***Ramiz M. Aliguliyev[1], Yadigar N. Imamverdiyev[2]***
Institute of Information Technology of ANAS, Baku, Azerbaijan
[1]a.ramiz@science.az, [2]yadigar@lan.ab.az
DOI: 10.25045/jpit.v08.i1.01

# CONCEPTUAL BIG DATA ARCHITECTURE FOR THE OIL AND GAS INDUSTRY

*Big data technologies provide important approaches and tools for the creation of data management systems in oil and gas industry. The paper proposes a conceptual architecture for a hybrid Big data platform for storing and analyzing large volumes of data gathered from oil and gas industry systems in real-time by deep analytics and machine learning methods in distributed cluster systems. We also consider the question of selection of necessary tools from Hadoop ecosystem for building of a viable Big data solution.*

***Keywords****: oil and gas industry, Big data, Hadoop, Apache Spark, MapReduce, Big data analytics, Big data architecture.*

## Introduction

At present, oil and gas industry is shifting from the light oil era (for its density and the ease of extraction) to the heavy oil era [1]. Oil resources do not exhaust in the Earth, they move to deeper layers and to more remote and difficult areas, which makes its extraction increasingly difficult. One of the main characteristic of this transition era is the extensive application of information technology in the production cycle chain; sometimes it is described as a scientific-technical revolution in oil and gas industry [2].

The leading line in this trend is the introduction of "digital field" (or "smart fields") technologies [3–5]. Philosophy of the digital field is the principle of "measure-model-decide-fulfill-and-control". Instrumental basis of the digital fields consists of fiber-optic sensors. The sensors installed in these oil wells enable the distributed measurement of temperature, pressure and other parameters. Information is transferred from these fiber-optic sensors to the focal points of the digital fields – real-time control center or to the monitoring center. Developing fiber-optic systems for real-time collection and transfer of geological information from the fields, production and technological data processing in the monitoring centers, real time 3D visualization of processes and the process data, and the introduction of robotic technology provide remote control of the operations and remote services [6].

The standard approach to the problem solution in monitoring centers includes the use of high performance and expensive computing resources, but they do not guarantee the desired pace of large-scale data processing. As an alternative, the use of solutions of High Performance Data Analytics based on Apache Hadoop software technology stack for Big Data processing [7]. Big data technologies enable distributed operative data stream processing and building real-time analytics systems used in the monitoring centers.

A number of important technological changes have occurred since the inclusion of the term "Big data" for the first time in 2008 [8]: unstructured data storage and processing have been transferred to the clouds, storage facilities, along with the increased volumes of storage devices, their prices have also fallen, and Hadoop ecosystem has formed. In expensive business-analytics tools and predictive analytics systems are now available to customers, a new storage category – analytical stocks have already appeared on the market and so on [9, 10].

Big data technologies ensure to gain new knowledge by real-time collection and processing of large volumes of data in various formats. Currently, along with other fields, these technologies are introduced to the oil and gas industry [11, 12].

There are too many general statements about how to transform data into the value with the help of big data technologies. The oil and gas industry is no exception. Marketing papers of many companies emphasize the great opportunities of Big Data for data mining and data aggregation in oil and gas industry. However, application of Big Data analytics in the oil and gas industry is still

on an experimental level [13]. Only a few companies are trying to use Big data technology [14]. For example, Chevron uses Hadoop for seismic data processing (IBM BigInsights). Shell has carried out a pilot project on the use of Hadoop Amazon Virtual Private Cloud (Amazon VPC) for the seismic sensor data. There are a few other examples proving the application of Big data concept in the oil and gas industry.

In addition to the exponential growth in the data volume, most available companies do not intensively apply to these data. Therefore, a question arises about how to economically justify the storage of non-required data without exception of a quick access to them [15]. At this point, in the oil and gas industry, centralized storage and distributed processing is of great importance.

This case study aims at developing a conceptual Big data architecture for the oil and gas industry. This paper proposes the oil and gas industry systems and machine learning methods to the analysis of data to support in-depth analytics. Big data platform architecture is proposed. This platform should efficiently accept and store data from any source and make them accessible for Big data analytics tools. The primary task is uploading large volumes of data into the distributed cluster system. To keep the balance between the volume of stored data and the request duration, it is necessary to develop hybrid approach. Selection of the necessary instruments from Hadoop ecosystem for the creation of a viable Big Data solution is studied.

