CSCI 5408 - DATA Management Warehousing Analytics

Assignment 2

YUGANTHI KRISHNAMURTHY

B00839935

yg617681@dal.ca

Section A. Summary

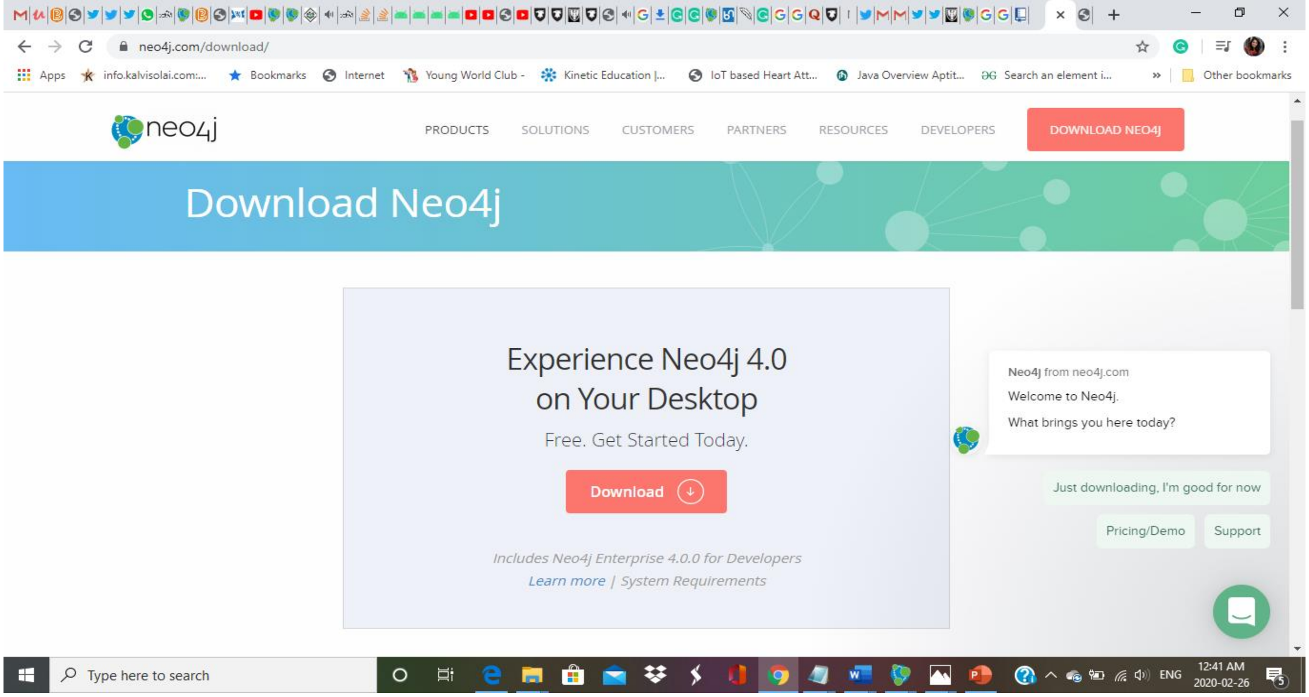
SUMMARY OF THE DNR CAMPING PARKS DATASET

The DNR Camping Parks Reservation Data 2016 gives information on the camping sites. These camping sites include information about the parks in different provinces. But the major focus is on the parks in the Nova Scotia region. The information provided in the dataset is collected from the reservation system used for the camping site reservation. The usage of the camping sites is major in the summer season. From the given dataset, it is understood that the season for major bookings for the provincial parks in the month of March to the month of October. This dataset is included under the forest's category. The information about the camping sites is included in the department of natural resources. This dataset contains information about parkName, state, country, adult, child, partySize, rateType, bookingType, equipment, bookingStartDate, bookingEndDate, Night, permits. From these attributes, we could gather the necessary data. By gathering the necessary data, we can analyze the parks in different provinces, the size of the park, the availability date for the park, booking type of the park. The booking rate differs for different aged people. This dataset can be considered as a vital one as it gives more information on the recreational spots. From the provided information the graph database can be created. The graph database provides a good analysis and visualization of the data about the Nova Scotia provincial parks. The parks are related to each other with the same dependency attribute. The dataset can be considered as a valid dataset as it contains information related to particular data. The data on NS parks helps us to visualize the complete information on the parks. The dataset on DNR Camping parks helps us to provide information on natural resources. It also helps the forestry department to find camping sites and maintain it.

