Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 100 (2016) 1071 – 1084
Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Big Data Analytics in Support of the Decision Making Process

Nada Elgendy[a*], Ahmed Elragal[a,b]

[a]*Department of Business Informatics & Operations Management, German University in Cairo (GUC), Cairo, Egypt*
[b]*Department of Computer Science, Electrical, and Space Engineering, University of Technology, Luleå, Sweden*

Abstract

Information is a key success factor influencing the performance of decision makers, specifically the quality of their decisions. Nowadays, sheer amounts of data are available for organizations to analyze. Data is considered the raw material of the 21st century, and abundance is assumed with today's 15 billion devices [aka Things!] already connected to the Internet. Accordingly, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such rapidly changing data of high volume, velocity, variety, veracity, and value by using big data analytics. This paper aims to research how big data analytics can be integrated into the decision making process. Accordingly, using a design science methodology, the "Big – Data, Analytics, and Decisions" (B-DAD) framework was developed in order to map big data tools, architectures, and analytics to the different decision making phases. The ultimate objective and contribution of the framework is using big data analytics to enhance and support decision making in organizations, by integrating big data analytics into the decision making process. Consequently, an experiment in the retail industry was administered to test the framework. Accordingly, results showed added value when integrating big data analytics into the decision making process.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Big data analytics; decision making; B-DAD framework; design science

* Corresponding author. Tel.: +201002063444; fax: +201002063444.
E-mail address: nada.el-gendy@guc.edu.eg

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.251

1. Introduction

Digital technologies have changed the way organizations are built and function, triggering the need for novel solutions and a wide array of functioning applications (Brunswicker et al., 2015). As storage capabilities have exponentially increased and methods of data collection have changed, enormous amounts of data have become easily available. Every second, more and more data is being created from various sources. This data needs new ways to be stored and analyzed in order to extract value. Furthermore, organizations need to get as much value as possible from the huge amounts of stored data (Elgendy and Elragal, 2014). Additionally, companies and individuals possess more technologies and devices, which create and capture more data in different categories. A single user nowadays, can own a desktop, laptop, smartphone, tablet, and more, where each device carries very large amounts of valuable data. These types of data are now being referred to as big data, or data with such volume, variety, and velocity that it becomes difficult to manage with current tools (Russom, 2011).

Big data can include text with social sentiments, clickstreams, audio and video, website log files, as well as spatial and geolocation data, multimedia, XML data, etc. (Chang et al., 2014). Such data requires a new type of big data analytics due to its size, variety, and rapid change, as well as different storage and analysis methods. Additionally, these enormous amounts of big data need to be properly analyzed, in order for valuable and pertaining information to be extracted. Therefore, with the increasing demand for utilizing big data and taking advantage of its opportunities, organizations are seeking clear and simple solutions and guidelines for big data management. Accordingly, the research question of this paper is "*How to integrate big data analytics into the decision making process*?" The aim of this research is to develop and test a framework for the integration of big data tools and techniques, into the decision making process. By adopting this framework, decision makers should be able to enhance the quality of the decision making process, and potentially the quality of the decision as a byproduct. The framework incorporates different important aspects of big data analytics, such as the data analytics lifecycle, necessary infrastructure and architecture, as well as required tools; all mapped to the different decision making phases.

2. Background

The term "Big Data" applies to datasets that grow so large that they become awkward to work with using traditional database management systems. Moreover, the size of big data has expanded beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time (Kubick, 2012). Three main features characterize big data: volume, variety, and velocity, or the three V's (Fan et al., 2015). The volume of the data is its size, while velocity refers to the rate with which data is changing, or how often it is created. Finally, variety regards the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Additionally, IBM added a 4th V, which is veracity (Jagadish, 2015). Additionally, the value of the data has also been considered by some researchers to be a 5th V (Chang et al., 2013).

Big data analytics is where advanced analytic techniques are applied on big data (sets). Analytics based on large data samples can help reveal and leverage business change. However, the larger the set of data, the more difficult it becomes to manage (Russom, 2011). Sophisticated analytics can substantially improve decision making, minimize risks, and uncover valuable insights from the data that would otherwise remain hidden. Sometimes decisions do not necessarily need to be automated, but rather augmented by analyzing huge, entire datasets using big data techniques and technologies instead of just smaller samples that individuals with spreadsheets can handle and understand (Manyika et al., 2011).

Moreover, the managerial decision making process has been an important and thoroughly covered topic in research throughout the years. Simon's four phases of decision making: intelligence, design, choice, and implementation, are popularly adopted by decision makers in different domains (Turban et al., 2007). Furthermore, according to Jagadish (2015), there are many steps to the big data analysis pipeline, and each step comes with its challenges and required decisions. These decisions range from what data to acquire, to how to represent the data in a suitable manner for analysis after extracting, cleaning, and integrating the data with other sources, to how to make decisions based on the results of the analysis. In order for the big data analysis to produce real value, all of these challenges and decision have to be effectively planned and accommodated for.

Decision makers are constantly on the lookout for chances to make more informed decisions, and they need to be able to understand and utilize big data in order to further enhance the (traditional) decision making process. Thus, research needs to cover how the big data analytics tools and methods can be integrated with the decision making process in order to enhance decision making and provide valuable insights for decision makers.

3. The B-DAD Framework

Our research follows the design science methodology, so accordingly Peffers et al.'s (2008) six stages design science process is adopted for building and evaluating the framework. The first two stages, identifying the problem and defining the objectives of a solution, were completed through exploratory research. Consequently, using the applicable knowledge from the knowledge base and the business needs of the environment, an artifact - namely the B-DAD framework – was developed. This is attained through perusing the literature and research, as well as testing some of the big data analytics technologies to add to the framework. Accordingly, both research rigor and relevance are attained.

Subsequently, after the B-DAD framework is developed, it is evaluated and demonstrated by using it to apply big data analytics in order to support decision making. This demonstration is fixed in the form of experiments on real data, and actual business cases in order to provide a sufficiently relevant context for evaluation. Finally, the evaluation of the framework was accomplished by observing how well the framework was in applying big data analytics throughout the decision making process in order to support making a more insightful decision. Additionally, the smoothness of the process and the applicability of the framework in the different scenarios were observed. As a result, we iterated back to the framework design and development phase in order to incorporate the modifications resulting from the experiments into the final B-DAD framework. The process is elaborated and communicated in detail below.

