A New Conceptual Model for Big Data Analysis

Thi Thi Zin[1(✉)], Pyke Tin[1], and Hiromitsu Hama[2]

[1] University of Miyazaki, Miyazaki, Japan
thithi@cc.miyazaki-u.ac.jp, pyketinll@gmail.com
[2] Osaka City University, Osaka, Japan
hama@ado.osaka-cu.ac.jp

Abstract. In today modern societies, everywhere has to deal in one way or another with Big Data. Academicians, researchers, industrialists and many others have developed and still developing variety of methods, approaches and solutions for such big in volume, fast in velocity, versatile in variety and value in vicinity known as Big Data problems. However much has to be done concerning with Big Data analysis. Therefore, in this paper we propose a new concept named as Big Data Reservoir which can be interpreted as Ocean in which all most all information is stored, transmitted, communicated and extracted to utilize in our daily life. As a starting point of our proposed new concept, in this paper we shall consider a stochastic model for input/output analysis of Big Data by using Water Storage Reservoir Model in the real world. Specifically, we shall investigate the Big Data information processing in terms of stochastic model in the theory of water storage or dam theory. Finally, we shall present some illustrations with simulation.

Keywords: Big Data · Stochastic model · Water Storage Reservoir Model

1 Introduction

From industry to consumer, banking to retail, from medical expert to patient and many other sectors have already been and have being embraced into Big Data - regardless of whether the information come from private or public. In terms volume, the scaling of petabyte data has been flowed into big data reservoir daily from web services, social media, astronomy, and biology science. Thus, big data can be defined as a collection of large datasets that may not be processed using traditional database management tools [1]. Mostly, problems of data storage, information manipulation, and especially searching key information from big data have been front line research areas which are widely researched and engineered by sizable researchers [2]. Specifically, the data flow into big data has two sources such as collective gathering and individual generation. The collective gathering big data includes smart city data, national geographic conditions monitoring data, and earth observation data [3]. Usually the collective gathering data are obtained by using statistical sampling techniques leading to high quality data. On the other hand, most of big data are generated by individuals by using social media data on the Internet. The individual data generation is more freedom giving low reliability and usability [4].

J.C.-W. Lin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 579,
https://doi.org/10.1007/978-981-10-6487-6_7

Collecting and analyzing data are commonly concerned with statistics which make able to judge on the basis of hypothesis. However, the big data technology is much advanced at the stages of data sampling, storage management, data computation, and data communication. In the traditional scientific paradigm, the theory is proved with the experiments. In the current scientific paradigm, the scientific finding is often obtained by computer simulation, and is mainly explored from multi-source observations from big data. In summary, we can say that big data have the characteristics of 4 V namely, volume, variety, velocity, and veracity. As the names describe the meaning we can explain these characteristics as follows. First V stands for volume of large amount of data and the second V represents variety or multiple-type or multi-source data. The third V representing velocity of generating data and processing data at high speed and the fourth V characterize veracity or value which refers to the high quality and value of captured and analyzed data. Data quality is comprehensively measured with inherent information content and its user demands satisfaction [5].

In this paper we propose a new concept and approach to the big data by introducing an analogy of big data with reservoir theory in the stochastic water storage processes. We then analyze the data inflows and outflows into buffers to investigate the insight patterns of big data to extract some important key information. The rest part of the paper includes some related works in Sect. 2, the overview and problem formulation in Sect. 3, illustrative simulations results in Sect. 4 and concluding remarks in Sect. 5.

