International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Some Survey on Challenges in Real Time Processing of Streaming Data

K. Saranya, S. Chellammal

Department of Computer Science, Bharathidasan University Constituent Arts and Science College, Affiliated to Bharathidasan University, Navalurkuttapattu, Tiruchirappalli, Tamilnadu, India

ABSTARCT: Now a days, the world is filled with emerging trends and technologies. Due to the immense amount of smart technologies, most of the domains are upgraded with smart devices in order to produce huge data with enormous speed. The data which flows continuously with respect to speed is termed as streaming data [1]. There are plenty of resources which are capable in generate those high speed events such as mobile devices, satellites, sensor networks, telecommunications, healthcare sectors, business organizations and so on. The data comes from variety of sources need high quality real time analysis for better decision making. Real time analysis, in the sense, the data should be processed and analysed at the time of arrival. But, there are some acceptable challenges encountered while dealing with streaming data having high data rate, vast amount of data, concept drift, data in any format, noisy data, abnormal events, out of order data etc. This survey mainly concentrates on the problem arises in handling streaming data and offers some remedies. Also, the paper gives some general suggestions and research directions.

KEYWORDS – Streaming data, real time analysis, concept drift, scalability.

I. INTRODUCTION

The volume of automatically generated data are constantly increasing. According to the Digital Universe Study, over 2.8ZB of data were created and processed in 2012, with a projected increase of 15 times by 2020 [2]. This growth in the production of digital data results from our surrounding environment which is being equipped with more and more sensors. People carrying smart phones produce data, database transactions which are being counted and stored. Streams of data are extracted from virtual environments in the form of logs or user generated content. A significant part of such data is volatile, which means it needs to be analyzed in real time as it arrives.

Streams refer to constant and continuous flow of data and the corresponding analysis of streams is one of the hot research fields [3]. A data stream is a huge, endless, temporally ordered, infinite and agile information. In recent years, the interest in this field has increased resulting in large volumes of literature has been published in past few years. As well as researches have been carried out on data streams mainly motivated by the many evolving applications which involve huge volumes of data generated from various sensors, data from supermarkets, telephone logs, and data from satellites and various other sources. There are various data stream related issues may arise at the time of processing and some of them are listed as frequently changing dynamic nature, huge volume and speed with which data is generated, memory requirements, etc.This paper identifies real world challenges for data stream research that are important but not yet solved.

In the remainder of the article, Section II provides a glance of literature review, section III discusses about open challenges in streaming data, Section IV highlights the summary of the survey and covers with research issues and future work.

II. LITERATURE REVIEW

In [2], open challenges have been discussed along with some applications. The aim of the work is to cover issues in streaming data such as incomplete data, irrelevant data, delayed information, dynamic nature of data. The work discusses challenges in preprocessing the streaming data as the data arrives continuously even with concept drift. Also, the paper talks about timely responses as a major concern in stream processing. The paper discusses different data stream algorithms for processing with various evaluation metrics. Finally, it concludes that there are some unaddressed research issues.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

A survey on data stream is represented in [3]. The paper discusses about the generation of high speed data in various domains with its challenges so far mentioned. The work explains about the traditional methodologies available in processing streaming data with some gaps. Later, the survey expands its view on data mining techniques in data streams. Finally, it highlights *issues related to memory requirements, concept drift and scalability*. These challenges are need to be solved in real time using effective techniques.

Issues in data streams have been elaborated in [4] with some techniques such as sliding windows, sampling, sketches, etc. This paper initially highlights issues related to limited memory, data rate at high speed, continuous data and the need for immediate response. Also, the paper highlights about *Data Stream Management System* (DSMS) where one time queries are replaced with continuous queries and it uses SQL query language for processing data streams. This work concludes that there may be lack of guarantee in order of data arrival as a open challenge.

The research work [5] briefly explains challenges associated with processing the data stream in real time. The paper emphasizes the need for minimal latency and the necessity for prediction & detection of abnormal events while processing. To resolve this issue, this work proposes a big data based environment which is capable of handling large amount of high speed data. Also, the work compares Hadoop and Apache Spark with Apache Storm and suggests that Apache Storm suits well in processing data in real time, where Hadoop and Spark are designed for batch and micro batch processing respectively. Hence, it is suggested to use Storm and Machine learning for processing data in real time.

