Acta Informatica Pragensia, 2015, 4(2): 108–121
DOI: 10.18267/j.aip.64
Peer-reviewed paper

AHP Model for the Big Data Analytics Platform Selection

Martin Lněnička*

Abstract

Big data analytics refers to a set of advanced technologies, which are designed to efficiently operate and maintain data that are not only big, but also high in variety and velocity. This paper analyses these emerging big data technologies and presents a comparison of the selected big data analytics platforms through the whole data life. The main aim is then to propose and demonstrate the use of an AHP model for the big data analytics platform selection, which may be used by businesses, public sector institutions as well as citizens to solve multiple criteria decision-making problems. It would help them to discover patterns, relationships and useful information in their big data, make sense of them and to take responsive action.

Keywords: AHP model, Big data analytics, Big data life cycle, Platform selection, Decision-making.

1 Introduction

The period of business decision-making processes based on simple reports generated from filtered, preselected and structured data is coming to an end. The emphasis is no longer placed on counting and comparing key performance indicators, but rather on finding a statistically significant connection between these indicators and related datasets. These have much less structure, or more complex structure, than the traditional models. Consequently, a novel view of large amount of structured as well as unstructured and semi-structured data can help gain a sustainable competitive advantage for businesses.

Although big data are characterized in terms of volume, velocity and variety, it is more practical to define big data in the context of management relationships and analytics and how they impact business decisions (Loshin, 2013). In many real-world situations, it is important to make accurate predictions based on the available information. Today, software architecture constrains the achievement of quality attributes such as high performance, usability and maintainability of a system (Daniluk, 2012). Different industry sectors and users will also tend to want to ask different types of questions. Big data is poised to add greater value to businesses, to solve science problems, to support modern medicine, etc. (Tien, 2013).

Thanks to developments in both hardware and software, the technology to store, interrogate and analyse data is improving rapidly (Lake & Drake, 2014). However, challenges vary for different applications as they have differing requirements of consistency, usability, flexibility,

* Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentska 84, 532 10 Pardubice, Czech Republic
✉ martin.lnenicka@gmail.com

compatibility or data flow (Shamsi, Khojaye, & Qasmi, 2013). Thus, to perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes imminent and choosing the right platform becomes a crucial decision. Researchers have been working on building novel data analysis techniques for big data more than ever before, which has led to the development of many different algorithms and platforms (Singh & Reddy, 2014).

As a result, the main aim of this paper is to propose an Analytic Hierarchy Process (AHP) model for the big data analytics platform selection based on the three defined use cases. Accordingly, some of the various big data analytics platforms are discussed in detail and their applications and opportunities provided in several big data life cycle phases are portrayed.

2 Research Methodology

The main goal of this paper is to propose the AHP model for the big data analytics platform selection to help businesses as well as public sector institutions and citizens, so they can make an informed decision. In addition, this paper offers added value by means of a classification of existing big data analytics platforms based on the big data life cycle.

The literature reviewed is selected based on its novelty and discussion of important topics related to big data analytics and platforms comparison in order to serve the purpose of this research. Method of the AHP is used to compare the defined criteria. The AHP is a multiple criteria decision-making (MCDM) tool that has been applied to many practical decision-making problems (Saaty, 1990; Saaty, 2008). It has been used in almost all the applications related with decision-making, including the capability of handling many criteria, mainly if some of the criteria are qualitative, as well as the evaluation of large sets of alternatives. This proves the versatile nature of the AHP, enabling to arrange the different alternatives according to the requirements of the decisions to be taken (Vaidya & Kumar, 2006).

3 Literature Review

3.1 Big Data and Big Data Analytics

For the most part, in popularizing the big data concept, the analyst community and the media have seemed to latch onto the alliteration that appears at the beginning of the definition (Loshin, 2013). Though, big data are attributed to have such characteristics as volume, velocity and variety (3V). Other authors, such as Demchenko, Grosso, Laat, and Membrey (2013) or Loshin (2013) added additional Vs (value and veracity) or variability respectively, intended to capitalize on an apparent improvement to the definition.

Big data fundamentally mean datasets that could not be perceived, acquired, managed and processed by traditional technologies within a reasonable time, which indicates that efficient tools and platforms together with suitable methods have to be developed to analyse and process big data (Chen, Mao, & Liu, 2014).

The popularity of big data analytics platforms, which are often available as open-source, has not remained unnoticed by big companies. Google uses MapReduce for PageRank and inverted indexes. Facebook uses Apache Hadoop to analyse their data and created Hive. eBay uses Apache Hadoop for search optimization and Twitter uses Apache Hadoop for log file analysis and other generated data (Saecker & Markl, 2013). Also various platforms have been developed mainly by the industry to support big data analytics, including Yahoo's PNUTS, Microsoft's SCOPE, Twitter's Storm, etc. (Guo, 2013; Zhao, Sakr, Liu, & Bouguettaya,

2014). Although Apache Hadoop has been very successful for most of the big data problems, it is not an optimal choice in some situations. Mostly because of the drawbacks such as special data processing requirements, difficulties to configure and manage the Hadoop cluster, etc. (Guo, 2013).

