BIG DATA AND ARTIFICIAL INTELLIGENCE

Djuro Klipa, Igor Ristic, Aleksandar Radonjic, Ivan Scepanovic[1]

Abstract: The Big Data embodies a technology that permits the storage, processing, and management of broad and complex data sets in which traditional data processing applications are not applicable. These sets of data are usually characterized by a substantial volume of information that they carry, variety, versatility in terms of the format in which they are written, as well as a high-speed ingress which is often greater than the speed of processing. A particular challenge is the data which are coming from the Internet of Things (IoT) "world" that is constantly expanding and which already consists of several billion devices that can measure various parameters in the environment, communicate, process and transmit information. The data streams emanating from these devices are changing traditional approaches to data management and contribute to the emergence of the Big Data paradigm. This paper discusses the characteristics of the IoT infrastructure in terms of vast scale sensor applications and a possibility of connectivity of sensor networks, as well as various techniques of collection, storage, archiving and processing of the data on the cloud.

Key words: big data, data, information, database.

1. INTRODUCTION

The evolution of the mobile business, IoT and social media technology leads to a rapid increase in the volume of the data. The paradox is that public institutions and businesses have access to an oversize amount of data, however, do not use them wholesomely. A set of technologies for the achievement of the infrastructure and services required for reliable, distributed and scalable storage and use of large amounts of data is called big data (Radenkovic et al., 2015). Sensors and sensor networks are implemented to track different parameters in the environment and generate excessive amounts of data. In a large number of cases, it is necessary to collect and integrate the data from diverse sensor systems and then analyse them accordingly. Sensory data are analysed in real time to deliver forecasts in a timely manner and in a suitable format to the users. Various forms of mobile communication devices are increasingly implemented to access myriad of applications, especially via mobile devices. In recent years, more information has been produced than in the whole history (Zikopoulos et al., 2012).

The phenomenon of digital interconnection is unmistakably observable in contemporary business, but it is growing prominent in private life. Business has already been using the Intelligent production systems which independently reveal the optimization potential. The IoT device monitoring platforms are suitable to recognize issues or even preventively react while fully automated logistic processes are not rare any more. In our private life, the number of intelligent followers increases by our side: fitness bracelets, or the "smart home" control by smartphones, face recognition systems in public institutions to identify individuals under police warrant, are some examples of a number of technological applications at present. If a significant amount of the data, available through these forms of networking, has been adequately selected and appropriately used, it would be possible to make the

[1] **Djuro Klipa**, Faculty of Management, Assistant Professor, **Igor Ristic**, Faculty of Management, Assistant Professor, **Aleksandar Radonjic**, Faculty of Management, Associate Professor, **Ivan Scepanovic**, Faculty of Management, Assistant Professor,
✉ corresponding author: risticig@famns.edu.rs
✉ djuro.klipa@famns.edu.rs; aleksandar.radonjic@famns.edu.rs; scepanovic@famns.edu.rs

processes more efficient through the IoT, spread knowledge and act preventively or predictively.

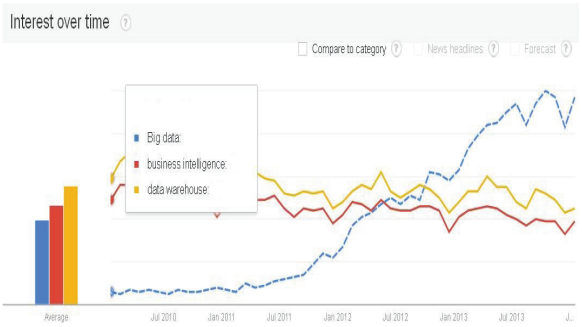

Figure 1. Interest in the Big Data term

2. THE BIG DATA TERM

The most widely applied the Big Data definition has been inferred from the Meta Group analysis which includes the information resources of considerable size, high speed and wide variety of data that require new and innovative methods of processing and optimization of Information, improvement of the insight into the content of the data and decision-making based on the processed data (Laney, 2001).

