ISBN 978-93-84422-79-0
9th International Conference on Recent Trends in Science Engineering, Computers and Technology (RTSECT-2017)
Singapore Aug. 10-11, 2017

Role of Big Data in Environmental Sustainability

Gajendra Sharma and Roop S.R. Bajracharya
Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal
Email: gajendra.sharma@ku.edu.np

Abstract: *Big Data has emerged as one of the most promising and game-changing entities in commerce, society, and politics. This paper looks into the role of Big Data in environmental sustainability by using the globally available, national datasets that measure a multitude of environmental issues, ranging from smart city management to air quality and forests. The presented literature and immense future possibilities clearly state the important role that big data plays in environmental sustainability. As the volume of data collected in big data is enormous, conventional methods of data analysis cannot be used to process and look-up the data. Big data technologies and frameworks like Hadoop, MapReduce and Hive are used to store, fetch and analyze the data.*

Keywords: *Big data, environmental sustainability, hadoop*

1. Introduction

1.1 What Is Big Data?

Big Data refers to extremely large amounts of data collected from various sources such as sensors and census that enable the use of this massive amount of data to reveal patterns, trends, and associations. Big data has been termed as, “the next frontier for innovation, competition, and productivity” (Manyika, 2011). McKinsey Global Institute also explained that big data classification changes not only over time as a result of technology enhancements, but also across different sectors due to the variation in dataset size, software, and tools that are common to each sector (Manyika, 2011).

Big data is characterized by traits known as the “four V’s” - volume, variety, velocity and veracity(IBM, 2014). Volume refers to the size of the data set that requires large amounts of data storage and computational power. Variety refers to the various numerous types of data. Velocity refers to the analysis of streaming data coming from real-time. Lastly, veracity refers to the uncertainty of this huge amount of data. Big data has the potential to enhance decision-making (Dumbill, 2013) and with the increases in the capabilities and power of technologies like cell phones, computers, and various sensor systems the amount of information now available, combined with the rapid growth of that information (Mayer-Schönberger & Cukier, 2013) has shifted a huge amount of focus to big data and its applications.

1.2. Big Data Applications

Many sectors like healthcare, retail, politics, and business have taken the advantage of big data to predict and yield large amounts of profits (Mayer-Schönberger & Cukier, 2013). Big Data has emerged as a game-changing presence in commerce and politics. Private sector companies are pushing the limits of Big Data for their targeted solutions. It is also stated that with further progress in big data applications can lead to, “increase sector-wide productivity by at least 0.5 per cent a year through 2020,” (Manyika, 2011). Big data has an enormous potential in all these fields.

https://doi.org/10.15242/DiRPUB.DIRH0817221

1.3.Big Data and Environment

Environmental sustainability is not yet part of the popular big data practices just yet. Big data is playing a transformative role in sectors such as retail, manufacturing, and healthcare but the environmental sustainability effort however requires much work and improvements. Furthermore, the health of ecosystems is in decline as indicated by the Millennium Ecosystem Assessment (Assessment, 2005). Big data could generate new forms of analytics and insight for problems regarding environmental sustainability. This paper looks into how big data is being applied toward the environmental sustainability effort. We also study to what extent big data is impacting environmental sustainability while other sectors are experiencing revolutionizing effects.

Over the last decade, an increasing number of environmental issues have required the analysis of data from hundreds of thousands of different locations, including the regional, national and global monitoring networks. As a result of the massive volumes and a wide variety of data types is now available to the scientists. Solutions to these kinds of problems increasingly require the tools and techniques of big data. Few of the current such applications like energy management in smart grids as mentioned by (D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015)and analysis of the efficiency in deploying the low-carbon vehicles (Gennaro, Paffumi, & Martini, 2015)are discussed in this paper.

2. Literature Review

As the big data technology and its associated fields continue to evolve, practitioners and researchers have yet to find much application of big data in the field of environmental sustainability. As a result, it is difficult for researchers to perform meaningful cross-study comparisons and build on the outcomes from the previous studies. The relevant literature to determine the current condition of big data application in the environmental field is studied in this paper.

