Obilikwu and Ogbuju *J Big Data* (2020) 7:95
https://doi.org/10.1186/s40537-020-00370-1

Journal of Big Data

RESEARCH Open Access

A data model for enhanced data comparability across multiple organizations

Patrick Obilikwu[1] and Emeka Ogbuju[2*]

*Correspondence: emeka.ogbuju@fulokoja.edu.ng
[2] Department of Computer Science, Federal University, Lokoja, Nigeria
Full list of author information is available at the end of the article

Abstract

Organizations may be related in terms of similar operational procedures, management, and supervisory agencies coordinating their operations. Supervisory agencies may be governmental or non-governmental but, in all cases, they perform oversight functions over the activities of the organizations under their control. Multiple organizations that are related in terms of oversight functions by their supervisory agencies, may differ significantly in terms of their geographical locations, aims, and objectives. To harmonize these differences such that comparative analysis will be meaningful, data about the operations of multiple organizations under one control or management can be cultivated, using a uniform format. In this format, data is easily harvested and the ease with which it is used for cross-population analysis, referred to as data comparability is enhanced. The current practice, whereby organizations under one control maintain their data in independent databases, specific to an enterprise application, greatly reduces data comparability and makes cross-population analysis a herculean task. In this paper, the collocation data model is formulated as consisting of big data technologies beyond data mining techniques and used to reduce the heterogeneity inherent in databases maintained independently across multiple organizations. The collocation data model is thus presented as capable of enhancing data comparability across multiple organizations. The model was used to cultivate the assessment scores of students in some schools for some period and used to rank the schools. The model permits data comparability across several geographical scales among which are: national, regional and global scales, where harvested data form the basis for generating analytics for insights, hindsight, and foresight about organizational problems and strategies.

Keyword: Collocation, Data comparability, Data aggregation, Data model, Big data

Introduction

Where multiple organisations are related, the need to compare their operational output requires that their independent heterogeneous operational datasets are compatible for analysis. The suitability of datasets from heterogeneous sources for analysis is referred to as data comparability and the nature of such analysis is referred to as cross-population analysis [1]. Data comparability is therefore an attribute of datasets that renders them usable for cross-population comparative analysis. A major difficulty in achieving data comparability is the heterogeneous nature of data obtained from multiple organisations and sources. The inherent heterogeneity is eliminated or reduced to their barest

minimum by cleansing and reformatting the datasets when the datasets are aggregated from their different sources. Aggregation is a step in the data value chain and the techniques involved in aggregating data are data cleansing, reformatting and data integration [2].

Aggregating data from different sources to support decision-making at the level of management or supervision is a very important task. Aggregated data is required for decision-support because a decision that is not based on verifiable data creates room for contention and controversy leading to opposition to such a decision. Inherent in decision-making that is supported with data, is the fact that arriving at such decision is preceded by rigorous comparative analysis. Where the data used for comparative analysis is acquired from disparate sources, the associated heterogeneity creates problems that render the data incomparable. Depending on the nature of the decision scenario, the associated comparison could be between multiple processes in an organization or even between multiple organizations. The organizations could, in turn, be located within the same country or in geographically disperse locations. Data comparability is global if it involves governments or Non-Governmental Organizations (NGOs) whose operational scope is worldwide. The United Nations (UN), for example, has set up United Nations Educational, Scientific and Cultural Organization (UNESCO) with a declared purpose to contribute to promoting international collaboration in education, sciences, and culture to increase universal respect for justice, the rule of law, and human rights to fundamental freedom as proclaimed in the UN Charter. The UN Charter sets up the basis of cooperation between member nations of the United Nations [3]. In this way, UNESCO promotes education in member countries on behalf of the UN and thus can be said to have oversight and monitoring functions over education in member countries on behalf of the UN. To achieve this aim, organs of UNESCO regularly collect data on educational practices in member countries. A fundamental requirement of data so collected and processed by the UN through the UNESCO Institute of Statistics is global comparability [2].

This paper is of the opinion that cross-population comparability is enhanced when an appropriate data model is used to store the underlying data used in the analytical process. The objectives of this paper are therefore threefold: (i) show how the current approach of storing data in disparate sources using Enterprise Resource Planning (ERP) applications introduces data heterogeneity which makes the stored data incomparable across multiple organizations; (ii) review how Big Data models have evolved from federated databases to data warehouses and a hybrid of both in response to the growth of data in the *V*-dimensions of Big Data when aggregated from multiple sources; (iii) show that the proposed data model enhances cross-population comparative analysis where multiple organizations are involved.

The concept of Big Data came up as a technology to cope with the storage of massive data sets that were being moved from diverse sources into data warehouses. The dimensions of Big Data were then known as volume, velocity, variety and lately veracity [4]. Volume refers to the size of data being created, Velocity is the speed at which data is created, captured, extracted, processed, and stored while variety connotes different data types and sources ranging from structured, semi-structured to unstructured data. Aggregating data from multiple organisations increases data volume to a level that it attains the status of Big Data. Velocity, variety and other dimensions of Big Data

usually come into play but when data is aggregated, volume is the Big Data dimension that must first be handled. Aggregating data from multiple sources has a consequence of data being lost in the process [5]. Data loss tampers with the veracity dimension of Big Data because veracity means the truthfulness, accuracy and integrity of data [4]. Veracity raises issues of quality, reliability, uncertainty, as well as incompleteness.

To put volume in a perspective that emphasizes its relevance to data aggregation and data comparability across multiple organizations, volume may be redefined as voluminosity, vacuum, and vitality–three additional *V*-dimensions of data as exposed by [4]. Voluminosity in volume states that there is already a very large set of data collected and even much more is available to be harvested. Between the volume collected so far and those yet to be collected there is a significant gap, making voluminosity a significant attribute of volume. In a nutshell, volume refers to the size of data being created from all sources in an organization including text, audio, video, social networks, research studies, medical data, space images, crime reports, weather forecasting and natural disaster [6]. The scale of data volume is now in terabytes, petabytes, and exabytes, a phenomenal challenge that is being addressed using a combination of big data technologies consisting of hardware and software [7]. The combination of commodity hardware and the Hadoop Distributed File System (HDFS) is an example of big data technology often deployed to address the issues of voluminosity.

The vacuum dimension of volume states that there is a strong requirement for empty spaces to store large volumes of data. Vacuum also refers to the creation of room to store, process and manage tremendous datasets from the existing datasets. This dimension of volume pops up the research question about how much storage space is available for incoming data rather than how much data has already been stored. The process of creating storage space for incoming data is equally as challenging as it is with managing vast sets of already stored data. The vacuum dimension of Big Data is concerned with creating space, by either augmenting storage devices or other techniques to compress the size of data [7]. The vitality of volume states that there is a massive amount of data actively served and unserved. Vitality emphasizes the survival of data in the storage environment and thus its reliability. In a large data bank, some data are actively used while some are not [7]. However, companies generate revenue from the actively used data only and the rest are stored for future uses. There is the risk that data stored for future use is abandoned or not properly maintained. As the risk of being abandoned gets higher, anything can happen to those datasets not currently in use. In other words, with less investment and attention to the unserved data, they are exposed to incidences of fire, earthquake, flood, war, and terrorist which are the prominent causes of data loss. Thus, vitality is a critical component of volume. The lack of vitality, in any case, is symptomatic of the absence of disaster management systems which decimates data reliability or leads to complete data loss. Apart from reliability, vitality also describes flexibility, dependability, and security. Vitality is an integral component of volume just as the volume is to Big-Data.

This work was also inspired by the concept of collocation in language processing. In corpus linguistics, collocation is used to describe a sequence of words that often occur together and can be extracted from a corpus. When applied to big data, collocation extraction involves finding interesting word combinations in large corpora [8, 9]. The

concept of collocation has also been applied to management of resources such as power, computing resources in a data center, placement of satellites and measuring instruments [10]. To cut costs, business entities often rent space for servers and other computing hardware in a data center. Typically, a data center operating this collocation arrangement provides the building, cooling, power, bandwidth and physical security while the customers provide servers and storage. In satellite technology, collocation is used to describe the placing of two or more geostationary communications satellites in an orbit nearby each other. In remote sensing technology, collocation is used to describe matching remote sensing measurements from two or more different instruments. In summary, collocation is the placement of several related entities in a single location or proximity. In a way similar to this generic definition of the word, collocation, the proposed Collocation Data model permits data from different sources to be cultivated (collocated) as footprints of users on a single sign-on application. The application also processes the data collocated without any further preprocessing step since the single sign-on application enforces a uniform data format on the data cultivated. The Collocation Data model views the locations from where the footprints of users used to cultivate big data come from as geographical scales [11].

