# Methodological Review On Big Datamodeling For Analytics

Pooja S.B[1], Shiva Balan[2]
[1] Student at NI university, [2]Assistant Professor,
[1] poojasnair100@gmail.com, [2] rvsivan@gmail.com,
Department of Software Engineering,
Kumaracoil, Thucklay, Kanyakumari, Tamilnadu, India - 629 180.

**Abstract-Modern technology has marshaled data outputs in a prodigious manner. The management of these generated data outputs in each of the fields has become a painstaking task. The realm of Big Data has now stepped up in to the next level of predictive analytics where data reduction has a major role. By utilizing a variety of statistical, modeling, data mining and machine learning techniques the recent and historical data can be used to predict possible outcomes, which would in fact help the stakeholders to analyze things in a better way. Big data modeling plays an important role in effective production of such prediction and its accuracy. In the field of Geographic Information Systems, These models help to manage the geospatial Big Data and provide a visual way to manage the massive amount of data resources that would in turn reduce the computation cost. Some of the data model can analyze the massive amount geospatial Big Data. There are different data models available, but these models faces different challenges such as time consumption, less accuracy, duplication etc. To overcome all those challenges and to meet all our need we need more such models. This paper studies various data models and their performance.**

***Keyword:*** **Big Data, Big Data modeling, Big Data analytics, Remote Sensing**

## 1. INTRODUCTION

"Any sufficiently advanced technology is indistinguishable from Magic" – As data is the wealth of our time it last longer than systems. Access to data is nowadays becoming ultimately competitive advantage i.e. why organizations like Google, Facebook etc try hard to render us things free and keep us logged in all the time. In such a competitive environment where things like Cost, Time, effort etc. counts, there is an immense need for proper modeling of the real life data to predict the future. Such efforts are being made since the civilization; there was always a question on 'what', 'why', 'what will' that existed in human which is evident from hisnotion of divine omniscience. Now all that we do is to give a scientifical approach to this existing practices. As the real life data really means big in nature all that we need to attain this is to have more investments into real life applications and to have better processing and better storage technologies.

Big Data as a term elucidates 'large data sets', which cannot be stored in the traditional database. The volume and the detail of data produced in many of the domains like social media platform, space research, agriculture research, media networks, meteorological, Laboratory, Internet of things etc. started increasing per diem. It may be either in structured or in unstructured format. Big Data help both the structured and unstructured data to store, process, analyze and manage effectively. To formulate the complex data, life cycle of Big Data is used. Based on the life cycle, Big Data is considered to have three main challenges data challenge, process challenge and Management challenge.

 

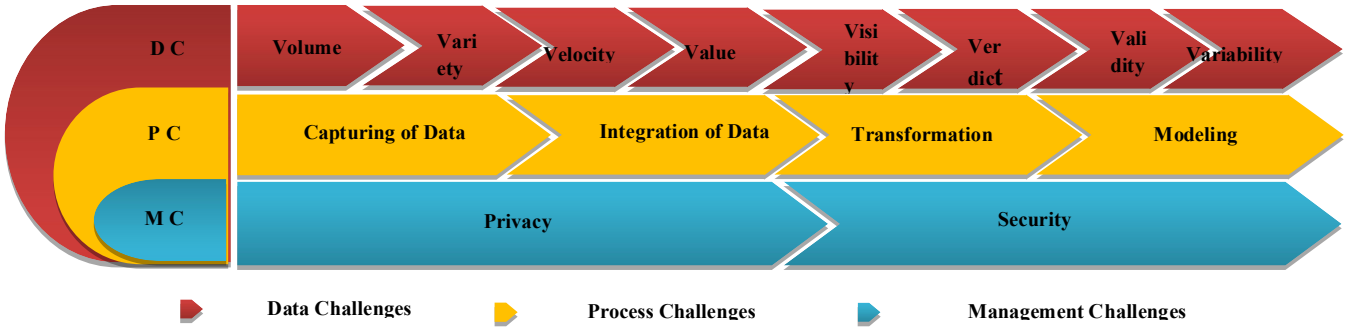


Fig 1: life cycle of Big Data

Data challenge relate to the data itself, which include"3vs" which is introduced by Gartner analyst Doug Laney in a 2001 metaGroup publication. Volume alludes to the massive amount of information generated by different fields every day i.e. gathering huge amount of data from different field. Example in business, social networks, in research etc would produce massive amount. Thus, the size of the data generated is said to be volume. Variety alludes to the data type that is generated by each field. The data may be structured, unstructured (audio, video, images etc) or semi-structured. Velocity: It refers to the speed at which the data is generated and the speed at which the data is transferred. Example messages from social media.IBM introduces fourth V Veracity refers to the uncertainty of the data. This may due to the lack of quality and accuracy. Value refers to the identification of hidden data from the large data set [1] [2].Process challenge includes the capturing of data, integration of data, transformation of data, and modeling. Management challenges include privacy, security, etc. [3].Now the number of V's is extended up to nine.

