Journal of Petroleum Exploration and Production Technology (2020) 10:2601–2611
https://doi.org/10.1007/s13202-020-00889-2

REVIEW PAPER - PRODUCTION ENGINEERING

Transforming petroleum downstream sector through big data: a holistic review

Harsh Patel[1] · Dhirenkumar Prajapati[2] · Dharamrajsinh Mahida[1] · Manan Shah[3]

Received: 27 November 2019 / Accepted: 9 April 2020 / Published online: 27 April 2020

Abstract

Big data refers to store, manage, analyze, and process efficiently a huge amount of datasets and to distribute it. Recent advancements in big data technologies include data recording, storage, and processing, and now big data is used in the refinery sector for the estimation of the energy efficiency and to reduce the downtime, maintenance, and repair cost by using various models and analytics methods. In the liquefied natural gas and city gas distribution industry, also, it is used in maintenance and to predict the failure of process and equipment. In this paper, authors have reviewed that how big data now used in the storage and transportation of oil and gas, health and safety in the downstream industry and to accurately predict the future markets of oil and gas. There are many areas where we can efficiently utilize big data techniques, and there are several challenges faced in applying big data in the petroleum downstream industry.

Keywords Big data · Data science · Oil and gas · Production · Refinery

Introduction

Technological advancements in petroleum industry in recent years cause generation of huge amount of datasets in different sectors of petroleum industry including upstream, midstream, and downstream. Big data is the technology which helps oil and gas companies to handle and process these huge datasets from upstream, midstream, and downstream. International Data Corporation Energy took one survey in 2012, based on that about 70% of US oil companies were unaware from big data and its applications in oil and gas industry. Recently, there was a survey done by General Electric and Accenture and executives in which they found that 81% of them considered big data in the top priorities of oil companies. (Mohammadpoor and Torabi 2018).

Even the renewable energy industry also uses the big data analytics to predict the energy generation by various sources such as solar, wind, hydropower. Recently, Ifaei et al.'s (2018) case study on renewable energy in Iran took place, where they use clustering analysis (K-means) method to analyze the renewable energy sources in Iran. As a result, they found that the share of solar, wind, hydro- and biogas sources had average shares of 55.7%, 25.7%, 12.7%, and 5.9%, respectively. So, Iran can implement the solar and hydropower renewable sources for their green energy generation.

There are various definitions of big data by different studies. Some of them are as below.

Big data is the amount of data beyond the ability of technology to store, manage, and process efficiently (Manyika et al.2011; Jha et al. 2019; Kakkad et al. 2019; Kundaliya et al., 2020). Big data is a term which defines the hi-tech, high speed, high volume, complex, and multivariate data to capture, store, distribute, manage, and analyze the information (Shah et al. 2019; Patel et al. 2020a, b; Ahir et al. 2020; Parekh et al. 2020).

Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization (Jani et al. 2019; Patel et al. 2020a, b).

Big data technologies are new generation technologies and architectures which were designed to extract value from multivariate high volume datasets efficiently by

✉ Manan Shah
manan.shah@spt.pdpu.ac.in

[1] School of Petroleum Technology, Pandit Deendayal Petroleum University, Gandhinagar, Gujarat, India

[2] Petrowatch, Ahmedabad, Gujarat, India

[3] Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, Gujarat, India

providing high speed capturing, discovering, and analyzing (Gantz and Reinsel 2011; Shah et al. 2020a, b; Pandya et al. 2020; Sukhadiya et al. 2020).

Hashem et al. (2015) defined big data by combining various definitions in the literature as follows:

In the cluster of methods and technologies, new forms are integrated to unfold hidden values in diverse, complex, and high volume data sets (Gandhi et al. 2020). The use of big data applications is increasing day by day in energy sector and that creates significant opportunity in energy conservation, energy management, environmental protection and energy consumption and generated production data. A huge amount of data generated from wide spectrum of energy sources includes smart meter reading data, weather–climate data, asset management data, energy performance certificates, building stock editing, sustainable energy, socioeconomic data, and comfort levels etc. (Marinakis et al. 2018).

In recent years, the global oil and gas industry has had to explore extremely turbulent waters after a long period of investment in super-capital assets and significant funding to help strengthen investments. In light of the recent surge in digital and data in different areas, oil companies need to consider digital technology advancements to take advantage of the additional benefits derived from the current income generation limit within the company. Digitization can be used to leverage additional resources to maximize growth in the oil and gas sector to create value in an interconnected energy system.

"Data is the oil of the new economy" has been the most recognized reference in recent years. Global financial meetings have even embraced him. Be that as it may, there is so much to say about the huge data generated by the oil sector and its upstream part in particular. In the upstream part, understanding and using information allow companies to stay focused on themselves by organizing, investigating, representing, and progressing in the field.

The 2D, 3D, and 4D geophysical organizations with oil and gas propulsion propelled by seismic sources generate remarkable information in the middle of the investigation phases. They almost control the execution of their operational resources. To do so, they use a large amount of information collection sensors in underground wells to provide consistent and consistent information to understand the benefits and environmental conditions. Unfortunately, these data come in different and progressively complex structures, making it a test for the collection, translation and use of divergent information. For example, Chevron's internal computer movement alone exceeds 1.5 terabytes every day.

Advances in huge information coordinate collections of normal and unique information to deliver the right data at the right time to the right leader. These capabilities enable organizations to track large volumes of information, change core responsive leadership to proactive, and enhance all periods of investigation, progression, and creation. In addition, enormous information offers different possibilities to guarantee safer and more reliable tasks. Another significant impact of this would be shared learning.

The oil and gas industry is extremely concentrated, and its condition is exceptionally controlled. Against this questionable condition described, these organizations must increase creation, improve costs, and mitigate the effects of natural hazards. The downstream oil and gas sector is a complex, information-driven business with exponentially increasing volumes of information. They need to capture and monitor more information than any time in recent memory and try to store, explore, and get useful data from this huge amount of information. In addition, from this information, organizations in the oil and gas sector can obtain quantifiable incentives. The purpose of this paper is to show how large-scale information can be used to increase the operational knowledge of the downstream oil and gas industry and to assist base managers in various segment exercises.

