Streaming Big Data Analytics- Current Status, Challenges and Connection of unbounded data Processing platforms

SK. Wasim Akram , M.Varalakshmi , J.Sudeepthi

Abstract— A strategy of examining immense dimensions of structured, un-structured, Semi-Structured data sets is referred as Big data Analytics. Streaming Big Data refers to data generated continuously from number of data sources like Internet-of-Things (IoT) devices, mobile applications, Embedded Sensors, web clicks and many more are needed to be store, processed and analyzed in a tiny interval of time in order to extract meaningful insights and take proper decisions in a timely fashion as the necessity arises. However analyzing streaming big data (continuous flow or unbounded data) is a very challenging problem. Continuous data streams have become essential prerequisite for numerous industrial and scientific applications, the current existing technology Hadoop-MapReduce is not appropriate for stream processing of big data. This paper discusses the challenges and benefits of streaming big data along with its architecture, and focuses on different open source streaming processing platforms that are existed to process the huge data at a high speed.

Keywords— Structured, Unstructured, Semi-structured, Big Data, streaming data, IoT, hadoop, MapReduce

I.INTRODUCTION

Big data [1] is an advancing term that designates any huge quantity of Un-structured, structured and Semi-Structured data sets that has the possible to be mined in order to get useful information. “Big data” is a term that refers to Petabytes and Exabyte’s of data and it is not bonded to a specific quantity. Big data describe an enormous capacity of data which is difficult to analyze and process with traditional mechanisms. The generation of data is moreover big that exceeds current processing capacity. For instance, size of data in YouTube [2] (or) in Facebook is fall under the class of big data because it required gathering and accomplishing the data on a regular basis. The basic characteristics of Big Data not only focuses on scale and volume, also involves many of the features like Variety, Velocity, Volume, and Complexity and many more.

Online data, streaming data [3], is become more essential in online social networks and many organizations. Applications aid the fact that electronic data should be acquired immediately and then made available to many other information systems. A best example is GPS coordinates that collects data via sensors, it provide the information to several users. This information is used in all kinds of context aware information systems, e.g. location based services. According to Gartner[4], a 3V model of Big Data is described in below diagram:

Classification of Big data:

Volume is a major dimension of big-data. Currently, the volume of data is increasing exponentially, from terabytes to petabytes and beyond.

Volume is a major dimension of big-data. Currently, the volume of data is increasing exponentially, from terabytes to petabytes and beyond.

Velocity includes the speed of data creation, capturing, aggregation, processing, and streaming. Different types of big-data may need to be processed at different speeds [15]. Velocity can be categorized a

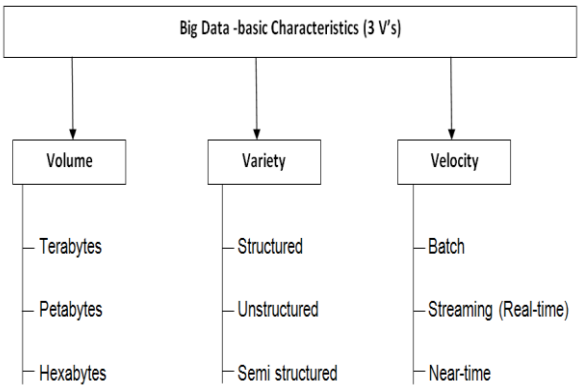


Fig1: Big data Basic Characteristics

Volume:

“It is one of the main measurements related to Big data sets, where the volume of Big Data is currently increasing from-terabytes, Petabytes, Hexabytes and beyond”.

Variety:

“Generally, big data comes in different forms and types. It may consist of structured, semi-structured, and unstructured data. A wide variety of data types may be produced by IoT such as text, audio, video, sensory data and so on”.

Structured big data comprises characters, numbers, floating points, and dates

Revised Manuscript Received on July 18, 2019.

SK. Wasim Akram , Assistant Professor, Dept. Of CSE, VVIT College, Guntur, A.P, India (E-mail: shaikwasimakram585@gmail.com)

M.Varalakshmi, Assistant Professor, Dept. Of CSE, VVIT College, Guntur,A.P, India.

J.Sudeepthi, Assistant Professor, Dept. Of CSE, KITS College, Guntur,A.P, India.

Velocity:

"The rate of IoT big data production and processing is high enough to support the availability of big data in real-time. This justifies the needs for advanced tools and technologies for analytics to efficiently operate given this high rate of data production".

