Journal of Information Systems and Informatics

Vol. 6, No. 2, June 2024 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

DOI: 10.51519/journalisi.v6i2.631 Published By DRPM-UBD

Optimizing Business Intelligence System Using Big Data and Machine Learning

James, G. G.[1], Oise, G. P [2], Chukwu, E. G.[3], Michael, N. A.[4], Ekpo, W. F. [5], Okafor, P. E [6].

[1, 2] Department of Computing, Faculty of Computing and Applied Sciences, Topfaith University, Mkpatak, Nigeria
[3] Federal University of Technology, Ikot Abasi, Nigeria
[4] Department of Mathematics and Computer Sciences, Ritman University, Ikot Ekpene.
[5] Supper Phone Technology Limited, Oron Road, Uyo, Akwa Ibom State University, Nigeria
[6] Department of State Services, Bayelsa Command
Email: [1] g.james@topfaith.edu.ng, [2] g.oise@topfaith.edu.ng, [3]chukwugabriel0010@gmail.com, [4]nsemike@yahoo.com, [5]gabresearch@gmail.com, [6]okaforpeace@gmail.com.

Abstract

The Business Intelligence (BI) and Data Warehouse (DW) system deployed in the Nigerian National Petroleum Corporation should provide cooperate decision makers with real-time information to help them identify and understand key business factors to make the best decisions for the situation at any given time. The relentless collection of data from user interactions have introduced both a high level of complexity, as well as a great opportunity for businesses. In addition to connecting not just people, but also machines to the internet, and then collecting data from these machines via sensors would result in an unimaginable repository of data. This ever-increasing collection of data is known as Big Data. Integrating this with existing Business intelligence systems and deep analysis using Machine Learning algorithms, Big Data can give useful insights into business problems and perhaps even to make suggestions as to when and where future problems will occur (Predictive Analysis) so that problems can be avoided or at least mitigated. This paper targets at developing a system capable of optimizing a business intelligence using big data and machine learning approach. The design of a system to optimize the Business Intelligence System using Machine Learning and Big Data at NNPC was successfully carried out. The System was able to automatically analyze the sample report under NNPC permission to use and it generated expected predictive outputs which serves as a better guide to managers. When applying Deep Learning, one seeks to stack several independent neural network layers that, working together, produce better results than the already existing shallow structures.

Keywords: optimized business intelligence, big data, machine learning approach, intelligent system, neural network.

1. INTRODUCTION

The goal of any organization is to achieve competitive advantage. This can only be possible if key decision makers within an organization have access to information, they can use to make informed decisions faster and quickly. Large scale organizations today produce very large data that needs to be interpreted for decision makers on a daily basis. Many of these organizations turn to data analysis and reporting tools otherwise known as Business Intelligence (BI) systems. However many of these systems do not offer predictive abilities. Many of these systems only look in-ward and analyze data within the organization's data warehouse and report. However organizations are surrounded by tons of external data known as big data, within the web that can be analyzed against their own data and used for predicting future trends that the organization can key quickly. This project will review and test some of algorithms and frameworks that can be used to achieve this purpose.

In the past, the concept of business intelligence (BI) was known as Executive Information Systems (EIS) and later known as Decision Support Systems (DSS), hence, Business intelligence (BI) according to Sahay and Ranjan [1] has two basic different meanings which relates to the use of the term intelligence. The primary, less frequently, is the human intelligence capacity applied in business affairs/activities. Intelligence of Business is a new field of the investigation of the application of human cognitive faculties and artificial intelligence technologies to the management and decision support in different business problems. And the second relates to the intelligence as information valued for its currency and relevance [2], [3]. The source of data in this wise was anything ranging from flat files to a normalized online transaction processing (OLTP) database system, while the end products are reports that allow end users to derive meaningful information by slicing and dicing the facts. However these data are often not available in a relational format. For example, in certain locations the data might be available in Comma Separated values (CSV) or Extensible Markup Language (XML) format which may not be consumed as is.

The organization under review is the Nigerian National Petroleum Corporation. "The Nigerian National Petroleum Corporation (NNPC)" is Nigeria's state oil corporation which was established in April 1, 1977. In addition to its exploration activities, the corporation was given powers and operational interests in refining, petrochemicals and products transportation as well as marketing. Between 1978 and 1989, NNPC constructed refineries in Warri, Kaduna and Port Harcourt and took over the 35,000- barrel shell Refinery established in Port Harcourt in 1965.
"The pipelines and product marketing company (PPMC) is the product distribution arm of NNPC. PPMC is directly responsible for the comparative ease with which petroleum products are sourced and distributed to all parts of the country, at a uniform price: a phenomenon which Nigerians have come to take for granted. PPMC, a subsidiary of NNPC, ensures, among other things, the availability of petroleum products to sustain our industries, automobiles and for domestic cooking". [4]

"PPMC receives crude oil from the NNPC unit called the National Petroleum Investments Management Services (NAPIMS). PPMC then supplies the crude oil to NNPC local refineries however, petroleum products are sometimes imported to supplement local production when the local refineries are unable to process enough for the country's needs. "Petroleum products either imported or refined locally are received by the PPMC through import jetties and pipelines, and distributed through pipelines to depots strategically located all over called bridging to designate retail outlets. There is also provision for using the rail system to move from some of the PPMC depots". [4]. The paper is concerned with developing a system capable of optimizing a business intelligence using big data and machine learning approach.

Business analytics makes extensive use of data, statistical and quantitative analysis, and explanatory and predictive modeling to help make actionable decisions and to improve business operations. There are many kinds of business analytics, real-time vs. non-real-time, strategic vs. tactic, planned vs. unplanned, and structured vs. unstructured [5]. Managers have used business analytics to inform their decision making for years. Now, they are using business analytics not only in analyzing past performance but also in identifying opportunities to improve future performance [6]. According to Chen et al. [7], business analytics consists of big data analytics, test analytics, web analytics, network analytics, and mobile analytics, many of which are unstructured and cannot be analyzed by relational database management tools [8] points out that there are three types of career categories for graduates majoring in big data business analytics: top pier management consulting, financial and risk analysts, and data scientists. Wills [9] Suggested that applications of big data analytics in healthcare industry may begin with small data analytics since it is much more appropriate for healthcare managers and organizations to translate data and to have actionable intelligence based on current infrastructure in the healthcare systems. Critical skills for business analytics include optimization analytics, descriptive analytics, and predictive analytics. For example, Saed et al. [10] suggest the following, communication skills, SQL and query skills, data mining and data warehousing, statistics skills, data visualization, text mining, and no SQL skills.

In the survey to described Business intelligence (BI) as the process of transforming raw data into useful information for more effective strategic, operational insights, and decision-making purposes so that it yields real business benefits [11]. According to them, the new emerging technique can not only improve applications in enterprise systems and industrial informatics, respectively, but also play a very important role to bridge the connection between enterprise systems and industrial informatics. The paper further intended a short introduction to BI with the emphasis on the fundamental algorithms and recent progress. In addition, they pointed out that business intelligence (BI) is the key technologies for users to efficiently extract useful information from oceans of data. The concept of BI was firstly introduced by Garter Group, and incipiently referred to the tools and technologies including data warehouses, reporting query and analysis.

Typical BI applications ensures the analysis and measurement of the customer's thoughts, behaviors, relationships, buying attitudes, choices, and many more parameters that form the backbone of effective strategy in building business operations management, customer relationship management, and other business operations [12]. Normally, Internet-of-Things (IoT) Big Data are highly unstructured, ambiguous, inconsistent, and incomplete, thereby creating many hazards in the analytic of potential knowledge granules related to BI applications. This creates potential challenges for data mining and knowledge granules that can be effectively clustered for further analysis and exploration [13].

