International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 1, 14-25
Available online at http://pubs.sciepub.com/iteces/4/1/3
© Science and Education Publishing
DOI:10.12691/iteces-4-1-3

An Experiential Study of the Big Data

Yusuf Perwej*

Department of Information Technology, AI Baha University, Al Baha, Kingdom of Saudi Arabia (KSA)
*Corresponding author: yusufperwej@gmail.com

Abstract The intention of this paper is to evoke discussion rather than to provide an experiential extensive survey of big data research. The Big data is not a single technology but an amalgamation of old and new technologies that assistance companies gain actionable awareness. The big data are vital because it empowers organizations to congregate, store, manage, and manipulate countless amounts data at the pertinent speed, at the pertinent time, to gain the pertinent intuition. Eventually big data solutions and practices are typically essential when eternal data processing, analysis and storage technologies and techniques are inadequate. In particular, big data addresses detached requirements, in other words the amalgamate of multiple un-associated datasets, processing of huge amounts of amorphous data and harvesting of unseen information in a time-sensitive genre. In this paper, aimed to demonstrate a close-up view about big data, including big data concepts, security, privacy, data storage, data processing, and data analysis of these technological developments, we also brief description about the characteristic of big data, big data techniques, technologies and tools emphasizes critical points on these issues.

Keywords: *datasets, big data, differential privacy, anonymization, diagnostic analytics, data storage*

Cite This Article: Yusuf Perwej, "An Experiential Study of the Big Data." *International Transaction of Electrical and Computer Engineers System*, vol. 4, no. 1 (2017): 14-25. doi: 10.12691/iteces-4-1-3.

1. Introduction

The big data are becoming one of the most necessary technology trends that has the forceful for dramatically transshipment the way organizations use information to enlarge the customer experience and transform their business models [1]. The flourish of technologies and services, the huge amount of data is produced that can be structured and unstructured from the various sources. This similar type of data is very arduous to process that contains the trillion records [2] of millions and millions people information that includes the social media, audios, web sales, images and so on. The necessity of big data comes from the larger companies like Google, Facebook, Yahoo, etc. The intention of analysis the huge amount of data which is in unstructured form [3]. Google contains the volumetric amount of information. Hence there is the necessity of big data analytics that is the processing of the complex and vast datasets. Big data analytics analyze the huge amount of information used to unwrap the cryptic patterns and the additional information which is advantageous and [4] essential information for the utilization [2]. Big data is a comprehensive term for any collection of data sets so enormous or complex that it becomes unintelligible to process them using customary data management techniques [5]. As companies begin to assess novel types of big data solutions, many recent lucky chances will unfold. For example, manufacturing companies may be able to keep an eye on data coming from machine sensors to find out how processes need to be modified before a objectionable event happens. In so far as it will be possible for retailers to observe data in real time to upsell customers respective products as they are executing a transaction.

Big data technology can be used in healthcare to determine the cause of an illness and endow a physician with guidance on treatment rational choice. Big data is not [2] a distinct solution, however, put into action a big data solution requires that the infrastructure is in place to endorse the [6] distribution, scalability, and management of that data. In as much as, it is important to put both a business and technical strategy in place to make use of this an essential technology trend. For many [7] essential reasons, we think that it is compulsory for you to understand big data technologies and be aware the ways that companies are using emerging technologies [8] such as Hadoop, MapReduce, and modern database engines to transform the value of their data. Big Data is a field faithful to the processing, analysis, and storage of huge collections of data that frequently originate from contrasting sources.

2. What is Big Data?

The Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for distinguishable and timely decision making. The big data mention to datasets whose size on the far side of the capability of typical database software tools to hold, store, manage, and analyze [2]. Big data delineate a holistic information management master plan that includes and integrates many recent types of data and data management side by side traditional data. Big data is a manifest that describes the huge volume of data both structured and

unstructured that inundates a business on a diurnal basis. However, it's not the amount of data that's essential [7]. It's what sort organizations do with the data that reason. Big data can be analyzed for insights that lead to preferable decisions and astute business moves. From above big data adds recent techniques [8] that zestful computational resources and move toward to execute analytic algorithms [9]. This shift is necessary as datasets continue to become bigger, more diverse, more sophisticated and [10] streaming pivotal. The big data introduce novel technology, skills to your information architecture, processes, and the people that design, operate, and utilize them.

3. The Big Data Concepts and Terminology

The big data are used as a concept that refers to the incapacity of traditional data architectures to dexterously conserve the new data sets. In this paper various fundamental concepts and Terminology are used in big data.

3.1. The Datasets

The collections or groups respective data are normally referred to as datasets shown in Figure 1. Every group or dataset member (datum) shares the identical set of attributes or properties as [11] others in the identical dataset. A data set is a collection of respective, discrete items of respective data that may be accessed singly or [12] in combination or managed as an entire entity. Few examples of datasets are firstly a collection of image files in a directory secondly the historical weather observations that are accumulated as XML files.

3.2. The Data Analytics

The data analytics is a comprehensive term that encompasses data analysis. Data analytics is a discipline that includes the management of the absolute data [13] life cycle, which surrounded by collecting, storing, analyzing, cleansing, organizing and governing the data. The term comprises the development of analytical methods, automated tools, scientific techniques [14]. In Big Data environments, data analytics have evolved methods that allow data analysis to happen through the use of exceedingly scalable distributed technologies and frameworks that are competent of analyzing huge volumes of data from dissimilar sources.

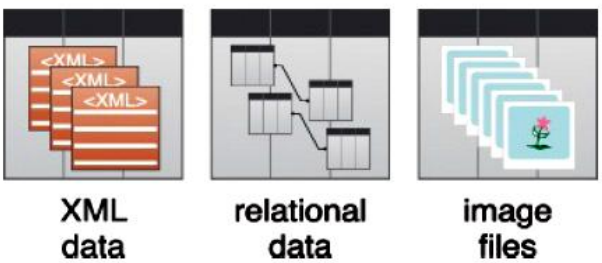

Figure 1. The Three Datasets Based On Three Different Dataformats

3.3. The Data Analysis

The data analysis is the procedure of investigating data to find reality, patterns, relationships, insights and/or trends. The collective pursuit of data analysis is to [13] support preferable decision making. A straightforward data analysis example is the analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the everyday temperature [14]. The outcome of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information. Again the carrying out data analysis assist establish patterns and relationships among the data being analyzed.

4. The Data Analytics

In this paper, I have discussed various analytics types leverage various techniques and analysis algorithms. This insinuates that there may be varying data, storage and processing exigency to facilitate the delivery of multiple types of analytic results. There are four usual [14] categories of analytics that are specific by the results they produce shows in Figure 2.

