THE DEVELOPMENT OF INFORMATION SYSTEM OF FORMATION AND USE OF INFORMATION RESOURCES FOR EVALUATION OF PARAMETERS AND EVALUATION OF RECOMMENDATIONS BASED ON BIG DATA TECHNOLOGY TOOLS: WORK WITH MONGODB

N. SAPARKHOJAYEV[1*], A. MUKASHEVA[2] and P. SAPARKHOJAYEV[3]

[1] *PhD, Associate Professor, Head of Department "Computer Engineering", Akhmet Yassawi International Kazakh-Turkish University, Sattarkhanov Avenue 29, Turkestan, Kazakhstan, nursp81@gmail.com
[2]PhD Doctorate student, Department "Information Technologies", Satbayev University, Satpayev Str. 22, Almaty, Kazakhstan, mukasheva.a.82@ gmail.com
[3]Candidate of pedagogical sciences, Professor, Department "Physics and Mathematics", The Korkyt Ata Kyzylorda State University, Kyzylorda, Kazakhstan, nurdash1552@mail.ru

***Abstract* - The main goal of this research work is to describe automation process of forming a relational structured database in the Hadoop ecosystem environment. Selection a source in the Internet environment and extracting information online, choosing an import tool, studying unstructured data in Hadoop are described. The use of tools (systems, utilities) such as MongoDB, Hadoop in this research work allows combining operational and analytical technologies.**

***Keywords* - MongoDB, BigData, Hadoop, store, execute, data.**

I. INTRODUCTION

Big Data incorporates all kinds of data and from a content perspective one can make the distinction between structured data, semi-structured data and unstructured data [1].

• Structured data are data that are part of a formal structure of data models associated with relational databases or any other form of data tables. They can be generated both by computer software or humans.

• Semi-structured data are data that are not part of a formal structure of data models. Examples are EDI, SWIFT, and XML and JSON data [2].

• Unstructured data are data that do not belong to a pre-defined data model and include data from e-mails, video, social media websites and text streams. They account for more than 80% of all data in organizations [3].

Until recently, software technology did not effectively support doing much with them except storing or analyzing manually. Just as with structured data, unstructured data are either machine generated (by computer or software) or human generated [2].

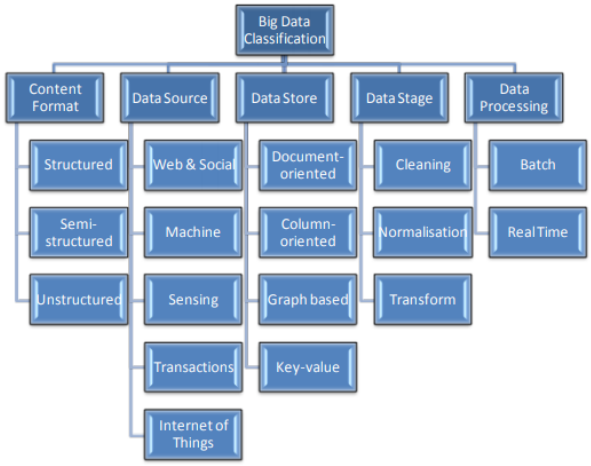

Figure 1: Big Data classification [4].

II. LITERATURE REVIEW

Hadoop vs MongoDB

Big data is getting bigger, and with it the complications in managing data. For many, tools such as Apache Hadoop, MongoDB and NoSQL singularly represent big data [5].

Hadoop.

Based on a comparative analysis of the distributions of Cloudera, Amazon, Azure, Google cloud and Hortonworks, a product of Hortonworks was chosen [6], because it does not require financial costs, the software is distributed on the basis of free downloads and is technically convenient for installing and working with it. Also Hortonworks distribution allows the programmer to additionally download other Hadoop ecosystem tools for working with large data arrays.

MongoDB.

For operational Big Data remaining burdens, NoSQL Big Data frameworks, for example, record databases have risen to address a wide arrangement of utilizations, and different models, for example, key-value stores, column family stores, and graphical databases are enhanced for more applications that are particular.