Chang H-T et al. [14], in their research work inspected the structural analysis and clustering of complex knowledge granules in an Internet of Things (IoT) big-data environment. In their work, they proposed a knowledge granule analytic and clustering (KGAC) framework that explores and assembles knowledge granules from IoT big-data arrays for a business intelligence (BI) application. Their work implements neuro-fuzzy analytic architecture rather than a standard fuzzified approach to discover the complex knowledge granules. Furthermore, they implemented an enhanced knowledge granule clustering (e-KGC) mechanism that is more elastic than previous techniques when assembling the tactical and explicit complex knowledge granules from IoT Big Data array in such a way as to present knowledge of strategic value to executives and enable knowledge users to perform further BI actions. The benefits of their framework design can help executives and knowledge users generate cognitive decisions, plans, and actuations for the effective monitoring of BI applications.

Other studies focus on standalone fuzzy/GA/neural - based knowledge analytic mechanisms to boost the multi-criteria decision-making systems for BI-service applications [15]. Asemi, A. et, al. [16] in their study addressed BI-service applications that implement either fuzzy based or fuzzy-GA or neuro-fuzzy-based knowledge analytic technologies to boost the multi-criteria decision-making process used to construct a business security strategy of a business-to-consumer (B2C) organization selling products and services. Song et al. [17] sought to characterize the relationship between big data and small data. They attempted to synthesize the literature on big data, as well as emerging literature on small data, and identify the connections between them. Key questions include what is the relationship between big and small data, how does big data become small data, and when is there a reciprocal relationship as small data is aggregated into big data?. The main contribution of this research work is building a Deep Neural Network model with most relevant features to loss prediction in the daily activities of the NNPC downstream sector. Hence, Deep learning algorithms are a subset of

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Machine Learning algorithms that typically involve learning representations at different hierarchy levels to enable building complex concepts out of simpler ones, [16], hence, a direct loss minimization approach to train deep neural networks was proposed by Duan and Xu [11] in their work of Training Deep Neural Networks via Direct Loss Minimization, a novel dynamic programming algorithm that can efficiently compute the weight updates was developed and the experiments showed that it is beneficial when compared to a large variety of baselines in the context of action classification and object detection, particularly in the presence of noisy labels but the work is challenged because there is no provision of direct loss minimization in the context of other non-decomposable losses.

Saldana et al. [18] proposed in his work entitled; Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation to develop a set of computationally efficient, yet accurate, methods to predict flash points of fuel mixtures based solely on their chemical structures and mole fractions. Two approaches were tested using data obtained from the existing literature: (1) machine learning directly applied to mixture flash point data (the mixture QSPR approach) using additive descriptors and (2) machine learning applied to pure compound properties (the QSPR approach) in combination with Le Chatelier rule based calculations. It was found that the second method performs better than the first with the available databank and for the target application. A novel equation was proposed, and they evaluated the performance of the result, fully predictive, Le Chatelier rule based approach on new experimental data of surrogate jet and diesel fuels, yielding excellent results. They further predicted the variation in flash point of diesel–gasoline blends with increasing proportions of gasoline.

Sung et al, [19] in their work of Hourly Water Level Forecasting at Tributary Affected by Main River Condition developed an hourly water level forecasting models with lead-times of 1 to 3h using an artificial neural network (ANN) for Anyang cheon stream. To consider the backwater effect from this river, an enhanced tributary water level forecasting model was proposed by adding multiple water level data on the main river as input variables into the conventional ANN structure which often uses rainfall and upstream water level data. The results indicate that the inclusion of multiple water level data on the main river to the network provides water level forecasts with greater accuracy at the Ogeumgyo gauging station of interest. However, the challenge to the work was the final best ANN model for forecasting 3h ahead was unsatisfactory, showing underestimation at many rising parts of the hydrograph.

Another approach to building a predictive analytics system is with Clustering algorithms which are useful tools for data mining, compression, probability density estimation, and many other important tasks. However, most clustering algorithms require the user to specify the number of clusters (called k), and it is not always clear what is the best value for k. When clustering a dataset, the right number k of clusters to use is often not obvious, and choosing k automatically is a hard algorithmic problem. Hence, Sahay and Ranjan [1] in their paper of learning the k in k-means presented an improved algorithm for learning k while clustering, called the G-means Algorithm. The G-means algorithm is based on a statistical test for the hypothesis that a subset of data follows a Gaussian distribution. G-means runs k-means with increasing k in a hierarchical fashion until the test accepts the hypothesis that the data assigned to each k-means center are Gaussian. Two key advantages are that the hypothesis test does not limit the covariance of the data and does not compute a full covariance matrix.

Wang et. al. [20] in his work of Support vector machines based on K-means clustering for real-time business intelligence systems introduced K-means Support Vectors Machines(KMSVM) to speed up the response of Support vector machines (SVM) classifiers by reducing the number of support vectors. Support vector machines (SVM) is often applied to build classifiers, which can help users make well-informed business decisions. Despite their high generalization accuracy, the response time of SVM classifiers was still a concern when applied into real-time business intelligence systems, such as stock market surveillance and network intrusion detection. They succeeded in optimizing the response SVM classifiers by the K-means SVM (KMSVM) algorithm as proposed in the paper. The KMSVM algorithm combines the K-means clustering technique with SVM and requires one more input parameter to be determined: the number of clusters. Experiments compare the KMSVM algorithm with SVM on real-world databases, and the results show that the KMSVM algorithm can speed up the response time of classifiers by both reducing support vectors and maintaining a similar testing accuracy to SVM.

Pelleg and Moore [21] proposed a regularization framework for learning k, which they call X-means. The algorithm searches over many values of and scores each clustering model k using the so-called Bayesian Information Criterion. Based on the foregoing studies, this project will attempt to implement a Big-data analysis technology on NNPC's existing BI infrastructure using the Machine-Learning Algorithm known as the k-means clustering to optimize the existing system for predictive analysis, acquisition, storage and analysis of big data to provide new insights into the nation's oil and gas value chain. The algorithms and techniques to be used here will be applicable to any organization.

2. METHODS

2.1. Research Methods

To achieve the set fits, the following methods was used to actualize the work:

1) Carry out a preliminary investigation of the existing Business intelligence system for improvement.

2) Obtain a realistic statistical field data for predictive analysis.
3) Adopt a big data and machine learning approach in the design of the proposed system.
4) Provide a new layer for predictive analysis, acquisition, storage and analysis of big data that can be integrated into the established Business Intelligence Processes.
5) Utilize an algorithm and techniques that can be applicable to any organization.
6) Implement the proposed system using a robust programming tools such as JavaScript, Power BI, Visual Studio Code IDE and Microsoft Excel for temporary data entry and upload.

This paper review Machine Learning Algorithms, Big Data strategies and frameworks that can be used to optimize the existing system in the downstream sector of NNPC-PPMC and focus more on predictive analytics on the probability of loss before commencement of daily truck-out activities from any tank of choice.

The relentless collection of data from user interactions have introduced both a high level of complexity, as well as a great opportunity for businesses. In addition to connecting not just people, but also machines to the internet, and then collecting data from these machines via sensors would soon result in an unimaginable repository of data. This ever-increasing collection of data is known as Big Data. Integrating this with existing Business Intelligence Systems and Deep Analysis using Machine Learning Algorithms, Big Data can give useful insights into business problems and perhaps even to make suggestions as to when and where future problems will occur (Predictive Analysis) so that problems can be avoided or at least mitigated.