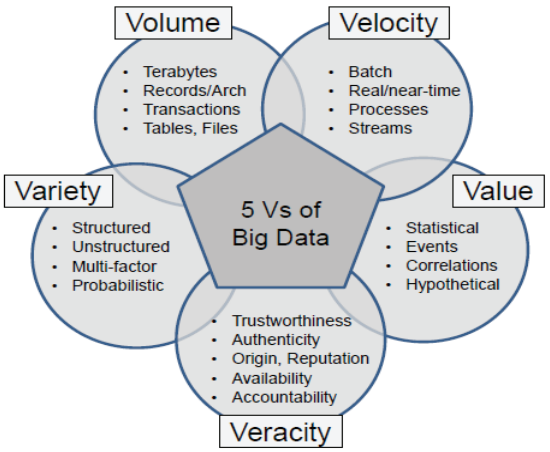

Figure 2. 5Vs of Big Data

Data in the IoT solutions

The Big Data term has often been explained by usage of the five 'V' models of which the main features are:

- Volume - exponentially increases having in view that the main sources of oversize amounts of data are sensors, mobile phones, transaction data, records,

video/audio recordings, social networks, etc.

- Variety - the data are by nature diverse and often unstructured and therefore do not fit well into the database (for example, multimedia data, documents, data collected from various sensors, logs, etc.).
- Velocity - the currentness of the data is often less than 1 second and therefore the data must be processed and analysed in real (or near to real) time, i.e. from the moment they begin to be collected
- Value - through efficient data mining and analytics a huge amount of data being collected during the business operations can be wholesomely used with an aim of improvement and efficiency thereof.
- Veracity – integration of multitudinous data systems brings forward veracity, i.e. correctness and accuracy. In the background of any data management appears the basic doctrine of reliability, availability, and originality of the data.

Which of the data may be gathered via IoT solutions has been presented in the Figure 3.

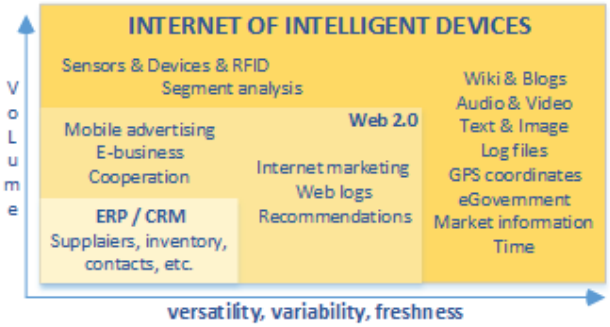

Figure 3. Data in the IoT solutions

3. BIG DATA INFRASTRUCTURE

Data are collected from a large number of sources and their volume is constantly increasing. The collected data are transferred into the large data file system by applying proper extraction, transformation and load processes. The data contained in the big data file system are analysed in real time by use of diverse tools, techniques and algorithms. The results of the data analysis are presented to users in the appropriate visual format (Figure 4).

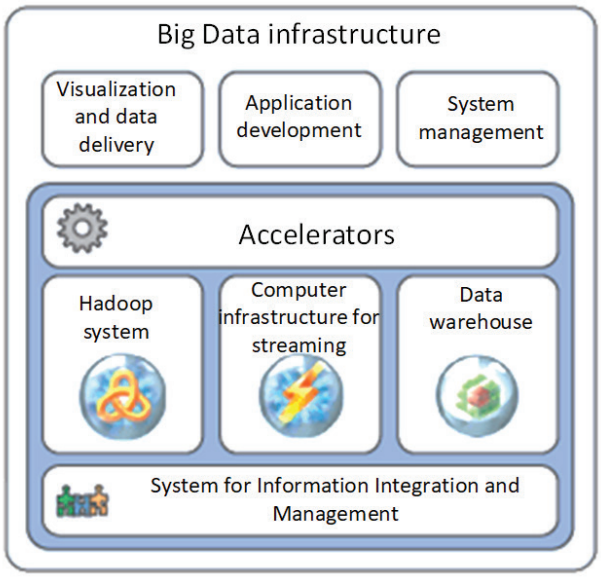

Figure 4. Big Data Infrastructure

4. DISTRIBUTED FILE SYSTEMS

In order to reliably and scalably store substantial amounts of data, it is necessary to provide the storage and management of files in a distributed environment. The distributed file systems ensure easy access to files in different geographic locations, file replication between servers and data compression optimized for transfer through a network with limited bandwidth.