**Large-scale data in oil and gas industry**

In the oil and gas industry, all data has always been important, and large-scale data has always been generated in various areas. However, today, extreme volumes of data are being generated at incredible speed. It is also linked to the successes recently achieved in the field of sensor technologies, the creation of structure sensors (seismic 4C (4-component), fiber-optic sensors for wells, oil and gas extraction, processing and transportation systems). Geophysics (4D seismic – the 4th dimension is time, seismic measurement sequence is taken by a certain time interval in the field), geology and oil and gas fields development (4D monitoring) are the main sources of large-scale data [16].

Two classes of the data are distinguished in oil and gas industry: 1) data obtained from the monitoring and control of technological processes; 2) data processed in the management processes of companies and operations. It is worth taking a brief look at some characteristics of these data classes.

At present, seismic exploration method is primarily applied for the search of oil and gas deposits. The essence of this method is to generate elastic waves artificially, for example, by the explosion on the surface of the ground (or sea), and then to register them with special seismic sensors on the ground surface. The volume of seismic data obtained during the exploration of a deposit can reach terabytes [16].

The reasons for the rapid increase in the volume of geological and seismic data are obvious. This is due to the transition from two- and three-dimensional seismic analysis to four-dimensional seismic analysis, which increases the volume of data by ten times. Moreover, the efforts are often made for the development of the fields in difficult conditions, which requires conducting seismic surveys that provides more detailed information [17].

In recent years, Logging while drilling (LWD), Measurement while drilling (MWD) and Seismic while drilling (SWD) technologies are used for drilling control. Passive seismic monitoring can be used both for vertical and horizontal drilling and hydraulic drilling of the layers [17].

The introduction of fiber-optic sensors allows measuring the temperature, pressure and other parameters each 100 m, 10 m and 1 m and even 10 cm of the well [18]. Real time control of the work and condition in the well through the sensors is also possible. The most demanding tool in practice can be a well visualizer, the application of which may realize the dream of oilmen in the field – to see the status of the mechanical and physical characteristics of each point of the well during the entire life-cycle of the field with the naked eyes. The daily volume of the data, obtained

from any such sensors in a well, reaches a few terabytes.

A large number of Magnetic flux leakage (MFL) sensors are used to control the oil and gas pipelines, and to determine the location and size of different types of defects of the pipelines. Sensors are placed at an equal distance along the bounds of the pipe, and they measure MFL signals each 3 mm. As a result, the gathered volume of data becomes quite large [19].

The satellites (for example, Gazprom "Yamal" a group of satellites) can be used to transfer large-scale information from the periphery to the center. However, when it comes to a real-time transfer of large-scale data, the power of satellite communication channels can be insufficient. In this case, fiber-optic communication channels for the transmission of very large volumes of data are used. For example, in late 2010, BP completed the installation of fiber-optic communication channel of 1,200 km length connecting all offshore platforms in the Gulf of Mexico. Implementation of the project cost 80 million USD. Norway has a similar experience, although, unlike BP, it is not a closed system, as the coastline is long, and optical fiber communication channel goes directly to the sea platforms. Furthermore, data are processed on the coast. There are similar projects in Russia for Stockman deposit and Lukoil offshore platforms in the Caspian Sea [20].

3D geological and hydrodynamic models of oil and gas fields also generate large volumes of synthetic data. During the geological modeling of oil and gas fields, 3D model of layers is developed, and based on this, the hydrocarbon reserves in the layer is estimated. The 3D hydrodynamic model based on geological model shows the changes in the layer properties and in the volume of reserves, as well as the pace of oil (gas) extraction by the wells. Special software of companies such as Landmark, Roxar, Schlumberger and TimeZYX is used for the geological and hydrodynamic modeling of the fields. Geological and hydrodynamic models are adapted based on the data received from sensors installed in the well. Different texts, tables, graphical reports are obtained based on geological and hydrodynamic models of oil and gas fields. Thus, in geological and hydrodynamic modeling of oil and gas fields, processing of terabytes, even petabytes of seismic, geophysics and mining data becomes a routine procedure [21].

Another source of data is the digitization of historical data gathered on the geological and geophysical surveys conducted in the company. The data collected on all the fields and wells (including drilling and production) of the company, and the technical passports of the objects are digitalized in compliance with industrial standards and are included in the centralized archive.

In addition to the complex automation of production processes of oil and gas companies, the informatization of management and business processes is also implemented on a large scale (projects for the introduction of SAP system in several companies). The data on the management processes are processed in budgeting systems, management reports, performance monitoring systems, consolidation of the financial reports, contracts, treasury, risk, human resources management systems. The data on the operational processes are collected in project management systems, repair works, production operations, procurement and order management, production planning, environmental protection, transportation management and logistics planning systems.

