Data Mining and Cloud Computing

Robert Vrbić

University Vitez, Travnik, Bosnia and Herzegovina

Critical review

DOI: **10.7251/JIT1202075V** UDC: **37.018.43+371.331]:004**

Summary: Cloud computing provides a powerful, scalable and flexible infrastructure into which one can integrate, previously known, techniques and methods of Data Mining. The result of such integration should be strong and capacitive platform that will be able to deal with the increasing production of data, or that will create the conditions for the efficient mining of massive amounts of data from various data warehouses with the aim of creating (useful) information or the production of new knowledge. This paper discusses such technology - the technology of big data mining, known as Cloud Data Mining (CDM).

Key words: data mining, cloud computing, cloud data mining, NoSQL

Introduction

Every day, as a consequence of business activities and, especially in recent years as a consequence of using on-line services, enormous amounts of data are being created. All this accumulated data is potentially hiding in (useful) information, such as the buying preferences, financial situation, interests, political views etc. of users or clients, which can significantly improve the decision-making. But, how to get to this hidden and potentially useful information, which is hidden in the "sea of data" when processing and storing large amounts of data, which is daily multiplying, represents a significant problem and reveals certain limitations of the traditional information and communication technologies and tools? The answer is, of course, the application of modern technology. Cloud infrastructure can be effectively used for intensive and demanding operations with data that is typical for processes of data mining. It is necessary to have available scalable data warehouses and scalable computing resources that are capable to accept, efficiently store and deeply analyze such large amounts of data, and Cloud offers that, without huge investments that are necessary if one wants to build a DM system within the IS company or organization.

Generally About Data Mining

For years, companies and other organizations "accumulate" large amounts of data and in the past few years the volume increased manifold. The question is: is some useful, some hitherto undiscovered information hidden in this data? The answer to this question can provide the application of Data Mining process (DM).

In essence, data mining is the process of discovering or finding some new, valid, understandable and potentially useful forms of data. The form of data refers to a discovered regularity among the data variables. If the detected regularity applies to all data, then it is about discovered model, if, however, the regularity can be correlated with the extent of data – it is a pattern or template.

Data mining is carried out over large volumes of data in order to „pull"new information out of them that will be the basis for making (better) business decisions.

DM is highly multidisciplinary field, which has its roots in statistics, mathematics, information theory, artificial intelligence, machine learning theory, data-

bases and in the whole series of other related fields. It can be said that Data Mining is the natural evolution of technology, which uses the concepts, methods and techniques of the disciplines above. That evolution has began at the same time when the data was stored for the first time in the computer system and it continues with the advancement of technology access to data, and in the last few years, with the development of modern technologies which enable customers navigation through the data in real time, this new field of in-depth analysis of data and discovering new information is reaching its peak.

DM field is closely associated with the technologies of data warehousing and systems for database management, whose development directly affects the progress in the field of data mining. DM involves activities of searching large databases and data warehouses with the aim to find the hidden, so far unknown facts, regularities or patterns. With mining, it is possible to identify the following types of information: classes, clusters (categories), conditional association events (e.g. customers who buy product A, in 70% of cases they buy the product A1), sequences, which are establishing events that in a certain probability follow one after the other and forecasts, which predict the future from the existing data.

Data mining is a complex and challenging activity, or set of activities, whose implementation requires experts from different fields. It is usual that in the DM project are participating:

- Computer scientists - their role is data preparation,
- Analysts - their task is the choice of method and methodological interpretation of the results of mining,
- Experts - they are familiar with the problem domain, define a business problem, choose relevant data and suggest activities on the basis of the obtained results.

