Analysis of big earth sciences data through cloud computing

Ramla Khan*

Open University, U.K

Abstract

Google Earth Engine (GEE) is a cloud-based platform that has revolutionized the field of remote sensing with its high-speed computing power and simple use. It is spread across 66,000 CPUs for parallel computation access to its users. It has a massive catalogue of datasets available for users that were once mined from different sites and has an efficient workbench environment for algorithm developments, processing the datasets, and crowdsourcing. GEE also helps in pixel-based image classification used in a variety of projects and can also run a number of sophisticated algorithms like random forest, regression tree, support vector machine, Naive Bayes, and GMO max entropy.

Keywords: Google Earth Engine, methodologies, Digital Earth.

Accepted on October 10, 2021

Introduction

The virtual depiction of the Earth we live on is referred to as Digital Earth. From data to model, it portrays the Earth in the digital world. To create the digital reality, data is collected and models are abstracted. Various sensors used to examine our home planet while developing Digital Earth create massive volumes of data. NASA researchers coined the phrase "big data" to characterise the vast volume of data that exceeds the limits of main memory, local disc, and even distant disc (Friedman 2012). "Big Data is a phrase used to characterise the vast quantity of data in the networked, digital, sensor-laden, information-driven world," according to the National Institute of Standards and Technology (NIST) (Chang and Grady 2015). This term relates to the abundance of digital data in the context of Digital Earth, which focuses on big data's geographical features of social information, Earth observation (EO), sensor observation service (SOS), cyber infrastructure (CI), social media, and commercial information (Guo 2017; Guo et al. 2017; Yang et al. 2017a, bSatellites, sensors, simulation models, mobile phones, utilities, cars, and social networks all capture digital Earth data in various formats, such as images, text, video, sound, geometry, and combinations of these (Yang et al. 2017a, b). Because of the variety of data sources and vast data volume, Digital Earth data are by definition big data.

The growing availability of large Earth data has opened up previously unimagined possibilities for understanding the Earth in the context of the Digital Earth. Big data has been defined by the 5 Vs (volume, variety, velocity, veracity, and value) in recent study (Gantz and Reinsel 2011; Zikopoulos and Barbas 2012; Marr 2015). Firican (2017) expanded the 5 Vs to include variables, validity, vulnerability, volatility, and visualisation in big data characteristics.Satellite and drone-collected remote sensing images easily exceeds the TB and PB thresholds. The Integrated Multi-satellite Retrievals for GPM (IMERG) data package, for example, captures global precipitation data every half hour, yielding up to 3.45 TB of data each year (Huffman et al. 2015). Other sources of location-based information, such as social media (e.g., Twitter) and geographic information systems (e.g., OpenStreetMap), are continually expanding.In the context of Digital Earth, data sources include sensors, digitizers, scanners, numerical models, cell phones, the Internet, videos, emails, and social networks. Data structures, frameworks, indexes, models, management methodologies, and techniques for all sorts of geographic data are all in need of improvement. These geographical data are also represented in a variety of data types, such as vector and raster, structured and unstructured.With the development of advanced techniques such as drone observation for disaster monitoring, the speed with which Earth data is collected and generated has increased. The data creation of IoT nodes, for example, is quick due to the large number of object-based sensors in the IoT. Most sensors continually provide data in real time.The precision of geographical data varies depending on the data source (Li et al. 2016). The quality of remote sensing pictures such as TRMM and IMERG, for example, is determined by sensor setup, calibration procedures, and retrieval algorithms

Even if the data acquired by rain gauges is limited, it is more accurate.Validity is concerned with the accuracy and correctness of Earth data for the intended use, similar to veracity. Data preprocessing, such as data augmentation, interpolation, and outlier detection, play an important role in uncovering information from big Earth data, in addition to data selection, in which data are chosen with appropriate spatial and temporal resolutions and variables for a specific application. The community may benefit from consistent data quality, uniform definitions, and metadata, resulting in high-validity Earth data.In the context of large Earth data, variability refers to the ongoing change in the meaning of data, particularly for data that depends on natural language processing. Twitter data, for example, has emerged as a new source of information for natural disaster management (Yu et al. 2018), since tweets written during catastrophes may be gathered to assist situational awareness. Words' meanings vary over time; for example, the term "Irma" may be a name, but it became to signify the Atlantic's greatest storm in most tweets around October 2017.Because some geographic data contains personal information or is sensitive, security is a concern. Cellular data, for example, has been widely used to assess human behaviours in smart city applications; nevertheless, displaying phone numbers may reveal people's personal information.Volatility relates to the timeliness and freshness of Earth data, or how long the data will be usable and relevant to applications, as well as how long it should be preserved. Due to the pace and amount of large Earth data, storing all of it in a live database without performance difficulties is unfeasible. For data currency, availability, and speedy retrieval, a set of criteria should be created (Firican 2017), for example, historical and less often visited Earth data might be kept on a lower-cost storage tier.Due to poorly scalable, low-functionality, and high-velocity datasets, visualisation of Earth data is a difficult undertaking with limited memory. When displaying geographical vector data, traditional methods may fail to show billions of points, polylines, and polygons; thus, graphical approaches, such as data clustering, parallel coordinates, cone tree, or circular network diagrams, should be employed to represent Earth data (Firican 2017).

References

1. Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. Acm SIGMOD Rec 22(2):207
2. Amirebrahimi S, Rajabifard A, Mendis P et al (2016) A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration. Int J Digit Earth 9(4)
3. Anderson A (2015) Statistics for big data for dummies. John Wiley & Sons, Hoboken,
4. Balakrishna C (2012) Enabling technologies for smart city services and applications. In: 2012 sixth International conference on next generation mobile applications, services and technologies. IEEE, Paris, France, 12–14 September 2012
5. Baumann P, Mazzetti P, Ungar J et al (2016) Big data analytics for earth sciences: the earthserver approach. Int J Digit Earth 9(1):3–29
6. Bereta K, Caumont H, Daniels U et al (2019) The copernicus app lab project: easy access to copernicus data. In: EDBT. pp 501–511

***Correspondence to:**

Ramla Khan

Phd student

Open University

UK