 DOI: 10.22616/rrd.24.2018.082

USE OF NOSQL TECHNOLOGY FOR ANALYSIS OF UNSTRUCTURED SPATIAL DATA

Monta Poļakova, **Gatis Vītols**
Latvia University of Life Sciences and Technologies, Latvia
monta.kivite@gmail.com; gatis.vitols@llu.lv

Abstract
Every day millions of new data records with spatial component are produced in the world, which provide valuable information to make decisions and solve business-related issues. However, a large part of this data is hardly analyzed because of their different structures and schemas. The aim of the paper is to improve the integration, processing and analysis of unstructured spatial data.
During the research, the author analysed geospatial data types and sources, explored NoSQL solutions for geospatial data processing and chose the open-source tools which are the most appropriate for the stated goals, as well as analysed the coverage of forest areas with protected zones using MongoDB database capabilities and visualized results in a map, using QGIS software.
MongoDB is a useful tool for geospatial data analysis and has a large number of embedded topology analysis functions and has drivers for widespread programming languages like JavaScript, Python, PHP, Java, Scala, C#, C, C + +, etc. QGIS has extensions that allow to make connections to databases, including a connection with MongoDB. Using these features, the developers can develop geographic information systems to analyse geospatial data – structured, semi-structured and unstructured.
Generally MongoDB is used for real-time data analysis, however, complicated analysis of large data sets can take up to hours and even days, so it is still necessary to find the best solution to get results in an acceptable time for users. Using MongoDB together with Apache Hadoop - the framework to support big data applications - could be a possible solution for this problem.
Key words: NoSQL, MongoDB, geographic information system, spatial data, database.

Introduction

A lot of spatial data is produced in business sector and also in everyday life. Not in all situations this data has an easily processable structure for relational databases. In most cases this data is unstructured or semi-structured. NoSQL solutions may provide possibilities to process this data in a more effective way and provide business sector with valuable information, which can be used to make decisions and solve issues related to business.

The amount of geospatial data created, collected and used is increasing because of observations and measurements of geo-sensor networks, satellite imageries, point clouds of laser scanning and location-based social networks. It has become a serious challenge for data management and analysis systems. Traditionally, relational database management systems (RDBMS) are used to manage and analyze geospatial data, but there are some situations, when these systems may not provide required efficiency and effectiveness. The authors (Amirian, Winstanley, & Basiri, 2013) consider that NoSQL solutions can provide the efficiency necessary for applications using geospatial data.

RDBMS requires exact schema definition for input data into database. Database schema should be changed by modifying or adding fields to it every time when the business requirements for application are changing. It can be a lengthy process. NoSQL databases are free from this restriction because nearly all these databases are schema-less. Without predefined schema, NoSQL can be used for various data types like structured data, semi-structured data and even unstructured data (Yue & Tan, 2017).

Several spatial extensions are included in multiple NoSQL databases, which enables NoSQL based systems to address the complexity of distributed data. This feature allows using NoSQL where traditional geographic information systems cannot offer highly distributed and scalable performance (de Souca Baptista *et al.*, 2014).

Latvia lacks a simple united system for everyday user where to combine, analyze, review and store various geographic data from different systems, for example, precise agricultural systems, Rural Support Service, State Forest Service and other geographic information systems (GIS) that are accessible to public. Currently in Latvia the user has an opportunity to retrieve GIS data, but has no chance to combine and search for interconnection of this data because national information systems and various other geographic information systems are not connected and most likely the user has no tools to analyze GIS data. For example, in the precision agriculture various machinery, tractors and devices equipped with sensors are used and every manufacturer has their own system for data processing, however, a simple GIS application, which would allow combining data from various systems, is missing. Problem is also that every data set can have a different structure. There is an online portal geolatvija.lv with some data from national institutions, however, there is no functionality implemented in the portal that would support users' own data uploads and analysis, including the execution of queries for the data.

