International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Big Data Knowledge Discovery Platforms: A 360 Degree Perspective

Neelam Singh, Devesh Pratap Singh, Bhasker Pant

Abstract: ***Big Datais a buzzword affecting nearly every domain and providing different set new opportunity for the development of knowledge discovery process. Although it comes with challengeslike abundance, extensiveness and diversity, timeliness and dynamism, messiness and vagueness, and with an uncertainty as all the data generated does not relates to any specific question and can be associated with another process or activity. To address these challenges are certainly cannot be handled by the traditional infrastructure, platforms and frameworks. New analytical techniques and high performance computing architecture came into picture to handle this explosion. These platforms and architecture are giving a cutting edge to the Big Data Knowledge Discovery process by using Artificial Intelligence, Machine Learning and Expert systems. This study encompasses a comprehensive review of Big Data analytical platforms and frameworks with their comparative analysis. A Knowledge Discovery architecture for Big Data Analytics is also proposed while considering the fundamental aspect of gaining insights from Big Data sets and focus of this analysis is to provide the open challenges associated with these techniques and future research directions.***

Keywords: Big Data, Knowledge Discovery, Artificial Intelligence, Expert Systems.

I. INTRODUCTION

Dataexplosionhas initiated Big Data phenomenon. The term “Big Data” originated into picture, in relation to present context, in the late 1990s,“Francis X. Diebold” in his first paper “Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting” in the year 2000 (published in 2003) marked the beginning of the much sought after topic of today namely “Big Data” although the acclaim of using the term is credited to John Mashey, the chief scientist for SGI, in a Silicon Graphics (SGI) slide deck through the heading of "Big Data and the NextWave of InfraStress". We are witnessing the Big Dataperiod, the issue here is not getting data but accurate data and deploying computing powers to boost our domain knowledge and also torecognize patterns that cannot be classified or exploredformerly.“Big Data” is identifiedas a phenomenon in which the traditional functional abilities of enterprises has become less effective and scalable to store, process, analyze and visualize the data.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Neelam Singh*, Assistant Professor, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun (Uttarakhand) India.

Dr. Devesh Pratap Singh, Professor & Head of Computer Science and Engineering department, Graphic Era Deemed to be University, Dehradun (Uttarakhand) India.

Dr. Bhasker Pant, Dean Research & Development and Associate Professor, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun (Uttarakhand) India.

Big Dataencompasses the gathering and dispensation of outsized data sets and relatedarchitectures and proceduresrequired to evaluatethem. Big data architectures comes in variety of paradigm spanning across multiple machines as cluster or distributed in nature with specialized processes to handle knowledge discovery process.

The integration of knowledge discovery process with Big Datadriveopens a range of unique opportunities for organizations in terms of future strategy, getting a competitive edge and many more. Yet, Big Data comes along with unidentified and distinctive architectural and algorithmic challenges.

Knowledge Discovery from Data (KDD) can be defined as a collection of processes integrated to excavatenovelfeatures and knowledge from multifaceted datasets. KDD is an interdisciplinary domain spanning its wings across BioInformatics, Astronomy, Computer Science, Statistics, IoT, Recommender Systems to name a few. Tools and techniques for Knowledge Discovery are taken from paradigms including distributed programming, machine learning, statistical inferences, visualization and high performance computing.

Colossal data sets i.e. Big Data comprises of hidden pattern and knowledge which is likely to be discovered from, knowledge discovery in databases (KDD) process, which conventionally performs data selection, preprocessing, subsampling, conversions, pattern discovery, post-processing and knowledge exploitation in a chronological order. Areas like business intelligence, medicine, bioinformatics, military, education and research are highly influenced and benefited by the application of data mining techniques. Advancements in this area like in classification, pattern matching has increased the potential to acquire domain specific unexplored knowledge and value.

II. LITERATURE REVIEW

The “National Institute of Standards and Technology (NIST)” [1] suggests that, “Big Data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant horizontal scaling for efficient processing.”

Big Data is accumulated from heterogeneous data producing sources. Like a smart wearable that produces the number of steps a person has walked throughout the day, along with a statistics of kcal burnt, heart rate, average speed and other activities like cycling, swimming etc., terabytes of data being produced bythe planned square kilometer array telescope. Petabytes of data is being accumulated and created by social networking sites like twitter, by scientific experiments and by sensors every day [2]. Owing to its given inherent characteristics Big Data pose the following challenges:

A. Variety and Heterogeneity

As stated by Han Hu et al. [3] variety and heterogeneity obstructs the path to achieve scalability from a Big Data system. Variety and heterogeneity of data poses problem to effectively gather and assimilate data from contrasting distributed systems with proven scalability. Real time and near-real time data assimilation, modeling, optimization and visualization is another core area to deal with.

B. Volume

Considering the volume characteristic, the inundation of data as input is the principalpoint that is much of a prime concern as it may hamper the data analytics process to a very large extent. Unlikecustomary data analytics, forevaluatingwireless sensor network data, the biggest blockage is to communicate this assimilated data to further layers for storage, processing etc as stated by Baraniuk [4].

C. Velocity

Chun-Wei Tsai et al.[5] stated that, real-time or streaming data comes with a rapid flow in a very short duration which is hard to be handled by the conventional systems and tools. This data with high velocity is difficult to control and manage and thus can create a bottleneck in the process of data analytics and thus will obstruct the efficiency drastically.

D. Value

Value refers to the quality and provenance associated with Big Data. The main impediment here is to recognize, extract, transmute and analyze this information to find the hidden value from it [5].

