Accepted Manuscript

Efficient resource management system based on 4Vs of big data streams

Navroop Kaur, Sandeep K. Sood

PII: S2214-5796(16)30090-9
DOI: http://dx.doi.org/10.1016/j.bdr.2017.02.002
Reference: BDR 60

To appear in: *Big Data Research*

Received date: 4 July 2016
Revised date: 31 December 2016
Accepted date: 20 February 2017

Please cite this article in press as: N. Kaur, S.K. Sood, Efficient resource management system based on 4Vs of big data streams, *Big Data Res.* (2017), http://dx.doi.org/10.1016/j.bdr.2017.02.002

Efficient Resource Management System based on 4Vs of Big Data Streams

Navroop Kaur* and Sandeep K. Sood
navonline98@gmail.com* and san1198@gmail.com

***Abstract* — Big data streams are generated continuously at unprecedented speed by thousands of data sources. The analysis of such streams need cloud resources. Due to growth of big data over cloud, allocating appropriate cloud resources has emerged as a major research problem. The current methodologies allocate cloud resources based upon data characteristics. But due to random nature of data generation, the characteristics of data in big data streams are unknown. This poses difficulty in selecting and allocating appropriate resources to big data stream. Solving this problem, an efficient resource management system is proposed in this paper. The proposed system initially estimates the data characteristics of big data stream in terms of volume, velocity, variety and variability. The estimated values are expressed in terms of a vector called Characteristics of Data (CoD). On the other hand, clusters of cloud resources are created dynamically with the help of Self-Organizing Maps (SOM). SOM uses CoD to create and allocate cluster to big data stream. Moreover, the topological ordering of clusters formed by SOM is used to reduce waiting time. The proposed system is tested experimentally. The experimental results show that the proposed system not only efficiently predicts data characteristics but also effectively enhanced the performance of cloud resources.**

***Index Terms* — Big data streams; Cloud computing; Self-organizing maps; Characteristics of Data (CoD).**

1. Introduction

Rapid development and adoption of smart objects in every sphere has amplified the prevalence of Internet of Things (IoT). The growing number of IoT devices has led to a drastic increase in data volume and data velocity. On the other hand, the heterogeneity of IoT devices enhances data variety. Consequently, IoT data is characterized by volume, velocity and variety.

According to Gartner IT Glossary [1], big data is defined as:

"*Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*."

With respect to the above definition, it can be stated that the data harvested by IoT devices has entered big data era.

Furthermore, in a smart environment, some of the IoT devices are periodic while the others are event triggered [2]. The periodic devices generate constant amount of data at regular intervals, thereby generating *big data streams*. On the other hand, event triggered devices are activated when certain event is encountered. The erratic nature of trigger event varies data flow rate. The variation in data flow rate is termed as variability. Therefore, in addition to volume, velocity and variety, IoT data satisfies another dimension of big data called variability.

The IoT devices can generate data stochastically. For example, one event triggered device may, in turn, trigger other IoT devices. In such a case, it is difficult to determine how many devices will be activated and how much data will be generated. Such stochastic nature leads to the generation of big data streams with unknown characteristics. Here, data characteristics imply volume, velocity, variety, and variability of data.

Apart from the IoT devices, big data streams are generated by other applications too, such as social media, click-streams, business transactions, GPS systems, and sensor networks. Intuitively, these applications generate huge volumes of data at high velocity. Moreover, data from these sources consists of images, video, audio and text which contribute to data variety. The trending topics on social media and daily/seasonal loads enhance variability. Therefore, big data streams from these sources are characterized by 4Vs: Volume, Velocity, Variety and Variability. It can be noted here that big data streams from most of the applications are generated randomly and therefore they too have unknown characteristics.

The incessant and unprecedented speed of big data streams escalates the problem of its real time analysis. Conventionally, cloud computing is used to tackle this issue. But with the growth of big data over cloud [3], selecting appropriate cloud resources for such real time analysis has emerged as major research problem. The current practices [4]–[6] allocates cloud nodes based upon user-defined memory size, GPU power and processing power. For such an allocation, the user must be acquainted with characteristics of data (or the 4Vs of data). For example, the user selects higher memory size for higher volume data; higher GPU power for video streams and higher processing power for higher velocity and higher variability. Alternatively, user selects cloud nodes based upon the 4Vs of data stream. Such resource selection is limited by the expertise of user. Moreover, even if the user is expert, the knowledge of data characteristics is necessary.

As stated earlier, the 4Vs of incoming big data streams are unknown to the user due to random data generation by IoT devices and other applications. Therefore, user is unable to determine appropriate cloud resources for real-time data analysis.