Ramiz M. Aliguliyev[1], Yadigar N. Imamverdiyev[2],
Fargana J.Abdullayeva[3]
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a.ramiz@science.az, [2]yadigar@lan.ab.az, [3]farqana@iit.ab.az

DOI: 10.25045/jpit.v07.i1.02

THE INVESTIGATION OF OPPORTUNITIES OF BIG DATA ANALYTICS AS ANALYTICS-AS-A-SERVICE IN CLOUD COMPUTING FOR OIL AND GAS INDUSTRY

The increase in volume of data collected in oil and gas industry has engendered the emergence of serious problems in this sector. The problems caused by large-volume data in oil-gas industry and the current state of application of big data analytics in this sector are analyzed in the article. The big data analytics platforms developed for oil and gas industry by large organizations and the practice of large world oil and gas companies in big data analytics are studied. The analysis of large-volume data for big data analytics and the purposes of usage of cloud computing platform are explored. The suggestions and recommendations are presented for the realization of big data analytics as Analytics-as-a-Service.

Keywords: *Big data analytics, OLAP, cloud computing, Analytics-as-a-Service, oil and gas exploration and extraction, Hadoop, MapReduce, data science.*

Introduction

The large implementation of sensor equipment has led to collection of big data set in fields such as bioinformatics, social networks, and oil and gas industry in digital era. Fast-track analysis of this type of mass data in lesser time is a major feature of competitiveness of modern business environment. However, the transformation of collected data into information cell has posed big problems for analytics [1].

In modern society, one of the fields exposed to big data problems is the government sector [2]. Serious measures are being carried out toward the investigation of complex problems caused by big data in government sector in developed countries. For this purpose, the Obama (the US president) administration has initiated an incentive regarding the study and development of big data [3].

The data obtained by the application of information and communication technologies in oil and gas sector is considered as *oil* of the economy of 21st century. In [4], it is mentioned that information technologies enable the production of more barrels of oil than any equipment (assets) in comparison with previous periods. The upgrade of equipment, automatization of processes and organization of several partnerships have accelerated the increase of data volume in oil and gas sector. Some experts reckon that the data volume grows five times each year [5]. In 2013 report of International Data Corporation, the volume of indicators for oil and gas sector was evaluated as 2,7 zettabytes. The data for oil industry is collected from sensors, space and global positioning system coordinates, meteorological services, seismic data and other measurement equipment [6]. Other sources of data collection are social media, electronic mail, text, images and multimedia. Here, the data is present in a structured or semi-structured form. Hence, storing data in a traditional data warehouse, regular request and analysis of these data is very complicated or expensive. Big data analytics, based on innovative analytics, has acquired a leading position in generating new solutions for the processing of this sort of data.

Big data analytics is considered as a technology, which provides the collection of valuable knowledge enabling the generation of competitive predominance, emergence of new ideas and high profitability [7].

The application of big data analytics in oil sector provides opportunities for the advanced decision-making in oil exploration and extraction, the acceleration of oil discovery and cost reductions in mentioned process.

Inability of all analytical systems to maintain required accessibility and scalability; lack of

containment of different user roles such as analytics, field experts, working process executors and request executors; in addition, the setup of analytical program application in organization's infrastructure; realization of procurement, configuration and administration of those by user as such; the inability of these systems to satisfy dynamically changing user requests are the factors necessitating the joint usage of big data analytics and cloud computing technologies. The use of analytics services (Analytics-as-a-Service, AaaS), presented by clouds for development of systems with listed features and provision of users with dynamic analytics tools, is considered as an invaluable instrument. AaaS is not a unique analytical system or program, but rather a platform introducing services. AaaS platform includes several analytical programs and it is usually interpreted as an analytical administration system based on service level contract. The main purpose here is the positioning of analytical systems in cloud in order to serve to mass of users acting in different roles by using the features such as the organization of clouds access from any location, data distribution and scalability.

Different approaches, which are devoted to the representation of big data analytics to users as AaaS service, have been put forward [8,9,10]. However, the investigation of realization opportunities of Big Data analytics as AaaS for oil and gas industry is not considered in suggested approaches.

In presented article, data analytics, advanced analytics, big data analytics, data science, and the essence of scientific notions of data are investigated. Wide analysis of suggested approaches is carried out toward the formation of big data analytics. Several suggestions and recommendations are introduced for realization of big data analytics as cloud service in oil and gas industry.

Notions in Big Data Analytics

Prior to defining big data analytics, the following notions must be clarified:

Analytics – covers the methods of analysis and extraction in order to support decision-making. These methods allow to identify from data the image of a particular importance and occurring changes, and to specify the reaction of people to these changes [11].

Advanced analytics – is a group of analytical methods used to predict future outcomes [12]. In this regard, advanced analytics is valued as an irreplaceable instrument in solution of issues such as the detection of best providers, determination of sale season and etc. Usually, predictive analytics, data mining, statistical analysis, complex SQL, data visualization, artificial intelligence, text analytics and etc. tools are included in advanced analytics group. In literature sources, detective analytics term is used instead of advanced analytics in some cases. All these techniques exist for a long period; some of them emerged in 1990's. These techniques are widely implemented in big data analysis, as they are well adapted to the analysis of large data set measured in multiterabytes.

Big data analytics – is the application of advanced analytics tools to big data set [13]. As seen from definition, big data analytics covers two objects: big data and analytics. On the other hand, big data analytics – is a new age technology developed for high-quality knowledge extraction from different type of data (structured and unstructured) and allows for the prompt realization of data mining, detection and analysis [14].

