Towards an Architecture for Handling Big Data in Oil and Gas Industries: Service-Oriented Approach

Farag Azzedin[1], Mustafa Ghaleb[2]
Information and Computer Science Department,
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Existing architectures to handle big data in Oil & gas industry are based on industry-specific platforms and hence limited to specific tools and technologies. With these architectures, we are confined to big data single-provider solutions. The idea of multi-provider big data solutions is essential. When building up big data solutions, organizations should embrace the best-in-class technologies and tools that different providers offer. In this article, we hypothesize that the limitations of the proposed big-data architectures for oil and gas industries can be addressed by a Service Oriented Architecture approach. In this article, we are proposing the idea of breaking complex systems to simple separate yet reliable distributed services. It should be noted that loose coupling exists between the interacting services. Thus, our proposed architecture enables petroleum industries to select the necessary services from the SOA-based ecosystem and create viable big data solutions.**

Keywords*—Service-oriented architecture; big data; Hadoop; oil and gas; big data architecture

I. Introduction

Petroleum industry, one of the pioneers in utilizing big-data-based technologies, has been for a long time using state-of-the-art devices including sensors and actuators, to collect and monitor oil wells [1], [2]. With dramatic changes in oil and gas industry combined with technological advances in how to gather and use massive data, an ideal solution for handling big data is becoming an ultimate goal [1], [3].

Handling large-scale data reduces costs and increases performance by using and integrating latest technologies initiated by service models such as IoT and Cloud solutions. Large petrochemical companies are presently active in utilizing big data technologies, tools, and data sets for processing huge amounts of data generated by their core activities. Data-intensive applications, such as seismic data processing, containing millions of records each with thousands of data values. These records together make seismic data huge traces each of which has few thousands amplitude values. To properly manage and process such large amount of data, appropriate retrieval and computing methodologies should be put in place.

Oil and gas companies can leverage big data technologies to collect, manage, and gain new insights that help increase core activities performance. In addition, big data technologies can help petroleum companies to optimize their business operations, reduce costs, and increase their competitive edge. Key players in big data solutions such as IBM [4], Hortonworks [5], Oracle [6], and Microsoft [7] proposed big-data-based architectures to efficiently accept and store data from any source and make them accessible for Big data analytics tools. These proposed big data architectures for oil and gas industries are being developed based on industry-specific platforms and hence limited to specific tools and technologies. With these architectures, we are confined to big data single-provider solutions. The idea of multi-provider big data solutions is essential. When building up big data solutions, organizations should embrace the best-in-class technologies and tools that different providers offer. For instance, a company might use Amazon for a subscription service but then look to Google for their AI functionality. Gartner predicted that the market for multi-provider solutions will spread out and will be the common strategy for 70% of enterprises by 2019.

In this article, we are motivated by the limitations of the existing proposed big-data architectures for oil and gas industries. We propose a Service Oriented Architecture (SOA) for oil and gas companies, where different services can be employed irregardless of the service provider. Oil and gas companies can use various services without knowledge of their internal processes. Furthermore, service providers will implement only those services that related to their expertise and interest. Service requesters select appropriate services to perform their tasks. SOA realizes many advantages for oil & gas companies including increased agility, improved workflows, extensible architecture, enhanced reuse, and a longer application life cycle. A service provider now is able to quickly and efficiently construct a big data solution reusing already existing services. A service provider can also provide its own services suitable for oil & gas domains. All of these services contribute to the big data software ecosystem. In a nutshell, service providers work together to achieve business objectives while participating in some other big data software ecosystems that target similar environments for different edge solutions.

The idea behind SOA is to create complex systems from a combination of simple parts. SOA is basically an architectural revolution of constructing reliable distributed service-oriented environments for the sake of delivering only functionalities. This comes with the emphasis on loose coupling between cooperating services as stated in [8], [9]. In addition, SOAs are independent of the implementation details of services. As such, utilizing SOA needs only certain standards defining the services and their inputs and outputs. These services can be provided as long as the standards are met. An oil & gas company can choose the best suitable service for its needs since service providers are loosely-coupled. This can encourage service providers to improve their QoS to enhance the oil & gas business.

Big data ecosystems introduce diversity and flexibility.

Flexibility is provided by bringing together different types of service providers to cooperate instead of compete. These different service providers enhance the services diversity available to many service requesters. The SOA notion was stimulated by the emergence of web services [10], [9] which are strong on standards. SOA systems service standards and message exchange. We hypothesize that combining SOA architecture and deploying this architecture as web services, will create flexible, ubiquitous, and liable service infrastructure.

To the best of our knowledge, the only key players in big data solutions, Microsoft and Hitachi, are proposing a solution using SOA. However, their SOA is still with one domain compared to our proposed architecture which is an intra-domain service-oriented architecture. This article proposes big data architecture based on SOA. The proposed architecture enables oil and gas industry systems to 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 while keeping balance between the volume of stored data and the request duration.

II. Related Work

Oil and gas companies relied for decades on data to make decisions in order to expand production and to be competitive in dealing with other companies. Oil & gas companies are trying to increase the effectiveness of analyzing massive data and use the latest technology tools. The main objectives for these companies are to improve production efficiency, reduce costs, and alleviate the impact of environmental threats. Because of substantial volume of data, these companies utilize sophisticated geophysics modeling and state-of-the-art simulations to support and monitor their operations. The collected data is captured by using tens of thousands of sensors in surface facilities and subsurface wells. These data-collecting sensors provide real-time and continuous monitoring of operational assets and environmental conditions [11]. Hence, solutions for handling massive data for oil & gas industries with unique architectures have been proposed.