Implementation of the distributed file systems

Some of the implementations of distributed file systems are:

- Google File System (GFS) – efficient and reliable access to the data in huge clusters, primarily to support Internet search services.
- Hadoop distributed file system (HDFS) - designed to store big files, a gigabyte or terabyte in size.
- GlusterFS - is frequently used to implement a cloud computing infrastructure and services.

Non-relational databases

IoT applications generate considerable amounts of data, and therefore scalable, distributed and reliable database models are required to be implemented:

- The Key-data model in which all data are stored in two columns (one column is the key and the other data is the value) and the data does not need to be atomic and its changes are monitored by a time stamp (Example: Amazon Dynamo)

• The BigTable (a relational data model) is abstracted by a large global table wherein each table element has a row identifier, a column identifier and a time stamp (Examples: Google Big Table and Cassandra).
• A document model with a central concept document that represents a group of semi-structured data being stored together with the metadata, providing flexibility, easy integration with applications, and reducing the size of the database (Examples: XML databases, MongoDB).
• A graph model that uses a graph structure for modelling semantic links among the graph nodes (usually people, businesses, etc.) other concepts about which the data are stored (Examples: Neo4j, Facebook uses the Graph Search mechanism to set up a query in natural language and search for users and connections among them.

Searching data in the Big Data

In 2010, Google patented the MapReduce algorithm (Figure 5) to search data (pairs: key, data - (k, v)) in two steps:
• The first step (Map) - parallel and distributed processing of each ordered pair from the first domain: Map (k1, v1) and selection of only those pairs that are mapped to pairs from another domain: sheet (k2, v2).
• The second step (Reduce) - parallel and distributed pair processing is performed (k2, sheet (v2)), which, based on the criteria for reporting, maps these pairs into the search result: sheet (v3).

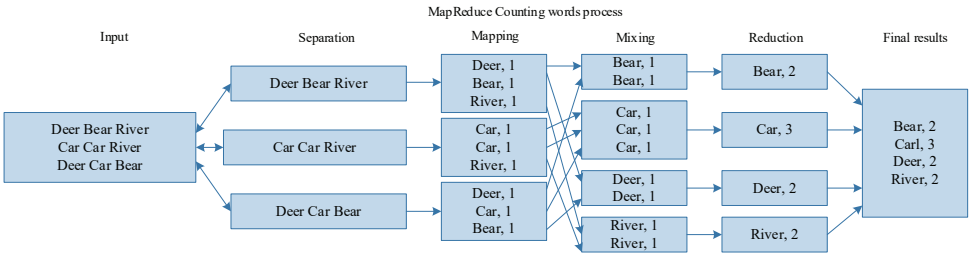

Figure 5. MapReduce algorithm

5. DATA ANALYTICS IN IoT

The real-time monitoring and management in an enormous IoT systems consisting of hundreds of thousands of sensors, creates problems such as:

- detection of risky devices in the system maintenance phase and their repair prior to being really broken,
- optimization of the functioning of smart devices - configuration, adjustment of interaction among end users, optimization of energy consumption,
- analysis of devices' defects and fixing in the upcoming versions.

The most important feature of proper tools for processing unstructured data is to be relatively easy to use and have the ability to quickly process the data with the tendency to process the data in real time.

Big Data Analytics

In the Big Data analytics (Figure 6), various techniques are applied, such as:

1) Cluster analysis – the method for determining relatively homogeneous groups of objects within a heterogeneous population.

2) Crowdsourcing – the technique for collecting data by a large group of people or communities through a media such as the web.

3) Joining Rules – the techniques to detect links between seemingly unrelated data in oversize databases.

4) Classification – the process of organizing information into categories (classes) so that the data can be more clearly analysed or understood.

5) Machine learning – constructing algorithms and computer systems that are capable to adapt to new circumstances and learn from the experience.