The purpose of big data analytics is to create suitable conditions for important decision-making by individuals. This is carried out by engaging data scientists who are capable to analyze large-scale transaction data. This, in turn, necessitates the emergence of *data scientists* and *data science*.

Data science – is constituted of different elements and constructed upon techniques and theories pertaining to several fields [15]. These techniques are constituted of mathematics, mathematical statistics, data engineering, image identification, advanced computing, visualization, modelling of uncertainty, data warehouse, and high performance computing (figure 1).

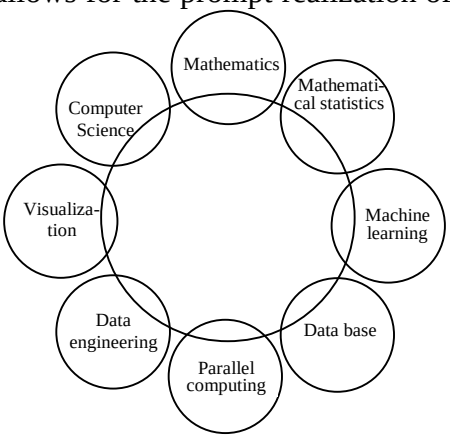

Figure 1. Techniques required for Data science

The purpose of data science is to obtain opinions in the form that can easily be understood by non-experts by using all accessible and appropriate data. A person applying the data science is called *a data scientist*. Data scientist must possess significant skills in analytics and must be able to solve data-related complex issues in cooperation with experts of other related fields.

Cloud technologies – is a type of parallel and distributed system, constituted of mutually related and virtualized computer masses, these computers are dynamically developed and introduced to clients on the basis of service level contracts or several unique computation resources constructed via the consent between service provider and the client [16].

Analysis of Studies on Big Data Analytics

Data analytics, existing until today, has already sustained its potential in the sphere of decision-making in financial, administrative and scientific fields by conducting complex computations and knowledge discovery for scientific inventions. Industrial and scientific society develops big data analytics by applying distributed file system (Hadoop), MapReduce programming model and other different technologies and by obtaining parallelization in algorithms for the purpose of encountering problems caused by big data.

At present, one of the issues considered by scientific society is to achieve parallelization feature of clustering. Parallel clustering is deemed most effective method in data clustering. Parallel clustering is carried out, in parallel, by conducting non-dependent computation procedures in several computation processors. Various parallelization strategies are investigated in [17], PIC (Power Iteration Clustering) approach is introduced in algorithm parallelization for big data clustering. DBCURE-MR has proposed density-based clustering algorithm [18]. This algorithm allows detecting clusters of different density, and it is based on MapReduce model in order to have the parallelization feature. Density-based traditional algorithms find clusters separately; in contrast, DBCURE-MR algorithm finds several clusters in parallel.

Traditionally, OLAP (On-Line Analytical Processing) technology cannot overcome the problems caused by big data. Thus, OLAP system named Haolap is developed for Big Data in [19]. Haolap (Hadoop based OLAP) is a multi-dimensional OLAP technology system and based on Hadoop distributed file system and MapReduce algorithm. This system makes use of a special multi-dimensional model for memorization of objects and computations. On account of MapReduce algorithm, the system can carry out OLAP operations on large-scale data.

In many sectors the graph structure is constructed in such dimensions that their processing requires specific techniques, mainly techniques with feature of parallelization. In order to test the process of parallel processing of large graphs, empirical comparison of existing techniques such as MapReduce, MapReduce2 (MR2) and Bulk Synchronous Parallel (BSP), which are available in Hadoop environment, has been conducted, [20]. As a result, BSP model was selected as the best in eliminating graph problems in comparison with MapReduce model. MapReduce encounters several problems during the processing of data formed as network. In order to eliminate these problems, Google Corporation has developed the system named Pregel. Pregel system applies BSP model for processing of graphs in clouds [21].

[22] explores the rules of parallel computation of Principal Component Analysis (PCA) model for the purpose of acceleration of the issue of data set summarization. PCA is a widely known technique allowing for the reduction of the volume of large-volume data and extraction of features from those. In this research work, it is attempted to transform the existing technique pertaining to the feature of sequence into parallel algorithm, and it is able to calculate PCA via one transition in big data set by using summarization metrics. The issue of integration of algorithm with DBMS (database management systems) is also reviewed in this article.

Ophidia system is one of the systems enabling the parallel processing of data analysis [23]. Ophidia acts as a big data analytics in electronic science sphere. The system contemplates a component in parallel analyzing the data for multi-dimensional scientific data management

pertained to spheres such as weather change, astrophysics, bioinformatics and etc. Internal storage and hierarchic arrangement technique of data was considered.

A model reflecting HACE (Heterogeneous, Autonomous, Complex, Evolving) theorem, which explains salient characteristics of big data and its process of data processing in terms of data mining, is presented [24]. According to HACE theorem, big data is enormous due to the heterogeneity of its source, autonomous while it is based on distributed and centralized management mechanism, it is complex and developing due to data links. These features cause serious impediments in acquiring necessary knowledge from big data. Hence, to eliminate these hindrances, a model, introduced by the authors, reflects the process of data processing in terms of data mining, which takes HACE characteristics into consideration.

In [25], it is claimed that the application of visual analytics approach provides more efficient analysis of big data collected in climate sphere in the process of climate study. In this regard, the application issue of EDEN visual analysis system to data set simulating location system is reviewed here.

One of the technologies considered in relation to the emergence of big data is a technology of text mining. The analysis of different aspects of text analytics and text mining technology such as legal, business analytics and security issues are analyzed in [26].

The analysis of existing problems, and wide analysis of techniques and technologies of big data are carried out in [27]. The current state of application of parallel and distributed systems on big data analytics are explored in [28].