4.1. The Descriptive Analytics

Descriptive analytics are an elementary stage of data processing that creates an essence of historical data to yield advantageous information and possibly prefabricate the data for afore analysis. Descriptive analytics keep together answering questions about events that have previously occurred. This form of analytics contextualizes the data to proliferate information [15]. Descriptive analytics is infrequently said to provide information about happened. The Descriptive analytics are mostly carried out via ad-hoc reports or dashboards. The reports are commonly fixed in nature and demonstration documented data that is offered in the form of data grids or charts. Descriptive analytics or data mining is at the nethermost of the big data value chain, but they can be valuable for exposing patterns that offer discernment. A common example of descriptive analytics [16] would be assessing credit risk using primal financial accomplishment to predict a customer's suitable financial performance.

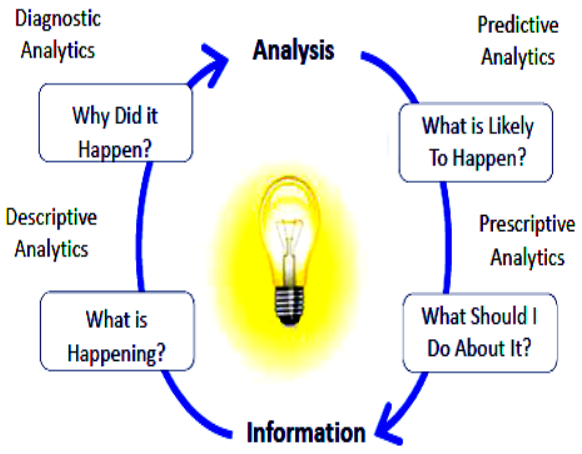

Figure 2. The Four Categories Of Analytics

4.2. The Diagnostic Analytics

The diagnostic Analytics are a form of advanced analytics, which investigate data or content to answer the question like Why did it happen?, and is also characterized by techniques such as data discovery, data mining, drill-down and correlations. The diagnostic analytics objective to determine the cause of an incidence that occurred in the previous using questions that focus on the reason behind the occurrence. The target of this category of analytics is to determine what information is respective to the incidence in order to enable answering questions that solicit to determine why something has occurred [16]. Diagnostic analytics are a far down look at data to afford to perceive the causes of incident and behaviors. Diagnostic analytics endow copious value than descriptive analytics, however, need a more state-of-the-art proficiency. Diagnostic analytics normally need to collect data from many sources and storing it in a structure that impart itself to performing roll-up and drill-down analysis. Diagnostic analytics are used to find or to determine why something befall. For example, in a social media marketing expedition, you can utilization descriptive analytics to evaluate the number of fans, page views, reviews, pins, posts, mentions, followers, etc.

There can be thousands of online mentions that can be refined into a single view to look what worked in your previous campaigns and what didn't.

4.3. The Predictive Analytics

The predictive analytics, which is used to recognize future probabilities and inclination, say endow information about what might befall in the forthcoming. The predictive analytics, information are [16] intensify with meaning to procreate knowledge that enlighten how that information is related. The potency and magnitude of the associations form the basis of models that are used to generate future predictions based upon bygone phenomena. Predictive analytics try to predict the consequence of phenomena, and predictions are made based on patterns, trends and exceptions found in decrepit and present data. This can headship to the identification of both risks and occasion. Predictive analytics use big data to identify previous patterns to predict the oncoming days. For example, some companies have gone one step further use predictive analytics for the [17] all sales process, analyzing headship source, social media, documents, CRM data, number of communications, types of communications, etc. This kind of analytics necessitates the use of huge datasets comprised of internal and external data and dissimilar to data analysis techniques. It comes up with greater value and requires a more advanced proficiency than both descriptive and diagnostic analytics.

4.4. The Prescriptive Analytics

The prescriptive analytics is well-becoming to try to identify the optimal outgrowth to events, given the parameters, and propose decision options for optimal take advantage of a forthcoming opportunity or detract a further avail e risk. Prescriptive analytics build upon the outcome of predictive analytics by make for actions that should be taken. Prescriptive analytic attempt to measure the effect of subsequent decisions in order to advise on possible outcomes before the decisions are actually made. The prescriptive analytics predicts not only what will happen, but also how come it will happen providing recommendations [16] regarding actions that will take gain of the predictions. Prescriptive analytics endow more value than any isolated type of analytics and correspondingly require the most state-of-the-art cuteness, as well as correlate with unique software and tools [18].

5. The Key Characteristics of Big Data

The term Big Data appear to be popping up ubiquitously these days. The majority of these data characteristics were at the beginning identified by [19] when he published an article delineate the influence of the volume, velocity and variety of e-commerce data on enterprise data warehouses. The veracity has been anthologized to account for the lower signal-to-noise ratio of unstructured data as differentiate to structured data sources. Eventually, the aim is to conduct analysis of the data in such a manner that high-quality outcome are delivered in a timely demeanor, which endow optimal value to the enterprise. To perceive the incidence in Figure 3 that is big data, it is generally described using five V's Volume, Velocity, Variety, Veracity and Value.

5.1. Volume

The big data imply spacious volumes of data. It used to be employees origination data. Presently that data are generated by networks, machines and human interaction with systems like social media the volume of data to be analyzed is enormous. Also, whether a particular data can in fact considered as a [20] big data or not, is a protege upon volume of data. The volume signifies to the vast amounts of data procreate every second. Just consider all the twitter messages, photos, video clips, emails, sensor data, etc. We genesis and share every second. We are not talking Terabytes, but Zettabytes or Brontobytes.

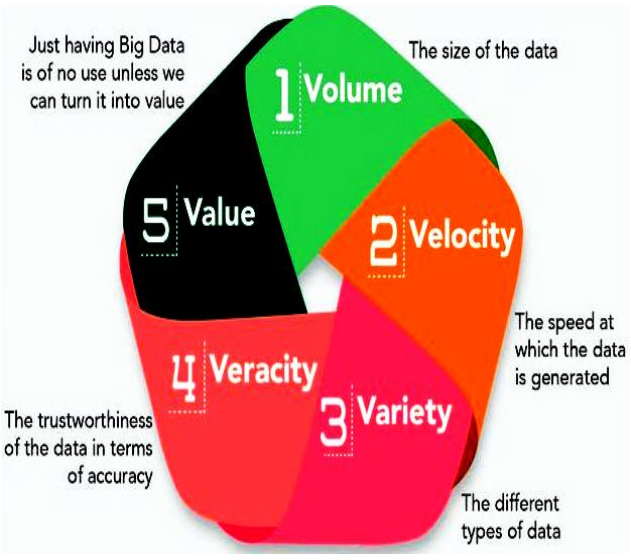

Figure 3. The Five V's of Big Data

5.2. Velocity

The velocity refers to the speed at which new data are originating and the speed at which data transmigration around. In big data environments, data can come at rapid speeds, and enormous datasets can assemble within very less periods of time. Velocity denote to the speed at which recent data are generated and the speed at which data transmigration around. From an enterprise viewpoint, the velocity of data translates into the amount of time it takes for the data to be processed as soon as it enters the enterprise's perimeter. Facing with the fast inflow of data be in need of the enterprise to design highly elastic and obtainable data processing solutions and corresponding data storage capabilities [19]. Big data technology allows us now to examine the data while it is being generated, without ever deposit it into databases. However, depending on the data source, velocity may not every time be high. For example, MRI scan images are not generated as repeatedly as log entries from an exalted traffic webserver.

