I.J. Information Technology and Computer Science, **2016, 8, 1-12**
Published Online August 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.08.01

Enhancing Big Data Value Using Knowledge Discovery Techniques

Mai Abdrabo
Information Systems Department, Faculty of Computers and Information, Suez Canal University, Ismailia, 41611, Egypt
E-mail: mai_abdrabo86@yahoo.com

Mohammed Elmogy, Ghada Eltaweel, Sherif Barakat
Information Technology Dept., Faculty of Computers and Information, Mansoura University, Mansoura, 35511, Egypt
Computer Science Department, Faculty of Computers and Information, Suez Canal University, Ismailia, 41611, Egypt
Information Systems Dept., Faculty of Computers and Information, Mansoura University, Mansoura, 35511, Egypt
E-mail: {melmogy@mans.edu.eg, ghada@ci.suez.edu.eg, sherifiib@yahoo.com}

Abstract—The world has been drowned by floods of data due to technological development. Consequently, the Big Data term has gotten the expression to portray the gigantic sum. Different sorts of quick data are doubling every second. We have to profit from this enormous surge of data to convert it to knowledge. Knowledge Discovery (KDD) can enhance detecting the value of Big Data based on some techniques and technologies like Hadoop, MapReduce, and NoSQL. The use of Big Data value is critical in different fields.

This survey discusses the expansion of data that led the world to Big Data expression. Big Data has distinctive characteristics as volume, variety, velocity, value, veracity, variability, viscosity, virality, ambiguity, and complexity. We will describe the connection between Big Data and KDD techniques to reach data value. Big Data applications that are applied by big organizations will be discussed. Characteristics of big data will be introduced, which represent a significant challenge for Big Data management. Finally, some of the important future directions in Big Data field will be presented.

Index Terms—Knowledge Discovery (KDD), Big Data, Hadoop, MapReduce, NoSQL.

I. INTRODUCTION

A. Big Data

Nowadays, data is becoming big in volume, variety and velocity. We aim to use KDD to reach the value of this big data. Big Data value strongly appears in applications section, however there are a lot of challenges faced the increasing of data.different data sources generate floods of data. Data is a gathering of qualities and variables related in a particular sense and varying in another, yet the measure of data has dependably been expanded [1]. Until 2003, 5 Exabytes (10^{18} bytes) of data were made by human, yet these days human is making this sum in just two days. In 2012, the volume of data was expanded to 2.72 zettabytes (10^{21} bytes). It is hoped to rehash at regular intervals, reaching out to about eight zettabytes of data by 2015 [2]. Data will become to 35.2 ZB by 2020, i.e. 37.6 billion hard drives of 1TB limits will be required to store these data [3]. This expansion drives the world to move from data to Big Data. The Big Data term has gotten the expression to portray the gigantic sum and different sorts of quick data. Big Data is a term for critical data sets getting to be bigger, more expanded, and confounded structure with the hindrances of storing, analyzing, and visualizing for procedures [2]. For instance, different sources produce biomedical data like a microscope, macroscopic world, and genomics in diverse structures. The greater part of the biomedical Big Data is created in genomics. Other than the issue of heterogeneous and distributed data, the noisy, missing, and conflicting data are found and must be taken in care. It leaves an enormous crevice between the current "dirty” data and the machinery to productively handle the data for the application purposes [4]. Consequently, Big Data is additionally characterized as a huge volume of data that needs new advancements and architectures. Analytics and mining are important to discover quality in Big Data utilizing proficient data revelation strategies [5]. KDD is the method of making a qualification high-minded, novel, valuable, and ultimately intelligible patterns from massive data repositories [6]. Big Data helps for getting to huge volumes of data, trying to increase basic bits of knowledge (insight) from repeated processing [7]. Connected to the KDD process, Big Data development offers numerous novel chances for organizations to profit by new bits of knowledge (insight) because of the trouble of analyzing such large datasets [8]. Big Data analytics is the strategy of exploration into a monstrous volume of data to distinguish concealed patterns and hidden relationships, yet KDD is more valuable than Big Data analytics because KDD incorporates distinctive analysis methods. Processing of Big Data changes data to information, from the information to knowledge, and knowledge to wisdom [5], as appeared in “Fig.1”. The human mind`s content can be stratified into five classes [9]:

- **Data:** Symbols that basically exist in the raw structure.

- **Information:** It is handled data to be valuable. It gives answers to "who", "what", "where", and "when" inquiries. It speaks to important data taking into account the method of relational connection.
- **Knowledge:** It is the connected data and information to reply "how" addresses. It is the suitable collection of information.
- **Understanding:** It is an intellectual, expository, interpretive, and probabilistic procedure that replies "why" questions.
- **Wisdom:** it is foreordained understanding.

A misinterpretation is that the "Big Data Revolution” is about the measure of the data, as appeared in Fig. 2. It is the utilization of data to achieve more profound, knowledge, and raising the conceivable outcomes that originate from improved data accessibility, analysis, and action. Big Data is a buzz term, which regularly contains warehousing data at a gigantic scale. In any case, the genuine inspiration - why organizations put so intensely in the majority of this, how gaining from that data [10].

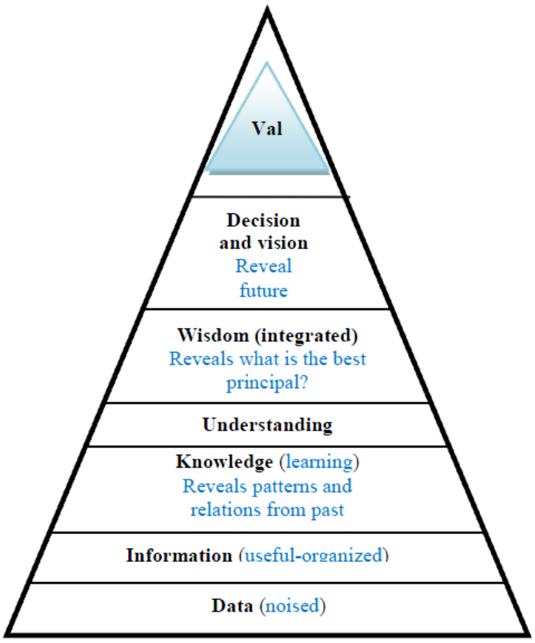
Val
Decision and vision
Reveal future
Wisdom (integrated)
Reveals what is the best principal?
Understanding
Knowledge (learning)
Reveals patterns and relations from past
Information (useful-organized)
Data (noised)

Fig.1. The Big Data processing is to implement: from data to information, from information to knowledge, and from knowledge to wisdom.