Lately, the increase in interdependence of grid and cloud technologies and big data analytics is clearly observed. Hence, the works included in 2014 special edition of "Future Generation Computer Systems" journal encompassed the issues of big data architecture, big data processing systems, big data application and modelling, MapReduce optimization, resources distribution, resources monitoring, energy-saving utilization of resources. Alongside covering new ideas and current techniques related to this field, these articles also commented upon the future research.

One of the events that emphasized the importance of cloud technologies in Big Data analysis was "Big data analytics for oil and gas" conference held on 17-18 October, 2013, which brought together leading data scientists of major oil companies such as British Petroleum, Shell, Anadarko and Chevron. In the conference, data scientists have submitted reports regarding the application rules of Hadoop ecosystem in seismological processes, the importance of cloud technologies in larger analysis of big data. Moreover, "Big Data Energy Innovation" Summit was held on 19-20 May, 2015. Main topics of the Summit covered big data contributions to energy sector.

Some studies have been conducted in this sphere. A conceptual model has been proposed [8] for creation of service-oriented decision support systems (SODSS) supporting the process of decision-making. Here, SODSS architecture is constructed of 3 layers: IaaS, SaaS and Business Process (BP). According to architecture, operating systems, information storage, online analytical processing and computer tools can be proposed to users as a service. Positioning ability of data in any location and availability of processing instruments during 24 hours are considered in Data as a Service block of architecture. In analytics service block, cloud technologies are used in solution of analytical problems. This service is sometimes called as Agile Analytics. Scalability and low costs are main advantages of agile analytics.

An event titled "Cloud Computing and Scientific Applications" was held in Ottawa, Canada in 2012 for the purpose of sharing the practice in the field of execution, modelling and monitoring of scientific research systems in clouds. Experts, who attempted to expand the implications of scientific research systems by using clouds, introduced the reports. The major topics of the conference were "Big data in clouds", "Scientific computations in clouds", "Social computations in clouds" and etc. While applying clouds in big data analysis, the scientists encounter several complex problems. A detailed literature related to these problems can be reviewed in [9].

The significant role of cloud technologies in efficiency growth of big data analytics has been

widely discussed in [10]. The importance of emergence of big data in cloud technologies environment, cloud technologies in big data, storage systems of big data and the links with Hadoop technology are discussed here. Scientific-research issues related to scalability, accessibility, data integrity, data transmission, data quality, data heterogeneity, security, legal and regulation issues of data, and data management, while using big data in clouds, is thoroughly investigated.

Big Data Analysis Methods

Big data analysis encompasses several directions such as mathematical statistics, data mining, machine learning, social network analysis, signal processing, image identification, optimization and visualization techniques (figure 2) [27]:

- *Optimization techniques.* It is applied in solution of problems requiring quantitative evaluation.
- *Mathematical statistics.* It is a science studying data collection, organization and interpretation. Statistical techniques are used for determination of correlation relations and causality links between various objects. It is also possible to find numerical comments via statistical techniques. However, standard statistical tools are not deemed sufficient in big data management issues. Hence, many researchers either propose approaches toward widening the classical techniques for the processing of this sort of data or develop completely new techniques.
- *Data mining*. It is the compilation of techniques that allows to derive necessary information (image) from data. Cluster analysis, classification, regression and the study techniques of associative rules are attributed to these sort of methods. Data mining utilizes the methods of machine learning and mathematical statistics. The analysis of big data becomes a complex issue while applying traditional data mining algorithms. Hence, the existing methods are attempted to be enhanced in order to process big data in this sphere.

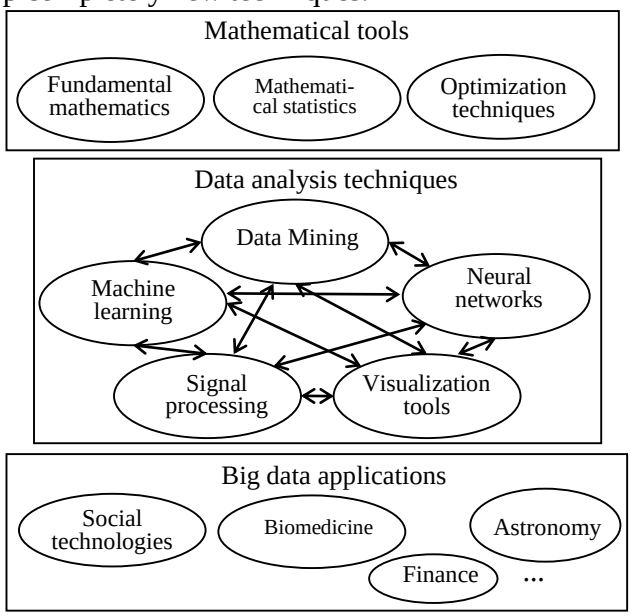

Figure 2. Big data analysis techniques

- *Machine learning.* It is an essential part of artificial intelligence, the purpose of which is to develop algorithms serving to detect the behaviors based on empirical data. The salient characteristics of machine learning are the knowledge discovery and automatized intellectual decision-making. In order to cope with big data, it is attempted to enhance the algorithms of machine learning of both types: supervised learning and unsupervised learning.
- *Visualization tools.* These are the tools used for generation of tables, images, diagrams and other intuitive images for the purpose of data comprehension. While volume, velocity and variety problems (these are also referred as 3V) pertain to big data [30,31], visualization of those is not as facile as the visualization of small-scale traditional relational data set. Several works have been carried out toward the enhancement of visualization methods. However, those methods are not deemed sufficient for big data processing. Hence, many researchers firstly attempt to reduce the size of big data significantly in order to visualize them by using feature extraction and geometrical modelling.
- *Social network analysis*. It is one of the major techniques of modern sociology and illustrates social relations in terms of network theory. Social network is constituted of hosts and links between them.