2.2 Proposed Architecture

The Proposed system is an integration of the machine learning Layer which uses Deep Neural Network to train the sample data for an optimized output. Below is the proposed NNPC BI system architecture and as Capture in Figure 1.

The architecture as presented in figure 1 is made up of the Machine Learning Layer, Extract, Transform and Load (ETL) Layer, The Enterprise Data Warehouse (EDW), and the Reporting Layer. Each of these components shall be discuss in detail in this section.

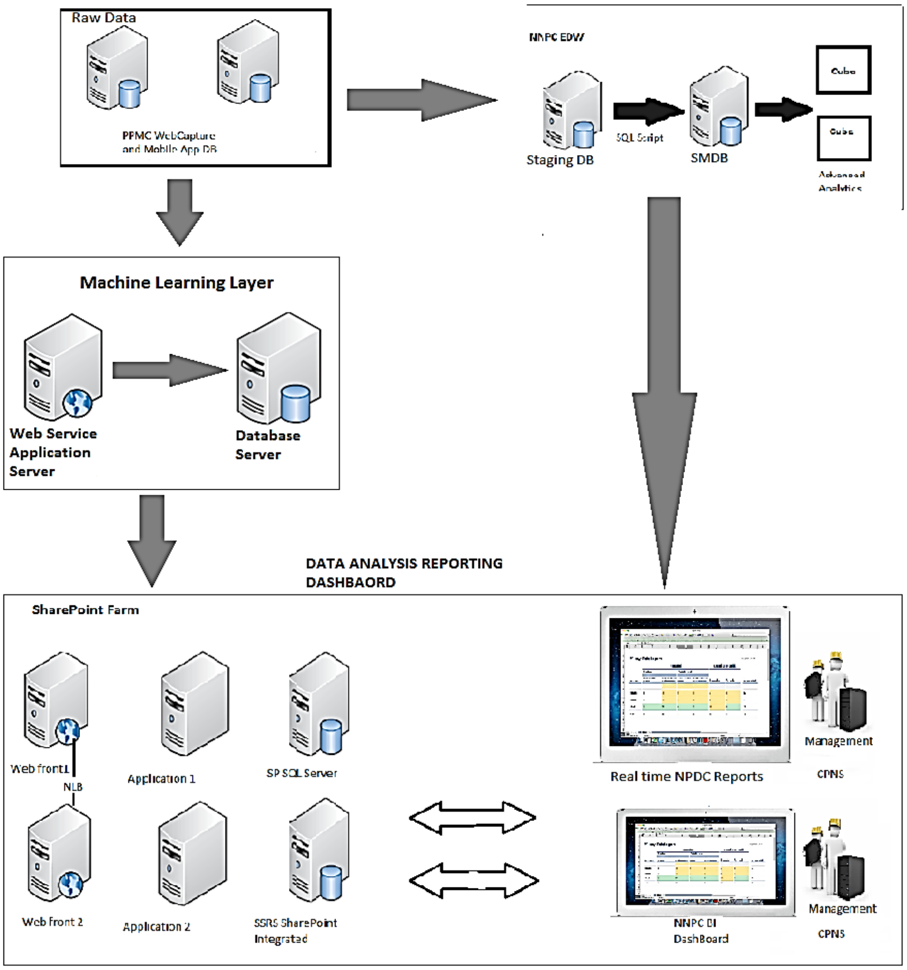


Figure 1. Showing the Architecture of the proposed BI System

2.2.1 Machine learning layer

The machine learning layer is an attempt at finding patterns from the data coming into EDW, these patterns hold the key to fully understanding the incoming data and providing true predictive analysis. A layer is the highest-level building block in deep learning. A layer is a container that usually receives weighted input, transforms it with a set of mostly non-linear functions and then passes these values as output to the next layer.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

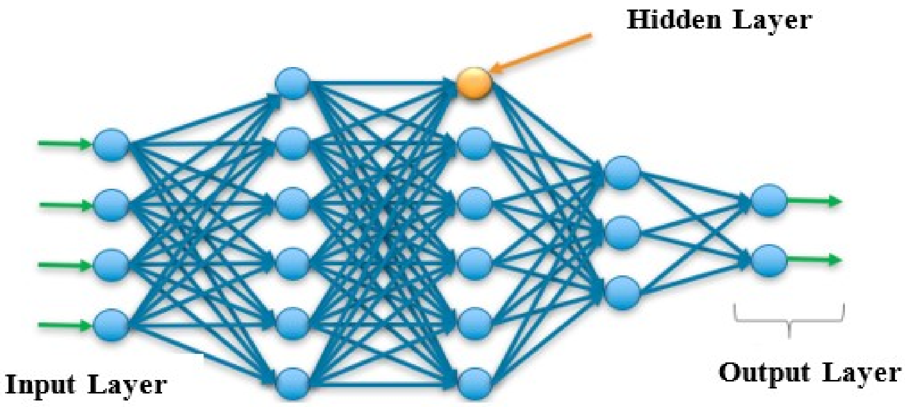


Figure 2. Showing the deep learning layer of the proposed BI System

This system will implement a key machine learning method known as the deep neural networks. The Machine Learning Layer consists of two primary components which includes: the Web Service Application Server and the Database Server and three other layers.

1) Application Server

Here, the implementation of machine learning program will be hosted in the cloud, the solution will be designed using JavaScript. The solution runs on Node. Js which is a runtime for running JavaScript on the server side.

2) Database server

In this case, the cloud storage acts as the database server which stores processed data from the machine learning application [22] [23]. In a typical enterprise environment, the source of data would be from series of excel files, tank gauging system sensors, web services or database tables. But for the purpose of this project, we limited our test scenario to an existing set of excel file that address a common pain point within the organization. Also, a test server was deployed to host the system in the cloud called the Heroku platform.

2.2.2 Extract, Transform and Load (ETL) Layer

The responsibility of the ETL layer is to extract data for varied sources and transform it into a normalized structured data that can be further loaded in a dimension model known as the data warehouse. If the source of the data is a structured OLTP system, the transformation required is minimal and is basically an extract and load operation [24] [25] [26] [27].

2.2.3 The enterprise data warehouse (EDW)

The Data warehouse (DW) layer of the solution is essentially an RDBMS database which is designed using dimension modeling techniques such as the Ralph Kimball Approach or Bill Inmon approach. In the DW layer, data is classified as dimensions or facts based on characteristics. A dimension gives the context to slice the data, while the facts are measures of interests.

2.2.4 Reporting layer

The reporting layer is the last layer in the chain. This layer is responsible for providing reports of analyzed data coming from the Enterprise data warehouse in form of charts and reports - a manner that is easily understood by the decision makers. Figure 3 and figure 4 shows the reporting charts of how the existing system presents its reporting output.

