A THEORETICAL EXPLORATION OF DATA MANAGEMENT AND INTEGRATION IN ORGANISATION SECTORS

Chisom E. Offia and Malcolm Crowe

School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley, Scotland

ABSTRACT

Big data development is a disturbing issue that will affect enterprise across various sectors. The increase of data volume, high speed of data generation and increasing rate of different data from heterogeneous sources have led to difficulties in data management. This paper first reviews different aspects of big data management, including data integration and traditional data warehouse, and their associated challenges. The problems include increase of redundant data, data accessibility, time consumption in data modelling and data movement from heterogeneous sources into a central database, especially in the big data environment. We then propose a logical data management approach using RESTview technology to integrate and analyse data, without fully adopting traditional ETL processes. Data that for governance, corporate, security or other restriction reasons cannot be copied or moved, can easily be accessed, integrated and analysed, without creating a central repository. Data can be kept in its original form and location, eliminating the movement of data, significantly speeding up the process and allowing for live data interrogation. It may not be the practical solution for every situation but, it is a feasible solution that is comparably cost effective.

KEYWORDS

Big Data; Data Integration; Data warehouse; RESTView

1. INTRODUCTION

The era of big data is broadly defined as information available from billions or trillions of data generated by people, machine, structured and unstructured form throughout cyber universe [1]. This big data era has now resulted to datafication meaning the digitization and analytics of big data [2]. Big data as defined by [3] is the integration of several disciplinary technologies which facilitate customers by gathering all incredible services to a point.

The world is currently facing an expansion in size and variety of digital data which are generalized by both human (user) and machine (system). Internet of Things communications have great impact on how big data have developed. Example of digital data varieties are big data in health sector where data is retrieved from different sources like Electronic records (ER), patient DB, clinical trials, medical measurement and imaging [4].

The introduction of data integration is to combine data from different sources or departments of big data to provide a unified access to user, furthermore, to deliver and utilize this information across different organization. Putting together big data and data integration makes the traditional data integration different from big data integration. This means that the big data challenges are now to be considered when integrating data.

The remaining part of the paper is organized as follows. Chapter 2 explains big data management and its challenges. Chapter 3 explores the overview of data integration and its possible challenges and chapter 4 is the overview of data warehouse. In chapter 5 the discussion focuses on different

data management approaches from various literatures and chapter 6 provides a brief introduction of proposed framework. Chapter 7 is conclusion and recommendations.

2. BIG DATA AND ITS CHALLENGES

2.1. BIG DATA

Big data is characterized by five V's; Volume (which refers to the size of data which has expanded beyond terabytes), Velocity (which refers to the speed of product of data to meet different demands), Variety (which refers to data generated from diverse sources and these data are unstructured which make it difficult to categorize), Veracity (which highlights the significance of addressing data uncertainty and managing it), Value (which refers to the significant advantage that big data can offer to business) [5], [6], [7]. Some authors in [7] added validity as the correctness or how valid the data is to the city.

Volume: This refers to the size of data which is generated every second [8]. The size is no longer reckoned in terabytes but petabyte. At this level, datasets become too large to manage, store and analyse using the traditional database technology because of its size. The reason for the large data size is because of the increasing number of people sharing, communicating and accessing data through different devices. Also, there are many sensors being connected as internet of things [9]. According to the research by [10], in four (4) billion number of people utilizing mobile phone have the percentage of 12 making use of smartphones whereby the percentage increases by twenty (20) each year. Another research [11] outlines that, thirty (30) million networked sensor nodes are currently in utilities, transportation, automotive, industrial and retail sectors. The rapid increase in large-scale data has led to new challenges in data mining techniques, data management, data integration, data processing, and data retrieving. This now requires novel approaches to addressing the big data issues [11].

MapReduce is a technique designed by Google which has the ability to control large amount of data using a cluster system. Hadoop is another framework created from MapReduce and Big-Table for the purpose of allowing parallel processing in groups of high latency and this system has tolerance requirement s for any failure and load assessment [12].

Variety: - Some of the data today are structured and unstructured. This means data variety is referred to as different type of data from different sources [13]. Examples are data sources from text messages and video [8]. In the past, structured data focused on data arrangements done easily by fitting it into rows, tables or relational databases. In a variety of big data, enterprises face issues when analysing. Enterprises continue to automate further business processes and this disrupts the design of data systems. Another challenge that enterprise faces is how to sort these different data to be readable by end users without creating ambiguous results [14]. Since different data comes from different sources, many cases report both relevant information and some information that needs proper filtering. Therefore, a reliable system is needed to identify impactful information. The size of data does not consistently follow a specific template or format and these different forms led to poor data quality which requires a clear indication that diversity is a natural property of big data [11].

Velocity: - This refers to how fast new structured data is generated and at what speed these data move around [13], [14], and [8]. For example: in social media services, we observe how fast it is generated from different sources. The challenges related to velocity include the requirement to manage the high influx rate of non-homogeneous data. Another example is Internet of Things (IoT) from smart city which is connected to Sensors and theses sensors will be constantly channelling tiny bits of data at near constant rate. And as the number of flow increases, so is the flow and speed of data. Data from mobile apps like google map for transportation generate floods of information that can be utilized through producing real-time data and also the pattern and

behaviour of data can be observed [11]. Cloud computing was designed with the promise to solve the data storage and its processing time challenges [12].

Veracity: - This is the management of data uncertainty, if data is not accurate it will create errors in analysis and that will be a problem for the organization [15]. With the generation of data which constantly changes, the quality and accuracy of data are uncontrollable. Therefore, big data technology has made the quality and accuracy less controllable. Understanding data from different sources is more important when relating to veracity. IBM analysed the characteristics of data which denotes untrustworthiness inherited in many sources of both structured and unstructured data [11]. In decision making, veracity and the frequency of how data is generated (velocity) is a great challenge;

Nevertheless, it is important that data should be sampled for operational decision making and veracity should be considered by identifying the risk and complexity it poses when constructing an operational decision [16].

Value: - Research considers value as an essential feature within data when extracted from multiple sources to make meaningful data out of them. The Centre for Economics and Business Research (CEBR) in UK predicted that big data analysis will be beneficial to the retail, manufacturing, financial and public sectors [17]; [9]. Value depends on thorough analysis of accurate and valid data because the data increases rapidly. It is messy seeing that data constantly changes in different thousands of formats which are worthless without analysis [18]. To ensure good practice of data utilization, it is important to have efficient big data management tools in place. These management tools should be able to recognize different formats of data from different sources, structuralize the unstructured data into groups, manage, classify and control them [7]; [19]; [20].

Valence: - This refers to the interconnection of data and it raises the challenge of data complexity in the aspect of analysis. In big data, valence measures ratio of the actual connected data sets compared to the possible number of connections that could happen within the process of collection [71].

2.2. BIG DATA CHALLENGES

Big data researchers face challenges on how data can be stored, managed, extracted and analysed constantly [1]. One method proposed by [1] is linked open data (LOD). LOD allows structured data to be interlinked and extract value out of the data through semantic queries. LOD applies resource Description Framework (RDF) and W3C for the purpose of data interchange on the web. This method shares information in a way that can be read automatically on communication between computers. However, the quality of interconnection relationship is still questionable. Another researcher [21] highlighted that big data do not only rely on archiving and conserving large quantity of data but accessing these data in an efficient way. The solution to this problem was the use of Hadoop Distributed File System (HDFS) which is stated that it is fault- tolerant because the loaded file is replicated three times [21]. The problem found with this HDFS model is that replication of data three times increases the demand for storage space.