- *Neural networks*. Neural networks comprise online-trained algorithms without necessitating the positioning of data in storage. Neural networks are suitable for processing both of flow data and big data stored in data warehouse. Neural network algorithms perform parallel computations very accurately, such that, this capability is not considered in machine learning technology.
- *Signal processing*. The use of electronic sensor devices in any location in modern world has turned the theory of signal processing into important sphere of big data. Signal processing is a technology encompassing the theory of processing or transmission of information existing in physical, symbolic and abstract forms, program applications and algorithms. It utilizes mathematical, statistical, computational, heuristic and linguistic methods for signal illustration, modelling, analysis, synthesis, detection, extraction and conduction of forensic investigation. While various sources generate data in continuous real time regime, data analysis must be carried out "instantly" without requesting past records. For this reason, enhanced research is carried out toward the adaptation of existing signal processing analytics such as principal component analysis, dictionary learning, compressive sampling and etc. to current large-scale data regime.

Above listed methods usually serve to a particular field to which data pertains. If data analytics experts related to the field of any subject-matter do not possess the required knowledge about that field, serious problems are encountered in data analysis process. Moreover, the selection of appropriate apparatus and program maintenance resources for data analysis can pose difficulties also for scientists acting as experts.

Cloud Computing and Big Data Convergence

At present, big data and cloud computing are information technologies incentives primarily valued by world organizations [7]. Big data technology is usually linked to cloud computing technologies. It is because, this technology requires the presence of platforms such as MapReduce for maintenance of integration and processing of data acquired from Hadoop and various sources to provide the storage of big data set in distributed clusters.

MapReduce is a robust system to denials with scalability feature, which allows for parallel processing of large-volume data in low-level computers [32]. MapReduce was first developed by Google Corporation in 2004. It has gained large attention of both scientific and industrial organizations due to its perspective opportunities, and it has been largely applied. MapReduce system is deemed better scalable and more efficient, as in the context of cloud technologies, the introduction of infrastructure resources is possible upon request. Simplicity, scalability, robustness to denials are the main three characteristics of MapReduce approach. Therefore, the utilization of MapReduce services such as Amazon EMR is feasible and beneficial for organizations and enterprises for big data processing.

Amazon EMR. Amazon Elastic MapReduce (Amazon EMR) – is a web service providing fast big data processing with low costs. Amazon EMR simplifies the process of big data processing by incorporating the Hadoop system. The function of Hadoop system is to distribute large-volume data to Amazon EC2 elements. Amazon EMR is capable to manage securely and reliably the spheres such as log analysis, web indexing, the organization of data warehouse, machine learning, financial analysis, simulation of scientific research and bioinformatics.

Hadoop. It is a platform of open code software providing the processing of big data set in distributed computation environment. The function of this processing tool is to distribute the data among balanced computer clusters, which perform in parallel.

Hadoop Distributed File System (HDFS). It is a component of file system of Hadoop platform. The appointment of HDFS is to keep metadata and program application data of file system in distributed form [33].

The reasons for use of cloud infrastructures in big data processing are as following [7]:

- *The sufficiently large volume of investments spent on big data analysis*. This requires the presence of effective and low-cost infrastructure such as clouds.

- *The positioning of big data in both internal and external infrastructures*. Usually, organizations locate their more vulnerable information in internal infrastructures and less important information in third party infrastructures. Cloud technologies can provide the analysis of data located in both internal and external infrastructures.
- *The necessity of the presence of analysis services for knowledge extraction from big data*. The use of Analytics as a Service is considered as a sole instrument providing the efficient management of IT budget.

Analytics as a Service

Analytics as a service in cloud is required when an organization lacks the skills in analytics field. AaaS is an interesting concept, which is useful in application for both small and large organizations.

In order to carry out the solutions of big data analytics in cloud, Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS) models can be applied [34].

1. *Data analytics software as a service*. It introduces analytics software to a final user as an Internet service.
2. *Data analytics platform as a service*. It provides a platform for users in order to create personal analytics programs for them. BigQuery service introduced by Google Corporation can be shown as an example for data analytics platform as a service,. BigQuery enables the use of enormous computer and storage opportunities of Google Corporation and real time decision-making by users for the analysis of large-volume data. BigML is another large-volume data platform as a service. BigML presents a cloud-oriented machine learning service. Large-volume data platforms usually consist of several modules, for instance, problem specification analysis block, data storage and management block, detection, visualization and etc. [35].
3. *Data analytics infrastructure as a service*. It provides virtualized resources to users. These resources, in turn, are used by program manufacturers as a computer infrastructure in order to execute their analytics programs.

Any AaaS must be constructed on following principles [11]:

- The sequence of steps for problem execution must be organized hierarchically. The steps such as hierarchic data structure, the specification of analytic models, configuration of analytic program must be included;
- Successive, parallel, iterative and selective flows must be supported;
- Data and management flows must be supported;
- The obstacles necessary for identity verification must be described in detail;
- Based on service level contract, the service quality for AaaS must be introduced;
- Scalable cloud resources must be used transparently for efficient execution of problem flow;
- Recommendation systems must be supported for the assistance to a specific field in construction of analytical and visualization models, and execution of problem flow;
- The role of user and his/her occupation aspects must be taken into consideration while providing services;
- The flow of succinct problem flow must be illustrated by large graph;
- The minimal data shift must be allowed;
- The privacy and security of data must be maintained.

