Global Scientific JOURNALS

GSJ: Volume 10, Issue 3, March 2022, Online: ISSN 2320-9186

www.globalscientificjournal.com

# SECURING BIG DATA AND ANALYSIS USING UTF BASE 64 ALGORITHM

**Adeleye S. A[1]., Oladele K. J[2]., Nwachukwu K. I[3]., Aworonye E. A[4]**

Department of Computer Science
Federal University of Petroleum Resources (FUPRE)

**Idama R. O[5]**
Department of Computer Science
Delta state University of Science and Technology (Ozoro)

**Abstract**
Nowadays, companies are starting to realize the importance of data availability in large amounts in order to make the right decisions and support their strategies. With the development of new technologies, the Internet and social networks, the production of digital data is constantly growing. The term "Big Data" refers to the heterogeneous mass of digital data produced by companies and individuals whose characteristics (large volume, different forms, speed of processing) require specific and increasingly sophisticated computer storage and analysis tools. This article intends to define the concept of Big Data, its concepts, challenges and applications, as well as the importance of Big Data Analytics by securing it with UTF base 64 algorithm in a large Business Enterprise. Mongo DB

**Keywords:** Big Data; Big Data Analytics; Hadoop; Internet; Security

## 1.0 INTRODUCTION

The digital data produced is partly the result of the use of devices connected to the Internet. Thus, smartphones, tablets and computers transmit data about their users. Connected smart objects convey information about consumer's use of everyday objects. Apart from the connected devices, data come from a wide range of sources: demographic data, climate data, scientific and medical data, energy consumption data, etc. All these data provide information about the location of users of the devices, their travel, their interests, their consumption habits, their leisure activities, and their projects and so on. But also information on how the infrastructure, machinery and apparatus are used. With the ever-increasing number of Internet and mobile phone users, the volume of digital data is growing rapidly. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information

plays a major role in the economical, cultural and political stage (Rieder, 2013).

## 2.0 WHAT IS BIG DATA?

A. Definition

The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society. The challenge is not only to deal with rapidly increasing volumes of data but also the difficulty of managing increasingly heterogeneous formats as well as increasingly complex and interconnected data.

Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases.

Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies, on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing (David, 2015).

### 2.1 Characteristics of Big Data

The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi-structured, and unstructured (variety) data, and arriving faster (velocity) than before. These are the 5V.

**a. Volume**

Volume is one of the most important attributes for data in the big data class. It depicts very large and ever growing amount of data ranging from terabyte ($10^{12}$ byte) to yottabyte which is trillions of gigabyte as expressed.

b. **Velocity**

Velocity is refers to real time availability of data for processing. Big data is also characterized by instantaneous arrival of enormous data for processing. It entails the rate at which data is circulated within the system e.g. the speed at which data arrives from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc.

**a. Variety:**

Variety is the analysis of big data consisting of data derived from various sources such as emails like machines, social networks, business transactions, mobile devices etc. data from different sources assume different forms such as emails, spreadsheets, photos, videos etc. Variety as a property of Big Data refers to the different forms of data from different sources.

Beyond volume, variety and velocity, which are the main characteristics of Big Data, researches in the field of Big Data have reveal more characteristics of Big Data, thereby adding to the V's of Big Data. These include:

**Veracity:** this refers to the truthfulness of the data. It deals the relevance of the data being stored ( and /or processed) to the problem being analyzed. It reveals the need to avoid accumulation of dirty data.

Volatility: this deals with the reasonable life span of stored data in the world of real time data processing. It investigates the validity of stored data to the current analysis.

**Validity:** Decisions are as valid as the data used in the analyses.

## 2.2 Big Data Analytics

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as "predictive analytics", 'data mining', 'text analytics' and 'statistical analysis'. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases (Bechmann, et al, 2015).