3.1. Framework Development

The B-DAD, or the "Big – Data, Analytics, and Decisions", framework was developed in order to map big data tools, architectures, and analytics to the different decision making phases. The "Big" is hyphenated, because it refers to the following three aspects as being big, not only the data, and additionally maps the incorporation of these aspects together. Hence, the data is big, the analytics are big, and the resulting decisions are also big. Thorough analysis and synthesis of relevant literature, investigating state-of-the-art technologies in big data (analytics) and practices, have contributed to the development of our framework. However, the framework is in no way inclusive of all the big data tools, technologies, and analytics, and rather serves as a conceptualization of some of the possible approaches to performing big data analytics in support of the decision making process. Additionally, the framework assumes that the decision domain is already known, and does not need to be first explored in order to extract a problem which needs to be solved, or a question which needs to be answered. The framework is depicted in Fig. 1.

The first phase of the decision making process is the intelligence phase, where data which can be used to identify problems and opportunities, is collected from internal and external data sources. In this phase, the sources of big data need to be identified, and the data needs to be gathered from different sources, processed, stored, and migrated to the end user. Accordingly, the first step in the framework is identifying the big data which will be used for the analysis. The main difference in this step from Fayyad et al.'s (1996) KDD process lies in the diversity of the types of data which will be identified, and their various sources. In addition to relational data and common transactional or operational data, there is social media data, text, images, and audio. Additionally there is data which results as the output of machines and devices, such as system log files, sensor data, satellite data, and mobile or GPS data. Moreover, geospatial data has become very important for analysis, along with internet data, clickstream files, and XML.

Such big data needs to be treated accordingly, so after the data sources and types of data required for the analysis are defined, the chosen data is acquired and stored, similar to the acquiring phase in Fisher et al.'s (2012) big data pipeline, and Oracle's (2015) integrated information architecture. The acquired data can then be stored in any of the big data storage and management tools. These tools can range from traditional DBMSs, such as the open source MySQL or PostgresSQL, to EDWs and columnar or MPP databases, such as Cassandra, PADB, and SAND. Additionally, a distributed file system like HDFS can be used for storing big data, as well as NoSQL databases, such

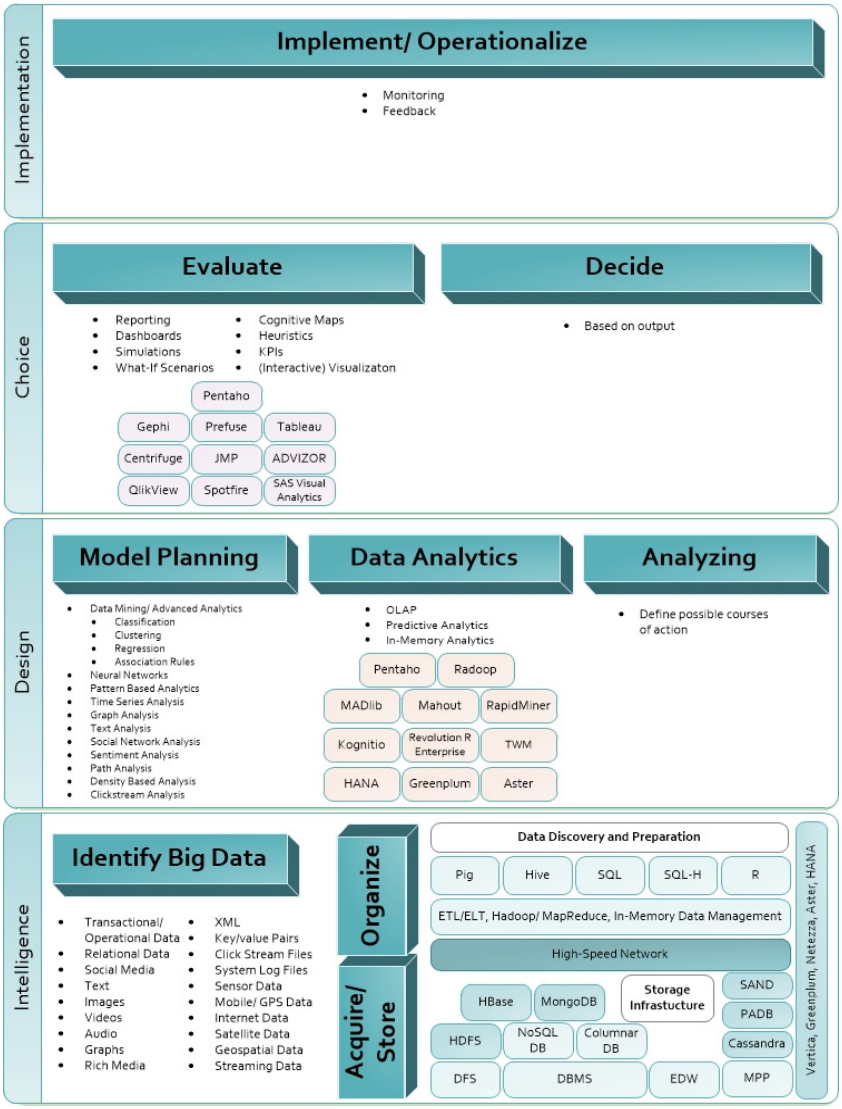

Fig. 1. B-DAD Framework

as MongoDB, CouchDB, or HBase which is built on top of HDFS. In the framework, the examples of specific storage tools are depicted in a slightly darker blue color than the generic technologies.

After the big data is acquired and stored, it is then organized, prepared, and processed, as in the data preparation phase in EMC's (2012) data analytics lifecycle, the processing and transformation phases in the KDD process, and the organizing phase in Oracle's (2015) integrated information architecture. This is achieved across a high-speed network using ETL/ELT or big data processing tools. Hadoop and MapReduce, as well as in-memory management can be used for data processing. Moreover, the data can be queried, and computations and processing can be applied using several different languages, ranging from Pig and Hive, to R for statistical computing, to SQL, and SQL-H for directly accessing Hadoop data. Such tools, along with others, can enable big data discovery and preparation for the desired analyses.