Impacts of technological advancement in downstream industry

There are many advancement and innovations took place in petroleum industry up till now. Technological advancement causes cost reduction in operations, increase in quality of output product, and reduction in negative impact of environment. Elatab (2016) researched that digitalization in oil field can cut the cost in operations up to 10–25%. Mills (2013) has researched that new hydraulic Fracturing technology helps to bring down the cost of shale oil production between $7 and $15 per barrel and efficiency and productivity of US's shale fields increased 200–300%. Bertocco and Padmanabhan (2014) have researched that advancement in big data analytics helps to enhance the oil production by 6–8%.

There were many innovations took place in petroleum refining industry. Earlier a simple distillation technique was used for the refining purpose. That process separates the crude in lighter crude, naphtha, kerosene, and heavier crude. But as the need of lighter product increased, for the profit purpose, they took new advancement in technology called thermal cracking that was used to convert the heavier crude into lighter one. That causes the increasing profit and output product quality. After thermal cracking, catalytic cracking was used for the refining purpose. After that, thermal reforming technique replaces the catalytic cracking. After that, catalytic cracking method is replaced by the catalytic reforming method. As per the study of Enos (1958), the profit per barrel of crude for thermal cracking, thermal reforming, catalytic cracking, and catalytic reforming was

$0.186, 0.038, 0.123, and 0.208 accordingly. So, we can see that the profit gradually increases as the innovation and advancement take place. As there are notable impacts of technological advancement and innovation in petroleum industry, in this paper, we are going to review or discuss about the impact of big data analytics in petroleum downstream industry. Yin (1994) developed a model called present value index to evaluate the economic advantage from radical innovations and incremental improvements. He compares the profit between earlier radical process of innovation and incremental improvements. They found that the incremental improvements gave more profits than earlier radical process.

Methodology

Like big data involves enormous amount of datasets and in some complex and complicated issues, it becomes necessary to have unique solutions and latest technologies to deal with those big data problem and issues. Technologies we are going to use should be speedy and precise processors. Below, section tools and technologies that are commonly used for big data analysis are mentioned and described.

Big data analytics tools

Apache Hadoop

This tool was developed by Doug Cutting and Mike Cafarella as an open-source framework. Hadoop is capable to process and solve the huge and complex datasets with scalable computing. This tool is made up of two major layers including Hadoop Distributed File System (HDFS) and MapReduce. It works in two phase. In first phase, data is stored under HDFS layer with its master server called name node and cluster of slaves called data nodes. In the second phase, various tasks such as tracking and executing jobs will be done under MapReduce layer. So basically the data processing and analysis in Hadoop have two phase called map phase and reduce phase (Fig. 1) (Ishwarappa and Anuradha 2015).

Job tracker is master node, and task tracker is slave node of MapReduce. MapReduce is able to handle the enormous amount of datasets using multiple clusters (Rehan and Gangodkar 2015).

Hadoop is used to track unstructured and sentiment data from various sources such as social media, geolocation, and data from sensor and machine. It is specially designed for the storage of voluminous data and to perform parallel process. Hadoop is also used to strengthen security and its compliance.

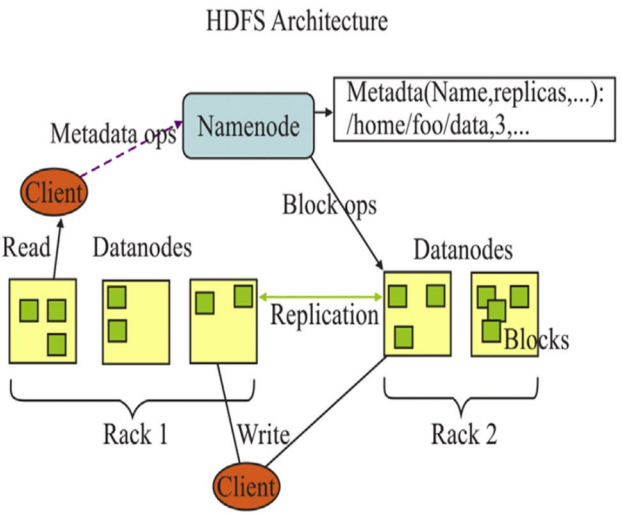

Fig. 1 HDFS architecture with name node and data nodes (Borthakur 2018)

MongoDB

MongoDB is document-oriented database technology which is written in C++ and Jason-based NoSQL technology. This technology can be able to control and handle disorganized and unformed data like documents, multi-, and social media. This technology also provides dynamic and flexible structure to be able to fit the requirement of different users. (Trifu and Ivan 2014).

MongoDB is most widely used to store primary data in web applications. MongoDB cannot be used in applications where durability, isolation, consistency, and atomicity are required such as in database-level transactions and to design system of core banking for a bank. As MongoDB is based on NoSQL data base, it should not be used in application where table joins are required.

Cassandra

Cassandra is same NoSQL Database technology as MongoDB. This technology was first a Facebook project, but now it is open-sourced technology. It is more useful where it is feasible to expend more time to study a complex system that gives much power and flexibility.

Cassandra cannot be used in transactional databases, but it is better choice if we have both unstructured and structured data and expecting rapid growth in database. Cassandra is used in streaming data, real-time analytics as well as in fraud and crime detection.

Predictive analytics methods

There are also some most widely used predictive modeling methods in big data analytics such as decision trees, multiple linear regression, k-cross-validation, and support vector machine which are illustrated below.

Multiple linear regressions

Multiple linear regression analysis is a modification of simple linear regression analysis which is used to define correlation between two or more independent and continuous single correlated variable. We can obtain a function Y using n parameters (X_1, $X_2 \ldots X_n$) with multiple linear regressions (Xie et al. 2019).

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where Y is a continuous variable, but we can write the formula in the following way when Y becomes a replica variable.

$$\ln\left(\frac{Y_i}{1 - Y_i}\right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

In this analysis, Y_i is the probability.

Here, if $\beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n > 0$, we consider it as Y_i; if $\beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n < 0$, we do not consider it as Y_i.

