March - April 2020
ISSN: 0193-4120 Page No. 2027 - 2032

Geospatial challenges in Big Data using Cloud Computing

V. Satish Kumar [1], R. Anil Kumar [2]

[1] Assistant Professor in Computer Science & Engineering department at Dr. K V.Subba Reddy College of Engineering for Women, Kurnool

[2] Assistant Professor in Computer Science & Engineering department at Dr. K V.Subba Reddy College of Engineering for Women, Kurnool

Article Info
Volume 83
Page Number: 2027 - 2032
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract

Big Data is one the emerging concept which gives new opportunities for research, development, innovation and business. It's characterized by four Vs: volume, velocity, veracity and variety and should bring significant value through the processing of massive data. The transformation of massive Data's 4 Vs into the 5th (value) may be a grand challenge for processing capacity. Cloud Computing has emerged as a replacement paradigm to supply computing as a utility service for addressing different processing needs with a) on demand services, b) pooled resources, c) elasticity, d) broad band access and e) measured services. The utility of delivering computing capability fosters a possible solution for the transformation of massive Data's 4 Vs into the 5th (value). This paper investigates how Cloud Computing can be utilized to deal with Big Data challenges to enable such transformation. We introduce and review geospatial scientific examples, including climate studies, geospatial knowledge mining, and dust storm modeling. The tactic is presented during a tabular framework as a guidance to leverage Cloud Computing for Big Data solutions. It was exhibited with some examples that the framework method supports the life cycle of massive processing, including management, access, mining analytics, simulation and forecasting. This tabular framework also can be referred as a guidance to develop potential solutions for other big geospatial data challenges and initiatives, like smart cities.

Keywords: *Big Data, Cloud Computing, Spatiotemporal data, Geospatial science.*

1. Introduction:

Model simulation and earth observation produce tera- to peta- bytes of data daily [14]. Geospatial data acquisition methods, like phone conversations, unmanned aerial vehicles and social media produce geospatial data at even faster speeds. [15]With an addition to the massive Volume, geospatial data exist in Variety of forms for various applications, their accuracy and uncertainty span across a good range as defined by Veracity, and data are produced during a fast Velocity through real time sensors [16]. With unprecedented information and knowledge embedded, these big geospatial data are often processed for adding Value to raised research project, engineering development and business decisions. They envisioned supplying innovation and advancements to improve our lives and understanding of the world systems when transformed from the primary four Vs to the last V (value) through advancements during a sort of geospatial domains.

Such transformations pose challenges to data management and access, analytics, mining, system architecture and simulations. For instance, the primary challenge is how to affect the variability and veracity of massive data to supply a fused dataset which will be utilized during a single decision network [6]. Another issue is how to deal with the speed of Big Data to possess scalable and extensible processing power based on the fluctuation of the info feed. [20] Supporting on-demand or timely data analytical functionalities also pose significant challenges for creating the worth.

Cloud Computing has emerged as a new concept to generate computing as a utility service with five advantageous characteristics: a) rapid and elastic provisioning computing power; b) pooled computing power to raised utilize and share resources; c) broadband access for fast communication; d) on demand access for computing as utility services; and e) pay-as-you-go for the parts used without a big upfront cost like that of traditional computing resources [15]. Service-oriented architecture is integrated in Cloud Computing and enables "everything as a service", including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). For big geospatial data problems in various geosciences and relevant domains, Cloud computing engages Big data enlightens potential solutions while redefining about geospatial science and digital earth [5].

Utilizing Cloud Computing to deal with Big Data issues remains in its infancy, and it's a frightening task on how the five advantageous characteristics can address the primary four Vs of massive Data to succeed in the 5th V. [8] This paper illustrates how Cloud Computing supports the transformation with scientific examples including climate studies, knowledge mining and dust storm simulation. These examples are highly representative and can be easily adapted to other environmental and concrete research fields, such as smart cities. The life cycle of big geospatial data (data management and access, analyses/mining, phenomena/scenario simulation) are examined through these examples.