Some vendors have also provided a variety of tools, platforms, or appliances to support big data across the storage and management, as well as the discovery and organization steps. These allow for a more comprehensive big data

solution with more features in a single package, rather than having to mix and match technologies. Examples include Vertica, Greenplum, IBM Netezza, Teradata Aster, and SAP HANA.

The next phase in the decision making process is the design phase, where possible courses of action are developed and analyzed through a conceptualization, or a representative model of the problem. The framework divides this phase into three steps: model planning, data analytics, and analyzing. In the model planning step, a model for data analytics is selected and planned. This is similar to the model planning phase in EMC's (2012) data analytics lifecycle, as well as the model selection phase in the KDD process. In this step, the models and algorithms which are found to be appropriate, based on the types of data available and the analyses or output intended, are selected and planned for. A variety of some of the models and analyses which can be chosen are depicted in the framework.

Traditional data mining and advanced analytics techniques, such as classification, clustering, regression, and association rules, can be chosen, along with machine learning and AI techniques such as neural networks, decision trees, and pattern based analytics. Moreover, time series analysis can be used for analyzing sequences of data points which represent values at successive times. Furthermore, text analysis, from documents or social media, social network analysis, and sentiment analysis, can also be selected if the big data is in the form of text, or we are dealing with social media data. Additionally, graph analyses can be used for representing complex networks, and path analyses can describe directed dependencies among variables. Moreover, density based or spatial analyses can be applied for clustering dense areas or dealing with spatial or geographical data, and clickstream analyses can be used for web data and analyzing mouse clicks.

Subsequently, in the data analytics step, the selected model is applied. It may also be accompanied with OLAP, and predictive analytics can be further applied to analyze current and historical data and results so as to make predictions about the future. Furthermore, in-memory analytics and processing can be used with big data in order to enhance and speed the access to and scoring of the analytic models. Several analytical tools and technologies can be used in this step, such as HANA, Greenplum, Aster, Kognitio, Revolution R Enterprise, which is built upon the R language, Teradata Warehouse Miner (TWM), MADlib, Mahout, RapidMiner, Radoop, which is a RapidMiner extension that integrates the data analytics capabilities of Hive and Mahout to provide a data analytics solution for Hadoop, as well as Pentaho, which can perform predictive analytics and OLAP, and can integrate Hadoop as well as NoSQL and analytic databases.

In the analyzing step, the output of the previous step and the results of the analytics are analyzed, similar to the analyzing step in Oracle's (2015) integrated information architecture. Accordingly, the possible courses of action to be taken are defined. These courses are then chosen from in the following phase.

Consequently, the next phase in the decision making process is the choice phase, where methods are used to evaluate the impacts of the proposed solutions, or courses of action, from the design phase. In the framework, this phase is divided into two steps, evaluate and decide. In the evaluate step, which is comparable to the KDD process' evaluation in the interpretation/evaluation step, the proposed courses of action and their impact are evaluated and prioritized. This could be done using reporting, dashboards, simulations of the solutions, what-if scenarios, cognitive maps, heuristics, KPIs, as well as advanced or interactive data visualization. Some of the big data visualization tools available include Gephi, which is mainly a graph-based visualizer and data explorer, Prefuse, Tableau, QlikView, Spotfire, SAS Visual Analytics, Centrifuge, JMP, and ADVIZOR. Additionally, Pentaho also provides big data visualization, as well as reporting and dashboard features.

Accordingly, the next step in the choice phase is to decide on the best course of action, similar to the decision step in Oracle's (2015) integrated information architecture. This is where the decision actually takes place based on the results of evaluating the possible courses of action, and finally choosing the best or most appropriate one.

Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented. In this step, the results of the choice are operationalized, or put to action, as in the last phase of EMC's (2012) data analytics lifecycle. Hence, big data tools and technologies can be used in monitoring the results of the decision, as well as in providing real-time or periodical feedback on the outcomes of the implementation.

