A Rigorous Investigation on Big Data Analytics

Kajal Rani and Raj Kumar Sagar

Abstract Nowadays Big Data becomes a new trend in science, technology, business, and marketing. Traditional data analytics techniques are not able to be applied straightforward toward big data. There is a requirement to developed high-performance platform that analyzes big data more efficiently. It is challenging for the organizations to unlock the patterns of information actionable value in massive volume of data, enabling great improvements in business and technical processes, customer analytics. Datasets are heterogeneous in granularity and accessibility. There are many issues and challenges that company faces while storing and handling Big Data. The skill to automatically store, organize, review, and analyze the data is essential. This paper will tell why there is need of big data? Why big data is such a big hype? What is the need of analytics? This paper describes need of Big Data and its analysis. Paper discusses a brief investigation on big data analytics. The use of tools like HADOOP, HIVE PIG, and SPARK in summarizing the data.

Keywords Big data analytics · Big data issues and challenges
Analytics techniques · Apache hadoop · Apache drill · Project storm

1 Introduction

Big Data is one of the most hyped business issues today. Big data refers to large volume of datasets produced by millions of users. Big Data analytics process uncovered unseen samples that help to make better decision. Big Data sets originated from multiple data sources in different forms. These datasets are varied of

K. Rani (✉)
Department of CSE, Amity University, Noida, Uttar Pradesh, India
e-mail: Er.kajalchauhan6apr@gmail.com

R.K. Sagar
Amity University, Noida, Uttar Pradesh, India
e-mail: rksagar@amity.edu

M. Pant et al. (eds.), *Soft Computing: Theories and Applications*,
Advances in Intelligent Systems and Computing 584,
https://doi.org/10.1007/978-981-10-5699-4_60

nature. Big Data comes in different formats. Traditional processing tools and technologies cannot cope with large datasets. Big Data Analytics defines the analysis of large collection of data that may be social media data, log files machine data, and enterprise data. New technologies are evolved to address large quantity data HealthCare, web traffic, enterprise data, sensor data, social data, business data, machine data, and global positioning system data. Researchers and traders purposed solutions to big data systems. Big Data is a proactive approach. Big data sets a point at which traditional technologies and tools are not sufficient for uncovered value or insights in cost-effective manner. Big data analytics model required to be re-evaluated. Big data has become very popular. Big Data analytics helps to gain more profit and productivity and improve efficiency of public and private sectors. Big data may be incomplete, inaccurate, and duplicate noisy. Here, we need to analyze these data. Big Data analytics indicates the starting of new form of technologies. Big data contains various features. Big data storage and management is very complicated task. Big data management handles large amount of data. New approaches, tools, and techniques need to be developed to analyze Big Data. Big Data is related to all aspect of human activity.

The definition of Big Data gives tools, set of methods, and technology to compare traditional data with Big data. Comparison is presented in Table 1. Big data is constantly update

It is semi-structured and unstructured data whereas traditional data is structured data. Data integration is easy in traditional database but difficult in big data.

This paper is organized as follows: Sect. 2 explains the features, characteristics of Big Data, discusses the motivation for adapting Big Data Analytics, and briefly highlights on Big Data Analytics Techniques; Sect. 3 focuses on Issues and challenges related with Big Data; Sect. 4 explains three open-source Big Data Analytics Framework and comparisons; and Section 5 is a conclusion to the study.

Table 1 Comparison between traditional data and big data

	Traditional data	Big data
Volume	GB	Constantly updated
Generated rate	Per hour, day	More rapid
Structure	Structured	Semi-structured or unstructured
Data source	Centralized	Fully distributed
Data integration	Easy	Difficult
Data store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch

2 Big Data Analytics Characteristics

In Big Data input, data could babble from web server logs, broadcast audios, social networking sites, mp3s of music web server logs and documents traffic flow sensors, scan of government documents, satellite imagery, the details of web pages, banking transactions PS trails, financial market data, telemetry from automobiles, etc. For the identification of information from Big data, it has to be analyzed through different aspects.

