Submitted on: 16.09.2019

Title of the Satellite Meeting: Data intelligence in libraries: the actual and artificial perspectives
Date: 22-23 August 2019
Location: Deutsche Nationalbibliothek, Frankfurt

Performance Optimisation in Sharing Big Geoscience Data

Basuti Bolo
Department of Computer Science and Information System, Botswana International University of Science and Technology (Private bag 16), Palapye, Botswana.
E-mail address: (basutibolo@gmail.com)

Thabiso Maupong
Department of Computer Science and Information System, Botswana International University of Science and Technology (Private bag 16), Palapye, Botswana.
E-mail address: (maupongt@biust.ac.bw)

Lesego P. Peter
Botswana Geoscience Institute (Private bag 0014), Lobatse, Botswana.
E-mail address: (peterl@bgi.org.bw)

Abstract:

Online Geoscience data sharing plays a key role in enabling organizations to make informed decisions. Geoscience data covers both spatial and non-spatial datasets. To maximize the benefits, structured data workflows and frameworks governing the data has to be continually experimented. Reproducibility of enhanced frameworks driving defined data provenance that govern the datasets enables efficiency of improved data. However, achieving data sharing of multiple geoscience datasets with different provenance in practice is problematic – previous findings signify issues of approaches used to outline data visualization due to uncontrolled changes on the input data used, internet of things (IoT) devises and algorithms used in ascertaining data correctness. The resulting problems fall within the category of big data issues. In this paper we present a framework that addresses geoscience big data solutions through data visualization based on the structured datasets. The framework is such that it addresses performance optimization, data input correctness and redundancy. The resulting approach is not constrained by the type of data format or size. The results are evaluated through an extensive set of experiments that validate the approach and highlight the key benefits of the proposed framework. This included ways of reducing data redundancy and correctness based on visualization approach.

Keywords: *Geoscience big data, Data sharing, Framework, Data workflows, Visualization.*

Introduction

Geoscience data is regarded as big data and covers both spatial and non-spatial datasets. These datasets can be structured or unstructured. Sharing of geoscience big data online is a challenge. This is because the data can be from different sources produced and stored in different formats. It is well known that big data in general suffers from storage, analysing, processing and visualization (Barik et al 2017). To maximize the benefits of sharing big data online, structured data workflows and frameworks governing the data has to be continually experimented upon. Reproducibility of enhanced frameworks driving defined data that govern the datasets enables efficiency of improved data.

However, achieving data sharing of multiple geoscience datasets with different provenance in practice is problematic. Data to be shared needs to be clearly structured, correct and must be of high quality. The datasets need to be viewed and assessed before sharing in order to categorize them according to their formats for easy selection and use. This can be achieved by visualizing the data and development of visualization framework which allows for easy sharing of big data.

This study uses advanced data visualization techniques to manage the geoscience datasets and developed a Geoscience Data Visualization Framework (GeoDVisF) based on structured geoscience datasets. As already outlined the framework is developed to improve the sharing of geoscience big data for better data management, sharing and accessibility. In the following we outline the general and specifics objective of our study.

General and Specific Objectives

The main objective of this study is to develop Geoscience Data Visualization Framework for easy sharing of online geoscience big data.

In order to achieve our main objective, we carry out the following;

1. We identify components and techniques for development of the framework
2. We analyze the data and components, build the relationships between them and develop our framework
3. Finally, we validate the framework through experiments.

A brief on the literature

Data visualization has become common in the areas of research (Li et al 2016, Barik et al 2016, Barik et al 2017**).** This is also evident in the fact that over the last few years, a large number of novel information visualization techniques have been developed, allowing visualizations of multidimensional data sets (Daniel et al 2004). Previous findings signify issues of approaches used to outline data visualization due to uncontrolled changes on the input data, used internet of things (IoT) devises and algorithms used in ascertaining data correctness. These resulting problems fall within the category of big data issues.

Big data is a set of large data sets which are useful to develop knowledge by revealing patterns, trends, and associations. Big data refers to both the structured and unstructured high volume of data. Moreover, the quality of data is more important than its quantity (Wadhwani 2017). In general, since big data is structured and unstructured datasets with massive data volumes it cannot be easily captured, stored, manipulated, analysed, managed and presented by traditional technologies sets (Li et al. 2016). Volume, Velocity and Variety (3Vs) has been used to describe dig data, and more recently Veracity has been added to describe data integrity and quality (Laney 2001). Words such as Variability, Validity, Volatility, Visibility, Value and Visualization (5Vs) have been used several times in the literature to describe and give characteristics of big Data (Li et al. 2016).

