Big Open Data for Environmental Information Systems

Donald Cowan, Paulo Alencar
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Email: dcowan/palencar@csg.uwaterloo.ca
Fred McGarry
Centre for Community Mapping
Waterloo, Ontario Canada N2L 2R5
Email: mcgarry@comap.ca
R. MarkPalmer, Trevor Boston, Rigel Rozanski
Greenland International Consulting Ltd.,
Collingwood, Ontario, Canada L9Y 1V5
Email: mpalmer/tboston/rrozanski@grnland.com

Abstract

The frequency of extreme weather events has accelerated, an apparent outcome of progressive climate change. Excess water is a significant consequence of these events and is now the leading cause of insurance claims for infrastructure and property damage. Governments recognize that plans for growth must reflect communities needs, strengths and opportunities while balancing the cumulative effects of economic growth with environmental concerns. For such a process to be effective it will be necessary to develop and operate cumulative effect decision support software (CEDSS) tools, and to work closely with all levels of government including watershed management authorities (WMAs) that supply environmental data. Such cooperation and sharing will require a new open and big data information-sharing platform, which is described in this paper as an open and big data platform for environmental analysis and management.

Keywords

open movement, big data, open data, environmental information systems

I. Introduction

The frequency of extreme weather events has accelerated over the last decade. Although excess water is one of the significant impacts of these weather events and is the leading cause of insurance claims for infrastructure and property damage [1], [2], lack of water can be just as problematic. Prolonged periods of drought can limit our access to drinking water and impair agricultural production, while also changing the landscape ecosystem, encouraging invasive species and putting terrestrial species at risk.

Economic development involving human activities such as construction of buildings and infrastructure, and resource development modifies the landscape and its response to extreme weather. Economic development is ***cumulative***, as changes in one area will at a minimum impact adjacent areas if not further afield. Thus, planning for adaptation to extreme weather events must recognize these ***cumulative effects*** [3], [4].

Thus, the spectre of extreme weather and cumulative effects has accelerated the need to build environmental information systems that predict the cumulative impact of development and then monitor the results to determine variances that might occur. Modelling supports the notion that based on past data we can predict the impact of future human activity on the landscape. In contrast, monitoring evaluates the current condition of the environment and supports alerts and alarms to detrimental effects and verification of modelling predictions.

Modelling requires a large amount of historical data on which to base predictions, while monitoring can also generate a huge amount of data, which can be and is used to support modelling. Where does this data originate and how can it be structured so as to be accessible to the various environmental tools that are being developed for both modelling and monitoring?

This paper explores this topic by determining the nature of environmental data and creating an architecture for big open environmental data.

II. Big and Open Data

Why big and open data for environmental modelling and monitoring? The data used in modelling is gathered through monitoring, so we need to look at this type of data.

Environmental monitoring gathers many different types of information such as:

- rainfall,
- land contours and floodplain mapping,
- land use,
- soil permeability, and
- water flow rate.

This information comes from many different sources such as federal, provincial, state or municipal governments and non-governmental organizations (NGOs) such as watershed management authorities (WMAs). The data is also gathered by many different means including humans, terrestrial, aircraft-based and satellite-based sensors. The latter two often using technologies such as digital photography or Lidar.[1]

Each organization captures data using different approaches. Typically federal authorities capture weather data, while provinces and states capture geographic information such as land contours. Floodplain mapping, land use and soil permeability often come from municipalities or WMAs. For example, a municipality will share jurisdiction of a land mass with one or more WMAs and each organization will have control over the use of land. For example, the municipality may allow paved parking lots, while each WMA may only allow permeable parking lots.

Some of the data such as that collected for weather is obviously big data because of the frequency of collection and its geographic extent. Other data such as that for geographic form is also big data but is not collected as frequently, as the land form is usually only altered by human activity such as development or extreme weather.

Environmental data has also become open data as all levels of governments and related agencies attempt to become more transparent in their interactions with business and the public. Open data is based on the concept that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. However, republishing does imply citing the original source not only to give credit but to ensure that the data has not been modified or results misrepresented.

