LARGE SCALE SIMULATION IN BIG DATA

Priyanka Gupta[1], Mahvish Jabeen[2]
[1]Assistant Professor, Shri Ramswaroop Memorial University, Lucknow
[2] Assistant Professor, Shri Ramswaroop Memorial University, Lucknow

ABSTRACT

The paper lays emphasis on the emerging need of simulation to handle the ever increasing data which is generated all over the globe. With the rise of technology, a digital revolution gas taken place. All the industries, varying from hospitals to small business to social media websites, are working on the concept of sending data to database, further on to servers, and getting back the response from the server that data has been processed either successfully or unsuccessfully through the reflected response message. This data is generated at an immense rate per second. And the best part, ironically, is that all the data has to be permanently stored so as to retrieve it from anywhere and at anytime. On the other hand, the worst part, again ironically, is that to handle the enormous data, a high tech server or the high tech workplace has to be generated. For the handling of this data, the concept of Big Data has come in limelight. The Big Data has rapidly developed and achieved success in almost all the working fields. It helps to synchronize the data sequentially as the action or the processing has been done. The implementation of Big Data has to be done through simulations, the concept of Artificial Intelligence. So, it's evident that handling data storage has migrated from manual processing to artificial intelligence. This paper aims to study the simulation of Big Data at a large scale and how it can help to store the enormous data virtually effectively.

Keywords: *Big Data, Digitalization, Simulation, Technology.*

1.0. Introduction

The concept of Big Data is emerging and paving its way through technological era. It is useful because it serves many purposes related with data storage. There are around 7.6 billion people all over the world (as of December 2017) who are surfing over 1 billion websites (as of October 2014, confirmed by NetCraft). From an individual to a large organization, there are "n" number of active users who play with the data. Every second, "n" number of data is created, modified, deleted or simply accessed. This is possible because of the huge servers, which is usually compilation of a large number of databases.

According to American Express, there is a virtual communication between a system and the server through the mode of request-response. A request from the user is sent to the server, server processes the request from the database, and then a response is generated from database to server, which is forwarded to the user in the readable format. Everytime, a data or a request is sent, it attaches itself with a unique value that is generated at the moment, just like OTP (One Time Password). These data may be in any format, so it is converted into structures (which include queues, dequeues, stacks, circular queues). This data, thus is permanently stored in database. The database can be understood as a data centre or a centre repository where all the generated data is compiled in its understandable format.

2.0. Literature Review

Manish Parashar in his work concluded that complex applications which on high-end computers generate a large amount of data which needs to be managed as well as analysed so as to get proper insights. The data costs including performance, energy, latency are dominating. Traditional method of data management or analytics pipelines is breaking down. Parashar also mentioned that there are many challenges for the Big Data such as programming, scheduling and mapping, control, automatic management of runtime. Some of the solution that he proposed, include hybrid data staging, dynamic code deployment or in-situ workflow execution,

Kangsun Lee and Joonho Park suggested, in their experiments, ARLS (After action Reviewer for Large-scale Simulation data) improved data processing time significantly comparing to the traditional output analysis tools.

Xiao Song, Yulin Wu, Yaoefi Ma, Yong Cui and Guanghong Gong suggested that Big Data can be studied and understanding complicated problems like acquisition of weapon systems, combat analytics, as well as military training. The researchers also reviewed that various military simulations can be used for producing the data to be used for varying purposes, at a large-scale.

Benoit Lange and Toan Nguyen introduced VELaSSCo project. They mentioned that simulations produce exponentially growing volumes of data, and it is not possible to store them anymore with existing IT systems. Therefore, VELaSSCo aimed to develop new concepts for integrated end-user visual analysis with advanced management and post-processing algorithms for engineering applications, dedicated to scalable, real-time and petabyte level simulation.

Shengcheng Yuan, Yi Liu, Gangqiao Wang and Hui Zhang focused on microscopic traffic simulation. They proposed a method of cross-simulation which can be applied to the data collected, usually, in normal circumstances into a large-scale traffic evacuations which provides with a better supporting insight for decision makers instantly.

3.0. Objectives of the study

- To create deep understanding of big data in large scale industries.
- To understand how the simulation can work best for Big Data.

