МРНТИ 20.53.21
УДК 004.042

DOI: https://doi.org/10.51889/2020-1.1728-7901.57

MODELING OF LARGE VOLUMES OF DATA WITH THE USE OF NoSQL

Zhapsarbek N.B.[1]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*

Abstract

In the modern world, specialists and the information systems they create are increasingly faced with the need to store, process and move huge amounts of data. The definition of large amounts of data, Big Data, is used to denote technologies such as storing and analyzing large amounts of data that require high speed and real-time decision making during processing. In this case, large volumes, high accumulation rate, and the lack of a strict internal structure of "big data" are considered. All of this also means that classic relational databases are not well suited for storing them. In this article, we showed solutions for processing large amounts of data for pharmacy chains using NoSQL.

This paper presents technologies for modeling large amounts of data using NoSQL, including MongoDB, and also analyzes possible solutions, limitations that do not allow this to be done effectively. This article provides an overview of three modern approaches to working with big data: NoSQL, DataMining and real-time processing of event flows. In this article, as an implementation of the studied methods and technology, we consider a database of pharmacies for processing, searching, analyzing, forecasting big data. Also, when using NoSQL, we showed work with structured and poorly structured data in parallel in different aspects and showed a comparative analysis of the newly developed application for pharmacy workers.

Keywords: big data, pharmacy, data processing, analysis, NoSQL, MongoDB, DataMining.

Аннотация

Н.Б. Жапсарбек[1]

[1]*Казахский национальный университет им. аль-Фараби, г.Алматы, Казахстан*

МОДЕЛИРОВАНИЕ ОБРАБОТКИ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ NoSQL

В современном мире специалисты и создаваемые ими информационные системы все чаще сталкиваются с необходимостью хранить, обрабатывать и перемещать огромные объемы данных. Big Data, используется для обозначения таких технологий, как хранение и анализ больших объемов данных, которые требуют высокой скорости и принятия решений в режиме реального времени во время обработки. В этом случае рассматриваются большие объемы, высокая скорость накопления и отсутствие строгой внутренней структуры «больших данных». Все это также означают, что классические реляционные базы данных плохо подходят для их хранения.

В статье мы показали решения обработки больших объемов данных для сетей аптек с использованием NoSQL. В работе представлены технологии моделирования больших объемов данных с использованием NoSQL, в том числе MongoDB, а также анализируются возможные способы их решения, ограничения, которые не позволяют сделать это эффективно. Приводится обзор трех современных подходов к работе с большими данными: NoSQL, DataMining и обработка потоков событий в реальном времени. В статье в качестве реализации изученных методов и технологии рассматривается база данных аптек для обработки, поиска, анализа, прогноза больших данных. Также при использовании NoSQL показали работу со структурированными и плохо структурированными данными параллельно в разных аспектах и показали сравнительный анализ нового разработанного приложения для работников аптек.

Ключевые слова: большие данные, аптека, обработка данных, анализ, NoSQL, MongoDB, DataMining.

Аңдатпа

Н. Б. Жапсарбек[1]

[1]*әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан*

ҮЛКЕН КӨЛЕМДІ ДЕРЕКТЕРДІҢ NoSQL АРҚЫЛЫ МОДЕЛЬДЕУ

Қазіргі әлемде мамандар мен олар құрған ақпараттық жүйелер үлкен көлемде деректерді сақтау, өңдеу және жылжыту қажеттілігіне көбірек ұшырайды. Үлкен мәліметтердің үлкен көлемін анықтау үлкен жылдамдықты және өңдеуді нақты уақыт режимінде қабылдауды талап ететін деректердің үлкен көлемін сақтау және талдау сияқты технологияларды білдіреді. Бұл жағдайда үлкен көлемдер, жинақталудың жоғары деңгейі және «үлкен деректердің» қатаң ішкі құрылымының болмауы қарастырылады. Мұның бәрі классикалық реляциялық мәліметтер базасы оларды сақтау үшін өте қолайлы емес дегенді білдіреді. Бұл мақалада біз NoSQL көмегімен дәріханалар тізбегі үшін үлкен көлемдегі мәліметтерді өңдеудің шешімдерін көрсеттік.