This can be used in permeability, porosity, and saturation prediction. We can clearly see that multiple linear regression can help the variables Y and X to be calcified. Because of this, it can help to reduce lithology data mining dimensions. This combined with decision tree k-cross-validation and vector supporting machines to help size reduction.

Decision trees

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision tree is being used as a predictive model in machine learning and data mining for mapping observations of the item's targeted value. In the tree structure, classifications are represented by the leaves and conjunctions of features are represented by the branches.

Decision tree helps to illustrate the data by identifying the most relatable variables and relations between more than one variables. Basically, it is used to categorize the data and to predict the objective like cost, profit, etc. However, it is useless for decision making in data mining.

K-cross-validation

Cross-validation is a resampling method used to evaluate models of machine learning on a restricted data sample. This method has a single parameter called k which represents the number of sets to be divided into a given data sample. The method is often referred to as k-fold cross-validation. When we select a particular value for k, it can be used in the model reference instead of k, such as $k = 5$ becoming five-fold cross-validation. Cross-validation is basically used in applied machine learning to predict and evaluate a machine learning model's ability on unseen data. That is, using a few sample to evaluate how the model is expected to perform in general when used to predict data not used during model training. K-cross-validation is easily understandable, and its less biased or less optimistic estimation of model nature made this method famous than many other methods.

Cross-validation is used specifically when a little amount of data are available as it uses all the data for the analysis. It is also used when we are working with dependent or grouped data as well as to choose the best parameters among various predictive analytics models.

Support vector machine

Support vector constructs are supervised learning models with related learning algorithms that analyze data used to analyze classification and regression. The support vector machine uses the kernel function for nonlinear cases, which places the input in higher-dimensional space and finds the appropriate classification. Similarly, in the form of the least square support vector machine, it can also be used for regression where it changes the inequality constraint to the constraint of equality. This can be seen in the optimization below:

$$\text{Objective function: } \min f(w) = \frac{w^2}{2} + \gamma 1/2 \sum_{K=1}^{N} e_k^2$$

$$\text{Constraint condition: } y_k w^T \emptyset\left(x_k\right) + b = 1 - e_k, k = 1, \ldots, N$$

By solving both the equations, we obtain the following regression

$$y(x) = \sum_{K}^{N} \propto_K K\left(x, x_k\right) + b$$

Basically, support vector machine is used in exact classification of unseen data. It is used in categorization of images, face detection, bioinformatics, generalized predictive control, handwriting identification, and protein remote homology detection. However, in petroleum industry, it can be

used to calculate porosity, saturation, and permeability. In many cases, data may have a nonlinear relationship due to complex situations in geology and logging. The support vector machines should generally be used when the nonlinearity is strong and the decision trees when the nonlinearity is not strong.

Big data analytics in refinery

In the world, major energy consuming sectors include petroleum refining sector with 4% of total global primary energy consumption. So, increasing energy efficiency of refineries becomes important option to reduce the greenhouse gas emissions and reduction in energy consumption by refining industry.

Ghaderi (2008) researched and reviewed two models for energy analysis in refinery. The first is data envelopment analysis (DEA), and the second is principal component analysis (PCA).

Data envelopment analysis

DEA method is used to estimate the efficiency in a given data set of decision-making units (DMUs). This type of model has two types: One is input oriented, and other one can be output oriented. For the refinery, we are using input-oriented DEA models because we want to evaluate the efficiency of refineries under a fixed structure. The input-oriented model of DEA gives efficiency scores of refineries that helps us to evaluate how efficient we can use electricity and fuels to produce refinery products.

There are various models in DEA such as CCR model and Andersen and Petersen model. CCR model is not capable enough to rank efficient units as it assigns a common index to all decision-making units. So generally, Andersen and Petersen model is used in refinery sector. AP model is much more efficient if the efficiency score of that decision-making unit is equal or greater than one.

Principal component analysis

PCA is most likely to use in multivariate statistics like factor analysis. Principal component analysis is done by reducing the number of variables which are in under study and ranking and analysis of decision-making units. It is used in various industries, medical facilities, towns, etc. PCA produces outputs from different inputs derived from various sources. Some of the examples of using PCA for the analysis are illustrated below.

Tokarek et al. (2018) used PCA analysis to measure air pollutants of Athabasca oil sands. In this analysis, they used varimax rotation along with PCA to illuminate 28 variables including volatile organic compounds (VOCs) and intermediate-volatility organic compounds (IVOCs), although ten of them were spotted and classified based on different kind of source.

Zhu (1998) evaluated and ranked some cities of China based on their economic performance, in which Zhu used both DEA (non-statistical method) and PCA (multivariate statistical method) analysis for performance evaluation of cities. As result, Zhu found consistency between the evaluation ranking of cities from both PCA and DEA method.

Thus, energy efficiency in refinery sector is analyzed by using DEA approach. Then, DEA is verified and validated by the PCA approach. These models applied to the Iran and OECD countries, in which they found more potential reduction in fossil fuel consumption than potential reduction in electricity.

Big data can be also used in prognostic foresight. Prognostic analytics methods helps to analyze the data more far beyond insight and hindsight for forecasting than predictive analytics methods. Prognostic analytics method helps refinery in various areas such as maintenance and repair, operations, finance, and life cycle management.

Prognostics analytics help refinery in maintenance and repair by long-term and short-term scheduling of maintenance and maintenance staff planning and allocation. In the operation section of refinery, prognostic analytics helps for production planning based on to the future availability profile to eliminate the downtime risk for field projects. As well as in finance section, it helps to decrease the maintenance costs, increase benefits from petroleum operators and also covers the insurance policy and costs. Last and important section is life cycle management which covers remaining useful life (RUL)—optimal exploitation, replacement, and retrofit planning.

As per one case study, petrochemical manufacturer wanted to utilize abundant condition and process data histories to better exploit equipment RUL, so that customer chooses Cassantec prognostic solution and applied it on four-stage cracked gas compressor. The main benefits of using prognostic solution were reduction in downtime cost and reduction in maintenance cost of the equipment (Plate and Ag 2018).

Big data analytics in natural gas industry

Big data is now covering almost every industry also including LNG and CGD industry. Natural gas industry includes the liquefaction and regasification of natural gas, transportation of LNG and NG through ship and pipeline, and distribution. How big data helping natural gas industry is as below.

