Examination of NoSQL Transition and Data Mining capabilities

Dimitrios Rousidis[1][0000-0003-0632-9731], Paraskevas Koukaras[1][0000-0002-1183-9878], and Christos Tjortjis*[1][0000-0001-8263-9024]

[1]The Data Mining and Analytics research group, School of Science and Technology, International Hellenic University
GR-570 01 Thermi, Thessaloniki, Greece
{d.rousidis, p.koukaras, c.tjortjis}@ihu.edu.gr

Abstract. An estimated 2.5 quintillion bytes of data are created every day. This data explosion, along with new datatypes, objects, and the wide usage of social media networks, with an estimated 3.8 billion users worldwide, make the exploitation and manipulation of data by relational databases, cumbersome and problematic. NoSQL databases introduce new capabilities aiming at improving the functionalities offered by traditional SQL DBMS. This paper elaborates on ongoing research regarding NoSQL, focusing on the background behind their development, their basic characteristics, their categorization and the noticeable increase in popularity. Functional advantages and data mining capabilities that come with the usage of graph databases are also presented. Common data mining tasks with graphs are presented, facilitating implementation, as well as efficiency. The aim is to highlight concepts necessary for incorporating data mining techniques and graph database functionalities, eventually proposing an analytical framework offering a plethora of domain specific analytics. For example, a virus outbreak analytics framework allowing health and government officials to make appropriate decisions.

Keywords: NoSQL, graph databases, Machine Learning (ML), Data Mining (DM).

1 Introduction

On March 6th, 1972 when the paper titled “Relational Completeness of Data Base Sublanguages” by E.F. Codd, was published few could expect its impact. The first paragraph of its abstract “*This paper attempts to provide a theoretical basis which may be used to determine how complete a selection capability is provided in a proposed data sublanguage independently of any host language in which the sublanguage may be embedded*” proved prophetic [1]. Since the late 70s with the launch of Oracle, the first commercially available Relational Database Management System (RDBMS) the Relational Model (RD) dominated software applications. Since then, nearly everything changed in IT, while nowadays the world is living the Big Data era. According to the ‘TechJury’ website, in 2020 every person will generate 1.7 megabytes in just a second, whilst Internet users generate about 2.5 quintillion bytes (2,5 Exabytes) of data each day [2]. Social Media (SM) and Cloud Computing skyrocketed the volume of data and forced the IT industry to search for alternative databases (DB) and Database Management Systems (DBMS). Along with the data, there is a simultaneous growth to the structured, encoded set of data than describes and aids to the discovery, management, assessment of the described entities, the metadata. However, metadata in DBMS are handled today in ad-hoc ways [19]. Therefore, NoSQL, a term used by C. Strozzi in 1998 [3], introduced a mechanism that enhances the functionalities of the typical tabular-based RDBMS for all simple and complex data and metadata.

The goal of this paper is to elaborate on benefits of NoSQL DBs as well as the opportunities and new possibilities by combining Machine Learning (ML) methods and supplying practitioners and researchers with enough arguments for the necessity of NoSQL DBs. Therefore, what types can be utilized based on various occasions and what are the Data Mining (DM) tasks that can be performed.

The rest of the paper is structured as follows: Section 2 presents the characteristics, and the categorization of NoSQL DBs. Section 3 provides an analysis of the most common DM tasks with the utilization of graph DBs. The last section introduces an ongoing research project that utilizes Graph NoSQL DB for the development of a virus outbreak decision making framework.

2 NoSQL Databases

2.1 Characteristics of NoSQL

NoSQL is based on the BASE (Basically Available, Soft State and Eventually Consistent) model in contrast to the ACID (Atomicity, Consistency, Isolation, Durability) model. Its main advantage is the ease of storing, handling and manipulating, providing access to huge volumes of data, becoming ideal for data intensive applications [4].

The main characteristics of NoSQL DBs are: 1. Non-relational: not fully supporting relational DB features, such as joins, 2. Schema-less (lacking a fixed data structure), 3. fault-tolerance as data are duplicated to multiple nodes, 4. Horizontally scalable (connecting multiple hardware or software entities to work as a single logical unit), 5. Open source (they are cheap and easy to implement), 6. Massive write-read-remove-get performance, 7. Strong Consistency (all users see the same data), 8. High Availability (all users have access to at least one copy of the requested data) and 9. Partition-Tolerance (the total system keeps its characteristics even when deployed on different servers).