The aim of the paper is to improve the integration, processing and analysis of unstructured spatial data. To achieve the aim, the following tasks were set:

- to analyze unstructured data types and typical sources;
- to explore NoSQL technologies and solutions for unstructured spatial data analysis;
- to analyze the coverage of the forest areas with protected areas using MongoDB database capabilities;
- to display the results of the analysis on the map using QGIS software.

Materials and Methods

Software

During the exploring of variety of NoSQL databases and their used technologies and solutions for spatial data analysis, MongoDB NoSQL database was chosen, which might be one of the best-known NoSQL databases in the market. MongoDB is free and open-source software, with the usual GNU (GNU General Public License is a free software license) and Apache-type licensing restrictions, which was designed more for analytics and complex data than it was for tables, ACID (Atomicity, Consistency, Isolation, Durability) compliance, and other standards and support requirements that come with ordinary relational databases. MongoDB is considered to be a document-oriented database which uses BSON (Binary-JSON) as a data description tool. BSON is an extension of JSON; it uses a length field to increase the efficiency of scanning (Pries & Dunnigan, 2015).

MongoDB database is called non-relational and schema-less, which makes working with data flexible because there is no predefined structure required in documents. MongoDB's data structure is not entirely lacking of schema, because it is still needed to define collections and indexes in database, but there is no need for predefined structure for documents added to database (Hows *et al.*, 2015).

For geospatial data analysis MongoDB Community Server 3.4.10 release (October 31, 2017) was used during this research. This version is compatible with Windows, Linux and MacOS operating systems. For data visualization on a map, QGIS Desktop 2.18.13 with GRASS 7.2.1 software was used. It was chosen because it is an open-source software and has plugins like Load MongoDB Layers, which links with MongoDB database and makes it possible to get spatial information directly from database into QGIS and then visualize it, as well as Save Data in MongoDB, what is usable for data saving back in MongoDB database after changes have been made to the spatial content. QGIS software allows to create, edit, visualize, analyze and publish geospatial information on Windows, Linux, Unix, Mac OSX and Android devices and supports numerous vector, raster, and database formats and functionalities (QGIS Development Team, 2018).

Data

For spatial data analysis with NoSQL, information about protected areas and forest areas in Latvia was used. The data of protected areas in Latvia is available online on https://www.daba.gov.lv/public/lat/iadt/ or https://data.gov.lv/dati/lv/dataset/ipasi-aizsargajamas-dabas-teritorijas-iadt government web sites. It is possible to download shapefile datasets from these sites. These datasets then should be converted in csv, tsv or GeoJSON file formats and geospatial data should be transformed in WGS84 coordinate reference system because MongoDB only supports WGS84 reference system for geospatial queries on GeoJSON objects. The data of protected areas was transformed in WGS84 coordinate reference system and converted to 17 csv files, which contain spatial information (polygons, points and lines) with ~ 120 thousands of protected area marks all over Latvia, and then imported into MongoDB database as GeoJSON objects. The information about forest areas of Latvia was provided by the State Forest Service of Latvia. The data was downloaded from the database into 1 csv file with ~2.5 million polygons of forest compartments of the whole Latvia in WGS84 reference system and imported into MongoDB database as GeoJSON objects.

Data analysis

For the unstructured spatial data analysis with available data, a NoSQL script was developed, which was searching and counting for every forest compartment its intersecting protected areas. In such a way, the forest compartments, where economic activity more likely will be restricted or limited, were found. Script were developed according to the scheme in Fig. 1. According to the scheme, it was determined that the number of processed compartments is 0 at the beginning and the first while loop is executed. This loop determines that operations will be continuously repeated in it while the processed compartment count will not be the same as the count of all compartments in the database. Then all unprocessed compartments are selected and in the next while loop operations are going through all these unprocessed compartments one by one. If there is any unprocessed compartment in this set, then its intersecting protected areas are searched. If any are found, the number of intersecting areas is saved in a separate field for compartment. Compartments with the highest number of protected areas are considered as the most restricted by the economic activity.