Like big data, the analytics associated with big data is also described by three primary characteristics: volume, velocity, and variety (ibm.com/bigdata/). There is no doubt data will continue to be created and collected, continually leading to incredible volume of data. Second, this data is being accumulated at a rapid pace, and in real time. This is indicative of velocity. Third, gone are the days of data being collected in standard quantitative formats and stored in spreadsheets or relational databases. Increasingly, the data is in multimedia format and unstructured. This is the variety characteristic. Considering volume, velocity, and variety, the analytics techniques have also evolved to accommodate these characteristics to scale up to the complex

and sophisticated analytics needed (Russom, 2011; Zikopoulos et al., 2013). Some practitioners and researchers have introduced a fourth characteristic: veracity (Ohlhorst, 2012). The implication of this is data assurance. That is, both the data and the analytics and outcomes are error-free and credible.

Simultaneously, the architectures and platforms, algorithms, methodologies, and tools have also scaled up in granularity and performance to match the demands of big data (Ferguson, 2012; Zikopoulos et al., 2012). For example, big data analytics is executed in distributed processing across several servers (nodes) to utilize the paradigm of parallel computing and a divide and process approach. It is evident that the analytics tools for structured and unstructured big data are very different from the traditional business intelligence (BI) tools. The architectures and tools for big data analytics have to necessarily be of industrial strength. Likewise, the models and techniques such as data mining and statistical approaches, algorithms, visualization techniques, etc., have to be mindful of the characteristics of big data analytics. For example, the National Oceanic and Atmospheric Administration (NOAA) uses big data analytics to assist with climate, ecosystem, and environment, weather forecasting and pattern analysis, and commercial translational applications. NASA engages big data analytics for aeronautical and other types of research (Ohlhorst, 2012). Pharmaceutical companies are using big data analytics for drug discovery, analysis of clinical trial data, side effects and reactions, etc. Banking companies are utilizing big data analytics for investments, loans, customer demographics, etc. Insurance and healthcare provider and media companies are other big data analytics industries.

The 5Vs are a starting point for the discussion about big data analytics. Other issues include the number of architectures and platform, the dominance of the open-source paradigm in the availability of tools, the challenge of developing methodologies, and the need for user-friendly interfaces. While the overall cost of the hardware and software is declining, these issues have to be addressed to harness and maximize the potential of big data analytics.

**Hasard et al 2013** proposed a RDB for data storage in EHR. The model was implemented with a MySQL server. Having examined other peoples work, this research looks at the hybridized approach to big data

storage and management. In this research, the HER system consist of NoSQL database developed with MongoDB and a relational database using MySQL. Structured data such as name, address etc. are stored in the MySQL database while the seemingly schemaless data are stored in the MongoDB database. This will give us a system that has the functionalities of NoSQL and that of relational databases. (Mohamed, 2016) carried out a research titled Big Data Query using Apache server and Web/ Internet technology.

**Qin Yao et al, 2015** published a work on transaction processing system titled Design and Development of a Medical Big Data Processing System. Based on Hadoop. they used Precision and recall algorithm which are the two metrics that are widely used in the field of information retrieval and statistical classification to evaluate their quality of results. Their work was conclusively described as a design and development of a Hadoop-based Medical Big Data processing system that can be applied for future uses of Medical Big Data. Which can resolves the problems of Medical Big Data collection, storage and analysis. Which resulted to build a Mahout based distributed recommendation engine to reveal the collective intelligence behind the Medical Big Data.

### 3.0 Materials and Methods

The existing system is based on Blessing and Asagba (2016) who adopted that big data analytics and the Apache Hadoop in the business enterprise which serves as an open source project are rapidly emerging as the preferred solution to business and technology trends that are disrupting the traditional data management and processing landscape. Enterprises can gain a competitive advantage by being early adopters of big data analytics. Even though big data analytics can be technically challenging, enterprises should not delay implementation. As the Hadoop projects mature and business intelligence (BI) tool support improves, big data analytics implementation complexity will reduce, but the early adopter competitive advantage will also wane. Technology implementation risk can be reduced by adapting existing architectural principles and patterns to the new technology and changing requirements rather than rejecting them. This analysis includes the stochastic, operational, and simulation techniques for modeling the dynamic behavior and for computing performance indicators of a variety of selected server systems. The techniques for

workload characterization are also presented. In addition to theoretical results, the research also includes experimental computer performance analysis techniques and laboratory experiments to build high performance server systems requires an understanding of the servers and what makes them fast or slow.