Chen et al. (2014) introduced the general background of big data and review related technologies, such as cloud computing, Internet of Things, data centres and Apache Hadoop. They emphasized the importance of big data analysis through the six key technical fields: structured data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis and mobile data analysis. Also Elgendy and Elragal (2014) analysed some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains. They concluded that big data analytics can be applied to leverage business changes and enhance decision-making by applying advanced analytics techniques on big data and revealing hidden insights and valuable knowledge. Vossen (2014) studied and reviewed the issues, techniques and applications of big data, with an emphasis on business intelligence architectures. Saecker and Markl (2013) then presented an overview of existing processing approaches to address the problems arising from big data analytics when using modern hardware architectures. They claim that models of distributed systems and modern hardware architectures employ a parallel model. Loshin (2013) suggested that as a way to properly ground any initiatives around big data, one initial task would be to evaluate the fitness of business as a combination of the five factors: feasibility, reasonability, value, integrability and sustainability.

Kaisler, Armour, Espinosa, and Money (2013) focused on the issues and challenges of big data. They suggested that there are three fundamental issue areas that need to be addressed in dealing with big data: storage and transport issues, management issues and processing issues. Demchenko et al. (2013) introduced the big data life cycle management model that includes all the major stages and reflects new challenges and specifics in the big data management. Data integrity, access control and accountability must be supported during the whole big data life cycle.

Chen et al. (2014) and Che, Safran, and Peng (2013) reviewed of state of the art frameworks and platforms for processing and managing big data as well as the efforts expected on big data mining. Also Singh and Reddy (2014) provided an in-depth analysis of different platforms available for performing big data analytics and assessed the advantages and drawbacks of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support. Lee et al. (2011) provided a timely remark on the status of MapReduce studies and related work to aim at improving and enhancing the MapReduce framework. They mostly focused on the overcoming of this framework's limitations. Zhao et al. (2014) provided a comprehensive survey for a family of approaches and mechanisms of large scale data processing mechanisms that have been implemented based on the original idea of the MapReduce framework. Sakr, Liu, and Fayoumi (2013) surveyed the MapReduce framework's variants and its extensions for large scale data processing.

3.2 Big Data Life Cycle and Related Platforms

A life cycle of big data can be divided in four phases (Chen et al., 2014): data generation, data acquisition, data storage and data analysis. Data generation and data acquisition are then an exploitation process, data storage is a storage process and finally data analysis is a production process that utilizes the raw material to create new value. Data analysis is the final and the

most important phase in the life cycle of big data, with the purpose of extracting useful values and providing decisions (Chen et al., 2014). Tien (2013) also identified four components of big data: acquisition (including data capture), access (data indexing, storage, sharing and archiving), analytics (data analysis and manipulation) and application (data publication).

The big data analytics process should support the whole data life cycle, which may include identifying the data analytics problems, collecting datasets, data analytics and also supporting tools. Therefore, the comparison of the related platforms is focused on these requirements. The basic phases were originally defined in Lněnička and Komárková (2014) and later modified based on the literature review presented above. These are: data acquisition, data aggregation and transfer, data storage and search (file systems, databases and programming languages) and data analysis (business intelligence and data mining).

3.2.1 Data Acquisition

Most of the big data acquisition comes from internal sources within the business, i.e. relational databases, data warehouses and file systems. Obviously, some of the data are structured, but many very important components are semi-structured or unstructured (Loshin, 2013). The unstructured data are more complex and cannot be compiled in older database format. Externally sourced data are mostly unstructured, they come from the Internet such as social data, spatial data, news data or public sector's data. New opportunities are also opening up in the form of open data and linked data, which may be relevant in meeting the needs of the business. These data can be found at open data portals. The related benefits and risks have recently been presented in Lnenicka (2015).

3.2.2 Data Transfer and Aggregation

The big data transfer requires specified tools to move large amount of data in the most efficient, scalable and also secure way possible. Also the big data aggregation is then needed for efficient use within big data analysis platforms, which are described below. While the market for solutions used to aggregate data from multiple sources is relatively limited, it is also characterized by a variety of very different approaches. These tools with their various inputs and outputs require explanation, which affects both the competitive strength and the portfolio attractiveness ratings. Most of these tools are closely connected with the concrete platform. Therefore, the following tools are mostly focused on the Apache Hadoop ecosystem and the MapReduce framework. These are e.g. Avro, Flume, Chukwa, Splunk, Sqoop or Tika.

3.2.3 Data Storage and Search

At the core of any big data environment are the database engines, which contain the collections of data relevant to the business. These engines need to be fast, scalable, and rock solid. They are not all created equal, and certain big data environments will fare better with one engine than another, or more likely with a mix of various engines (Cattell, 2011; Loshin, 2013).

Madden (2012) discussed the differences between traditional databases and large-scale data management platforms and concluded that these databases don't solve all aspects of the big data problem. Traditional data management and analysis systems are based on the relational database management system (RDBMS). However, such RDBMSs only apply to structured data, other than semi-structured or unstructured data (Chen et al., 2014). Databases to manage data of this size are known as NoSQL databases. There are several solutions, ranging from distributed systems and massive parallel processing databases for providing high query

performance and platform scalability, to these non-relational and in-memory databases, which have been used for big data (Elgendy & Elragal 2014; Guo, 2013). Cattell (2011) examined a number of scalable SQL and NoSQL databases to provide a comprehensive survey. Chen et al. (2014) identified three main NoSQL databases: key-value, column-oriented and document-oriented databases. Based on the literature review, the author updated this categorization by adding graph databases and object databases focused on big data.