Visualization is a way of making visible features of a given set of data or system from the graphical analysis of scientific data through the infographics (Vickers). It allows an overview of the data for the user to identify interesting subsets. Visualization of features has been described as the analysis process in which the data can be explored in order to build hypotheses. (Zhang et al. 2012, Li et al. 2016). The technology can also be defined as communication tools for hypotheses, results and other ideas in big data (Li et al. 2016). Visualization combines a range of skills and disciplines such as statistics, maps, graphic design, and computer graphics. The exploration of data visualization technology is based on three steps; overview, zoom and filter (Shneiderman 1996). The user needs to get an overview of the data and to identify interesting patterns. For the analysis process, the data need to be accessed for accuracy (Daniel 2002)

Visualization analysis helps in identifying patterns, build relationships and draw new hypotheses for further computational analysis. Analysing and visualizing big data is to view patterns and relationships of the datasets in order to understand, correlates and represent it in a visual context (Van 2009, Evangelidis 2014, Barik et al 2017). There are also several visualization tools that are especially designed for different datasets such as geospatial and non-spatial data (Barik et al 2017).

Materials and Methods

Generally big data research results are difficult to present (Li et al 2016). The data used are complex and represented in words, text or images. A better presentation of the results is by building the structures. In this research it was done by building a framework. The result presentation was the Geoscience Data Visualization Framework (GeoDVisF) based on the structured geoscience datasets. GeoDVisF is a framework based on advanced data visualization technologies and processing tools such as Geographical Information System (GIS), GIS software and geoscience data. The GeoDVisF framework was used to explain the linkage of geoscience data, processing, presentation, display, visualization output and data repository. The framework is capable of describing the relationship between components. A conceptual concepts represented graphically has been used to build the GeoDVisF. Conceptual framework is a map to show the components and their relationships. According to Vaughan, a conceptual framework is defined as; A graphically written statement that explains main components of researched and their relationships (Vaughan 2008).

A conceptual concept was constructed by incorporating components of the structures and build new framework, the GeoDvisF. This research has built a new framework based on the new technologies and knowledge. The components of the conceptual framework are; Geoscience data, Data processing and tools, Presentation and Display, Visualization output, Metadata and Data Knowledge Repository. In this work the framework was based on technical knowledge, data, related theories and research. Key words were identified, and drew the key components and build the relationship between them using flow charts and shape based diagram to show the linkages flow.

The validation of the developed GeoDvisF was based on the geoscience Big data, and the data used was left unchanged. Data incorporated from different datasets was used as original in order to allow further sharing when collating and producing different geological thematic map layers. The geoscience data was processed and transformed into useful geospatial information in a map format for easy visualization.

Results

This study developed a visualization geoscience data framework of the structured data. The GeoDVisF is a framework developed consists of six components namely; Component 1 (C1) Geoscience Data, Component 2 (C2) Data Processing, Component 3 (C3) Presentation and Display, Component 4 (C4) Visualization Output), Component 5 (C5) Metadata Retrieval, and component 6 (C6) Data and Knowledge Repository, as shown in Figure 1.