2.1 Big Data Characteristics

To simplify the characteristics of Big Data, the five V's of volume, velocity, veracity, value, and variety are mainly used to organize varied aspects of big data. They are different aspects which work as the lens that picture the behavior of data and the software platforms available to develop them. Most possibly you will compete with the every V's of Big Data characteristics with one another for the better analysis of input data.

Volume—The main area of interest of analytics is processing complex and large volume of data. Volume depicts the astounding challenge to traditional IT structures. The distributed approach of querying and measurable storage is performed using this feature. Large numbers of archived data, which can be in the structure of logs, are becoming difficult for the companies to process it as they lack that ability.

Velocity—The significance of the velocity of the data is the pace at which the more and more data is generated. Immediately, data is provided to user when required. In the 8 years of 2005–2013, the digital universe spreads out from 130 million to 40 trillion. The use of smartphones has increased the rate of streamed data flow. The velocity is not just the problem but also the storage of streaming fast data inflow for later batch processing. There are two main reasons for the consideration of processing of streaming. First one is when the storage of input data in their loyalty is fast enough this implies that some level of assessment should appear as the streams of data for the storage requirements to become practical. The second reason for the consideration of streaming is the range of data originated from various resources is batch to real time.

Variety—It is least expected that data portrays itself in a much organized manner which is ready for processing. The big data of source data is diverse, and it does not lay itself into the appropriate relational structures [1]. The structures can be in a form of text from social networking sites, a raw feed directly from a sensor source and an image data from the websites [2]. This data does not come prepared for the integration into an application varied browsers send a number of data to users having information perhaps using varying software versions to exchange information with you, and there will be inaccuracy if human involvements is there. Flaws and inconsistency will also immerge.

Veracity—Due to the increasing rate of data, there is no time to spend in the cleaning of the massive data before using it. Analyzing of the data for the business process to enhance the productivity and decision making the organization requires a mechanism that should deal with the imprecise data. The processed and unprocessed, the clean and unclean, and the precise and imprecise combine to form big data.

Value—By processing and analyzing high-volume data with increased velocity, variety and veracity, here comes the need to unlock the hidden patterns to uncover the useful information or data for the business organizations to understand the habits of consumers and to facilitate the requirements of them. This unlocked data represents the actionable value of big data. The actionable value of data is that factor that can transform their business processes from the strategic view in revenue (Fig. 1).

2.2 Motivation for Big Data Analytics

Traditional technologies and tools are not satisfactory to reserve and collect large volume of datasets. Statistics [8] shows data grows rapidly. Analytics of Big data is the process that is used to collecting, storing, and analyzing high-volume high-velocity heterogeneous data. Big data Analytics process helps to extract hidden useful patterns and meaningful insights. Big data analytics allows to extract valuable insights it provide better understanding of customer behavior and understanding. Further, the result is used for decision making which helps in business grow, online browsing, online business, social media, network weather forecast, and telecom. Big data Analytics can help to make better decision it helps in future

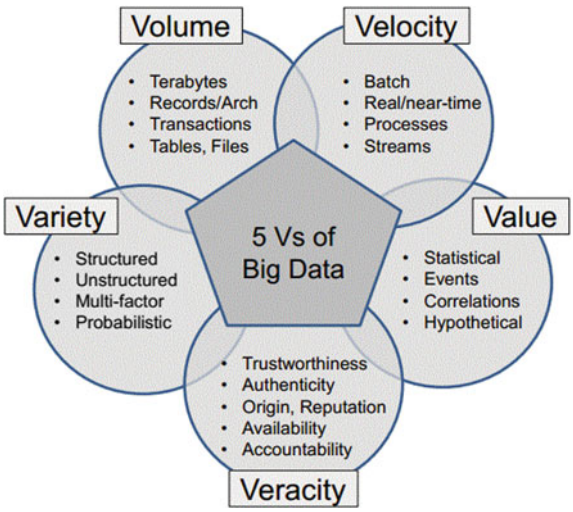

Fig. 1 5 V's of big data [8]

prediction of sales, in predicting customer demands. Due to heterogeneity of data, analytics become difficult. Analytics can be divided into three categories predictive analytics that uses statistical data, descriptive analytics that contain historical data which is related with business intelligence, prescriptive analytics is used to find out the optimized solutions for the concerned problem.

