Hadoop and Big Data Readiness in Africa: A Case of Tanzania

Augustine Malero[1] and Hassan Seif[2]
[1, 2] College of Informatics and Virtual education, University of Dodoma, Dodoma, Tanzania

Abstract—Big data has been referred to as a forefront pillar of any modern analytics application. Together with Hadoop which is open source software, they have emerged to be a solution to the processing of massive generated both structured and unstructured data. With different strategies and initiatives taken by governments and private institutions in the world towards deployment and support of big data analytics and hadoop, Africa cannot be left isolated. In this paper, we assessed the readiness of Africa with a case study of Tanzania in harnessing the power of big data analytics and hadoop as a tool for drawing insights that might help them make crucial decisions. We used a survey in collecting the data using questionnaires. Results reveal that majority of the companies are either not aware of the technologies or still in their infancy stages in using big data analytics and hadoop. We identified that most companies are in either awakening or advancing stages of the big data continuum. This is attributed by challenges such as lack of IT skills to manage big data projects, cost of technology infrastructure, making decision on which data are relevant, lack of skills to analyze the data, lack of business support and deciding on what technology is best compared to others. It has also been found out that most of the companies' IT officers are not aware with the concepts and techniques of big data analytics and hadoop.

I. INTRODUCTION

Big data has been referred as a forefront pillar of any modern analytics application [1]. With the recent attention this study has enjoyed from researchers and practitioners, it has carried multiple definitions in the literature, Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software [2]. In a research report provided by McKinsey[3] a definition in a context of user activity modeling is given. McKinsey defines big data as "data-sets" which for practical and policy reasons cannot be "processed, stored and analyzed" by traditional data management technologies and require adaptation of work-flows, platforms and architectures.

Apache Hadoop [4] or simply Hadoop, is an open source software developed by Apache project licensed under the Apache v2 licence with an implementations of a distributed file system [5] and MapReduce that were inspired by Google GFS [6] and MapReduce [7] projects. The Hadoop ecosystem also includes projects like Apache HBase [8] which is inspired by Google BigTable [9], Apache Hive [10], a data warehouse built on top of Hadoop, and Apache ZooKeeper [11], a coordination service for distributed systems. Apache is being built and used by a global community of contributors worldwide [12].

It is also noted that the real power of Hadoop is in the number of compute nodes in the cluster instead of the compute and store capacity of each individual node [13]. According to Dumitru, A. [13], the strength of the Hadoop is due to the following;

- It is highly scalable—Yahoo runs Hadoop on thousands of nodes
- It integrates storage and compute—the data is processed right where it is stored
- It supports a broad range of data formats (CSV, XML, XSL, GIF, JPEG, SAM, BAM, TXT, JSON, etc.).
- Data doesn't have to be "normalized" before it is stored in Hadoop.

Despite the existence of other platforms in big data analysis, Hadoop has been chosen due to the large share it enjoys and being an open source with a thriving community (the Big Data industry leaders, such as IBM and Oracle embrace Hadoop as an integral part of their products and services)[14]. It is thus an ideal solution for developing countries.

II. LITERATURE REVIEW

Big data can be categorized into three types, namely structured, un-structured and semi-structured [15]. Structured big data is the one which has high degree of organization that is found in databases, data warehouses and enterprise solutions. Un-structured big data is raw data that has been extracted from applications on the Internet but has not been processed into productive and more meaningful formats. Semi-structured data is depending upon user view point, where structured and unstructured data meet.

Big Data is not one thing, and it does not have magical powers to provide advanced business analytics on its own. Big Data is not a technology. It is, however, a shift in the thinking on how to gain insight from data with increasing volume and varying formats [16].

With close to 90% of Fortune 500 companies investing in Big Data projects, the commercial market has already witnessed the benefits of investing in large-scale data processing [17]. According to Syncsort, eponymous software companies specializing in high speed sorting products, European organizations are very keen to explore the possibilities of Big Data through Hadoop implementations. In a survey carried out in the EU, about 58% of respondents said that they had Hadoop projects starting at their company in the near term [18]. Opportunities in big data have been spotted in financial services, healthcare, retail, Web/Social/Mobile, manufacturing and in government [19]. It is not surprising that Big Data is seen as a major impetus for international development [14].

