Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data Analytics: A Perspective View

Suman Pandey*
Department of Computer Science & Engineering, KNIT, Sultanpur,
Uttar Pradesh, India

Abstract--Big data is a term coined for massive data sets with more varied and complex structure and also having the difficulties of storing, analysing and visualizing it for further usages. The process of diving into large amounts of data to discover patterns and disguised correlations is named as big data analytics. These are certainly useful to companies or organizations by gaining richer and deeper insights to compete in market. Thus big data implementations need to be analysed and executed efficiently as possible. Big data analytics is methodology which automates the gathering, organizing, contextualizing, processing and analysing Big data i.e. large set of data to capture patterns help make better decisions. Big data analytics challenges the situation of the present infrastructure of data storage management and also statistical data estimation. This paper studies the content, scope, methods, advantages and challenges of big data and also discusses privacy issue concern on it. The motive of the proposed study is to provide better and significant insights from research prospects and also lays an overview of data analysis methodologies and tools which are currently being utilized or proposed in literature. This work will be quite useful for the future researchers in this domain and facilitate the development of optimal techniques to address Big data.

Keywords: Big data; MapReduce; batch processing; real-time processing; stream processing.

I. INTRODUCTION

In past few years also known as digital decade there has been an explosion of data. In the digitation process various organizations has generated large amounts of data owing to scientific experiments, various website and sensors networks which coined the term Big Data to identify those data which cannot be created, managed, captured or processed by traditional tools in a reasonable amount of time. The following features are used to describe Big data.

- Variety

With various sources like social media sites, web Pages, documents, Web Log Files, e-mail, sensor devices the Big data is semi structured and comprised of non-structured, semi structured and even unstructured raw data. Thus it is very difficult to be processed by existing traditional analytics.

- Volume

The term Big is used to represent the large volume of data and presently it is in petabytes and might increase to zettabytes in near future. These days' social sites are producing data in order of terabytes and which cannot be managed by the yester analytical systems.

- Velocity

The velocity defines not only speed of the data being generated from various sources but also speed at which data flow e.g. data from the sensor devices is constantly adding to the database store. Thus our traditional systems are unable to perform the analytics on such data which is constantly in motion.

- Variability

The term variability defines the inconsistencies of the data flow. With certain events the usage of the social media causes peak in data loads and that become challenging to be maintained by the traditional tools.

- Complexity

The complexity of Big data is in response to link, match, cleanse and transform data across systems coming from various sources. Thus it is necessary to correlate and connect relationships, hierarchies otherwise multiple data linkages causing unmanaged control.

- Value

User extracts important results from the filtered data obtained by running queries against the stored data and these reports may help find the business trends and implement their strategies.

With this growing amount of data as being generated every day without any ends in sight, there is a strong requirement for storing, processing and analysing such voluminous data. There are efficient tools capable of storage management to data processing and also being developed to this specific purpose. But to process this large volume of data, efficient computing power and space is required. Having specialized hardware e.g. supercomputing infrastructures is not feasible economically most of the time.

In the time of cheap and powerful commodity servers, it seems unrealistic that the big data needs moving from the PC world and getting into newer computing platforms better to deal with big data's specific requirements. Large

computing commodities may be economical alternative to required computing power i.e. using large number of low-cost, low-performance commodity computers perform in parallel instead of using fewer high-performance and high-cost computers. But such large number of commodity hardware obtrudes challenges that are not with traditional high-end hardware and thus frameworks for big data must be specifically developed in order to overcome these issues.

The organization of paper is as follows. The section II represents the background, fundamental concepts and describes the rapid growth of data volume whereas section III discusses emergence of Big data analytics. The Section IV describes the challenges of Big data analytics while section V studies the enabling technology of Big data with supporting techniques described in section VI. The section VII discusses tools of analysis and in final section VIII, the work is concluded with mentioning relevant future scope.

II. BACKGROUND

Issues and challenges in Big data are discussed in [1]. The paper [2] discusses some of traditional databases and concluded that databases may not answer every aspects of the Big data issues. The machine learning algorithms need to be more robust and easier for users. There is need a data management system extending these algorithms so that users can produce and indulge their data, ensure its consistency and may browse, visualize and understand their results/produce of these algorithms. The author has discussed in [3] different architecture for Big data and comes to the point that with varied architectures and design decisions the data analytics always tries to scale out, elasticity and high availability. The work studied in [4] concepts of Big data along with the available market solutions used to handle and explore the unstructured large data.

The observations showed that analytics has added value for the social business. The work proposed in [5] is Scientific Data Infrastructure (SDI) a generic architecture model. This model is a basis for building interoperable data with the help of available modern technologies and the best practices. The authors have shown that the models proposed can be easily implemented with cloud based infrastructure services provisioning model. In paper [6] the author has investigated the difference in Big data applications and how they are different from the traditional methods of analytics existing from a long time. In paper [7] authors have analysed Flickr, Locr, Facebook and Google+ social media sites and concluded that the privacy implications and also geo-tagged social media an emerging trend in social media sites.