Big Data Analytics in Healthcare—Big data analysis gives important insight into medical field. Hospital and medical researchers store large- and high-volume data about patient's historical state, current state data, medication, and other details. Drug manufacturing firms store large volume of data. Analysis of Big data helps in both public and private medical services. In health sector, Big Data analytics help to detect which department needs to reorganize. The analysis of Big Data supports medical decision makers. Analysis of Big data helps to monitor and assess quality of medical services. Medical Data analytics helps to provide new life to patients gives valuable insights and discovery of new solutions to the doctors. It helps in measure the performance of medical units. It offers decision support tools and reduces high cost of medical sector. Analysis of Big data helps doctors and practitioners to find new solutions and ways to genetic and hereditary attacks.

Big Data Analytics in Intelligence Services—Many intelligence firms collects high-volume data from heterogeneous sources such as web sources, sensor data, publicly available sources, social networking websites, signal intercepts they make analysis based on these gathered data. Linking and connecting all the data and information makes the available information discovered. Thefts can be detected. Sound Analytics technique can handle high volume of unstructured data.

Environments and Big Data Analytics—Perception of Environment conditions in a better way needs large size of data from various sources as water being monitored using sensors and air quality system, mixture of gasses present in environment. These examples shows that adaptation of technologies, methods, tools, and frameworks to provide better valuable insights hidden from data sources.

Big Data Analytics and Marketing Campaigns—By analyzing Big Data Organizations help to discover new insights and valuable information that is present in large amount datasets. Insights can be used for predicting Customer behavior. Customer gives their reviews and ratings to the product, so Organizations can analyze this big amount of data to increase profits and values. Organizations can forecast the behavior of customers. Big Data Analytics lead to targeted market analysis. Organizations can developed new strategy and provide high-level satisfaction to the customer.

2.3 Big Data Analytics and Techniques

Analytics of Big Data uses advanced technologies. Big Data analytics process analyze high amount datasets and discover hidden patterns, meaningful information, insights, market tendency, recognized relationships. Analytics process is useful to make better decisions. It provides business benefits, business values, new

strategies, and enhanced efficiency. Big Data Analytics uses wide variety of modern techniques (Table 2).

Blackett et al. Categorized Big Data Analytics into three categories: predictive analytics, descriptive analytics, and prescriptive analytics. All categories give valuable insights and profits to the organizations. In Big Data System, the first step is to capture lots of data and information from various sources.

Predictive Analytics—Predictive analytics predicting future trends based on statistical techniques. The nature of predictive analytics is probabilistic. Predictive analytics can only predict and forecast future probabilities. Predictive analytics uses logistic and linear regression to predict the future trends and outcomes. It can only forecast what might happen in the future. Regression and logistics techniques are used to extract patterns from big datasets. Predictive analytics uses data mining, statistical methods, and machine learning to make predictions about the future trends. Predictive analytics gives answer what will happen.

Descriptive Analytics—Descriptive method is the simplest method of analytics. The purpose of descriptive analytics summarizes historical data to what come about. This analytics answers the questions what happened? Descriptive analytics is connected with business intelligence. Descriptive analytics looks historical data and understands the reason behind success or failure. Marketing operations and sales uses descriptive analysis.

Prescriptive Analytics—Prescriptive analytics is the third business analytics. It is useful for decision making. Prescriptive analytics summarizes the big data, business rules, computational science, and then make predictions to business analytics. Business analytics adopts these predictions to make better profits and satisfying the customers. Prescriptive analytics take new current data every time. Prescriptive analytics improve accuracy and provide better decisions. Prescriptive analytics takes hybrid datasets as input and prescribe how to take advantages of this to predict future.