Section B. Programming

I have used the java programming language to clean the data in the excel sheets. I have created three csv files to satisfy each criterion and store the data after data cleaning. I have attached the java files to the zip file. The folder file1 contains the code for the file1 containing data on all parks of Canada. File2 folder contains the java programming for data on ParkName, State, PartySize, BookingType, RateType, Equipment. File3 folder contains the java code for the data where all "less than" is replaced with "LT". The folder java source file contains only the main java source code files. File1 contains the java code extracting data on all parks. File2 contains data on ParkName, State, PartySize, BookingType, RateType, Equipment. File3 contains data where all "less than" is replaced with "LT". Csv files folder contains all the csv files after data cleaning.

Section C. Installation



Download Neo4j

Experience Neo4j 4.0
on Your Desktop

Free. Get Started Today.

Download

Includes Neo4j Enterprise 4.0.0 for Developers
Learn more | System Requirements

Neo4j from neo4j.com
Welcome to Neo4j.
What brings you here today?

Fig 1: Neo4j installation from google

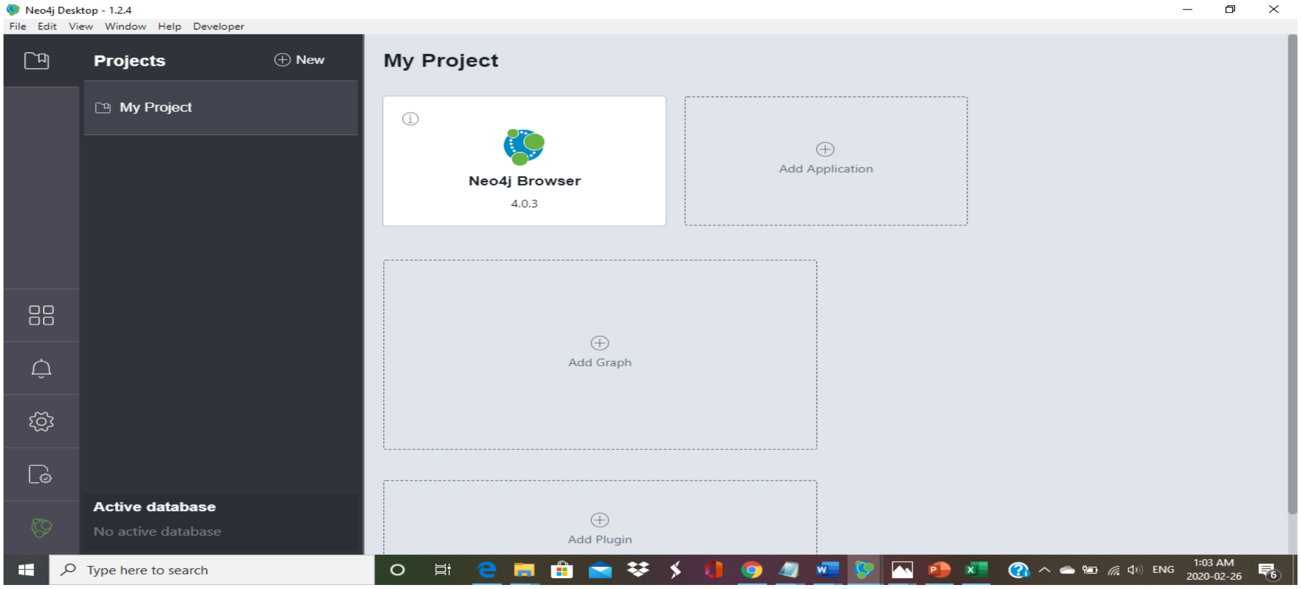


Fig 2: opening Neo4j in the desktop

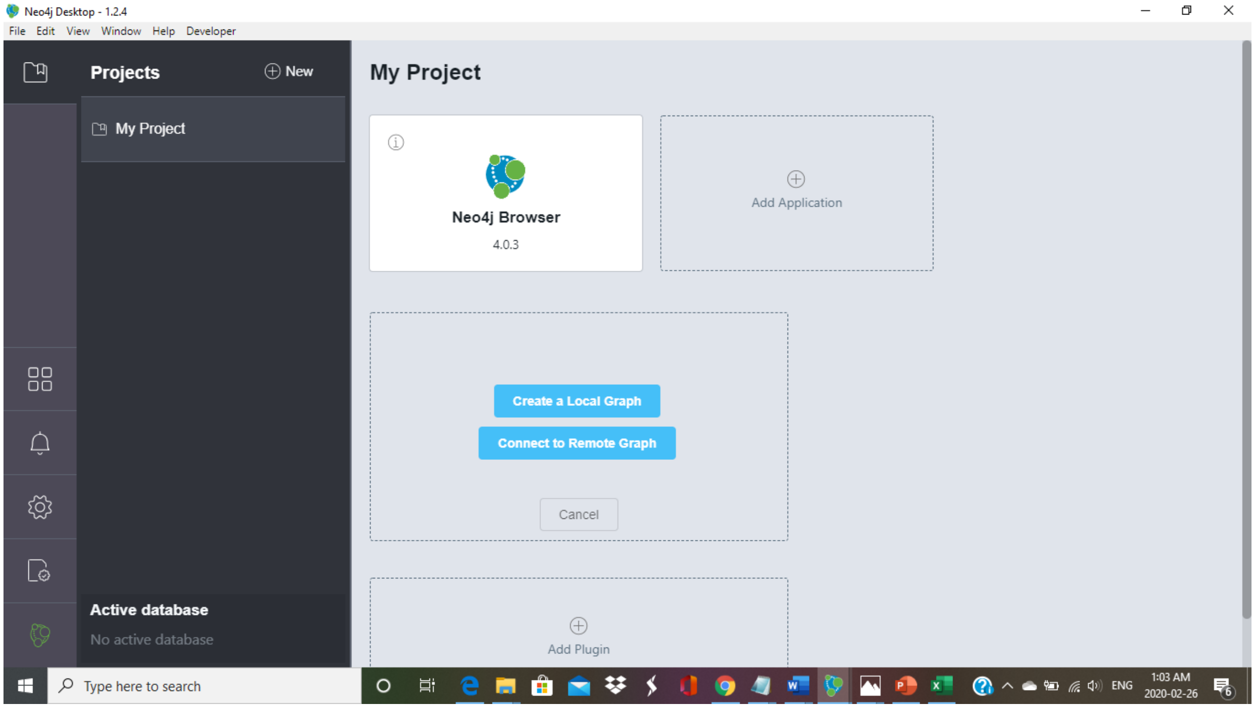


Fig 3: Creating the local graph

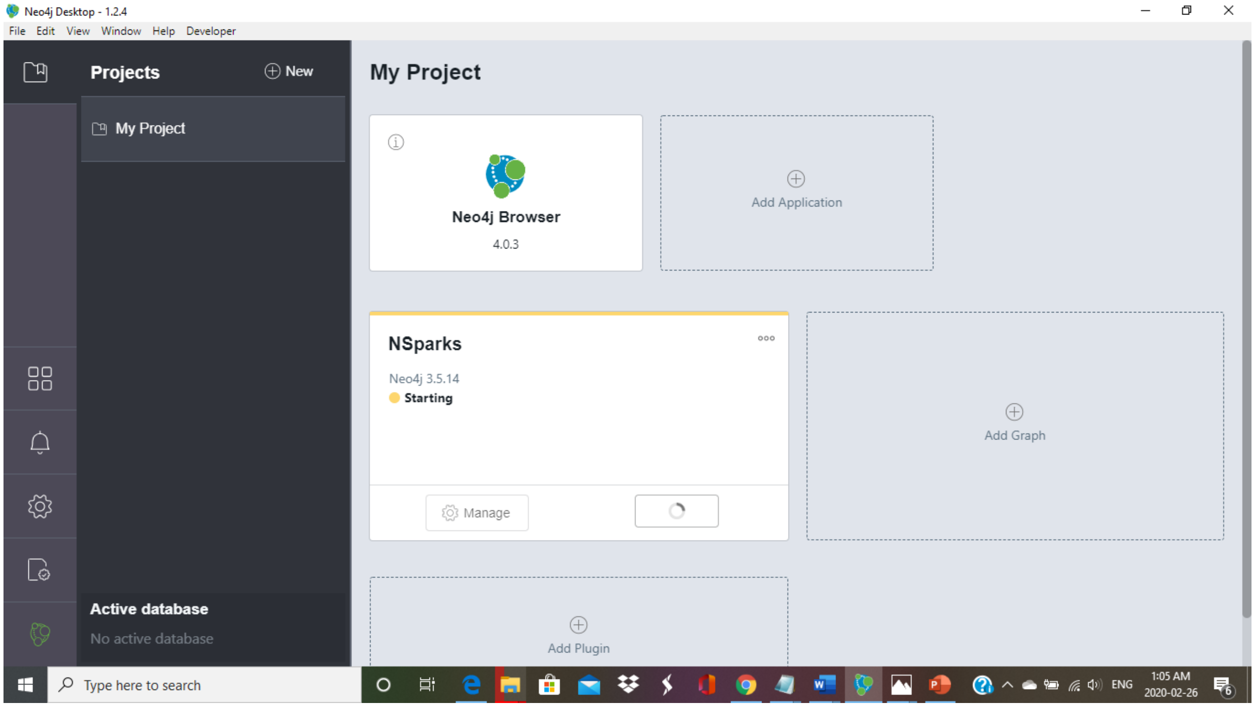
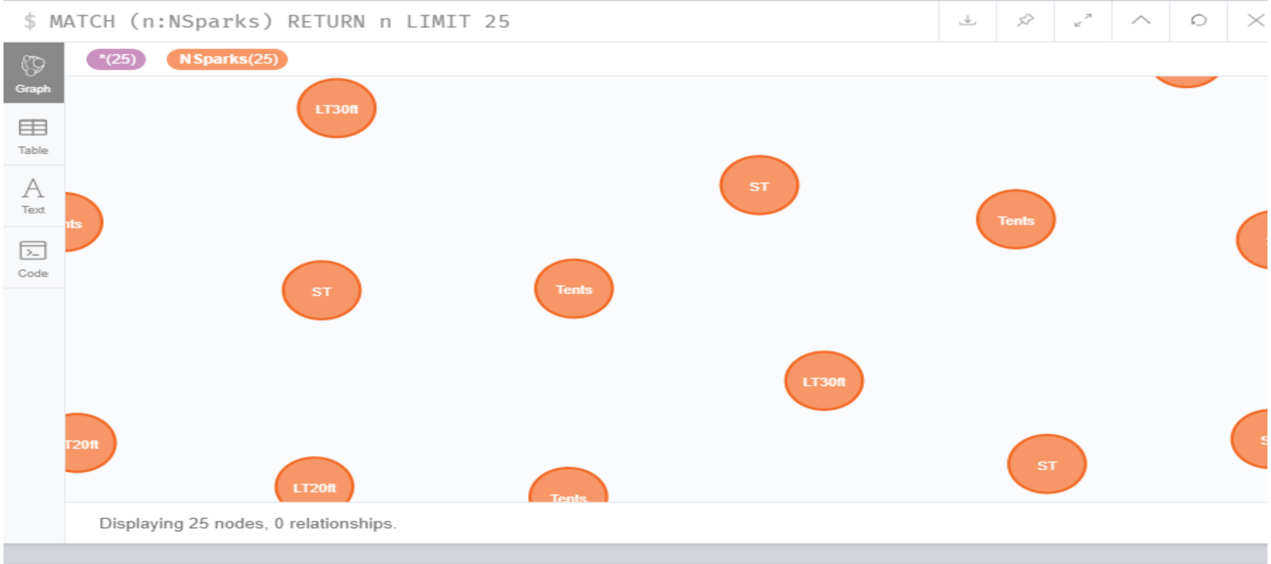