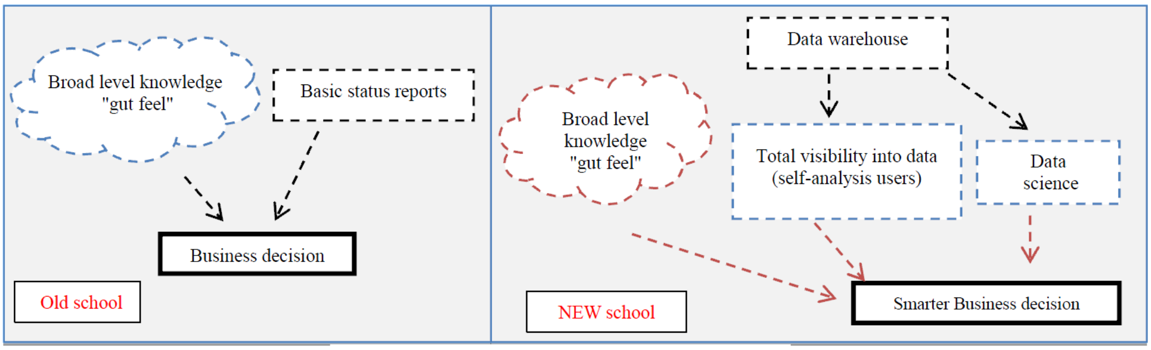
Broad level knowledge "gut feel"
Basic status reports
Business decision
Old school
Data warehouse
Broad level knowledge "gut feel"
Total visibility into data (self-analysis users)
Data science
NEW school
Smarter Business decision

Fig.2. The Big data revolution [10].

These instruments and limits might incorporate [10]:

- **Dashboards**: visualize data to make understanding of data easy.
- **Real-time measurement**: understand what is going on the business now.
- **OLAP tools:** enable anybody to go deeply into the data.

Data integral to a broad range of choice making, performance management, and business process management. Devices of OLAP help end client for revelling through data exceptionally well. It accomplishes this by utilizing a particular system that permits adaptable “querying" of information with quick execution time. The outcome is simplicity for business clients to get custom perspectives of the data. For instance, an advertising supervisor can without much of a stretch get full permeability into week-over-week offers of a particular sort of item getting through a particular securing channel, just by clicking a couple catches on an OLAP device.

This dashboarding/OLAP system additionally makes noting data addresses more clear for some sorts of analysis (e.g. marketing analysts and financial analysts).
With these instruments, analysts can jump profoundly to comprehend business components without specialized difficulties working with crude, unstructured information in a data warehouse [10]. Necessary pieces of Big Data work together to get business value:

- Data Warehousing Technology
- Business Intelligence (BI)
- Data Science

As the world has turned out to be more digitized, now big organizations are working with data stores at a petabyte scale. To continue enhancing with these necessities, a bunch of inventive innovations have been gained that give a framework to handle such tremendous masses of data. Not just the storing of this huge data is the challenge, however finding the effective approaches to preparing and processing it to deliver noteworthy information. Including relational database systems, NoSQL database systems, and software ecosystems for distributed computing (e.g. Hadoop/MapReduce) examined in points of interest later.

Then again, BI is the capacity of integration with data and whatever remains of the organization. In particular, it is a significant association between the data warehouse and business leadership/business analysts, which are empowering full transparency in the subtlety of what is going on in the employment. The BI bunch at a venture satisfies this by obtaining and supporting differences of instruments that end-clients get a handle on the majority of the data in an edible medium

Self-Serve analytics implies data can be gotten to and comprehended by everybody. It is robust because it makes KDD is taken to the following level by data science utilizing a deep learning from data, advanced techniques like predictive modelling, and pattern recognition through machine learning. Its professionals are known as data scientists and thought to be high workers in any organization with Big Data desire. Data science is enlisting between the lines and profound induction from data mining out key knowledge that is let go behind the racket and growing intense data-driven capacities. By the day's end, the objective of data science is to give worth to revelation by transforming data into gold [10]. Big Data is recognized by developing the volume (a measure of data), velocity (speed of data in and out), and a variety (scope of data sorts and sources).

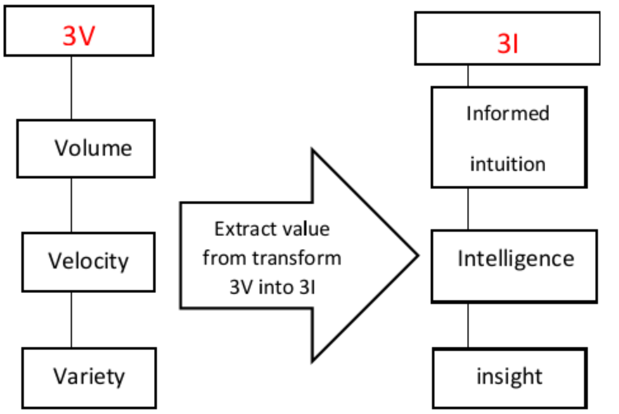

Fig.3. The three Vs transform into three Is.

Three Vs change into three Is to extricate the value of Big Data [11], as appeared in “Fig. 3”:

- **Informed intuition:** Means how to predict future occurrences and what are successful actions.
- **Intelligence:** Looking at what is going now on real time (or close to real-time) and determining the action to take.
- **Insight:** Knowing what has happened and identifying the actions to take.

Notwithstanding, with the development of data to be greater and greater it has more characteristics as takes after:

- **Volume** (size quickly developing): Enterprises are producing data that are ascending at an exponential pace. The consideration of data size is moving from Terabytes to Zettabytes [12].
- **Variety** (unstructured data`s era): Multiple sources produce data in heterogeneous configurations. In this manner, the data has no schema that is moving structured and semi-structured data storage for altogether unstructured data [12].
- **Velocity** (streaming data`s era): The rate of data generation is quick that has moved from data sets (batch) to streaming data [12].
- **Value** (The era of cost associated with data): While the data is being produced, gathered, and analyzed from diverse quarters, it is important to say that today's date costs [12].
- **Veracity** (The era of data pollution that needs cleansing): There is the need to check the accuracy of the data by eliminating the noise through procedures, for example, data family and disinfection to guarantee data quality [12].
- **Variability:** It is viewed as a challenge for data streaming and data loading to be kept up particularly with the expansion in the utilization of the social media which causes a top in data loads with specific occasions happening [5].
- **Viscosity (Consistency):** During data following, there is resistance (slow down) from business rules, and even be a limitation of innovation, yet it is critical to quantify this resistance. For instance, social media checking falls into this class for offers ventures to comprehend their business some assistances with impacting and opposes the utilization of the data [7].
- **Virality:** Measuring and portraying how rapidly data is shared in individuals to-individuals (peer) network system. The rate of spread is measured in time. For instance, re-tweets that are shared from a unique tweet is a decent approach to take after a point or a pattern [7].

Big Data can be characterized by 8Vs and the complexity and ambiguity as well.

- **Ambiguity (Uncertainty):** An absence of metadata makes uncertainty in Big Data. For instance, in a photograph or a group, M and F can depict gender or can delineate Monday and Friday [7].
- **Complexity:** It is an undertaking to link, match, cleanse, and transform data across systems coming from several origins. It is likewise necessary to tie in and correlate relationships, power structures, and multiple data linkages or data can rapidly spiral out of control [5].

