Redefining the Concept of Big Data: A Ghanaian Perspective

Eleanor Afful[1], Kofi Sarpong Adu-Manu[2], Grace Gyamfuah Yamoah[2], Jamal-Deen Abdulai[2], Nana Kwame Gyamfi[2], Edem Adjei[2], Isaac Wiafe[2] and Ferdinand Apietu Katsriku[2*]

[1]KACE AITI
[2]Department of Computer Science, University of Ghana
*Corresponding author: fkatsriku@gmail.com

ABSTRACT

The world is on the verge of a data tsunami. Voluminous amounts of unstructured data are being generated using different technologies. To manage the huge amounts of data being generated, a new concept of 'Big Data' has evolved. The emergence of 'Big Data' is leading to real transformation in the business world. Governments and commercial enterprises on the African continent are beginning to take an interest in the use of technologies associated with Big Data for the analysis of enormous amount of data they currently generate and they wish to do so in real time. The advances being made in big data technologies have fuelled this uptake. Until recently companies in Ghana did not realize the utility of big data analytics due mainly to lack of knowledge and the limited penetration of these technologies. Increasingly, however, these companies now realize the difference in value that data analytics could make to their decision making process and to develop strategies that will give them competitive advantage. It has become clear to many of these corporate organizations that they are in possession of large volumes of data which, if properly analysed, can provide them with a wealth of knowledge to run their businesses more efficiently and productively. The analytic necessary to the understanding of these wealth of data are provided by big data technologies. This paper seeks to redefine the concept of big data and reviews its development, the potential impact that big that can have on a developing economy, the sectors of the economy of Ghana that stand to gain most from adoption of big data technologies and how these can be achieved. We propose that big data concept be defined more objectively by the use of a function. The paper shows how big data can be leveraged for rapid economic advancement. The paper additionally examines the investment prospects of adopting big data technologies for the economic environment of Ghana and some of the issues that organizations must resolve to successfully implement the technologies in Ghana.

Keywords: Big Data, Analytics, Economic development, Big Data Architectural, Ghana

Introduction

Traditional data modelling and organization methods have proved useful and appropriate for varied functions in the past few decades and this is attested to by the phenomenal success of relational database systems. These traditional methods are coming under huge strain with the exponential growth in data and in most cases the traditional systems are unable to cope. This growth in data has been fuelled in particular by the success of internet companies such as Google and Facebook.

Huge volumes of information are still available and unexploited because the existing data modelling and management tools are not well suited to handling such information. Data aggregation, transforming data scattered across multiple sources into a new summary, is one of the key features used in databases, especially for business intelligence (e.g. extract, transfer and load (ETL), online analytical processing (OLAP) and analytics /data mining).

For databases built on Structured Query Language (SQL), aggregation is used to prepare and envision data for a more profound level of analysis. Such an operation is however difficult and often impossible to perform on

enormous volumes of data in terms of the memory and time requirements (memory-and-time-consumption).

Database maintenance and optimization is a key activity for relational databases. As the number of queries from across multiple sources increases, optimizing query execution becomes difficult to handle. For databases bigger than relational ones, a key requirement is that they be maintained and optimized for continuous optimal performance; such a task thus becomes less than trivial.

Additionally, the data residing in the database must be highly structured and cleansed. Businesses spend significant effort to extract, transform and load the data between data warehouse and relational databases. Enormous costs are involved in doing these and greatly limits the breadth of data available for analysis. The current systems are not easily scalable and do not scale up to the combined increase in velocity, volume and variety as defined for big data.

This paper proposes a more objective definition of the concept of big data, looks briefly at its development and what impact these technology can have on a developing economy. The paper examines those sectors of the Ghanaian economy which could possibly benefit the most from application of big data technologies and how the aforementioned benefits might be achieved, based on new architectures.

Background

Initially Big data is normally defined using the three V's, Volume, Velocity and Variety. However of late, two other parameters have been included: Veracity and Value. The big data concept may then be depicted as shown in Figure 1. Variety is assured through the numerous and diverse data sources, each generating some quantity of data per unit time to the data volume. The amount of data generated per unit time may not be static but dynamic and subject to change over time. The data being generated from these sources may either be structured or unstructured. Volume is the summation of all the data coming from diverse sources per unit time and arriving at one processing centre. In the literature volume has been defined as how much data there are, velocity, the rate at which new data are created and how quickly the data are processed and variety is defined in terms of the format of the data, whether structured, semi-structured or unstructured.

The two new dimensions are veracity, used to refer to the trustworthiness of the data and value, which refers to what gain businesses' can derive from the data have been added lately. Other dimensions have also been used, notably volatility and validity; however these have not gained widespread acceptance and use. It is noteworthy that these definitions only provide a qualitative view of what is described as big data. Some researchers have sought to define 'big data' in terms of a fixed volume such as petabytes or zettabytes. There is however no consensus on exactly what quantity of volume would constitute big data. Velocity may be defined as the rate of change in the volume of data generated and transferred to the enterprise office. Value is derived from the analytics performed on the data. Ultimately, the volume and velocity are intimately linked to the processing capacity of the system under consideration and hence the business needs. Assigning a numerical value to what will qualify in terms of volume as big data is thus not very useful. What may qualify as big for one enterprise may not be so for another enterprise. A helpful definition will be "when the data arriving begins to exceed the processing capacity of the conventional database and data warehouse solutions available to an enterprise". For many businesses therefore, big data becomes a moving target as they need to constantly evolve new solutions for the data they process. The volume of data will depend on the rate at which new data are being generated and the rate at which there are arriving. As such volume and velocity are intricately linked. What has not been mentioned as far as velocity is concerned is the rate at which the data are being processed. This constitutes another aspect of velocity not intricately linked to volume. Even though it may be argued that the processing rate affects the volume of data yet to be processed, it does not affect the total volume of data an organization has.

