ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 10(42), DOI: 10.17485/ijst/2017/v10i42/120344,* November *2017* ISSN (Online) : 0974-5645

Real- Time Mining Techniques: A Big Data Perspective for a Smart Future

S. Sanila[1], D. Venkata Subramanian[2] and S.Sathyalakshmi[2]

[1]Department of CSE, Hindustan University, Padur, Kelambakam – 603103, Chennai, Tamil Nadu, India; ssanila@gmail.com
[2]School of Computing Sciences, Hindustan University, Padur, Kelambakam – 603103, Chennai, Tamil Nadu, India; dvenkat@hindustanuniv.ac.in, slakshmi@hindustanuniv.ac.in

Abstract

Background/Objectives: Big Data refers to a set of data that is too big to be managed without using new algorithmic technologies. It is a term used to describe datasets of huge sizes that cannot be handled with the typical data mining software tools. It is critical in today's modern, agile and ever changing environment for systems to extract useful information from data in real-time. Data warehouses are valuable sources of information, but their value becomes enhanced multifold when information is collected and analyzed on the go. Real- Time data analytics are techniques which involves the evaluation of streaming data. **Methods:** Real-Time analytics will enable visualization of the changes as and when they happen. The development of technologies and the deployment of new generation applications influenced the social environment. One side it is tempting the whole nation to development and the other side it is affecting social bonding and responsible attitude of the youth. Any application area becomes "smart" when it is highly advanced in terms of overall infrastructure, sustainable energy and communication mechanisms which will provide essential services to providers. **Findings:** Mining from these real-time data is interesting and extremely challenging as all the predictions, classifications and complicated analysis have to be done on the go itself. **Improvements and Applications:** The in-depth analysis of the existing technologies needs to be carried out thoroughly and its applications have to be modified accordingly.

Keywords: Cloud Computing, Hadoop-Map Reduce, Real-Time Data Analytics, Smart Applications, Spark, Stream Processing

1. Introduction

In today's world the day to day increase in data is by about 5 Exabyte. This is primarily due to the use of emails, audio, video and data streams, health information, various queries, social networks, scientific data and ongoing mobile phone applications. This large amount of data is known as Big Data. The process of exploring interesting and useful information in this mountain of data is known as Data Mining from Big Data. But processing of this large amount of data is one major challenge in research point of view. Cloud computing with high performance is now widely used for large-scale real-time data processing applications.

In order to ensure efficient mining of the Big Data, the need of the hour is to boost research aimed at exploring new analytic methods and development of powerful software tools[1]. To manage the data that is currently generated from sources like sensor networks, measurements in sensor networks and traffic management, log records or click streams in web monitoring, meaningful processes, call detail records, emails, blogs, twitter posts and various others, we need efficient data analytics. The arrival of data at high speed is the major characteristics of a data stream. Any data is a snapshot of a data stream taken at a particular moment of time. The high speed movement of the data stream poses two major challenges to any algorithm that aims at processing this data are the constraints of space

*Author for correspondence

and that of time. Any algorithm that aims to tackle data streams should be capable of using limited resources of time and memory and also did with the data's nature to change its distribution with the passage of time.

In data stream mining there are 3 main areas of interest; namely accuracy, amount of space (computer memory) needed and the time required making learned predictions. All these areas are symbiotic in nature-adjusting the time and space used by an algorithm can influence accuracy. If we gave the algorithm the luxury of referring to more pre processed data and look up tables, the space occupied increases but the time needed for computing decreases. Similarly the longer the system has to analyze data the more accurate the result. In short the concept of green computing which advocates the study and practice of using computing resources efficiently is the need of the hour.

1.1 Big Data

Big Data refers to a set of data that is too big to be managed without using new algorithmic technologies. It is a term used to describe datasets of huge sizes that cannot be handled with the typical data mining software tools[2]. Entire big data system stores and distributes structured, semi-structured and unstructured data. Initially big data had a file sharing distributed framework, to store the collected data. Users can interact with it in a modified C++ called Enterprise Control Language (ECL). Figure 1 Illustrates the basic flow of processes involved in big data mining.

