Journal of Health Informatics in Africa

10th Health Informatics in Africa Conference (HELINA 2017)

Peer-reviewed and selected under the responsibility of the Scientific Programme Committee

Design of a Data Analytics Model for National Health Insurance Scheme

Terungwa Simon Yange [a*], Hettie Abimbola Soriyan[b], Oluoha, O[c].

[a]Department of Mathematics/Statistics/Computer Science, Federal University of Agriculture, Makurdi, Nigeria
[b]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
[c]Department of Computer Science, University of Nigeria, Nsukka, Nigeria

Background and Purpose: The need for better and faster decision-making based on data is stronger than ever; and being able to use massive amounts of data is a necessary competitive advantage. This has necessitated the urgent need for a sophisticated data analytics model for the effective transformation of data into actionable information to enhance quality decision-making. For instance, the healthcare domain is faced with unnecessary delays in the processing of the data submitted to the National Health Insurance Scheme (NHIS).
Methods: To address this, a data analytics model based on deep learning was designed in this research using unified modelling language.
Results: This model is intended to be implemented using Apache Hadoop and MySQL. When implemented, the model will make it easier to consolidate, cleanse, analyse, and publish data, so that all stakeholders will get information that they can act on, in the format they need.
Conclusion: Thus, the stakeholders will access the information more easily, which will enable them to plan, evaluate, and collaborate more effectively.

Keywords: Hadoop, Deep Learning, Data Analytics, Health Insurance

1 Introduction

Data has become an increasingly important resource for organisations as it is the mother of information technology. Due to the improvement in information technologies and the growth of internet, organisations are able to collect and store huge amount of data. Data however is not the same as information as it need to be analysed and extracted before it is useful. This becomes more difficult as the amount of data, data type and analytical dimensions increased. In order to support decision making, data needs to be converted into information and knowledge [1]. The concept of applying a set of technologies to turn data into meaningful information is what is known as data analytics. Organisations are faced with a number of problems when attempting to analyse their data. There is generally no lack of data. In fact, many businesses are drowning in data; they are unable to turn it into actionable information as it is a big challenge to determine relationships, predict future events, spot bad data, and allow for its analysis.

The core methodology in data analytics is machine learning, which is the area of computer science that aims to build systems and algorithms that learn from data so as to aid the processing and modelling of large amounts of data to discover previously unknown relationships. A variant of this area is known as deep learning which applies machine learning techniques to both structured, semi-structured and unstructured data [2]. It is based on learning multiple levels of representation. The multiple levels of representation correspond to multiple levels of abstraction. A significant feature of deep learning which makes it more promising in data analytics, is the learning of high level representations and complicated structure automatically from huge amounts of data to obtain useful information. Also, deep learning provides high-accuracy results, avoids the expensive design of handcrafted features, and utilizes the

*Corresponding author address: Department of Mathematics/Statistics/Computer Science, University of Agriculture, Makurdi, Nigeria.
Email: lordesty2k7@ymail.com, Tel: +2348064067803
 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

massive unlabelled data for unsupervised feature extraction [3]. This makes it more suitable for analytics of huge data generated by information intensive industries; e.g., oil and gas sector, financial institutions, health insurance sector etc.

The healthcare insurance is among the most information intensive industries. Its data, information, knowledge and insights keep growing in every second and the ability to extract useful information that will improve the quality of healthcare services rendered is very crucial. The primary purpose of the analytics of health insurance data is to track payments done by beneficiaries to providers for healthcare services, beneficiaries' contributions (premium), the cost of health services and also to address fraud [4]. This is the major source of the cost associated with healthcare which is vital to addressing the economic challenges associated with modern healthcare system. It provides for the payments of benefits as a result of sickness or injury which includes insurance for losses from accident, medical expense, disability, or accidental death and dismemberment.

