Big Data Analytics in Oil and Gas: An Empirical Study

Santosh Kumar Sahu, Oil and Natural Gas Corporation Limited
Vanshika, Mody University, Laxmangarh, Rajastha
Suman K Batra , Mody University , Laxmangarh, Rajasthan
Sanjay Kumar Jena, National Institute of Technology, Rourkela

sahu_santosh@ongc.co.in

Keywords

Big Data Analytics, Predictive Analysis, Machine Learning on Oil & Gas Data

Summary

As data-intensive decision making is being increasingly adopted by businesses, governments, and other agencies around the world. Now-a-days, it's a big challenge to acquire, process, and interpret the extensively large and variety of data. Big Data Analytic Techniques is used to deal with high Volume, Variety and Velocity of data. It effectively stores, process and visualize the data in an effective manner. Drawing upon implementation experiences of early adopters of Big Data technologies across multiple industries, this paper focuses the issues and challenges faced during data analytics by oil and gas companies. It also described the data flow from raw to processed data format used for predictive analysis. The most advanced scalable machine learning approaches are applied and visualized using Apache Spark and Mahout. In our experiment, intrusion datasets are used as input data due to lack availability of public data related to Oil and Gas. The processing of data using Spark and visualization are same for other data set. To assess the predictive model, the popular model evaluation parameters are calculated and depicted.

Introduction

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates an enterprise on a day-to-day basis. However, it's now not the quantity of facts that's vital. It's what agencies do with the facts that matters. Big data facts can be analyzed for insights that cause higher decisions and strategic enterprise movements.
While the term "large information" is particularly new, the act of gathering and storing large amounts of facts for eventual analysis is ages antique. The concept gained momentum inside the early 2000s while industry analysts articulated the now-mainstream definition of large statistics as the three Vs:
Volume – agencies collect facts from an expansion of assets, including enterprise transactions, social media and information from sensor or system-to-device records. In the past, storing it would've been a hassle – however new technology (such as Hadoop) has eased the load.
Velocity– records streams in an unprecedented manner of speed and need to be treated in a timely manner. RFID tags, sensors and smart metering are driving the want to deal with torrents of records in near-actual time.
Variety – records are available in all kinds of codes – from structured, numeric information in conventional databases to unstructured text files, email, video, audio, stock ticker facts and monetary transactions.
Big data technologies are vital in supplying more correct analysis, which may additionally cause extra concrete decision-making resulting in greater operational efficiencies, cost discounts, and decreased risks for the commercial enterprise.
To harness the energy of huge facts, we require an infrastructure that could manage and manner huge volumes of structured and unstructured information in real-time and can maintain security and protection.

Oil industry

Big data is emerging era in oil and fuel enterprise. In oil and fuel enterprise, drilling, exploration, preservation, manufacturing most of these activities produce considerable quantity of statistics. Currently it has been too sizeable to manner efficiently also these industries are going through fundamental demanding situations, the prices of extraction are rising and the turbulent state of international politics provides to the problems of exploration and drilling for new reserves.

With the proper technology answers, those businesses can circulate beyond conventional real-time monitoring to greater agile actual-time prediction. through swiftly reading incoming technical and commercial enterprise information—and applying that facts to complex fashions in actual time they can generate tactical insights that assist increase drilling and production performance at the same time as preventing issues. by means of fast searching and analyzing a huge quantity and type of aggressive intelligence, including information approximately mergers, acquisitions or new investments, they can appreciably improve strategic choice making.

Literature survey

Drilling

While identifying and selecting ability oil deposits and capability drill sites with seismic video display units and other device, large amounts of generated statistics require a successful platform for storage and processing. This information is beneficial when looking for, preparing, and making choices about drilling web sites, because it has direct impact on drill fulfilment, safety, and charges. Hadoop can be used as an agency records hub (EDH) for storing and processing seismic statistics, properly statistics, industry news, climate, soil, and system statistics, and is a extra cost-effective solution than conventional legacy structures.

Interpretation of seismic data

Seismic facts is also hired in manufacturing for reservoir capturing, creating 3D models of the reservoir in subsurface. Simulations are then executed to assess how much oil have to be produced in a nicely manner, once in a while methods are altered to increase yield and meet the determined forecast. Hadoop can help manufacturing engineers use the seismic and manufacturing information to optimize output.

Reservoir engineering

Oil businesses require some technology to examine the oil availability and reserves before they make investments sources into drilling by integrating real-time statistics into mechanical earth models, they may be able to analyze the statistics saved on Hadoop, and increase more secure and extra sustainable drilling strategies.

