**Title: Understanding Big Data and NoSQL**

**Final Project Overview**

Lisa Frazier

BIAM530 – Week 8

December 11, 2023

**Introduction:**

NoSQL databases have come forth as important modern technologies for IT. In an enterprise landscape awash in data, non-relational databases enable better insights and sharper analysis. In this paper you will get an overview of the types of data issues that businesses face today. One of the major differences between SQL and NoSQL DBs, storing unstructured data. You will learn that you need to apply some structure when it comes to using the data.

Today, an additional data factual business mutiny to new attained of social media, mobile, and sensing element or machine-generated data, hold the potential to reattach an organization. With that said, companies are established to perceive that Big Data is moreover about business variation than IT transformation. Hence, Big Data is answering questions they could not previously answer and make opportune decisions at exceptional level of constancy than before which in turn, capitulate new discernment that can provide businesses distinction and recently discovered functioning planning.

The big new data place of origin added to newly discovered modern logical capabilities, excessive allegiance answers to these questions, and prepare an increased absolute understanding of customers outcome, and performance that can propel business consequences across various business ramification, such as the questions below:

What is NoSQL database, specifically Mongo DB and Neo4J?, Hadoop Vs. Spark, Hadoop DB/RoboMongo recommendations and alternatives, how do Big Data storage and Management work and what are the pros, and cons and finally, what would I recommend using MongoDB or Neo4J.

**Step 1. NoSQL | MongoDB | Neo4J Database:**

Big Data is defined as the large volume, variety, and velocity of data that exceeds the capabilities of traditional data processing methods for documents. The three V’s consist of a blend of written work, illustration, videos, and auditory attribution integrated into intricate formation generated by original provocations for data operations that necessitate the entire attributes; quantity of data stored, momentum at which details are entered the network and the diversity in the structures of the data to be stored.

Hence, NoSQL is considered as a non-traditional database technology that develop to address challenges pertaining specifically to documents oriented databases, the term "NoSQL" encompasses a diverse set of database technologies, each designed for specific data models and use cases. MongoDB is a modern and versatile approach in handling data particularly when it comes to scalability, flexibility, and performance attribute. Neo4J’s strength is its ability to depict complex query relationships mostly for graphs.

**Step 2. Hadoop Vs. Spark:**

The significance of Big Data in uncovering patterns, trends, and insights that drive decision-making and innovation across various industries is **Hadoop** because it is a Map-Reduce framework, which is acceptable for diverse Big Data. The term, “HDPS, Hadoop Distributed File System, “supply repository to contribute to the process. For example, imputing data and median outcome. These characteristics are organized in structures and disperse to Map tasks. However, **Spark** has surprisingly progressive “Map Reduce Framework” that is designed to overcome issues exhibited by Hadoop for example: high execution times, in algorithms that only share data through a distributed file system (HDFS), and

**Step 3. Hadoop and MongoDB/RoboMongo:**

**Would I recommend it?**

Yes, I would recommend both because each one has its own advantage as well as challenges. **Hadoop** is designed for its scalability, batch processing, cost-effective storage, diverse data types and community ecosystems, dealing with large and growing datasets. On the other hand, **MongoDB** has flexible data modeling, horizontal scaling, real-time access, and development speed. So, depending on what type of project you are working on would determine the best suited system.

**Would I recommend an alternative?**

I have not had extensive training in either of the databases above, and certainly no experience with other systems. However, according to Professor Sid Amend and classmates on Webex, the grafts are better in Hadoop, and no other system was mentioned to be better than or worse.

While using Adventure works, I did not notice much about the disadvantages, I was concerned only that I can accomplish the same results. If I have to choose one, I will go with Hadoop.

**Step 4. Big Data Storage and Management Solutions:**

**Pros:**

1. **Data storage:**

Storing and maintaining substantial volumes of data can be pricey since it requires equipment, computer programs, preservation, and data merge replication for relieving faults. Condense algorithms have prominent compaction rate that can aid to save storage space.

1. **Data processing:**

Big data demands individualized implements and procedures, such as dispersed extraction methods for managing the mass of data in an appropriate demeanor. Also, including cleaning, organizing, aggregating, and analyzing data to extract insights and support decision making.

1. **Data integration:**

A collection of data is customarily generated by consolidating dissimilar incidents and documentation. As they advance from a variety of origin, it is tedious to analyze them. Furthermore, they contain many inconsistencies that need to be normalized along the pipeline. This requires data purification, modification and adjustment techniques that can assure the stability of the dataset configuration.