With the prised benefits of the scheme outlined above, results from past researches have demonstrated the dissatisfaction of beneficiaries with the scheme for reasons such as additional fees on the pretext of no-inclusion of particular services in their insurance plan and poor customer service. It has also been discovered that there is poor general knowledge of the scheme among those that have enrolled for it [6][7]. Most of these issues are due to the lack of proper data processing facilities to provide good insights from the data collected from various health service providers in varying formats in the course of administering the scheme. This data (i.e., data about the enrolment of beneficiaries and service providers, claims submitted, registration updates, complaints, enquiries, sanctions of service providers) ranges from the structured, semi-structured to unstructured data which are manually collected from varying data sources posed a great challenge due to the manner it is collected and presented. The processing of this data is done by an exhausting manual task carried out by a few personnel who have the responsibility of approving, modifying or rejecting these requests within a limited period from their reception. This has called for a sophisticated analytics tool for proper processing, but as stated by [7], there is no such tool available in NHIS hence the manual processing of the data available which caused unnecessary delay in the process.

These delays in processing the transactions of the scheme have been a discouraging factor in embracing the scheme. Those who have dared to register with the scheme complain of delay in having their documentations regularized to enable them reap the full benefits of the scheme. Therefore, many who are yet to register get discouraged by the experiences of those who have started the process of documentation as it takes ninety (90) to three hundred and sixty-five (365) days. The implication is that the number embracing the scheme tends to be reducing or not encouraging. Service providers complain of excessive delay in processing their claims leading to delay in payments of their bills and this could lead to total withdrawal of service or provision of substandard services. In fact, the general complain in Nigeria is that service providers are excessively owed and payment delayed for a period of six (6) to twelve (12) months [5]. It must be understood clearly that the sustainability of the scheme is largely dependent on the commitment of service providers; therefore, prompt payment of their bills is likely to raise their morale and guarantee some level of commitment.

1.1 Related Works

Data analytics had empowered businesses to make better decisions. With the amount of digital and analog data increasing at an enormous rate, rigorous research is carried out in an effort to extract value from these datasets, to convert them into a form that can be used to make smarter decisions for improving business results. Data analytics has integrated the state-of-the-art computational and statistical techniques to extract business value from a rapidly expanding volume of data. In the business world where the gap between winners and losers is narrowing down, companies are increasingly turning to data analytics to gain a competitive advantage in productivity, profitability and sustainable manufacturing processes for better products and better services. In order to exploit the potential value in data, it is of crucial importance that the data be processed and analysed so as to derive meaningful insights in order to achieve it set objectives.

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

1.2 Implementation of Data Analytics

Large-scale data (big data) refers to data that exceed the capability of traditional data processing technologies. This is differentiated from small data which is process by traditional technologies in many ways: the amount of data (volume), the rate of data generation and transmission (velocity), the types of data: structured, semi-structured and unstructured (variety), the trustworthiness of the data (veracity), value and complexity [8][9]. The rate of data creation has increased so much that ninety (90) per cent of the data in the world today has been created in the last two years alone. This acceleration in the production of data has created a need for new technologies to analyse these massive data sets.

The framework for implementing big data is the Apache Hadoop. This is a fast-growing platform that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer. Hadoop has capabilities to store and handle huge amounts of unstructured data within a smaller timeframe in an economically responsible way [9][10]. Two important parts of the Hadoop ecosystem are the Hadoop Distributed File System (HDFS) for making the partition of data and computing across many nodes possible; and the MapReduce which is the framework that understands and assigns work to the nodes in a cluster, and is able to distribute data workloads across thousands of nodes with machine learning as its core.

Traditional machine learning algorithms were designed to make machines recognize and understand the real world to learn a new knowledge and experience in limited dataset by some special customized methods. But this has become a challenging task for these algorithms to learn and analyse huge amount of data, complicated structure and wide range of varieties. The key limitation of traditional machine learning is that it can't efficiently generate complicated and non-linear patterns from raw input data [8][9][10]. Hence, deep learning (a variant of machine learning) is a very promising method to solve analytics problem in large-scale data.