Figure 3.1 illustrate the big data analyses of data in a large enterprise for proper storage of records to align with efficient records keeping system of the market dividend and profit. The records are therefore classified a both structured and unstructured data to be process so as to make a valid decision.

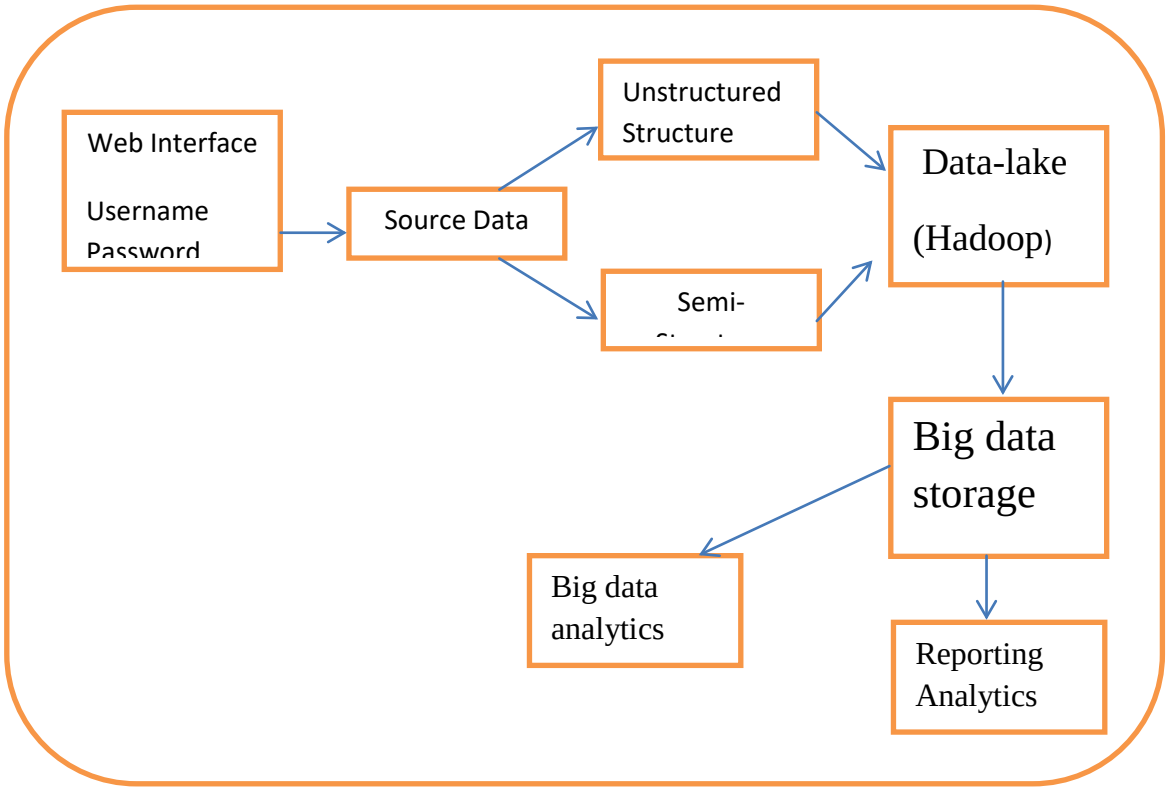


Figure 3.1: Architecture of Existing System

**b. Proposed System**

The proposed system is to improved on the security issues of the existing system on big data performance analysis and storage capacity in other to identify online theft that are capable of extracting data without proper authorization. The goal of this proposed system is to create a three (3) ways

authentication (Username and Password, Generation of token and finger tomb as a security mechanism to be used to secure the platform of the Big data), so that data were captured and analyzed in order to gain actionable insights from the data at a much lower cost.

**The algorithm for the Proposed System**

The algorithms used in the proposed system in UTF base 64 to secure big data storage and performance which are concerned with the following steps, where each step is expected to convey momentous improvement.