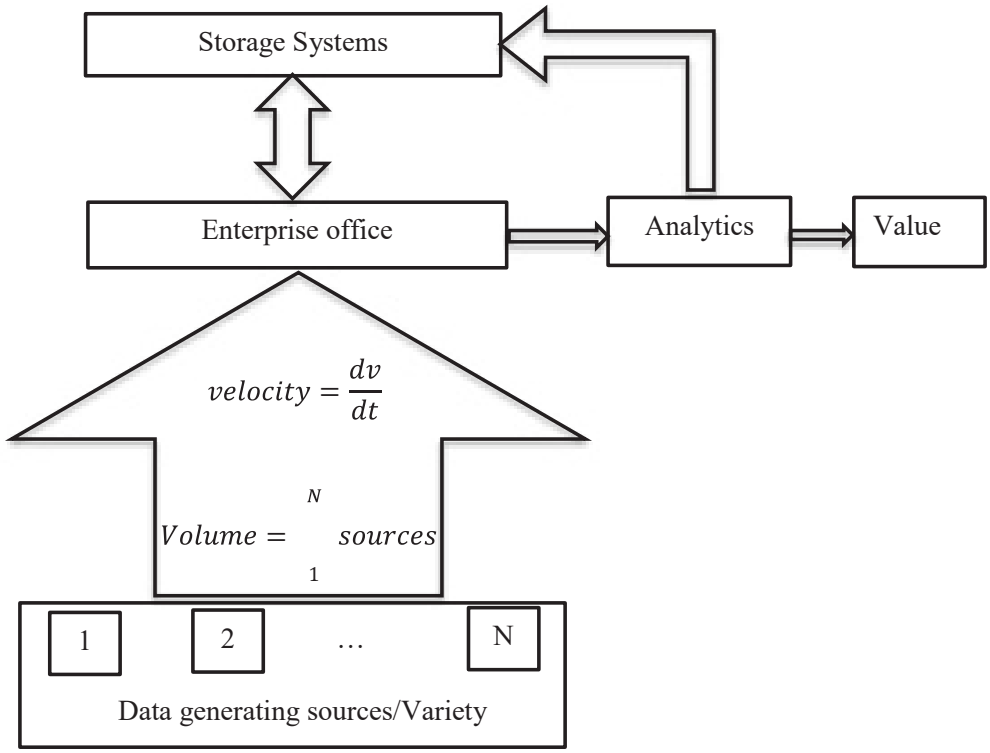

Fig. 1: Big data concept and interaction between the v's

Veracity of data is also not a new concept. Businesses have always sought to ensure the quality of data they obtain. What is new is the fact that with big data, a substantial proportion of the data is from external sources and hence additional measures now need to be taken to ensure the quality of the data. This, coupled with the wide variety of data, necessitates new requirements. Data cleansing thus becomes an additional and important element of the processing cost.

Creating business value out of data has always been a driving motivation for many businesses, so this is not a new concept. What provides a business with a competitive advantage is what insights it can derive from its data warehouses to enable it make better, real time, and smart decisions. This implies a real need for detailed and in-depth knowledge and hence the demand for new analytical tools.

The key new dimension now is therefore variety. Businesses have always been interested in data for the value that may be derived from insights such data provide on analysis. When the data grow beyond their current processing capability, they simply upgrade their systems to more powerful ones and in most instances the same relational database technology. As such, growing data volumes has always been an issue that businesses have to contend with. The rate at which new data are being generated and arriving and the variety of formats the data take are hence the key issues. Whilst in the past businesses had to deal with data of the same type and form, the explosion in new data sources mean that data are constantly being generated from new sources and taking different forms. Businesses now have to integrate different forms of data; unstructured, graph data, voice, images, video, etc. What is the relation between the value we derive from big data and the other dimensions namely volume, velocity and variety? It is proposed that a relation be established between the value derived from big data and the other dimensions, namely volume, velocity and variety. In our opinion, it would be helpful to

view big data and the value obtained from it as a function dependent on the parameters of volume, velocity and variety. Such a relation may be expressed as a function dependent on the parameters of volume, velocity and variety.

Given some of these challenges, it is necessary to develop new technological means of managing, analysing, visualizing and extracting meaningful information from the large and complex heterogeneous data sets that are being generated from diverse and distributed sources. Progress in this area will enhance scientific innovation and provide new paths to scientific inquiry. The development of new data analytic tools and algorithms will also be an important outcome of any progress made. Other benefits will include the development of scalable data infrastructure and architecture which will ultimately lead to a better understanding of social processes and interactions for greater security, economic growth, and in general, an improved quality of life and the wellbeing of people.

To efficiently model the data requirements for business intelligence and analytics, a new technology has emerged, called NoSQL (Non-Structured Query Language), a distributed non-relational database with variations in implementation. NoSQL was designed to cater for demands of data that were being generated by the web. The origins of NoSQL can be traced to the work done at Google in building a proprietary database, Big Table. The Big Table was designed to overcome some of the inherent limitations of Relational Database Management Systems (RDBMS). Some of these limitations are the need for specialized and robust servers that were less prone to failure; the length of time required to process queries; and the need to have structured data. Since this early work, many companies have also turned their attention to developing such systems that promise low implementation cost in terms of the hardware requirements and the ability to massively scale up horizontally i.e. by adding thousands of nodes so that storage and retrieval would be distributed across them using parallel processing techniques, thereby reducing storage and retrieval time. All these allow the setting up of server farms very quickly and cheaply. These systems also have a high level of fault tolerance since the same data are stored on different machines and they do not have any limitations on structure, which means one could store almost anything together.