Deep learning is based on learning multiple levels of representation of data features. The multiple levels of representation correspond to multiple levels of abstraction. A significant feature of deep learning, also the core of data analytics, is to learn high level representations and complicated structure automatically from massive amounts of data to obtain meaningful information [10]. This massive data provides training dataset for deep learning algorithms to learn more complicated features and to improve the state-of-the-art performance. Mining and extracting meaningful patterns from input data for decision-making, prediction, and other inferencing is at the core of data analytics [11].

1.3 Health Insurance Data Analytics

The major drive of the analytics of health insurance data is to track payments done by beneficiaries to providers for healthcare services, beneficiary contributions (premium), the cost of health services and address fraud as stated earlier [4]. Most works done in the analytics of healthcare insurance data has been in the area of combating fraud [12] and hence, there are serious issues with regards to delays in the processing of the data. In Nigeria, we have a peculiar situation; this is because the analytics in the NHIS is completely manual [7]. This has given rise to most of the issues we have in the scheme today. Fraud is one out of the many issues in NHIS and therefore a proper analytic tool should take into cognizance the other issues (delay in registration/update, payment, response to complaints, etc.) affecting the processing of data in addition to fraud.

As stated above, most researchers in this area focused on the detection of fraud in health insurance; for instance, [13] develop a data mining technique for fraud detection in health insurance scheme using knee-point k-means algorithm. They considered NHIS as the case study for the work. The research did not classify the fraud detected, whether it is provider, consumer or insurer frauds; it uses only the unsupervised technique (K-Means algorithm) for clustering; and the data was collected from only one HMO which cannot yield a perfect result. Consequently, in order to improve the health status of Nigerian via this scheme, the analytics tool should go beyond identifying and preventing fraud [14]. A crucial and peculiar issue in the Nigerian National Health Insurance Scheme is the high level of corruption in the sector, lack of accountability and clear sense of irresponsibility [5].

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

2 Methods and Materials

To achieve the objective of this paper, an investigation of the system was carried out via the review of existing works on data analytics and NHIS, observation, and data collection from different HMOs. The data analytics model which employed deep learning was designed using unified modelling language (UML) tools. The implementation of this model designed is intended to be done using Apache Hadoop and MySQL. The Apache Hadoop which comprised of Hadoop Distributed File System (HDFS) and MapReduce will be used for storing and analysing the data (i.e., the HDFS will be used for storing the data that will be collected from the various sources; and the MapReduce which is the Programming Platform will use Java technology to implement the data analytics model for the analysis of complex data in the health insurance scheme). The MySQL will be used to store processed data which will subsequently be made available for visualisation/reports. The model implementation is outside the scope of this paper

2.1 Study Site

In this research, we considered ten (10) HMOs in Nigeria and they include: Hygeia HMO Ltd, Total Health Trust Ltd, Clearline International Ltd, Healthcare International Ltd, Medi Plan Health Care Ltd, Multi Shield Nigeria Ltd, United Healthcare International Ltd, Premium Private Health Trust Ltd and Ronsberger Nigeria Ltd. These are accredited by NHIS.

2.2 Data Collection

Data was collected from stakeholders in the HMOs and listed in section 3.2. The data captured five (5) different aspects of healthcare insurance. First is the claims data submitted by different providers; second is the provider enrolment data; third is beneficiary enrolment data; fourth is a list of blacklisted providers that have been sanctioned for various dishonest behaviours with the scheme; and fifth is complaint/inquiry data. The claims, beneficiary enrolment and the provider enrolment data were gotten from transactional data warehouses. Each claim, consists of several data elements with information about the beneficiary, provider, the health condition (or diagnosis), the service provided (procedure or drug), and the associated costs. Note that the providers typically are affiliated to each other through organizations such as hospitals. This information and additional data about the providers is present in the provider data.

2.3 Data Format

The formats of data collected include: comma-separated values (CSV), spreadsheet, portable document format (PDF), scanned images (.png, .jpeg, .gif, .bmp, etc.), emails and email attachments, web contents, structured data from relational databases (Oracle, MySQL, PostgreSQL, MSSQL, etc.), text files, etc.