This paper by [2] outlines challenges relating to the management of big data in government. These challenges are storage and computational complexity, ensuring transparency and lastly the formidable analysis of voluminous data. The researcher further explained that it is difficult for data analysis to be done properly since government data are stored in silos which are being managed by different departments.

Another challenge of big data in analytical aspect spotted by [9] is the access and sharing of information. They further explained the reason why the access to data and data sharing is a challenge is because of security issues of the clients' data and also one getting to know the operation of the company [9]. Data collected for analysis lacks basic structure and some values of these data are missing. This is as a result of data that is managed by people who do not

understand the particular data and this has also been a challenge [6]; [2]. Variety in big data has led to data inconsistency resulting to analytics sprawl and also poor data quality [6]. The high velocity of data has generated challenges relating to speed handling the speed of new data or updated data that are constantly generated.

Cities face challenges when relating to big data and these challenges are in different forms; they are privacy and security, data quality, collection of data, data sharing, analysis of data from different sources, accessibility of data, management of data and many others.

Table 1. A brief listing of big data challenges and suggestions from researchers as solutions to big data problems [6]; [22]; [23]; [24].

Big Data Environment	**Challenges related to**	**Suggested solutions**
Volume	Data storage, Data scale, Data reduction	Data record, Sensors, Web Scrapping, NoSQL, Apache Hadoop, sharing data based on cloud
Variety	Data access,	Structure data
Velocity	Data collection,	Data as value
Veracity	Data retrieval	Streamline Data, Deep learning, Cross validation.
Data	Data inconsistency, Data heterogeneity,	Data authenticity, Emotion recognition.
Data	Non-aligned data structure.	Legal provision for the use of data and also use of data codes, Web traffic and communication monitoring, Implementing privacy protection law and Segregation of the networks.
Human factor	Data visualization	Benchmarking the application,
Big Data Environment	Data processing speed	Formation of functional group, Digital certificate, Data encryption.
Volume	Data abnormality, Data quality, Data accuracy	Open platform, Government
Variety	Dirty data.	Suggested solutions
Velocity	Privacy	Data record, Sensors, Web Scrapping, NoSQL, Apache Hadoop, sharing data based on cloud
Veracity	Security	Structure data
Data	Human Involvement	Data as value
Data	Challenges related to	Streamline Data, Deep learning, Cross validation.
Human factor	Data storage, Data scale, Data reduction	Data authenticity, Emotion recognition.

Every department like e-government, city agencies has their own data warehouse or silo of confidential information [7]. The collection of data, the processing stage, data transaction, data integration, transformation, data storage, data computation, and understanding data form the big data challenges with relation to analytics. In the aspect of organization, data acquisition is a challenging task which can lead to complexity, robustness and security issues. The use of different technologies, devices and IoT, gave birth to the need of deep understanding of data. This prompted the need for data transformation. When these data are not in real-time, it produces

incomplete data which in return compromises data accuracy. Therefore, data extracted from heterogonous sources of different physical devices need fast and real-time analysis [3].

Privacy and security of data is another challenge on its own in relation to big data. Sometimes there might be errors in the process and analysis of data which can lead to exposure of information about its users. That is the reason why individual access to data is impracticable. Suggestions have been made by some researchers [6] that encryption should be used to ensure safe transfer of data over the network. Databases may contain sensitive data which government uses and this data demands high level of security policies and effective mechanism to protect it [7].

In reducing the challenges of access to data together with its security measures, policies are needed to ensure that all information is accessible fairly to consumers while validity and monitoring processes continue. Some people tend to make assumptions that access to data is a violation of person's privacy rights thereby creating high demand to clearly notify and protect privacy rights of citizens, organizations and various departments of smart cities.

3. DATA INTEGRATION

Data integration started early in the 1980s whereby combining the different varieties of data sources is mostly known as information silos [25]. The first method of data integration was carried out by university of Minnesota in the year 1991 and was used for the Integrated Public Used Microdata Services (IPUMS). This applied the traditional data warehouse technology and ETL approach to integrated data into unified schema [26].

In 2009, the approach of data integration changed to access and retrieving data directly from the source database by applying the Service Oriented Architecture (SOA) which rely on mapping. This mapping is done in two methods which are the GAV and LAV approach. The reason for the change of approach was as a result of changes to organization demands in accessing real-time data through a mediated means [27].

In 2010, semantic conflicts arose and researchers sought to resolve this problem between heterogeneous sources. One common way to resolve this issue was applying the use of ontologies which is represented as ontology-based data integration [27.] In 2011 data isolation problem arose which resulted to the development of desperate data model. One method proposed by [28] to eliminate this issue was enhanced data modelling. This model was designed to reorganized data model by augmentation and by structuring meta-data in a standard form of data entities.

Theoretically, data integration (I) is defined by the global schema (G) known as the mediated schema, defined by the heterogeneous set of source schema (S) and by mapping of queries between sources (M). This can be expressed as

$$I = GSM \tag{1}$$

The global-as-view (GAV) is designed in a way the mapping (M) associates with each element of global schema (G) over the source (S). In local-as-view (LAV) the source database is designed as set of views over G, this means that M associates with each element of S over G. In the GAV the process is simple and straightforward but the problem occurs when new sources of data are to join the system. In LAV the association between G and S is not defined and as that, the modelling is easy because new sources can be added [20].

The goal of data integration is to provide unified access to data. For example using airport data source; if data reside in different databases and it is disintegrated then, customers will not be able to benefit from using a single mode of access to acquire diverse information related to the particular airport, but if data is integrated into single source like Air info, (where the information about the particular airport, airline, flight schedule, departure time, arrival time, run way used, airfare, and many more) it will be beneficial for a client to use one single mode to view the

required information. These researchers [29] also elaborated that integrating multiple data sources into a single system can be difficult to achieve and that will require some amount of manual effort to understand the format of the data present in each source.

The advancement of organizational needs in workflow includes data management, storage, maintenance, and data integration. Data integration is therefore, a complex issue for organization in deploying big data architecture because of the heterogonous data form used by them [30]. Data integration usually focuses on the need of a firm (or organization), where the data is currently in silos in different departments of the organization. Data integration currently is often used by database vendors to combine data bases into one larger system, using ETL= Extract, Transform, Load approach in creating the new combined data sets. ETL (Extract Transform Load) tool is one common form of data integration software used by many organizations. This is not always appropriate because of data protection issues and redundant data, and in any case, often results in data no longer being managed by people who understand it: the ETL process often loses important features of the data [30]; [31].

3.1. QUERY PROCESSING

In a data integration system, there is a schema which poses user to insert queries. The schema is known as a mediated schema or global schema which answers queries posed by users using the data sources. The system needs a mapping function that describe the semantic relationship the mediated schema and the schema of the sources [32].

In data integration query processing is represented using conjunctive queries and data log. The conjunctive query is a logical task which is exploited to relationship of database. It is shown as F (A, B) where A<B. Technically, integration of data gets to modify queries which are represented by the views in order to allow the result to be equivalent to user's query. Query processing expatiate the sub-purpose of the user's query based on the basic agreement identified in the mediator. In GAV, the system designer writes the mediator code to specify the query rewriting and in LAV the query passes through a more thorough process of rewriting due to the fact that it has no mediator to adjust to user's query.

Query processing in LAV, the process is called view-based query processing and it uses two (2) approaches which are view-based query rewriting and view-based query answering [32]. The problem of view-based query process is the computation of answer to a query based on a set of views rather than raw data in the database. The two approaches are to redevelop the query into an expression language that is fixed to refer views only and provide an answer.