Speaking clearly MongoDB is built for the cloud. Its local scale-out engineering, empowered by 'sharding',

adjusts well with the even scaling and deftness managed by cloud computing. Sharding consequently disperses information equally over mult i-node clusters and equalizations questions over them [7]. First, it is the fastest-growing new database in the world that provides a rich document oriented structure with dynamic queries. Second, it allows compartmentalizing data into collections in order to divide data logically. MongoDB can manage data of any structure without expensive data warehouse loads, no matter how often it changes. Thus, we can cheap new functionality without redesigning the database [8].

MongoDB can join any kind of information – any structure, any format, and any source – no matter how regularly it changes. Your analytical engines can build based on its comprehensiveness and in real-time.

Nowadays utilizing MongoDB for analytics since it lets them store any kind of information, analyze it in genuine time, and alter the pattern as they go. MongoDB's archive show empowers you to store and prepare information of any structure: occasions, time arrangement information, geospatial arranges, content and double information, and anything else. You'll be able adjust the structure of a document's pattern fair by including unused fields, making it simple to bring in modern information because it gets to be accessible [9].

III. The Description of Proposed System and Its Architecture

Installation of Hadoop

The Hortonworks distribution offers the following installation procedure [10]: Apache Ambari is selected, installed and launched. Ambari provides user interface management with its own RESTful APIs. Ambari allows system administrators to manage, provide work, and control a Hadoop cluster, as well as integrate Hadoop with existing enterprise infrastructure.

Ambari provides step-by-step installation of Hadoop services for any number of hosts. Ambari supports the configuration of Hadoop services for a cluster. Ambari provides centralized management of the start, stop, and configuration of Hadoop services for the entire cluster.

At the beginning, Apache Ambari is installed in the following sequence:

- The latest version of Ambari HDP 2.6.5 is used;

- CentOS 7 operating system [11] with the following tools: yum and rpm package manager; tools like scp, curl, wget, unzip, tar; programming language Python [12] and Java (JDK 8+ Open Source or Oracle) [13]. Also it is needed to pre-configure the CentOS 7 operating system as well as NTP (Network Time Protocol) must be installed, because it is necessary that all cluster members can synchronize their internal clocks via the Internet according to their time zone.

- availability of a database management system (DBMS) for using tools such as Oozie or Hive [10]. Apache Ambari HDP allows the installation process to select and install one of the default DBMS packages, usually MongoDB [14]. We can also choose and deploy one of the DBMSs on the server and then, during the installation process, using the built-in configurator, specify the already installed DBMS as the main one that Apache Ambari HDP will use.

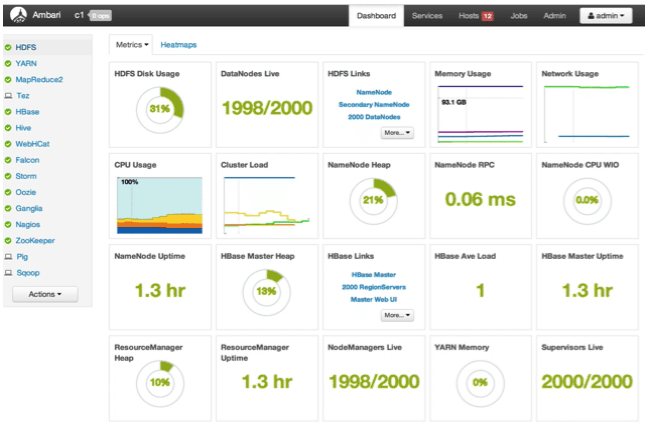

Figure 2: The window of prepared cluster.

For the convenience of changing configuration files, in the Apache Ambari environment, you must select the service on the panel in the browser.

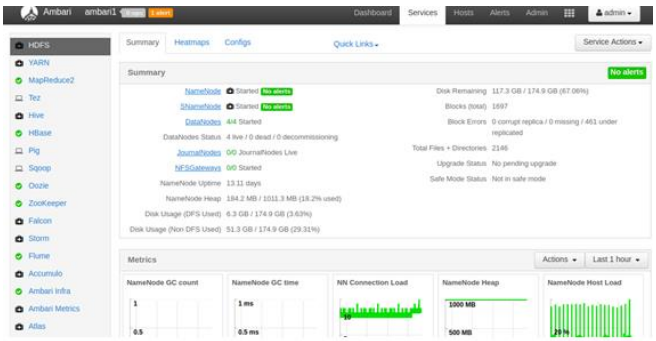

Figure 3: Choosing a service to change configuration files.