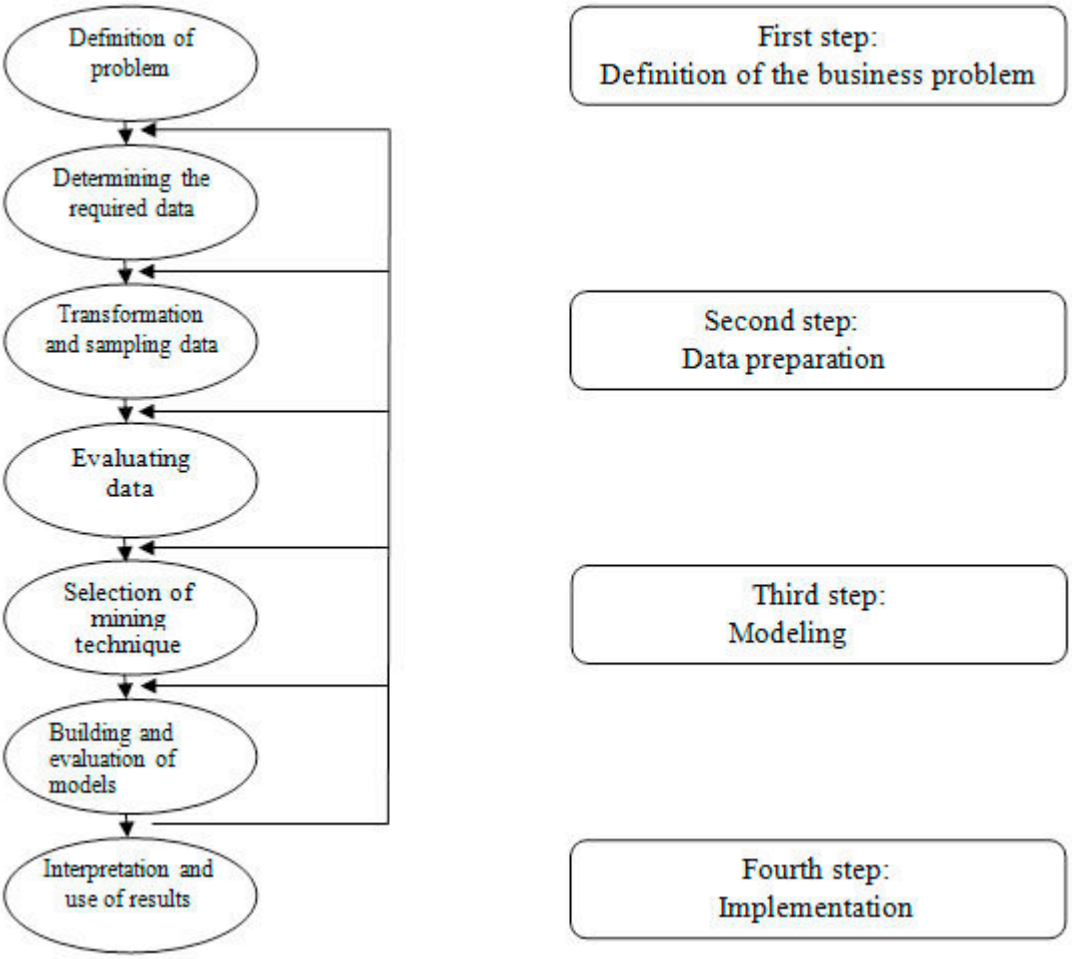


Figure 1: Phases of the data mining process [2]

It is recommended that in the project team there is one more person- the project manager, whose role is to coordinate and lead organizational DM project.

There is no prescribed procedure for data mining, which will surely and always result in finding valuable information. However, it is possible with the planning approach, following the standard steps (phases) of data mining process, significantly increase the probability of success.

Building a model is particularly important step in the process of data mining. It is a complex process that involves several activities: selections of data mining technique, identifying the case, the choice of entities that need to anticipate, identify data for analysis, optional creating dimension and virtual cube from the resulting model and processing of model and collecting the results. When creating the DM model, the biggest problem is how to apply different techniques (and different algorithms) to different sets of data, with the aim to find interesting, important and useful patterns.

A huge amount of complex and disparate information does not permit the application of the same algorithms, or technique of mining, so the role of the analyst – an expert in the field of data mining is particularly important because it is in his competence the decision on the choice of tools, techniques and methods that will be used in specific cases. In one data mining project it is possible to choose by using multiple methods, where the procedure itself is carried out in the same way as in the case of choosing one method. If it is decided that the chosen method, or methods, is (or are) inappropriate, parameters of chosen method can be changed or the selection of the new method can be made. Some of the most used methods and techniques of DM are: classification, associations, sequential analysis, clustering, prediction, neural networks, fuzzy logic, decision tree, market basket analysis and memory based reasoning.

In the context of this work, the application of data warehouses is very important. Specifically, for the purposes of modern companies that operate in the global market, and whose IS does not end at the front door of the company, it is required such form of organizing (and management) data which is based on the concept of data warehousing. Data warehouses merge, or integrate the data from different sources, historical data on the management of company and the data (of interest) from the environment. Data warehouses, according to the technical requirements and content, are significantly different from the "standard" transaction-based systems and designed to provide one with an easier data search, their analytical processing and reporting. Data warehousing is an important concept of effective decision support system which is extensively developing in the last few years. It brings the idea of active finding and offering the information needed in the decision (business) making process It uses the procedures of analytical processing, data mining and knowledge discovery from data. By mentioning concepts and methods based on information technology, the aim is to achieve intelligent management of the company in today's complex market conditions.

As already noted, the data enters into DW from various sources, including the company's transactional systems. The most important and most comprehensive work in the process of data storing is the integration of such data and organizing the content. These activities are part of ETL (Extract, Transform, Load) process, whose task is to capture or collect data from heterogeneous sources, transform it in the appropriate format and fill the warehouse with such prepared and filtered data. Although data mining can

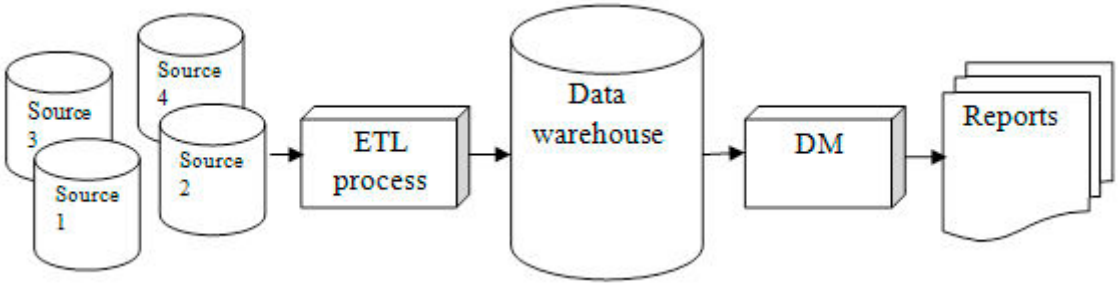


Figure 2: A simplified representation of DW as a source of data for the DM process

be conducted without the data warehouse, it is clear that its establishment and application significantly increases the chances of success of the DM process.