Big data challenges include Security, Scalability, visualization, storage, data transformation, data quality, data integrity, and availability [4]. The data transfer is affected by network characteristics, end-system characteristics, and dataset characteristics. It becomes more complicated during the transfer of heterogeneous file. By using pipelining, parallelism and concurrency, the data transfer can be enhanced [5]. The decision model determines the weight of the data transferred from a big data. To acquire the knowledge from Big Data some of the techniques such as data mining, knowledge discovery, and ontological approach are used [6].

## 2. BIG DATA ANALYTICS

In 2009, Kryder projected that if hard drives were to continue to progress at their then-current pace of about 40% per year, then in 2020 a two-platter, 2.5-inch disk drive would store approximately 40 terabytes (TB) and cost about $40.The volume of data generated in all the fields is increasing day by day, which is said to be the Big Data. To analyze, store this massive amount of data is quite difficult. Big Data analytics is also known as Big Data mining. Many methodologies have been used for Big Data analytics e.g. MapReduce[7].Big Data analytics helps to construct valuable information from the Big Data.

Data identification is the first step carried out in which the data are collected. Then the data is gone through the preparation stage, which means the collected data will contain some redundancy and may be inconsistent so to make the data feasible this step is used. In preparation stage, preprocessing is also done which consist of integration, transformation, cleaning, normalization, missing value imputation, noise identification and data reduction. Next step is modeling in which different modeling methods is evaluated

 

and then the best model that suits for our work is selected. After modeling deployment is done that is said to be converting the data to our need. Last step is prediction that is done to bring out more Accuracy (ACC) [8].

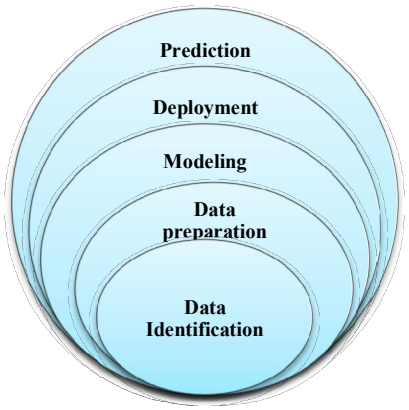


Fig 2. Steps involved in Big Data Analytic process

$$ACC = \frac{\text{Number of cases correctly classified}}{\text{Total number of test cases}}$$

To improve the efficiency of big data analytics and to avoid the issues in job scheduling genetic algorithmcan be used. For this Non-dominated Sorting Algorithm (NSGA-II) is used [9].To obtain the hidden information from the Big Data Analytics scalable parallel processing infrastructure is used (MapReduce). Clustering, classification, association rules and sequential pattern are some of the algorithm that are used to find out the hidden information from the raw data. Hadoop implement the MapReduce framework in Big Data Analytics. Data proliferation can be analyzed in a short span of time by using some of the methods which include parallel distributed algorithm, divide and conquer, clustering, sampling, reduction methods etc. To speed up the computational time of big data analysis sampling is used[10]. To find out the relation between the data association rule, linear regression, sequential pattern etc are used.

Data with different formats is generated from different sources. After understanding the different features of the data, which may include the type, trait, sources and accuracy the data is selected. The quality of the data can be evaluated by using Data Quality-in-Use model. This model consists of three characteristics contextual adequacy, temporal adequacy and operational adequacy, which is used to study the internal and external characteristics of Big Data [11].

To extract the information from the data machine learning is the primary step [12]. Traditional machine learning algorithm is difficult to use in the conventional environment because it has the capability to use a small amount of dataset. To handle the huge amount of data MapReduce that is a distributed processing algorithm is used. Map would perform filtering and sorting where reduce will perform grouping and aggregation. Map will split the dataset into <key-value> pair feature of the data is extracted by means of spatial analysis[13][14]. Pixel fusion, decision level fusion and feature level fusion are some of the traditional data fusion methods. These traditional methods cannot use for remote sensing data since the remote sensing data is of

 

different format [15]. Steps involved in data fusion are data identification, data collection, data preprocessing, and then data fusion is done then the target parameter estimation is done [16].

To store and manipulate Big Data NoSQL (Not Only SQL) and NewSQL is used. To bring out the storage phase in Big Data management data clustering, replication, indexing etc are used. To compress the large volume of data into groups clustering is done and the entities that have the similar features are placed together. Clustering would convert the large volume of data into small volume. To reduce the storage space Buza, Nagy propose an algorithm Storage-Optimizing Hierarchical Agglomerative Clustering (SOHAC) that is based on hierarchical agglomerative strategy, which would create different cluster for each object. Indexing process helps to improve the performance and future efficient retrieval of stored data [17].There are different tools available in the big data analytics but some of the most commonly used tools are [18] [19].

| SL. No | The Computer Science | The Mathematical Science | IT Challenge | The Multi-disciplinary approach |
|---|---|---|---|---|
| **1.** | Design Of Algorithm | Statistics | Storage | Contextual Problem Solving |
| **2.** | Conceptualization | Optimization | Computational power | |
| **3.** | Scalability (Machine Learning, Network & Graph analysis, streaming of Data and Text mining) | Uncertainty Quantification | | |
| **4.** | Distributed data | Model development (statistical, Ab Initio, simulation) | | |
| **5.** | Data Revolution, Reduction and Implementation | Analysis and Systems Theory | | |
| **6.** | Frameworks &Architectures | | | |

