A Model Design of Big Data Processing System using Hace Theorem

Anthony Otuonye I. Ph.D*, Udunwa Ikenna A.1 and Nwokonkwo Obi C. Ph.D1

1Department of Information Management Technology, Federal University of Technology Owerri, Nigeria.

*Corresponding author email id: tnonyeotuonye@yahoo.com

Abstract – In this research paper, we have developed a new big data processing model using the HACE theorem to fully harness the potential benefits of the big data revolution and to enhance socio-economic development of in developing countries. The paper proposes a three-tier data mining structure for big data storage, processing and analysis from a single platform and provides accurate and relevant social sensing feedback for a better understanding of our society in real-time. The whole essence of data mining is to analytically explore data in search of consistent patterns and to further validate the findings by applying the detected patterns to new data sets. Big Data concern large-volume, complex, and growing data sets with multiple, autonomous sources. Our data-driven model involves a demand-driven aggregation of information sources, mining and analysis to overcome the perceived challenges of the big data. The study became necessary due to the growing need to assist governments and business agencies to take advantage of the big data technology for the desired turn-around in their socio-economic activities. The researchers adopted the HACE theorem in the model design which characterizes the unique features of the big data revolution. The Hadoop’s MapReduce technique was also adopted for big data mining, while the k-means and Naïve Bayes algorithm was used to ensure big data clustering and classification. By this model, the suggestions of various IT scholars and authors has been achieved who observed the need to revisit most of our data mining techniques and suggested distributed versions available methods of data analysis due to the new challenges of big data.

Keywords – Big Data, HACE Theorem, Data Mining, Open Source, Real-Time.

I. INTRODUCTION

1.1. The ‘Big Data’ Concept

Big Data is a new terminology that describes the availability and exponential growth of data, which might be structured, semi-structured and unstructured in nature. Big data consists of billions and trillions of records which might be in terabytes or petabytes (1024 terabytes) or exabytes (1024 petabytes). According to [11], the amount of data produced in our society today can only be estimated in the order of zettabytes, and it is estimated that this quantity grows at the rate of about 40 percent every year. The presidential debate between former President Barack Obama and Governor Mitt Romney on 4 October 2012, for example, triggered more than 10 million tweets in 2 hours [6], which exposed public interests in healthcare among other things.

Different researchers have defined big data in several ways. According to [10], big data is defined as the large and ever-growing and disparate volumes of data which are being created by people, tools and machines. Both structured and unstructured information are being generated on a daily basis from a number of sources including social media, internet-enabled devices (such as smart phones and tablets), machine data, and video and voice recordings. According to [10] also, our society today generates more data in 10 minutes than all that all of humanity has ever created through to the year 2003; and as [7] puts it, “… 90% of all data in the world today were created in the last two years”. Big data has so grown and continues to grow exponentially beyond the capability of our commonly used software tools, which can no longer capture, process and manage data within the tolerable time space.

The Big data revolution is on and corporate organizations around the world can leverage on this revolution for global economic recovery. Big data can provide business owners and corporate managers with innovative technologies to collect and analytically process the vast data to derive real-time business insights that relate to such market forces as consumers, risk, profit, performance, productivity management and enhanced shareholder value.

Many IT scholars have opined that most traditional restraints of the relational database system can be overcome by the big data technology in a cost-effective manner, which have opened opportunities for storage and processing of all types of data coming from diverse sources.

Apart from the inability of the traditional system to manage disparate data, it is equally very difficult to integrate and move data across organizations which is traditionally constrained by data storage platforms such as relational databases technology and batch files, which has limited ability to process very large volumes of data, data with complex structure (or with no structure at all), or data generated at very high speeds.

With the recent growth in technology, coupled with the ability to harness and analyze disparate volumes of data, and the increased statistical and predictive modeling tools available for today’s business, big data will no doubt bring about positive changes in the way businesses compete and operate.

Today, the world is in need of technologies that can provide accurate and relevant social sensing feedback to better understand our society in real time. In some cases also, the knowledge extraction process has to be very efficient and close to real time since it is almost infeasible to store all observed data. The unmatched data volumes require an effective data analysis tool and a prediction platform to achieve fast response and real-time classification for data.