Бұл жұмыста NoSQL, соның ішінде MongoDB қолдану арқылы деректердің үлкен көлемін модельдеу технологиялары ұсынылған, сонымен қатар оны тиімді орындауға мүмкіндік бермейтін шешімдер мен шектеулер талданған. Үлкен деректермен жұмыс істеудің үш заманауи тәсілдеріне шолу жасалынған: NoSQL, DataMining және оқиғалар ағынын нақты уақыт режимінде өңдеу. Бұл мақалада зерттелген әдістер мен технологияны қолдану ретінде біз үлкен деректерді өңдеуге, іздеуге, талдауға, болжауға арналған дәріханалар базасын қарастырамыз. Сондай-ақ, NoSQL-ді қолдану кезінде біз параллель құрылымдалған және нашар құрылымдалған мәліметтермен жұмысты әр түрлі аспектілерде көрсетіп, дәріхана қызметкерлері үшін жаңадан жасалған қосымшаның салыстырмалы талдауын көрсеттік.

Түйін сөздер: үлкен көлемдегі деректер, дәріхана, деректерді өңдеу, талдау, NoSQL, MongoDB, DataMining.

Big data is a variety of tools, approaches and processing methods for both structured and unstructured data in order to use them for specific tasks and goals. Today, under this simple term, only two words are hidden - data storage and processing.

Despite the frequency with which this term is used in discussions of modern computer technology, it does not have a single universally accepted definition. Most of the definitions of the term "big data" used can be attributed to one of the three main classes [1].

The development of forecast models is especially relevant for the business sector, the main task of which is to have knowledge that can increase efficiency, reduce costs, and / or increase sales. It is the Big Data sphere that is the provider of effective predictive solutions - according to statistics, only 0.5% of accumulated digital data is currently being analyzed, the rest contains a huge amount of “hidden” knowledge that could potentially be a source of huge profit and superiority over competitors.

At present, there is no generally accepted definition of this term, nor an authoritative body that would propose such a definition, so we can only discuss some general properties of databases that belong to the NoSQL category of modern information-analytical systems. [3].

The data presented as a NoSQL data warehouse demonstrates an additional phenomenon: they usually retain considerable flexibility in data with limited use of the concept of a scheme, as is usually the case in databases [4]. MongoDB was used for open source NoSQL databases. The REST API CRUD or RESTful API is also widely used in MongoDB and is widely used as unstructured data representations to support several types of multimedia such as text, HTML, JSON, etc.

All processing of large amounts of data using NoSQL scale. Therefore, the main essence of storing big data is an online database, which NoSQL can manage better, use a horizontal scaling strategy and provide similar flexibility [2]. It is also worth noting that for analytical data, CRUD means create, read, update, and delete operations.

The relevance of the study - this work considers one of the possible ideas for applying the big data paradigm - the ability to create a website for pharmacies based on publicly available data

The aim of the study is to build a model and develop on its basis a system that allows the processing of large amounts of data using NoSQL.

The object of study is the technology for modeling large amounts of data using NoSQL, including MongoDB.

The subject of the study is - methods, algorithms and circuit solutions for the implementation of basic data, modification and processing of unstructured pharmacy data based on Data mining.

The statement of the problem of data analysis in the general case is as follows:

• There is a fairly large database containing information on the target area of knowledge - hereinafter referred to as the “Training Sample”.

• It is assumed that there is some “hidden knowledge” in the database. It is necessary to develop methods for detecting knowledge hidden in large volumes of raw data. In the current conditions of global competition, it is precisely the regularities (knowledge) found that can be a source of additional competitive advantage.