5.3. Variety

Variety refers to the dissimilar types of data we can now use. The Data variety refers to the various formats and types of data that necessity to be supported by big data solutions. In variety distinct sources and the nature of data, both structured and unstructured. The early days, spreadsheets and databases were the only sources of data considered by most of the applications. At present, data in the form of videos, monitoring devices, PDFs, emails, photos, audio, pics, are also being considered in the analysis applications. Data variety brings defiance for enterprises in terms of data unification, variation, processing, and storage.

5.4. Veracity

The big data veracity refers to the biases, noise and dissimilarity in data. The data that is being stored, and mined well-becoming to the problem being analyzed. Veracity refers to the perfection or allegiance of the data. The data that enters big data environments needs to be assessed for perfection, which can lead to data processing activities to sort out invalid data and stripping noise [20]. In the veracity, data can be a portion of the signal or the noise of a dataset [21]. Noise is data that cannot be transformed into information and thus has no value, whereas signals have value and lead to significant information. The data with an exalted signal-to-noise ratio has more veracity than data with an infirm ratio. In your big data scheme, you need to have your team and partners work to help keep your data distinguishable, and processes to keep 'messy data' from deposit in your systems.

5.5. Value

The value it is all well and nice having access to big data, but however we can turn it into value it is unnecessary. You can nicely argue that value is the most crucial V of big data. The value specialty is intuitively respective to the veracity specialty in that the utmost the data fidelity, the more value it holds for the business. Value is also incumbent on how prolonged data processing takes because analytics outcome have a shelf-life. The ultimate challenge of big data is to convey value [20].

5.6. Other Characteristics

There are several other V's that get added to these depending on the context.

- **Valence:** This refers to how big data can restriction with each other, forming connections between or else disparate datasets.
- **Validity:** The big data veracity is the matter of validity, meaning is the data precise and immaculate for the intended use. The distinctly valid data is key to making the proper decisions.
- **Variability:** This refers to the ridiculousness which can be shown by the data at times, thus impede the process of being able to control and manage the data effectively. The transformation in the [20] data leads to wide transformation in quality. Moreover, resources may be needed to process, identify, or filter, obscure quality data to make it more advantageous.
- **Volatility:** The big data volatility refers to how protracted is the data valid and how protracted should it be stored. In the real world of real time data you need to determine at what point is data no longer episodic for the present analysis.

6. The Categories of Big Data

The data processed by big data solutions can be human-generated or machine-generated, in spite of the fact that it is finally the responsibility of machines to generate the analytic outcome. The human-generated data is included of the spreadsheets, [8] presentations, images, emails, Word documents, audio, and video files that we create and share with other human beings day-to-day. Human beings generated data is the consequence of human interaction with systems, such as digital devices and online services. The Machine-generated data are information that is the evident outcome of an application process, computer process, or created without human interference. Machine-generated data is originated by software programs and hardware devices in reaction to the real-world occurrence. For example, a log file outshine an authorization judgment made by a safety service, and a point-of-sale system originate a transaction opposed to inventory to reflect product purchased by a customer. From a hardware viewpoint, for example, machine-generated data would be information transport from the a lot of sensors in a cell phone that may be reporting information, together with the position and cell tower signal strength. As signified, human-generated and machine-generated data can come from a heterogeneity of sources and be represented in various formats or types in Figure 4. The primary types of data are:

- The Structured Data
- The Unstructured Data
- The Semi-structured Data

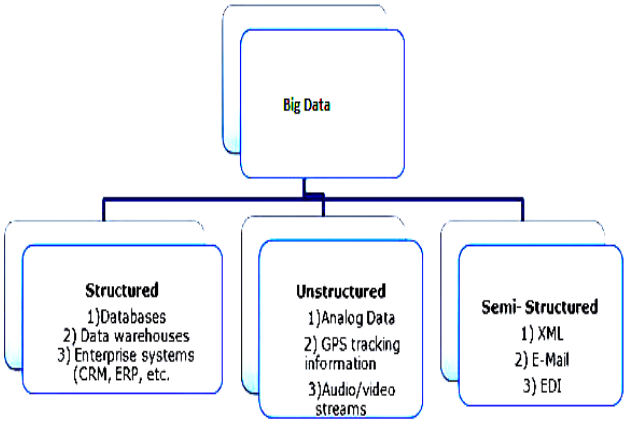

Figure 4. The Three Different Types Of Data

These data types refer to the inner organization of data and are sometimes called data formats.

6.1. Structured Data

Any data that can be stored, retrieved and processed in the form of definite format is termed as a 'structured' data. The structured data refer to kinds of data with an advanced level of organization, such as information in a relational database. When information is extremely structured and predictable, search engines can more comfortably organize and display it in imaginative ways. In the structured data markup [1] is a text-based organization of data that is incorporated in a file and served from the web. Structured data are continually cultivating by enterprise applications and information systems like CRM and ERP systems.

6.2. Unstructured Data

Any data with unacquainted form or the structure is classified as unstructured data. From above to the size being large, unstructured data poses various challenges in terms of its processing for obtaining value out of it. The unstructured data are the usual label for delineate data that is not contained in a database or some different type of data structure [8]. The unstructured data may be textual or non-textual. Textual unstructured data are generated in the media similar as PowerPoint presentations, Word documents, email messages, collaboration software and instant messages. Non-textual unstructured data are generated in media similar as MP3 audio files,GIF images, JPEG images, and Flash video files. Unstructured data cannot be straight, processed or queried using SQL. If it is requisite to be stored within a relational database and it is stored in a table as a Binary Large Object.

6.3. Semi-structured Data

The semi-structured data is data that has not been organized into a unique repository, such as a database, however, in spite of that has associated information, such as metadata, that makes it more responsible for processing than raw data. Semi-structured data is data that is neither raw data, nor typed data in a traditional database system. Alternatively, semi-structured data is hierarchical or graph-based. This variety of data is normally stored in files that contain text. The semi-structured data generally have special pre-processing and storage requirements, in particular if the underlying format is not text-based. For example, of pre-processing of semi-structured data would be the validation of an XML file to make sure that it abide by to its schema definition. One more example of semi-structured data [8] would be BibTex files or a Standard Generalized Markup Language (SGML) document. That files are semi-structured may accommodate rational data made up of records, but that data may not be organized in an identifiable structure. Few fields may be Traceless or contain information that can't be effortlessly described in a database system.

7. The Big Data Security and Privacy

The data are currently one of the most essential assets for companies in every field. Big Data originated new matter concerned with not only to the volume or the variety of the data, but also to data security and privacy. The multiformity of data format, data stream, data source, and infrastructures may reason peerless security vulnerabilities on big data. Security and privacy measures in big data essential scale nonlinearly.