Currently, the available data requiring to be analyzed can be data at rest, streaming data, near-real-time data or real time data, which is not only huge but also comprehends heterogeneous datatypes[6]. Big Datacomes withdistinctivetraits of being “massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous,” and canalter the result from statistical inferences and data explorationmethodologies [7].

It may encompass more imprecise or inconsistentdata. For illustration, multiple digital ids of a single entity, or service or an account type used by multiple users, severly affecting or degrading the precision of the mining results [8]. Blackett et al. [9] categorized data analytics under three categoriesdepending on the complexity of analysis: Descriptive analytics, predictive analytics and perspective analytics.

Big Data may be created from a wide array of heterogeneous sources like smart wearable devices, handheld devices like smartphones, social networks, RFID, sensors and Internet of Thingsand applications that showcase thestructures like velocity,volume and variety. Taking this into account, the entire data analysis method needs to be re-inspected[5].

E. Non-scalable

Most of the customary data analysis proceduresweren notintended to consider large-scale and composite dataset. These analysis methods are not scalableenough because of their architectural limitation and boundationfailing to handleoutsized or composite datasets.

The strategy and methodology adopted by traditional data analysis methods works on the objective that all the computation and processing will be saturated typically to a particular machineassumingwhole of the data to be in memory for the data analysis progression. Considering this point it can be assumed that volume of Big Data restraints the performance of conventional data analytics.

F. Non-dynamic

Dynamic scalability that is to get adjusted according to the situations is not a property of most conventional data analysis thus limiting them toinvestigate the input data on-the-fly or real time streaming data. For example, in a classification model the classifiers are customarilystaticand they cannot be automatically transformed.

A technique called“incremental learning” [10] was introduced which focus on the dynamic adjustment of the classifiers required in a training process in spite of limited resources. As most of the conventional analysis methods are non-dynamic in nature they cannot tackle with the Big Data velocity issue.

G. Uniform data structure

All the conventional data mining techniques follow a predefined schema and format, but in case of Big Data there is feature called variety, wherein data comprises of different schemas and format. Also data suffers from issues like incompleteness, missing values or noisy. So most of the techniques fails or gives unfamiliar results while handling Big Data.

The literature work was examined for Big Data architectures. The findings reveals that Big Data architecture contains the following components:

- Hardware and software modules, building and development standard and unsurpassedmethodologies. All sources outlines the fundamental of a Big Data architecture along with the given modules also required a parallel processing engine likeHadoop, a file system(e.g. distributed file system like HDFS), and usually a NoSQL database (e.g. HBase). Also the architectures comprises of data sources, data mining procedures, synchronization and formation engines, database, observing etc.

The following modules have a place in majority of the literature that describes a Big Dataarchitecture [11]:

- A querying engine;
- A predictive analytics engine;
- A statistical analysis or machine learning engine;
- A data importing / collecting / ETL engine;
- A real-time/stream/complex event-processing engine.

Several architecture principles like e.g. SAS, Kimball etc. aims at “Close-to-source data processing”, which states that data should be analyzed as early as possible to reduce storage costs and processing time. MicroStrategy believes in retrieving and storing as many data as possible and performing analytics at a relatively late stage [11].

- Big Data analytics is been supported by various analytical platforms provided by different organizations being categorized as[12] (1) Processing or Compute: examples include “Hadoop” [13], “Nvidia CUDA” [14], “Twitter Storm” [15], (2) Storage including frameworks like “Titan or HDFS”, and (3) Analytics: “MLPACK” [16] or “Google BigQuery”.

There are various viabletools and architectures which focuses on finding methods and techniques to effectively and efficiently find meaning from the large pool of data collected through a huge array of resources.

- Big Data processing requires architectures which supports parallel programming model like MapReduce which is a parallel batch-oriented computing model, as a single machine may not work proficiently on Big Data.MapReduce is implemented on machine learning and data mining algorithms[17]. Public availability of these architecture can be provisioned using cloud based platforms in the form of services.
- Optimizing model parameters or to get solution statistics Data mining algorithms scans through the testing data. Chu et al.[18] suggested a general-purpose parallel programming implementation on multi-core processorsthrough MapReduce methodology to improve the efficiency of machine learning procedures and devicedtentypical data mining algorithms through the structure, namely k-Means, naive Bayes, locally weighted linear regression,logistic regression,linear SVM, Gaussiandiscriminant analysis, the independent variable analysis, expectation maximization and back-propagation neural networks.
- Gillick et al. [19] proposed an improvement in the execution mechanism of MapReduce in Hadoop, along with evaluating the algorithms' performance ofsingle-pass learning, query-based learning and iterative learning in the improvised framework, theirstudy introducedmethodology for data sharing amongst computing nodes playing their role in parallel learning algorithms and to manage storingdata in distributedarrangement andprovided results using medium-sized clusters for large scale data mining acitivities for the MapReduce mechanisms.
- A multi-core and multi-processor system based on parallel programming paradigm following MapReduce named Phoenix was given byRanger et al. [20], and comprehended three data mining algorithms on the framework i.e. k-Means, PCA, and linear regression.
- Das et al.[21] performed a review on the in-depth assimilation of "R (an open source statistical analysis software)" with"Hadoop" to push data computation to parallel processing, in order to tackle the scalability parameter of customary analysis software and Hadoop. Nawsher Khan et al.[22] studied and outlined the limitations of hadoop by considering feature like data redundancy, framework complexity, sql support and optimization.