Table: 1 Domain Specific Challenges to Big Data Analytics

| Big data analytic software tool | Data storage & Management software tool | Data cleaning tool | Data and Oracle data mining tool | Data visualization tool | Data Integration tool |
|---|---|---|---|---|---|
| Accenture | Hadoop | OpenRefine | RapidMiner | Tableau | Blockspring |
| Alpine Data | Cloudera | Datacleaner | IBM SPSS Modeler | Silk | Pentaho |
| Alteryx | MongoDB | Trifacta | Teradata | CartoDB | Stitch |

Table 2: Shows the different software tool used in Big Data

### 2.1 Hadoop

The Hadoop is an open-source java based frame work which can store large sets of data. It uses MapReduce programming language. Apache Hadoop project provides open-source software for high performance, scalable and distributed computing which allows for the distributed processing of huge data sets across group of computers with simple programming models. Distributed File System that is the storage part is used by the Hadoop (HDFS)]. It concentrate on real time data analytics. Hadoop uses innovative 'schema on read' for data analysis which makes it suitable to deal with highly unstructured data

 

inputs. To process large amount of data, Apache Hadoop is used. Companies such as Amazon, Microsoft, Google, Cloudera, Alacer use Apache Hadoop. To process huge data Skytree server is used Graphical environment is provided by Talend Open Studio to perform big data analysis it is used. To produce the report from the database column jaspersoft can be used. By using Dryad the parallel and distributed program can be improved. The report can be generated from the huge data set by using Pentaho[26,27]. Data warehousing is typically a single RDBMS where Hadoop spread across clusters of machines which has an advantage of data locality. It can handle huge amount of data surpassing the ability of a single machine. Hadoop does not expect the data to be structured it has Schema on Read which makes it possible to handle Unstructured data. Using Hive structures can be made on top of it. So it is flexible. Hadoop is cost effective as it is open-source technology. Big Data Decision theory is used to process decision making these techniques are classified into statistical technique, data analysis technique and visualization technique. Mathematical techniques include static and optimization methods. Data curation and analysis is supported by static method. Optimization methods has elucidate as efficient techniques. Due to the nature of quantitative implementation. Data analysis technique includes data mining, machine learning, artificial neural networks and signal processing. Data mining including classification, regression and clustering, regression, statistical learning, association analysis and linking mining [28].visualization technique helps to make the data meaningful [29].

On the top of the Hadoop cluster, MapReduce is implemented. In multiple computing node the data are stored. In the form of MapReduce, distributed algorithm is implemented. To handle large data set MapReduce is used [30].The advantages and disadvantages are discussed below in table 2.

| Advantage of Hadoop | Disadvantage of Hadoop |
|---|---|
| Distributed data processing | Single master node |
| Easy to handle | Cluster management is too hard |
| Independent task | Unobvious configuration of the node |
| Linear | Restrictive programming model |
| Simple programming model | Joining of multiple data set which make slow down the system |

Table 3: Advantage and disadvantage of Hadoop

### 2.2 MapReduce

The large data set can be effectively handled by Hadoop and the programming model used in Hadoop is MapReduce [31] which was proposed by Dean and Ghemawat at Google. Hadoop was inspired by Google papers which uses MapReduce Programming Model. It is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. This concept was emerged from the 'Divide and Conquer' method.

The input data is initially loaded into HDFS database here after initial partition of complex data into smaller units they are mapped. These mappers process the data and produce intermediate results to reducers. Reducers are used to collect the intermediate results to produce the final result which is again written to HDFS. A typical Hadoop job involves running several mappers and reducers across different nodes in the cluster. MapReduce is a YARN-based programming paradigm for

 

parallel processing of massive data sets. The process of MapReduce jobs includes the Map phase and the Reduce phase. Each phase has *key-value* pairs as input and output, the types of which may be chosen by the programmer through specifying two functions: the Map function and the Reduce function [32, 33].The figure below shows an illustration of mapping and reducing in Map reduce Programming language.