Big data needs the storage of a massive amount of these data, thus, this makes it a necessity for advanced storage infrastructure, which is designed to scale out on multiple servers, often with the use of cloud computing technologies. There are several file systems, which can be used together with the concrete platform, e.g. Hadoop Distributed File System, Google File System, GlusterFS, Quantcast File System, Ceph, Lustre, XtreemFS or MooseFS.

To search big data in the database, Lucene (high-performance, full-featured text search engine library written entirely in Java), Solr (enterprise search platform, which is highly scalable, supporting distributed search and index replication engine) or Elasticsearch can be used.

3.2.4 Data Analysis

The most widely used big data analytics platform is Apache Hadoop (Elgendy & Elragal 2014; Zhao et al., 2014). It is an enabling technology for working with huge datasets that provides both distributed storage and computational capabilities (Guo, 2013). Nowadays, Apache Hadoop is consist of tens related projects such as Phoenix (a relational database layer over HBase), Drill and Hive (SQL query and manage engines for Hadoop and NoSQL), Zookeeper (a centralized service for maintaining configuration information, naming, providing distributed group services and synchronization), HCatalog (a storage management layer for Hadoop that enables users with different data processing tools), Oozie (a workflow scheduler system to manage Hadoop jobs), real-time distributed systems Kafka, Spark and Storm, etc. Compared with the open source Hadoop releases, enterprise Hadoop distributions are easy to configure, adopt and maintain, and sometimes new features are added. Some of these platforms, while highly relevant to big data, are not big data specific and have been around for a while (Zhao et al., 2014).

The other vendors include Actian, Infobright, Kognitio and Platfora, which have centred their big data stories around database management systems focused entirely on analytics rather than transaction processing. Cloudera, Hortonworks, Pivotal, MapR, and others are working on ways to do SQL analysis, in-memory analysis, and even streaming analysis on top of Apache Hadoop. Other platforms that do not follow the MapReduce / Hadoop route are e.g. GridGrain, High Performance Cluster Computing (HPCC), Sector / Sphere, SCOPE / Cosmos or Dryad (Lněnička & Komárková, 2014).

The most widely used business intelligence platforms are, e.g.: BIRT, Jaspersoft, OpenI, Palo Suite / Jedox, Pentaho, SpagoBI or Talend. The most widely used big data mining platforms are e.g. Giraph, GraphLab, IKANOW, KEEL, KNIME, Apache Mahout, Orange, PEGASUS, RapidMiner, Rattle, SAMOA, SPMF or Weka. These are mostly offered in a community open source edition as well as under several commercial editions with broad support for various databases and data sources, including NoSQL and other big data sources.

3.3 Hardware, Software and Services Evaluation and Selection

Whatever the claims of hardware manufacturers and software suppliers, the performance of hardware and software must be demonstrated and evaluated based on the various attributes of quality. Large companies frequently evaluate proposed hardware and software using the

processing of special benchmark test programs and test data (Marakas & O'Brien, 2013). It requires a series of decisions based on a wide range of factors and then each of these decisions have considerable impact on the evaluation of performance, usability and maintainability for overall success of the most suitable platform selection (Daniluk, 2012).

Benchmarking simulates the processing of typical jobs on several computers and evaluates their performances. Users can then evaluate test results to determine which software package displayed the best performance characteristics. Notice that there is much more to evaluating hardware than determining the fastest and cheapest computing device. As an example, the question of obsolescence must be addressed by making a technology evaluation. The factor of ergonomics and social perspective is also very important. Ergonomic factors ensure that computer hardware and software are user-friendly, that is, safe, comfortable, and easy to use (Marakas & O'Brien, 2013). Bengtsson and Bosch (1998) evaluated the software platform quality attributes specifically for maintainability. The most useful method for maintainability is change scenario method as compared to other methods such as simulation, mathematical modelling and experience-based assessment. Connectivity is another important evaluation factor, because so many network technologies and bandwidth alternatives are available to connect computer systems to the Internet, intranet and extranet networks (Marakas & O'Brien, 2013).

The evaluation has a great impact on the quality of attributes. Valacich, George, and Hoffer (2012) proposed several the most common criteria to choose the right platform. These are: cost, functionality, efficiency, vendor support, viability of vendor, response time, flexibility, documentation and ease of installation. Lake and Drake (2014) emphasize the importance of the computational complexity factor and the increased efficiency of algorithms in the big data era. Marakas and O'Brien (2013) propose these hardware evaluation factors:

- Performance – What is its speed, capacity, and throughput?
- Cost – What is its purchase price? What will be its cost of operation and maintenance?
- Reliability – What is the risk of malfunction and what are its maintenance requirements? What are its error control and diagnostic features?
- Compatibility – Is it compatible with existing hardware and software? Is it compatible with hardware and software provided by competing suppliers?
- Technology – In what year of its product life cycle is it? Does it use a new untested technology, or does it run the risk of obsolescence?
- Ergonomics – Has it been "human factors engineered" with the user in mind? Is it user-friendly, designed to be safe, comfortable, and easy to use?
- Connectivity – Can it be easily connected to wide area and local area networks that use different types of network technologies and bandwidth alternatives?
- Scalability – Can it handle the processing demands of a wide range of end users, transactions, queries, and other information processing requirements?
- Software – Are system and application software available that can best use hardware?
- Support – Are the services required to support and maintain it available?