Oracle [6] proposes a reference architecture[1] for improving oil & gas performance with big data that meets the needs in oil and gas market. Oracle shows the key components of the typical information architecture and how Oracle products can fit in the architecture. Various characteristics are considered in the architecture such as processing methods, format and frequency, data types and consumer applications. In addition, state-of-the-art engines have been added to support real-time processing and the latest big data handling technologies.

IBM [4] introduces a big data platform with broad capabilities designed for oil & gas industries to optimize their operations. IBM built its solution using open-source Hadoop framework with their unique innovations to enhance business performance and streamline their strategic decision making. IBM offers a family of Hadoop distribution offerings that extend the value of open source Hadoop for data processing, warehousing and analytics. IBM products such as InfoSphere BigInsights are introduced as tools in this architecture to enable organizations to turn big, complex data volumes into meaningful data. Using these IBM tools, firms can discover and analyze new business insights hidden in large volumes of structured and unstructured data. Hence, oil & gas industries will be able to ingest and analyze collected data in real-time.

Microsoft [7] proposes an upstream reference architecture[2] to provide a reliable foundation to ensure the interoperability across components and improve analytic and operational efficiencies. This architecture imitates a service-oriented computing environment that includes and integrates business productivity tools, domain applications, and back office applications. It has built a partner ecosystem targeted for oil & gas industries to accelerate their operations and decision-making.

MapR [12] is considered one of the big data technology leader because of its reliable architecture as an enterprise-grade solution. MapR proposes an architecture for oil & gas industries and has its own file system namely, MapR-FS. MapR also employs its own NoSQL database and MapR-DB combined with Hadoop. MapR supports batch and real-time processing applications. MapR's features include large number optimization, consulting and partnership programs, and a free version with limited functionalities. The MapR Converged Data Platform (MCDP) enables oil & gas industries to increase production profitably and tap into all data sets and transform them into one platform for processing and analysis. In 2015, Mtell and MapR provide a new big data platform called Mtell Reservoir[3] that incorporates Mtell Previse Software, MapR Distribution. Hadoop, and Open time-series database software technology. The new system is targeted towards historical and real-time sensor data as well as maintaining data produced from data sources such as oil rigs, mining, chemical plants, water, and waste water.

Hortonworks [5] provides an enterprise ready data platform that helps companies in adopting a modern data architecture. In Hortonworks Data Platform (HDP) architecture, all kinds of data are transferred to Hadoop Distributed File System (HDFS). Many operations are performed on the stored data on HDFS utilizing Yet Another Resource Negotiator (YARN) operating system. Finally, by utilizing their specific tools, data can be visualized. This open source solution is based on Apache Hadoop and supports real-time analysis. HDP has developed many unique modules and added them to the original open source project. HDP with Hive as a central data warehouse layer is used in building dynamic and unified structures. The notable advantage HDP/Hive based architecture with regards to oil & gas industry is scalability. Another advantage is the parallel processing of massive data.

Cloudera [13] is the market leader and known player in the Hadoop space to release the first commercial Hadoop distribution. Cloudera provides data management and analytics platform built on Apache Hadoop and open source technologies. It combines the Hadoop ecosystem under cloudera manager and develops other products such as Impala database. Cloudera and Hortonworks merged recently to become a next generation data platform and deliver industry's first enterprise data cloud. By taking cloudera's investments in machine

[1] http://www.oracle.com/us/technologies/big-data/big-data-oil-gas-2515144.pd

[2] https://news.microsoft.com/download/archived/presskits/industries/manufacturing/docs/UpstreamArchitecture.pdf

[3] https://mapr.com/company/press-releases/mtell-and-mapr-deploy-big-data-platform-oil-and-gas-manage-real-time-and/

learning and data warehousing with Hortonworks' investments in end-to-end data management, this merger will offer cloud-based deployments and allow users to download distributions to be deployed on private as well as on-premises clouds.

Hitachi [14] proposes a reference analytics architecture for oil & gas industry which is developed based on SOA. The architecture contains three main layers: data lake, Hitachi's oil & gas analytics platform, and remote operations center applications. The data lake layer contains MySQL cluster, MonogoDB and file system components. The second layer contains various services such as data access, data ingestion and transformation, feature extraction, process models, knowledge management, events processing, visualization, OLAP, and administration services. The last layer provides applications for oil & gas phases such as exploration, drilling, completions, production, distribution, and maintenance. The architecture lacks supporting real-time processing and the latest DFS big data technologies.

Authors in [11] propose a conceptual big data architecture for oil & gas industry for storing and analyzing acquired data in real-time. This architecture uses a service bus to coordinate data flows, reduce transfer costs, enable data storage, and provide information about the status of transferred data. The architecture uses a broker that acts as a data transmission channel between consumers and producers. The architecture uses specific products and does not support SOA.

As a summary, most of the existing architectures for handling big data in oil & gas industries do not support SOA. Only Microsoft [7] and Hitachi [14] use the concept of SOA to build their architectures. These existing SOA architectures are still inter-domain in nature and hence do not utilize the full advantages of SOA.