In our case, HDFS will be selected for configuration. After we select the HDFS service in the panel, the metrics that indicate the state of the HDFS file system will appear on the screen, as well as we will see all the necessary information about the number of nodes that host the HDFS file system. We can change the storage location of metadata for Namenode and Datanode as well as increase the amount of RAM for tasks performed by HDFS.

Figure 4: Menu for selecting advanced settings.

This way it is possible to change configuration files not only for HDFS but also for many other tools.

Installation of MongoDB

The mongodb-org package is not in the CentOS official repository. However, MongoDB supports a special separate repository that can be added. Using a text editor, create a .repo file for yum, CentOS Packet Manager [14]:

Sudo vi /etc/yum.repos.d/mongodb-org.repo

Open the official MongoDB documentation (Install on Red Hat section) and add the latest stable release information to the file [14-17].

[mongodb-org-3.2]
name=MongoDB Repository
baseurl=https://repo.mongodb.org/yum/redhat/$releasever /mongodb-org/3.2/x86_64/
gpgcheck=1
enabled=1
gpgkey=https://www.mongodb.org/static/pgp/server-3.2.asc

Save and close the file. Then you need to make sure that yum sees the MongoDB repository. To do this, use the repolist command [14-17].

Yum repolist
…
repo id repo name
base/7/x86_64 CentOS-7 – Base
extras/7/x86_64 CentOS-7 – Extras
mongodb-org-3.2/7/x86_64 MongoDB Repository
updates/7/x86_64 CentOS-7 – Updates
…
Install package mongodb-org:

sudo yum install mongodb-org

After running the command, two requests will appear.

Is this ok [y/N]:

The first is a request to allow the installation of a MongoDB package, and the second is to import a GPG key to confirm the integrity of the downloaded packages. Type Y and press Enter [14-17]. Then launch MongoDB service:

sudo systemctl start mongodb.

Loading data into MongoDB

An article in .pdf format will be taken as data source.

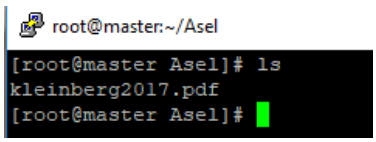

Figure 5: Loading data into MongoDB.

Before uploading data into MongoDB, a database needs to be created. At first stage logging into the MongoDB management console is needed to be performed. To do this, type the mongo command in the console.

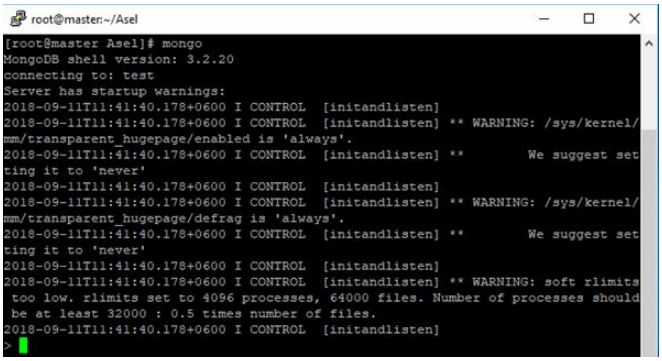

Figure 6: Creation of DB.

First you need to check which databases already exist in MongoDB, for this you need to type the command show dbs.

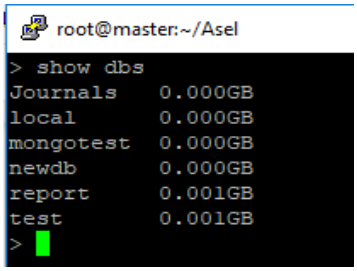

Figure 7 – Checking created DB.

To create a new database, type the command use report, report the name of the new database. Data about the new database will appear only after we load any data into it.

