronomical Journal, 122, 1861

Szalay A. S., Kunszt P. Z., Thakar A. R., Gray J., Slutz D., 2000, in Manset N., Veillet C., Crabtree D., eds, Astronomical Society of the Pacific Conference Series Vol. 216, Astronomical Data Analysis Software and Systems IX. p. 405

Tasca L. A. M. et al., 2014, ArXiv e-prints

Teorey T. J., Yang D., Fry J. P., 1986, ACM Computing Surveys, 18, 197

Tojeiro R., Heavens A. F., Jimenez R., Panter B., 2007, MNRAS, 381, 1252

Tojeiro R., Wilkins S., Heavens A. F., Panter B., Jimenez R., 2009, ApJS, 185, 1

Trujillo I., Aguerri J. A. L., 2004, MNRAS, 355, 82

Tyson J. A., 2002, in Tyson J. A., Wolff S., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4836, Survey and Other Telescope Technologies and Discoveries. pp 10–20

van den Bergh S., 1960a, ApJ, 131, 558

van den Bergh S., 1960b, ApJ, 131, 215

Vázquez-Mata J. A., Hernández-Toledo H. M., Park C., Choi Y.-Y., 2010, in Peterson B. M., Somerville R. S., Storchi-Bergmann T., eds, IAU Symposium Vol. 267, IAU Symposium. pp 464–464

Weinzirl T., Jogee S., Khochfar S., Burkert A., Kormendy J., 2009, ApJ, 696, 411

Williams R., , 2006, "Data Powers of Ten"

Yao S. B., ed. 1985, Principles of Database Design, Volume I: Logical Organizations. Prentice-Hall

York D. G. et al., 2000, Astronomical Journal, 120, 1579

Zavala J., Avila-Reese V., Firmani C., Boylan-Kolchin M., 2012a, MNRAS, 427, 1503

Zavala J., Avila-Reese V., Firmani C., Boylan-Kolchin M., 2012b, MNRAS, 427, 1503

Zhang J., Chen C., 2010, ArXiv e-prints