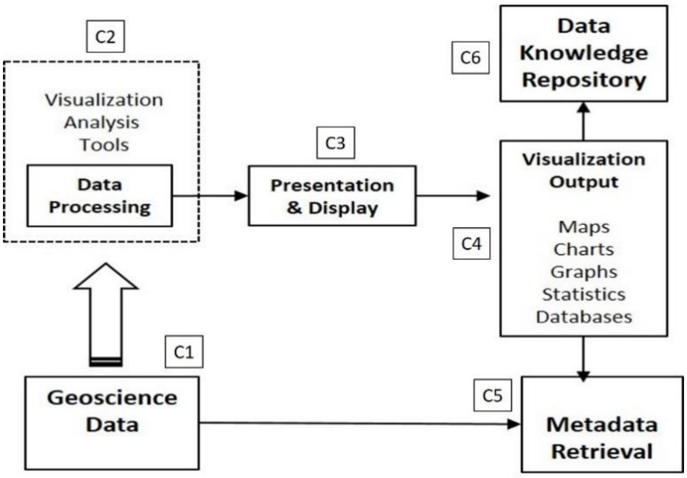

Figure 1: Geoscience Data Visualization Framework

Component 1 (C1) namely Geoscience Data is the main source of geoscience data. The data can be a raw data or processed data. These data have to be attached to a metadata, which is information about the data. There is a need to include a metadata in the framework being Component 5 (C5) as it is displayed in Figure 1. This is because, metadata guides the user with the information about the datasets. For visualization to be clear, the data had to be processed and structured according to a standard format in order to be presented easily. Processing the data also reduce datasets errors and improve the data quality. A set of dataset had to be categorized according to thematic groups and formats, these reduces the problem of data redundancy. In the data processing, visualization analysis tools had to be used to process the data.

Visualization tools such as Geographical information system (GIS) and the GIS software tools such as QuantumGIS (QGIS) and ArcGIS had been identified as the visualization tools used for processing structured geospatial datasets. This process has to be done in Component 2 (C2), namely Data Processing. After Data Processing stage, the results had to be presented and displayed for visualization, namely Component 3 (C3) Presentation and Display. The presentation and display of the results could be in different formats such as maps, charts, graphs, statistics and databases. The output is called Visualization output (Component 4 (C4). Maps and charts are useful to the users because it shows clearly the type of data and its coverage area. It also reduces time of selection, overlapping and wrong choice of data sets. Therefore, the visualization output could be stored online for the users and decision making, as presented in Component 6 (C6) of the framework. This improve data sharing of geoscience datasets and Visualization had been used as a tool for better sharing of these data.

Validation of the geoscience datasets was validated by processing the data and transforming it into useful information. The output was a map as it is shown in Figure 2. A geological map of Kolobeng of Botswana was produced based on the developed framework and displayed visually.

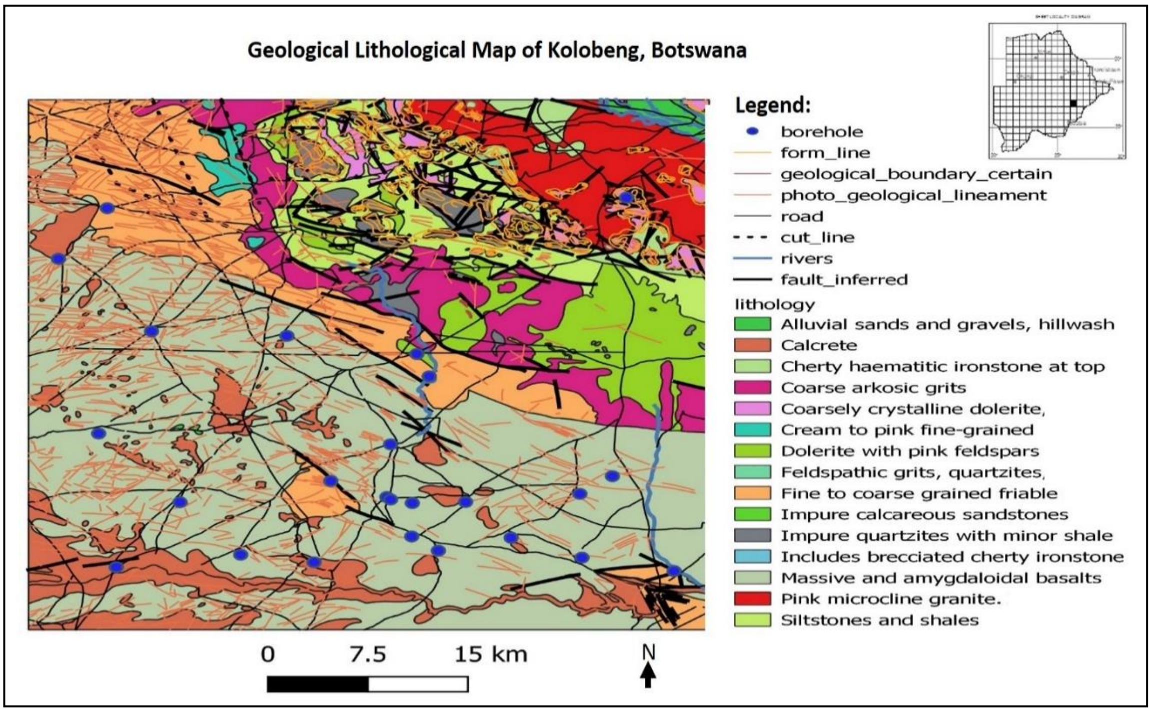
Geological Lithological Map of Kolobeng, Botswana
Legend:
borehole
form_line
geological_boundary_certain
photo_geological_lineament
road
cut_line
rivers
fault_inferred
lithology
Alluvial sands and gravels, hillwash
Calcrete
Cherty haematitic ironstone at top
Coarse arkosic grits
Coarsely crystalline dolerite,
Cream to pink fine-grained
Dolerite with pink feldspars
Feldspathic grits, quartzites,
Fine to coarse grained friable
Impure calcareous sandstones
Impure quartzites with minor shale
Includes brecciated cherty ironstone
Massive and amygdaloidal basalts
Pink microcline granite.
Siltstones and shales
0
7.5
15 km
N