Cloud Computing

In recent years, one of the most important and interesting topics of the ICT world is certainly - Cloud Computing, therefore here will be presented only basic information and features of this technology or business model.

NIST (National Institute of Standards and Technology) defines Cloud Computing as a model that provides a ubiquitous, simple and on-demand network access to a shared set of resources (e.g. network resources, servers, data storage, applications and services) that can be readily available for use, or, if necessary, shut down, and all with minimal intervention of service providers [1]. Influential Gartner & Forrester provides the following definition: "CC is the area of computing in which scalable and highly resilient IT facilities provide in the form of services delivered via the Internet to numerous external customers." A large number of experts believe that this is about a new business model and technology platform for accommodation, launching and usage of various IT services and products. Seen from the users' point of view, Cloud Computing can be defined as a new, cheaper and safer (?) way of using software solutions which will be leased as needed. On the other hand, from the aspect of the service provider, Cloud Computing can be defined as a new way, new technology and different distribution channel primarily of IT products and provision of IT services.

Despite the large number of definitions, which have as a focus different aspects of this, and which are still regarded as controversial, business and technology model is possible, on the basis of everything previously said about them, to know the basic idea, and the possibilities offered by the concept of Cloud Computing which will surely mark, to a lesser or greater extent, the world of information and communication technologies. CC has already become a phenomenon that engages, in one way or another, the whole world of ICT. The fact that the largest (and richest) IT companies like Microsoft, Google, Oracle and Cisco are now standing behind this concept represents a clear sign of the direction in which the world of information technology is moving, at least in the next few years. CC concept, according to NIST, has five key characteristics:

- On-demand self-service,
- Broad network access,
- Resource pooling,
- Rapid elasticity,
- Measured service [1].

Cloud service delivery models and models of its implementation

Providing CC services is divided into three elementary architectural models and different derivative combinations of the basic models. These three basic classifications are known as SPI (Software, Platform, and Infrastructure) model.

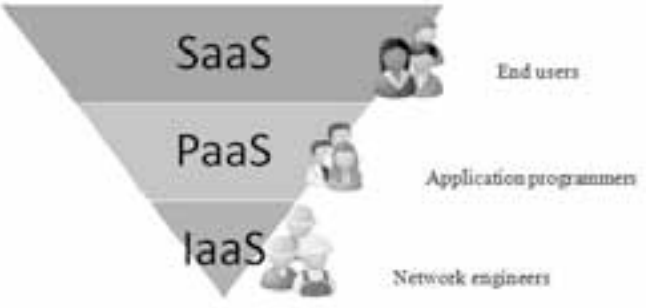


Figure 3: SPI model [9]

Thus, the basic models of providing CC services are:

- SaaS (software as a service) - a technology platform that allows access to applications via the Internet in the form of services that are hired as needed, instead of buying a separate software programs that must be installed on the user (office and /or home) computers;
- PaaS (Platform as a Service) - model is a variation of SaaS structure that, as a service delivers environment development. Allows the user to build his own applications that run on the provider's infrastructure. Applications are delivered to users through the servers' interface accessible via the Internet.
- IaaS (Infrastructure as a Service) - provides the ability to use computer infrastructure (mainly virtual platforms). Users do not buy servers,

software, data storage or network equipment, but they buy these resources as an external service.

Regardless of the type of service delivery models (SaaS, PaaS, or IaaS), there are four basic models of implementing Cloud Computing services, including:

- Public Cloud - platform available and open to the public, regardless of whether they are individuals or organizations;
- Private Cloud - CC infrastructure accessible to only one organization. It can be managed by the organization itself or someone else who is doing that for the organization (outsourcing);
- Community Cloud - model of implementation that provides the ability for more organizations to share the same CC structure. Infrastructure supports special communities that have common interests, needs and security requirements;
- Hybrid Cloud - Model, which consists of two or more previously, discussed types of the establishment of CC structure which remain unique and independent entities, but with a certain kind of reciprocal link, in order to achieve mobility of data between them.

Advantages and disadvantages of Cloud Computing

Like any other technology, and CC, in spite of many advantages, has some (significant?) disadvantages. The table below gives an overview of them.

Table 1: Advantages and disadvantages of CC

Advantages	Disadvantages
possibility for significant cost reduction	problem of availability (lack of availability)
reducing the need for maintenance software support	safety problem
reduction of IT departments in companies	management problem
Scalability	the possibility of sudden termination of the provider
focusing on the primary business	
availability and independence of the unit	
saving energy and contribution to the conservation of the environment	

Cloud Data Mining - CDM

CDM (Cloud Data Mining) offers tremendous potential for analyzing and extracting the (useful) information in various fields of human activities: finance, banking, medicine, genetics, biology, pharmacy, marketing, etc. The application of this technology should enable that with just a few clicks of the mouse one can reach the desired information about customers, their habits, interests, purchasing power, frequency of purchases of certain items, location and so on. Cloud should enable everyone to use this potential, providing, in the form of service, what was recently reserved only for the big (and rich) companies. Small and medium-sized companies that do not have sufficient funds to invest in (too) expensive systems, now have the opportunity to rent a Cloud service for efficient analysis of all the data in the organization, as well as the data out of it, and which is of interest to the organization.