There is no form of integrity constraint in GAV query processing, which means that the views in GAV are assumed to be the same. The problem is that the GAV system becomes complex when the language used for expressing the global schema allows integrity constraint [20].

3.2. BIG DATA INTEGRATION

The process of extracting, transforming and integrating data are time often spent by developers to run analysis and visualization program. The difficult aspect is writing this program for reshaping the data but it is important because each analytical tool expect the data to be organized in a specific form. This process is more challenging when the angle of big data is involved and this is because the sources can be large from millions and billions of data streaming and in heterogeneous form. With this, developers cannot possibly review all the data [33].

Big Data Integration is quite different from the traditional data integration due to the addition of volume, velocity, variety, and veracity. Therefore, big data challenges are applied to the traditional data integration after observing data sources are now in millions. Also, the speed rate at which newly collected data is made accessible, many of the data sources change frequently with time, and also the amount of data source is rapidly exploding. It is observed that data sources

are enormously different in their format, unveiling large variety even for significant similar entities. Lastly, Data sources are of different attributes with outstanding varieties in the coverage, accuracy and timeliness of data provided [29].

3.3. CHALLENGES OF DATA INTEGRATION

The challenge with data integration falls on query optimization, inadequate resources (that is the cost and lack of skilled experts who understands the new data), skilled professional who are now difficult to find since data integration requires high professionals who understand the model of data and which right tools to use [30]. Other challenge known is scalability which crops up when new information from numerous sources are integrated with data from legacy systems and as a result influences the efficiency of legacy systems. This happens due to some system modification to fit the specification of new technologies. Another challenge is the ETL process as each data item passes through an ETL process, the transformation of huge data set will affect the storage ability of databases [30].

A research by [29] further explained that, there are three major challenges faced with data integration. One of which is Semantic Ambiguity where the form at which the data is stored in their sources are too broad to understand and these data formats are written differently in their different sources. The second challenge is the instance representation ambiguity, this is whereby an instance A might represent Airline1 in Airport1 and Instance A might be represented as Airfare in Airport2. Therefore, when the instances are represented differently from different sources, it will be difficult for integration. Lastly, data inconsistency is part of the challenge of data integration because one system may be outdated and the other system may have current date which may result into integrating data with diverse different date and time.

Another challenge of data integration is structural integration. This is related to the heterogeneity of a data model which is often related to legacy system. This legacy system known as out-of-date data cannot accept new requirements and technologies and this is because they are not managed by people or experts that understand it [31].

A researcher also pointed out that, the time and impact required to create the data source description and semantic mapping between the data source and the mediated schema is one major challenge in setting up a data integration application [32].

Other challenges of data integration are lack of data management expertise, unanticipated costs, uncertainty of data management, extracting value from data and bad data.

4. TRADITIONAL DATA WAREHOUSE

4.1. OVERVIEW

Data warehousing (DW) is a proper system which is used for reports and analysis and is widely considered an essential element for decision processes. In 1970s Bill Inmon was known as the father of DW and others (Barry Devlin, Ralph Kimball and Paul Murphy) who developed a theory of data warehousing with good examples [34]. The analysis of large quantity of data collected from different sources without powerful tools is of great challenge which has led to how data should be stored, integrate, managed and analysed. These requirements are led to the need of data warehousing and data mining [35]. Decision support has different requirements on database technology compared to the traditional online transaction processing application. DW plays a crucial role in decision making by providing reliable and efficient tools to decision makers. The data warehouse integrates a wide range of heterogeneous data sources in multidimensional structures that support decision making [36]. Data warehousing applies ETL technology [36]; [34]. With respect to the integration process of different sources, this paper by [36] suggested the use of the contextual model to conceptual design (CMCD). The paper explained that the CMCD will produce a unified approach in extracting automatically the DW multidimensional model in order to meet business framework. This CMCD is organized in 4 stages which are Contextual

modelling, Extraction of fact and dimensions, Extraction of measures and time granularity Extraction of non-functional requirements. With this model, they concluded that it is helpful in reduce the risk of inaccuracies between contextual business requirements analysis and DW modelling design.

Data warehousing aims to help the knowledge operative make better and faster decisions for present and future purposes. Data warehouse is typically maintained separately in a separate mode from the organization's operational databases. It also functions as a support to online analytical processing. As seen in figure 4.1 data warehousing uses summarized and consolidated data which are more useful for decision making than individual records. Exploration of a multidimensional model helps to identify the most significant dimensions for decision making. Any decision process requires data trends and predictions that are not immediately available in operational databases and also requires historical data whereas the operational database stores only current data and this will lead to challenges of missing information in the decision process [34].

The development cycle of a data warehouse includes requirements gathering, design dimension model, testing and maintenance phase. The design phase is divided into three. These are the (i) Conceptual design, (ii) Logical design implementation and (iii) physical design [35]; [72].
Traditional DW job is performed in batches during offline periods and since near-real time is becoming an essential requirement recently, DW experts think it is impossible to give end users near-real time data [34]. A researcher [37] suggested the Change Data Capture (CDC) approach to achieve near-real time DW. Other approaches suggested are Real Time Data Warehousing (RTDW) and Real Time Data Cache (RTDC).

Data warehouse architecture in figure 1 describes the process from data acquisition to data integration, to data repository, to analytics and finally to presenting. Data acquisition are controlled by software program used to acquire data from heterogeneous sources or hardware like sensors (detecting pollution from different location) so this data are taken from real-world to the hardware. Basically, data warehouse is like relational database which is used for analytical purposes. It is a centralized database where multiple data sources are stored together. Data repository refers to specific type of storage entities. It is like data mart which reserve a particular population of data isolated so it can be used for greater insight. They are like groups of database except for various purposes.

There are challenges faced with traditional data warehouse. These are data extraction and cleaning, data transformation and integration, efficient retrieval of result, testing of voluminous data from different sources, data security, the management of data warehouse, data optimization of complex queries, generic transformation for generating ETL physical schemas, and estimation of volume of data for testing [35]. Data warehouse is a tool widely used for analytical purposes but as the web based data increases rapidly, the relational database platform is not enough to handles problem arising from it. Relational database platform which was widely used has its advantages however, the recent business requirements and applications are changing frequently and it is observed that the relational database is no more suitable the newly evolved applications. Also research has observed that there are limited researches done to generate data warehouse in the environment of big data. There are some tools that are developed to solve big data issues in data warehouse, for example MongoDB but it has challenges in implementation due to the unstructured nature of data [38].

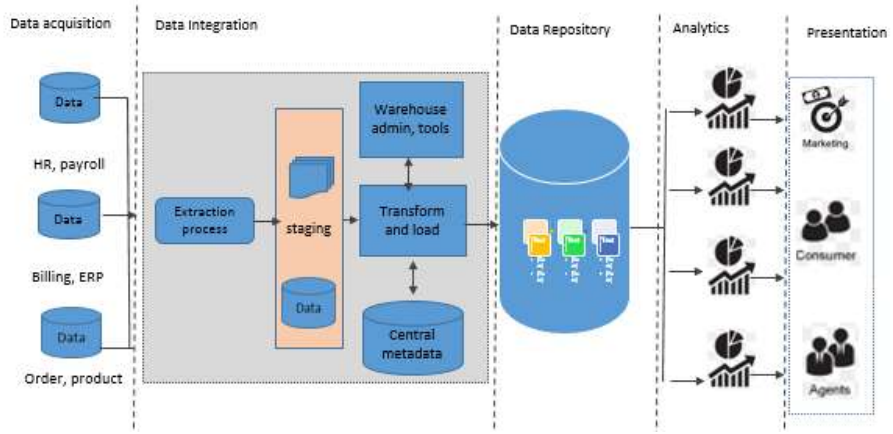

Figure 1 - The process of traditional data warehousing [34].