The rest of this paper is organized as follows: "Related works" section reviews the development of data models from federated systems up to data lakes and warehouses and how the latter is applicable to big data sourced from multiple organisations. Data comparability is reviewed as a desirable quality of data when sourced from multiple sources or organisations that are related by regulatory bodies. In "Methodology" section, the proposed model is discussed as the Collocation Data Model. The model is a big data store, designed using the novel Entity-Case diagram, a variant of the Use-Case diagram. Entity Relationship diagrams are subsequently used to decompose the Entity-Case diagram to lower-level details describing data derived from multiple organizations. In "Results and discussion" section, it is shown how the Collocation Data Model seamlessly processes the performance data of students cultivated from the footprints of users from multiple schools. Use-Case diagrams are used to show the design of a single sign-on application that implements all the features required to cultivate the collocated data and analyze them. Finally "Conclusion" section concludes the paper and outlines current and future lines of research.

Related works

Over the years, models that describe enterprise datasets have been developed and can be referred to as enterprise data models. Experience with big datasets show that the enterprise data models do not effectively model big data. This led to big data models beginning with federated systems, then data lakes and data warehouses. Enterprise data models have traditionally been implemented using relational database management systems (RDBMS). This was easy to do then because enterprise data models are proprietary to the applications to which they serve as backends and hence the volume of data is usually within manageable size. Big data models, on the hand are implemented using a polyglot of database management systems as no one database management system exemplify completely all the *V*-dimensions of big data [12].

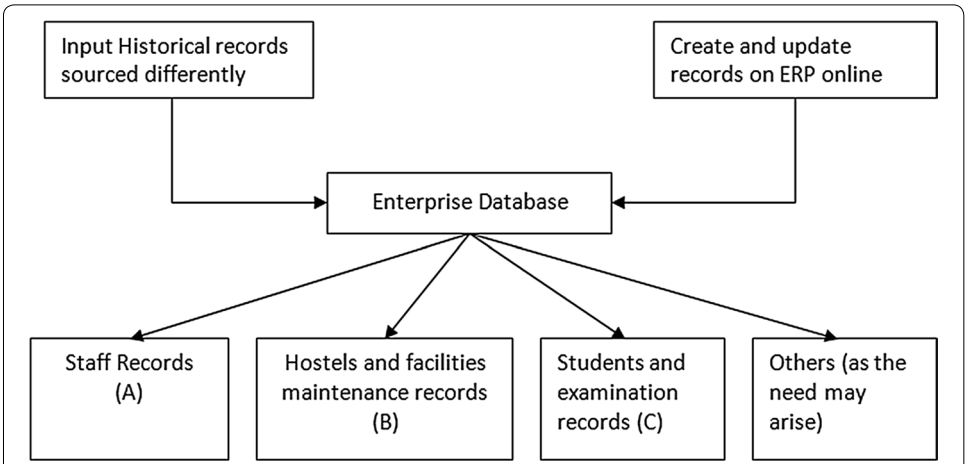

Fig. 1 ERP Architecture of Tertiary Education adopted from [13]

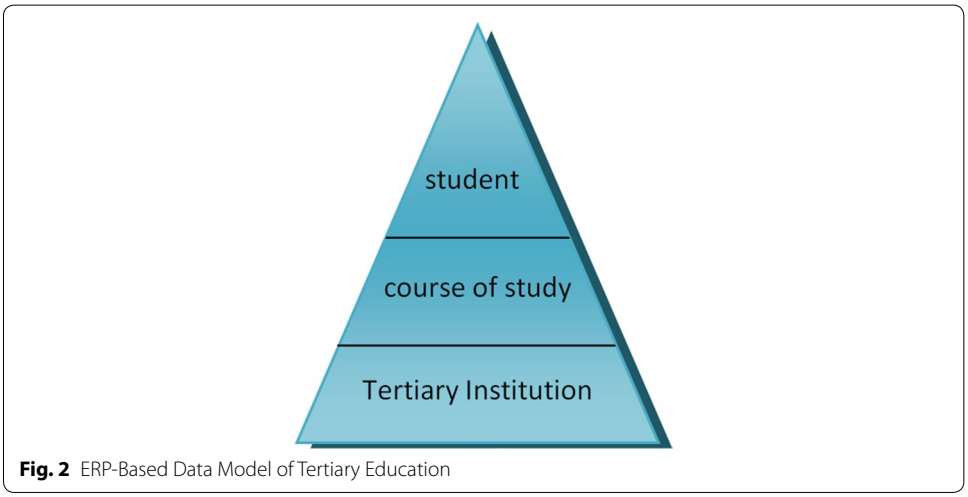

Fig. 2 ERP-Based Data Model of Tertiary Education

The enterprise data model

An enterprise data model used to model data about staff and students in a Nigerian University was implemented as a database backend to an Enteprise Resource Planning (ERP) system [13]. The data model was reported 12 as being applicable to other Universities supervised by the Nigerian Universities Commission (NUC). With the ERP, each University independently manages its administrative and academic programs. The architecture of the ERP as adapted from [13] is depicted in Fig. 1.

An Enterprise-Bound Data Model of educational institutions limits the scale of the model to the institution as an enterprise entity and hence not internationalized or even regionalized. The model scale is limited to harvesting data about students and the courses they are studying in an institution as depicted in Fig. 2.

The ERP data model gives an idea of the nature of data that can be sourced from an organisation. The model has however been grossly found wanting in modeling situations where multiple applications and databases are involved. This is because the ERP data model is implemented as an independent database, making it lack uniformity in terms

of data structure. The need for a data model that achieves uniformity in data collection and supports data analytics has been emphasized by relevant agencies in different countries and sectors of national life. In New Zealand, for example, the government has an Official Statistics System. To improve the quality of data on sexual orientation, a conceptual framework of sexual orientation was developed as part of this statistical system [14]. Veracity, the quality of data on this system, is measured by its timeliness, accuracy, reliability, and comparability.

Aggregating data from multiple organisations such as universities under the same supervision or enterprise databases such as the Official Statistics System obtainable in New Zealand produces data in Big Data dimensions. The aggregation process requires the knowledge of the several ERP data models associated with each organisation and a corresponding Big Data model that describes the aggregated data. In the subsequent sections of this review, existing big data models are discussed in terms of their historical development. The gaps observed from this review form the basis for a novel big data model that enhances data comparability across multiple organisations.

Big data models

In many respects, aggregating data from several sources promotes data heterogeneity because each institution will implement the ERP data model using its preferred database schemas. The resultant heterogeneity makes cross-institutional comparative analysis tedious, requiring that the aggregated data is cleansed and reformatted. In the particular scenario of Universities, data comparability issues that arise create problems for agencies that regulate University education, who must source data from the independent enterprise databases for analysis. The aggregation operation introduced when data from several ERP data models are aggregated and linked elongates the data value chain as depicted in Fig. 3.

The aggregation operation creates data in V-dimensions of big data and hence is a major focus of research into big data models. The first attempt to model big data was the federated model and the game plan was to create a model that permitted a snapshot view of several enterprise data models. McLeod and Heimbigner were among the first to define a federated big data model, as one that "defines the architecture and interconnectivity of databases that minimize central authority yet support partial sharing and coordination among database systems" [15]. Over the years, federated systems came to be known as a collection of cooperating (distributed) systems that are autonomous and possibly heterogeneous [16]. The federated system requires a lot of network bandwidth and stability to ensure online connectivity to the central

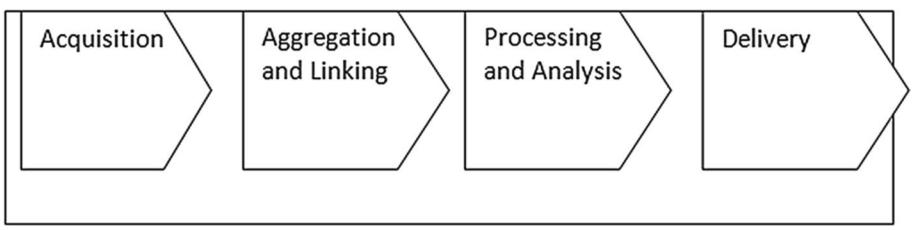

Fig. 3 Data Value Chain [2]

system since it cannot guarantee that the autonomous databases will be in the same location in all cases. The federated databases create mirrors of the original data being aggregated and do not create copies at the point of aggregation. This makes physical copies that can be historically referenced unavailable at the aggregation point.

A data warehouse is a big data model aimed at solving the problems associated with federated systems. A data warehouse was initially defined as a method of storing historical and integrated data for read-only view in decision support systems [17]. It was subsequently viewed as a repository of data sourced from multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision making. Building the data warehouse includes cleansing the data, reformatting the data to fit the unified schema (data integration) and mining data [18]. Data mining is the process of discovering interesting knowledge from the large amounts of data stored in the data warehouses. Data mining comes with the ability for Online Analytical Processing (OLAP) which enables information to be viewed from different perspectives in the form of dashboards to facilitate decision-making. Early implementations of the data warehouse model had storage limitations until Big Data stores were introduced. With the introduction of Big Data stores, the concept of OLAP was also extended to mean Big Data Analytics [19, 20].