**Cons:**

1. **Cost of the two Big Data Solutions:**

More equipment does not warrant a prolonged remedy for ameliorate execution, as the quantity and fitting together of the data may eventually oppress the system. Hardware is costly and is nowadays undergoing a deceleration in progression, generation, and correcting of the blueprint of our data channel that is consequently fundamental.

1. **Data security and privacy:**

In recent years, simultaneously, consumers and organizations are progressively troubled about the privacy and security of their data. In addition to the obvious need of securing and elucidating data, the new ordinance requires companies to store sensitive user information as close to their location as possible.

1. **Infrastructure faults:**

Hardware systems can be dependable only for an unquestionable amount of time before they will have to be upgraded if they no longer perform at the desired level or break down and need to be replaced. In addition to the inability to deliver their tasks in that preservation period, companies cannot afford to permanently lose that data in case of failure.

**Cost**

Because of the issues above, companies usually depend on two or three more duplicate clusters that can control traffic in case the main cluster fails. Although this plan of action can help with load distribution, the cost will increase both for maintenance and running expenses, since all the data must be duplicated into two or three extra clusters. Sample pricing below.**A screenshot of a computer

Description automatically generated**

[**https://www.mongodb.com/pricing**](https://www.mongodb.com/pricing)

**Step 5. Comparative Analysis for MongoDB and Neo4J:**

Challenges related to storage, processing, analysis, and visualization of massive datasets. MongoDB is well-suited for various use cases, while Neo4j excels in scenarios where relationships are central to the data structure of graphs.

**Below is a breakdown Table of MongoDB Vs. Neo4j**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Step 6. MongoDB Vs. Neo4J: Recommendation:**

Recommendations related to this project aims at building a robust and reliable system suitable for organizations data that have become complicated and time consuming as data amount increases enormously every second, every day. So, I would recommend that a great strategy would be to have an automated fall over, proactive monitoring, and regular testing of documentation and security of the infrastructure. Hence, MongoDB can execute the requirements that are needed for documentation, whereas Neo4J are tailored to graph. Depending on

**Conclusion:**

The key points about Big Data and NoSQL are emphasized by their transformative potential in the data-driven era and traditional data processing systems, which are the three V’s. There is transformative potential such as Hadoop and Spark processing of a mass amount of data is by distributing data information across computers. Data is tedious to extract, store, and process, but also extremely voluminous in that they overpower current technologies and challenge us to create new techniques for managing them.

Consequently, big data analytics can make timely improved performances, and informed decisions to say the least. Therefore, empowering organizations to predict patterns, trends, and future outcomes. Meaning, regular updating and maintenance, implementing regularly scheduled software updates and staying current with all laws and policies by staying in compliance with state rules

**References:**

Understanding Big Data and NoSQL Databases. (2016, May 09). *SQL Server Pro,*<https://devry.idm.oclc.org/login?url=https://www.proquest.com/magazines/understanding-big-data-nosql-databases/docview/1795436432/se-2>

Pelucchi, M., Psaila, G., & Toccu, M. (2018). Hadoop vs. Spark: Impact on Performance of the Hammer Query Engine for Open Data Corpora.*Algorithms, 11*(12), 209. <https://doi.org/10.3390/a11120209>

<https://ng.cengage.com/static/nb/ui/evo/index.html?deploymentId=5835902453346576351460624129&eISBN=9781337627924&id=1921067817&snapshotId=3691321&>

Băicoianu, A., & Scheianu, I. V. (2023). Managing and Optimizing Big Data Workloads for On-Demand User Centric Reports.*Big Data and Cognitive Computing, 7*(2), 78. <https://doi.org/10.3390/bdcc7020078>

Nagireddy, V. (2023). *Job Recommendation System With NoSQL Databases: Neo4j, MongoDB, DynamoDB, Cassandra and Their Critical Comparison*(Order No. 30426689). Available from ProQuest One Academic. (2827919287). <https://devry.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/job-recommendation-system-with-nosql-databases/docview/2827919287/se-2>

DeVry University Cengage-MindTap / Book: Developing and Managing Database for Business / Module 14.1, 2, 3 <https://ng.cengage.com/static/nb/ui/evo/index.html?deploymentId=5835902453346576351460624129&eISBN=9781337627924&id=1921067817&snapshotId=3691321&>