2. Climate analytics

In order to know climate change and its impacts to environmental and concrete issues, the climate data observed within the past and simulated for the longer term should be well managed and analyzed [17]. However, both observation and simulation produce Big Data. For instance, subsequent IPCC report is going to be supported100+petabytes of knowledge and NASA will produce 300+petabytes of climate data by 2030. This data is entirely different in format, study objective and in spatiotemporal resolution [23]. Big data can help advance the understanding of climate phenomena and can help to identify how impacts of global climate changes on society and ecosystems can be remedied. Detecting global temperature anomalies and investigating spatiotemporal distribution of utmost weather events, especially over highly populated regions [5].

The many petabytes of climate data can only be managed during a distributed and scalable environment. Cloud Computing could help the management as follows: a) provisioning on-demand flexible virtual machines (VM) consistent with the quantity of climate data; and b) automatically deploying HDFS, Hadoop Distributed filing system, on the VMs to create a distributed file system. Data are often maintained in native format rather than sequenced text for saving [12] space for storing. Logical data architecture is additionally built to facilitate fast identification, access, and analyses.
The core architecture is a spatiotemporal index for the multi-dimensional climate data stored on HDFS. The index maps data content onto the file, node and byte levels within the HDFS. Total nine components are used for the index that includes: shape, space and time information describe about grid's logical information which correlates to data query, byte offset, node list, byte length, compression code and file path identify specific location on the HDFS. Users can directly access and locate the data with content description and exact spatiotemporal from the index.

Table 1
The Big Data challenges as illustrated in the four examples are addressed by relevant cloud advantages to reach the Big Data Value and achieve the research, engineering and application objectives.

	On-demand Self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
Volume	2.1	4.1	2.1	2.1, 3.1, 3.2, 4.1, 4.2, 4.3, 5.1	4.1
Veracity	2.1	3.1, 5.3			
Velocity			2.1	4.1	4.3
Variety		3.1, 5.2	2.1		
Value	2.1, 3.2			2.1	3.2

March - April 2020
ISSN: 0193-4120 Page No. 2027 - 2032

3. Knowledge mining from big geospatial data

From different spatiotemporal stamps and resolutions, we have gathered big geospatial data for environment and concrete studies using various methods, e.g., Global Positioning System (GPS), remote sensing, and Internet-based volunteer. The increment in volume, velocity, and sort of the spatiotemporal data poses a grand challenge for researchers to get and access the right data for research and decision support. One method of addressing this Big Data discovery challenge is to mine knowledge from the [24] large geospatial data and their usages for query expansion, recommendation and ranking. The mined knowledge includes but isn't limited to domain hot topics, research trends, metadata linkage and geospatial vocabularies similarity.
Volume, velocity and variety of big data have been challenged by this process (Table 1). Such a mining process poses two challenges: a) the way to divide Big Data into parallelizable chunks for processing with scalable computing resources; and b) the way to utilize computing resources for processing the divided Big Data with an adaptable number.

On-demand resources within a virtual cluster

Cloud Computing facilitates automatic virtual cluster with a dynamic number of VMs .More computing resources can be transferred to process the big historic data, [29] while the dynamic number of VMs are often provisioned to handle real-time data streams. On-demand computing resources are necessary to satisfy the need of dynamical log data volumes. For instance, within the January 2014 PO. DAAC log mining [27] task with more VMs within the cluster, less time interval was spent on finishing the task. The two time-based partition and IP-based partition are accelerated dramatically for mining processes (Fig.1). Log processor changed the sessions which are generated from time-based partitions.

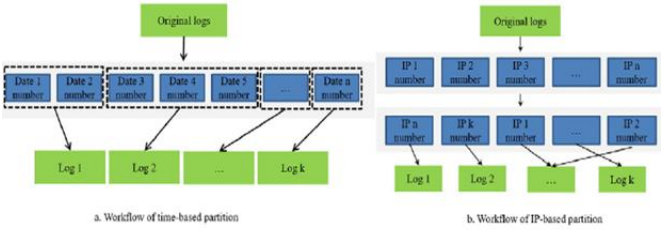

Fig. 1 The workflow of time-based partition (a) and IP-based partition (b).

4. Dust storm forecasting

Dust storms are serious hazards to health, property, and therefore the environment worldwide, especially urban areas. [28] Visibility has been decreased rapidly due to the increase of accidents during and after a duster; air quality and human health are compromised when dust particles remain suspended within the atmosphere; when dust interferes with the energy captures mechanics are applied then the efficiency of renewable energy sources are reduced. Therefore, it's crucial to predict an upcoming dust event with high spatiotemporal resolution to mitigate the environmental, health, and other asset impacts of dust storms [29]. A typical requirement for such prediction requirement is to simulate at some point phenomena within a two-hour computational time.

