Aplicación de bases de datos orientadas a grafos en la Astrofísica

Cynthia Alejandra Martínez Pinto
Instituto Tecnológico de Ciudad Guzmán
cynthia_amp@hotmail.com

Rosa María Michel Nava
Instituto Tecnológico de Ciudad Guzmán
michel91_3@hotmail.com[2]

Resumen

Gracias a los avances de la tecnología, cada día se puede tener información rápida y oportuna sobre cualquier área del conocimiento humano. Las bases de datos relacionales tienen una amplia aceptación y han demostrado su efectividad. Sin embargo, actualmente la manipulación de grandes cantidades de datos ha resultado difícil y se han implementado otras estrategias para manejar la información.

Un caso específico es la representación de datos en forma de grafos, dado que las bases de datos relacionales tienen ciertos inconvenientes al momento de representar este tipo de estructura.

Este artículo está orientado a presentar una solución mediante el uso de bases de datos orientadas a grafos, que permitan mantener de forma eficiente, rápida y sencilla, relaciones múltiples entre distintos nodos, lo cual es ideal para la manipulación de grandes cantidades de información (*Big Data*).

Bajo este concepto, se trabajó en conjunto con el doctor Miguel Ángel Aragón Calvo, de la Universidad Johns Hopkins para desarrollar una base de datos capaz de almacenar y manipular los terabytes de información que se guarda en archivos de datos binarios en los servidores del departamento de Física y Astronomía de la Universidad.

La solución fue el desarrollo de una herramienta que genera una base de datos orientada a grafos la cual permite tanto la creación como la manipulación de datos astronómicos, siendo la primera que se utiliza para estos fines y representando un excelente auxiliar en las

investigaciones y simulaciones que realizan los investigadores y profesionales en el área de la astronomía y Astrofísica a nivel internacional.

Palabras clave:

Introducción

La tecnología ha avanzado mucho en las últimas décadas, principalmente con la invención de Internet del cual la ciencia ha estado sumamente beneficiada, ya que también ha tenido sus pasos agigantados proporcionando un gran beneficio para las personas de esta y otras épocas.

La creación de nuevas tecnologías, hablando del área de las Ciencias Computacionales, permite que otros campos de estudio las usen de manera específica, tal es el caso de las bases de datos y su diversidad de tipos. Las bases de datos orientadas a grafos tienen la capacidad de manipular grandes cantidades de información para procesarla y obtener resultados de acuerdo a estudios específicos realizados por investigadores de distintas áreas, esto se conoce como *Big Data* la cual se define como grandes conjuntos de datos. Es una tecnología emergente, motivo por el cual este proyecto comenzó su desarrollo.

En la Astronomía también se generan grandes cantidades de información en el momento que los astrofísicos realizan cierto tipo de simulaciones conocidas como *N-Body*, para estudiar más de cerca la creación e interacción de las galaxias en el universo real. Para guardar esta información las bases de datos relacionales o tradicionales, no se adaptan al tipo de estructura que forman los datos, resultando de baja eficiencia en su manipulación.

Se pensó en las bases de datos orientadas a grafos las cuales tienen la capacidad de guardar grandes cantidades de información manteniendo la naturaleza de su estructura y el cómo está relacionado un tipo de información con otro.

Desarrollo

Tipo de investigación

Se presenta una investigación aplicada con el objetivo de crear una nueva manera de almacenar datos astronómicos resultantes de simulaciones realizadas para representar la relación entre galaxias que forman el universo, conocido como Webcósmica.

La representación natural de una galaxia es a través de grafos, por lo tanto es difícil tratar de guardar estas estructuras en una base de datos relacional. Los trabajos realizados hasta la fecha solo han podido representar dicha información a través de árboles, pero aun así, resulta complicado plasmar estos árboles en una base de datos relacional. El presente proyecto elimina estas complicaciones al utilizar bases de datos orientadas a grafos.

Tipo de muestra

En este caso se utilizó un muestreo no probabilístico, debido a que las pruebas realizadas se llevaron a cabo con el doctor Miguel Ángel Aragón Calvo, Astrofísico Investigador de la Johns Hopkins University, miembro de la comunidad de astrónomos y Astrofísicos con quien se trabajó directamente en el desarrollo de la investigación.

Muestra

Como muestra se considera un grupo de Astrofísica ubicado en el edificio Bloomberg, Departamento de Física y Astronomía de la Johns Hopkins University, conformado por un grupo de 20 Investigadores e Ingenieros.