III. METHODOLOGY

Big Data Knowledge Discovery requires an intricate study and knowledge of architecture, tools and technology that can focus and target the underlying characteristics specific to Big Data. In this paper we adopted the given methodology to tackle this aspect:

Phase 1(Section IV):

- Anextensivequalitative study of Big Data Analytics frameworks and platforms were conducted on the basis of various parameters like Data models, structures &type, Big Data Management, Big Data analytics & tools, Big Data Infrastructure and Big Data security.
- Architecture, Frameworks or Models reviewed are MPI (Message Passing Interface[24]), Map Reduce (Hadoop[26]), Lambda Architecture[31], Kappa Architecture[34] and Spark with their respective pros and cons are enlisted in table1.

Phase 2 (Section V):

- Big Data Knowledge Discovery has its origins in Machine Learning as well. This phase enumerates the challenges of handling Big Data with Machine learning like Heterogeneity, complexity, overfitting and high dimensionality.

Phase 3 (Section VI):

- Considering the given qualitative study a neutral Knowledge Discovery Architecture for Big Data Analytics (KDBDA) is proposed with layered approach.
- All layers namely Data Sources, data collection/ ingestion layer, data storage, data analytics, data security and data management & monitoring has been explained with their roles.

Phase 4 (Section VII):

- Result analysis is done as a qualitative comparison of all the architectures and the proposed architecture and is enlisted in table 2 based on various parameters like scalability, data I/O performance, fault tolerance, data size and formats etc.
- Open issues and research directions are provided by analyzing the entire study.

IV. BIG DATA ANALYTICS FRAMEWORK AND PLATFORMS

Big Data architecture contains the entities including hardware and software modules, architecture standard and associated algorithms. These architecture and frameworks supports Big Data Knowledge Discovery process and they are intended to necessarily lay emphasis on the given criteria:

A. Data Models, Structures, Types

It defines the data format to be used within the framework, whether it will use relational or non-relational approach and the type of file system to be used etc.

B. Big Data Management

The framework specifies the Big Data Lifecycle (Management) Model and also it deals with Big Data transformation/staging – association, supervisionand authorizationinvolving various tasks like curation, and archival etc.

C. Big Data Analytics and Tools

Main concern is on the applicationarea of the Big Data, the target uses, presentation and conception.

D. Big Data Infrastructure (BDI)

It comprises of loading, assimilating, manipulating, (High Performance Computing,) Network etc. It also defines the target or actionable strategies and formulates the Big DataFunctional support.

E. Big Data Security

This is one of the biggest challenging are where security aspect related to Big Data are monitored like whether it is Data security in-rest or in-motion and what are the various trusted processing environments.

Machine learning is amongst the principalpractices for data analytics. Big Data has revolutionized the analytics process requirements and generated new demands for the existing paradigms including machine learning (ML). Now with Big Data in picture with millions of parameters and features and complex models to be ingested in a Machine Learning process requiring powerful predictive analytics techniques.

Big Data Knowledge Discovery implementation has its roots in Machine Learning. Conceptualizing a Big Data Knowledge Discovery process without machine learning is next to impossible. All the elements of these frameworks needs support from machine learning.

Architectural models also should be designed to support Big Data knowledge Discovery process. Keeping in view four architectures are studied that provides computational support andenable processing ofenormousquantities of data within a sensible time periodagreeing to S. Wadkar et al [23]:

- "Massively parallel processing" (MPP) database system including"EMC's Greenplum" and "IBM's Netezza".
- "In-memory database systems"for example"Oracle Exalytics", "SAP's HANA" and "Spark ".
- MapReducedispensation model and frameworkslike "Hadoop" and "Google File System (GFS)".
- "Bulk Synchronous Parallel (BSP) systems"for example"Apache HAMA" and "Giraph".

Based on the literature table 1 illustrates the pros and cons of various Big Data analytics framework and architecture.

Table-I: A relativereview of some of the Big Data architecture and frameworks.

Sno.	Architecture/Frameworks/Models	Concept	Pros	Cons
1	***MPI(Message Passing Interface) [24]***	Buffer –to-buffer communication	▪ the state conserving process eliminates the need to read the same data again and again as the processes can residetill the system is executing.[25] ▪ Well suited to handle iterative processing ▪ Hierarchical master/slave paradigm supporting dynamic resource allocation.	▪ Fault intolerant. ▪ Complex and faces optimization issues.
2	***Map Reduce (Hadoop) [26]***	Batch Processing	▪ An easy, simple but communicative model. ▪ Data model and schema independent approach. ▪ MapReducehas an affinity in working with diverse storage layers such as Big Table as it is architecturallyneutral ofavailable storage layers.. ▪ MapReduceinterprets the data at processing time making it highly suitable for processing unstructured or semi-structured data. ▪ MapReduce is linearly scalable because it can work on any amount of data and the algorithm scales linearly. ▪ MapReduce is extremelytolerant to faults and errors. For example, Google testified that with an average of 1.2 failures per analysis job,MapReduce still continues to perform its work[27]. ▪ MapReduce offers excellent scalability, as stated by Yahoo! In 2008 they are able to scale out over 4,000 nodes in their Hadoop gear[29].	▪ In Map Reduce data integrity is lower and data is more susceptible to errors because Map Reduce model interprets localized data which does not need normalized data as it would force the program to look elsewhere for important data repeatedly. ▪ Map Reduce suffers from performance degradation as each input item is parsed during read and for data processing these input items are transformed into data objects.[26] ▪ The original dataflow of Map Reduce is simple yet fixed, restricting many complex algorithms to be implemented on it. Also multi input algorithms are not well supported by it [28]. ▪ Due to batch processing nature of MapReduce, its operation are rarely optimized for I/O proficiency and exhibit latency problem because for processing stage all of the inputs for an MR job in advance required to be organized[26]. ▪ Also not able to manage and support real time data due to latency issues. ▪ MapReduce is not designed to support the execution of iterative algorithms. MapReducearchitectural limitation make it unsuitable for iterative processes.