(Keeso, 2014)has concluded that organizations like CI, WRI, BT, Anthesis, and the US government are making big data as a part of their work and efforts toward environmental sustainability. Studies like these show how the world is slowly progressing towards the future where big data can be used to analyze the enormous environmental data that is available to work for a better environment. (D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015)show the use of big data in energy management in smart grids for proper energy management which leads to a reduction in wastage of energy. They have used methods like dimensionality reduction as shown in (A. Dieb Martins, 2013) to reduce the massive amount of data obtained from the smart meters.Similarly, (Gennaro, Paffumi, & Martini, 2015)investigates the activity patterns of large-scale samples of fuel vehicles data to understand the mobility in urbanized areas to analyze the efficiency in deploying the low-carbon vehicles.

Although big data has a huge potential in environmental sustainability the present condition of this field is limited only to a few of the research projects. Through further studies and advancement in technology the environment field is also sure to gain many benefits from big data like other fields.

3. Big Data Use Cases

Environmental big data refers to the massive volumes and a wide variety of data types available regarding water supply lines, wind directions and monitoring data. We now see various ongoing projects and possible application of big data in the field of environmental sustainability.

3.1 Analyzing the Road Travel Data to Enforce Low-Carbon Transport Policies

(Gennaro, Paffumi, & Martini, 2015) provides an overview of the applications of a big data in the field of road transport policies in Europe. They use the datasets of driving and mobility patterns collected by means of navigation systems and onboard GPS systems to perform develop its core algorithms. (Gennaro, Paffumi, & Martini, 2015) analyzed large-scale mobility statistics to look into the potential of electric vehicles in replacing conventional fuel vehicles and related modal shift and also analyses the energy demand coming from electric

https://doi.org/10.15242/DiRPUB.DIRH0817221

vehicles. This paper shows how big data can inspire smart policies to enable the large-scale deployment of the next generation of green vehicles, offering an unprecedented opportunity to shape policies for future mobility and smart cities.

The results of the data processing platform designed for supporting EU transport policies via big data are presented in (Gennaro, Paffumi, & Martini, 2015). The platform is natively conceived for multi-purpose applications, and it is made of 6 modules: pre-processor, statistical mobility module, modal shift and vehicle usability module, energy demand module, infrastructure design module, vehicle to grid applications and gaseous emission module. It has the objective to investigate the activity patterns of large scale samples of conventional fuel vehicles or GPS data to understand the mobility in urbanized areas, evaluating the potential of deploying low-carbon vehicles, and deriving the modal-shift introduced by these technologies. This paper shows the effective development of a methodology being able to handle large amount of data, to perform customized analysis of different kind and identify non-obvious relations among large datasets using big data.

3.2 Dynamic Energy Management

(D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015) shows the use of big data techniques to use the large volumes of data generated by the vast amount of smart meters to optimize the economic efficiency, reliability and sustainability of power flow in smart electricity grid. These kinds of implementation help in preventing the energy loss through efficient energy management.

Most of the currently available power grid systems focus on modeling of traditional network components, i.e., the generation systems, loads, and transmission network. But a distribution grid test-bed which can be used to test the designs of integrated in-formation management systems has been proposed in(N. Lu, 2011) . This test-bed represents the correlation and interdependency among data sets, aiming to efficiently monitor the status of the smart grid and detect abnormalities. Interestingly enough, extensive sets of smart grid's detailed trial data, which can be used in order to test the designed schemes, can be easily acquired, thus facilitating the research in this area. The registered users can also access the public model, including key functions, assumptions, and analytical tools.