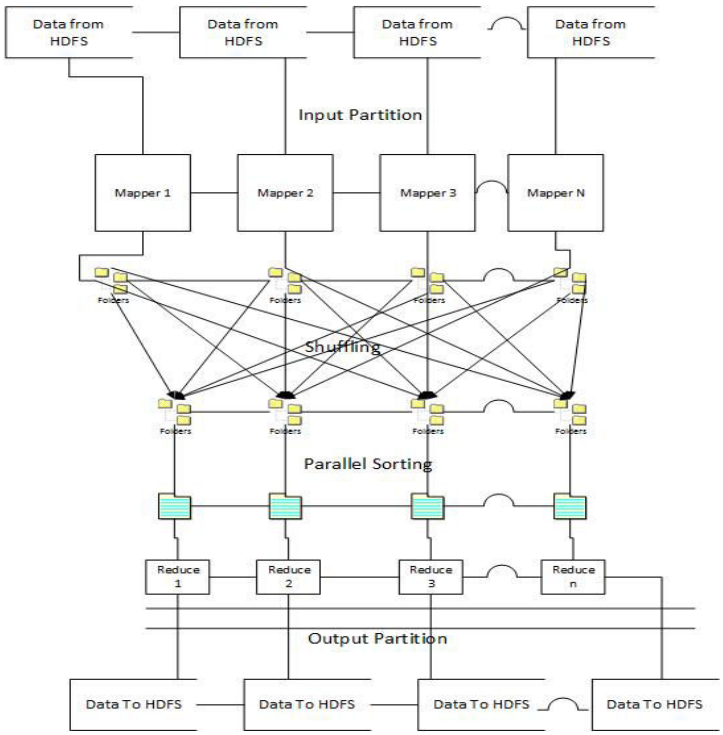


Fig: 3 Illustration of Map & Reduce Process

### 2.3 Apache Spark

Spark is a de facto framework of upcoming paradigm for big data processing developed by the University of California at Berkeley. Spark is a fast framework that used in both business and research. Spark stands as an alternative to Hadoop. Spark consists of different components, which includes Spark core, GraphX, Spark streaming and Spark SQL. Each of them has its own usage like machine learning, graphic analysis, storage etc . It is structured to overcome the disk I/O fault and to enhance the performance of earlier systems. It has an ability to perform in-memory computation, which makes it unique from all others. It permits the data to be cached in memory, thus avoiding the Hadoop's disk overhead limitation for iterative tasks. Spark supports Java, Scala and Python and can run large data systems. In some cases, it is tested to be up to 100× speeder than Hadoop where the information is store in the memory, when data resides on the disk it is 10× speeder. It can run on

 

HadoopYarn manager and the data can be taken form HDFS which makes it extremely versatile to run on different systems [32, 33].

## 3. BIG DATA MODELING

To manage the huge amount of data we make use of the Big Data modeling. Data modeling used to design the data base data. Data modeling is the first step done in the data base development. For the implementation of data base design, both the data model and the data base model are used. Big Data model is constructed by creating data blocks, which include storage, data type, relationship, read write requirement etc, then by using the modeling application the models are maintained

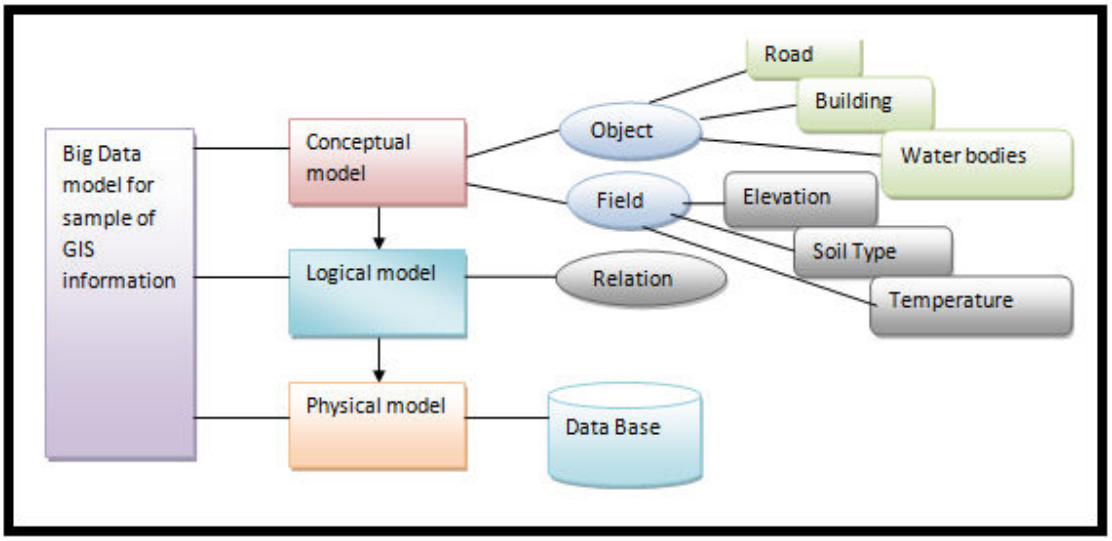


Fig 4: Big data model for GIS information

The Big Data model consists of three different layers, which include conceptual model, logical model and physical model. Conceptual model include the object filed etc. In GIS the object (discrete) include road, building, water bodies etc To detect and recognize object 3-D modeling (text/sign analysis) is done. The individual tree structure i.e. width, height, location, crown can be extracted from the lidar data by using Canopy Height Model (CHM). Lidar provide the direct measurement of the forest structure. To produce a remote sensing image with high-resolution lidar data and temporal data MODIS can be linked together[34]. The field is continuous that contain elevation, soil type, temperature etc. To represent the reality pixel is used. To represent the object smaller than the pixel size the pixel is into smaller chunks. In each pixel, the physical phenomena such as ground phenomena are changed [35]. Logical model bring out the relation between the conceptual object. Physical model is the last level of the data modeling it comprises the storage.Based on the distributed approach large storage space can be provided by the HDFS. Large data can be parallel processed by Hadoop and MapReduce [36].