3.	***Lambda Architecture[30]***	Batch and real-time processing	▪ Capable enough to process uninformedBig Dataassignments in real time. ▪ Retain the input data unchanged. ▪ Lambda architecture takes in account the problem of reprocessing data [31]. ▪ Designed to focus instantaneously, on both the Volume and the Velocity constraint of Big Data [32]. ▪ Provide better data resiliency [33] ▪ Better Conflation of queries and data because data can be stored in a normalized mode and as per requirement can be easily de-normalized.[33] ▪ Complexity of application can achieve better scalability [33].	▪ Lambda architecture tends to be inflexible as maintaining code that needs to produce the identicaloutcome in two multifaceted distributed systems is generally difficult. ▪ It increases the system complexity as the code designed becomes specific to the framework it is executed upon. Also the operational complexity increase to run and debug two systems [34].
4.	***Kappa Architecture[34]***	Real time streaming	▪ Unlike lambda architecture no need to maintain two distinct code base for batch and speed layers as in lambda architecture. ▪ Reprocessing is done only when the code is changed.	▪ Maintaining states at the time of failures is a difficult task in stream processing system. ▪ Synchronization is a major issue. ▪ Requires extra temporary storage when reprocessing is done.
4	***Spark***	Batch and real-time processing	▪ Minimize the disk I/O limitations. ▪ Distributed memory abstraction, resilient distributed datasets (RDDs), assisting coarse-grained transformations ▪ Able to perform in-memory computations. ▪ Adaptablemaking it efficient to run on diverse systems. ▪ Faster batch processing ▪ Spark is appropriate choicefor iterative processing, interactive processing and event stream processing.	▪ "In-memory" capabilities can lead to a hindrance in the cost proficient processing of Big Data. ▪ Requires to be integrated with Hadoop or other cloud based data platform as it does not have a file system of its own. ▪ Takes large number of resources.

V. BIG DATA AND MACHINE LEARNING: A PERFECT ASSOCIATION OR MYTH

A. Challenges

▪ Machine Learning has extended its boundaries for over a decade fulfilling the requirements of discovering and mining the information hidden in the data set generated by varying application domains like biology, bioinformatics, astronomy to name a few as they are composite data-intensive areas.

▪ On the other hand, with advancement in technology and rapid progression in nearly every field of science and technology data is generated and collected at an unprecedented scalewith high degree of complexity and volume.Big Data is demanding to deal by means of conventional learning methods as these well-known process of learning were not designed to work properly with extraordinarymagnitudes. For instance, maximum conventional machine learning models and algorithms are devisedto process data that is required to be entirely loaded into memory [35], that is practically impossible in terms of Big Data.

▪ Enormouscapacities of Big Dataimposes restriction to customary machine learning methodologiesthat are designed to be qualified using a single processor and storage. Instandby, distributedparallel computing frameworks are chosen. To perform parallel and distributed large-scale data processing "Alternating direction method of multipliers (ADMM)" [36], [37] wasformulated which serve as a computing structure to orchestratescalable, distributed, online convex optimization algorithms. To diminish the challenges related to high volume of Big Data for machine learning, apart from distributed theoretical framework, some realistic parallel programming approaches are also formulatedwithimplementation on learning algorithms. MapReduce [38], [39], a commanding programming framework, performsspontaneous parallelization and provision of computation applications using commodity machines on large clusters.

▪ Cloud computing [40], [41] elasticity feature can be utilized to enhance the computing and storage capabilitiesrequired to manageoutsized scale data analytics making machine learning algorithms more scalable. As per thisviewpoint, distributed GraphLab, animplementation for machine learning on cloud, has been projected [42].

▪ Real time applications are necessarily time sensitive. The value of analysis is dependent directly on time. The significance and worthiness of result is calculated on the basis of time. E.g. stock market prediction and agent-based autonomous exchange (buying/selling) systems.