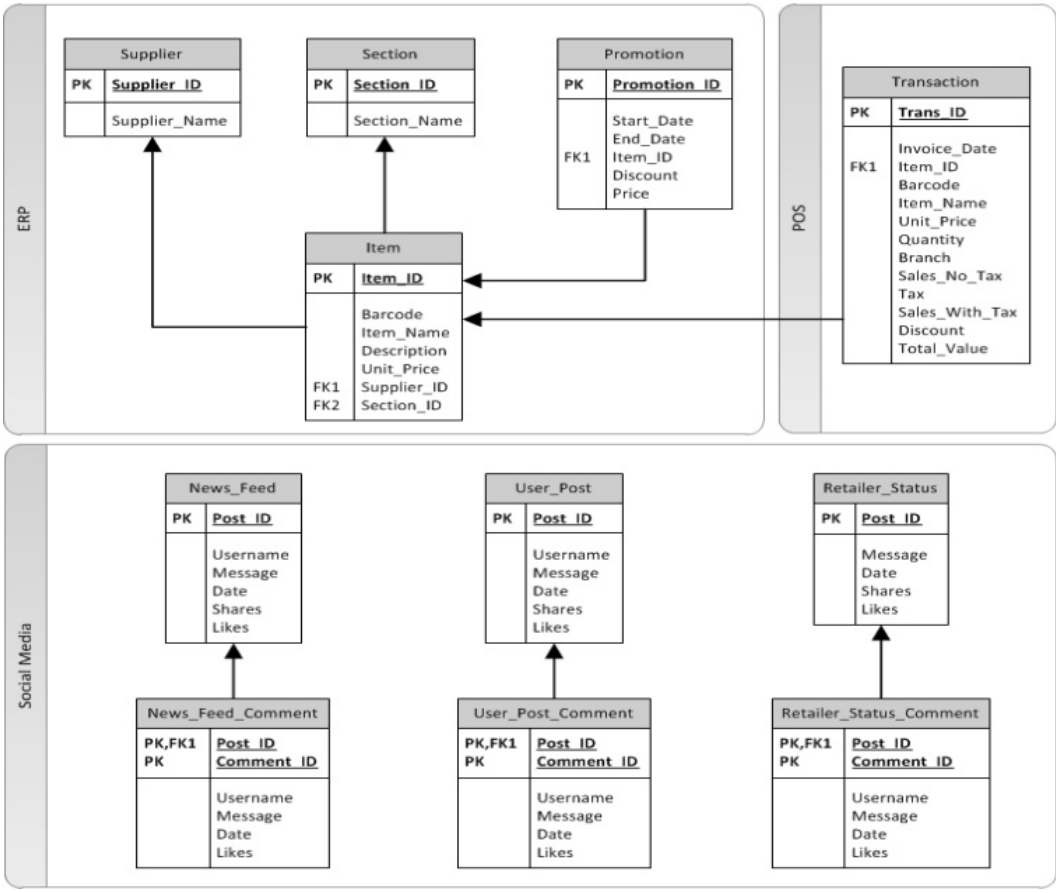

Fig. 2. Data Model

3.2. Framework Evaluation: Retail Experiment

After the development of the IT artifact, the framework then needs to be evaluated. Therefore, an experimental evaluation method was chosen in order to test the B-DAD framework. Hence, an instantiation of the framework was taken and applied to actual data. Moreover, some of the available solutions, such as Aster, Cloudera HDFS, TWM, RapidMiner, Pentaho, Gephi, Tableau, and several DBMSs, were tested in a lab experiment, and the integration and flow between varieties of the provided tools were examined.

The experiment was performed in the area of retail in order to evaluate the B-DAD framework. Accordingly, the decision domain is testing promotion effectiveness and the impact of sentiments and social media on sales. Therefore, the decision would be which products should promotions be offered on, when should they be offered, and whether social media marketing campaigns are efficient and should be focused on or not. By analyzing the available data about item purchases, as well as the feedback and posts of customers in addition to their response to social media, such knowledge should be gained in order to support our decisions. Each phase of the framework is elaborated below. However, the implementation phase was not tested, as in that case the decision would have to actually be executed and monitored over time, which would not be feasible within the scope of our experiment.

1) Intelligence Phase:

In the first phase of the framework, the intelligence phase, the big data which will be used needs to be collected. In this experiment, a mix of relational data, social media data, and text is used. The data model is shown in Fig. 2. In order to get relational data about retail purchase, we chose one of the largest hypermarkets in Egypt to acquire data from their Point-of-Sales (POS) and Enterprise Resource Planning (ERP) systems. The hypermarket deals with more than 80,000 items, and has two large branches, divided into 30 different sections. The number of daily visitors can reach approximately 50,000 visitors. Thus, for our experiment we took a sample of their POS and ERP data, covering the six months, from January 2013 to June 2013. However, the data was in Arabic, which made it very difficult to deal with due to the encodings of the files and the records. The data was then stored in a Teradata DBMS using Teradata loading tools.

As for the social media data, we needed customer posts and comments from the hypermarket's Facebook fan page. No tweets related to the hypermarket were found on Twitter. Consequently, we used the Facebook API in order to extract the posts and comments within the time duration of the sales data we have available. Accordingly, we gathered and stored the fan page's statuses, posts and related comments, and the fans' posts and their comments for the specified time period.

2) Design Phase:

Subsequently, the second phase in the framework tested in the experiment is the design phase. This is where the model planning takes place, and the relationships which need to be identified are defined. In order to find the relationships between the different retail attributes we have, such as the discounts and purchases, several models were planned for. The models and analyses used are briefly described below.

a) Visualization Analysis:

First of all, we needed to understand the distribution of the discounts, quantities purchased, and the total value of item purchases across the different sections and branches, as well as over the six month time period, so we started by using Tableau for visualizing the relationships.

Through visualization, we could see the hypermarket branches and department sections with the most sales, discounts, and profit, as well as the branches. Moreover, a time-series analysis was performed to visualize the relationship between the discounts offered, quantities sold, and the total value of sales across the given time period. Some interesting information, such as the products and times affecting the peak sales, as well as their relationship with the promotions and discounts offered across time was extracted.

b) Correlation and Regression Analysis:

In order to further explore the relationship between the variables, a correlation analysis was performed on TWM. In the resulting correlation matrix, we saw that the variables with the highest correlation to the discounts are the quantity and the sales values. Moreover, logistic regression was performed on TWM to measure the relationship between whether or not there is a discount, and the remaining independent variables. Accordingly, the prediction of having a discount or not was found to be based on nine independent variables, which were determined to be used in the model.

c) Cluster Analysis:

Subsequently, a cluster analysis was performed in order to group the items together based on their similarities, according to the discounts. We started by using the cluster analysis in TWM, with k-means being the chosen algorithm, and the number of clusters being two, since we are focusing on where to provide the discounts, which have the Boolean values of 0 and 1 (no discount and discount). The clustering resulted in "Cluster 1" which has 78% of the items and "Cluster 2" which has 22% of the items.

Afterwards, clustering is performed with Weka, using the same k-means algorithm with the k also equal to two clusters, according to the categories of the discount. The clusters are also scored, using the classes to clusters evaluation, where the discount is the class. Accordingly, after the cluster model is created from the training data, it is evaluated by assigning each value of the class to a cluster, and evaluating the correctness of clustering the points based on whether the discount value of the point matches that of the class it was clustered in or not. Here "Cluster 0" has 43% of the instances, and "Cluster 1" has 57% of the instances, which differs from the results of TWM. This may be as TWM uses the Mahalanobis distance in calculating the distance between the points and the mean of each cluster, while in Weka, the Euclidean distance was used.