- Segment analysis – the application for language processing, identification, and extraction of information from the textual material.
- Integration of the data – the techniques that integrate and analyse the data from multiple sources aiming to develop more efficient and more precise data processing.
- Genetic algorithms – a stochastic search method that mimics the biological process of evolution.

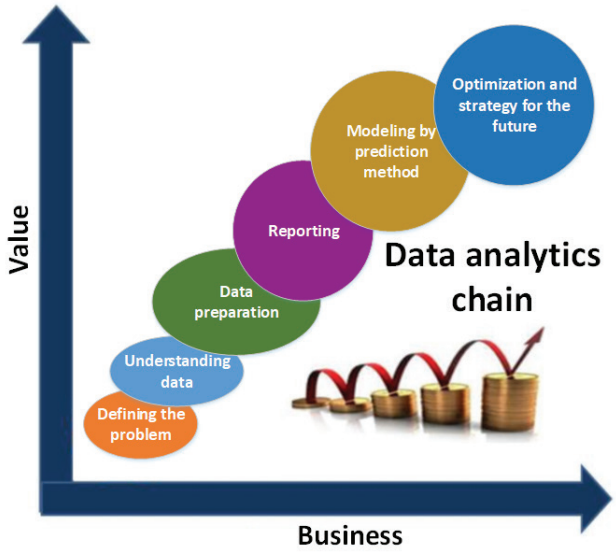

Figure 6. Big Data analytics

Hadoop is a software framework (Figure 7) of the open code for storage, search and analysis of huge amounts of the data written in the Java programming language. Hadoop is designed for appropriate support for the so-called batch data processing and as such is not suitable for real-time data processing

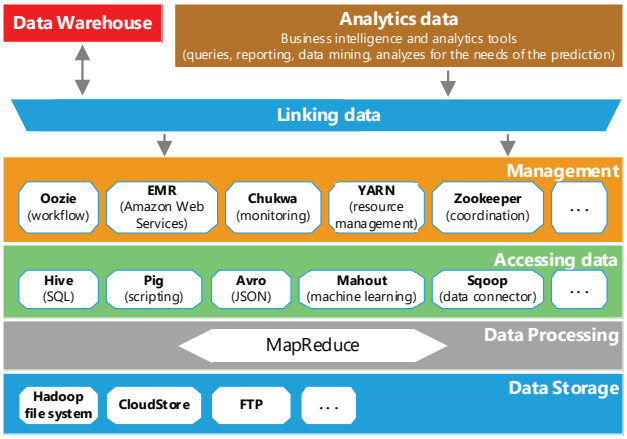

Figure 7. Hadoop system

6) Neural networks – can process the data in parallel of which components are independent of each other.
7) Network analysis – the techniques used to detect connections among at-first-glance unconnected nodes in a graph or network.
8) Optimization – a set of numerical techniques for the reorganization of complex systems and processes to improve their performance according to one or more criteria.

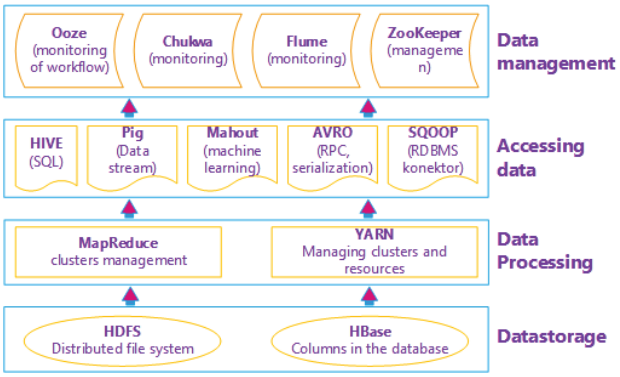

Figure 8. Hadoop framework for Big Data

Components of the Hadoop system

The basic elements of the Hadoop system (Figure 8) are:

1) Hadoop Common packages to support other modules

2) Hadoop Distributed File System, which provides applications to access data fast.

3) Hadoop YARN module for managing resources in a cluster and for managing the execution of tasks.

4) Hadoop MapReduce for parallel processing of voluminous datasets using the MapReduce algorithm.