With NoSQL it is possible to efficiently and cost-effectively build massive computing systems capable of handling the exponential growth in the volume of large data sets as is currently experienced. Independent of its format structured, unstructured or semi-structured the techniques underpinning NoSQL, ensure that the limitations imposed by RDBMS on data size, format and speed are eliminated, leading to fast and efficient ways of processing and analysing variety of data in real-time, bringing real benefits to businesses.

With regards to the velocity with which new data are being generated, NoSQL has the capability to process terabytes and exabytes amounts of data in real time. The new techniques implemented in NoSQL process, extract, load and transform data in the database eliminating the need for the data to be transferred in and out of the database of the data warehouse. The advances in processing and storage capabilities in computing technology in the last decade with increased speed has effectively eliminated data size as a constraint.

Platforms and tools for big data

Even though arguments have been made to the effect that most of the data generated today are either unstructured or semi-structured, emerging big data technologies could be divided into two categories- structured and unstructured data. To handle structured big data, a number of customized technologies have emerged. These technologies are aimed at storing and retrieving the large amounts of data associated with big data. The Google File System (GFS) is an internet scale file system, a robust and scalable system which provides the sort of reliability required for internet applications. Object-store techniques aim to improve on redundancy and data availability. The Amazon Simple Storage System, OpenStack Swift and Nirvanix cloud storage are examples

of this approach. Underlying many of these solutions is the massive parallel processing (MPP) technique. MPP is based on a distributive processing architecture consisting of a series of nodes controlled by a master. When engaged the master distributes a query across the nodes for maximum processing efficiency. Similarly, the system can do autofast data import and export through the same underlying mechanism. Almost all the vendors operating in this domain use either software or hardware combined into a single compliance. This ensures consistency in the hardware and that is crucial to obtaining optimum performance. Data locality plays an important role in obtaining high performance in big data analytics. By processing the data as close as is possible to its generating source, we minimize the highly prohibitive costs of data transfers. MapReduce exploits the concept of data locality to give an improved performance. A variant of MapReduce is Hadoop. Hadoop is an open source implementation of MapReduce. It is based on the Hadoop Distributed File System (storage) with distributed processing architecture consisting of a series of nodes controlled by a master. When engaged the master distributes a query across the nodes for maximum processing efficiency. This programming paradigm allows for massive scalability across hundreds or thousands of servers in any Hadoop cluster of nodes much like Google core infrastructure, which requires different skills sets. Building analytic solutions requires knowledge of a new set of Application Programming Interfaces (API) and this is one major drawback. MapReduce is typically controlled by Java programming language, the term is used to refer to two separate and distinct jobs that Hadoop programs perform. In the first task the program does a mapping of input data and then processes it to produce key/value pairs. The reduce function takes those key/value pairs and then combines or aggregates them to produce the final results. The name, MapReduce gives a clue of the order in which the tasks are carried out, the reduce job is always performed after the mapping has been done. Combining the use of data warehousing, data mining and relational database alongside techniques such as optimization, simulation, visualization and predictive analysis for big data sets provide better strategies to obtain insight from massive data sets enabling better decision.

Implementation Areas of Big Data

Key/Value Store is a fundamental data model used for example in Hadoop, Voldemort, DynamoDB and Memcached. Key-value databases are lightweight, schema-less, relationship-less and transaction-less data stores used primarily for storing temporary data in memory. Examples of such formats of key value database used for very large scale storage systems include Riak, Redis and MemcachedDB. The key can be synthetic or auto-generated while the value can be String, JavaScript Object Notation (JSON), BLOB (basic large object) etc. The key value type basically, uses a hash table in which there exists a unique key and a pointer to a particular item of data. There can be matching keys in different containers which are made up of logical group of keys. Performance is enhanced to a great degree because of the cache mechanisms that accompany the mappings. Key/Value pairs however fail to offer ACID (Atomicity, Consistency, Isolation, Durability) capability, as they fail on consistency. This capability must be provided for by the application itself. To read a value one needs to know both the key and the bucket because the real key is a hash (Bucket + Key). The Key Value Store database model is popular because it is easily implemented. A major weakness of this scheme is that it becomes increasingly difficult to maintain unique value keys as the volume of data grows. To address this challenge, complex schemes are introduced to generate unique character keys for very large sets.

Document Oriented Database

The idea here is to aggregate the data, mainly in the form of key value pairs, this is then compressed into a searchable record format. XML (Extensible Markup Language), JSON (Java Script Object Notation) and BSON (which is a binary encoding of JSON objects) are some of the typical encoding schemes available. One significant distinction between a key-value store and a document store is that a document store has associated with it the attribute metadata related to the stored content. This provides a means of querying the data based on the stored content. Unlike traditional relational

databases where data and relationships are stored in tables, in this scheme they are simply a collection of documents independent of each other. Document style databases are schema-less and this makes a simple task of adding fields to JSON documents without having to first define the required changes. The most commonly used document-based databases are CouchDB, Apache and MongoDB. To store data CouchDB employs JSON with JavaScript as the querying language and MapReduce and Hypertext Transfer Protocol (HTTP) to implement the application programming interface.

Column Family Database

A column-family database provides the capability to organize the rows as groups of columns. This capability implies that each single row of a column-family database now has the capacity to contain several columns. All the columns which are related are grouped together as column families providing the capability to retrieve the columnar data for multiple entities. This is achieved through an iterative process. The flexibility that column family provides applications enables a wide range of complex queries and data analyses to be performed. This is reminiscent of the functionalities supported by a relational database. This design enables them to store massive volumes of data running into billions of rows with each row containing hundreds and possibly thousands of columns. Significantly, a column family database can still provide very fast access to these vast quantities of data due mainly to a most efficient storage mechanism. If a column-family database is well-designed then it will be fundamentally faster and have greater scalability than an equivalent relational database holding the same volume of data. This performance is achieved at a cost, it can only support a specific set of queries unlike the queries in a relational database which are more generalized. Designers of column-family databases must ensure that column families are designed optimizing for the most commonly use queries for the applications under consideration. In contrast, majority of relational Database Management Systems (DBMS) store their data in rows. Storing data in columns as done in column families has the benefit of allowing fast search/access as well as providing for efficient data aggregation. In relational databases a single row is stored as a continuous disk entry. As a result, different rows of data may be stored as different entries on the disk. On the other hand, columnar databases store all the cells which correspond to a particular column as a contiguous disk entry; this makes the search/access time much faster than can possibly be achieved in a relational database.

