Graph-Based Database Management System Use Cases for Retail Titans - Based on Walmart and eBay Utilizing Neo4j

Yi-Hsueh Yang
Heinz College, Carnegie Mellon University
Master of Science in Public Policy and Management – Data Analytics
Pittsburgh, PA, United States
yihsuehy@andrew.cmu.edu

***Abstract*—With database management systems appearing in succession in response to the exponential growth in data and information, the level of compatibility between the database and customers' pain points greatly speaks for the performance of companies. NoSQL database systems are bringing revolutionary changes into the data storage world, with key-value pair, document-base, column-family-based, and graph-based these four types of storage methods, many business problems are gradually conquered by utilizing them. This research report aims to break down the process and upshot for a solution in each of the companies after transforming from a relational database to Neo4j. The ultimate goal is to bring out a thorough discussion and analysis of how this new technology works for these companies and whether there may be some potential hazards that existed for them.**

Keywords—***Walmart, eBay, Neo4j, NoSQL, relational database***

I. Introduction

This research concentrates on the use cases of two retail titans - Walmart Inc. and eBay Inc., referred to hereinafter as "Walmart" and "eBay", in an identical NoSQL database management system - Neo4j in detail. The background, the methodology, the effectiveness of using, and the comparison of these two use cases on one database management tool are discussed. The purpose of this research is to dive into the methodologies and the pros and cons of these two use cases and provide a general overview of how the non-traditional approach help solves the problem that evolved in recent year. Moreover, meanwhile looking at the outstanding parts of the technology, and how the breakthrough dominates the current e-commerce market, I also want to evaluate whether the downsides of the technique are going to be viewed only as a temporary solution if the market and people's demand continue to grow. In short, I'm trying to understand whether the solution of using Neo4j as their backend database entirely solves the problem of growing users without any concern in the future.

II. Type of business

A. Walmart

Walmart is an internationally-known retail company that was established in 1962. As a family-owned business, Walmart has become the largest private company with 2.2 million employees, owning more than 10k stores worldwide and collecting about 36 million dollars from its US stores every day. Besides operating a chain of supercenters, Walmart has an enormous e-commerce service that facilitates online ordering and shipping. There are currently 10 websites operated by Walmart to service online retailing to its customers. For handling emerging digital customers and potential future use, Walmart has built a data center called Area 71 in Jane, Missouri. The data center keeps Walmart's data secure within its own facilities and aids the retail titan to store terabytes of data and conduct big data analysis for its products and customers.

B. eBay

Unlike Walmart, eBay is purely a company that focuses one hundred percent on e-commerce, it has neither supercenters nor physical stores, it performs as a marketplace allowing sellers and buyers to trade. Originating in 1995, in San Jose, California, eBay gained its ground after the doc.com foam, it is now to be said one of the five largest e-commerce platforms in the world. The business model at eBay is mainly charging selling fees from the sellers that are putting up goods to sell.

There are two main parts of the fee, one charge whenever you start your business on eBay and list over 250 items a month, and another charge 10% to 15% of the final whenever a trade happened. For increasing revenue, eBay tries its best to deliver goods to their customer within 3 days to meet customers' needs. However, challenges coming up with more and more e-

commerce platforms are emerging, a way to stay competitive is to reduce the delivery time from seller to buyer, that is the time when eBay is constantly improving their delivery, which results in options like 2-day shipping and 1-day shipping appears.

III. Previous data use and Storage Method

A. Walmart

Making the website sit still passively isn't a huge problem, but it is not ideal to expect a big leap in revenue. Therefore, Walmart rapidly introduced the recommendation system on its website, trying to increase its selling performance by recommending products and stimulating people to buy more while browsing its website. Initially, Walmart utilizes the batch process real-time online recommendation method which is a way to recommend products by throwing batches of data into the recommendation calculation model to generate products with a higher potential of being chosen. The batch method had its impact, however, accompanied by the increasing amount of customer visits and requests sent, the drawbacks gradually emerge: the system is experiencing more and more latency, and some manpower has to be assigned to do the analysis work.