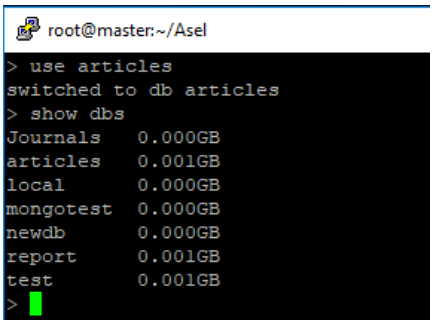

Figure 8 – The name of the new DB.

Next, you need to upload our file to the MongoDB database, for this you need to use the mongofiles command.

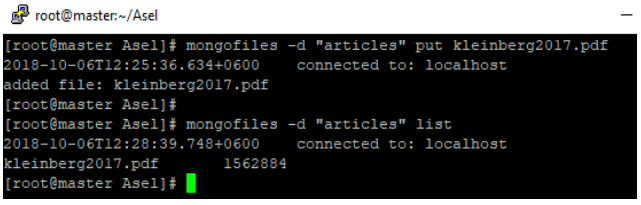

Figure 9: The information about the loaded file.

To upload a file from MongoDB it is necessary to do the following.

Figure 9: The process of uploading a file from MongoDB.

IV. Conclusion and Future Work

The main part of the software used in this research work for working with BigData is open source. This allowed us to produce work with structured and unstructured data. Illustrations of this model incorporate MongoDB (by MongoDB, Inc.) and Hadoop (by Cloudera and others) [7].This research work is initial step in big project that deals with BigData technology tools and Data Mining algorithm allowing user to work, control and analyze huge amount of data. Next step in this project is to perform MapReduce applications that connects to DB, which we built both in MongoDB and Hive, after this step Data Mining algorithms used in MapReduce applications will be able to manipulate thru data stored in MongoDB and allow users to to receive appropriate data according to requests. Finally, authors plan to use all acquired skills and expertise in applying BigData technology in the medicine for helping doctors in their hard work.

References

[1] Kambatla, K., Kollias, G., Kumar, V., Grama, A. (2014). *Trends in big data analytics*. Journal of Parallel and Distributed Computing, 74 (7), 2561-2573.

[2] *White Paper BIG DATA*, Version 1.2 – November 2016.

[3] Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R. (2013). *Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field*. In Holzinger, Andreas; Pasi, Gabriella. Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science. Springer. Pp. 13–24.

[4] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., Khan, S. U. (2015). *The rise of "big data" on cloud computing*: Review and open research issues. Information Systems, 47, 98-115.

[5] *Harnessing the Big Data – Hadoop vs MongoDB*. Available: https://www.happiestminds.com/blogs/harnessing-the-big-data-hadoop-vs-mongodb/

[6] *Hortonworks.* Available: https://hortonworks.com/.

[7] *What Is Big Data?* Available: https://www.mongodb.com/big-data-explained

[8] Abbes, H., & Gargouri, F. (2016). *Big Data Integration: A MongoDB Database and Modular Ontologies based Approach.* Procedia Computer Science, 96, 446–455. Doi:10.1016/j.procs.2016.08.099.

[9] *MongoDB Makes It Easy*. Available: https://www.mongodb.com/use-cases/real-time-analytics

[10] *Apache Ambari*. Available: https://ambari.apache.org/

[11] *CentOS Documentation*. Available: https://www.centos.org/docs/

[12] *Python.* Available: https://www.python.org/

[13] *Java*. Available: https://java.com/ru/download/

[14] *Installing MongoDB in centos 7.* Available: https://www.8host.com/blog/ustanovka-mongodb-v-centos-7/

[15] *Install and protect MongoDB in ubuntu 16.04.* Available: https://www.8host.com/blog/ustanovka-i-zashhita-mongodb-v-ubuntu-16-04/

[16] *Online magazine for professional web designers and developers*. Available: http://www.coolwebmasters.com/databases/3778-webdev-with-mongodb-part

[17] *Mongodb manual*. Available: https://docs.mongodb.com/manual/tutorial/install-mongodb-on-red-hat/#configure-the-package-management-system-yum