Graph Database

The graph database is the final variant of NoSQL database management systems that is considered in this work. Unlike the other models, the graph based DBMS models represent the data based on tree-like structures and using edges to connect the various nodes such as is used in graphs. Just as in mathematics, certain operations are much simpler to perform using these types of models. These databases are commonly used by applications where it is necessary or required to establish boundaries for connections. For example when you register on a social network of any sort, your friends' connection to you and their friends' relation to you are much easier to work with using graph-based database management systems. An example is Neo4J, the most widely used graph store apart from RDF (Resource Description Framework) triple stores.

Architectural (Conceptual) framework

In order for Ghana to leverage big data for economic development, a conceptual framework supporting the activities of all stakeholders (individuals, private and public sector) should be developed. This architectural or conceptual framework should take into consideration the role of companies or organizations, policy makers, institutions, and individual users towards the adoption of big data for economic development.

Several frameworks have been discussed in the literature (Manyika *et al.*, 2011; Wamba *et al.*, 2015; Global Pulse, 2012). We propose in this paper an architectural framework that can support the use and implementation of big data to boost the Ghanaian economy as depicted

in Figure 2. This framework seeks to point out the benefits of the use of big data in driving the economy of Ghana. From the framework, a collaboration between the public and private sectors in Ghana is a step towards an integrated economy and this can boost productivity significantly with the implementation of big data. Companies/industries in Ghana are expected to provide incentives to enhance the economy and also for users in the form of rewarding innovation. Big data analytics offer a huge economic impact for organizations (Gangadharan, 2014).

Policy makers produce and use data to facilitate enhanced policymaking processes, it is encumbered upon them to also play their part by promoting and fostering data-driven innovation and growth throughout economies (Andrade *et al.*, 2014). For big data to realize its potential in Ghana, innovation which is driven by advances in technology, policymakers need to articulate coherent guidelines, standards and policies on the use of data and the associated technologies. A possible way of achieving this objective is through openness and transparency; ensuring that using open data formats public data is accessible, promoting legislation which is balanced and which takes into consideration the competing needs of all sectors of the economy; and supporting education that focuses on equipping students with data science skills and competencies (Andrade *et al.*, 2014).