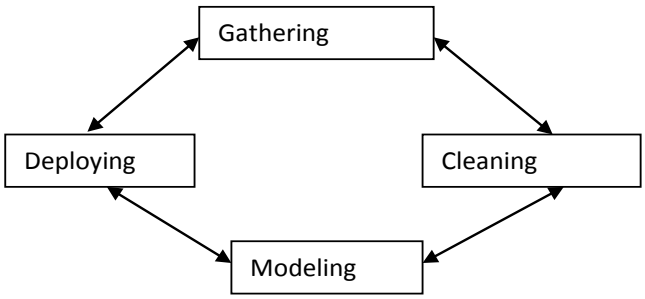

Figure 1. Standard approach of big data.

Storage and processing of these large data on the internet is not possible using traditional storage approaches. Big Data is fast becoming an important and growing tool in organizations around the world to improve efficiency and quality. Due to the importance Big Data technologies have become essential to today's competitive environment. All the manipulation of data is done by new generation technologies. Industries main concern is to maintain speed in processing large datasets in terms of waiting time between queries and running the program.

Big Data can be handled in two ways:

1. Sampling: This is based on obtaining approximate solutions using a subset of the most suitable examples drawn from the dataset which is originally too large. It can crunch a huge volume of data using a distributed parallel processing paradigm[3]. The better the quality of the sample that are drawn from the dataset, the more accurate the results. This sampling reduces the amount of memory and time needed to process the data.
2. The distributed system: The MapReduce methodology[3] started in Google has grown to become the most popular distributed system in use today. This framework was very successful and Apache implemented this algorithm as an open sourced project known as Hadoop. It implements batch processing system and is an excellent offline data processing platform for big data. Hadoop originating from Yahoo is an open source MapReduce system that is popular in handling non streaming data[4]. Inspired by ideas in functional programming the MapReduce model divided the algorithms into two main steps- Map and Reduce. The mappers are fed with input data obtained by splitting data from data sets. This output from the mappers is then fed to the reducers that will produce the final output of the algorithm. In today's business environment speed and agility is a better combination than size and scale. The ability to gather real time data, interpret and gather knowledge from it, react on the basis of this knowledge and most importantly predict new business opportunities in what makers any Business a winner today and a leader tomorrow.

The revolution being brought about by Big Data mining is not restricted to the industrialized world. The main reason for this is the mobile devices spreading in developing countries. 80% of them are located in the developing world and responsible for the generation of vast volumes of data. The big data analyses of the developing nations are being undertaken by a UN initiative known as Global Pulse[5]. They are using the Big Data Analysis to:

1. Develop faster responses in times of a crisis by the detection and analysis of anomalies in the usage of digital media,
2. Fine tune various public policies with a closer representation of reality, and
3. *To check* the success and failure of various policies in Real-Time and to make the necessary changes needed to ensure their success[5].

2. Materials and Methods

If a group of transactions are collected over a period of time from a large data set, batch processing is efficient. For that the needed data is collected first, then entered for analysis, processed over time and then results are produced. For this purpose new and efficient open source technologies have been invented. As the volume variety and veracity of data changes over time, traditional databases become insufficient for decision making and analysis. Distributed processing allows handling of large volumes of both structured and unstructured data efficiently. These huge mountains of data are known as Big Data and are handled effectively with the help of some distributed open source software platforms like Hadoop, and other related software. Hadoop is commonly employed in batch processing and is found to be effective in all recent big data strategies[6].

2.1 Apache Hadoop

Hadoop is the most popular open source implementation of the Map Reduce framework. The ease of use, scalability and failover properties of Hadoop makes it even more popular in big data scenario. Many data mining algorithms are flowing towards Hadoop due to its massive parallel processing ability. No SQL is a distributed database which is most commonly used for Hadoop implementation[7]. Hadoop Distributed File System (HDFS) differs from the existing systems by its sole properties of fault tolerance and low cost hardware. It is a part of the Hadoop core.

2.2 Apache H Base

It is a massively scalable, distributed column based database built on Hadoop to handle billions of messages per day. The flexibility of H Base is because of its column-orientation. If we want to maintain large amount of data and flexibility, H Base is ideal. It is a non relational database runs on top of HDFS. It offers fault tolerant storage and quick access of large data[7].

