# A Survey on Big Data in Real Time

Rasika Ashokrao Dugane
Computer Science and Engineering
HVPM's COET
Amravati, Maharashtra, India
*dugane.rasika@gmail.com*

Prof. A. B. Raut
Computer Science and Engineering
HVPM's COET
Amravati, Maharashtra, India

***Abstract*-**Big data is data whose characteristics force us to look beyond the traditional methods that are prevalent at the time. Online news, micro-blogs, search queries are just a few examples of these continuous streams of user activities. Evolving data streams methods are becoming a low-cost, green methodology for real time online prediction and analysis.Heterogeneity,scale,timeliness,complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata.

***IndexTerms-*** *Big Data, Data Streams, Hadoop, Sensors, Twitter*

*****

## I.INTRODUCTION

Most data generated is originally streaming data. This fact is especially true for data representing measurements, actions and interactions, such as the one coming from sensor networks or the Web. In-rest data is just a snapshot of streaming data obtained from an interval of time. In the streaming model, data arrives at high speed, and algorithms must process it in one pass under very strict constraints of space and time. Streaming algorithms use probabilistic data structures in algorithm give fast approximated answers. However, sequential online algorithms are limited by the memory and bandwidth of a single machine. Achieving results faster and scaling to larger data streamsrequires to resort to parallel and distributed computing. Map Reduce is currently the de-facto standard programming paradigm in this area, mostly thanks to the popularity of Hadoop1, an open source implementation of Map Reduce started at Yahoo!. Data stream real time analytics are needed to manage the data currently generated, at an ever increasing rate, from such applications as: sensor networks, measurements in network monitoring and traffic management, log records or click-streams in web exploring, manufacturing processes, call detail records, email, blogging, twitter posts and others. In fact, all data generated can be considered as streaming data or as a snapshot of streaming data, since it is obtained from an interval of time. In the data stream model, data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time. Consequently, data streams pose several challenges for data mining algorithm design. First, algorithms must make use of limited resources (time and memory). Second, they must deal with data whose nature or distribution changes over time. We need to deal with resources in an efficient and low-cost way. Green computing is the study and practice of using computing resources efficiently. A main approach to green computing is based on algorithmic efficiency. In data stream mining, we are interested in three main dimensions:Accuracy, Amount of space (computer memory) necessary. The time required to learn from training examples and to predict

### 1.1 WHAT IS BIG DATA?

Big data is a buzzword, catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.

**Fig: 1.1.** Big Data

Big data is typically described by the first three characteristics. The term big data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely structured data. Big data analytics requires capturing and processing data where it resides. This paper explores the value of data at the edge of networks, where some of “biggest” big data is generated. As the use of sensors and devices as well as intelligent systems continues to expand, the potential to gain insight from the flood of data from these sources becomes a new and compelling opportunity. Businesses that can harness the power of big data at the edge and unlock its value to the organization will outperform their competitors with greater capabilities to innovate creatively and solve complex problems whose solutions have been out of reach in the past.

### 1.2 THE ELEPHANT IN THE ROOM

Hadoop and big data have come to become synonyms in the last years. Hadoop has evolved from a simple open source clone of Map Reduce to aflourishing ecosystem of related projects for data storage, management, representation, processing and analysis. Map Reduce and streaming are two fundamentally different programming paradigms, albeit related from a theoretical point of view. In recent years, mostly because of the popularity of Hadoop, there have been various attempts to shoehorn streaming and incremental computation on top of Map Reduce, such as Hadoop Online Prototype. However, all these systems are adaptations and hybridizations rather than principled approaches, and therefore present limited support for proper streaming computation.

### 1.3 BIG DATA STREAMS: VOLUME + VELOCITY

Big data is often understood along 3 dimensions, called 3 V's: Volume, Variety and Velocity. Big data streams are characterized by having high volume and high velocity. Additionally, when dealing with web and social streams variety is given for granted. The key to deal with such complex data is, in our opinion, to combine streaming with distributed computing and open source. The streaming paradigm is necessary to deal with the velocity of the data, distributed computing to deal with the volume of the data, and being open source for the variety. While the first two points are easy to understand, the latter deserves some explanation. No two companies or individuals have the same needs, and open source is a guarantee of openness and adaptability. Indeed, with an open source solution any skilled person can modify the code to suit their needs. For the reasons above, we exclude from our consideration existing commercial SPEs such as IBM Info Sphere Streams2, Microsoft StreamInsight3, StreamBase.Opensource, distributed Stream Processing Engines (SPEs) have been the focus of much research and development recently. One of the first ones to be available as open source was Borealis. Currently, S4 and Storm5 are the State-of-the-art.They draw inspiration from both traditional SPEs and Map Reduce. Similarly to Map Reduce, in these systems data is seen as a sequence of records which can be routed to the right processing element based on the value of a key. Similarly to SPEs, records are processed one by one, and any aggregation (such as the reduce phase in Map Reduce) is left to the user.

## II.BIG DATA TECHNOLOGY

Big-Data Technology: Sense, Collect, Store and Analyze. The rising importance of big-data computing stems from advances in many different technologies:

1.Sensors: Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages.

2. Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet.

3. Data storage: Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, costs around $100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes.

4. Cluster computer systems: A new form of computer systems, consisting of thousands of “nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing, where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms.

5. Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can rent storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and computing become utilities that are ubiquitously available.

