Design Big Data Analysis System - Bigdeepexaminator

Janusz Bobulski[1], Mariusz Kubanek[2]
Institute of Computer and Information Science,
Czestochowa University of Technology,
Czestochowa, Poland
[1]januszb@icis.pcz.pl
[2]mariusz.kubanek@icis.pcz.pl

ABSTRACT

Big Data is a term used for such data sets, which at the same time are characterized by high volume, di-versity, real-time stream inflow, variability, complexity, as well as require the use of innovative technolo-gies, tools and methods in order to extracting new and useful knowledge from them. Big Data is a new challenge and information possibilities. Correct interpretation of data can play a key role in the global and local economy, social policy and enterprises. We present a data analysis system design with the use of ar-tificial intelligence that will help in obtaining valuable information from big data.

CCS Concepts

• **Information systems→Data streams**

Keywords

Component; Big data; intelligent systems; data pre-processing multi-data processing.

1. Introduction

Loads of new digital data are generated from various sources of information every day, and problems of processing Big Data resources have emerged. The processing of large data sets is associated with many challenges that the person supervising the whole process must cope with. These include: the problem with understanding and acceptance of large data sets, a variety of Big Data technologies, high system costs, scalability, quality of data and transforming large sets of data into valuable information. [1, 2 and 3]

Insufficient understanding and acceptance of large data sets causes problems with the implementation of projects in many companies. Lack of knowledge about what large data is, what infrastructure is necessary, etc. causes companies to lose a lot of time and resources, which causes delays in the company's progress and development. Big data and related knowledge should be obtained and accepted by the company's management, and then transferred down the structure. To ensure understanding and acceptance of large data at all levels, IT departments must organize numerous training and workshops. In order to better accept changes in data management, all Big Data implementation processes must be monitored and controlled on an ongoing basis.

Currently, there is a large variety of technologies used on the market. In the course of their analysis, the questions arise: do I need Spark, or the speed of Hadoop MapReduce is enough, can it better store data in AWS or Cassandra? Finding the right solution is difficult, especially if you lack the knowledge you need. It is worth to hire an expert in this situation or turn to a professional company that provides Big Data solu-tions.

This solution will also reduce the costs of implementing new technologies in the company. Open-source solutions seem to be cheaper at the beginning, but considering all the additional costs, ie new equipment and personnel, it may turn out that this solution is more expensive and less productive in the long run, but it will certainly slow down the company's development.

In justified cases, one should use the offer of suppliers offering ready solutions operating in the cloud, thanks to which we will also avoid some problems with scalability. If we additionally choose the right architecture, we will guarantee the proper system performance. An additional important issue from the point of view of scalability is the adequately design algorithms of large data sets, including the possibility of expansion.

In Big Data there is a big problem with the integration and homogeneity of data, because they come from different sources and there are different formats, for example, the format of dates in the US and in Europe, whether upper and lower case letters in the name of their own. This results in the need to control their condition [3,4]. The key factor here is the correctness and quality of the data, because in the case of their lack the results of the system will be unsatisfactory, according to the principle: garbage at the input-garbage at the output. Therefore, the system should be equipped with a data control and conditioning module, analogous to signal processing systems where the input signal is filtered to eliminate interference [5, 6, and 7].

Transforming large data into valuable information is the biggest challenge in Big Data. It is a difficult task and requires the development of new algorithms. One of the widely known techniques used for data analysis is Deep Learning. It is a powerful tool that requires large amounts of data to learn. It may seem that the Big Data and Deep Learning connection will be an ideal solution, but as always the problem lies in the details, and more specifically in the diversity of data.

An appropriate system for analyzing data streams should be created, the analysis of which will provide necessary insights and ensure the provision of valuable information not visible at first glance.

ICAIP 2019, November 8–10, 2019, Chengdu, China

ACM ISBN 978-1-4503-7675-4/19/11…$15.00
https://doi.org/10.1145/3373419.3373425

The article presents a project of such a system allowing Big Data processing and Deep Learning analysis.