For some obvious reasons, the oil and gas companies are always in the focus of the media, commercial structures, as well as the state authorities, and therefore, collection and analysis of large volumes of unstructured information from external sources is also very important. These sources include the mass media, websites, social media, e-mail, various reports, photos, and multimedia. It should be noted that this is unstructured or semi-structured data. For this reason, their storage in the traditional data warehouse, regular availability and analysis are very complicated.

The complexity of the organizational structure, complex industrial processes scattered throughout the wide territory, and variety of areas requires solutions from the oil and gas companies, which connect all the data necessary for effective enterprise management in an unified information space. Traditional IT infrastructures, particularly the infrastructures of the unrelated data storage systems operating for various areas or sections are increasingly diversifying, and their service is getting complicated and costs are increasing, and their management is getting difficult.

The limited capabilities of these systems do not provide the level of economic efficiency required from the current oil and gas industry institutions. This is one of the main reasons proving the urgency of the use of Big Data technologies in the oil and gas industry.

**General information about Hadoop ecosystem**

Hadoop ecosystem is considered as synonym to Big Data technologies. Initially, Hadoop was a tool for data storage in clusters and MapReduce parallel processing, but now it is a big stack of technologies related to large-scale data processing (not only through MapReduce).
Hadoop core includes [22]:

- **Hadoop Distributed File System (HDFS)** – distributed file system that enables the storage ofpractically unlimited volumes of data.
- **Hadoop YARN** (*Yet Another Resource Negotiator*) – platform for the control of cluster resources and problems.
- **Hadoop MapReduce** – distributed platform for programming and execution of MapReduce-computing.
- **Hadoop Common** – set of utilities and libraries used by other modules in Hadoop ecosystem. For example, HBase uses Java archives (JAR files) stored in Hadoop Common to access HDFS modules.

Currently, there are many Apache projects directly associated with Hadoop, but not included in the Hadoop core:

- **Hive** – Software for SQL queries on large databases (turns SQL-queries into a series of MapReduce-tasks);
- **Pig** – programming language for top-level data analysis. In this language, a single line of program code can be converted into a series of MapReduce-tasks;
- **Hbase** – column data base realizing Big table paradigm;
- **Cassandra** – high performance distributed "key-value" database;
- **ZooKeeper** – service for distributed configuration maintenance and the synchronization of changes made to configuration;
- **Mahout** – machine-learning program library on large-scale data.

First of all, when we speak about Hadoop, we focus on its file system - HDFS. At a first glance, it is made up of an ordinary file system, file descriptors' tables and data fields. Instead of table, HDFS uses a particular server -NameNode, and the data is distributed across a large number of DataNodes. The data are divided into blocks (usually 64 Mb or 128 Mb), the server saves the route of each file, the list of blocks and block replicators. HDFS system has a classic tree structure of UNIX directories; users triple rights and even console commands are similar.

The key feature of HDFS is its reliability. Classic configuration of Hadoop cluster contains Name Node, MapReduce master (called JobTracker), DataNode, and a set of working computers TaskTracker. MapReduce consists of two stages [23, 24]:

*Map* – executed in parallel and locally (where available) on each data block. The application is sent to the server, where the data is located, instead of carrying large volumes of data to the application's location, and it realizes data processing.

*Reduce* – main node collects pre–processed data from the working nodes, combines them and generates the problem solution.

Hadoop MapReduce is designed for the data packets processing, it solves the problem consecutively, which slows down the system operation. Therefore, Hadoop is often used for data maintenance, rather than their processing.

**Hadoop distributives**

Hadoop project is a top–level project of Apache Software Foundation, so Apache Hadoop is

considered the main distributive and central repository for all other developments. However, a number of difficulties in practice accompanies this distributive: to install Hadoop in cluster the computers should be pre-configured, packages to be installed, many configuration files to be adjusted and other operations are required. Human carries out the work and the supporting documents are often incomplete or absent. Therefore, in practice, distributives are used to automate this work, and the distributives of three companies: Cloudera, Hortonworks, MapR are mostly used.

CDH (*Cloudera Distribution including Apache Hadoop*) – combines the most popular tools of the Hadoop ecosystem under Cloudera Manager. Cloudera Manager is responsible for cluster setup, installation of all components and their subsequent monitoring. Along with CDH, the company is developing its other products as Impala (database).

HDP (*Hortonworks Data Platform*) – distinguishing feature is developing Apache products rather than its own ones (Table 1). For example, Apache Ambari is used instead of Cloudera Manager, and Hive – instead of Impala. As a result, HDP is more open, and cluster control system Ambari enables to develop solutions for both Linux and Windows, and provides to migrate to Azure HDInsight cloud environment. Accordingly, HDP can be selected as the base distributive to build the monitoring center, and it replicates solutions to the maximum number of platforms.