III. Background

Big data refers to the increased volume of data that is difficult to gather, store, process, and analyze efficiently utilizing traditional database technologies and software techniques [15], [16], [10]. Big data is generally characterized by six Vs: Volume, Variety, Velocity, Veracity, Variability, and Value [15], [16], [10]. As shown in Fig. 1, big data chain value starts by generating data from huge volume data sources such as sensors, social media, reports, and transactions [17]. This data is then captured and transported into data storage. Data capturing can be done based on the selected solution. For example, big data can benefit from Flume Hadoop module in collecting, integrating, and migrating large data volumes from different data sources into HDFS or any other centralized data storage. Data transportation depends on the data center location. If the data center is local, then transportation will be done in one phase utilizing the same network infrastructure. However, if the data centre is remote, transportation takes two phases. First, inter-datacenter which delivers data from the data source into the edge of datacenter network and then intra-datacenter transportation. Data is then stored depending on the structure of the data. The vast majority of big data is unstructured and therefore is handled with Not NoSQL Hadoop modules such as Cassandra or Mango DB and then processed and analyzed to extract needed information which will help decision makers predict and take proper actions.

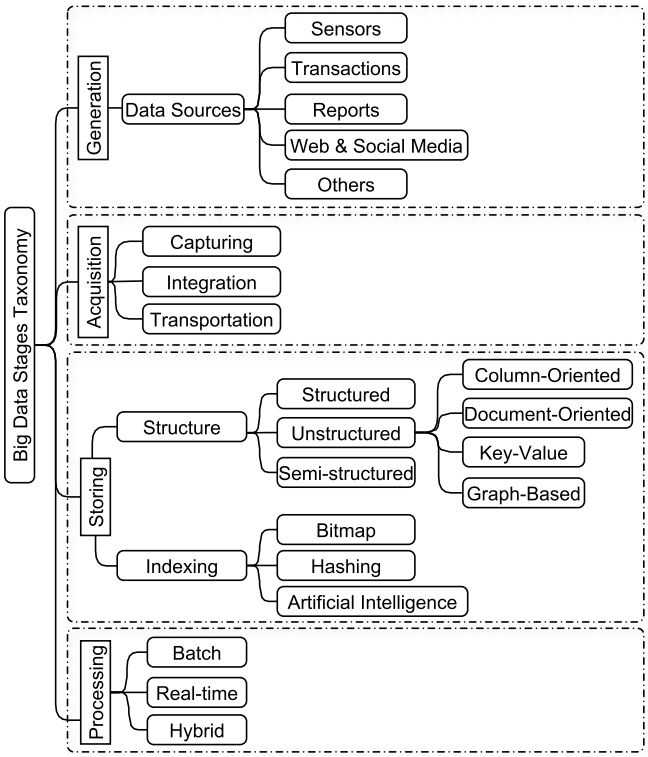

Fig. 1. Big Data Stages Taxonomy.

Cloud computing promises reliable hardware, software, and infrastructure as services provided through the Internet and distributed data centers. Cloud services have been proven to be powerful means of performing complicated large-scale computing tasks. They extend a wide range of IT functionalities from database storage and manipulation to application services. Storing, processing, and analyzing big datasets are making it imperative for many organizations to adopt cloud computing architecture [18]. Big data and cloud computing are connected to facilitate commodity computing for processing distributed queries and returning result sets in a timely manner across multiple datasets. Cloud computing technology solutions offer required infrastructure, tools and technologies to handle big data. Briefly, we can say that there is a mutual impact between cloud computing and big data; i.e., cloud computing offers a perfect solution to big data. On the contrary, massive volume of data comes from development and spread use of cloud computing will increase the potential of big data [17], [16].

Handling a complex, large and fast moving information is difficult using everyday data management tools. For example, in oil & gas industry, big data sources are heterogeneous and also these sources can be previously untapped and relatively new, such as weather patterns, seismic input, and social media. Data that comes from these multiple sources is often captured and was not always retained for long-term use. Combining the data from various sources and using similar previously archived data can lead to improved decision making [3]. The oil & gas industry should benefit from big data technologies in order to optimize operations, supply extra insights, provide better monitoring and extra revenues. Moreover, oil & gas or-

ganizations can utilize big data to improve oil exploration and production while increasing safety and reducing environmental risks.

A. Big Data Solutions

The most common solution widely used in handling big data and considered as a de facto standard solution is Hadoop [19]. Hadoop was developed by Yahoo's engineers before it had been adopted by Apache as an open source. It provides a very efficient solution for data storing and processing and for system management and integration different modules. Hadoop is the major software infrastructure platform for developing Internet-based applications similar to MapReduce and Google's file system. Hadoop comprises of two main parts: HDFS and MapReduce framework. HDFS is the foundation for core data storage of all Hadoop applications. It is a distributed file system which serves as data storage source of MR, and which runs on commercial hardware. HDFS distributes and stores files in data blocks of 64MB to various nodes of a cluster in order to enable parallel computing for MapReduce [20]. The HDFS implementation environment might have hundreds or even thousands of servers storing only some part of the whole file system data. This is prone to hardware failures because more servers result in more hardware and hence the probability of failures increase. As such, services such as fault detection and recovery are fundamental architectural HDFS goals.

Hadoop has proven to be a powerful technology to solve many challenges of big data. In the analysis and management domain of big data, Hadoop introduces many advantages in various areas such as expandability, high-cost efficiency, strong flexibility, and high fault-tolerance. Recently, Hadoop was utilized widely in big data industrial applications such as clickstream analysis, spam filtering, social recommendation and network searching. Hadoop commercial support and execution are provided by many organizations including MapR, Cloudera, IBM, Oracle, and EMC.

Currently, there are three types of Hadoop distributions. Commercial distributions namely, MapR, Hortonwork, and Cloudera. These distributions are not cost-effective but they have better performance and better deployment flexibility. The second type of Hadoop distribution is Apache open source, considered to be cost-effective and widely deployed in the industry. The cloud hadoop distribution is deployed at Amazon, Google and, Microsoft. All of these distributions are not ready made solutions. They need to be customized based on business strategy as well as business current technology. Prior to that customization, a kind of big data maturity assessment should be conducted to know to which degree industry relevant IT is ready to deploy such big data solutions [3].