Environment safety

By using the use of big data information from numerous resources anomalies in drilling can be diagnosed in actual time. So problems may be solved before they end up very severe and drills may be close down proactively before any environmental dangers. This statistic can be used to increase environmental fitness and protection of oil rigs and drills through identity of styles and outliers earlier than any catastrophic incidents take area.

Advanced Analytics

Making plans and forecasting of production statistics, records generated from nicely and set of information produced throughout seismic acquisitions are examples of superior analytics. So essentially superior analytics helps in enhancing exploration and drilling efforts.

Data input process

Like in oil industry more data from the better geophone sensors, better algorithms and more compute power to change the oil world forever.

Even with the 3D seismic acquisition techniques, drilling still requires proper information about precisely what the head of the drill is chewing through.

Cheaper, faster and better is a familiar trajectory. Seismic richer data streams enable the emergence of better algorithms and better dynamic geophysics models that accommodate the complex mix of plastic and brittle behavior of rocks in subsurface.

This will play directly into big data analytics where correlations become possible now across the entire suite of hydrocarbon information- surface survey maps, drill-bit sensors, pressures etc. Better and richer data flows will drive the emergence of new big data analytics. The Figure 1 depicts the data flow diagram of our experiment.

such determination of drilling coordinates through oil shale to optimize the number of wellheads needed for efficient extraction of oil, optimization of drilling resources by not over drilling well site, reducing waste of drilling exploration wells, etc.

So, for inputting our data we will use Scalable Machine learning approach using spark or mahout technology to handle big data. For inputting our data, we used Scalable Machine learning approach using Spark or Mahout technology to handle big data for this we have to first prepare our data.

Preparing of data involves

• **Data Cleaning-** It includes removing of noise and unwanted data.

• **Data Transformation and Reduction** -The data can be transformed in appropriate form so that we can easily apply our algorithm on it.

It can be done in 2 ways:

Normalization The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range.

Generalization The data can also be transformed by generalizing it to the higher concept. For this purpose, we can use the concept hierarchies.

After all these steps we input our data apply predictive analysis on it and get the required output.

Due to unavailability of public well data we have considered the basic intrusion data set in which there are two class labels i.e. Attack and Normal.

Result Discussion

The proposed approach using scalable machine learning by combining supervised (Support Vector Machine) as well as unsupervised (K-Means) is given in Figure 2. The K-Means algorithm first form two clusters (as our dataset contains two class labels) and then the second supervised approach predict the labels of the clusters. The combination approach provides a stable and better result as compare to their individual's performance. The time taken for the experiment also less. Because the SVM only predicts by taking three instances of each cluster and then it labels the cluster either as normal or as attack.

To evaluate the performance of the model, we have obtained the confusion matrix of the datasets.

Table 1: Confusion matrix of NSLKDD Train Dataset

	1	-1
1	TP=5967	FN=692
-1	FP=105	TN=19031
Accuracy=**96.91**		

Table 2: Confusion matrix of NSLKDD Test Dataset

	1	-1
1	TP=2492	FN=76
-1	FP=21	TN=2162
Accuracy=**97.95**		

By considering the Table 1 and 2, we will calculate the accuracy of the model on the two training and testing dataset. During the training and testing, the ROC curve also generated and given in Figure 3.

Conclusions

The oil & fuel area is transitioning to a datacentric enterprise. While big data still needs to prove its effectiveness in oil & gas, the industry is starting to understand its capability. Hadoop is a complete information storage and processing energy lets in oil and fuel agencies the capacity to carry out risk detection, platform synchronization, and pattern analysis to reduce mistakes, streamline protection protocol and permit equipment for properly use and maximum performance. big information technology has been a beneficial tool in making oil drilling and production procedures extra green, safe, and has resulted in a sizeable boom in manufacturing. This no longer handiest helps our national economic system more sustainable, but it helps decrease dependence on different international locations for fuel and oil. As greater records is accrued, greater patterns are decided, the oil pulling process improves. This could lessen tax and finances spent on uploading fuel and oil.

References

S. K. Sahu, S. Sarangi and S. K. Jena, "A detail analysis on intrusion detection datasets," 2014 IEEE

International Advance Computing Conference (IACC), Gurgaon, 2014, pp. 1348-1353.

Waller, M. A. and Fawcett, S. E. (2013), Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. J Bus Logist, 34: 77–84.

Bertocco, Riccardo, and Vishy Padmanabhan. "Big data analytics in oil and gas." (2016).

Schwartz, Norbert, and G. L. Clore. "How do I feel about it." The informative function of (1988).

Holdaway, Keith. Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models. John Wiley & Sons, 2014.

Hems, Adam, Adil Soofi, and Ernie Perez. "How innovative oil and gas companies are using big data to outmaneuver the competition." Microsoft White Pap. (2013).