7.1. Security

The security is the exercises of defending information and information assets by means of the use of technology, processes and training from Unauthorized access, Disruption, Modification [22], Inspection, Disclosure, Recording, and Destruction. The security focus on more protecting data from despiteful attacks and the wrong use of stolen data for profit. Big data security challenges can be split into four categories; data management, integrity and reactive security, infrastructure security, data privacy in Figure 5. Infrastructure security be made of secure distributed programming and security exercises in non-relational data stores. Next data privacy refers to privacy preserving analytic, encrypted data center and granular access control. The data management implies transaction logs, secure data storage, auditing and [23] data origination. In integrity and reactive security includes filtering, validation, and real time monitoring. A valuable portion of information security attempt goes into monitoring and analyzing data about phenomena on servers, networks and additional devices. In big data analytics at present applied to security monitoring, and they make able both broader and more in-depth analysis. In several ways, big data security analytics and analysis are an extension of security information and event management (SIEM) and respective technologies.

Next the big data security analysis tools usually extension of two functional categories first SIEM, and second performance and availability monitoring (PAM). SIEM tools typically reckon on event management, log management, [23] and behavioral analysis, as well as database and application monitoring. PAM tools, mainly relate to operations management. In spite of, big data analytics tools are more than just SIEM and PAM tools incorporate with together they are designed to collect, integrate and analyze huge volumes of data in close by real time, which requires various additional potential. The

SIEM, big data analytics tools have the potential to accurately explore devices on a network.

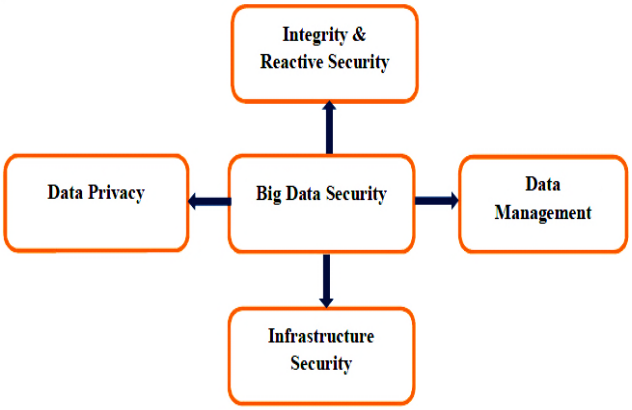

Figure 5. The Big Data Security

7.2. Privacy

The information privacy is the privilege to have some rein over how the particular information is collected and used. Information privacy is the competence of a personal or group to prevent information about themselves from becoming familiar to people other than those they give the information to [24]. The significant issue of user privacy is the recognizance of personal information during transmission over the Internet. Data privacy is an intention on the use and governance of personal data things like setting up policies in place to make certain that consumers personal information is being collected, shared [25] and utilized in pertinent ways. Some conventional methods for privacy preserving in big data are described in brief here in this paper.

7.2.1. De−Identification

De-identification is a conventional technique for privacy-preserving data mining, where in order to intercede personal privacy, [26] data should be first decontaminate with generalization and suppression before the deliverance for data mining. De-identification is a vital tool in privacy protection, and can be wayfaring to privacy preserving big data analytics.

7.2.2. K-Anonymity

In the freeing of data is said to have the k-anonymity property if the information for each person contained in the freeing cannot be perceived from at least k-1 individuals whose information demonstrate in the freeing. In the circumstances of k-anonymization issue, a database is a table which consists [26] of n rows and m columns, where each row of the table signify a record relating to a [27] distinctive individual from a population and the entries in the different rows need not be peerless. The values in the [28] dissimilar columns are the values of attributes associated with the members of the populace.

7.2.3. L-Diversity

This is a form of group based anonymization that is making use of preservation, privacy in data sets by decrease the granularity of data representation. This reduction is a trade-off that consequence in some loss of viability of data management or mining algorithms for gaining some privacy [27]. The l-diversity model is a dispersion of the k-anonymity model which diminishes the granularity of data representation [29] make use of methods, including popularization and suppression in a way that any given record maps onto at least k various records in the data.

7.2.4. T-Closeness

This is a further refinement of l-diversity group based anonymization that is used to intercede privacy in data sets by reducing the granularity of a data representation. This deficiency is a trade-off that outcome in some loss of satisfactoriness of [26] data management or mining algorithms in order to advantage some privacy [29]. The t-closeness model refinement the l-diversity model by treating the values of an attribute patently by taking into account the delivery of data values for that attribute.

7.2.5. Differential Privacy

The differential privacy is a technology that endows researchers and database analysts a facility to achieve the useful information from the databases that contain personal information of people without revealing the private identities of the person. This is done by introducing a slightest discomposure in the information provided by the database system. The distraction introduced is huge sufficient so that they protect [30] the privacy and at the identical time small sufficient so that the information provided to the analyst is still useful. Differential Privacy promissory note to provide the solution to this problem as shown Figure 6. In Differential Privacy analyst is not endowed the direct access to the database containing personal information. A negotiator piece of software is initiated between the database and the analyst to protect the privacy. This mediate software is also called as the privacy sentinel.

- **Step 1:** The analyst can make a query to the database through this interstitial privacy sentinel.
- **Step 2:** The privacy Sentinel takes the query from the analyst and assess this query and other preliminary queries for the privacy risk. After evaluation of privacy risk.
- **Step 3:** The privacy sentinel then gets the answer from the database.
- **Step 4:** Add some deformation to it according to the assess privacy risk and in the end provide it to the analyst.

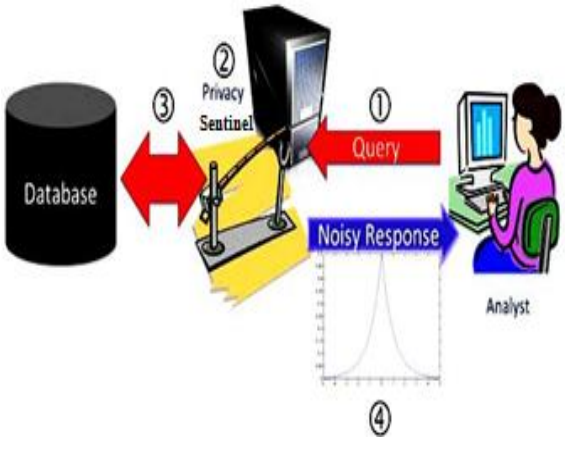

Figure 6. The Big Data Differential Privacy As A Solution To Privacy-Preserving In Big Data

7.2.6. Identity Based Anonymization

These techniques come up against issues when successfully combined anonymization, privacy protection, and big data techniques to analyze usage data while protecting the recognizance of users. The Intel Corporation Human Factors Engineering team wanted to use web page ingress logs and big data tools to enlarge the facility of Intel's heavily used internal web portal [31]. To protect Intel employees' privacy, they were requisite to alienate personally identifying information (PII) from the portal's usage log pantry, but in a way that did not impact the utilization of big data tools to do analysis or the capability to re-identify a log entry in order to inquire into abnormal behavior. There exists a large expansion for ahead of research in privacy preserving methods in big data.

