Seminar Predictive Analytics in Big Data WS 2016/17

Jasim Waheed Ansari

RWTH Aachen University, 52056 Aachen, Germany
jasim.waheed@rwth-aachen.de

Abstract. Predictive Analytics (PA) has replaced the idea of ad-hoc data analysis by fact based decisioning. PA include statistical methods from machine learning and data mining to build data model for regression, prediction, neural networks, classification purposes. In order to incorporate big data criterion, the frameworks and tools that perform modeling using previously stated methods has to address the 4Vs of Big Data, namely volume, veracity, velocity and value. This paper is aimed at providing an overview of big data analytics framework and then highlighting few major tools, comparison and assessment of their characteristics. The second aim is to go through the major issues and challenges that are faced while carrying out PA in big data. We will also address the possible solutions in form of best practices concerning those problems. The third aim is to provide an example from ERP systems to highlight its predictive analytics capabilities using big data systems. The fourth aim is to present a conceptual framework of integrating Complex Event Processing and PA called Predictive Complex Event Processing. We will observe that the given framework would be a generic design pattern for future work. On an ending note, we will demonstrate a case study to get a lucid idea on a practical level.

1 Introduction

In recent years, there is a flood of data that is shared or generated from various sources. This data is usually presented in an unstructured format such as videos, audios, texts, images. The presented data could or could not be related to each other or there could be a possibility of having some hidden patterns. To perform any sort of analysis, analysts cannot use such unstructured data directly. Such data requires conversion into a well-formed format (similar to a relational data), which could be used by data scientists for purpose of analysis on them. As a result, this structured data can then be used to decipher hidden meanings or pattern which supports prediction of market behavior, enterprise needs, enabling precision based decisioning using technique called Predictive Analytics.

1.1 What is Predictive Analytics?

Predictive Analytics is the process of predicting useful information from a set of data (structured,unstructured or semi-structured). It surrounds techniques from statistics, data mining, machine learning and artificial intelligence.

Predictive Analytics steps: As per general guideline, the steps along with percentage of time spent on each of the step are as follows:

1. Understanding the domain (5-10 percent)
2. Understanding the data (5-10 percent)
3. Preparing the data (50-60 percent)
4. Modeling (5-15 percent)
5. Evaluation (5-10 percent)
6. Deployment (10-15 percent)

It is worth noting that each step could have as many iterations as needed. Adding to that, core effort in processing is emphasized at the data.
Figure 1 shows Predictive Analytics process which is followed actively by analysts across various industries [7]

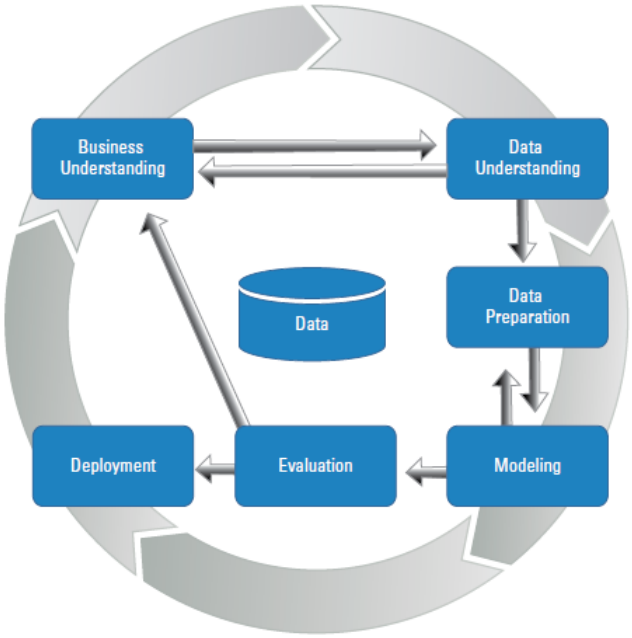

Fig. 1. Predictive analytics process

Techniques available: From predictive analytics process, modeling and evaluation steps are where analysts use algorithms to produce key information they

desire. Predictive modeling algorithms [8] falls under Supervised Learning. In Supervised Learning, the predictor value or target variable is *supervisor*, it is the column in dataset that is used for predicting value from other column values. It is divided into two sets:

1. Classification: class based or discrete target variable
2. Regression: continuous target variable

Figure 2 shows Predictive Analytics techniques used for modeling[9].