Instrumentos

En el desarrollo de la investigación se utilizó Java en su versión 1.6 y superior, Neo4j 1.9 SNAPSHOT y Eclipse IDE como entorno de desarrollo.

Aparatos

La Johns Hopkins University proporcionó un espacio de oficina así como una computadora básica compuesta por 2 Gigabytes de memoria RAM y Sistema Operativo Linux, desde la cual se hacía conexión mediante el protocolo SSH a una computadora denominada “gwln1” que consiste de

130 Gigabytes de memoria RAM y 16 procesadores, esta fue la principal computadora en la realización de las pruebas finales.

Procedimientos

El proyecto desarrollado tiene como objetivo la implementación de una Base de Datos Orientadas a Grafos que permita guardar la información proveniente de simulaciones hecha por Investigadores y Astrofísicos de la Universidad Johns Hopkins.

Para esto se utilizó el modelo incremental que consta de las fases de análisis, diseño, codificación y pruebas, las cuales se mencionan a continuación con cada una de las actividades realizadas en cada fase.

Análisis

Dentro de la fase de análisis se realizaron actividades de búsqueda de un Sistema de Base de Datos Orientada a Grafos, el análisis de los requerimientos establecidos para el funcionamiento de la herramienta, el estudio de la información disponible para conocer la manera en que se encuentra estructurada, así como algunas pruebas preliminares que ayudaron a comprender cómo funciona la base de datos elegida.

Para generar las bases de datos se tomó la decisión de implementar Neo4j, ya que presenta características de alto rendimiento, funciona en las plataformas de sistemas operativos más utilizados y está hecho con Java, cuenta con una amplia *API* para poder manipular Neo4j desde código, no requiere de una extensa instalación lo cual facilita su portabilidad, la versión más actual de Neo4j cuenta con la capacidad de almacenar 34.4 billones de nodos, 34.4 billones de relaciones, y 68.7 billones de propiedades, gracias a estas capacidades puede soportar un grafo completo de 262,144 nodos y 34’359,607,296 relaciones.

Un factor importante fue el analizar los requerimientos de los Investigadores y Astrofísicos de la Universidad Johns Hopkins los cuales tenían la necesidad de implementar una Base de Datos Orientada a Grafos y una herramienta que les permitiese manipular la información, almacenarla y realizar consultas de acuerdo a criterios propios de los Investigadores y Astrofísicos, estos elementos debían tener las siguientes características:

- Base de Datos Orientada a Grafos.

- Lenguaje de programación libre.
- Alto rendimiento y eficiencia.

Diseño

El sistema desarrollado se denomina MIP, este viene del latín *Multum in parvo*, que significa "Mucho en poco" o "Muchas cosas en un lugar pequeño", así bautizado por el docgtor Miguel Ángel Aragón Calvo (para ver más información puede consultar la siguiente dirección: http://skysrv.pha.jhu.edu/~miguel/MIP/index.html#).

En la fase de diseño se elaboraron diferentes diagramas para plasmar la estructura del sistema. Uno de ellos es el que se muestra en la figura 1, que se refiere al diagrama de actividades, que ilustra la lógica que se sigue para crear una base de datos en MIP.