Table 2 Big data analytics techniques

Big data analytics	Techniques
Sql analytics	Count, mean, OLAP
Descriptive analytics	Univariate distribution, central tendency, dispersion
Data mining	Associations rules, clustering, feature extraction
Predictive analytics	*Classification, regression, forecasting, spatial, machine learning, text analysis*
Simulation	Monte Carlo, agent-based modeling
Optimization	Linear optimization, non-linear optimization

3 Issues and Challenges Related with Big Data Analytics

Big Data is diverse in nature, Organizations face challenges companies and organizations, and traditional systems are facing problems to store and analyzing hybrid data to make useful decision. This is a challenging task to companies and organizations. Big data analysis needs to store efficient data and requires queries to large datasets. Big data contain heterogeneity and incompleteness. Data comes from different data sources while machine language algorithm works on homogenous data. Big data challenges will be difficult to resolve as data is generated continuously. There is always a need of efficient, cost effective, appropriate tools and technologies. Some challenges are given below.

Privacy, Security—Big data is new technology trend: Privacy and security are two important aspects of any organization. Mostly, information present in such datasets is important. So there is a need of security mechanisms like encryption to secure the data from unauthorized external sources. Traditional resources are not sufficient to dealing with Big Data to ensure privacy and security. The analysis of Big data would make the system safer. Advanced techniques need to develop in terms of infrastructure, application, and data.

Big Data Management and Sharing—Big Data Management is important for any organization and companies to build new business model. Big Data management helps organization to organize the data and uses this data for future purposes. Analytics of Big Data extract meaningful insights from the use of Big Data management. This managed data is used by the companies to gaining more profits in business and market. Big Data Analytics contain large dataset. The datasets should be detectable, approachable, and usable. The agencies must persist to privacy laws, thigh having these urges. The present tendency toward the open-source datasets has noticed an importance on the development of such datasets ready for access of the public. The agencies should pay some more concern on making data conventional, standardize. It always permits them to use it and for the collaboration of privacy laws to the maximum extent possible.

Analytical Skills and Technology—Big data Analytics has given a maximum effort on ICT provider to develop new technology and tools to handle more complex datasets. The current technologies and tools are not able to process, store, and analyze huge volume of discrete datasets. Developers and vendors of such a huge data system suggest some solutions for making reliable effective, useful tools to reduce the complexity of huge datasets. Big data analytics helps to discover tools for Big Data integration and tools for manage resources.

Data Representation—Big Data sets comes from heterogeneous sources. Big data is varied in nature. Big Data represents huge amount data that is found in structured and unstructured forms. Datasets are heterogeneous in granularity and accessibility. An Efficient Data Representation should be designed to contemplate the hierarchy, collection of such huge amount of data. Integration technique should be designed in such a manner that performs operations efficiently.

Big Data Quality—As Big Data grows, data quality is important and challenging. Currently, quality assessment standards and methods are lacking. Data quality is an important aspect. Data quality defines as a set of quality attributes. Many factors affect data quality at different levels of processing. It is very difficult to find out manually generated data quality errors. There is a need to maintain data integrity. This includes the following:

How can be data preprocessed in order to improve quality of the data and result?
How can be data preprocessed so as to improve efficiency?
How to confirm the integrity of data?
How to calculate the worth of information in large datasets?

Complexity—Big Data deals with large records of structured and unstructured data, and this large datasets contains complex relationship between them it increase complexity of Data. Data records contain complicated relationships. As data grows rapidly, more data increases more complexity. Incompleteness is also a big challenge to deal with Big Data Analytics.

4 Analytics Frameworks for Big Data

Big Data required several types of frameworks to run several types of data analytics. A large number of different frameworks available to processing enterprise data.

