Ceyhun Ozgur Journal of Engineering Research and Application *www.ijera.com*
ISSN : 2248-9622, Vol. 8, Issue 7 (Part -II) July 2018, pp 27-32

RESEARCH ARTICLE OPEN ACCESS

A Comparative Study of Network Modeling Using A Relational Database(E.G. Oracle, Mysql, SQL Server) Vs. Neo4j

Ceyhun Ozgur,Jeffrey Coto,David Booth
Valparaiso University

Date of Submission: 21-06-2018 Date of acceptance: 09-07-2018

I. INTRODUCTION

According to Mansaf andKashish (2016), big data growth is very high. It is very difficult to manage due to various characteristics, which is why Hadoop and neo4j are both critical to using large datasets and applying knowledge from them to real-world scenarios. Hadoop was developed from Google's techniques in analyzing large datasets whereas neo4j is graphical software. There are two basic functions when using Hadoop. These functions are how the program stores files and how the program processes data. Hadoop is capable of storing files larger than the disk space available on a single node in a network or server and is capable of storing multiple large files at once.

Shifting focus away from the technical language of Hadoop's usage to its applicability to modern business problems and interests can make discussions of the software more accessible to those without data-science backgrounds. Parms (2015) states in his article that businesses need not be aware of the finer points of Hadoop's operating protocols, only that they must understand how data are stored, removed, and processed upon Hadoop's installation and use.

Hadoop is an open source program package, meaning that the code of the program is available to all for modifications with little to no restrictions. Thus, anyone can use the program as a platform upon which to build his or her operations. In addition to anyone being able to access Hadoop, its open source nature allows for high degrees of personal modifications to best suit specific datasets. The four most important modules of Hadoop are the Distributed File-System, MapReduce, Hadoop Common, and YARN(Dhyani, B. &Barthwal, A. 2014). All of these modules are necessary for a basic understanding of the program and smooth application thereof. According to Zujie, Jian, Weisong, Xianghua, and Min (2014), the inefficiency of Hadoop fair scheduler for handling small jobs motivates us to design the Fair4S, which introduces pool weights and extends job priorities to guarantee the rapid responses for small jobs.

Distributed File-System

The Distributed File System is what makes storing and accessing data from multiple locations, as well as linking said locations, work smoothly. Normally, the individual node would determine a file system, as they are typically part of a computer's OS. Hadoop, however, uses its own file system instead of the one on the node, which allows access to data from a computer using any supported OS.

MapReduce

MapReduce provides the basic data examination tools for Hadoop. It is named after the key functions it performs: reading the data from the database, putting it into a format that can be analyzed (called a map) and carrying out mathematical operations, such as sum or mean, within the dataset (reduce).

Hadoop Common

The third module is Hadoop Common. This provides the tools via Java needed for the user's OS to read the data stored in the Hadoop file system.

YARN

The fourth module is YARN, which manages the resources of the particular nodes involved in storing and analyzing data. This allows for diagnostics to be run on nodes with failures or errors as well as monitoring overall resource usage while working with large data sets.

Hadoop Inauguration

Hadoop, released in 2005 by the non-profit organization Apache Software Foundation, began development when software engineers realized that datasets necessary for smooth business operations were larger than what could practically be manipulated on a single storage device, such as a node or external hard drive. This is partially due to the fact that larger storage devices, while able to handle such datasets, are also less time efficient. The information being read out from the data takes much longer in larger devices to reach specific segments than in standard commercial hardware.

Instead, using multiple devices in a parallel fashion gives efficiency without sacrificing data storage options.

The Usage of Hadoop

Hadoop's lack of rigidity in terms of both software and hardware usages means that companies are able to modify or create systems as necessary, saving money in the long run by allowing choice of vendor. It has become the most widely used analytics software for nodes that are not specially created for large data processing. This makes it popular among businesses, as it is practical as well as time and cost efficient (Dhyani, B. &Barthwal, A. 2014). Most large online presences use Hadoop, as anyone is able to download and modify the software (Dhyani, B. &Barthwal, A. 2014).These modifications made by either individuals or businesses are often shared with Hadoop's development team (Dhyani, B. &Barthwal, A. 2014),allowing the product to be improved upon given real-world scenarios and concerns. This collaboration in development is critical not only to the maintenance of open-source software, but also to paving the way for innovative changes in big data analytics.