Other important Hadoop projects and modules are:

1) Ambari web tool for tracking and managing Hadoop system
2) Avro system for data serialization
3) Hcatalog storage management system
4) Hbase scalable distributed database for storage of the structured data
5) Hive system for the data warehousing and ad hoc data queries
6) Mahout Library for Machine Learning and Data Mining.
7) Oozie workflow management (workflow, scheduling).
8) Pig framework for performing parallel calculations
9) Sqoop data transfer tool between Hadoop and relational databases
10) Spark software model for application support such as ETL, machine learning, streaming, etc.
11) ZooKeeper Service for Coordination of Distributed Application

6. ARTIFICIAL INTELLIGENCE

Artificial intelligence is a broader concept of machine learning addressing the use of computer systems to mimic human knowledge functions. Artificial Intelligence is the process wherein machines perform tasks by complying with the algorithms based on an “intelligent” approach. Machine learning is a subset of Artificial Intelligence, its focal point being to ensure the machine capability to obtain a set of data and learn, alter the algorithms while learning more about the information they are processing. Deep learning is an additional level of more complex analytics and considered as a subset of machine learning. Current solutions that can analyse and interpret the IoT data, artificial intelligence and machine learning, can also extract more advanced insights into the data and more rapidly extract the relationships among these metadata. Artificial Intelligence identifies uncommon tendencies more precisely, thus eliminating the necessity to screen the data by the use of advanced techniques. The business benefits of IoT are wholesomely established, however, the enormous benefits of intelligent things are just emerging. Companies may expect new and innovative ways of operation.

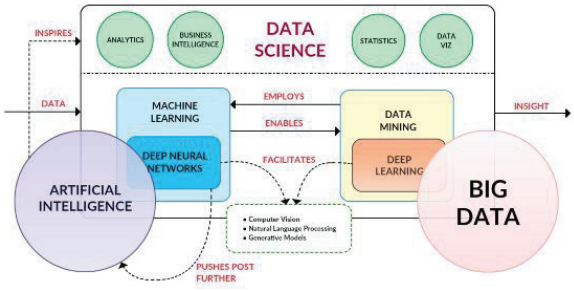

Figure 12. Artificial intelligence and Big data

7. BIG DATA APPLICATION

Business

Internet intelligent devices technologies ensure that businesses have a huge amount of data on clients: personal data, movement data, data on habits, interests, needs, and others. As data sources, the data collected from social networks, mobile phones, and other sources are used. By efficacious application of these data, it is possible to better anticipate the needs of clients, create personalized products and services, and increase customer satisfaction.

Medicine

The usage of sensors to monitor the medical condition of patients is on the rise. Storage and analysis of the collected data can be applied for diagnostics improve, patient monitoring, predicting reactions to therapy, analysis of biochemical processes in the body, provision of remote health services, etc.

Transport and Traffic

Smart train control systems can automatically adjust the speed of a train based on the input data such as weather conditions, terrain topography, distance from a destination point, distance from other trains, and further on. In the future, it is expected that the use of the large data technologies applied in smart cars will ensure safe driving without a driver.

Meteorology

Sensors may be used for the collection of meteorological data (Figure 9). Expeditious analysis of the historical data in combination with the current data can contribute to the better prediction of weather disasters, such as hurricanes, and better protection of populations in vulnerable areas.

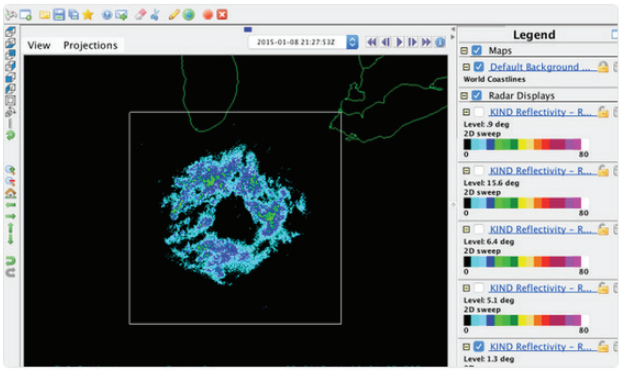

Figure 9. Application in meteorology

Education

Smart classrooms in educational institutions contain a greater number of sensors, video cameras, microphones, and other interactive devices. The data collected from intelligent devices can be stored in the information system of the educational institution and used for monitoring students' activities, measuring the quality of tuition and student satisfaction, determining the typical behaviour of students and creating personalized services, supporting the scientific and research work of teachers and students, etc.