Batch Analytics for Historical Data—Map Reduce and associate technology are very useful for batch analytics on Big data. Map Reduce is a framework using which we can write application to process huge amounts of data. Map Reduce is also a programming model. In this every data processing is divided into Map Reduce step. Map Reduce works in parallel fashion. Map Reduce framework breaks large data set into smaller ones and performs parallel processing which are executed on slave nodes. Map Reduce computes nodes independently. Map reduce contains master node (job tracker) and slave node (task tracker). Map Reduce includes several phases of process such as input phase, map phase, shuffle, and reduce phase. The result of map phase is intermediate key value pair. Map Reduce is not for random data. It works on sequential data. Map Reduce is fault tolerance framework. Map Reduce framework plays an important role in Big Data Analytics. Batch processing is widely adopted, and it gives faster response to real-time applications [3].

Stream Processing for Current Online Data—Stream processing takes input in the form of stream data with storm being a representative framework [4]. It is used for online analytics. It takes infinite data size as input. Data quality is an important aspect. Data comes in streams. It requires only few passes over streams to find approximation results. It provides result as quickly, and it takes milliseconds. Data stream processing is highly active technique.

- Interactive ad hoc queries and analysis with apache drill.

4.1 Apache Hadoop

Big Data deals with Apache Hadoop. It is available publically and open-source framework with no licenses fees that enables distributed processing of large volume of data. Apache Hadoop is basically based on Map Reduce programming model. Apache Hadoop supports java language. Apache Hadoop is highly scalable. Apache Hadoop supports distributed file system. IT companies and Organizations need to analyze huge volume data set they need more advanced technologies apache hadoop provide solution to the organizations. As Hadoop has become popular platform, it assures high availability at application layer [5]. Apache Hadoop includes listed sections.

Apache Hadoop includes listed sections:

(a) HDFS: It provides high throughput. HDFS supports large number of datasets. HDFS supports Master Slave Design.
(b) Hadoop core: It contains common utilities that support other modules.
(c) Hadoop YARN: Yet Another Resource Negotiator: framework for scheduling the job and cluster management. Hadoop Map Reduce model: programming model for massive volume of datasets.

Figure 2 represents how to manage data with Apache Hadoop. Here, user submits a query. Apache Hadoop framework is fault tolerant and robust. Hadoop is based on master (name node)/slave (data node) architecture. Master node manages file system policies and namespace and provides access to files. Job trackers

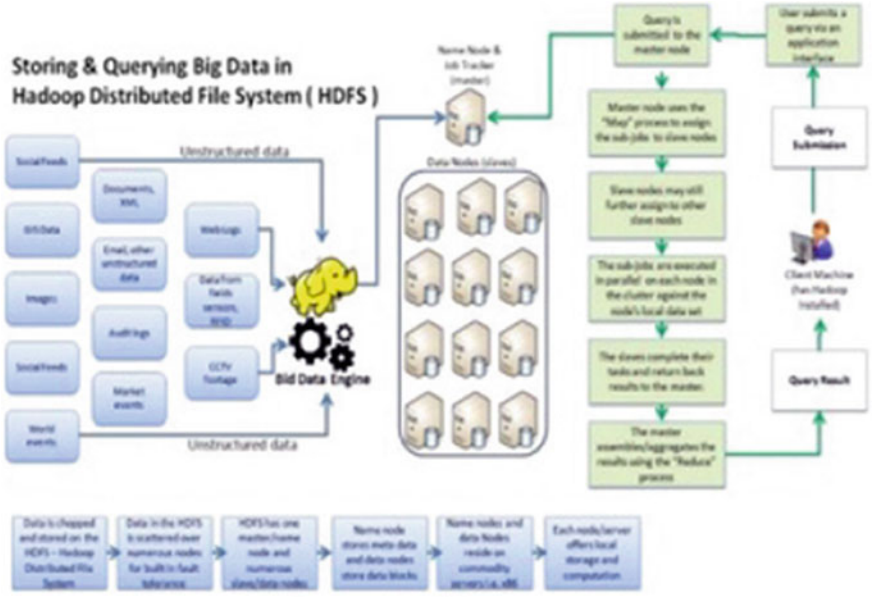

Fig. 2 Data store and retrieval operation in Apache Hadoop [9]

assigning the jobs to data nodes, data nodes are responsible for read and write operation from client sites. The data nodes also execute creation, deletion, and replication operations. The name node performs open, close, and rename operation. Data node stores each file, which is split into a sequence of blocks.