They also defined these software evaluation factors (Marakas & O'Brien, 2013):

- Quality – Is it bug-free, or does it have many errors in its program code?
- Efficiency – Is the software a well-developed system of program code that does not use much CPU time, memory capacity, or disk space?
- Flexibility – Can it handle the business processes easily, without major modification?
- Security – Does it provide control procedures for errors, malfunctions, improper use?

- Connectivity – Is it Web-enabled so it can easily access the Internet, intranets, and extranets, on its own, or by working with Web browsers or other network software?
- Maintenance – Will new features and bug fixes be easily implemented by software developers?
- Documentation – Is the software well documented? Does it include help screens and helpful software agents?
- Hardware – Does existing hardware have the features required to best use this software?
- Other Factors – What are its performance, cost, reliability, availability, compatibility, modularity, technology, ergonomics, scalability, and support characteristics?

Traditional evaluation methods often focus only on the system functionality or on a single non-functional requirement, e.g. high-performance, real-time or reusable systems (Bengtsson & Bosch, 1998; Daniluk, 2012). Therefore, it is necessary to propose a robust model for the big data analytics platform selection.

3.4 Multiple Criteria Decision-Making and Analytic Hierarchy Process

Real-world decision-making problems are complex and no structures are to be considered through the examination of a single criterion, or point of view that will lead to the optimum and informed decision (Vaidya & Kumar, 2006; Zavadskas & Turskis, 2011). MCDM offers a lot of methods that can help in problem structuring and tackling the problem complexity because of the multi-dimensionality of the sustainability goal and the complexity of socio-economic, environment and government systems. Therefore, Zavadskas and Turskis (2011) present a thorough historical review and classify and illustrate the primary steps of MCDM methods. MCDM can be roughly separated into Multi-Objective Decision-Making (MODM) and Multi-Attribute Decision-Making (MADM) components. MODM then includes Multiple Objective Programming (MOP), Goal Programming (GP) and compromise solution methods. These problems can be solved using many methods including single level, fuzzy, multi-stage and dynamic methods. MADM includes structure relation methods (e.g., Interpretive Structural Modelling (ISM), Decision Making Trial and Evaluation Laboratory (DEMATEL) or fuzzy cognitive map), weight analysis (e.g. AHP, Analytic Network Process (ANP) or entropy measure) and performance aggregated methods (e.g. Simple Additive Weight (SAW), Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) or grey relation for additive types and fuzzy integral for non-additive types) (Liou & Tzeng, 2012).

The AHP is a MCDM tool that has been used in almost all the applications related with decision making (Vaidya & Kumar, 2006). The AHP is a powerful, flexible and widely used method for complex problems, which consider the numeric scale for the measurement of quantitative and qualitative performances in a hierarchical structure (Saaty, 1990). This is an Eigen value approach to the pairwise comparisons. It is one of the few MCDM approaches capable of handling many criteria (Zavadskas & Turskis, 2011; Liou & Tzeng, 2012). The most important characteristic of the AHP is combining knowledge, experience, individual opinions and foresights in a logical way (Vaidya & Kumar, 2006).

A case study where the AHP method was employed to support the software selection can be found in Karaarslan and Gundogar (2009), Lai, Wong, and Cheung (2002), Silva, Goncalves, Fernandes, and Cunha (2013) or Wei, Chien, and Wang (2005). Authors mostly focused on the business environment and their findings may be applicable in the following model. In the Czech Republic, the use of the AHP is promoted by e.g. Brožová, Houška, and Šubrt (2013).

4 Criteria Definition and Description

Based on the literature review above, these criteria are selected to choose the most suitable platform satisfying the requirements of various big data analytics challenges. They are under three categories based on their feasibility and integrability:

1. technical (hardware and resources configuration requirements) perspective:
    - 1.1 availability and fault tolerance – networks, servers, and physical storage must be both resilient and redundant, this criterion has the values of: Poor (1) / Fair (2) / Good (3) / Very Good (4) / Excellent (5),
    - 1.2 scalability and flexibility – how to add a more scale for unexpected challenges, the criterion has the values of 1, 2, 3, 4, 5,
    - 1.3 performance (latency) – data processing time, based on a single transaction or query request, the criterion has the values of 1, 2, 3, 4, 5,
    - 1.4 computational complexity – extensions such as data mining and business intelligence tools, the criterion has the values of 1, 2, 3, 4, 5,
    - 1.5 distributed storage capacity and configurations – to work with different storage systems, how much data needs to be available in storage nodes at the same time, how much data is required to be archived on a periodic basis, etc., the criterion has the values of 1, 2, 3, 4, 5,
    - 1.6 data processing modes – time aspect of data (how often are data managed), real-time and stream processing against historical data and time series data sources, this criterion has the values of: Transaction processing (1) / Real-time processing (2) / Batch processing (3),
    - 1.7 data security – level of security and tools offered, data are protected, more or less valuable, platform is subject to strict security, compliance or governance requirements, the criterion has the values of 1, 2, 3, 4, 5,
2. social (people and their knowledge and skills) perspective:
    - 2.1 ease of installation and maintenance – command line interface or graphical user interface, skills and knowledge needed for the deployment of a new solution, the criterion has the values of 1, 2, 3, 4, 5,
    - 2.2 user interface and reporting – usability and complexity of features, the criterion has the values of 1, 2, 3, 4, 5,
    - 2.3 documentation and support – to simply describe each feature of the tool, technical and customer support, the criterion has the values of 1, 2, 3, 4, 5,
3. cost and policy perspective,
    - 3.1 cost – what a customer wants, how much can be spent on, the criterion offers these options: Open source (1) / Trial version (2) / Commercial release (3),
    - 3.2 sustainability of the solution – the cost associated with the skills maintenance, configuration, and adjustments to the level of agility in development, how much data will the organization need to manage and process today and in the future, the criterion has the values of: Low (1) / Medium (2) / High (3),
    - 3.3 policy and regulation – related to the deployment of the selected solution such as privacy policy, law conflicts and restrictions of the use, etc., the criterion has the values of 1, 2, 3, 4, 5,

