International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Architectural Structures and Viewpoints of Streaming Big Data

Aysh Alhroob

Abstract*: The Streaming big data is one of shifting trend of technology, from data cycle from external behaviour (Hardware) that is low-level low language to digital level, through virtual memory to be implemented as datasets. The proposed model in this preliminary study impending to track data from a different platform and considered important parameters in establishing regulate the storage machinery, storage format, and the pre-processing tools. Moreover, the source for the data type is unstructured data that request organized data in different levels to able forwards process data from machine level to MetaData. The type of storage is virtual memory and the important matter is the capacity is limited. The tuple is third level context the digital data from each sensor through template are recorded by counter time.

Keywords: Streaming data, Big Data, Volume, Velocity, Preprocess, Smart City, and IoT.

I. INTRODUCTION

The Big data can be defined as the rapid growth of raw data by using technology to transform and analysis of data-heavy workloads. Big data is a new trend in the data mining community; it can be used with abroad number of fields such as bioinformatics, marketing, traffic, medicine, weather and climate. Big data frequency and size are considered important parameters in determining the storage appliance, storage arrangement, and pre-processing tools. Data occurrence and size hang on the data sources which can be classified as follow: on-demand as the data used with social media data, continuous feed data as the big data resulting from traffic, real-time data as weather data and traffic data, and time-series data as time-based [1].

Traffic data is growing rapidly as it starts to save pictures and videos in centre place to use it, and the cities' security needs to extend the period of saving traffic huge data to support public security and decision-makers. The increasing amount of traffic data possess [2][3], the traffic management is the issue because counter by time to swap among colours (Red, Yellow, and Green) , so these are the initial parameters even through colours the time is the major scale, meanwhile the counter of the traffic in configuration mood is movement from more than a one still the challenge looks to solve into different aspects. Inner to challenge from Big Data aspect led to finding out Global Position System (GPS) [4], [5]: Since the GPS device is compatible and useful in Hardware and Software [6]. Consequence the level of challenges is affected in different steps as follows:

Data collection: called raw data which generated and produced based on time, the weak points are found in two aspects, the first aspect is incomplete data, missing data, duplicated data, and reduction data; the second aspect is a fly relationship which the relationship based on the time that sensitively to create datasets. Data combination: based on the type of distribution system to transfer repository. Analysis data: utility to the aggregated data.

The challenges show impact to storage capacity especially in data collection during generating the raw data which the streaming data treat with virtual memory (Buffer) [7]. On the other hand, the process to generate a dataset with clean data and setting the relationship need preprocess technique [8].

The rest of this paper arrange as follows. Section 2 discusses the background and literature review. Section 3 describes the proposed Model. Section 4 concludes the work and future work.

II. BACKGROUND AND RELATED WORKS

Big Data Streaming

As illustrated in [6], streaming data is appropriate to data that has no discrete starting or finish. For case, the data rely on traffic-light is nonstop and no "begin" or "finish." Streaming data is the method of transfer data records ceaselessly instead of in bunches. By and large, information gushing is valuable for the sorts of information sources that send data in little sizes (frequently in K-bytes) in a nonstop stream as the data is created. This may incorporate a wide assortment of data sources such as telemetry from associated gadgets, log records created by clients utilizing your web applications, e-commerce exchanges, or data from social systems or geospatial services.

Customarily, data is moved in bunches. Clump planning frequently shapes sweeping volumes of data at the same time, with long periods of inaction. For outline, the method is run each 24 hours. Though this can be a capable way to handle expansive volumes of data, it doesn't work with data that are suggested to be spilt since that data can be old by the time it is taken care of. The streaming of Data is ideal for time arrangement and recognizing designs over time. For illustration, web session length followings as an example. The IoT data mostly is well suited to streaming data principles. Traffic sensors, health sensors, exchange logs, and action logs are all great applicants for streaming of data.

Manuscript published on 30 September 2019
* Correspondence Author
Aysh Alhroob, Department of Software Engineering, Faculty of Information Technology, Isra University, Amman, Jordan. Email: aysh@iu.edu.jo

This spilt data is regularly utilized for real-time accumulation and relationship, sifting, or examining. Streaming of data permits you to investigate data in real-time and gives you bits of knowledge into a wide extend of exercises, such as metering, server action, relocation of devices, or site clicks.

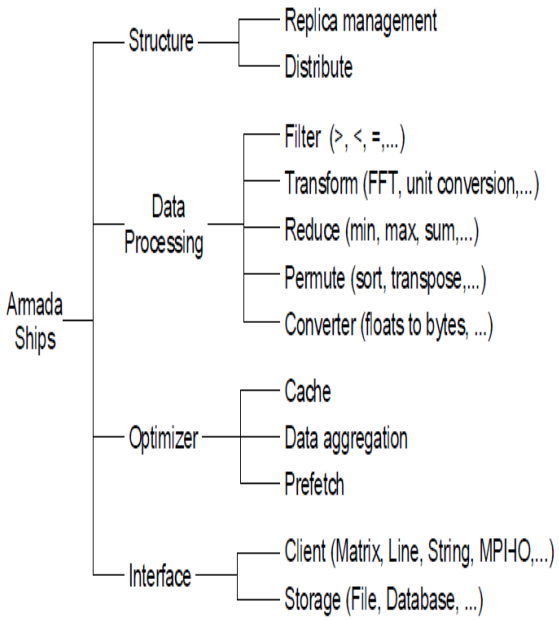

Fig. 1.Data sequence to preprocess

The authors in [7] found that scalability, privacy and stack adjusting issues, as well as an observational examination of huge data streams and innovations, are still open to encourage inquire about endeavours. They found that even though, noteworthy investigate endeavours have been directed to the real-time investigation of big data stream not much consideration has been given to the preprocessing arrange of enormous data streams.