Most of them are implementing or have already implemented big data tools and technologies. However, in big data analytics, companies face variety of challenges, including infrastructure and data governance also with policy issues. Despite this, IT managers are excited about big data analytics and consider it a top IT priority for their organizations. IT managers engaged in big data analytics share common important characteristics. One-third of companies which we surveyed are working with very large amounts of data such as with 500 TB or more of data per week. Thus they have prioritized big data analytics and working from a formalized strategy for big data. Currently processing unstructured data sources. The top data source continues to be business transactions while data comes from a variety of sources both structured and unstructured. Most of IT managers cited business transactions in a database as the top source.

Currently, applications such as health care analytics (e.g., personalized genomics), business process optimization, and social-network-based recommendations demand for efficient analyses of large datasets. However prognosis suggests that data growth will largely exceed future improvements in the cost and density of its storage technologies, the processing power and the associated energy requirements. For example, in the period from 2002 to 2009 data grew 56-fold, compared to a corresponding 16-fold increase in processing power and data centres grew in size by 173% per year [8]. Extrapolating these trends in data growth, it will take about 13 years for a 1000-fold increase in computational power (or theoretically 1000× more energy). However, energy efficiency is not expected to increase by a factor of 25or more over the same duration. This creates a wide gap of almost a 40-fold increase in the data analytics energy requirements [9,10].

A comprehensive study of big-data workloads can help understand their impact on hardware and software design to process such huge data set. Inspired by the seven dwarfs of numerical computation [11], Mehul Shah et al. [12] tries to define a set of data processing kernel that cover current and future data-centric workloads. Drawing from an extensive set of workloads, a set of classifying attributes such as response time, working set, access pattern, data type, processing complexity, read vs write and conclude that five workload models could satisfactorily cover data-centric workloads as in near future: (i) distributed sorting at petabytes scale, (ii) in-memory index search, (iii) high load processors and regular communication patterns, (iv) sequential-access based data de-duplication and (v) video uploading and streaming server at interactive response rates. However, general business documents, e-mail, sensor or device data and imaging data all unstructured data were in the top five. Adoption of big data tools and technologies continues to grow. Slightly less than half are implementing or have already adopted big data tools or technologies, while one-third of them report that they are in the evaluation stage.

III. EMERGING BIG DATA ANALYTICS

The usages of Big data analytics in collecting and analysing large amounts of data in variety of application domains is motivated by following factors.

- The complex systems need beyond the data-driven modeling and hypothesis generation to understand system behaviour and interactions and such applications depends upon big data analytics to study environments (the sciences), social interactions, and engineered systems, among others.
- The commercial and business enterprises the importance of big-data analytics is well accepted in efficient guiding decision processes.

- Big Data analytics have tremendous potential for prognostic interventions, novel therapies and in shaping lifestyle and behaviour of human beings and is also key to cost efficiencies and sustainability of the healthcare infrastructure.
- In client-oriented or interactive environments, big-data analytics guide the systems interface with the clients. Examples of businesses shaping their operational environment to optimize client experience (and business outcomes) are well understood.
- Ubiquitous environments such as smart homes are now emerging, simultaneously optimizing for living spaces as well as energy footprint and cost.
- In complex systems that lend themselves to shaping and control, analytics enable controlled evolution and design. The aforementioned application areas are, by no means, exhaustive.

IV. BIG DATA ANALYTICS: CHALLENGES

Challenges in Big data analysis are described in following as.

A. Heterogeneity and Incompleteness

In case of human beings, a great deal of heterogeneity in information is comfortably tolerated owing to richness of language and natural reasonable understanding. But, machine analysis algorithms require homogeneous data and cannot fill up the incompleteness. Thus data must be carefully structured before starting the data analysis. Efficient representation, access, and analysis of semi-structured data require further work. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and errors must be managed efficiently before starting data analysis and this is in fact a challenge.

B. Scale

Of course, size of Big Data needs further efforts to be processed. Managing large and rapid growing volumes of data is a challenging issue. In the past, this problem was scaled down by getting faster processors employing Moore's law and provided the resources needed to cope with increasing volumes of data. But, in today's scenario volume of data is scaling faster than computing resources. Just reviewing the last five years it is seen that the processor technology has been advanced rather as processors are doubling their clock cycle frequency every 2 years and due to power constraints, processors are being built with increasing numbers of cores instead of increasing the clock speed.

In previous decade, large data processing systems worried about parallelism across nodes in a cluster but now, parallelism within a single node is successfully being implemented. The technique of parallel data processing that were applied in the past for processing data across nodes cannot be directly applied within the node, since the architecture are different e.g. many more hardware resources such as many processor caches and processor memory channels shared across cores are required in a single node. These unprecedented changes require us to rethink how to design, build and operate data processing components.

Towards the solution, dramatic shift is underway and that is move to cloud computing, which now aggregates multiple workloads with varying performance goals such as interactive services demand for quick answer e.g. within a fixed response time and this level of sharing of resources in large clusters requires new ways to determine, how to process data so that it meets the goals of each workload cost-effectively. To deal with system failures, which occur more frequently as we operate on larger and larger clusters.