Big data security is an important issue that needs to be considered. Any architecture for handling big data should have cross-layer security services. In handling big data, end-users as well as customers should be insured that ethics and other security requirements will not be violated. The big data architecture has to protect CIA (Confidentiality, Integrity, and Availability) security requirements and other non-CIA security requirements like anonymity, access control, and accountability.

IV. Big Data in Oil and Gas Upstream

The data volume in oil & gas industry is coming from seismic data, spatial/GPS coordinates sensors, weather services, and different measuring devices. Specific applications handle structured data and these applications are utilized to manage all upstream activities such as surveying, imaging and processing, exploration development, reservoir modelling, production, and other activities. The data that is generated through these upstream activities is semi-structured or unstructured such as spreadsheets, emails, images, word processing documents, voice recordings, data market feeds, and multimedia. This means that it is costly or hard to either store, query, or analyze such data. To this end, suitable tools and technologies for big data need to be utilized [3].

The entire upstream process begins with the acquisition of seismic data across a potential area of interest in search of petroleum sources. The focus area is identified for feasibility in exploration, drilling and production of crude oil & gas. Once the data is successfully collected, the acquired data is processed and interpreted to determine a location for drilling. The drilling of exploratory wells is then initiated to record technical data that will collect accurate statistics in terms of available reserves. If large enough reserves are proven, the field development is started including installation of production facilities, pipelines, storage facilities, and transportation. Upon successful completion of these activities, the midstream sector takes over. Thus, the upstream sector can be classified into three important segments: exploration, development and production. The exploration phase consists of two important tasks, seismic data acquisition and processing. The development phase consists of several activities including seismic and geological interpretation, reservoir modelling and simulation, exploratory drilling, facilities and reservoir engineering. Finally, the production phase spans reservoir drilling and testing, production development and optimization, and supervisory control and data acquisition (SCADA). Fig. 2 lists the main phases of oil & gas upstream and how are they related to big data.

During the exploration phase, advanced geophysics modeling and simulation techniques are conducted to support seismic operations. With the help of big data technologies in the exploration phase, experts and managers can accomplish operational and strategic decision-making to enhance exploring efforts, new prospects assessment, seismic traces identification, and new models building [3]. Every oil Company uses their own format for storing and processing of seismic data. But the society of exploration has a standard for storing acquired and processed seismic data using tape as either SEG-D or SEG-Y format [21]. Since the acquired large set of data is mostly unstructured, impure, redundant, and in varying formats, it is necessary to process this data using proper data mining and analysis techniques.

In the upstream development phase, the focus is on data analysis and interpretation, provision of standardized tools, and detection of drilling and production problems. Large oil companies, such as Saudi Aramco, have used specialized tools (OilField Manager - OFM) for well and reservoir analysis to automate and dynamically integrate engineering requirements for production optimization. These requirements include remedial well analysis, water management, reservoir management

Big Data	Oil & Gas Upstream: Exploration	Reservoir Engineering & Development	Drilling and Completion	Production
Volume	Seismic acquisition SEGD	Facilities Reservoir engineering	Sensors: - Flow - Pressure - ROP	SCADA sensors: - Flow - Pressure
Variety	Structured data: - SEGD - Pre-stack - Post-stack Semi-structured: - implantation	Structured data: - WITSML(XML) - PRODML - RESML Unstructured data: - Log curves/ Drilling & Test/ Lithology/ Cores…	Structured data: - WITSML(XML) Semi-structured data: - Final well report, - Daily drilling report Unstructured data: - Drilling log/ Gas log …	Structured data: - PRODML - RESML Semi-structured data: - Crude analysis report
Velocity	Real time data acquisition: - Wide azimuth data acquisition		Real time data acquisition: - Mud logging/ LWD/ MWD	Real time data acquisition: - SCADA sensors
Veracity	Seismic processing	Reservoir modeling	Sensor calibration	Sensor calibration
Variability	Seismic interpretation Geology interpretation	Reservoir simulation Combination of seismic drilling and production data	Data interpretation & optimization	Data interpretation
Value	Navigation Visualization & Discovery Run integrated asset models	Improve drilling program Drive innovation with unconventional resources (shale gas, tight oil)	Reduce costs Reduce non productive time Reduce risks Improve HSE performance	Increase speed to first oil Enhancing production

Fig. 2. Big Data vs Oil & Gas Upstream, adapted from [3].

and surveillance, and production data monitoring. So, big data can help to assess and improve drilling programs and drive innovation with unconventional resources.

On the other hand, two main aspects during the upstream drilling namely, drilling interpretations as well as understanding subsurface play a vital role in any big data solution. First, tremendous cost can be saved if big data solutions are used to recognize anomalies that negatively affect drilling and thus causing misleading interruptions. Second, big data solutions can help in drilling to better understand earth subsurface so affordable energy can be delivered safely [3].

Big data technologies also play a role in upstream production. Using such technologies can shift assets to further productive areas. Technologies also can provide business intelligence to reservoir engineers by enabling future prediction based on historical results and by integrating and analyzing data from seismic, drilling, and production processes. Furthermore, big data can help in enhancing oil recovery from existing oil wells, improving performance forecasting, optimizing real-time production, increasing safety measures, and preventing risks.

V. Big Data in Oil and Gas Midstream

Midstream includes monitoring transportation, monitoring the environment, crude assay and predictive maintenance [22]. Monitoring transportation methods include pipelines, rails, barges, oil tankers, or trucks. Monitoring transportation of oil & gas involves collecting data of the transported oil & gas. Mainly the transportation methods are simple and generate a small amount of data. However, pipelines can use complex distribution systems that involve real-time sensors to generate a large amount of data and hence big data.