In this research paper therefore, we propose a big data processing model from the data mining perspective. Our model will make use of HACE theorem which characterizes the features of the Big Data revolution. The challenges of big data are broad in terms of data access, storage,
searching, sharing, and transfer. Therefore, these challenges will be born in mind as we make our propositions in this research paper.

1.2. Aim and Objectives of Study

This research paper aims at developing a Big Data Processing Model using the HACE theorem to enhance socio-economic activities in developing countries. The study will seek to achieve the following specific objectives:

i. Use HACE theorem to fully explain the characteristics and features of the Big Data revolution.

ii. Propose a Big Data processing system using the HACE theorem and a three-tier data mining structure to provide accurate and relevant social sensing feedback to better understand our society in real-time.

iii. Make recommendations for a way forward.

II. LITERATURE REVIEW

2.1 Scholarly Views on Existing Big Data Processing Systems and Models

A good insight into the various models of big data systems was given by [11], in their paper titled “Data Mining and Information Security in Big Data Using HACE Theorem. They equally presented a model of big data from the data mining perspective but focused generally on security and privacy issues in big data mining using the AES algorithm. The table 2.1 presents the achievements and scholarly views of other researchers in the area of big data technology.

Table 2.1. Comparative study of various big data processing models

<table>
<thead>
<tr>
<th>S/n</th>
<th>Technique used</th>
<th>Description</th>
<th>Drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>AES Algorithm [11]</td>
<td>Used AES Algorithm to model big data security systems with a focus on data mining and privacy issues.</td>
<td>System could not provide accurate and relevant social sensing feedback in real-time.</td>
</tr>
<tr>
<td>3.</td>
<td>Parallelization strategy [2]</td>
<td>Used SVM algorithm, NNLS algorithm, LASSO algorithm to convert the problems into matrix-vector Multiplication.</td>
<td>Suitable for medium scale data only, and no form of security was provided,</td>
</tr>
<tr>
<td>6.</td>
<td>Decision Tree Learning [15]</td>
<td>Converts original sample data sets into a group of unreal data sets, from where the original samples cannot be reconstructed without the entire group of unreal data sets.</td>
<td>Centralized, Storage Complexity, and Privacy Loss.</td>
</tr>
</tbody>
</table>

2.2 Historical Background and the Five ‘Vs’ of Big Data

Though big data is a relatively new concept but the operations involved in data gathering, storage, and analysis for value-added business decisions are not new. In fact, data gathering and analysis has been there from the time of that human beings began to live together and engage in socio-economic activities. Nevertheless, the big data concept began to gather momentum from 2003 when the ‘Vs’ of big data was proposed to give foundational structure to the phenomenon which gave rise to its current form and definition. According to the foundational definition in [10], big data concept has the following five mainstreams: volume, velocity, value, variety and veracity.

Volume:

Organizations gather their information from different sources including social media, cell phones, machine-to-machine (M2M) sensors, credit cards, business transactions, photographs, videos recordings, and so on. A vast amount of data is generated each second from these channels, which have become so large that storing and analyzing them would definitely constitute a problem, especially using our traditional database technology. According to [10], facebook alone generates about 12 billion messages a day, and over 400 million new pictures are uploaded every twelve hours. Users’ comments alone on issues of social importance are in millions. Opinions of product users generated by pressing the “Likes” button are in their trillions. Collection and analysis of such information have now become an engineering challenge.

Velocity:

By velocity, we refer to the speed at which new data is being generated from various sources such as e-mails, twitter messages, video clips, and social media updates.
Such information now comes in torrents from all over the world on a daily basis. The streaming data need to be processed and analyzed at the same speed and in a timely manner for it to be of value to businesses and the general society. Results of data analysis should equally be transmitted instantaneously to various users. Credit card transactions, for instance, need to be checked in seconds for fraudulent activities. Trading systems need to analyze social media networks in seconds to obtain data for proper decisions to buy or sell shares. Big data technology gives the ability to analyze data while it is being generated, and has the ability to move data around in real-time.

**Variety:**

Variety refers to different types of data and the varied formats in which data are presented. Using the traditional database systems, information is stored as structured data mostly in numeric data format. But in today’s society, we receive information mostly as unstructured text documents, email, video, audio, financial transactions and so on. The society no longer makes use of only structured data arranged in columns of names, phone numbers and addresses that fits nicely into relational database tables. According to [7], more than 80% of today’s data is unstructured. Big data technology now provides new and innovative ways that permit simultaneous gathering and storage of both structured and unstructured data [10].