MapR – unlike the previous two companies, the main source of income of which is consulting and partnership programs, it directly deals with its products sale. Its advantages include a large number of optimizations and Amazon partnership program. Disadvantage is the limited functionality (M3) of free version. In addition, MapR company is both the main ideologist and creator of Apache Drill.

Table 1

| **Processing stage** | **Software** |
|---|---|
| Receiving input data flow | Kafka data broker |
| Indexing, search, classification, clustering, simple statistical processing | Solr indexing |
| Online processing of intensive data flow | Apache Spark Streaming, Storm |
| Large-scale linear data packet processing | MapReduce, Apache Spark, Apache Spark DataFrame |
| Large-scale graphics data packet processing | GraphX, Apache Giraph |
| Statistical processing, machine learning, predictive analysis | Apache Spark ML |

**Big data analytics**

Large-scale data collection, management, analysis and visualization tools and technologies are used in several areas: statistical analysis, computer technologies and applied mathematics. Some of them were previously used to work with small data, and subsequently, were successfully adapted to large-scale data; while others appeared due to scientific issues, and were used by the companies (first of all, Google, Amazon, Yahoo, Facebook, etc.) aimed at working with large-scale data.

The Big Data storage and software platforms, included in Big Data ecosystem, provides the technological base for Big Data, and they ensure collection, storage and control of data received from different sources. Big Data analytics tools are built based on data mining, machine learning, and text mining [25].

Big Data ecosystem problems can be divided into three directions:

1. **Data storage and control** – hundreds of terabytes or petabytes of data can't be stored and managed in traditional relational databases.
2. **Unstructured data processing** – Big Data are mostly unstructured: text, video, audio, images, multimedia and so on. Organization of unstructured data processing and analysis is one of complex research issues. Unstructured Data Mining is a relatively young field of scientific research, whereas more studies have been conducted in the field of text

mining thus far [26–29].

3. **Big Data analysis** – Big Data Analysis uses statistical analysis, Data Mining, machine learning, simulation models, optimization methods, data visualization, data aggregation and integration and so on. Predictive analytics is considered as a separate field of study [30].

Visualization tools, such as color identifiers, semi-transparency, data placing by layers, building isolines and diagrams and etc. should be widely used to reflect easily the perception of all the data necessary for making operational management decisions [31].

Software established within Apache Hadoop technology provides distributed data processing in all phases of data analysis using horizontal scaling, which is one of the main advantages of these systems. Existing algorithms are capable to work effectively even in a cluster, consisting of one server, and can be scaled hundred times connecting to servers in an emergency mode when the volume of processed data is increased.

Big Data technologies are used to process large volumes of quasi-structured data, to perform automated horizontal scaling through proportionally increasing the performance and volume of processed data, to implement prompt search minimally using joining operation, and to operatively process large intensity events flow and so on.

**Conceptual Big Data architecture for the oil and gas industry**

Analytics system of the situation center of the oil and gas industry has to process large-scale data and provide a convenient interface to reflect both incoming data and the results of their analysis. Proposed conceptual Big Data architecture for the oil and gas industry is shown in Figure 1. In the conceptual architecture, selected Apache Hadoop, Apache Kafka, Apache Spark and Apache Spark streaming technologies enable us to process hundreds of terabytes and petabytes of data [32, 33]. Solr indexing server is used to access the processing results promptly.

**Service Bus**. The data received by the oil and gas industry system are coming from the various information systems and digital field sensors, where they are indexed, archived, and the object's condition is estimated and so on. **Service Bus** is used to ensure the coordination of all data flows and to reduce their transfer costs. **Service Bus** enables reception and storage of data from different sources and their centralized distribution to consumers. Relational DBCS (database control system) can be used as a key element of the **Service Bus**; it should provide the storage of the transmitted data and the service information about the status of the data transmission by each source and consumer. In distributed processing, in cluster mode, due to the complexity of mechanisms to ensure large loadings and guaranteed data delivery, the **Service Bus** errors (denials), as well as the problems related to data synchronization and transmission may occur. A variety of solutions (ActiveMQ, RabbitMQ, Apache Kafka etc.) created specifically for high performance distributed processing of data flow are available. The information included in the Kafka data broker of Apache Hadoop ecosystem removes most of the shortcomings listed above.

Information systems are mainly tending to follow two models – queue and publication/subscription. Kafka uses the concept summarizing these two models. Apache Kafka is a distributed publication/subscription queue. The queue for different categories is called topic. Kafka names data publishers - producers, and the readers - consumers. Data queues provide a specific buffer. Kafka topics can be flows. Brokers, who make up Kafka cluster servers, act as a data transmission channel between producers and consumers.