8. Big Data Storage Concepts

Big data storage is a storage infrastructure that is designed in particular to store, manage and repossess huge amounts of data, or big data. The big data storage entitles the storage and sorting of big data in such a way that it can comfortably be accessed, used and processed by applications and services operational on big data. Big data storage is also able to flexibly scale as requisite [32]. Big data storage in the first instance supports storage and input/output operations on storage with a very huge number of data files and objects. A typical big data storage architecture is made up of a unusable and scalable supply of direct concatenate storage pools, clustered or scale-out network concatenate storage or an infrastructure based on object storage format. The storage infrastructure is associated with computing server nodes that enable rapid processing and retrieval of big amount of data. The necessity to store big data datasets, often in various copies, inventive storage strategies and technologies have been created to instate economical and highly scalable storage solutions. The big data storage refers to the storage and management of huge-scale datasets. The following topics are talking about in this paper.

- **Clusters:** A cluster is a decisively coupled collection of servers, or nodes. These servers ordinarily have the same hardware specifications and are connected simultaneously via a network to work as a single unit. The each node in the cluster has its own committed resources, such as memory, a processor, and a hard drive. A cluster can execute a task by segmenting it into miniature pieces and distributing their execution onto different computers that be suited to the cluster.
- **File Systems:** The file system is the technique of storing and organizing data on a storage device, such as hard drives, flash drives, DVDs. A file is an atomic unit of storage applied by the file system to store data. A file system endows a logical view of the data stored on the storage device and confer it as a tree structure of directories and files.
- **NoSQL:** In Not-only SQL (NoSQL) database is a non-relational database that is extremely scalable, fault-tolerant and in particular designed to house semi-structured and unstructured data. The NoSQL database often endows an API-based query interface that can be called from inside an application [33]. NoSQL databases also endorsement query languages other than Structured Query Language (SQL) in view of the fact that SQL was designed to query structured data stored inside a relational database.
- **Sharding:** The sharding is the process of horizontally partitioning a huge dataset into a collection of miniature, more manageable datasets called shards. The shards are distributed across various nodes, where a node is a server or a machine. The every shard is stored in a separate node and each node is accountable for only the data stored on it. The every shard shares the same schema, and all shards collectively represent the complete dataset. Sharding is often transparent to the client, but this is not a necessity. Sharding allows the distribution of processing loads across various nodes to obtain horizontal scalability.
- **Replication:** In the replication stores several copies of a dataset, known as replicas, with several nodes. Replication provides scalability and availability due to the reality that the same data is replicated on different nodes. The imperfection tolerance is also achieved since data redundancy make sure that data is not vanished when an individual node bust up.
- **CAP Theorem:** The Availability, Consistency, and Partition Tolerance (CAP) theorem, also known as Brewer's theorem, indicate a triple restriction related to distributed database systems in Figure 7 shows. It's betting that a distributed database system, running on a cluster. The CAP theorem is at the heart of conversations about different models for data distribution in computer systems. The CAP theorem really is an outcome of a set of deeper reality in distributed systems about shared knowledge.

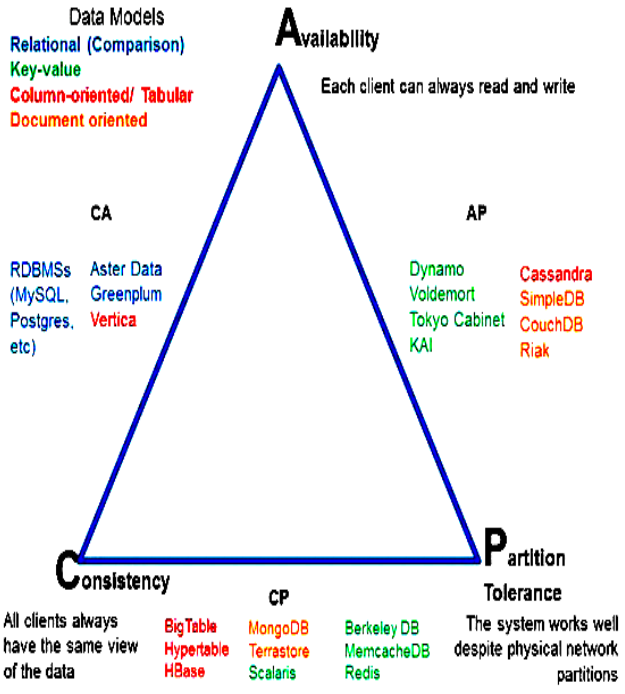

Figure 7. The CAP Theorem

9. Big Data Processing Concepts

The big data processing techniques analyze big data sets at terabyte or even petabyte scale. When taking into account the relationship between a data warehouse and its associated data marts, it becomes apparent that partitioning a huge dataset into a miniature one can speed up processing. The big data datasets stored on distributed file systems or within a distributed database are previously partitioned into miniature datasets. The key to understanding big data processing is the achievement that [34] unlike the centralized processing, which occurs within a conventional relational database, big data is mostly processed in parallel in a distributed trend at the location in which it is stored. Bring forward the discussion of big data processing concepts.

- **Parallel Data Processing:** The parallel data processing involves the at the same time execution of multiple sub-tasks that collectively comprise a huge task. The aim is to detract the execution time by dividing a single huge task into multiple miniature tasks that run concurrently. In spite of the fact that parallel data processing can be achieved through multiple networked machines, it is more typically receiving within the confines of a single machine with several processors or cores.
- **Batch Processing:** The batch processing, also called offline processing, involves processing data in batches and ordinarily imposes delays, which in turn outcome in high-latency responses. Batch workloads typically involve huge quantities of data with sequential read/writes and reckon, in the groups of read or write queries.
- **Flink:** In apache flink is an engine which processes streaming data. Flink can handle large volume data streams while keeping a short processing latency, and its DataStream API has added a huge number of state-of-the-art features, like at accessory out-of-order streams, accessory Event Time, and a very user-friendly, customizable windowing component. If you are processing streaming data in actual time, Flink is the preferable preference.
- **Transactional Processing:** The transactional processing is also called online processing. Transactional workload processing follows a method whereby data are processed interactively without delay, resulting in very short-latency responses. Transaction workloads involve small amounts of data with unsystematic reads and writes.
- **Distributed Data Processing:** The distributed data processing is closely linked to parallel data processing in that the similar principle of "divide-and-conquer" is applied. In spite of, distributed data processing is invariably achieved through physically distinct machines that are networked jointly as a cluster.
- **Cluster:** In the same customs that clusters provide essential support to create horizontally scalable storage solutions, clusters also provides the mechanism to empower distributed data processing with straight scalability. So far as clusters are highly scalable, they provide a surpassing environment for big data processing as huge datasets can be divided into miniature datasets and then processed in parallel in a distributed style.
- **Hadoop:** A Hadoop is an open-source framework for volumetric scale data storage and data processing that is consistent with commodity hardware. The Hadoop framework has instituted itself as a de facto industry platform for contemporaneous big data solutions. It can be used as an ETL engine or use an analytics engine for processing huge amounts of structured, semistructured and unstructured data [35]. In the an analysis viewpoint, Hadoop implements the MapReduce processing framework.
- **Spark:** Spark is rapid becoming one more popular system for big data processing. Spark is favorable with Hadoop (helping it to work acute), or it can work as a standalone processing engine. The Hadoop's software works with Spark's processing engine, supersede the MapReduce section. This, in turn, can superiority to a variety of substitute processing scenarios, which may comprise a mixture of algorithms and tools for the two systems.
- **Samza:** The samza also processes distributed flux of data. Samza is constituted on Apache Kafka for messaging and uses YARN in cluster resource management. Samza uses a normal API, and unlike the majority of short -level API messaging systems, it offers a straightforward, callback based, process message. When a computer in the cluster, lets go on, the YARN component transparently transacts the tasks to one more computer. Nowadays LinkedIn uses Samza, stating it is critical for their members have a definite experience with the notifications and emails they receive from LinkedIn. As an alternative each application sending emails to LinkedIn members, all emails is sent via a central Samza email distribution system, modulate and organizing the email requests, and then sending a summarized email, based on windowing criteria and distinguished policies, to the subscriber.