Scientific research

Collecting data from the universe is often based on remote sensing technologies (Figure 9). In surveys of the space, the substantial amounts of the data are collected on a daily basis whilst numerous algorithms for analysing enormous amounts of data are being developed in order to analyse the collected data.

Smart cities

The integration of the data collected through different smart city systems (Figure 11) may improve the quality of the city administration services. The Big Data and internet intelligent devices can be implemented into big city systems that have ICT preconditions to adopt these principles. An example of possible implementation is the city of Belgrade which owns the ICT infrastructure, has a considerable number of business systems under its jurisdiction and has approximately two hundred electronic services offered to its citizens. With the achievement of the aforementioned advanced options, it is necessary, for the future BelgradeSmartCity Intelligent System, to define city registers that will structure the generated data of internal systems (at all app/sensor levels) into relevant information to manage and optimize the entire city system. The parameters that will consequently be generated as a Big Data infrastructure can be introduced as a part of the city resource management system whilst the designed benefit would be cost reduction, optimal use of resources, generation of new services

for the needs of citizens and the economy, as well as generation of new and advanced Big Data analytical tools.

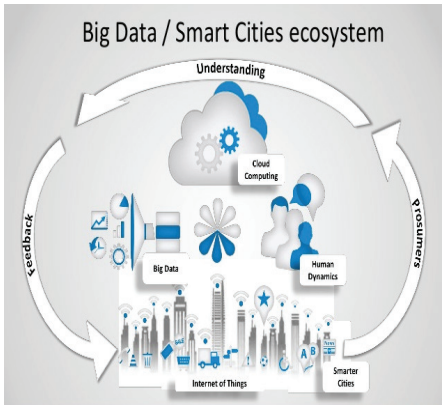

Figure 10. Application of the smart city systems

8. CONCLUSION

The Big Data represents a new concept in storing, searching and analysing substantial amounts of data. The data collected by sensors are everywhere, they are often different in structure and must be used in real time. The essence of Big Data technology is not to provide access to a sizeable amount of the data from sensor networks but in activities that contribute to their usability. The main objective is to develop efficient and reliable methods for obtaining usable information from the data gathered in real time. The application is possible in various fields: business, medicine, scientific research, smart cities, meteorology, transport, and traffic, etc. The Big Data is a building block in business intelligence that establish new values and new knowledge. With the application of business intelligence, we improve the making of business decisions. The Big Data is not a substitute for the business intelligence system but it can introduce a new value through existing systems (Pavlovic, et al., 2014).

References

B. Radenković, M. Despotović-Zrakić, Z. Bogdanović, D. Barać, A.Labus, Electronic business, FON, 2015

Đ. Klipa , R. Dragović, Security and technological aspects of social networks

D. Laney, „Application Delivery Strategies“, Meta Group, 2001

K.Simić, M.Despotović-Zrakić, Ž.Bojović, B.Jovanović, Đ.Knežević „ A platform for a smart learning environment, Facta Universitatis-series:Elektronics and Energetics,2016

P.C. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, J. Giles, „Harness thePower of Big Data“, McGrawHill, 2012

R. Dragović, D. Dragović, B. Perović, Đ. Klipa, Strategic management in the judiciary based on decision support systems, Yuinfo 2013

R. Dragović, M. Ivković, B. Perović, Đ. Klipa, Dataveillance and data mining as a technology support to the process of investigation, Telfor 2011, IEEE

R.Pavlović, R.Dejanović „Big Data i Business inteligence„ INFOTEH Jahorina„ 2014
S. Hansen How Big Data Is Empowering AI and Machine Learning, 2017

Received: 15-10-2022
Accepted: 03-11-2022