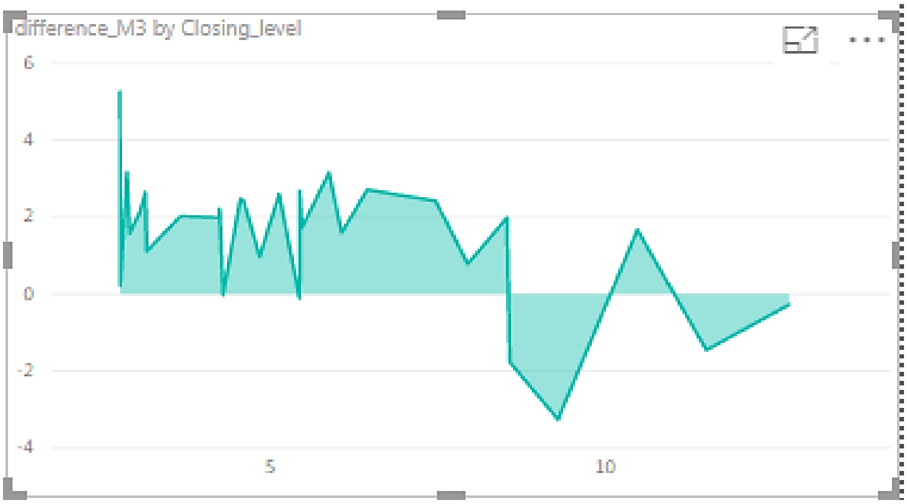


Figure 3. Reporting chart showing difference in closing level

Figure 3 presents the chart showing the differences that existed in closing level. Whilst Figure 4 present the report chart which carried out the comparison between sales and PRA closing gross.

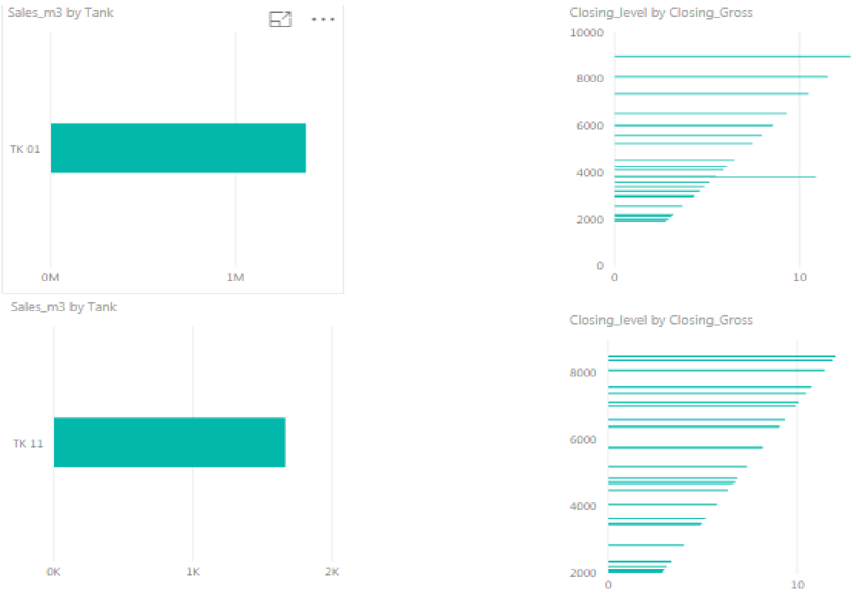


Figure 4: Report charts comparing Sales and PRA closing Gross.

2.3 System Flow Diagram

Figure 5 shows the flow of activities during the manipulation of the system by the users (Manager/Superintendent/Operator) loss prediction.

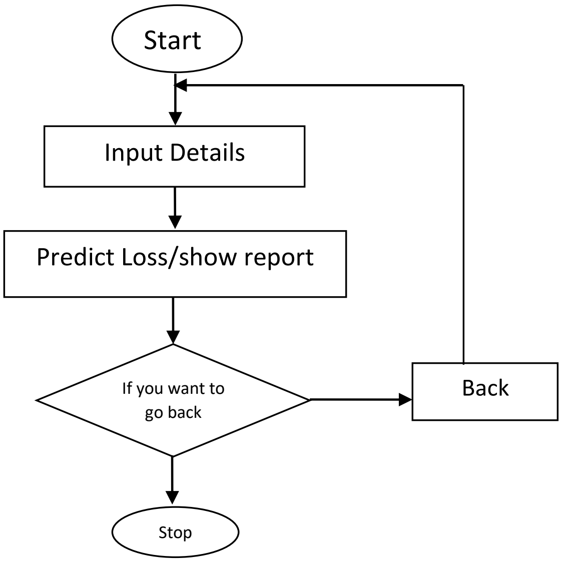


Figure 5. System Flow Diagram

2.4. Use Case Diagram

Use case diagrams are used to visually describe the interaction between users and the Business Intelligence system. Typically, they are used to describe how users would perform their role using the system and as such, form an essential part of the development process. There are two basic elements that make up a use case:

Actor: The actor is the user with a specific role which describes how he or she interacts with the system.

System: Use cases capture functional requirements that specify the intended behavior of the system.

Use cases are modeled using unified modeling language and are represented by ovals containing the names of the use case. Actors are represented using lines with the name of the actor written below the line. To represent an actor's participation in a system, a line is drawn between the actor and the use case. Boxes around the use case represent the system boundary [25].

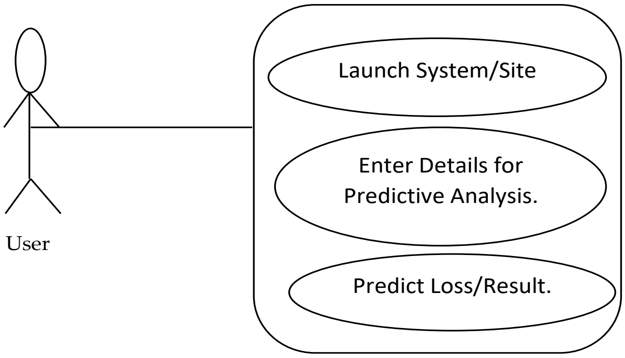


Figure 6. Use Case Diagram

2.5 Deep Neural Network (DNN)

Several researchers in this field have defined Deep Neural Networks as networks that have an input layer, an output layer and at least one hidden layer in between. Each layer performs specific types of sorting and ordering in a process that some refer to as "feature hierarchy." One of the key uses of these sophisticated neural networks is dealing with unlabeled or unstructured data. The phrase "deep learning" is also used to describe these deep neural networks, as deep learning represents a specific form of machine learning where technologies using aspects of artificial intelligence seek to classify and order information in ways that go beyond simple input/output protocols.

Many Deep Neural Network employ the feed-forward training approach where each hidden unit, j, typically uses the logistic function (the closely related hyberbolic tangent which is also often used and any function with a well-behaved derivative can be used) to map its total input from the layerxj, to the layeryjthat it sends to the layer above.

2.6 Deep Neural Network (DNN) Algorithm

Deep Neural Network is mostly applied in cases where the amount of data to be computed is more than that which could be handled by ordinary linear equations or regressive mathematical models. From the graph below in figure 7, we see that the performance and efficiency of deep learning Algorithm increases proportionally with the amount of data input thus validating the use of big data.