Cloud provides technology that can "handle" huge amounts of data, which cannot be processed efficiently and at reasonable cost using standard technologies and techniques. Analyzing data which coursesq to social networks, pattern recognition, processing of large-scale images, encryption and description and, of course, data mining is just one of the examples of the tasks that are ideal for implementation in the Cloud.

Data mining in Cloud (CDM) is, from a technical point of view, a very tedious process that requires a special infrastructure based on application of new storage technologies, handling and processing. Big Data/Hadoop is the latest hype in the field of data processing. Based on the algorithms and technologies developed by large Internet companies, there is a

quite widespread ecosystem of solutions for processing and analysis of huge amounts of data.

Big Data and NoSQL bases (storages)

Huge data production, in the last few years as a result of business activities, the activities within the social network etc., implies the need for efficient storage and analysis of this data. Big Data is (relatively) new term for large and complex data sets that cannot be processed and maintained by using traditional tools for managing databases. Big Data involves the use of so-called NoSQL database that proved ideal for storing very large amounts of data in distributed systems. Relational databases are based on strict principles, that means that the stability, reliability and failure resistance is insured. However, in the Cloud, where it is necessary to provide a base that has to be fast, scalable and easily extensible, relational databases deal with the problem. Of course, this does not mean that the relational model is inferior to non-relational models, but the complexity that brings relational model cannot provide the required efficiency and speed in terms of processing very large volumes of data, and the lack of scalability of RDBMS is the major cause of new (and different) mechanisms or ways of managing data - NoSQL (Not Only SQL) database. Large Internet companies such as Google, Twitter, Facebook, and Amazon, which work with very large amounts of data, have created a technology for their storage and processing in the Cloud in order to maintain distributed systems and scalability of database. Such databases, non-relational, of course, do not support the ACID (Atomicity, Consistency, Isolation, and Durability) properties in full; actually they represent pure data warehouses with simple mechanisms of data control and transactions.

NoSQL concept relies on the following grounds:

1. Scalability - ability to automatically respond (giving the required major resources) in accordance with the increase in the application;
2. Replication - data in the case of distributed databases is stored in multiple nodes;
3. Partitioning Data – means data sharing in a way that the different parts of the database are in different nodes. The goal of the partitioning data is to improve performance when reading and writing data.

Possible occasional inconsistencies of the NoSQL base data "compensate" by providing much greater flexibility and ensuring scalability, that in the Cloud environment represents a fundamental requirement. Compromises in terms of ACID properties are necessary in CC environment because they can overcome certain limitations of relational databases and provide better performance in the following areas of application:

- Storage and processing of very large amounts of data,
- Storage and processing semi-structured and unstructured data, all with low latency reading operations and automatic scalability.

There are several key factors that influenced the appearance and development of NoSQL databases, including:

- Continued growth of data production,
- Growing demands for processing semi-structured and non-structured data,
- Avoiding complex and costly object-relational mapping,
- Cloud Computing requests,
- Effective work, efficiently storing large amounts of data and its processing,
- Scalability,
- Indispensable compromise in relation to the ACID properties.

In the last few years NoSQL solutions are developing quickly, so that today there is a significant number of them. Although there is no single definition which defines what is included in the term NoSQL, in practice there are the following classes of NoSQL databases: Key-Value, Document oriented, Graph, Column oriented.

Apache Hadoop

Apache Hadoop, an open source project, is seen as a framework for the development of distributed and scalable applications that work with very large amounts of data (measured in petabytes). It is based on Google's MapReduce algorithm and a special

data management system HDFS (Hadoop Distributed File System), which also derived from Google's File System. Hadoop was developed in Java, so it is about a cross-platform product [3].

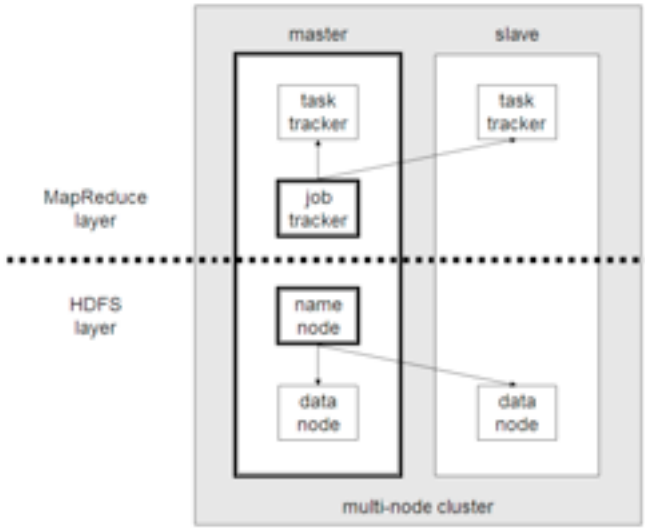


Figure 4: Hadoop cluster [3]

It works in a manner that the tasks, needed to be done, are allocated per cluster computer and then manage those computers in order to perform tasks as quickly and reliably as possible. Hadoop framework supports the ability to perform a huge number of calculations and performs processing of "naked" unstructured data. Hadoop, among others, use Google, Facebook, IBM, Yahoo, Twitter, Amazon, Adobe and, more recently Microsoft as a part of its Azure Cloud platform. For implementation of data warehouse and in-depth analysis and data mining, additional modules Hive and Pig are used.