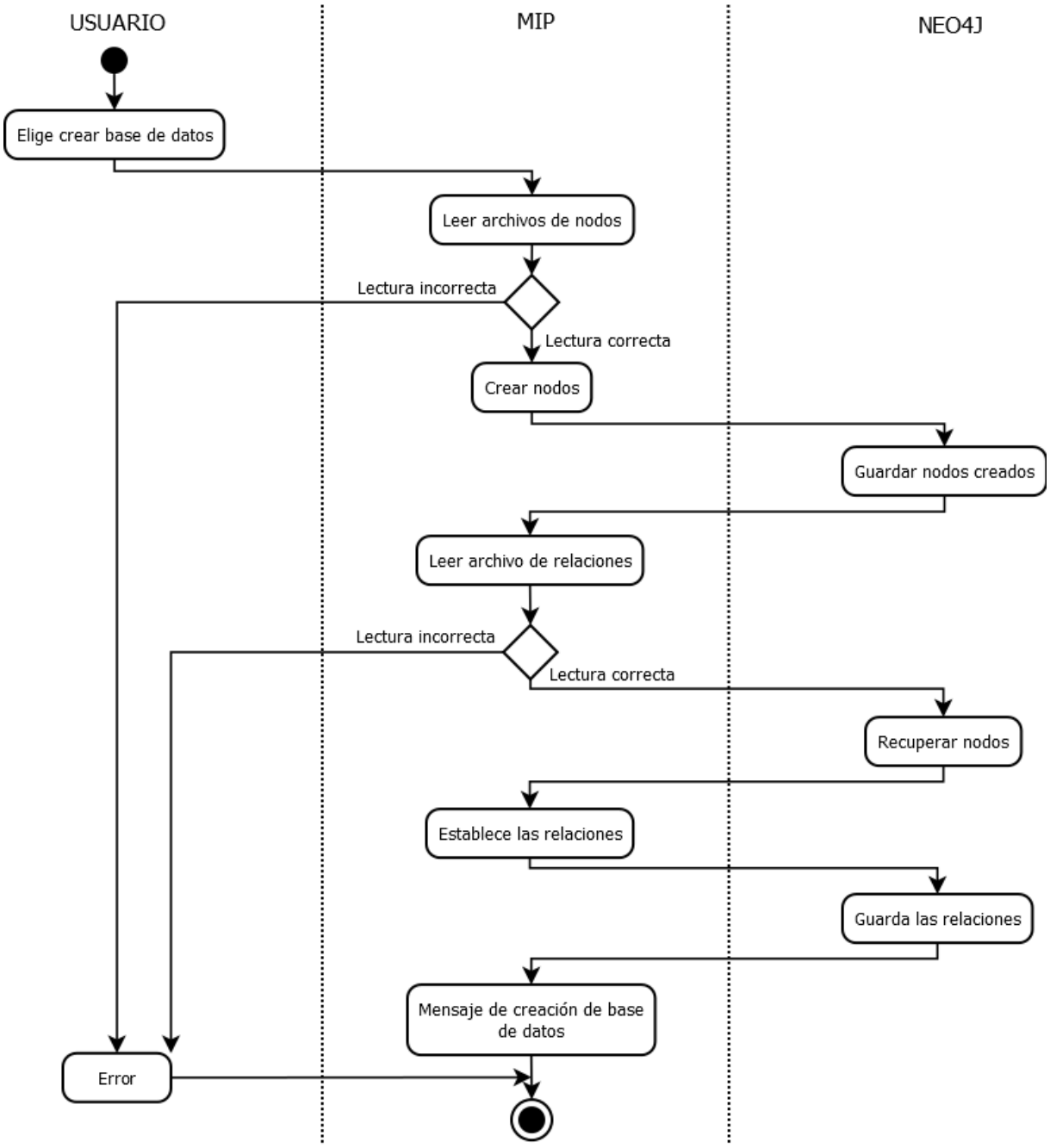

Figura 1. Diagrama de actividades para el caso de uso “Crear base de datos”

Codificación

Desarrollo de una herramienta para la creación de grafos y ejecución de consultas

Se desarrolló una herramienta que hace posible la creación de los grafos en la base de datos y la realización de consultas, esta herramienta es implementada en los equipos de la Universidad Johns Hopkins permitiendo a los astrofísicos continuar con pruebas basadas en sus simulaciones e investigaciones posteriores.

Consta de dos módulos por separado, el primero es para el almacenamiento de información en la base de datos, ya que Neo4j permite añadir nodos y relaciones en caso de que previamente ya existan, o en caso contrario, crear la base de datos y guardar la información, el segundo es para la ejecución de las consultas que el usuario desee introducir y una rutina establecida para obtener el progenitor más masivo de una galaxia, estudios de suma importancia para los astrofísicos.

Pruebas

Implementación de pruebas con datos de una simulación real

Para comprobar el funcionamiento de la herramienta creada, se realizaron pruebas con datos de una *simulación* real la cual cuenta con un total de 37,359 nodos y 36,687 relaciones, la información fue almacenada en archivos XML.

Después de varias pruebas y correcciones al código, se logró realizar la creación de la base de datos en un tiempo de 1 minuto con 45 segundos. Desde la lectura del archivo XML hasta la creación de la base de datos.

Una manera de comprobar que el proceso fue realizado correctamente fue activando el *Webadmin* de Neo4j, donde se muestran los nodos y sus relaciones (Figuras 2 y 3).

