Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1135–1143
DOI: 10.15439/2017F442
ISSN 2300-5963 ACSIS, Vol. 11

A view on the methodology of analysis and exploration of marketing data

Maciej Pondel
Wrocław University of Economics,Poland
Unity S.A. Wrocław
Email: maciej.pondel@ue.wroc.pl

Jerzy Korczak
Wrocław University of Economics,Poland
ICT4EDU, Wrocław
Email: jerzy.korczak@ue.wroc.pl

Abstract—**The paper proposes a methodology for the development of a marketing decision support system using Big Data technology and data mining techniques. The approach was inspired by the CRISP-DM methodology, which is not oriented towards Big Data projects. Therefore, we have modified this methodology with respect to the purpose and technological requirements of the project. The proposed methodology was tested during development of RTOM (Real Time Omnichannel Marketing) project. Project tasks focus on the analysis and exploration of large and heterogeneous data sets. The paper presents the phases of the project implementation according to the extended CRISP-DM methodology, taking into account the specifics of the analysis and exploration processes of large real-time marketing databases. Examples of project steps are also provided to illustrate the approach.**

I. Introduction

DATA exploration is a process of automatic detection of non-trivial, unknown, and potentially useful relationships, rules, patterns, similarities, or trends in large data sets [1]. Generally speaking, the task of exploration is to analyze data and processes it in order to better understand and use it in decision-making processes. Data mining is a multi-disciplinary area that integrates a range of research fields such as information systems, databases and warehouses, statistics, artificial intelligence, parallel computing, operational research, visualization, and computer graphics. Exploration systems use a broad range of information and communication technologies, Web technologies, information retrieval methods, and geolocation techniques, as well as signal processing and bioinformatics.

In this paper, an approach to development methodology of the analysis and exploration of marketing data is presented, adopted in a Real Time Omnichannel Marketing (RTOM) system. In the project, the data is collected mainly in real time and huge sets of data are processed, with high heterogeneity of data sources, formats, volume, and intensity of inflow. The user of RTOM (manager, marketing analyst, etc.) expects acquisition of non-trivial, new and useful knowledge that can be used in the decision-making process. In addition, the knowledge, extracted from the collected data, should be used automatically in customer communication processes to optimize the selected parameters of business process such as purchase probability, customer satisfaction, customer retention risk, product margin, and more. Therefore, our project is not a typical task for most classic Business Intelligence systems, whose implementation is relatively well known [2].

Taking into account the complexity of the project, its innovative character as well as the multiplicity of skills and competences involved in it, and the inherent application of modern information technologies, it was necessary to adopt a uniform methodology of project implementation. In literature, a wide range of descriptions of data mining algorithms applied to generate insightful business analyses can be found, but there is much less information about the methodology of Big Data exploration [1], [2]. This methodology supported by software should enable teams to more efficiently and effectively implement projects entailing real-time knowledge acquisition from very large databases.

So far, several data mining methodologies and process models have been developed. They have achieved varying degrees of success in business applications. According to Gartner, in 2015, 85% of Fortune 500 organizations failed to execute Big Data projects! Those who succeeded were characterized by a high degree of organizational maturity and a good methodological approach [3].

Recent studies of the usage of methodologies in large database exploration projects indicate that the CRISP-DM methodology, proposed by MIT, dominates (42% of applications), followed by own methodologies (19%), while the SEMMA methodology proposed by SAS ranks third (13%) [4]. The other methodologies such as KDDProcess, My Organizations, and domain-oriented methodologies accounts only for a few percent of the market [3],[5]

When selecting the methodology for our project, the following considerations were taken into account:

1) the specificity and complexity of the project, in particular the process of exploring large databases in real time,

2) the need for a pragmatic approach to deliver an application focused on specific sales management and marketing issues,

This work was supported by Regional Research Program, Wrocław, Poland. Grant RPDS.01.02.02-02-0079/15-00

3) organizational maturity and competence of Unity S.A. in the areas of Big Data applications, modern analytical tools and information technology.

As a result of the studies and discussions, we chose the CRISP-DM methodology as a framework. Despite many usage areas, it is not a methodology oriented towards Big Data projects. Therefore, this methodology has been modified to meet our needs, the purpose of the project as well as its technological requirements in mind. In the following sections of this paper, we describe in detail the phases of the project development process, taking into account the specificity of the analysis and exploration processes of large real-time marketing databases.

II. RTOM PROJECT OUTLINE

Real-Time Omnichannel Marketing (RTOM) provides automated and personalized real-time consumer interaction based on the collection and processing of empirical consumer data in a multi-channel sales and marketing model using artificial intelligence and geotargeting algorithms.

The basic assumption of a multi-channel sales strategy is based on the fact that a single customer transaction can be carried out using more than one customer contact channel with the supplier. In a classic multi-channel approach, the seller has multiple customer-facing channels (e.g. bricks & mortar shopping centers, website, on-line commerce systems, mobile application, contact-center, and many more). The omnichannel approach is intended to improve the customer experience. Deplyment of the omnichannel approach requires full integration of off-line channels with those on-line at the business and IT level. Today's Consumer Journey involves a variety of activities and it is carried out in multiple communication channels, as shown in Figure 1.