Apache Hive

Hive is a data warehouse infrastructure built on the top of Hadoop framework and allows analyzing data and generating queries in a way similar to SQL queries in RDBMS (HiveQL). Hive was initially developed specifically for Facebook, but today it is used and developed by others, such as Netflix and Amazon (as a part of Amazon Elastic MapReduce platform) [4].

Pig

Pig is a platform designed for high levels of Hadoop, which is responsible for making MapReduce programs. Pig makes it easy to write MapReduce code introducing a special language - Pig Latin and the environment for the execution of such code. Pig translates the code from higher-level language (Pig Latin) into MapReduce code that is then executed in a cluster computer.

```
INPUT = LOAD '/tmp/my-copy-of-all-pages-on-internet';
-- Extract words from each line and put them into a pig bag
-- datatype, then flatten the bag to get one word on each row
WORDS = foreach INPUT generate flatten(TOKENIZE((chararray)$0)) AS word;
-- filter out any words that are just white spaces
FILTERED_WORDS = FILTER WORDS BY word matches '\\w+';
-- create a group for each word
WORD_GROUPS = GROUP FILTERED_WORDS BY word;
-- count the entries in each group
WORD_COUNT = foreach WORD_GROUPS generate COUNT(FILTERED_WORDS) AS COUNT, GROUP AS word;
-- order the records by count
ORDERED_WORD_COUNT = ORDER WORD_COUNT BY COUNT DESC;
store ORDERED_WORD_COUNT INTO '/tmp/number-of-words-on-internet';
```

Figure 5: Program code that will generate a parallel execution of tasks in a distributed environment (in thousands of computers) Hadoop cluster for counting the words in the huge sets of data [5].

MapReduce

MapReduce is a module that is used for highly distributed processing of large data sets using thousands of computers. Introduced in 2004 by Google, MapReduce can be seen as a framework or system for the execution of a query in the background. Regardless of the amount of data, the system processes the entire data set for each query. Processing is defined by two functions:

- Map - transparently reading "raw data" from a distributed file system, filtering and generating pairs of key - value;
- Reduce - processing of associated and sorted pairs generated Map functions and generating output in the key - value format.

MapReduce is a fundamental concept of processing in Hadoop environment. Subsystem for performing MapReduce programs in Hadoop makes a major node, which is called „job tracker“, and a set of node's workers is called „task tracker“. MapReduce program sent to perform an action is called "job". Hadoop divides the job into a set of tasks. Entrance to the MapReduce program is a set of data stored within the distributed file system. Hadoop shares data in the partitions of the same size which are then allocated to Map functions, or to say it performs the mapping data. Map functions generate k-v pairs that the system merges and sorts by key. When all Map functions are finished with the task, Reduce functions perform tasks on sorted and allied pairs.

Performing of tasks is completely under the control of the main node. Before the performance of specific tasks, "job tracker" must choose to which job task it belongs, that will run. Anticipated job scheduler selects the first job that comes into the job queue. After selecting the job, job tracker assigns tasks that make him free worker. Task tracker periodically reports its state to head node, where the situation represents information on the number of available slots for Map/Reduce tasks. After Map / Reduce tasks are granted, significant optimization is accomplishing. Specifically, the Map tasks are assigned to nodes' workers that contain their own data that handles just the assigned task. This is extremely important because in this way we avoid the (expensive) network communication. The job ends when a node worker that performs the last task is presented to the head node as the one that has completed the assigned task [6].

CDM solutions in application

Considering that mining in the Cloud is something new, still there is no large number of solutions that are fully completed and available to users; however, new products are coming, and soon, a significant number of solutions for data mining which will exploit the potential of Cloud Computing will appear on the market,. Here, only some of the existing solutions will be briefly presented:

- Google BigQuery Service (Dremel),
- Amazon Elastic MapReduce (EMR) and
- MS SQL Server Data Mining for the Cloud.

Google BigQuery

Google's Cloud service BigQuery is one of the "freshest" services of this type. Namely, after eleven months in the previous year have passed, Google announced the limited available test version of this tool, and this service is only a few months ago (from 1.5.2012) publicly available.

Basic features: speed (analyzes billions of records in one second), scalability, simplicity (communication through simple and accessible SQL-like language), the possibility of group work, security (SSL is used to access), various possibilities of use (through the web user interface - BigQuery browser tool, over the command line, BigQuery command-line tool or through REST API). Google has provided client libraries for virtually all major programming platforms, along with scripts and examples of ready applications. Google has offered the possibility of using this Cloud service completely free of charge with a limit of 100 GB of data that can be stored and analyzed on a monthly basis [7].

Here will be shown some screenshots that illustrate the operation and features of the CDM tool. For access, it is necessary to have a Gmail account and the process of logging into the service is fast and intuitive. For testing, I used two existing warehouses that Google has made available to users just for testing purposes (natality and wikipedia).