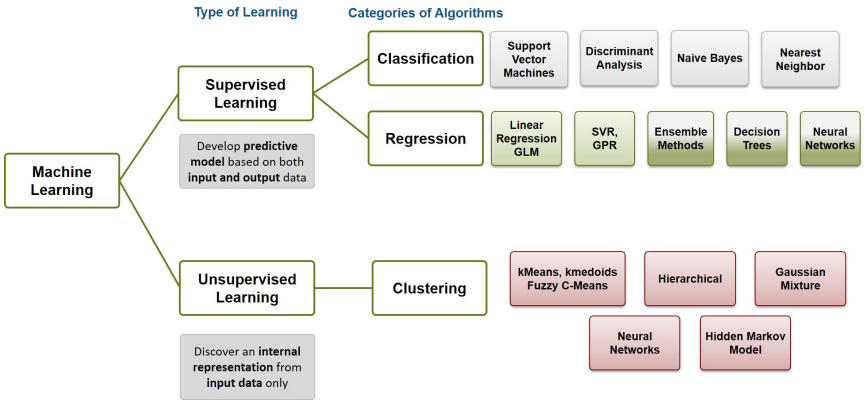

Fig. 2. Predictive analytics techniques

1.2 Predictive Analytics: Harness the power of Big Data

Almost every decision making process, be it in any enterprise, involves predictive analytics to drive their business better and helps gaining competitive edge in the market. Decision making in current era is based on day-to-day operational business data rather than on only special projects or scenario. Current tools and technologies are unsatisfactory and not up to standards to process such huge chunks of operational data. They are also unable to make in-depth insight and generate value.

Thanks to Big Data technology and tools, predictive analytics can be applied to deluge of data at enterprise level, sometimes also called as Big Data Analytics. We have now many ways to tackle and test different predictive models at various level of framework for business strategies. Figure 3 shows delineation of characteristics of Predictive Analytics techniques in transactional databases vs big data technologies. The former is based on prediction through historical structured data whereas big data platform can be used for modeling in real-time with structured, unstructured or semi-structured data. The common methods involved in transactional databases are data mining algorithms such as decision trees, regression, clustering, association rules, etc. In big data mediums, advanced techniques such as speech recognition, mobile based analytics, etc. [[8]].

In the upcoming sections, we will be providing the conjunction of Predictive Analytics with Big Data technology and bring the following topics in coherence:
(1) understanding big data architecture for analytical perspective
(2) comparison and assessment of big data analytical tools used for predictions
(3) address few of the challenges which collides interest of predictive analytics with big data. One of the key challenges being privacy- arising from ever increasing data from online services and personal devices such as mobile phones. These are subjected to risk of personal information being exposed for illicit uses.
(4) discuss example from Enterprise Resource Planning (ERP) systems, and explore the opportunities of predictive analytics on ERP system's integrated big data hub.
(5) discuss about Complex Event Processing which deals in identifying complex events based on the rules dictated by users of the system. In order to avoid the manual intervention of users for providing progressive rules, we will discuss about the framework that includes Predictive Analytics technologies in Complex Event Processing tools and applications.
(6) case study demonstrating the power of predictive analytics coupled with big data technology.