The data warehouse model of data aggregation creates copies of data from different sources while the federated database model allows a central view of autonomous databases. Since data in the warehouse are copies, they do not necessarily need to be aggregated real-time but the resultant data volume is usually big in the dimensions of Big Data. Both federated databases and data warehouses allow organizations to dramatically improve their capability to manage massive datasets. For one reason or the other, the federated databases may fail to give a complete view of the autonomous databases.

The Hadoop MapReduce framework is typical of the data warehouse big data model. The Hadoop component of the framework provides a distributed storage system managed by the Hadoop Distributed File System (HDFS). HDFS supports the safe and rapid big data processing architecture of the MapReduce component of the Hadoop MapReduce framework. In this way, the framework supports the design of applications that generate big data analytics for business intelligence among others. The data integration capability of Hadoop enables it to integrate data from different sources, in different formats and characters into data lakes that support data input and output between multiple data sources and databases [21]. Bagui and Dhar [22] created a data warehouse using the Hadoop MapReduce framework on the Amazon AWS EMR. The data warehouse had 1.5 GB real life transactional dataset from 1.7 million web html documents mainly written in English and sourced from several websites. The dataset was downloaded from the FIMI website (https://fimi.uantwerpen.be/data/). As a pre-processing step, the documents were filtered to remove the html tags and common stop words. A stemming algorithm was then applied to convert each document into a distinct transaction containing a set of all distinct terms (items) that appeared in the document. Each data set is subsequently replicated and the replication combined with the dataset to make bigger datasets of 6 GB, 12 GB and 18 GB which were then used in the experiments. Voss et al. [5] referred to this pre-processing step as ETL (Extract Transform and Load).

Google Collaboratory (Colab) is an online framework where one can write, execute Deep Learning and Machine Learning codes [23]. Colab uses the Jupyter Notebook technology to connect to Google Drive on the cloud or to the hard drive of a local computer depending on where the dataset is located. For bigger datasets, Colab can download such datasets directly from external sources to google drive at very high speed once authentication to Google is confirmed. Loading data from external sources is a step towards a better architecture as it is a way of separating data storage from the notebook. Different versions of python and different runtime environments are available in the Colab framework and they offer many possibilities to connect with external data stores. In particular, the AWS S3 and versions of relational databases that scale horizontally to accommodate the volume, variety and velocity of the big data dimensions of data have been used for file storagein Colab. Like existing big data models, Colab does not grow the dataset but requires that the dataset be sourced externally, loaded into storage, pre-processed and then analysed for analytics. Carneiro et al. [24] used Colab as a tool for accelerating deep learning computer vision and other GPU-centric applications. The experimental dataset, made up of images were loaded into Colab and processed for predictive analytics.

In summary, the existing big data models have a pre-processing step in the aggregation process that leads to data losses. While this may not be avoidable in instances where the data sources are unrelated, the reverse is the case when multiple organisations related by operational procedures, management or supervision are involved. Beyond the advantage that these big data models help manage massive datasets, the data losses associated with them tamper with the veracity of the aggregated data. This paper proposes that this problem can be solved by developing a single sign-on application that cultivates data from multiple sources as footprints of users. This approach has the advantage of eliminating data losses, in the process of which the veracity of the aggregated data is improved and data comparability enhanced. The design of the proposed big data model will be such that the geographical locations (scales) where the data sources are domiciled are reflected.

Footprints-based big data models

Two big data models, namely the federated databases and the data warehouse models have been reviewed. It was exposed that both techniques aggregate data in large volumes in the dimensions of Big Data but cannot guaranteed veracity of the aggregated data due to data losses often involved. In addition to this problem, the design tools used in modeling existing big data models do not indicate the sources from which the data was aggregated from to form big data.

Capturing the data sources when modeling big data is important because the increase in data volume results from the many number of autonomous sources often localized in geographical categories of states, regions, country among others. These geographical categories or scales are hierarchically arranged. For example, states are components of a country and a group of states form a region in a country. A global view of an agricultural data model proposed in [11], described data collected in respect of the operations of farmers and their personal details. The data description starts with the level of the farm field, then the agro-ecological zone, regional, national and global levels. Within the context of each of these levels (geographical scales), actors are identified and the functions

they perform within the agricultural ecosystem are defined. The hierarchical nature of the geographical scales makes them factors or multiples by which aggregated data is scaled [11]. For ease of technical expression, the geographical scales will be referred to as data scaling factors or simply data scales henceforth.

In a way similar to the agricultural big data model, an application used by a multinational conglomerate can cultivate data about its operations on a global scale while the payroll of a local government area or city administration will have a narrower scale. Obviously, the higher the data scale, the more the volume dimension of data comes to play and vice versa. Data aggregated at the global data scale which is hierarchically higher than a country will be bigger than data aggregated at the country data scale. A larger data scale also means a larger user base with users requiring data in different formats, thereby prompting the variety-dimension of data. A large user base also means high traffic across the internet or whatever form of communication that is used for connecting users in the several geographical scales defined by the data model.

Within the context of aggregated data, each data scale represents an additional level of data aggregation. At the farm level of an agricultural data model, data about several fields are defined. At the ecological zone level, data about several farms are aggregated. At the regional level, data about the farms in the region are aggregated. At the national scale, data about the farms in a country are aggregated. On the global scale, data about the farms world-wide are aggregated. The farms and their embedded fields are the entities about which agricultural data is collected. While the farm is the enterprise unit, the field is an embedded entity and both can be referred to in generic terms as the entity. Data collected about the entity are the atomic data units that must be aggregated from the multiple sources within the context of the data scales defined. Using this approach, it is possible to collocate the entire data of multiple organizations within the data scales using a uniform data format for all the organizational data. It is also obvious that the uniform format eliminates data heterogeneity and therefore achieves cross-organizational data comparability seamlessly.

Data comparability

The review of related works so far has shown that aggregating data using footprints solves the problem of data loss which guarantees veracity. With data veracity achieved, further comparative analysis can be said to be reliable. In other words, with veracity achieved, data comparability is enhanced. Data comparability is a quality of data whose importance in cross-population analysis cannot be over-emphasized. Data comparability is a very useful concept in the ranking of universities and educational institutions in general [25]. Databases provided by Thomson Reuters and Elsevier have been used in global ranking schemes of universities [26–28]. Within the context of Big Data, these databases are expandable to take on more relevance. Considered as a justification for a Big Data-driven comparability model of international tertiary education are the challenges faced by the World Education Indicators (WEI) Programme [29]. Data on foreign students are produced jointly by Organisation for Economic Co-operation and Development (OECD) and UNESCO (Institute for Statistics) under this Programme and since 1997, OECD and 19 non-OECD countries have been asked to report their number of foreign students along with other information on their

education systems to OECD/UNESCO. Many countries that are now, student destinations such as China and South Korea are not yet included in the WEI project but will be included in the future as OECD and UNESCO have plans to expand coverage. Very problematic and worrisome is the fact that the data reported to OECD are not always consistent with other statistics published by countries themselves or reported in other sources. This, in no small way, has negatively affected the comparability of these statistical data.

In a way similar to the operations of UNESCO, Food and Agricultural Organisation (FAO) operates on behalf of the UN in the agricultural sector and World Health Organisation (WHO) operates in the health sector [30]. Of course, several other organs of the UN monitor and regulate other aspects in member countries. Several other Regulatory bodies and NGOs operating at global, regional and national levels exist to equally regulate and monitor the activities of subordinate agencies. In all these areas including agriculture, data comparability or consistency is a requirement for achieving cross-nationally comparable analytics [31].

The International Financial Reporting Standards (IFRS) has been globally accepted as a regulatory standard by firms that have adopted IFRS and the standard helps to provide useful information for global decision-makers, by enabling the comparison of the performance of two or more companies from different countries that adopt the standard [32]. The International Financial Reporting Standards (IFRS) is a set of guidelines that give a framework for reporting the performance of companies, to properly assess the financial health of organizations. IFRS aims to unify the different national financial statements to create the comparability of such statements [33]. Investors, regulators, academics, and researchers have all emphasized the importance of the comparability of financial data. An analysis of two or more firms with similar comparable parameters implies that the results of one study apply to another. Financial statement comparability increases the overall quantity and quality of information available to analysts about the firm and lowers the cost of acquiring information [34]. Several factors point towards the importance of financial data comparability. According to the Securities and Exchange Commission (SEC) [35], when investors judge the merits of investments and *comparability* of investments, efficient allocation of capital is facilitated and investor's confidence nurtured. The usefulness of *comparable* financial statements is underscored in the Financial Accounting Standards Board (FASB) accounting concepts statement which states that investing and lending decisions essentially involve evaluations of alternative opportunities, and they cannot be made rationally if *comparative* information is not available [33].