For understanding the cause of the problem, we can dive deeper into the batch processing method. The concept of the batch processing method works in two ways, one is it processes at a given time, and another is it processes when the amount of data has reached its threshold. The data are stored in a conventional relational database which data analysts or data scientists extracted data from whenever the premise mentioned above is met. This leaves a huge risk to the system whenever the visits to the website are huge of a sudden, the webpage might not be able to react and provide a high-accuracy recommendation in real-time. The failure of the system is expected to be more serious if the website is constantly meeting its maximum capacity of visits, and the company might start to lose its client and resulting in a loss of sales and revenue. Therefore, to prevent the potential hazard, Walmart shifted its focus to the NoSQL database management system, referred to hereinafter as "DBMS", Neo4j in particular, to transform the way they had their data stored and improve the recommendation system.

B. eBay

Shutl, a UK-based company was acquired by eBay and renamed "eBay Now", facilitating the shipping procedure for eBay in the UK initially but expanding its service territory to the US in 2014. Shutl acting like a coordinator between customers, couriers, and suppliers. They use data like product information, inventory information of suppliers, couriers' availability, and address from both ends to come up with a better delivery strategy and timing. Promising customers to get their goods in time, ensuring a sufficient number of couriers, and keeping orders running for suppliers to sell more, what Shutl is doing, is the most important work in the system, therefore, optimization is considerably important. However, all the data are stored in MySQL DBMS initially. The relational database can handle it at first but not until when the number of requests greatly increased, the problem of the relational database was then unveiled. Joins are too time-consuming to operate, especially when the query grows larger and larger. [1]Volker Pacher, a senior developer at eBay, even stated that the slower query they have, about 15 minutes, is longer than the fastest delivery they offer. These reasons are propelling eBay to find a new way to restructure their data and they turn to the graph-based DBMS - Neo4j.

IV. Use Cases

Neo4j is a graph-based NoSQL DBMS where there are three most important features: nodes, features, and edges. Nodes are entities that can hold key-value pairs and be tagged with labels, similar to rows in conventional RDBMS. Features are the characters that usually come along with nodes and are helpful for users to identify or categorize them. From the perspective of RDBMS, It can be viewed as columns that are defining the rows underneath them. Edges are usually called relationships which is the most important difference to mention, not existing in any other DBMS, a completely different model to store data and provide clear connections between two existing nodes.

There are many ways for different graph databases to store data, some of them are based on the relational engine and have their data stored in tables while some use local graph storage which Neo4j falls into this category. Graph storage provides magnificent efficiency when searching for data since the model setting makes it a database index alike, so you can see it as a huge piece of index contributing to incredibly high-speed indexing and giving out the data users require. With all of these features, Neo4j is known to be the solution to handling complex connection problems. Companies value three characteristics when choosing DBMS: intuitiveness, speed, and agility.

A. Walmart

The design for Neo4j is straightforward with the term of node and edge(relationship), it helps the user easily understand the relationship within the linked list of entities. The design itself also avoids excessive joins than relational databases to get relationships between 2 entities. Other kinds of NoSQL databases can also greatly lessen the number of joins in the query, however, those have to go through the searching process of using indexes which inherently makes Neo4j a better solution in acquiring relationships. As with the other NoSQL DBMSs, the ability to alter data dynamically is never a problem, the effect towards it is substantially smaller than relational DBMS. Neo4j can even remain its status available for computing while altering the graph which makes it extra efficient in performing real-time calculation results.

In addition, older versions of recommendation systems are mostly built on keywords that users insert into the search engines, indicating the recommendation are surrounding the

[1] *Case Study: eBay Now Tackles eCommerce Delivery Service Routing with Neo4j* (NeoTechnology, 2014). https://dist.neo4j.com/wp-content/uploads/Neo4j_CS_eBay.pdf.

keyword, not including other factors like environment and using conditions. For example, if the word 'coffee' is detected, the system will probably recommend products like coffee drinks or coffee beans, however, it overlooked the situation when one would like to look for the coffer brewer instead. It is extremely essential to link as much relevant information as it can to provide a more ad-hoc recommendation. To be said, Walmart achieved this goal by implementing artificial intelligence to recognize the needs of customers and rapidly export relevant products linking data like product information, use environment, and use condition to their system with Neo4j. In this way, the recommendation no longer looks identical to one another but customer-oriented demand recommendation is shown and make the system more lively and thoughtful.