The users of AaaS can be the following individuals:

- Scientific analytics. Individuals possessing the knowledge related to analytic programs and methodologies. These individual may or may not possess domain knowledge, for example, experts on data mining technology.
- Domain experts. These individuals possess domain knowledge, they understand the meaning of various data values, usually instruct analytics and assist in decision-making

from obtained outcomes. These individuals, for example, can be doctors, and meteorologists.

- Practitioners. They practically conceive the value of data and execute beforehand assigned problem flow as a final user. Nurses can be shown as examples for this type of individuals.
- Administrators. They obtain business values by carrying out alternate markup and perform administrative measures based on this. Sales or working staff administrators can be shown as an example.
- Managers. They carry out strategic planning based on company goals and request the specific information from different aspects, accordingly.

AaaS possesses the following capabilities [7]:

- Collection and extraction of structured or unstructured data from reliable sources.
- Data management under policy instructions and the organization of control.
- Realization of data integration, analysis, conversion and visualization for the purpose of transmission of necessary information in required time point to required location.

IBM was the first corporation that attempted to provide AaaS services to users. AaaS platform belonging to IBM is illustrated in figure 3. The primary aim of this platform is to reduce the financial burden of organizations devoted to private analytics. This platform enables the users to upload their (structured or unstructured) data to third party infrastructure for analysis purposes.

Another leading organization performing in Aaas services sector is SAS (Statistical Analysis System) organization. The SAS organization introduces AaaS to users based on predictive analytics.

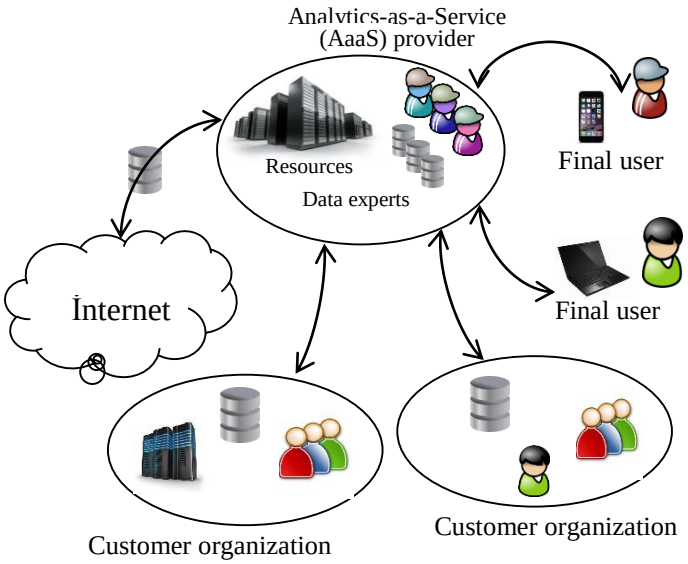

Figure 3. Analytics as a Service

The analysis of current state of big data analytics application in oil and gas industry

The following problems have been specified for oil and gas exploration and extraction [36]:
Drilling:

- The necessity of automatization in real-time decision-making;
- The necessity of knowledge extraction from real-time data archive after the completion of drilling works;

Oil production:

- The necessity to detect the problems in oil wells before they have reached a serious condition;
- The necessity of rapid optimization for quality maximization of output results;

Data management:

- The necessitation of big data techniques as data volume reaches astronomical levels;
- The impediment of real-time analysis by data volume;

It is possible to increase the effectiveness of oil and gas exploration and production by application of big data analytics to mentioned problematic spheres.

Big data technology can be applied to all phases (exploration, production, processing, retail sales) of oil and gas extraction [37]:

- Enhancing searches. The proposal of new ideas to working group by maintaining the integration of enterprise data with real-time production data.
- The site evaluation and creation of new perspectives. The creation of competitive exploration by using geographical information, newscast and other syndicated information sources.
- Improving engineering studies. Early and low-risk determination of commercial perspectives by using complex geological models and detailed technical investigation in wells.
- Optimizing subsurface understanding.

Knowledge extraction from data acquired in oil and gas sector with big data analytics provides the following opportunities for organizations of this sector [36]:

- Maintenance of competitiveness by planning, exploration, production and development of oil wells;
- Maximization of oil production by management and forecasting;
- Reduction of time allocated to first extraction of oil and gas, reduction of costs on operations and productivity increase during life period of oil wells;
- The timely provision of direct and automatized accessibility of credible information for working group.

Investing on the Big Data is valued as big Return on Investment (ROI).

Hadoop-based solution based on big data technology is proposed in [38] for data management of oil - gas extraction and production. Solution architecture reflects the process of data processing acquired in oil wells by big data analytics. Architecture is constructed in five phases consisting of blocks such as data collection, data storage, data aggregation, data analytics and data visualization. Firstly, data related to oil extraction from various wells is collected. Collected big data is, then, stored in Hadoop distributed file system, and collected data is aggregated in different bases (Hive/Hbase, NoSQL) and started to be analyzed at the next phase. At the stage of data analytics, analysis is carried out for the purpose of recognition of drilling images, and obtaining lithology content, based on different parameters of log files of oil wells. At the stage of data visualization, output results at analytics phase is transmitted to business-analysis system, evaluation table is formed here and decision-making is carried out after table-based data visualization.

Big data technology can improve the quality of operations in following sectors of oil industry [39]:

Exploration and exploitation:

- *Path Identification*. The use of improved analytics (for instance, image identification) in big data set collected during seismic exploration for the purpose of recognition of previously overlooked seismic pattern signatures.
- *Enhancing exploration efforts*. Introduction of new knowledge to operative groups by integrating real-time collected oil extraction data and data collected in an enterprise.
- *Creating new perspectives*. The conduction of competitive exploration by using analytics applied to geographical data, and oil-gas information.