Omnichannel is a big business and IT challenge, but first of all a chance to fully understand customer needs and behaviors (see [6], [7]).Therefore, data mining tasks and Big Data technology must employed to attain the full implementation of the strategy [10]. The basic requirements for the RTOM system are the following:

1) Building a unified customer profile based on the Master Data Management concept [11], with various types of references between entities implemented, e.g.:

2) Shopping preferences: what size of clothes the customer buys, what colors / styles they choose, their favorite brands, etc.,

- Channels in which the customer contacts the retailer / purchases products / picks products up / gives feedback,
- Time of purchase (e.g. birthday / occasions, holiday, particular season , etc.)
- Final receiver (whether the client buys for themselves, partner / spouse, child, another person).

3) Ability to seamlessly incorporate new artificial intelligence models. Currently available recommendation are based mostly on statistical analysis or simple association rules. In RTOM, we will implement unsupervised learning methods: various clustering algorithms, multi-level associative rules, and also supervised learning methods such as classifiers and predictors. The system must allow the final user to design their own predictive models.

4) Ability to perform analysis of behavioral data not only describing store transactions, but also characterizing the way visitors navigate the website, preform searches and filter data , etc., and how they interact with off-line channels (store visit records, complaint registers, communication with contact centers). The analysis is supported by domain-specific knowledge of the industry / characteristics of the products offered by a selected retail network, for example:

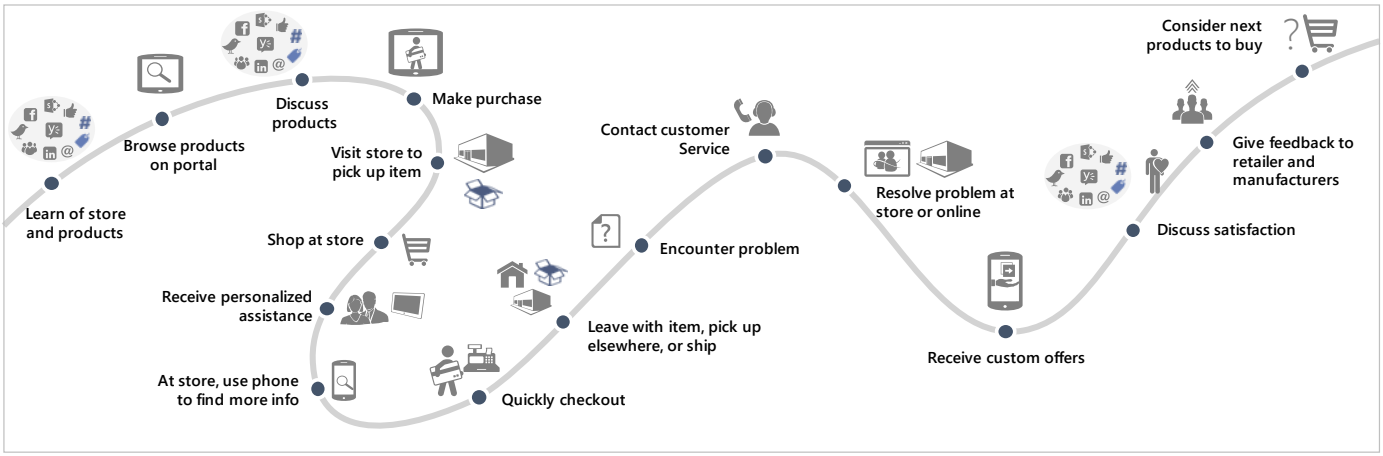

Fig. 1. Customer Experience Journey Map

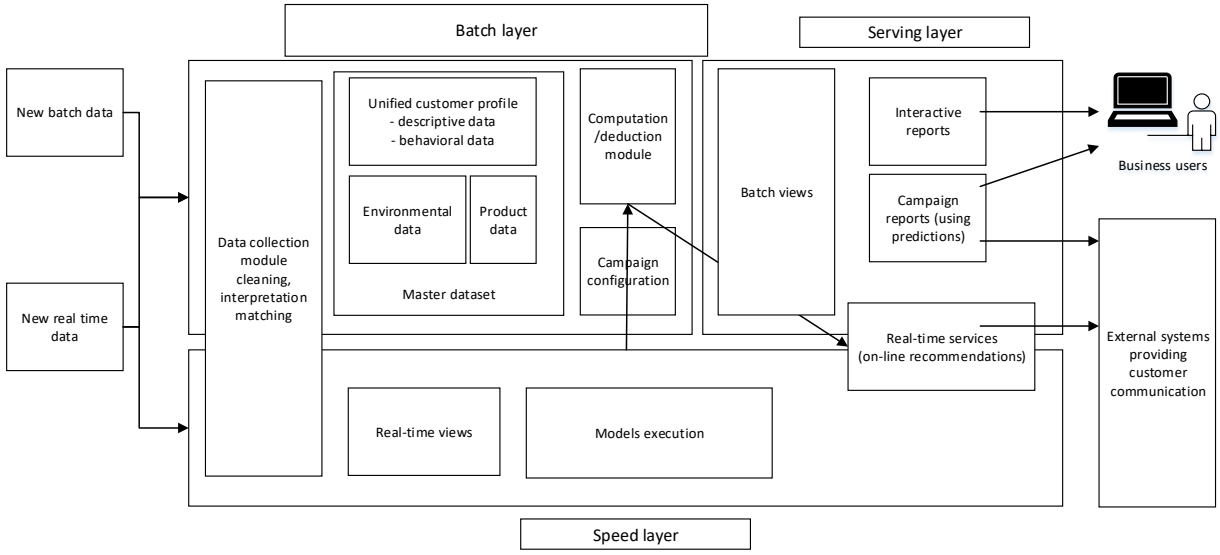

Fig. 2. RTOM general architecture

- Product should be identified by its characteristics rather than on SKU (e.g., the information that the client viewed white running shoes of a given size and brand is far more important for us than the fact that the product's id was 343202043
- Product's designation e.g. the season regards to casual shoes or jackets, and is irrelevant in regard to a wallet and maybe partly relevant to a skirt or t-shirt.
- For whom the product is designed (male, female, youth, infants), which matters in the case of clothes or books, but not for TV sets.