The spatial data model include the vector data mode, network data model, topology data model, geometry data model, raster data model, tessellation data model, irregular data model etc.[37]. Raster and vector model is the basic data model. Only after encoding, the geographical data is used. The real world contains the spatial features form that the spatial feature identified.

 

Then the needed area is selected. The selected area is then delineated in the conceptual data model by choosing either raster or vector model. To encode the data choose the appropriate data model.After choosing the data model, the spatial data structure selected to store the model. Geospatial data stream does not have the capability to store the connectivity and adjacency but the network and topological model need to store the connectivity and adjacency. Thus, the spatial data indexing is used.

### 3.1 Vector and raster data model

The vector data model help to describe the road, building, lake etc i.e. it helps to store the individual features so it is said to be discrete data model. Point, line and polygon are the three main geometric shape used to represent the real world entities. When similar data is stored in raster and vector then the vector data require less space for storage when compared to the raster data.

Raster data model used to describe continuous features. Square shaped row and columns with equal size can represent it. Each cell contains a value the value can either be in integer or in decimal (raster model contain value which is stored in the form of matrix). The raster itself does not contain any color; the color assigned by the software for interpretation this done by the value taken from the cell. To build the GIS image in the raster model the individual cell used to represent the point, network, line, area. The raster data structure integrated with the remote sensed image. After performing the modeling, the next step is the classification. There are different methods used for classification i.e. Support Vector Machine, decision tree, Naïve Bayesian classification. The Support Vector Machine (SVM) is the method used for performing supervised and unsupervised classification. that helps to create group after analyzing the data pattern with the help of statistical learning theory SVM classification include pattern matching, text classification etc.[38].

One of the disadvantages of the raster model is that it takes more space than the vector data model. The compression techniques play a very vital role with development of GIS technology. It would represent the information with more accuracy by using the least storage space. The memory space taken by the raster model to store the spatial data is also large. To diminish the size of the raster data set and to avoid the redundancy the compression method is used to save the storage space and for the effective transaction on the network. Compression is of two types lossless compression and lossy compression.

The encoded data can be decompressed in lossless compression or decoded without any loss of the information and we will thus get the exact data. The decompressed file remain as same as that of the original one. Because in the lossless compression the redundant data is eliminated, thus we will get the exact data after the decompression. Lossless compression is made used by some of the image file formats like PNG (Portable network graphics) or GIF (Graphics Interchange Format). Whereas some other file formats like TIFF (Tagged Image File) and MNG (Multiple-image Network Graphics) may use either lossless or lossy method. Thus, this technique helps to reduce the size of the data for storage purpose. Some of the example of lossless compression method are run length encoding, Huffman coding, Prediction by Partial Matching (PPM) etc.

In lossy compression when the compressed data is decompressed the data that are relived will not remain as same as that of the original. This is due to the algorithm would eliminate irreverent information. The advantages of lossy compression are in some cases lossy compression can produce smaller compressed file than lossless compression method. Some of the lossy methods are JPEG, PCM and MPEG.

 

**3.2 Geometric and Topological model**

To find out the structural features of the data set topological method is used. It is done before performing supervised or unsupervised analysis. To analyze massive amount of data geometric and topological model is used. To study the distance function geometric function can be used that deal with the point cloud data set. The finite sample taken from the geometric object with noise is said to be the point cloud. For qualitative mathematics, topology provides a formal language on the other hand geometry is quantitative.

In topology, the proximity of nearness relation is studied without using distance. After defining topology based on the notation of nearness the properties such as continuity, connectedness, and closeness are introduced. The input is in the form of the point cloud and the output is the data summary collection that helps to estimate the topological features Algebraic topology is used to make the topological features into simpler groups.

**Mathematical background of Topology:** Morse theory helps to relate the global features with the local features of critical points. Morse theory identifies points in which the topological changes have occurred.

Let h :M → R

is taken as a continuous function defined on a domain M.

For each scalar value a ∈R,

the level set $h^{-1}$(a) = {x ∈M | h(x) = a}

may have multiple connected components

A set of points collectively join to form a topological space. Open set means the collection of subset [26]. After introducing topology, some of the properties such as continuity, connectedness and closeness introduced. U is a set and it is open when a point starts from U and goes on any direction inside that set.

A topological space is a set X and a set τ of subsets of X satisfying

The following axioms:ϕ and x are in τ

if U1, U2, . . . , Un are in τ, then so is $\cap_{i=1}^{n} U_i$

if $U_i$, i ∈ I are in τ, then so is $U_{i\epsilon I} U_i$

**3.3 Spatial and Temporal model**

The spatial and temporal characteristics of drought are studied Palmer Drought Severity Index (PDSI) is the meteorological drought indexes used all over the world [39]. To find out the severity of drought across the different climate PDSI is the first procedure used. Based upon the primitive water balance model PDI is used. To accelerate the query performance in the temporal Big Data temporal index plays an important role. The current temporal index faces difficulty to handle the variety of query. Thus for temporal Big Data a new segmentation based hybrid index B+- tree (SHB+-tree) is proposed [40]. This would improve the construction and maintenance performance. At first according to the time order, the temporal data present in the temporal table is divided into small chunks. By joining the temporal and object index, the hybrid index is built in each of the divided fragments and then the temporal Big Data is shared by them. The construction and maintenance performance is improved by using segmented storage strategy and bottom- up index construction in each part of the hybrid index. The database model is

 

implemented by using the PostgreSQL, which is an open source relational database management system. PostGIS extension is used for the spatial functions [41].