3 Results

Data analytics is not an isolated set of activities but a continuum; hence the model described a cohesive set of solutions for data analytics: from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems for ongoing analysis. It also shows an integrated architecture for data analytics which makes it easier to perform various types of activities and to move data among these components. Implementing this model will improve data management and reduces cost incurred in the processing of data in the NHIS. Figures 1, 2 and 3 shows different views of the model.

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

4 Discussions

The structure of the analytics model comprised of the input, the Hadoop (HDFS and MapReduce), the Relational Database Management System-MySQL, and the output components. With this arrangement, raw data generated from the various health facilities with varying data formats - comma-separated CSV, spreadsheet, PDF, scanned images (.png, .jpeg, .gif, .bmp, etc.), emails and email attachments, web contents, structured data from relational databases (Oracle, MySQL, PostgreSQL, MSSQL, etc.), text files, etc. are loaded into HDFS which is a scalable fault-tolerant distributed system for data storage. It can store any type of data – structured, semi-structured and unstructured. This data is less dense and more information poor at this unrefined stage, and is not fit immediately into the predefined relational database or data warehouse; and if forced into a relational database, valuable information will be lost in the process. This data is placed in a non-relational database-HDFS which stores data in files that accept varying formats of data. This explains why loading data into Hadoop can be faster than loading data into a relational database.

Within Hadoop, the data would be interpreted using the MapReduce platform. It processes the data into a structured manageable form which would be evaluated as well. The MapReduce implement the deep belief network and back propagation algorithms which aid in the data processing and subsequently storing it in MySQL. The output is fetched from the relational database technologies for business intelligence, decision support, reporting etc. This is the final stage which visualized the insights uncovered and these visualizations are automatically updated.

The architectural flow of the model as shown in figure 3 depicts a high level view of the model and how it processes the data. In step 0, data is extracted from the HDFS; and at step 0.1, the MapReduce engine is responsible for splitting the data extracted from HDFS. The engine then caches the split data for the subsequent MapReduce invocations. Every algorithm has its own engine instance, and every MapReduce task will be delegated to its engine (step 1). Similar to the original MapReduce architecture, the engine will run a master (step 1.1) which coordinates the mappers and the reducers.

The master is responsible for assigning the split data to different mappers, and then collects the processed intermediate data from the mappers (step 1.1.1 and 1.1.2). After the intermediate data is collected, the master will in turn invoke the reducer to process it (step1.1.3) and return final results (step 1.1.4). Processed results are stored in MySQL (step 3) and reports are generated based on the user's request (step 4) is used for business intelligence, decision support, reporting etc. This could visualize the insights uncovered and these visualizations are automatically updated.