2.3 Apache Cassandra

It was developed and designed to handle large amount of data and also to ensure the delivery of high availability without a single failure. It is a distributed No SQL database developed by Face book. It is an open source DBMS intended for online applications which requires fast response and no down time[8].

2.4 Apache Pig

It was developed by Yahoo Inc. It allows us to write complex Map Reduce transformation using a simple scripting language called Pig Latin, which connects things together. It can work with complex data structures and can operate on different levels of nesting. Pig translates Pig Latin to Hadoop so that it can be executed within HDFS[8].

2.5 Apache Avro

A framework performs remote procedure calls. It provides serialization of data. Interesting fact is that it is used to port data from one program or language to another, so as to make the data self describing.

2.6 Apache Chukwa

An open source data collection and monitoring system for large distributed systems built on top of HDFS. It inherits scalability and robustness properties of Hadoop.

2.7 Apache Drill

It is an efficiently processing nested data system for analyzing large data sets. The distributed behavior of Drill allows the system to process petabytes of data in seconds. It enables interactive queries instead of using Map Reduce process.

2.8 Apache Flume

Flume provides services for collecting, aggregating and transporting huge amounts of data efficiently in a distributed and reliable environment. It has tunable reliability, failover and recovery mechanisms which allow online analytic data manipulation.

2.9 Apache S4

It is a distributed stream computing platform developed by Yahoo. It serves as a general purpose, fault tolerant, pluggable platform that allows individuals to easily create applications which processes unbounded streams of data[9]. Apache S4 has three basic components:

1. Processing Element (PE): It can send and receive events;
2. Processing Node (PN): *Acts as hosts to PEs*; and
3. Adapter: Feeds events to S4 cluster.

Traditional decision support technologies are inadequate for a real-time environment. Applications nowadays require all types of data to be processed. So, continuous analysis needed to be conducted in memory. Responsiveness is critical in case of handling large amount of unstructured data. As we are moving forward it is an acceptable fact that traditional databases and data warehouses are becoming history. The data has its value only if it is collected and analyzed during the actual interaction. It has to occur in real-time. Hadoop and all other platforms discussed so far are still efficient for batch processing unbound amounts of data. The emergence of Cloud Computing, Internet of Things and location based services had led the amount of data to grow at an amazing speed. It has now become a great challenge to handle and process large amount of real-time data[9].

We have to develop systems that should have the ability to process data and generate the results strictly within certain time frame. In this way the responses are posted in the order of milliseconds or microseconds depending upon the requirements. The decision to choose a particular working platform must depends on data size, speed, throughput, model development and effective optimization tools on which it is going to work[10]. This must have the capability to handle huge amount of data which are coming at high speed in parallel. Analysis has to be done on the moving stream in a fault tolerant way. Spark and Storm provides these functionalities along with the capabilities of Hadoop and Map Reduce.

2.9.1 Smart Future

Today digitization has become the norm rather than the exception. As such the private sectors as well as the government agencies have taken the steps towards adopting real time mining techniques. The aim is not only economic savings but also effective utilization of available resources. Real time data mining will enable to achieve substantial inroads towards performance improvements in the Health, Transportation, Energy, Performance and Water supply sectors[11]. Organizations are waiting for methods to employ efficient, cost effective ways to provide better services thereby increasing responsiveness. The data for these includes record about citizens, business analytics, bank transactions, warehouses, broadcasted images, traffic data, and geospatial and weather information. To overcome the existing challenges in the areas of crime pattern analysis, unemployment, education, surveillance systems, cyber security, sensors, transportation department, police department, social program agencies, tax agencies etc.

Five steps recommended to accomplish the tasks are:

1. Define;
2. Assess;
3. Plan;
4. Execute; and
5. Review.

2.9.2 Proposed Model

The amount of data that has been generated is invaluable information and could be used for various applications and those data contain lots of benefits and yields revolutionary data management mechanisms. These huge data exceeds the processing capability of conventional database systems. Statistics shows that on the inclusion of real time data from various sources, the amount of data produced would be 31GB per hour. Our traditional architecture couldn't cope up with the rapid development of data[12].

2.9.2.1 Algorithm

Step 1: Input - unstructured raw data from the storage for pre-processing

(The data set can be defined as d1, d2, d3…dn)

Step 2: Pre Processing - the dataset contains separate header

Information to differentiate both offline data and the real time.