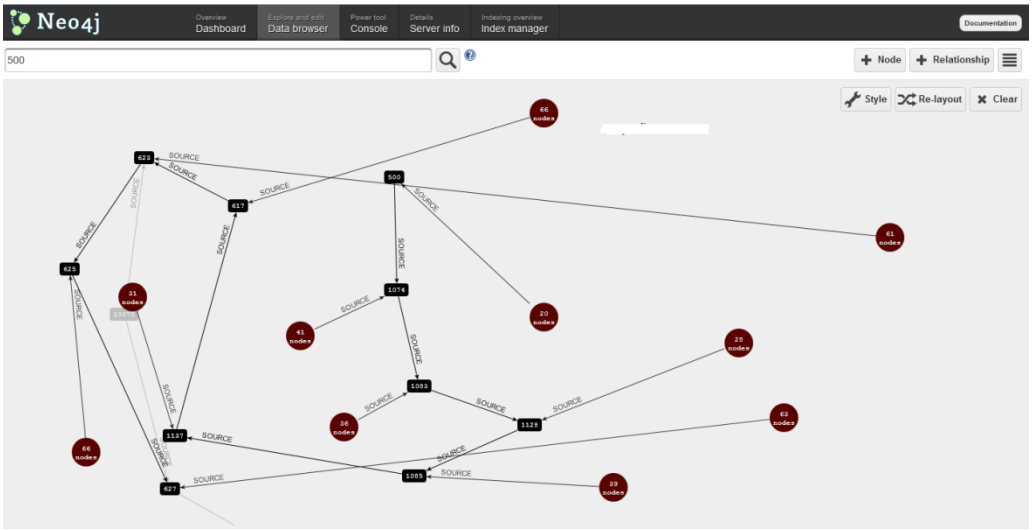

Figura 2.Grafo visto desde Webadmin tras realizar la prueba del sistema

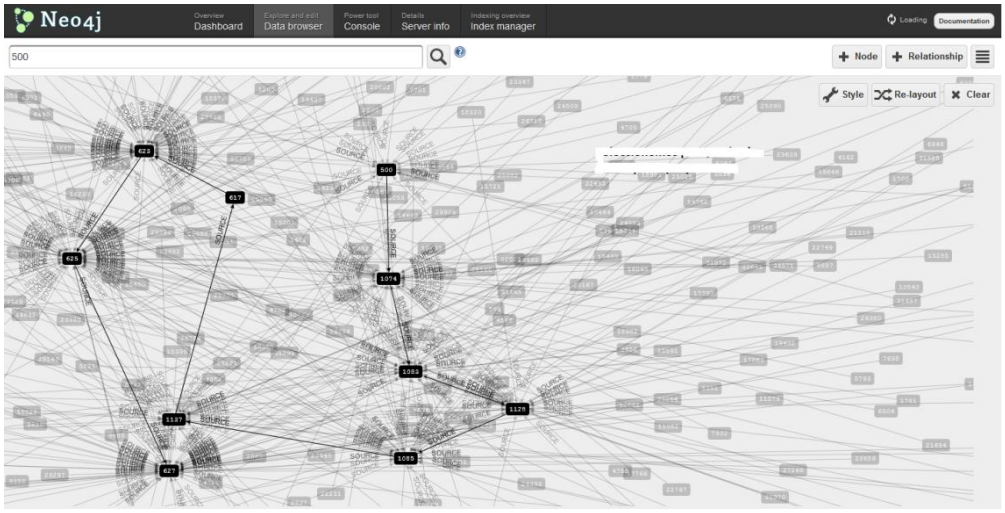

Figura 3. Grafo ampliado en la selección de nodos

Resultados

Una vez desarrollado el software y después de realizar las pruebas pertinentes, el licenciado en Informática Pablo Ramírez Solórzano (quien desarrolló este software), se trasladó a la ciudad de Baltimore Maryland, Estados Unidos (del 20 de Noviembre al 20 de Diciembre 2012), a implantar el sistema y realizar pruebas con la información que reside en los servidores de la Universidad Johns Hopkins en conjunto con el doctor Miguel Ángel Aragón Calvo.

Las pruebas se realizaron con la primera simulación que contiene poco más de un millón de galaxias con sus respectivas relaciones. El proceso de la creación de la base de datos se va mostrando en la misma terminal, donde el usuario puede darse cuenta si existiera algún error o cuando el proceso haya terminado correctamente, así lo muestra la figura 4.

Figura 4. Proceso de crear la base de datos

En la herramienta Webadmin de Neo4j se puede observar los elementos de la base de datos, como son el número de nodos, número de relaciones, propiedades y otros (ver la figura 5).