Figure 6: Basic details of used "tables"

Compose Query

```
SELECT title, num_characters
FROM publicdata:samples.wikipedia
WHERE title CONTAINS "cloud computing" and num_characters>2000
ORDER BY num_characters DESC;
```

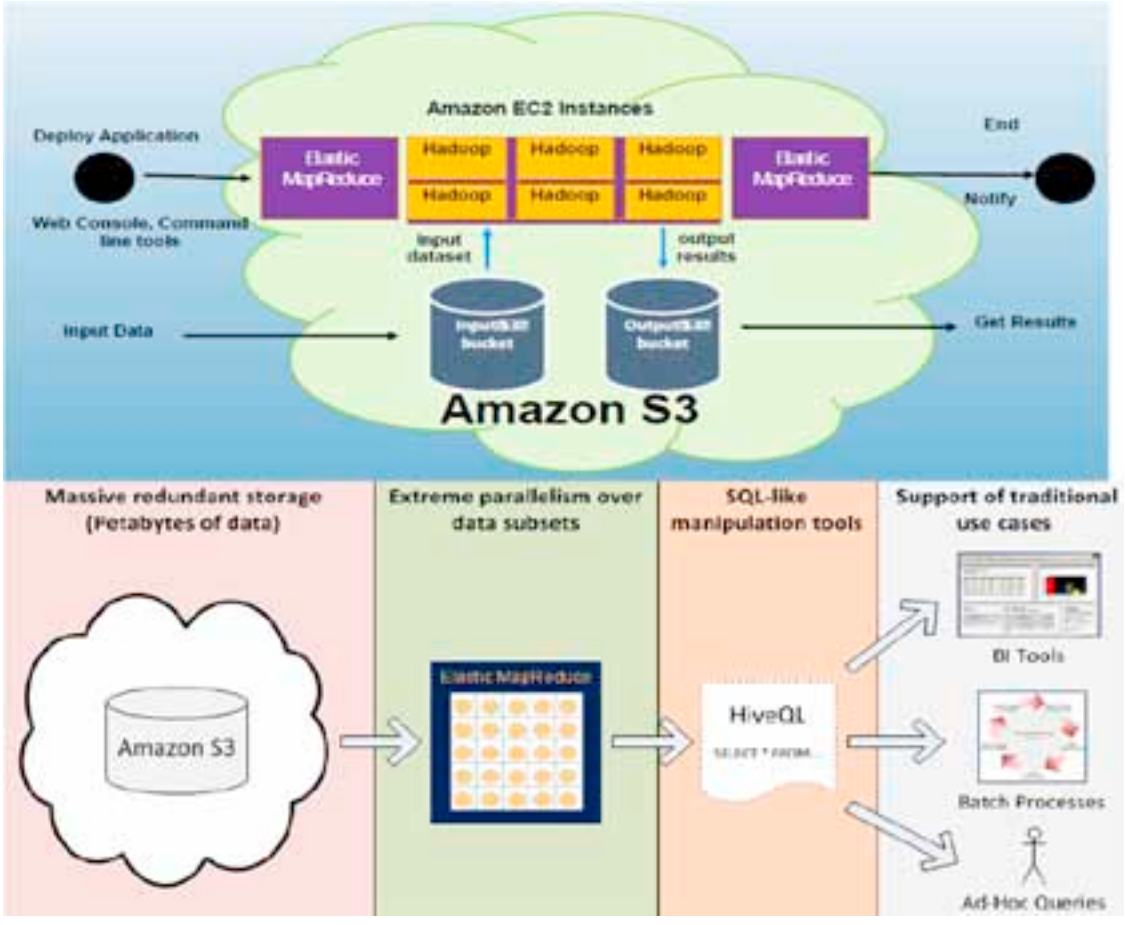
RUN QUERY Query complete (3.0s elapsed, 9.13 GB processed)

Query Results 9:51am, 11 Jul 2012 Download as CSV Save as Table

Row	title	num_characters
1	Nimbus (cloud computing)	2709
2	Nimbus (cloud computing)	2688
3	Nimbus (cloud computing)	2658
4	Nimbus (cloud computing)	2632
5	Nimbus (cloud computing)	2632
6	Nimbus (cloud computing)	2417
7	Nimbus (cloud computing)	2409
8	Vertebra (cloud computing framework)	2216
9	Vertebra (cloud computing framework)	2208
10	Vertebra (cloud computing framework)	2208
11	High performance cloud computing	2159
12	High performance cloud computing	2159
13	High performance cloud computing	2148
14	High performance cloud computing	2127

Figure 7: Query and Query execution results which find how many articles of Wikipedia in the title have the words "cloud computing" and that the article contains more than 2000 characters

Figure 8: Amazon EMR architecture [8]



Amazon Elastic Mapreduce (EMR)

EMR is a platform for developing applications that enable analysts, developers and researchers that in a relatively simple and fast way, and without large expenses, analyze massive amounts of data from different data sets. EMR is based on Hadoop and executes on a scalable infrastructure Amazon EC2 and Amazon Simple Storage Service (Amazon S3).