The objectives of this study:

Based on the foregoing, this work is aimed primarily at solving such problems as:

• Analysis of the current state of the Big Data area

• Identification of the advantages and disadvantages of each approach

• Creating a data model for pharmacies intended for analysis

• Implementation of the system based on the model of existing pharmacies

• Development based on a system that allows the processing of large amounts of data using NoSQL

The results of the study:

According to the task, a website was created for pharmacy workers with these criteria. A MongoDB database has been created and contains information about the following objects:

• Employees - last name, first name, patronymic, address, date of birth, position, salary, information about the transfer (position, reason for transfer, number and date of order).

• Assortment of medicines - name of the medicine, form of packaging, price per package, quantity.

Business rules.

- Each medicine has a list of substitutes that can be recommended to customers in the absence of the main medicine.
- Each medicine can be a substitute for many medicines.
- Each medicine can be issued in various packings.
- The price of the medicine is determined by the packaging.

During the study of this topic, used technologies and analysis methods applicable to Big data. This is Data Mining and Statistical Analysis.

Data Mining (DM) - literally, these words mean "mining, excavating, extracting data." [8]

Statistical analysis - measurements, monitoring, analysis of mass statistical data and their comparison, the study of the quantitative side of mass social phenomena in numerical form.

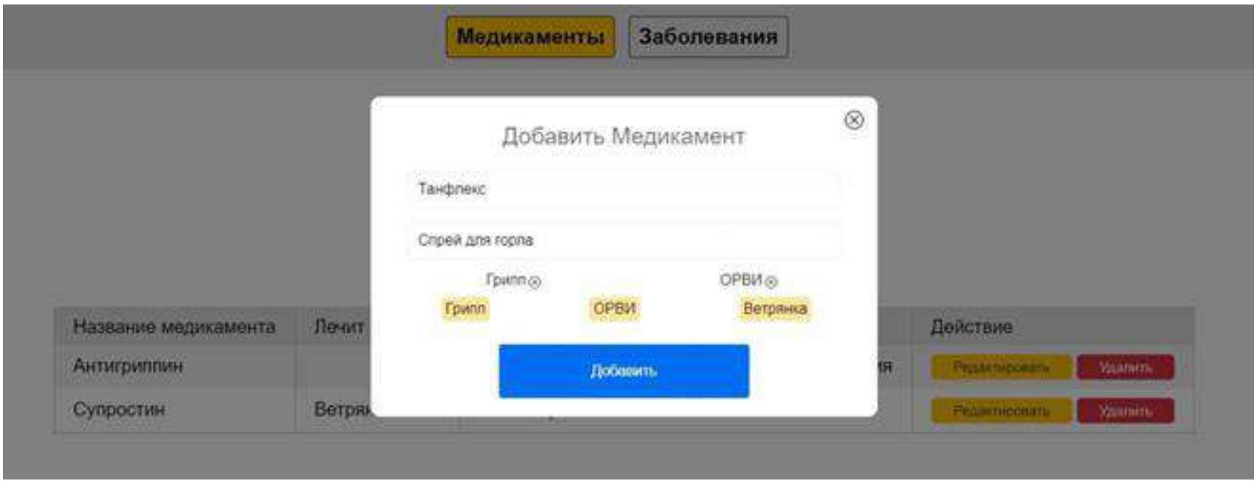
When creating this site, I took into account that the site works with big data and requires an important role for processing. For the analysis of disease and for the management of medicines, we use the method of statistical data analysis. The following resources showed pages for adding and finding medication and disease. Data can be added, edited or deleted. During the study, a site with a base of pharmacies for pharmacy workers has been implemented. The database is implemented in a document-oriented database management system MongoDB. This is not a bad result for processing large amounts of data, you can also increase the database for a good analysis of work.

In general, the developed system solves the following tasks:

- Collection of data for analysis - involves accessing information sources, gaining access to event logs containing drug data.
- Conducting data analysis - the task is to identify data laws, the basis is the task of classification into five classes. This problem is solved using machine learning methods.
- Formation of analysis results for users of the system - classification results containing a system decision on the user class can solve applied business problems, for example, creating contextual advertising, or generating recommendations.