Figura 5. Información de la base de datos a través de Webadmin

Una vez creada la base de datos, se pueden realizar diferentes consultas, con el objetivo de obtener información con un determinado criterio, según sea requerido por el usuario. Este era un requisito muy importante para los doctores en Astrofísica, es decir, ellos necesitaban libertad para consultar la información, por lo tanto se desarrolló un módulo de consulta, que a través de escribir algunos comandos en Cypher (lenguaje para realizar consultas Neoj4) se ejecuten consultas diferentes, conforme la necesidad del usuario. En la figura 6 se muestra un ejemplo y los resultados serán visualizados a través de la consola.

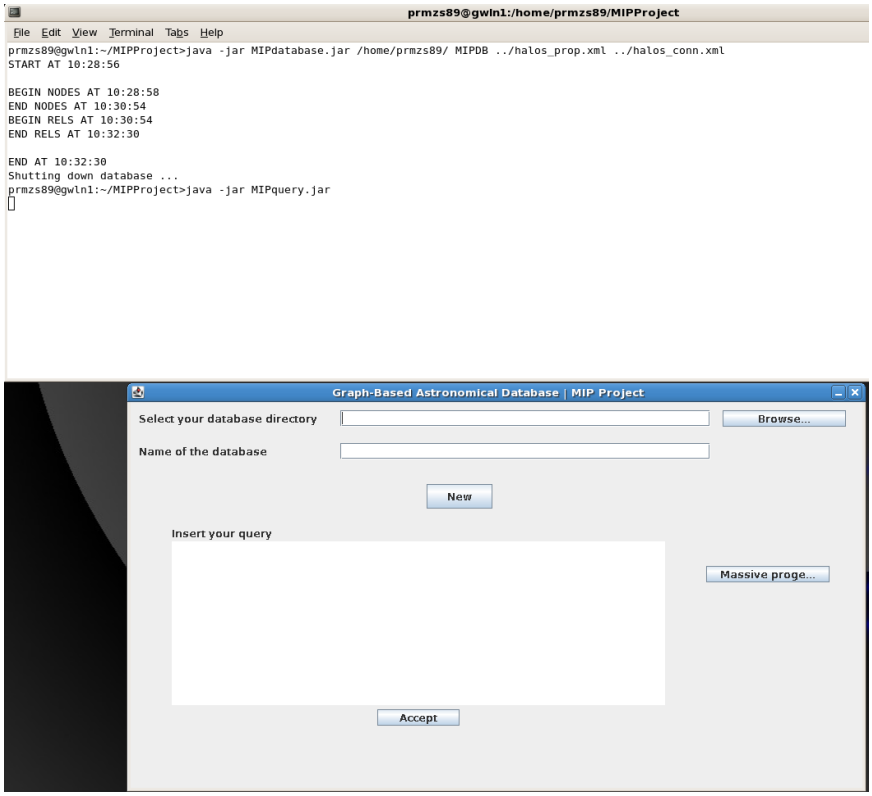

Figura 6. Módulo de consultas en ejecución

Es necesario seleccionar el directorio donde se encuentra la base de datos e introducir el nombre de la base de datos, como se muestra en la figura 7.

Figura 7. Selección de la Base de Datos

Con los datos proporcionados ahora, es posible ejecutar una consulta insertando el comando de lenguaje Cypher, así como los resultados son mostrados a través de la consola, al mismo tiempo se genera un archivo de texto que se guarda en el mismo directorio donde se encuentra la base de datos con el nombre 'QueryResult.txt' (figura 8).