Variety of dust model input investigation

With the rise of spatiotemporal resolution of a dust forecast model, the biggest problem is to access dynamic data with different formats, content and uncertainties [26]. The potentiality of broad network access of Cloud Computing can serve the preprocessing of the model input file with advanced network bandwidth and scalability. Huang, Yang, Benedict, Chen et al.(2013) [17] and Huang, Yang, Benedict, Rezgui et al. (2013) [18] showed that Amazon cloud instances can complete most of the forecasting tasks in less time than HPC clusters (Fig. 2), indicating that Cloud Computing has potential to resolve the concurrent intensity of the computing demanding applications.

March - April 2020
ISSN: 0193-4120 Page No. 2027 - 2032

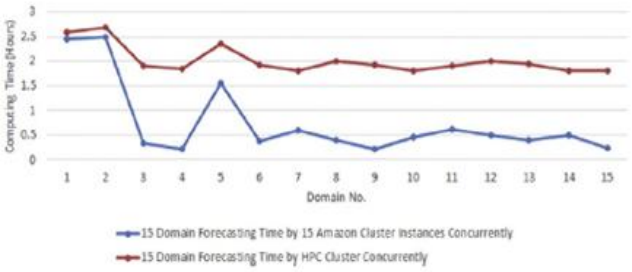

Fig. 2 NMM-dust execution time for 15 forecasting tasks on Amazon EC2 and HPC cluster (Huang, Yang, Benedict, Chen et al., 2013; Huang, Yang, Benedict, Rezgui et al., 2013).

Data veracity of dust forecast improvement

One of the foremost significant factors affecting the veracity of model output is that the uncertainty of model initial condition. These uncertainties are often investigated and characterized through sensitivity tests using various model variables. To revert the uncertainty of the initial conditions, data assimilation techniques are applied to correct initial conditions of the model from the observations. In order to improve model accuracy and also to reduce model uncertainty, regularly conduct sensitivity tests and data assimilation to keep the efforts of preprocessing and integration into the model.

Conclusion

In this paper we proposed geospatial data challenges in big data by introducing scientific examples like climate studies, geospatial knowledge mining and dust storm modeling. In order to exhibit the challenges, we presented a framework method that supports big data processing, management, access, simulation and mining analytics. Some of our other contributions can be summarized as follows:

- Spatiotemporal Big data processing requires real-time data processing, information extraction and automation to extract information and knowledge. More scalable spatiotemporal mining methods should be developed to require advantage of the elastic storage and computing resources.
- High number of tools are require to measure the usage of resources, including computing resources, data for pricing purposes and also to guide the usage of Cloud Computing services.
- To spot and stop attacks for tracking and to maintain trust information, we have to do more research that we will be addressed.

References:

[1] Ammn, N., & Irfanuddin, M. (2013). Big Data challenges. International Journal of Advanced Trends in Computer Science and Engineering, 2(1), 613–615.

[2] Batty, M. (2013). Big Data, smart cities and city planning. Dialogues in Human Geography, 3(3), 274–279.

[3] Bulkeley, H., & Betsill, M. M. (2005). Cities and climate change: Urban sustainability and global environmental governance. 4. (pp. 1–2). Florence: Psychology Press, 1–2

[4] Chen, C. P., & Zhang, C. -Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314–347.

[5] Das, M., & Parthasarathy, S. (2009). Anomaly detection and spatio-temporal analysis of global climate system. Proceedings of the third international workshop on knowledge discovery from sensor data (pp. 142–150) ACM

[6] Fan, J., & Liu, H. (2013). Statistical analysis of Big Data on pharmacogenomics. Advanced Drug Delivery Reviews, 65(7), 987–1000.

[7] Gordon, M. I., Thies, W., & Amarasinghe, S. (2006). Exploiting coarse-grained task, data, and pipeline parallelism in stream programs. ACM SIGOPS Operating Systems Review, 40(5), 151–162.

[8] Huang, Q., Yang, C., Benedict, K., Chen, S., Rezgui, A., & Xie, J. (2013a). Utilize Cloud Computing to support dust storm

March - April 2020
ISSN: 0193-4120 Page No. 2027 - 2032

forecasting. International Journal of Digital Earth, 6(4), 338–355\

[9] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and its technical challenges. Communications of the ACM, 57(7), 86–94.