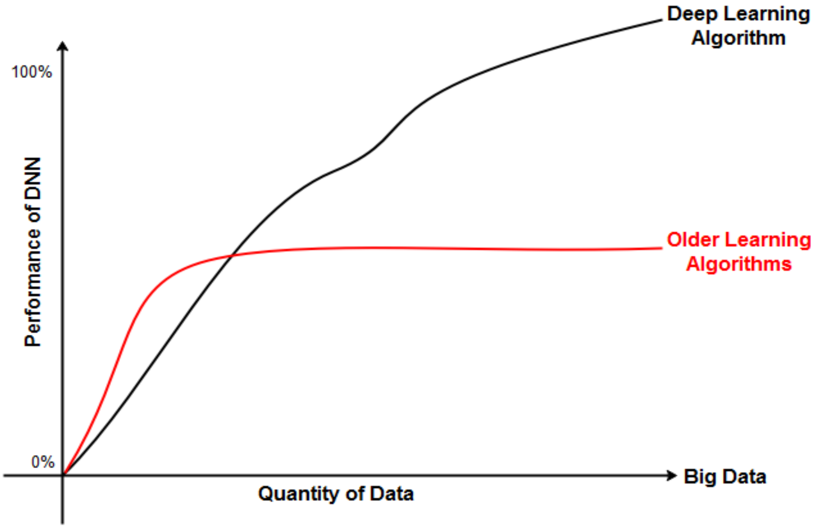


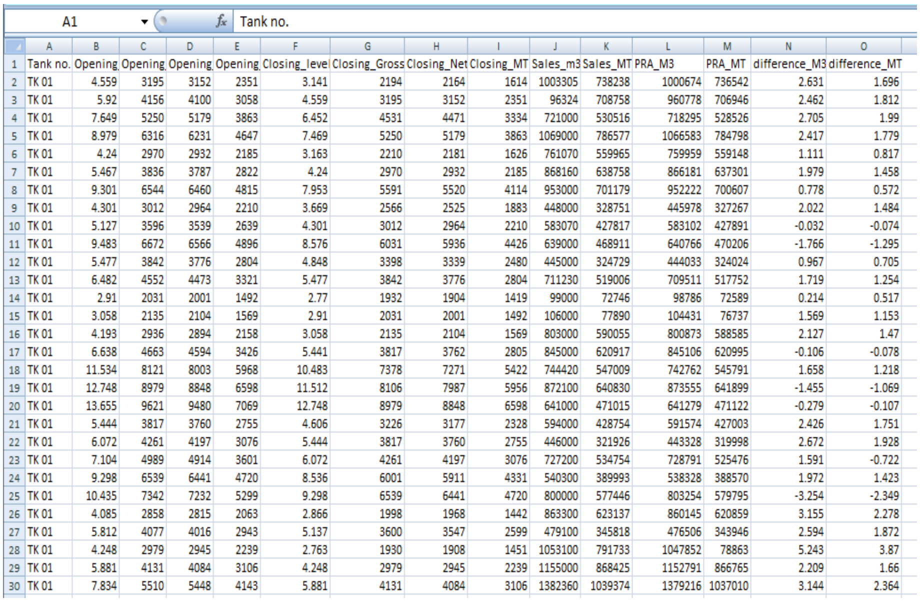
Figure 7. DNN Performance versus Big Data

The best way to describe neural networks is by using the conventional mathematical notations that illustrates the calculations that take place on each layer and within it neuron on the network. A deep neural network as already stated above is made up of an input layer, hidden layer(s) and an output layer. The input layer, consisting a number of neurons which receives the input sample compressed into formats that can be processed by the neurons. The units from the first hidden layer to the output layer compute the inputted sample data with respect to the defined weights and biases of the network synapses. We adopted a topology or network configuration that connects the input layer to the first hidden layer which is also fully connected to the second hidden layer. This connection is continued to the output layer.

2.7 Staging and Data Warehouse Database

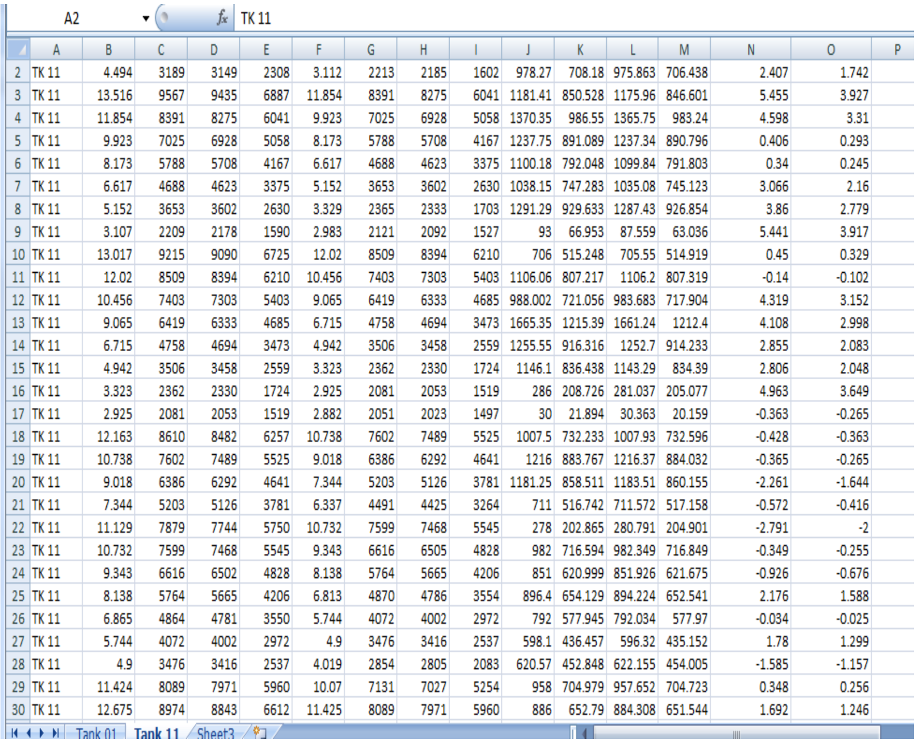
A staging database is an internal data store used in transforming and preparing the data obtained from the source systems before the data is loaded to other data stores in a data warehouse. In our solution we are extracting our data from the staging database of the data warehouse and the tank gauging systems using SQL Server integration services (SSIS) into the database (cloud) to be processed by our Machine Learning System. SSIS is a platform for data integration and workflow application. It features a fast and flexible data warehousing tool used for data extraction, transformation, and loading (ETL). The tool may also be used to automate maintenance of SQL Server databases and updates to multidimensional cube data.