According to [5] the main use of NoSQL in industry is: 1. Session Store (managing session data), 2. User Profile Store (enabling online transactions and user-friendly environments), 3. Content and Metadata Store (building a data and metadata warehouse and storage of multitype data), 4. Mobile Applications, 5. Internet of Things (aiding the concurrent expansion, access and manipulation of data from billions of devices), 6. Third-Party Aggregation (with the ease of managing huge amounts of data, with access by third-party organizations), 7. E-commerce (storing and handling enormous volumes of data), 8. Social Gaming, 9. Ad-Targeting (enabling tracking user details quickly).

The main advantages of NoSQL DBs are: i) non-relational, ii) schema-less, iii) data are replicated to multiple nodes and can be partitioned, iii) horizontally scalable, iv) provide a wide range of data models, v) database administrators are not required, vi) less hardware failures, vii) faster, more efficient and flexible, viii) high pace evolution, ix) less time writing queries, x) less time debugging queries, xi) code is easier to read, xii) Big Data compliant (high data velocity, variety, volume, and complexity) xiii) they have huge volumes of fast changing structured, semi-structured, and unstructured data. On the other hand, there are some disadvantages i) Immaturity, ii) no standard query language, iii) Some DBs are not ACID compliant and iv) no standard interface [6].

2.2 Categorization

There is a debate in the bibliography about the number of categories that NoSQL DBs can be grouped. Most sources group them in four categories, but there is a number of experts that group them in five categories which are: i) Column, ii) Document, iii) Key-value, iv) Graph and v) Multimodel, the latter being the extra one added to the four category division. In [7] authors refer to 15 categories; the five main ones (1 to 5) and 10 others, which are denoted as Soft NoSQL Systems (6 to 15): 1. Wide Column Store / Column Families, 2. Document Store, 3. Key Value / Tuple Store, 4. Graph DB, 5. Multimodel DB, 6. Object DB, 7. Grid & Cloud DB Solutions, 8. XML DB, 9. Multidimensional DB, 10. Multivalue DB, 11. Event Sourcing, 12. Time Series / Streaming DB, 13. Other NoSQL related DB, 14. Scientific and Specialized DBs, 15. Unresolved and uncategorized [8].

The five most popular categories are: (i) *Key-Value stores.* Key-Value, based on Amazon's Dynamo paper [9] are designed to handle massive associated arrays which consist of pairs of keys and values, and they also can retrieve values as long as a key is known. The most popular key value DB are Redis, Amazon DynamoDB, MS Azure Cosmos DB (considered a multi-model DB), Memcashed, and Hazelcast.

(ii) *Wide column stores.* Also called Extensible Record Stores, are based on Google's BigTable paper [10], can store data in records than can hold huge numbers of dynamic columns. Their data model consists of a collection of column families, key and value where the value is a set of related columns and they are indexed by the triple combination of row key, column key and timestamp. The most popular ones are Cassandra, HBase, MS Azure Cosmos DB, Datastax Enterprise and MS Azure Table Storage.

(iii) *Graph databases.* Data are represented by graphs, inspired by graph theory. Their data model consists of nodes and edges linked with relationships. The most popular graph DBs are Neo4j, MS Azure Cosmos DB, ArangoDB, OrientDB, and Virtuoso.

(iv) *Document stores.* In a document store, data are stored in so-called documents. The term "documents" refers to arbitrary data in a schema-free organization of data. The most popular document DBs are MongoDB, Amazon DynamoDB, Microsoft Azure Cosmos DB, Couchbase and CouchDB.

(v) *Time Series.* A Time Series Database (TSDB) is a DB optimized for time-stamped or time series data; each entry is associated with a timestamp. A TSDB is used for measuring change over time and the properties that distinguish them are data lifecycle management, summarization, and large range scans of many records. The most popular TSDBs are InfluxDB, Kdb+, Prometheus, Graphite and RRDtool.