B. Knowledge discovery process

Field expert in society can rely on manual analysis to turn data into knowledge for supporting decision support by useful patterns. Today, a variety of names has been afforded for this operation, including mining data, extracting knowledge, information discovery, harvesting information, data archeology, and processing data pattern. In classic knowledge, the discovery process is featured by various steps beginning with a selection of data, pre-processing, transforming data, mining data, and interpretation. In this context, data mining is a subsection of KDD. It is respectable to note that KDD can be viewed as a process and transacts the complete value-added chain from data. There is a novel approach to combine Human-Computer Interaction (HCI) & KDD. The fundamental reason of HCI-KDD is to encourage end clients intuitively to discover and represent useful and available data previously. It may be specified in the classical sense as the procedure of identifying novel data patterns, with the goal of interpreting these figures.

Datasets are possessed by an area expert. They might have the capacity to recognize, extract, and understand useful information to gain new, and previously unknown knowledge [3]. Through the integration of data, it can profoundly separate knowledge. By utilizing such new knowledge, data can be handled progressively to comprehend and apply the data, to make keen judgements and all around educated choices. Data

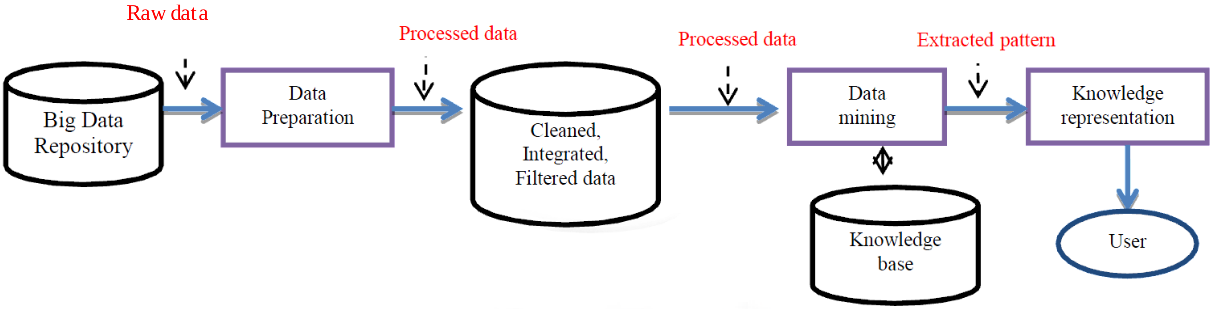

Fig.4. The KDD process [6].

Intelligence is the capacity to get a more inventive, precise, and complete learning to take care of a particular issue through an in-depth analysis of the gathered data. It is a capacity to comprehend and tackle issues quick, adaptable, and efficiently. Utilizing a variety of cutting edge method`s comes to more profound data intelligence to create more value of data [5]. The cleaned and integrated data is stored in databases or data warehouses. It is important to note that data mining can be performed without the presence of a data warehouse through data warehouses significantly improve the efficiency of data mining. Knowledge Presentation: Presentation of the knowledge extracted in the data mining step in a format easily understood by the user is an important issue in knowledge discovery. To deal with different types of complex problem domains, specialized algorithms have been developed that are best suited to the particular problem that they are designed for [6]. The work of KDD can be grouped into preparation data, mining data, and presentation knowledge. Data mining is the center step where the algorithms for extricating the helpful and interesting patterns are connected. The primary motivation behind mining biological Data is to use automated databases to store, compose, and index of data. This data empowers the discovery of new organic bits of knowledge. Run of the mill issues of bioinformatics where digging systems is required for extracting meaningful patterns [6]. In this sense, data preparation and knowledge presentation can be considered, separately, to be preprocessing and post-preparing ventures of data mining, as appeared in “Fig.4”.

In the data preparation step, data is first cleaned to reduce noise, erroneous, and missing data as far as possible. Once the data is cleaned, it may need to be integrated since there could be multiple sources of the data. After integration, moreover, redundancy removal may need to be taken away.

II. Big KDD Techniques and Technologies

Concerning illustration demonstrated over “Fig.5”, there may be a review from claiming applicable technologies will realize. There are two traditional architectures to manage Big Data for providing insights (Hadoop ecosystem, HPCC system) [2].

A. *Hadoop ecosystem*

Doug Cutting created Hadoop as two core services. The first is a reliable, distributed file system called Hadoop Distributed File System (HDFS). The second is the high-performance parallel data processing engine called Hadoop MapReduce. The mix of HDFS and MapReduce presents a software framework for processing massive amounts of data in parallel on large clusters of hardware (suitable for scaling thousands of nodes) in a fault-tolerant manner. Hadoop has witnessed in environments where massive server farms to gather data from different sources. Effort and time required for loading data into another system can be reduced by Hadoop [15].

Concerning illustration indicated done “Fig.6”, that Hadoop community includes Hive is an SQL dialect.

Furthermore, Pig is a data stream dialect by that making MapReduce employments are stowed away behind higher-level abstractions all more fitting to client objectives. Federating for benefits are executed by Zookeeper, what's more, an arrangement planning will be performed toward Oozie [15].

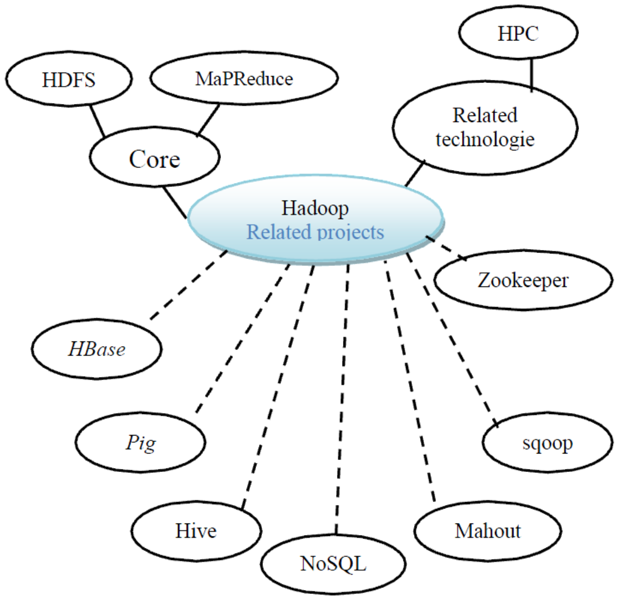

Fig.5. The Big Data Technologies.