Liquefied natural gas plants

As per one of case study, General Electric Company used diagnostic analytics on its gas turbines which were used in LNG liquefaction and regasification plant (Fig. 2). The major

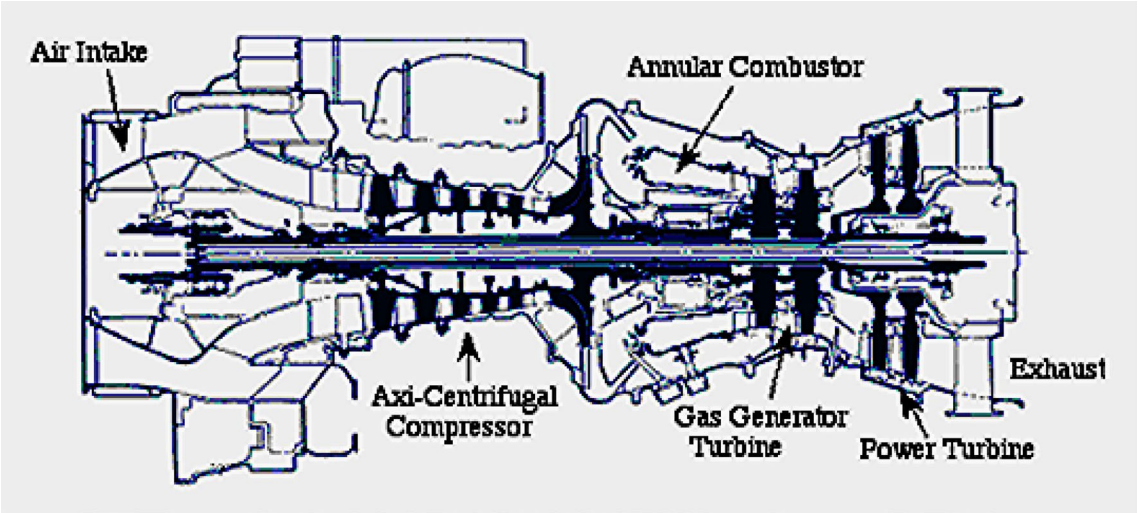

Fig. 2 Gas turbine cross section

factors to improve the efficiency of plant work are uptime, downtime, and maintenance. Big data helps to increase profitability gas turbine to maximize the uptime and minimize the downtime. The maintenance plays an important role in profitability of any industry (Fig. 3). Sometime unexpected maintenance came out in equipment and stops the production flow and disrupts the whole cycle in plant. Big data helps to avoid maintenance shutdowns by analyzing the data to predict the necessity of maintenance and its time cycle.

City gas distribution sector

In future, CGD will play an important role in world energy basket. As per one survey, it is expected around 20% share of natural gas in India's energy basket by 2025. The usage of analytics in CGD sector is lower than other industry. So to complete the target of CGD expansion in India, CGD companies should have to use holistic approach by embedding analytics in all possible business processes. Analytics in CGD sector will help to improve the operations and address the issue in PNG and CNG and improve infrastructure and management of assets.

Storage and transportation

Many researches show that we can improve the transportation and shipping performance by applying big data analytics methods on it. In research of Anagnostopoulos (2018), he shows that propulsion power can help to improve the ship performance and also can reduce emission of greenhouse gases. In this study, they conducted data analytics by using eXtreme Gradient Boosting (XGBoost) and multilayer perception (MLP) neural networks methods. Multilayer perception is updated version of scikit-learn Python Library's neural network models (big data technique). The data in this study were taken over three months of period from the sensor throughout a Large Car Truck Carrier M/V.

Efficiency of natural gas pipeline is measured by the changes that occur in energy and volume of gas. Mu-wei et al.(2019) developed data envelop analysis model to analyze the variation of input and output energy and volume and to determine the efficiency of the natural gas pipeline. They analyzed that the growth of efficiency is inversely proportional to the amount of natural gas transmission.

Fig. 3 GE big data diagnostics process for the LNG liquefaction industry segment

They conclude that transmission and economic efficiency consists trade-off relationship between them.

As per Wang et al.'s(2017) study, they carried out a research on the storage equipment to measure and explore the risk-level and forewarning management from the data modeling and mining as shown in Fig. 4. They carried research on various models such as K-means cluster analysis and logical regression fitting. The figure shows the specific process used in this research, in which they found linear regression model best fits in risk assessment of oil and gas storage. Apart from that they got idea of online as well as intelligent forewarning management information system, which helps in management that includes big data monitoring, intelligent assessment, automatically forewarning, and the advance of contingency.

Big data analytics in health and safety

Nowadays, big data analytics are used by many companies for their health and safety purpose. For example, shell company now uses predictive data analytics to identify potential work hazards in their field operations. Ajayi et al. (2019) developed a model for the health hazard analytics, which consists six stages. First is the data preparation, which identifies and fixes up the errors within the data. Second is exploratory analytics and model selection in which the suitable analytical model picked up for the particular datasets. In the third stage, analytical models are designed to identify and evaluate the health and safety. So as, in the fourth stage, parameter extraction and model execution are done. Then, after predictive analysis, health forecasting is a fifth stage and prescriptive analysis is sixth stage of this proposed model.

Tanabe et al. (2010) proposed a safety model design for the onshore natural gas liquefaction plant. This model consists of the safety critical design basis matrix to identify the external common cause of failure more accurately. This model provides shield to the liquefied natural gas cryogenic spill. This model maximizes the natural air circulation to prevent the plant accidents.

In study of Xie et al. (2019), they proposed a structure framework to detect the factor which influences and predicts the failure rate of SIS equipment. This proposed framework includes statistical models and data-driven model such as PCA and PLSR to predict the failure of rates. In this case, they applied the framework on the shutdown valves at six oil and gas facilities and they found that the size and medium are the main factors for the safety influencing factor (Fig. 5).

In the study of Cadei et al. (2018), they used the H_2S concentration to predict the hazard event and developed the prediction software to forecast operational upsets during oil and gas production and hazard events. They took the data from different sources such as real-time stories, historical data, operator data, and maintenance reports, and they processed the data by modeling (using artificial neural network) and model validation.