Storing and processing the huge amount of data generated by the smart meters, requires improved platforms, appropriate for big data analytics, such as Hadoop, Cassandra, and Hive (M. Mayilvaganan, 2013). Hadoop is a promising platform for the distributed processing of large smart grid's data sets. It is s a collection of open source tools and includes the concept of MapReduce. Cassandra database, which supports the cloud infrastructure, can be used in order to store the large data sets. Also, Hive data warehouse software, which uses a simple SQL-like language, can be used to query datasets that are stored in a distributed environment with ease.

The smart electricity grid enables flow of power and data between suppliers and consumers in order to facilitate the power flow optimization. By using the big data methods for real-time exploitation of large volumes of data generated by the vast amount of smart meters the smart grid system can be made more energy efficient. Hence, robust data analytics, high performance computing, efficient data network management, and cloud computing techniques are critical towards the optimized operation of smart grids (D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015).

3.3 Effect of Human Activity on Groundwater Quality

Big data can be used to evaluate the potential effect of human activity on groundwater quality (Rominger, 2015). This would require gathering information on massive monitoring well networks. This kind of voluminous data gives rise to different storage and computation power related problems. Also the variety of data types and the need to link these data together is an ongoing challenge for the scientists.(Rominger, 2015) also introduces the use of Geographic Information System (GIS) database which could be an efficient and cost effective solution to organize and assess huge amount of environmentally generated data. But "GIS databases are only as good as the information on which they are based" (Rominger, 2015).

3.4 Other Possibilities

(Allouche, 2017) lays out various possibilities of application of big data for the environmental benefits. Deforestation is one of the greatest concerns for the environment. Big data provides alternative solutions deforestation lowering the carbon footprint and decreasing the negative impact on the ecosystem. By implementing big data, scientists can effectively monitor plant and animal species in danger of extinction through collection of related information from numerous sources such as satellite data, animal population and even social media. Using such information plans can be created early enough to fight the problem and preserve the species (Allouche, 2017).

Poaching is another serious environmental issue. With Big Data, however, their reach is significantly expanded. Trouble areas and animals can be pinpointed and actions taken to prevent poaching.

Big data can also be used to detect the broken and ineffective sprinklers that cause problems like overwatering and water wastage. Data incoming from different sensors embedded into these sprinklers can be used to detect and solve residential water waste. "Big Data is essential in order to forge ahead and provide our progenitors with a sustainable future." (Allouche, 2017)

4. Implementation

An approach identical to (J. & E.A.Mary, 2015)can be used to implement the application of big data for the environmental data obtained from numerous sources. We look into the basic architecture of Hadoop which can be used for the analyzing the collected environmental data to produce any helpful results. In this section we also look at how the data is collected, stored and analyzed using Hadoop.

4.1. Big Data Ecosystem for Environment

All the environmental data gathered through the various processes are heterogeneous which don't make much sense while looking at holistically. These data need to be processed for further analysis and to yield any useful information.