Steaming Data

Data streaming [5,6] is a analyzing giant sets of Big data is to be processed instantaneously to deliver the results instaneously, by this user will get the real-time information and help them to make faster and better decisions. Some of the real time examples of data streaming are Geospatial services, User log files form websites and mobile activities, purchase from online stores, In-game activities, In an ecosystem the information is shared among various connected IoT devices, financial ports and Information on your Facebook, Twitter, Instagram and other social websites.

Table 1: Difference between streaming data and batch processing data:

Batch processing	Stream Processing
It handles a large batch of data	It handles individual records or micro batches of few records
Latency of Batch processing will be in a minutes to hours	Latency of Stream processing will be in seconds or milliseconds
Data generated on mainframes is a good example of data that, by default, is processed in batch form	It is useful for tasks like fraud detection.
It works well in situations where you don't need real-time analytics results	Stream processing is key if you want analytics results in real time
In this processing Model, the operations are to be applied onto the entire dataset	In stream processing, the operations will be applied to each individual data item as it passes through the system
Apache Hadoop and its MapReduce is exclusively used for batch processing	Apache Spark streaming, Storm, Samza and Flink are used for stream processing

Challenges and Benefits of Streaming Big data Analytics:

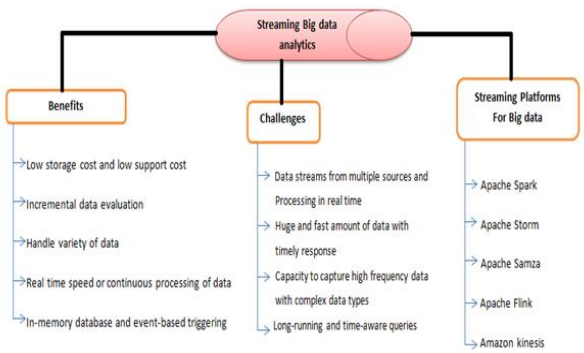


Fig 2: Streaming Big data Analytics: Challenges, benefits and platforms

1.3 Data stream Processing Frameworks:

The systems which handle unconstrained extent of datasets are referred as Frameworks of streaming big data. These Frameworks can process only one record at a time. Various Stream processing Frameworks of Apache are:

1. Spark Streaming
2. Storm
3. Samza
4. Flink

Spark Streaming:

5. "Apache Spark is an open source framework of big data processing built at the base of Hadoop MapReduce to perform sophisticated analysis and designed for speed and ease of use. Spark has lightning fast performance and speed up processing times because it runs in-memory on clusters. Spark can keep running as either independently or on top of Hadoop YARN, where it can read data specifically from HDFS. Organizations like Yahoo, Intel, eBay Inc., Hitachi solutions are utilizing it".

Storm

"It is a distributed real-time computation system that claims to do for streaming what Hadoop did for batch processing. It can be used for real-time analytics, machine learning, continuous computation, and more. The cool thing is that it was designed to be used with any programming language. It keeps running on top of Hadoop YARN[12] and can be utilized with Flume to store information on HDFS. Storm is as of now utilized by any likes of WebMD, Yelp, and Spotify"

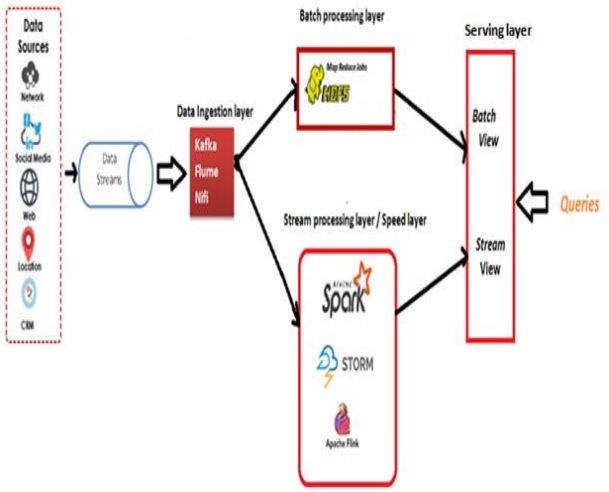
Samza:

"It is a distributed stream-processing framework that depends on Apache Kafka and YARN. It gives a basic callback-based API that is like MapReduce, and it incorporates preview administration and adaptation to non-critical failure in a tough and versatile way"

Flink:

"Apache Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Flink has been designed to run in all common cluster environments, perform computations at in-memory speed and at any scale".