The project is intended to provide a retailer or a client with real-time recommendation of a product purchase, discount or marketing activity in order to maximize a selected customer experience factor (purchase probability, customer satisfaction, customer retention risk, product margin). The recommendations should be delivered by models based on the knowledge gathered from collected data sets and supported by experienced specialists' expertise.

The project also includes features of generating knowledge from collected data in the form of:

- interactive reports facilitating confirmation or denial of hypotheses,
- recommendations for marketing messages directed to individual customer segments resulting from the predictive model but not necessarily generated in a real-time,

Considering the heterogeneity of data sources mentioned earlier, the enormous amount of data and the need to generate real-time response, we decided to base the RTOM architecture on Lambda architecture. Lambda is a reference architecture for scalable real-time data processing systems [9], [10]. As shown in Figure 2, the platform consists of 3 layers typical for Lambda architecture, namely:

- batch layer - storing immutable append-only set of raw data, describing: customer features and customer's behavior (a unified customer profile). This collection is called the master dataset from which we generate batch views. This repository is based on Apache Hadoop and HDFS file system. We use Hadoop based data retrieval mechanisms mainly:
  - Hive[1] (data warehouse software),
  - Impala[2](low latency and high concurrency analytic database for BI/analytic queries on Hadoop)
  - HBase[3] (non-relational, distributed database inspired by Google's Bigtable approach),
  - Cassandra[4], (distributed NoSQL database) etc..
- serving layer – storing indexed batch views, which enables to generate reports in a low-latency and ad-hoc way. It also stores predictive models defined in our project.
- speed layer - real-time views storing recent data only to compensate the batch views with real time data.

The Lambda Architecture aims to satisfy the needs for a robust system that is fault-tolerant, both against hardware failures and human mistakes, being able to serve a wide range of workloads and use cases, and in which low-latency reads and updates are required. The resulting system should be linearly scalable, and it should scale out rather than up [12]. Although we are aware that Lambda Architecture is questioned [13], we decided to use it as a reference architecture but we carefully follow it's indicated drawbacks to avoid potential problems.

[1] https://hive.apache.org/
[2] https://impala.incubator.apache.org/
[3] https://hbase.apache.org/
[4] http://cassandra.apache.org/

III. CRISP DM METHODOLOGY – PROPOSED EXTENSIONS

Many of the mentioned methods and technologies were used to analyze marketing data for the purposes of the project and implementation of the RTOM platform. Unlike most existing CRM systems, we were more focused on analyzing heterogeneous, semi-structured data available in real-time. This required not only broad adoption of Big Data technology, artificial intelligence, and the mobile technology, but also consistent assumption of the appropriate methodology for the design and implementation of the platform. As we noted, the adopted methodology is largely founded on the CRIPS-DM methodology.

The CRISP-DM methodology assumes that each data mining project develops in a specific lifecycle. Unlike the original CRISP-DM version, where the process of project development is divided into six phases, in our approach two stages of CRISP-DM: understanding and data preparation are integrated into one. Figure 3 shows a diagram of the RTOM platform development process. The arrows in the diagram shown below indicate the relationships between the different phases. It should be pointed out that improvement and enhancement of the existing solutions usually follows the five phases. The circle surrounding them symbolizes the continuous adaptation of the solutions to new environments.

The **first phase** consists in defining and understanding the project requirements from a business perspective and pre-planning activities to achieve the project goal. Understanding business considerations includes:

- clear formulation of the goal and requirements of the project using business terminology,
- use of defined objectives and constraints to detail the problem,
- formulation of the initial hypotheses and methods of their validation,
- collection of opinions about the proposed solutions put forward by managers, shareholders and domain experts,
- identification of sources of data acquisition and the scope of required data,
- identification of the necessary tools and information technologies,
- definition of the initial schedule of activities to be undertaken so as to achieve the goals.

In the approach, we assume that the formulated hypotheses are pre-validated on a sample of source data by a data analyst, using the Orange [5]data mining platform . The analyst should document their work and provide the first version of the models with an I/O description (including the definition of variables and required data preprocessing).

The milestone of this phase is the elaboration of documentation containing answers to the above-mentioned points and documentation of the pre-model (models) developed on the Orange platform.