Fig 4: Starting the new graph database

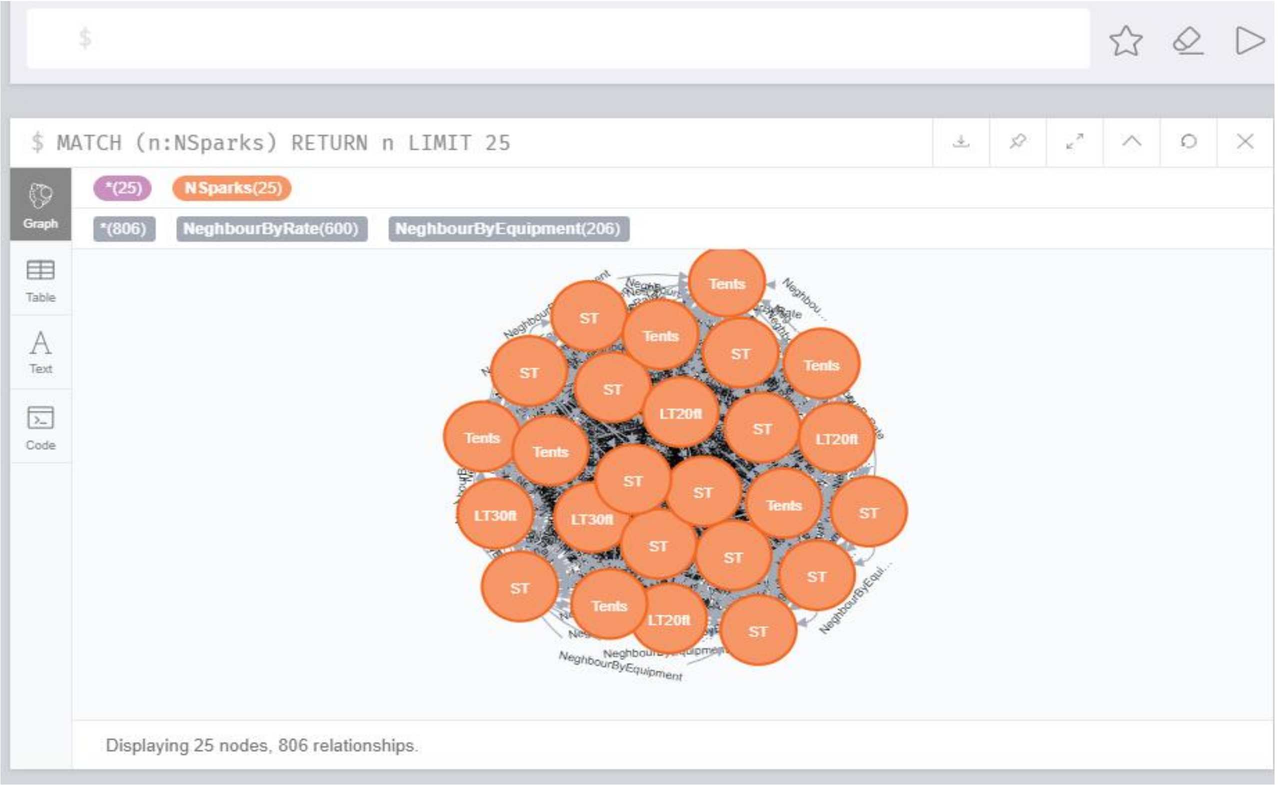
Section D. Building the graph using neo4j



```
$ MATCH (n:NSparks) RETURN n LIMIT 25
```

Displaying 25 nodes, 0 relationships.