In [6], big data stream analysis has been discussed as a major topic. In that, batch computing and stream computing have been compared with some basic qualities such as input, data size, storage, processing, latency and applications. The paper describes about various issues namely *scalability, integration, fault tolerance, timeliness, consistency, incomplete data and high throughput* that may occur while handling stream analysis. The paper proposes some tools and techniques for collecting and processing streaming data in real time. The paper also lists out the methods such as Spark Streaming, Apache Storm, Kafka, Apache Samza, NoSQL, etc., are used for processing big data streams. Also, a comparison table is provided in this work to make better understanding of the tools and techniques. The survey concludes the need to consider the above mentioned issues while dealing with high speed data.

III. CHALLENGES IN HANDLING STREAMING DATA

The recent evolution of sensor technologies, wireless communications and smart devices leads to produce data at high speed with enormous amount. Since, it is fast, it is mandatory to process such data in real time which brings new challenges. From the investigation done, it is clear that the challenges that occur while handling streaming data vary from domain to domain. Following are some of the common challenges:

A) **Failed to process variety of data:** Many of the domains such sensor networks, satellites, healthcare, IoT devices, telecommunication, social media are in practice to generate stream of events as real time data. The type of data may be of any form such as structured or unstructured means the stream may contains audio, video, text or images as content. The traditional techniques viz., sliding window, sampling, histograms and sketches are unable to handle streaming data as mentioned in [7]. Also, [7] discusses the need for more efficient mechanisms to process speedy data

B) **Stream data analysis is difficult:** Analysis of large dataset becomes a open challenge in processing streaming data as the collection of data may contain structured and unstructured data [8]. It is unknown to detect the format of data. It requires more computational task as one has to handle inconsistencies in the dataset. Handling inconsistencies in streaming data is an open issue

C) **Infinite dataset:** It is heavily difficult to store and process all the continuous flow of data in real time. In batch system, the size of the dataset is known and it is quite fair to process such kind of data. In streaming environment [9], the size of the dataset is infinite in nature as the data is on the fly. It is mandatory to process such kind of infinite data with efficient techniques which are capable of handling large amount of data with high speed. Hence, it becomes an another issue while working with data streams in real time.

D) **Dynamic nature of data:** Another major challenge deals with streaming data is concept drift. The underlying contribution of various features may change with time or distribution of values of features itself may change over time. Concept drift [10] is a essential aspect to be considered when managing data streams. As the data

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

keeps on changing, it is difficult to analyse those data with high speed. Most of the traditional techniques are suitable in handling static data. When move on to data streams, it becomes complicated to handle concept drift as it is a critical issue of data streams.

E) **Timely analysis:** Streaming data needs real time analysis. That is one has to analyze the data at the moment of data arrival itself. For instance, data is stored in traditional database is often processed with batch processing. It leads to longer latency and more computational cost. This situation becomes completely different in case of streaming data where the data needs to be analyzed and accordingly processed simultaneously as it arrives [11]. Here processing and analyzing the data in real time as it arrives is a key challenge.

F) **Scalability:** As the volume of input streams or the complexity of data increases, there is a need to partition the arrived data in a kind of distributed data storage to meet the required scalability [11]. This feature automatically reduces the load over the current node and balances the larger load with other nodes. This in turn compels the need for parallel computing. Here, the need for parallel processing needs to be considered as an another issue

G) **Missing value, imperfect data:** Batch processing is suitable for handling data imperfections such as missing value and noisy data as the data is stored in somewhere. But, it is quite complicated in real time streaming data as it never stored in anywhere before processing the data. At the time of stream processing, there is a greater need to have built-in mechanisms to support elasticity against stream imperfections [12] such as missing data, delayed or out of order data which are commonly present in real time data streams.

H) **Failed to detect outliers:** Data streams may evolve with time. Outlier is represented with an abnormal data or suspicious data over entire dataset [10]. As the data keeps on changing with high speed, it is difficult to detect an outlier. A number of methods have already been proposed to detect and classify anomalies for static data set. The applicability of those methods heavily depends on data dynamics. For applications where data is on the move, the challenge of detecting anomalies has become harder over time, as data can dynamically evolve nature. In such cases, being able to detect anomalies in real-time becomes a crucial feature.