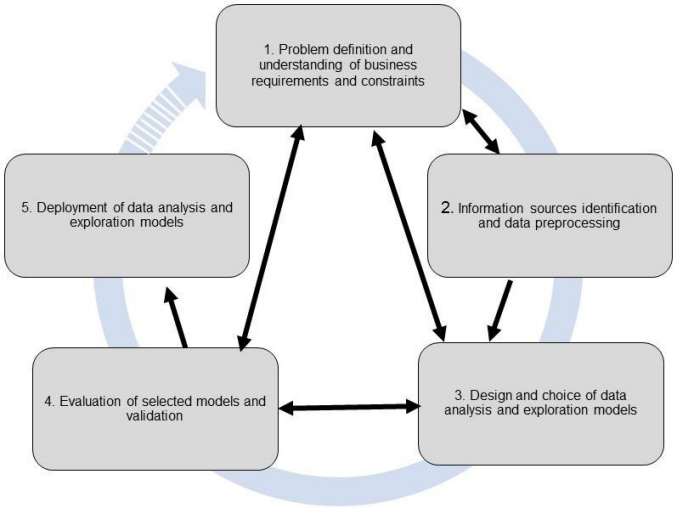

Fig. 3 Phases of the platform development methodology

To illustrate the approach, we apply the example of one of the tasks solved on the RTOM platform, i.e. customer clustering. In this application, clustering can concern customers, products, transactions, and customer contacts with store web pages. For example, in our database we have several thousand customers, each described by several dozen attributes of various importance. The goal of clustering is to find clusters of similar customers to whom we can send an offer or whom we can target when promoting specific products. We require clusters to have specific statistical characteristics (such as minimum variance) and usefulness in marketing decision making (e.g. determining loyal customer groups). Clustering is expected to ensure that promotion of store products becomes more effective, which will be specifically expressed in sales profitability ratios. In this phase, the data have to be identified; in our case, they are transactional systems, CRM, geolocation data, social networks and logs of store web services.

Working on the problem, it is extremely important to formulate preliminary hypotheses and to assess the proposed methods of achieving the goals set by the company's managers, shareholders and domain experts. What is innovative, in terms of methodology, is to develop a prototype of a model and perform initial validation on a simplified case, using an easy-to-use data mining tool. One such tool is an open source visual programming platform Orange. The clustering process diagram is shown in Fig.4.

The obtained results together with the cluster visualization allow one not only to better understand the problem and to clarify business objectives, but also to perform an initial validation of the solution.

[5] The Orange platform is an easy-to-use data mining tool with a rich graphical interface and functions for data analysis, classification, clustering and prediction. The visual design of the data exploration process together with the ability to expand functions in Python make Orange a tool used very often by analysts. More information about Orange can be found on the web site of the University of Ljubljana http://orange.biolab.sl.

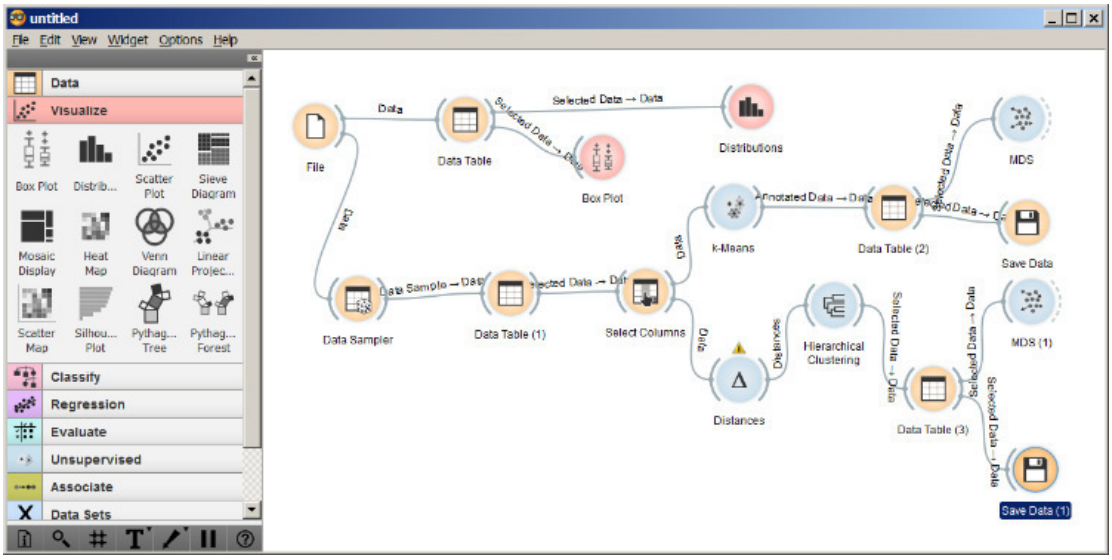

Fig. 4 Visual diagram of customer clustering

The **second phase** concerns the identification, understanding and preparation of data. In our approach, compared to the original CRISP-DM, we integrated two stages: the understanding and preparation of data. From all the phases, it is the most iterative and time consuming work. The main task is to collect pre-process data to be used by the tools and data mining algorithms. In the context of Big Data technology, the data is collected in so-called data sandboxes. Technically, a data sandbox consists of massively parallel processors, extensive memory, and I / O mechanisms that ensure the scalability of the data collection processes and the independence of the operational database systems [14]. Thanks to this, the sandbox provides the ability to carry out complex data analyses without interrupting the operation of the company's information systems. The collected data may be of heterogeneous types; they can come from transactional systems, mobile devices, OLAP cubes, telephone logs, Web logs, and the Internet. It should be taken into account that the size of a data sandbox may exceed many times the size of a company data warehouse.

It should be noted that although the sandbox data is shared by data analysts and exploration modules, at the same time the sandbox platform has to ensure data security and confidentiality.

The second important task of this phase is the preparation and transformation of data according to the Extract-Load-Transform (ELT) scheme. The value of the ELT is that it preserves data in its original form in the database. Therefore, the analyst may freely convert it or leave it unchanged. As far as this job is concerned, it is important to control the quality of the collected data and provide statistically useful measures. The last task is to organize and design the transformation process of raw data. Typical transformation operations include attribute analysis, , data cleaning and normalization, completion of missing information, etc.

