A Systematic Approach For Data Cleansing Process of Geospatial Data to Perform Application Specific Data Analytics

M S Satyanarayana[1], Dr. Nageswara Guptha M[2], Dr. Vasanthi Kumari P[3]
{satya.satya2011@gmail.com[1], mnguptha@yahoo.com[2], vasanthipvk@gmail.com[3]}

Research Scholar, SVCE, Bengaluru[1], Associate Professor Senior Grade 2, VIT Bhopal University[2],Sehore, Associate Professor, Dayanand Sagar University, Bengaluru[3]

Abstract. Data Analytics is the key word of today's era. Huge data is getting generated day by day from various resources starting from social networking sites to sensors then machines. How this can be handled in effective manner to get some value out of it, this is the biggest question in front of all engineers today. Geo Spatial Data, this data is another type of data which is getting produced because of the objects on the surface either they are static or dynamic. As per the statistics every year there is a 20% increase in Geospatial data production. And this Geospatial Data can be used for multiple purposes in various applications like autonomous vehicles, location based services, identifying the object in surface etc..., but the biggest challenge faced here is how this data can be analyzed and stored for future purpose. This data may be live data or stored data, it might be structured, un-structured or quasi structured data, it might be with duplicates or without duplicates and with null values or without null values. The challenge here is how this data can be used to perform data analytics and produce the results which can be used for future use. In the proposed research the main concentration is on how Geospatial data can be cleaned and made ready to use for data analytics for future use in applications like driverless vehicles, Location Based Services etc.., the first step in performing data analytics is collecting the Geospatial data then cleaning the same for further use. Once it is cleaned and ready to use the data analytics will be performed for further decision making.

Keywords: Data Analytics, Geospatial Data, Structured, Quasi Structured, Un-Structured.

1 Introduction

Geospatial data is getting accumulated in different forms each and every day. As geospatial data is going to be the next generation data which would be used in marketing, sales, identifying the people of same mindset etc.., the geospatial data will play vital role in the next generation business. The geospatial data is going to be the heart of decision making in next generation business either IT or Non IT Industries. Most of the companies are working on processing this geospatial data either for their own purpose or for the customers. Every year the size of the geospatial data is getting increased on an average of 20%. The biggest challenge here is how this data can be handled further for the performing analytics and to take final decisions.[15]

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308858

The geospatial data is entirely different from normal data which is getting accumulated every day. The normal data might consist of text, images or videos which can be easily analyzed for further data processing. Coming to geospatial data the biggest challenge is storing the longitude and latitude of an object. Once the data of the object stored it can be further used. But here the biggest challenge is if the object is dynamic, continuously the location information has to be stored. Sometimes more than the exact information, nearby equal information has to be recorded to get the required output of the application. The geospatial data has to be processed effectively in order to get the final results. [14]

Here in the proposed research the main concentration is on developing a frame work in order to process geospatial data analytics with the operations like Geospatial data gathering, Analyzing, Cleaning, Enriching and Storing the data for further use. The first step in analyzing the geospatial data is to remove the spaces and null values. And also avoid the duplicate data.

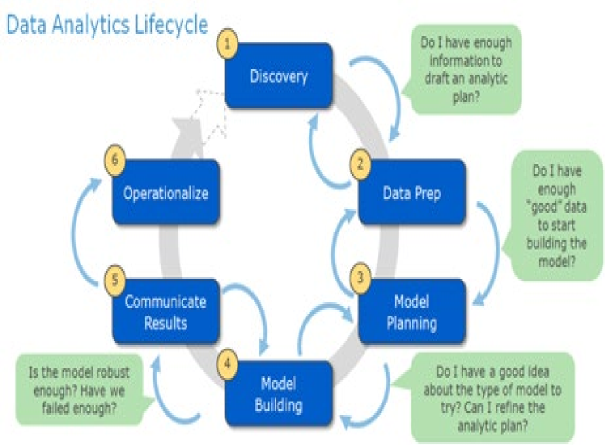
Data Analytics Lifecycle
1
Discovery
Do I have enough information to draft an analytic plan?
2
Data Prep
Do I have enough "good" data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed enough?
5
Communicate Results
6
Operationalize

Fig.1. Data Analytics Lifecycle

By referring the default data science life cycle (Referred from Google) Fig.1. it is clearly evident that to build any model for data analytics firstly the data has to be analyzed and pre processed. Once the data is ready as per the requirement the model can be built effectively. [13]

In the same way geospatial data analytics also has the process to be followed while processing the data to build a model. By understanding, the Fig.1. As reference the model for Geo Spatial Data Analytics will be build effectively with accurate results.