Based on the literature review of the possible advantages and disadvantages of various big data analytics platforms, eleven tools were selected as alternatives to be compared. Some frameworks such as SCOPE / Cosmos or Dryad are omitted, because there are no stable implementations of them. Also most of the literature is concerned with the data analysis as the most important phase. Therefore, the AHP model is focused on the big data analytics

platforms, which offer tools for data analysis. A decision table with the values for the selected alternatives can be seen in the Tab. 1. The data used are from 2015. The AHP model's structure is a hierarchy of four levels constituting goal, criteria, sub-criteria and alternatives as can be seen from the Fig. 1.

ALTERNATIVES	CRITERIA AND THEIR TYPE												
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**
	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MAX	MIN	MIN	MAX
Amazon Kinesis	4	4	4	3	4	2	5	4	4	3	3	3	4
Apache Hadoop	5	4	3	5	4	3	3	3	3	4	1	1	2
Apache Spark	4	4	5	3	3	3	3	3	3	3	1	1	2
Apache Storm	4	3	4	2	3	2	2	2	3	2	1	2	2
Cloudera	5	4	4	4	4	2	4	5	5	4	2	3	3
GridGrain	4	3	5	2	3	2	4	2	3	3	1	2	3
Hortonworks	5	5	4	4	3	2	4	4	4	3	1	2	3
HPCC	4	4	5	3	4	3	3	3	4	3	1	2	4
InfoSphere Streams	4	3	4	3	3	2	4	3	4	3	2	2	4
MapR	4	4	4	3	4	3	5	4	4	3	3	3	4
Sector/ Sphere	4	3	5	1	3	3	4	2	3	2	1	1	2

Tab. 1. *Decision table for the big data analytics platform selection. Source: Author.*

Three following use cases are designed to meet the various users' needs. These use cases are focused only on the platforms, which offer data analysis tools. However, most of these tools can be integrated with several data transfer, storage and search platforms to support the whole big data life cycle and related phases.

Use case 1 – scientist or advanced user

A high scalable and fault tolerance platform, which offers a high computational complexity and number of techniques implemented, is required. Batch processing platform is more important than real-time processing. Data security is not required, data are available mostly for the testing purposes as open data. User has also a very good knowledge and programming skills. The selected platform has to be open source with no policy and regulation conflicts.

Use case 2 – medium-sized business

The business needs a highly available, flexible, scalable and fault tolerance platform with a good computational complexity to store a big amount of data. It requires a real-time processing platform with a very good data security. Platform has to be easy to deploy with a wide customer support. The business has very good financial resources. Privacy policy options and SLA are very important.