[10] Jiang, Y., Li, Y., Yang, C., Armstrong, E. M., Huang, T., & Moroni, D. (2016). Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. ISPRS International Journal of Geo-Information, 5(5), 54.

[11] Knippertz, P., & Stuut, J. B. W. (2014). Mineral Dust. Dordrecht, Netherlands: Springer. Korf, R. E. (2011). A hybrid recursive multi-way number partitioning algorithm. IJCAI proceedings-International Joint Conference on Artificial Intelligence, 22(1), 591.

[12] Li, Z., Yang, C., Huang, Q., Liu, K., Sun, M., & Xia, J. (2014). Building model as a service to support geosciences. Computers, Environment and Urban Systems. http://dx.doi.org/ 10.1016/j.compenvurbsys.2014.06.004.

[13] Manuel, P. (2015). A trust model of Cloud Computing based on quality of service. Annals of Operations Research, 233(1), 281–292.

[14] Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. Computers, Environment and Urban Systems, 34(4), 264–277.

[15] Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 18–33). Berlin Heidelberg: Springer.

[16] Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phonecall data. Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development (pp. 11) ACM.

[17] Einav, L., & Levin, J. D. (2013). The data revolution and economic analysis (no. w19035). National Bureau of Economic Research.

[18] Marr, B. (2015). Big Data: Using SMART Big Data. Analytics and metrics to make better decisions and improve performance. Wiley 258pp.

[19] Yang, C., Xu, Y., & Nebert, D. (2013). Redefining the possibility of digital Earth and geosciences with spatial Cloud Computing. International Journal of Digital Earth, 6(4), 297–312

[20] Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. Communications of the ACM, 57(3), 78–85.

[21] Mell, P., & Grance, T. (2011). The NIST definition of Cloud Computing. Meng, X., Bradley, J., Yuvaz, B., Sparks, E., Venkataraman, S., Liu, D., ... Xin, D. (2016). Mllib: Machine learning in apache spark. JMLR, 17(34), 1–7.

[22] Rosenzweig, C., Solecki, W. D., Hammer, S. A., & Mehrotra, S. (Eds.). (2011). Climate change and cities: First assessment report of the urban climate change research network (pp. xvi). Cambridge: Cambridge University Press.

[23] Skytland, N. (2012). Big Data: What is NASA doing with Big Data today. (Open. Gov open access article).

[24] Jiang, B., & Thill, J. C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. Computers, Environment and Urban Systems, 53, 1–3

[25] Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). Spatiotemporal data mining in the era of big spatial data: algorithms and applications. Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data (pp. 1–10) ACM.

[26] Krämer, M., & Senner, I. (2015). A modular software architecture for processing of big

March - April 2020
ISSN: 0193-4120 Page No. 2027 - 2032

geospatial data in the cloud. Computers & Graphics, 49, 69–81.

[27] Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 33(6), 403–408.

[28] Knippertz, P., & Stuut, J. B. W. (2014). Mineral Dust. Dordrecht, Netherlands: Springer.

[29] Wilkening, K. E., Barrie, L. A., & Engle, M. (2000). Trans-Pacific air pollution. Science, 290(5489), 65.

[30] Zhao, C., Liu, X., Leung, L. R., Johnson, B., McFarlane, S. A., Gustafson, W. I., Jr., ... Easter, R. (2010). The spatial distribution of mineral dust and its shortwave radiative forcing over North Africa: Modeling sensitivities to dust emissions and aerosol size treatments. Atmospheric Chemistry and Physics, 10(18), 8821–8838.

[31] Niu, T., Gong, S. L., Zhu, G. F., Liu, H. L., Hu, X. Q., Zhou, C. H., & Wang, Y. Q. (2008). Data assimilation of dust aerosol observations for the CUACE/dust forecasting system. Atmospheric Chemistry and Physics, 8(13), 3473–3482.

[32] Darmenova, K., Sokolik, I. N., Shao, Y., Marticorena, B., & Bergametti, G. (2009). Development of a physically based dust emission module within the Weather Research and Forecasting (WRF) model: Assessment of dust emission parameterizations and input parameters for source regions in Central and East Asia. Journal of Geophysical Research. Atmospheres, 114(D14).