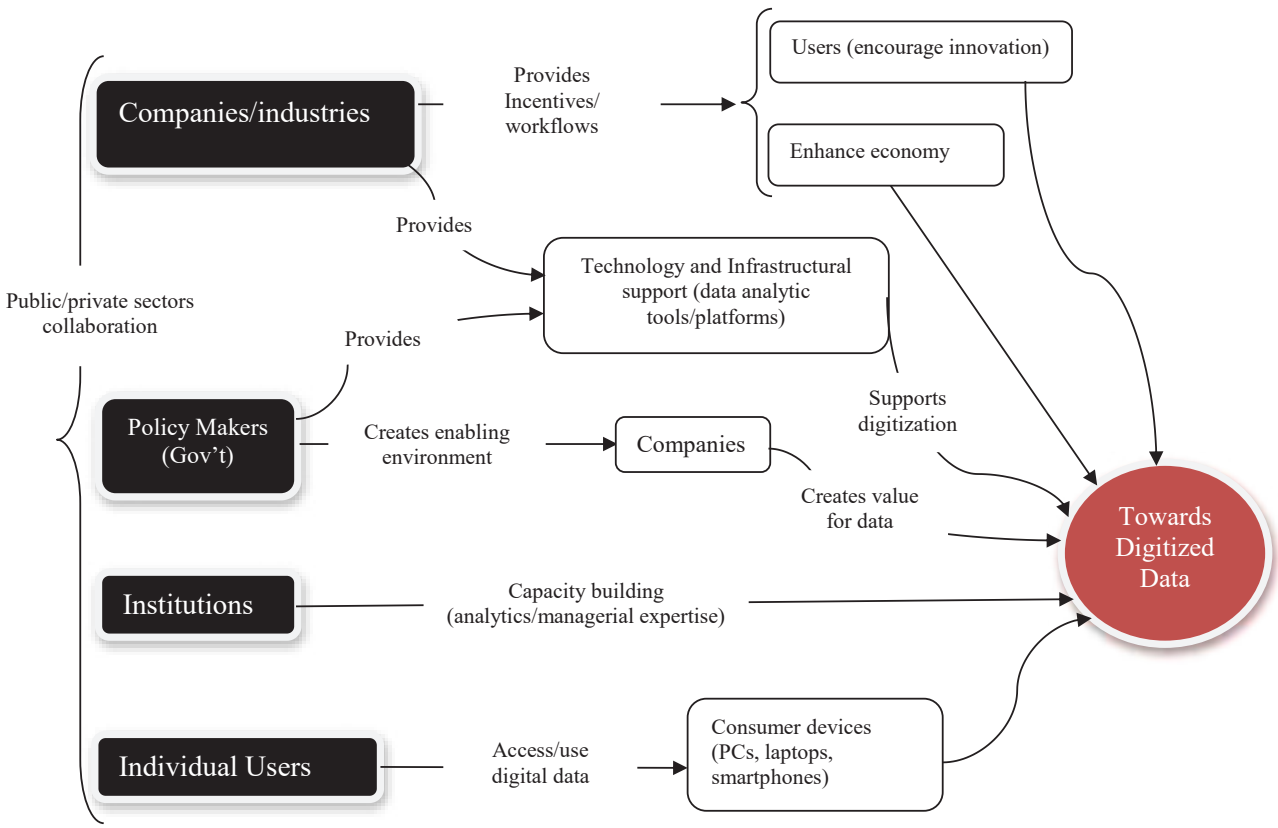

Fig. 2: Architectural framework for big data implementation in Ghana

Institutions also have a role to play in the progressive development and implementation of big data in Ghana. One of the major challenges is lack of expertise. Individuals with technical know-how in big data analytics are in short supply. According to Manyika *et al.*, (2011), a major limit on realizing value from big data is the acute shortage of skills and talent, especially of people with profound capability and proficiency in statistics and machine learning, and managers and analyst who know how to operate companies by using insights from big data. This places a huge responsibility on educational institutions and Information Technology training centres to build the right capacity for the nation to have a cream of talents ready to take up the task of analytics. This requires that the relevant departments be strengthened to enable them to fulfill those expectations. The role that Computer Science can play in transforming economies in the 21st century and beyond is well known and the arguments in support of this have been well made. This must however be driven from the highest echelons of government. A conscious effort must be made on the part of government to move towards an economy driven by advances in computing technology and hence the need to support computer science education in the country.

Global Pulse (2012) discussed the foremost apprehensions and challenges raised by big data for development and suggested probable ways of addressing some of them. They went ahead to discuss the sources of data for development in a growing environment such as Ghana. For a developing country like Ghana to benefit from the full potential of big data, these data sources should be taken into consideration. Most sectors of the economy in Ghana still depend to a great extent on paper-based record-keeping, as such, the data source is not automated and hence, easily digitized. The data sources can be digitally generated, passively produced, automatically collected, geographically accessible and continuously analysed (Global Pulse, 2012). These data sources are relevant for big data for economic development. Ghana generates massive amounts of digital data from different streams of the economy (online data) from different organizations and online platforms. This is however nowhere near what is possible. In a big data environment, the data are expected to be generated or created digitally such that they could be processed or manipulated by an electronic device. The recent Police and Fire Service e-Recruitment drive is an example of how such migrations could be achieved (Ghana Police Service 2016 e-recruitment, Ghana Fire Service e-recruitment 2016). The data produced should have the ability to interact with other digital services. They should be collected automatically after they have been produced. The location or time span for one operation should be available/accessible and should be analysed in real-time with no difficulty. Until the country meets some or all the sources of big data for development, Ghana will not be able to leverage big data for economic development.

Big Data and Ghana's Economy

The world has reached a stage where data are all around us. This data can be obtained from digital images, social media streams, financial and banking transaction records, wired and wireless sensors, GPS signals, and a myriad of other sources. Today, approximately 12 terabytes of data are generated from tweets alone on a daily basis. The flow is quickening and shows no signs of abating; with nearly 90% of the data in the world today created in only the last two years. We are truly facing a data tsunami; and there will be 44 times more of the data currently available by the year 2020 (Manyika *et al.*, 2011 as qtd by Gobble 2013). The advent of disruptive technologies such as Internet of Things plays a huge contributory role in the phenomenal growth in data that we now witness.

Recently, the Ghanaian economy has seen great boost in the emergence of companies and organizations that collect increasing amounts of digitized data from clients and employees. Some of these sectors of the economy are the oil and gas industry, healthcare industry, financial services (banks), telecommunication industries, government agencies, retail shops and other data driven businesses. In this paper, only a few of these areas will be discussed along with the potential impact of big data analytics. The increase in telecommunication network providers in the country is an indication that the

majority of Ghanaians have subscribed to one or more of these telecom networks. Ghana can take advantage of the opportunities big data offers that can be leveraged to create a better environment for its citizens and organizations.

In 2011, the Government of Ghana introduced some of its services online. The online services are made available at the Government of Ghana web portal. The web portal promises to serve as one-stop window for services and information offered by all Ministries, Departments and Agencies (MDA), MMDAs and other relevant government of Ghana agencies. The portal consists of four sub-portals, categorized as Citizens, Non-Citizens, Businesses and Governments as shown in Figure 3. This is a clear indication that substantial structured and unstructured data will be obtained by the government (Ghana Government e-Services Portal, 2011). Since Ghana performs these services online it is possible that most of these data that are being generated and collected are from different devices and are of different formats (photos, videos, text, audio, etc.). This makes the data unstructured.

In recent times, Ghana has seen a major shift from paper-based to electronic record keeping in most of the agencies and ministries. For example, recently the National Health Insurance Authority (NHIA) introduced biometric data collection for all clients on their scheme. Other agencies such as the National Identification Authority (NIA), the Electoral Commission (EC), the Ghana Education Service (GES), the Social Security and National Insurance Trust (SSNIT) and the Ghana Health Service (GHS) are all transitioning from the traditional data collection and progressing to electronic data processing and collection. Figure 4 depicts how these agencies access their individual databases for their day to day transactions.