Managing and manipulating big and open data presents challenges to any information system. Typically big data is an evolving term that focuses on three highs: volume, velocity and variety, while open data is usually published as flat files by the agencies that make it accessible. Thus, the release of big and open data raises numerous issues that need to be addressed. Some key questions are:

- How do you find big and open data that has been released to public view?
- How can big and open data be prepared so that it can be easily accessed?
- How do you develop applications so that viewing, comparison and use of big and open data can be made accessible to the broader public?
- How do governmental and non-governmental organizations with limited resources decide what data to make open and how to maintain it, once it is public?
- Will new technological approaches be required to access and use big and open data?

The next section looks at big and open data in a specific domain namely environmental modelling and monitoring and describes an architecture that should address many of these problems.

III. Data Architecture

Developing an architecture that can house big and open data is a problem and there are many attempts particulary with big data to define such an architecture and tools such as Hadoop [5], MapReduce [6] and Neo4J [7]. In contrast there seems to be little work on this problem for open data. The kind of tools that have been developed for big data focus on a certain class of problem [8] characterized by the data rather than more directly on problems related to domains of science and engineering such as the environment.

A. Converting Environmental Big Data to Manageable Data

Fortunately the terrestrial environment domain has an inherent structure that provides a direction for the data architecture. The world is naturally divided into watersheds[2] and terrestrial environmental models usually predict behaviour in a watershed. If a

[1]Lidar is a remote sensing technology that measures distance by illuminating a target with a laser and analyzing the reflected light. Lidar is used to make high-resolution digital elevation maps.

[2]A watershed is defined as an area of land where all the surface water drains into the same place, whether it's a creek, a stream, a river or an ocean. Therefore, all precipitation, such as rain or snow, that falls on a watershed ends up flowing to the same place. A watershed may be open or closed, in an open watershed all water eventually drains into the ocean, whereas in a closed watershed the water escapes through evaporation or seeping into the earth [9]

model spans two or more watersheds then the model must be run twice, once for each watershed, since although the water may flow into the same lake or ocean the one watershed does not directly impact the other.

Thus, our data architecture can initially be based on geography where data is organized by watershed. All the different data required for a model or gathered from monitoring can be attached to the geography of a watershed or sub-watershed. The actual granularity and size of the data set will determine the size of the watershed that is used.

Of course watersheds may cross political boundaries and environmental rules and regulations can be different. Thus, the architecture must be divided by political boundaries with watersheds under those boundaries. If watershed crosses a boundary then it will appear at least twice and the data about the watershed must be assembled although it could probably be pre-assembled with an indication of the jurisdictional source of the data.

The structure of the data model would start with the world and then be subdivided into countries, which are further subdivided into provinces or states. Each province or state contains watersheds where a watershed may appear more than once because it is cross-jurisdictional. Data can easily be found using a map interface to indicate the watershed of interest and the source of the data where the watersheds are indexed by geography for efficiency of access.

By using the concept of a watershed as a way of organizing environmental data it has become possible to convert a big data problem into one that is manageable and further where the data should be able to be stored and accessed through an SQL database. This approach has several advantages including the fact that the level of abstraction for manipulation of data is much higher than that supported by current big data structures [10].

B. Open Data - Managing Data from Multiple Sources

Why should the data architecture contain a provision for open data? Currently the private sector works with data from government and WMAs, so why is there a need for an Open Data platform to assess environmental impacts and monitor the outcomes? To answer this question we should first examine the process that is currently followed in modeling and discover its deficiencies. Then we can illustrate a new form of cooperation between the public and private sector that will be far more effective in addressing issues related to the interactions between the cumulative effects of economic growth and the environment.

How is environmental impact analysis conducted currently? The following steps outline the process for a specific geographic area.

1) Identify the modelling procedures that are required.
2) Find or develop the modelling tools to implement the procedures.
3) Acquire the necessary mapping and environmental data from all the different government agencies and WMAs for the analysis and simulation.
4) Do the simulation and analysis, and prepare a report.