Topics are divided into partition. Each partition is a data set sequence. The data is continuously added to the partition. Each partition is given a unique serial number (called offset). Any information can be referred by its partition number. In each partition a server acts as the leader. The leader manages all reading and writing requests in the partition. And the leader's followers replicate it. When any partition fails, the producers and consumers automatically connect to another server.

The partitions have two purposes - scaling and paralleling. Topics can be divided into several

partitions, each partition can be located in different servers (scaling). Consumer can read various partitions at the same time (paralleling). Kafka performance for data stream transfer can reach tens of thousands of data per second. Due to linear scaling, as in LinkedIn, tens of millions of data can be processed.

**Gobblin** – downloads data from Kafka to HDFS (developed by LinkedIn). Initially, LinkedIn used its Camus for this purpose (billions of data were downloaded per day). Gobblin is a universal platform for the "digestion" (*Extract, transform, and load, ETL*) of large-scale of data from various sources in Hadoop.

**Data Indexing**. Solr server practically performs real-time data stream indexing (events fixing in the index is performed each few seconds), and using various types of records in different cores (tables), in cloud mode, unlimited volumes of data indexed by the horizontal scaling and rich language of search requests are provided [34]. Furthermore, a quick search by most requests (a few milliseconds), faceting, queried data grouping and simple statistical analysis, as well as stored data clustering is provided. The main feature of Solr is its horizontal scalability, high speed search and processing of large-scale data (hundred millions of records). The data main stream comes to the Solr entrance from the chargers, and the chargers receive data indexed by flows (topics) from Kafka broker. The data may also be received from the processes of real-time or offline analysis. A part of data can be adjusted, added or deleted during the quality improvement process.

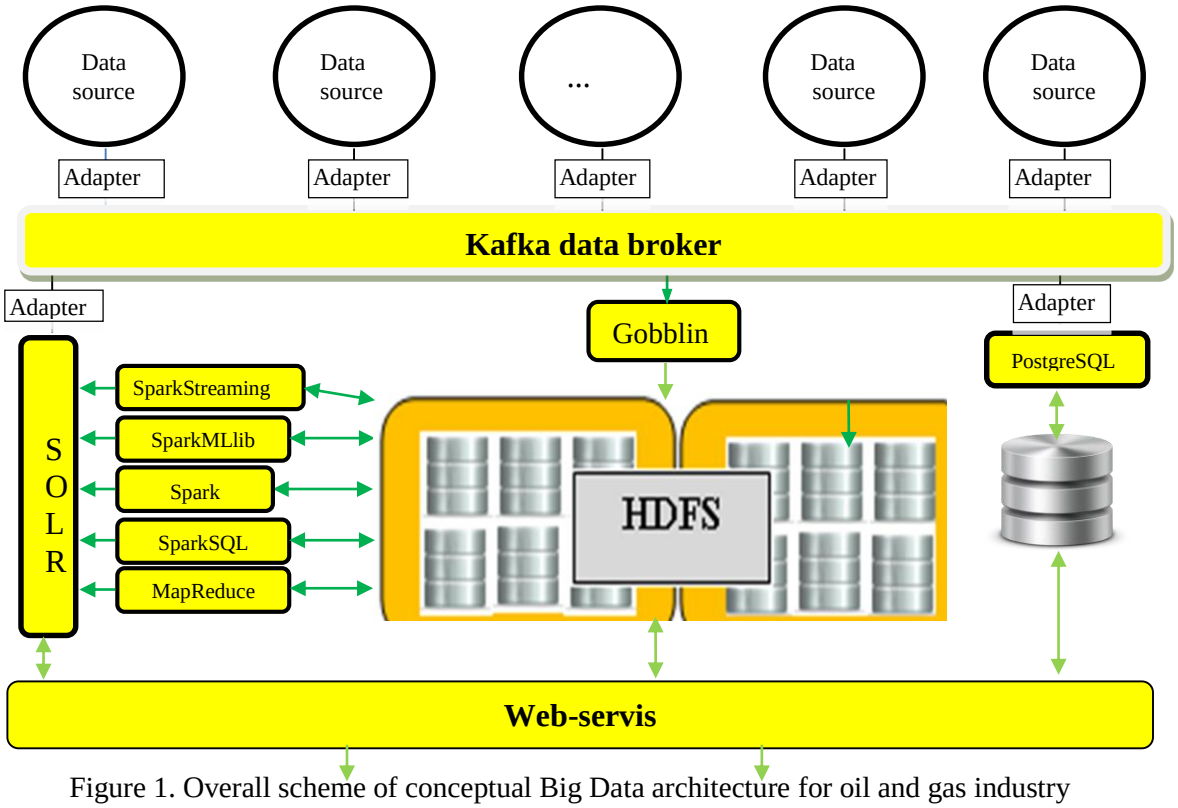


Figure 1. Overall scheme of conceptual Big Data architecture for oil and gas industry

**Large-scale data analysis.** Various Hadoop-based platforms are available for large-scale data analysis – Spark, Storm, GraphMap, H2O and others.