Imanian et al. (2018) researched about the method which includes statistical process control and engineering process control to detect, observe, and control the major causes including bottomhole pressure while drilling operations. However, SPC uses data-driven control charts and statistical

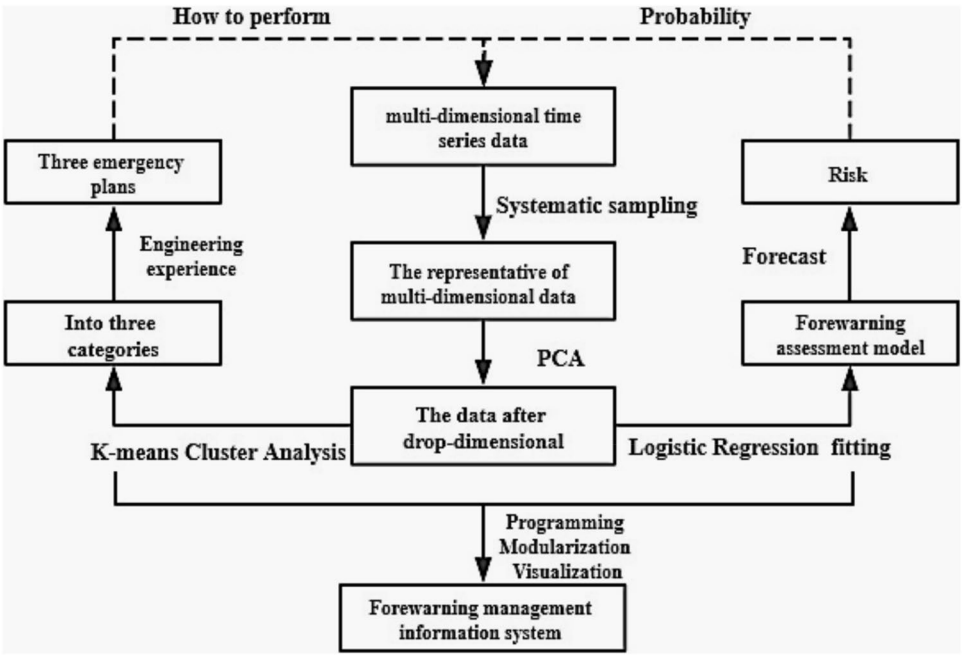

Fig. 4 Technical framework map (Wang et al. 2017)

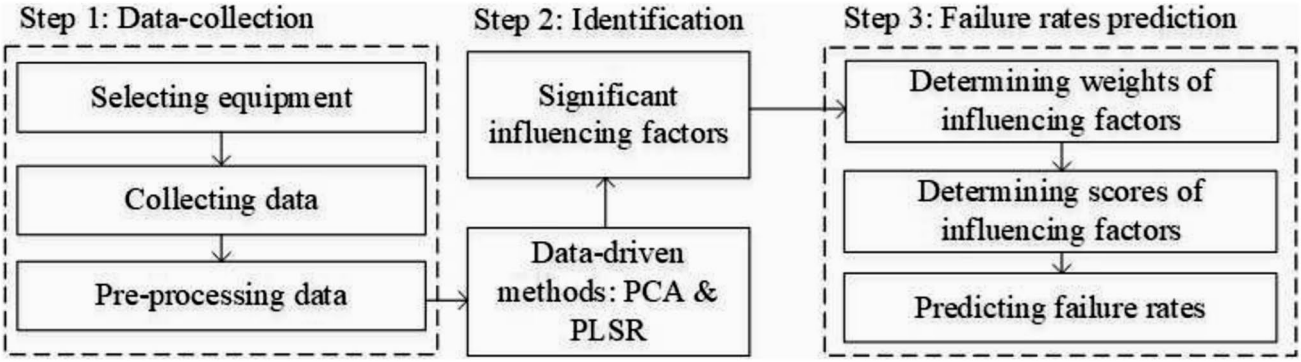

Fig. 5 Framework for predicting failure rates (Jani et al. 2019)

methods to analyze real-time situation, mainly surge and swab operations as well.

Big data analytics in future markets

Yu et al. (2019) have researched various factors and trends, which were related to oil consumption prediction. In this research paper, they showed that how a Google trends can help to predict the future oil consumption through a big data-driven forecasting model. They developed a forecasting model driven based on big data for improvement in prediction of future oil consumption. The model has two main big steps: First is relationship investigation, and second is prediction improvement. To investigate relationship between Google trends and oil consumption, cointegration test and a Granger analysis were carried out. Then in the both statistical and artificial intelligence (AI) model, the selected Google trends were applied to improve the assumption of oil utilization (Fig. 6).

In the study of Panja et al. (2018), they have researched about the methods that help us to predict the future oil and gas production from shale using the artificial intelligence. Future production will give us idea about the future markets of oil and gas. They proposed three types of surrogate models a response surface model (RSM), a least square support vector machine (LSSVM) model, and an artificial neural networks (ANNs) model. These models were developed under time-based (90 days, 1 year, 5 years, 10 years, and 15 years) and rate-based constraints (5 bbl/day/fracture) and compared in this study. As the result of the study, they found that LSSVM and RSM have better capability to predict the future oil and gas production than the ANN model. So, by using, we can indirectly forecast the oil and gas markets and manage the demand and supply of hydrocarbon (Fig. 7).