C. Timeliness

The other side of larger data size is speed of processing. The larger the size of data set to be processed, the longer it will take to be analysed. The design of a system analytics that effectively deals with size must be able processes it faster also. However, it is not just this speed that usually matters when one speaks of velocity in the context of Big data rather, there is an acquisition rate challenge or a timeliness challenge that comes in. There are situations where immediate analytics is required. For example, in a fraudulent credit card transaction it should ideally be flagged before the transaction is completed or preventing the transaction from taking place at all. In a large data set, it is often necessary to find elements of a specified criterion and this type of search is likely to occur repeatedly in course of data analysis. Scanning the whole data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding of qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criterion.

With new analyses using Big data, there are specific criterion but which needs new index based structures that support such criteria. Consider a traffic management system which maintains thousands of vehicles and manages local hot spots on roadways. Such systems need to predict potential congestion points along a route chosen by a user and also suggest alternatives. Developing so requires evaluation of many spatial proximity queries that must be resolved with the trajectories of moving objects. These require new index structures supporting such queries and designing such structures becomes particularly challenging when the data volume scales much larger and also tight time limits.

D. Privacy

The privacy of data is of important concern but it becomes crucial in the context of Big data. However, there is great fear regarding the inappropriate use of personal data particularly through linking of data from multiple sources. Managing secrecy or privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

Let consider data is being collected from various location-based servers. These new architectures need to share location with the service provider which creates privacy issues and hiding the user's identity without hiding its location would not suffice to completely address secrecy issue. A potentially malicious attacker of location-based server can grasp the identity of the source queried from their position information. A user's location information can be obtained through several stationary communication points (e.g., cell towers). It is important to hide user location and is much more challenging than hiding its identity as location-based services need location of the user to have successful data access or data collection, while the identity of the user is not necessary. There are certain additional challenging problems to be resolved. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases.

Practically, Big data is always getting larger and changes over time and none of the available techniques results in any useful content being released in this scenario. Yet another very important issue is to address security of data and information shared over internet applications. Nowadays may online services need us to share private information over the internet but beyond record-level access we do not understand technical concept of data sharing that how the shared data can be shared and how it gives users fine control over this sharing.

E. Human Collaboration

With tremendous advancements made in computational analysis, there are many patterns that humans can easily detect but machine algorithms have a hard time finding. Ideally, analytics for Big data is not at all computational rather it will be designed explicitly to have a human interactions in the processing loop. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big data analysis system must take input from multiple human experts and explore results shared. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input and support their collaboration. A popular but new method of harnessing human ingenuity is to solve problems through crowd sourcing. It also needs a framework to use in analysis of such crowd-sourced data with conflicts. Here, issues of uncertainty and error become more pronounced. The extra challenge is the inherent uncertainty of the data collection devices. This is probable that collected data are spatially and temporally correlated and which can be pre-processed to better assess their correctness.

F. System Architecture

These days' companies appreciate value of business intelligence. Business data needs to be analysed for efficient growth of business such as log analytics and social media analytics have been in use to assess risk, customer retention, brand management and so on. Traditionally, such distinct tasks are handled by separate sub systems even if each system includes common steps of information extraction, data cleaning, relational aggregations, statistical and appropriate exploration, predictive modelling and visualization tools. Using separate systems is not advisable in Big data not only due to cost of systems themselves but time to load data into multiple systems costs much. Big data needs run heterogeneous workloads on a single infrastructure which must be sufficiently efficient to handle all these varied workloads. Here, the challenge is not to design a system which ideally suits to all processing tasks instead system architecture must be flexible enough to efficiently run these different workloads.

V. BIG DATA ANALYTICS: ENABLING TECHNOLOGIES

A. Granular Computing

This is a very basic and general concept to construct an efficient computational model for complex applications with large scale data, information and knowledge and which works by reducing data size into different level of granularity [13]. In case of loss of hidden information owing to data size reduction process, this theory may be unable to perform correctively [14].

B. Cloud Computing

Cloud computing offers more values from their data enabling blazing-fast analytics in a fraction of previous costs. This motivates companies to acquire and store even more data, creating more need for processing power and driving a virtuous circle [15]. Apart from its flexibility, cloud computing addresses one of the challenges related to transferring and sharing data among others. Thus solution of Big data problems will resort to cloud computing to achieve high level of development [16].

C. Storage Technologies

The increasing volumes of data needs an efficient storage techniques employing data compression and storage virtualization technologies such as Solid-State Drive (SSD), Direct-Attached Storage (DAS), Phase Change Memory (PCM), Network-Attached Storage (NAS), and Storage Area Network (SAN) [17, 18].

D. Biological Computing

Biological computing models are also effectively used tools for big data analytics and offers efficient mechanisms to organize access and process data in ways more practical to the ranging and nearly infinite input data we deal with every day. As near future is very prone to use bio-inspired technologies so a large amount of funds and human resources are poured into related research activities [19, 20].