We extracted the data from staging database of the data warehouse into an excel file for test purposes and cleaning before pushing it into the SQL Database of our machine learning system. The sample data for each of the report shown in Figure 8 and Figure 9 was integrated to a new database setup for the machine-learning engine to read from.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Tank no.	Opening	Opening	Opening	Opening	Closing_level	Closing_Gross	Closing_Net	Closing_MT	Sales_m3	Sales_MT	PRA_M3	PRA_MT	difference_M3	difference_MT
2	TK 01	4.559	3195	3152	2351	3.141	2194	2164	1614	1003305	738238	1000674	736542	2.631	1.696
3	TK 01	5.92	4156	4100	3058	4.559	3195	3152	2351	96324	708758	960778	706946	2.462	1.812
4	TK 01	7.649	5250	5179	3863	6.452	4531	4471	3334	721000	530516	718295	528526	2.705	1.99
5	TK 01	8.979	6316	6231	4647	7.469	5250	5179	3863	1069000	786577	1066583	784798	2.417	1.779
6	TK 01	4.24	2970	2932	2185	3.163	2210	2181	1626	761070	559965	759959	559148	1.111	0.817
7	TK 01	5.467	3836	3787	2822	4.24	2970	2932	2185	868160	638758	866181	637301	1.979	1.458
8	TK 01	9.301	6544	6460	4815	7.953	5591	5520	4114	953000	701179	952222	700607	0.778	0.572
9	TK 01	4.301	3012	2964	2210	3.669	2566	2525	1883	448000	328751	445978	327267	2.022	1.484
10	TK 01	5.127	3596	3539	2639	4.301	3012	2964	2210	583070	427817	583102	427891	-0.032	-0.074
11	TK 01	9.483	6672	6566	4896	8.576	6031	5936	4426	639000	468911	640766	470206	-1.766	-1.295
12	TK 01	5.477	3842	3776	2804	4.848	3398	3339	2480	445000	324729	444033	324024	0.967	0.705
13	TK 01	6.482	4552	4473	3321	5.477	3842	3776	2804	711230	519006	709511	517752	1.719	1.254
14	TK 01	2.91	2031	2001	1492	2.77	1932	1904	1419	99000	72746	98786	72589	0.214	0.517
15	TK 01	3.058	2135	2104	1569	2.91	2031	2001	1492	106000	77890	104431	76737	1.569	1.153
16	TK 01	4.193	2936	2894	2158	3.058	2135	2104	1569	803000	590055	800873	588585	2.127	1.47
17	TK 01	6.638	4663	4594	3426	5.441	3817	3762	2805	845000	620917	845106	620995	-0.106	-0.078
18	TK 01	11.534	8121	8003	5968	10.483	7378	7271	5422	744420	547009	742762	545791	1.658	1.218
19	TK 01	12.748	8979	8848	6598	11.512	8106	7987	5956	872100	640830	873555	641899	-1.455	-1.069
20	TK 01	13.655	9621	9480	7069	12.748	8979	8848	6598	641000	471015	641279	471122	-0.279	-0.107
21	TK 01	5.444	3817	3760	2755	4.606	3226	3177	2328	594000	428754	591574	427003	2.426	1.751
22	TK 01	6.072	4261	4197	3076	5.444	3817	3760	2755	446000	321926	443328	319998	2.672	1.928
23	TK 01	7.104	4989	4914	3601	6.072	4261	4197	3076	727200	534754	728791	525476	1.591	-0.722
24	TK 01	9.298	6539	6441	4720	8.536	6001	5911	4331	540300	389993	538328	388570	1.972	1.423
25	TK 01	10.435	7342	7232	5299	9.298	6539	6441	4720	800000	577446	803254	579795	-3.254	-2.349
26	TK 01	4.085	2858	2815	2063	2.866	1998	1968	1442	863300	623137	860145	620859	3.155	2.278
27	TK 01	5.812	4077	4016	2943	5.137	3600	3547	2599	479100	345818	476506	343946	2.594	1.872
28	TK 01	4.248	2979	2945	2239	2.763	1930	1908	1451	1053100	791733	1047852	78863	5.243	3.87
29	TK 01	5.881	4131	4084	3106	4.248	2979	2945	2239	1155000	868425	1152791	866765	2.209	1.66
30	TK 01	7.834	5510	5448	4143	5.881	4131	4084	3106	1382360	1039374	1379216	1037010	3.144	2.364

Figure 8. Sample data from tank 01.



A2 ▾ *fx* TK 11

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
2	TK 11	4.494	3189	3149	2308	3.112	2213	2185	1602	978.27	708.18	975.863	706.438	2.407	1.742	
3	TK 11	13.516	9567	9435	6887	11.854	8391	8275	6041	1181.41	850.528	1175.96	846.601	5.455	3.927	
4	TK 11	11.854	8391	8275	6041	9.923	7025	6928	5058	1370.35	986.55	1365.75	983.24	4.598	3.31	
5	TK 11	9.923	7025	6928	5058	8.173	5788	5708	4167	1237.75	891.089	1237.34	890.796	0.406	0.293	
6	TK 11	8.173	5788	5708	4167	6.617	4688	4623	3375	1100.18	792.048	1099.84	791.803	0.34	0.245	
7	TK 11	6.617	4688	4623	3375	5.152	3653	3602	2630	1038.15	747.283	1035.08	745.123	3.066	2.16	
8	TK 11	5.152	3653	3602	2630	3.329	2365	2333	1703	1291.29	929.633	1287.43	926.854	3.86	2.779	
9	TK 11	3.107	2209	2178	1590	2.983	2121	2092	1527	93	66.953	87.559	63.036	5.441	3.917	
10	TK 11	13.017	9215	9090	6725	12.02	8509	8394	6210	706	515.248	705.55	514.919	0.45	0.329	
11	TK 11	12.02	8509	8394	6210	10.456	7403	7303	5403	1106.06	807.217	1106.2	807.319	-0.14	-0.102	
12	TK 11	10.456	7403	7303	5403	9.065	6419	6333	4685	988.002	721.056	983.683	717.904	4.319	3.152	
13	TK 11	9.065	6419	6333	4685	6.715	4758	4694	3473	1665.35	1215.39	1661.24	1212.4	4.108	2.998	
14	TK 11	6.715	4758	4694	3473	4.942	3506	3458	2559	1255.55	916.316	1252.7	914.233	2.855	2.083	
15	TK 11	4.942	3506	3458	2559	3.323	2362	2330	1724	1146.1	836.438	1143.29	834.39	2.806	2.048	
16	TK 11	3.323	2362	2330	1724	2.925	2081	2053	1519	286	208.726	281.037	205.077	4.963	3.649	
17	TK 11	2.925	2081	2053	1519	2.882	2051	2023	1497	30	21.894	30.363	20.159	-0.363	-0.265	
18	TK 11	12.163	8610	8482	6257	10.738	7602	7489	5525	1007.5	732.233	1007.93	732.596	-0.428	-0.363	
19	TK 11	10.738	7602	7489	5525	9.018	6386	6292	4641	1216	883.767	1216.37	884.032	-0.365	-0.265	
20	TK 11	9.018	6386	6292	4641	7.344	5203	5126	3781	1181.25	858.511	1183.51	860.155	-2.261	-1.644	
21	TK 11	7.344	5203	5126	3781	6.337	4491	4425	3264	711	516.742	711.572	517.158	-0.572	-0.416	
22	TK 11	11.129	7879	7744	5750	10.732	7599	7468	5545	278	202.865	280.791	204.901	-2.791	-2	
23	TK 11	10.732	7599	7468	5545	9.343	6616	6505	4828	982	716.594	982.349	716.849	-0.349	-0.255	
24	TK 11	9.343	6616	6502	4828	8.138	5764	5665	4206	851	620.999	851.926	621.675	-0.926	-0.676	
25	TK 11	8.138	5764	5665	4206	6.813	4870	4786	3554	896.4	654.129	894.224	652.541	2.176	1.588	
26	TK 11	6.865	4864	4781	3550	5.744	4072	4002	2972	792	577.945	792.034	577.97	-0.034	-0.025	
27	TK 11	5.744	4072	4002	2972	4.9	3476	3416	2537	598.1	436.457	596.32	435.152	1.78	1.299	
28	TK 11	4.9	3476	3416	2537	4.019	2854	2805	2083	620.57	452.848	622.155	454.005	-1.585	-1.157	
29	TK 11	11.424	8089	7971	5960	10.07	7131	7027	5254	958	704.979	957.652	704.723	0.348	0.256	
30	TK 11	12.675	8974	8843	6612	11.425	8089	7971	5960	886	652.79	884.308	651.544	1.692	1.246	

Figure 9. Sample data from Tank 11.