In the study of Nazari et al. (2019), they proposed new method to predict and evaluate the economic growth by using the panel autoregressive distributed lag model

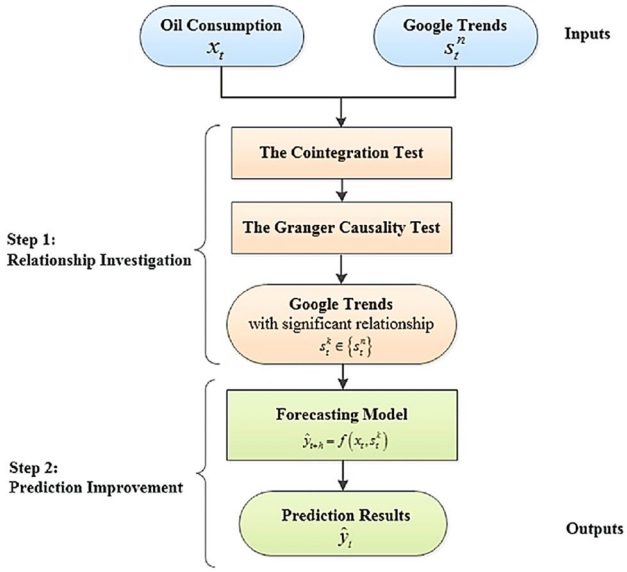

Fig. 6 General framework of an online big data-driven forecasting model using Google trends for oil consumption (Pandya et al. 2020)

(ARDL) in the occurrence of unreliability about production and oil revenues. This model consists of pooled mean group (PMG) and mean group (MG) estimation methods in case of both with and without uncertainty. ARDL is data-driven model which uses data analytics and regression equations to predict the future consumption of energy with help of current and past energy consumption data.

Big data analytics challenges

One of the main problems or challenges of using big data in various industries including petroleum industry is to handle high amount of cost in data recording, data storage, and data analysis. Based on the research of Mounir et al. (2018) and Beckwith (Beckwith 2011), all these issues

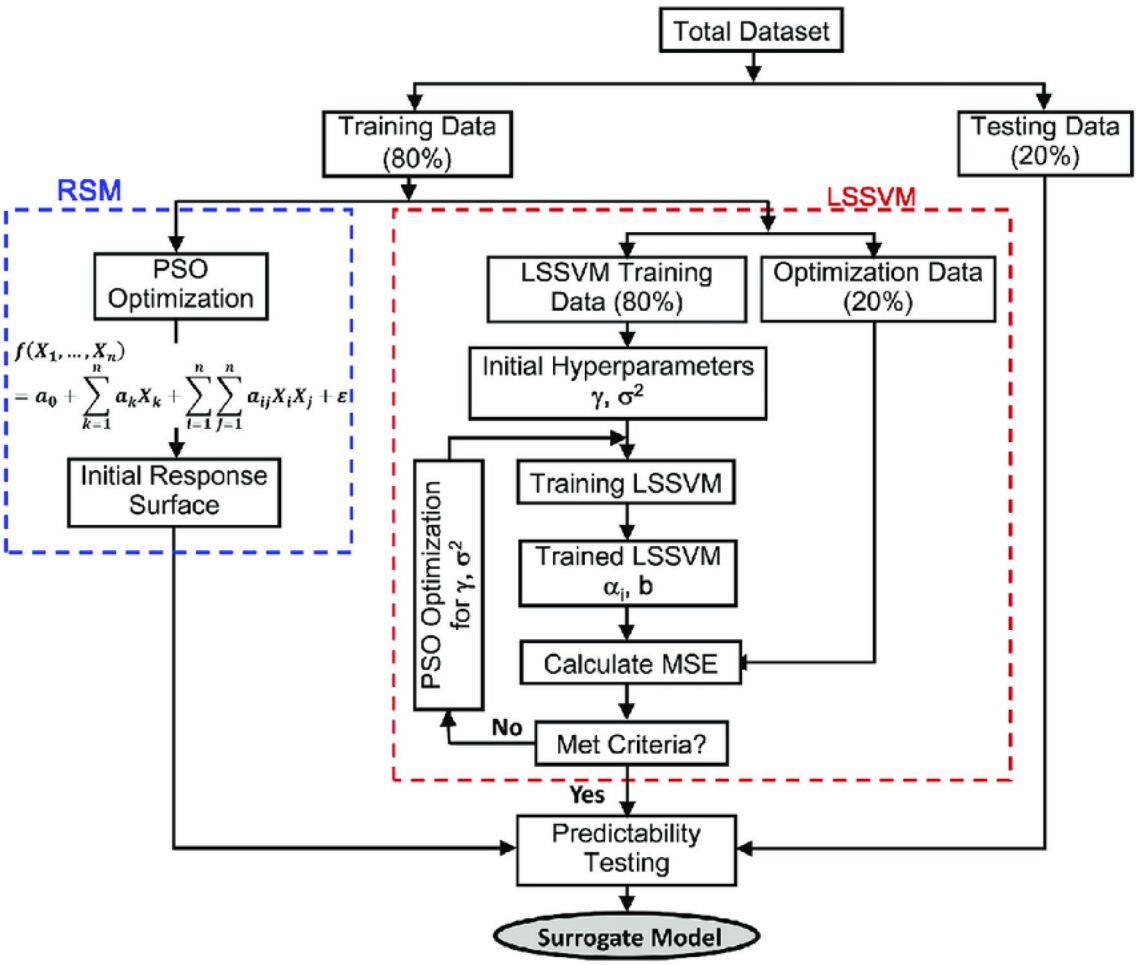

Fig. 7 Workflow used to develop RSM and LSSVM (Sukhadia et al. 2020)

can be resolved by the recent advancement in technologies such as cloud computing, Internet of things, and fog computing to manage, analyze, and store the big crucial and seismic data in petroleum industry.

Based on the survey of Feblowitz (2013), the main problem of using or applying the big data technologies in petroleum and energy sector is lack of knowledge awareness, skilled professionals, and cost of big data technologies. So, first, we have to aware the staffs of petroleum and energy-related companies with the big data and its applications and technologies to successfully implement the big data in petroleum and energy sector.

Maidla et al. (2018) researched that due to the poor quality and inaccurately aggregated and classified data, data mining was wrongly analyzed. Another drawback was the accuracy in recording the data by the data recording sensors. To understand and evaluate the physics of data was also one of the issues, which can be solved by the collaboration between the data scientist and petroleum engineers.

In the study of Preveral and Petit (2014), they suggest individual company to invent their own big data tools, data transferring, data mining, and storage facilities which will help them in cost cutting of software ownership.