The importance of data quality especially as it regards data comparability is substantiated by global initiatives aimed at improving the global comparability of data [36, 37]. To give verve to these initiatives, many countries have made them part of the objectives of their diplomatic missions. The Global International Waters Assessment (GIWA) is one such global initiative. GIWA is a global recognition of the inextricable links between the freshwater and coastal marine environment. Its assessment framework collects environmental and socio-economic information to determine the impacts of a broad suite of influences on the world's aquatic environment [38]. GIWA provides a global perspective of the world's transboundary water by assessing

66 regions that encompass all major drainage basins and adjacent large marine ecosystems. GIWA was, therefore, a holistic, globally comparable assessment of all the world's transboundary waters based on comparable data from each region assessed.

Data comparability makes it easy to report analytical results using dashboards. The concept of a dashboard is extensively used by UN member countries to present reports across the countries. The Sustainable Development Goals (SDG) dashboards are used by the UN to monitor the implementation of the 2030 agenda of transforming the world in line with the 17 goals for sustainable development in member countries [39]. The use of these dashboards unifies data sourced from different countries and makes them analytically meaningful. In Chile, a single window for environmental reporting has been developed [40]. Environmental operators use one unique portal to comply with all reporting requirements. In this way, duplication of reports is avoided and the data being of a single format makes it easier to generate reports to other international standards. For example, comparable data on school enrollment and gender parity used by UNESCO to measure progress on Goal 4 of the SDG of the UN is based on the Education 2030 targets and compliance with the International Standard Classification of Education 2011 (ISCED 2011).

Comparing energy performance requirements for appliances across countries is difficult because of variations in product definitions, misaligned energy test procedures, and divergent efficiency metrics. This has made the global landscape of test procedures and energy efficiency metrics complex, thus requiring concerted efforts across countries aimed at improving the global comparability of appliance energy efficiency standards and labels [41]. Dashboards have been used extensively to solve this problem. Dashboard reformat data and, in this instance, it has been found very useful in reporting results of global comparative analysis of energy efficiency.

With the heightened focus on sustainability around the world, the need for a database of global toxic releases that can identify pollutant hot spots and potential needs for environmental regulation has led to the promulgation of the Pollutant Release and Transfer Registers (PRTRs) in more than 50 countries. The PRTR initiative is aimed at developing the capacity of countries to gather information in the form of chemical inventories or lists, supplemented by a means for disseminating the gathered information (information exchange) as required in many international agreements. But there are significant barriers to comparing data from individual PRTRs, including varying thresholds for reporting requirements and substantial differences among the lists of included toxics. To enhance the comparability of existing PRTR data sets, five recommendations have been made [42]

1. Rather than try to compare entire PRTR datasets, identify specific chemicals and/or sectors where comparisons can be made
2. Identify chemical classes to compare across countries with existing PRTRs
3. Identify normalizing factors to facilitate comparisons
4. Pursue a "relative comparison" approach
5. Create a global PRTR

Each of the recommendations is targeted at reducing the heterogeneity of the data sets from individual PRTRs. The fifth recommendation reduces data heterogeneity to the barest minimum as data in a global register will also have uniform formats in so many respects. The concept of a global register for individual PRTR tallies with the concept of data aggregation whereby data is harvested from multiple PRTRs. With a global PRTR, it is easy to increase the scale at which data is aggregated. The higher the scale, the less of the comparability issues experienced.

It has been demonstrated that aggregating data achieves data comparability across multiple organizations which in turn makes it easier to produce reports and when the data grows large enough, it makes it possible to perform data analytics on data in the three dimensions of sight, namely hindsight, insight and foresight as captured in Fig. 4.

The three dimensions of sight, amount to the value proposition of Big Data Analytics. Whereas hindsight is entirely descriptive and partly diagnostic, insight is both diagnostic and predictive; and foresight is partly predictive but largely prescriptive. Data aggregation is fundamental to cultivating data big enough for Big Data Analytics. Prediction models assume that data in quantities large enough to guarantee prediction accuracy. Prediction accuracy is required if the full value proposition of the three dimensions of sight is to be realised.

Methodology

Literature has exposed several ways by which data comparability based on aggregation can be achieved across multiple organizations. Data is often lost in the process thereby reducing the suitability of the aggregated data for comparative analysis. This constraint is eliminated when data about the operations of multiple organizations are cultivated and harvested in a single database backend, under the control of a common application. Data cultivation and harvesting is a concept first introduced in [43] to refer to growing data over some time, to make it become big enough for predictive analytics. The data may belong to different organisations but they are grown using a single sign-on application.

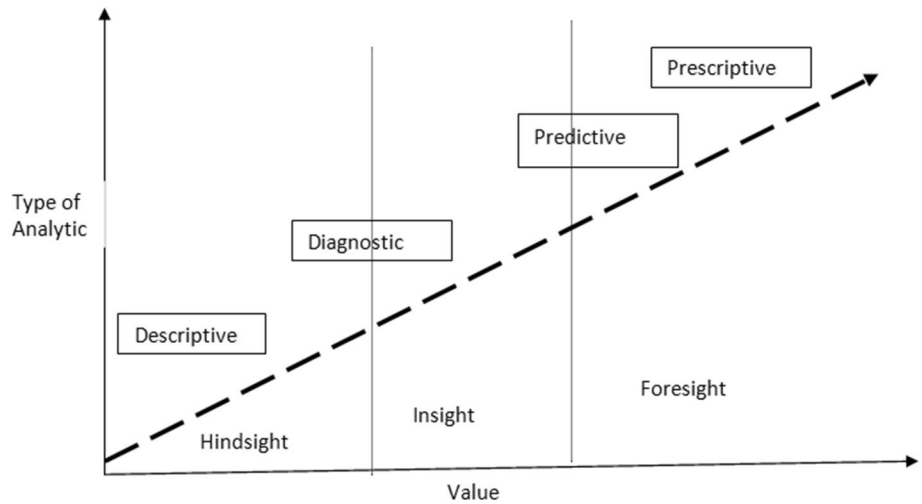

Fig. 4 The three dimensions of sight (adapted from [20])

In the subsequent subsection, use-cases have been used to demonstrate the features of the application. The Big Data backend of the application is designed using entity-cases (an adapted form of the use-case paradigm) combined with ER-diagrams. A proof of the novel concept of data collocation is also done.

Use-cases

Use cases are an abstraction of a system's behaviour and hence they are used for documenting user requirements. They may also be used for communication between various participants in a software project, i.e. system developers, its future users, and owners. Klimek and Szwed [44] noted that Use Cases are relatively easy-to-understand, even for people not familiar with information technology. Use cases enable the understanding of the system though they do not show lower-level implementation details. Use cases can also describe the system requirements and be used for formal verification of the requirements. This is often done at the initial phase of the system modelling so as to reduce production costs throughout the whole software development cycle.

The authors in [44] show that the functional requirements of a system under construction can be depicted by a Use Case Diagram, Use Cases and their relationships are shown in Fig. 5.

Big data backend

In the ERP data model as depicted in Fig. 1, data exists in isolated institutional databases otherwise referred to as institutional repositories. Data about a single organization is reported to be occurring in petabytes getting to exabytes [30]. These statistics among others imply that the data volume generated by each ERP application used by an organization multiplied by the number of organizations about whom data comparability is required can only be imagined within the context of big data. The big data backend that stores the data can be implemented as p-stores, c-stores and NoSQL [12, 45]. In these Big Data stores, data about multiple organisations can be stored without necessarily storing them according to organizations or sources from where they are generated. Should there even be a need to store according to organizations, such arrangements are transparent to the processing application and the user. This approach is an improvement on data warehousing given that it stores historical data as well as the current data being harvested from footprints.

The experimental data for this research is a dataset of multiple secondary schools. The availability of this dataset is stated in the declarations section of this manuscript. Existing

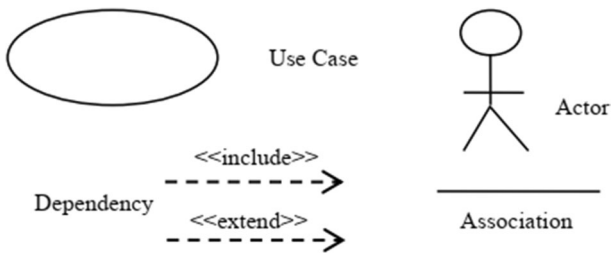

Fig. 5 Use Case Diagram symbols [44]

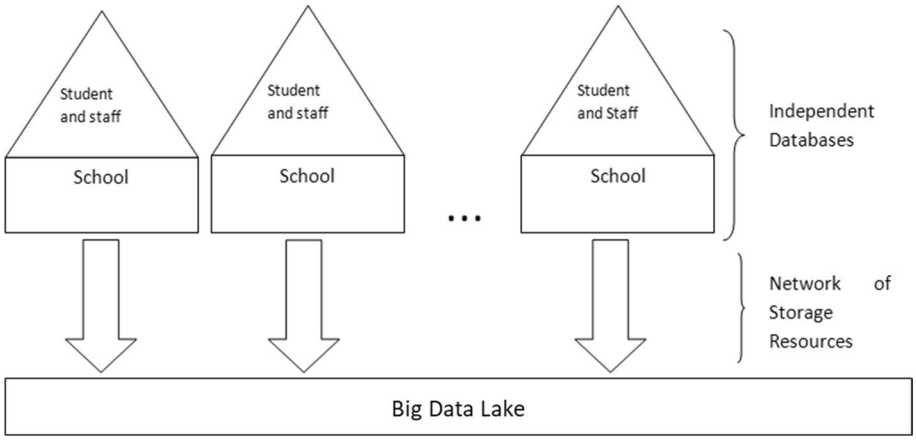

Fig. 6 Data Warehouse Model of Multiple Schools Data

ERP data models implement the dataset as consisting of autonomous databases. Figure 6 shows the independent databases of the ERP data model of the schools. Data about staff and students of each school are stored in the independent databases and mined into a warehouse implemented as a Data Lake. A Data Lake is a storage repository that holds a vast amount of raw data in its native (as-is) format until it is needed.