- Processing value is also calculated in terms of data freshness. Machine learning algorithms are sensitive to data on fly or streaming data i.e. non-stationary data is a challenge to machine learning algorithm.In [43] the author gives a comparative overview of batch processing paradigms through streaming processing theory and also reviewed the mechanizationtargeting to verify the effectiveness of streaming technology. Representative streaming processing classificationsconsist of"Borealis" [44], S4 [45], "Kafka" [46], and several other contemporary architectures proposed to provide real-time analytics onBig Data [46]-[48].
- With Big Datacomes uncertainty and incompleteness as beingcollected from incongruent sources making it difficult and practically impossible to be processed by machine learning algorithms which were in the past were usually provided with fairly perfect data from recognized and quite restricted sources, making the learning outcomes to be unmistaken. Sample distributions are exposed to statistical methodologies like means and variances. Distribution-based approach in [49] constructs a decision tree considering the thoroughinformation undertaken by the probability distributions.
- Inadequate and deficient data set knowledge discovery is tough as nearly all machine learning architectures and algorithms works well on clean and preprocessed data. Chen and Lin [35] gave an assessment on how noisy and incomplete data can be curated by the employment of deep learning methods.
- Streaming data and high-dimensional data limits the capabilities of machine learning methods as learning from these multifaceted deep architectures is a problematic optimization job.
- Value exposes a significantattribute of Big Data. A challenge accompanyingthe value associative toBig Data is the miscellany of data meaning, which meansthe value associated with the data varies owing to the scenarios. To incorporate agility and intelligence to the existing system cognition-assisted learning technologies are required.Value from Big Data can be generated by the selectionof machine learning techniques with specific nature and purpose is still a difficult task. Ontology based and semantic web methodologies can beadvantageousforBig Data analysis, but their applicability to machine learning methods are under infancy.
- Big Data often comes in different variety from heterogeneous data sources. This generation ofdissimilar, unrelated, high-dimensional, and nonlinear data available indiverserepresentative forms, affects the knowledge discovery process due to extraordinary degree of complication, data integration issues and requirement for data reduction.XW Chen et al. suggested that individual data source should be examined to learn good representative features and later on applying representation learning to integrate them at varying levels [35].Two-dimensional spectrum heterogeneous data is proposed to be curated from data fusion theory based on statistical learning [50]. Along with all these, deep learning methods also considered for integrating data from different sources. All these methods still face overfitting problem and optimization issues.
- High-dimensional data can comprise of features which are not of much importance.Dimensionality reduction is one of thecompetent method to treat high-dimensional data. Data dimension reductioncan also be done using feature selection or extraction. For example, Sun et al. [51] formulated a local-learning-based feature selection algorithm for supporting multifaceted dimensional data analysis. The prevailingrepresentative machine learning algorithms for data dimensionality reduction embrace"Principal Component Analysis (PCA)", "Linear Discriminant Analysis (LDA)", "Locally Linear Embedding(LLE)", and "LaplacianEigenmaps"[46]. Low-rank matrix is also employed to manage wide-scale data analysis and dimensionality reduction [52], [53]. Most of these algorithms shows effective performance on huge-dimension however the effectiveness and precisionweakensconsiderably when applied on parallel distributed architecture.
- Machine Learning algorithm capabilities are constrained by overfitting, which makes it a major concern while designing machine learning algorithms. Big Data provides a considerable large sample size for learning minimizing the risk of overfitting.Overfitting can be handled by using simple models having less parameters to be tuned
Conversely, the parameter tuning constraints will take a different ideology with Big Data. Training a model with billions of parameters can now be considered as we have sufficientBig Data, expedited by powerful computational paradigms enabling the efficient training of such models [54].
- Volume of the space increases so fast that the available data becomes sparse. Intuition fails at high dimensions. Exceedingly complex models are procreated, which assimilate the training data but do not generalize well to unseen data.
- Big Data ispromptlyintensifying the entire science and engineering domains.Major transformation and substantial opportunities can be entailed by various sectors as a learning outcome from these colossal data set.
- Most of the Machine Learning mechanisms shows degradation in efficiency and scalability while managing data with enormous capacity, varying categories and forms, elevated speed, vagueness and deficient, and low value density. These factors requires a major change in machine learning strategiesin terms ofBig Data processing.

VI. PROPOSED KNOWLEDGE DISCOVERY ARCHITECTURE FOR BIG DATA ANALYTICS (KDBDA)

Big-data applications involve the accumulation of distinct and extraordinary methods, tools and techniques to handle data effectively from the preliminary phase, to gather data till the concluding phase to generate value from data, due to the innate features of Big Data. Thus an architecture is required that will adapt according to the characteristics of Big Data by employing suitable tools and techniques. Based on the review of various architectures discussed in the previous sections, knowledge discovery architecture for Big Data analytics is presented and described. The complete process to attainknowledge from Big Data can be divided into two fundamental processes: Data management and Data analytics.
Data management is related to the procurement, governing, assimilation, fortification and storage of data for data analysis.

Data analytics comprises of data modelling, investigating and understanding to gain valuable knowledge from raw data. Based on the results attained from the systematic review of the available architectures a reference knowledge discovery architecture for Big Data is proposed as shown in Figure 1. Every component can be elective in nature specific to the requirements and application domain.The layers and components are selected and put together in the architecture to consider the various knowledge discovery, data analytics and data management necessities and their interactions.

1. **Data Sources**: The momentous volume of data produced by the varied and huge number of data sources is not only too capacious but also too fast wide range of data sources can be deployed to assimilate data that can be structured, semi-structure or unstructured in nature.
2. **Data Collection/ IngestionLaer**: Collection and Ingestion is a crucial step in knowledge discovery process as it enables data from disparate sources to be transferred to internal systems in an efficient, reliable, fault tolerant and resilient manner. This layer will take care of Big Data volume and velocity characteristic by employing tools like "Apche Kafka".
3. **Data Storage Layer:** Big data storage requires sophisticated architectures to handle different type of data that can be batch, real time or streaming data. Distributed architecture are one possible solution for this.No SQL database are also playing a prominent role for Big Data storage requirements.
4. **Data Analytics:** This layer comprises of four sublayers namely: Data Preprocessing, Querying layer, Data Analysis layer and Visualization Layer.

- **Data Preprocessing** involves data cleaning, transformation and data reduction for better analysis
- **Querying Layer** supports in data filtering for choosing relevant data to assist in data analysis.
- **Data Analysis**: This layer involves developing an analytical model by inspecting the data sets to surmise knowledge by finding patterns and drawingconclusions with the assistance of expert tools and algorithms. Technologies like predictive modelling, deep learning, machine learning can be utilized for this purpose.
- **Visualization Layer**: To get a better understanding of the knowledge gained this layer provides a pictorial or graphical format for data representation considering the inherent characteristics of Big Data i.e. volume so that data should not collapse or condensed.