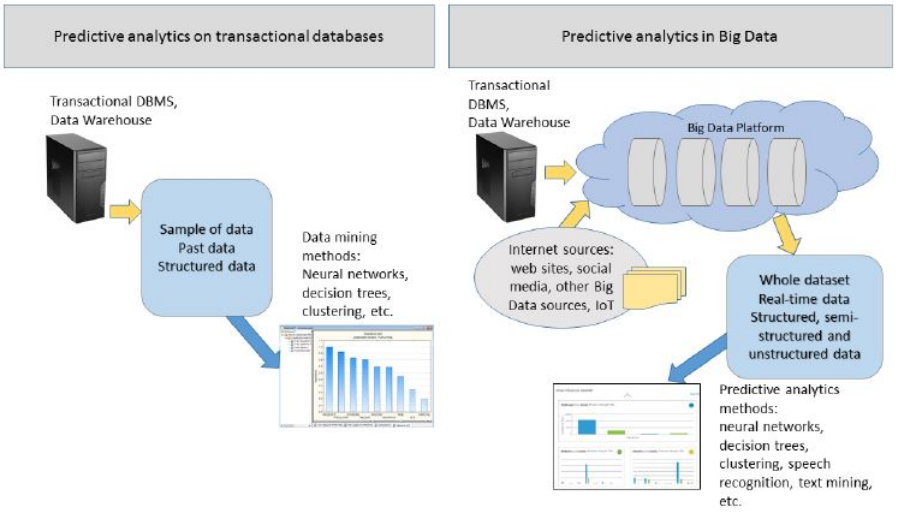

Fig. 3. Predictive analytics in traditional databases vs Big Data

2 Big Data System Architecture

A big data system is immensely complex. In order to demonstrate the architecture of predictive analytics in big data, we should firstly understand the catego-

rization of two paradigms of big data analytics with respect to processing time requirements:

- Stream Processing: Streaming processing data is dependent on fresh data i.e. it needs to be analyzed as soon as it is available. Some famous open source systems available are Storm and Kafka.
- Batch Processing: In batch processing, data is stored as chunk and then analyzed. One of the most famous models available is Map Reduce.

In this section, we will discuss architecture based on batch processing types, since it is widely adopted in industries. We will show one such framework placed on value chain for big data analytics[5]. Big data value chain framework involves four stages i.e. generation, acquisition, storage and analytics. In this work, we will see leading technologies in analysis phase.

2.1 Big Data Value Chain architecture

This architecture is based on system engineering approach, well recognized in industries, the notion is to break down the big-data system into four continuous phases in horizontal axis as shown in Figure 4 [5].

Data Analysis: It is the main and concluding step from big data value chain that focuses on the aspect of analytics. This phase supports mechanisms or tools to investigate, transform, and modeling data to discover insights. We will discuss hereby some classification metric of big data analytics. Furthermore, we will describe common methods in big data analytics that together build the pillar for making predictions comprehensively. We will then give an overview picture of nomenclature of big data analytics with respect to data type.

Categories: Blackett et al [10] proposed classification of big data analytics into three levels as follows:

1. Descriptive Analytics: analyses historical dataset to describe what happened. It is widely used with business intelligence or ERP solutions.
2. Predictive Analytics: focuses on future predictions and trends, uses supervised learning algorithms to understand trends, and data mining exacts insight.
3. Prescriptive Analytics: focuses on decision making. For instance, simulation is used to analyze the systems and recognize various optimization techniques to get efficient solutions.

3 Predictive Analytics Technologies

The technology that is adept in supporting big data includes new kinds of database platforms(for example, Hadoop and Spark) as a central core of big

