International Journal of Trend in Research and Development, Volume 6(1), ISSN: 2394-9333
www.ijtrd.com

Big Data Streaming: An Introduction

[1]Matthew N. O. Sadiku, [2]Damilola S. Adesina and [3]Sarhan M. Musa,
[1,2,3]Roy G. Perry College of Engineering, Prairie View A&M University, Prairie View, TX, United States

Abstract: Nowadays most data is generated in the form of a stream. Streaming data is one that is generated continuously by several data sources. It includes data such as ecommerce purchases, information from social networks, sensor networks, search engines, e-mail clients, social networks, computer logs, and financial trading floors. It is key to turning big data into fast data. This paper provides a brief introduction to big data streaming.

Keywords: *Big Data, Big Data Streaming, Big Data Analytics*

I. INTRODUCTION

Data is the new currency in today's digital economy. We are witnessing the exponential growth of data due to the constant generation of new information that has to be collected, stored, and processed. How data can be processed faster (batch versus streaming processing) is an interesting debate. Some data naturally comes as a never-ending stream of events. A stream is a table data in the move. Examples of such data include data from traffic sensors, health sensors, customer transactions, website visits activity logs, manufacturing processes, email, blogging, twitter posts, and almost all IoT data. To perform batch processing on continuous data requires stopping data collection at some time, storing it, and processing it. Stream processing naturally handles never-ending data streams.

Stream processing is a big data technology which enables users to query continuous data stream and detect conditions fast within a small time period. It is also known as real-time analytics and streaming analytics. Its key strength is that it can provide insights faster, often within milliseconds to seconds [1]. Information derived from data streaming analysis gives companies visibility into many aspects of their business and customer activity. Streaming data is an analytic computing platform that is focused on speed because applications may require a continuous stream of often unstructured data to be processed.

II. USES OF STREAM PROCESSING

Streaming data processing is useful when dynamic data is gene rated on a continual basis. It is better suited for real-time monitoring and response functions. For example, an online gaming company collects streaming data about player-game interactions, and feeds the data into its gaming platform. As another example, in the IoT, massive data streams that are continuously being generated at high speed, and algorithms that process it must do so under very strict constraints of space and time.

Stream processing plays a key role in a data-driven organization. Following are some of the use cases [2].

- Stock market surveillance,
- Smart patient care
- Monitoring a production line
- Supply chain optimizations
- Intrusion, surveillance and fraud detection
- Most smart device applications: Smart car, smart home
- Smart grid
- Traffic monitoring and transport management system
- Sport analytics
- Context-aware promotions and advertising
- Computer system and network monitoring
- Traffic monitoring
- Predictive maintenance
- Geospatial data processing

When organizations are planning for their future, they need to be able to envisage how changes impact the products and services will offer. To do this may require analyzing lots of data. Streaming data is beneficial when analytics have to be done in real time while the data is in motion.

III. USING DATA STREAMING FOR BIG DATA

Unlike traditional data, the term big data refers to large growing data sets with heterogeneous formats (structured, unstructured and semi-structured data) from heterogeneous sources. Big data is characterized by the five "Vs" which are volume, variety, velocity, veracity and value. Volume relates to the data's size (terabytes, petabytes, or zettabytes). Variety refers to different types of data and their sources (sensors, devices, social networks, the Web, mobile phones, and so on). Velocity refers to the speed with which the data is generated from sources (for hourly, daily, weekly, monthly, or yearly). Veracity implies truthfulness and credibility of the data thus collected. Value refers to the actual use of the data collected [3]. The five V's are illustrated in Figure 1 [4].

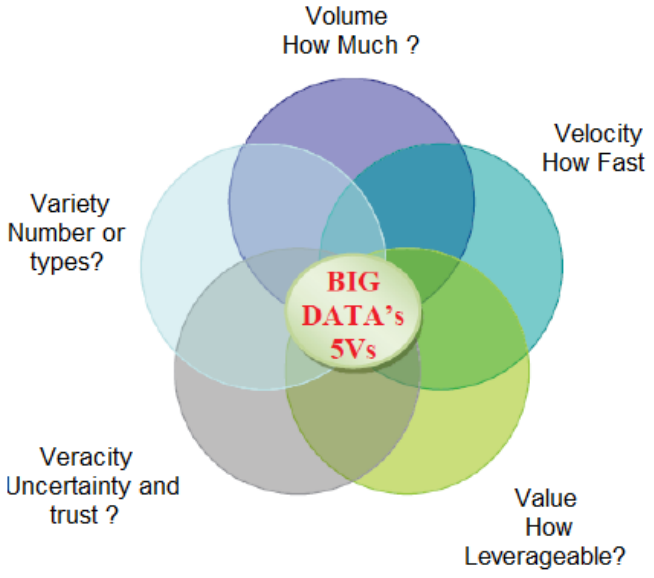

Figure 1: Big data's 5V [4].

Different data sources such as social networks, sensors, and satellites generate big data. Big data associated with time stamp is called big data stream. Examples of big data streams include sensor data, call center records, social media, stock exchange, retail data, sales and shipment, and healthcare data [5]. With the explosion of big data, processing big data streams has become a crucial requirement for many scientific and

industrial applications. Processing big data in real time has become essential for enterprises to garner general intelligence and avoid biased conclusions. But processing such data cannot be done by a traditional database. Some features of big data streams include [6]:

- Real-time: Generate data and feedback results in real-time.
- Unpredictable: The arrival rate and order of data are uncertain, and it is timevarying.
- Volatile: It is impossible to store all of data, and most of them will be discarded.
- Disorder: Data objects are disorder in a same data stream or between data streams.
- Infinite: Data are continuously generated with high rates and throughput, and it is unlimited.