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

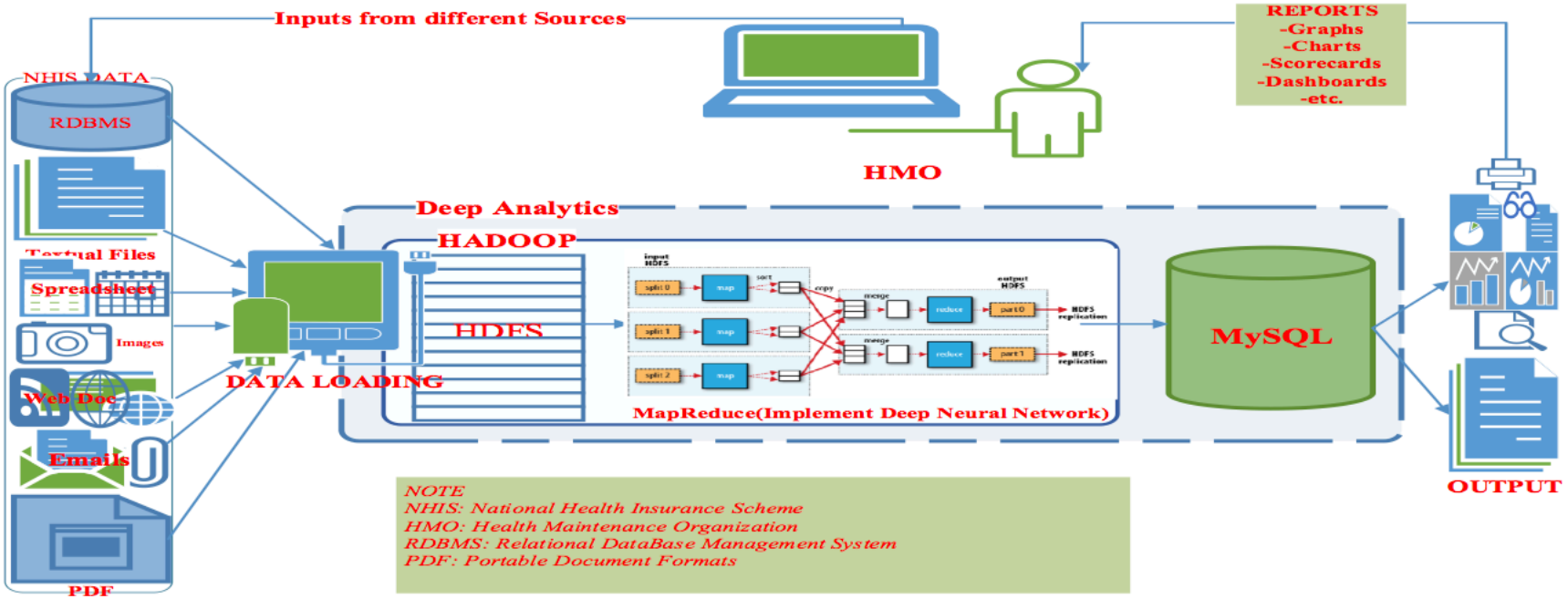

Figure 1. Conceptual View of the Model

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

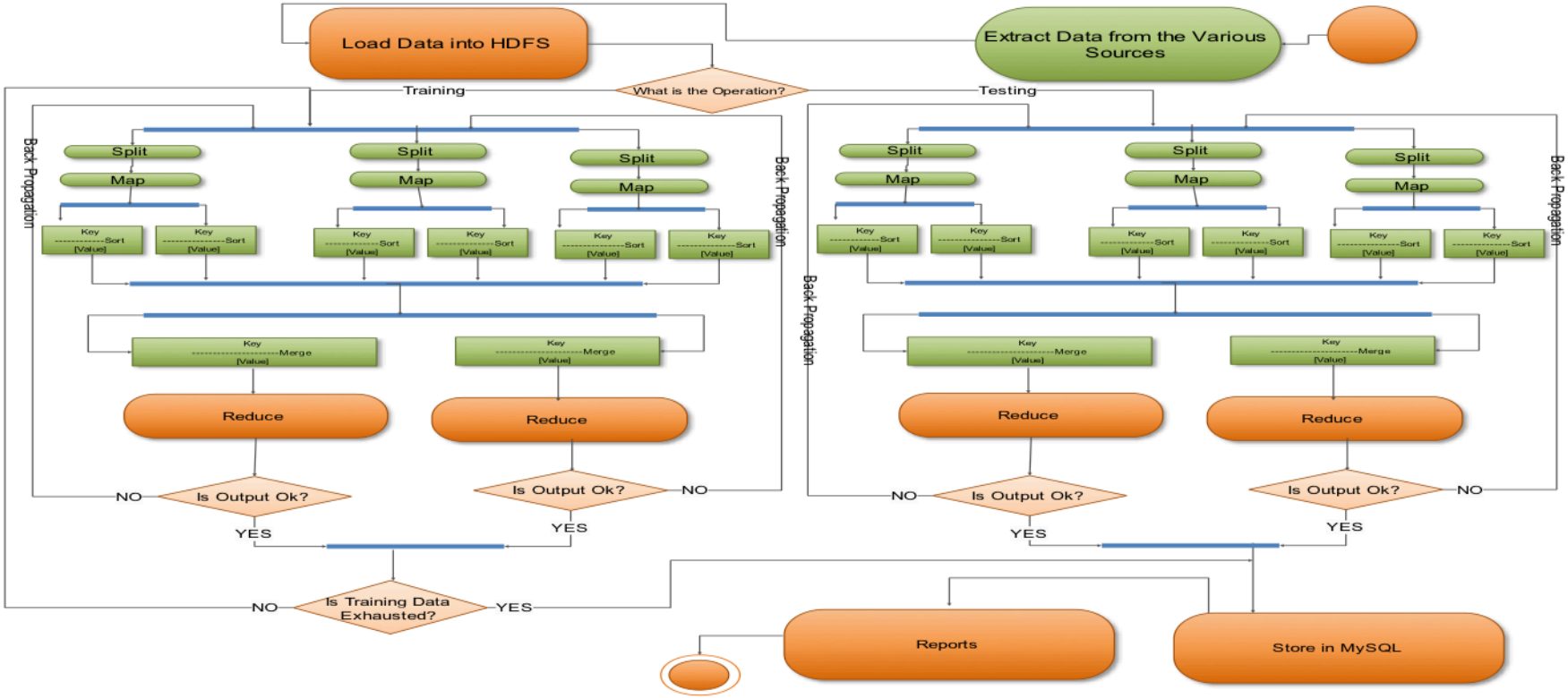

Figure 2. Activity Diagram of the Model

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

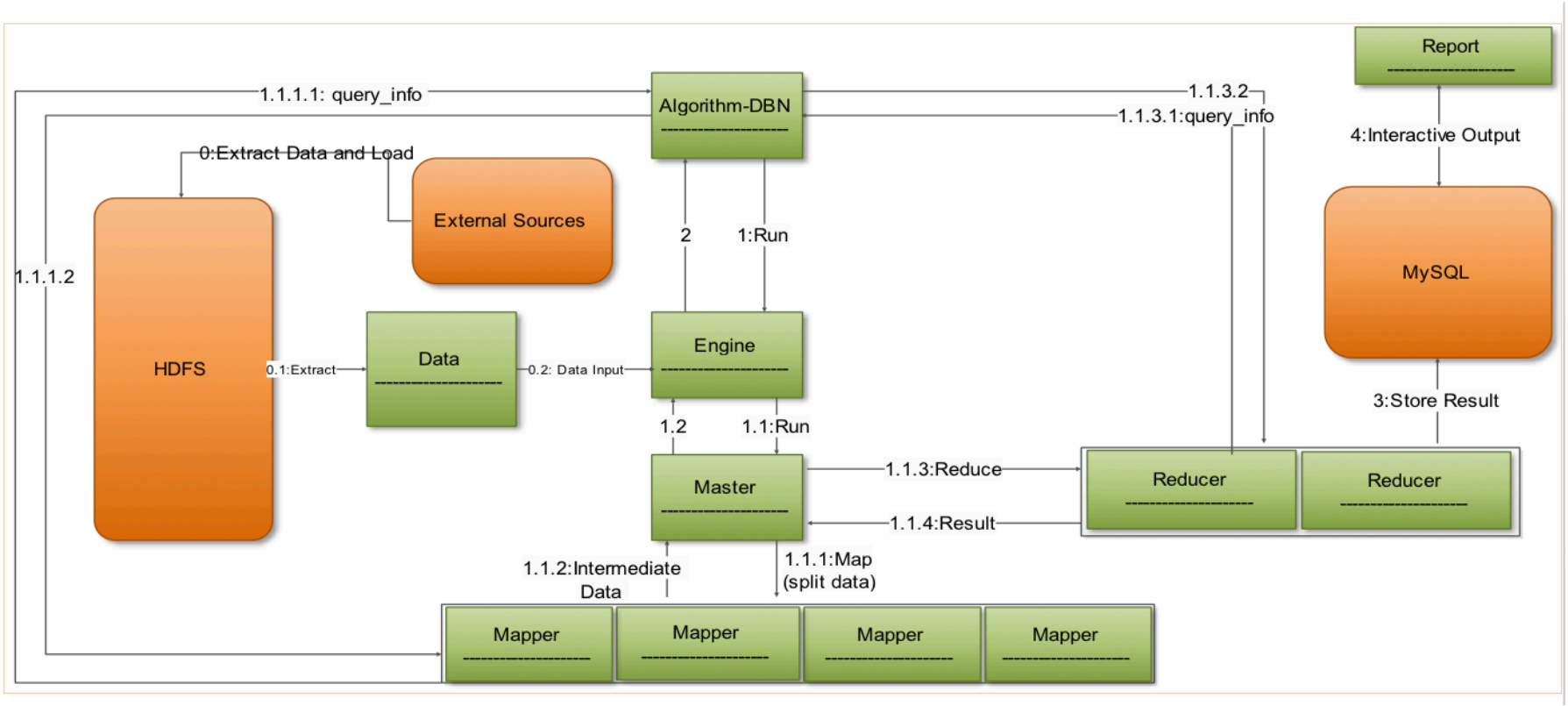

Figure 3. Architectural flow of the Model

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163

5 Conclusion

This paper designed a data analytics model based on deep learning techniques. This model when implemented will aid in the processing of data for NHIS. This will significantly reduce the time required to manage, query, and process data in NHIS. That is, it will reduce the time it takes in processing document of registered beneficiaries, claims submitted by providers, payment of claims and the remission of contributions to NHIS, HMOs and HSPs, which will also shrink the negative impact of significant losses owing to the delay. Thus improving the health status of Nigerian

References