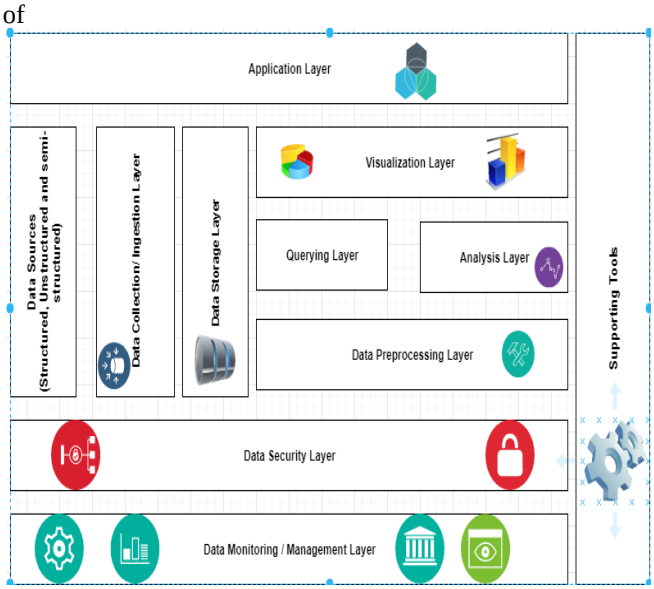

Fig.1. Proposed Knowledge Discovery Architecture for Big-Data Analytics

4. **Data Security Layer:** This layer handles the most challenging aspect of Big Data architecture involving the security of data. Tools like Apache Atlas comes equipped with a set of scalable and extensible governance services that can be utilized to handle security measures.

5. **Data Monitoring and Management Layer**: Monitoring the entire system for any issue and also to manage the entire framework is the role of this layer. Various products available in the market can be utilized to handle the functionality of this layer for e.g. Apache Ranger provides monitoring and security framework in Hadoop platform to enable monitoring. Apache Zookeeper is a service programming tool helps in distributed coordination across nodes.
6. **Application:** An abstraction layer for specific implementation of Big Data applications to suit customized end-user requirements.
7. **Supporting tools: A**part from the standard tools and service this layer helps to incorporate additional tools and platforms to handle specific and time bound requirements.

B. Results and Research Directions

- **Result Analysis**

The proposed Knowledge Discovery architecture for Big Data can be formatted based on the user requirement to fit in the requisite tools in any of the given layer. It's a neutral architecture.

Table 2 Comparison of different platforms/frameworks with the proposed framework based on various characteristics

Characteristics/ Platforms		MPI	MapReduce	Lambda Architecture	Kappa Architecture	Spark	GPU	KDBDA
Scaling	Vertical	*****	****	***	**	***	****	****
	Horizontal	***	***	***	****	***	****	****
Data I/O Performance		****	****	***	***	****	*****	***
Iterative processing support		****	**	**	***	****	****	****
Fault Tolerance		*	****	***	***	****	****	***
Data Processing Support	Real Time	****	**	****	*****	****	*****	****
	Batch	***	*****	****	*****	****	*****	****
	Stream	*	**	*	****	***	*****	****
Data Format	Structured	****	****	***	***	****	*****	****
	Unstructured	****	*****	***	***	****	****	****
	Semi Structured	****	*****	***	***	****	****	****
Data Size		*****	****	***	****	****	**	****
Security		**	***	***	***	***	**	**
Privacy		**	***	**	***	****	**	**

Table 2 gives a comparison between different platforms and the proposed KDBDA platform taking into account some fundamental characteristics required for Big Data processing platforms: scaling with reference to both horizontal as well as vertical scaling, data I/O performance, Iterative processing support, Fault Tolerance, Data processing support in terms of real-time, batch and streaming data, data format support categorized as structured, unstructured and semi structured, Security and privacy. Scalability, Data I/O performance, Security and Fault tolerance are platform/ architecture dependent attributes while Data Format, Iterative processing support, Data size and privacy are specific to application or algorithm.

A qualitative comparison between the platforms with proposed KDBDA architecture is given in the table where three stars for a given platform indicates that it is better than the platform with two stars. This rating table is a qualitative review of the strengths and weakness of various platforms available and proposed for Big Data Analytics.

- **Research Directions**

Based on the study conducted on the available architecture and looking at the Big Data characteristics and features following parameters should be considered while adapting and selecting a specific tool and technology.

- **Big Data Architecture**

1) As per Big Data Analytics is concerned data I/O is the data reading or writing rate to memory(disk) or rate at which the data transfer occurs amongstnodes in a cluster. Most of the platforms exhibit poor data I/O performance.

2) Much of the Big Data analytics process is expected to take place on in-memory infrastructure to provide faster transfer between storage and processing. This can become problematic in distributed environments due to the need to quickly compile data from diverse sources and generate results to users who may be some distance away from the processing center.

3) Optimization is another point to focus as time taken to generate the result by most of the architectures is considerably much more althoughlarge scale datasets are handled efficiently. There is a strong need to optimize speed or throughput to help analytical architecture to work on real time or streaming data processing.

- **Big Data Machine Learning**

1) Training of ML techniques takes place on a specific labeled datasets or data realm and thus may not be appropriate for another set or domain of data. This approach is one of the prime concerns while dealing with Big Data. Thus finding a suitable ML algorithm to analyze Big Data is an issue to be resolved.

2) Big Data comes with huge volume and variety. ML algorithms are trained using certain number of class types while in the case of Big Data the class type grows dynamically leading to inaccurate classification results. Thus to find appropriate class types from silos of Big Data that can be used for Machine learning algorithm is another challenging issue.