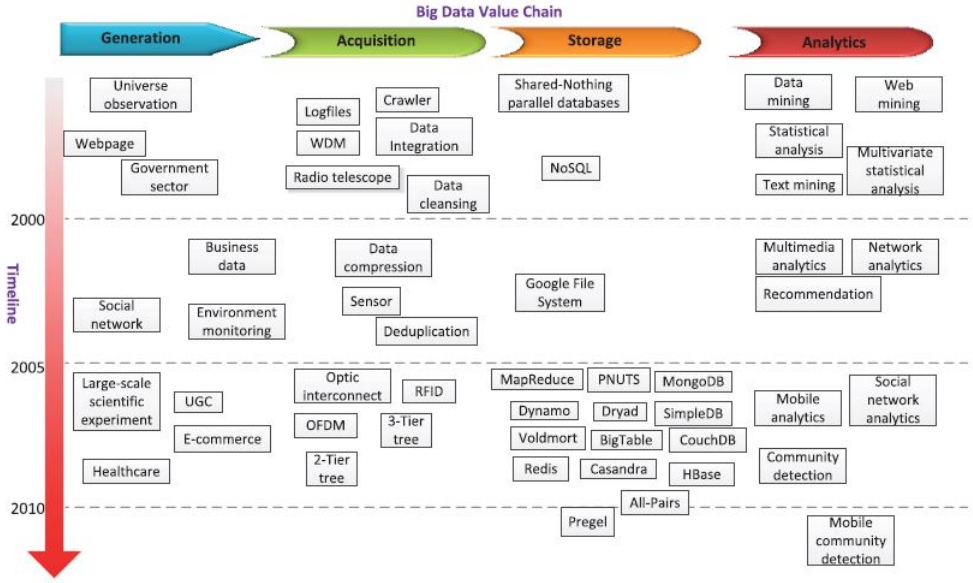

Fig. 4. Big data value chain architecture. Horizontal axis composes of four continuous phases for data life cycle. For each stage, we demonstrate cutting edge technologies available past ten years in vertical axis

data systems, predictive tools are built on top of them to perform analytics in cloud, thus enabling enterprises with scalability and on-demand usage. We will see software stacks as framework for predictive analytics in Hadoop and Spark respectively.

3.1 Hadoop Software Stack

It is a massive framework encompassing different modules, including Hadoop Distributed File System (HDFS) & Hbase for data storage, MapReduce as computation core for analysis on data, Flume & Sqoop as integration tools. This framework is in accordance with big data value chain architecture and it is used as a powerful solution for batch-type processing applications. The layered core software libraries is shown in figure 4. On top of this framework is the data mining library called *Mahout*.

Mahout: It contains major algorithms responsible for clustering, classification and recommendation (collaborative filtering) to support large-scale predictive analytics model. One of the most famous ways to operate on data needed for such a model is to deploy Mahout in machine that is already executing Hadoop[11]. Hadoop nominates a master system coordinates with the other systems (i.e. Map systems and Reduce systems) applied in its distributed processing.

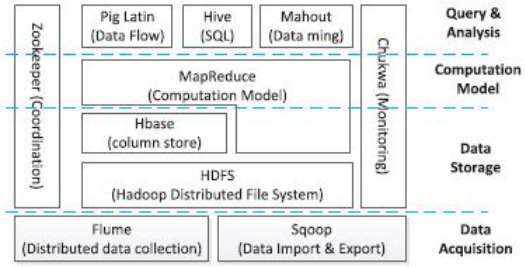

Fig. 5. Hadoop Software Stack, covering the major function of big data value chain, inclusive of data import, storage and processing

3.2 Spark Framework for real-time analytics

Another famous open-source platform for processing big data that is best applicable for stream-oriented data processing. Spark is viewed as a strong contender to replace MapReduce capabilities of Hadoop. Spark sits on top of existing Hadoop cluster which relies on YARN for efficient resource management and job scheduling.

MLlib: It is effective for iterative-based large scale machine-learning applications for predictive analytics. The algorithms are written in Scala and the linear algebra libraries uses native C++. MLlib includes support for JAVA, Scala, and Python APIs. They are released as part of Spark project under Apache 2.0. Figure 6 shows Apache Spark ecosystem, where MLlib is built on top of Spark.