2. BIG DATA ARCHITECTURES

2.1 Lambda Architecture

One of the two architectures used in Big Data systems is the Lambda architecture, which provides paral-lel processing of large data sets and the possibility of continuous access to them in real time. The idea be-hind this architecture is to create two separate flows, where one is responsible for data processing in batch mode, while the other one for accessing them in real mode [8]. A steady stream of data is directed to both layers (Figure 1).

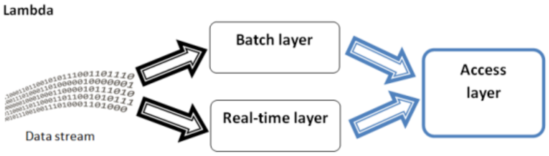


Figure 1. Idea of Lambda architecture.

In the batch layer, calculations are performed on the entire data set. It happens at the expense of time, but the data received in return contain the full history and high quality. It is assumed that the data set in the batch layer has an undivided form, which should only be expanded rather than delete data from it. In this way, you can ensure data consistency and access to historical data.

The real-time layer processes incoming data in real mode. The short time of access to data in this layer translates into the possibility of faster information retrieval. Unfortunately, the lack of access to historical data means that not all calculations are possible. Often the quality and faith-dignity of data from the real time layer is not as high as from the batch layer. The latter should be considered more reliable, but due to the longer time needed to load them, the real-time layer proves to be an invaluable help in order to guarantee the possibility of processing data in real mode.

The access layer is responsible for creating views based on the batch and real-time layers. The data is aggregated in such a way that the end-user sees it as a single, coherent whole. Views should be prepared in such a way as to ensure the possibility of performing all kinds of ad-hoc analyses, while enjoying fast access to data.

The Lambda architecture concept provides many advantages, primarily a perfect compromise between batch and real-time processing. The biggest and most often mentioned disadvantage is the need to maintain two independent applications - one for powering the batch layer, and the other for the real-time layer. The tools used in each layer are different, so it's hard to choose one that can be used for two purposes. Unfortunately, this architecture is more complicated and more expensive to maintain, so for our system we chose the second of the popular architectures used in Big Data - Kappa.

2.2 Kappa Architecture

Kappa's architecture appeared as a response to criticism related to implantation and maintenance of systems based on Lambda architecture [8]. The new architecture was based on four main assumptions:

1. Everything is a stream - the stream is an infinite number of completed data packages (batches), so that each data source can be a data stream generator.
2. Data is immutable - raw data is persisted in the original form and does not change, so that you can use them again at any time.
3. The KISS rule - Keep is short and simple. In this case, using only one engine for data analysis instead of several, as in the case of Lambda architecture.
4. The ability to restore data state - calculations and their results can be refreshed by retrieving historical and current data directly from the same data stream at any time.

It is critical for the above rules to ensure that the data in the stream remains unchanged and original. Without this condition being met, it is not possible to obtain consistent (deterministic) calculation results.

The idea of Kappa architecture is shown in Figure 2, and we may see a simplified structure in relation to Lambda architecture.

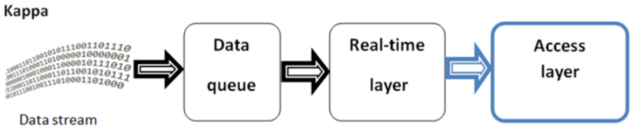


Figure 2. Idea of Kappa architecture.

Just like in the Lambda architecture, we have a Real-Time layer and an Access layer that performs the same functions here. In contrast, there is no layer of the Batch, which has become redundant because history can be reconstructed at any time from the data stream at the Access layer using identical data processing engine.

3. BIGDEEPEXAMINATOR SYSTEM PROJECT

The name of BigDeepExaminator system is an abbreviation of the names of the component system tech-nologies that is: Big Data,, Deep Learning and Examinator of data.