Monitoring the environment is regulated by governments' policies and companies' protocols. The objective are to protect society and monitor the emissions which could be harmful to people and the environment at large. This monitoring phase is very important and includes real-time collection of sensor data to help analyze and predict environment living conditions based on the levels of pollution emissions. Crude assay is a service provided to the refining sector where it provides information about the expected oil & gas before it arrives to the refineries. This helps to reduce set up time by understanding the quality of the crude oil expected to be processed. Predictive maintenance means identifying the problems in advance to save time.

Escalating demand for midstream infrastructure puts pressure on midstream companies to continue building new infrastructures such as pipelines. Midstream companies also modify existing pipelines to move oil from the well site to a refinery, processor, or storage facility. Many midstream companies consider their data output as big data since it is massive and contains structured and unstructured data. The bit rate of this output is also high. Thus, oil & gas companies are starting to invest in big data relevant solutions such as Hadoop to manage transportation fuel cost, monitor pressure efficiently, and forecast supply and demand [23], [7], [24].

As we know that pipeline pressure fluctuates as a result for either normal or abnormal activities. In both cases, there should be an automated system detecting and responding to these activities. Traditional SCADA systems are not enough for such situations because they are not capable to differentiate between anomalies and standard causes. On the other hand, Hadoop relevant solutions possess the capabilities to automate the process of detecting and responding to these events. Nowadays, pipeline companies are linking between variable producers and end-users raising the complexity of pipeline companies and bringing new challenges. Many oil & gas companies start investing in installing sensors inside and outside their pipelines to measure pressure, temperature, volume, and vibration. This process is generating new trend of unseen data that accordingly is useful for monitoring and decision making [25].

There are also more evidences showing that output data from oil & gas pipeline companies are considered big data. For example, every $150,000$ miles of pipeline creates 10 terabytes of data [17]. This obviously means that output data from oil & gas pipelines is huge enough to say that the first V "volume" is met . Also, pipeline companies such as TransCanada and Enbridge are utilizing four technologies that mainly see, feel, smell, and hear various aspects of their oil pipelines [17]. This means that the generated data will be in different types and formats which comply with the other V "Variety". Furthermore, since data generated from sensors is sent in real-time to help decision makers act proactively in case of any unexpected event, the third V "Velocity" is met.

VI. Proposed Architecture

Oil & Gas industries are required to invest more in proper tools and technologies that support various big data architectures. So, in this field, the need for a unified big data architecture is required for seamless handling exploration, drilling, and production big data. We have proposed a big data block architecture for oil & gas industries, shown in Fig. 3, which shows a set of capabilities that petroleum companies should consider as they enter the big data space. It contains capabilities around data generation, integration, management, security, operations, analytics, and visualization. The architecture enables finding, managing, visualizing and understanding all traditional and big data to be represented as one entity to enhance decision making through many exercises such as exploring new data sources in oil & gas industries for potential value, mining the relevant data to the industry, assessing the business value of the content, detecting patterns, visualizing and reporting the outputs. The architecture consists of three tiers, namely, data generation (data sources), data management, and analytics and visualization.

A. Data Generation Tier

The Bottom tier represents data sources including traditional and non-traditional data. It highlights the importance of considering all data sources. The reason beyond this is to accommodate new data sources such as the data generated from the IoT devices. This will also increase the potential of extracting deeper, bigger, more complex, and frequent data. Thus, enhancing accurate insights and discovering hidden patterns and values will be achieved.

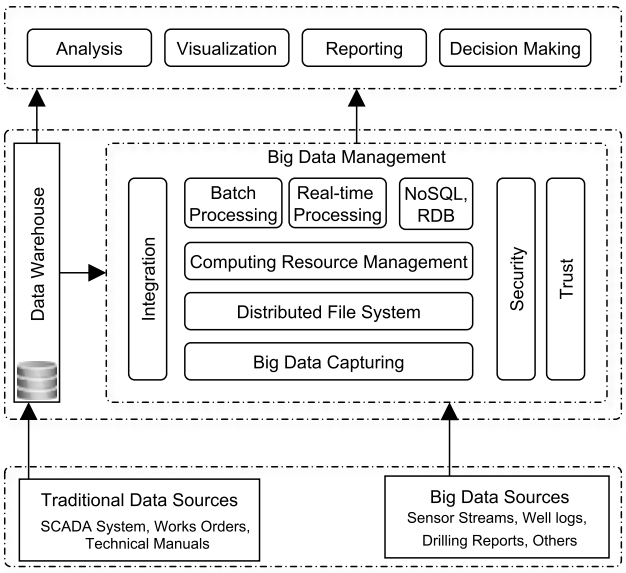

Fig. 3. Big data block architecture for oil & gas industries.

Data generated from sources such as sensors, well logs and drilling reports are considered unstructured. This data is captured by specific big data tools and then moved into new aggregated format to be ready for processing. These various massive data generators from different oil & gas sectors (upstream, midstream, and downstream) need to be considered and utilized by the architecture to ease the process of monitoring pump pressures, RPMs, temperatures, and flow rates.

Traditional data sources represent the data that is available in the organization's repositories, typically stored in a well-defined format such as relational databases and flat file formats and it is mostly structured. In oil & gas industries, the traditional data sources can be from SCADA, work orders, or technical manuals.