2.8 Mathematical Equations

From the work of Larochelle et al. [28], the functionality of the Deep Neural Network algorithm is illustrated thus. Given an input x, the value of the *j-th* unit in the *i-th* layer is denoted$\hat{h}_j^i(x)$, with $i = 0$denoting the input layer, $i = l + 1$ denoting the output layer. The size of a layer is denoted as$|\hat{h}^i|$. The default activation function is calculated by an internal bias $b_j^i$of that neuron. The set of weights W_{jk}^ibetween $\hat{h}_j^{i-1}(x)$ in layer $i - 1$ and unit $\hat{h}_j^i(x)$ in layer$i$ determines the activation of unit $\hat{h}_j^i(x)$. The DNN algorithm is explained thus:

$$\hat{h}_j^i(x) = sigmoid\left(a_j^i\right)$$

$$where a_j^i(x) = b_j^i + \sum_k W_{jk}^i \hat{h}_j^{i-1}(x) \forall i \in \{1, \ldots, l\} with \hat{h}^0(x) = x \qquad (1)$$

The $sigmoid(\cdot)$is the sigmoid squashing function: $igmoid(a) = \frac{1}{1+ e^{-a}}$, alternatively, the sigmoid function could be replaced by the hyperbolic tangent). Given the last hidden layer, the output layer is computed similarly thus;

$$o(x) = \hat{h}_j^{i-1}(x) = f\left(a^{l+1}(x)\right)wherea^{l+1}(x) = b^{l+1} + W^{l+1}\hat{h}^{l}(x)$$

The activation function $f(\cdot)$depends on the supervised task defined for the network to achieve. For a regression problem, an identify function is required however since the problem discussed in this work is to achieve maximization of resources, the maximization function is used thus:

$$f_j(a) = max_j(a) = \frac{e^{a_j}}{\sum_{k=1}^{K} e^{a_k}} \qquad (2)$$

When an input sample x is fed into the network, $equation(1)$is applied at each layer and this will generate a pattern of activities in the other layers of the network [28].

3. RESULTS AND DISCUSSION

3.1 System Implementation

The completed code was compiled into an executable package using Microsoft MS Build from the visual studio IDE. We deployed the executable file into the production server. Next, we set up a scheduled task to execute the machine learning algorithm every five minutes. This means that every 5 minutes the SSIS will pull data from the staging database (Gauging systems and databases) automatically into the data warehouse and load the analyzed data into the machine learning database (cloud) for reporting and analysis. One of the goal of BI is to ensure that quality data would be presented to stakeholders and customers. As data is brought together from different sources, there is a possibility of finding a broad range of data quality issues that require attention. For example, one of the most challenging data quality issues is duplication of data. For the purpose of this project, the data quality approach would be employed in the SSIS. The data quality approach we will use include: profiling, cleansing and auditing.

3.1.1 Profiling

Data profiling would help us to proactively assess whether a source data extract meets the baseline quality standards of the solution. This involves identifying any condition or issues that would cause the ETL packages not to be executed successfully and that would force us to start again from the beginning. Some of the activities that would carry out here includes: checking of records that might have been missing or NULL values for important columns which could cause a serious problem or failure in the source system. For data profiling approach, we will have to choose the right SSIS transforming components, e.g. the conditional split transformation which would help to filter the SSIS data flow based on specific conditions for each quality indicator, such as NULL data values, duplicate identifiers, or invalid data ranges.

3.1.2 Cleansing

Data cleansing enforces the business and schema rules of an application for each source record on a column-by-column and record-by-record basis. The major cleansing activities that would be carried out in the ETL would involve:

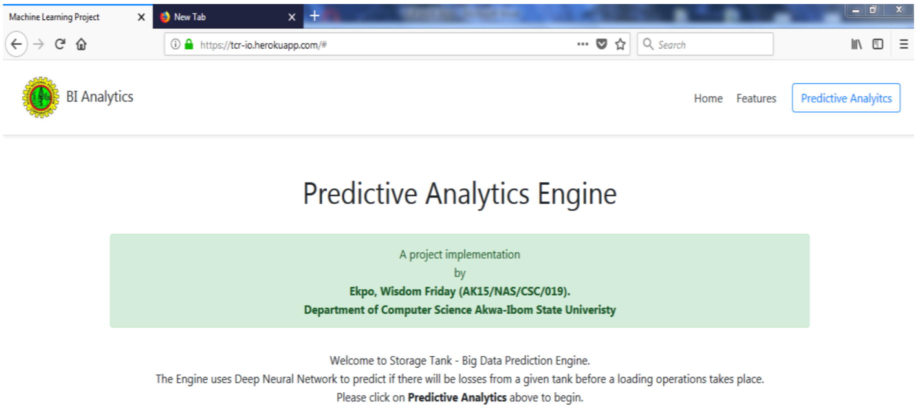
1) Using unique identifier such as composite key, to spot and overcome data duplicates.
2) Using SSIS Transforming components such as derived column and Data Conversion to overcome the problem of inconsistent data formats.

3.1.3 Auditing

Auditing was used to provide proof that all data integrated solution satisfies necessary business, technical, and regulatory standards. These are:

1) The ability of the system to track all data integration operations.
2) The ability to ensure that all data have been successfully processed.
3) The ability to track the success, failure and execution duration of every component of the integration solution.

Figure 10 shows the interface of the optimized Business Intelligence System for predictive analytics.

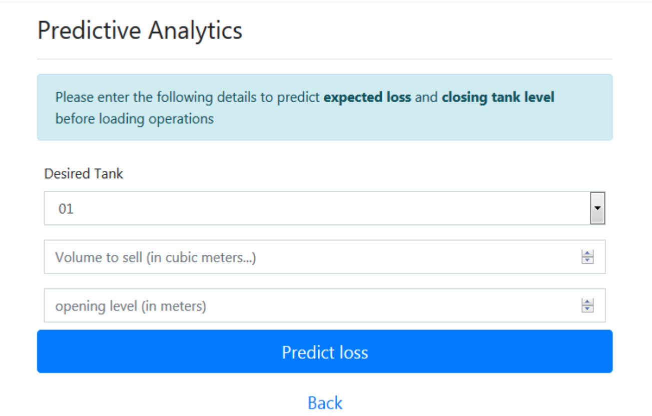


Predictive Analytics Engine

A project implementation
by
Ekpo, Wisdom Friday (AK15/NAS/CSC/019).
Department of Computer Science Akwa-Ibom State Univeristy

Welcome to Storage Tank - Big Data Prediction Engine.
The Engine uses Deep Neural Network to predict if there will be losses from a given tank before a loading operations takes place.
Please click on **Predictive Analytics** above to begin.

Figure 10. Showing the interface of the Implemented System



Predictive Analytics

Please enter the following details to predict **expected loss** and **closing tank level** before loading operations

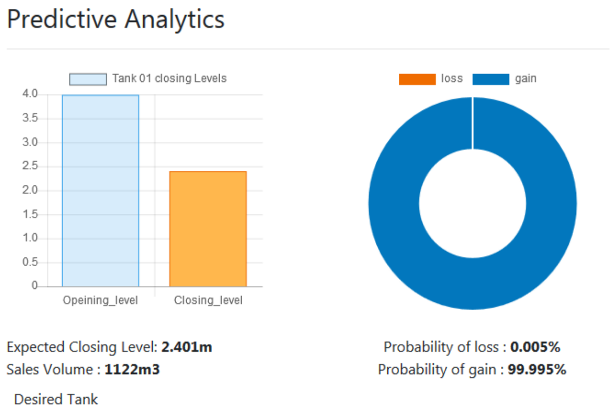
Desired Tank

01

Volume to sell (in cubic meters...)

opening level (in meters)

Fig 11: Showing the input Interface for users.



Predictive Analytics

Expected Closing Level: **2.401m**
Sales Volume : **1122m3**

Desired Tank

Probability of loss : **0.005%**
Probability of gain : **99.995%**

Fig 12: Predictive analytics Engine result

Fig 13: Different types of tanks embedded with Tank gauging systems to monitor product flow rates, Temperature, Density, Level and report to the display interface

3.2 Discussion

The database engine was designed using the Deep Neural Network (DNN) whose sole purpose is to train a network of neurons with data alone. In the case of this project, DNN training and inference happens in the cloud. For example, when the required data is uploaded through the SSIS on your Desktop or Smartphone, the data set as uploaded is sent up to the cloud for processing on a Machine Learning server. Once the inference processing has occurred, a result is sent back to the desktop or Smartphone. At the end of the analysis done by the machine learning Predictive Analytics engine, it outputted the results into two (2) text files representing;

1) The probability percentage of Loss.
2) The probability percentage Gain.
3) The expected Closing Level which guides operators to end of operations closing deep off.

