Explainable GeoAI Real Time Data Model for Heterogeneous Datasets: Graph Database Approach

Abraham Tula *, Firaol Geleta **

* Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
** Ethiopian Construction Design and Supervision Works Corporation (ECDSWCo), Addis Ababa, Ethiopia

Abstract. Spatial activities are described and linked to the identified place or location. In the age of the Internet of Things (IoT), a vast collection of spatial datasets is emerging. The introduction of GeoAI into spatial data analytics is changing the scope and perspective of analytical capabilities in many ways. Since GeoAI is the merging application of spatial data science, artificial intelligence, and geospatial information science, and is the highest and most advanced application of geo-enrichment, intensive heterogeneous data sources have been used. Due to the extensive open data sources generated by mobile devices, sensor data streams from static or moving sensors, satellites, the availability and sharing of data via standard APIs have now increased immensely. In this article, a graph database approach is intensively emphasized to develop an object oriented based explainable GeoAI data model in its various applications. In addition to the available data sources, large amounts of data are currently being generated by various institutions. The issue of sharing and reusing data between institutions is receiving more and more attention for various reasons. Linking datasets between different platforms creates ambiguities for both machine and human. In this article, the research mainly analysis the problems in real-time generated data management of heterogeneous spatial data in the application of GeoAI and provided recommendations.

Keywords. GeoAI, graph database, moving features, heterogeneous datasets, real time data model

Published in "Proceedings of the 18th International Conference on Location Based Services (LBS 2023)", edited by Haosheng Huang, Nico Van de Weghe and Georg Gartner, LBS 2023, 20-22 November 2023 Ghent, Belgium.

This contribution underwent single-blind peer review based on the paper. https://doi.org/10.34726/5742 |

1. Literature Review

In the era of geospatial "Big Data" [1], up to 80% of big data is "spatial" with locational components attached [2]. With the advanced development in remote sensors, GPS-enabled applications and the popularity of mobile devices, as well as increasingly affordable data storage and computational technologies, geospatial big data are produced from a wide range of disciplines from commercial business to scientific research and engineering at a very fine spatial, temporal and spectral resolutions[3][4][5]. Such geotagged data in large volume, high velocity, and abundant variety that exceed the capacity of current common spatial computing platforms are defined as spatial big data [5]. The recent breakthrough in machine learning, or more generally AI and more specifically deep learning, enables a new research paradigm of data-driven science to analyses, mine, and visualize massive spatial big data (SBD) that are difficult to handle using traditional spatial analysis methods [5].

An increased availability of geospatial big data and real time generated dataset, the advancement of artificial intelligence (AI) and the availability of high computing power have created a momentum for the digital exploitation of geospatial big data real time analysis and in turn combines discipline in spatial science, artificial intelligence methods in machine learning (e.g. deep learning), data mining to extract knowledge from spatial big data. This all together emerge the new scientific discipline called Geospatial Artificial Intelligence (GeoAI) [6] [7].

The emergence of GeoAI has a significant role by developing conventional technologies AI and innovating new technologies to use the high potential geospatial big data to address ever developing new complex challenges faced in our day to day activities [6]. As the main component of the GeoAI infrastructure, an appropriated technologies can be applied to improve certain steps in the heterogeneous data management and intern maximizing the return on geospatial big datasets [6].Even though a drastic improvement have been achieved in geospatial big data analysis and geospatial artificial intelligence both in practical implementation and hypothesis analysis, the lack of high trained and labelled quality data and appropriate data management is remain the main challenge for GeoAI [17] [8].

The key challenge in GeoAI applications is scaling the integrated system to complex data scenarios. Even though GeoAI in its first inception certainly apply the AI technologies, especially those based on Deep Learning, i.e. usually require huge collections of layered and training data, GeoAI is an ideal opportunities in spatial based research challenge solutions [6]. Based on the fact that an explainable model could be the crucial variable in a predictive model, data structure is an essential factor in developing an explainable GeoAI model [6] [9] [10].