The data warehouse model of the dataset requires too much effort for data to be aggregated from individual schools taking note of updates. Above all, the process often leads to data losses. A single sign-on application with a uniform backend can be created to grow the dataset from the footprints of users drawn from all the schools. The resulting Big Data set is then modelled appropriately.

Big data modelling

One of the compelling qualities of Big Data is that it transcends geographical zones and locations in terms of sources from which it is generated. Big Data models must therefore reflect such geographical scales from which big data is sourced. The Collocation data model reflects the geographical scales of the sources from which it cultivates big data using data scales as earlier described. In this paper, an adapted form of the Use-Case diagram is used to model the data scales. The Use-Case diagram has been used to model functions performed by actors in applications [44]. In this work, the Use-Case diagram is adapted such that as an entity-case diagram in which case, the actors are the data scales representing the organisations or sources from which the big data has been generated and the Uses are referred to as entities that store data elements about organisations. The data scales are the basis of aggregation and so can be referred to as data aggregator. This novel design methodology for big data stores is depicted in Fig. 7.

The proposed model

The data warehouse big data model has over the years helped reduce data heterogeneity, a major cause of incomparability of data harvested across multiple organizations but has not solved the problem of data veracity arising from data losses. The

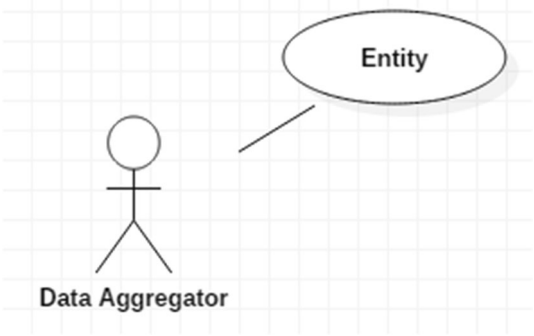

Fig. 7 Entity-Case Diagram

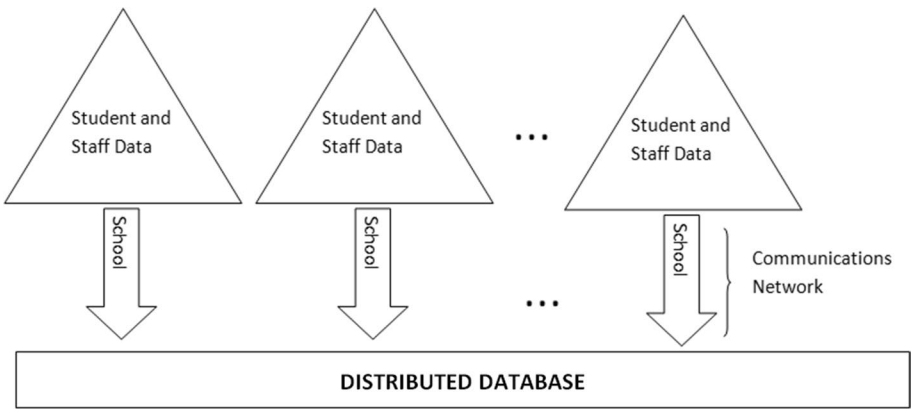

Fig. 8 Architecture of a Data Collocation Model of Multiple Schools Data

effort and energy dissipated in deploying data warehousing and mining techniques are reduced when the data about multiple organisations are collocated.

Data Collocation in the context of this paper describes a scenario, where multiple organizations related by common management or supervisory agency have their data stored in a single database and processed by a single sign-on application common to all of them. In a data warehouse, the data from heterogeneous sources are identified as coming from independent sources. In the collocation model, data is stored in distributed databases in which case, the location of any piece of data is transparent to both application and user. When the volume of data becomes big in the dimensions of Big Data, repeatedly increasing the capacity of existing hardware components to accommodate the increase in volume becomes unrealistic [7]. Rather adding commodity systems connected in a network has been found to be more useful. With a network of commodity systems in place, Big Data of any size can comfortably be organized as a distributed database and distributed among the computer resources.

The warehouse data model of the experimental data is depicted in Fig. 6 as independent databases whose copies are piped into a data lake. The Collocation Data model of the dataset views each school as a data entry point from where data about staff and students are persisted into a database remotely using some form of

communication network. The size of the resultant data is such that a distributed database is proposed for the model. The architecture of the collocation data model of the dataset is depicted in Fig. 8.

Collocation is a virtual concept where there are no obvious demarcations between the data belonging to each of the organizations participating in the collocation arrangement. Physically there are no guarantees that data of the same institutions will be located on the same storage resource because the collocation data model comes with an inherent capacity to distribute data across local and global networks. The collocation data model will generate data in a volume that can only be cultivated, stored and harvested using big data stores which by default take care of the concept of data aggregation [12]. In Fig. 5, the entity-case diagram was introduced as a data design model that can be used to design big data stores. When the agricultural big data was given as an example, the enterprise unit was the farm while farm fields were embedded entities. The data aggregators were the ecological zone, regional, national and global scales. In considering the multiple schools' scenario, the school is the enterprise unit and student, teacher, management among others are the embedded entities all referred to generically as the entity. The data aggregators can be defined as the local government area, state, and country in which the schools are sited. Putting the entity and data aggregators together, the design of a data store applicable to the architecture in Fig. 8 is depicted as an entity-case diagram in Fig. 9.

The entity-case is further decomposed using an ER-diagram in Fig. 10.