2 Literature Survey

In this paper the main takeover from this paper is how spatial data will be analysed and visualized with respect to cloud. Terrafly Cloud a wlak through has been done to understand how the data will be processed for further use. [19]

This paper has given insight into how data can be utilized to intercommunicate amongst the vehicles for smooth and intelligent transportation. The data processing techniques has been understood to process the collected data while taking the decision. [20]

This paper gave a deep insight into how Location Based Services using GMAPS can be used for identifying the near by vehicles. And how algorithms like KNN can be used to preprocess and take decision on the LBS. [7]

This paper gave insight of how to identify the Natural Disaster based on the previous data and also which can be used to process and predict current scenario. [8]

This paper gave insight of how to build a framework for moving objects like Vehicles using service oriented architecture. And how to use cloud based intelligent system to build the advanced automation system for autonomous vehicles. [9]

The main takeover of this paper is to understand how geospatial data can be combined with big data for further processing. How machine learning or AI can be combined with geospatial data for further processing. [18]

The main insight in this paper is how big data can be combined with geospatial data for further processing to perform analytics. [17]

3 Proposed Method:

After extensive research on the existing methodologies, frameworks, algorithms etc…, used for data analytics either on the normal data or Geospatial Data the proposed frame work will be developed in order to perform data analytics on the geospatial data. Performing this analytics on geospatial will surely help the industries in the growth of their organization, to do next level marketing, identify the potential customers and identify the potential market, trace the customers when they move around etc…,[5]

The proposed method consists of step by step process as shown below before performing the analytics. The geospatial data has to be cleaned by following step by step process manually i.e. line by line but the problem here is time consuming. So there are tools which will be used for further processing like tableau, MS-Power BI for further processing in the form of heat maps. But the major problem here is it can only process the data which is already structured. [4]

The biggest challenge here is how to analyze and make structured data out of the initially gathered geospatial data which might be having duplicates, null values, redundant values etc.., most of the times the data collected will be Un-structured or Quasi Structured Data. [6]

The proposed research mainly concentrates on the data cleaning and enriching activities of geospatial for further use. [3]

The following are the steps followed in doing the process.

- Problem definition
- Requirement gathering
- Data acquisition
- Data Fusion, Filtering and pre-processing[3]
- Data Extract & Store
- Data Cleansing& Quality Assurance
- Data Partitioning
- Spatial Data Analytics

The above steps are considered while developing the proposed frame work for geospatial data. All these steps has to be performed every time to make sure that the quality and structures data is getting generated which is further used for the data analytics as well as to build the model.[12]

Common Issues faced in the existing system are:

Huge geospatial data is getting generated – The problem here is

- ➢ The entire data is treated as a one data set. Then the comparison and segregation will take huge time.
- ➢ Geospatial data clusters are not created based on the problem definitions.[10]
- ➢ The data is still at macro level not processed at micro level.
- ➢ Searching the required geospatial data in the entire data sets is very difficult.
- ➢ Normal data will be stored in the form of rows and columns which can be easily tracable, but same method with respect to geospatial data wont workout as it is not going to have fixed data and it is application specific.[2]

After the continuous research on the existing methodologies the key observations made are.

- ➢ There are so many tools available to perform the data analytics and provide the results.
- ➢ The main issue is the available tools are not application specific they are very generic in nature.[2]
- ➢ Segregation of geospatial data as per the need of the customer is very much essential.

On top of the above key observation the biggest challenge is providing the proper input geospatial data to the tools. That means there is a need for system or application which can easily segregate the data and make the geospatial data as structured data before giving it as input to the tools for analytics. [1]

4 Proposed Methodology:

In the proposed method the main concentration is given on creating platform to segregate data based on following steps as shown in Fig.2.