4.2 Storm Project

Hadoop can process and store large amount of data that previously unthinkable by the use of its related technologies have made it possible to process and store huge amount of data. Storm provides real-time analytics. Storm is fault tolerant and very simple to use. Storm processes streams of data. Processing of real-time data at such an immense scale going to be big requirements for business. Storm defines rules for real-time computation like how Map Reduce can make it more ease in writing of parallel batch processing. Parallel real-time computation can become easy using storm's rule [4].

Figure 3 shows storm cluster architecture. A storm cluster is pretty likely to Hadoop cluster, but on Hadoop cluster, you can run map reduce jobs on storm, and you run topologies, jobs, or topologies that are very different [4].

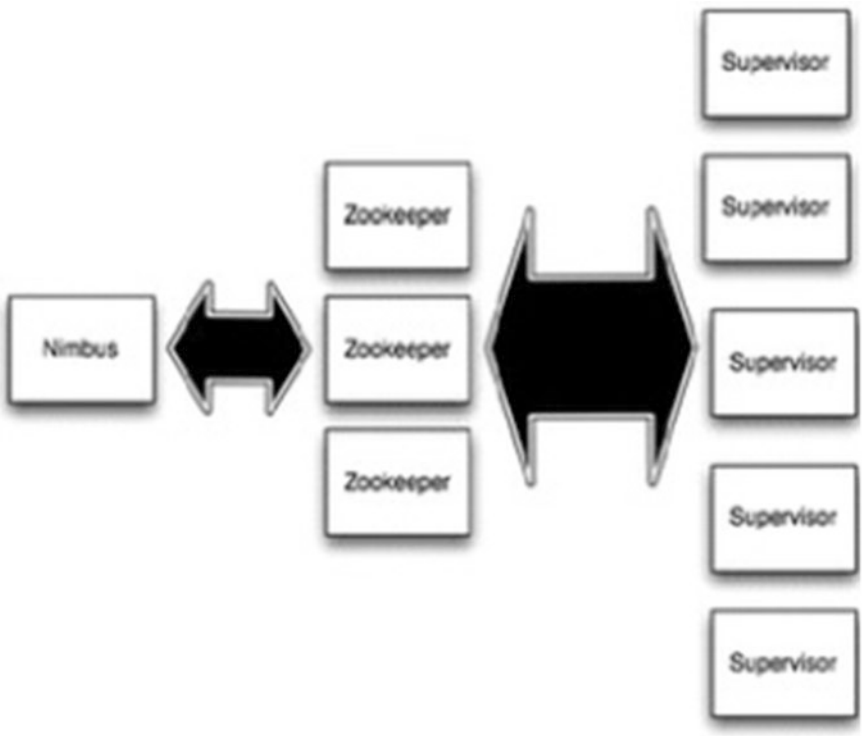

Fig. 3 Architecture of storm cluster [10]

Basically, nodes are divided into two types in storm cluster: (1) master node and (2) worker nodes. The master node runs a demon thread, i.e., "Nimbus," which is similar to Hadoop's "job tracker." Nimbus is mainly responsible for distributing code across the cluster and assigns tasks to different machines and monitor for failures occurring. Each worker node executes a daemon known as the supervisor. The supervisor takes account for work given to its machine and performs start and stop worker processes as required based on what Nimbus has assigned to it. Each and every process runs a part of a topology, a topology that is working and has many worker processes spread around many machines. All coordination between Nimbus and the supervisor is done through Zookeeper [6]. The Nimbus daemon and supervisor daemons are failing fast and stateless. All state is kept in zookeeper or on local disk. The design leads to storm clusters being incredibly stable.