Elements of the system

The general scheme of the system is shown in the Figure 3. Main parts of the systems are:

1. System: A system for processing any type of data that is able to handle the system, having its own language and grammar, can learn, is intelligent and able to find solutions to new problems; universal for tasks requiring reasoning in a general sense, using the knowledge base; a system capable of learning and gathering knowledge, as well as data mining and extracting new knowledge from the knowledge possessed; the goal of the system is to gather knowledge, not data.
2. User: a person, device or system having the ability to communicate with the system, having the ability to handle system input / output streams. This applies to formulating questions, characterizing the analysed problem, and interpreting the results of the system's operation.
3. Data: a set of information in the form of a data stream as well as a file. The system should process data in both known formats and new types.
4. External commands: queries from the user in natural language, possibly as simple as possible, constituting correct sentences of a subset of natural language defined

in accordance with syntax - grammar, i.e. alphabet, and construction rules.

5. Methods library: a set of functions that will work on a set of external data provided to the system, based on algorithms of artificial intelligence, deep learning, application algorithms, data processing and analysis; this library will be used by the processing unit.

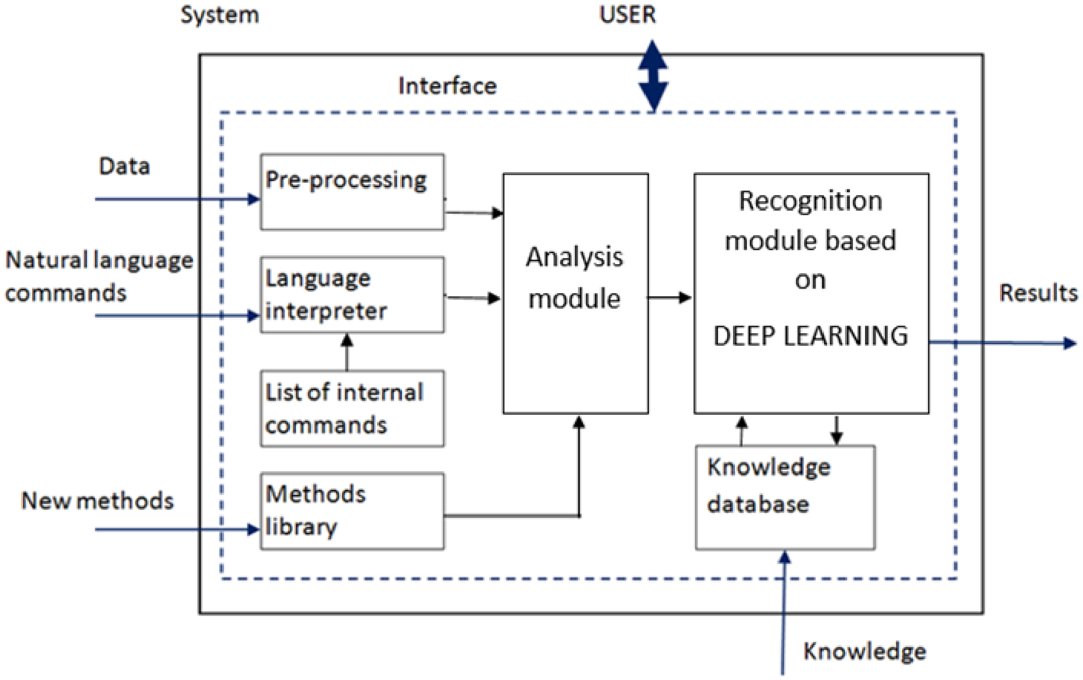


Figure 3. Idea of BigDeepExaminator system

6. Pre-processing: recognition and interpretation of input streams or streams in order to select optimal methods of information processing; recognition of data forms and their processing to the necessary internal form with the analysis of correctness and error signalling.
7. Language interpreter: program that verifies the correctness of external commands, intelligent, interactive parser and generator of executable code or an internal language suitable for the system to perform in the processing; the interpreter can be in the form of an agent or bot using preliminary analysis and algorithms based on fuzzy sets and artificial neural networks, speech recognition.
8. List of internal commands: a set of internal system instructions that are not available to an external user; using a library of methods, its own internal data format and its own language and grammar.
9. Knowledge database: Knowledge of the system stored in accordance with specific rules; objects and relations between them introduced from outside or deduced by the system and stored in the system in a specific way (human and system-readable) enabling quick access to them and their use in taking partial actions in solving problems and in requesting; knowledge base will use elements 3-9.
10. Analysis module: transformation of input streams into output streams based on elements 3-9; it is a part of the system that enables the data to be processed using all of its isolated elements; execution of the code or internal language generated by the language interpreter and features extraction for recognition module.
11. Recognition module: artificial intelligence sub-systems based on Deep Learning neural networks; recognizes patterns and interprets responses.
12. formulated by the user and presented in a way that is understandable to him; they also supply the knowledge base; this is every effect / result of the system operation.
13. System input: there are data on the system input - see point 3.
14. Additional data and information entered into the system: the system has the option of downloading additional external data in the form of knowledge or new methods and information about new data formats.
15. System output: a data stream obtained in the results of data processing by the system.