[1] Meen V. Data analytics at work: Introducing a Data Mining Process Framework to Enable Consultants to Determine Effective Data Analytics Tasks. Unpublished MSc. Thesis submitted to the Faculty of Technology, Policy and Management Delft University of Technology: 2012.
[2] Hoyt RE, Yoshihashi A. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals Sixth Edition., FL; Pensacola: 2014.
[3] Sun K, Wei X, Jia G, Wang R, Li R. Large-scale Artificial Neural Network: MapReduce-based Deep Learning. NerveCloud in the course of EEL-6935: Cloud Computing and Storage, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, 32611 USA. 2015; 1-9.
[4] Agba MO, Ushie EM, Osuchukwu NC. National Health Insurance Scheme (NHIS) and Employees' Access to Healthcare Services in Cross River State, Nigeria. GJHSS. 2010; 10(7): 9-16.
[5] Eteng FO, Ijim-Agbor U. Understanding the challenges and prospects of administering the national health insurance scheme in Nigeria. IJHSSR. 2016; 2(8): 43-48.
[6] Oyegoke TO. Development of an Integrated Health Management System for National Health Insurance Scheme. An Unpublished M.Sc. Thesis Submitted to the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria: 2015.
[7] Chen CLP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. JIS. 2014; 275: 314–347.
[8] Zhang K, Chen X. Large-Scale Deep Belief Nets with MapReduce. IEEEA-PI. 2014; 2: 395-403.
[9] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep Learning Applications and Challenges in Big Data Analytics. JBD. 2015; 2(1): 1-21.
[10] Leelavathi MV, SahanaDevi KJ. Efficient Deep Learning for Big Data: A Review. IJCSE. 2016. 4(3), 30-35.
[11] Bagul PD, Bojewar S, Sanghavi A. Survey on Hybrid Approach for Fraud Detection in Health Insurance. IJIRCCE. 2016; 4(4): 6918-6922.
[12] Fashoto SG, Owolabi O, Sadiku J, Gbadeyan JA. Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm. AJBAS. 2013; 7(8): 140-144.
[13] Moyano LG, Appel AP, de Santana VF, Ito M, dos Santos TD. GraPhys: Understanding Health Care Insurance Data through Graph Analytics. WWW '16 Companion, April 11–15, 2016; Montréal, Québec, Canada, 2016. 227-230.
[14] Yunusa U, Irinoye O, Suberu A, Garba AM, Timothy G, Dalhatu A, Ahmed S. Trends and Challenges of Public Health Care Financing System in Nigeria: The Way Forward. IOSR-JEF. 2014; 4(3): 28-34.

 J Health Inform Afr. 2017;4(1):42-50. DOI: 10.12856/JHIA-2017-v4-i1-163