E. Quantum Computing

Quantum computing is used to harness and exploit the powerful laws of quantum mechanics to process information and data. A quantum system encodes the zero and the one into two distinguishable quantum states. Certain problems can be solved much faster by larger-scale quantum computers more efficiently and faster than traditional ones [21].

F. IPv6 and IOTs

The incontrovertible protocol has such as a zillion IP(v6) addresses. IETF standards have adopted two new major protocols in IPv6, for IoTs and they are Low-Power and Lossy Networks (RPL) in Routing Protocol [22] and the Constrained Application Protocol (CoAP) [23]. IPv6 Internet connects all things of a human to monitor even at his work place and results in generating Big-Data.

G. Wireless and Sensor Networks

Data communication and networking technologies such as optical/MPLS, IP networks, Passive Optical Networks (PONs), mobile wireless access technologies are embedded with sensing technology. With the potential benefits of large-scale deployment of low-cost, energy efficient, and multiservice capabilities, the WSN applications are responsible for huge amounts of heterogeneous data generated from a wide variety of application domains. This ultimately needs to develop novel approaches to ensure guarantees over data transmission delay and loss. The ability to discover, store, clean, analyse and model big sensor data still needs to be developed.

VI. SUPPORTING TECHNIQUES

The techniques for big data analytics encompasses number of areas such as data mining, statistics, neural networks, machine learning, social network analysis, pattern recognition, signal processing, optimization methods and visualization approaches which are summarized here.

A. Statistics

To collect, organize and interpret the big data the old techniques of statistics are used. To exploit the casual relationship and correlation ship among distinct objectives, efficient approximate algorithms are proposed for large-scale multivariate monotonic regression. In this approach the estimating functions which are monotonic with respect to input variables are considered. The basic trends in data-driven statistical analysis focus on scale and parallel implementation of statistical algorithms. The statistical learning and statistical computing are the two hot research areas [24].

B. Optimization Methods

Solutions of quantitative problems employ optimization methods. In most of the research works large-scale optimization are performed using co-evolutionary algorithms and Big Data applications need real-time optimization [25].

C. Data Mining

Data mining is used to extract useful patterns from data. The common classification and clustering analysis, association rule mining, regression and discriminate analysis are basic analytics being used. It uses techniques from statistics and machine learning. The Big Data mining is a Challenging issue. Mostly the extensions of available techniques are based on analysing a particular amount of samples of Big Data to derive a partition for the overall data. The genetic algorithms are also applied to clustering as optimization criterion [26].

D. Social Network Analysis

Social Network Analysis (SNA) views social relationships in terms of network theory; consists of nodes and ties. Visualization Approaches are used to create diagrams, tables, images and other intuitive display methods are used to understand data [27].

E. Artificial Neural Network and Deep learning

Artificial Neural Network (ANN) has a wide scope in big data analytics. A successful application may be utilized in pattern recognition, image analysis, adaptive control and other areas. Mostly ANNs being employed are based on statistical estimations, classification optimization and control theory [28]. The deep learning method based on neural networking, has good potential for solving business problems. It enables to recognize items of interest in large quantities of unstructured and binary data and deduce relationships without needing specific models or programming instructions [29].

F. Visualization

In order to extract knowledge from large and complex datasets this derives a set of methods or techniques to combine the computing capability of computers with imaginative and perceptive power of humans. This technique is based on user interaction and visual system. It is extension of data visualization. Thus IVA is a suitable technique to analyse multi-dimensional data comprising of large number of data points and where simple graphing and non-interactive techniques results into insufficient understanding of the information.

Visualization techniques are used to create tables, images, diagrams and others to understand data [30] and for large-scale data visualization many researchers use feature extraction and geometric modelling to reduce the data size significantly before actual data rendering [31].

G. Column-Oriented Databases

Since the big data is unstructured and column-oriented they need methods for huge data compression and timeliness. The downside to these databases is to allow batch updates only and having a much slower update time than traditional models [32].

H. Schema-Less Databases or NoSQL Databases

A NoSQL (Not-only-SQL) database is basically designed to distribute, store and access data using methods which are different from relational databases (RDBMS's). NoSQL technology was originally developed and used by internet applications such as Facebook, Google, Amazon and others which needs a DBMS that could read and write data anywhere in the world. Scaling and delivering performance across massive data sets and millions of users . In todays' time almost every organization has to use cloud applications that personalize their experience with their business and NoSQL is database technology for powering such systems. NoSQL databases systems offers key-value storage focuses on the scalability of data storage with high-performance, provides low-level access mechanism and very flexible for data modelling, and easy to update application developments and deployments [32].

VII. BIG DATA ANALYTICS: TOOLS

Big Data tools can be classified into three major categories [33] as follows.

A. Tools for Batch Processing

Batch processing is an efficient method of processing high volumes of data where a group of transactions is collected over a period of time known as batch. Data is entered, collected and formed into batches and then processed to get results. Hadoop perform batch processing. Batch processing uses distinct programs for input, process and output such as in payroll and billing systems.

In today's time only a fraction of the potential of repositories of Big data can be exploited using traditional batch oriented approaches but often value of data decays quickly and high latency becomes intolerable in some applications.