10. Big Data Analysis Techniques

Big data analysis can assist an organization to mix structured, semi-structured and unstructured data at a stretch. Big Data analysis is the new comprehension in IT world. It consist of huge raw data which could be in any form and can be advantageous in an enterprise. Big Data analysis, mixture conventional statistical data analysis approaches with computational ones [36]. The statistical sampling from a population is best possible when the entire dataset is available, and this prerequisite is typical of conventional batch processing scenarios. Nevertheless, big data can transference batch processing to actual time processing due to the requirement to make sense of streaming data. Accompanied by streaming data, the dataset huddle over time, and the data is time-ordered. In the streaming data places an prominence on at the right time processing, for analytic outcome have a shelf-life. An organization will operate its big data analysis engine at two speeds firstly processing streaming data as it arrives and secondly performing batch analysis of this data as [37]

it collect to look for patterns and trends. In this section, I am discussing the following basic types of data analysis.

- **Quantitative Analysis:** The quantitative analysis is a data analysis technique that peculiarity on quantifying the patterns and correlations found in the data. The based on statistical practices, this technique commingle analyzing a huge number of observations from a dataset. Primarily the sample size is huge, the outcome can be applied in a generalized manner to the complete dataset.
- **Data Mining:** Data mining, also known as data quest, is an exclusiveness form of data analysis that goal huge datasets. In relation to big data analysis, data mining, commonly refers to automated, software-based techniques that sift through massive datasets to recognize patterns and trends. In particular, it involves extracting hidden or unfamiliar patterns in the data with the intention of pick out earlier unfamiliar patterns. Data mining forms the foundation of predictive analytics and business intelligence (BI).
- **Qualitative Analysis:** In qualitative analysis is a data analysis technique that attention on delineate various data qualities using words. It entails analyzing a miniature sample in greater profundity compared [36] for quantitative data analysis. These analysis results cannot be generalized to a whole dataset due to the miniature sample size. They cannot be measured numerically or used for numerical equivalence.
- **Statistical Analysis:** Statistical analysis uses statistical technique based on mathematical formulas as a medium for analyzing data. The statistical analysis is most over and over again quantitative, but can also be qualitative. This genre of analysis is normally used to describe datasets via summarization, like providing the mean, median, or mode of statistics allied with the dataset. It can also be used to conclude patterns and relationships within the dataset, like regression and correlation.
- **Correlation:** The correlation is an analysis technique used to take the plunge whether two variables are associated with each other. If they are found to be associated, the next step is to decide what their relationship is. The use of correlation assistance to develop an understanding of a dataset and discovery, relationships that can assist in explaining a phenomenon. Correlation is therefore normally used for data mining where the identification of link between variables in a dataset leads to the find of patterns and anomalies.
- **Visual Analysis:** In visual analysis is a form of data analysis that involves the graphic representation of data to make able or enlarge its visual perception. In the based on the premise that humans can comprehend and draw conclusions from graphics more speedily than from text, visual analysis acts as a find tool in the field of big data. The goal is to use graphic representations to develop a deeper [38] comprehend of the data being analyzed. Extraordinarily, it assistance, identify and highlight hidden patterns, anomalies, correlations. Visual analysis is also straight related to exploratory data analysis as it embolden the formulation of questions from dissimilar angles.
- **Machine Learning:** The Humans are nice at spotting patterns and relationships within the data. Regrettably, we can not process huge amounts of data very fast. Machines, on the contrary, are very adept at processing huge amounts of data fast, but only if they know how. If human intellect can be combined with the processing speed of machines, machines will be able to process huge amounts of data without requiring much human interference.
- **Sentiment Analysis:** Eventually the sentiment analysis is a specialized form of text analysis that concentrate on determining the partiality or emotions of individuals. This form of analysis decides the viewpoint of the author of the text by analyzing the text within the context of the natural language. Sentiment analysis not only endows information about how individuals believe, but also the intensity of their affection. All this information can then be amalgamated into the decision-making process. The sentiment analysis incorporates identifying client satisfaction or dissatisfaction early, gauging product success or lack of success, and disclosure new trends.

11. Big Data Techniques, Technologies and Tools

The extensive diversification of techniques and technologies has been developed and adapted to aggregate, analyze, manipulate, and visualize big data. These techniques and technologies draw from various fields, together with computer science, applied mathematics, statistics, and economics. This means that an organization that motive to derive value from big data has to outshine a flexible, multidisciplinary viewpoint.

11.1. Big Data Techniques

There are numerous techniques that draw on disciplines such as computer science and statistics that can be used to analyze datasets. In this paper, we confer a list of some categories of techniques suited across a range of industries. Again, I note that not all of these techniques rigidly require the use of big data some of them can be applied influential to smaller datasets in spite of all of the techniques we list here can be well-becoming to big data and huge and more diverse datasets can be used to genesis a lot of and insightful outcome than smaller, less diverse ones.

- **A/B Testing:** This technique in which a control group is compared with a diversity of test groups in order to determine what treatments will ameliorate a given goal variable, e.g., marketing reaction rate. This technique is also known as echeloned testing or bucket testing. Big data enable large numbers of tests to be executed and analyzed, make sure that groups are of sufficient size to detect significant dissimilarity between the rein and treatment conglomeration. While more than one variable at the same time manipulated in the treatment, the

multivariate popularization of this technique, which enforce statistical modeling, is usually called A/B testing.