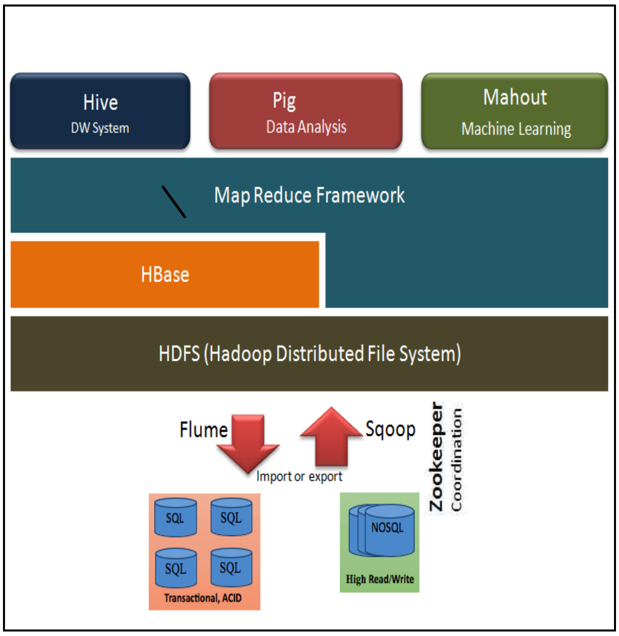

Fig.6. The Hadoop ecosystem [15].

- **MapReduce**

MapReduce is the most utilized with an extensive variety. It displays an adaptable and versatile establishment for analytics, toward universal reporting weight with the leading-edge machine taking in algorithms. In MapReduce model, "job" may be performed the following decay under littler "tasks" on circulating clusters [15].

- **Pig**

Pig could extract, transform, load (ETL), process, and dissect substantial data sets. It will be a stage that utilization pig Latin dialect. It likewise offers data control operations in the grouping, joining, also sifting [15].

- **Hive**

Hadoop needs to utilize hive on it is an SQL-based data warehouse framework. Hives` reductions would support those outline for data, ad-hoc queries, and the dissection for substantial datasets put away Previously, HDFS, MapR-FS, S3, furthermore some NoSQL databases. Hive does not think about a social database, yet all an inquiry motor that enhances extraordinary SQL parts to inquiry information utilizing extra enhancements for composing new tables alternately files, However not upgrading single person records. Hive employments are optimized to versatility. It does not oblige intricate ETL forms [15].

- **NoSQL**

NoSQL may be an umbrageous expression database administration frameworks that set the requirements of (RDBMS) relational database administration frameworks on accomplishing objectives for all more expense profit investigation scalability, adaptable tradeoffs about accessibility. That NoSQL community is gigantic. Around for it; those well-known databases are HBase, Cassandra. Hadoop is nearly tied with HDFS more than others [14].

- **Cassandra**

Cassandra recognizes those the vast majority well-known NoSQL database for gigantic datasets. It is a key-value, bunched database that performs column-oriented storage, sharing by entering ranges, what's more, excess stockpiling to versatility in both data sizes and read/write execution [14].

- **HBase**

HBase recognizes a distributed, column-oriented database that helps Big-table such as competencies once highest priority on Hadoop. SQL queries (but not updates) need aid improved utilizing Hive, However with Helter Skelter inactivity. However, HBase displays Helter Skelter also compose execution and will be utilized in a few extensive applications, for example, Facebook's informing stage. Toward default, HBase Yet stockpiling done HDFS may be wanted for use for Pig [14].

- **Machine learning in over Hadoop:**

- **Mahout:** Mahout is utilized to fabricating versatile machine Taking in libraries. Mahouts` primary calculations for clustering, classification, what's more, clump built collective shifting are performed utilizing those MapReduce for Hadoop. For a fact, three regular machine-learning utilization situations would basically underpin via Mahout [16]:
- **User-based recommendations,** with an anticipating new inclination for client that employments to client practices data mining.
- **Clustering** searches for similarities between data points by a user-specified metric, to identify

clusters (groups of points) in data that appear more similar to each other than to members of other groups.

- **Classification** continuous value can be predicted from previous examples by applying discrete labels to data.

Data can be inputed to hadoop core system based on sqoop and flume:

- **Sqoop:** This is an open source tool designed for efficiently transferring data between Apache Hadoop and structured, relational databases. Sqoop is recommended for importing datasets to HDFS.
- **Flume:** This is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data to HDFS. Flume is able to read data from most sources, such as logfiles.

The thing that is setting around Big Data foundation [13], Concerning illustration demonstrated for table 1.

Table 1. The big data infrastructure technologies

RDBMS	NOSQL	Hadoop, MapReduce and Massively Parallel Computing
Traditional relational database management system	Mentioned to as "Not Only SQL."	Hadoop is ideally a software ecosystem that permits for massively parallel computing, but not a type of database
It is database management throughout the age of the internet	Presents an entirely different framework of databases that allows for high-performance, processing quickly of massive scale data.	It handles types of NoSQL distributed databases (such as HBase), by spreading data across thousands of hosts with little diminution in functioning.
The architecture of RDBMS such that data is coordinated in a highly-structured manner that keeps the warehouse very "neat".	To handle the biggest data warehouses on the planet that needs the NoSQL distributed database infrastructure to be the answer i.e. the likes of Google, Amazon. Unstructured data may be stored in multiple processing nodes where the distributed databases concept surrounded by NOSQL across multiple hosts (servers).	The Hadoop ecosystem consists of MapReduce, a computational model that calls for intensive data processes and distributes the computation across endless number of hosts (Referred to as a Hadoop cluster).
Data must be well-structured because of performance with data declines' size gets bigger.	NoSQL databases are unstructured in nature to trade of rigid consistency requirements for speed and agility.	A large data procedure that might get 20 hours of processing time on a centralized relational database system. This large data may only involve 3 minutes when distributed across a large Hadoop cluster of commodity servers, all processing.

B. HPCC Systems

As shown in "Fig.7" HPCC Systems (High-Performance Computing Cluster) architecture includes the Thor and Roxie clusters. It is also used common middleware components, an external communication layers.

- **Client interfaces** give acceptable both end-user benefits and framework administration devices. Assistant parts provide checking and improving stacking, what's more, storing about file-system information starting with Outside wellsprings [17].
- **Thor (the refinery information Cluster)** is answerable for transforming, linking, what's more, indexing massive volume of data.

It capacities Likewise a dispersed record framework with parallel preparing control spread crosswise over those hubs. A bunch camwood scale from a solitary hub should be many hubs [17].

- **Roxie (the inquiry Cluster)** displays separate high-octane on the web inquiry transforming and competencies from claiming information warehouse [17].
- **ECL (Enterprise control Language)** will be those capable modifying dialect that is suitableness to taking care of Big Data that mixes information representational and calculation execution [17].