In the project, the data sandbox platform is run under Linux; we apply the NOSQL database technology available on the Hadoop platform, and processing compliant with the MapReduce paradigm available in the Spark engine [15].

The phase milestones are the development of technical documentation and the creation of the sandbox for RTOM. For example, in the RTOM project, the main source of data is the transaction processing system and customer logs with the store's web application. The database schema is illustrated in Fig.5.

In addition to transactional data, the sandbox collects data from all marketing channels, which include customer geo-location data or data pertaining to customer activity in social networks.

The **third phase** of the process focuses on the design and choice of the data mining model. While in the previous phase we were concentrated more on data quality, at this phase we undertake the problem of discovering the relationships between variables in the area of specific business problems. We use the documentation of the preliminary version of the model (models) previously prepared on the Orange platform. It is important here to engage the domain experts who might suggest variables that can influence the solution and to accept or reject the hypotheses defined in the first phase. In particular, it may concern interpretation of correlation and causal relationships.

The choice of attributes is crucial for the performance of data exploration. The analyst must be open to the prospect of examining various algorithms, their parametrizations, and the composition of the input vectors. The design of the input vector and the data mining models is an iterative processs. Model learning and testing on all possible variables is usually impractical. In order to reduce the dimensionality of space, analysts can consult the experts who will suggest important variables or use algorithms that rank variables according to criteria such as the Gini index, information gain, χ^2, ANOVA, or the rate of entropy reduction.

There are many data mining models. Generally, they fall into three categories: classification, prediction and

clustering[6]. In the RTOM project, we restricted the offer to the models available in Apache Mahout [7], Spark MLlib [8], Tensorflow Core[9] and Pandas[10] [16], [17].

To provide an example of our work, we applied the clustering models available in Apache Mahout[11], Spark MLlib and Tensorflow Core libraries [8], [15]. From the available clustering models, the k-means model [1], [10] was chosen. The following fragment of the code illustrates a part of the model specification