1) Apache Hadoop and Map/Reduce

Apache Hadoop is one of the most well-established system used for data-intensive distributed applications. It employs the computing method based on Map/Reduce. Apache Hadoop platform comprises of the Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS).Pioneered by Google and developed by Yahoo! The map/reduce [34] is a programming model and used for processing and generating large volume of data sets. Map/Reduce works on divide and conquer method and recursively breaks down a complex problem into many sub-problems until these sub-problems is scalable for solving directly. Then these sub-problems are assigned to a cluster of working nodes and are solved in separate and parallel ways. Finally, the results obtained of sub-problems are combined to produce solution to the original problem. By the addition of Map/Reduce the Hadoop works as a powerful software framework [35,36] for applications which process large quantities of data in parallel on large clusters which comprises of perhaps thousands of nodes of commodity hardware in a reliable, fault-tolerant manner.

2) Dryad

The another popular programming models the Dryad [37] which executes parallel and distributed programs and scale up processing power from a very small cluster to a large cluster. It uses dataflow graph processing [38]. Dryad infrastructure provides a platform for high performance distributed execution engine with good programme ability. A Dryad programmer firstly develops several sequential programs and connects them using one way channels. Numerous functionalities of Dryad includes generating the job graph, scheduling the processes on the available machines, handling transient failures in the cluster, collecting performance metrics, visualizing the job, invoking user defined policies and dynamically updating the job graph.

3) Apache Mahout

The Apache Mahout [39] is scalable and commercial machine learning technique for vast unstructured data set as intelligent data analytics. Many internet applications such as Google, Yahoo!, Amazon, IBM, Facebook and Twitter have implemented this technique in their projects. Core algorithms of Mahout's includes clustering, pattern mining, classification, regression, batch based collaborative filtering run on top of Hadoop platform through the Map/reduce framework [40].

4) Jaspersoft BI Suite

The Jaspersoft package is used to produce reports from database columns. This is a scalable Big data analytical tool. Jaspersoft [41] has a capability of fast data visualization on popular storage platforms such as MongoDB, Redis, Cassandra, Riak, and CouchDB. Hadoop is also represented using JasperReports and using a Hive connector to HBase as it can be integrated with all the leading Big data platforms. One important property of Jaspersoft is that it can quickly explore Big data without extraction, transformation and loading. As directly connected to main Big data stores provided with a SQL interface or not it explores large scale data by HTML5. Visualization is powered by a terabyte-scale, columnar-based in-memory engine which optimizes performance by pushing down query processing to the underlying data store if necessary.

5) Pentaho Business Analytics

Pentaho [41] is another platform for Big Data which produces reports from both structured and unstructured large datasets. Pentaho acts as platform for business analysis of Big data to provide professional services to businessmen with facile access, integration, visualization and exploration of data. Therefore, Pentaho can help businessmen make

data-driven decisions with positive effect. The underlying techniques include good security, scalability and accessibility. Similar to JasperSoft, this can also be piped to the most popular NoSQL databases such as MongoDB [42] and Cassandra [43]

6) Skytree Server

The first general purpose machine learning that is Skytree Server [41] is advanced analytics system developed to process huge datasets at high speeds. It provides many advance algorithms for machine learning. Being very user friendly this has five specific cases such as recommendation systems, anomaly/outlier identification, predictive analytics, clustering and market segmentation and similarity search. The Skytree is optimized to run many machine learning algorithms on Big data and handles structured and unstructured data from HDFS, RDBMS, flat files, common statistical packages, and machine learning libraries.

7) Tableau

Tableau [44] has three main components as Tableau Desktop, Tableau Sever, and Tableau Public to process Big data. Tableau Desktop visualize data shows it in a different but intuitive way. Tableau Server is a business intelligence engines provides browser-based analytics and Tableau Public creates interactive visuals. Tableau is also an embed Hadoop infrastructure. It also uses hive to organize the queries and cache the information for in memory analytics. Caching helps fast the processing inside a Hadoop cluster. Thus it provides an interactive mechanism to Big data applications.

8) Karmasphere Studio and Analyst

Karmasphere [41] is also based on Hadoop system and provides a novel technique to self-service in efficient, fast and collaborative way. Karmasphere is natively developed for Hadoop platform and provides users an user-friendly workspace to process their Big data applications. It has capability to discover business insight from large datasets including data ingestion, iterative analysis, visualization and reporting. Karmasphere Studio is a set of plugins built on top of Eclipse. In this well-designed integrated platform users can easily write and execute their Hadoop jobs. Karmasphere Analyst grows up rapidly the analytical process on Hadoop clusters. This also embeds Hive connector to process structured and unstructured data on cluster. SQL programmers, technical analysts an administrator of database experiment with Hadoop in graphical environment.

9) Talend Open Studio

Talend Open Studio [44] is an open source which provides users visual environment to process their analysis of Big data. Apache Hadoop has developed and involves HDFS, Pig, HCatalog, HBase, Sqoop or Hive. Users solve Big data problems without writing complicated Java code as required in Hadoop. By using users can build up their own tasks through dragging and dropping varieties of icons onto a canvas. Visual programming seems like a superordinate goal but the icons can never represent the mechanisms with enough detail to make it possible to deeply understand.