Passing to the point of analysis of the apparatus of mathematical statistics for solving the problems of the intellectual analysis of big data, the consideration of initially theoretically substantiated approaches to the problem of data analysis begins. A review and analysis of this and the following data processing methods will also focus on the efficiency of the algorithms, their computational performance, complexity of implementation and applicability to the task.

Despite the apparent simplicity of Big Data ideas, there are many problems that one has to deal with when solving data analysis problems. In addition to the high cost, there are problems, firstly, of the choice of the processed data: that is, the determination of which data should be extracted, stored and analyzed, and which should not be taken into account. Also, huge amounts of information compared to other areas leave their imprint on the computing capabilities of such solutions - the most powerful super-computers have been solving other problems for years. In the course of achieving the goal set in this work creating a method and system of effective classification for a pharmacy, a number of tasks were performed, in particular, the analysis of the subject area of data analysis was carried out, the existing Big Data approaches were analyzed, and the applied analysis system was directly implemented.



Тамфлекс

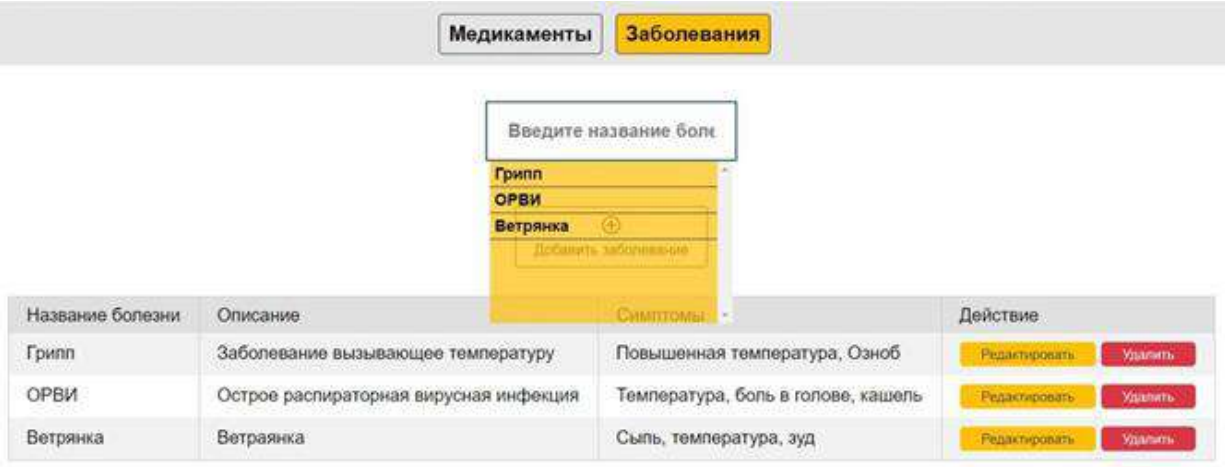
Спрей для горла

Антигриппин

Супростин

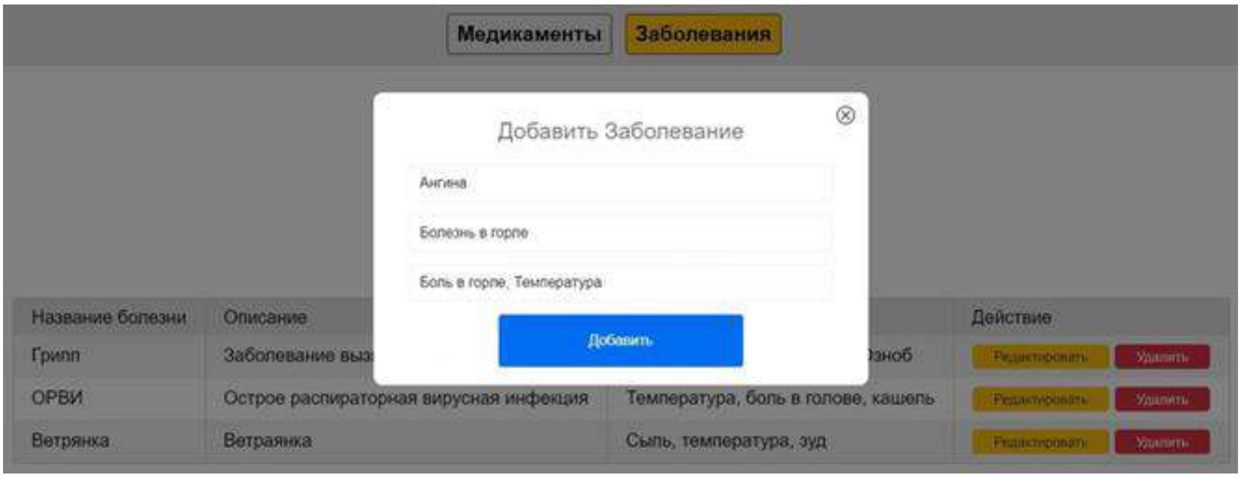
Ветрян

Figure 1. The page for adding drugs



Название болезни	Описание	Симптомы	Действие
Грипп	Заболевание вызывающее температуру	Повышенная температура, Озноб	Редактировать Удалить
ОРВИ	Острое распираторная вирусная инфекция	Температура, боль в голове, кашель	Редактировать Удалить
Ветрянка	Ветраянка	Сыпь, температура, зуд	Редактировать Удалить

Figure 2. Search Page - Pharmacist Handbook



Ангина

Болезнь в горле

Боль в горле, Температура

Название болезни	Описание		Действие
Грипп	Заболевание выз	Озноб	Редактировать Удалить