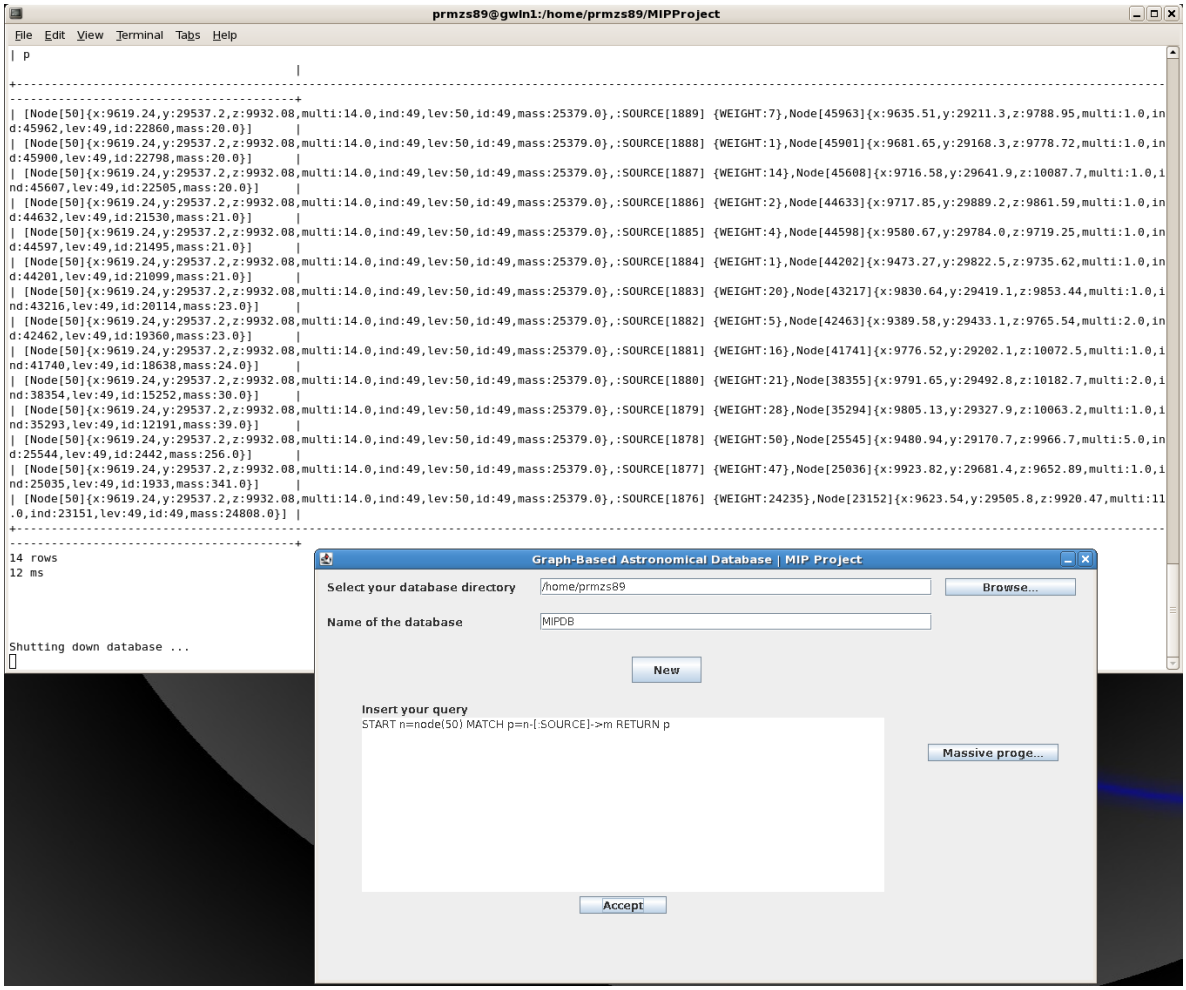

Figura 8. Resultado de una consulta

La función de encontrar el progenitor más masivo fue establecida por la razón de que es información muy importante para el estudio de las galaxias en este tipo de simulaciones, se introduce el número de un nodo inicial, se buscan los nodos conectados al inicial que se encuentran en el siguiente nivel para tomar el que contiene mayor masa y lo mismo sucede con cada uno de ellos, esto dice mucho de la historia de formación de la galaxia. En la figura 9 se muestra el momento en que se introduce el número del nodo inicial y en la figura 10 muestra el resultado del progenitor más masivo.

