Information Technologies for Building a Data and Knowledge Warehouse for Science and Technology Forecasting and Research of Critical Infrastructures

Young researcher's note

E.P. Khayrullina[*]

Melentiev Energy Systems Institute of Siberian Branch of Russian Academy of Sciences, Irkutsk, Russia.

***Annotation* – The paper is concerned with the application of the Open and Big Data technology in energy forecasting. The focus is made on the issues related to the creation of a data and knowledge warehouse for science and technology forecasting, and research of critical infrastructures.**

***Keywords* – big data, data and knowledge storage, science and technology forecasting, critical infrastructures.**

I. Introduction

Currently, the world is actively developing intelligent information technologies that support the innovative development of many industries and economic areas.

The importance of the intelligent energy technologies is undeniable [1]. As evidenced by the experience of many countries, these technologies are becoming the foundation for the development of electric power industry. In the USA, in particular, this is the main direction of improving the economy, China sees these technologies as a way of national strategic development. The European Union countries use innovative technologies as a basis for a new energy policy [2]. One of the components of the intelligent energy systems will be science and technology forecasting [3].

Science and technology forecasting implies the identification and preliminary assessment of trends in the development of science and technology, the prediction of major scientific and technical solutions capable of making quantitative changes in the overall science, technology and production potential of the country, in social relations, and world politics.

The information for science and technology forecasting will be formed on the basis of Open Data and Linked Open Data. The data will also be collected from such sources as state information systems that integrate the data on science and technology projects and developments (CITIS, RFBR, FIPS, etc.), as well as various commercial systems such as SCOPUS, Web of Science, RSCI, and Science Index [3,4].

An analysis of the information to be collected from the above-described sources can indicate general trends in the development of scientific and engineering thought [4, 5].The technologies and tools to be developed can be used to analyze the threats and assess the risks of cyber security incidents in critical infrastructures, including energy facilities [6].

II. Specific features of the data and knowledge warehouse for energy forecasting

Energy forecasting necessitates accumulating and analyzing a large amount of data. The more information is processed, the more accurate the forecasting results will be, as the increase in the data amount enhances the completeness of information and makes it possible to assess the reliability and consistency of the data obtained.

An exponential increase in various information in the modern world affects the creation of data and knowledge warehouse. The volume of information in the world increases by 30%, annually.

The data and knowledge warehouses are primarily designed for science and technology forecasting based on the analysis of heterogeneous, unstructured data sets of large volumes that need scale-out software and distributed computing for their effective processing and storage.

[*] Corresponding author.
E-mail: Lena-Skoklenyova@yandex.ru

http://dx.doi.org/10.25729/esr.2018.03.0012
Received August 15, 2018. Revised November 28, 2018.
Accepted December 18, 2018. Available online Jan. 25, 2019.

The accumulated volume of data structured in the warehouse can be used to solve related problems in energy research: linked data on information technologies used at energy facilities can be used to analyze their cyber security.

III. Big Data Technologies

In this paper, the term Big Data means some technologies, tools, and methods for processing structured and unstructured data of large volumes, that allow distributed processing of information.

Principles of working with big data:

1. Horizontal scalability, i.e. with a rise in the storage volumes, the system should be able to support an increase in the number of servers;
2. Fault tolerance;
3. Data locality, i.e. data should be processed and stored at the same machine, otherwise the data transmission effort may exceed the data processing effort.

To solve the problem of data locality, Google proposed the MapReduce concept. MapReduce is a model of distributed computing, used for parallel processing of large amounts of information [7]. MapReduce assumes that the data are organized in the form of some records. Data are processed in 3 steps. The block diagram of this model is presented in Figure 1.

1. Map step. The input data of the problem to be solved represent a large list of values that are preliminarily processed at the step Map. To this end, the master node of the cluster receives this list, divides it into parts and sends it to the slave nodes. Then, each of the slave nodes converts the elements of the resulting collection to zero or several intermediate key-value pairs.
2. Shuffle step is unnoticeable for the user. At this step, the intermediate results are grouped.
3. Reduce step. At this step, the master node receives intermediate responses from the slave nodes and transfers them to the free nodes to perform the next step. The system sorts and groups all “key-value” pairs by key and then, for each pair “key-group of values”, collapses the values often into one value or an empty list. The obtained result is a solution to the originally stated problem [7, 8].