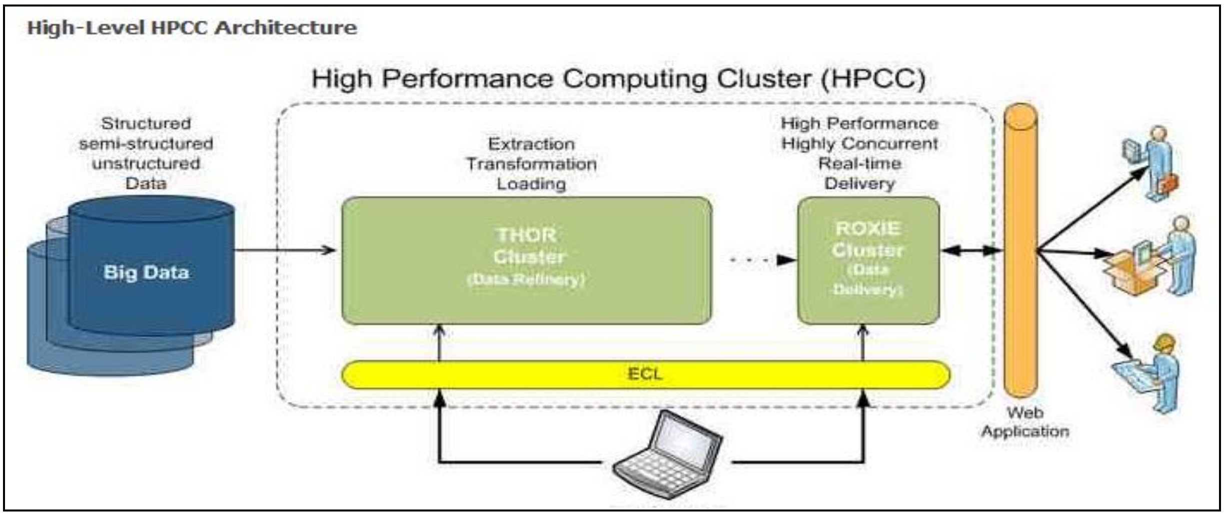

Fig.7. The HPCC System architecture.

Comparison between traditional architectures to manage Big Data for providing insights (Hadoop ecosystem, HPCC) shown in Table 2:

Table 2. Comparison between Hadoop ecosystem and HPCC.

Comparison items	HPCC	HADOOP
Clusters	Performed in Thor and Roxie	Perform with MapReduce processing
Programming language	ECL is programming language	MapReduce based on Java language.
Data Warehouse	HPCC uses Roxie for structural queries	Hive present the abilities of data warehouse and enable data to be loaded into HDFS.
Performance for a 400-node system	success is 6 minutes 27 seconds HPCC faster than Hadoop	success is 25 minutes 28 seconds

III. Related Work

Discovery of new knowledge could occur, assuming that those correct performing artists would introduce also adjusted. Components for example, such that intuition, acuteness, and the likelihood of claiming perception need aid troublesome with control. For example, Begoli furthermore Horey [8] intended three standards that need aid broadly worried will boost the controllable factors. KDD from enormous data performs in three standards utilizing Hadoop. Guideline person said that KDD includes a mixture of analysis techniques like distributed programming, pattern recognition, data mining, natural language processing, sentiment analysis, statistical, visual analysis, and human-computer interaction. Therefore, those structural engineering must back different techniques and explanatory strategies. Measurable Investigation will be intrigued by summarizing gigantic datasets, comprehension data. Also, characterizing models to prediction data mining correlates with the finding of service models clinched alongside enormous datasets by itself. Machine learning in combines for data mining also measurable techniques making a difference machines with seeing all the datasets. Visual dissection is a creating zone over which huge datasets administration to clients On testing routes will have the capacity with getting it associations. Guideline said that a generally speaking, KDD structural engineering must obtain to have furthermore worked that procedure line. Preparation of data also clumps analytics will be aggravated by troubleshooting errors, forgetting values, and unusable configuration. Transforming organized and semi-structured data. Furthermore, guideline three said that settling on the comes about approachable also idiot proof making a difference succeed data issues [2].

KDD starting with Big Data could see tolerance to have deeper insights. The past three standards that have been honored for KDD starting with Big Data toward ORNL (Oak Edge National Laboratory).

Gosain and Chugh [18] exhibited two new standards dependent upon security furthermore auspicious Investigation for KDD starting with enormous data. The guideline particular case gave those issue from claiming timeliness. Timeliness alludes of the rate for data securing Furthermore data analysis, guideline two Ensures Big Data security will be significant a result security and security ended up a critical issue. The additional those data will be available; the preferred is those analytics. Anyway it is a greater amount powerless on dangers. These standards might help associations will accomplish better outcomes about Big Data analytics. Emulating these principles, possibility profits of Big Data camwood a chance to be understood in the practically proficient way. Cost, ongoing also future fill in will at present deliver concerning illustration enormous data issues.

Previously, data put away done data warehouses taken after some scheme, what's more, Institutionalization that prompt, effective data mining. However, to later years' data seen fluctuating. NoSQL databases have been recommended toward Lomotey also Deters [12] on the suit the data. Not a number devices would be accessible on performing data mining also analytics from such storages. IBM Scrutinize need identifier AaaS (Analytics-as-a-Service) as a territory that might offer benefits of the business worth. It will be a direct result AaaS camwood support in the conversion of unstructured data for benefits of the business production ventures; future development expects during incorporating Taking in what's more versatility Characteristics. It will further encourage the capacity should aggravate proposals on clients gave their expression, what's more, subject mining prerequisites.

Joining about data starting with different heterogeneous wellsprings under a serious data model that permits canny querying a chance to be a paramount open issue in the range of huge data. Traditionally, Extract-Transform-Load (ETL) strategy need to be utilized to joining data in the business. Therefore, Bansal [19] suggested a semantic ETL skeleton. It utilization semantic advances to prepare rich what's more serious data. It In light of data joining also furnished semantic data distributed on the web. The great production of the scheme will test those recommended technique, which utilized the semantic ETL procedure will incorporate a couple of state funded information sets. This procedure will incorporate taking a gander under ongoing data, what's more, entryway semantic ETL could provide assistance for its integrative should Fabricate requisitions to different domains, for example, healthcare, training with rundown a couple.

Assuncao et al. [20] aggravated an essential proposition of a dispersed group classifier algorithm given those well-known irregular Forests to Big Data. That suggested algorithm expects with finer that effectiveness of the calculation eventually Tom's perusing and conveyed preparing model known as MapReduce, lessen that arbitrariness sway by staying should stochastic mindful irregular Forests algorithm (SARF). That calculation may be secured ahead two crucial components: SARF to raise those nearby models also MapReduce on methodology extensive scale data on an spread (and parallel) mode.

Liao what's more in length [24] recommended an acclimatization of Big Data mining parallel algorithms-MRPrePost. MRPrePost is a parallel algorithm depended on Hadoop stage. It evolves PrePost eventually Tom's perusing method for including a prefix pattern, and once this support in the parallel configuration thoughts. That mrprepost algorithm could adjust to mining huge data's affiliation decides. Trials need to be demonstrated that MRPrePost algorithm is a greater amount predominant over PrePost furthermore PFP viewing performance, and the solidness and adaptability.