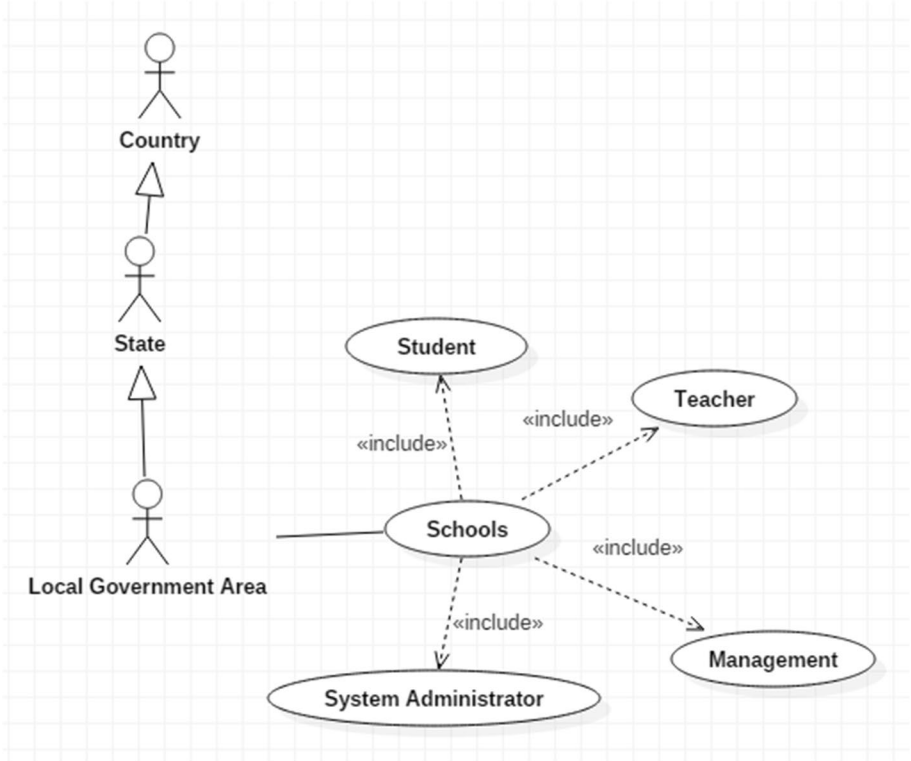

Fig. 9 Entity-Case diagram of the collocation data model of Schools

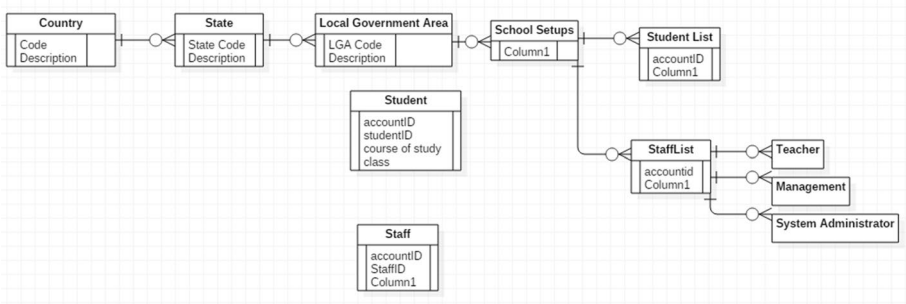

Fig. 10 E-R Diagram of the Collocation Data Model of a Group of Schools

Developing applications with the database backend implemented using the Collocation Data model has the advantage of maintaining uniform data formats that enhance data comparability. It is the Data collocation concept that has made Facebook the behemoth it is today. Using big data technologies, multiple user data across the globe are collocated such that reports and analytics are generated on Facebook to describe user likes and preferences [46]. Recommender Systems and the Amazon Recommender system, in particular, uses the same model. Shopping activities of users are harvested in the millions and stored as big data. The big data is then used to predict the user's choice of products on sale [47]. Amazon can compare the shopping preferences of one shopper with another based on what one had bought previously. The poser is that, if Facebook and Amazon achieved data comparability in real-time by collocating data about multiple users in one application, then such can also be attempted for multiple organizations.

The design of the proposed Collocation Data Model has been demonstrated as consisting of architecture, an entity-case diagram, and the Entity-Relationship diagram. The backend implementation of the Collocation Data model, on the other hand, is a distributed database whose implementation is based on the concept of database scalability. Partition schemes are used to achieve database scalability. The partition schemes are mainly vertical and horizontal partitioning schemes [48]. The partitions are arrived at using partition predicates. In either case, the invariant is the fact that the sum of the partitions distributed makes up the entire distributed database.

Proof of concept

Collocation creates a natural big data from multiple organizations, implying a uniform data format since the data are hosted in a single database, though distributed. The distribution of data in a collocation data model creates partitions based on the partition predicates used. In a case where the partition predicate is based on each organization, for example, a simple predicate is used to filter the data according to the individual organizations and distributed accordingly. In actual practice, the partition predicate could be such that partitions the data according to some other attributes that are common to the organizations such as products, departments, the accounting cycle, and many others. The application of the partition predicate ensures that the cardinality of relations is kept below the threshold value at all times.

Table 1 Relation (assessmentscores)

Tuple	School ID	Student ID	Assesment type	Subject	Session ID	Score
T_1	SCH1	6823	CA	MATHS	2017/18	45
T_2	SCH1	6823	EXAM	MATHS	2016/17	70
T_3	SCH2	2677	CA	MATHS	2017/18	73
T_4	SCH2	2677	EXAM	MATHS	2016/17	24
T_5	SCH3	9418	CA	MATHS	2017/18	90
T_6	SCH3	9418	EXAM	MATHS	2016/17	65

Table 2 Big data partitions of assessment scores

School ID	Session ID	Partition name	Tuples
SCH1	2017/18	SCH12017/18	$T_1, \ldots$
SCH1	2016/17	SCH12016/17	$T_2, \ldots$
SCH2	2017/18	SCH22017/18	$T_3, \ldots$
SCH2	2016/17	SCH22016/17	$T_4, \ldots$
SCH3	2017/18	SCH32017/18	$T_5, \ldots$
SCH3	2016/17	SCH32016/17	$T_6, \ldots$

The experimental data for this research consists of the assessment scores of the students in three schools collocated in a relation called *assessment scores*. As earlier mentioned, the availability of this dataset is stated in the declarations section of this manuscript. A big data strategy that implements the *assessment scores* relation as a big data store partitions the assessment scores data according to school and academic sessions. *Assessment scores* relation is abstracted in Table 1.

At the implementation level, there can be as many schools depending on the data aggregator and number of participant schools at each level of aggregation. The scores of the students, as well as the number of students, have the potential of growing infinitesimally as the academic sessions go by. The generated big data is partitioned horizontally and each partition distributed accordingly. Assuming SchoolID and SessionID are chosen for use as partition attributes, then the partition predicates are formed from the distinct values in the value set associated with the partition attributes. The distinct values are SCH1, SCH2, and SCH3 for SchoolID and 2016/17 and 2017/18 for SessionID. The distinct values produce the partition predicates, SchoolID = ' SCH1' and SessionID = ' 2016/17', SchoolID = ' SCH1' and SessionID = ' 2017/18′, SchoolID = ' SCH2' and SessionID = ' 2016/17', SchoolID = ' SCH2' and SessionID = ' 2017/18′, SchoolID = ' SCH3' and SessionID = ' 2016/17', SchoolID = ' SCH3' and SessionID = ' 2017/18'. Each of the conjunctive predicates is concatenated to produce a code used to describe each of the partitions as depicted in Table 2.

The SessionID changes each academic session implying that new partitions are created each session and the previous partitions archived. The expectation is that assessment scores of students in a school within a session will not exceed a volume threshold value that can become a challenge to the database management system. In this way, the volume is taken care of as proposed by the Collocation data model.

Proof

Theorem *Given P_1, P_2 ... P_n as the partitions of a data set R, then $R = \{P_1, P_2, ..., P_n\}$ where n = the number of distinct values in the value set associated with the partition key that generated P_1, P_2 ... P_n.*

Axiom The following axioms are applicable:

1. A partition key has a value set, V whose element cannot be null
2. The number of distinct values of V is n = number of partitions produced

Proof: Let ϭ be the partition predicate associated with a distinct value of V, and then Card (ϭ) is the cardinality of the tuples filtered by ϭ.

Given any value of n, there exists $ϭ_1, ϭ_2, ..., ϭ_n$, where.

$ϭ_1$ filters all tuples in P_1 from relation R,

$ϭ_2$ filters all tuples in P_2 from relation R, and.

$ϭ_n$ filters all tuples in P_n from relation R,

Since the elements of V cannot be null, then Card(V) = Card (R).

Since $ϭ_1, ϭ_2, ..., ϭ_n$ filter the tuples of R according to the distinct values of V, it follows that.

$\text{Card}(V) = \text{Card}(ϭ_1) + \text{Card}(ϭ_2) + ... + \text{Card}(ϭ_n) = \sum_i^n Card(\sigma_i)$

This implies that $\sum_i^n Card(\sigma_i) = \text{Card (R)}$ since n is the number of distinct values of V defined in R.

This shows that $R = \{P_1, P_2, ..., P_n\}$ since $ϭ_1, ϭ_2, ..., ϭ_n$ filter the tuples of R. **QED**.

Results and discussion

Longitudinal studies are used primarily in applied research aimed at discovering trends or patterns in groups of individuals over time. Many scientific disciplines, such as medicine, sociology, technology and education have used longitudinal studies extensively. Voss et al. [5] created a data repository of longitudinal healthcare databases that were aggregated from several observational datasets into a data warehouse referred to as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The source datasets were in varying formats, namely comma-separated text files and databases (MySQL, SQL Server, ORACLE, PostgreSQL). Constructing the OMOP Common Data Model data warehouse involves the ETL process made up of extracting, transforming, and loading tasks. The process uses an open-source tool called WhiteRabbit to analyze the structure and content of each observational dataset. The analysis involves listing all tables, fields, and distinct values in the fields of each observational dataset. In a subsequent stage, a CDM Builder program transforms the raw observational data based on the analysis done by WhiteRabbit into the CDM. Each CDM Builder has properties that are unique to the observational datasets transformed. Voss et al. [5] reported that the aggregated CDM improves data quality, increases efficiency, and facilitates cross-database comparisons, but at the cost of information loss across all six observational databases.

The Collocation data model, on the other hand does not require an ETL process. Assuming the Collocation data model was used in the longitudinal studies described above, the associated single sign-on application would have cultivated and harvested the

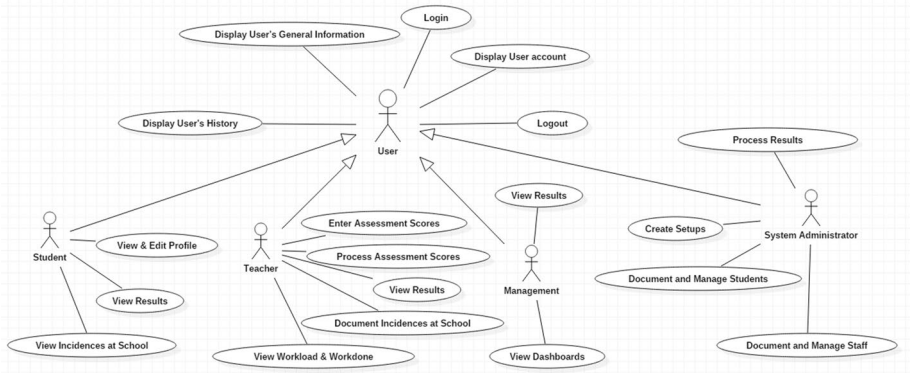

Fig. 11 Use-Case Diagram of the Functionalities of a School Management Software

Table 3 Statistical Features of the Dataset

Location (State)	Number of students	Distribution of Schools
Benue	842	X,Y
Kogi	204	Z
Total	1044	

observational dataset in Big Data dimensions as footprints of users from the multiple healthcare units from which the observational datasets were drawn. This would have been an improvement on the entire process. This fact is extrapolated to mean that the Collocation data model solves the problems associated with existing big data models, namely the federated databases, data lakes and data warehouses. In existing models, big data is aggregated from diverse sources and formatted to attain a uniform format that enables the data to be analysed comparatively. In the process of cleansing the data this way, data losses are experienced. Existing models do not also reflect the geographical scales from which the data is aggregated.