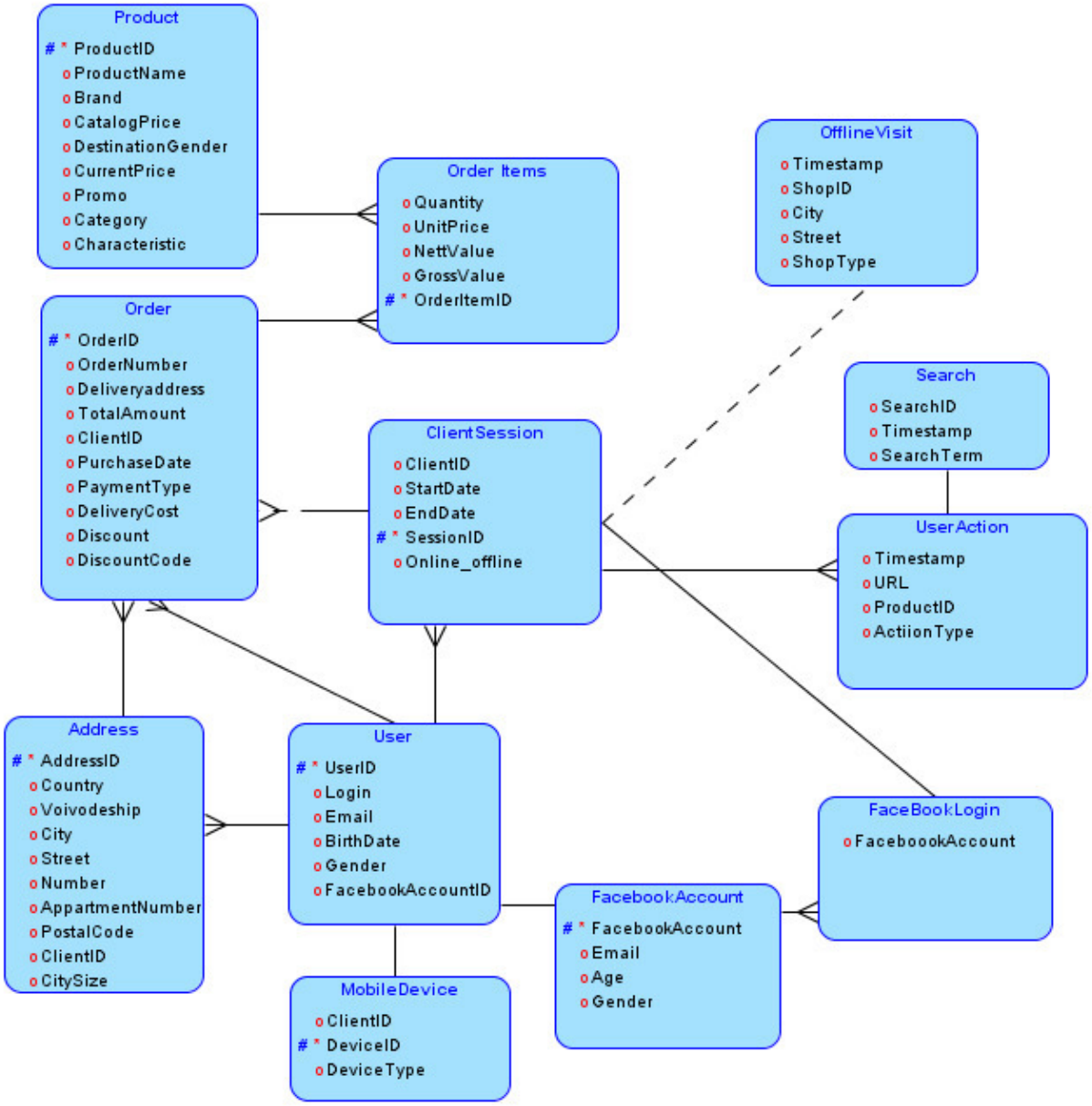

Fig. 5 Conceptual diagram of the database

Presented model code is intended to build clustering model of customers. Before we start we need to calculate aggregate values describing clients' behavior. In the input dataset one row represents one client for whom we select his or her birthday, gender and we calculate the date of the first client's order, number of orders and their total values in 2016 and 2017 as well as discounts in 2016 and 2017.

[6] Classification and prediction are very similar and generally related to the type of data used to build a given model. If the decision attribute is categorical, then the predicate problem of the value of such an attribute is presented as a classification problem. If the decision attribute is continuous (numeric), the problem is called a prediction problem.

[7] http://mahout.apache.org/users/basics/algorithms.html
[8] http://spark.apache.org/docs/latest/ml-guide.html
[9] http//www.tensorflow.org/
[10] http://pandas.pydata.org/

```
  %%spark -o vector_df
  from pyspark.ml.feature import VectorAssembler
  from pyspark.ml import Pipeline

  def create_vector_assembler(col):
      vector_column_name = 'v_' + col
      input_cols = [col]
      return  VectorAssembler(inputCols=input_cols,
outputCol=vector_column_name)

  columns = [
       'min_order_place_date',
          'birth_date',
          'number_of_orders',
          'promotion_counts',
          'sum_2016',
          'avg_disc2016',
          'sum_2017',
          'avg_disc2017',
      'gender'
  ]
vector_columns = map(lambda c: 'v_' + c, columns)
vector_assemblers  =  map(create_vector_assembler,
columns)
pipeline = Pipeline(stages=vector_assemblers)
vector_df                                          =
pipeline.fit(df).transform(df).select(vector_colum
ns)
  vector_df.cache()
  from numpy import array
  from math import sqrt

from   pyspark.mllib.clustering   import   KMeans,
KMeansModel

def to_training_point(data_frame):
    return array([getattr(data_frame, column_name)
for column_name in column_names])

  column_names = d.columns
  training_data = d.map(to_training_point)
  clusters   =   KMeans.train(training_data,   8,
maxIterations=10,
          runs=10, initializationMode="random")
```

To interpret the obtained clusters of customers, we built a decision tree using the library *pyspark.mllib.tree*. The tree rules allowed interpretation of customer groups in marketing terms. The following fragment of code illustrates the process of decision tree generation.

```
from   pyspark.mllib.tree   import   DecisionTree,
DecisionTreeModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.regression import LabeledPoint

ttd=sc.parallelize(map(lambda(p,l):
LabeledPoint(l,p),cl))

tree=DecisionTree.trainClassifier(ttd,numClasses=8
,categoricalFeaturesInfo={},impurity='entropy',
maxDepth=5,maxBins=4,minInstancesPerNode=10,
minInfoGain=0.1)
  names   =   dict(map(lambda   (k,p):   ('feature
{}'.format(k),       '{}       ({})'.format(p,k)),
enumerate(d.columns)))
  res = tree.toDebugString()
  print(res)

  for f,t in names.iteritems():
      res = res.replace(f,t)
  mytree                                           =
sqlContext.createDataFrame(sc.parallelize([(res,)]
), ['c'])
mytree.show()
```

The milestones of the phase are documentation of data exploratory models, a set of models along with specifications of the data used in the learning, testing and validation processes.

The aim of the **fourth phase** is the assessment of the quality of the developed data mining models. The prerequisite for the task is to clearly define the evaluation criteria. The evaluation problem is a multi-criterial one [18]. In addition, however, it often happens that managers, shareholders, and experts impose supplementary priorities during the model evaluation.

In general, the models should be evaluated in terms of quality and efficiency before being implemented on a sample of sandbox data. Two-step model testing is recommended here: first – on a pilot trial, later – on full information resources. As a result, the cost / modification time of the model can be reduced due to simple errors or oversights, thereby diminishing the risks associated with testing and validating the production version of the platform. It is advisable to gradually extend the scope of assessment, e.g. to product categories, selected sale channels or market regions. When launching a model in the real life environment, the assessment should first focus on detecting anomalies in the input data before running the model. The model's performance is evaluated not only in terms of quality and efficiency but also in terms of mutual cooperation with other platform resources. This action permits one to formulate operational recommendations for the model deployment under real conditions

It is very important to prepare datasets for model building and evaluation (model learning, testing and validation.) The quality of selected models is assessed according to predefined business criteria and generally accepted assessment criteria for each category of data exploration models.

In the example discussed here, the proposed clustering models were evaluated. In general, the measures of evaluation can be divided into two categories: internal assessment of clustering results and evaluation based on external criteria.