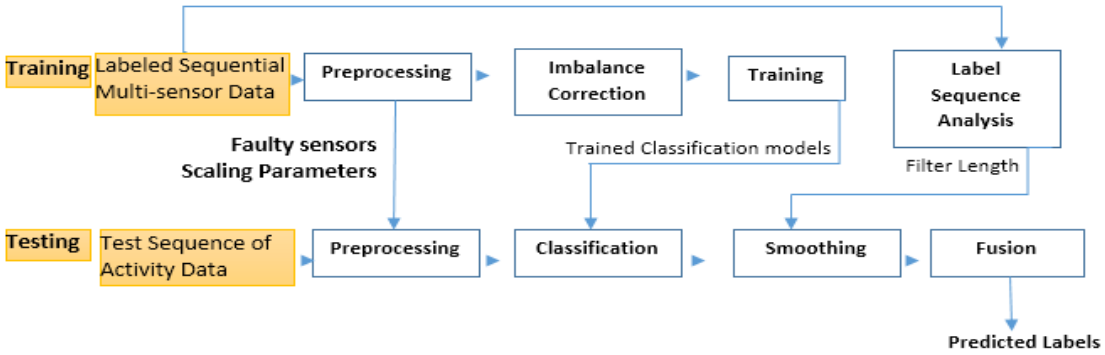

Fig. 2.Block Diagram of the Integrated Framework

Data preprocessing

Data preprocessing is a major method to handle raw data through analysis data processing as shown in Table 1 .The preprocessing methods for several models based on features, which affect the optimization and accuracy; the activity of data preprocesses is a milestone for acquiring information and smoothing to enhance the raw data by dividing data into directions are training and testing as shown in figure 2.

The Armada framework is the development of widely distributed network of heterogeneous systems and devices which focus the efficiency into parallel I/O, the data collection “raw” is often stored in structured data that request a significant preprocessing or filter to allow computation at the right place; armada presented hybrid sequential processes and parallel represented nodes as shown in figure 1 the layers of data.

Table- I: Data Preprocessing Methods

Model/Algorithm	Data Preprocessing Methods
Bayesian neural network learning	multi-layer Perceptron
Bilevel fuzzy optimization	Origin-Destination
Nonaxial models	Potentially Visible Set
Semantic Preprocessing for Mining Sensor Streams	Ontology-based preprocessing
Video database benchmarking	Video preprocessing toolkit

The filter data process federated dataset into two administrative concerns: first call blueprint which concerned the layout of nodes depending on the interface of dataset and preprocessing conditions, which data stream is closer to data server; second API ship services provides processing and distributed data to near data sources (control flow, data flow) [8]. A novel integrated framework that is considered for classification strategies is based on Structure-Preserving Oversampling (SPO) by using standard non-sequential machine learning which evaluated by sequential nature of sensory data. The main stages in this framework are data collection, data transfer, and classification stages (trained, tested), as presented in Figure 2. The training phase through the data sequence to preprocess to handle the missing data collected from the sensor device, the test phase concerned with activity data to enhance the performance of classification based on filter length signal that applies Class Independent (CL) to analyses training label and Class Dependent (CD) to activity class in testing.

III. PROPOSED MODEL

The proposed model expresses the streaming data that is collected from traffic (different locations). Meanwhile, the data is adjusted through time as the primary key to allow store records the reason rely on time because the time can simulate between hardware and software platforms. The source for the data type is unstructured data that request organized data in different levels to able forwards process data from machine level to MetaData.

The data collection from sensors device find some issues as duplicate, incomplete, inaccurate, and missing data. In this proposed stage, the data is generated as shown in Figure 3.

