A Comparative study of Rasdaman and Open Data cube for Time Series Geospatial Big Data

[1]Jayati Gandhi,[2]Nekita Chavhan,[3]T.P.Girish Kumar

[1]P.G.Research Student,[2]Assistant Professor,[3]Scientist/Engineer
[1]Department Of Computer Science and Engineering,
[2]Department Of Computer Science and Engineering,
[1]G.H.Raisoni College Of Engineering, Nagpur, India,
[2]G.H.Raisoni College Of Engineering ,Nagpur, India,
[3]RRSC-ISRO(Central),Nagpur, India.

Abstract : Earth Observation (EO) has been constantly generating large amount of Geospatial data over the last few years which is used in resource monitoring, environment protection, and disaster prediction. The applications like Ground surveying, remote sensing and mobile mapping produces geo-spatial data. The growth of EO data has great challenges in recent approaches for data management and processing. By applying the traditional data analysis tools many issues arises when we use these huge amount of data. Therefore the Array Database technologies are used in managing and processing EO Big Data's. Array database technologies are mainly used to support multi-dimensional data management and analysis. The Array database technologies such as Rasdaman and Open data cube are used. It provides flexible, fast, scalable geo services for multi-dimensional data. The main aim of this paper is to implement the efficient way of storing, retrieving and processing of temporal satellite data by doing the comparative analysis of Open data cube and Rasdaman array database.

Index Terms - **Remote sensing, Earth Observation Data, Array Database, Open data cube, Rasdaman.**

I. Introduction

Earth Observation is done for the following activities such as gathering, managing, processing, observing and representing the physical, chemical and biological information pertaining to earth system by using remote sensing techniques. It has tremendous applications for environment monitoring [1]. Earth Observation satellite produces regular stream of multi-spectral, multi-resolution, multitemporal remote sensing images due to the development of sensor technologies.

Many platforms like satellites, planes and vehicles have been used as sensor carriers to collect various data of earth and generates large amount of data types and formats [2]. Today's excessive availability of Earth observation (EO) datasets gives clear and improved understanding of the environment on different scales like regional, continental, and planetary. EO datasets is freely available by various space agencies. This includes data like MODIS and Landsat, from currently launched satellites like the Sentinels. Earth observation satellites generate petabytes of geospatial data.

Big data like geospatial big data is also used in the society for various purposes like meteorology, diagnostics, disaster management, logistics, and so on. In today's world 80% of the data is geo referenced which shows the huge importance of handling geospatial big data [3]. The quantity and importance of big geospatial data hosted on cloud environments is continuously growing [4]. Remote sensing data are collected and used in different industries and research institutes due to the development of earth observation and GIS techonology [5].

Geospatial data is the information of an object, defined by values in a coordinate system. In general language, geospatial data is used to represent the shape, size and location of an object on earth which includes country, rivers and towns.

Flat files or Database Management Systems (DBMSs) are used to store the large amount of data. For managing remote sensing images different data base technologies like relational database and array database are used. Currently array database technologies are tremendously used for managing remote sensing images [6]. The array database is created and implemented as a common database service as it provides flexible and efficient storage and retrieval of multidimensional array data like sensor, image, simulation or statistics data [7].

It has gained attention of various data scientist from different industries. For storing this large amount of data Open Data cube and Rasdaman are used. They are some of the open source libraries and have array db implementation. They are used to store time series satellite data as multidimensional arrays. The Open Data Cube project (ODC) provides open-source tools for setting up infrastructures and provide access to satellite images as data cubes [8]. The implementation is done in Python and supports the method of simple image indexing .The data stored in Rasdaman, provides a flexible processing and accurate service which is based on Open Geospatial Consortium (OGC) standards.

The Australian GeoScience Data Cube (AGDC) project has established EO Big Data storage and processing framework by using different technologies like multidimensional array-based storage and HPC technologies [9].Most of the Geospatial data such as remote sensing images are usually multidimensional arrays as per the terms of data structure, so it naturally follows an approach for storing and managing data in an array database. Array-oriented management solutions have been available for several decades for the development of Hierarchical Data Format (HDF) [10] and Net CDF [11] data formats and libraries started in the late 1980s.The main objective of this paper is to do the comparison of Rasdaman and Open data cube array database..