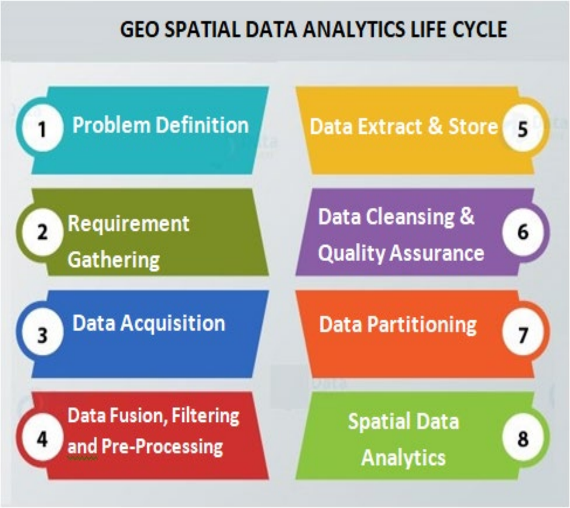

Fig.2. Geospatial Data Analytics Life Cycle

5 Problem Definition

Before simply processing the data, the first step is need to analyze what kind of application this data will be used and check whether it required entire data or only portion of the data.

This is the initial phase but crucial phase to define the problem so that the delay in further process will be reduced. [4]

Requirement gathering

Once the problem is defined, the next step is to gather the requirements as per the problem statement. What are the required main parameters and supporting parameters based on that the data will be gathered either live data or stored data. [2]

Data Acquisition

Collecting the data of the objects based on the movement on the surface and converting them into required format either in the form of tables, paragraphs etc.., and [1]

Data Fusion, Filtering and pre-processing

In this step data fusion will be done by collecting data from multiple sources and integrating the required data as per the nee of the application. [3]

Once Data Fusion is done, the filtering operation is carried out where the data collected in fusion will be considered in the form of sub sets as a portion of data and segregated data based on the requirement.

As a last step the data pre processing is carried out to make the geospatial data into understandable format for further processing. And data will be stored in the tables or data base. [4]

Data Extract & Store

Once the data preprocessing activity is performed, only meaningful data will be available for further use. Based on the requirement the data extraction process is carried out and stored into temporary tables for performing analytics to make decisions. [4]

Data Cleansing& Quality Assurance

In this step in order to make sure that there is no inconsistent data, the data will be verified for incomplete data, inconsistent data, null values, duplications etc.., to make sure that the data is qualitied for further processing. [2]

Data Partitioning

In the above step the entire data is processed made readily available for further processing, in this step the data will be partitioned further based on the macro level need to serve their micro solutions. Means based on the need it will be partitioned and stored, based on the requirement to gather data it will be directly searched in that portion in order to save the time and get the accurate data for further processing**.** [3]

Spatial Data Analytics

This is the next step after entire input spatial data is ready as per the problem definition. This spatial data can be further used for the analytics to solve real time problems. [11]

Algorithm - Selection & Clustering:

```
1. Start
2. Select the Problem Statement/Application
3.  Collect the required information based on Step 2
4. Analyse the required data after collecting information
5. Apply the process of Data Cleansing and Fusion
6. After step 5 cluster the data based on the need.
7. Apply the Analytics and store the result data for further use
8. Repeat Step 4 to 7 until the final results are obtained
9. Consider the final output as input to the final deciaion making.
10. Stop
```

The above algorithm will give an exclusive procedure to be gollowed while performing spatial data analytics.

6 Conclusion:

After extensive survey and research on geospatial data it is concluded that there should be a systematic procedure based on platform to analyse the geospatial data in very effective manner for further decision making at high level in an organization. This geospatial data can be used for various applications like sentimental analysis, autonomous vehicles, business intelligence etc..,

Future Enhancement:

In future the same system can be enhanced with efficient and optimized way to identify the natural disasters. It helps especially in the field of autonomous vehicles to move further one step. And also sentiment analysis for the machines can be implemented to test the efficiency and accuracy of the system.