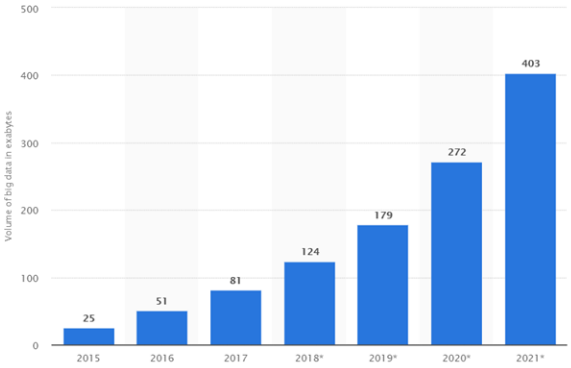
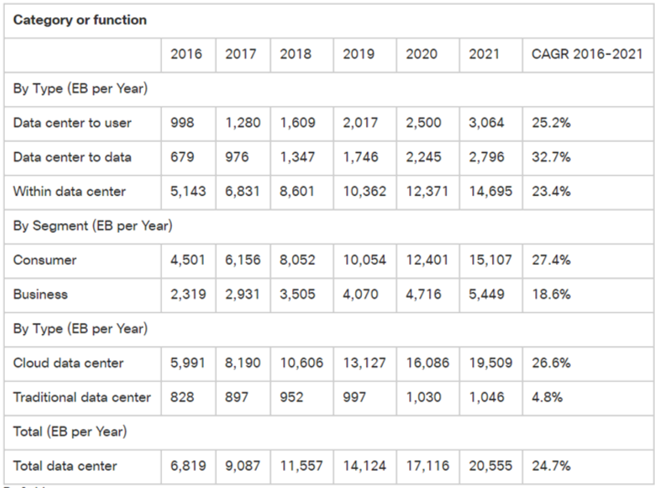


Figure 4. Volume of big data in data centre storage worldwide [11].

Table 1. Global data centre traffic, 2016–2021 [12]



Category or function							
	2016	2017	2018	2019	2020	2021	CAGR 2016-2021
By Type (EB per Year)							
Data center to user	998	1,280	1,609	2,017	2,500	3,064	25.2%
Data center to data	679	976	1,347	1,746	2,245	2,796	32.7%
Within data center	5,143	6,831	8,601	10,362	12,371	14,695	23.4%
By Segment (EB per Year)							
Consumer	4,501	6,156	8,052	10,054	12,401	15,107	27.4%
Business	2,319	2,931	3,505	4,070	4,716	5,449	18.6%
By Type (EB per Year)							
Cloud data center	5,991	8,190	10,606	13,127	16,086	19,509	26.6%
Traditional data center	828	897	952	997	1,030	1,046	4.8%
Total (EB per Year)							
Total data center	6,819	9,087	11,557	14,124	17,116	20,555	24.7%

16. The data will not be collected. They will be used to extract knowledge. The "raw" data being the basis of the knowledge discovery process is very often characterized by the following features: measurement errors, missing values in data, disruption during sampling, human mistakes. The system should be resistant to such errors. Therefore, replacing such data with empty values is more justified than with null values, because "zero" is also a value and can lead to incorrect conclusions being drawn.