Conclusion

This paper gives a complete review which includes application of big data analysis in petroleum downstream industry like refineries, city gas distribution, liquefied natural gas, oil and gas storage, transportation etc. This paper shows how various big data technologies like Apache Hadoop, MongoDB, Cassandra, and predictive mathematical modeling such as linear regression, decision trees, K-cross-validation, and support vector machine can be effectively used in petroleum downstream industry. Refineries are using PCA and DEA models and prognostic foresight models for the energy analysis in plants. Big data helps refinery in reduction in electricity and fossil fuel consumption of plant and

many other areas of it like maintenance, repair, etc. In the LNG and CNG, Big data helps to predict maintenance to avoid unwanted shutdowns, operations, address issues and improve infrastructure. Big data helps in storage of oil and gas as well as in transportation of it. It helps us via improving the transportation ship performance and evaluating the risk and forewarning system of petroleum product's storages. Recent advancements in big data help in health and safety to increase the hazard prediction by using various safety models, predictive analysis, PCA, PLSR, and H2S concentration to detect the hazard and failure rates. Using the models such as ANN, RSM, LSSVM, statistical models, and AI model, we can predict the future trends and markets of oil and gas. Despite of having many applications of big data in petroleum downstream, there are several challenges related to cost and lack of knowledge and professional experience in the big data field made us to resolve those challenges for the future advancement.

Acknowledgements The authors are grateful to S & P Global, School of Petroleum technology and Department of Chemical Engineering, School of Technology, Pandit Deendayal Petroleum University, Petrowatch for permission to publish this research.

Authors Contribution All the authors make substantial contribution in this manuscript. DD, SP and MS participated in drafting the manuscript. DD wrote the main manuscript, all the authors discussed the results and implication on the manuscript at all stages.

Availability of data and material All relevant data and material are presented in the main paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

Ahir K, Govani K, Gajera R, Shah M (2020) Application on virtual reality for enhanced education learning, military training and sports. Augm Hum Res 5:7

Ajayi A, Oyedele L, Delgado JMD, Akanbi L, Bilal M, Akinade O, Olawale O (2019) Big data platform for health and safety accident prediction. World J Sci Technol Sustain Dev 16(1):2–21. https://doi.org/10.1108/WJSTSD-05-2018-0042

Anagnostopoulos A (2018) Big data techniques for ship performance study. In: Proceedings of the 28th international ocean and polar engineering conference, pp 887–893

Beckwith R (2011) Managing big data: cloud computing. J Pet Technol 63:42–45. https://doi.org/10.2118/1011-0042-JPT

Bertocco R, Padmanabhan V (2014) Big data analytics in oil and gas: converting the promise into value. http://www.bain.com/Images/BAIN_BRIEF_Big_Data_analytics_in_oil_and_gas.pdf. Accessed 11 Oct 2016

Borthakur D (2018) HDFS Design. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. Accessed 7 Aug 2018

Cadei L, Montini M, Landi F, Porcelli F, Michetti V, Origgi M, Tonegutti M, Duranton S (2018) Big data advanced anlytics to forecast operational upsets in upstream production system. Abu Dhabi Int Pet Exhib Conf Soc Pet Eng Abu Dhabi 1:1–14. https://doi.org/10.2118/193190-MS

Elatab M (2016) 5 Trends in oil & gas technology, and why you should care. http://venturebeat.com/2012/03/28/5-trends-in-oil-gas-technologyand-why-you-should-care/. Accessed 18 Nov 2019

Enos JL (1958) A measure of the rate of technological progress in the petroleum refining industry. J Ind Econ 6(3):180–197

Feblowitz J (2013) Insights IDCE, Analytics in oil and gas: the big deal about big data, pp 5–7. http://refhub.elsevier.com/S2405-6561(18)30142-1/sref12

Gandhi M, Kamdar J, Shah M (2020) Preprocessing of non-symmetrical images for edge detection. Augment Hum Res 5:10. https://doi.org/10.1007/s41133-019-0030-5

Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iview 1142:9–10

Ghaderi SF (2008) Energy efficiency modeling and estimation in petroleum refining industry—a comparison using physical data 1(6):123–128

Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of "big data" on cloud computing: Review and open research issues. Inf Syst 47:98–115

Ifaei P, Farid A, Yoo C (2018) An optimal renewable energy management strategy with and without hydropower using a factor weighted multi-criteria decision making analysis and nation-wide big data—Case study in Iran. Energy 158:357–372. https://doi.org/10.1016/j.energy.2018.06.043

Imanian Mahdi, Ghassemi Aazam, Karbasian Mahdi (2018) Monitoring and control of bottomhole pressure during surge and swab operations using statistical process control. Energy Sources Part A Recov Util Environ Effects 1:1. https://doi.org/10.1080/15567036.2018.1464613

Ishwarappa J, Anuradha J (2015) A brief introduction on big data 5Vs characteristics and Hadoop technology. Procedia Comput Sci 48(2015):319–324

Jani K, Chaudhuri M, Patel H, Shah M (2019) Machine learning in films: an approach towards automation in film censoring. J Data Inf Manag 1:1. https://doi.org/10.1007/s42488-019-00016-9

Jha K, Doshi A, Patel P, Shah M (2019) A comprehensive review on automation in agriculture using artificial intelligence. Artif Intell Agric 2:1–12

Kakkad V, Patel M, Shah M (2019) Biometric authentication and image encryption for image security in cloud framework. Multiscale Multidiscip Model Exp Des 1:1–16. https://doi.org/10.1007/s41939-019-00049-y

Kundalia K, Patel Y, Shah M (2020) Multi-label movie genre detection from a movie poster using knowledge transfer learning. Augment Hum Res 5:11. https://doi.org/10.1007/s41133-019-0029-y

Maidla WM, Rigg J, Crumrine M, Wolf-zoellner P (2018) Drilling analysis using Big data has been misused and abused. In: IADC/SPE Drill. Conf. Exhib., Fort Worth. https://doi.org/10.2118/189583-MS

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, McKinsey, Global Institute (2011) Big data: the next frontier for innovation, competition, and productivity, pp 1–156

Marinakis V, Doukas H, Tsapelas J, Mouzakitis S (2018) From big data to smart energy services: an application for intelligent energy management. Fut Gen Comput Syst 1:15. https://doi.org/10.1016/j.future.2018.04.062