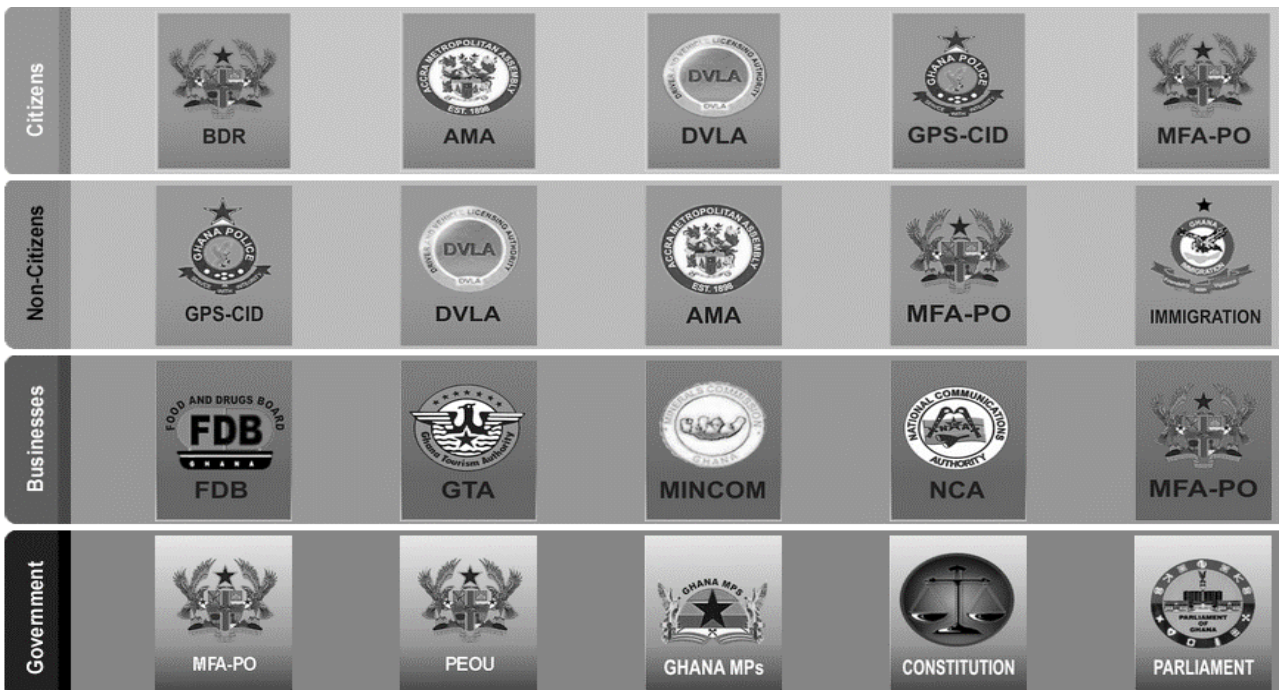

Fig. 3: e-Services Portal of the Government of Ghana

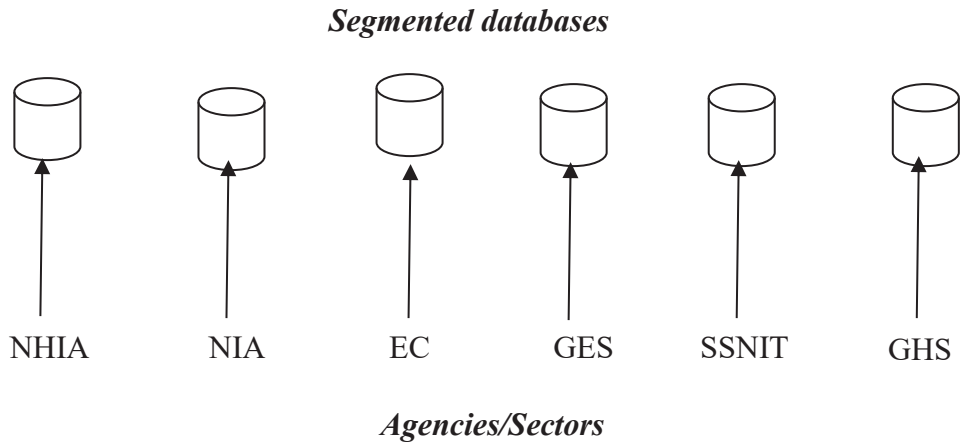

Fig. 4: Shift from paper based records to electronic records

One major challenge here is that these agencies do not share their data. Thus though they produce and collect a lot of data, they are sitting in data warehouses without being put to much use. Users are unable to gain remote access to this data, and when they do get access, there are no analytical tools for any meaningful information to be extracted from it. It is imperative that policies be put in place to regulate sharing and accessing data from a common platform within the public sector; the government could create high value for such data. Rather than look for solutions outside of the country, departments of computer science should be strengthened to provide the research leadership necessary for its realization.

The financial services arena has also seen a sharp increase in its online transactions. The use of mobile communication devices has become routine for personal communications and also for financial and business transactions including money transfer, job search, buying and selling of goods, as well as for the transfer of data such as school grades, examination results, stock levels and prices of various commodities, medical information (Global Pulse, 2012). The introduction of new technologies is helping to drive a wave of innovation across the African financial services sector as banks create new and accessible banking channels and take banking services to previously unbanked parts of society (Suresh, 2012). In the health industry, the introduction of biometric registration of patients and employees has attested to the data acquisition within that sector. Each of these agencies generate huge amount of data about their clients especially when there are transactions to be processed. A shift in the Ghanaian economy from paper-based to digital data is in the right direction. This shift cannot be accomplished solely on the basis of imported technologies but must be driven by local content achieved through research leadership by our academic institutions.

Healthcare

For any health information system to be effective it is important that it has access to all health data pertinent to the case under consideration in real time. In many countries of the world this data would come from many different and unconnected systems, Hammond *et al.* (2010). Lewis *et al.* (2012) reported that in low and middle income countries 42% of health institutions use ICT to extend geographic access to health care, whilst 38% use it to improve on data management. According to Raghupathi (2010) as citied by Raghupathi and Raghupathi (2014), the healthcare industry historically has generated large amounts of data, driven by record keeping. Ghana is not very different. The health industry

in Ghana generates millions of data records, but most of these are stored in hard copy form, whereas the current trend is toward rapid digitization of these large amounts of data. A number of health facilities are now moving towards digital records, but currently all these efforts are segmented and disjointed. To derive benefit from the digitization process, these efforts by individual facilities have to be coordinated and centralized (Asangansi and Braa, 2010). Effort must be made to implement an architectural platform onto which individual agencies can simply 'plug in and play'. Health policy makers in Ghana must provide vision and develop the required strategies necessary to achieving a fully integrated health information system for the country.