4.2. LOGICAL DATA WAREHOUSE

Logical data warehouse (LDW) is new data management architecture which is used for analysis. This approach combines the strength of traditional repository warehouse, alternative data management and access strategy [39]. This paper [40] explained LDW as a single view of data. The logical layer of the data warehouse delivers different techniques for viewing data both in warehouse store and across different organizations without transporting or transforming data. With logical data warehouse views are provided and these views act as an interface into different data and its sources [41]. Logical data warehouse is needed because organizations are not yet satisfied with the traditional data warehousing, enterprises are seeking for a better management of data like LDW because the traditional data warehouse is slow in process and it has also increased within end users' organizations [40]. With logical data warehousing different enterprise data will be treated as one singular data warehousing for large-scale information management.

With the use of data virtualization, access is granted easily to data and federated data. Data virtualization layer are capable of initiating externally managed processes [40]. Element that forms the logical data warehousing are Repository management, Data Virtualization, Distributed Processes, Auditing Statistics and performance evaluation services, SLA (Service Level Agreement), management, taxonomy or ontology Resolution and Metadata management [39]. Logical data warehouse also responds faster to your ever-changing analytics and business intelligence.

4.3. VIRTUAL VIEWS

A view is a saved SQL query. It is a virtual table that executes SQL statements. A view provides up-to-date data. View permission allows a user see the metadata of which the permission is granted [42]. Views can limit the amount of disclosure of the information in the main table to the public, with views multiple table can be joined into a single virtual table. Views can appear as aggregate table meaning it can sum-up data and present a calculated result as part of the data; therefore, it hides the complexity of data.

5. DATA MANAGEMENT TECHNIQUES AND LIMITATIONS

There is a large research about data management techniques which involves data integration and data warehousing and its challenges. Different techniques and frameworks have been proposed by various researchers. In this section, the paper overviews different literatures that focused on the big data management and analytics techniques, and also the application of each framework is critically reviewed.

This paper [43] explored the application of data integration and big data techniques on smart healthcare. The framework which is mediator environment for multiple information sources (Momis) is proposed to collect, integrate and display data in more efficient way. This framework allows users to search among clinical records in order to expand brief reports and discover new relations on data collected. This is for the purpose of proper analysis to avoid errors. The Momis framework applied integration which provides a unified access to data from different databases without migrating all data to a single system. In comparison to this paper by [44], they implemented a technology called Dr.warehouse which is an open source data warehouse to manage clinical reports. The framework Dr.warehouse, provides search engines and it uses graphical user interface to explore text for search results. In both frameworks Dr.warehouse and Momis, the data collected was not in real-time and therefore, merging with new data may produce errors in analysis. Also in this framework Dr.warehouse, misspelling can occur and this may lead to error in responding to request.

In their research [45], they focused on the navigation plans when building data integration as a challenge. The paper explained further that the schema used for integration are challenging when data sources are Web of data. The research therefore proposed the use of global-local-as-view (GLAV). He highlighted that using the local-as-view (LAV) schema and the global-as-view (GAV) schema alone is not efficient as it has its own consequences. The GLAV has a flexible schema that does not depend on any particular transformation or details from the source. With this framework, the algorithm for answering queries is not efficient enough if additional constraints are stated on the mediated schema. Friedman, Levy and Millstein [45] pointed out that the use of GAV and LAV is not sufficient enough. Another researcher [20] explained the theoretical aspect of data integration which are the Local-as-view (LAV) and Global-as-view (GAV) semantics have been discovered that there is a need for more investigation in deep understanding of the relationship between LAV and GAV approach. He also pointed-out that there are still challenges associated with the algorithm design and complexity of new-based query processing which relates to integrity constraints in the global schema.

Another research paper applied big data technology to transportation sector using Markov techniques to match the transportation demands with city service provision [16]. This method shows how sharing transport load in a smart city can improve the efficiency in meeting city service demands. It is observed that, with this approach an in-depth modelling in implementing big data initiatives is required. While another paper by [46] explored the application of data integration on biomedical informatics explaining that, there are different sources of data in this angle. The researchers explained different steps (data integration, data standardization, data mining and knowledge discovery) for supporting decision making in translational bioinformatics for basic analysis. Also the research by [47] applied big data solution on predicting risk of readmission for congestive heart failure patients. Big data solution was used because clinicians recently struggle to use the traditional system to store data and analyse data. The paper further explained that extracting data from different sources independently is difficult. Therefore, the researchers used the application of Apache Hadoop as a big data solution to the challenging issues mentioned above [47] Adopting the use of Apache Hadoop has its own limitation; it has slow processing speed because MapReduce processes large data sets. It does not process streamed data which lead to the overall performance being slower.

Bansal and Kagemann [1] explained different current ETL tools which do not include a meaningful semantic relationship to the integration of data from heterogeneous sources. They proposed a framework which is semantic ETL; this applies semantics to various data fields and allows effective data integration. The framework generates a semantic model of the dataset to be integrated and finally generates semantic linked data that complies with the data model.

This research [48] analysed different challenges of big data and its integration process. This research argued that the existing techniques of data warehouse are now inefficient to manage these data that come from heterogeneous sources and how organization are striving to find new

solutions to these challenges. They reviewed different papers and organized different problem arising in big data integration (the integration of crowd sourcing data, inadequate tools to be provided in managing data sources, indexing techniques to improve data access performance and many more). The researchers [48] suggested that, new techniques and framework are needed to solve the issues in big data environment, while [33] identified other issues in big data integration and these include; normalization, integration and transformation of data from many sources into a format required to run large scale analysis and visualization tools. The paper proposed an approach which maps diverse sources into shared domain ontology. The system designed allows a developer to easily define correct data, reshape plans that transform data and restructure the output of one tool into the input of another tool.

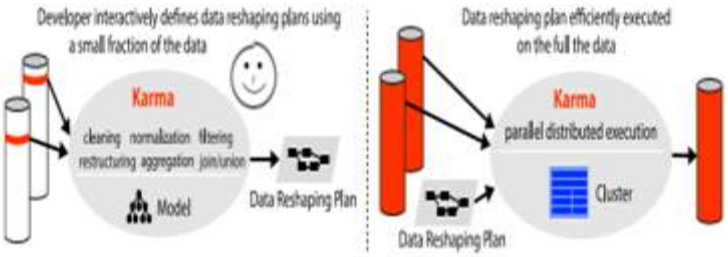

Figure 2 - Diagram of Karma model [33].

The problem with this approach is that, the developer can access all data which is not feasible and secured. The system also lacks the decision control on the quality of data needed, to build correct and reliable analysis. This therefore, makes the quality of data is relatively poor when using this approach.

The author [49] explained how transactional data are represented in different formats emphasizing that XML format is recognized worldwide and organization often adopt it. The paper then focuses on integrating XML data which is based on multiple related schemas. The researchers then discovered that more than one data warehouse can be identified from single related XML schema. The proposed method of achieving this was the use of schema graph which represented entities found in the XML schema. This framework identified multiple schemas from XML schema and was able to convert the XML schema to ROLAP data warehouse schema. It is challenging for this framework when semi-structured data are added to XML. Furthermore, there will be need for reconstruction of schema structure if the source XML changes.