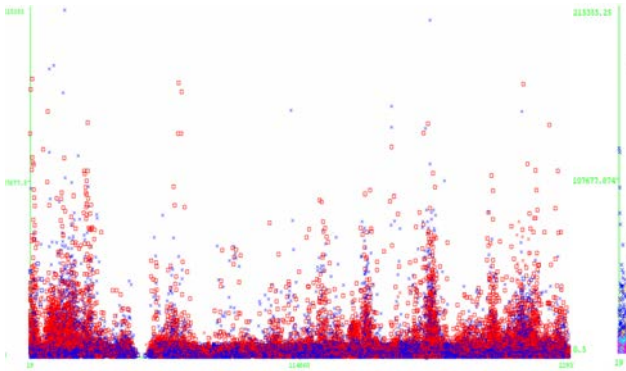

Fig. 3. K-Means Clustering Visualization

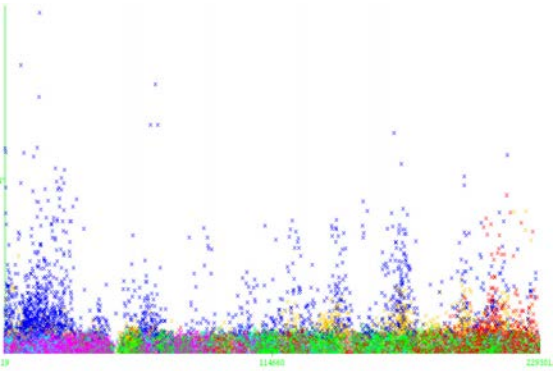

Fig. 4. EM Clustering Visualization

Fig. 3 represents the visualization of the clustered instances. The x-axis represents the item code, the y-axis represents the total value, and the colors represent the discount category. The red points are the ones that were assigned to "Cluster 0", where they should not have a discount, while the blue points are the ones that were assigned to "Cluster 1", where they should have a discount. However, due to the classes to clusters evaluation, the squares are the incorrectly clustered instances, where the red squares represent the items that actually do have a discount, but were put in the no discount cluster, while the blue squares represent the items that don't have a discount, but were put in the discount cluster. Thus, promotions could be put on the blue squares. Moreover, the errors, or increase in the amount of red squares, may be due to the aspect that about 80% of the promotion items which the hypermarket discounts are determined through a promotion plan with the suppliers, rather than by the hypermarket itself.

Additionally, by changing the clustering algorithm, we can find different results. Since having only two clusters was not very representative, we can use the Expectation Maximization (EM) method instead of k-means. The EM clustering method goes through each instance and assigns it with a probability distribution, in order to check the probability of that instance belonging to each of the clusters. The number of clusters was determined by cross-validation. The algorithm divided the data into ten different clusters, where the overall percentages of the cluster distributions are balanced.

Fig. 4 represents the visualization of the clustered instances based on the EM algorithm. The x-axis represents the item code, the y-axis represents the total value, and the colors represent the ten clusters. As we can see, the blue "Cluster 0" instances have a higher item code, the magenta "Cluster 5" instances have a lower total value, and the red "Cluster 7" instances have a higher total value.

d) Association Analysis:

The next analysis performed on the data, is association rule mining in order to discover interesting relations and detect the most common combinations between the variables of our data. We first converted the numerical code attributes to nominal attributes, and then discretized the remaining numerical variables, using equal frequency binning, in order to convert them to categorical variables for the analysis. Consequently, we then performed association rule mining using Weka and its Apriori algorithm, with the specified minimum support being 0.1, and the minimum confidence being 0.25. If the support was raised, then no rules of value would result.

Consequently, several interesting rules were extracted. For example, if the quantity purchased is low, then there is no discount, with a confidence of 84%, and the branch is the second branch, with a confidence of 45%. Additionally, the "Candy, Chocolate & Gum" section doesn't have a discount on 80% of the products. Moreover, if the branch is

the first branch, then there is no discount on the products 74% of the time, while if it is the second branch, then it doesn't have a discount on the products 72% of the time. Furthermore, 60% of the time if the purchases are made from the "Cosmetics" section, then they were made in the first branch. Many more association rules can be extracted from the results, however we could not extract rules based on the purchases of certain items together or perform a market basket analysis, since the data did not include individual shopping cart purchases and was rather aggregated daily invoices.

e) Decision Tree:

Finally, a decision tree was built as a type of classification analysis. TWM was used to build the tree model, by splitting on the gain ratio. The dependent variable chosen was the boolean discount variable, while the independent variables were the branch, item code, quantity of the item purchased, the section and supplier code of the item, the unit price, and the sales and total values. The confusion matrix of the decision tree shows that the accuracy of the model is high, with the percentage of correct classifications being 77.21%, and the percentage of incorrect classifications being 22.79%. Accordingly, the resulting rules from the decision tree showed the factors that affect the decision of whether or not an item should have a discount or not.

f) Social Media Analysis & Text Mining:

Afterwards, we needed to perform text mining on the Facebook data to analyze what people were saying about the hypermarket and its products, as well as their responses to the hypermarket's posts and marketing. Accordingly, we needed to find the most frequent words used in the posts, as well as the associations between these words. Hence, we could gain knowledge about what sentiments, products, or branches people were discussing, and what they were saying about them. For example, if we found that people were frequently dissatisfied with, or complaining about, a certain product, or they were happy with certain promotion, this knowledge can be added to our previous models in order to enhance our decision.

We started by using RapidMiner in order to perform the text mining. The process starts by reading the data from the database and selecting the desired attributes for the analysis and appending the selected data into an example set. Afterwards, the data is processed as a document, where the text is tokenized, and an operator is used for filtering the Arabic stop words, and removing them from the document. Next, the n-grams are generated, and the tokens are filtered by length to remove too long or too short tokens. After the document processing, the numerical and nominal attributes need to be converted to binomial in order to perform association rule mining. Finally, the FP-Growth algorithm is used to extract the frequent itemsets, and an operator is used to create association rules.

However, after several trials of days of running, the process never got past the FP-growth operator. The longest run time was over two days, yet no results were created. This is most likely due to the Arabic language being in Unicode, and being very difficult to process. Additionally, the FP-Growth operator may not be able to work properly with Arabic terms.