Financial Services (Banking)

There is convincing evidence that business has now recognized the ascendancy of data in the business sphere. In a survey recently conducted by Capgemini and *the Economist* of over 600 global business leaders, three-quarters of business leaders agreed that their organizations were data driven, and 90% of them, besides land, labour and capital, recognized information as the fourth factor of production (Gobble, 2013).

Ghana's banking sector has transformed from traditional walk-in and operate transactions to online and electronic banking operated venture where the presence of the customer is not really needed. The sector has expanded substantially over the last decade. The financial sector generates and stores massive amounts of data about customers. According to Suresh (2012), data from the banking industry indicate that banks in the Ghanaian markets spend up to10 percent of their operating income on data management.

In Ghana, despite the challenges in managing and securing customer data, Fidelity Bank, a mid-sized financial firm that has grown over the last ten years to become one of Ghana's leading financial institutions, has invested in a comprehensive, Big Data solution (Suresh, 2012).

Another key consideration associated with the significant growth in data volume of financial institutions is risk in the form of fraud. Constant vigilance and deterrence through technology is the key to protection and employing big data technology is a key measure to prevent attacks (Kothai, 2015). In implementing Big Data in the financial industry, the proposed framework in Figure 5 is proposed.

Challenges in implementing Big Data in Ghana

In practice, Big Data as a technology faces many challenges, one of which is heterogeneity and incompleteness. Since computer systems work most efficiently if they can store multiple items that are all identical in size and structure, the efficient representation, access and analysis of unstructured or semi-structured data poses analytical and storage difficulties. Another challenge is with the volume of data to be worked on within an organization. Managing large and rapidly increasing volumes of data can be challenging and requires that faster processing components and storage systems be designed and built. Also, with large data sets to be processed, speed could be an issue to deal with.

This is because the larger the data set to be processed, the longer it will take to analyse. Another challenge is privacy. For instance, there are strict laws governing what can and cannot be done with electronic health records. Big data raises concerns and fears regarding the inappropriate use of personal data, particularly through linking of data from multiple sources.

The implementation of Big Data in Ghana comes with its own challenges apart from the ones discussed above. Big Data has much potential for development. Big Data for development has been defined by Global Pulse (2013), to mean the identification of sources of big data relevant to policy and planning of development programmes. This concept is distinct from both "traditional" development data concept and what the private sector and mainstream media refer to as Big Data.