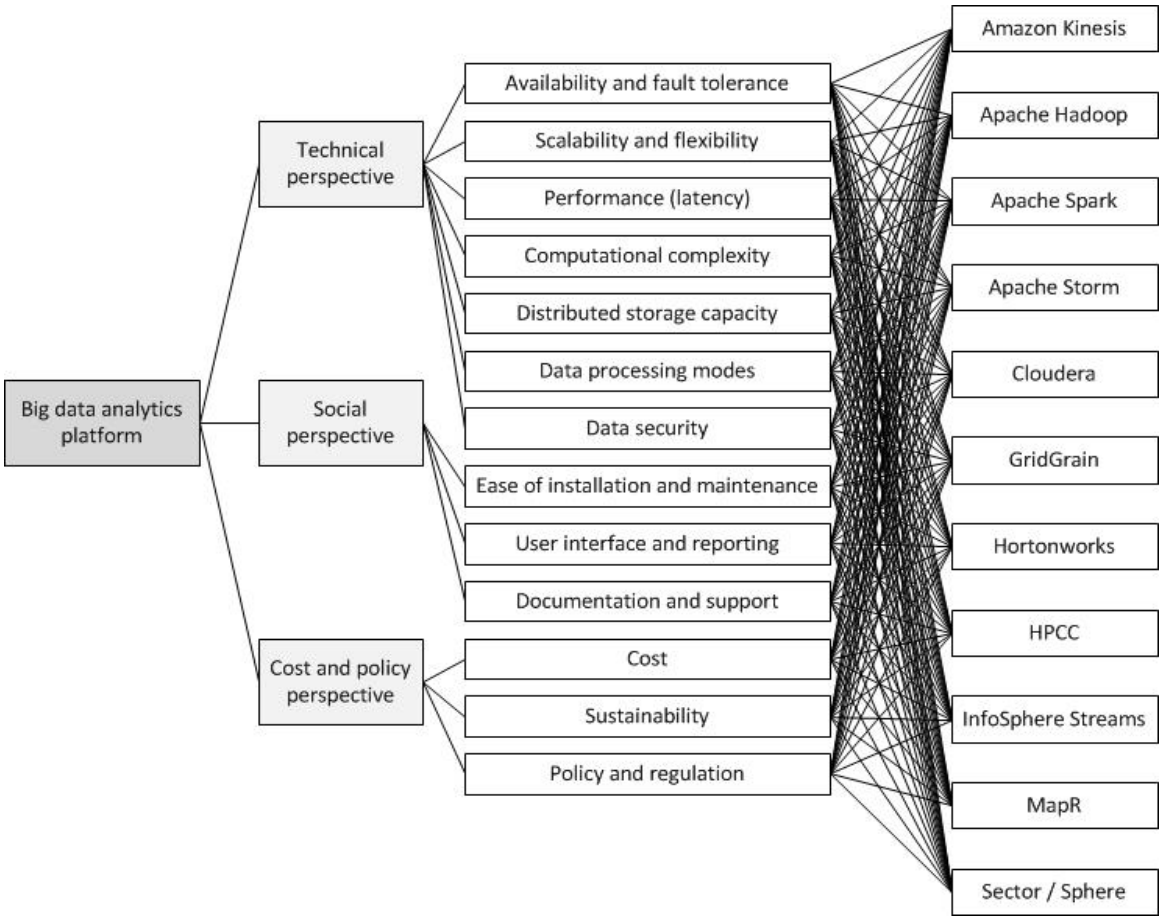

Fig. 1. *The AHP model for the big data analytics platform selection. Source: Author.*

Use case 3 – public sector institution

An available, flexible and fault tolerance platform, which offers a high variety and flexibility of computational complexity extensions is required. Batch processing and open source platform with a graphical user interface is preferred. It should be easy deployed as a small cluster. No personal data will be processed, however, there should be some security tools available. It requires a very good documentation and reference manual to easy deploy and maintain the selected platform.

5 Results and Discussion

In the Tab. 2, weights for the defined criteria for each use case are shown. Following the AHP methodology, paired comparisons of the alternatives on each attribute and the inter-attribute relative importance were made and converted to a fundamental scale of absolute numbers based on their intensity of importance. The scale then ranges from 1/9 (least valued than), to 1 (equal), and to 9 (absolutely more important than) covering the entire spectrum of the comparison. Then, all the calculations were performed to find the maximum Eigen value, consistency index, consistency ratio and normalized values for each criterion / alternative. If the maximum Eigen value, consistency index and ratio are satisfactory then decision is taken based on the normalized values, else the procedure is repeated till these values lie in a desired range (Saaty, 1990; Saaty, 2008). More details about this method, its steps and requirements can be found in Saaty (1990) or Saaty (2008).

HIERARCHY OF CRITERIA	WEIGHT		
	Use case 1	Use case 2	Use case 3
1. Technical perspective	0.540	0.493	0.493
1.1 Availability and fault tolerance	0.118	0.170	0.177
1.2 Scalability and flexibility	0.206	0.144	0.227
1.3 Performance (latency)	0.071	0.114	0.087
1.4 Computational complexity	0.358	0.110	0.169
1.5 Distributed storage capacity	0.071	0.081	0.087
1.6 Data processing modes	0.151	0.122	0.184
1.7 Data security	0.025	0.259	0.069
2. Social perspective	0.297	0.196	0.311
2.1 Ease of installation and maintenance	0.297	0.500	0.327
2.2 User interface and reporting	0.540	0.250	0.260
2.3 Documentation and support	0.163	0.250	0.413
3. Cost and policy perspective	0.163	0.311	0.196
3.1 Cost	0.443	0.413	0.474
3.2 Sustainability	0.169	0.260	0.150
3.3 Policy and regulation	0.388	0.327	0.376

Tab. 2. *Criteria and their weights for each use case. Source: Author.*

In this study, each use case reported a very low value of consistency ratio: use case 1 (0.018), use case 2 (0.037) and use case 3 (0.023), which is much better than the recommended 10% acceptable margin (Saaty, 1990). The only inconsistency was found in the cost and policy perspective where, especially in the use case 2, the importance of cost and sustainability of the solution is dealing with uncertainty about the way things will happen in the future.

In all the cases, the technical perspective is the most important issue. Use case 1 and 3 then prefer the social perspective. For the use case 2 (medium-sized business), the cost and policy perspective is the second most important perspective, together with the data security. Fig. 2 shows the final weights for the selected alternatives for each use case. Based on the needs of the user defined in the use case 1, these three most suitable big data analytics platforms are selected: Apache Hadoop (19.3%), Hortonworks (15.4%) and Cloudera (13.2%). For the use case 2, the choice is: MapR (14.4%), Amazon Kinesis (13.3%) and InfoSphere Streams (10.4%). For the use case 3, the choice is: Hortonworks (15.7%), Apache Hadoop (15.4%) and Cloudera (11.6%). In this case, the HPCC system is on the fourth place with 11.1% and may be used as an alternative to the MapReduce-based platforms.