Even in its raw state and without the complicated data analysis tools, Hadoop itself can be incredibly complex and challenging to work with (Dhyani, B. &Barthwal, A. 2014). This is why Apache, such as Cloudera developed several commercial versions of the software (Dhyani, B. &Barthwal, A. 2014). These software packages make the entire process of utilizing Hadoop simpler, from installation to usage and troubleshooting. These commercial packages also include training and support to streamline the usage process (Dhyani, B. &Barthwal, A. 2014).

Companies are free to expand their usage of Hadoop when they strategically expand. This includes adding new software as well as reimagining the use of old packages. The support garnered from the analytics community as well as from those who use Hadoop for their own personal needs, has led to the software being accessible for everyone.

You might also want to read:

- Big Data: What is Spark - An Explanation For Anyone
- Spark Or Hadoop - Which Is The Best Big Data Framework?
- Big Data: R Explained in less than 2 Minutes, to Absolutely Anyone
- How is Big Data Used In Practice? 10 Use Cases Everyone Must Read

Neo4j

Despite the abundant applications of relational database management systems, the trend ofusage has moved towards graph databases such as Neo4j. This goes beyond the needs of the dataanalytics community. Graph databases are able to perform operations on more complex data andrelationships, which are not necessarily removable from the original data records. One of thedesirable applications of Neo4j is its ability to do symmetric processes in a number of systemssimultaneously. Adding Neo4j to the repertoire of products used by a company will combineextant Oracle relational database management systems with the computational power of a graphdatabase to uncover previously unexplored relationships, reduce time to market, and manipulatethe data in a more time-efficient manner.Neo4j is a highly scalable native graph database thatleverages data relationships as first-class entities, helping enterprises build intelligentapplications to meet today's evolving data challenges.This article explores the differencesbetween relational databases, graph databases and data models. In addition, it explains how tointegrate graph databases with relational databases and how to import data from a relational store.The demand for engineers with Neo4j skills is growing tremendously. Tomorrow's jobs require NoSQL and graph database skills and these technical skills are needed for career advancement. Whether evaluating a graph database for a enterprise application, growing a startup business or even working on an afternoon project, one would not need to fully download and install the application to determine whether it is the right fit.

Neo4j has the largest and most vibrant community of graph database enthusiasts that contributes to the Neo4j ecosystem, as seen below.

- 1,000,000+ downloads, adding 50,000 downloads per month
- 20,000+ graph education registrants
- 20,000+ meet-up members
- 500+ Neo4j events per year
- 100+ technology and service partners
- 200 enterprise subscription customers, including 50+ of the Global 2000

Neo4j delivers lightning-fast read and write performances, while still protecting data integrity.It is the only enterprise-strength graph database that combines native graph storage, scalable architecture optimized for speed, and ACID compliance to ensure predictability of relationship-based queries. Neo4j makes it easier to load data into with the following capabilities:

- Staggering loading speed of huge data sizes, with very low memory footprint
- Choose how much and which data to import, without worrying about volume

Whiteboard-friendly Data Modeling to Simplify the Development Cycle

- The logical model is the physical model
- 1/10 the time-to-production by closing the gap between the business and IT
- Make changes on-the-fly as business requirements change

Neo4j has been hardened through years of production deployments and rigorous ongoing testing, so you can trust it. Plus, you can engage with the graph experts providing world-class support, at the right level for your organization. Neo4j is the only graph database recognized by key analysts such as Forrester and Gartner(Agricola, 2016)to have enough production applications to warrant inclusion in reports. Clustering and data replication demanded by transactional and operational applications.