In the experiments performed, the assessment scores of students in multiple schools were collocated thereby eliminating the need to cleanse the data. Both staff and students in the schools used a single sign-on application to perform their administrative and academic tasks as depicted using the Use-Case diagrams in Fig. 11.

From the footprints of these users, data was cultivated. The resulting database is a Big Dataset whose design followed the data structures demonstrated in Tables 1, 2. Characteristic of the Collocation data model, the Big Dataset has an inherent uniform format. There are also no pre-processing steps involving data cleansing and hence there are no data loses. The design made the Big Dataset cross-organisationally comparable as well as facilitated the processing of the dataset using a central application. Using data aggregators, the geographical scales in which the data sources are located were reflected in the Collocation Data model.

Several schools were collocated in the experiment performed. The data about three of the schools are compared for the purpose of ranking them. The schools have been

named X, Y, and Z to keep their real names anonymous. Those selected for this research were marked as such in the database such that only data about them show in the results. Two of the selected schools are in the same geographical category (state) while the third school is in another state. Using state as the basis, the statistical features of the dataset are depicted in Table 3:

The sessional average scores of students in the nominated schools are compared to one another in mathematics and the English language and used to rate the schools' academic performance. The initial datasets used for this research include the assessment scores which added up gives the termly totals. The termly totals are summed up and its average for each session computed for each school. A session is an academic year made up of three terms (semesters). The summation and average of the three sessional scores of each school gives the 3-year average scores. A sessional score in a subject is the average of the termly scores. A good sessional score therefore shows consistent performance through all the terms. The 3-year averages computed this way are presented in Tables 4, 5 for each of the three classes in the first year of junior secondary school.

A 3-Year average of the sessional scores as computed for each organization and represented in Table 6. The data in Table 6 is presented graphically in Fig. 12.

It is observed from these results that the students' performance in Mathematics seems to be at par in schools, X and Y. It is however poor in school, Z. School Z is at its best in English Language, followed by School X with school Y trailing. To get a broader picture of the performance of the schools, a simple ranking algorithm is used to compute the average scores in the two subjects as shown in the fourth column of Table 6 and depicted graphically in Fig. 13. School X is ranked the best, followed by School Y, then School Z.

Table 4 Sessional average in Mathematics in Junior Secondary School 1

Session	School, X	School, Y	School, Z
2016/2017	61.25	55.665	53.5
2017/2018	48.5	57.335	48.5
2018/2019	61.665	58.085	44.165
3-year average	57.14	57.03	48.72

Table 5 Sessional average in the English Language in Junior Secondary School 1

Session	School, X	School, Y	School, Z
2016/2017	55.165	60	59.835
2017/2018	56.665	53.5	52
2018/2019	56	44.5	57.835
3-year average	55.94	52.67	56.56

Table 6: 3-Year average scores

Subject	Mathematics	English language	Overall ranking
School, X	57.14	55.94	56.54
School, Y	57.03	52.67	54.85
School, Z	48.72	56.56	52.64

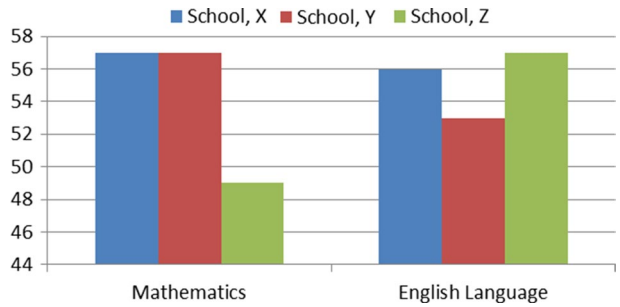

Fig. 12: 3-Year Average Scores

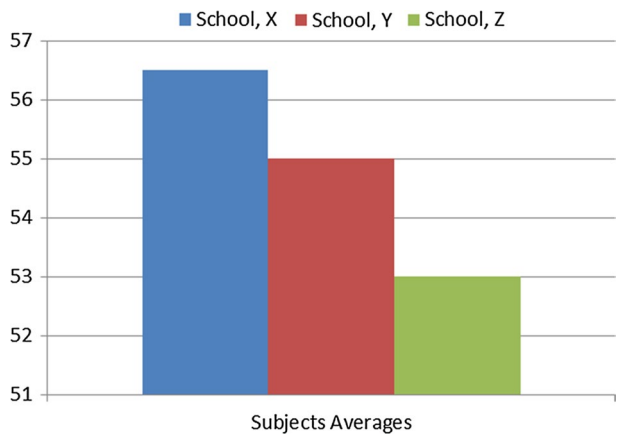

Fig. 13: 3-Year Ranking

The analysis done so far can as well be done for every other class and subject offered in the schools. A ranking based on the performance in more subjects other than just the two used in this instance is possible and the best students in a school can be compared across the multiple schools among other analytics that are possible.

The experimental dataset has been modeled using the Collocation data model. This facilitated the ranking analysis which was done without the need to aggregate data from autonomous data sources. In addition to this, the Collocation data model eliminated the need for data cleansing and reformatting, a requirement in the aggregation step in the data value chain with existing big data models namely the federated and data warehouse big data models. Data cleansing and reformatting are expensive and also lead to data loss. By eliminating the occurrence of data losses, data veracity is improved and the comparability of the dataset is enhanced.

Inherent in the Collocation model is a novel big data design methodology that combines two existing design methodologies, namely the Use-Case and Entity-Relationship diagrams to reflect the multiple organisations from which data is generated from as well as the geographical scales in which the multiple data sources are located. The Collocation Data model by so doing standardizes the design of big datasets used by data-intensive applications. Facebook, for example is one such data-intensive

application and it cultivates data about multiple individuals across the globe from their footprints as they interact with friends. Given this analogy, the Collocation Data model standardizes current practices whereby footprints of users are cultivated and harvested as big data.

Conclusions

The implementation of the Collocation Data model revealed that the aggregation step in the data value chain is eliminated when data from multiple organisations or sources is cultivated and harvested as footprints of users. This has been done using the experimental dataset to show that there was no need to aggregate data from the three sources independently, an operation that would have required that the data is cleansed and reformatted to enable cross-population analysis such as ranking. This implies a cost saving directly proportional to the cost of the aggregation step. The elimination of the aggregation step in this study also means non-tangible cost-savings associated with the inconveniences of data loss that was avoided. The primary objective of this study which was to demonstrate that when data is collocated, data comparability is enhanced thereby making analytics seamless was therefore met. The experimental data is still at its infancy as footprints of student's scores have been cultivated for only three academic sessions. As the academic sessions increase in number, data will grow and can be harvested in larger volumes. The increase in volume will make other predictive analytics aside ranking possible in the near future. The accuracy of the predictive analytics will also increase as the dataset size increases. This research has also shown that the ability to analyze massive amounts of real-time data and predict the future behavior of organizations that are related is critical and useful to stakeholders especially those placed in a position of performing management and oversight functions over multiple organizations.

Abbreviations

NGO: Non-Governmental Organization; UN: United Nations; UNESCO: United Nations Educational, Scientific and Cultural Organization; WEI: World Education Indicators; OECD: Organisation for Economic Co-operation and Development; FAO: Food and Agricultural Organisation; WHO: World Health Organisation; IFRS: International Financial Reporting Standards; CDM: Common data model; ETL: Extracting, transforming, and loading; SEC: Securities and exchange commission; FASB: Financial accounting standards board; GIWA: Global International Waters Assessment; SDG: Sustainable development goals; PRTR: Pollutant release and transfer registers; ERP: Enterprise resource planning; GIS: Geographic information systems; AWS EMR: Amazon Web Service Elastic Map Reduce; HDFS: Hadoop distributed file system; OMOP: Observational medical outcomes partnership.

Acknowledgments

We thank Mr. Chukwuma Benjamin Opara of the Federal University Lokoja for providing the professional proofreading service for this work.

Authors' contributions

PO designed the model diagrams, analyzed the datasets used in the proof of concept, and was a major contributor in writing the manuscript. EO validated the database frameworks, interpreted the diagrams and provided the initial logical arrangement of the sections including the references. Both authors read and approved the final manuscript.