Mills MP (2013) Big Data and microseismic imaging will accelerate the smart drilling oil and gas revolution. viahttp://www.forbes.com/sites/markpmills/2013/05/08/big-data-and-microseismic-imaging-will-accelerate-the-smartdrilling-oil-and-gas-revolution/#6f6e6e9548b2. Accessed 18 Nov 2019

Mohammadpoor M, Torabi F (2018) Big data analytics in oil and gas industry: an emerging trend. Petroleum 1:1–8. https://doi.org/10.1016/j.petlm.2018.11.001

Mounir N, Guo Y, Panchal Y, Mohamed IM, Abou-sayed A, Abou-Sayed O (2018) Integrating Big Data: simulation, predictive analytics, real time monitoring, and data warehousing in a single cloud application. In: Offshore technology conference, pp 1–14. https://www.onepetro.org/conference-paper/OTC-28910-MS

Mu-wei F, Chu-chu A, Xiao-rong W (2019) Comprehensive method of natural gas pipeline efficiency evaluation based on energy and big data analysis. Energy 188:116069. https://doi.org/10.1016/j.energy.2019.116069

Nazari M, Asadi E, Imanian M (2019) Uncertainty, budget deficit and economic growth in OPEC member countries. Energy Sources Part A Recov Util Environ Effects 1:1. https://doi.org/10.1080/15567036.2019.1668510

Pandya R, Nadiadwala S, Shah R, Shah M (2020) Buildout of methodology for meticulous diagnosis of K-complex in EEG for aiding the detection of Alzheimer's by artificial intelligence. Augm Hum Res. https://doi.org/10.1007/s41133-019-0021-6

Panja P, Velasco R, Pathak M, Deo M (2018) Application of artificial intelligence to forecast hydrocarbon production from shales. Petroleum 4(1):75–89. https://doi.org/10.1016/j.petlm.2017.11.003

Parekh V, Shah D, Shah M (2020) Fatigue detection using artificial intelligence framework. Augm Hum Res 5:5

Patel D, Shah Y, Thakkar N, Shah K, Shah M (2020a) Implementation of artificial intelligence techniques for cancer detection. Augm Hum Res 5(1):1. https://doi.org/10.1007/s41133-019-0024-3

Patel D, Shah D, Shah M (2020b) The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports. Ann Data Sci 1:1. https://doi.org/10.1007/s40745-019-00239-y

Plate MV, Ag C (2018) SPE-181037-MS big data analytics for prognostic foresight new dimension of petroleum asset management, pp 6–8

Preveral AT, Petit N (2014) Geographically-distributed Databases: a big data technology for production analysis in the oil & gas industry. In: SPE Intell. Energy Conf. Exhib, Society of Petroleum Engineers, Utrecht, pp 1–9. https://www.onepetro.org/conference-paper/SPE-167844-MS

Rehan M, Gangodkar D (2015) Hadoop, MapReduce and HDFS: a developers perspective. Procedia Procedia Comput Sci 48:45–50. https://doi.org/10.1016/j.procs

Shah G, Shah A, Shah M (2019) Panacea of challenges in real-world application of big data analytics in healthcare sector. Data Inf Mana 1(3–4):107–116. https://doi.org/10.1007/s42488-019-00010-1

Shah D, Dixit R, Shah A, Shah P, Shah M (2020a) A comprehensive analysis regarding several breakthroughs based on computer intelligence targeting various syndromes. Augm Hum Res 5:14. https://doi.org/10.1007/s41133-020-00033-z

Shah K, Patel H, Sanghvi D, Shah M (2020b) A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augm Hum Res 5:12. https://doi.org/10.1007/s41133-020-00032-0

Sukhadia A, Upadhyay K, Gundeti M, Shah S, Shah M (2020) Optimization of smart traffic governance system using artificial intelligence. Augm Hum Res 5:13. https://doi.org/10.1007/s41133-020-00035-x

Tanabe M, Miyake A (2010) Safety design approach for onshore modularized LNG liquefaction plant. J Loss Prev Process Ind 23(4):507–514. https://doi.org/10.1016/j.jlp.2010.04.004

Tokarek TW, Odame-Ankrah CA, Huo JA, McLaren R, Lee AKY, Adam MG, Willis MD, Abbatt JPD, Mihele C, Darlington A, Mittermeier RL, Strawbridge K, Hayden KL, Olfert JS, Schnitzler EG, Brownsey DK, Assad FV, Wentworth GR, Tevlin AG, Worthy DEJ, Li S-M, Liggio J, Brook JR, Osthoff HD (2018) Principal component analysis of summertime ground site measurements in the Athabasca oil sands with a focus on analytically unresolved intermediate-volatility organic compounds. Atmos Chem Phys 18:17819–17841. https://doi.org/10.5194/acp-18-17819-2018

Trifu MR, Ivan ML (2014) Big data: present and future. Data Syst J 5:32–41

Wang T, Li T, Xia Y, Zhang Z, Jin S (2017) Risk assessment and online forewarning of oil & gas storage and transportation facilities based on data mining. Procedia Comput Sci 112:1945–1953. https://doi.org/10.1016/j.procs.2017.08.052

Xie L, Håbrekke S, Liu Y, Lundteigen MA (2019) Operational data-driven prediction for failure rates of equipment in safety instrumented systems: A case study from the oil and gas industry. J Loss Prev Process Ind 60:96–105. https://doi.org/10.1016/j.jlp.2019.04.004

Yin JZ (1994) Managing process innovation through incremental improvements: Empirical evidence in the petroleum refining industry. Technol Forecast Soc Change 47(3):265–276. https://doi.org/10.1016/0040-1625(94)90068-X

Yu L, Zhao Y, Tang L, Yang Z (2019) Online big data-driven oil consumption forecasting with Google trends. Int J Forecast 35(1):213–223. https://doi.org/10.1016/j.ijforecast.2017.11.005

Zhu Joe (1998) Data envelopment analysis vs principal component analysis: an illustrative study of economic performance of Chinese cities. Eur J Oper Res 111(1):50–61. https://doi.org/10.1016/S0377-2217(97)00321-4