B. Data Management Tier

The data management tier consists of two main components: data warehouse and big data management. These two components are integrated to be seen as one entity and utilized for data reporting and visualization. Data Warehouse exists in every enterprise to model and capture the essence of the business from their enterprise systems. The structured data generated from traditional data sources such SCADA systems, work orders and technical manual is transformed and stored in the data warehouse to model and support interactive business intelligence actions.

Data integration is a combination of business and technical processes utilized to combine data from various sources into valuable and meaningful information utilized by tier three. A comprehensive data integration solution includes discovery, monitoring, cleansing, and transforming data from different sources. In the proposed architecture, the integration process is done to traditional data coupled with semi-structured and unstructured massive data sources from big data sources to

increase the success rate of potential projects and key analytical initiatives. Oil & gas companies want to integrate while continuing their governance and data quality best practices. The integration process deals with very large files and provides reliability, fault tolerance, efficiency and scalability.

The computing resource management component works with multiple processing models such as real-time and batch processing. This layer makes this architecture flexible in terms of design and implementation by supporting multiple processing models. Data can be processed in a batch or a real-time mode depending on the data source and the business goal. Batch and real-time paradigms fit well with petroleum industries. Batch big data processing uses large data volumes of which sets of transactions are captured over a wide span of time. Data is collected and processed producing batch results. Batch processing needs separate services for the input phase, the process phase and the output phase. In contrast, real-time data processing needs a continual input phase, process phase, and output phase. Data must be processed in a small time domains.

The real-time processing paradigm fits well with petroleum industries because many incidents such as pump pressures and temperatures need to be monitored to take a quick reaction such as corrective action. In the oil & gas industries, not having real-time intelligence can lead to safety issues, poor decisions, maintenance issues and can cost money. Accurate real-time data prevents repeat failures and streamlines maintenance, land processing and acquisition, drilling and exploration plans. The structured data is stored in Relational Database (RDB) whereas unstructured data in NoSQL database products. A NoSQL database mechanism is used to store and retrieve large amount of distributed data and provides useful features including replication support, schema-free, simple API, and flexible and consistent modes. Common NoSQL database types are key-value, document-oriented, and column-oriented which provide major support for big data handling.

Security and trust management cross all big data management components to ensure that collecting, storing, processing, and accessing data are handled by secure and trustworthy entities. Security and trust management modules are integrated with other data management modules by offering the necessary APIs and interface to manage, monitor, provision, and operate the solution clusters at scale.

C. Business Intelligence Tier

This tier brings intelligence data and functionality closer to users. This tier provides interfaces for analysis, reporting, visualizing analyzed information to provide value to the petrochemical companies to take the right decision in their business. This tier provides an environment for the business intelligence products such as Spreadsheets, reporting and querying software, and information delivery portals. These products run and communicate with users by sending data to and receiving data from the user's middle tier, which relies on intelligent servers to perform processing, including data query and analysis.

This tier enables both existing and new application to analyze, report, and visualize analyzed information to provide value to the petrochemical companies in channeling their decisions. The resulting real–time and pre-computed models are merged for visualization and prediction purposes. Reports can be provided on an hourly, daily, weekly, monthly, or yearly basis. Also, users can generate interactive reports based on their needs utilizing available data and other ad-hoc reports. After getting analyzed information, users can display it using some visualization tools either in graphical or tabular format.

D. Service-Oriented Approach

Existing handling big data architectures are product-based. Our proposed architecture is service-oriented based combining suitable services from different providers regardless of the products as long as the services are provided.

Fig. 4 shows the required services for the basic operations for handling massive data in oil & gas industries. The selected different vendor services can exist at different locations and can still communicate and cooperate with each other to achieve global business objectives. The discovery service has association, dissemination, and matchmaking sub-services responsible for service registration and ensuring that registered services are legitimate and available to service requesters. In our proposed architecture, association service helps service providers make their services available on the ecosystem. A service provider needs to associate itself, connect, and cooperate on the network. The dissemination service propagates available services to other service requesters by advertising summary information about available services. There is a collaboration between the association and dissemination services for advertising the presence of association service and information of the providing services. The matchmaking service should be available to answer service queries with the list of highly recommended service providers. The matchmaking service can interact and cooperate with other services in order to provide various priority levels for other services based on service requirements such as security and trust.

When a service requester gets a candidate list of service providers willing to give the service and might meet the QoS needs nominal by the service requester, the service requester must choose the required services that best address its issues. The service requesters request the selection service to decide for the best and suitable service for its business purposes.

The quality assessment service provides multiple services that can be invoked from other services in the system to assess numerous QoS attributes of any given system service. Quality assessment services prioritize and highlight the service providers that provide trustworthy and secure services. These quality assessment services, such as security and trust, play a vital role towards the success of any service-oriented ecosystem.

As shown in Fig. 5, the functional architecture is presented and the aim is to show the segregation of functionalities across the different layers of the architecture. On top of the data sources layer, the data acquisition layer which enables to capture data and integrate multiple data from different sources and transfer the aggregated data to data storage management. The core component layer is built on top of data acquisition layer. This layer, the core component layer, sets the core functionalities of the architecture to handle and get value from massive data. As depicted in the Fig. 5, there are 2 core components and each component has sub-components.