4.0. What is Big Data?

Let’s take a simple example. You go to a hotel. You order the waiter for some food, say Chowmein and Manchurian. The waiter takes the same order to the kitchen. The kitchen receives the order and processes the food. The kitchen gives the food to the waiter. And then your order is served at your table. This the simple way in which request-response communication has taken place.

Now, let’s understand how data is generated. Let us take the example of Facebook, a social media channel. There are over 2.07 billion monthly active Facebook users all over the world, as of 1st November, 2017. There are 1.15 billion mobile daily users and 1.37 billion daily users who log onto Facebook. Every 60 seconds on Facebook: 510,000 comments are posted, 293,600 statuses are updated and 136,000 photos are uploaded (Source: The Social Skinny). Just imagine the quantity of data that is generated per second, per minute, per hour, per day, per week, per month and per year. It’s enormous, simply enormous. If we talk about other social media channels and their daily users, Twitter has 284 million active users, WhatsApp has 500 million active users, Instagram has 600 million active users, and so on and so forth.

So, to handle this enormous data, there was a requirement of something more developed and efficient method than the traditional methods used for the data storage. Earlier, the data generated was confined to some GBs (gigabytes) only. But now, there are millions of (TB) terabytes data that is generated. Moreover, it has to be processed instantly. So, the concept of Big Data came into play.

Data scientists that operate in Big Data are using N-V, that is, volume, velocity, variety, etc.

Volume. If size of the Big Data has to be defined, it is simply impossible. It differs according to various fields and moreover, it is expanding over the time. There are different big companies which are generating terabytes, which is a huge amount, of fresh data on a daily basis. It is estimated that the database which is being used may also exceed petabyte. Therefore, huge amount of volume stands as one of basic properties in Big Data. In accordance with International Data Corporation or IDC, volume of the digital data generated in the world may reach even zetabytes which amounts to 2^{30} terabytes. Not much to a surprise, it is doubling within the span of two years.

Velocity. The trending term “data explosion” refers to the speed of data generation. It is quick and needs to be processed from time to time. As for E-commerce applications, processed information need to be available in fraction of seconds; else value of the data will be lost.

Variety. With great volume of Big Data, it has various types as well as formats. Traditionally, structured data which was saved in the database had regular format, like, time, date, string and amount, whereas unstructured data, on the contrary, possess formats like image, audio, text, video. This unstructured data also can be blog, web page, photo, or comment on some commodity, filling questionnaire, collection of sensor data, application log, etc. So, unstructured data is used to express all the human-readable materials but, at the same time, cannot be understood by machines. Generally, these are stored in some file system or rather NoSQL (also known as Not Only SQL) database. This has data model which is simple in nature like Key-Value Pair. Also, variety in the data refers that there are various sources of Big Data. For instance, data for the traffic analysis might come from a fixed networked camera, taxi, bus, or any subway sensors.

Value. It is more than obvious that Big Data produces valuable information for its owner. This information helps in forecasting future, create fresh chance, as well as reducing risk or any cost involved. Thus, Big Data helps to change and improve the life of people.

Veracity. The term "veracity" means using the only trustworthy data, else decision maker might make improper decisions with the incorrect knowledge. For instance, online review of a commodity from the customer is essential for the ranking system, and suppose, if few give fake comments purposely for mere profit, results will be affected negatively and so the ranking too. Veracity, thus, helps to detect that fake data and remove it before analysis.

5.0. Simulation in Big Data

Big Data is the concept that is "Beyond Data". From simple records of criminals to GPS facility in mobile to Google Translate, each require the use of Big Data. In short, it has now become a part of daily life of all the fields. To process the Big Data, simulation, another concept derived from artificial intelligence, is required. Simulation helps to recreate the data wherever and whenever asked for, from the server or the database indeed. It is useful in decision making but is somewhat difficult to be applied by layman.

Simulation model systems are very simple agents and their interaction varies from small numbers to large numbers. It provides the "right" information to make better decisions: predict and how to influence

If we conduct simulation in "traffic and infrastructure system with reference to business", which is considered as one of the large-scale simulations, simulation is confined to three basic questions:

1. What is the best preparation for a major event?
2. How do changing travel pattern affect businesses?
3. Which incentives give young people greater access to the housing market?