2 Some Related Works

In this section, we shall present some research works of others which are related to this paper. Although a tremendous amount of research works concerning with Big Data analysis and Theory of Storage so called the stochastic reservoir theory has been appeared in the literature, we will describe some works which are in line with our works of this paper. The theory of storage with respect to probability concepts was first introduced by P.A.P Moran in 1954. Since then many researchers had examined and extended the works of Moran in theoretical aspect as well as application aspects. Among them, we would like to refer some works of Phatarfod [5] about stochastic reservoir theory and its extensions [6–9]. In this concern, a common concept is that a reservoir is built for preventing floods or irrigation use in which water inflows into the reservoir and released a certain amount for optimal regulation of a system in which the inflows and outflows are formed a sequences of random variables satisfying laws of probability so that the name becomes a stochastic reservoir theory. Even though we name a stochastic reservoir theory, there are many parts which demand the use of statistical techniques such as time series analysis of inflows to estimate an optimal size of a reservoir. Also, by using statistical regression analysis, we can find the probability distributions of inflows so that the input-output analysis is done for computing various important quantities including storage size, optimal regulation policies, overflow and emptiness probabilities and the respective times to be take. Those quantities are very useful to draw some analogies between the stochastic reservoir theory and buffer data storage system in the Big Data Analysis.

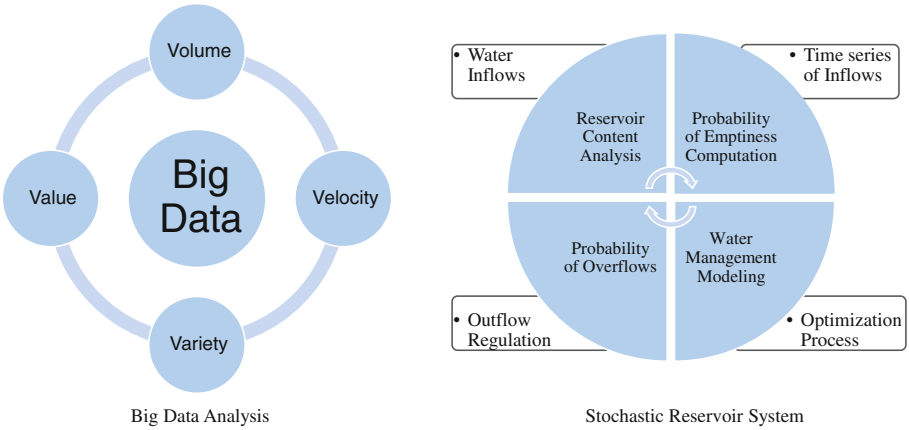

Fig. 1. Stochastic Reservoir and Big Data Models

In the context of Big Data analysis, it has been well recognized that various sources such as social networks generate a huge amount of data having big in volume, fast in velocity, vary in variety and high in values. It will be worthwhile to note concerning with Big Data where the generation process of data is done by the users as well as providers so that most of data are unstructured making the analysis attractable for extracting some reliable and useful information. In order to extract important key information, we make big data processing by establishing some kinds of buffer storage systems where multiple streams of multiple data flow into the buffers of the big data. This is the analogy to be taking into account between the stochastic reservoir theory and big data analysis which is illustrated in Fig. 1.

There is so much innovation in data platforms to enable the efficient processing of new type analytics through novel data structures which is termed as data reservoir which needs a constant flow of new data where ever it comes from existing sources or new sources in order to investigate the usefulness and reliability of the information. The data in the data reservoir can be stored in the data reservoir's index or catalog. The catalog defines the origin, owner, and the characteristics of the data. Similar to water reservoir, the data reservoir can provide the regulation of inflows and out flow of the data for optimum usages. The data reservoir is designed to offer access the data for analytics [8]. On the other hand, big data reservoirs are large catchment areas where all kinds of data can be stored and analyzed. This fact is the foundation of a natural information systems of future developments. Also, as the data volumes are large, velocity high and the variety huge, it is essential for data reservoirs to be cost effective and flexible to other methods.

Searching data usually provides an automatic and effective way for reservoir evaluation. In [10], the authors considered a technique of data searching for logging reservoir evaluation and verify its performance. Compared with traditional evaluation methods, it is efficient and not so dependent on expertise. In the following we propose a framework which would be able to analyze.