Using the internal criteria, we evaluate the clustering hierarchy taking into account the similarity of instances within clusters and the similarity between clusters. The following measures can be applied [1],[18]:

- Davies-Bouldin index

$$DB = 0.5n \sum max ((\sigma_i + \sigma_j)/ d(c_i,c_j)$$

where n is the number of clusters, c_i and c_j are cluster centers, σ_i and σ_j are the standard deviations, and d is a distance between cluster instances and centroids.

The algorithm that generates the lowest value of the DB index is considered the best according to this measure.

- Dunn index

$$D = min (d (i, j) / max d'(k)$$

where *d (i, j)* is the distance between clusters *i* and *j*, and *d'(k)* is the distance measure within clusters *k*.

Dunn's index focuses on cluster density and the distance between clusters. Algorithms preferred by the Dunn index are those that reach high values of the measure.

In external evaluation methods, clustering results are assessed using external data, not taken into account during the clustering process. For instance, such data concern the customers who were previously assigned to the clusters by experts. In this case, the clustering results generated by the algorithm are compared with the clusters determined by the experts. The following can be cited among the measures :

- cluster homogeneity index calculated according to the formula:

$$WJK = 1/N \sum max \mid m \cup d \mid$$

where *M* is the number of clusters created by the algorithm, and *D* is the number of the expert's clusters.

- Jaccard index measures the similarity between two sets of observations according to the following expression:

$$WJ = TP / (TP + FP + FN)$$

where *TP* means True Positive, *FP* False Positive and *FN* False Negative rates.

For two identical sets $WJ = 1$.

- Rand index is sensitive to false clustering decisions and calculated according to the formula:

$$WR = (TP + TN) / (TP + FP + FN + FN).$$

The Rand index, as the previous ones, is based on a comparison with benchmark classes given by the expert. It provides information about the assessment of similarity between the correct decisions of the clustering algorithm and those on the benchmark.

Apart from these metrics other measures can be also applied, such as F-score, Fowkes-Mallows index, etc.

Marketing analysts often map clustering results in the form of Multi-Dimensional Scaling (MDS) diagrams; an example is shown in Figure 6. The MDS diagrams not only ensure easy visual assessment of clusters and their dispersion, but also indicate outliers. The described measures allow determining whether the selected models meet all the business requirements and demonstrate the hypotheses defined in the first phase of the methodology. In the event of positive evaluation by managers, shareholders and analysts, it is possible to deploy the model and disseminate the results.

The milestone of this phase is an evaluation report of data mining models containing the above described indexes and their interpretation.

The final phase of the methodology is the deployment of positively assessed models and the RTOM platform. The deployment takes place in two steps. First, the pilot version of the platform is implemented in the real production environment and the results are evaluated in terms of content, usability and performance. The reports are assessed by managers and business analysts for their correctness, completeness, and usefulness in decision-making. At the same time, the platform is monitored by designers and future system administrators. The monitoring is mainly about the computational efficiency, and the use of computational and memory resources. Earlier validation of the pilot version allows us to limit the risk of the full version's interference with all other components of the company's information system . It also provides time for adjustments and fine-tuning before implementing the full version of the platform.