3) Machine learning algorithm often faces synchronization issues and overheads, although they are error tolerant, as states synchronization is often been neglected.

- **Big Data High Dimensionality**

1) Big Data is considered as heterogeneous, high dimensional and non-linear in nature. In ML algorithms labeled patterns play very crucial role, more patterns means higher accuracy but they come with computational time and cost especially when Big Data is concerned. It is difficult to find how many patterns are needed for training? This overfittingproblem is one of the foremost constraints in Big Data machine learning.

- **Big DataSecurity:** Security is one of the major concern for most of the Big Data Analytics platforms.
- **Big Data Privacy:** Preserving privacy of Big Data is another prime issue as it can lead to problems like sensitive data breach etc.

VII. CONCLUSION

This study provides a summary of the customary Big Data analytical platforms and frameworks with their challenges, taken after by a comparison between these methods based on their processing capabilities, I/O handling, and memory handling mechanisms.A tools and technique architecture neutral knowledge discovery architecture for Big Data Analytics is also proposed in the paper. In this study we also examined and outlinedthe concerns and challenges associated to Big Data in relative to Machine Learning. Big Data analysis is a very demanding field requiring sophisticated architecture with high degree of scalability, resiliency, processing capabilities, and architectures which are capable enough to handle this explosion of data. This study outlined various open research directions and issues which are required to be considered in order to get complete advantage from this Big Data revolution.

REFERENCES

1. 1.M. Cooper and P. Mell, “Tackling Big Data,” National Institute of Standards and Technology, U.S., June 2012. [Online] Available: http://csrc.nist.gov/groups/SMA/forum/documents/june 2012presentations /f%csm_june2102_cooper_mell.pdf.
2. Jianqing Fan, Fang Han, and Han Liu, “Challenges of Big Dataanalysis,”National Science Review, vol.1, Dec. 2014, pp.293-314.
3. Han Hu et al., "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol. 2, no., pp. 652-687, 2014. doi: 10.1109/ACCESS.2014.2332453.
4. Baraniuk RG, “More is less: signal processing and the data deluge,” Science.2011;331(6018):717–9.
5. Chun-Wei Tsai , Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos, “Big Data analytics: a survey,” Journal of Big Data, Springer International Publishing, pp. 2-21, 2015. doi : 10.1186/s40537-015-0030-3.
6. Russom P,” Big Data analytics,” TDWI, Tech. Rep. , 2011.
7. C.Ma et al., “Machine learning for Big Data analytics in plants,” Trends Plant Science, vol.9, Issue 12, pp.798–808, Dec. 2014.
8. D. Boyd andK.Crawford, “Critical questions for Big Data,” Information Communication Society, 15(5), pp. 662–79, 2012.
9. G. Blackett, “ Analytics Network-O.R. Analytics,” 2013. [Online]. Available: http://www.theor society.com/Pages/SpecialInterest /AnalyticsNetwork_anal%ytics.aspx.
10. P.Laskov et al., “Incremental support vector learning: analysis, implementation and applications,” Journal of Machine Learning Research. vol.7 2006, pp. 1909–36.
11. B. Geerdink, "A reference architecture for Big Data solutions introducing a model to perform predictive analytics using Big Data technology," *8th* International *Conference for Internet Technology and Secured Transactions (ICITST-2013)*, London, 2013, pp. 71-76. doi: 10.1109/ICITST.2013.6750165
12. M.Pospiech and C. Felden, “Big Data—a state-of-the-art,” In: Proceedings of the Eighteenth Americas Conference on Information Systems, Seattle, Washington, August 9-12, 2012, pp. 1-23 [Online]. Available: http://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22.s
13. Apache Hadoop, February 2, 2015. [Online]. Available: http://hadoop.apache.org.
14. Cuda, February 2, 2015. [Online]. Available: http://www.nvidia.com/object/cuda_home_ new.html.
15. Apache Storm, February 2, 2015. [Online]. Available: URL: http://storm.apache.org/.
16. R.R. Curtin et al., “MLPACK: a scalable C++ machine learning library,” Journal of Machine Learning Research, vol. 14, Issue 1, pp. 801–805, 2013 .
17. Wu Xindong et al., “Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, Issue 1, pp. 97-107, January 2014.
18. C.T.Chu et al., “Map-reduce for machine learning on multicore,” In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS '06), MIT Press, 2006, pp. 281-288.
19. Gillick et al., “MapReduce: Distributed Computing for Machine Learning,” Berkley, December 18, 2006.
20. Ranger et al., “Evaluating MapReduce for multi-core and multiprocessor systems,” In: Proceedings of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA '07), 2007, pp. 13-24.
21. Das et al., “ Integrating R and Hadoop,” In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10), 2010, pp. 987-998.
22. N. Khan et al., “Big Data: survey, technologies, opportunities, and challenges,” The Scientific World Journal, 2014, pp. 1–18.
23. Sameer Wadkar, MadhuSiddalingaiah, “Pro Apache Hadoop,” 2nd ed., Apress, 2014.
24. “MPI: A Message-Passing Interface Standard Version 3.0,” www. mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf, Message Passing Interface Forum, 2012.
25. Dominique LaSalle and George Karypis, “MPI for Big Data: New Tricks for an Old Dog,” Parallel Computing, vol. 40, Issue 10, 2014. pp.754-767.
26. A. Pavlo et al. “A comparison of approaches to large-scale data analysis,” In Proceedings of the ACM SIGMOD, 2009 , pp. 165–178.
27. J. Dean et al ., “MapReduce: Simplified data processing on large clusters,” Communications of the ACM, vol.51, Issue 1, 2008, pp. 107–113
28. Kyong-Ha Leeet al. “Parallel Data Processing with MapReduce: A Survey”. SIGMOD Record, vol. 40, No. 4, December 2011, pp. 11-20
29. A. Anand, “Scaling Hadoop to 4000 nodes at Yahoo! ,”http://goo.gl/8dRMq, 2008.
30. Marz Nathan and Warren James, “Big Data: Principles and best practices of scalable realtime data systems,” Manning Publications, 2013.
31. Michael Hausenblas ,”Applying the Big Data Lambda Architecture,” Dr.Dobb’s, Rep. Nov.12, 2013.
32. D. Laney, “3D data management: Controlling data volume, velocity, and variety,” Technical report, META Group, February 2001.