2.3 Popularity

The rise of the popularity of NoSQL DBs can be demonstrated by their extensive development and use by IT colossi. For instance, Apache's Cassandra is used by Facebook, Reddit, Twitter, Digg and Rockspace, amongst others. Baidu is using Hypertable. Google has developed BigTable. Amazon has developed DynamoDB and LinkedIn is using Project Voldemort. According to the ranking of DB-Engines, "*an initiative to collect and present information on DMS*" which ranks DBMSs according to their popularity (updated monthly), NoSQL DB are constantly on the rise, whereas relational DBs, although still on top, remain unchanged or face minor decline. DB-engines methodology for measuring the popularity of a system is based on the following 6 parameters: 1) Number of mentions of the system on website, 2) General interest in the system, 3) Frequency of technical discussions about the system, 4) Number of job offers, in which the system is mentioned, 5) Number of profiles in professional networks, in which the system is mentioned, and 6) Relevance in social networks.

According to the DB-engines ranking (https://db-engines.com/en/ranking) for the period from November 2012 until September 2020, it is evident, that despite occupying the first four positions, the popularity of relational DB is disputed. NoSQL DBs are on the rise. MongoDB is in 5th place demonstrating a 339% increase on popularity, the most popular Key-Vale DB is Redis (7th place), Cassandra, the most popular Wide Column DB, is 10th overall, Neo4j is leading the Graph DBs in 21st place and finally InfluxDB is the most popular Time Series DB in 29th place.

3 Data Mining Tasks utilizing graphs

Complex Information Networks are an emerging field in this era of powerful complex data organizations and web-based media mining. The DM tasks linked with Heterogeneous Information Networks (HIN) ought to adapt to the new demanding requests on this field of studies. The main DM tasks utilizing graphs are being presented while being categorized as follows [11]:

Similarity is a method for discovering how similar objects are. It offers the foundations for a plethora of other DM techniques, like clustering, classification, web search etc.

Clustering is notable for carrying DM tasks that require big data objects to be fragmented and grouped into smaller clusters that share a degree of similarity, but at the same time maintaining dissimilarity from objects in neighboring clusters. Modern datatypes and objects, like networked data diverge from the 'traditional' data where clustering is based on the unique and consistent object characteristics [12].

Classification is useful when possible class marks need to be ascertained, which is attainable through a classifier or an appropriate new model. In ML, classification is carried out on indistinguishably structured objects. However, the new emerged needs of modern object types, require to also take into account their relationships (associations). Hence, a linked based object classification occurs when entities related with each other are structured in this way, forming unique graphs. Conventional strategies are regularly reused or stretched out to have the option to deal with this sort of associations [13].

Link Prediction is one of the most demanding DM tasks. It investigates whether possible connections between nodes exist, utilizing rules, such as: a) the examination of

nodes and b) hub attributes. Literature refers to link prediction by examining the structural attributes of social networks with predictors, or attribute information [11].
Ranking features are significant since they can quantify an object's significance within a social network. For instance, RankClus manages bipartite networks creating clusters of objects maintaining the equality of significance both on clustering and ranking. NetClus, is an optimal solution for star-type schema clustering, whilst other popular frameworks are HeProjI, OcdRank etc. [11].
Recommendation and related systems comprise of a wide range of algorithms from various domains. The goal is to recommend suitable services and objects to users. This can be accomplished using similarity features. In contrast to older recommendation systems that were utilizing user specific feedback information measurements, recent techniques have become more astute and functional, by utilizing collaborative filtering, matrix factorization or circle-based techniques [11].
Information Fusion is one of the main concerns that characterize HINs. The goal is to combine data from many variant HINs and improve intricacy and scrutiny of the information retrieved. Robust algorithms combine objects regardless if they belong to the same networks or they have identical semantic meaning. SM networks are brimming with this type of data, making them proper candidates for this kind of task [11].