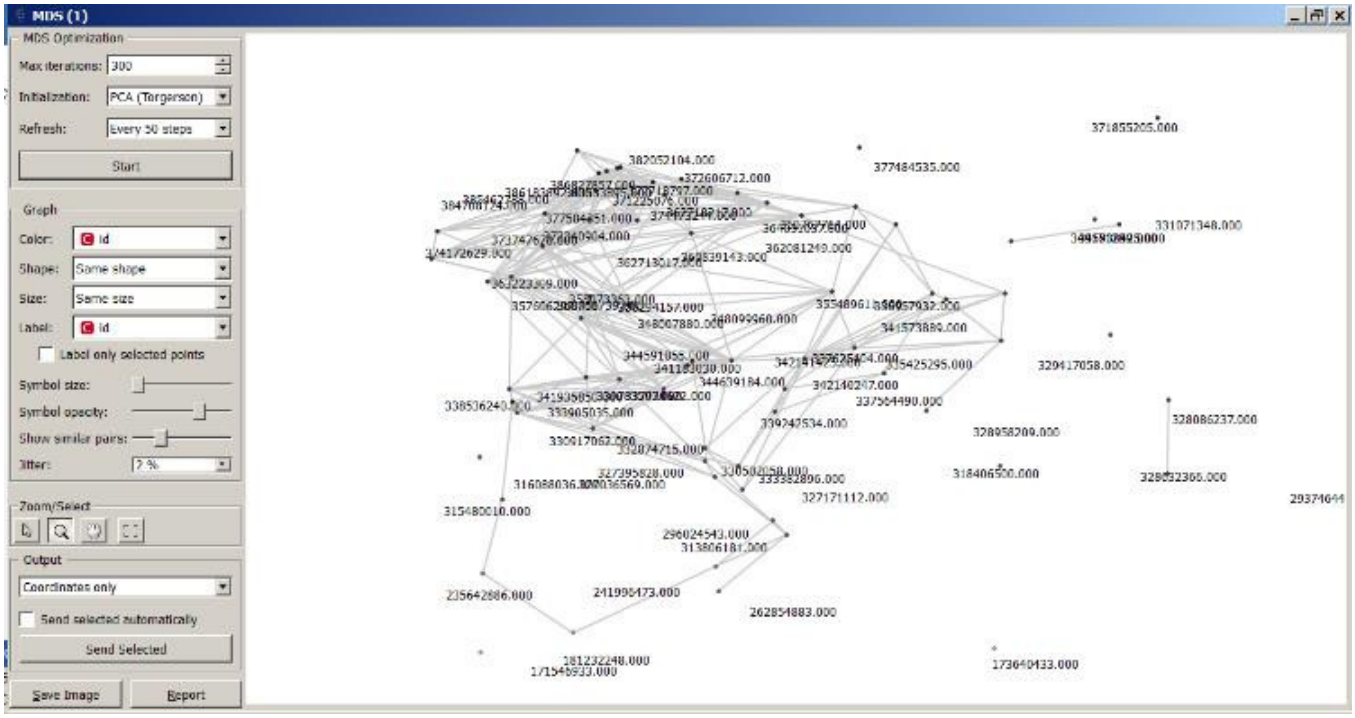

Fig. 6 MDS diagram of customers

In the second step of this phase, the platform runs in a full production environment. Performance results are disseminated to users, who often have to undergo additional training. Also, new organizational roles are then defined and new specialists are employed. It should be pointed out that the new business and technological solutions are revolutionizing the marketing practices and data processes hitherto in use.

The process of improvement of decision making systems never ends. With the advancement of new information technologies, the data mining methods impose improvements of information systems. Therefore , after the deployment, we should think and plan future updates and enhancements. In Figure 1, further development of the platform is illustrated by the dashed arrow leading to the first phase of the process.

The main milestones of the phase are the following:

- the application deployment plan and dissemination of results
- the schedule of platform monitoring and maintenance,
- the final report and technical documentation.

IV. Some comments on Big Data in the context of the proposed data mining methodology

The proposed methodology was presented in the context of the RTOM development, entailing Big Data technology and real-time business data processing. The phases of exploration discussed above show that the approach is different from that applied in Business Intelligence type solutions. In these systems, despite apparent resemblance, we do not deal with massive data streams coming in real time [10], neither do we have to solve technological problems related to the scalability of the application and the heterogeneity of data. The problem of integrating various software components and the efficiency of the exploration processes is also less important. Therefore these aspects were what we we tried to emphasize in the methodology adopted for the development of the RTOM platform.

Summing up, certain key issues for the RTOM platform development have to be highlighted: namely:

- Data quality and volume of data. The studies have shown that as the data streams from different sources increase, their quality deteriorates. Therefore, the processes of data collection and preparation are extremely important in the RTOM project. In consequence, data quality determines the quality of exploratory models as well as the usefulness of the generated results. This particularly applies to cleaning and noise filtering processes and algorithms for completing the missing data stored in the sandbox.
- Availability of models. Today most algorithms and models of data exploration are available in software libraries, some references were cited in this paper. Therefore, there was no need for presenting a full specification of models and programming them from scratch. More important from the user's perspective was to describe algorithm profiles ***with their*** parameterization and interface for various, useful components of the RTOM platform, for example related to model evaluation or visualization of data and results.
- Hadoop, the open source Apache product, is not a data mining platform; it is one of the tools for management and operation on very large data sets [14]. Undoubtedly, Hadoop, MapReduce and HDFS components improve the performance of systems on large, distributed datasets. It should be noted, however, that Hadoop works well on linear case studies, while most business applications are nonlinear problems. Therefore, in our methodology we extensively used, among other things, Apache Mahout and Apache Spark MLlib, which provide efficient data mining tools using Hadoop.
- Interpretation of results and their use in decision-making. any of the data mining models are rated for quality, accuracy and performance. In business applications, we must take into account the economic criteria of the cost, and the specific measurable and non-measurable effects of the model. Apart from those, features also important for managers include the ease of understanding and interpretability.

References

[1] Witten I., Frank E., Hall M., Pal C., (2017) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufman
[2] Shmueli G., Bruce P., Stephens M, Patel N.,(2017), Data Mining for Business Analytics, Wiley
[3] Piatetsky-Shapiro G., (2014), KDnuggets Methodology Poll
[4] Shearer C., (2000), The CRISP-DM model: the new blueprint for data mining, J Data Warehousing; 5, pp. 13-22
[5] Azevedo, A. and Santos, M. F., (2008), KDD, SEMMA and CRISP-DM: A parallel overview [In] Proceedings of the IADIS European Conference on Data Mining, pp. 182-185
[6] Frazer, M., Stiehler, B. E. (2014,). Omnichannel retailing: The merging of the online and offline environment. In Proceedings of the Global Conference on Business and Finance (Vol. 9, No. 1, pp. 655-657).
[7] IBM (2011), Introducing Apache Mahout". ibm.com. 2011
[8] Rigby, D., (2011), The Future of Shopping. Harvard Business Review, December 2011.
[9] Karau H., Konwinski A., Wendell P., Zaharia M., (2015), Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly
[10] Marz N., Warren J., (2015), Big Data: Principles and best practices of scalable realtime data systems, Manning Publ.
[11] Chorianopoulos, A. (2016), Effective CRM using predictive analytics. John Wiley & Sons.
[12] http://lambda-architecture.net/
[13] https://www.oreilly.com/ideas/questioning-the-lambda-architecture
[14] White T., (2015), Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, O'Reilly
[15] Ryza S., Laserson U., Owen S., Wills J., (2015), Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly,
[16] Laserson U., Owen S., Wills J., (2015), Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly
[17] Owen S., Anik R., Dunning T., Friedman E., (2012), Mahout in Action, Manning Publ.
[18] Shmueli G., Patel N., Bruce P., (2010), Data Mining for Business Intelligence, Wiley