B. Tools for Stream Processing

In past years as efficiency is mandatory for any application to cope with large volume of data several distributed data processing systems have been evolved which are different from the batch processing and handle data as they arrive and thus acknowledging the ever growing importance of timeliness and velocity in Big data analytics.

1) Storm

Storm [41] is a scalable, fault-tolerant and distributed real-time analytical system to process streaming data in contrasts to Hadoop system being used for batch processing. Its set up is very easy and can be operated easily and also a reliable system. Storm is as efficient as process million instructions per second per node. Therefore it is used in interactive operation system, real-time analytics, on-line machine learning, distributed RPC continuous computation, and ETL. A Storm cluster is also similar to a Hadoop cluster whereas users run different topologies for different Storm tasks. A Storm cluster has two kinds of working nodes one master and several worker nodes. The master node and worker nodes implement two kinds of daemons such as Nimbus and Supervisor respectively. The two daemons are with same functionality in Map/Reduce framework. Nimbus is in charge of distributing code across the Storm cluster while scheduling tasks to worker nodes and monitoring the whole system. In case of failure in the cluster, the Nimbus will detect it and re-execute the corresponding task. The supervisor complies with tasks assigned by Nimbus, and starts or stops worker processes as necessary based on the instructions of Nimbus. The whole computational topology is partitioned and distributed to a number of worker processes, each worker process implements a part of the topology.

2) S4

S4 [45] is a general-purpose, distributed, scalable, fault-tolerant, pluggable computing platform for processing continuous unbounded streams of data [46,47]. It was initially released by Yahoo! in 2010 and has become an Apache Incubator project since 2011. S4 allows programmers to easily develop applications and possesses has several competitive properties including robustness, decentralization, scalability, cluster management and extensibility [48]. The core of S4 is written in Java. S4 also employs Apache ZooKeeper to manage its cluster, like Storm does. S4 has been in production systems at Yahoo! for processing thousands of search queries and good performances show up in other applications.

3) SQLstream

SQLstream [41] is another Big data platform designed for real time processing large-scale streaming data. It focus is on intelligent and automatic operations of streaming Big data. SQLstream is efficient to catch patterns from large set of unstructured data file, sensor, network or other machine generated data. The new release of SQLstream s-Server 3.0 performs better in real-time data collection, transformation, sharing and supporting real-time Big data management

and analytics. The standard SQL language is still used in the underlying operations. SQLstream works very fast due to in-memory processing which also termed as "NoDatabase" technology. The data will not be stored in the disks instead the arriving data are regarded as streams and processed in memory using streaming SQL queries. Streaming SQL is extension of SQL takes advantage of multi-core computing and performs parallel streaming data processing.

4) Splunk

Splunk is intelligent platform for exploiting real time information from machine generated Big data. It has adopted in many well-known companies such as Amazen, Heroku and Senthub. Splunk combines the cloud technologies and Big data to help users to search, monitor and analyse their machine generated data using a web interface. It presents the results in intuitive way using graphs, reports and alerts. Splunk provides metrics to many applications, diagnose problems for system and IT infrastructures. Splunk Storm [41] is based on cloud version of Splunk's Big data analytics and is is very different from the other stream processing tools. Its efficiency includes indexing of structured or unstructured machine generated data, real-time searching and reporting analytical results.

5) Apache Kafka

Kafka [49], developed at LinkedIn is a high-throughput messaging system. It is a tool to manage streaming and operational data using in memory analytical techniques for making real-time decisions. Kafka, a distributed publish-subscribe messaging system has four main functionality (1) persistent messaging with O (1) disk structures, (2) high-throughput, (3) support for distributed processing, and (4) support for parallel data load into Hadoop.

It has vast scope in number of reputed organizations as data pipelines and messaging tools. Activity data is the record of various online human actions such as webpage content, clicklist, copy content and searching key words. It is advised to log these actions into canned file and aggregate combine them for subsequent analysis. Operational data defines the performance of servers e.g. request times, CPU and IO usage, service logs, etc. The knowledge of operational data is helpful for real-time operation management. Kafka combines off-line and on-line processing data to provide real-time computing and produce ad hoc solution.

6) SAP Hana

SAP Hana [50] is Big data analytics platform which performs real-time analysis on predictive analysis, business processes and sentiment data processing. SAP HANA database is core of the real-time platform. It is somehow different from other database systems. Data warehousing, operational reporting and predictive and text analysis on Big data are three HANA specific real-time analytics. SAP Hana works with variety of applications either from SAP or not such as demographics and social media interactions.

C. Tools for Interactive Analysis Processing

The tools in the interactive analysis allow users to undertake their own analysis of information and directly interact with the computer in real time. The data can be reviewed, compared and analyzed in tabular or graphic format or both at the same time.