Step 3: Let Oi for online data and Ri for real time stream

Step 4: if dataset contains Ri info then move the specific data set for real time computation RT or send to offline computation OL.

Step 5: While (RT = true)

{

Step 6: If (RT = stream)

Step 7: then move dataset (RT, stream)

Step 8: else if (RT = strearn event type)

Step 9: Then move_ dataset (RT, strearn)

}

Go to Step 5 for further analysis if else goes to step 10

Step 10: If data set contains Oi data move the specific data set for offline computation OL

Step 11: If (OL= true)

{

Step 12: If (OL= batch)

Step 13: then move_dataset (OL, batch)

}

Go to step 10 for periodic analysis

In the proposed architecture, unstructured data which are collected should be analyzed first through a filtration process as shown in the Figure 2. In the real time paradigm, data acquisition is carried out by collecting raw data and analyzed in two ways such as online and offline. For online analysis or real time analysis, a centralized server which further connected to cloud is used. In offline computations batch processing or distributed processing is deployed using storage system appropriate for it.

The Real-Time Analytics involves the extraction, manipulation and analysis of the given data on the go itself. So latency and faults won't be occurring at any point of time. The system must have fault tolerant behavior so that at any point of time if the data got corrupted it should get retrieved from the alternative source. As with all the Hadoop based system which supports distributed nature, there will be another node anywhere in the cloud with the copy of the same data. Spark and Storm also follows the same scheme[13].

2.9.2.2 Apache Storm

It is a real-time computation system which processes unbounded real-time streams of data, like what Hadoop did for batch processing. It can be used with any programming languages. For real-time data analytics, Storm can be used on YARN, a resource management provider for Hadoop. It has the ability to replay the data which was processed unsuccessfully in the first pass of a process. Storm was proposed to develop object detection systems where objects are detected in real-time from large amount of data. More reliable resources and infrastructures are required to handle these continuous parallel streams of data[14]. Storm focused on fault tolerance and management. It implements guaranteed message passing which removes intermediate queuing and allows messages to flow directly between the tasks themselves. Though Storm is stateless it manages distributed environment and cluster state via Apache Zookeeper. It reads raw stream