.

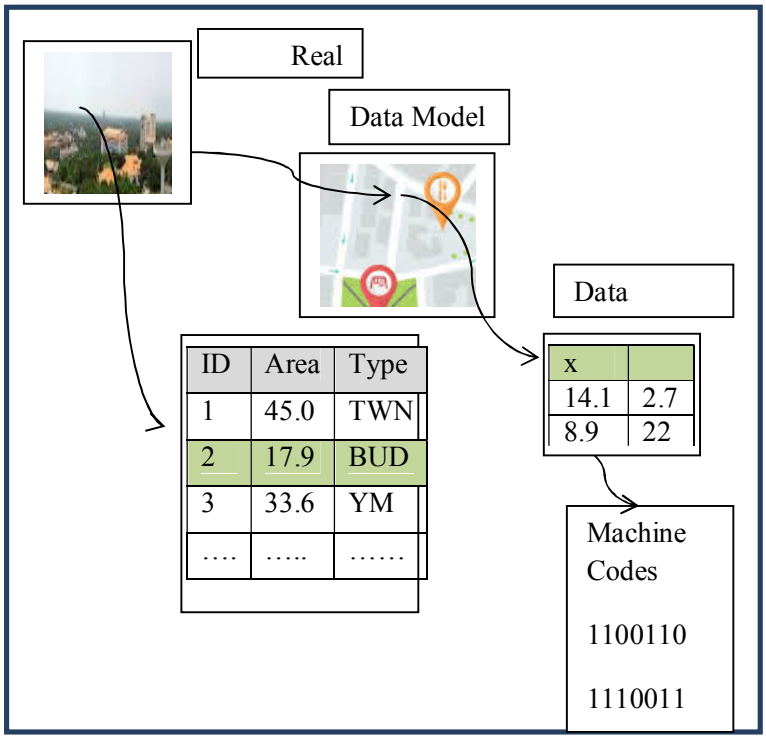


Fig 5: Representation of Spatial entities

## 4. CASE STUDY:

**Remote Sensing Field:**

The main goal of the remote sensing application is to extract the needed information. Massive amount of remote sensing data generated every day from all over the world This remote sensing data can be used in many applications which includes disaster management i.e. to identify flood, drought, to identify the forest mapping, to identify the road, building, city planning, GPS navigation, water prone area, Agricultural, National Wasteland Monitoring etc. The remote sensing data received from the satellite contain flaws and deficiencies. The correction of these deficiencies can be gone through preprocessing stage that involves geometric correction, radiometric correction, atmospheric correction noise removal etc. Issues in remote sensing data include the security, data compression, and data retrieval. The remote sensing big data analytics architecture includes the Remote Sensing Big Data acquisition Unit (RSDU), Data Processing Unit (DPU) and Data Analysis Decision Unit (DADU). In the first step, the data is been acquired from the satellite and send it to the base station. In the second step DPU include filtration, load balancing and parallel processing. The final step DADU performs compilation, storage and generation of result[30].Challenges involved in remote sensing data are Uncertainty and incompleteness, temporal variability, spatial autocorrelation, spatial heterogeneity and multi-resolution and multi-scale changes [41].Many algorithms such as image de-noising, image restoration, image fusion, change detection, feature extraction and image interpretation use wavelet transform for multi-resolution representation [42].

 

| GIS Data Types | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Attribute Data | | | | Spatial Data | | |
| Description | Describes the characteristics of spatial features | | | | Describes the absolute and relative location of geographic features. | | |
| Models | Database Models | | | | Data Models | | |
| Formats | Record Based | | | | Vector | | Raster |
| Characteristics | Tabular Model | Hierarchical Model | Network Model | Relational Model | Topology Data Model | Computer Aided Drafting Data Model | Raster Data Model |
| Data Structure | 1. Every record shares the same set of variables.<br>2.Every row has the same set of column headers | 1. One to Many or One to One relationships<br>2.Based on parent child relation Ship | 1.Allowed the network model to have many to many relationship<br>2. A record can have many parents as well as many children | 1.One to One, One to Many, Many to Many relations<br>2.Based on relational data structures | 1.Intelligent data structure<br>2.Records adjacency information | 1.Listing elements data structure<br>2.String of vertices | 1.Grid cell data structure<br>2.Regularly spaced grids |
| Arrangement/ Principles | Sequential Data files with Fixed Formats | Tree Structure (Hierarchy of tables) | Plex Structure | Organized by Rows and Columns | Feature adjacency and connectivity | Alpha numeric and Graphical | Data Accuracy and resolution based. |
| Drawbacks | 1. Fewer Capabilities<br>2. Does not supports relations | 1.Redundancy<br>2.Fails to handle many relations efficientl y | 1. No independency between objects<br>2.Difficult to design relations | 1. High hardware cost<br>2.Minute level design leads to Complexity in database | 1.More complex,Inefficient for high spatial variability<br>3.Overlay operations are difficult | 1.More complex,Inefficient for high spatial variability<br>3.Overlay operations are difficult<br>4. Lacks definition of spatial relations between features | 1.Less Compact, 2.topological relations are difficult to represent<br>3. Output graphics are less aesthetic |