On the other hand, the results of the document processing were analyzed, where the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm gives the relative importance of a word to a given document, compared to the importance of that word to all of the documents. Accordingly, we could see the number of occurrences of the most frequent words in the Facebook posts and comments. For example, several interesting words occurred frequently, such as "promotion", "quantity", "valid until", "while goods last", "oil", "offer", "good", "sugar", "Ramadan 10 branch", "oranges", "beautiful", "prices", "Friday", "olives", "olive oil", "lavender", "kilo", "watermelon", "pomegranate", "magazine", etc.

g) Sentiment Analysis:

Next, we performed sentiment analysis on RapidMiner in order to discover the positive and negative opinions of the users on the hypermarket's Facebook page. We started by creating a set of labeled data by taking a sample of 100 posts and storing them in documents in different polarity folders, based on whether they are positive or negative posts. The documents are then processed, and the words are tokenized, filtered by length, stemmed, using light Arabic stemming, and the Arabic stop words are removed. Additionally, the word vectors are pruned to get rid of very common and very infrequent terms, which occur less than 3% or over 95% of the time. Subsequently a vector wordlist is created from the data and stored, as well as a Naïve Bayes classification model using cross-validation. Accordingly the data is divided into a training set and a test set, and the model is built using the training set, and is then applied on the test set and its performance is measured by evaluating its accuracy.

Out of the 64 documents which were predicted to be negative, 40 were correctly classified as negative, leading to a precision of 62.5%, and a recall of 100%, since all of the negative documents were classified as negative. On the other hand, out of the 60 positive documents, 36 documents were classified correctly, resulting in 60% recall, while all of the 36 documents which were predicted to be positive actually were, resulting in 100% precision. Therefore, the total accuracy of the model is 76%.

Next, the built model was applied to the unlabeled posts in order to classify them as positive or negative. However, despite this being a traditional and commonly used form of sentiment analysis, it was not very practical in our case. It requires storing each post in a separate document, and manually labeling the positive and negative training set. Moreover, after applying the model and classifying the documents as positive or negative, we have to manually go back to the document and open it to see the post that was classified.

Accordingly, Repustate is a sentiment analysis and social media analytics website. It has an API for sentiment analysis in five different languages, including Arabic. By using a trial of Repustate on the hypermarket's Facebook page, sentiment analysis was performed on the posts. Accordingly, we could see the negative posts, which can be filtered, for example by time, in order to see which times throughout the day most of the negative comments were posted. From a simple visualization of the negative posts, along with the users who posted them, we can easily find out the people who are complaining about products or services, as well as their negative feedback and opinions. For example, we can see a user providing feedback on the poor delivery, in addition to the post that the quality of the second branch is worse than that of the first branch.

Additionally, we were able to view the positive posts, which are filtered according to the gender of the users who posted them. Furthermore we can view the neutral posts, and the posts can also be filtered by the date on which they are posted, as well as by the device used for posting, whether it is by phone, or by other devices. Finally, we can view the number of positive or negative posts, in order to monitor user sentiments over time, and take actions when posts reach a certain limit on a day. Moreover, the Repustate promoter score calculates how likely users are going to speak positively about the hypermarket, with 10 being the best score and 0 being the worst, score. Accordingly this can allow for taking corrective or preventive measures in order to increase positive posts and reduce negative posts.

3) Choice Phase:

In the choice phase, we need to choose a proposed solution from the results of the previous phase, thus in our case when, on what items, and through which means to provide promotions. However, we do not have enough background knowledge or details about the hypermarket's functions or KPIs to use them for evaluation. Therefore, we chose to use visualization as the main means for the evaluation step in this experiment, as it was found to be the more value adding method in our case.

Moreover, each of the analyses in the previous phase resulted in visualizations which could additionally be used for the choice phase. Additionally, we created new visualizations in this phase in order to see the relationships between the different variables, as well as their fluctuation over time, especially during certain periods. Moreover, to assess the impact of social media advertising on customer purchasing, we took particular instances as examples. For instance, Tableau was used to visualize the more effective method of promotion for a certain juice, in addition to the effect of customer posts and opinions on other purchases.

Additionally, the user interactions with the Facebook posts over time were also visualized, so that we can see the Facebook activity of the users, and focus on days with high activity for adding posts, or for analyzing the user buzz. Furthermore, we can view the number of likes and shares for each post, and we can understand the posts which are most interesting to users, as well as the comments and feedback which grab their attention, or which they support. Moreover, the number of likes on a promotion post portrays user interest or a positive opinion about this particular promotion, and lack of user interaction with posts can portray disinterest.

Therefore, by analyzing the sales of items and discounted items over time, and by incorporating the sentiments and feedback of users, we were able to determine the effectiveness of online promotions and sentiments on purchasing patterns, as well as determine the dates and items during and upon which promotions should be offered.

3.3. Experiment Findings

In this experiment, it was shown that we could make our decision of what items to offer promotions on, and when, as well as using social media as a context for marketing, and its effect on purchasing and gaining customer feedback. Moreover, our framework was followed by using the mapped big data analytics tools and methods within their designated phases of the decision making process. Accordingly, the big data was stored and processed, the analytics were performed and the results were analyzed, and the decision making was enabled, as well as enhanced. Furthermore, the decision was supported by additional information extracted due to big data analytics. Overall, the steps of the framework went smoothly and were valuable and insightful. The necessary modifications are elaborated below.

However, there were still some limitations and drawbacks, not related to the framework, faced during the experiment. These are mainly due to the Arabic language being very difficult to work with. First of all, like several other languages, it is debatable whether a word or phrase is positive or negative, and it depends on the context, and whether the Arabic is formal or slang. However, Arabic is tougher than other languages in respect to the variety of forms a root word can take, based on the tense, context, and sentence grammar. Additionally, the same form of a word can have several meanings. For example, the word "helwa" can be a noun meaning candy, or an adjective describing taste, meaning sugary or sweet, or even an adjective meaning nice or great. Therefore, in our case for example, we cannot know if the intended usage refers to an item, the taste of an item, or an expression of liking. Saad and Ashour (2010) highlighted the complexity of the Arabic language, in that it has a very complex morphology, and it is a highly derivational language, with widespread synonyms as well as variations in the lexical category, whether noun, verb, etc., in different contexts. Moreover, the encoding of the Arabic characters poses a problem, as it has different encodings according to the machine platform, and text preprocessing, mining, and information retrieval can lead to incorrect results if the encoding is not correct. Furthermore, as in our text mining case, Arabic encodings are very difficult in processing, and may not be supported by several tools.