It is also important to note that, there are challenges faced by big data implementation. Peglar et al., [19] contend that; there are ten big data problems which includes: modeling true risk, customer churn analysis, Recommendation engine, Ad targeting, Pos transaction analysis, Analyzing network data to predict failure, Threat analysis, Trade surveillance, Search quality and Data "sandbox".

Collaborative consulting [20] argues that in the process of translating Big Data hype, organizations should recognize

there are three paths towards Big Data adoption on which two paths are the extremes of the adoption scale. This includes; non-adoption path, rapid adoption path, slow-and-steady path. Hence organizations will follow one of these paths based on four factors: a compelling business case, supporting data governance, maturing vendor capabilities, and Big Data skills available in the workforce. For African countries, the well-known caveats of the Big Data debate, such as privacy concerns, interoperability challenges, and the almighty power of imperfect algorithms, are aggravated by long-standing development challenges like lacking technological infrastructure and economic and human resource scarcity. This has the potential to result in a new kind of digital divide: a divide in data-based knowledge to inform intelligent decision-making [14].

Different strategies have been implemented by businesses and governments to support big data analytics research and development. The US government in particular has shown its commitment and readiness in Big data by launching the long-term 2012 Big Data Initiative, investing $250 million a year in Big Data among its federal agencies, in addition to similar efforts within the Departments of Defense, Homeland Security, Energy, the Intelligence Community, NASA and more [21][22].

III. METHODOLOGY

A. Design of the study

A design which was adopted in this study is exploratory case study. Since the study was aimed on investigating the readiness of Africa in big data and hadoop a case of Tanzania. Case study is an ideal methodology when a holistic, in-depth investigation is needed [23].

B. Population

The targeted population for this study was defined as the totality of all stakeholders involved in the development of the ICT sector and aware of the opportunities that big data analytics and hadoop represents for Africa. Respondents from different entities all over the country were contacted and their responses recorded. The population of interest is usually defined by the purpose of the research and the research question itself [24].

C. Sample of the study

The sample was deployed with proportionate split among various sectors. Reasons for this were based on the factor that; to ensure at least each sector has been contributed on the findings which would be obtained.

D. Sampling technique

Probability sampling technique was used in this study. In order to enhance assurance in the consequent findings; the study area was divided into various parts/sectors. Then, proportionate split of respondent was done for each part. Lastly, random sampling of participants was done from a defined population of interest [25].

E. Data collection method and procedure

In this research, questionnaire was used as a research instrument. It was prepared and submitted to various respondents. Before collecting the data, a questionnaire was fully examined and pre-tested by the researchers. Then, was cross-checked for apparent mistakes and unclear signals, and numbered for easy computerization and identification of each respondent.

F. Data analysis

The aim of the study was to determine the readiness of Africa in Hadoop and Big data with a case study of Tanzania. The use of various analysis technique supports on analyzing the data for this study. Statistical analysis software package that was used in analyzing collected data is called Statistical Package for Social Sciences (SPSS) for Windows. Descriptive statistics and other inbuilt modules were used to statistically analyze the survey data for this study.

IV. RESULTS AND DISCUSSION

A. Challenges associated with data growth

It has been noted that, most of the companies' data are growing. About 98.5% of respondents report to have a growing companies'/organizations' data. In response to this massive generation of data, companies face a number of challenges. The challenges preponderant out by respondents as a result of data growing includes; Deciding what data is relevant, Cost of technology infrastructure, Lack of IT skills to manage big data projects, Lack of skills to analyze the data, Lack of business support and deciding what technology is best. Fig. 1 depicts percentage of the respondents on the challenges faced by the companies found in Tanzania as a result of data growth.