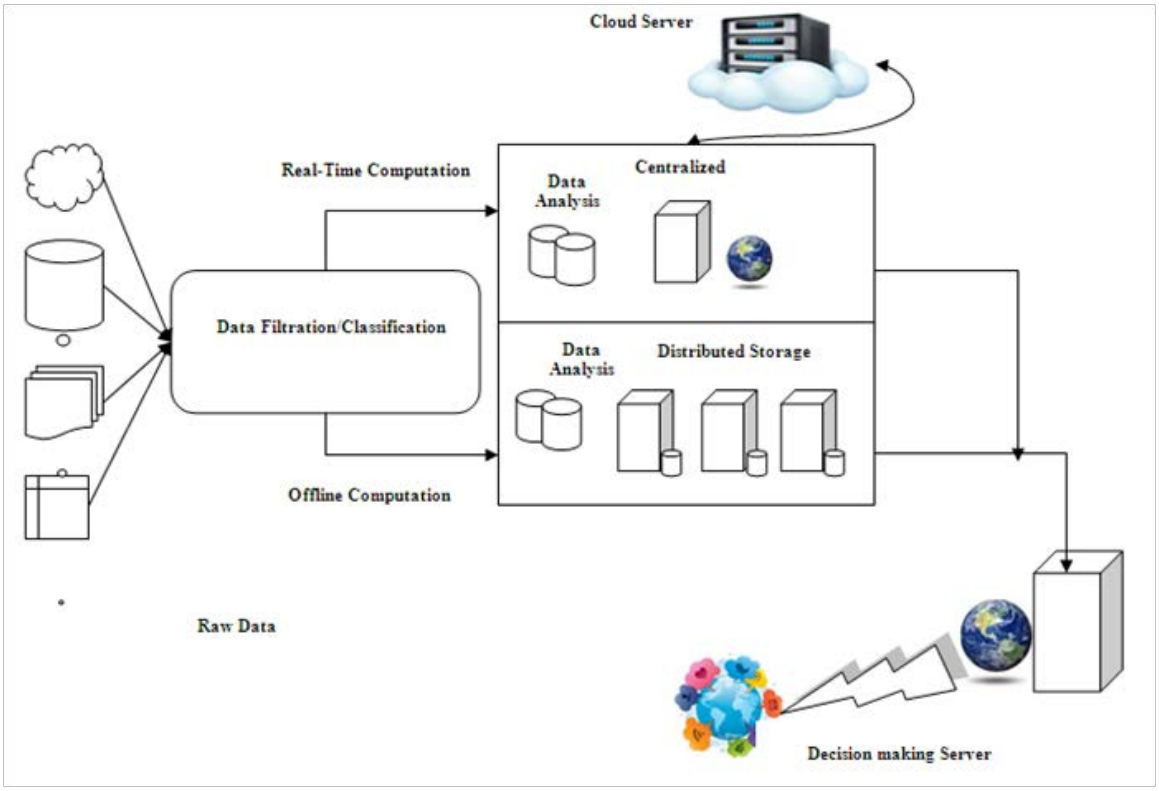

Figure 2. Real-time big data paradigm.

of real-time data from one end and passes it through a sequence of small processing units and output the processed or useful information at the other end.

2.9.3 Apache Spark

Spark is not a modified version of Hadoop. It is not dependent on Hadoop because it has its own cluster management. Hadoop is one of the ways to implement Spark. Map Reduce handles large amounts of data but cannot provide low latency processing. Spark is an alternative which performs at speeds up to 150 times faster than Map Reduce. Spark employs iterative algorithms and interactive data mining. There are two ways to create RDD's- parallelizing and referencing a data set[14]. Spark houses in-memory cluster computing for its rapid performance. To tackle with wide range of online data processing scenarios. Spark combines SQL, streaming and complex analytics together. Spark can be used with Hadoop or it can works in the cloud. It has the ability to access various data sources like HDFS, Cassandra, HBase or S4. Spark provides main memory abstraction with the help of Resilient Distributed Datasets (RDD).

RDD is a fundamental data structure of Spark. It is an immutable distributed collection of objects. It is a collection of parallel nodes within the cluster and it performs two types of operations viz; transformations and actions. Resilient in the sense that it supports both distributed as well as fault tolerant nature in the database itself. Parallelism allows it to speed up the processing thereby increasing the throughput[15].

3. Further Plans and Discussion

3.1 Real Time Mining with Big Data

Data streams correspond to the 'V' of velocity of Big Data, and the process of extracting knowledge instantaneously from rapidly and dynamically changing data examples are known as Real-Time Data Mining. The ability of analyze big data has enabled the world to move from forecasting which was made on the basis of models built on past data- to now casting-which is based on collecting data and predicting what is happening now[15].

Real time mining of data can provide significant advantages and benefits to enhance the quality of life, reduce risks, increase profitability and even help saves thousands of lives in times of crisis[16]. Real Time analytics visualizes the changes along with the run because of the in-memory transfers. Algorithms have to be developed for scaling linearly up or down based on the available memory requirements. Developing technologies to process big data drives new innovations and creative discoveries promoting social progress and allowing us to find ways to solve many problems which was considered to be very hard or even impossible earlier. Real-time denotes the ability to process data as it arrives, rather than storing the data and retrieving it at some point in the future. If the processing time of a transaction exceeds the customer's attention span, the merchant doesn't consider it real time. Real-time architecture in a system will not have all components at first. So on the go we build a system appropriate to an initial application, with the ability to expand to incorporate any missing components in the future.

3.2 Streaming Data Analysis

Streaming data differs from other kinds of data as it is always flowing, loosely structured and it needs high cardinality storage. We need to develop infrastructures and algorithms. Real- Time data analytics are techniques which involves the evaluation of streaming data[17]. This will enable visualization of the changes as and when they happen. In order to achieve a real time data analysis we require a platform to make sense of the constantly changing analysis, applications to process the unstructured stream of data along with continuous in memory interactions.

4. Conclusion

Streaming data differs from other kinds of data as it is always flowing, loosely structured and it needs high cardinality storage. We need to develop infrastructures and algorithms. Real-Time data analytics are techniques which involves the evaluation of streaming data[18]. This will enable visualization of the changes as and when they happen. In order to achieve a real time data analysis we require a platform to make sense of the constantly changing analysis, applications to process the unstructured stream of data along with continuous in memory interactions. Responsiveness is the most crucial element in cases where handling of real-time data is considered[19].