GraphMap – focused on the parallel implementation of machine learning algorithms. Map stage defines the computings to be performed independently from each other (in various nodes), and Reduce stage combines the results.

Storm was announced by Twitter as an open source project in 2011. It is presented as "Hadoop for real-time data processing", and geared towards data stream processing and continuous computations (the results are transmitted by the stream method as they are received).

The proposed conceptual architecture uses Apache Spark for large-scale data analysis. This choice can be justified as follows. The main advantages of MapReduce are scaling, ease of use

and denial sustainability. However, there are a number of shortcomings of MapReduce implementation in Hadoop. The main disadvantage is the low performance of iterative algorithms' implementations (Machine Learning).

Standard MapReduce is designed so that all the results – both end and intermediate results are written to disk. As a result, the duration of writing and reading operations may a few times be longer than the duration of computing. Spark eliminates this problem. Spark also uses data locality idea, but it directs most of calculating directly to RAM instead of the disc. Compared to MapReduce, this enables increasing performance significantly and performing iterative algorithms effectively.

Given this, in conceptual architecture, Apache Spark technology is preferred for solving the complex analytical problems that require multistage complex processing for the oil and gas industry [36]. The main advantage of the Spark is the fast data stream processing. Spark does not support problem distribution, but instead uses Mesos cluster manager, which is responsible for the isolation of resources and their distribution throughout the network.

RDD (*Resilient Distributed DataSet*) is a basic notion in Apache Spark. RDD-based distributed computing model was developed at the University of California (Berkeley). RDD is a set of distributed unchangeable objects on node sets [37]. Most RDD operations are carried out without computations; the computations are used only on demand. When a part of data set is lost, they can be restored, and therefore the application is called resilient set.

In addition to RDD-based Spark, the specific distributed systems – data storage (*Shark*), graphic data processing system (*GraphX*), data stream processing system (*Spark Streaming*), machine learning algorithms library (*MLlib*), and the frame sets for SQL access to processed data (*SparkDataFrames*) have also been developed [37].

One of the interesting features of Spark is that it does not use Hadoop YARN to perform operations, it has own stream API and short-term independent packet processing procedures. The lack of distributed data storage system is one of shortcomings of Spark. Therefore, to launch the Spark, it uses HDFS's data storage capacity. In fact, Spark complements Hadoop and can work together with Hadoop file system.

Apache Spark supports a wide class of operations that provide data processing distributed across cluster servers, which ensures the maximum effective use of processing power of cluster. If the partitions are correctly distributed, the interim data transfer through the network can be to minimized, which practically provides the linear increase of the cluster performance by increasing the number of servers.

Incoming data stream processing based on Apache Spark Streaming module is used for the solution of the issues of the operative nature. Process is enabled for each operative issue and the data stream from Kafka broker is read and depending on the issue, its analysis is carried out.

**Hadoop-SQL tool**. There are a few SQL-oriented application software in Hadoop infrastructure [38]: Hive, Spark SQL and others. Hive should be applied for the compatibility of Hadoop with SQL [39]. Hive historically is the first and currently the most popular database management system in Hadoop platform. HiveQL Query is used as the query language, though it is a concise SQL dialect, it provides quite complex queries on the data stored in HDFS. Its recent version has shifted from the classic MapReduce to Tez platform [40], which has made it practical for interactive analytics by accelerating it many times. ORC (Optimized Row Column) is used in Hive [41].

**Database for time series**. Data processing in oil and gas industry systems is divided into two streams – real-time processing and packet processing of incoming data. The vast majority of the incoming data is processed in real time. In most cases, the data coming from the sensors often have time series properties. Time series can be performed in relational databases. However, if the incoming data speed is very high, then NoSQL solution is required. For time series, in conceptual level, NoSQL system refuses data model and offers a choice depending on the model. At present, OpenTSDB is one of the solutions often used in NoSQL systems to work with the time series [42].

OpenTSDB itself is a collection of TSD (Time Series Daemon) daemon and command utilities. Daemons use HBase to store the data, and also support open protocols for data access. In other words, OpenTSDB – is a user scheme (architecture) of HBase to store time series. This scheme includes interface elements (TSD) and data description model in HBase. Daemons are not interdependent, and this ensures the horizontal scaling when the number of data streams increases. Time series use Graphana as the user interface for data access and visualization.