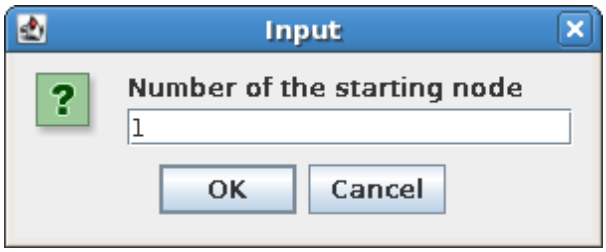

Figura 9. Inserción del nodo inicial

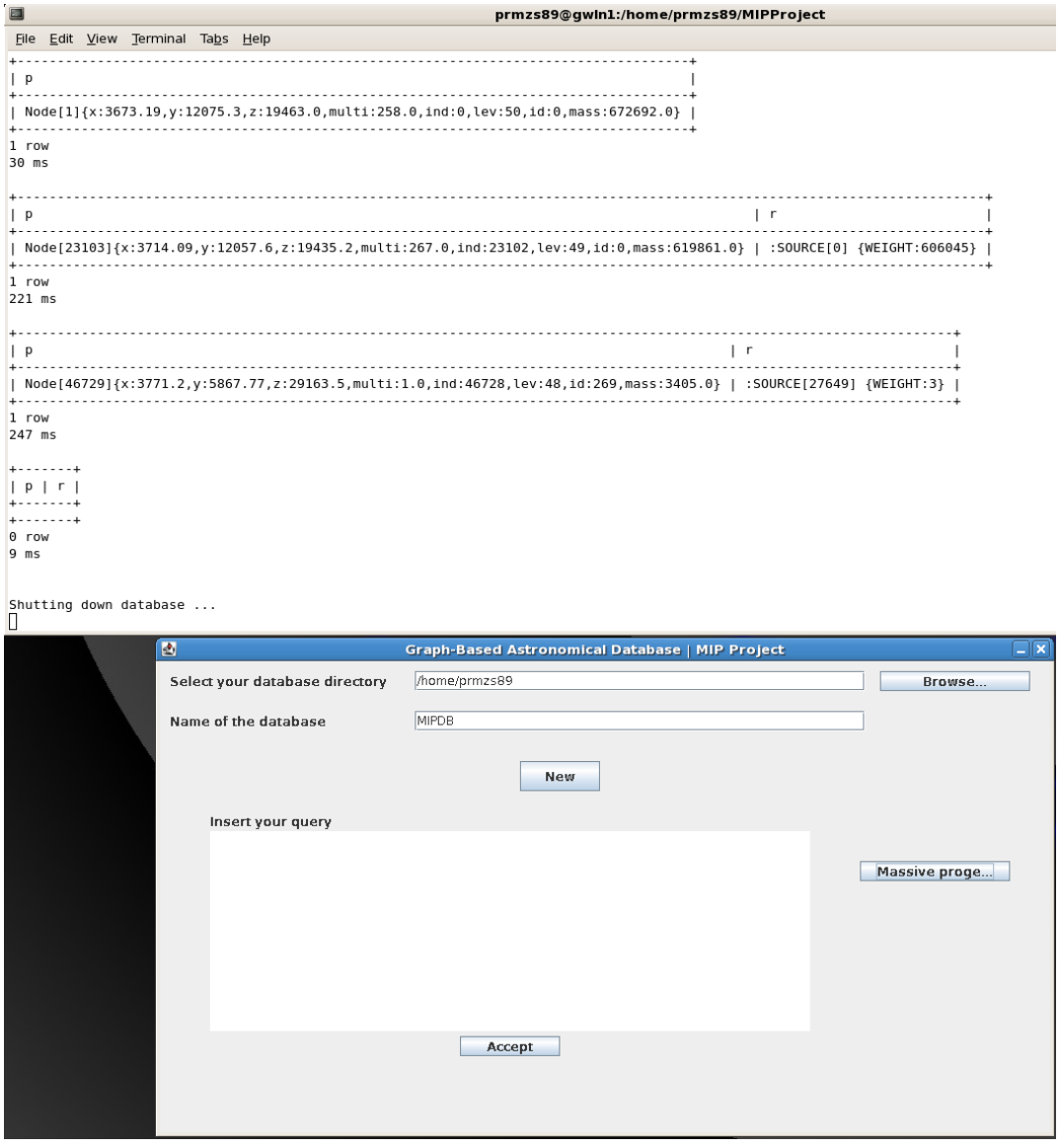

Figura 10. Resultado del progenitor más masivo

Conclusiones

El objetivo principal del proyecto fue proporcionar a los Investigadores y astrofísicos de la Universidad Johns Hopkins la capacidad de almacenar grandes cantidades de información en su estado natural y consultarla de manera rápida y eficiente. Esto se logró a través de la implementación de las Bases de Datos Orientadas a Grafos.

El proyecto ha mostrado buenos resultados con la creación de la base de datos y la realización de las consultas, así mismo, se pueden añadir datos de simulaciones futuras, para obtener una base de datos completa con todo el catálogo de galaxias con que cuenta la Johns Hopkins University. El sistema ha sido diseñado con la capacidad de que se agreguen datos a los ya existentes, estableciendo las relaciones correspondientes, permitiendo que la base de datos crezca según sea necesario.