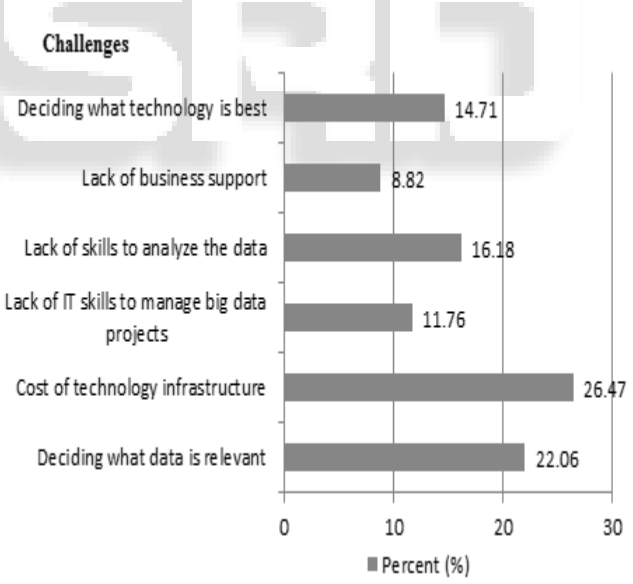

Fig. 1: Challenges faced by the company as result of data growing

It has been noted that; cost of technology infrastructure contribute the most (26.47%), followed by deciding on what data is relevant (22.06%), lack of skills to analyze the data (16.18%), deciding on what technology is the best for analyzing such data (14.71%), lack of IT skills to manage big data project (11.76%) and the last one is the lack of business support (3.82%).This implies that, most of the companies found in Tanzania are faced by problem of cost technology infrastructure as the major challenges on handling the large volume of data.

B. Awareness of big data analytics and hadoop

It has been found that even though there is a vast amount of data which are produced, most of the companies are struggling with the tools to handle their data (36.76%), don't really have a data strategy (14.71%), and don't get enough value of their data (11.76%). This means that companies generate large amounts of data in their everyday activities but do not make any use of the same. The study revealed that about 36.76% only of the companies surveyed reported to have a good strategy and right tools to deal with their data. This can be depicted by Figure. 2

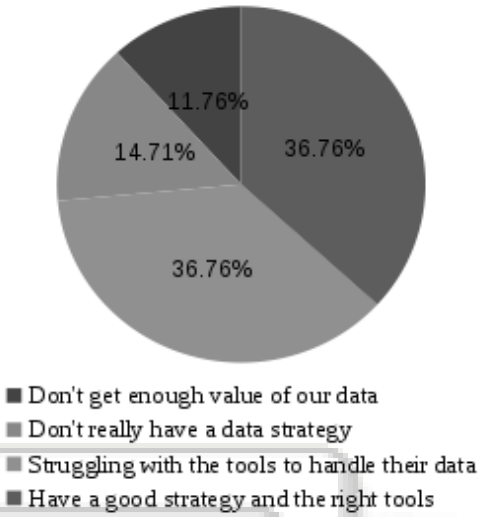

Fig. 2: The way company/organization dealing with their data

It was also found that; in Tanzania most of the respondents surveyed were not aware with the concept of big data and hadoop. This means that; it will not be possible for them to get insights from their daily generated massive data sets, a crucial process for getting knowledge that might help them make crucial business decisions. This can be proved by the question which was asked from the questionnaire on whether they ever heard of big data analytics or hadoop. Only about 27.94% of respondents reported to be familiar with big data analytics and hadoop, 29.41% of respondents had an idea of big data only, 0% has an idea about hadoop only while 42.6% of the entire respondents had never heard about big data analytics and hadoop. This can be depicted by the following bar chart.