Steps of the design of proposed system:

1. Ensure the proper login
2. If successful login
3. Go to 5
4. Else, stop.
5. Accept big data (database) file internally from the system as input
6. For each of the input big data files (encrypt with UTF base 64)

a. Search & retrieve basic details of the file.

b. Test the servers to know the time used, memory consumed and jobs accomplished.

c. Query the database

d. Decrypt the cyphertext to plaintext

7. Output servers' test results

**3.3.2 Benefit of the Proposed System**

1. Data are secured from an unauthorized users
2. Teradata, as a Hadoop resource, is unmatched for ease-of-implementation, time-to-value and advanced analytics.

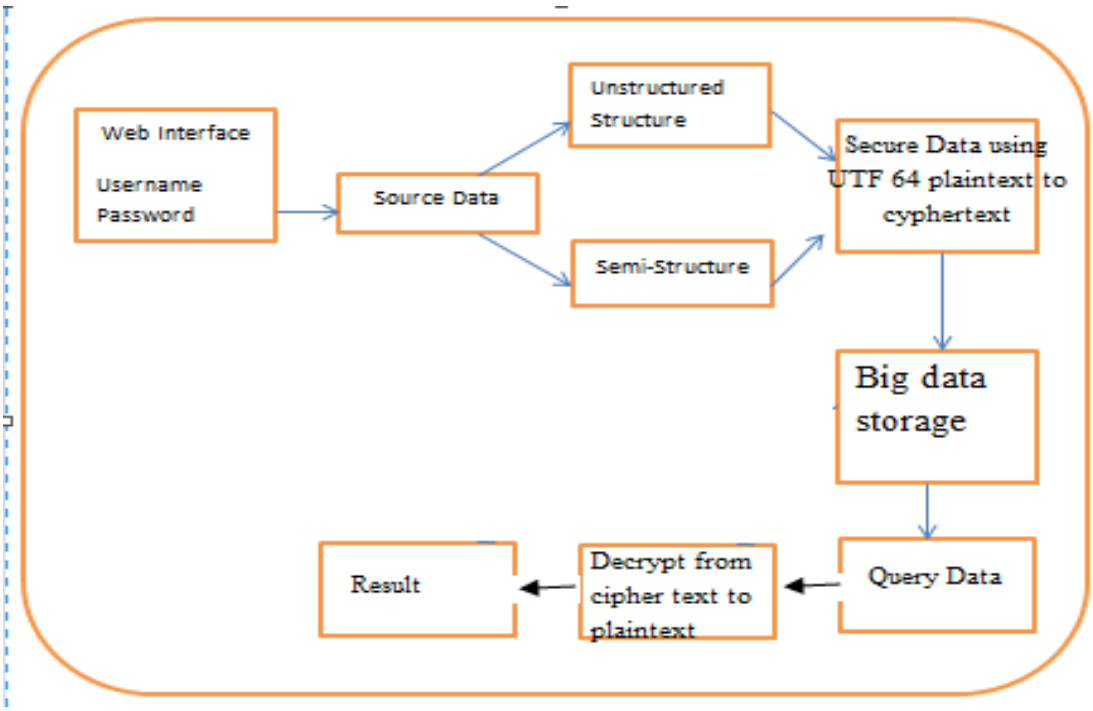


Figure 3.2: Architecture of the Proposed System

**4.0 Performance evaluation**

The performance evaluation of a software is measured with the matrices used. Several matrices can be used in the evaluation of a software ranging from speed, time, efficiency and so on. In Big data analysis for Business Enterprise system, the matrices used in the evaluation testing are the file size, the time to search for query of data. This section explained in details how the performance for query the data of the proposed system compares to the existing system is faster and the throughput which are shown in the table illustrated in table 4.3

Table 4.3: Comparison between Existing and Proposed System

| Business enterprise | Number of jobs | Existing system | Proposed system |
|---|---|---|---|
| MTN mobile network | 1000 | 0.400 | 0.24 |
| GLO mobile network | 1000 | 0.520 | 0.300 |
| Sensor Data (temperature) | 1000 | 0.440 | 0.280 |
| | | | |