**Value:**

Data is only useful if value can be extracted from it. By value, we refer to the worth of the data. Business owners should not only embark on data gathering and analysis, but understand the costs and benefits of collecting and analyzing such data. The benefits to be derived from such information should exceed the cost of data gathering and analysis for it to be taken as valuable. Big data initiative also creates an understanding of costs and benefits.

**Veracity:**

Veracity refers to the trustworthiness of the data. That is, how accurate is the data that have been gathered from the various sources? Big data initiative tries to verify the reliability and authenticity of data such as abbreviations and typos from twitter posts and some internet contents. Big data technology can make comparisons that bring out the correct and qualitative data sets. There are new approaches that link, match, cleanse and transform data coming from various systems.

2.3 Types and Sources of Big Data

There are basically two types of big data which are usually generated from social media sites, streaming data from IT systems and connected devices, and data from government and open data sources. The two types of big data are structured and unstructured data.

**Structured Data:**

These are data in the form of words and numbers that are easily categorized in tabular format for easy storage and retrieval using relational database systems. Such data are usually generated from sources such as global positioning system (GPS) devices, smart phones and network sensors embedded in electronic devices.

**Unstructured Data:**

These are data items that not easily analyzed using traditional systems because of their inability to be maintained in a tabular format. It contains more complex data in the form of photos and other multimedia information, consumer comments on products and services, customer reviews of commercial websites, and so on. According to [11]), sometimes, unstructured data is not easily readable.

2.4 HACE Theorem (Modeling Big Data Characteristics)

The HACE theorem is a theorem that is used to model big data characteristics. Big Data has the characteristics of being heterogeneous, large volume, autonomous sources with distributed and decentralized control, and a complex and evolving relationships among data. These characteristics pose enormous challenge in determining useful knowledge and information from the big data.
To explain the characteristics of big data, the native myth that tells the story of blind men trying to size up an elephant, usually come into mind. The big elephant in that story will represent the Big Data in our context. The purpose of each blind man is to draw useful conclusion regarding the elephant (of course which will depend on the part of the animal he touched). Since the knowledge extracted from the experiment with the blind men will be according to the part of the information he collected, it is expected that the blind men will each conclude independently and differently that the elephant “feels” like a rope, a stone, a stick, a wall, a hose, and so on.

To make the problem even more complex, assume that:

i. The elephant is increasing very quickly in size and that the posture is constantly changing

ii. Each blind man has his own information sources, possibly inaccurate and unreliable that give him varying knowledge about what the elephant looks like (example, one blind man may share his own inaccurate view of the elephant with his friend), and this information sharing will definitely make changes in the thinking of each blind man.

Exploring information from Big Data is equivalent to the scenario illustrated above. It will involve merging or integrating heterogeneous information from different sources (just like the blind men) to arrive at the best possible and accurate knowledge regarding the information domain. This will certainly not be as easy as enquiring from each blind man about the elephant or drawing one single picture of the elephant from a joint opinion of the blind men. The difficulty stems from the fact that each data source may express a different language, and may even have confidentiality concerns about the message they measured based on their country’s information exchange procedure.

HACE theorem therefore presents the key characteristics of Big Data to include:

a. **Huge with Heterogeneous and Diverse Data Sources**

Big data is heterogeneous because different data collectors make use of their own big data protocols and schema for knowledge recording. Therefore, the nature of information gathered even from the same sources will vary based on the application and procedure of collection. This will end up in diversities of knowledge representation.

b. **Autonomous Sources and Decentralized Control**

With big data, each data source is distinct and autonomous with a distributed and decentralized control. Therefore in big data, there is no centralized control in the generation and collection of information. This setting is similar to the World Wide Web (WWW) where the function of each web server does not depend on the others.

c. **Complex Data Relationships and Evolving Knowledge Associations**

Generally, analyzing information using centralized information systems aims at discovering certain features that best represent every observation. That is, each object is treated as an independent entity without considering any other social connection with other objects within the domain or outside. Meanwhile, relationships and correlation are the most important factors of the human society. In our dynamic society, individuals must be represented alongside their social ties and connections which also evolve depending on certain temporal, spatial, and other factors. Example, the relationship between two or more facebook friends represents a complex relationship because new friends are added every day. To maintain the relationship among these friends will therefore pose a huge challenge for developers. Other examples of complex data types are time-series information, maps, videos, and images.