B. eBay

Walmart uses Neo4j to greatly reduce the processing time and make real-time recommendations smoothly while eBay uses it to perform more efficiently in its logistics and implement same-day delivery to its customers. For storing every piece of useful data from products, customers, and their logistics partners, eBay had them originally stored in normalized RDBMSs. Whenever there is a request coming up from the customers, the system would start to collect data related to the order, such as addresses, products, product warehouses, carriers, etc. Since data of those types are stored in different tables and sometimes in different databases, it creates a large burden for the system. That is to say, the system is joining many tables only for one specific piece of information for a huge proportion of the time. Inherently, joining tables is a time-consuming job for DBMS to perform.

Although the system might handle a small number of jobs with ease at first, however, the processing speed of the system back-end gradually decreases as more products are stored, more customers' information is stored and more requests came. For tackling millions of requests at the same time, the concurrency calculation result is important but never an easy task for eBay anymore. It is at that time, eBay renew its architecture and introduced Neo4j to provide better calculation speed to satisfy its customers with its goal of achieving same-day delivery. The data model completely changes the performance of the service, both customers and eBay have more options on choosing or offering when it comes to product delivery. More delivery options give flexibility to more user groups to cover all the different preferences one may have. eBay, therefore, easily keeps its customers sticking to its platform and generates more profit from its e-commerce business.

Gaining more money externally from the firm perspective, Neo4j also helps the interior personnel to do their job more efficiently. Cypher is a query language designed mainly for the use of graph databases, especially Neo4j, it adopts the features of Neo4j: nodes, and relationships, allowing it with high readability since the syntax can be translated into human-spoken language easily. The creation of it is based on the concept of SQL language also providing it with strong intuitiveness which makes it easy to learn. The combination of Neo4j with the Cypher query language brought out a transformative impact on eBay's engineers, shortening its coding lines from plausible hundreds of lines of code mostly consisting of JOINS to Cypher code under 10 lines that can solve identical tasks.

eBay imported Neo4j for increasing the speed of database calculation and to implement the service of same-day delivery. The idea behind this is to provide new calculations to both the seller and the logistics unit whenever a new request comes or a new update is requested. The importance of seller giving out its product to the right logistics unit that is currently available within the nearest range of the seller is crucial, after receiving the product, distributing all the products to different customers with the most efficient route is equally crucial. Both the sellers and logistics rely heavily on the accuracy of the plan to provide adequate service, so the customers can enjoy the service and remain shopping at eBay.

Recall from the features that Neo4j has, the agility of it has made eBay combats easily the exponential growth of customers, new nodes, features, and relationships can be constantly added to the database without any great impact on the data searching speed; the unseen processing speed has also made eBay working without hardness while streamlining routes and carrier for high demand delivery. This not only brings back the competitiveness of eBay in the e-commerce market but also provides customers with some amazing delivery services.

V. COMPARISON

A. Similarities

Both companies turned to identical NoSQL database management systems, there are numerous kinds of DMBS on the market the reason why they both chose Neo4j is highly connected to the data structure these companies are having. Relationship is an important feature of Neo4j, with its help, incredible process, and response speed of picking out data can be performed. It is equivalently important to both companies to present the versatile relationship between thousands and millions of products and the requirement to find the shortest path and vacant courier timely. The Similarities of both use cases from Walmart and eBay are:

- Facing the Same difficulties: Both companies have their database architecture transformed because of facing a high quantity of growth in their requests from customers.
- Helping Speed up: Neo4j is helping them speed up their queries and avoid joins by using Cypher to substitute traditional SQL language.
- No Constant Changes: Attributes of data like product information and shipping address, do not alter often for the retail industry, only the function of adding and deleting is encountered, so Neo4j can perfectly suit their needs.