II. STREAM PROCESSING MODEL & RESUTS



This model is taken from basic architecture of lambda. It is divided into 3 layers: batch layer, Stream processing layer and serving layer. In this model, the real time data can scrutinize by Speed layer by considering only the current data into account. It delivers the actual sights that are continuously up to date and stores them in a fast store. The stream processing layer can be appreciated with data streaming technologies such as Spark Streaming, Storm, SamzaandFlink.

1.4 Comparison of Streaming Data Processing platforms:

The following table focuses on tools[13,14] that are used to process batch(bounded) data and streaming(unbounded) data, along with various key features are to be compared.

Table 2: comparison of streaming data processing platforms

Features	Hadoop-MapReduce	Spark Streaming	Storm	Samza	Flink
Processing Framework	batch	Batch and Stream	Stream	Stream	Stream, batch processing
Architecture	Master/slaves	Master/slaves	Peer-Peer	Peer-Peer	Peer-peer
latency	High	A few seconds (< 1s)	Less than a second (< 100ms)	Latency is low	Low latency is Low and high throughput
Storage	In Disk	In memory	In memory	In memory	In memory speed at any scale
Coordination tool	Zookeeper	Zookeeper	Zookeeper	Zookeeper	Zookeeper
Fault tolerance	Yes	Yes	Yes	yes	Yes
Source Model	Open source	Open source	Open source	Open source	Open source
Fault tolerance	Yes	Yes	Yes	Yes	Yes

III. CONCLUSION

In data streaming, before data is actually placed in a disk it should be processed among multiple clusters of servers when it is in motion. The data is divided and sent in terms of pieces (chunks) of size kilobytes and it is processed as per records. Analytics is performed concurrently, getting the results, apply the actions and make decisions that are useful for fruitful benefits in near future. Speed is one of the key elements of data streaming that makes it differs with batch processing. This survey tried to focuses on challenges and benefits of bigdata streaming analytics and a thorough comparison of data stream processing tools is made. In future, we will implement and perform analytics on streaming data with this proposed architecture.

IV. REFERENCES

1. Debating big data: A literature review on realizing value from big data" by wendy Arianne Gunther, Mohammad H. Rezazade, Journal of Strategic Information Systems, no-26, 2017.
2. "Big data analysis on youtube using Hadoop and Mapreduce" by soma hota, IJCERT, Vol-5, Issue-4, 2018.
3. Real-time Data Stream Processing Challenges and Perspectives" by OUNACER Soumaya, TALHAOUI Mohamed Amine, ARDCHIR Soufiane , DAIF Abder rahmane and AZOUAZI Mohamed, in IJCSI, volume 14, Issue 5, September 2017.
4. "Analysis of media Datasets using Hadoop MapReduce"Usha.G.R,Aishwarya S, Kavitha, IJSDR, Vol-2, Issue-6, June 2017.
5. K. LEBOEUF, "2016 Update_ What Happens in One Internet Minute_ - Excelacom, Inc." [Online].Available: http://www.excelacom.com/resources/blog/2016-updatewhat-happens-in-one-internet-minute.
6. "A SURVEY ON STREAM PROCESSING AND STREAMING ANALYTICS FOR REAL - TIME BIG DATA" by B. Srivani, Dr. N. Sandhya and S. Renu Deepti in IJLTET –ICRACSC, 2016.
7. "Survey paper on Big data and Hadoop" by Varsha B.Bobade in IRJET, vol-3, issue-1, Jan-2016.
8. Predective analysis Concepts in Big Data-A survey" by S. Banumathi, A.Aloysius, in IJARCS, Vol-8,No-8, sep-oct-2017.
9. "Big data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", IJCST, Vol-4, issue 1, Feb 2016.
10. "Technologies to handle Big Data: A survey" by sabia and Love Arora in ICCCS, 2014.
11. "A review paper on big data and Hadoop" by Harshawardhan S.Bhosale, Devendra P. Gadekar, IJSRP,Vol 4, Issue 10, Oct 2014.
12. Umapavankumar.K, Dr.B.Lakshmareddy ," Various Computing models in Hadoop eco system along with the perspective of analytics using R and Machine learning" Vol. 14 CIC 2016 Special Issue International Journal of Computer Science and information security:
(IJCSIS)https://sites.google.com/site/ijcsis/ ISSN 1947-5500.
13. www.cloudera.com
14. "Comparative analysis of Big data Technologies" by Rachit Singhal, Mehak jain, Shilpa Gupta, IJAER, Vol-13, No-6, 2018