Prabha and Sujatha [22] discussed that multiple source produces a mass production of data creating Big Data. The main purpose is to extract useful information from large data volume. Clustering Incremental Weighted Fuzzy C-Means (IWFCM) introduced weight that described the importance of each object in the clusters. IWFCM produced cluster with a minimum run times and high quality. The e-book dataset is performed over the Hadoop environment that implements over MapReduce framework and data is reduced using IWFCM. A powerful best approach will bunch that enormous volume from claiming information may be IWFCM to loadable furthermore unloadable datasets. The disseminated nature's domain need to be set up the place the gigantic datasets require should have a chance to be trimmed. Those The progressive fuzzy clustering camwood an opportunity to be propelled eventually Tom's perusing included indexes of the dispersed platform similar to Hadoop to recovery for data. Those run duration of the time camwood makes abbreviated because of indexing.

Shuliang et al. [3] talked about spatial data mining frameworks that point with make spatial data bit by bit summarized under spatial knowledge. Through those mix of space data, profoundly extracting knowledge. Eventually, Tom's perusing utilizing such new learning to transforming knowledge to ongoing will comprehend, apply furthermore settle on canny judgments furthermore well-informed choices. Space learning might be self - learning, self-enhance, universal, also effectively perceived. It Might serve Similarly as a foundation to the choice backing. In organizations take a full point from claiming spatial knowledge, it will a chance to be that is only the tip of the iceberg exacts, what's more, dynamic to people should learn, work, life, also accomplish intelligence state. It will serve as move forward asset utilize and gainfulness level. Borne [21] talked about learning discovery`s center technique furthermore era of quality starting with data will be a data science. A standout amongst those The greater part significant methodologies may be facts. The relationship between causation depended on looking into factual considering viewed as the universe for Big Data.

Fania furthermore Mill [23] examined the mining capability also, dissect Big Data. It grants organizations deeper, what's more, richer insights under benefits of the business design also pattern. It additionally aids operational efficiencies furthermore aggressive focal point in distinctive fields about distinguishing those most up to date key advantage furthermore significant insights. Fundamental classes need aid enormous database holds structured data that would excessively little. Profound analytics used to discover replies will complex, Big Data visualization, also analytics devices assistance should benefit important insights through progressive refinement furthermore generalization. A Stage In light of a blending of the massively parallel preparing (MPP) data warehouse machine also groups about industry-standard servers running Apache Hadoop.

IV. Applications

Effectiveness furthermore intensity for enterprises` generation might be a chance to be improved by provision of Big Data.

Looking into marketing, correlation analysis of Big Data could assist the Association to foresee the purchaser conduct technique furthermore find new business models, for example, bargains arranging after correlation for enormous data that aides associations with streamline their costs.

Toward utilizing Big Data in the supply chain, enterprises might direct stock optimization, logistic optimization, and supplier coordination should close the hole between supply, what's more, demand, control budgets, furthermore enhance benefits.

Big Data analysis aides clinched along back. To example, China vendors Bank (CMB) with use eventually Tom's perusing distinguishing exercises as “Multi-times score accumulation” and “score return in shops” are functional to attracting personal satisfaction clients. Toward building a client drop out cautioning model, the bank camwood offers high-yield money related results. Dissecting customers’ transaction records camwood a chance to be proficiently distinguished toward applying huge information investigation. It aides respectable execution additions were attained [25].

Previously, 2008, Farecast might have been bought Tom's perusing Microsoft eventually, Forecast, need an aerial shuttle ticket conjecture framework that anticipates the developments also rising/dropping ranges from claiming air transport ticket value. That framework needs to be coordinated under the Bing web index toward Microsoft. Eventually, Tom's perusing 2012, that framework need to be spared about 50 USD for every ticket for every passenger, with the forecasted precision concerning illustration Helter Skelter as 75 % at present [25, 27].

Applying predictive analysis for Big Data aided the what's to come for the organization. What's more, the lion's share of corps parts does not stay in their starting work areas once their comm. Santa Clause Cruz Police section will uncover that wrongdoing. By dissecting SNS, the police division could find wrongdoing patterns furthermore wrongdoing modes, furthermore actually foresee the wrongdoing rates to real regions) [27].

For April 2013, more than particular case a million American utilized from claiming Facebook. In this way Wolfram Alpha (a registering also internet searcher organization) investment around mulling over the theory from claiming social conduct technique by examining social data. As stated by that analysis, practically from claiming Facebook clients become hopelessly enamored done their initial 20s, furthermore get locked in when they are around 27 a long time old, then get hitched when they need aid over 30 a considerable length of time old. Finally, their marriage connections show moderate transforms between 30 also 60 quite sometimes old [25].

Applying in learning will be the engineering that employments spatial Big Data mining technique. It extracts a while ago unknown, possibly useful, what's more, extreme frisbee intelligible standards [3].

That tourism industry camwood profit from learning revelation strategies will Figure concealed data. Starting with examining on the web tourists’ profiles, this learning might be a chance to be concentrated. Bring of shortages from claiming analysis procedure serves will recognize visitors’ behavior, move forward offices, what's more, administrations furthermore meet different inclination from claiming visitors [26].

It camwood be connected will anticipate race result, for example, mining Twitter, Big Data will foresee 2013 Pakistan race victor, what's more, Twitter, need to be investigated in the 2008 us presidential races. Topsy investigated tweets something like both the presidential hopefuls (Obama, what's more, Romney) also computed their Notoriety Score given that assumption available in the users' tweets. A Big Data investigation venture to a Stanford course embraced a related approach of the 2012 us presidential races [28]. Requisitions likewise show up in movement data framework that produces huge movement data with gathering ongoing GPS (Global Positioning System) data, matching positions to produce stream movement majority of the data. It acquires incredible business sectors previously, China [30].

Previously, social insurance, what's more, medicinal administrations therapeutic enormous data provisions are quickly developing withhold numerous abundant furthermore different majority of the data values. The requisition from claiming enormous therapeutic data will profoundly impact the medicinal services business [25]. Big Data analytics and health awareness restorative professionals store a tremendous add up of data over patients’ therapeutic history, medication, also different points. Medication manufacturing endeavor saves the enormous sum from claiming data. These data are altogether perplexing over way what's more here and there professionals cannot associate with other data. Thus, brings about the incredulous majority of the data remain Hidatsa. Toward applying propelled systematic techniques, this stowed away data might be extracted, which brings about customize the solution. Propelled analytics systems camwood additionally accumulate knowledge under inherited, what's more, ecological developments from claiming ailments [28]. For example, throughout the 2009 influenza pandemic, Google procured auspicious data by analysis Big Data. Google discovered that Throughout those spread of influenza, sections every now and again looked during its hunt engines should figure those are spreading for flu furthermore actually identified the starting place to spread the flu. The related analysis comes about bring been distributed to nature [25, 27].

V. Challenges

A standout amongst the remarkable tests is the change of enormous volumes of quantitative data under the qualitative majority of the information that help over generally speaking existence fulfillment [31].