Fig 5: Loading each park as a node

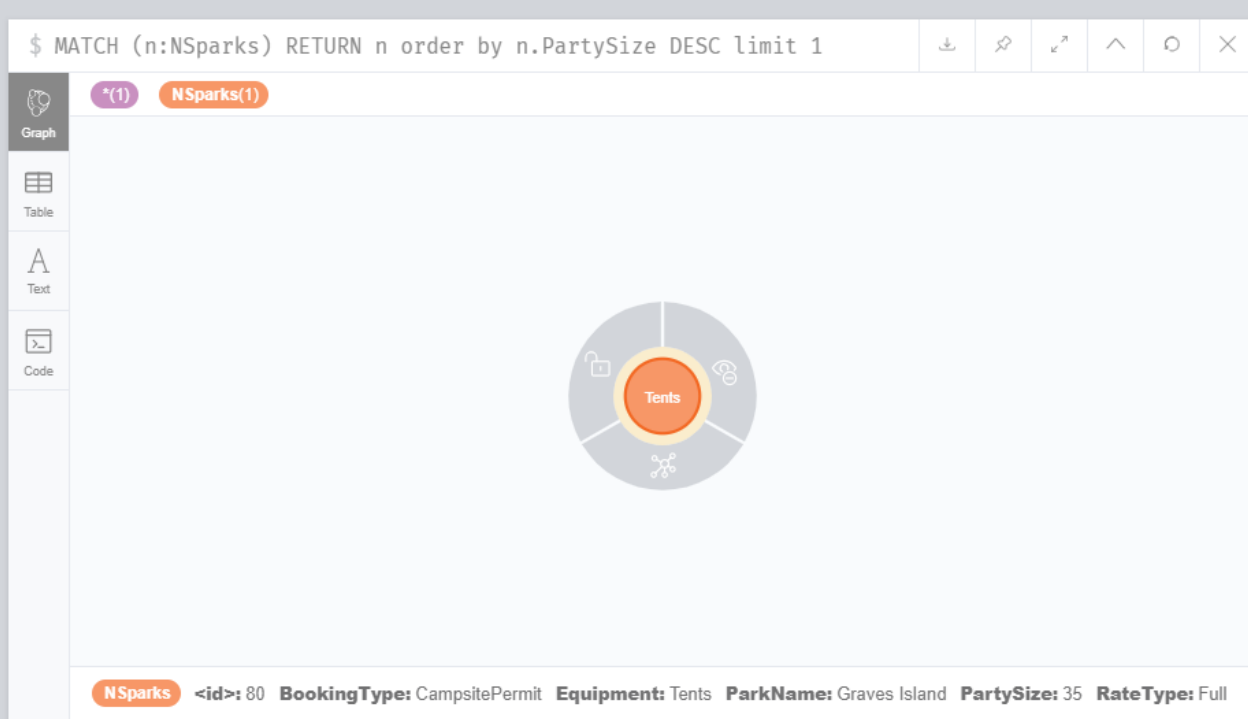


```
$
```

```
$ MATCH (n:NSparks) RETURN n LIMIT 25
```

Displaying 25 nodes, 806 relationships.

Fig 6: Nodes connected using edges



```
$ MATCH (n:NSparks) RETURN n order by n.PartySize DESC limit 1
```

NSparks **<id>:** 80 **BookingType:** CampsitePermit **Equipment:** Tents **ParkName:** Graves Island **PartySize:** 35 **RateType:** Full

Fig 7: Parks that support highest number of partySize

Section E. Technical report

Technical report on NEO4J

Graph Database

Neo4j is a graph database tool. A graph database is used to provide the information about the tables and data in the tables in the form of graph. The graph includes nodes and edges. The nodes are connected with the help of the edges. The main advantage of using the graph database is that it is very easy to visualize data. The user can easily handle the data by accessing the nodes and finding the relationship between the nodes with the help of the edges. The main purpose of using the graph database is because it is simple. It allows the user to include the new data without loosing the existing data. Since the nodes are connected without any index reference, the data can be retrieved as quick as possible. The graph database is also used to handle the complex data. The hidden relationships between the data and the nodes can also be easily identified with the help of the graph database. It also helps to understand the complex queries. Thus, it helps developers to handle the major debugging errors with the complex queries.

Neo4j

Neo4j was developed by Neo4j, Inc. The main advantage of the neo4j is that it provides online backup. Neo4j uses ACID property. ACID stands for Atomicity, Consistency, Isolation, Durability. Neo4j majorly deals with the database that follows the ACID property. ACID property includes sharing the data safely. Cypher query language is used to query in the Neo4j. Neo4j is an open-source software or tool. Neo4j stores database values in the local storage. The navigation between the nodes happens with the help of the pointers. The reason for using the pointers to store the data in the neo4j is because the retrieval of the data is faster.

As it uses the pointer to store the data, it is considered as the index free. If indexes are used to store the data in the database, the retrieval time is faster.