3 Proposed Stochastic Model for Big Data Analysis

In order to make an optimal operation of data reservoir networks for the big data we propose a new concept of stochastic model from water storage systems to analyze bag data reservoir networks systems in the big data. In typical models, the inputs to the reservoir are data released from upstream reservoirs and stochastic inflows from external sources (like social media and smart phones), while the outputs correspond to the amount of information to be transmitted during a given time period (e.g., a day or an hour or a month). The dynamics for the single reservoir can be modeled by a state equation where the amount of information at the beginning of period $t + 1$ reflects the flow balance between the information that enters (upstream releases and stochastic inflows) and the information that is released or transmitted during period t. The amount of information to be released during a given time period from each reservoir is chosen to minimize some possibly nonlinear cost (or maximize benefit) function related to the releases. Thus, optimal management of the reservoir network can be formulated as an optimization problem, in which the aim is to determine the quantity of information releases that minimize a total cost over a given horizon of T time periods (e.g., a year).

Let X_t for $t = 0, 1, 2, \ldots$ be the amount of data flows into the data reservoir during the time interval $(t, t + 1)$ and at the of interval $t + 1$, a certain amount of data say m is taken out then the data content Z_t left in the reservoir at the end of interval after the release is given the following stochastic state equation:

$$Z_{t+1} = \min(K, Z_t + X_t) - \min(m, Z_t + X_t) \tag{1}$$

The Eq. (1) can be solved when the probability distribution of $\{X_t\}$ and constant m are known. In this paper we shall consider the case of independent and identical distribution for $\{X_t\}$. In order to do so, we first state Wald's fundamental identity in sequential analysis.

Let us define $Y_t = X_t - m$. Then $\{Y_t\}$ is also a sequence of independent and identically distributed random variables. We also have from Eq. (1) that $Z_N = \sum_{t=1}^{N} Y_t$. In this, there are two absorbing barriers at $-u$ and $K - m$ (>0), for the random walk starting at the origin. Here u is the dam content between 0 and $K - m$.

Let n be the smallest positive integer such that $Z_n \geq K - m - u$ or $Z_n \leq -u + m$. Then, if $G(\theta)$ is the probability generating of the distribution of Y:

$$E\left[e^{-\theta} Z_n \{G(\theta)\}^{-n}\right] = 1 \tag{2}$$

for all θ such that $|\mathrm{G}(\theta)| \geq 1$.

It is known by Wald that there is one dominant root θ_0 such that $G(\theta) = 1$. Substituting θ_0 in Eq. (2), we can obtain the probability of absorbing at the barrier which is approximately equal to:

$$P_u = \frac{1 - e^{-(K-m+1-u)\theta_0}}{e^{(u+1-m)\theta_0} - e^{-(K-m+1-u)\theta_0}} \tag{3}$$

By assuming the capacity of data reservoir as K and the initial data level as u, with unit release $m = 1$,we then have the probability of data emptiness and flooding data of overflow as described in Eqs. (4) and (5).

$$P_u = \frac{1 - e^{-(K-u)\theta_0}}{e^{u\theta_0} - e^{-(K-u)\theta_0}} \tag{4}$$

$$Q_u = \frac{1 - e^{u\theta_0}}{e^{-(K-u)\theta_0} - e^{u\theta_0}} \tag{5}$$

From the duality theorem of random walk namely as $P_u = F(K - u)$, we obtain the stationary distribution of data reservoir content as shown in Eq. (6).

$$F(x) = \text{Prob (Data content } \leq \text{x)} = \frac{e^{x\theta_0 - 1}}{e^{K\theta_0} - 1} \tag{6}$$

Now the problem become to evaluate the value of K or the size of data buffer such that:

$$F(K) = \text{Prob (reservoir content } \leq K - 1) = P \text{ where } P \text{ is to be given} \tag{7}$$

4 Experimental Simulation Results

First we made a test how the approximate results given in Eq. (7) are reasonably acceptable or not by comparing the ground truth done by Prabhu [11] for the water content in the theory of storage. The comparison results are presented in Table 1.