There are certain features of SQL operators implementation in time series, for example, the reading operation (*SELECT*) is directly related to processing methods. Continuously incoming data processing is always associated with a certain window – a part of data of short time interval. As a result, in relational database, where time series are located, the queries to read the data in successive portions prevail rather than the reading queries with complex conditions. These and other processing features of time series require the use of special systems for their processing and storage. Large-scale time series data processing needs the use of Hadoop SQL tools. However, "wide table" and "blob" formats used in Open TSDB make SQL-based data access difficult. In this case, there are some advantages of working with time series data through Spark SQL.

The most important aspect of NoSQL systems working with time series is support of MQTT (Message Queuing Telemetry Transport) protocol [43]. MQTT is the most popular communication protocol for support of machine-to-machine communication (M2M) and Internet of Things (IOT). MQTT is used as a transport protocol for the implementation of publication/subscription model. Sensors (measurement tools) publish their data, and databases act as subscribers and read them to store published data. MQTT bus support may be required in connection with the time series.

## Conclusion

Cloud technology and Big Data analytics is the basis of the infrastructure of the future oil and gas industry, and they help to organize a world-class infrastructure of this area by ensuring more efficient management of relationships, and provide full use of most precious value – data. Conceptual Big Data architecture proposed in this paper can be considered in practical projects of different scale in oil and gas industry.

## References

1. Campbell C.J., and Laherrère J.H. The end of cheap oil // Scientific American, 1998, vol.278, no.3, pp.78–83.
2. Cross L.R. The technology revolution in oil and gas, 2014. http://www.worldservicesgroup.com/publications.asp?action=article&artid=6496
3. Saputelli L. A., Bravo C., Moricca G., Cramer R., Nikolaou M., Lopez C., Mochizuki S. Best practices and lessons learned after 10 years of Digital Oilfield (DOF) implementations // SPE Paper 167269, SPE Kuwait Oil and Gas Show and Conference, 2013, p.1. http://dx.doi.org/10.2118/167269-MS.
4. Dickens J., Feineman D., & Roberts S. Choices, changes and challenges: lessons for the future development of the Digital Oilfield // Society of Petroleum Engineers. 2012. http://dx.doi.org/10.2118/150173-MS.
5. Feineman D. R. Digital Oilfield Implementation: Learning From the Ghostbusters // Society of Petroleum Engineers, 2014. http://dx.doi.org/10.2118/167831-MS.
6. Holland D. Exploiting the Digital Oilfield: 15 Requirements for Business Value. Xlibris, 2012.
7. Imamverdiyev Y.N. great potential and challenges of Big Data technologies // Problems of information society 2016, No1, pp.23–34.
8. Editorial: Community cleverness required // Nature, 4 September 2008, vol.455, no.7209, p.1.
9. Alguliyev R.M., Hajirahimova M.S. Big Data phenomenon: Challenges and Opportunities // Problems of Information Technologies, 2014, No2, pp.3–16.