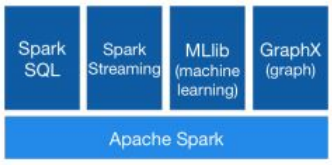

Fig. 6. Spark Ecosystem

3.3 Big Data Predictive Analytics Tools

Extremely strong knowledge in statistics and technical skills were previously required to carry out predictive modeling, due to complexities in the usage of

statistical models and tools. Advanced abilities were also required in analyzing the result. However, due to the increasing growth of technology the tools are no more focused to advanced experts. Business experts from various organizations can use tools marketed by proprietary software enterprises and open source organizations to predict their business growth. Figure 7 [6] shows Predictive Analytics tools classification with its characteristics and examples.

	Open Source and Freeware Tools	Proprietary Tools	Software API Tools	Predictive Pricing Solutions	Customer Churn Software Tools
Characteristics	Has no cost involved and are available with free license	These are the licensed tools provided by various vendors, offering better support	Application programming interfaces accessing machine learning algorithms for prediction using RESTful services	Using Machine Learning to produce cutthroat pricing depending on various factors such as competitor pricing, market, demand and supply	Deciding which customers about the abandon the service, retention plans to avoid churn, analyze the real value of loss
Examples	R, RapidMiner, Weka, Apache Spark Mahout, Apache Spark MLlib, KNIME, sci-kit learn	IBM Watson Analytics, SAS Enterprise Miner, SAP Predictive Analytics, Oracle Advanced Analytics	Google Cloud Prediction API, GraphLab Create, Microsoft Azure Machine Learning, PredictionIO	Blue Yonder Dynamic Pricing, PROS Pricing Analytics, Retelon Dynamic Price Management System	Lattice Engines, Linkedin Sales Navigator, AgilOne, SalesPredict, Adobe Recommendations

Fig. 7. Predictive Analytics Tools Classification and Examples

Comparison of open source tools The comparison matrix[12] in the given figure 8 explains some of the open source tools (described previously) capabilities with respect to data science techniques. It needs to be observed that Weka offers the maximum support on an open-source level. In the later section of this paper, we will highlight the case study of predicting chronic kidney diseases using Weka tool.

4 Issues in predictive analytics

Predictive analytics are built on set model solutions, and if these are not optimized and addressed properly, they may present problems such as slow down the progress and alter the business analytics solutions. We will address few of the issues associated in this section.

– Data quality: Data is fundamental to predictive analytics, and it is highly recommended to have a deeper knowledge on the available data before investing in any predictive analytics technologies. Andreescu, et al figured out

	Orange	Tanagra	Rapid Miner	KNIME	R	Weka
K-means Clustering	Yes	Yes	Yes	Yes	Yes	Yes
Association Rule Mining	Yes	Yes	Yes	Yes	Yes	Yes
Linear Regression	Yes	Yes	Yes	Yes	Yes	Yes
Logistic Regression	Yes	Yes	Yes	Yes	Yes	Yes
Naive Bayesian Classifiers	Yes	Yes	Yes	Yes	Yes	Yes
Decision Tree	Yes	Yes	Yes	Yes	Yes	Yes
Time Series Analysis	No	No	Some	Yes	Yes	Yes
Text Analytics	Yes	No	Yes	Yes	Yes	Yes
Big Data Processing	No	No	No	No	Yes	Yes
Visual WorkFlows	Yes	Yes	Yes	Yes	No	Yes

Fig. 8. Comparison matrix showing few open source tools supporting common predictive analytics techniques

that the poor data could lead to erroneous results. In order to maximize the quality, data should be subjected to data preparation, cleaning and formatting. All of which are significant part of data mining steps.

- Model complexity: Analytic solutions developed for predictions are generally known for taking overly complex models and delivering easy-to-understand results from them. But, companies get carried away and started to expect much more beyond the tools capabilities. This eventually end up slowing down the delivering of report. To maximize the result while minimizing the time to deliver reports, an effective model life-cycle management approach is needed to be adopted.