These researchers [50] explained the increasing growth of linked data and the challenges associated with the development of link data application. Therefore, they [50] produced a framework known as linked data integration framework (LDIF) which is used as a component within linked data application to collect linked data from web and translate the data collected into a clean local target representation while ensuring that data provenance is intact. For the translation, it uses an expressive mapping language. This research [51] focused on three key challenges of linked open data (LOD) integration which are the data quality assessment, conflict resolution and quality improvement. The researcher also reviewed other different techniques proposed by other papers in tackling these three challenges and proposed method to evaluate the data quality challenge. They proposed a workflow for data quality assessment. These are Resource selection, Data quality problem profiling, Source mapping, Data quality evaluation and data fusion, and data quality improvement.

Most of the techniques in the existing literature still copy a lot of data for analysis and the level of agreement in data collection is not automated. These techniques do not support data sharing from different sources in the form of logical data warehouse.

Table 2. This table presents a critical review on the different data management techniques.

Techniques Used?	**Application on?**	**Method of data collection**	**Data integration involved?**	**In Real-time?**	**Reference**
Analysis using Hadoop	Electric Vehicles	Automated (sensor)	Yes	Not specified	[52]
Qualitative approach	Smart cities	Not applicable	No	No	[2]
Extract Transform Load approach	IPUMS (Integrated Public Use Microdata)	Not Specified	Yes	No	[26]
Dr.Warehouse approach	Clinical Reports	Manual data-entry	Yes	No	[44]
MOMIS approach	Facioscapulohumeral dystrophy genetic disease patient's data	OPenClinica via Electronic Data Capture (EDC)	Yes	No	[43]
Not specified	Airline information	Not known	Yes	No	[29]
Using VLC (visible light communication) an Foglet	Road asset reporting with smart vehicles	Automated	Not known	No	[53]
Using AVL (automatic vehicle location) approach	Transportaton (vehicle routes prediction)	Automated (Sensor)	Yes	Yes	[54]
Using GIS (geographical information system) for analysis	Transportation (traffic accident)	Not automated	Yes	No	[55]
Using wearable sensor device	Air quality	Automated	Yes	Not known	[56]
Using Alphasense techniques	Air quality monitoring	Automated	Yes	Yes	[57]
Qualitative approach of data integration	Smart city	Not applicable	N/A	N/A	[58]
Using semantics ETL	House-hold and fuel economy data	Data collected in CSV format	Yes	Not known	[1]
Markovan approach	Transportation management	Not specified	Yes	No	[16]
Data integration using cytoscape techniques	Biological system	Not specified	Yes	No	[59]

6. PROPOSED FRAMEWORK

The proposed framework in Figure 3 depicting how the traditional process that uses the Extract Transform Load techniques to extract data from the source eliminated or replaced. This figure 3 explains that rather than copying the whole information from different sources to single data repository, integration can still be done by creating a single line of communication between the clients and contributors data. Here, data that is needed to work with are provided in the views based on the service level agreement with each data contributors and access is made to the views only with the permission granted from each contributors.

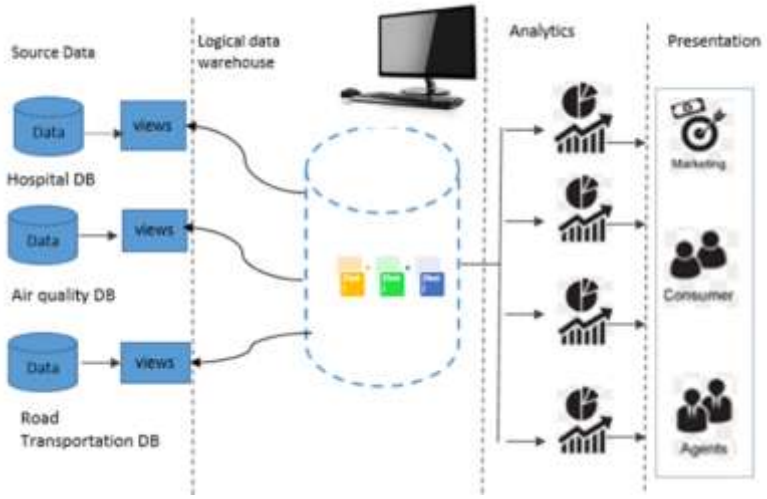

Figure 3 – showing virtual data integration

In this proposed framework, figure 4 explained the flow chart of the implementation. When an analytics is to be conducted which required two or more organizations' data, the government agency will provide certain things on paperwork which he requires from each company and the data format required. There will be a service level agreement (SLA) between the government and the contributors of data. After the agreement is made, contributors will provide the views with the format features then the government agency accesses the view through URL (HTTP concept) with specific aggregated SQL query. The government agency's database will have to adopt the RESTView technology which will aid in accessing the view using HTTP concept and the result appears in JSON format which can be convert. This new form of data integration do not support copying all of data, it rather access data provided and also access data when it is needed.

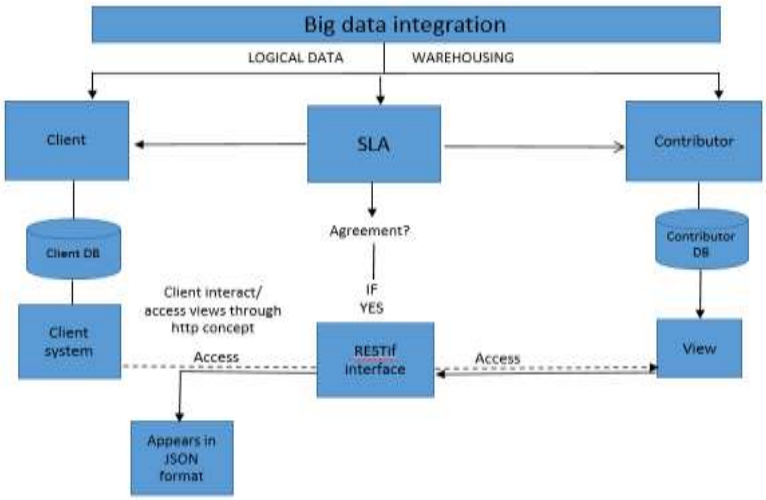

Figure 4 – Flow chart of the proposed model

6.1. PROPOSED METHOD OF BIG DATA INTEGRATION

The proposed framework is applying logical data warehouse to data integration for analysis whereby specific data can be accessed from individual database by sending queries to the contributor through Http concept whereby the customer's database adopts REST-view interface. This implies that, the application of REST interface on database for integration will be the mediator between the data source and the consumers' data warehouse where the process of integration is done [67]. The RESTView technology makes it possible to extract data without copying the whole data from database and this denote a logical data warehouse. This project attempts to use REST interface on any of the database tools or data integration tools where data can be gotten through a certain query made using the HTTP concept. The query sent by customer is done by specific selected aggregated information and it is when required, together with certain level of agreements between the contributor and customer.

Logical data warehouse which is a new data management is used for the process of analytics. With logical data warehousing different enterprise data will be treated as one singular data warehousing for large-scale information management.

With the use of data virtualization, access is granted easily to data and federated data. Element that forms the logical data warehousing are: Repository management, Data Virtualization, Distributed Processes, Auditing Statistics and performance evaluation services, SLA (Service Level Agreement) management, taxonomy or ontology Resolution and Metadata management [39]. It also responds faster to your business intelligence and ever-changing analytics.