Drilling and completion:

- *Forecasting successful drilling.* Based on limited data, real-time application to large-volume drilling data for anomaly detection and forecasting the probability of success of drilling alongside the monitoring and notification.

***Extraction and operations*:**

- *Productivity forecast.* The forecast of productivity in ten thousands of oil wells. Should the forecast outcomes in old wells not exceed the specified productivity threshold, urgent restoration of those must be notified.
- *Enhanced oil production*. The application of various methods is considered in order to increase the volume of crude oil extractable from oil wells. In this regard, the application of analytics, used in the analysis of different type of big data, to seismic, drilling and production

data is also deemed possible in order to increase crude oil volume.

- *Preventive technical service*. It is a technique for forecasting the time necessary for conduction of technical service in a device. Pressure, volume and temperature are concomitantly analyzed, compared with the date of previous breakdown; potential failures are forecasted by applying improved analytics.

Big data application in oil and gas industry has following advantages [40]:

- *Improving operations*. Obtaining collective knowledge from unique management platform comprising structured, unstructured and real-time data. Reduction of costs of regulatory acts through non-productive time (NPT) and the presence of real-time risk management mechanism.
- *Achieving high rate in oil extraction*. The acceleration of oil production by minimizing data bottlenecks.
- *Upgrading assets*. Prolongation of exploitation period and forecast of exploitation conditions of a device.
- *Creating a unified ontology for oil and gas.*

Big data Analytics platforms for oil and gas industry

One of the programs carrying out big data analysis in oil and gas sector is InfoSphere BigInsights program developed by the IBM [41]. InfoSphere BigInsights is a Hadoop-based platform and carries out the collection, processing, analysis and management of large-volume and different sort of data related to oil and gas sector. BigInsights program comprises InfoSphere Streams, InfoSphere Data Explorer, IBM PureData System for Analytics and IBM PureData System for Operational Analytics programs. Those cannot perform as autonomous programs separately and are executed under InfoSphere BigInsights program management. A specific cloud in analytics sphere of an enterprise is called Blue Insight. Blue Insight is capable of gathering data from 100 various bases and conducts analytical operations on data volume exceeding petabyte. "Blue Insight" infrastructure of the IBM is introduced to foreign customers as "IBM Smart Analytics Cloud".

Another organization engaged in big data analytics solution development is Hitachi enterprise. The enterprise has established Global Center for Innovative Analytics (Hitachi Global Center for Innovative Analytics, HGC-IA) in 2013 in order to introduce innovative solutions to customers, based upon big data facilities at international level. The structure of the Center is comprised of a set of software serving to development of big data analysis solutions, data scientists, architects and advisors. The scientific-research department called Global Big Data Innovation Lab (GBDIL) is engaged in coordination of big data-related research activities worldwide.

GBDIL has proposed a general architecture, which can act as an instruction for other enterprises in creation of big data solutions. Architecture consists of components carrying out big data storage and management, analysis and visualization. The solutions, created on the basis of proposed architecture, can be executed both on unique host and big cluster. Alongside the architecture approach, HGC-IA and GBDIL are also involved in creation of customer solutions for communication, oil and gas, healthcare and other industrial sectors. The schematic representation of big data analytics of Hitachi organization for oil and gas industry is introduced in [42]. In order to enhance the interest of Hitachi toward oil and gas technologies sector, "US Big Data Lab" has initiated collaboration with Energy and Environmental Research Center (EERC) functioning under the University of North Dakota. EERC is an enterprise engaged in development of energy and environmental technologies. Hitachi and EERC have directed their research efforts toward the solution of complex problems encountered by operators during formation of Bakken Field. "Bakken Production Optimization" program of the enterprise serves for two main assignments:

- Optimization of oil well productivity by accurate analysis of geology of Bakken Field;
- Reducing surface impacts of drilling by increasing the efficiency of drilling works and using water, gas and other tools.

Program solution carries out the above mentioned goals by obtaining knowledge from big data.

One of the systems serving to exponentially increasing data management in oil and gas sector is Lustre system developed by joint efforts of Hitachi (Hitachi data systems) and Intel enterprises. The term Lustre was generated by the synthesis of names "Linux" and "cluster". Lustre is a high-productivity file system assigned for computer clusters. Lustre can be executed in ten thousands of computer hosts and in storage with memory exceeding petabyte, and can process hundreds of gigabytes of data per second. The system is capable to process all sorts of data collected from various sources [43].

iGATE is planning to develop Oil Well Log Analytics system [40]. This system serves to integrate oil exploration and production data in an integrated platform. It is planned that the system will use the cloud infrastructure of iGATE Corporation. Here, data indexation and storage, data scrubbing, data clustering, migration, standardization and analysis of data (structured, unstructured and real-time) gathered from various sources is carried out on a platform equipped with Hadoop ecosystem and R analytics revolution analytics).

Microsoft Corporation has developed instructions encompassing the application rules of new technologies for big data processing in oil and gas industry [37]. It proposes MURA architecture (Microsoft Upstream Reference Architecture), which is capable of acting as a guidance in generating big data solutions for oil enterprises.

The extreme level of technical obstacles and data security problems in oil and gas industry impedes the process of transmission of big data collected in this sector to general clouds. For this purpose, in [44], necessity of using specific and hybrid cloud solutions is justified in oil sector.

Experience of large oil and gas companies in big data sphere

The larger the seismic data gathered by oil companies, the higher is the capability of a company in positioning natural gas zones.