Table 4: Overview on existing GIS Systems

**Road Extraction**: It is too complex to model a road structure using a general structural model. Therefore, at first, the road features and the road model are analyzed then the classification of road extraction method is done[33]. In road feature, image enhancement is done to extract the needed information from the remote sensing image. To extract the road more effectively road model is used. Road extraction method used different algorithmsuch as classification- based, knowledge based dynamic programming active control etc. Classification based method uses geometric, photometric, texture features of a road. The classification accuracy is too

 

low. This classification-based method is divided into supervised and unsupervised classification. To train the labeled sample supervised classification method is used. The supervised classification method includes Artificial Neural Network (ANN), Support Vector Machine (SVM), Markov Random Fields (MRFs), and Maximum Likelihood (ML). Unsupervised classification does not need any training samples. The accuracy of unsupervised classification method is less when compared to the supervised classification method. To perform parameter description and knowledge discovery unsupervised classification is done. Commonly used algorithms are mean shift, graph theory, K-mean, Spectral etc.

## 5. CONCLUSION

After studying the existing data models in depth, it has understood that the data modeling requirements changes according to the domains on which the analysis is performed. The structuring& the source of raw data collection are very much essential for effective analytics. Even though there are different data models to bring out meticulous result, the complexity still exists. To handle the large data sets new data models are very much required. Data modeling helps to mold the data in a prescribed form according to the need of the user. In future, a new data model for predictive analytics in the agriculture sector by using raw data from GIS, Meteorology, Central Survey Organizations etc., which would enhance the planning and policy formation can be perform.

## REFERENCE

[1] Baumann, P., Mazzetti, P., Ungar, J., et al. 2015,"Big Data Analytics for Earth Sciences: the EarthServer approach", Digital Earth.

[2] Marjani, M., Nasaruddin, F. and Gani. A, 2017, "Big IoT Data Analytics: Architecture Opportunities and Open Research Challenge",IEEE Access , 5,5247-5261.

[3] Sivarajah, U., Kamal, M.M. andIrani, Z.2016, "Critical analysis of Big Data Challenges and analytical methods", Business Res, 263-286.

[4] Hashem, I., Yaqoob, I ., Anuar, N.B., 2014, "The raise of big data on cloud computing: Review and open research issues", Information System, 98-115.

[5] Yildirim,E., Arslan, E., Kim, J. et.al. 2015,"Application-Level Optimization of Big Data Transfers through Pipelining, Parallelism and Concurrenc"IEEE Trans on Cloud Computing, 4(1), 63-75.

[6] Wu, C., Chen,Y. and Li, F., 2016,"Decision Model of Knowledge Transfer in Big Data Environment", China Communications, 13(7), 100-107.

[7] lv, Z., Song, H., Basanta-val, P. et al.,2017, "Next Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics" , IEEE Trans on Indus Inform, 13(4), 1891-1899.

[8] Singh,D and Reddy C, K., 2014, " A survey on platform for big data analytics big data", J Big data, 2:8

[9] Lu, Q; Li, S; Zhang, W; et.al., 2016, "A genetic algorithm-based job scheduling model for big data analytics wireless communication and networking". EURASIP J Wireless Comm Network, 152

[10] Tsai, C., Lai, C., Chao, H., et.al. 2015, "Big data analytics: a survey", J Big Data, 2:21

 

[11] Jorge,M., Ismael,C., Bibiano, R., et.al. 2015,"A Data Quality in Use model for Big Data", Future Generation Computer Systems, 63,123-130.

[12] Xing,E.P., Ho,Q., Dai,W.,et al. 2015, "Petuum: A new Platform for Distributed Machine Learning on Big Data, IEEE Transactions on Big Data, . pp. 49-67.

[13] Grolinger,K, Hayes, M; Wilson A, et.al 2014 " Challenges for MapReduce in Big Data" Services, IEEE World Congress on Services , 182-189.

[14] Wang,M., Wu, Y., Yen, N. et.al. 2016,"Big Data Analytics for Emergency Communication Networks: A Survey", IEEEComm Surveys Tutor 18(3), 1758-1778.

[15] Chi, M., Plaza, A., Benediktsson, J, A., et.al., 2015, "Big Data for Remote Sensing Challenges and Opportunities" , Proc. of the IEEE, 104 (11),2207 - 2219

[16] Xiong, G; Zhu, F; Dong. X, et.al., 2016, "A Kind of Novel ITS Based on Space-Air-Ground Big-Data, IEEE Intell Transport Sys Magaz. 8(1) 10-22.

[17] Siddiqa,A., Hashem, I,A,T., Yaqoob,I., et.al., 2016, " A Survey of Big Data Management: Taxonomy and State-Of-the-Art" J Network Computer App.71, 515-166.

[18] Rehman, M,H., Chang, C., Batool, A, et.al., 2016, "Big Data reduction framework for value creation in sustainable enterprises" ,Int J Inform Management, 36(6), 917-928.