Whiteboard-friendly Data Modeling to Simplify the Development Cycle

- The logical model is the physical model
- 1/10 the time-to-production by closing the gap between the business and IT
- Make changes on-the-fly as business requirements change

Most Neo4j customers find their total cost of ownership decreasesthrough optimization of the production environment and increased efficiency. With Neo4j, one can choose the license and bundle package that needed, and add clustering and data replication capabilities that make sense for your deployment and your organization.Building a graph of your data is fairly simple as the graph structure represents the real world much better than columns and rows of data. GraphGists are teaching tools, which explore how data in a particular domain would be modeled as a graph, and see some example queries of that graph data. Any developer can create a GraphGist by visiting portal.graphgist.org.

Neo4j, the most utilized database, aids businesses by both creating products and services as well as through innovative use of extant tools. Neo4j has the advantage of providing the opportunity to build new applications, developing new techniques to be used in existing applications, committing resources to decrease the development time of new and more complex applications, and lowering the price of its tools in order to maintain affordability as compared to competitors' database services.

Customers have enjoyed recommendations for modified continued use of Neo4j's products as well as the speed at which they are able to make comments and receive information and recommendation about the latest consumer preferences and business trends. Neo4j is orders of magnitude quicker than MySQL, requiring one-tenth to one-one hundredth fewer lines of code while being versatile enough to add new functions into the software, according to Volker Pacher, Senior Developer(Dayaratna, 2014). The software is able to utilize these functions and this efficiency by making data relationships the foremost factor in its operation, making it a relational database. All of these factors combine to make Neo4j optimal when processing a variety inputs and file sizes while maintaining data integrity, regardless of whether the file is locally or globally created. Neo4j is capable of supporting large files in a multitude of languages and has support options to match the varied uses of the program.

Oracle's innovations to its cloud-based service continue to afford businesses competitive opportunities while keeping usability at the core of its design. Advances in data management make the transfer of data files of all sizes to the cloud easily and securely. The cloud's ability to scale to data files as well as to rapidly analyze them helps provide faster delivery of results, greater time for innovation, and saves not only time but money. Even if your use of the cloud is development-based rather than geared towards storage, you will have instant access to high-quality database options. Oracle Database Cloud services generate profitability and value by utilizing the hybrid development strategy to allow clients to have the same experience wherever they go.

SQL Server is a relational database management system (RDBMS) created by Microsoft. It has all the standard features of such software while still maintaining a competitive edge against Oracle Database (DB) and MySQL. Like other RDBMS programs, SQL Server supports ANSI SQL, the standard language.

Hadoop's Apache program allows multiple computers to work on large data sets at the same time via programming models. It can scale from one server up to potentially thousands of machines, offering local computation and storage at each individual computer. The library is designed to address issues in the application rather than using hardware, to deliver a highly reliable service.

Maintaining a competitive edge during this time of increased demands and diverse software is challenging. Several functions and applications are coming together to create a new landscape of analytics. Some of these factors include diverse data or large amounts of data coming in at once, whether generated automatically or manually, and the majority of this data is not structured. This means that current software or applications are put under a large burden when attempting to analyze the rapid-fire intake of complex data. As companies attempt to find a

solution to these problems, Hadoop's file system (HDFS) is able to meet their demands. It is increasingly becoming more affordable and is capable of processing low levels of structure in data, making its services a key contender in today's analytics field. Using advanced analytics, HDFS is able to find insights even within massive data files; its algorithms have developed to learn and predict trends within the data, aiding in computations for predictive analytics with large data sets.Text analytics is an up-and-coming method of explaining the data, which may decrease interpretation errors as well as a computer failure network in unstructured data clusters. Analytics within memory and databases have decreased computation times, helping businesses stay competitive as they analyze large files.

As enterprises look to embrace big data and Hadoop, they have numerous questions: "How can I deal with data preparation on Hadoop?" "How does utilizing Hadoop impact visualization and other kinds of analysis?" "What kind of analytical techniques are available to analyze Hadoop data?" "How do I use Hadoop with in-memory processing?" (Halper, 2014).

Using in-memory analytics sets runs the calculations on RAM rather than the disk, which avoids the time-consuming I/O concerns. This is a huge advantage when working with large data sets as in-memory processing can be orders of magnitude faster than accessing the data from the disk, which benefits companies and consumers by cutting down on iterations. In-memory processing can also be distributed, which means it can handle the multi-pass-through data and iterative workloads without getting bogged down, all while allowing communication among independent units to take advantage of parallel processing.