In line with the increasing use of computer analysis in artificial intelligence, more and more humans and integrated machines will work together to improve our understanding of spatial Big Data [6]. Since the biggest challenge in this emerging field of science is reliability and robustness, the degree of difficulty in model interpretability must be emphasized [6] [9] [10] [17].

To process the ever-increasing large heterogeneous open-source geospatial datasets, an advanced digital infrastructure that enables highly scalable data processing is required to provide a standardized, time-appropriate solution technology

for real time heterogeneous, computationally intensive geospatial data problems [10] [11].

Figure 1. Three main components of GeoAI. (Source [3])

2. Objective

- Study on real time moving objects spatio temporal data management
- How efficient graph database for explainable GeoAI application
- Verify and explain the integration capabilities of graph database applications of GeoAI

3. Research Question

- How is data about moving objects represented in real time?
- Is explainable GeoAI better in graph database technology than in other database technologies?
- How easily is a non-spatial dataset discoverable?
- Can researchers determine and analyze the appropriate database technology for explainable GeoAI?
- How can data interoperability and automated routing be verified in a graph data model?

4. Challenges in explainable GeoAI application

In the era of geospatial big data and digital worlds, datasets from multisource devices for machine-to-machine communication are expected. How efficient in terms of computing capability, prediction model, and analysis is the key challenge when it comes to explainableGeoAI.The heterogeneity of the dataset from sensors and devices has to be edge computing capability. In this case, integrations and information exchange in the data exchange layer are traceable and identifiable [11] [13] [16].

The dataset that is currently collected and analyzed is mainly compiled in the framework of a relational database. However, the dataset that requires specifically the application of GeoAI is large-scale spatiotemporal data. A fixed data structure as a relational database for these types of datasets is challenging. A large, heterogeneous, and scalable data source is not efficient in data manipulation and analysis [12].The data management and their structure in the application of GeoAI for all kind of database illustrated that the difficulty and challenges can be easily identified. In the subsequent part of this article, a detailed practical example identified.

Efficiency, memory usage, and energy consumption could also be an extension of this article in a separate research project task. However, the challenge in explainable GeoAI is mainly the heterogeneity of the data source, data integration, and model accuracy. In the following unit, the comparisons among various types of databases are described based on scalability and performance [6] [7].

5. Database comparisons

Different types of database technologies are developed for specific and general purposes. The main database types used in research data management and analysis are relational/SQL databases, time series databases, and NoSQL databases. Even though each database is designed for its subsequent operation and goal, whether they are open or commercial tools, users and various industry and research projects have preferred their best interest and suitability for the purposes desired. The selection of a suitable database type in terms of analysis capability, computation time, scalability and integrability is distinguished in Table 1. Managing heterogeneous large spatio-temporal real time data with the appropriate application of GeoAI the outer most achievement and goal [7] [14] [15]. To give an overall view, the following table summarizes the most important comparison criteria.

Comparison Criteria	**Relational /SQL Database**	**NoSQL Graph Database**
Scalability	rigid	Flexible
Performance (Transaction)	good	good
Performance (Deep analytics)	poor	excellent

Table 1. Database type and comparison criteria (source [14])

As big geospatial datasets are generated from heterogeneous data sources such as mobile devices, social networks, and the Internet of Things (IoT), the memory is flooded with data at every moment of collections. As the cornerstone of this study, a real-time unstructured Big Data database and graph technology are explored to improve the applicability of GeoAI. NoSQL unstructured database is more weight than relational database in its characteristics such as flexibility, dynamism, agility, and explicitly integrate with hetrogenopues data source without significantly changing the whole designed setup and data integrations[13] [14].

6. Methodology

The application of explainable GeoAI was identified with respect to the geospatial research and development activities of the Leibniz Center for Agricultural Landscape Research (ZALF) and the Ethiopian Construction Design & Supervision Works Corporation (ECDSWCo). Both institutions rely mainly on spatial Big Data sources. In particular, the real-time geospatial data requirement plays a key role in crop yield prediction, soil moisture, flowering stage detection, flood forecasting, and real-time data monitoring. In this paper, we focus on case studies of soil moisture and flood prediction and monitoring. One of the goals of this research project is to develop a graph database to manage real-time spatio-temporal data on moving objects for selected projects.