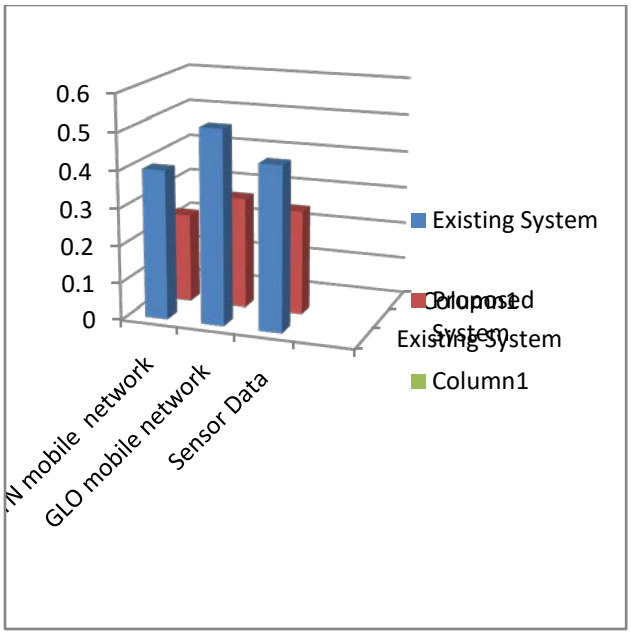


**Fig.4.9: Graph of the comparison**

## 5.0 Conclusion

In conclusion, the new approach to Querybig data using IIS as a servers with the Mongo Database system as an open source software and implemented with VB.NET being a simple, multi-paradigm, structured, object-oriented, modern and event-driven high level programming language. This enabled us as researchers to determine the right server in handling big data. Accuracy in big data may lead to more confident decision making. And better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Software is needed to manage this data flow: Tera-data supports such queries through software called Query-Grid both on hardware appliances and in the cloud. Query-Grid also allows you to feed data from the relational or Hadoop stores to analytical platforms. In Tera-data's suite, that role is played by Aster, which offers graphs and other advanced analytical tools. Aster has its own big data solution for executing queries, called SQL-MapReduce. Query-Grid has connectors to a number of other databases as well, such as Oracle and MongoDB.

Further, metadata management and data lineage play a key role in data governance. Teradata Loom scans all data entering Hadoop and adds metadata, including its lineage (when the data was generated and by whom). The metadata from Loom not only helps users find data they need, but helps

them judge its reliability and the purpose for which it was collected. You also can structure Hadoop or HDFS data in such a way that it's accessible to SQL as well.

**6.0 References**

Bechmann, J. Jensen L., and Vahlstrup P, 2015."Studying social media data across platforms on planned event: Facebook, Instagram and Twitter data patterns at music festival," paper presented at *Users across media conference, University of Copenhagen*.

Blessing J.and AsagbaP.O(2016) : Big data storage using relational database and Mongo database, ijs, 34-56

David F (2015): "The mismeasurement of privacy: Using contextual integrity to reconsider privacy in HCI," CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 367–376.

Ferguson, M. (2012). Architecting a big data platform for analytics. *A Whitepaper prepared for IBM*, *30*.

Mohammed, A. F., Humbe, V. T., & Chowhan, S. S. (2016, February). A review of big data environment and its related technologies. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1-5). IEEE.

Ohlhorst, F. (2012): Big Data Analytics: Turning Big Data into Big Money. New York: John Wiley & Sons.

Rieder S.D., (2013): "Numbers have qualities too: Experiences with ethno-mining," Ethnographic Praxis In Industry Conference, 9(1), 123–140.

Yao, Q., Tian, Y., Li, P. F., Tian, L. L., Qian, Y. M., & Li, J. S. (2015). Design and development of a medical big data processing system based on Hadoop. *Journal of medical systems*, *39*(3), 1-11.

Zikopoulos, P.C., deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., and J. Giles(2013):. Harness the Power of Big Data—The IBM Big Data Platform. New York: McGrawHill.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1-34.