- Communication gap: There may be a gap of conveying the insights from data scientists to business users. In order to overcome such problems, enterprises should have skilled enough people not only in analyzing trend but also in documenting it and presenting this info in crisp and compelling terms.

- Privacy concerns: This is one of the key issues of all. As the big data continues to grow bigger, the analytical capabilities on them poses severe risk of discovering insights in form of individual data. Therefore, such analytical insight could never rule out to make individual data anonymised. We look into some best practices available for preserving privacy of the user data.

4.1 Best practices available for preserving privacy

In this section, we discuss three privacy preservation methods [13].

1. Data Anonymization: It is the process of masking or changing data that will be published in such a way that doesn't conceal sensitive information such as full name, social security number etc. The problem with anonymization is that the sensitive information can still be de-identified through linkability with external sources. Three privacy methods with respect to data anonymization are as follows:
   – K-Anonymity: A dataset is K-Anonymous if for any row in the dataset with given attributes, there exists atleast K-1 similar records. It is achieved by suppression or generalization. In suppression, attributes are masked by some constant values 0,* etc. In generalization, identifiers are masked by a more generic value from levels up the hierarchy.
   – L-Diversity: It tries to bring diversity in sensitive data record. It maintains that each equivalence class has atleast L disparate values of sensitive columns.
   – T-Closeness: A dataset is T-closed if distance between distribution of sensitive column data in a particular equivalence class and distribution in whole dataset is no more than a threshold value of T.
2. Notice and Consent: It is one of the most common method for web applications and services. The consumer needs to approve the notice before using the services. It imposes a privacy preservation on the user.
3. Differential Privacy: It is a method empowering data scientists to extract the overall picture from datasets but holding strong privacy on individual data. In contrast to anonymization, there is no modification done to data but there is an added firewall-like interface above the dataset to calculate the results and add much needed inacurracies.

5 Example: Big Data Predictive Analytics in ERP Systems

Enterprise Resource Planning(ERP) systems are business process management software, which integrates applications across different departments of an enterprise to provide a holistic view of employees, which corresponds to financial impact on the business. Adding predictive analytics capabilities to ERP systems render progressive guideline to yield effective informed decision.

5.1 Enterprise Platform for Decision Management in ERP System

Predictive analytics on enterprise level enables unlocking insights from structured and unstructured data, identifying classification, associations, and segmentations. Figure9 shows deployment of Decision as a Service in ERP system landscape. This framework is potent for transforming large set of transactional data into mathematical form for being usable by predictive models.

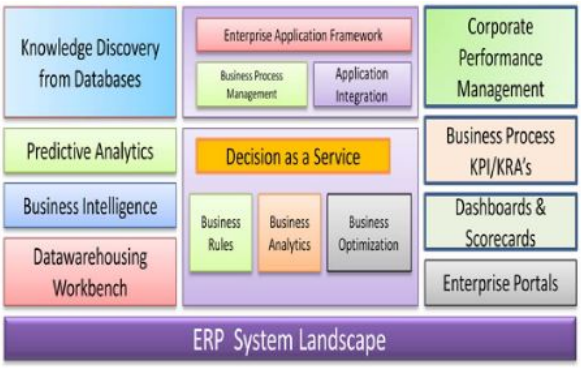

Fig. 9. Enterprise Platform for Decision Making [2]

5.2 ERP with Big Data Using Predictive Analytics

In current business situations, integrating [2] big data solutions to ERP helps in using big data from diverse sources such as web logs, documents, call center transactions, sensors etc. Figure 10 discusses big data predictive analytics framework aimed at automating operational decisioning processes in enterprises using SAP ERP system.