One of the possible solutions for organizing distributed big data storage and processing is the NoSQL class databases. They come in 4 types:

1. A key-value store. It is a database that uses the key to access the value. The examples of such stores are Berkeley DB, MemcacheDB, Redis, Riak, and Amazon DynamoDB.
2. A Bigtable storage. In this storage, data are stored as a sparse matrix, with its rows and columns used as keys. The examples of databases of this type are: Apache HBase, Apache Cassandra, Hypertable, SimpleDB.
3. A document-oriented database. It serves to store hierarchical data structures. The examples of this type of databases are CouchDB, Couchbase, MarkLogic, MongoDB, eXist, Berkeley DB XML.

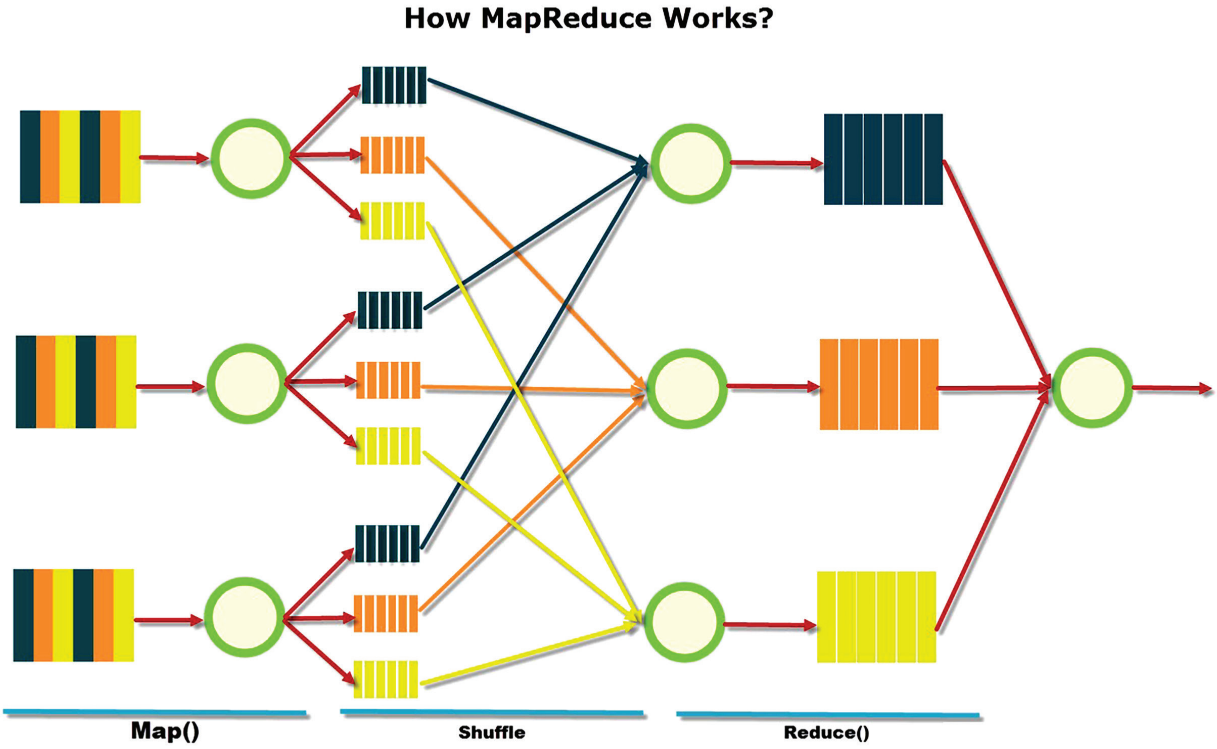

Figure 1. The MapReduce model (http://blog.sqlauthority.com).