From the viewpoint of data lifecycle, the simulation process can be divided into three consecutive phases: data generation, data management, and data analysis. Data generation is concerned with what kinds of data should be created and how to create valid data in the given amount of time. Data management is concerned with collecting huge volume of data and that too without disturbing the process of normal simulation. It provides a good amount of storage as well as its processing capability is efficient. Data analysis uses different analytic methods so as to get value from result of simulation.

6.0. Challenges with Simulation in Big Data

The biggest problem with Big Data is it's beyond the understanding of humans. Day by day, Big Data is getting bigger. And, humans are being trapped in their own web of data.

The discovery of ubicomp and a large number of extremes communicating in own feedback loops within the cloud is causing data to increase exponentially. These extremes include all wearables and handhelds, or mobile sensors. They communicate without servers uselessly in the cloud. So, now one can understand how big are the Internet things and their dependency on the back end or "the cloud". These machines are similar to insensitive people and operate in the cloud without any interference from humans.

Google, around some one decade ago, developed a means that Yahoo is using to clone the data spread in huge clusters and which started to mine big datasets on ad-hoc batch cost effectively. That way has developed as Hadoop. On the traditional database front, many ways are there to measure analytics using the technology of non-relational as well as modified relational database.

Only a part of the population is aware enough of these methods to make use of Big Data. Many layers are there of understanding that humans need to build with enormous data they are generating. To a much surprise, only a part of the layer is exposed to the whole population. A great amount of work is required.

So, imagine this situation as a pile of challenges. Among these challenges are identification, discovery, simulation and modelling, semantics, analysing, storage and processing.

These disciplines are just the surface of the problem. There are sub-challenges under challenges. And each challenge requires its own special level of understanding. We are not efficient at tapping useful resources to address challenges of Big Data, reason being the increasing generated data creating a larger problem. Every individual working on the problem is able to see only small piece of larger problem.

Also it’s obvious to understand what humans want and need to begin with, or what the natural world needs to sustain life at scale. After all, those are the more fundamental problems we're all trying to deal with.

7.0. Conclusion

Storing, or querying, as well as maintaining of Big Data is extremely expensive. There are basically three requirements of Big Data to be worthy of its cost: it must be voluminous in nature, of high velocity, with a huge amount of variety.

Digital product companies like Facebook, PayPal and Ebay need to store as well as recall each and every record of the voluminous data to render their products. For these companies, Big Data is worthy asset. Also, their data fulfils three criteria, that are, there are n number of users and file types that should be queried, with everyone posting different data at every time.

Any company which doesn't make use of Big Data to make core products might find it difficult to justify its cost.

Querying, storing as well as maintaining large data sets is expensive and is time consuming. So, till very necessary and till these three conditions are fulfilled, Big Data is not at all useful.

References

[1] B. P. Zeigler and H. S. Sarjoughian, *Guide to Modeling and Simulation of Systems of Systems*, Springer, London, UK, 2013.

[2] Begum, K. (2016). *Big Data and Large Scale Methods In Cloud Computing*. International Journal Of Engineering And Computer Science.

[3] Benoit Lange, Toan Nguyen, *Big Data architecture for large-scale scientific computing*

[4] Brackstone, M., McDonald, M.: Car-following: a historical review. Transp. Res. Part F Traffic Psychol. Behav. 2(4), 181–196 (1999)

[5] C. J. Clark and P. Hallenbeck, “Data and event visualization,” in *Proceedings of the ITEA Live-Virtual Constructive Conference*, El Paso, Tex, USA, 2009..

[6] D. Kr´ol,M.Wrzeszcz, B. Kryza, Ł.Dutka, and J. Kitowski, “Massively scalable platform for data farming supporting heterogeneous infrastructure,” in *Proceedings of the 4th International Conference on Cloud Computing, Girds, and Virtualization (CLOUD COMPUTING ’13)*, pp. 144–149, Valencia, Spain, May-June 2013.