A. Geospatial Big Data Analysis

According to Morais ''80% of data is geographic", which means that the presence of Georeferenced data in the real world is very large in quantity [12]. These large volumes of data present the importance of handling big data techniques and tools. The Geospatial data has the coordinates of location that describes objects and things with respect to geographic area. Evans et al. [13], the geospatial data is called as "Big data" as it has the characteristics of big data. Shekhar et al. says that, spatial data is regularly increasing and represents the characteristics of huge size, high variety and high update rate of datasets [14]. These data needs appropriate efforts to understand new data processing and data management technologies as it is using the mechanisms of spatial computing. The big data is increasing the expectations of researchers to manage huge data in both ways of increasing speed and observing capacities as well [15].

B. Array Database

For storing and operating multi-dimensional discrete data (MDD) array is designed. There are two methods to understand the concept of array. First is from the point of function mapping, an array A is considered as a function f (a): D → V, mapping is done from an index domain D to a value domain V [16]. In this regard, Array provides a convenient and efficient approach to gain the values from indexes. Second is the set theory approach that treats an array as a collection of same elements ordered in a discretized space [17]. Each element in the space is called cell and each cell contains a value. The coordinate is actually a vector that is used for identifying the particular position of a cell in the space. The length of coordinate is called dimension. In the real world, array data are usually represented as images, multimedia, simulation or statistics data appearing on Earth and space. The array database technologies such as Rasdaman and Open Data cube are used which provides flexibility and scalability in storing large amount of data. The storage process for array databases may differ, Open Data cube provide a SQL-like query language and Rasdaman provide a rasql query language.

1. Rasdaman: For the applications of array data Rasdaman provides a flexible, high- performance and scalable DBMS. It follows an approach of client/server architecture in which the queries queries are processed on the server side. For raster data storage a base RDBMS is used to support BLOB (Binary Large Object). Arrays in Rasdaman are divided into tiles which are the basic unit for data storage and then they are access and stored as BLOBs in the base DBMS [18]. The Rasdaman server act as a middleware which is used for mapping the array semantics into the relational table semantics. Rasdaman provides a SQL-like language called as Rasdaman Query Language (RasQL) to change raster data. Thus RasQL queries are parsed and run by Rasdaman servers, which are used in retrieving data from the base RDBSM. Rasdaman also provides a Web application called as petascope [19], which implements some OGC (Open Geospatial Consortium) Web Service interfaces including Web Coverage Service (WCS) and Web Coverage Processing Service (WCPS). With the help of this application, both geospatial raster data and geoprocessing functions can be shared on the Internet. Rasdaman is largely used in the domain of geospatial science.
2. Open Data Cube: ODC provides an on open and freely accessible exploitation tool to enchance the impact of satellite data and to help a community to develop some applications. An implementation of the ODC is made up of three things at the technical level i.e data, an index and software.

Data is generally file based, either in local directories of GeoTIFFs or NetCDF files, but it can be anything that GDAL can read, including Cloud Optimized GeoTIFFs stored on AWS' S. PostgreSQL is used as a database to store a data type, like Landsat 8. The index enables a user to ask for data at a time and location, without requiring to know where the required files are stored and how to access those files. The Software of the ODC is a library of Python that enables a user to index , ingest and to query data and a wide range of other functions related to managing data