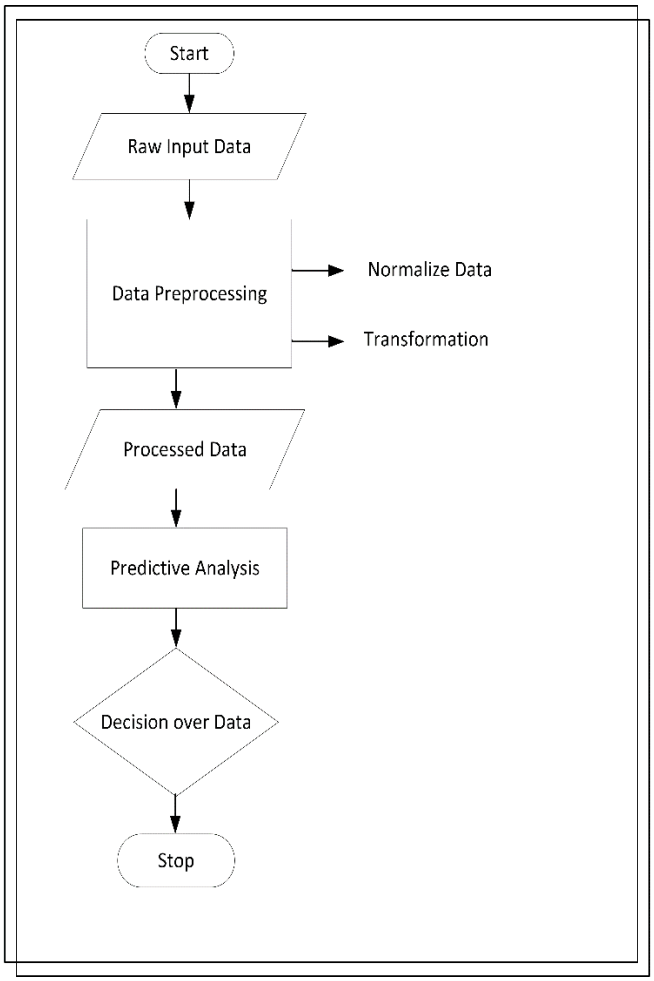

Figure 1: Data flow diagram of the proposed approach

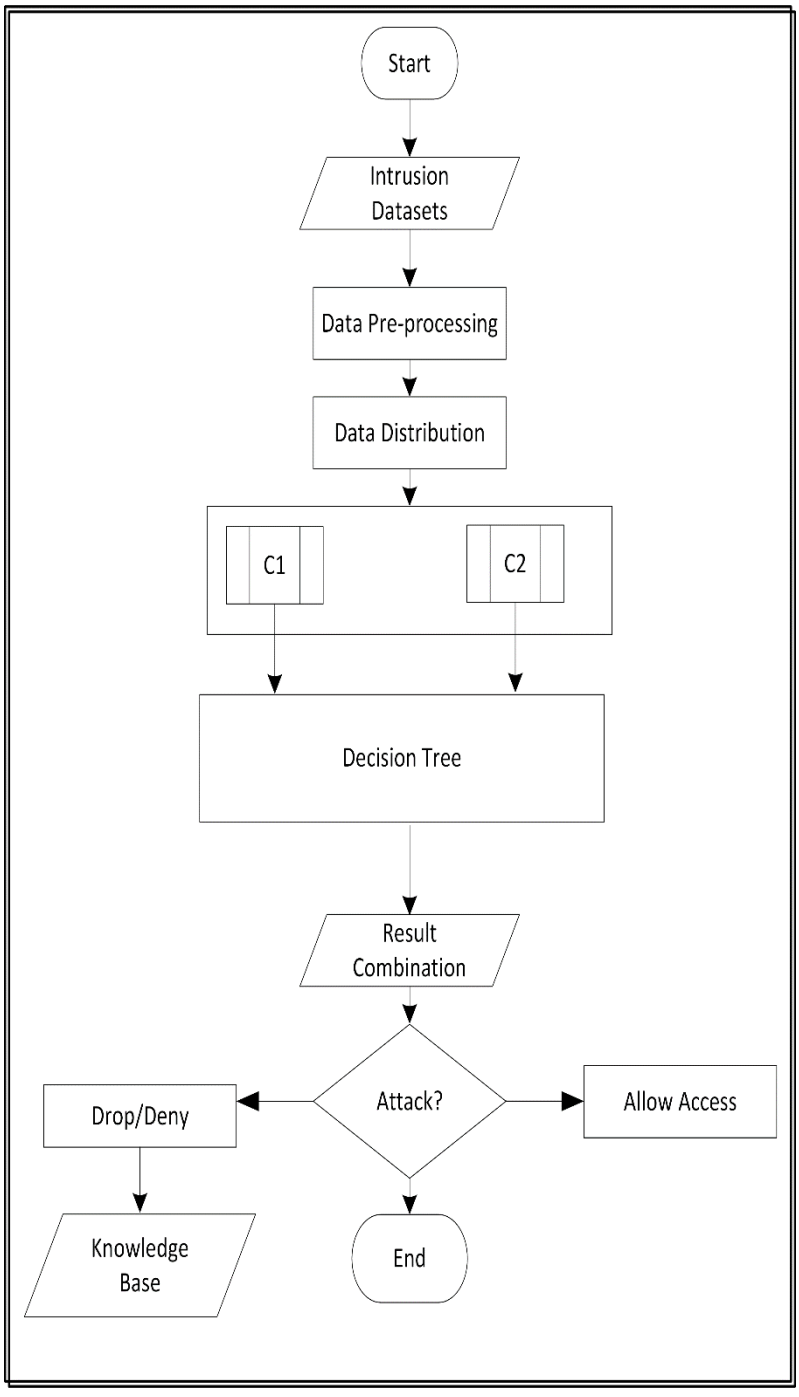