B. Differences

Although importing the same kind of technology and adopting the same database to improve their system, there are still many different perspectives touched due to their different

situation faced. From the primary usage and the target that the company is aiming for, Neo4j is used in several different ways in these two cases. There are two main differences: usage scenario and relationship determined.

Walmart uses Neo4j to achieve a customer-based ad-hoc recommendation system on their e-commerce webpage. It brings out more in-depth relationships and brought out information rapidly to form a customized recommendation list, trying to keep their customers longer on their webpage shopping for their needs to transform into revenue for the company. Whereas eBay uses Neo4j to optimize its calculation speed to achieve same-day delivery in their merchandise shipping service. It utilizes the advantage of Neo4j for adding nodes and deleting nodes and make the best use of linking relationship between multiple nodes to form the shortest path for sellers, customers, and couriers. Providing better service to all three ends and expecting to have a higher customer stickiness to generate more promising profit in long term.

The relationship is an indispensable part of the database, it provides connections to nodes that are correlated to each other, not only providing an RDBMS-joined-table-like function but also allowing the system to retrieve the connection of connection, not just one layer anymore but multiple layers, in an unbelievable speed. It is utilized and defined in different ways for those two implementations in retail companies. On one hand, product information, user condition, and user scenario are transformed into relationships in the Neo4j database in Walmart, the relationships are treated as more customer-oriented, so to allow more ad-hoc recommendations. On the other hand, at eBay, relationships are used to record the connections between customers, sellers, and logistics, more specifically, product inventory, addresses, and status of idle delivery couriers. It allows eBay to update its product and logistics situation with real-time data constantly, thus, it can be more capable to offer more variety of shipping services and make sure those services are right on track.

VI. Potential Drawbacks

The introduction of Neo4j helped Walmart and eBay level up their business can are capable to deal with a larger number of requests and customers, it won't seem like a problem if their users aren't increased like before. Even though it seems to be the ideal solution for e-commerce, the speed can be raised super high too, relationships can be found in an unrealistic amount of time, and everyone got their advantages and disadvantages, including Neo4j. Here we can discuss some of the problems lying along with its data model and its characteristic:

A. Doesn't Support Distribution System

The data model itself is not designed as a way to be computed on a distributed system, so the application with Neo4j can only be done locally. It is still possible to utilize it through a distributed computer but eventually is not an easy task since it is not designed to do so. Due to the fact, there is no parallel computing, the data cannot be shared, and there is always a single node solely computing in a graph database. Replication is done for every newly added data to enable index-free adjacency and to keep its referential integrity safe. Although we are amazed by its incredible speed of retrieving data, it still has its limit upon facing an enormous number of requests simultaneously. The only way to increase its speed and capacity is by scaling the hardware vertically, upgrading with memories and Solid-State Drives with greater capacity, however, there is obviously an upper limit, and therefore there comes the bottleneck.

B. Doesn't Scale Well Horizontally

Neo4j does not scale as well as the other NoSQL DBMS out there in the market. To be frank, it is absolutely doable to scale the Neo4j database into some gigantic chunk of datahub that consists of millions of nodes and relationships by sharding. However, if the database grows to a size that big, it slows down the processing speed while node hopping around connections. The database can handle some degree of traversal but not to be increased unlimitedly, so the Neo4j database can perform extremely well on go relatively local graphs with provided local queries but might run into some problems when encountering large graphs.

C. Bad Input Performance

The graph data structure leads to poor writing performance, real-time reading and writing are hard to keep up, and importing large amounts of data is troublesome. The performance of the official methodology of loading CSV files is not ideal whereas Neo4j-import is good, but it can only be used for database initialization.

D. [2]No Data Protection

Another problem for the Neo4j database is that all data are stored in a single database, this can be viewed as a mechanism for design to increase its speed in retrieving data, but it also satisfies the option to compute in parallel. Though it operates the same way as RDBMS on a local computer, it doesn't provide data encryption. In other words, the process of inserting a username and password to log in that we are pretty familiar with does not exist in Neo4j databases, instead, the user would have to utilize Java or JVM to encrypt their data.