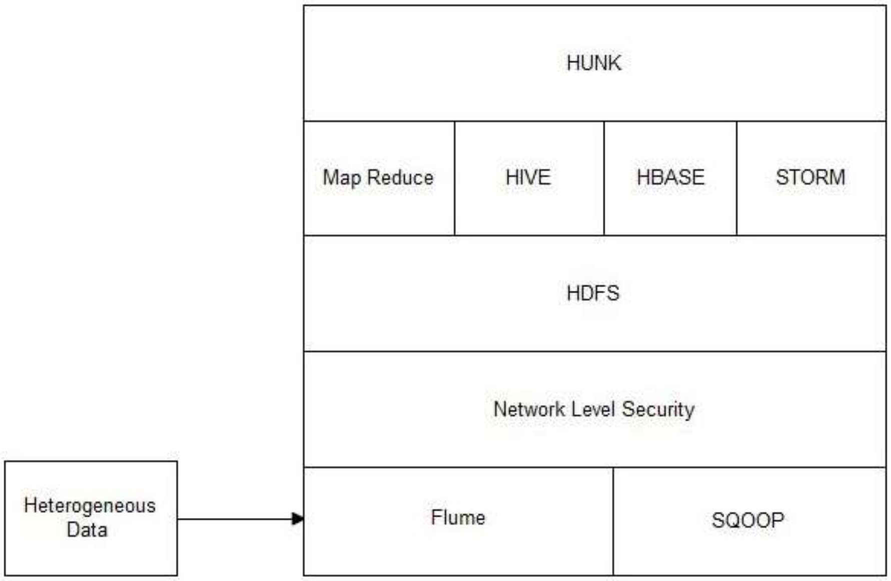

Fig.1: Big Data Architecture for Environmental Data

Big Data Ecosystem is a complex system that constitutes of components and technologies to handle large-scale data processing and analytics on it as shown in (Shvachko, Kuang, Radia, & Chansler, 2010). It includes getting the data from various sources, store them in HDFS (Hadoop Distributed File System), process the data using Hadoop components such as Map-Reduce, perform analysis using PIG and generate Business Intelligence reports.As shown in figure 1 the data is given as input to HDFS through flume (a service for supplying log data into Hadoop) and SQOOP (tool to transfer bulk data between Hadoop and Datastores).Analysis of the data is performed using Map-Reduce and HIVE by implementing machine-learning algorithms which help in analyzing the similar pattern of data.

This helps in predicting the various patterns and relevant information from the available data. So produced results can be stored using Hbase which is used for storing the multi-structured data. STORM is used to perform live streaming of real-time data such as weather data.

Finally the report is generated through HUNK (a tool to make data more usable, accessible for the people who need at the speed they need) that provides powerful user interface and dashboard for scientists to explore, analyze and visualize datafrom Hadoopas mentioned in (Shvachko, Kuang, Radia, & Chansler, 2010)

4.2.Big Data Lifecycle

(J. & E.A.Mary, 2015) has proposed the big data lifecycle to formulate the flow of data from the point of its collection to the delivery phase. An identical approach can be used in the case to process the environmental data.

4.2.1 Data Collection

This phase involves collecting relevant data from different sources such as sensors and satellite imagery and storing it in the HDFS. Every big data application relies on a set of data, and the correct pre-processing of the data to harness the data potential and application. In order to develop an effective pre-processing methodology two large datasets of driving and mobility patterns collected by means of GPS have been analyzed as pilot study in (Gennaro, Paffumi, & Martini, 2015). The driving pattern databases for the Italian provinces of Modena and Firenze have been acquired from a private company Octo Telematics.

Similarly, data generated from Energy management systems (EMSs) in SGs which include data from real-time wide-area situational awareness (WASA) of grid status through advanced metering and monitoring systems, consumers' participation through home EMSs, demand response algorithms, and supervisory control through computer-based systems mentioned in (K. le Zhou, 2013) is used in (D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015).

4.2.2 Data Cleaning

These enormous amounts of data collected through heterogeneous sources are very clustered and full of junk data. These data need to be removed. (Gennaro, Paffumi, & Martini, 2015)applied a double filter to the raw data in order to delete the vehicles driven for more than 50% of the trips outside the province borders and all the trips with a length less than 30 meters and/or duration less than 30seconds.

As shown in (D.Diamantoulakisb, M.Kapinasb, & K.Karagiannidisa, 2015), smart meters generate large volume of data.Processing this amount of data is inefficient in terms of communication cost, computing complexity, and data storage resources utilization.Methods like dimensionality reduction has been applied in(A. Dieb Martins, 2013), can be used to provide a reduced version of meters' original data via random projection. It is shown that processing the produced summarized version of data instead of the original stream of data leads to an acceptable relative error. This will have the advantageof scalability, complexity reduction, and increase in execution speed.

4.2.3 Data Classification

This phase involves classifying the data based on their structure. (Gennaro, Paffumi, & Martini, 2015)has their transportation data sub aggregated by day, week and month to produce better results from the data.