What happens to the tools and data that have been gathered and generated for the analysis and simulation relating to a specific land use change, development or infrastructure plan? Typically a report is produced and the data used to generate the report are stored in a computer system or a desk drawer. The next time a similar analysis or simulation is required the process is repeated, all the data are assembled again and the raw results of the original analysis and simulation are not accessible. Thinking of the cost of repeating all these operations, it is clear there must be a better way. In addition, because the results of previous simulation or analysis are not readily available it is impossible to show the cumulative effects when projects changing the environmental landscape build upon one another.

If we could retain and keep the open data current then subsequent simulations and analyses would be simpler and much less expensive. There a number of steps that need to be undertaken to make the platform amenable to using open data.

Once the data is identified, collected and structured for the first time it must be kept current. The currency of data is dependent on what it measures and so may even have a “best before date.” For example, weather data must be updated quite frequently perhaps daily, while land shape and soil permeability will not change significantly over a period of years and may only need to be examined annually.

However, no matter what the time is between updates, keeping current will need to be automated. Scripts will have to be produced that are activated on a variable schedule to check if data needs to be updated and perform the update when necessary.

Further, the production of such scripts will have to be automated as both the open data produced by a government agency or NGO could change form or location or the receiving data structures may also change over time. Software will have to be built that “walks” over both source data and the destination data structures and automatically creates the scripts.

IV. Conclusions

Assessing the impact of human activity on the terrestrial environment requires access to both big and open data. Big in that many aspects of the environment such as weather produce large amounts of data and open as holders of environmental data such as governments and watershed management authorities (WMAs) are making it accessible to the general public.

Although access to this data is supporting environmental modelling and monitoring the data has to be structured and managed carefully. Fortunately the terrestrial environmental domain is structured into watersheds thus allowing the big data problem to be converted to a “smaller” data problem and making this data manageable by normal SQL database structures.

However, dealing with open data is far more complex than appears on the surface. Data will need to be combined from all levels of government and with data from NGOs, businesses and other sources such as university research labs. Maps will also be important as most open data has a geo-spatial component. There are many issues relating to data sustainability/maintainability and privacy that are very important.

This paper is an attempt to outline issues associated with big and open data in the environmental domain and to suggest a direction toward a solution.

ACKNOWLEDGMENTS

The authors thank the Ontario Research Fund and FedDev Ontario for support in developing the concepts described in this paper. The authors also thank Doug Mulholland and Anthony Robins, members of the University of Waterloo Computer Systems Group for their work in implementing the software that tested the initial ideas on big and open data.

REFERENCES

[1] Insurance Bureau of Canada, "The financial management of flood risk," http://assets.ibc.ca/Documents/Natural%20Disasters/The_Financial_Management_of_Flood_Risk.pdf, May 2014.

[2] KPMG, "Water Damage Risk and Canadian Property Insurance Pricing," http://www.cia-ica.ca/docs/default-source/2014/214020e.pdf, February 2014.

[3] Evan R. Ross, "The Cumulative Impacts of Climate Change and Land Use Change on Water Quantity and Quality in the Narragansett Bay Watershed," http://scholarworks.umass.edu/masters_theses_2/111/, May 2014.

[4] The Resource Innovation Group, "Toward a Resilient Watershed," http://www.theresourceinnovationgroup.org/storage/watershed-guide/Watershed%20Guidebook%20final%20LR.pdf, January 2012.

[5] T. White, *Hadoop: The Definitive Guide*, 3rd ed. O'Reilly Media/Yahoo Press, May 2012.

[6] J. Lin and C. Dyer, *Data-Intensive Text Processing with MapReduce*. Morgan & Claypool Publishers, 2010.

[7] I. Robinson, J. Webber, and E. Eifrém, *Graph Databases*, 2nd ed. O'Reilly, 2015.

[8] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2012.

[9] Canadian Geographic, "Watershed 101," http://www.canadiangeographic.ca/watersheds/map/?path=english/watershed101, 2011.

[10] M. L. Braun, "Levels of abstractions in big data," http://blog.mikiobraun.de/2012/09/big-data-abstraction-dsl.html, September 2012.