- **Classification:** This technique to recognize the categories in which recent data points be suited to, based on a training set containing data points that have previously been classified. In an application is the prediction of segment-specific customer behavior (e.g., buying decisions) where there is an apparent hypothesis or purpose result.
- **Association Rule Learning:** A set of techniques for exploring the interesting association, i.e., "association rules," among variables in [39] huge databases. These techniques consist of a variety of algorithms to originate and test presumable law.
- **Network Analysis:** These techniques used to mark out relationships among discrete nodes in a graph or a network. However, the social network analysis, connections between personal in a community or organization are analyzed, in other words, how information travels, or who has the most efficacy over whom.
- **Crowd sourcing:** A technique for collecting data submitted by a huge group of people or the crowd through [40] an open call, generally through networked media that is to tell the Web. This is a type of mass collaboration and an illustration of using Web 2.0.
- **Data Fusion And Data Integration:** This technique that brings together and analyze data from various sources in order to develop insights in ways that are more potent and believable, more accurate than if they were developed by analyzing alone source of data.
- **Predictive Modeling:** A techniques in which a mathematical model is evolved or select for optimal predict the probability of the result. For example an application in customer relationship management is the use of predictive models to evaluate the probability that a customer will brainstorm or the probability that a customer can be cross-sold an additional item.
- **Spatial Analysis:** These techniques, few applied from statistics, which analyze the geometric, topological, or geographic properties paraphrase in a data set. Generally the data for spatial analysis come from geographic information systems that hold data, including location information, addresses or latitude/longitude coordinates.
- **Time Series Analysis:** In this technique from both statistics and signal processing for analyzing a series of data points, representing values at successive times, to quotation significant characteristics of the data. In time series forecasting is the use of a model to predict subsequent values of a time series based on known bygone values of the alike or additional series.
- **Statistics:** This technique science of the collection, interpretation of data, organization, and including the design of surveys and experiments. In Statistical techniques are generally used to make judgments about what relationships between variables could have happened by chance and what relationships between variables likely outcome of underlying causal relationship. Statistical techniques are also used to make little the likelihood of Type I inaccuracy (erroneous positives) and Type II inaccuracy (erroneous negatives).

11.2. Big Data Technologies

There are an increasing number of technologies used to, manipulate, manage, aggregate and analyze big data. We have detailed some of the most renowned technologies, But this list [41] is not comprehensive, particularly as more technologies continue to be developed to support big data techniques, some of which we have listed.

- **Cassandra:** This is a database management system designed to control large quantities of data on a distributed system. This system was primordially developed at Facebook and is now managed as a project of the Apache Software foundation.
- **Business Intelligence (BI):** A kind of application software designed to report, analyze, and present data. This tool is over and over again used to read data that have been already stored in a data mart or data warehouse. BI tools can also be used to generate excellence reports that are generated on at fixed interval basis, or to show information on real-time management dashboards.
- **Cloud Computing:** A computing instance in which extremely scalable computing resources, usually configured as a distributed system, are provided as a service through a network. The cloud computing, often designated to as simply "the cloud," is the distribution of on-demand computing resources.
- **Stream Processing:** This system designed for action, huge real-time streams of incident data. Stream processing empowers applications such as algorithmic trading in fraud detection, process monitoring, financial services, RFID incident processing applications and location-based services in telecommunications.
- **Distributed System:** In this technique several computers communicating through a network, used to extricate a normal computational issue. The issue is divided into several tasks, each of which is extricate by one or more computers working in parallel. The avail of distributed systems includes a surpassing performance at a scanty cost, surpassing reliability and to a greater extent scalability.
- **Data Warehouse:** This is a uniqueness database optimized for reporting, frequently used for storing huge amounts of structured data. The data are transposed using ETL tools from operative data stores, and reports are over and over again generated using business intelligence tools.
- **Extract, Transform, And Load (ETL):** This software tool used for quotation data from external sources, transform them to suitable operational needs, and load them into a database or data warehouse.
- **R:** This is programming language and software atmosphere for statistical computing and graphics. The R language has happen a virtually standard among statisticians for developing statistical

software and is extensively used for statistical software development and data analysis. The R is part of the GNU Project, a collaboration that endorsement open source projects.

- **Visualization:** This Technologies used for fabricating images, diagrams, or animations to communicate a message that are over and over again used to synthesize the outcome of big data analyses.
- **HBase:** It is also an open source, distributed, non-relational database modeled like on Google's Big Table. It was fundamentally developed by Powerset and is presently managed as a project of the Apache Software foundation as part of the Hadoop.
- **MapReduce:** A software framework proposed by Google for processing volumetric datasets on [42] certain kinds of difficulty in a distributed system and also implemented in Hadoop.
- **Mash-up:** This application that uses and modulates data presentation or functionality from two or additional sources to create recent services. These applications are mostly made available on the Web, and frequently use data retrieve through open application programming interfaces or from open data sources. Mashups are over and over again defined by the type of content that they overall. The mashup, for example, could overlay traffic data from one source on the Internet upon maps from Yahoo, Microsoft, Google or any content provider.

11.3. Big Data Tools

Big data are huge data sets, which are arduous to capture, create, manage and process with the conventional database models within a satisfactory time. The big data tools provide the capacity to analyze a miscellaneous of Information, analyze Information in motion [43] on ad hoc basis, analyze extreme volumes cost effectively, Now we are discussing some big data tools.

- **RapidMiner:** RapidMiner gives businesses a centralized solution that features a greatly powerful and robust graphical user interface that enables users to deliver, create, and maintain predictive analytics. RapidMiner not only assistance you understand and find value in your data, but enables you to create models and plans so that you can extract critical statistics and information on which you will base your decisions and action plan.
- **MarkLogic:** MarkLogic is built to deal with bulky data loads and allow users to access it through real-time updates and alerts. It confer geographical data that is combined with content and location relevance along with data filtering tools. This tool is ideal for those looking at paid content discovery app development.
- **Tableau:** The Tableau is a data visualization tool whose mainly focus is on business intelligence. With Tableau, you have the ability to create scatter plots, bar charts, maps, and more in the absence of programming.
- **Talend:** This tool open studio is that it is open source, which means that improvements will keep on rolling out as the community tweaks the tool. This tools also includes products for developing, testing and deploying data management and application integration products.
- **Splunk:** This tool specialize typically in harnessing machine data created from a number of various sources, like websites, sensors and applications. The company also enables developers to write code using any technology platform, language or framework.
- **Hive:** This software tool provide make easy managing and querying large datasets residing in the distributed storage. Apache Hive provides a mechanism that helps project structure into this data and then query it using HiveQL an SQLlike language.

12. Conclusion

In recent years, the term big data have emerged to describe a new instance for data applications. The technologies being introduced to support this instance, have an extensive variety of interfaces, making it arduous to construct tools and applications that harmonize data from several big data sources. In this survey paper, we review the background and state-of-the-art of big data. We first introduce the general background of what Is big data, big data concepts and terminology and review related different analytics types, such as descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. The next section shows the key characteristics of big data and categories of big data used in world wide. In this paper, we have investigated the security, privacy in big data. This paper also presents recent techniques of privacy preserving in big data like the De-identification, k-anonymity, T-closeness, and L-diversitym, and Differential privacy, Anonymization etc. Again then focuses on the data storage, data processing and data analysis. We finally examine the big data techniques, technologies and tools. The big data also mean big systems, big opportunity, big challenges and big profits, so more research work in these sub-fields are necessary to extricate it. I am luckily is witnessing necessary the birth and development of big data. The capital investments, human resources and inventive ideas are basic components of development of big data.