E. Response to Drawbacks

Neo4j is especially suitable for the relationship between community or network websites in which the relationship between users including friendship, relatives, friends, colleagues, etc., each person is regarded as a node, and the relationship between users is regarded as an edge, the whole relationship or connections form into a gigantic network of relationships which is both easy to read and provide extra

[2] Pethuru Raj, Ganesh Chandra Deka, *A Deep Dive into NoSQL Databases: The Use Cases and Applications, Advances in Computers* (Elsevier, 2018) Volume 109, Chapter 2, Pages 49-76, ISBN 9780128137864, https://doi.org/10.1016/bs.adcom.2018.01.004.

convenience for engineers to utilize the data. Leveraging the speed that it can provide and the potential disadvantages this NoSQL DBMS is holding, there are several checkboxes to provide measurements if Neo4j is a great alternative for your architecture.

1) Whether your current RDBMS is either becoming slower gradually or can't handle all the requests and calculations anymore;

2) Whether the dataset you have is something that relies heavily on the storage of network relationships;

3) Whether your organization has the resources or abilities to keep all the data safe by any external measures;

If the answers to all three of the questions above are positive or close to positive, then the migration from RDBMS to Neo4j is worthy to give it a makeover. Some successful examples like the transformative transition from RDBMS to Neo4j in Walmart to provide a more accurate recommendation to its customers which brings a huge amount of revenue to it and made it one of the biggest e-commerce platforms nowadays and also the shift of architecture from eBay allowing it to make instant calculations precisely and handle many other emergencies to keep their promise with their speedy delivery service, both of them have proven that the changed to Neo4j is somehow a big move for both of their business.

However, there is also a hidden requirement or precaution worthy to keep in mind before an action. It is always better for the organization or the database user to leverage the confidential level and the amount of data since more than one database might be required for decreasing the risk of it crashing out or being hacked. It can simply be split into different databases if the purposes are different and data from different databases have little connection with each other.

VII. Conclusion

Overall, Neo4j is a graph database specializing in storing data of networks and relationships. It is a powerful tool with a bunch of outstanding advantages that attract many fortune 500 companies, like Walmart and eBay per se, to utilize it as one of their database infrastructures. First, it is known for its rapid-fire speed of database operations as if the database is big enough. Different data are distributed into different tables in the traditional RDBMS, and since it is mostly normalized, there is only one record for each piece of data, indicating many joins would have to be operated each time a search is needed. Since JOINS are relatively time and memory-consuming, Neo4j does a really good job of speeding up and saving a lot of time as the volume or requests get larger. Second, the database itself is pretty intuitive, the nodes represent every object in the real world and the edges represent the interaction or the relationship between two nodes which is both readable and understandable. Although the supporting language Cypher is completely different from SQL, it is well designed and suits the data model of Neo4j well, in other words, although it may take some additional time to learn Cypher, it is still worth the time since it requires fewer lines of code to perform some intuitive functions.

Third, it is undoubtedly more agile. No matter what kind of new data needs to be added to the database, if it is not a node it is surely an edge. The only part that might need the user to pay a little more attention is the attribute of either the edge or node. In contrast to RDBMS, a new table is carried out when data needed to be added, in addition, the connections, that is the primary key and foreign key relation would have to be also put into consideration. Last but not the least, under a certain level of size of the database, thanks to its special data storage structure and specially optimized graph algorithms, the processing speed is not going to decrease significantly.

Other than all the advantages of Neo4j, some disadvantages would need to be considered upon taking action. First, it can only be stored locally, no distributed system is supported, therefore, a huge infrastructure with a high level of storing ability would be required. Second, due to the design of its data model, it is especially hard to scale horizontally like any other NoSQL DBMS, it is capable to deal with scaling, but a limit exists. The amount of data nowadays might not yet reach the upper ceiling of it, but if precautions can be made to reduce the harm when the problem comes, it is always recommended to do so. Third, the input might only be an easy task at the start of the database, thus, a whole migration plan should be set up perfectly before acting, otherwise, some extra effort might need to be put in after the initialization of the database. Lastly, the database itself cannot be protected virtually, no authentication is needed when utilizing the data in it, which brings out the importance of physical protection of the data like building your data warehouse just like Walmart did or using other third-party software to help with the encrypting.