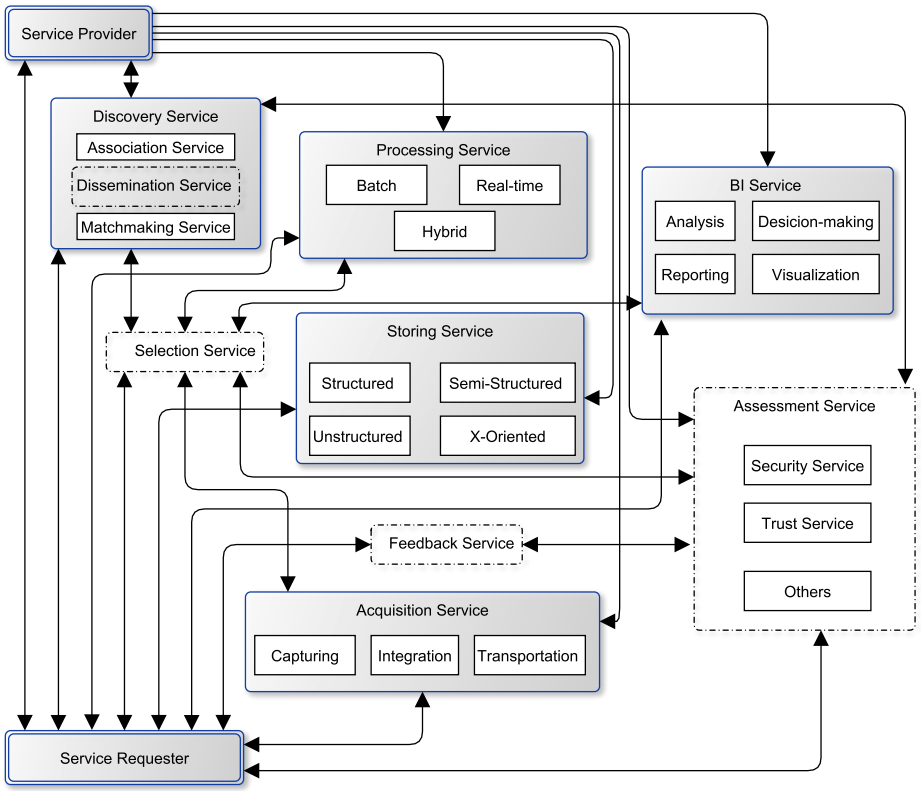

Fig. 4. Big data service-oriented architecture.

These core components are related to processing and storage. The quality assessment layer ensures the transactions and interactions are done in a secure way. Also, this layer ensures the trustworthiness of the data sources and services. Furthermore, it facilitates the discovery of the best registered services and contains a feedback module to receive feedback of all transactions. Finally, the application layer is also called business intelligence layer is presented on top of all layers to analyze, visualize, and report the gained outputs. This layer also allows the decision maker to make a decision based on some business purposes.

In Fig. 6, we show how operations are employed to accomplish functionalities. The data can be generated from either in traditional data sources such as SCADA system or from big data sources such as sensors. The capture task is responsible to capture data from big data sources and the data warehouse keeps the historical data. So, we need to integrate all data and transform it into the distributed file system and this step done by integration and transportation tasks. The distributed file system allows users to store and share their data and make the data more protected from a node failure. It does not serve to data processing directly but is an essential part for data processing tasks. Processing data tasks can be either batch or real-time or hybrid. The batch processing happens when you process the data that have already been stored over a period of time in the distributed file system. While the real-time processing takes place when you process data in real-time that comes from data sources. The hybrid processing is a mix of batch and real-time data processing tasks. The security, trust, and feedback are alongside of all tasks to provide a secure environment for sharing and processing data. Finally, business intelligent tasks which include analysis, visualization, reporting and decision-making tasks to get insight and extract valuable information to ease the process of take decisions.

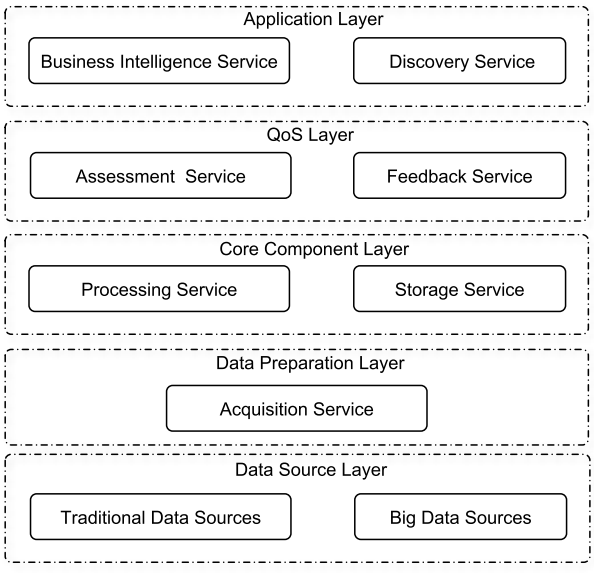

Fig. 5. Big data service-oriented architecture: Functional architecture.

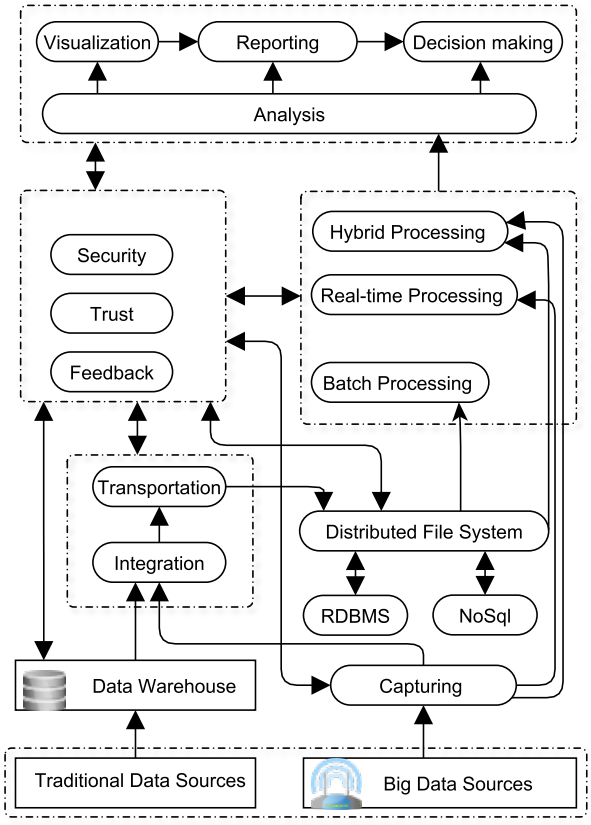

Fig. 6. Big data service-oriented architecture: Operational architecture.