Traditional decision support technologies are inadequate for this environment. Achievement of the above

mentioned aspects can happen only by developing the ability and techniques to handle heavy streams of data in parallel, deployment of more efficient scheduling and resource management, development of framework for in-memory data management and up gradation of already existing algorithms. However before any of these applications can be put to effective use there are certain challenges that need to be overcome. The most pressing challenges of these are those caused by the real time requirements that demand utmost attention to fast collection, transfer and the processing of big data. These challenges include the need for those applications to collect current events, analyze the collected data, decide based on these analysis results and respond in appropriate time levels.

The digitization drive with focus on improving the quality of life of one and all occupying today's ever growing modern world can be made "smart" by adopting real time mining techniques. The tools that will move us towards achieving the goal of being smart are the adoption of real time streaming data processing platforms and up gradation of existing algorithms coupled with extensive use of knowledge gathered from experience[19]. Development of efficient frameworks and up gradation of existing algorithms is mandatory to overcome existing challenges and to provide this world a safe real time environment without any threats and security faults. By 2020 the amount of information stored worldwide is expected to be 50X larger than today's.

Big data analytics extracts and enriches the available structured semi structured or even unstructured data using the intellectual techniques thereby exploring and visualizing user defined contexts in a real- time fashion.

5. References

1. Zhigao Zheng, Ping Wang, Jing Liu, Shengli Sun. Real-Time big data processing framework: Challenges and solutions, Applied Mathematics and Information Sciences. 2015; 9(6):3169–319.
2. O'Driscoll A, Daugelaite J, Sleator RD. Big data hadoop and cloud computing in genomics, ELSEVIER Biomedical Informatics. 2013; 46(5):774–81. Crossref.
3. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters, Communication ACM. 2008; 51(1):107–13. Crossref.
4. Reichman OJ, Jones MB, Schildhauer MP. Challenges and opportunities of open data in ecology, Science. 2011; 331(6018):703–5. Crossref. PMid: 21311007.
5. Chen YC, Peng WC, Lee SY. Efficient Algorithms for Influence Maximization in Social Networks, Knowledge and Information Systems. 2012; 33(3):577–601. Crossref.
6. Fan W, Bifet A. Mining big data: current status, and forecast to the future, Journal of Engineering Research and Applications SIGKDD Explorations. 2012; 14(2):1–5. Crossref.
7. Aggarwal CC. Managing and mining sensor data. Advances in Database Systems, Springer; 2013. p. 1–547. Crossref, Crossref.
8. Letouz E. Big data for development: opportunities and challenges. Global Pulse: New York 10017; 2012. p. 1–47.
9. Massive Online Analysis. Date accessed: 22/03/2017. http://moa.cms.waikato.ac.nz/.
10. Nuaimi EA, Neyadi NA, Mohammed N, Jaroodi JA. Applications of big data to smart cities, Journal of Internet Services and Applications; 2015. Crossref.
11. Smart Cities. Date accessed: 13/05/2016. https://india.gov.in/spotlight/smart-cities-mission-step-towards-smart-india.
12. Im DH, Cho CH, Jung I. Detecting a large number of objects in real time using Apache Storm. IEEE Transaction. Information and Communication Technology; 2014. p. 836–38. Crossref.
13. Maske M, Prasad P. An introduction to real time processing and streaming of wireless network data, IJARCCE. 2015; 4(1):1-4. Crossref.
14. Ballou K. Apache Spark and Storm; 2017.
15. Apache Hadoop. Date accessed: 20/07/2017. http://hadoop.apache.org.
16. Gartner. Date accessed: 11/07/2017.http://www.gartner.com/it-glossary/bigdata.
17. Bockermann C, Blom H. The streams framework. Technical Report 5, TU Dortmund University; 2012.
18. Lu HP, Sun ZY, Qu WC. Big data- driven based real-time traffic flow state identification and prediction, Hindawi Journal on Discrete Dynamics in Nature and Society. 2015; 2015:11.
19. Barlow M. Real-time big data analytics: Emerging architecture. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol: CA; 2013. p. 1–32.