### III. Methodology

This study will make use of the Data mining technique and the HACE theorem which characterizes the unique features of the big data revolution. Implementing the HACE theorem with data mining technologies will provide a model of big data that ensures accurate social sensing feedback and information sharing in a real-time fashion.

#### 3.1 Data Mining Technique

For mining in big data, the Hadoop's MapReduce technique will be used, while the k-means and Naive Bayes algorithm will be used for clustering and dataset classification. We shall consider the suggestions of [8] that observed the need to revisit most of the data mining techniques in use today and proposed distributed versions of the various data mining methods available due to the new challenges of big data.

#### 3.2 HACE Theorem Implementation

HACE theorem models the detailed characteristics of the Big Data which include: Huge with Heterogeneous and Diverse Data Sources that represents diversities of knowledge representation, Autonomous Sources and Decentralized Control similar to the World Wide Web (WWW) where the function of each web server does not depend on the other servers, and Complex Data Relationships and Evolving Knowledge Associations, which therefore suggests that each object is treated as an independent entity with consideration on their social connection with other objects within the same domain or outside.

A popular open source implementation of the HACE theorem is Apache Hadoop, which is equally recommended in this research paper. Hadoop has the capacity to link a number of relevant, disparate datasets for analysis in order to reveal new patterns, trends and insights, which is the most important value of big data.

#### 3.3 Challenges Facing Big Data Processing Systems

Big data systems face a number of challenges. The amount of data generated is already very large and increasing daily. The speed of data generation and growth is also increasing, which is partly driven by the proliferation of internet connected devices. The variety of data being generated is also on the increase, and current technology, architecture, management and methods of analysis are now unable to cope with the flood.

#### 3.3.1 Data Security and Privacy Concerns

Most Governments around the world are committed to protecting the privacy rights of their citizens. Many countries including Australia have passed the Privacy Act,
which sets clear boundaries for usage of personal information. Government agencies, when collecting or managing citizens’ data, are subject to a range of legislative controls, and must comply with a number of acts and regulations such as the Freedom of Information Act (in the case of Nigeria), the Archives Act, the Telecommunications Act, the Electronic Transactions Act, and the Intelligence Services Act. These legislative instruments are designed to maintain public confidence in the government as a secure repository and steward of citizen information. The use of big data by government agencies will therefore add an additional layer of complexity in the management of information security risks.

3.3.2 Data Management and Information Sharing

Every economy thrives by information, and no society can survive without access to relevant information. Government agencies must strive to provide access to information whilst still adhering to privacy laws. Apart from making data available, data must also be accurate, complete and timely for it to support complex analysis and decision making. Qualitative data will save costs, enhance business intelligence, and improve productivity. The current trend towards open data is highly appreciated since its focus is on making information available to the public, but in managing big data, government must look for ways to standardize data access across her agencies in such a way that collaboration will only be to the extent made possible by privacy laws.

3.3.3 Technological Initiatives

Government agencies can only manage the new requirements of big data efficiently through the adoption of new technologies. If big data analytics is carried out upon current ICT systems, the benefits of data archiving, analysis and use will be lost. The emergence of big data and the potential to undertake complex analysis of very large data sets is, essentially, a consequence of recent advances in technology.

IV. MODEL FORMULATION AND DISCUSSIONS

4.1 Our Proposed Big Data Mining Model

We begin by proposing a High Level Model that serves as conceptual framework for big data mining. It will follow a three-tier structure as shown in figure 4.1.

![Three-tier structure in big data mining](image)

**Fig. 4.1. Three-tier structure in big data mining**

4.1.1 Interpretation of Major Elements in Big Data Mining Model

**Tier 1:**

Tier 1 concentrates on accessing big datasets and performing arithmetic operations on them. Big data cannot practically be stored in a single location. Storage in diverse locations will also increase. Therefore, effective computing platform will have to be in place to take up the distributed large scale datasets and perform arithmetic and logical operations on them. In order to achieve such common operations in a distributed computing environment, parallel computing architecture must be employed. The major challenge at Tier 1 is that a single personal computer cannot possibly handle the big data mining because of the large quantity of data involved. To overcome this challenge at Tier 1, the concept of data distribution has to be used.