The study of big data technology in oil and gas industry is at experimental stage thus far [45]. Big data technology is applied by few number of oil companies [46].

In 2010, "Shell" oil company encountered large energy costs against the worldwide distributed 3 and regional 400 data centers. The company found a solution in use of clouds and developed own cloud infrastructure by using Virtual Special Cloud service introduced by Amazon Corporation. The massive feature of geographical data acquired by seismic sensors carrying out oil discovery has overcomplicated the process of decision-making. In order to eliminate this problem, the enterprise applies Hadoop system, introduced by cloud infrastructure of Amazon corporation, for processing of big data collected from thousands of oil wells. Chevron oil company applies Hadoop technology in seismic data processing as well.

Seismic Hadoop project is developed by Claudera company. This project is developed for the purpose of seismic data storage in Hadoop cluster and processing.

PointCross is one of the organizations specialized in generation of big data solutions for oil and gas industry. PointCross is a significant worldwide company in technology and services. It is engaged in generating corporative program solution and application for drilling, seismic and production data.

According to forecasts, majority of data of the world will be processed via Hadoop System in the nearest future.

Big data problems in oil and gas industry

The following big data problems are specified in an oil and gas industry [40]:

- Exponential growth of data volume (structured, unstructured, real-time) in various sources during well exploitation;
- Oil and gas organizations usually bear substantial costs devoted to exploration and production data management, and the processing of non-conforming data flow in different sources.
- The difficulties related to the use of data for flexible and effective response to users' requests.

- Utilization of mix of different software products for data interpretation and decision-making by geologists and geophysicists.
- The large volume of problem-oriented information included in each data cluster.

Suggestions and recommendations

The maintenance of following provisions is necessary during realization of AaaS service in Cloud Computing platform of Big Data analytics for oil and gas industry:

1. The application of platforms, program application and infrastructure resources introduced by Clouds with minimum risk;
2. The determination of the degree of robustness of cloud providers against security risks, while locating oil and gas extraction data in clouds;
3. The organization of efficient data integration located in separate locations;
4. The development of instruments that organize the integration of structured and unstructured data of oil extraction located in virtual clouds and analysis of those in real-time regime.
5. The organization of customer data base and analytical solutions via AaaS environment;
6. The development of analytic tools capable of carrying out the continuous analysis of data constituting a flow in oil and gas sector;
7. Transformation of existing data mining algorithms into cloud-oriented instruments and application of those for knowledge discovery from large-volume data in oil and gas sector;
8. The monitoring of oil and gas extraction in current analytic solutions supported by Clouds, development of a function, which generates continuous analytical decisions from events occurring in real-time regime, in order to trace the traffic activity and to detect occurring events.
9. Organization of coordination between data and information flow;
10. Enhancing management of primary processing practice and assets without acquiring them;
11. To endeavor to minimize the time spent on decision-making;
12. The maintenance of security of data, information and analytic models;
13. Development of effective and efficient management models of decision-making systems;
14. To advise on the changes occurring in business operations and respond instantly.

Conclusion

At present, majority of oil companies are highly dependent upon the data owned and management services carrying out high-quality collection, storage, management and analysis of data. In order to eliminate the problems caused by big data in oil and gas industry, data strategies are being developed in several enterprises.

The methodology of installation of data analysis systems in oil and gas sector is based on development of program applications and lacks in having a service-oriented feature. A user is responsible for the installation, configuration and management of program applications.. Moreover, these systems cannot satisfy dynamically changing needs of a user. Hence, the application of AaaS services, introduced by cloud technologies, can substantially improve the responsiveness of the process of big data processing gathered in oil and gas industry. It is to be mentioned that the application of cloud technologies in big data processing can also enable the optimization of overall data volume of organization.

The presented article has explored the current state of application of big data analytics in cloud computing platform in oil and gas sector and the analysis of several analytics programs and solutions in this sector has been carried out. Furthermore, several recommendations and suggestions were proposed for the maintenance of effective big data analysis collected in oil and gas sector via AaaS service.