Among all the pros and cons being discussed above, I think any other e-retailers or even retailers who have enough data requirements and might meet the migration checkboxes mentioned previously, after reviewing the probable threat of the technology which I believe could be solved in ways like how eBay and Walmart did, it is strongly recommended to think of Neo4j and give it a chance to bring incredible efficiency to companies' performance and enhance their service to the customers.

References

[1] "How Big Data Analysis helped increase Walmarts Sales turnover?" ProjectPro.io.
https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109#toc-1 (accessed Oct. 7, 2022).

[2] Shumba, Rose, "Exploring the Use of Graph Databases to Catalog Artifacts for Client Forensics" (2018).
Annual ADFSL Conference on Digital Forensics, Security and Law. 5.
https://commons.erau.edu/adfsl/2018/presentations/5

[3] "What is eBay" ecommercenect.org.
https://www.ecommercenext.org/kb/what-is-ebay/ (accessed Oct. 1, 2022).

[4] Kamille Nixon, "Sustainable Competitive Advantage: Creating Business Value through Data Relationships" Neo4j.com.
https://go.neo4j.com/rs/710-RRC-335/images/wp_sca_neo4j.pdf (accessed Sep. 30, 2022).

[5] Jessica Thiele, "Batch vs. Real Time Data Processing: What's The Difference?" vlomni.com.
https://vlomni.com/real-time-vs-batch-data-integration/ (accessed Sep. 29, 2022).

[6] Laura Shiff, "Real Time vs Batch Processing vs Stream Processing" bmc.com.
https://www.bmc.com/blogs/batch-processing-stream-processing-real-time/ (accessed Sep. 29, 2022).

[7] Dave Packer, "How Walmart Uses Neo4j for Retail Competitive Advantage" Neo4j.com.

[8] https://neo4j.com/blog/walmart-neo4j-competitive-advantage/#:~:text=Walmart%20is%20now%20using%20Neo4j,IT%20team%20based%20in%20Brazil. (Accessed Oct. 2, 2022).

[9] A. Munoz-Arcentales, A. Montoya, M. Chalen and W. Velásquez, "Improve customer experience based on recommendation and detection of a pattern change in eating habits," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018, pp. 221-225, doi: 10.1109/CCWC.2018.8301622.

[10] "Case Study: eBay Now Tackles eCommerce Delivery Service Routing with Neo4j" Neotechonology.com.
https://dist.neo4j.com/wp-content/uploads/Neo4j_CS_eBay.pdf (accessed Sep. 22, 2022).

[11] Venkata Sai Raja Bharath Vadlamannati Lakshmi, "Application of Active Rules on Graph Database," M.S. project, Dept. CS., California State University., Sacramento, California, United States, 2018. Available: https://csu-csus.esploro.exlibrisgroup.com/esploro/outputs/graduate/Application-of-active-rules-on-graph/99257831048201671

[12] "What is a graph database" Neo4j.com.
https://neo4j.com/docs/getting-started/current/get-started-with-neo4j/graph-database/ (accessed Sep. 18, 2022).

[13] "Why graph database, why Neo4j?" cnblogs.com.
https://www.cnblogs.com/rubinorth/p/5846140.html (accessed Oct. 5, 2022).

[14] "Cypher Query Language" Neo4j.com.
https://neo4j.com/developer/cypher/ (accessed Oct. 4, 2022).

[15] Hao Shen, "Data Mining: the ideology of graph database, Neo4j" Sohu.com.

[16] https://www.sohu.com/a/358283815_715776 (accessed Oct. 5, 2022).

[17] Siddhartha Duggirala, Chapter Two - NewSQL Databases and Scalable In-Memory Analytics, Editor(s): Pethuru Raj, Ganesh Chandra Deka, Advances in Computers, Elsevier, Volume 109, 2018, Pages 49-76, ISSN 0065-2458, ISBN 9780128137864,
https://doi.org/10.1016/bs.adcom.2018.01.004.
(https://www.sciencedirect.com/science/article/pii/S0065245818300135)