[7] Dia, H.: An agent-based approach to modelling driver route choice behaviour under the influence of real-time information. Transp. Res. Part C Emer. Technol. 10(5), 331–349 (2002)

[8] Gipps, Parker, D., Lajunen, T., Summala, H., 2002. Anger and aggression among drivers in three European countries. Accident Analysis & Prevention, 34(2), 229-235

[9] J. Kołodziej, H. Gonz´alez-V´elez, and L. Wang, “Advances in data-intensive modelling and simulation,” *Future Generation Computer Systems*, vol. 37, pp. 282–283, 2014.

[10] Kangsun Lee, Joonho Park, *A Hadoop-Based Output Analyzer for Large-Scale Simulation Data*

[11] L. Gong, B. Huang, and Z. Liu, “Research on technique of data storage inmass battle simulation,” *Computer&Digital Engineering*, vol. 40, no. 2, pp. 52–55, 2012 (Chinese).

[12] Manish Parashar, *Addressing Big Data Challenges in Simulation-based Science*

[13] O. Savas, Y. Sagduyu, J. Deng, and J. Li, “Tactical big data analytics: challenges, use cases, and solutions,” *ACMSIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 86–89, 2014.

[14] P.A.Wilcox, A. G. Burger, and P. Hoare, “Advanced distributed simulation: a review of developments and their implication for data collection and analysis,” *Simulation Practice and Theory*, vol. 8, no. 3-4, pp. 201–231, 2000.

[15] Parker, D., Lajunen, T., Summala, H.: Anger and aggression among drivers in three European countries. Accid. Anal. Prev. 34(2), 229–235 (2002)

[16] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, “The changing paradigm

of data-intensive computing," *Computer*, vol. 42, no. 1, pp. 26–34, 2009.

[17] S. M. Sanchez, "Simulation experiments: better data, not just big data," in *Proceedings of the Winter Simulation Conference*, pp. 805–816, IEEE Press, Savanah, Ga, USA, December 2014.

[18] Shengcheng Yuan, Yi Liu, Gangqiao Wang, Hui Zhang, *A Cross-Simulation Method for Large-Scale Traffic Evacuation with Big Data*

[19] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Springer, Berlin, Germany, 2012.

[20] Tawfik, A.M., Rakha, H.A., Miller, S.D.: Driver route choice behavior: Experiences, perceptions, and choices. In: Intelligent Vehicles Symposium (IV), 2010 IEEE, pp. 1195–1200. IEEE (2010)

[21] Treiber, M., Helbing, D.: Memory effects in microscopic traffic models and wide scattering in flow-density data. Phys. Rev. E 68(4), 046119 (2003)

[22] Tu, H., Tamminga, G., Drolenga, H., de Wit, J., van der Berg, W.: Evacuation plan of the city of Almere: simulating the impact of driving behavior on evacuation clearance time. Procedia Eng. 3, 67–75 (2010)

[23] V. Turner, D. Reinsel, J. F. Gantz, and S. Minton, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet ofThings*, IDCAnalyze the Future, 2014.

[24] X.F.Hu, X. Y. He, andX. L. Xu, "Simulation in thebig data era— reviewof newideas and newtheories in the 81stAcademic Salon of China Association for Science and Technology," *Scientia Sinica Informationis*, vol. 44, no. 5, pp. 676–692, 2014 (Chinese).

[25] Xiao Song, Yulin Wu, Yaoefi Ma, Yong Cui, Guanghong Gong, *Military Simulation Big Data: Background, State of the Art, and Challenges*

[26] Y. Ma, H. Wu, L. Wang et al., "Remote sensing big data computing: challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[27] Y. Zou, W. Xue, and S. Liu, "A case study of large-scale parallel I/O analysis and optimization for numerical weather prediction system," *Future Generation Computer Systems*, vol. 37, pp. 378–389, 2014.

[28] Y.Wu, X. Song, andG. Gong, "Real-time load balancing scheduling algorithm for periodic simulation models," *Simulation Modelling Practice andTheory*, vol. 52, no. 1, pp. 123–134, 2015

[29] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y.: T-drive: driving directions based on taxi trajectories. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 99–108. ACM (2010)

[30] https://gigaom.com/2012/08/22/facebook-is-collecting-yourdata-500-terabytes-a-day/.