Advanced analytical techniques such as data or text mining and machine learning can benefit tremendously from in-memory processing, especially by reducing the time spent on analytics. This means that the statisticians are free to fine-tune models or explore new approaches without worrying about the potential loss of time, leading to innovative designs and ultimately increased productivity. For example, a single iteration for a predictive model can be reduced from hours to minutes, meaning more and better models can be built and utilized to examine data, giving a competitive edge.

Once data has been stored in memory, it can be accessed more efficiently. If someone is able to quickly build a model, they are also able to share and test the model with others at the same rapid pace. The model can be adapted as suggestions arise; creating a better and faster iterative process that yields an accurate model and benefits businesses using Hadoop. These benefits are conferred whether a business relies on Hadoop for memory capabilities, processing, or single pass analytics or not.Begin with the Single Node Setup, which shows you how to set up a single-node Hadoop installation. Then move on to the Cluster Setup to learn how to set up a multi-node Hadoop installation.

Ambari™

A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.

HadoopDB began as a research effort in 2008 to transform Hadoop --- a batch-oriented scalable system designed for processing unstructured data --- into a full-fledged parallel database system that can achieve real-time (interactive) query responses across both structured and unstructured data. In 2010 it was commercialized by Hadapt, a start-up that was formed to accelerate the engineering of the HadoopDB ideas, and to harden codebase for deployment in real-world, mission-critical applications.

HadoopDB

HadoopDBcombines ideas from the Hadoop and database system communities even though research challenges have emerged as HadoopDBdeployed in the real world. Many of these challenges involve loading data into structured storage. Although this loading of data can greatly accelerate query execution times, the upfront cost of this load is antithetical to the Hadoop premise, which is data does not need to be organized, cleaned, and pre-processed before being available for query processing. Therefore, there are two approaches to reducing these costs: (1) an invisible loading technique where data is incrementally loaded into structured storage over time, based on users' patterns of data access and (2) a queue-based locality scheduling technique that, when data had been loaded in a heterogeneous manner across the nodes in a cluster, improves upon Hadoop's greedy scheduler and more efficiently assigns tasks to nodes that have the data stored locally.

A fundamental task in data integration and data exchange is the design of schema mappings, that is, high-level declarative specifications of the

relationship between two database schemas. Several research prototypes and commercial systems have been developed to facilitate schema-mapping design; a common characteristic of these systems is that they produce a schema mapping based on attribute correspondences across schemas solicited from the user via a visual interface. This methodology, however, suffers from certain shortcomings. In the past few years, a fundamentally different methodology to designing and understanding schema mappings has emerged. This new methodology is based on the systematic use of data examples to derive, illustrate, and refine schema mappings. Example-driven schema-mapping design is currently an active area of research in which several different approaches towards using data examples in schema-mapping design have been explored. After a brief overview of the earlier methodology, this tutorial will provide a comprehensive overview of the different ways in which data examples can be used in schema-mapping design. In particular, it will cover the basic concepts, technical results, and prototype systems that have been developed in the past few years, as well as open problems and directions for further research in this area.

Trust and Reputation have become key enablers of positive interaction experiences on the Web. These systems accumulate information regarding activities of people or peers in general, to infer their reputation in some context or within a virtual community. Reputation information improves the quality of interactions between peers and reduces the effect of fraudulent members. In this tutorial we motivate the use of trust and reputation systems and survey some of the important models introduced in the past decade. Among these models, we present our work on the knot model, which deals with communities of strangers. Special attention is given to the way existing models tackle attempts to attack reputation systems. In a dynamic world, a person or a service may be a member of multiple communities and valuable information can be gained by sharing reputation of members among communities. In the second part of the tutorial, we present the CCR model for sharing reputation across virtual communities and address major privacy concerns related to it. In the third part of our talk, we discuss the use of reputation systems in other contexts, such as domain reputation for fighting malware, and outline our research directions on this subject.