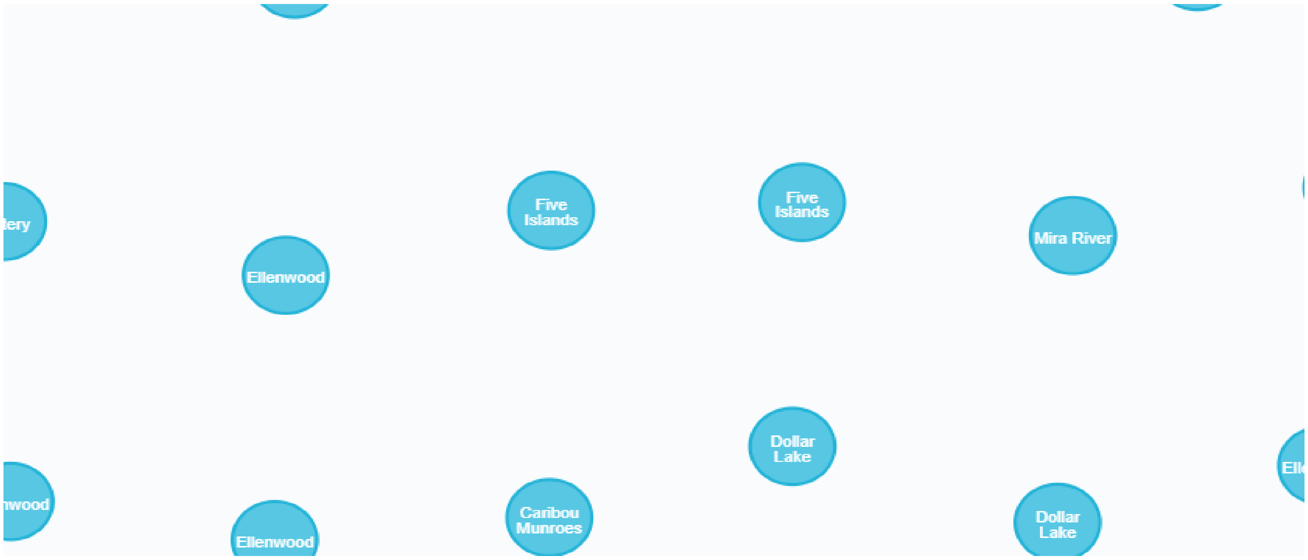


Fig 8: Graph created with nodes using the Neo4j graph database tool

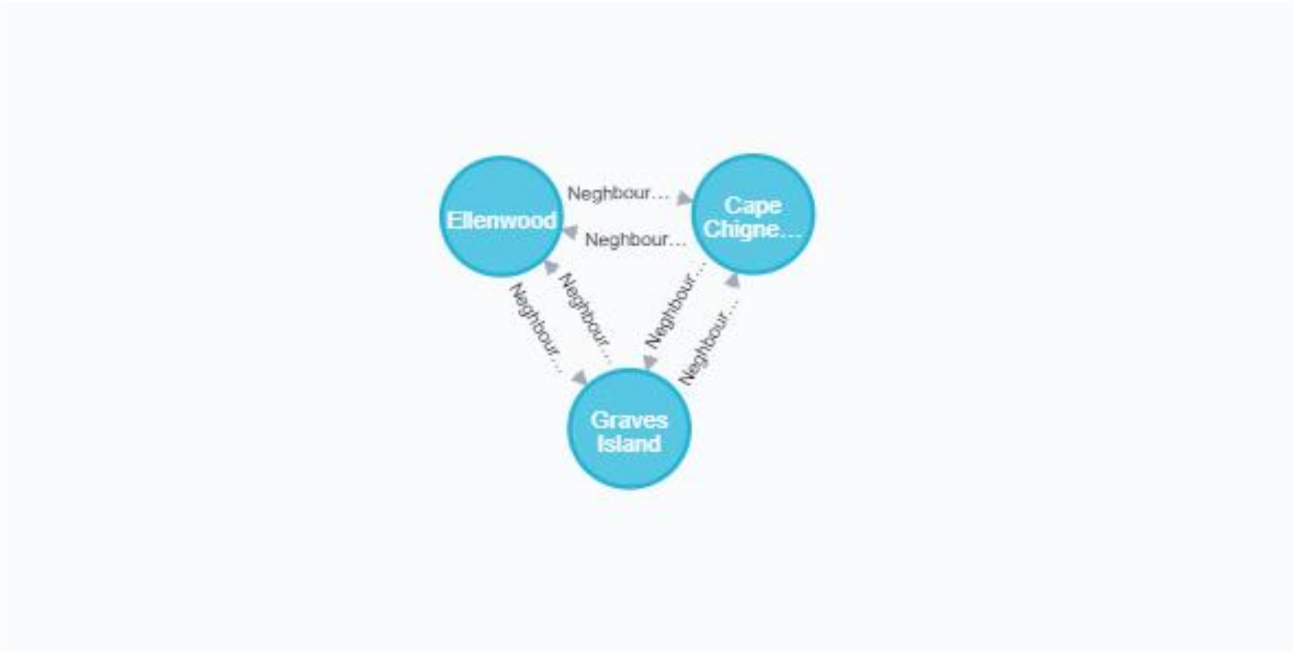


Fig 9: Graph created with relationships using the Neo4j graph database tool

Significance of Neo4j

Neo4j has many significances. It stores the data in the local data storage. It uses ACID property and ensures safe transactions and access to the data. As the data used everyday increases, it is difficult to handle the numerous data with the help of the tables in the database. Instead, the neo4j tool can be used to handle the huge data with the help of the graph. The data can be accessed easily with the help of the graphs. One major advantage of the Neo4j is it is an open-source software tool. It also provides fast reading and writing operations. Many organizations and industries are using neo4j. Some of the most significant organizations which use neo4j are Facebook, eBay, etc. These organizations use neo4j because they handle an enormous amount of data. The execution time is very fast in neo4j. It is faster than the MySQL database. It can also handle complex queries. As a result, it can debug complex queries too. There are no limits on the amount of data that can be handled by the neo4j. It uses high-speed frameworks. These frameworks can be used to handle faster transactions. Neo4j can be handled with the cypher query language. It is used to write complex queries with great ease.

Limitations of Neo4j

Neo4j faces few limitations. Every tool has its own significance and limitations. Neo4j does not allow the user to insert, delete, update values as Mysql workbench. Neo4j does not uses the index values. Instead it uses the index-free values. Though the retrieval speed is faster, the correct identification of the data, its relationship with the other data is quite difficult. Neo4j can be used only with the cypher query language. So, it provides difficulty for the user to learn a new programming language. When a node is