II. LITERATURE SURVEY

The Planet Server system is a service component of the EU-funded Earth Server project. It aimed at providing and observing planetary data online. Earth Server project1 has created an on demand online open access and ad hoc infrastructure for huge amount of Earth observation data. The WCPS is proved to be efficient for quickly processing large amounts of data and delivering of finished products to the end user at a very extremely low cost by mixing a full hyper spectral unmixing chain as a part of the NASA Web Sensor suite of web services and by mixing standard processing and vegetation analyzing methods for agricultural applications [20]. For the production of valuable geo-information cloud based platforms and high performance computing forms the single way forward for applying image analysis task and data analytics task over the web[21],[22],[23]. Inorder to store big EO data, some advanced algorithms are required to retrieve, store and classify information from large datasets [24]. Retrieval from the dataset of satellite image has been created which is based on the method of semi-supervised for the annotation of images [25] and on the enrichment of metadata, of the semantic annotations, and the image content [26].Therefore these multidimensional data are generated which needs to be stored in database system. On the relational model, a relational database management system (RDBMS) , which was not appropriate for multi-dimensional data.Since RDSMSs have been largely used, and there are some shortcomings of storing spatial data thus researchers developed fixes for storing spatial data efficiently. PostGIS is an extension for PostgreSQL and it is a free spatial database. It which allow users to create their own backend for various purposes like mapping, raster analysis and routing applications. It also allow users to create their own queries in SQL format.. To store MODIS fire archive Davis has successfully implemented a use case by using PostGIS extension [27]. To manage large raster data MYSQL is used as a backend for WebGIS. It is an application of RDBMSs for Geospatial data [28]. For measuring how a database supports big data scalability is used which is considered as another indicator [29]. RDBMSs scaled up with costly hardware, but did not work accurately in parallel with commodity hardware [30].The Apache Hadoop project developed open source software to overcome this problem which allow the distributed processing of large datasets for working on clusters of commodity computers by using simpler programming models. Since the traditional RDBMSs could not efficiently handle the huge data like satellite images and weather simulation data. For big scientific data management Array DBMSs have become a great area of research [31].To create these logical views on EO data various technical solutions have rapidly gained traction over the past few years. Array database technologies are used for storing large amount of data. Open data cube and Rasdaman are used for storing large amount of multidimensional data. The first national scale EO data cube was established in Australia [32], whose technology is now the support of Digital Earth Australia [33] and the Open Data Cube (ODC) [34]. ODC is free and open source technology. It stands behind other operational EO data cubes, such as in Switzerland [35], Colombia [36], Vietnam [37], the Africa Regional Data Cube [38] and some other nine national or regional initiatives are under development [39]. Rasdaman is an array database system that has been used since the mid-1990s. For storing large amount of data Rasdaman is another leading technology. In order to process datasets of images along with grid-based database structure Pioneering Array DBMS (PICDMS) is used [40]. Another pioneering Array DBMS is called as Rasdaman which is also called as raster data manager.

III. COMPARISON OF RASDAMAN AND OPEN DATA CUBE

For accessing, managing, and observing large quantities of Geographic Information System (GIS) data called as Earth observation (EO) data open data cube is used. It represents a common analytical framework composed of a number of series of data structures and tools which facilitate the organization and analysis of large gridded data collections. The Open Data Cube was developed for analysing earth observation data and it's very flexible thus allows other meshed data collections to be included and analyzed.

Open data cube is designed to analyzed large amount of Earth Observation Data. It provides python API for executing high performance queries also provides flexible data access. Open Data Cube also allows scalable continent for scaling processing of stored data.

It provides an environment of data analysis for analyzing earth observation data from multiple satellites. Open data cube provides SQL query language for supporting retrieval, manipulation and data definition. It provides Python API.
The multi-dimensional arrays, such as sensor, image, simulation, and statistics data which appears on earth and space can be stored by Rasdaman. It is world's leading array analytics engine and it is different from others because of its flexibility, performance, and scalability.

Rasdaman can process arrays present in file system directories as well as in databases. For supporting retrieval, manipulation, and data definition Rasdaman provides a query language called rasql. Different data loading functions based on GDAL can be implemented by Rasdaman for supporting different raster data formats like GeoTiff, NetCDF, and HDF. Thus there is no need of preprocessing the original datasets because it is automatically loaded as multi-dimensional arrays.

It also provides users to modify their functions by C++, Java, and Python API. But it cannot distribute the input data across the cluster automatically, and users have to specify which data are loaded into which node. The Configuration and set up of Open Data Cube is very complicated as compared to Rasdaman. Rasdaman is more flexible and cost saving than Open Data Cube. The processing of queries is complicated in Open Data Cube as compared to Rasdaman. Thus it is the world's most flexible and scalable array engine.

IV. CONCLUSION

This paper presents the comparison of Rasdaman and Open data cube for storing, retrieving and manipulating the satellite data. Earth Observation (EO) has been constantly generating large amount of Geospatial data, so for storing these huge amount of data Array Database Technologies are used. Array data base technologies such as Rasdaman and Open data cube proves to be efficient in storing large amount of Geospatial big data. Hence they allow storing and querying large amount of multi-dimensional arrays like sensor, image, simulation, and statistics data. Rasdaman proved to be the world's most flexible and scalable Array Engine. The processing of queries and configuration of Rasdaman is much more easier than that of Open data cube.

.