High performance data analysis is a required competitive component, providing valuable insight into the behavior of customers, market trends, scientific data, business partners, and internal users. Explosive growth in the amount of data businesses must track has challenged legacy database platforms. New unstructured, text-centric, data sources, such as feeds from Facebook and Twitter do not fit into the structured data model. These unstructured datasets tend to be very big and difficult to work with. They demand distributed (aka parallelized) processing.

Hadoop, an open source software product, has emerged as the preferred solution for Big Data analytics. Because of its scalability, flexibility, and low cost, it has become the default choice for Web giants that are dealing with large-scale click stream analysis and ad targeting scenarios. For these reasons and more, many industries that have been struggling with the limitations of traditional database platforms are now deploying Hadoop solutions in their data centers. These industries are also looking for the economy. According to some recent research from Infineta Systems, a WAN optimization startup, traditional data storage costs $5 per gigabyte, but storing the same data costs roughly 25 cents per gigabyte using Hadoop.

II. CONCLUSION

This paper reviewed two of the common modern information methods that are capable of handling big data. The first is Neo4j, which is graph based data modeling software and the second is Hadoop, which goes beyond the analysis of big data.Connected information is becoming more and more common as single-user data grows in complexity and volume. Neo4j was built to efficiently store, handle, and query highly connected elements in such datasets. With powerful and flexible data models, the real-world scenarios are fairly represented without a loss of richness. The property graph model is easy to understand and manipulate, especially for object-oriented and relational database developers.Hadoop developed from a necessity to handle the explosion of Internet-originate data and grew beyond the capabilities of commercially available systems to handle its analyses. It was initially inspired by papers published by Google, which outlined its approach to handling massive data input, and has since become the de facto choice for storing and analyzing up to petabytes of data.

REFERENCES

[1]. Agricola, A. (2016, January 20). Graph database leader Neo Technology declares record2015.Neo4jNews.Retrievedfrom:https ://neo4j.com/news/graph-database-leader-2015/
[2]. Baoan L. (2014). Knowledge management based on big data processing. Informational Technology Journal 13(7). 1415-1418
[3]. Dayaratna, A. (2014, March 23). Neo4j adopted by retail giants eBay and Walmart

for real time, e-commerce analytics. Cloud ComputingToday.Retrievedfromhttps://cloudcomputingtoday.com/2014/03/23/1068051/

[4]. Dhyani, B. &Barthwal, A. (2014). Big data analyticsusing Hadoop. International Journal of Computer Applications 108(12). 0975-8887

[5]. Strang, K.D., &Sun, Z. (2016). Analyzing relationships in terrorism big data using hadoop and statistics. The Journal of Computer Information Systems. doi: 10.1080/08874417.2016.1181497

[6]. Halper,F.(2014,March). Eight considerations for utilizing big data analytics with Hadoop.TDWI Checklist. Retrieved from: https://www.sas.com/en_gb/offers/14q2/bigdata_hadoop/register.html

[7]. Hoffman, B.L. (2015, August 20). What's the difference between Apache Hadoop and Apache Spark? Big data and analytics on IBM power systems.[Web log comment]. Retrievedfromhttps://www.ibm.com/developerworks/community/blogs/f0f3cd83-63c2-474490219ff31e7004a9/entry/What_s_the_Difference_Between_Apache_Hadoop_and_Apache_Spark?lang=en

[8]. Mansaf,A.&Kashish,A.S.(2016, September). Big data analytics in cloud environment usingHadoop.Retrievedfromhttps://arxiv.org/abs/1610.04572

[9]. Neo4j. (2017 February 23). Neo4j: The world's leading graph database. Retrieved from:https://neo4j.com/product/

[10]. Zujie,R. Jian, W., Weisong, S., Xianghua, X., & Min, Z. (2014).Workload Analysis, Implications, and Optimization on a Production Hadoop Cluster: A Case Study on Taobao." IEEE Transactions on Services Computing 7(2), 307-321.

Ceyhun Ozgur "A Comparative Study of Network Modeling Using A Relational Database(E.G. Oracle, Mysql, SQL Server) Vs. Neo4j"International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.7, 2018, pp.27-32