IV. DATA ANALYTICS OF BIG DATA STREAMS

Big data analytics is the core of the big data processing because it is the decisive factor in the decision-making process. All of the five famous big data Vs (Volume, Variety and Velocity, Veracity, and Value) are involved in processing data streams. Big data analytics can be classified into two different approaches: in batches and in streams [7]. Batch processing has been long investigated and has become increasingly mature. Batch processing platforms (e.g. Hadoop and Spark) tackle huge static input datasets. Stream processing has emerged recently as another approach to tackle the unbounded data. Streaming enables the organization to collect, store, manage, and manipulate vast amounts data at the right speed and time, to gain the best insights. In order to process big data streams in real time, several technologies have been proposed and developed.

Popular techniques include MapReduce, Big Table, Apache Kafka, Apache Storm, Apache Flink, Yahoo!S4, and Hadoop. Today, Apache Kafka is most popular architecture used for processing the stream data. It is the de facto architecture to stream data. It is developed at LinkedIn and available as an open source project. It has the following features [8]:

1. Scalability: This framework scale easily without down time.
2. High-volume: It is designed to work with high volume of data.
3. Reliability: Kafka is partitioned, replicated, distributed and fault tolerance.
4. Data Transformations: This frame work should provide provision for ingesting the new data stream from producer.
5. Low latency: To focus on traditional messaging, requires low latency.

Companies such as Twitter, LinkedIn, Yahoo!, and Netflix are using Apache Kafka. Technological advances have promoted big data streams common in many applications, including mobile Internet applications, internet of things, industry production process, social media analysis, financial analysis, video annotation, surveillance, IoT-based monitoring systems, MOOCs, and medical services. Big data stream systems are used in monitoring prisoners and people with dementia. To achieve these applications, numerous hardware and software solutions have been proposed [9].

V. CHALLENGES

Stream processing can work with a lot less hardware than batch processing. However, necessity to handle huge amount of data brings new opportunities and challenges. To facilitate analysts to find the outliers hidden in big data streams, several major challenges have to be addressed. Detecting anomalies for streaming applications remains a challenge. Accuracy and privacy are conflict properties for big data streaming algorithms. Preserving the privacy of big data streams requires accessing big data streams, which is problematic [10]. Streaming data is difficult to be analyzed and processed in real time due to high velocity and huge size of data.

CONCLUSION

With the exponential growth of the interconnected world to the Internet, a very large amount of data is produced coming in a form of continuous streams Big data is flowing into every area of our life. It is regarded as datasets whose size is beyond the ability of typical software tools to capture, store, and analyze. Processing such big data stream in real time is a crucial issue for several applications. New mining techniques are necessary due to the volume, variety, and velocity of such data. More information about big data streaming can be found in the books in [11,12].

References

[1] "What is stream processing?" https://medium.com/stream-processing/what-is-stream-processing-1eadfca11b97

[2] "What is stream processing in big data and what does it do?" https://www.quora.com/What-is-stream-processing-in-big-data-and-what-does-it-do

[3] M. N.O. Sadiku, M. Tembely, and S.M. Musa, "Big data: An introduction for engineers," Journal of Scientific and Engineering Research, vol. 3, no. 2, 2016, pp. 106-108.

[4] H. Asri et al, "Big data in healthcare: challenges and opportunities," Proceeding of International Conference in Cloud Technologies and Applications, June 2015.

[5] E. Mohammadian, M. Noferesti, and R. Jalili, "FAST: Fast anonymization of big data streams," Proceedings of the 2014 International Conference on Big Data Science and Computing, Beijing, China, August 2014.

[6] L. Chen, S. Gao, and X. Cao, "Research on real-time outlier detection over big data streams," International Journal of Computers and Applications, October 2017.

[7] Y. Jiang, Z. Huang, and D. H .K. Tsang, "Towards max-min fair resource allocation for stream big data analytics in shared clouds," IEEE Transactions on Big Data, vol. 4, no. 1, January-March 2018, pp. 130-137.

[8] B. R. Hiraman, C. M. Viresh, and K. C. Abhijeet, "A study of Apache Kafka in big data stream processing," Proceedings of International Conference on Information, Communication, Engineering and Technology, Pune, India, August 2018.

[9] K. Kanoun et al., "Big-data streaming applications scheduling based on staged multiarmed bandits," IEEE Transactions on Computers, vol. 65, no. 12, December 2016, pp. 3591-3605.

[10] A. Cuzzocrea, "Privacy-preserving big data stream mining: Opportunities, challenges, directions," Proceedings of IEEE International Conference on Data Mining Workshops, 2017, pp. 992-994.

[11] P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, 2011.

[12] P. C. Zikopoulos et al., Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data. New York: The McGraw-Hill Companies, 2012.

About the authors

Matthew N. O. Sadiku (sadiku@ieee.org) is a professor at Prairie View A&M University, Texas. He is the author of several books and papers. He is a fellow of IEEE.

Damilola S. Adesina (damiadesina87@gmail.com) is a doctoral student at Prairie View A&M University. His research work is currently focused on optimizing communication systems using machine learning and deep learning.

Sarhan M. Musa (smmusa@pvamu.edu) is an associate professor in the Department of Engineering Technology at Prairie View A&M University, Texas. He has been the director of Prairie View Networking Academy, Texas, since 2004. He is an LTD Sprint and Boeing Welliver Fellow.