Table 1. Distribution function of reservoir content for independent gamma inputs with parameter α and 100 units' release; size of the reservoir, ∼ = 1000

Reservoir content	α = 1/0.9	α = 1/0.9	α = 1/1.8	α = 1/1.8
x	Exact	Approximate	Exact	Approximate
0	0.031	0	0.005	0
100	0.068	0.042	0.016	0.0132
200	0.115	0.094	0.033	0.0323
300	0.172	0.158	0.059	0.0614
400	0.242	0.237	0.097	0.1045
500	0.331	0.333	0.156	0.1716
600	0.435	0.452	0.245	0.2717
700	0.565	0.598	0.379	0.4238
800	0.721	0.778	0.581	0.6529
900	1	1	1	1

From the Table 1, it can be seen that the approximate results are good while the content lies between 200 and 600. That means if the content is between the range of (K/400, 3K/400), the results get better. On the other hand, the results are not very much good when the content is near to zero. This fact does not make much effect for big data because the big data will never become shortage of data. Therefore, we regard that the approximate results for the data reservoir content will make sense for use to determine the reservoir size optimally.

In order to do so, we note that the reservoir size K satisfies the equation $F(lK) = P$ specified probability say p in Eq. (6). Putting $e^{lK\theta_0} = z_0$ into Eq. (6),

$$\frac{e^{y\theta_0 - 1}}{e^{K\theta_0} - 1} = \frac{z_0 - 1}{z_0^{(1/l)} - 1} = P \tag{8}$$

It is known from the solution of a polynomial equation, we can obtain the unique solution of (8) other than non-zero. We then have the optimal reservoir size as:

$$K = \frac{n \log z_0}{\theta_0} \tag{9}$$

By the simulation results we note that for reservoirs with the same values of l and P, $K\theta_0$ is a constant. Thus, to compare sizes of reservoirs with the same critical level and critical probability, we need only compare the values of θ_0. Thus we can compare the values of θ_0 by varying the size of data reservoirs so that we can obtain the optimal size of data reservoir of computing buffer size in the big data. This will make efficient computing effect to investigate insight information for the big data. However, in this paper we can give the outline procedures only.

5 Conclusion

In this paper we had proposed a new concept of big data reservoir by using the concept of stochastic water storage to investigate the insight information from the big data. We have used Wald's fundamental theorem in sequential analysis which was very popular in various fields of research such as queuing theory, dam theory and other operation research problem. We have presented only simulation so far. More works to be done in the future.

Acknowledgment. This work is partially supported by the Grant of Telecommunication Advanced Foundation.

References

1. Hilbert, M.: Big Data for development: a review of promises and challenges. Dev. Policy Rev. **34**(1), 135–174 (2015)
2. Wang, L., et al.: Bigdatabench: a big data benchmark suite from internet services. In: Proceedings of 20th IEEE International Symposium on High Performance Computer Architecture, pp. 488–499 (2014)

3. Li, D.R., Yao, Y., Shao, Z.F.: Big Data in the smart city. Geomatics Inf. Sci. Wuhan Univ. **39**(6), 630–640 (2014)
4. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K.: Efficient machine learning for big data: a review. Big Data Res. **2**(3), 87–93 (2015)
5. Phatarfod, R.M.: Some aspects of stochastic reservoir theory. J. Hydrol. **30**(3), 199–217 (1976)
6. Bohling, G.: Stochastic simulation and reservoir modeling workflow. Aust. J. Basic Appl. Sci. **3**, 330–341 (2005)
7. Karacan, C.Ö., Olea, R.A.: Stochastic reservoir simulation for the modeling of uncertainty in coal seam degasification. Fuel **148**, 87–97 (2015)
8. Browning, C., Kumin, H.: Stochastic reservoir systems with different assumptions for storage losses. Am. J. Oper. Res. **6**(5), 414 (2016)
9. Archibald, T.W., McKinnon, K.I.M., Thomas, L.C.: An aggregate stochastic dynamic programming model of multi-reservoir systems. Water Resour. Res. **33**(2), 333–340 (1997)
10. Thomas, A., McMahon, T.A., Pegram, G.S., Vogel, R.M., Peel, M.C.: Revisiting reservoir storage-yield relationships using a global stream flow database. Adv. Water Resour. **30**, 1858–1872 (2007)
11. Prabhu, N.U.: Some exact results for the finite dam. Ann. Math. Stat. **29**(4), 1234–1243 (1958)