References

[1] Gandomi, A., & Haider, M. Beyond, "*The hype: big data concepts, methods, and analytics*," International Journal of Information Management, 35(2), 137-144, (2015).

[2] Sagiroglu, S.; Sinanc, D., "*Big Data: A Review*", 20-24 May 2013.

[3] S. Kaisler, F. Armour, J.A. Espinosa, and W. Money, "*Big Data: issues and challenges moving forward*," in: Proceedings of the 46th IEEE Annual Hawaii international Conference on System Sciences (HICC 2013), Grand Wailea, Maui, Hawaii, January 2013, pp. 995-1004.

[4] Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G., "*A Big Data implementation based on Grid Computing*", Grid Computing, 17-19 Jan. 2013.

[5] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., "*Shared disk big data analytics with Apache Hadoop*", 18-22 Dec., 2012.

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

[7] Mayer-Sch¨onberger V, Cukier K, "*Big data: a revolution that will transform how we live, work, and think*" Eamon Dolan/Houghton Mifflin Harcourt, 2013.

[8] O. R. Team, "*Big data now: current perspectives from OReilly Radar*", OReilly Media, 2011.

[9] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) "*Big data: the next frontier for innovation, competition, and productivity"*, McKinsey Global Institute.

[10] Yuki Noguchi, "*Following Digital Breadcrumbs To 'Big Data' Gold"*, November 29, 2011.

[11] Sagiroglu, S.; Sinanc, D., (20-24 May 2013), "*Big Data: A Review"*.

[12] Mayer-Schonberger V, Cukier K. "*Big data: a revolution that will transform how we live, work, and think",* Boston: Houghton Mifflin Harcourt; 2013.

[13] K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012.

[14] Fisher D, DeLine R, Czerwinski M, Drucker S. "*Interactions with big data analytics",* Interactions. 2012; 19(3):50-9.

[15] Russom P., "*Big data analytics*" TDWI: Tech. Rep ; 2011.

[16] J. Fan, F. Han, H. Liu, "*Challenges of big data analysis*", National Science Review, 1 (2) (2014), pp. 293-314.

[17] Nyce, Charles (2007), "*Predictive Analytics White Paper (PDF)",* American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America.

[18] A. Labrinidis, H.V. Jagadish, "*Challenges and opportunities with big data",* Proceedings of the VLDB Endowment, 5 (12) (2012), pp. 2032-2033.

[19] Laney D. "*3D data management: controlling data volume, velocity, and variety*", META Group, Tech. Rep. 2001.

[20] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "*Big Data Processing in Cloud Computing Environments*", International Symposium on Pervasive Systems, Algorithms and Networks, 2012.

[21] Russom, P.. "*Big Data Analytics*" In: TDWI Best Practices Report, pp. 1-40 (2011).

[22] Cloud Security Alliance Big Data Working Group, "*Expanded Top Ten Big Data Security and Privacy Challenges*", 2013.

[23] A.A. Cardenas, P.K. Manadhata, S.P. Rajan, "*Big Data Analytics for Security",* IEEE Security & Privacy, vol. 11, issue 6, pp. 74-76, 2013.

[24] T. Omer, P. Jules, "*Big Data for All: Privacy and User Control in the Age of Analytics*", Northwestern Journal of Technology and Intellectual Property, article 1, vol. 11, issue 5, 2013.

[25] De Cristofaro, E., Soriente, C., Tsudik, G., & Williams, A. "*Hummingbird: Privacy at the time of twitter. In Security and Privacy (SP)",* 2012 IEEE Symposium on (pp. 285-299), 2012.

[26] Li N, et al. "*t-Closeness: privacy beyond k-anonymity and L-diversity*", In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

[27] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam *"M. L-diversity: privacy beyond k-anonymity"* In: Proc. 22nd international conference data engineering (ICDE); 2006. p. 24.

[28] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. "*Protection of big data privacy"* In: IEEE translations and content mining are permitted for academic research. 2016.

[29] Sweeney L. "*K-anonymity: a model for protecting privacy",* Int J Uncertain Fuzz. 2002; 10(5): 557-70.

[30] Xu L, Jiang C, Wang J, Yuan J, Ren Y. "*Information security in big data: privacy and data mining*", IEEE Access. 2014; 2: 1149-76.

[31] Sedayao J, Bhardwaj R. *"Making big data, privacy, and anonymization work together in the enterprise experiences and issues*", Big Data Congress; 2014.

[32] Zhang, Xiaoxue Xu, Feng, (2-4 Sep. 2013), "*Survey of Research on Big Data Storage*".

[33] Zhang, Y., Feng, H., Hao, W., et al.: "*Research on the storage of file big data based on NoSQL*", Manufact. Autom. 6, 27-30 (2014).

[34] Ji, C., Li, Y., Qiu, W., Awada, U., and Li, K. (2012) "*Big data processing in cloud computing environments. Pervasive Systems"*, Algorithms and Networks (ISPAN), 2012 12th International Symposium on, pp. 17-23, IEEE.

[35] Bu, Y., Howe, B., Balazinska, M., and Ernst, M. (2010). "*Haloop: Efficient iterative data processing on large clusters*", Proceedings of the VLDB Endowment, 3, 285-296.

[36] Hu, H., et. al. (2014). "*Toward scalable systems for Big Data analytics: A technology tutorial*" Access IEEE, 2, 652-687.

[37] Purcell, B. *"The emergence of big data technology and analytics"* Journal of Technology Research, Holy Family University, pp. 1-6.

[38] Yang, X., & Sun, J. (2011). *"An analytical performance model of MapReduce"* In IEEE Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 306-310.

[39] R. Agrawal, T. Imielinski, and A. Swami, "*Mining association rules between sets of items in large databases,"* SIGMOD Conference 1993: 207-16; P. Hajek, I. Havel, and M. Chytil, "The GUHA method of automatic hypotheses determination," Computing 1(4), 1966; 293-308.

[40] Jeff Howe, "*The Rise of Crowdsourcing,"* Wired, Issue 14.06, June 2006.

[41] C.L. Philip Chen, Chun-Yang Zhang, *"Data intensive applications, challenges, techniques and technologies: A survey on Big Data",* Information Sciences, www.elsevier.com/locate/ins, January 2014.

[42] Jeffrey Dean and Sanjay Ghemawat, "*MapReduce: Simplified data processing on large clusters,"* Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December 2004.

[43] Almeida, Fernando; Santos, Mário. 2014. *"A Conceptual Framework for Big Data Analysis. In Organizational, Legal, and Technological Dimensions of Information System Administration",* ed. Irene Maria Portela, Fernando Almeida, 199-223. ISBN: 9781466645264. USA: IGI Global.