Amazon EMR allows making DM applications and / or analytical scripts, which are made in the SQL-like languages like HiveQL or Pig. If, however, one wants to create sophisticated applications in Java, C++, Perl and other languages Amazon has provided quality support in the form of examples with complete source code and related tutorials. The principle of operation of the EMR can be described through the following four stages:

1. Creating scripts or applications;
2. Transfer of data and / or applications in the Amazon S3 environment;
3. Running Map / Reduce job through the management console system (AWS Management Console) where one gives the number of EC2 instances and determines the location of data and applications on the S3 platform;
4. Observation of the given activities till obtaining the final result of mining.

Amazon with this service provides, for many, the leading market position in the field of providing CDM services, or as some authors call it "Analytics as a Service." With EMR, Amazon targets companies that operate with huge amounts of data and companies that need elastic and flexible infrastructure for storage and in-depth analysis (mining) data.

What offers EMR?

First of all, EMR is an on-demand service, which can be classified as a category of SaaS (Software as a Service) and PaaS (Platform as a Service) solutions, depending on the implementation by the user. EMR offers flexible resources, programmability, payment according to the standard CC principle only used resources, geographically dislocated EC2 infrastructure and in most cases an increased level of security [10].

Amazon EMR integrates a wide array of tools of other, independent producers like Karmasphere An-

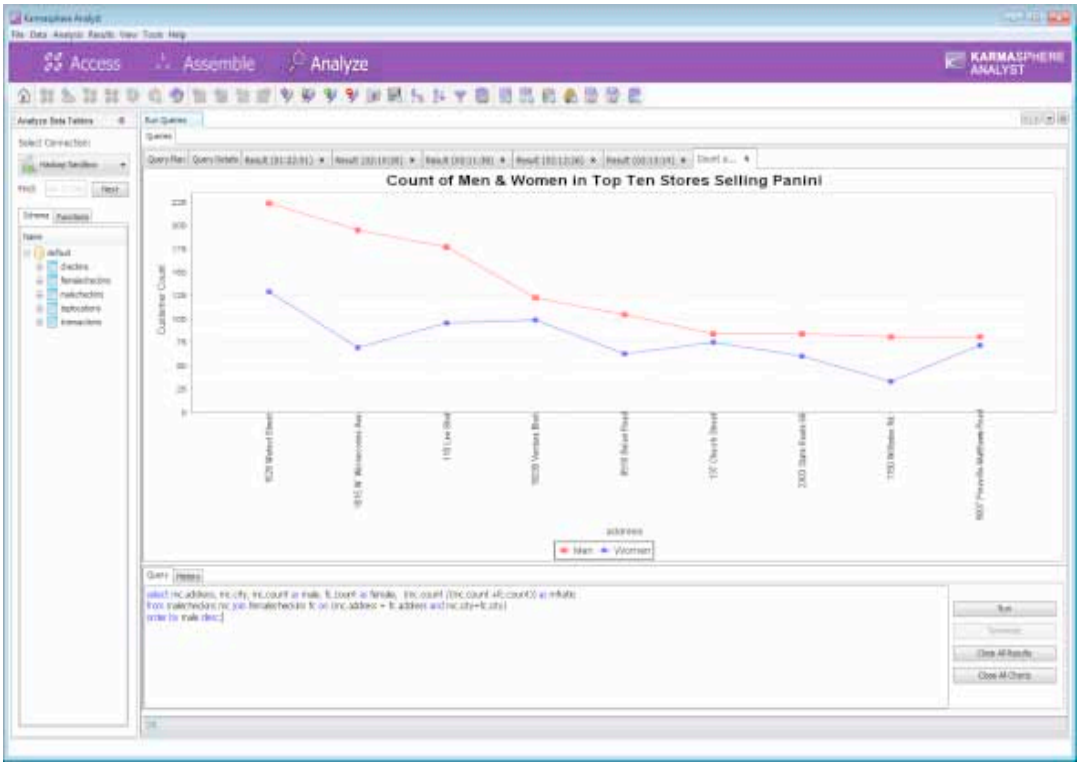


Figure 9: Karmasphere Analyst [13]

Figure 10: Currently (still) are available only three DM tools

alyst, the intuitive integrated environment designed primarily for professional analysts.

Karmasphere Analyst provides support for the processes of "big data" in-depth analysis. It executes through the "4A" (Access, Assemble, Analyze, Act) activities or phases.

Access phase connects to the Hadoop cluster. One can create, configure and test the connection and save the connection settings for later use.

Assemble phase involves organizing structured, semi-structured and unstructured data in different formats and their preparation for the next page. The result of this phase is one or more tables.

Analyze activity allows one to perform iterative analysis, based on HiveQL language. It has a number of tools that help to define the query and its modification. When in the process of analysis it starts to recognize trends and patterns, it enters into new iterations with obtained results, which can additionally format, filter and sort.

Act is the final stage of "big data" analysis. It is conducted by obtaining results and includes certain activities of the results, such as saving in the form of database table, Hive table, in Excel format (.xls) or as a graph [11] [12].

SQL Server Data Mining for the Cloud

SQL Server Data Mining for the Cloud is a Microsoft service for performing data mining in the Cloud, and is developed as a WCF (Windows Communication Foundation) application. WCF is a set of APIs in .NET framework for developing service-oriented applications. It enables users' access to the service via a special tool or application that runs within a web browser and allows users to set up user data on the server / servers, selection of DM tools, configuration of tools and review of results. The connection with the service ends when the client application displays the results. The service is publicly available (with limited functionality, or with a limited number of DM tools) at http.//clouddm.msftlabs.com. For its use no registration is required.