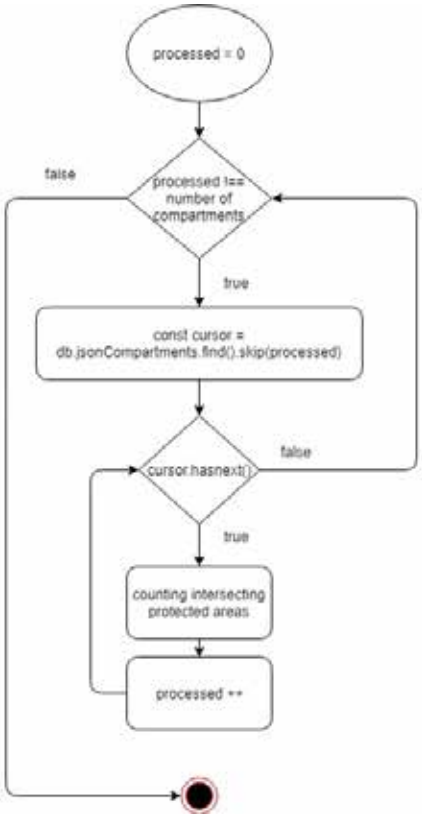

Figure 1. Search scheme for intersecting protected areas.

Results and Discussion

Data analysis results were imported into QGIS software using plugin – Load MongoDB Layers. Then the results were visualized using the map in Fig. 2. The darkest colors indicate places (forest lands) with the highest density of protected areas. The darker the green color is, the larger is the number of protected areas that intersect with this forest compartment, so it can be estimated that in these places the economic activity will be the most restricted. In turn, the lightest – yellow areas – characterize forest territories where the economic activity will not be restricted by the protected areas. White areas in the map are places where there are no forest lands – cities, agricultural land etc.

If the results of the analysis will be appended with features of roads and traffic information, thus lightening those forests which are easier to access, the map can be used to make a decision to purchase this forest land or not. If the land is hardly accessible and economic activity more likely will be restricted then probably the potential land buyer will decide not to buy this forest. In addition, if only 'Latvia's State Forests' (LVM) forests will be showed on the map together with traffic information, and national roads along with forest roads, this information can be used to create an application for people to make a decision, based on the frequency of visits to forest by other people as well as road accessibility, if it is worth to go to this forest to look for a Christmas tree in winter time or pick mushrooms or forest berries in summer.

MongoDB is commonly used for real-time data analysis. Complex analysis of large amount of perpetual data sets can take up to hours or even days. As the spatial data analysis with NoSQL took a long time (several days), the best solution to achieve results within a reasonable time still have to be found. A possible solution for this problem could be using MongoDB together with Hadoop - a software technology designed for storing and processing large volumes of data (as well as Big Data). Hadoop can be used as a data warehouse where larger data sets from MongoDB and other data sources can be uploaded. Then data analytics can use MapReduce or other programming models to create queries on these large data sets and return results back to MongoDB database. MongoDB together with Hadoop solutions for Big Data analysis are used by such well-known brands as Ebay and FourSquare (MongoDB, Inc., 2018).

According to the research and time used for analysis, it can be concluded that if the GIS application for data analysis from various data sources without predefined schema is created for small data sets and small areas (for example, within the borders of owners one or a couple of properties), then MongoDB will be