Funding

Not Applicable.

Availability of data and materials

The data that support the findings of this study are available from [www.teekler.com] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [www.teekler.com].

Competing interests

The authors declare that they have no competing interests.

Author details
[1] Department of Mathematics and Computer Science, Benue State University, Makurdi, Nigeria. [2] Department of Computer Science, Federal University, Lokoja, Nigeria.

Received: 6 February 2020 Accepted: 1 November 2020
Published online: 10 November 2020

References

1. UNESCO Report. Thematic indicators to monitor the education 2030 agenda. Report of the Technical Advisory Group established by UNESCO to develop recommendations for education indicators. 2015
2. Petrova-Antonova D, Georgieva O, Ilieva S. Modeling of educational data following big data value chain. International Conference on Computer Systems and Technologies (CompSysTech'17). 2017.
3. United Nations, Full Text of UN Charter. https://www.un.org/en/sections/un-charter/un-charter-full-text/. Accessed 6 June, 2020
4. Storey VC, Song I. Big data technologies and management: What conceptual modeling can do? Data Knowl Eng. 2017;108:50–67.
5. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A, Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Informat Assoc. 2015;22:553–64. https://doi.org/10.1093/jamia/ocu023.
6. Khan MA, Uddin MF, Guptam N. Seven V's of big data: Understanding big data to extract value. Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1). 2014.
7. Patgiri R, Ahmed A. Big Data: The V's of the game changer paradigm. International Conference on High-Performance Computing and Communications. 2016.
8. Koyama Y, Masako Y, Yoshimura K, Kosho S. Large scale collocation data and their application to Japanese word processor technology. COLING-ACL, 1998.
9. Kessler R, Béchet N, and Berio G. Extraction of terminology in the field of construction, 2019 First International Conference on Digital Data Processing (DDP), London: United Kingdom, 2019, pp. 22-26. https://doi.org/10.1109/DDP.2019.00015.
10. Manjula L, Shylaja D. A collocation method for power aware resource management in cloud computing. Int J Scientific Res Develop. 2015;3(4):3–5.
11. Jones JW, Antle JM, Basso BO, Boote KJ, Conant RT, Foster I, Godfray HCJ, Herrero M, Howitt RE, Janssen S, Keating BA, Munoz-Carpena R, Porter CH, Rosenzweig C, Wheeler TR. Brief history of agricultural systems modeling. The agricultural model intercomparison and improvement project (AgMIP). 2016
12. Stonebraker M, Çetintemel U. One size fit all: An idea whose time has come and gone. ACM Digital Library. 2018. https://doi.org/10.1145/3226595.3226636.
13. Bamiro OA. Enhancing the quality of leadership and governance of Nigerian universities towards sustainable management and optimal performance. Abuja: Executive Development Programme for Council Members of Nigerian Universities; 2016.
14. Frank P, Alistair G, Veale JF, Binson D, Randell LS. Toward global comparability of sexual orientation data in official statistics: A conceptual framework of sexual orientation for health data collection in New Zealand's official statistics system. J Environ Publ Health. 2013. https://doi.org/10.1155/2013/473451.
15. McLeod D, Heimbigner D. A federated architecture for Information Management. ACM Transact Informat Syst. 1985;3(3):253–78.
16. Sheth AP, Larson JA. Federated database systems for managing distributed, heterogenous, and autonomous databases. ACM Comput Surv. 1990;22(3):183–236.
17. Inmon WH, Hackathorn R. Using the data warehouse. New York: Wiley; 1994.
18. Han J, Kamber M. Data mining: concepts and techniques. Burlington: Morgan Kaufmann Publishers; 2006. p. 1–42.
19. Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM SIGMOD Record. 1998. https://doi.org/10.1145/248603.248616.
20. Corcoran M. The five types of analytics. https://docplayer.net/. Accessed 6 Dec 2019
21. Lakshmi JVN, and Sheshasaayee A. Machine learning approaches on map reduce for Big Data analytics, 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), Noida, 2015, pp. 480–484. https://doi.org/10.1109/ICGCIoT.2015.7380512.
22. Bagui S, Dhar PC. Positive and negative association rule mining in Hadoop's MapReduce Environment. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0238-8.
23. Bisong E. Google Collaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley. https://doi.org/10.1007/978-1-4842-4470-8_7
24. Carneiro T, Medeiros Da Nóbrega RV, Nepomuceno T, Bian B, de Albuquerque VHC, Filho P. Performance analysis of google collaboratory as a tool for accelerating deep learning applications. IEEE Access. 2018;6:61677–85.
25. Emeka O, Sunday EA, Virginia E, Edward O. Web mining: Cybermetrics analysis of the nine (9) newly established federal universities in Nigeria in 2011. Int J Adv Res Comput Sci Softw Eng. 2015;5(8):904–13.
26. Frey JS. International credentialing of tertiary education, principles, questions, and concern. 2018. https://www.aic.lv/ace/ace_disk/Recognition/exp_text/frey.pdf. Accessed 20 Jan 2020
27. Rust V, Kim S. Globalization and Global University Rankings. In Zajda, J. (2015). (Ed.). Second International Handbook of Globalization, Education, and Policy Research (pp. 167–180). Dordrecht: Springer. Taylor-Sakyi K. (2015): Big Data: Understanding Big Data, ArXiv
28. Robertson SL, Olds K. World University Rankings: on the new arts of governing (quality). Centre for Globalisation. Bristol: Education and Societies, University of Bristol; 2012.

29. Kritz MM. Globalization and internationalization of tertiary education. International Symposium on International Migration and Development. Turin: United Nations Secretariat; 2006.
30. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Informat Sci Syst. 2014. https://doi.org/10.1186/2047-2501-2-3.
31. Marsh T, Fischer M. Accounting for agricultural products: US versus IFRS GAAP. J Busin Econ Res. 2013. https://doi.org/10.19030/jber.v11i2.7620.
32. Ichiro M. IFRS application and the comparability of financial statements. J Account Fin. 2017;17(5):7–8.
33. Franco GD, Kothari SP, Verdi RS. The benefits of financial statement comparability. J Account Res. 2011;49:895–931. https://doi.org/10.1111/j.1475-679X.2011.00415.x.
34. Tamas D, Agota K. Measurement of agricultural activities according to the International financial reporting standards. Proced Econ Finan. 2015;32:777–83.
35. SEC concept release: International accounting standards. 2000. https://sec.gov/rules/concept/34-42430.htm. Accessed 10 January 2020
36. Weyl DK, Pirie MF, Klinkenberg F. Improving global comparability: An analysis of appliance energy efficiency standards and labels. CLASP & The Policy Partners. 2014
37. Zschunke A. Global Comparability of analytical results. Analytica Conference '98, Symposium 1: Trends of accreditation and licensing of laboratories, Munich, 1998
38. Thorton JA. Evaluation of the Global International Water Assessment (GIWA) Project. United Nations Environment Programme (UNEP) Project Number GF/1100–99–01. 2006
39. Boza ME. SDG Dashboards: The role of information tools in the implementation of the 2030 Agenda. Report of research collaboration between UNDP-SIGOB and the UNDP Bangkok-Hub. 2017
40. Rodrigo P, Gariazzo Á, Shee S. Chile's National Environmental Accounts Plan. Chile: Publication of the Ministry of the Environment; 2016.
41. Klinckenberg F, Pirie M. F, McAndrew L. Improving global comparability of appliance energy efficiency standards and labels. A publication for the Collaborative Labeling and Appliance Standards Program (CLASP) in collaboration with The Policy Partners. 2014
42. Meyer T, Ostrum B, Peterka A, Reyes-Jones C, Stone M. Enhancing PRTR comparability to address global comparability needs. A publication of the United States Environmental Protection Agencies in collaboration with George Washington University, the Policy Partners. 2017
43. Ogbuju E, Taiwo K, Ejiofor V, Onyesolu M. Big data-driven e-government framework in Nigeria. Bayero J Pure Appl Sci. 2018;11:252–9.
44. Klimek R, Szwed P. Formal analysis of use case diagrams. Comput Sci. 2010;11:115–31. https://doi.org/10.7494/csci.2010.11.0.115.
45. Naren J, Elakia, Gayathri, Aarthi. Application of data mining in educational database for predicting behavioural patterns of the students. Int J Eng Technol. 2014; 5: 4469–72
46. Hines K. 6 Facebook reporting tools for in-depth analysis of fan pages. 2016. https://www.postplanner.com/6-facebook-reporting-tools-in-depth-analysis-fan-pages. Accessed 10 August 2016
47. Putnam J. A Complete Review of the Amazon Shopping Cart Experience. 2016. https://rejoiner.com/resources/amazon-shopping-cart-experience. Accessed 11 Sept 2016
48. Tailor U, Patel P. A survey on comparative analysis of horizontal scaling and vertical scaling of cloud computing resources. Int J Sci Adv Res Technol. 2016;2(6):2395–1052.

Publisher's Note