10. Gasimova R.T., Big Data analytics: current approaches, problems and solutions // Problems of Information Technologies, 2016, No1, pp.75–93.
11. Hajirahimova M.S. Opportunities and challenges of big data in oil and gas industry //Proceedings of the National Supercomputer Forum (NSKF 2015), Russia, Pereslavl-Zalesskiy, 24–27 November, 2015.
12. Alguliyev R.M., Imamverdiyev Y.N., Abdullayeva F.J.Studying the opportunities of Big Data analytics for the oil and gas industry cloud computing platform as analytics-as-a-service // Problems of Information Technologies, 2016, No1, pp.11–26.
13. Feblowitz J. The Big Deal about Big Data in upstream oil and gas. IDC Energy Insights. October 2012.
14. Baaziz A., Quoniam L. How to use Big Data technologies to optimize operations in Upstream Petroleum Industry // International Journal of Innovation, 2013, vol.1, no.1, pp.19–29.
15. Sangvai P. Impact of Big Data in oil and gas industry // Proc. of the 10th Biennial International Conference & Exposition, 2013, pp. 439–440.
16. Onajite E. Seismic Data Analysis Techniques in Hydrocarbon Exploration. Elsevier Inc., 2014.
17. Hyne N. Dictionary of Petroleum Exploration, Drilling & Production. 2nd Edition. 2014.
18. Zhang M., Ma X., Wang L., Lai Sh., Hongpu Zhou H., Zhao H., Liao Y. Progress of optical fiber sensors and its application in harsh environment // Photonic Sensors, 2011, vol.1, no.1, pp.84–89.
19. Shi Y., Zhang C., Li R., Cai M., Jia G. Theory and application of magnetic flux leakage pipeline detection // Sensors, 2015, vol.15, pp.31036–31055.
20. Bravo C.E., Saputelli L., Rivas F., Perez A. G., Nickolaou M., Zangl G., De Guzman N., Mohaghegh S., Nunez G. State of the art of artificial intelligence and predictive analytics in the E&P Industry: a technology survey // Society of Petroleum Engineers, 2013. http://dx.doi.org/10.2118/150314-PA.
21. Kamal S. Z., Williams J., Liddle J. Continuous improvement of assets through existing and new digital oilfield technology // Society of Petroleum Engineers, 2014. http://dx.doi.org/10.2118/167908-MS.
22. White T. Hadoop: the definitive guide. O'Reilly Media, Inc., 2012.
23. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters // Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation, 2004, vol.6, pp.137–150.
24. Lee K.H., Lee Y. J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey // ACM SIGMOD Record, 2012, vol.40, no.4, pp.11–20.
25. Karthik 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.
26. Fan W., Bifet A. Mining big data: current status, and forecast to the future //ACM SIGKDD Explorations Newsletter, 2013, vol.14, no.2, pp.1–5.
27. Weiss Sh.M., Indurkhya N., Zhang T., Damerau F. Text mining: predictive methods for analyzing unstructured information. Springer; 2005, 260 p.
28. Aliguliyev R.M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications, 2009, vol.36, no.4, pp.7764–7772.
29. Alguliev R.M., Aliguliyev R.M., Isazade N.R. Multiple documents summarization based on evolutionary optimization algorithm // Expert Systems with Applications, 2013, vol.40, no.5, pp.1675–1689.
30. Siegel E. Predictive Analytics: The power to predict who will click, buy, lie, or die. Wiley; 1st edition, 2013, 320 p.

31. Mittelstadt S., Behrisch M., Weber S., Schreck T. et al. Visual analytics for the big data era - a comparative review of state-of-the-art commercial systems // Proc. of the IEEE Conference on Visual Analytics Science and Technology, 2012, pp.173–182.
32. Chardonnens T. et al. Big data analytics on high velocity streams: a case study // IEEE International Conference on Big Data, 2013, pp.784–787.
33. Jones M.T. Spark, an alternative for fast data analytics. IBM developerWorks, November 2011.
34. Wang H., Wang H., Liu Y., Yang F. Design and implementation of SOLR-based information retrieval system for value-added service // The Journal of China Universities of Posts and Telecommunications, 2008, vol.15, pp.51–54.
35. Douglas K., Douglas S. PostgreSQL: a comprehensive guide to building, programming, and administering PostgreSQL databases. – SAMS publishing, 2003.
36. Fazelat R. A Comprehensive analysis - data processing part Deux: Apache Spark vs Apache Storm, January 2016. https://www.linkedin.com/pulse/comprehensive-analysis-data-processing-part-deux-apache-fazelat
37. Tian X., Lu G., Zhou X., Li J. Evolution from Shark to Spark SQL: preliminary analysis and qualitative evaluation. Big Data Benchmarks, Performance Optimization, and Emerging Hardware, 2015, pp.67–80.
38. Abadi D., Babu S., Özcan F., Pandis I. SQL-on-hadoop systems: tutorial // Proc. of the VLDB Endowment, 2015, vol.8, no.12, pp.2050–2051.
39. Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Anthony S., Liu H., Wyckoff P., Murthy R. Hive A Warehousing Solution Over a MapReduce Framework // Proc. of the VLDB Endowment, 2009, vol.2, no.2, pp.1626–1629.
40. Saha B., Shah H., Seth S., Vijayaraghavan G., Murthy A., Curino C. Apache Tez: a unifying framework for modeling and building data processing applications // Proc. of the ACM SIGMOD International Conference on Management of Data, 2015, pp.1357–1369.
41. Huai Y., Chauhan A., Gates A., Hagleitner G., Hanson E. N., O'Malley O. , Zhang X. Major technical advancements in Apache Hive // Proc. of the ACM SIGMOD International Conference on Management of Data, 2014, pp.1235–1246.
42. Prasad S., Avinash S.B. Smart meter data analytics using OpenTSDB and Hadoop // Innovative Smart Grid Technologies-Asia, 2013, pp.1–6.
43. Hunkeler U., Truong H.L., Stanford-Clark A. MQTT-S – a publish/subscribe protocol for Wireless Sensor Networks // Proc. of the 3rd IEEE international Conference on Communication Systems Software and Middleware and Workshops, 2008, pp.791–798.