References

[1] Hema M.S., Maheshprabhu R., Nageswara Guptha M. (2018) Data Access in Heterogeneous Data Sources Using Object Relational Database. In: Venkataramani G., Sankaranarayanan K., Mukherjee S., Arputharaj K., Sankara Narayanan S. (eds) Smart Secure Systems – IoT and Analytics Perspective. ICIIT 2017. Communications in Computer and Information Science, vol 808. Springer, Singapore. https://doi.org/10.1007/978-981-10-7635-0_3

[2] Hema M S and Dr. Nageswara Guptha M, "Service Oriented Quality Driven Ontology Based Data Federation with User Feedback", Proceeding IEEE Second International Conference Convergence in Technology, 2017. (IEEE explorer)

[3] Hema M S and Nageswara Guptha M, "Data fusion in data federation using modified discriminative Markov logic networks" International Journal of Advanced and Applied Sciences, 2016, Vol. 3, No. 8 pp. 78-84. DOI: 10.21833/ijaas.2016.08.013

[4] M. Nageswara Guptha, P. T. Rajan, A. Chitra, "Data Sourcing using Federated Database System in Service Oriented Architecture", CSI Communications, April – 2009, Vol 33, 1, 12-18

[5] T. Ortner, J. Sorger, H. Steinlechner, G. Hesina, H. Piringer and E. Gröller, "Vis-A-Ware: Integrating Spatial and Non-Spatial Visualization for Visibility-Aware Urban Planning," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 2, pp. 1139-1151, 1 Feb. 2017

[6] A. Elliethy and G. Sharma, "A joint approach to vector road map registration and vehicle tracking for wide area motion imagery," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 1100-1104

[7] M. F. Rahman, S. Bin Suhaim, W. Liu, S. Thirumuruganathan, N. Zhang and G. Das, "ANALOC: Efficient analytics over Location Based Services," 2016 IEEE 32nd International Conference on Data Engineering (ICDE), Helsinki, Finland, 2016, pp. 1366-1369, doi: 10.1109/ICDE.2016.7498346.

[8] J. Wang, Y. Wu, N. Yen, S. Guo and Z. Cheng, "Big Data Analytics for Emergency Communication Networks: A Survey," in IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1758-1778, thirdquarter 2016,

[9] Z. Yang et al., "Web service-based SMAP soil moisture data visualization, dissemination and analytics based on vegscape framwork," 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 2016, pp. 3624-3627, doi: 10.1109/IGARSS.2016.7729939.

[10] Amir Gandomi, Murtaza Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management Volume 35, Issue 2,2015, Pages 137-144,

[11] Mengfan Tang, P. Agrawal, S. Pongpaichet and R. Jain, "Geospatial interpolation analytics for data streams in eventshop," 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 2015, pp. 1-6, doi: 10.1109/ICME.2015.7177513.

[12] D. Guo and Yi Du, "A visualization platform for spatio-temporal data: A data intensive computation framework," 2015 23rd International Conference on Geoinformatics, Wuhan, China, 2015, pp. 1-6,

[13] L. J. Klein et al., "PAIRS: A scalable geo-spatial data analytics platform," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 1290-1298, doi: 10.1109/BigData.2015.7363884

[14] A. Andrejev, D. Misev, P. Baumann and T. Risch, "Spatio-Temporal Gridded Data Processing on the Semantic Web," 2015 IEEE International Conference on Data Science and Data Intensive Systems, Sydney, NSW, Australia, 2015, pp. 38-45,

[15] C. Zhou, P. Meysman, B. Cule, K. Laukens, and B. Goethals, "Discovery of spatially cohesive item sets in three-dimensional protein structures," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 5, pp. 814–825, 2014.

[16] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[17] X. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," IEEE Access, vol. 2, pp. 514–525, May 2014.

[18] A. Katal, M. Wazid, and R. Goudar, "Big data: Issues, challenges, tools and good practices," in Proc. 6th Int. Conf. Contemp. Comput. (IC3), Aug. 2013, pp. 404–409.

[19] Yun Lu, Mingjin Zhang, Tao Li, Yudong Guang, and Naphtali Rishe. 2013. Online spatial data analysis and visualization system. In Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA '13). Association for Computing Machinery, New York, NY, USA, 71–78.

[20] W. Chen, F. Guo and F. Wang, "A Survey of Traffic Data Visualization," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 2970-2984,