A SURVEY ON BIG DATA ANALYSIS, APPROACHES AND ITS APPLICATION IN THE REAL WORLD

[1]Ashika A, [2]Dr. Mohan Kumar S
[1]Student, [2]Associate professor
[1]Department of Information Science and Engineering,
[1]New Horizon College of Engineering, India

Abstract : Big Data is data of high volume and high assortment being delivered or produced at high speed which can't be stored, managed, processed or analyzed utilizing the current conventional tools, strategies and structures. These days, a large portion of information stored in organizations are unstructured models. According to the IDC estimation 90% of data is unstructured data which is growing faster than all types of data. Over 80% of all potentially helpful business information is unstructured information which is in the form of sensor readings, comfort logs et cetera. In this paper we discuss fully structured, semi-structured and unstructured data. Moreover, we also explain the tools/techniques available for them.

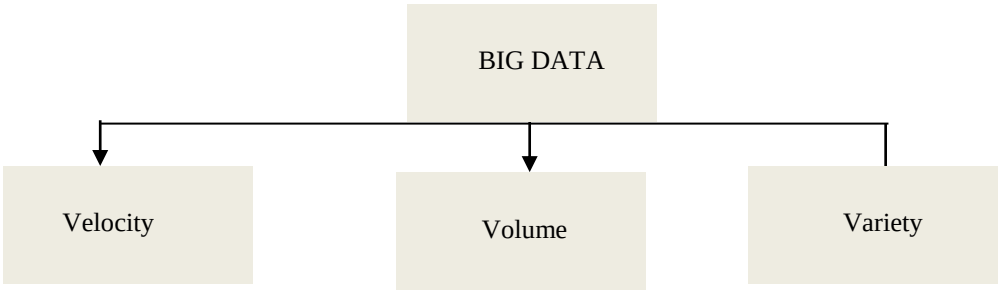
***IndexTerms* - Big Data, Structured data, Unstructured data, Semi-structured data, Text Analytics, Social Media Analytics.**

I. INTRODUCTION

Regularly data is produced, gathered in tremendous sum. However, many-a-times it remains unutilized without drawing helpful information. These experiences are crucial in vital and operational basic leadership processes like-promoting, client engagement, branding, and so on. Information created by different channels like promoting, appropriation, client engagement, social channels and web substance is in various organized structured and unstructured in numerous frameworks.

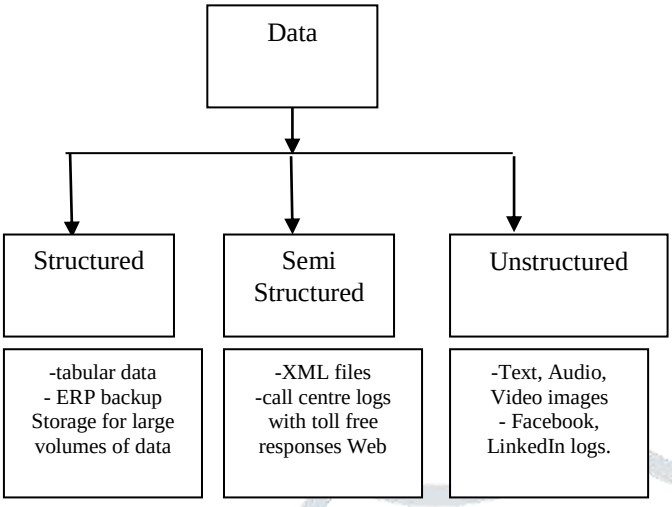
"Big data technologies describe a new generation of technologies and architecture designed to economically extract value from very large volumes of a wide variety of data, enabling high velocity capture, discovery and/or analysis" as defined by IDC. TechAmerica Foundation defines big data as "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information." [1].

The description of big data is incomplete without mentioning the three V's that are its characteristics: volume, variety and velocity.
(i)*Volume*: The massive scale and development of unstructured information outpace customary capacity and diagnostic arrangements. It refers to the high magnitude of big data which is in the order of terabytes to petabytes and more. For instance, some earlier estimates suggested that 20 petabytes of storage space were used to store 260 billion Facebook photos. In 2010, it was reported that up to one million photographs were processed by Facebook per second. Twitter generates 12 terabytes of data daily. In 2012, Facebook stated that 2.7 billion "likes" and "comments" were registered daily by the users.
(ii)*Variety*: Huge amounts of data generated from many sources such as e-mail, social media, videos, images, blogs, sensor has heterogeneous nature. Data can be structured, semi-structured and unstructured data. Structured data has a fixed format whereas unstructured data has no fixed schema or form.
(iii)*Velocity*: Big data is being generated continuously at an exponential rate. The high rate of development of big data presents enormous opportunity and will yield huge monetary additions if effectively used. Social media generates data explosively [1, 5].