Data virtualization falls under metadata and compiler-oriented approach. By using a proper specified meta-data description language, the database developer can effectively understand the format of datasets which are generated and used by the application. A virtual view and SQL query from relational database provides a very suitable means for identifying subsets of interest (metadata). A data analyst built databases by implementing an internal data representation to showcase heterogeneous raw data-set and thereby, allowing SQL query to be comprehensive enough to combine heterogeneous data sources into a single understanding query language [68].
Data virtualization is of great use in the business intelligence system because it is easier to change the system, new statistics or reports and existing ones can adapt quickly and easily. Data virtualization applies technologies like data abstracting, data integration, encapsulation and enterprise information integration. New sources of data can now be found out of the boundaries of organization and this has led to organization's interest in combining their own internal data with these new data sources for analytical and reporting abilities. When data virtualization is applied, it stands as an intermediary that hides most of the technical aspects of how and where data is stored in an application. Encapsulation as related to data virtualization does the hiding of the important aspect of data-store like the API, location, storage format, and access mechanism. With this it makes an application swift and maintainable. With respect to abstracting, it makes the consumer only see the data in an aggregated form not necessarily showing other columns that are not needed to the consumer. Data virtualization as data integration does not show how the integration process happens. It only produces a unified access to data.

Therefore, data virtualization is defined as a technology which offers a unified view of data to consumers, which hides the facts that data is integrated, also hides the technical access details and provides data on an aggregated level [69]. Data virtualization as defined by [69] is a technical class of platform used by consumers to combine data from disparate sources into single virtual data layer which provide integrated information services to various users in real-time.
What is required in the proposed model are:

- Http concept
- RESTView interface
- Service Level Agreement
- Virtual views

6.2. RESTVIEW INTERFACE

Representational Sate Transfer (REST) is a web design concept. It provides standard steps on how it should be used in developing web applications. REST is associated with HTTP. The aim of REST is to create a single / general interface whereby applications share the same convention [70]. This means one application knows how to communicate to all other app and vice versa. It provides independent deployment of component. This means if one REST interface is implemented on any application, you can always implement and deploy the REST interface without rewriting or modifying existing ones. One can make an application REST-like application by just using or adding REST interface on its application.

6.2.1. CONCEPT OF REST

Any resources can access URL when REST over HTTP is implemented in terms of request and response. The interaction between the client (requester) and sever (contributor) is organised with request from client to server and vice versa and this can be achieved by each request containing representation of the resource.

The REST representation is the current status of a resource. This can be when a client make new request and thereby updating the resource which is the current status at that time is the representation.

6.2.2. PRINCIPLES OF REST

State of Resource remains internal, Client make request or update to the server, the server will not store status of the client. The client request server responds immediately and do not save the context of the request made by event. Client request must contain all information enough for the server to process it. Information between the logical server and the logical client remained on client's side.

6.3. BENEFIT OF THIS FRAMEWORK

It is not always easy to access corporate data and make value out of it and that is the greatest challenge to producing effective reporting. In some organizations, information is difficult to access or its availability is limited due to unwillingness of some firm to consolidate their data into a single data repository. Using this framework, data can be left in their original source rather than copying this data into one central database, the framework will allow one to access data through a mediator (Restif acting as micro-services) which creates a line of communication thereby leaving data in their original sources. It also reduces complexity in data modelling and this is because Pyrrho Database which has RESTView technology does not need to re-write programming code. It promotes data transparency since there will be an agreement between the client and the contributor in providing the views according to SLA and therefore, the client works with the views provided. This framework can also allow one to interrogate live data. This framework reduces data redundancy since it doesn't encourage one to copy all information from different sources to one central data repository.

In terms of security issues, when dealing with integration of data from many sources, security mechanisms should be considered at various levels. It is a major requirement for data scientist to think or analyse on how a development of tools will co-operate with diverse data sets and their security demands [73]. Some of the traditional security mechanisms adopted when using relational database is access control which can be role based or credential based access control and this can feature in authentication, authorization and auditing [74] but these are not enough in the era of big data.

In big data analytics (process of data integration and warehouse), some of the security issues and challenges can vary from the distinction of data source and pattern and understanding the nature of data acquisition from diverse sources. The security challenges include how database storage can be protected; worrying about the end-point input validation; the real-time security and compliance monitoring; and how to detect metadata dependencies for confidentiality and security purpose [75]. With the proposed method of big data integration, the security challenge related to the data storage database will be minimised since the data owners are in-charge of their own data and using the logical data warehouse concept, data retrieved from different data source will not be store anywhere. Another security advantage it has is that, data from various sources is secured and the clients only have access to the views in which the data owners are willing to give access to base on service level agreement.

7. CONCLUSIONS

In this paper, we addressed the different data management process and its' challenge. We reviewed different paper which adopted different techniques and stated the problem with each technique. After reviewing different paper, it is observed that many process of data extraction involves the traditional method and do not have a virtualization or logical process of data analytics copying all of the data for analytics can lead to data redundancy and also some techniques do not have easy access to data for decision processes. In this era of big data, the challenges expanded to the volume, velocity and variety of data. Therefore, the implementation of big data integration involved the big data challenges and this is a high security risk if all data are being extracted from the sources in making analysis and decision. Thus, introducing logical data warehouse to traditional data integration provides a logical view of data and then data can only be used based on the specific reason and specified queries sent to the contributor. The quality of information is a problem when considering data sharing among different organization. Using our proposed model encourages easy access to data due to data sharing, reduces redundant data (by avoiding copying the whole data for analysis), reduces data complexity because it adopts easy semantics and for security purpose which uses logical warehouse whereby after data are used for the specific purpose it do not store. In this framework, more research is needed in the security aspect to discover the security issues it might face using it.

REFERENCES

[1] Bansal, S. and Kagemann, S. (2015). Integrating Big Data: A Semantic Extract-Transform-Load Framework. Computer, 48(3), pp.42-50.

[2] Saxena, S. and Kumar Sharma, S. (2016). Integrating Big Data in "e-Oman": opportunities and challenges. info, 18(5), pp.79-97.

[3] Khan, M., Wu, X., Xu, X. and Dou, W. (2017). Big data challenges and opportunities in the hype of Industry 4.0. 2017 IEEE International Conference on Communications (ICC).

[4] Karafiloski, E. and Mishev, A. (2017). Blockchain solutions for big data challenges: A literature review. IEEE EUROCON 2017 -17th International Conference on Smart Technologies.

[5] Crowe, M., Begg, C., Laiho, M. and Lau, F. (2016). Data validation for Big Live data.[online] Available at: https://www.researchgate.net/publication/315686427_Data_Validation_for_Big_Live_Data

[6] Chauhan, S., Agarwal, N. and Kar, A. (2016). Addressing big data challenges in smart cities: a systematic literature review. info, 18(4), pp.73-90.

[7] Al Nuaimi, E., Al Neyadi, H., Mohamed, N. and Al-Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6(1).

[8] Rabiul, I., Islam, R., Musfiqur, R. and Abiduzzaman, R. (2016). Big Data Characteristics, Value Chain and Challenges.

[9] Almeida, F. and Calistru, C. (2013). The main challenges and issues of big data management. International Journal of Research Studies in Computing, 2(1).