Eliciting serious data from this data is not those best test, however, should acquire data also, obscure knowledge discovery, search for patterns, and with bode well of the information. The stupendous test may be to prepare suitable data to what's more utilized by those end client. Perhaps, the fundamental issue will be that interaction, in light of it will be the humanity's end client who needs the problematic fathoming intelligence, henceforth that ability about asking canny inquiries regarding information. The issue in the term sciences will be that (biomedical) data models would describe toward critical complexity, settling on manual analysis toward those limit clients [4].

Overloading from claiming the majority of the biomedical data is the present challenge. The requirement to gather data an enormous volume for structured, semi-structured, unstructured data will streamline workflows, courses also guidelines, to expansion ability same time cutting costs and creating efficiencies will help a short overview from claiming intuitive also coordinated results to data disclosure also the majority of the data mining. That mossy oak huge challenges, including, the have on got ready also, applies novel methods, algorithms, also devices for that integration, fusion, pre-processing, mapping, expository considering furthermore translation for complex biomedical data with the plan on should recognizing testable hypotheses [4].

For universal, management of enormous data confronts with various challenges as impostor of data, expanding of data each day, speed of data is expanding snappier over ever, a grouping of data has a need to aid in inflating, volume (storage, what's more, get to precise substantial data), veracity (managing data) [28].

A. **Privacy,** security, what's more, trust - for associations every one security, what's more, security related go about to uplift that security for and situated reasonable limits for utilization of personage information. Dependence on the framework necessities on is safeguarded as those heft from claiming data holding builds. The certainty clients endure previously, these organizations also their abilities on safely hold information of a personage could effortlessly be influenced Tom's perusing spillage eventually from claiming data alternately data under people in general area [28]. Security is very important issue, companies invest a lot of efforts for securing data and keep customers` privacy. Data mining helps in intrusion detection [29].

B. **Data Management and Sharing** - Data Management and Sharing - organizations understand that for data will have whatever value, it obliges will a chance to be discoverable, also open accessible. Orgs must accomplish these requirements, be that as still adhering to security laws [28].

C. **Technology and analytical skills** - enormous data also analytics set parts of anxiety ahead ICT suppliers on building up novel instruments also innovation should deal with complex information. Current instruments furthermore building sciences are unabated on the store; transform also dissects that massive amount of the different majority of the information. Marketers, what's more, developers about Big Data frameworks also results including open hotspot product are getting more fit devices with improving the tests for enormous data analytics [28].

D. **Data Storage and Retrieval -** right now accessible advances would be skilled for taking care of data passage also information memory. Best those instruments planned to transaction transforming that will add, update, and hunt for Big Data [28].

E. **Data Growth and Expansion** - As the organizations increase their services, their data are also anticipated to rise. Few organizations also consider data expansion because of data grow in richness, data evolved with new techniques [28].

F. **Speed and scale -** when that volume of data increases, it is difficult with accomplishing knowledge under information inside the period. Procuring knowledge under data may be that is only the tip of the iceberg significant over preparing finish situated from claiming the majority of information. Transforming close to ongoing data will continuously require transforming interim to process fulfilling yield [28].

G. **Structured and unstructured data** - Transition between structured data stored in clear tables and unstructured information (pictures, pictures, text) required for depth psychology will bear on close to terminate processing of data. The innovation of new non-relational technologies will offer more or less flexibility in data representation and processing [28].

H. **Data ownership** - precise inconceivable measure for data resides on the servers from claiming social media administration suppliers. They do not own this data, in any case, they store data for their clients. The genuine holder of the page is those one who need to make those page or report card [28].

VI. Future Directions to Face Challenges

The HCI-KDD approach: it may be a blended from claiming methodologies over two areas, Human-Computer Cooperation (HCI) also KDD & Data Mining. It displays flawless states towards settling Big Data tests. Its objective may be to upgrade humanity's discernment action with machine brainpower. KDD transform to this depiction of a few issues also tests for machine of the humanity's - isolated under four ranges [4]**:**

A. **Intelligent information integration, the majority of the information fusion, also pre-selection performed on data sets**: it will be conveyed extensive amounts from claiming data in the gigantic multifaceted nature. Extensive medical data experiences situated of issues that could make part of three classes:

- Data sources are heterogeneous
- The data is in high degree of complexity (high-dimensionality).
- Noisy, uncertain data, dirty data, the discrepancy between data, information, and knowledge of Big Data sets.

It is a great challenge to integrate and fuse the biological data together with classical patient records and medical image data.

B. **Sampling, cleansing, preprocessing, mapping, the issue for blending numerous data sets concerning regular substances** will often be encountered clinched alongside KDD, often known as blend issue. Purifying data to impurities prompted that development about an extensive variety of routines will upgrade the exactness also thereby that approachability about existing majority of the information utilizing huge numbers machine learning algorithms.

C. **Advanced data mining methods, pattern discovery many data mining methods are designed for collections of objects well-represented in rigid tabular formats.** Advanced data mining approaches include:

- Graph-based data mining.
- Entropy-based data mining.
- Topological data mining.

Broadly speaking, information theory relates to quantifying data and to investigate communication processes using previous approaches.

D. **Intelligent media visualization, HCI, analytics.Finally, those effects on the provision of complex publicizing for secondary dimensional data:** we could pronounce at the same time, dimensional mathematically could be exceeded, we might recognize and bring down measurements that contribute to the meaning about visualization similarly to the mapping those higher, the more level dimensional space, a transform that continuously bears the hazard about demonstrating artifacts. Despite visualization may be developped for a foundation from claiming a few decades, a significant issue may be that nonattendance of finish devices that helps all analysis tasks**.**

E. **Horizontal area: Privacy, data protection, data security, data safety:** Dealing with data issues of privacy, data security, information security and data safety and the fair employment of data are of paramount importance. Its importance involves data accessibility, temporal limits, legal restrictions (such as copyright or patents may be relevant), confidentiality and data provenance.

VII. CONCLUSION

KDD is the method of making valuable and ultimately intelligible patterns from Big Datasets. Big Data expands every moment. Big Data applications cause big profit for banking, business organizations (Microsoft), tourism, elections and medical fields. It is well known that expansion of data has caused the huge volume of data from different sources in various structures.

The data generates in a rapid rate of speed with fewer data quality. All earlier cases represent huge challenges to apply KDD methods and algorithms for integration, preprocessing and management data. So, there is an emerged need to use new working environments like Hadoop to simulate the heterogeneous of data sources. Applying KDD algorithms help to extract the value of data in the intelligent method. Intelligent integration, selecting relative features, cleansing, preprocessing, and big representation data will be the important future research points.

REFERENCES

[1] Gupta, Richa. "Journey From Data Mining To Web Mining To Big Data", International Journal of Computer Trends and Technology 10.1, 18-20. Web, (2014).

[2] SAGIROGLU, and SINANC, "Big Data A Review", Collaboration Technologies And Systems (CTS), 2013 IEEE International Conference on San Diego, CA: IEEE, (2013).