References

1. Cuzzocrea A., Song I.Y., Davis K.C. Analytics over large-scale multidimensional data: the big data revolution! / Proc. of the ACM 14th International Workshop on Data Warehousing and OLAP, 2011, pp. 101–104.
2. Chen C.L. Zhang C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data // Information Science, 2014, vol.275, pp.314–327.
3. Obama Administration Unveils "Big data" initiative: Announces $200 million in new R&D investments, 2012, 4 p.
4. http://www.technologyreview.com/news/427876/big-oil-goes-mining-for-big-data/
5. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. Big Data: The next frontier for innovation, competition, and productivity, 2011, 156 p.
6. Brulé M., Tapping the power of Big Data for the oil and gas industry, IBM Software White Paper for Petroleum Industry, 2013, 8 p.
7. Big Data in the Cloud: Converging technologies, Intel IT Center, 2015, 12 p.
8. Demirkan H., Delen D. Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud // Decision Support Systems, 2013, vol.55, pp.412–421.
9. Pandey S., Nepal S., Cloud computing and scientific applications – big data, scalable analytics and beyond // Future Generation Computer Systems, 2013, vol.29, pp.1774–1776.
10. Hashem I.A., Yaqoob I., Anuar N., Mokhtar S., Gani A., Khan S.U. The rise of "big data" on cloud computing: Review and open research issues // Information Systems, 2015, vol.47, pp.98–115.
11. Zulkernine F., Bauer M., Aboulnaga A. Towards cloud-based analytics-as-a-service (CLAaaS) for big data analytics in the cloud / Proc. of the IEEE International Congress on Big Data, 2013, pp.62–69.
12. What is advanced analytics? http://www-01.ibm.com/software/data/infosphere/what-is-advanced-analytics/
13. Russom P. Big data analytics, TDWI research, 2011, 35 p.
14. Feblowitz J. Big data in upstream oil and gas, IDC energy insights, 2013, 45 p.
15. Liebowitz J. Business analytics: an introduction, 2013, 288 p.
16. Buyya R., Yeo C.S., Venugopal S., Broberg J., Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility // Future Generation Computer Systems, 2009, vol.25, no.6, pp.599–616.
17. Yan W., Brahmakshatriya U., Xue Y., Gilder M., Wise B. p-PIC: Parallel power iteration clustering for big data // Journal of Parallel and Distributed Computing, 2012, vol.73, no.3, pp.352–359.
18. Kim Y., Shim K., Kim M., Lee J.S. DBCURE-MR: an efficient density-based clustering algorithm for large data using MapReduce // Information Systems, 2014, vol.42, pp.15–35.
19. Song J., Guo C., Wang Z., Zhang Y., Yu G., Pierson J. HaoLap: Aa Hadoop based OLAP system for big data // Journal of Systems and Software, 2015, vol.102, pp.167–181.
20. Kajdanowicz T., Kazienko P., Indyk W. Parallel processing of large graphs // Future Generation Computer Systems, 2014, vol.32, pp.324–337.
21. Malewicz G., Austern M., Bik A., Dehnert J., Horn I., Leiser N., Czajkowski G. Pregel: A system for large-scale graph processing / Proc. of the International Conference on Management of Data, 2010, pp.135–146.
22. Ordonez C., Mohanam N., Garcia A.C. PCA for large data sets with parallel data summarization // Distributed and Parallel Databases, 2014, vol.32, no.3, pp.377–403.
23. Fiore S., D'Anca A., Palazzo C., Foster I., Williams D.N., Aloisio G. Ophidia: Toward big data analytics for eScience / Proc. of the International Conference on Computational Science, 2013, pp.5–7.

24. Wu X., Zhu X., Wu G., Ding W., Data mining with big data // IEEE Transactions on Knowledge and Data Engineering, 2014, vol.26, no.1, pp.97–107.
25. Steed C.A., Ricciuto D.M., Shipman G., Smith B., Thornton P.E., Wang D., Williams D.N. Big data visual analytics for earth system simulation analysis // Computers & Geosciences, 2013, vol.61, pp.71–82.
26. Truyens M., Eecke P.V. Legal aspects of text mining // Computer Law & Security Review, 2014, vol.30, no.2, pp.153–170.
27. Chen C.L., Zhang C.Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data // Information Science, 2014, vol.275, pp.314–327.
28. Kambatla K, Kollias G, Kumar V., Grama A. Trends in big data analytics // Journal of Parallel and Distributed Computing, 2014, vol.74, no.7, pp.2561–2573.
29. Hsu C., Li G., Niu W., Batten L., Dorronsoro B., Danoy G., Bouvry P., Katz D.S., Zhang Z. Intelligent big data processing // Future Generation Computer Systems, 2014, vol.36, 452 pp.
30. Alguliyev R.M., Hajirahimova M.S. The phenomenon of "Big data": Problems and opportunities // Information Technology Problems, 2014, №2, pp.3–16.
31. Alguliyev R.M., Hajirahimova M.S. "Big data" technologies / The problems of electronic government building, 1st Republican scientific-practical conference proceedings, 2014, pp. 214-127.
32. Dean J, Ghemawat S. MapReduce: a flexible data processing tool // Communications of the ACM, 2010, vol.53, no.1, pp.72–77.
33. Shvachko K., Hairong K., Radia S., Chansler R. The Hadoop distributed file system / Proc. of the IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010, pp.1–10.
34. Talia D. Clouds for scalable big data analytics // Computer, 2013, vol 46, no.5, pp.98–101.
35. Zheng Z., Zhu J., Lyu M.R. Service-generated big data and big data-as-a-service: An overview / IEEE 2nd International Congress on Big Data, 2013, pp.403–410.
36. Sangvai P. Impact of big data in oil and gas industry / Proc. 10th Biennial international Conference & Exposition, 2013, pp.439–440.
37. Hems A., Soofi A., Perez E. Drilling for new business value. How innovative oil and gas companies are using big data to outmaneuver the competition, A Microsoft White Paper, 2013, 12 pp.
38. Taneja P., Wate P. Big Data enabled digital oil field / Computer Society of India Communications, 2013, pp.18–20.
39. Baaziz A., Quoniam L. Big data in upstream oil and gas, IDC energy insights, How to use Big Data technologies to optimize operations in Upstream Petroleum Industry // International Journal of Innovation, 2013, vol.1, no.1, pp.1–9.
40. Big Data for the oil and gas industry, Issue 5/4, TechConnect, 6 p.
41. Tapping the power of big data for the oil and gas industry, IBM Software, 2013, 8 pp.
42. Dayal U. Akatsu M, Gupta C. Expanding global big data solutions with innovative analytics // Hitachi Review, 2014, vol.63, no.6, pp.333–339.
43. Big data in oil and gas: how to tap its full potential, Hitachi, WebTech Q&A Session, 2013, 45 pp.
44. Perrons R.K., Hems A. Cloud computing in the upstream oil & gas industry: a proposed way forward // Energy Policy, 2013, pp. 732–737.
45. Feblowitz J., The big deal about big data in upstream oil and gas. Paper & Presentation, IDC Energy Insights, 2012.
46. Nicholson R., Big data in the Oil & Gas Industry, IDC Energy Insights, 2012.