Figure 2: Proposed ensemble approach

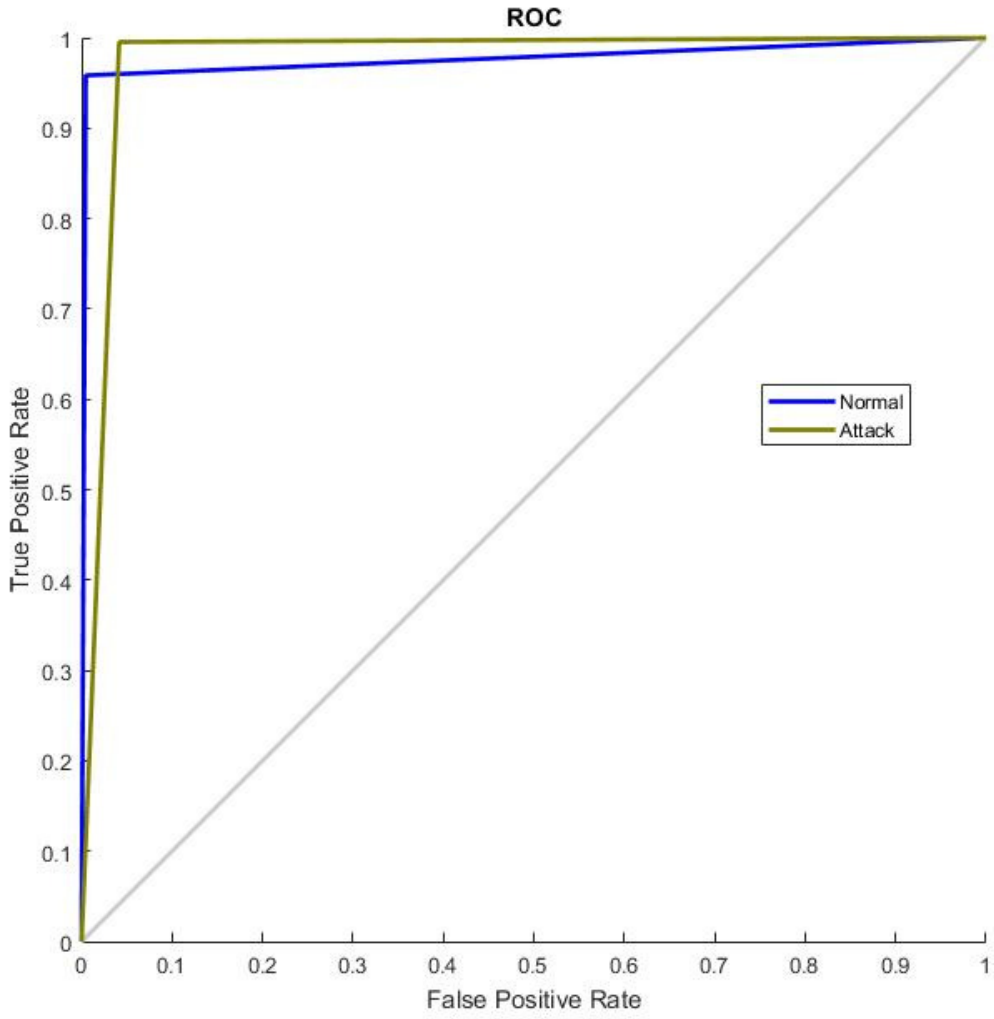

Figure 3: ROC of the proposed approch using NSLKDD Dataset