El trabajo desarrollado en esta investigación, beneficia a la comunidad astronómica que se dedica a la realización de simulaciones en las cuales los datos son representados en forma de grafos, ya que se eliminan los problemas ocasionados en las bases de datos relacionales al tratar de representar este tipo de estructuras, así mismo significa un ahorro de recursos ya que se encuentra desarrollado con software libre, que puede ser usado, modificado y distribuido libremente por lo que no requiere compra de software de ningún tipo.

Se espera que el desarrollo del proyecto aporte grandes beneficios a los investigadores y astrofísicos de la Universidad Johns Hopkins y continué su posterior desarrollo para incrementar las capacidades y funciones que actualmente tiene, revolucionando la manera de hacer ciencia con la Astronomía Computacional.

Glosario

API: Es el conjunto de funciones y procedimientos que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción, también denominadas librerías.

Astrofísica: Es una rama muy sólida de la astronomía que estudia la naturaleza y la estructura física de los cuerpos celestes, tanto próximos como lejanos. La

Astrofísica es una ciencia tanto experimental, en el sentido que se basa en observaciones, como teórica, porque formula hipótesis sobre situaciones físicas no directamente accesibles. La Astrofísica también estudia la composición y la estructura de la materia interestelar, nubes de gases y polvo que ocupan amplias zonas del espacio y que en una época eran consideradas absolutamente vacías.

Astronomía: Es la ciencia que se ocupa de los cuerpos celestes del Universo, incluidos los planetas y sus satélites, los cometas y los asteroides, las estrellas y la materia interestelar, los sistemas de estrellas llamados galaxias y los *cúmulo*s de galaxias.

Big data: Una referencia a los sistemas que manipulan grandes conjuntos de datos. Es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable.

Simulación: Es una técnica numérica para conducir experimentos en una computadora digital. Estos experimentos comprenden ciertos tipos de relaciones matemáticas y lógicas, las cuales son necesarias para describir el comportamiento y la estructura de sistemas complejos del mundo real a través de largos periodos de tiempo.

Web: Se refiere a la World Wide Web o Red informática mundial, es un sistema de distribución de información basado en hipertexto o hipermedios enlazados y accesibles a través de Internet.

Bibliografía

Angles, R. and Gutierrez, C. (2008) *Survey of graph database models.* ACM Comput. Surv. 40, 1, Article 1.

Ceballos, Fco. Javier. (2006) *Java 2, Curso de programación, tercera edición.*Alfaomega Grupo Editor S.A. de C.V.

Coss Bu, Raúl.(2003) *Simulación, un enfoque práctico.* Editorial Limusa S.A. de C.V.

Giudici E., Reinaldo, BrisLluch, Ángeles. (n. d.) *Introducción a la teoría de grafos.* Equinoccio, Ediciones de la Universidad SimonBolivar.

Gulbransen, David. (2002) *Using XML, second edition.* byQue Publishing.

Martín Sierra, Antonio J. (2010) *JAVA 2, Curso práctico, tercera edición.* Alfaomega, Grupo Editor.

Papa, David A., Markov, Igor L. (n. d.) *Hypergraph Partitioning and Clustering.*University of Michigan, EECS Department.

Pressman, Roger S. (n. d.)*Ingeniería del Software, Un enfoque práctico, sexta edición.* Mc Graw Hill.

Caicedo Barrero, Alfredo, Wagner de García, Graciela, Méndez Parra, Rosa María. (n. d.) *Introducción a la teoría de grafos, primera edición.* Ediciones Elizcom.

Referencias electrónicas

Anónimo 1. (n. d.) *Astrofísica*. Consultado el 10 de Octubre desde http://www.astromia.com/glosario/astrofisica.htm

Anónimo2. (n. d.) *N-Body Simulation*.Consultado el 11 de Octubre desde http://en.wikipedia.org/wiki/N-body_simulation

Anónimo 3. (n. d.) *Programación dirigida por datos*. Consultado el 11 de Octubre desde http://es.wikipedia.org/wiki/Programaci%C3%B3n_dirigida_por_datos

Neo Technology. (2012) *The Neo4j Manual v1.9-SNAPSHOT*. Consultado el 11 de Octubre desde http://docs.neo4j.org/pdf/neo4j-manual-snapshot.pdf