ОРВИ	Острое распираторная вирусная инфекция	Температура, боль в голове, кашель	Редактировать Удалить
Ветрянка	Ветраянка	Сыпь, температура, зуд	Редактировать Удалить

Figure 3. Page for adding disease

Nowadays, the Big Data theme inextricably follows hand in hand with the concept of loosely structured data. More and more new indicators, measurements, survey results appear daily at a high speed, bringing a chaotic set of values to a completely structured look is a difficult, time-consuming task. Significantly less effort will have to be made, there is a post-processing system that will satisfy poorly structured information, working with which in automated mode is not much more difficult than with completely structured ones. In conclusion, I want to say that this work is intended to improve the work of pharmacy workers, processing large amounts of data using MongoDB.

References:

1 Patricia B. Cerrito, Introduction to Data Mining. – SAS Institute Inc., 2006. - C.459

2 Snijders C., Matzat U., Reips U. D. 'Big Data': Big gaps of knowledge in the field of Internet. // International Journal of Internet Science 7 (2012). P. 1–5

3 Daniel T. Laros, Data Mining methods and models. – Department of Mathematical Sciences, 2006. – C.385

4 P. Zikopoulos, C. Eaton Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data // New York, NY, USA: McGraw-Hill, 2011

5 Свинарев Сергей. Еще раз о росте популярности NoSQL. [Электрон.ресурс] — 2015. — URL: http://www.jetinfo.ru/stati/ silnye-i-slabye-storony-nosql

6 Peter Bakkum Kyle Banker Shaun Verch, Douglas Garrett, and Tim Hawkins. MongoDB in Action (Second edition) - Manning Publications Co., 2016. — P. 481— ISBN: 9781617291609

7 Borland Bo. Pentaho Analytics for MongoDB. Packt Publishing, 2014. — P. 146

8 Королева О. В., Демьяненко А. И., Золотов А. Д. Разработка базы данных для информационно-справочной системы по поиску лекарств в аптеках // Молодой ученый. [Электрон.ресурс]— 2016. — №10. — C.27— URL https://moluch.ru/archive/114/30219/