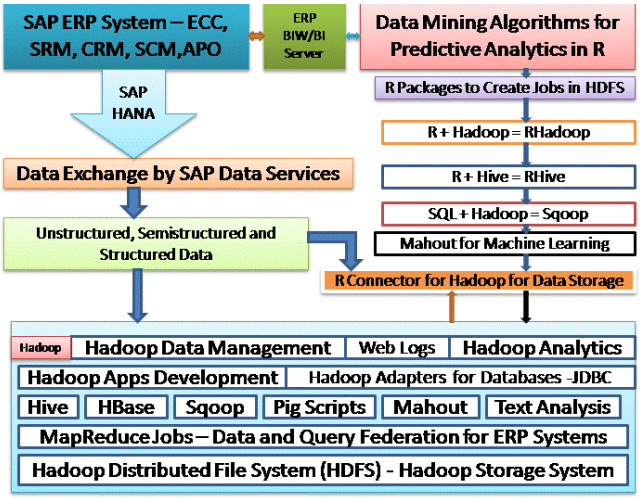

Fig. 10. Application Framework for Big Data Predictive Analytics in ERP Systems [2]

5.3 Predictive Analytics steps in ERP systems

The key factor for performing analytics for prediction is to build a predictive model. Following 11 explains the steps involved in model development: The au-

Step	Modeling Activity	Implementation Activity
1	Defining scope of the analytics project	Identified business processes and their outcomes
2	Data exploration	Data sources identification
3	Data cleansing and data integration	Using SAP ETL workbench for ensuring data quality
4	Building predictive model	Using SAP APD for modeling
5	Big Data Integration	Using Application Integration Connector
6	Assimilate analytics into business processes	Adopting closed-loop analytics to utilize step 4
7	Monitoring the model	Tracking results from different models

Fig. 11. Model Development Tasks

thor emphasizes that decisioning framework for ERP system should be designed to be agile and adaptive. Future work of the explained application framework is aimed at automating strategical decision making.

6 Conceptual Framework: Predictive Complex Event Processing

6.1 Background

6.2 Exploiting the combined value of Complex Event Processing and Predictive Analytics

6.3 Conceptual CEP-PA framework

6.4 Proof of Concept

7 Case Study: Predicting Election Trends using Twitter: Hillary Clinton vs Donald Trump

7.1 Building the network of Twitter users

7.2 Inferring the opinion of users

7.3 Predicting Election Trends

7.4 Results

8 Conclusion

References

1. Chandarana, P., Vijayalakshmi, M.: Big data analytics frameworks. In: Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on. (2014) 430–434
2. Babu, M.S.P., Sastry, S.H.: Big data and predictive analytics in erp systems for automating decision making process. In: Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on. (2014) 259–262
3. Earley, S.: Big data and predictive analytics: What's new? IT Professional **16**(1) (2014) 13–15
4. Fülöp, L.J., Beszédes, A., Tóth, G., Demeter, H., Vidács, L., Farkas, L.: Predictive complex event processing: A conceptual framework for combining complex event processing and predictive analytics. In: Proceedings of the Fifth Balkan Conference in Informatics. BCI '12, New York, NY, USA, ACM (2012) 26–31
5. Hu, H., Wen, Y., Chua, T.S., Li, X.: Toward scalable systems for big data analytics: A technology tutorial. IEEE Access **2** (2014) 652–687
6. : (Predictive analytics tools)
7. Wessler, M.: Predictive analytics for dummies. Alteryx Special Edition. Wiley (2014)
8. Zekić-Sušac, M., Has, A.: (Predictive analytics in big data platforms–comparison and strategies)
9. Pilotte, P.: (Analytics-driven embedded systems, part 2 developing analytics and prescriptive controls)
10. Blackett, G.: (Analytics network-o.r. analytics)

11. Bari, A., Chaouchi, M., Jung, T.: Predictive Analytics For Dummies. For Dummies (2016)
12. Wimmer, H., Powell, L.M.: A comparison of open source tools for data science. Journal of Information Systems Applied Research **9**(2) (2016)
13. Gosain, A., Chugh, N.: Privacy preservation in big data. International Journal of Computer Applications **100**(17) (2014) 44–47