deleted, the user can insert the data into the index of the node that was deleted. This results in pointing to different node which has index that was initially assigned to a different node which was already used by the deleted node. This results in causing the confusion and chaos. There are few issues with cypher query language which is to query and retrieve the data in the neo4j. Whenever the merge statements are used, the uniqueness of the data is not always maintained.

Alternatives of Neo4j

The visualization and analyzation of the given dataset can be done by the various alternative tools to neo4j. Some of the neo4j tools which can be used as the alternative to the neo4j are JanusGraph, Titan, Neptune, Cassandra, MongoDB, etc. The analyzation of the data can be best done with the graph database approach. Apart from the other methods, graph database is the best method to visualize and analyze the huge data.

References

[1]"What Is a Graph Database and Property Graph | Neo4j", *Neo4j Graph Database Platform*, 2020. [Online]. Available: https://neo4j.com/developer/graph-database/. [Accessed: 26- Feb- 2020]

[2]"What is the graph database? What is data model? - Bitnine Global Inc.", *Bitnine Global Inc.*, 2020. [Online]. Available: https://bitnine.net/blog-graph-database/what-is-the-graph-database/?gclid=CjwKCAiAhc7yBRAdEiwAplGxX1Qks-nOKFhDKKU--UlKFdyVp-cpDO4yBRiQoc5hnQRnF_nOFvuGOBoCxwEQAvD_BwE. [Accessed: 26- Feb- 2020]

[3]"Welcome to the Dark Side: Neo4j Worst Practices (& How to Avoid Them)", *Neo4j Graph Database Platform*, 2020. [Online]. Available: https://neo4j.com/blog/dark-side-neo4j-worst-practices/. [Accessed: 26- Feb- 2020]

[4]"Reading and Writing CSVs in Java", *Stack Abuse*, 2020. [Online]. Available: https://stackabuse.com/reading-and-writing-csvs-in-java/. [Accessed: 26- Feb- 2020]

[5]"How to Read CSV File in Java - Javatpoint", *www.javatpoint.com*, 2020. [Online]. Available: https://www.javatpoint.com/how-to-read-csv-file-in-java. [Accessed: 26- Feb- 2020]