4 Discussion – Future Work

The aim of this paper is to offer an insight about NoSQL DBs. A brief background on the reasons for their introduction and development is given. Next, their powerful characteristics and features are highlighted. Then, their many advantages over mainly RDs, especially their enhanced data and metadata management, along with their disadvantages are being grouped and analyzed [20]. The main categories with the most popular DB by category are presented, as well as the most widely used categories. Finally, according to popularity statistics, relational DBs are losing their users as four out of five demonstrate a decrease in their popularity (three of them from 10% up to 15%) and at the same time these users are transferring their trust to NoSQL DBs. This trend is also demonstrated by the fact that enterprise and IT colossi like Amazon, Apache and Google are leading their development.
This paper presents ongoing research related with the use of SM as a source for information retrieval and forecasting with the aid of DM techniques [14-17]. The next step of the project is to incorporate the use of graph DB (Neo4j) to provide a forecasting mechanism in healthcare [18]. The framework to be created will take into account more than 30 different parameters such as population characteristics (gender percentages, life expectancy, density, etc.), indexes (economic and medical, freedom of press etc.), policies applied (lockdown), including sentiment analysis related to COVID-19on data retrieved mainly from Twitter, as well as other SM platforms. The goal is to assign weights to these parameters, to provide a hands-on formula and mechanism to health and government officials, enabling them to make appropriate decisions during a pandemic.

References

[1] Codd, E. F. (1972). Relational completeness of data base sublanguages (pp. 65-98). IBM Corporation.

[2] Petrov, C. (2020, September 10). 25 Big Data Statistics - How Big It Actually Is in 2020? Retrieved August 3, 2020, from https://techjury.net/blog/big-data-statistics/.

[3] NoSQL. (2020, August 1). Retrieved August 4, 2020, from https://en.wikipedia.org/wiki/NoSQL.

[4] Moniruzzaman, A. B. M., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. arXiv preprint arXiv:1307.0191.

[5] Vaghani, R. (2018, December 17). Use of NoSQL in Industry. Retrieved August 5, 2020, from https://www.geeksforgeeks.org/use-of-nosql-in-industry.

[6] Nayak, A., Poriya, A. and Poojary, D., 2013. Type of NOSQL databases and its comparison with relational databases. *Int'l Journal of Applied Information Systems*, 5(4), pp.16-19.

[7] NoSQL Databases List by Hosting Data - Updated 2020. (2020, July 03). Retrieved August 5, 2020, from https://hostingdata.co.uk/nosql-database/.

[8] Zollmann, J. (2012). Nosql databases. Retrieved from Software Engineering Research Group: http://www. webcitation. org/6hA9zoqRd.

[9] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., & Vogels, W. (2007). Dynamo: amazon's highly available key-value store. *ACM SIGOPS operating systems review*, 41(6), 205-220.

[10] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems* (TOCS), 26(2), 1-26.

[11] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2016). A survey of heterogeneous information network analysis. *IEEE Trans. on Knowledge and Data Engineering*, 29(1), 17-37.

[12] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

[13] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[14] Koukaras, P., Tjortjis, C., & Rousidis, D. (2020). Social Media Types: introducing a data driven taxonomy. *Computing*, 102(1), 295-340.

[15] Koukaras, P., & Tjortjis, C. (2019). Social media analytics, types and methodology. In *Machine Learning Paradigms* (pp. 401-427). Springer, Cham.

[16] Rousidis, D., Koukaras, P., & Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9), 6279-6311.

[17] Koukaras, P., Berberidis C., & Tjortjis C. (2020, August). A Semi-supervised Learning Approach for Complex Information Networks. In *Proc. 3rd Int'l conf. Intelligent Data Communication Technologies and Internet of Things* (ICICI 2020), Springer Lecture Notes on Data Engineering and Communications Technologies (pp. 1-13).

[18] Koukaras, P., Rousidis, D., & Tjortjis, C. (2020). Forecasting and Prevention Mechanisms Using Social Media in Health Care. In *Advanced Computational Intelligence in Healthcare-*7 (pp. 121-137). Springer.

[19] Gupta, I., Raghavan, V., & Ghosh, M. (2015, June). Leveraging metadata in no SQL storage systems. In *2015 IEEE 8th Int'l Conf. on Cloud Computing* (pp. 57-64). IEEE.

[20] Lofstead, J., Ryan, A., & Lawson, M. (2019, June). Adventures in NoSQL for Metadata Management. In *Int'l Conf. on High Performance Computing* (pp. 227-239). Springer.