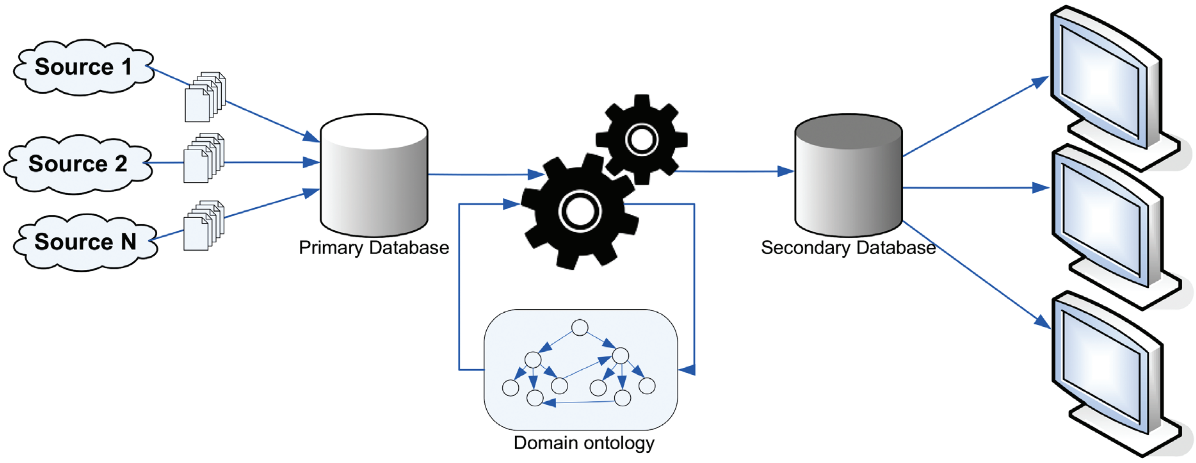

Fig. 2. The proposed architecture of the system.

4. A graph database. It is used in tasks with the data having a large number of links (social networks, fraud detection). The examples of the graph databases are: Neo4j, OrientDB, AllegroGraph, Blazegraph (RDF-storage formerly called BigData) [7].

IV. The proposed architecture of the system

The architecture presented in Figure 2 is proposed to implement the system of science and technology forecasting. The documents (articles, monographs, etc.) and relevant metadata are extracted from the resources specified in advance, such as RINC, Scopus and others. Based on the extracted documents, the primary document storage is formed. Then, a predetermined ontology of energy terms is used to make a semantic analysis, and the obtained results are transferred to a graph database. After this step, the results will be available to the end user.

V. Conclusion

Nowadays the Big Data technologies make it possible to store and process the information distributed on many servers and computers. The proposed solution will support an increase in the storage size in accordance with the amount of the information required for the tasks of science and technology forecasting in the field of energy and the study of critical infrastructures.

The results were obtained during the implementation of the scientific project of the fundamental research programs of SB RAS III.17.2.1, reg. № AAAAA17-117030310444-2, III.17.1.4, reg. № AAAA-A17-117030310436-7, with partial financial support of the RFBR grants № 17-07-01341, № 18-37-00271.

References

[1.] N.I. Voropai, “Electric power systems expansion planning: Methodology, models, methods, and their use,” Novosibirsk, Nauka, 2015, p. 448. (in Russian)

[2.] C. Cagnin, “Future-Oriented Technology Analysis,” *Strategic Intelligence for an Innovative Economy*, Springer, p. 170, 2008.

[3.] A.V. Mikheev, “A semantic-based approach to energy technology forecasting,” *Proceeding of International Workshop “Contingency management, intelligent, agent-based computing and cyber security in critical infrastructures,”* CM/IAC/CS/CI-2016, Russia, Irkutsk, MESI, pp. 41-42, 2016.

[4.] A.N. Kopaigorodsky, “Information support of collective expert activity for forecasting of innovative energy development,” *Proceeding of International Workshop “Critical Infrastructures: Contingency Management, Intelligent, Agent-based, Cloud Computing and Cyber Security.”*

[5.] E.P. Khairullina, A.N. Kopaigorodsky, “Application of ontologies in the design and implementation of information systems,” *Informatization and visualization of economic and social life*, pp. 200-204, 2015.

[6.] D.A. Gaskova, “An analysis of cyber security violations in the energy sector,” System Research in Energy, Proceedings of young scientists of MESI SB RAS, vol. 47, Irkutsk, MESI SB RAS, 2017, pp. 101-107, (in Russian.)

[7.] K. S. Manoj, K. G. Dileep, “Effective big data management and opportunities for implementation,” Information Science Reference, America, 2016 p.440

[8.] E. Dumbill, “Planning for Big Data,” *O’Really Radar Team*, America, p.78, 2012.

Elena P. Khayrillina, graduated from Irkutsk National Research Technical University in 2017. She is a postgraduate at Melentiev Energy Systems Institute SB RAS. Her main research interests are ontology, tech mining and Big Data.