DM tools available to users are:
- Analyze Key influencers,
- Forecasting and
- Prediction Calculator

Service allows the upload of user's data using "Load Data" tool, which should be in .csv format or using existing tables to test - "BikeBuyer Sample" and "Forecasting Sample"

Case Study – Facebook

Intensive data mining on Facebook is impossible without the use of CC solution. With about 500 million users and an average of one billion page views per day, this most popular social network (and Cloud application) daily generates and accumulates huge amounts of data. One of the biggest challenges, practically from the start, was (and still is) solving the problem of efficient storage, processing and analyzing (mining) of this data. To solve this problem, engineers and analysts needed powerful tools for mining and manipulating such large (huge) data sets. None of the servers have the capacity that could satisfy these needs, and the use of relational databases and RDBMS is no longer an option. Facebook, therefore, in order to continue its growth and business, had to develop and implement technology that will allow daily processing and storage of about 15 terabytes of new data; data that is unstructured, in

different formats, in different languages, and from different platforms ... Facebook needed extremely powerful, massive framework with the possibility of parallel processing and with the ability of a reliable and secure storing of huge amounts of data. In addition, it had to ensure an efficient way (model) of mining this data. Such extreme requirements traditional ICT infrastructure cannot fulfill in a satisfactory manner. But CC can! For example, Cloud, for Facebook needs, delivers 8500 CPU cores, and gives the option of using petabytes (250B) of storage space. Such power and capacity provides (for now) the possibility of performing rich in-depth data analysis on a wide range of mining parameters.

Conclusion

We live in a time of information that is the most important and most expensive resource. Huge amounts of data daily produce and in themselves hide potentially useful information. The data that is processed does not originate only from multiple information system of companies, giant amount of it comes from "on-line" environment, with a variety of services that users use for both commercial and private purposes. This data contains significant potential, and out of it invaluable information about, for example, buying preferences, financial situation and clients (users) interests can be drawn.

The task of ICT is to create methods and tools for efficient data processing. Today, that is not an easy task, on the contrary; processing and storage of vast amounts of data that are multiplying daily, represents a significant problem and reveals the limitations of the traditional information and communication technologies and tools. For some time, and even presently, a significant problem represents a general lack of funds. Companies are no longer able to invest great funds in the development of their IT sectors. On the other hand, the need for treatment, demanding deep processing and analysis of data has never been greater.

Where is the solution?

One of the solutions, surely, can offer the integration of in-depth analysis of data (data mining) and Cloud Computing. Huge storage and processing potential of CC, and well-known techniques and methods of data mining, which have "moved to the Cloud," create a powerful platform for analyzing vast amounts of data that is produced daily and in itself it hides much (useful) information, which is the basis for new knowledge and better business decisions, which, in return, is ultimately the main goal. By developing cloud based data mining solutions accessing data mining services every time and everywhere and from various platforms and devices will be made possible. Ultimately, the application of CDM solutions can provide a sort of knowledge discovery eco-system built of a large numbers of decentralized data analysis services.

Also, a significant moment that should be noted is that the creation and giving the service of data mining in the Cloud, today a critical business activity, which, otherwise, requires significant financial and technical resources, becomes accessible to the less affluent, small and medium-sized companies that have not used so far the advantages of the applying this segment of business intelligence.

Authorship statement

Author(s) confirms that the above named article is an original work, did not previously published or is currently under consideration for any other publication.

Conflicts of interest

We declare that we have no conflicts of interest.

References:

[1] Mell, P. and Grance, T. (2011). The NIST Definition of Cloud Computing (Draft). Recommendations of the National Institute of Standards and Technology, NIST.

[2] Pejić Bach, M. (2005). Data mining in the banking industry. Proceedings of Faculty of Economics, University in Zagreb.

[3] http://en.wikipedia.org/wiki/Apache_Hadoop (Accessed: July 2012)

[4] http://en.wikipedia.org/wiki/Apache_Hive (Accessed: July 2012)

[5] http://en.wikipedia.org/wiki/Pig_%28programming_language%29 (Accessed: July 2012)

[6] http://en.wikipedia.org/wiki/MapReduce (Accessed: July 2012)

[7] https://developers.google.com/bigquery/docs/overview (Accessed: July 2012)

[8] http://practicalanalytics.wordpress.com/2011/08/13/analytics-as-a-service-understanding-how-amazon-com-is-changing-the-rules/ (Accessed: July 2012)

[9] http://samisa-abeysinghe.blogspot.com/2011/07/cloud-computing-explained.html (Accessed: July 2012)

[10] http://aws.amazon.com/elasticmapreduce/ (Accessed: July 2012)

[11] http://aws.amazon.com/elasticmapreduce/karmasphere/ (Accessed: July 2012)

[12] https://karmasphere.com/kscoldnogood/Karmasphere-Analyst-User-Guide-v1.8/karmasphere-analyst-user-guide-v18.html (Accessed: July 2012)

[13] https://karmasphere.com/Getting_Results_Tutorial/Default.htm#07%20Lesson%20Seven%20-%20Visualizing%20Your%20Results/Visualizing%20Your%20Results. Htm (Accessed: July 2012)

Submitted: October 25, 2012.
Accepted: November 30, 2012.