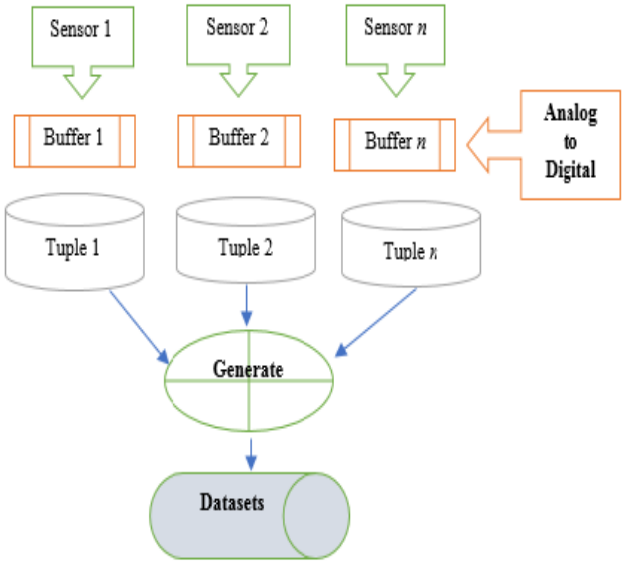

Fig. 3.Data Collection Stage

The data collection stage includes three levels, the first level considers to sensors that collect the data from different locations as setup early, the data collect present as analogue which rate between 0 to 1. This level needs extra device before transfer to the next level toward convert to digital. The second level named buffer for each sensor that is played as storage. The type of storage is virtual memory and the important matter is the capacity is limited. The tuple is third level context the digital data from each sensor through template are recorded by counter time, this level able to determine the issues of the data came from sensors as mentioned previously. The aggregated data from all sensors are combined in one storage with similar time condition that generated datasets.

The datasets illustrate the functions of sensors activity but in high-level data. Thus, able to monitor and control data that determinable to find out issues for data.

The procedures express as sorting functions:

```
IF time == t;
{
FOR Sensors.1 to Sensors.n
{
        Switch (sensors)
        Sensors.1== Buffer.1( Tuple.1)
        Break: Sensors.2== Buffer.2( Tuple.2)
        Break: Sensors.n== Buffer.n( Tuple.n)
}
Return 0;
        Fetch to datasets;
}
```

The data collection stage potential treats with obstacle flying relationship aspect through time which indicates the primary relationship and sensor name. Moreover, the tracking data will be simple refer to item indexing are time and tuple number. While datasets generate and transfer to the repository the principle of big data be appear through volume and velocity dimension of it.

This is a highly abstract model establishes for minimize the volume dimension rely on storage capacity, forwards to limitation size storage that requests preprocess function to ignore missing data and incomplete data, as well as I, can get the benefits from monitor and control process that implements by For Loop. Meanwhile, reduce the time complexity.

IV. CONCLUSION AND FUTURE WORK

The streaming big data one of trend shift of technology from different filed such as IoT and smart city, since the big data dimensions are core volume and velocity which the pinpoint into transfer data, aggregated data, and time complexity the significance of proposed method how to trace the source data from sensor before and after combined as datasets. However, the later researches in this field have empowered rising technologies and solutions to develop novel methods for Big Dat Streaming application and use cases scenarios, there's be that as it may a gap in giving productive and versatile strategies that empower real-time preparing and interpretation of streaming sensory and social media data in numerous situations. In this paper, principles of large-scale data analytics are discussed for real-time data handling and interpretation and examine how different sources of raw tactile data can be combined and prepared to extract actionable-knowledge that can be utilized by citizens and/or decision support systems.

REFERENCES

1. W. Jum'ah Al_Zyadat, F. Y. Alzyoud, A. M. Alhroob, and V. Samawi, "Securitizing big data characteristics used tall array and MapReduce," Int. J. Eng. Technol., vol. 7, no. 4, pp. 5633–5639, 2018.
2. W. J. Alzyadat, A. Alhroob, I. H. Almukahel, and R. Atan, "Fuzzy Map Approach for Accruing Velocity of Big Data," An Int. J. Adv. Comput. Technol., vol. 8, no. 4, pp. 3112–3116, 2019.
3. I. Almukahel, W. Alzyadat, and M. Alfayomi, "Hybrid Approach Using Fuzzy Logic and MapReduce to Achieve Meaningful Used Big Data," Int. J. Eng. &Technology, vol. 7, no. 4, pp. 6997–7001, 2018.
4. B. Akil, Y. Zhou, and U. Rohm, "On the usability of Hadoop MapReduce, Apache Spark & Apache Flink for data science," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-Janua, no. March, pp. 303–310, 2018.
5. S. Jin, J. Peng, and D. Xie, "Towards MapReduce approach with dynamic fuzzy inference/interpolation for big data classification problems," Proc. 2017 IEEE 16th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2017, pp. 407–413, 2017.
6. G. Alley, "what is data streaming," 25 Aug 2019. [Online].Available: https://dzone.com/articles/what-is-data-streaming.
7. T. Kolajo, O. Daramola and A. Adebiyi, "Big data stream analysis: a systematic literature review,", Journal of Big Data, Vol.6, No.1 P.47, 2019.
8. R. Atan and W.J.Alzyadat, "An Approach of Dynamic Filter for Weather Environment Dependent on Time", JATIT, Vol. 18, No. 1, P. 6, 2010.

AUTHOR PROFILE

Aysh M. Alhroob is an associate professor of Software Engineering in Isra University, Jordan. PhD (2010) from University of Bradford, UK. 2010. Aysh joined to Isra University in Jordan as Assistant Professor in Faculty of Information Technology, Software Engineering Department. Aysh has published more than 20 research papers in international journals and conferences.