REFERENCES

[1] Guo, H.; Liu, Z.; Jiang, H.; Wang, C.; Liu, J.; Liang, D. "Earth big data: A new challenge and opportunity for Digital Earth's development". Int. J. Digit. Earth 2017, 10, 1–12.

[2] Di, L.; Moe, K.; van Zyl, T.L." An overview of Earth observation sensor web". IEEE J. Sel. Top. Appl. Earth Observation Remote Sens. 2010, 3, 415–417.

[3] Hong Shu (2016) " Big data analytics: six techniques", Geo-spatial Information Science, 19:2, 119-128, DOI: 10.1080/10095020.2016.1182307

[4] Wang, X., Zhao, J., Zhou, Y., & Li, J.. " The geospatial data cloud: An implementation of applying cloud computing in geosciences". Data Science Journal, 13 (2015), 254–264. https://doi.org/10.2481/dsj.14-042

[5] Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C., Buytaert, W. 2015. Web technologies for environmental Big Data. Environmental Modelling & Software 63 (2015) 185-198

[6] G. Planthaber, M. Stonebraker, and J. Frew, "EarthDB: Scalable analysis of MODIS data using SciDB," in Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, ser. BigSpatial '12. New York, NY, USA: ACM, 2012, pp. 11–19.

[7] Tan, Z.; Yue, P." A comparative description to the array database technologies and its use in flexible VCI derivation". In Proceedings of the 2016 Fifth International Conference on Agro-Geoinformatics,Tianjin, China, 18–20 July 2016; pp. 1–5

[8]Open Data Cube. Available online: https://www.opendatacube.org (accessed on 23 May 2019)

[9] Lewis, A.; Oliver, S.; Lymburner, L.; Evans, B.; Wyborn, L.; Mueller, N.; Raevksi, G.; Hooke, J.;
Woodcock, R.; Sixsmith, J.; et al. The Australian geosciences data cube—Foundations and lessons learned.
Remote Sens. Environ. 2017, doi:10.1016/j.rse.2017.03.015

[10] The HDF Group, "HDF group history," https://support.hdfgroup.org/ about/history.html, [accessed: 2018-01-23].

[11]Unidata,"NetCDF,"http://www.unidata.ucar.edu/software/netcdf/, Boulder, CO: UCAR/Unidata Program Center, [accessed: 2018-01-23]

[12] Morais, C.D., 2012. '‘80% of Data is Geographic form" IEEE J.Sel 45-513.

[13] Evans, M.R., Oliver, D., Zhou, X., Shekhar, S., 2014. "Spatial big data: Different Case studies on volume, velocity, and variety. In: Karii, H.A. (Ed.), Big Data: Techniques and Technologies in Geoinformatics. CRC Press, pp. 149–176.

[14] Shekhar, S., Evans, M.R., Gunturi, V., Yang, K., Cugler, D.C., 2014. " Description of Bench marking of big data." In: Rabl, T., Poess, M., Baru, C., Jacobsen, H.-A. (Eds.), "Geospatial Big Data Benchmarks". Springer, Berlin Heidelberg, pp. 81–93.

[15] Gomes, L., 2014. "Machine-Learning Methods on the Delusions of Big Data and Other Massive Engineering Efforts", October 20, 2014. 2012; pp. 1–7

[16] A. R. van Ballegooij, "RAM: A study on multidimensional array DBMS," in Proceedings of the 2004 IEEE Conference on Current Trends in Database Technology, ser. EDBT'04.

[17] Berlin, Heidelberg: P. Baumann Guidance of Query language,"IEEE Conference on Rasdaman 2017 pp 120-124

[18]H. Pirk, Y. Zhang, S. Manegold, and M. Kersten. (2013)" Some queries on arrays in Monet data base" IEEE Conference on SQL queries 2016,pp 20-24

[19] A. Aiordachioaie , " Open Geo Consortium and WCS Geo Service Standards implementation ". Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 160–168.

[20] K. Karantzalos, A. Karmas, ": Big earth observation data processing for precision agriculture," in Proc. Eur. Conf. Precis. Agric., 2015, pp. 421–428.

[21] K. Evangelidis, K. Ntouros, S. Makridis, and C. Papatheodorou, "Services of Geospatial in the cloud," Comput. Geosci., vol. 63, pp. 116–122, 2014.

[22] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, "Big data in cloud computing" Inf. Syst., vol. 47, pp. 98–115, 2015.