4.2.4 Data Modeling

In this step the classified data is analyzed upon. The data is classified based on location, geography, type, etc. The traditional centralized frameworks for acquiring, analyzing and processing data require huge exchange of information among the remote like smart grids and centralized processors which is inefficient in terms of telecommunication resources management and economic cost. To this end, the authors in (R. Mallik, 2011) present several distributed data analysis techniques that can be successively used for energy demand prediction.

4.2.5 Data Delivery

This step includes generating of reports based on the data modeling. Tools like Hive data warehouse can be used with a simple SQL-like language, to query datasets that are stored in a distributed environment as shown in (M. Mayilvaganan, 2013).

5. Conclusion

As the problem of environmental degradation begins to rise, we must work towards conserving the environment. The rise in big data technologies has made much progress in fields like healthcare, politics, social media as well as weather forecasting. However environmental sector has yet to take benefits from big data. As mentioned in this paper we see how innovative use of big data in smart grid management and low carbon policy enforcement is implemented. Although slow there is a definite progress in use of big data in this sector.

The problem is not the lack of data but the lack of information that can be used to support decision-making, planning and strategy. The environmental sector can heavily benefit from utilizing big data technologies. To successfully identify and implement big data solutions and benefit from the value that big data can bring scientists and people need to devote time, allocate budget and resources to visioning and planning. With the help of open source platforms like Hadoop these goals are within arm's reach.

6. References

[1] A. Dieb Martins, E. G. (2013). Processing of smart meters data based on random projections. *IEEE PES Conference on Innovative Smart Grid Technolo-gies Latin America, ISGT lA*, 1-4.

[2] Allouche, G. (2017). 5 Ways Big Data Can Protect Our Planet. *https://datafloq.com/read/5-ways-big-data-protect-planet/82*.

[3] Assessment, M. E. (2005). *Ecosystems and Human Well-Being: Synthesis.* Washington DC: Island Press.

[4] D.Diamantoulakisb, P., M.Kapinasb, V., & K.Karagiannidisa, G. (2015). Big Data Analysis for Dynamic Energy Management in Smart Grids. *http://dx.doi.org/10.1016/j.bdr.2015.03.003*.

[5] Dumbill, E. (2013). Making Sense of Big Data. Big Data. Vol. 1, No. 1. *Mary Ann Liebert, Inc.*

[6] Gennaro, M. D., Paffumi, E., & Martini, G. (2015). Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities. *http://dx.doi.org/10.1016/j.bdr.2016.04.003*.

[7] IBM, C. (2014). The Four V's of Big Data. Accessed on July 20, 2015. *http://www.ibmbigdatahub.com/infographic/four-vs-big-data*.

[8] J., A., & E.A.Mary, A. (2015). A Survey Of Big Data Analytics in Healthcare and Government. *2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)*.

[9] K. le Zhou, S. l. (2013). A review of electric load classification in smart grid environment. *Renew. Sustain. Energy Rev. 24*, 103-110.

[10] Keeso, A. (2014). Big Data and Environmental Sustainability: A Conversation Starter. *Smith School Working Paper Series*.

[11] M. Mayilvaganan, M. S. (2013). A cloud-based architecture for big-data analytics in smart grid. *IEEE International Conference on Computa-tional Intelligence and Computing Research, ICCIC*, 1-4.

[12] Manyika, J. (2011). Big Data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.

[13] Mayer-Schönberger, V., & Cukier, K. (2013). *"Big Data: A Revolution that will Transform how we Live, Work and Think".* John Murray. London, UK.

[14] N. Lu, P. D. (2011). The development of a smart distribution gridtestbed for integrated information management systems. *IEEE Power and Energy Society General Meeting*, 1-8.

[15] R. Mallik, N. S. (2011). Distributed data mining for sustainable smart grids. *ACM SustKDD'11*, 1-6.

[16] Rominger, J. (2015). The Role of Big Data in Solving Environmental Problems. *Gradient Trends: Risk, Science and Application Issue 64*.

[17] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File. *IEEE*.