[10] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute Reports, 5, 15-36.
[11] Sivarajah, U., Kamal, M., Irani, Z. and Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, pp.263-286.
[12] De Oliveira Veras, A., de Sá, P., da Costa Pinheiro, K., Barh, D., Azevedo, V., Jucá Ramos, R. and da Costa da Silva, A. (2018). Computational Techniques in Data Integration and Big Data Handling in Omics. Omics Technologies and Bio-Engineering, pp.209-222.
[13] Mishra, S., Dhote, V., S. Prajapati, G. and Shukla, J. (2015). Challenges in Big Data Application: A Review. International Journal of Computer Applications, 121(19), pp.42-46.
[14] TOLE, A. (2013). Big Data Challenges. Database Systems Journal, vol. IV, p.no. 3.
[15] Trifu, M. and Ivan, M. (2014). Big Data: present and future. Database Systems Journal, 5(1), pp.32-41.
[16] Mehmood, R., Meriton, R., Graham, G., Hennelly, P. and Kumar, M. (2017). Exploring the influence of big data on city transport operations: a Markovian approach. International Journal of Operations & Production Management, 37(1), pp.75-104.
[17] CEBR (2012). Data equity: unlocking the value of big data. Centre for Economics and Business Research White Paper, 4, 7-26.
[18] McNulty, E. and Freeman, H. (2014). Understanding Big Data: The Seven V's - Dataconomy. [online] Dataconomy. Available at: http://dataconomy.com/2014/05/seven-vs-big-data/.
[19] Chaudhuri S. What next?: a half-dozen data management research goals for big data and the cloud. In Proceedings of the 31st symposium on Principles of Database Systems. ACM; 2012. pp. 1–4.
[20] Lenzerini, M. (2002). Data integration. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02.
[21] Merelli, I., Pérez-Sánchez, H., Gesing, S. and D'Agostino, D. (2014). Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. BioMed Research International.
[22] George, G., Osinga, E., Lavie, D. and Scott, B. (2016). Big Data and Data Science Methods for Management Research. Academy of Management Journal, 59(5), pp.1493-1507.
[23] Wang, L. and Alexander, C. (2015). Big Data in Distributed Analytics, Cybersecurity, Cyber Warfare and Digital Forensics. Science and Education Publishing, Vol. 1, pp.pp 22-27.
[24] Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. and Salehian, S. (2018). The 10 Vs, Issues and Challenges of Big Data. Proceedings of the 2018 International Conference on Big Data and Education - ICBDE '18.
[25] Smith, J., Bernstein, P., Dayal, U., Goodman, N., Landers, T., Lin, K. and Wong, E. (1981). Multibase. Proceedings of the May 4-7, 1981, national computer conference on - AFIPS '81.
[26] Ruggles, S., Hacker, J. and Sobek, M. (1995). General Design of the Integrated Public Use Microdata Series. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 28(1), pp.33-39.
[27] Ray, S., Bandyopadhyay, S. and Pal, S. (2009). Combining Multisource Information Through Functional-Annotation-Based Weighting: Gene Function Prediction in Yeast. IEEE Transactions on Biomedical Engineering, 56(2), pp.229-236.
[28] Mireku Kwakye, M. (2011). A Practical Approach to Merging Multidimensional Data Models. [online] Dx.doi.org. Available at: http://dx.doi.org/10.20381/ruor-5072.
[29] Dong, X. and Srivastava, D. (2015). Big Data Integration. Synthesis Lectures on Data Management, 7(1), pp.1-198.
[30] Kadadi, A., Agrawal, R., Nyamful, C. and Atiq, R. (2014). Challenges of data integration and interoperability in big data. 2014 IEEE International Conference on Big Data (Big Data).
[31] Ziegler, P. and Dittrich, K. (2008). Data Integration — Problems, Approaches, and Perspectives.
[32] Halevy, A., Rajaraman, A. and Ordille, J. (2006). Data integration: the teenage years. VLDB '06 Proceedings of the 32nd international conference on Very large data bases, pp.9 - 16.
[33] Knoblock, C.A. and Szekely, P., 2013, November. Semantics for big data integration and analysis. In 2013 AAAI Fall Symposium Series.
[34] Mukherjee, R. and Kar, P. (2017). A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight. 2017 IEEE 7th International Advance Computing Conference (IACC).
[35] Chandra, P. and Gupta, M. (2017). Comprehensive survey on data warehousing research. International Journal of Information Technology, 10(2), pp.217-224.

[36] Chakiri, H., Mohajir, M. and Assem, N. (2017). CMCD: A data warehouse modeling framework based on goals and business process models. 2017 IEEE AFRICON.
[37] Sultan, U. (2016). Literature review on real time data warehousing. Karachi, Pakistan: Institute of Business Administration Karachi.
[38] Maity, B., Sen, S. and Debnath, N. (2018). Challenges of Implementing Data Warehouse in MongoDB Environment. Journal of Fundamental and Applied Science, pp.222 - 228.
[39] Cisco (2017). Logical Data Warehouse | Cisco. [online] Compositesw.com. Available at: http://www.compositesw.com/solutions/logical-data-warehouse/
[40] Imagesrv.gartner.com. (n.d.). Logical Data Warehouse for Big Data. [online] Available at: http://imagesrv.gartner.com/media-products/pdf/samples/sample3.pdf.
[41] Russom, P. (2015). The Logical Data Warehouse: What it is and why you need it | Transforming Data with Intelligence. [online] Transforming Data with Intelligence. Available at: https://tdwi.org/webcasts/2015/06/the-logical-data-warehouse-what-it-is-and-why-you-need-it.aspx.
[42] Docs.microsoft.com. (2017). Views. [online] Available at: https://docs.microsoft.com/en-us/sql/relational-databases/views/views?view=sql-server-2017.
[43] Orsini, M., Calanchi, E., Magnotta, L., Gagliardelli, L., Govi, M., Mele, F. and Tuplert, R. (2017). The Italian FSHD registry: An enhanced data integration and analytics framework for smart health care. 2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI).
[44] Garcelon, N., Neuraz, A., Salomon, R., Faour, H., Benoit, V., Delapalme, A., Munnich, A., Burgun, A. and Rance, B. (2018). A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. Journal of Biomedical Informatics, 80, pp.52-63.
[45] Friedman, M., Levy, A. and Millstein, T. (1999). Navigational Plans for Data Integration. AAAI-99 Proceedings.

[46] Yan, Q. (2017). Data Integration, Data Mining, and Decision Support in Biomedical Informatics. Translational Bioinformatics and Systems Biology Methods for Personalized Medicine, pp.41-52.
[47] Zolfaghar, K., Meadem, N., Teredesai, A., Roy, S., Chin, S. and Muckian, B. (2013). Big data solutions for predicting risk-of-readmission for congestive heart failure patients. 2013 IEEE International Conference on Big Data.
[48] Arputhamary, P. and Arockiam, L. (2014). A Review on Big Data Integration. International Journal of Computer Applications (0975 – 8887) Advanced Computing and Communication Techniques for High Performance Applications (ICACCTHPA-2014).
[49] Sen, S. (2012). Integrating XML Data Into Multiple Rolap Data Warehouse Schemas. International Journal of Software Engineering & Applications, 3(1), pp.197-206.
[50] Andreas, S., Mendes, P., Bizer, C., Becker, C., Matteini, A. and Isele, R. (2012). LDIF -A Framework for Large-Scale Linked Data Integration. [online] Available at: http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Schultz-et-al-LDIF-WWW2012-DevTrack.
[51] Ahmed, H. (2017). Data Quality Assessment in the Integration Process of Linked Open Data (LOD). 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA).
[52] Bolly, V. (2014). System For Delivering Electric Vehicle Data. Open Access Theses, p.406
[53] Hussain, F., Farahneh, H., Fernando, X. and Ferworn, A. (2017). VLC Enabled Foglets Assisted Road Asset Reporting. 2017 IEEE 85th Vehicular Technology Conference (VTC Spring).
[54] Karbassi, A. and Barth, M. (2003). Vehicle route prediction and time of arrival estimation techniques for improved transportation system management. IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683).
[55] Shafabakhsh, G., Famili, A. and Bahadori, M. (2017). GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. Journal of Traffic and Transportation Engineering (English Edition), 4(3), pp.290-299.
[56] Rizea, D., Olteanu, A. and Tudose, D. (2014). Air quality data collection and processing platform. 2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference.
[57] Hojaiji, H., Goldstein, O., King, C., Sarrafzadeh, M. and Jerrett, M. (2017). Design and calibration of a wearable and wireless research grade air quality monitoring system for real-time data collection. 2017 IEEE Global Humanitarian Technology Conference (GHTC).
[58] An, X., Deng, H. and Sun, S. (2016). 11th International Conference on Cyber Warfare and Security - Data integration in the development of smart city china. Towards a digital continuity model. Pp.13 – 20.