Preliminary analysis of data should lead to the formulation of only such statistical statements regarding data for which the value of "true" is the invariant of correct transformations, e.g. median, mean or range. Therefore, the pre-processing unit should control the quality of data, because according to the GIGO principle (garbage in - garbage out) - entering erroneous data results in erroneous results - conclusions. Poor quality data make it difficult to draw correct conclusions and, as a result, knowledge exploration and rational decision making. Loaded data and dependencies derived from them can have serious consequences when it comes to formulating laws and rules.

Pre-processing should include data cleaning and transformation to prepare for exploration. It is estimated that the initial data processing is 70-80% of the knowledge discovery process. Data after verification of, for example, the range, will be converted to the internal format of meta-based tags. Thanks to this we will obtain a unified data structure. An additional benefit of this stage of data processing will be their standardization. A detailed description of the data format will be developed in further works and parallel processing of multiple data streams is also planned.

4. CONCLUSION

Big Data is a great challenge for scientists (Table 1. and Figure 4.). A large amount of data from a variety of sources is a factor that is a difficult problem to solve [13-16]. Adequate data conditioning has a decisive impact on the quality of results generated by data processing systems. New technologies of data collection and processing force interdisciplinary research and the need to combine existing solutions. Future large IT systems will be based on techniques that use Big Data, so new technologies should be developed that can process large amounts of data and extract useful knowledge from them, which will require artificial intelligence and Deep Learning techniques.

5. ACKNOWLEDGMENT

The project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019 - 2022 project number 020/RID/2018/19 the amount of financing 12,000,000 PLN..

6. REFERENCES

[1] Buhl, H., Röglinger, M., Moser F., Heidemann J.,: Big Data, Business & Information Systems Engineering, April 2013, Volume 5, Issue 2, 2013, pp 65–69.

[2] Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., Lapis, G.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, USA, 2012.

[3] Bekker A., 7 major big data challenges and their solutions, ScienceSoft, https://www.scnsoft.com/blog/big-data-challenges-and-their-solutions, April 2019

[4] Jinchuan, C., Yueguo, C., Xiaoyong, D., Cuiping, L., Jiaheng, L., Suyun, Z., Xuan, Z.: Big data challenge: a data management perspective, Frontiers of Computer Science, SP Higher Education Press, vol. 7, issue 2, 2013, 157-164.

[5] Katal, A., Wazid, M., Goudar, R.H.: Big Data: Issues, Challenges, Tools and Good Practices, 2013 Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, 2013, pp. 404-409.

[6] Doug, L.: Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies, META Group, Gartner, 2011.

[7] Maslankowski, J.: Analysis of the quality of data obtained from websites using Big Data solutions, Annals the Collegium of Economic Analysis, Issue 38, 2015.

[8] Marz N., Warren J.: Big Data Principles and Best Practices, Manning Publications Co.,2015.

[9] Zadeh, L.A.: Computing with Words: Principal Concepts and Ideas. Springer Publishing Company, Incorporated,2012.

[10] Schank, R.C.: Conceptual Information Processing, Yale University, New Haven, Connecticut, 1975

[11] https://www.statista.com, April 2019.

[12] https://www.cisco.com, April 2019.

[13] Bobulski, J.: Multimodal face recognition method with two-dimensional hidden Markov model, Bulletin Of The Polish Academy Of Sciences-Technical Sciences, vol. 65, issue: 1, 121-128.

[14] Bobulski, J.: Hidden Markov models for two-dimensional data, Advances in Intelligent Systems and Computing, 226, 2013, 141-149.

[15] Bobulski, J: HMM and WT fusion for face identification, Conference: 4th International Conference on Computer Recognition Systems (CORES 05) Location: Rydzyna Castle, Poland, May 22-25, 2005, Computer Recognition Systems, Proceedings, Advances In Soft Computing, 2005, 759-766.

[16] Bobulski, J.: Hidden Markov Models For Two-Dimensional Data, Conference: 8th International Conference on Computer Recognition Systems (CORES) Wroclaw, Poland, May 25-27, 2013, Proceedings Of The 8th International Conference On Computer Recognition Systems Cores 2013, Advances in Intelligent Systems and Computing vol. 226, 2013, 141-149