VII. Conclusion

In this article, we are motivated by the limitations of the existing proposed big-data architectures for oil and gas industries. We propose a SOA for oil and gas companies, where different services can be employed irregardless of the service provider. Oil and gas companies can use various services without the knowledge of their internal processes. Furthermore, service providers will implement only those services that related to their expertise and interest.

We proposed an architecture where complex systems are created from a combination of simple parts. Thus, services can be provided as long as the standards are met. Hence, oil & gas companies can choose the best suitable service for their needs since service providers are loosely-coupled. Since each organization is unique, solutions are tailored for individual organizations by providing the architecture as services.

Acknowledgment

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM). This project is funded by King Abdulaziz City for Science and Technology (KACST) under the National Science, Technology, and Innovation Plan (project number 13-INF2452-04).

References

[1] H. Hassani and E. S. Silva, "Big data: a big opportunity for the petroleum and petrochemical industry," *OPEC Energy Review*, vol. 42, no. 1, pp. 74–89, 2018.

[2] S. L. Nimmagadda, T. Reiners, and A. Rudra, "An upstream business data science in a big data perspective," *Procedia Computer Science*, vol. 112, pp. 1881–1890, 2017.

[3] A. Hems, A. Soofi, and E. Perez, "How innovative oil and gas companies are using big data to outmaneuver the competition." 2013.

[4] M. Brulé, "Tapping the power of big data for the oil and gas industry," *IBM Software white paper for petroleum industry*, 2013.

[5] S. Justin, "Modern oil & gas architectures built with hadoop," https://hortonworks.com/blog/modern-oil-gas-architectures-built-hadoop/, accessed: 2018-06-14.

[6] J. Hollingsworth, "Big data for oil & gas," *Oracle Oil & Gas Industry Business Unit*, 2013.

[7] A. Hems, A. Soofi, and E. Perez, "How innovative oil and gas companies are using big data to outmaneuver the competition. microsoft white paper," 2014.

[8] J. E. Hannay, K. Brathen, and O. M. Mevassvik, "Agile requirements handling in a service-oriented taxonomy of capabilities," *Requirements Engineering*, vol. 22, no. 2, pp. 289–314, 2017.

[9] M. Abdellatif, G. Hecht, H. Mili, G. Elboussaidi, N. Moha, A. Shatnawi, J. Privat, and Y.-G. Guéhéneuc, "State of the practice in service identification for soa migration in industry," in *International Conference on Service-Oriented Computing*. Springer, 2018, pp. 634–650.

[10] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big data and cloud computing: innovation opportunities and challenges," *International Journal of Digital Earth*, vol. 10, no. 1, pp. 13–53, 2017.

[11] R. M. Aliguliyev and Y. N. Imamverdiyev, "Conceptual big data architecture for the oil and gas industry," *Problems of information technology*, pp. 3–13, 2017.

[12] MapR, "Predictive maintenance using hadoop for the oil and gas industry," https://mapr.com/resources/predictive-maintenance-using-hadoop-oil-and-gas-industry/assets/mapr_whitepaper_predictive_maintenance_oil_gas_051515.pdf, 2015.

[13] J. Russell, *Cloudera Impala.* " O'Reilly Media, Inc.", 2013.

[14] R. Vennelakanti, A. Sahu, and U. Dayal, "Winning in oil and gas with big data analytics," *Hitachi Review*, vol. 65, no. 2, pp. 884–888, 2016.

[15] J. J. Seddon and W. L. Currie, "A model for unpacking big data analytics in high-frequency trading," *Journal of Business Research*, vol. 70, pp. 300–307, 2017.

[16] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, 2014.

[17] I. Hashem, I. Yaqoob, N. Anuar, and S. Mokhtar, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, 2015.

[18] H. Liu, "Big Data Drives Cloud Adoption in Enterprise," *IEEE internet computing*, 2013.

[19] T. White, *Hadoop: The definitive guide*, 2012.

[20] M. Chen, S. Mao, Y. Zhang, and V. Leung, *Big Data-Related Technologies, Challenges and Future Prospects*, 2014.

[21] E. Onajite, *Seismic data analysis techniques in hydrocarbon exploration*, 2014.

[22] PSAC, "Industry overview," http://www.psac.ca/business/industry-overview/, accessed: 2018-07-20.

[23] K. Kohlleffel, "The power of advanced analytics for midstream oil and gas," https://hortonworks.com/blog/the-power-of-advanced-analytics-for-midstream-oil-and-gas/, accessed: 2018-07-21.

[24] D. Cowles, "Oil, gas, and data," https://www.oreilly.com/ideas/oil-gas-data, accessed: 2018-06-30.

[25] A. Slaughter, G. Bean, and A. Mittal, "Connected barrels: Transforming oil and gas strategies with the internet of things," https://www2.deloitte.com/insights/us/en/focus/internet-of-things/iot-in-oil-and-gas-industry.html, accessed: 2018-07-20.