[19] Lu, X.F.,Cheng,C,Q., Ya,G,J et.al., 2011, "Review of data storage and management technologies for massive remote sensing data" , Sci China Technol. Sci. 54(12) 3220-3232.

[20] Huang, H, G., Lian,J., et.al., 2016, " A 3D Approach to Reconstruct Continuous Optical Images using Lidar and MODIS" , Forest Ecosystems, 2:20.

[21] N. Q. Mehmood, R. Culmone, and L. Mostarda, "Modeling temporal aspects of sensor data for MongoDB NoSQL database," J. Big Data, 2017.

[22] D. Xia, H. Li, B. Wang, Y. Li, and Z. Zhang, "A Map Reduce-Based Nearest Neighbor Approach for Big-Data-Driven Traffic Flow Prediction," IEEE Access, vol. 4, pp. 2920–2934, 2016.

[23] I. Yaqoobet al., "Big data: From beginning to future," Int. J. Inf. Manage., vol. 36, no. 6, pp. 1231–1247, 2016.

[24] X. L??, C. Cheng, J. Gong, and L. Guan, "Review of data storage and management technologies for massive remote sensing data," Sci. China Technol. Sci., vol. 54, no. 12, pp. 3220–3232, 2011.

[25] Y. Chen, H. Chen, A. Gorkhali, Y. Lu, Y. Ma, and L. Li, "Big data analytics and big data science : a survey," vol. 12, no. March, 2016.

[26] H. Wang, Z. Xu, H. Fujita, and S. Liu, "Towards felicitous decision making: An overview on challenges and trends of Big Data," Inf. Sci. (Ny)., vol. 367–368, pp. 747–765, 2016.

[27] M. NaimurRahman, A. Esmailpour, and J. Zhao, "Machine Learning with Big Data An Efficient Electricity Generation Forecasting System," Big Data Res., vol. 5, no. February, pp. 9–15, 2016.

[28] J. Archenaa and E. A. M. Anita, "A Survey Of Big Data Analytics in Healthcare and Government," Procedia - ProcediaComput. Sci., vol. 50, pp. 408–413, 2015.

[29] K. Grolinger, M. Hayes, W. a. Higashino, A. L'Heureux, D. S. Allison, and M. a. M. Capretz, "Challenges for MapReduce in Big Data," Proc. Serv. - IEEE World Congr. Serv., no. Services, pp. 182–189, 2014.

[30] J. Wang, Y. Wu, N. Yen, S. Guo, and Z. Cheng, "Big data analytics for emergency communication networks: A survey," IEEE Commun. Surv. Tutorials, vol. 18, no. 3, pp. 1758–1778, 2016.

[31] S. Salloum, R. Dautov, X. Chen, P. Xiaogang, and J. Z. Huang, "Big data analytics on Apache Spark," Int. J. Data Sci. Anal., vol. 1, no. 3, pp. 145–164, 2016.

[32] Dilpreet Singh, Chandan K Reddy. "A survey on platforms for big data analytics", Journal of Big Data 2014

 

[33] J. Vijayaraj, R. Saravanan, P. V. Paul, and R. Raju, "A COMPREHENSIVE SURVEY ON BIG DA TA ANALYTICS TOOLS CHARACTERISTICS DEVELOPER PROGRAMMING LANGUAGE CURRENT VERSION COMMUNITY SUPPORT," 2016.

[34] Lefevre, S., Tuia, D., Wegner, J, D., 2017, "Toward Seamless Multiview Scene Analysis From Satellite to Street Level",Proc. of the IEEE, 105 (10), 1-16.

[35] Bhatta,B ., 2013, " Research method in remote sensing" , Springer.

[36] Chen, H,M; Chang, K,C; Lin,T,S., 2016 "A Cloud-based system framework for performing online viewing, storage and analysis on big data of massive BIMs" , Autom in Construct. 71, 34-48.

[37] Dragicevic, S, S., Castro, F, A., Sester, M; et.al., 2016,"Geospatial Big Data handling theory and methods: A review and research challenges", ISPRS J Photogramm, 115, 119-133.

[38] Cavallaro, G., Riedel, M., Richerzhagen, M; et.al, 2015, " On Understanding Big Data Impacts in Remotely Sensed Image Classification using Support Vector Machine Methods" , IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens., 8[10] 4634-4646.

[39] Snasel, V., Nowakova, J., Xhafa; et.al., 2015,"Geometric and topological approaches to Big Data" , Future Gen. Comp Sys. pp:286-296.

[40] Xul, X, Xie.F, Zhou,X, 2016."Research on spatial and temporal characteristics of drought based on GIS using Remote Sensing Big Data" , Cluster comp.19(2), 757-767.

[41] Wang, M., Xiao,M., Peng,S., and Liu, G., 2016, "A hybrid index for temporal Big Data" Future Gen. Comp Sys.,72: 264-272

[42] Wieland, M and Pittore,M., 2017. "A Spatio-Temporal Building Exposure Database and information Life-Cycle Management Solution". ISPRS Int. J. *Geo*-Inf, 6, 114-118

 