4.3 Apache Drill

Apache Drill is a user- and developer-friendly software. Drill is an important project of Apache. Apache Drill is open-source software with no licenses fees. Apache Drill is available publicly. Apache Drill is a query engine. The Apache Drill is simple to use and scale to petabytes of data. Apache Drill provides connectivity to many data stores.

Figure 4 shows the architecture of Apache drill. It contains following parts.

User—it provides command line interfaces, and JDBC and ODBC and RESET interfaces for human.
Processing—allowing for pluggable query languages.
Data sources—pluggable data sources either local or in cluster setup, providing data sources.

Apache Drill is mainly focused on non-relational database and ad hoc queries. Apache Drill supports RDBMS. It can deal with multiple databases and file formats. Drill is capable to process millions of records in seconds. Apache Drill supports NoSQL databases. Apache Drill supports ANSI SQL standards. ANSI SQL can be used to get better result quickly. There is no requirement to define any schema. Apache drill is a query layer that works with underlying multiple data sources. Apache drill provides a flexible query execution framework that enables quick aggregation of statistics to explore data analytics. Big data analytics has become more accessible to a big range of users. Many times user need to run ad hoc queries against business application. Apache drill provides the solutions for that kind of issues [7].

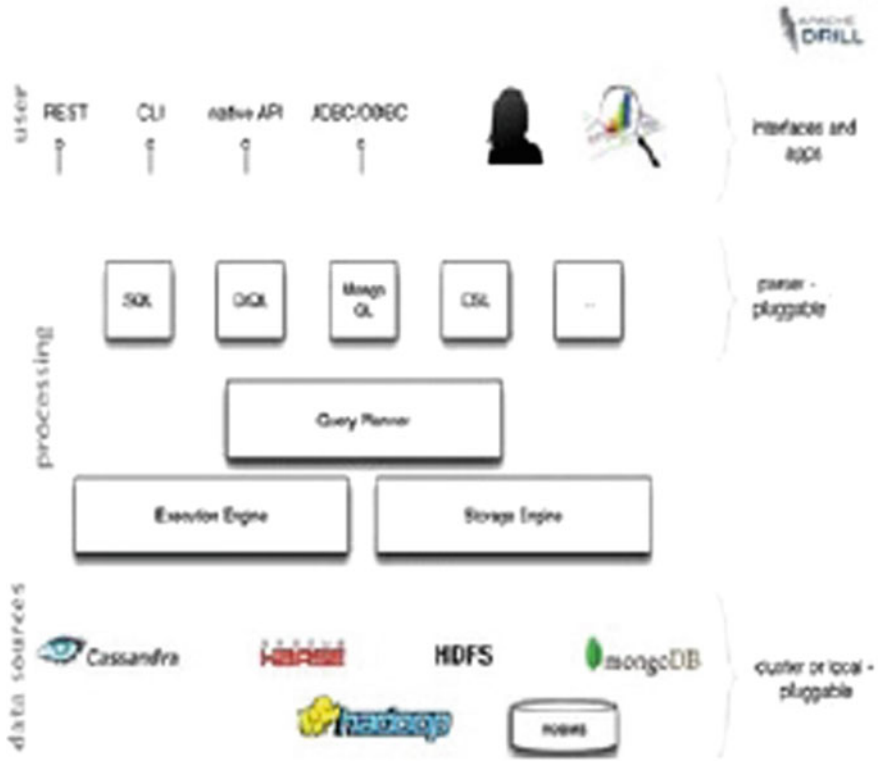

Fig. 4 Architecture of Apache drill [7]

5 Conclusion and Future Scope

Detailed study and analytics of Big Data has been accomplished, and comparisons between different frameworks are described below.