6.1. Distributed Spatiotemporal Graph Database

The use of big data is the cornerstone of GeoAI. The design of the model and the implementation of the framework are the focus of this paper. This section makes the largest contribution to the development of handling heterogeneous datasets for explainable GeoAI [1] [15] [16].

In the GeoAI data platform, the data sets are mainly unstructured and come from different sources. The ideal database platform that can handle unstructured and heterogeneous data sources is NoSQL graph database. This type of database is not only characterized by its flexibility and scalability, but also allows linking non-spatial and spatial dataset features [1] [15] [16].

Currently, unstructured and semi-structured data are becoming the mainstream of advanced spatio-temporal data management. Several institutions use observations and measurements from various sensors, such as geospatial data from location-based services (LBS) and social networks, which can be stored in semi-structured and unstructured databases. NoSQL databases provide a highly accessible and scalable way to efficiently manage semi-structured or unstructured geospatial data. Therefore, the concept and framework focused on establishing a distributed spatio-temporal graph database [1] [13] [16] [17].

Figure 2. Distributed Graph Database and GeoAI digital Infrastructure

6.2. Data preparation

Real-time heterogeneous data management is implemented to justify the hypothesis based on the designed system selection matrix index. Therefore, the study mainly focused on soil moisture monitoring and flood forecasting. Meanwhile, datasets of any subject can be integrated and used for the demonstrable ability of handling heterogeneous data types and real-time spatio-temporal data management. Data sets from sensors and mobile devices required for soil moisture monitoring are modeled in a graph database structure. Devices such as radiometers, spectrometers, spectroradiometers, and soil moisture sensors are the main source of sensor data sets [15].

The data from moving objects, and UAVs produced large sets of information in crop yield prediction and blooming stage prediction. Variables of time-location paired knowledge stored and ready to use from graph database. These datasets

require highly computing devices for processing. The volume and velocity of these big datasets are considered the scalable and integrable features of graph databases [3] [6].

Data from moving objects and UAVs provided large amounts of information for crop yield and flowering stage prediction. The variables of time-location paired knowledge data are stored in a graph database and can be used immediately. These datasets require high computational power for processing. The volume and speed of these large datasets are the characteristics of graph databases such as scalability and integrability, which can facilitate the storage and processing of large real-time datasets [3] [6].

6.3. Explainable GeoAI Analysis – real time spatio temporal data

Geospatial features of collected datasets are processed to instantiate the integration between non-spatial datasets for their maximum traceability and queriability [11]. The graph database design for explainable GeoAI is sympathetic for prediction and simulation model task. An explainable model can provide the capability to reviled the key variables and minimize ambiguity in the outcome. Connected features improve traceability and are interlinked with other features. From the geospatial data point of view, some variables could not be interlinked in the relational database. The major drawback of a relational database in explainable GeoAI application is that non-spatial data could not solely be retrieved and ready for analysis [3] [6] [10]. As depicted in the previous chapter, the usage of big data is the pillar of GeoAI .Large-size real-time dataset data cleaning, retrieving, and processing are highly efficient in graph database design.

To achieve and present AI explainability, explainable data in graph databases improve the knowledge of the data to be trained for a selected model. In addition, explainable data support the model with reasoning. We can generalize and conclude that the graph database for automated real-time ETL (extract, transform, and load) data pipeline can support data cleaning, quality detection, and data processing tasks. As shown in Figure 2, the main task of this tool is also data integration and quality control. The basic ETL pipeline is developed using the Python programming language to automate data collection from heterogeneous sources. A case study of flood prediction and soil moisture content determination of large geospatial data is used to develop the real-time ETL pipeline [19].