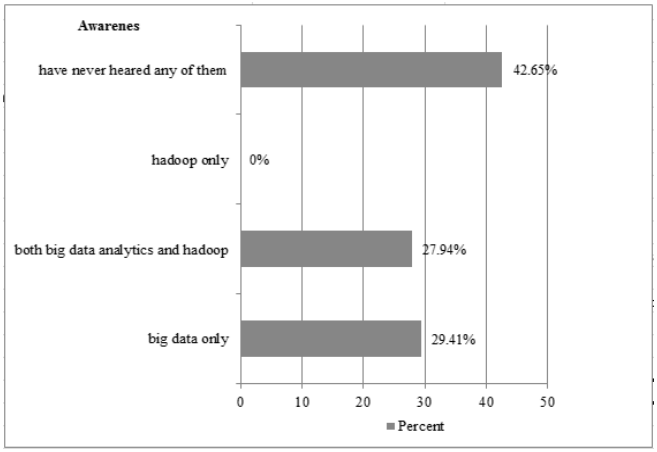

Fig. 3: Awareness of big data analytics and/or hadoop

C. A measure of a company's maturity around usage of big data

This part is concerned with a measure of a company's maturity around usage of big data especially around its usage for integration. The company can fall in any of the states among of the five states that have been inspired from Syncsort's Big Data Continuum [18]. This can be awakening, advancing, plateauing, dynamic, and evolved. In awakening state, a company primarily hand codes to process data by using basic ETL tools. By saying dynamic it means that the company is standardized on a data integration platform. Plateauing state means hitting limits (cost, growth) with a data integration platform. Dynamic state means that it experimenting with the hadoop while evolved state it means that there is a usage of hadoop as an enterprise data management platform. Based on these states it was found that most of the companies were at initial stages i.e., awakening or advancing. This can be proved by response from the respondents on the questions which was asked about a measure of a company's maturity around usage of big data. About 33.82% of respondents responded that they were at an awakening stage and 35.29% were at advancing stage. Only 25% of the respondents were at dynamic stage and 4.41% were at plateauing stage. There was no company which was at evolved stage (i.e., 0%).This means that; currently there is no any company using hadoop as an enterprise data management platform among those surveyed and only about 25% are experimenting with hadoop. This can be depicted by the following pie chart.

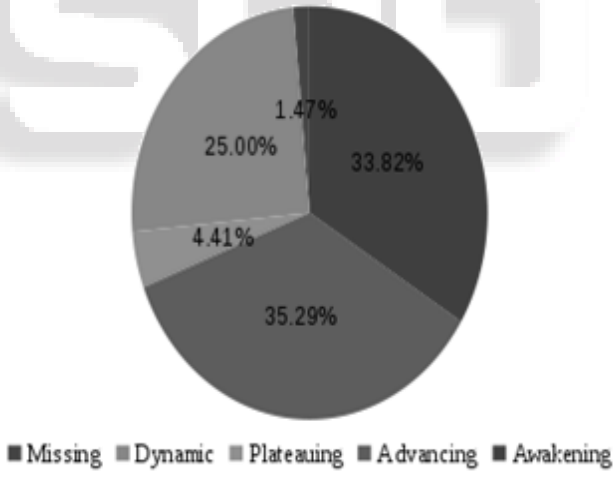

Fig. 4: Position of the company/organization in big data continuum

D. Adoption of data analytics

Results from various respondents showed that; most of the companies are not ready to adopt in big data analytics and hadoop. It has been found that; about 17.65% of respondents will reach mainstream adoption of big data analytics by 2014-2015, 16.18% by 2016-2017, 13.24% by 2018-2019 and 4.41% by 2020 and beyond while 48.53% so far have no plans for adoptions of big data analytics. Due to this fact it shows that most of the companies have no plans for adopting big data analytics while only about 17.65% will reach mainstream adoption of big data analytics by 2014-2015 a number so small. This can be depicted by Fig. 5