However, despite the complexity, we were still able to extract several important insights from the social media analysis. Accordingly, by merging the results of the different analyses, we could gain unprecedented insights upon which to base our decisions. While we were able to understand the relationships between the attributes, such as quantities, discounts, total values, branches, and sections from performing analytics on the relational data on its own, it was highly supported by the Facebook posts and comments. Otherwise, we would not have been able to understand the effect of social media on the customers' purchasing patterns, and we would have viewed the spikes in sales at certain times, after posting the online promotions, without understanding the underlying reasons. Additionally, we would not otherwise have been able to incorporate the user sentiments into our decisions, and understand how they feel about certain services, promotions, and items.

4. Results

From the experiment, several observations and enhancements regarding the framework have been identified. First of all, as previously stated in the framework development section, the framework serves as a conceptualization of some of the possible approaches to performing big data analytics in support of the decision making process. It was not intended to be inclusive of all the big data tools, technologies, and analytics. There are already several of these available, and they are constantly increasing, so there cannot be a comprehensive list of all possible solutions. Additionally, several solutions, which are not released as big data tools, can be used as well for certain means. In our experiments, although Weka is not - strictly speaking - a big data analytics tool, and is rather an open source machine learning tool, it has several, very useful and simple to use, features and analyses. Accordingly, it is not included in the framework as a big data tool; however it was used in the experiments for additional knowledge, perspectives, and visualizations.

Moreover, it was found that visualization, which was intended for use in the intelligence phase during data discovery, and in the choice phase during evaluating the possible courses of action, was found to serve as an analysis on its own in the design phase, from which valuable insight can be extracted. Consequently, we could visualize important relationships beforehand, that could be further incorporated into the additional analyses, or visualize the results of the analyses for simpler and more comprehensive understanding. Accordingly, it needed to be added to the

design phase in the framework. The same goes for statistics, which were intended to be applied in the intelligence phase, during data discovery. However, it was found that statistical analyses can be used in the design phase as a model or a form of analysis, on its own, or integrated with the other analyses. Hence, in the data analytics step, not only predictive analytics can be used to value and gain insight, but descriptive analytics as well can be applied.

Furthermore, it was found that there does not need to be two different steps, in two different phases for analyzing and evaluating the results of the big data analytics, and that they can rather be merged into a single step for simplicity. Additionally, the analyzing and evaluation should be on the analytics itself, and relating it to the decision domain, rather than on the possible courses of action. This is due to the aspect that after the analytics are performed, the results are analyzed in order to gain insights which can add valuable knowledge and aid in making the necessary decision. However, it is not always the case in which the best scenario would be to first define the possible courses of action, and then evaluate each one in order to select the best one in the decision. Sometimes, there are too many possible courses, or these possible courses are known beforehand.

Also, big data analytics differs from the traditional, more structured, method of finding a business problem, getting data, and analyzing it, and rather goes for a more unstructured approach, where we try to gather all sorts of data, perform analyses to extract whichever knowledge and information we can out of it, and accordingly see how we can apply these unprecedented insights and how they can help in our decision domain. Therefore, we do not always have to know the decision which will be made beforehand, and this case needs to be supported by the framework. Thus, both analyzing and evaluating should be merged into a single step, where either one, or both, could be applied on the results of the analyses.

Additionally, whilst the framework is not intended to be a tool for making the optimal structured and informed decisions, it is a conceptualization of mapping the different tools to the decision making process, and how big data analytics can be used for decision support. Moreover, the framework was found to be a somewhat more flexible, or iterative process, rather than sequential. Accordingly, we should be able to move back and forth between the phases and steps. Since, as previously stated, big data analytics is not very predictable beforehand, and may follow an unstructured approach where we do not know in advance the decision to be made, it makes sense to have to go back to prior stages at times. For example, while performing the analyses in the design phase, we might find that we need more data of different to enhance our models or add additional knowledge. Thus the framework should allow for flexibility between moving back and forth through the steps.

Conclusively, the findings of the experiment were incorporated into the modified framework. The new, tested B-DAD framework is shown in Fig. 5. As depicted, the intelligence phase is the same, without any modifications. However, the design phase now only includes the model planning and the big data analytics steps. Statistics and visualization have also been added as analyses. Moreover, descriptive analytics has been added to the analytics step. Furthermore, the analyzing and evaluating steps have been merged into a single step in the choice phase. Here, the step does not refer to analyzing and evaluating the possible courses of action, but rather analyzing the results of the