[59] Smoot, M., Ono, K., Ruscheinski, J., Wang, P. and Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics, 27(3), pp.431-432.
[60] Salinas, S. and Nieto Lemus, A. (2017). Data Warehouse and Big Data Integration. International Journal of Computer Science and Information Technology, 9(2), pp.01-17.
[61] Xu, K., Zhen, H., Li, Y. and Yue, L. (2017). Comprehensive Monitoring System for Multiple Vehicles and Its Modelling Study. Transportation Research Procedia, 25, pp.1824-1833.
[62] Alzoubi, H. (2018). Natural language processing and machine learning techniques applied on Electronic Health Records. IEEE journals of biomedical and health informatics.
[63] Ma, Z., Zhao, Z. and Yan, L. (2018). Heterogeneous fuzzy XML data integration based on structural and semantic similarities. Fuzzy Sets and Systems.
[64] Rajeswari, C., Basu, D. and Maurya, N. (2017). Comparative Study of Big data Analytics Tools: R and Tableau. IOP Conference Series: Materials Science and Engineering, 263, p.042052.
[65] Regueiro, M., Viqueira, J., Taboada, J. and Cotos, J. (2015). Virtual integration of sensor observation data. Computers & Geosciences, 81, pp.12-19.
[66] U ddin, M., Lie, H. and Li, H. (2017). Hybrid Cloud Computing and Integrated Transport System. 2017 International Conference on Green Informatics (ICGI).
[67] Crowe, M. and Begg, C. (2017). REST and the Management of Distributed Data.[online] Available at https://www.researchgate.net/publication/331558558_REST_and_the_Management_of_Distributed_Data
[68] Karpathiotakis, M., Alagiannis, I., Heinis, T., Branco, M. and Ailamaki, A. (2015). ust-In-Time Data Virtualization: Lightweight Data Management with ViDa. Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR). [online] Available at: http://wp.doc.ic.ac.uk/theinis/publication/just-in-time-data-virtualization-lightweightdata-management-with-vida/
[69] Denodo. (2018). Data Virtualization: How It Works. [online] Available at: https://www.denodo.com/en/data-virtualization/how-it-works.
[70] Fielding, R. (1999). Architectural Styles and the Design of Network-based Software Architectures. [online] Ics.uci.edu. Available at: https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf
[71] Big Data: Security Issues, Challenges and Future Scope. (2016). International Journal of Research Studies in Computer Science and Engineering, 3(3).
[72] Jindal, R. (2012). Comparative Study of Data Warehouse Design Approaches : A Survey. International Journal of Database Management Systems, 4(1), pp.33-45.
[73] (Jayasingh, B., Patra, M. and Mahesh, D. (2016). Security issues and challenges of big data analytics and visualization. 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I))
[74] Chaudhari, S. and P, V. (2013). Conceptual Framework for Geospatial Data Security. International Journal of Database Management Systems, 5(5), pp.29-35.g.
[75] Dev Mishra, A. and Beer Singh, Y. (2016). Big data analytics for security and privacy challenges. 2016 International Conference on Computing, Communication and Automation (ICCCA).
[76] Vijayaraj, J., Saravanan, R., Victer Paul, P. and Raju, R. (2016). A comprehensive survey on big data analytics tools. 2016 Online International Conference on Green Engineering and Technologies (IC-GET).
[77] Mackie, R. (2012). Application of service oriented architecture to finite element analysis. Advances in Engineering Software, 52, pp.72-80.
[78] Herak, S. (2016.). Service-oriented architectures (SOA) and their application and usage in healthcare. [online] engr.uconn.edu. Available at: http://www.engr.uconn.edu/~steve/Cse300/sen.pdf.
[79] Nadal, S., Romero, O., Abelló, A., Vassiliadis, P. and Vansummeren, S. (2018). An integration-oriented ontology to govern evolution in Big Data ecosystems. Information Systems.
[80] Shao, X., Tan, C., Voss, C., Li, S., Deng, N. and Bader, G. (2010). A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain–peptide interaction from primary sequence. Bioinformatics, 27(3), pp.383-390.
[81] Adjei, E. (2015). L - Fuzzy Structured Query Language. [online] Dr.library.brocku.ca.
[82] Zolfaghar, K., Verbiest, N., Agarwal, J., Meadem, N., Chin, S.C., Roy, S.B., Teredesai, A., Hazel, D., Amoroso, P. and Reed, L. (2013). Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach. arXiv preprint arXiv:1306.2094.

[83] Gazali, Kaur, S. and Singh, I. (2017). Artificial intelligence based clinical data management systems: A review. Informatics in Medicine Unlocked, 9, pp.219-229.
[84] Zolfaghar, K., Meadem, N., Teredesai, A., Roy, S., Chin, S. and Muckian, B. (2015). predicting risk-of-readmission for congestive heart failure patients on Big data solutions. International Journal of Engineering Development and Research, Volume 3,(Issue 2), pp.ISSN: 2321-9939.
[85] Elgendy, N. and Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. Advances in Data Mining. Applications and Theoretical Aspects, pp.214-227.
[86] Durairaj, M. and Poornappriya, T. (2016). A Review on Big Data Analytics Tools for Telecommunication Industry. International Journal of Circuit Theory and Applications, 9, pp.185-193.

AUTHORS

Offia Chisom Ernesther studied her bachelor's degree on BSc computer Engineering and graduated with Second class Upper Division in 2015 at Ghana Telecom University College, Accra Ghana. She furthered her Education in Teesside University; Middlesbrough, England United Kingdom on MSc Computer Security and Networks and graduated with Merit. She started her PhD degree in the year 2017 at the University of the West of Scotland, Paisley Scotland with the research focus on achieving logical data warehouse in the process of data analytics (Data Integration).

Malcolm Crowe was born in Dublin 13 January 1948. His first and only employment was at Paisley College in 1972, now the University of the West of Scotland, where he is an Emeritus Professor. He received his D.Phil from Oxford University in 1978 with a thesis entitled "The Connective k-theory of the Infinite Symmetric Group". Computing rather than Mathematics is his real passion. During the 1980s and 1990s he worked on several ESPRIT projects and co-authored several books on CSCW, Information Systems, and Interdisciplinary Research.Since 2006 he has been developing new ideas for how to improve the implementation of relational DBMS, where he regards the maintenance of the transaction log as the only guarantee of accountability and transactional behaviour. He developed a strongly typed optimistic RDBMS called Pyrrho (www.pyrrhodb.com) to try out several technical additions to relational technology including role-based namespaces and semantic and distributed databases. In 2016 Malcolm began to focus on "Big Live Data" (Virtual Data Warehousing), and his collaborators on this concept include Fritz Laux who presented a paper on this concept at DBKDA 2017.