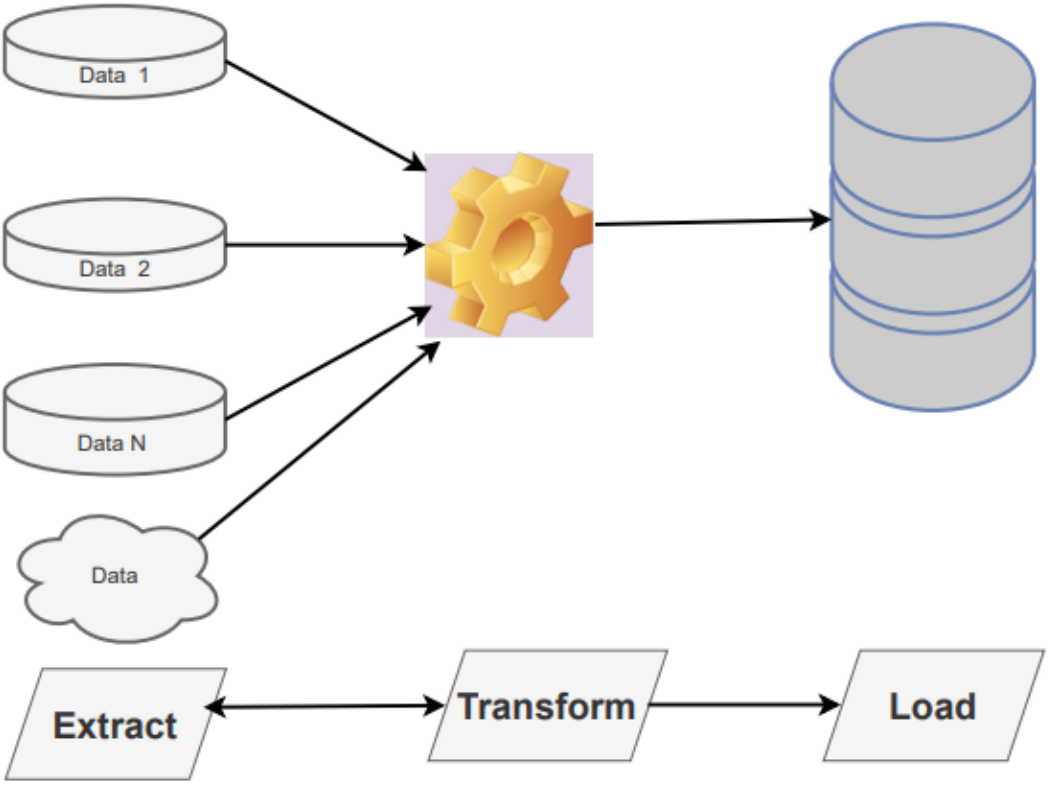

Figure 3. Real time ETL pipeline basic process

7. Conclusion

Heterogeneous data source object can be linked with each other. Knowledge derived from the large collection of the dataset discovered in support of one another. As explainable GeoAI consumes large dataset in a several layered feature, can support and provide well understandable answer to complex search queries. In the production of live stream data analysis output with an efficient performance, a distributed spatio temporal graph database set up a prominent digital real time spatio temporal infrastructure for the application of Explainable GeoAI. In the meantime, a standard data access API is maintained through the developed infrastructure and strengthen institutional inter cooperation. At the global level, where the data source in some application generated and collected worldwide, prepare a basis for the standard OGC moving features [18].

Note:

Since this project is an ongoing research project, the design of the graph database and the analysis of the real data are still in progress. The project encompasses a somewhat broader concept. The full version of the project will be available by the end of December 2023.

For your comment and feedback, Abraham.tula@zalf.de

References

[1] Rob Kitchin (2014), Big Data, new epistemologies and paradigm shifts, Big Data & Society April–June 2014: 1–12 DOI: 10.1177/2053951714528481

[2] Agnieszka Leszczynski and Jeremy Crampton (2016), Introduction: Spatial Big Data and everyday life Big Data & Society July–December 2016: 1–6 DOI: 10.1177/2053951716661366

[3] Wenwen Li (2020), Journal of Spatial Information Science, Number 20 (2020), pp. 71–77, doi:10.5311/JOSIS.2020.20.658