Then using Microsoft Power BI, the data produced was analyzed and visualized using scatter diagrams to ascertain the probability of loss and Gain of each Tank Operation.

The distinctive performance of the neural network is immediately apparent. Starting from the input to the output layer, It indicates that the data at each level in the storage tank, the chances of losses is dependent on the quantity for truck out, thereby disclosing the chances of loss due to tank behavior. Finally, decision makers would not want to use data if it output when there is higher percentage in the probability of loss to make business decisions. Of immediate interest to decision makers would be the result of data with the highest percentage in probability of gain. The probability percentage of losses = 0.005% and the probability percentage of gain = 99.995%. Clearly if the organization under review would ensure that data generated by their organization averagely stays within the high gain probability, they definitely make profit and perform above average all things being equal.

4. CONCLUSION

Finally, the researchers successfully carry out preliminary investigation of Business intelligence system and obtained a realistic statistical field data for predictive analysis. The work adopted a big data and machine learning approach in the design of the proposed system as well provide a new layer that can be applicable to any organization. It make use of the Machine-Learning Algorithm known as the k-means clustering to optimize the existing system for predictive analysis, acquisition, storage and analysis of big data to provide new insights into the nation's oil and gas value chain. The system was implemented with JavaScript, Power BI, Visual Studio Code IDE as front-end programming tools and Microsoft Excel for temporary data entry and upload. This system was able to automatically analyze the sample report under NNPC permission to use and as well generated expected predictive outputs which serves as a better guide to managers It was discovered that whenever there is need to apply Deep Learning Algorithm, one seeks to stack several independent neural network layers that when working together, could produce better results than the already existing shallow structures. The work comprehensively explored some of the main tasks normally performed when manipulating Time-Series data using deep neural network structures.

REFERENCES

[1] B. S. Sahay and J. Ranjan, "Real time business intelligence in supply chain analytics," *Inf. Manag. Comput. Secur.*, vol. 16, no. 1, pp. 28–48, Mar. 2008, doi: 10.1108/09685220810862733.

[2] G. G. James, E. G. Chukwu, and P. O. Ekwe, "Design of an Intelligent based System for the Diagnosis of Lung Cancer," *Int. J. Innov. Sci. Res. Technol.*, vol. 8, no. 6, pp. 791–796, 2023.

[3] C. Ituma, G. G. James, and F. U. Onu, "A Neuro-Fuzzy Based Document Tracking & Classification System," *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 10, pp. 414–423, Feb. 2020, doi: 10.33564/IJEAST.2020.v04i10.075.

[4] N. Okwuchukwu, D. Eseme, and A. Charlyn, "Commercialisation Of Public Enterprises In Nigeria: A Study Of The Nigerian National Petroleum Corporation (Nnpc)," 2023.

[5] S. Biswas and J. Sen, "A Proposed Architecture for Big Data Driven Supply Chain Analytics".

[6] S. LaValle, Lesser, E., Shockley, R., Hopkins, M. S., and Kruschsitz, N., "Big data, analytics and the path from insights to value," *MIT Sloan Manag. Rev.*, pp. 21–32, 2011.

[7] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact".

[8] Fitzgerald, Michael, "Training the Next Generation of Business Analytics Professionals," vol. 56, no. 2, p. 1, Winter 2015.

[9] Wills, Mary J., "Decisions Through Data: Analytics in Healthcare," *Journal of Healthcare Management*, vol. 59, no. 4, pp. 254–262, Aug. 2014.

[10] K. Saeed, A. Sidorova, and A. Vasanthan, "The Bundling of Business Intelligence and Analytics," *J. Comput. Inf. Syst.*, vol. 63, no. 4, pp. 781–792, Jul. 2023, doi: 10.1080/08874417.2022.2103856.

[11] L. Duan and L. D. Xu, "Business Intelligence for Enterprise Systems: A Survey," *IEEE Trans. Ind. Inform.*, vol. 8, no. 3, pp. 679–687, Aug. 2012, doi: 10.1109/TII.2012.2188804.

[12] G. George, M. R. Haas, and A. Pentland, "Big Data and Management," *Acad. Manage. J.*, vol. 57, no. 2, pp. 321–326, Apr. 2014, doi: 10.5465/amj.2014.4002.

[13] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data Mining with Big Data".

[14] H.-T. Chang, N. Mishra, and C.-C. Lin, "IoT Big-Data Centred Knowledge Granule Analytic and Cluster Framework for BI Applications: A Case Base Analysis," *PLOS ONE*, vol. 10, no. 11, p. e0141980, Nov. 2015, doi: 10.1371/journal.pone.0141980.

[15] Amir Atiya, "Learning Algorithms for Neural Networks," Retrieved from California Institute of Technology Pasadena database, 1991.

[16] Asemi, A.; Baba, M.S.; Haji Abdullah, R.; Idris, N., "Fuzzy multi criteria decision making applications:," in *Proceedings of the 3rd International Conference on Computer Engineering & Mathematical Sciences (ICCEMS*, Langkawi, Malaysia, 2014.

[17] Y. Song, A. G. Schwing, R. S. Zemel, and R. Urtasun, "Training Deep Neural Networks via Direct Loss Minimization".

[18] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, and B. Creton, "Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation," *Energy Fuels*, vol. 27, no. 7, pp. 3811–3820, Jul. 2013, doi: 10.1021/ef4005362.

[19] J. Sung, J. Lee, I.-M. Chung, and J.-H. Heo, "Hourly Water Level Forecasting at Tributary Affected by Main River Condition," *Water*, vol. 9, no. 9, p. 644, Aug. 2017, doi: 10.3390/w9090644.

[20] J. Wang, X. Wu, and C. Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems," *Int. J. Bus. Intell. Data Min.*, vol. 1, no. 1, p. 54, 2005, doi: 10.1504/IJBIDM.2005.007318.

[21] Dan Pelleg; Andrew Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proceedings of the 17th International Conference on Machine Learning*, School of Computer Science, Carnegie Mellon University, Pitsburgh, PA 15213, USA, 2000, pp. 727–734.

[22] Onu F. U.; Osisikankwu P. U.; Madubuike C. E.; James G. G., "Impacts of Object Oriented Programming on Web Application Development," *Int. J. Comput. Appl. Technol. Res.*, vol. 4, no. 9, pp. 706–710, 2015.

[23] G. Gregory and O. A. Ejaita, "The International Journal of Science & Technoledge," vol. 4, no. 7.

[24] M. Kubat, *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-63913-0.

[25] Raudel Ravelo Suárez, "Extensión Del Almacén De Datos Empresarial En Cimex," 2015, doi: 10.13140/RG.2.1.3533.6089.

[26] R. S, “System Analysis and Design,” *J. Inf. Technol. Softw. Eng.*, vol. 02, no. 05, 2012, doi: 10.4172/2165-7866.S8-e001.

[27] Honeywell, “Selecting the Right Technology for Tank Level Gauging for Custody Transfer Applications”, 2027.

[28] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, “Exploring Strategies for Training Deep Neural Networks”, *Journal of machine learning research*, vol. 10, no. 1, 2009.