The precision with which decision-makers can provide a paired comparison may be limited by their knowledge, experience, and even cognitive biases, as well as by the complexity of the big data analytics platform selection problem. To solve this problem, the decision-makers have to be trained to understand the details, strengths, and limitations of the AHP method as well as the related platforms (Wei et al., 2005).

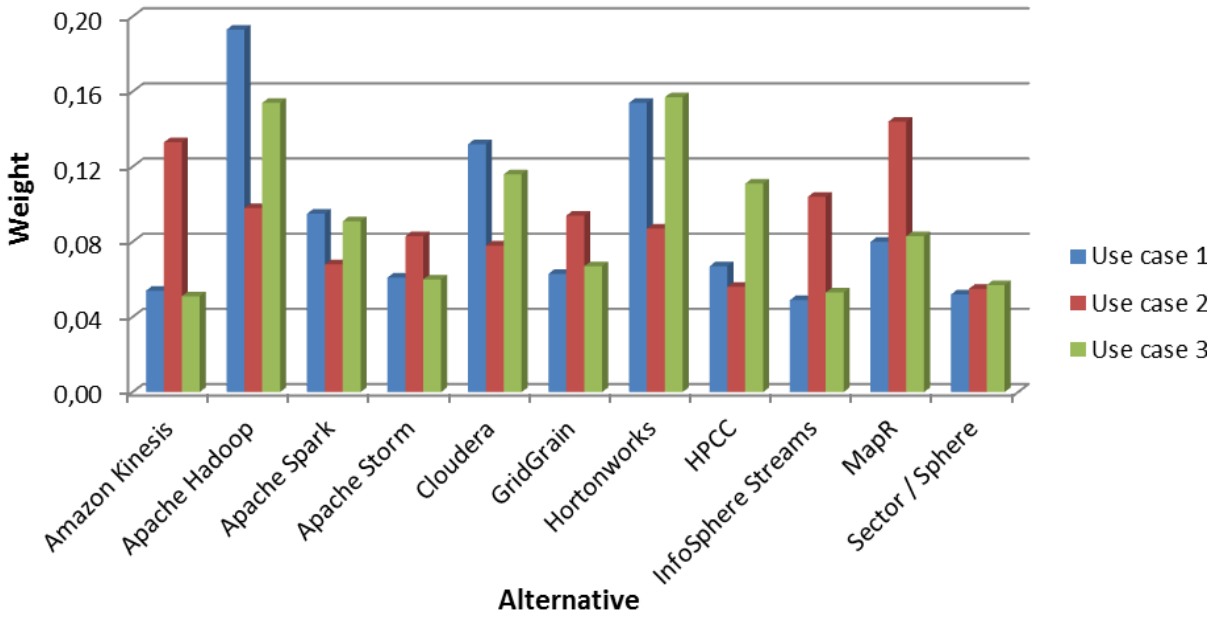

Fig. 2. *Weights of the alternatives for each use case. Source: Author.*

It has to be also noted, that the usage of the AHP method is not a new discovery in the selection of the most suitable big data analytics platform. However, the main contribution of this paper lies in providing a new hierarchy of criteria, which reflects the actual trends in the software evaluation in the big data era.

6 Conclusion

In this paper, the literature is reviewed in order to provide the overview of the big data analytics platforms and to propose the AHP model, which offers a simple but important evaluation method that can help businesses and public sector institutions in selecting the most suitable big data analytics platform. This approach is also flexible enough to incorporate extra attributes or decision-makers in the evaluation. Special attention is paid to the whole life cycle of the big data analytics. By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision-making and support informed decisions. The new AHP model can not only reduce cost during the selection phase, but also decrease the resistance and invisible cost in the implementation stage.

The results provided in this paper represent the first step to select the most suitable big data analytics platform based on the user's needs. Quantitative performance measures of the selected platforms will be the next step to evaluate and compare these platforms more precisely. Also the number of alternatives should decrease to five or less to clearly describe the differences between these platforms. The comparison presented above also helped to eliminate some of the unsuitable platforms for the defined use cases. Choosing the right platform for a particular big data application and combining of multiple platforms to solve various decision-making problems are planned for the future research.

References

Bengtsson, P., & Bosch, J. (1998). Scenario-based software architecture reengineering. In *Proceedings of the Fifth International Conference on Software Reuse* (pp. 308-317). New York: IEEE.

Brožová, H., Houška, M., & Šubrt, T. (2013). *Modely pro vícekriteriální rozhodování*. Praha: Česká zemědělská univerzita v Praze.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), 12-27. doi: 10.1145/1978915.1978919

Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. In B. Hong et al. (Eds.), *Database Systems for Advanced Applications* (pp. 1-15). Heidelberg: Springer.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), 171-209. doi: 10.1007/s11036-013-0489-0

Daniluk, A. (2012). Visual modeling for scientific software architecture design. A practical approach. *Computer Physics Communications*, 183(2), 213-230. doi: 10.1016/j.cpc.2011.07.021

Demchenko, Y., Grosso, P., Laat, C., & Membrey, P. (2013). Addressing Big Data Issues in Scientific Data Infrastructure. In *2013 International Conference on Collaboration Technologies and Systems* (pp. 48-55). New York: IEEE.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. In P. Perner (Ed.), *ICDM 2014. LNAI, vol. 8557* (pp. 214-227). Heidelberg: Springer.