1) Google proposed an interactive analysis system known as Dremel [51] and which is scalable for processing nested data. Dremel having different architecture with Apache Hadoop is a successfully complements Map/Reduce based computations. It runs queries (aggregated form) over trillion-row tables in seconds by combining multi-level execution trees and columnar data layout.
2) Apache drill [41] is another system used for interactive analysis in similar manner to Dremel but with more flexibility to support varied query languages, data sources and formats.

VIII. CONCLUSION AND FUTURE SCOPE

Big data involves big systems, big challenges and big profits and so demands big research works to harness the goodness. The Big data analytics is still in the initial stage of development as very limited Big data tools are available to solve the real Big data problems completely. The results of big data analytics will help the business organizations in assessing the growth, risk and catching the non-profitable data in the beginning and find the ways to nullify it. Reasoning to the above this paper has detailed the concept of Big data, its issues and challenges in more detail to review the concept. This scientific paradigm demands more research to solve Big data problems and enhance its scope. From literature survey it is found that this field require more supporting advanced storage, architecture and input output techniques. More progressive platform for Big data infrastructure requires efficient data-intensive techniques such as cloud computing, social computing and biological computing and in future these technologies can optimize the big data analytics.

REFERENCES

[1] S. Kaisler, F. Armour, J. Alberto Espinosa, W. Money, "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conference on System Sciences, 2013.
[2] S. Madden, "From Databases to Big Data", IEEE, Internet Computing, May-June 2012.
[3] K. Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE , Aerospace Conference, 2012.
[4] S. Singh, N. Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.
[5] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, C. de Laat, "Addressing Big Data Challenges for Scientific Data Infrastructure", IEEE , 4th International Conference on Cloud Computing Technology and Science, 2012.
[6] M. Courtney, "The Larging-up of Big Data", IEEE, Engineering & Technology, September 2012.

[7] M.Smith, C. Szongott, B.Henne, G. von Voigt, "Big Data Privacy Issues in Public Social Media", IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.

[8] Innovation at Google " The Physics of Data,2012, http/ www.parc.com/ event/936 /innovation -at- google.

[9] Scalable, Energy-Efficient Data Centers and Clouds, 2012, http://iee.ucsb.edu/ Data_Center_Report.

[10] The International Technology Roadmap for Semiconductors, 2010, http:// www.itrs.net/.

[11] K. Asanovic, R. Bodik, B.C. Catanzaro, J. J. Gebis, P. Husbands, K. Keutzer, D. A. Patterson, W. L. Plishker, J. Shalf, S. W. Williams, K. A. Yelick, "The Landscape of Parallel Computing Research: A View from Berkeley", Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley, 2006.

[12] M. Shah, P. Ranganathan, J. Chang, N. Tolia, D. Roberts, T. Mudge, "Data Dwarfs: Motivating a Coverage Set for Future Large Data Center Workloads", Technical Report HPL-2010-115, Hewlett Packard Laboratories, November 8, 2010.

[13] L.Jay, B.Behrad, "Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics". IEEE Int. Conference on Industrial Informatics, 2014.

[14] W. Pedrycz, Andrzej, V. Kreinovich, "Handbook of Granular Computing", WILEY, 2008.

[15] M. A. Nielsen, I.L. Chuang, "Quantum Computation and Quantum Information", Cambridge University Press, 2009.

[16] A. u. R. Khan, M. Othman, S. A. Madani and S. U. Khan, "A Survey of Mobile Cloud Computing Application Models",IEEE Communications Surveys & Tutorials, 16(1) , 393-413, First Quarter 2014.

[17] E.E. Schadt, M. D. Linderman, G. P. Nolan, "Computational solutions to large-scale data management and analysis", Nat. Rev. Genet. 11(9) 647–657, 2010.

[18] L. Hutchinson, "Solid-State Revolution: In-Depth On How SSDS Really Work", Ars Technica (2012).

[19] A. Pirovano, A.L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, R. Bez, Scaling analysis of phase-change memory technology, IEEE Int. Electron Dev.Meeting (2003) 29.6.1–29.6.4.

[20] Josh, Bongard, "Biologically Inspired Computing", Journal of Computer 42(4) ,95–98, 2009.

[21] Nakano, "Biological Computing Based On Living Cells and Cell Communication", 13th International Conference on Network-Based Information Systems (NBiS),42–47, 2010.

[22] D. R. Simon, "On The Power Of Quantum Computation", SIAM Journal on Comput. 26,116–123, 1994.

[23] T. Winter and P. Thubert, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," in RFC 6550, 2012.

[24] Z. Shelby, K. Hartke, and C. Bormann, "Constrained Application Protocol (CoAP)", In Internet Draft: Intended status: Standards Track, 2013.

[25] P.Pebay, D.Thompson, "Design and Performance of a Scalable, Parallel Statistics Toolkit", IEEE International Symposium on Parallel and Distributed Processing, 1475–1484, 2011.

[26] Vikas C. Raykar, R.Duraiswami, "A Fast Algorithm for Learning a Ranking Function from Large-Scale Data Sets", IEEE Trans. Pattern Anal. Mach. Intell. 30(7) , 1158–1170, 2008.

[27] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, **26(1)** , 97-107, Jan. 2014.