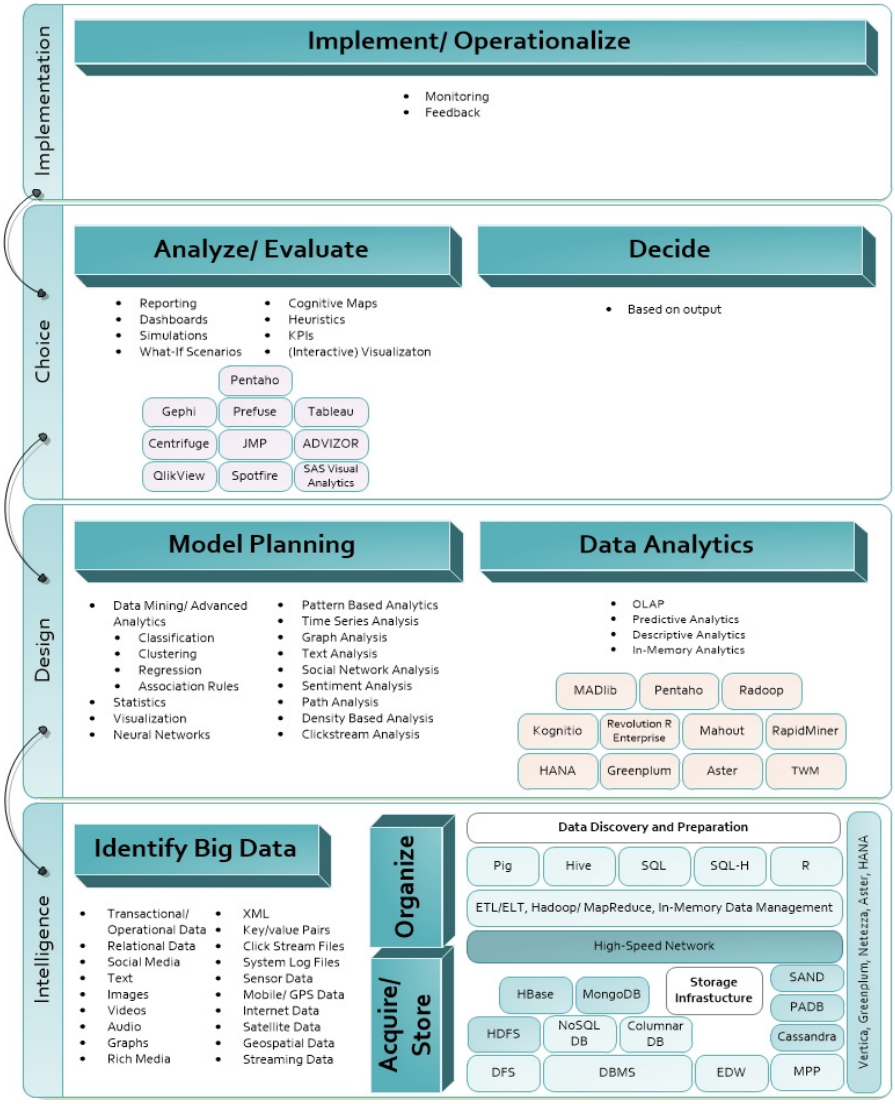

Fig. 5. Modified B-DAD Framework

prior analyses and evaluating their impact on the decision domain. Finally, two-sided arrows have been added between the phases to represent the flexibility of moving back and forth throughout the framework.

5. Conclusion

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge.

By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions. Accordingly, we followed the design science methodology in order to answer

our research question of "*How to integrate big data analytics into the decision making process?*" Consequently, the contribution of this research is the developed and tested B-DAD framework, which guides us through the decision making process supported by big data analytics and the various big data tools and methods.

Design science research should provide additions to the knowledge base, as well as applications in the appropriate environments. Hence, this research contributes to theory and to the knowledge base by perusing the literature and collecting various theories, methods, frameworks, and data analysis techniques from previous research in the knowledge base in order to build the B-DAD framework. Accordingly, it aggregates several parts of these aspects, and integrates and incorporates them into a single framework. Moreover, testing it on a real organizational scenario, which adds rigor and conceptuality, strengthened the evaluation of the framework.

Additionally, the research also provides contributions to the environment. The B-DAD framework can be applied, not only in research, but also in the industry and within organizations. One of the prolonged and constant aims of decision makers and practitioners within organizations is to enhance decision making and gain the highest levels of unprecedented knowledge and insights possible. Accordingly, this research has provided the people and the organizations with the B-DAD framework, which shows them how to integrate and apply big data analytics throughout the phases of the decision making process, in order to provide enhanced and more insightful decisions.

While it was shown that big data analytics could enhance decision making and enable the extraction of unforeseen insights and knowledge, it is no easy task. Other than the time and resource limitations related to most research, one of the main difficulties faced in our case was access to [big] data.

We believe that big data analytics is of great significance in this era of data overflow, and can provide unforeseen insights and benefits to decision makers in various areas. If properly exploited and applied, big data analytics has the potential to provide a basis for advancements, on the scientific, technological, and humanitarian levels.

References

1. Brunswicker, S., Bertino, E., Matei, S. (2015), Big Data for Open Digital Innovation – A Research Roadmap. In: Big Data Research, Vol. 2, pp. 53-58
2. Chang, R.M., Kauffman, R.J., Kwon, Y. (2014), Understanding the paradigm shift to computational social science in the presence of big data. In: Decision Support Systems, Vol. 63, pp. 67-80.
3. Fan, S. Lau, R., Zhao, J.L. (2015), Demystifying Big Data Analytics for Business Intelligence Through theLens ofMarketing Mix. In: Big Data Research, Vol. 2, pp. 28-32.
4. Elgendy, N., Elragal, A. (2014), Big Data Analytics: A Literature Review Paper. In: Advances in Data Mining: Applications and Theoretical Aspects, Springer International Publishing, pp. 214–227.
5. EMC (2012), Data Science and Big Data Analytics. In: EMC Education Services, pp. 1-508.
6. Fayyad, U., Piatetsky-Shapiro, G., Padhraic, S. (1996), From Data Mining to Knowledge Discovery in Databases. In: American Association for Artificial Intelligence, pp. 37-54.
7. Fisher, D., DeLine, R., Czerwinski, M., Drucker, S. (2012), Interactions with Big Data Analytics. In: ACM Interactions, Vol. 19, No. 3, pp. 50-59.
8. Jagadish, H.V. (2015), Big Data and Science: Myths and Reality. In: Big Data Research, Vol. 2, pp. 49-52.
9. Kubick, W.R. (2012), Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26-28.
10. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H. (2011), Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp.1-156.
11. Oracle (2015), An Enterprise Architect's Guide to Big Data: Reference Architecture Overview. In: Oracle White Papers in Enterprise Architecture, pp. 1-54.
12. Peffers, K., Tuunanen T., Rothenberger, M.A., Chatterjee, S. (2008), A Design Science Research Methodology for Information Systems Research. In: Journal of Management Information Systems, Vol. 24, No. 3, pp. 45-77.
13. Russom, P. (2011), Big Data Analytics. In: TDWI Best Practices Report, pp. 1-40.
14. Saad, M.K., Ashour, W. (2010), Arabic Morphological Tools for Text Mining. In: International Conference on Electrical and Computer Systems (EECS), pp. 1-6.
15. Turban, E., Aronson, J. E., Liang, T., Sharda, R. (2007). Decision Support and Business Intelligence Systems (8th ed.), Prentice Hall Publications, Upper Saddle River, NJ, USA