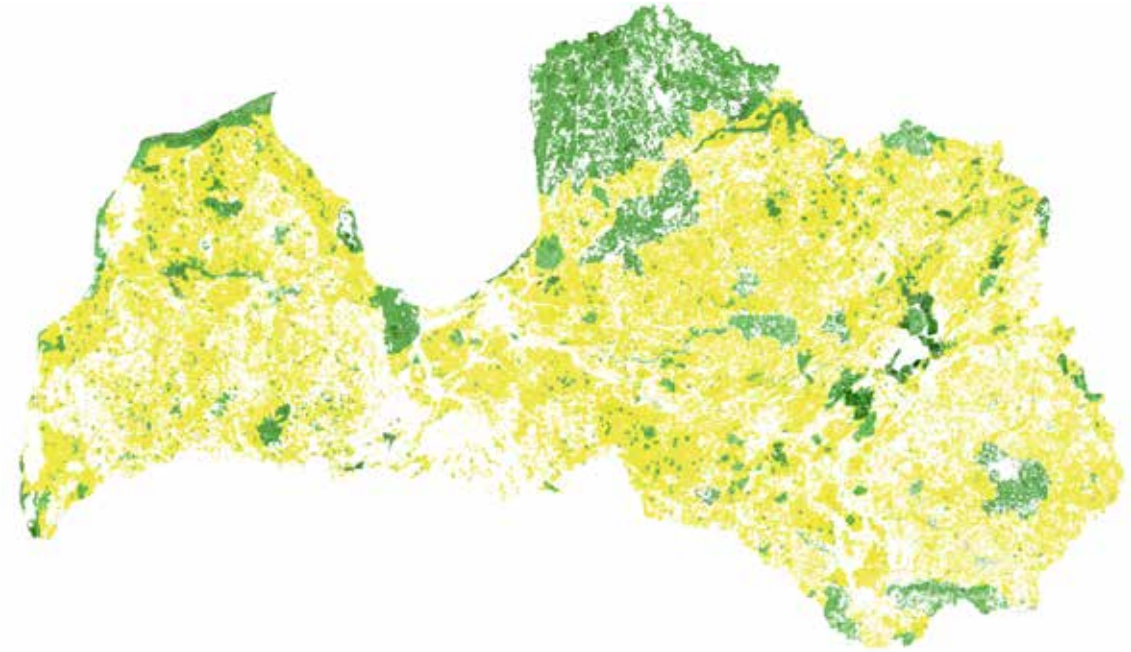

Figure 2. NoSQL data analysis of Latvia forest land intersections with the protected areas, visualized on a map.

a good solution for data analysis tasks. But if the data sets are large and cover large areas (for example, large regions, states or countries), then using MongoDB by itself for analysis tasks is not the best solution.

Conclusions

1. NoSQL solutions is a suitable tool for storage and analysis of data sets from various data sources without predefined schema or united structure of data, since there is no need to predefine table, structure, data types and columns for data import into database.
2. MongoDB can be used as a GIS database because of its integrated spatial functionality with fairly large spatial data analysis capabilities, however, there is a restriction that data should be transformed in WGS84 coordinate reference system before importing into database, otherwise spatial indexing will not work.
3. Results of this study show that MongoDB can be used for spatial data analysis with NoSQL, but for large areas and data sets it took a long time to obtain results, so the best solution for analysis of large data sets still need to be found.

References

1. Amirian, P., Winstanley, Ad., & Basiri, A. (2013). *NoSQL storage and management of geospatial data with emphasis on serving geospatial data using standard geospatial web services*. Maynooth: Department of Computer Science, National University of Ireland.
2. de Souca Baptista, C., Santos Pires, C.E., Batista Leite, D.F., de Oliveira, M.G., & de Lima Junior, O.F. (2014). NoSQL Geographic Databases: An Overview. In E. Pourabbas (Eds.), *Geographical Information Systems Trends and Technologies*, (pp. 73–104), Boca Raton: CRC Press Taylor & Francis Group.
3. Hows, D., Membrey, P., Plugge, E., & Hawkins, T. (2015). *The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB.* New York: Apress Media LLC.
4. MongoDB, Inc. (2018). Retrieved March 11, 2018, from: https://www.mongodb.com/hadoop-and-mongodb.
5. Peng, Y., & Zhenyu, T. (2017). GIS Databases and NoSQL Databases. *Comprehensive Geographic Information Systems*, 1(GIS Methods and Techniques), 50–80.
6. Pries, Kim H., & Dunnigan, R. (2015). *Big Data Analytics: A practical Guide for Managers*. Boca Raton: CRC Press, Taylor & Francis Group.
7. QGIS Development Team (2018). *QGIS – The Leading Open Source Desktop GIS.* Retrieved February 25, 2018, from: https://qgis.org/en/site/about/index.html.