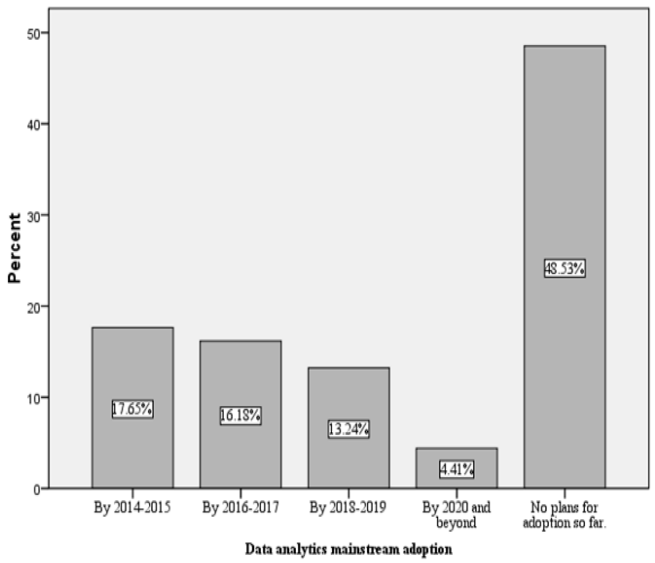

Fig. 5: Mainstream adoption of data analytics

V. CONCLUSION AND RECOMMENDATIONS

Even though most of respondents realize that there are massive volumes of both structured and unstructured data that are generated on their daily activities such as in social networks, transaction logs, sensor networks, Radio Frequency Identification (RFID) readers, call detail records, documents and other sources, they do not make proper use of the same. Most businesses are not ready to cope with changes by adapting to big data analytics and hadoop as a means for knowledge and insights acquisition. This is due to the fact that most of the companies are hindered by some of the challenges such as lack of IT skills to manage big data projects, cost of technology infrastructure, making decision on which data are relevant, lack of skills to analyze the data, lack of business support, and lastly deciding on what technology is best compared to others. It has also been found out that most of the IT practitioners surveyed are not aware with the concepts and techniques of big data analytics and hadoop. This makes most of the companies to be either in awakening or advancing stages of the big data continuum.

Since there is a significant lack of awareness to the IT professionals about big data analytics and hadoop, it is recommended that many seminar/workshops be conducted so as to raise awareness and knowledge to the people about the subject. Furthermore infrastructure should be improved so as enable in handling and processing of big data, which would be resulted in improving business processes by gaining insights and making better decisions.

REFERENCES

[1] LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. 52, 21–32 (2011).
[2] Chris Snijders, UweMatzat, Ulf-Dietrich Reips. "Big Data": Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science 2012, 7 (1), 1–5 ISSN 1662-5544.
[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity. Mckinsey Glob. Inst. 1–137 (2011).
[4] Apache Hadoop. Available at http://hadoop.apache.org
[5] Apache HDFS. Available at http://hadoop.apache.org/hdfs
[6] The Google File System. Available at http://labs.google.com/papers/gfs-sosp2003.pdf
[7] MapReduce: Simplified Data Processing on Large Clusters. Available at http://labs.google.com/papers/mapreduce-osdi04.pdf
[8] Apache HBase. Available at http://hbase.apache.org
[9] BigTable: A Distributed Storage System for Structured Data. Available at http://labs.google.com/papers/bigtable-osdi06.pdf
[10] Apache Hive. Available at http://hive.apache.org
[11] ZooKeeper: Wait-free coordination for Internet-scale systems. Available at http://www.usenix.org/events/usenix10/tech/full_papers/Hunt.pdf
[12] Apache Community. Available at http://wiki.apache.org/hadoop/PoweredBy
[13] Dumitru, A., "Dell | Hadoop White Paper Series: Hadoop Enterprise Readiness", Dell Inc, 2011.
[14] Hilbert, Martin, Big Data for Development: From Information- to Knowledge Societies (January 15, 2013). Available at SSRN: http://ssrn.com/abstract=2205145 or http://dx.doi.org/10.2139/ssrn.2205145.
[15] Mind Commerce, Big data Opportunities and Telecom, Mind Commerce, 2013.
[16] Collaborative consulting, "Making sense of big data: A collaborative point of view", 2013
[17] Big Data for Defence and Government 2013. Available at http://www.bigdatafordefense.com/
[18] Syncsort, "The European Big Picture on Big Data and Hadoop in 2013", Syncsort, 2013.
[19] Peglar, R and Isilon, E., "Introduction to Analytics and Big Data- Hadoop", Storage Networking Industry Association, 2012.
[20] Collaborative consulting, "Making sense of big data: A collaborative point of view", 2013.
[21] Big Data across the Federal Government. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf
[22] Big Data is a Big Deal. Available at http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal
[23] Feagin, J., Orum, A. and Sjoberg, G. A case for case study, University of North Carolina Press, Chapel Hill, NC. 1991.
[24] Marczyk, R. G., Dematteo, D. and Festinger, D., Essentials of research design and methodology, John Wiley & Sons. 2005.
[25] Saunders, M., Lewis, P. and Thornhill, A. , Research Methods for Business Students, Fourth Edition, Prentice Hall Financial Times, Harlow, Pearson Education, 2007