As Table 3 represents that Apache Hadoop is suitable for batch processing. Apache Hadoop has high latency. Hadoop can absorb any size of data. Apache Hadoop is more scalable and cheaper. Hadoop works on Map Reduce model. Project storm works with real-time computation and stream analysis. Project storm is easy to handle and use. It provides fast processing and process unbounded stream of data with high velocity. Storm is not suitable for batch processing of high-volume data. Apache drill is complex. It works well for interactive and ad hoc applications. Apache drill has low latency.

Table 3 Comparisons between big data analytics frameworks

Features	Apache Hadoop	Project storm	Apache drill
Owner	Community	Community	Community
Workload	Batch processing	Real-time computation, steam analysis	Interactive and ad hoc analysis
Source code	Open	Open	Open
Low latency	No	Yes	Yes
Complexity	Easy	Easy	Complex

5.1 Future Scope

Cost-effective and efficient tools are needed to analyze Big Data sets in real time. Data is generated continually. Here is a need to develop Big Data analytics frameworks for IT companies and Organizations. Data Analytics is hastily changing area. Traditional techniques need to revolution that can cope with Big Data. As data grows continuously, it is required to revisited current policies and system to manage large, hybrid datasets. Different frameworks need to analyze that practices with Big Data.

A new set of integration techniques should be designed. Traditional approaches are sufficient to extract value from huge amount of Data. Modified paradigms are required to develop.

Many organizations and companies face many challenges to cope with diverse Big Data. They need better tools and technologies to make better business decision.

For batch processing, it is not easy to adapt hastily growing data volume and real-time requirements. A big amount of time is wasted during batch processing transmission. Data Analytics provides new opportunities for real-time applications.

Security and privacy is also an important challenge for organizations dealing with Big Data. Big Data Analytics should find effective data control and quality mechanism. Big Data Analytics required facilitating new data processing techniques.

Big Data Analytics research needs more attention. Research on Big data can create more benefits for Business and organization. Big Data Analytics build a robust model that is able to analyze all size of data.

References

1. StructuredData: http://www.webopedia.com/TERM/S/structured_data.html
2. Semi structured Data: http://en.wikipedia.org/wiki/Semistructured_data
3. Apache-Hadoop: http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F
4. Project Storm: http://storm-project.net/

5. Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December 2004
6. Apache Zookeeper: http://zookeeper.apache.org/Big%20Data%20statistics-wikibon.org/blog/big-data-statistics
7. Hausenblas, M., Nadeau, J.: Apache Drill ad-hoc interactive analysis at scale, June 2013
8. Characteristics of Big Data: http://www.datatechnocrats.com/tag/bigdata/
9. Storing and querying data Big Data in HDFS: http://ecomcanada.wordpress.com/2012/11/14/storing-and-querying-bigdata-in-hadoop-hdfs/
10. Storm cluster: https://github.com/nathanmarz/storm/wiki/Tutorial
11. Katal, A., Wazid, M., Goudar, R.H.: Big data: Issues, challenges, tools and good practices. In: Sixth International Conference on Contemporary Computing (IC3) (2013)
12. Stephen, K., Frank, A.J., Alberto, E., William, M.: Big data: Issues and challenges moving forward. In: IEEE, 46th Hawaii International Conference on System Sciences (2013)
13. Conference on Communication, Information & Computing Technology (ICCICT), 19–20 Oct 2012
14. Michael, K., Miller, K.W.: Big data: New opportunities and new challenges. IEEE Technol. Soc. Mag. **13**
15. Sergey, M., Andrey, G., Jing Jing, L., Geoffrey, R., Shiva, S., Matt, T., Theo, V.: Dremel: Interactive analysis of web-scale datasets. Google (2013)
16. Apache-Dri: https://cwiki.apache.org/confluence/display/DRILL/Apache+Drill+Wiki
17. Apache HBase: http://hbase.apache.org/