Figure 2: Geological Lithological of Kolobeng Map.

The Map shows clearly what is presented on the Map. The distribution of the boreholes can be seen clearly across the entire map. The map shows that most of the area on the South West is covered by massive and amygdaloidal basalts. It shows easy visualization. When the map is shared online, the user can select the area of interest easier, fast and can leave the datasets that already had.

Discussion and Conclusion

Visualization could be one of the most powerful tool to capture and present the data in order to provide quality information. Visualizing the data had shown that it can provide a unique perspective that is able to reveal new patterns and trends making the information easier to be understood. The data had been clearly and whole presented completely presented visually based on the produced map of Geological area of Kolobeng area of Botswana.

Data visualization is about managing and communicating the data to provide valuable information that can be understood and shared by all. Visualization of data has proved that it can be used for easy management and sharing of structured data online because the data is digitally stored. The data that was processed and visualized had proved a way of viewing and assessing the data easier. In Visualization, patterns and relationships in the data can be observed and be used for decision making (Barik et al 2017). Visualizations also help built correlations from the big data analysis. It also closes the gap of machine data learning (Li et al 2016).

Recommendations

The concept defined by this paper has laid a foundation through which the aspects of Big data issues are addressed. Further work which we are working on involves outlining an algorithm which involves machine data learning, addressing data correctness based on the metadata. In order to ascertain that data visualisation does not distort the originality of the dataset. This will also take into consideration internet of things (IoT) issues, enlightening best ways towards addressing challenges tied to sharing geoscience big data.

References

1. A. K. Daniel, P. Christian, and S. Mike, "Visual Data Mining of Large Spatial Data Sets".2004.

2. A. K. Daniel and M. Ward. "Visual Data Mining Techniques, Book Chapter in: Intelligent Data Analysis, an Introduction: by D. Hand and M. Berthold", Springer Verlag, 2 edition, 2002.

A. K. Daniel. "Visual exploration of large databases", Communications of the ACM, 44(8):38-44, 2001.

3. A. K. Daniel A., et al. "Challenges in visual data analysis." Tenth International Conference on Information Visualisation (IV'06). IEEE, 2006.

4. B. Shneiderman. "The eye have it: A task by data type taxonomy for information visualizations. In Visual Languages", 1996.

5. Barik et al. 2017. "Investigation into the efficacy of geospatial big data visualization tools. International Conference on Computing", Communication and Automation (ICCCA2017).

D. Laney. "3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies". 2001.

6. K. R Barik, Dubey, Harishchandra, A. B Samaddar, R. D Gupta, and P. K Ray, "FogGIS: Fog Computing for Geospatial Big Data Analytics," 3rd IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, 2016.

7. K. Wadhwani, "Big Data Challenges and Solutions", Technical Report, DOI: 10.13140/RG.2.2.16548.88961. 2017.

8. K. Evangelidis, K. Ntouros, S. Makridis, and C. Papatheodorou, "Geospatial services in the cloud," Computers and Geosciences, Vol. 63, pp. 116–122, 2014.

9. N. M Van, H. J Scholten and R Van de Velde, "Geospatial technology and the role of location in science," Springer Netherlands; 2009.

10. P. Vickers, Member, IET Joe Faith, and Nick Rossiter. "Understanding Visualization: A Formal Approach using Category Theory and Semiotics", 2013

R. Vaughan, "Conceptual Framework". 2008.

11. S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein and T. Cheng. "Geospatial big data handling theory and methods: A review and research challenges," ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 115, pp.119-133, 2016.