Guo, S. (2013). *Hadoop Operations and Cluster Management Cookbook*. Birmingham: Packt Publishing.

Kaisler, S., Armour, F., Espinosa, J.A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *46th Hawaii International Conference on System Sciences* (pp. 995-1004). New York: IEEE.

Karaarslan, N., & Gundogar, E. (2009). An application for modular capability-based ERP software selection using AHP method. *The International Journal of Advanced Manufacturing Technology*, 42(9-10), 1025-1033. doi: 10.1007/s00170-008-1522-5

Lai, V. S., Wong, B. K., & Cheung, W. (2002). Group decision making in a multiple criteria environment: A case using the AHP in software selection. *European Journal of Operational Research*, 137(1), 134-144. doi: 10.1016/S0377-2217(01)00084-4

Lake, P., & Drake, R. (2014). *Information Systems Management in the Big Data Era*. London: Springer.

Lee, K. H., Lee, Y. J., Choi, H., Chung, Y.D., & Moon, B. (2011). Parallel Data Processing with MapReduce: A Survey. *SIGMOD Rec.*, 40(4), 11-20. doi: 10.1145/2094114.2094118

Liou J. J. H., & Tzeng, G.-H. (2012). Comments on "Multiple criteria decision making (MCDM) methods in economics: an overview". *Technological and Economic Development of Economy*, 18(4), 672-695. doi: 10.3846/20294913.2012.753489

Lnenicka, M. (2015). An In-Depth Analysis of Open Data Portals as an Emerging Public E-Service. *International Journal of Social, Education, Economics and Management Engineering*, 9(2), 589-599.

Lněnička, M., & Komárková, J. (2014). An Overview and Comparison of Big Data Analytics Platforms. In *Sborník příspěvků z mezinárodní vědecké konference MMK 2014* (pp. 3446-3455). Hradec Králové: Magnanimitas.

Loshin, D. (2013). *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*. Waltham: Elsevier.

Madden, S. (2012). From Databases to Big Data. *IEEE Internet Computing*, 16(3), 4-6. doi: 10.1109/MIC.2012.50

Marakas, G. M., & O'Brien, J. A. (2013). *Introduction to Information Systems*. New York: McGraw-Hill/Irwin.

Saaty, T. L. (1990). How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1), 9-26. doi: 10.1016/0377-2217(90)90057-I

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83-98. doi: 10.1504/IJSSCI.2008.017590

Sakr, S., Liu, A., & Fayoumi, A.G. (2013). The Family of MapReduce and Large Scale Data Processing Systems. *ACM Computing Surveys (CSUR)*, 46(1), 1-27. doi: 10.1145/2522968.2522979

Saecker, M., & Markl, V. (2013). Big Data Analytics on Modern Hardware Architectures: A Technology Survey. In M. A. Aufaure & E. Zimányi (Eds.), *Business Intelligence* (pp. 125-149). Berlin Heidelberg: Springer.

Shamsi, J., Khojaye, M. A., & Qasmi, M. A. (2013). Data-Intensive Cloud Computing: Requirements, Expectations, Challenges, and Solutions. *Journal of Grid Computing*, 11(2), 281-310. doi: 10.1007/s10723-013-9255-6

Silva, J. P., Goncalves, J. J., Fernandes, J., & Cunha, M. M. (2013). Criteria for ERP selection using an AHP approach. In *2013 8th Iberian Conference on Information Systems and Technologies* (pp. 1-6). New York: IEEE.

Singh, D., & Reddy, C. K. (2014). A survey on platforms for big data analytics. *Journal of Big Data*, 1(8), 1-20. doi: 10.1186/s40537-014-0008-6

Tien, J. M. (2013). Big Data: Unleashing Information. *Journal of Systems Science and Systems Engineering*, 22(2), 127-151. doi: 10.1007/s11518-013-5219-4

Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, 169(1), 1-29. doi: 10.1016/j.ejor.2004.04.028

Valacich, J. S., George, J. F., & Hoffer, J. A. (2012). *Essentials of Systems Analysis and Design*. New Jersey: Prentice Hall.

Vossen, G. (2014). Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science*, 1(1), 3-14. doi: 10.1007/s40595-013-0001-6

Wei, C. C., Chien, C. F., & Wang, M. J. J. (2005). An AHP-based approach to ERP system selection. *International Journal of Production Economics*, 96(1), 47-62. doi: 10.1016/j.ijpe.2004.03.004

Zavadskas, E. K., & Turskis, Z. (2011). Multiple criteria decision making (MCDM) methods in economics: an overview. *Technological and Economic Development of Economy*, 17(2), 397-427. doi: 10.3846/20294913.2011.593291

Zhao, L., Sakr, S., Liu, A., & Bouguettaya, A. (2014). Big Data Processing Systems. In *Cloud Data Management* (pp. 135-176). Heidelberg: Springer.