[23] C. Lee, S. Gasster, A. Plaza, C.-I. Chang,, " Recent Remote sensing techniques " IEEE Journal. Sel. . vol. 4, no. 3, pp. 508–527, Sep. 2011.

[24] A. Krizhe5. Lewis, A.; Oliver, S.; Lymburner, L.; Evans, B.; Wyborn, L.; Mueller, N.; Raevksi, G.; Hooke, J.; Woodcock, R.; Sixsmith, J.; et al. "Australian geo science data cube Foundations". Remote Sens. Environ. 2017, 202, 276–292. [CrossRef]

[25] P. Blanchart , "An algorithm based on semi-supervised techniquen in satellite image dataset for auto annotation and unknown structures ," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 3, no. 4, pp. 698–717, Dec. 2010.

[26] Han, D.; Stroulia : "A review of geospatial data model". In Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing (CLOUD), Santa Clara, CA, USA, 28 June–3 July 2013; pp. 910–917.

[27] D.K.; Ilavajhala, S.; Wong, M.M.; Justice, C.O. "Resource management system for fire information". IEEE Trans. Geosci. Remote Sens. 2009, 47, 72–79.

[28] Zhong, Y.; Han, J.; Zhang, T.; Fang, J. "Distribution of geospatial data storage and processing framework for large-scale WebGIS". In Proceedings of the 2012 20th International Conference on Geoinformatics (GEOINFORMATICS), Hong Kong, China, 15–17 June or big data analytics: A technology tutorial. IEEE Access 2014, 2, 652–687.

[29] Yang, C.; Yu, M.; Hu, F.; Jiang, Y.; Li, Y. Addressing big geospatial data challenges by cloud computing. Comput. Environ. Urban Syst. 2017, 61, 120–128.

[30] Hu, H.; Wen, Y.; Chua, T.S.; Li, X. " Big data scalable analytics: A technology tutorial. IEEE Access 2014, 2, 652–687.

[31] Rusu, F.; Cheng, Y. An array storage survey for processing query languages. arXiv 2013, arXiv: 1302.0103

[32] Lewis, A.; Oliver, S.; Lymburner, L.; Evans, B.; Wyborn, L.; Mueller, N.; Raevksi, G.; Hooke, J.; Woodcock, R.; Sixsmith, J.; et al. The Australian Geoscience Data Cube—Foundations and lessons learned. Remote Sens. Environ. 2017, 202, 276–292. [CrossRef]

[33] Dhu, T.; Dunn, B.; Lewis, B.; Lymburner, L.; Mueller, N.; Telfer, E.; Lewis, A.; McIntyre, A.; Minchin, S.; Phillips, C. Australia Earth digital—A new value is unlocked from earth observation data. Big Earth Data 2017, 1, 64–74. [CrossRef]

[34] Killough, B. Overview of the Open Data Cube Initiative. In Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 8629–8632.

[35] Giuliani, G.; Chatenoux, B.; Bono, A.D.; Rodila, D.; Richard, J.-P.; Allenbach, K.; Dao, H.; Peduzzi, P. Making an Earth Observations Data Cube, IEEE Conference 2017, 1, 100–117.

[36] Ariza-Porras, C.; Bravo, G.; Villamizar, M.; Moreno, A.; Castro, H.; Galindo, G.; Cabera, E.; Valbuena, S.; Lozano, P. CDCol: Colombian Needs meets geo science data cube requirements. In Proceedings of the Advances in Computing, Cali, Colombia, 19–22 September 2017; Springer: Cham, Switzerland, 2017; pp. 87–99.

[37] Cottom, T.S. An Examination of Vietnam and Space. Space Policy 2019, 47, 78–84. [CrossRef]

[38] Group on Earth Observations (GEO). Digital Earth Africa: Project Overview.IEEE Conference 2016 pp 78-80

[39] Baumann, P.; Dehmel, A.; Furtado, P.; Ritsch, R.; Widmann, N. The multidimensional database system RasDaMan. In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 1–4 June 1998; ACM: New York, NY, USA, 1998; Volume 27, pp. 575–577.

[40] Chock, M.; Cardenas, A.F.; Klinger, A. Database structure and manipulation capabilities of a picture database management system (PICDMS). IEEE Trans. Pattern Anal. Mach. Intell. 1984, 6, 484–492.