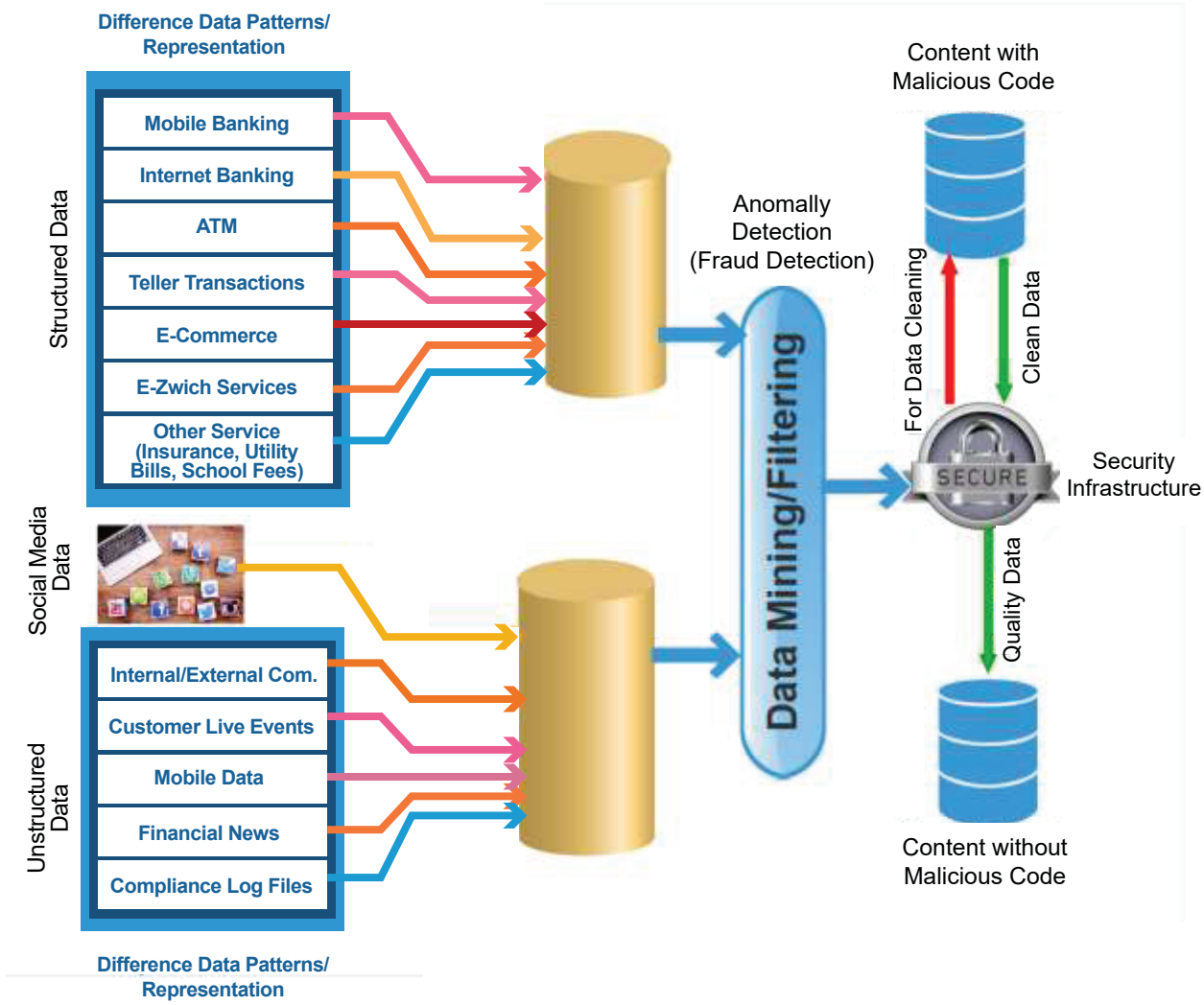

Fig. 5: Framework for big data implementation in the health sector

The lack of infrastructural support and the right technology is also a challenge to the implementation of big data in Ghana. Another challenge is the unavailability of the skilled personnel with the knowledge and skills related to big data analytics. The Commonwealth of Australia in 2013 suggested that since there is a shortage of university degrees that have a curriculum focused on big data analytics, it is important for education providers to design courses geared towards education and training in the area of big data scientists.

Research leadership has a key role to play in realization of the benefits big data has to offer for the Ghanaian economy. Research funding should be provided to educational institutions to run courses and training programmes aimed at producing the cadre of personnel with core skills to drive development in the field. Collaboration between government agencies and research/academic institutions will bring more opportunities for skills development and training and also the Government agencies should create procedures and practices that provide an enabling environment for responsible data analytics (Commonwealth of Australia, 2013).

Privacy and Security is an essential part of every society. Individuals and organizations have data that they protect. For anyone wishing to explore Big Data for development, privacy and security is a primary concern, since it has implications for all areas of work, from data acquisition and storage to retention, use and presentation (Global Pulse, 2013). Data anomalies are normally not detected at the early stages of data analysis and very often they are not discovered in real time. It is imperative to take note of the type of technology used in order to combat the anomalies.

Conclusion

This paper aimed to identify the potential for development and use of big data in current information administration in Ghana. The adoption of such practices by various institutions or organizations and government agencies in Ghana will enable these organizations to take full advantage of Big Data technologies. This will permit agencies to deliver better-quality and integrated services, improve policy development and identify new services and opportunities to make use of the national information assets, that is, Ghana government data and other data collected by the various agencies in the country. We have reviewed some of the technologies currently being used and proposed a functional definition of big data. We conclude that if the government harnesses the potential of big data to analyse data sets that are generated by the different agencies in the Ghanaian economy, this could improve government operations, policy development and service delivery for rapid economic development. There is the need also to strengthen research institutions to provide the leadership required to drive this effort.

References

Andrade P.L., Hemerly, J., Recalde G. and Ryan P.S. (2014). From Big Data to Big Social and Economic Opportunities: Which Policies Will Lead to Leveraging Data-Driven Innovation's Potential? The Global Information Technology Report – World Economic Forum. p.81-86

Asangansi I, Braa K. (2010). The emergence of mobile-supported national health information systems in developing countries. Stud Health Technol Inf 2010;160(Pt 1):540–4. [PubMed]

Commonwealth of Australia. (2013). Big Data Strategy — Issues Paper. Available at: https://www.finance.gov.au/files/2013/03/Big-Data-Strategy-Issues-Paper1.pdf

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. (2015). "How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study," *International Journal of Production Economics*. 165, p.234-246

Gangadharan J. (2014). 7 Ways to Leverage the Data Goldmine with Big Data and Analytics. *Global Practices, Domain Expertise, Customer Experience*

Global Pulse, (2012). Big Data for Development: Challenges and Opportunities. Available at http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

Global Pulse. (2013). Big Data For Development: A Primer Harnessing Big Data for Real-Time Awareness. Available at: http://www.unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf

Gobble M.M. (2013). Big Data: The Next Big Thing in Innovation. *Research-Technology Management*, Vol. 56, No. 1. p. 64-66

Government of Ghana e-Services Portal. (2011). Available at: http://www.eservices.gov.gh/SitePages/Portal-Home.aspx

Hammond W.E, Bailey C, Boucher P, Spohr M, Whitaker P. (2010). Connecting Information To Improve Health. Health Aff (Millwood) 2010. Feb 1;29(2):284–8.[PubMed]

Kothai M. (2015). How to use big data to combat fraud. *Big Data Science and Technology. World Economic Forum.* https://agenda.weforum.org/2015/01/how-to-use-big-data-to-combat-fraud/

Lewis T, Synowiec C, Lagomarsino G, Schweitzer J. (2012) E-health in low- and middle-income countries: Findings from the center for health market innovations. Bull World Health Organ; 90(5):332–40. [PMC free article] [PubMed]

J. Manyika J., M. Chui M., B. Brown B., J. Bughin J., R. Dobbds R., C. Roxburgh C. and A.H. Byers A.H. (2011). Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute*. Available at: (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

Raghupathi W. (2010). Data Mining in Health Care. In *Health care Informatics: Improving Efficiency and Productivity*, (Edited by Kudyba S.) Taylor & Francis p.211-223

Raghupathi W. and Raghupathi V. (2014). Big data analytics in healthcare: promise and potential, *Health Information Science and Systems* 2 (1) (2014), p. 3

Suresh K L. (2012). Managing the Big Data Challenge in Ghana's Banking Sector. *Building a Smarter* Planet. Available at http://asmarterplanet.com/blog/2012/11/21189.html