For processing of big data in a distributed environment, we propose the adoption of such parallel computing models like the Hadoop’s MapReduce technique [1].

**Tier 2:**

Tier 2 focuses on the semantics and domain knowledge for the different Big Data applications. Such information will be of benefit to the data mining process taking place at Tier 1 and to the data mining algorithms at Tier 3 by adding certain technical barriers and checks and balances and data privacy mechanisms to the process. Addition of technical barriers is necessary because information sharing and data privacy mechanisms between data producers and data consumers can be different for various domain applications. [4].

**Tier 3:**

Algorithm Designs take place at Tier 3. Big data mining algorithms will help in tackling the difficulties raised by the
Big Data volumes, complexity, dynamic data characteristics and distributed data. The algorithm at Tier 3 will contain three iterative stages: The first iteration is a pre-processing of all uncertain, sparse, heterogeneous, and multisource data. The second is the mining of dynamic and complex data after the pre-processing operation. Thirdly the global knowledge received by local learning matched with all relevant information is feedback into the pre-processing stage, while the model and parameters are adjusted according to the feedback.

4.2 Overall Design Structure of the Proposed System using HACE Theorem

The overall design structure of the proposed data mining system is shown in figure 4.2, which depicts the various activities involved in the mining process of big data.

4.3 Flowchart of the Proposed System

The flowchart of the proposed system is presented in figure 4.3.

4.4 Discussions

Figure 4.2 shows the proposed big data processing model that provides accurate and relevant social sensing feedback for a better understanding of our society in real-time. The model shows all the phases of the big data mining concept. The major elements of the system include: Big data sources, Admin, User interface, the Hadoop’s MapReduce system, and the Hadoop’s K-means and Naive Bayes algorithm.
4.4.1 The Admin
The Admin will be responsible for querying the system based on request. He will interact directly with the graphical user interface to supply his needs, while the query will be processed by the Hadoop System.

4.4.2 The Hadoop’s MapReduce Program
The MapReduce program is a programming model used to separate datasets and subsequently sent as input into independent subsets. It contains the Map() procedure which performs filtering and sorting on datasets. It also contains the Reduce() procedure that carries out a summary operation to produce output.

4.4.3 The K-means and Naive Bayes Algorithm
After the MapReduce operation, the output is transferred to the K-means or Naive Bayes algorithm to carry out clustering and classification of the datasets through an iterative procedure. The result is stored in separate data stores for easy access.

V. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion
In this research work, we have developed a new big data processing model using the HACE theorem in order to fully harness the potential benefits of the big data revolution and to enhance socio-economic activities in developing countries. The research also proposed a three-tier data mining structure for big data that provides accurate and relevant social sensing feedback for a better understanding of our society in real-time. The whole essence of data mining is to analytically explore data in search of consistent patterns and to further validate the findings by applying the detected patterns to new data sets.

5.2 Recommendations
Based on the big data model design in this research, and the subsequent discussions on the model, a full implementation of the ideas is highly recommend especially in developing countries that are yet to embrace the new big data technology. The benefits of the big data revolution are enormous and have the capacity to enhance economic activities in these countries. There are available big data technologies with affordable, open source, and distributed platforms (such as the Hadoop), and relatively easy to deploy.

REFERENCES

AUTHOR’S PROFILE
Dr. Anthony Otuonye
Dr. Anthony Otuonye obtained a Bachelor of Science Degree in Computer Science from Inno State University Owerrri, Nigeria, and a Master of Science Degree in Information Technology from the Federal University of Technology Owerrri. He also holds a Ph.D. in Software Engineering from Nnamdi Azikiwe University Awka, N-igeria. Currently, Dr. Anthony is a Lecturer and Researcher with the Department of Information Management Technology, Federal University of Technology Owerrri, Nigeria. He is a member of some professional bodies in his field and has authored several research articles most of which have been published in renowned international journals. He is married to Mrs. Chidimma Edith Otuonye and they live in Owerrri, Ino State of Nigeria.

email id: anthony.otuonye@futo.edu.ng; +234(0)8060412674