I) **Must generate predictable outcomes:** For a group of processing data, it is possibly to predict an outcome [11] in advance based on historical data is potential in nature. Considering a stream processing system, the earlier techniques must process time series events in a predictable way. When the same input stream is applied for reprocessing it should yield the same outcome which looks like the previous one except the time of execution. Most of the approaches support batch system to generate the predictable outcomes with the help of convenient mechanisms. While move on to streaming system, it becomes strictly complicated in order to provide an assurance for predictable and repeatable outcomes. Hence, data stream needs perfect tools to solve this kind of challenge in real world data.

Based on the review on other literatures, some of the research issues have been represented as follows:

- The dynamic characteristics of data flowing over time requires *adaptive* algorithms for analysis
- The data elements in the stream arrive online.
- The system has no control over the order in which data elements arrive to be processed, either within a data stream or across data streams.
- Data streams are potentially unbounded in size.
- Once an element from a data stream has been processed it is discarded or archived. It cannot be retrieved easily unless it is explicitly stored in memory, which typically is small relative to the size of the data streams.
- Data stream may contain incomplete, delayed and outliers.
- Stream of events need real time prediction and detection methods.
- The system needs real time visualization and scalability.

IV. CONCLUSION AND FUTURE DIRECTIONS

In this paper, various issues that are associated with handling streaming data in real time are presented. Open challenges or research topics in streaming data analytics are discussed. The need for adaptive algorithms is mentioned. Conventional databases and processing techniques are unable to handle the streaming data. So, there is a necessity to

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

move on to advanced methodologies which is able to analyze the streaming data in real time stream. As discussed earlier, processing streaming data in real time is a complex and interesting concept. Recent research works start developing new models for analyzing streaming data using big data techniques and frameworks. Those have to explored with associated pros and cons and ultimately effective solutions are to be arrived.

REFERENCES

[1] Alassandro Margara and Tilmann Rabl, *"Definition of Data Streams"*, pp.1-5, July 2018.
[2] Georg Krempl, Indre Zliobaite, Dariusz Brzezinski, et al., "Open Challenges for Data Stream Mining Research", July 2014.
[3] Shazia Nousheen M, Dr. Prasad G R, "A Survey paper on Data Stream Mining", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 5, Issue 08, pp. 321-325, August-2016.
[4] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani and Jennifer Widom, "*Models and Issues in Data Stream Systems"*, pp.1-30.
[5] Ounacer Soumaya, Talhaoui Mohamed Amine, Ardchir Soufiane, Daif Abderrahmane and Azouazi Mohamed, *"Real time Data Stream Processing Challenges and Perspectives"*, International Journal of Computer Issues (IJCSI), Online ISSN: 1694-0784, Print ISSN: 1694-0814, vol.14, Issue 5, pp.6-12, September 2017.
[6] Taiwo Kolajo, Olawande Daramola, Ayodele Adebiyi, "Big Data Stream Analysis: A Systematic Literature review", Journal of Big Data, pp. 1-30, 2019.
[7] Umesh Kokate, Arvind Deshpande, Parikshit Mahalle and Pramod Patil, *Data Stream Clustering Techniques, Applications and Models: Comparative Analysis and Discussions,* Article in Big Data and Cognitive Computing, pp.1-30, 17 October 2018.
[8] D. P. Acharjya and Kauser Ahmed P, *"A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools"*, International Journal of Advanced Computer Science and Applications (IJACSA), vol.7, No. 2, pp. 511-518, 2016.
[9] Fatih Gurcan and Muhammet Berigel, "*Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks and Challenges"*, 2nd International Symposium on Multidisciplinary Studies and Innovatve Technologies (ISMSIT), Online ISBN: 978-1-5386-4184-2, 19-21 October 2018.
[10] Joao Gama, "Issues and Challenges in Learning from Data Streams Extended Abstract", 2007.
[11] Michael Stonebraker, Ugur Cetintemel and Stan Zdonik, *"The 8 Requirements of Real Time Stream Processing"*, ACM Publications, vol. 34, Issue. 4, December 2005.
[12] Dmitry Namiot, "*On Big Data Stream Processing"*, International Journal of Open Information Technologies, ISSN: 2307-8162, vol.3, pp. 48-51, January 2015.