6. Data analysis algorithms: The enormous volumes of data require automated or semi-automated analysis – techniques to detect patterns, identify anomalies, and extract knowledge. Again, the "trick" is in the software algorithms - new forms of computation, combining statistical analysis, optimization, and artificial intelligence, are able to construct statistical models from large collections of data and to infer how the system should respond to new data. For example, Netflix uses machine learning in its recommendation system, predicting the interests of a customer by comparing her movie viewing history to a statistical model generated from the collective viewing habits of millions of other customers.

## III. PHASES IN THE PROCESSING PIPELINE

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline

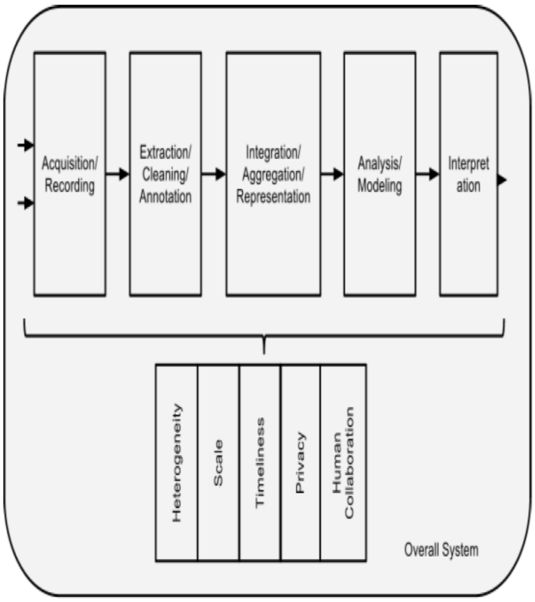

Fig: The Big data analysis pipeline

3.1 Data Acquisition and Recording: Big Data does not arise out of a vacuum. It is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude

3.2 Information Extraction and Cleaning: Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements(possibly with some associated uncertainty), and image data such as x-rays. We cannot leave the data in this form and still effectively analyze. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent (e.g., what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). In addition, due to the ubiquity of surveillance cameras and popularity of GPS-enabled mobile phones, cameras, and other portable devices, rich and high fidelity location and trajectory (i.e., movement in space) data can also be extracted.

3.3DataIntegration, Aggregation, Representation: Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

3.4QueryProcessin,DataModeling,Analysis:Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions.

3.5 Interpretation:Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Big Data is also enabling the next generation of interactive data analysis with real-time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot-lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it.

## IV. NEW APPLICATION: SOCIAL NETWORKS

A future trend in mining evolving data streams will be how to analyze data from social networks and micro-blogging applications such as Twitter. Micro-blogs and Twitter data follow the data stream model. Twitter data arrive at high speed, and algorithms that process them must do so under very strict constraints of space and time. The main Twitter data stream that provides all messages from every user in real-time is called Firehouse and was made available to developers in 2010. This streaming data opens new challenging knowledge discovery issues. In April 2010, Twitter had 106 million registered users and 180 million unique visitors every month. New users were signing up at a rate of 300,000 per day. Twitter's search engine received around 600 million search queries per day, and Twitter received a total of 3 billion requests a day via its API. It could not be clearer in this application domain that to deal with this amount and rate of data, streaming techniques are needed. Sentiment analysis can be cast as a classification problem where the task is to classify messages into two categories depending on whether they convey positive or negative feelings. For a survey of sentiment analysis, and for opinion mining techniques. To build classifiers for sentiment analysis, we need to collect training data so that we can apply appropriate learning algorithms. Labeling tweets manually as positive or negative is a laborious and expensive, if not impossible, task. However, a significant advantage of Twitter data is that many tweets have author-provided sentiment indicators: changing sentiment is implicit in the use of various types of emoticons. Smileys or emoticons are visual cues that are associated with emotional states. They are constructed using the characters available on a standard keyboard, representing a facial expression of emotion. Hence we may use these to label our training data. When the author of a tweet uses an emoticon, they are annotating their own text with an emotional state. Such annotated tweets can be used to train a sentiment classifier.

## V.NEW TECHNIQUES: HADOOP, S4 or STORM

A way to speed up the mining of streaming learners is to distribute the training process onto several machines. Hadoop Map Reduce is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. A Map Reduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job. Apache S4 is a platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time. Storm from Twitter uses a similar approach. Ensemble learning classifiers are easier to scale and parallelize than single classifier methods. They are first most obvious candidate methods to implement using parallel technique.

## VI.CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. We have only begun to see its potential to collect, organize, and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment. We have outlined new areas for research. These include structured classification and associated application areas as social networks. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation.

## REFERENCE

[1] A. Bifet and E. Frank. Sentiment knowledge discovery in Twitter streaming data. In Proc13th International Conference on Discovery Science, Canberra, Australia, pages 1{15. Springer, 2010}.

[2] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[3] D. J. Abadi, Y. Ahmad, M. Balazinska, M. Cherniack, J.-h.Hwang, W. Lindner, A. S. Maskey, E. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik. The Design of the Borealis Stream Processing Engine. In CIDR '05: 1stConference on Innovative Data Systems Research, pages 277{289, 2005.

[4] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011

[5] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[6]Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.

[7] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed stream computing platform. In ICDM Workshops, pages 170{177, 2010.

[8] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):