33. "Building An Efficient Microservices Architecture," White Paper © 2016 Newt Global Consulting, LLC. Online available at http://newtglobal.com/White/Building%20an% 20efficient %20Microservices%20architecture.pdf
34. Jay Kreps, "Questioning the Lambda Architecture," O'Reilly Media, Inc. , Sebastopol, California, Rep. July 2014. [Online] Available from: https://www.oreilly.com/ideas/questioning-the-lambda-architecture.
35. X.W. Chen and X.Lin, "Big Data deep learning: challenges and perspectives," IEEE Access vol. 2, pp. 514–525, 2014
36. F Andersson et al., "A new frequency estimation method for equally and unequally spaced data," IEEE Transaction on Signal Processing, vol.62, Issue 21, pp.5761–5774, 2014.
37. F Lin et al., "Design of optimal sparse feedback gains via the alternating direction method of multipliers," IEEE Transaction on Automatic Control, vol. 58, no.9, pp. 2426–2431, 2013.
38. J Dean et al., "MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 5, no. 1, pp. 107–113, 2008.
39. J Dean and S Ghemawat, "MapReduce: a flexible data processing tool. Communications of the ACM, vol. 53, no.1, pp. 72–77, 2010.
40. M Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, no.4, pp. 50–58, 2010
41. MD Dikaiakos et al., "Cloud computing: distributed internet computing for IT and scientific research," IEEE Internet Computing, vol. 13, no.5, pp. 10–13, 2009.
42. Y Low et al., "Distributed GraphLab: a framework for machine learning and data mining in the cloud," Proceedings of the VLDB Endowment, vol.5, no.8, pp. 716–727 2012.
43. N Tatbul, "Streaming data integration: challenges and opportunities," in Proceedings of the 26th IEEE International Conference on Data Engineering Workshops (ICDEW) Long Beach, 2010, pp. 155–158.
44. DJ Abadi et al., "The design of the borealis stream processing engine," in Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, 2005, pp. 277–289.
45. L Neumeyer et al., "S4: Distributed stream computing platform," in Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW), Sydney, 2010, pp. 170–177.
46. K Goodhope et al., " BuildingLinkedin's real-time activity data pipeline," IEEE Data Engineering Bulletin, vol.35, no.2, pp.33–45,2012.
47. W Yang et al., "Big Data real-time processing based on storm," in Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Melbourne, 2013, pp. 1784–1787.
48. B SkieS, "Streaming Big Data processing in datacenter clouds," IEEE Cloud Computing, vol. 1, pp. 78–83, 2014.
49. S Tsang et al., "Decision trees for uncertain data" IEEE Transaction on Knowledge Data Engineering, vol.23, , no.1, pp. 64–78, 2011.
50. Q Wu et al., "Spatial-temporal opportunity detection for spectrum-heterogeneous cognitive radio networks: two-dimensional sensing," IEEE Transaction on Wireless Communications, vol. 12, no. 2, pp.516–526, 2013.
51. Y Sun et al., "Local-learning-based feature selection for high-dimensional data analysis," IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 32, no. 9, pp. 1610–1626 , 2010.
52. LJP van der Maaten et al., "Dimensionality reduction: a comparative review," Journal of Machine Learning and Research, vol. 10, no. 1-41, pp. 66–71, 2009.
53. M Mardani et al., "Subspace learning and imputation for streaming Big Data matrices and tensors," IEEE Transaction on Signal Processing, vol. 63, no. 10, pp.2663–2677, 2015.
54. Omar Y. Al-Jarrah et al., "Efficient Machine Learning for Big Data: A Review," Big Data Research, Big Data, Analytics, and High-Performance Computing, vol. 2, no. 3, Sep. 2015, pp. 87–93.

AUTHORS PROFILE

Neelam Singh, Assistant Professor, Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India. She has received M.Tech (CSE) degree in Computer Science and Engineering, GEU, Dehradun, Indiain 2014. Her research interests include Big Data, Machine Learning, Mining and Artificial Intelligence.

Dr. Devesh Pratap Singh, Professor and Head of Computer Science and Engineering department at Graphic Era Deemed to be University Dehradun India. He has received M. Tech degree in Computer Science and Engineering from Uttarakhand Technical University Dehradun India in 2009. He has also received Ph.D. in 2015. His research interests include Information Security, Wireless Sensor Networks, Internet of Things and Soft Computing. He has published more than fifty research papers in his area of expertise. He is the member ofACM.

Dr. Bhasker Pant, Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal. Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded). and 5 candidates are in advance state of work. He has also guided 28 MTech. Students for dissertation. He has also supervised 2 foreign students for internship. He has more than 70 research publication in National and international Journals.He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.