[28] W. Tan, M. B. Blake, I. Saleh and S. Dustdar, "Social-Network-Sourced Big Data Analytics", IEEE Internet Computing,**17(5)** , 62-69, Sept.-Oct. 2013.

[29] Yan-Jun Liu, Chen, Wen, "Adaptive Neural Output Feedback Tracking Control for a Class of Uncertain Discrete-Time Nonlinear Systems", IEEE Trans. on Neural Networks, **22(7)**, 1162–1167, 2011.

[30] X. W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, **(2)**, 514-525, 2014.

[31] S. Simoff, M. H. B, A. Mazeika, "Visual Data Mining: Theory, Techniques and Tools for Visual Analytics", Springer, 2008.

[32] D. Thompson, J. A. Levine, P.P. Pebay, "Analysis of Large-Scale Scalar Data using Hixels", IEEE Symposiumon Large Data Analysis and Visualization, 23–30, 2011.

[33] J. Han, E. Haihong, G. Le, J. Du, 'Survey on NoSQL Database", 6th International Conference on Pervasive Computingand Applications, 363–366, 2011.

[34] J. Deam, S. Ghemawat, "Mapreduce: Simplified Data Processing On Large Clusters", Commun. ACM 51 (1) (2008) 107–113.

[35] C.Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems", IEEE 13th International Symposium on High Performance Computer Architecture, 2007, HPCA 2007, 2006, pp. 13–24.

[36] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Diane Cerra, second ed., 2000.

[37] M. Isard, M. Budiu, Y.Yu, A. Birrell, D. Fetterly, "Dryad: Distributed Data-Parallel Programs From Sequential Building Blocks", EuroSys '07 Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, vol. 41(3), 2007, pp. 59–72.

[38] H. Li, G. Fox, J. Qiu, "Performance Model for Parallel Matrix Multiplication With Dryad: Dataflow Graph Runtime", Second International Conference on Cloud and Green Computing, 2012, pp. 675–683.

[39] G.Ingersoll, "Introducing Apache Mahout: Scalable, Commercial-Friendly Machine Learning for Building Intelligent Applications", IBM Corporation (2009).

[40] R. Esteves, C. Rong, "Using Mahout for Clustering Wikipedia's Latest Articles: A Comparison Between k-Means and Fuzzy C-Means in the Cloud", IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom), 2011, pp. 565–569

[41] C.L. Philip Chen, Chun-Yang Zhang, "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data", Information Sciences 275 (2014) 314–347.

[42] E. Plugge, T.Hawkins, P. Membrey, 'The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing", first ed., Apress, 2010.

[43] E. Capriolo, "Cassandra High Performance Cookbook", Packt Publishing, 2011

[44] J. Brooks, "Review: Talend Open Studio Makes Quick etl Work of Large Data Sets", 2009.

[45] L. Neumeyer, B. Robbins, A. Nair, A. Kesari, "S4: Distributed Stream Computing Platform", IEEE Data Mining Workshops (ICDMW), Sydney, Australia, 2010, pp. 170–177.

[46] K.P. Lakshmi, C.R.K. Reddy, "A Survey On Different Trends In Data Streams", International Conference on Networking and Information Technology (ICNIT), 2010, pp. 451–455.

[47] Y. Mao, F. Wang, L. Qiu, S. Lam, J. Smith, "S4: Small State And Small Stretch Compact Routing Protocol For large Static Wireless Networks", IEEE/ACM Transactions on Networking 18 (3) (2010) 761–774.

[48] J. Chauhan, S. A. Chowdhury, D.Makaroff, 'Performance Evaluation of YAHOO! S4: A First Look", seventh international conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2012, pp. 58–65.

[49] A. Auradkar, C. Botev, S.Das, D. DeMaagd, A.Feinberg, P. Ganti, B. Ghosh L. Gao, K. Gopalakrishna, B. an Harris, J. Koshy, K. Krawez, J. Kreps, S. Lu, S. Nagaraj, N. Narkhede, S. Pachev, I. Perisic, L. Qiao, T. Quiggle, J.Rao, B.Schulman, A.Sebastian, O. Seeliger, A.Silberstein, B.Shkolnik, C. Soman, R.Sumbaly, K. Surlaker, S. Topiwala, C.Tran, B. Varadarajan, J.Westerman, Z.White, D.hang, J. Zhang, "Data Infrastructure at Linkedin', IEEE 28th International Conference on Data Engineering (ICDE), 2012, pp. 1370–1381

[50] S.Kraft, G.Casale, A.Jula, P. Kilpatrick, D.reer, Wiq, "Work-Intensive Query Scheduling for In-Memory Database Systems", 2012 IEEE 5th International Conference on Cloud Computing (CLOUD)", 2012, pp. 33–40.

[51] S.Melnik, A. Gubarev, J. J. Long, G. Romer, S.Shivakumar, M. Tolton, T. Vassilakis, "Dremel: Interactive Analysis of Webscale Datasets", Proc. of the 36th Int'l Conf. on Very Large Data Bases (2010), vol. 3(1), 2010, pp. 330–339.