[3] Shuliang, Gangyi, and Ming,"Big Spatial Data Mining", 2013 IEEE International Conference on Silicon Valley, CA: IEEE, (2013).

[4] Holzinger, Dehmer, and Jurisica. "Knowledge Discovery and Interactive Data Mining in Bioinformatics, Future Challenges and Research Directions", BMC Bioinformatics, (Last accessed on 2016).

[5] Katal, Wazid, and Goudar, "Big Data: Issues, Challenges, Tools and Good Practices, Contemporary Computing (IC3), Sixth 2013 IEEE International Conference on Noida: IEEE, (2013).

[6] Bandyopadhyay, Sanghamitra, "Advanced Methods for Knowledge Discovery from Complex Data", New York, Springer, (2005).

[7] Krishnan, "Data Warehousing in the Age of Big Data". Print., ISBN 978-0-12-405891-0.

[8] Begoli, and Horey," Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), Joint Working IEEE/IFIP Conference on Helsinki: IEEE, (2012).

[9] Data, Information, Knowledge & Wisdom", http://www.systems-hinking.org/dikw/dikw.htm,Systems-thinking.org,2005, (Last accessed on 2016).

[10] https://datajobs.com/what-is-big-data, Frank, Lo. "What Is

Big Data: The Complete Picture, Beyond The 4 V's." Datajobs.com", 2015. Web.(Last accessed on 2015).

[11] https://www.ida.gov.sg/~/media/Files/Infocomm%20Landscape/Technology/TechnologyRoadmap/BigData.pdf, (Last accessed on 2015).

[12] Lomotey, and Deters, "Towards Knowledge Discovery in Big Data", Service-Oriented System Engineering (SOSE), 8Th, 2014 IEEE International Symposium on Oxford: IEEE, (2014).

[13] https://datajobs.com/what-is-hadoop-and-nosql, Frank, Lo, "What Is Hadoop and NoSQL?", "Datajobs.com", 2015. Web (Last accessed on 2015).

[14] https://thinkbiganalytics.com/leading_big_data_technologies/nosql/, Thinkbiganalytics.com, "NoSQL | Think Big Analytics", 2014, Web. (Last accessed on 2015).

[15] https://thinkbiganalytics.com/leading_big_data_technologies/hadoop/, Thinkbiganalytics.com, "Hadoop Ecosystem: Think Big Analytics, 2014, (Last accessed on 2015).

[16] https://thinkbiganalytics.com/leading_big_data_technologies/machine-learning-in-hadoop-with-mahout/, Thinkbiganalytics.com, "Machine Learning in Hadoop: Think Big Analytics", 2014, (Last accessed on 2015).

[17] http://hpccsystems.com/Why-HPCC/How-it-works#ecl, Hpccsystems.com, "How It Works | HPCC Systems", 2015, (Last accessed on 2015).

[18] Gosain, and Chugh. "New Design Principles For Effective Knowledge Discovery From Big Data", International Journal of Computer Applications 96.17 (2014).

[19] Bansal, Srividya, and Kagemann, "Integrating Big Data: A Semantic Extract-Transform-Load Framework", Computer 48.3 (2015).

[20] Assuncao, et al," “Distributed Stochastic Aware Random Forests - Efficient Data Mining For Big Data". Big Data Congress, 2013 IEEE International Congress on Santa Clara, CA: IEEE, (2013).

[21] http://statisticsviews.com/details/feature/4911381/statistical-truisms-in-the-age-of-big-data.html, Borne, Kirk, "Bias in Randomised Factorial Trials", (Last accessed on 2015).

[22] Prabha, Sujatha, “Reduction of big data sets using fuzzy clustering”, International Journal of Advanced Research in Computer Engineering & Technology, (2014).

[23] Fania, Miller, “Mining big data in the enterprise for better business intelligence”, Intel white paper, www.intel.com/it, (2014).

[24] Liao, Long, “MRPrePost-A parallel algorithm adapted for mining big data”, Electronics, Computer, and Applications, 2014 IEEE Workshop on Ottawa, (2014).

[25] Chen, Mao, Y. Liu, “Big Data: A Survey,” Springer Science+Business Media New York, (2014).

[26] Aghdam, Kamalpour, Chen and Sim, "Identifying Places of Interest for Tourists using Knowledge Discovery Techniques", Industrial Automation, Information and Communications Technology (IAICT), 2014 International Conference on Bali, (2014).

[27] Mayer, Cukier,"Big Data: A Revolution That Will Transform How We Live, Work and Think", ISBN: 1848547927, UK, (2013).

[28] Chandarana, Vijayalakshmi, "Big Data Analytics Frameworks ", Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on Mumbai, (2014).

[29] Azad, Jha," Data Mining in Intrusion Detection: A Comparative Study of Methods, Types and Data Sets", I.J. Information Technology and Computer Science, (2013).

[30] Yu, Jiang, Zhu, "RTIC: a big data system for massive traffic information mining ", Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on Fuzhou, (2013).

[31] http://futurememes.blogspot.com/2014/03/big-data-becomes-personal-knowledge.html, Blogga, _La. "Broader Perspective: Big Data Becomes Personal: Knowledge_into_Meaning", Futurememes.blogspot.com, (2014), (Last accessed on 2015).

Authors’ Profiles

Mai Abdrabo is a demonstrator at the Faculty of Computers and Information, Suez Canal University. She obtained her bachelor's degree from Faculty of Computers and Information Suez Canal University in 2013. Her current research interests are Big Data and knowledge discovery.

Mohammed Elmogy is an associate professor at Information Technology Dept., Faculty of Computers and Information, Mansoura University, Egypt. He had received his B.Sc. and M.Sc. from Faculty of Engineering, Mansoura University, Egypt. He had received his Ph.D. from Informatics Department, MIN Faculty, Hamburg University, Germany in 2010. He has authored/coauthored over 70 research publications in peer-reviewed reputed journals, book chapters, and conference proceedings. He has served as a reviewer for various international journals. His current research interests are Computer Vision, Machine Learning, Pattern Recognition, and Biomedical Engineering.

Ghada Eltaweel is an associate professor at computer science Dept., Faculty of Computers and Information, Suez Canal University, Egypt. She had received her B.Sc. from faculty of science Cairo University. She had received her M.Sc. from Helwan University, Egypt. He had received her Ph.D. from faculty of computers and informatics Cairo University. Her current research interests are Machine Learning, and biomedical image processing.

Sherif Barakat an associate professor of Information Systems, Faculty of Computers and Information, Mansoura University, Egypt.

How to cite this paper: Mai Abdrabo, Mohammed Elmogy, Ghada Eltaweel, Sherif Barakat, "Enhancing Big Data Value Using Knowledge Discovery Techniques", International Journal of Information Technology and Computer Science (IJITCS), Vol.8, No.8, pp. 1-12, 2016. DOI: 10.5815/ijitcs.2016.08.01