[4] Wenwen Li, Michael Batty & Michael F. Goodchild (2020) Real-time GIS for smart cities, International Journal of Geographical Information Science, 34:2, 311-324, DOI: 10.1080/13658816.2019.1673397

[5] Pengyuan Liu, Filip Biljecki (2022) A review of spatially-explicit GeoAI applications in Urban Geography, International Journal of Applied Earth Observation and Geoinformation, Volume 112, 2022, 102936, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2022.102936

[6] Alastal, A.I. and Shaqfa, A.H. (2022) GeoAI Technologies and Their Application Areas in Urban Planning and Development: Concepts, Opportunities and Challenges in Smart City (Kuwait, Study Case). Journal of Data Analysis and Information Processing, 10, 110-126. https://doi.org/10.4236/jdaip.2022.102007

[7] Ontotext (2023), what is NoSQL Graph Database? , retrieved from https://www.ontotext.com/knowledgehub/fundamentals/nosql-graph-database/ , last retrieved date, 28/05/2023.

[8] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu & Budhendra Bhaduri (2020) GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, International Journal of Geographical Information Science, 34:4, 625-636, DOI: 10.1080/13658816.2019.1684500

[9] VoPham et al. Environmental Health (2018) 17:40 https://doi.org/10.1186/s12940-018-0386-x

[10] Wanyan T.Y., et al. (2021): Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. Data Intelligence 3(3), 329-339 (2021). doi: 10.1162/dint_a_00097

[11] Janowicz, K., Hitzler, P. ., Li, W. ., Rehberger, D. ., Schildhauer, M. ., Zhu, R. ., Shimizu, C. ., Fisher, C. ., Cai, L. ., Mai, G., Zalewski, J., Zhou, L., Stephen, S. ., Gonzalez, S. ., Mecum, B. ., Carr, A. ., Schroeder, A. ., Smith, D. ., Wright, D. ., Wang, S. ., Tian, Y. ., Liu, Z. ., Shi, M. ., D'Onofrio, A. ., Gu, Z. ., & Currier, K. . (2022). Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence. *AI Magazine*, *43*(1), 30-39. https://doi.org/10.1002/aaai.12043

[12] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu & Budhendra Bhaduri (2019): GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2019.1684500

[13] Gong et al. (2015): Real-time GIS data model and sensor web service platform for environmental data management. International Journal of Health Geographics 2015 14:2.

[14] Chad Vicknair et.al (2010) ACM SE '10: Proceedings of the 48th Annual Southeast Regional ConferenceApril 2010Article No.: 42Pages 1–6 https://doi.org/10.1145/1900008.1900067

[15] Omar H. et.al (2018): Journal of Engineering and Applied Sciences 13(17):7323-7328 DOI: 10.3923/jeasci.2018.7323.7328

[16] Li, W.; Hsu, C.-Y. (2022) GeoAI for Large-Scale Image Analysis and Machine Vision: Recent Progress of Artificial Intelligence in Geography. ISPRS Int. J. Geo-Inf. 2022, 11, 385. https://doi.org/10.3390/ ijgi11070385

[17] The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B4-2022 XXIV ISPRS Congress (2022 edition), 6–11 June 2022, Nice, France

[18] OGC Standards, OGC moving features, retrieved from https://www.ogc.org/standard/movingfeatures/, last retrieved date, 14/05/2023.

[19] Andreas Kretz (2019) The Data Engineering Cookbook, V1.1

Abbreviations

No.	Acronym	Description
1	**GeoAI**	geospatial artificial intelligence
2	**ZALF**	Leibniz Centre for Agricultural Landscape Research
3	**ECDSWCo**	Ethiopian Construction Design and Supervision Works Corporation
4	**SBD**	spatial big data
5	**LBS**	Location Based Service
6	**GPS**	Global Positioning System
7	**AI**	Artificial Intelligence
8	**NOSQL**	not only SQL
9	**UAV**	An unmanned aerial vehicle
10	**IoT**	Internet of Things
11	**OGC**	Open Geospatial Consortium
12	**ETL**	Extract, Transform and Load