II. STRUCTURED, UNSTRUCTURED AND SEMI-STRUCTURED DATA

Different types of big data are:



1) Structured data

Data that can be easily organized is structured data. It is usually stored in relational database system (RDBMS) and can be analyzed faster. Structured data has a predefined schema. In such a schema, data conforms to its specification. Due to its fixed nature, efficient search is possible on web for focused content by search engines[3,4].
Some examples of structured data are:

1) Machine Generated:
   Sensory Data - manufacturing sensors, medical devices and GPS data
   Point-of-Sale Data - Credit card information, product information, etc.
   Call Detail Records - Caller and recipient information, time for call.
   Web Server Logs - Page requests, server activities
2) Human Generated:
   Input Data – The data that is given as an input to a computer: age, gender, code, etc.

II. Semi-structured data

Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure[6].

III. Unstructured data

Unstructured data has no identifiable internal structure. Data also has no predefined model. Social media generates huge amount of unstructured data. Only 5% of data is structured, remaining is unstructured. Most of the organization's knowledge is unstructured and it is very important to utilize this massive amount of information in a useful manner. Different standards for unstructured data are open XML, SMTP, SMS, CSV and Information and content exchange. They can be broadly classified into textual and non-textual data[3].
Some of the examples are:

(i) Textual: documents, presentations, spread sheets, scanned images, etc.
(ii) Imagery: multimedia files, streaming video, etc.
(iii) HUMINT: reports, audio files, and gestures
(iv) Sensors: seismic, acoustic, magnetic, sonar, etc.
(v) Environmental: weather

	Unstructured	Fully Structured	Semi Structured
Technology	Character and binary data	Relational database tables	XML / RDF
Transaction Management	No transaction management, no concurrency	Matured transaction management, various concurrency techniques	Transaction management adapted from RDBMS, not matured
Version management	Versioned as whole	Versioning over tuples, rows, tables etc.,	Not very common versioning over triples or graphs is possible.
Flexibility	Very flexible, absence of schema	Schema-dependent, rigorous schema	Flexible, tolerant schema
Scalability	Very scalable	Scaling DB Schema is difficult	Schema scaling is simple
Robustness	-	Very robust, enhancements since 30 years	New technology not widely spread
Query- performance	Only textual queries possible	Structured query allows complex joins	Queries over anonymous nodes are possible.

The above table shows the comparison between unstructured, structured and semi-structured [2].

III. TOOLS/TECHNIQUES FOR HANDLING UNSTRUCTURED DATA

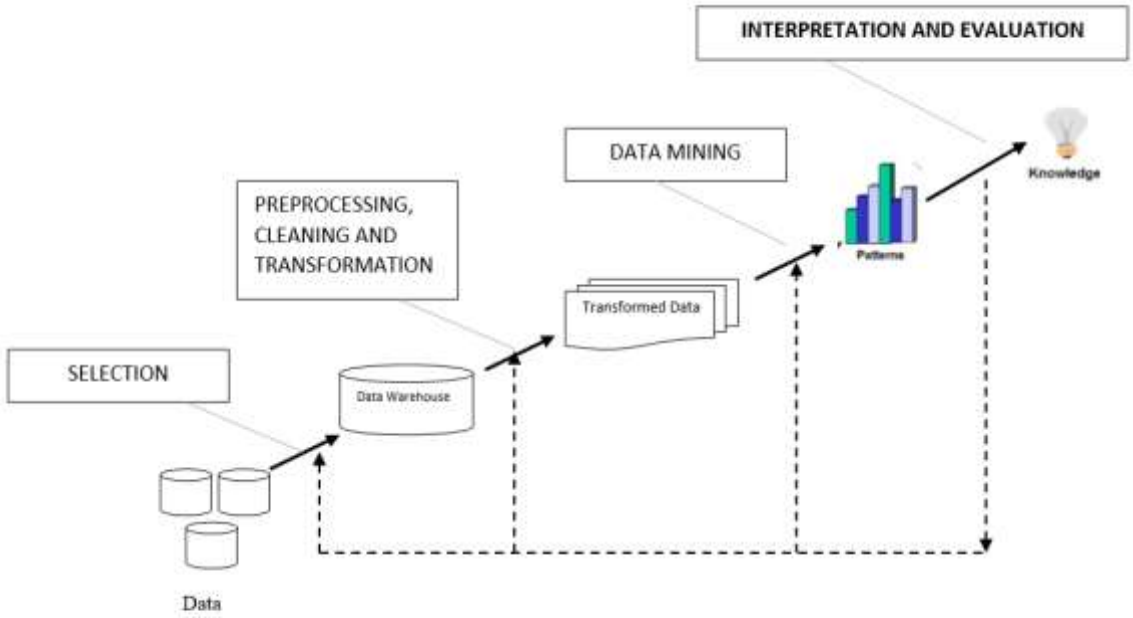
The different **techniques** used to search analyze and deliver unstructured data are
a) Data Mining
b) Text Analytics
c) Social Media Analytics

a) Data Mining

It is a process of discovering interesting knowledge, such as patterns, associations, changes, anomalies from large amounts of data stored in database, data warehouse or other repositories. Data mining is an analysis step of KDD (Knowledge Discovery in Databases)

Steps involved in Knowledge Discovery in database are:

1) Data Cleaning – inconsistent, irrelevant data and noise is removed in this step
2) Data Integration – multiple, heterogeneous data sources are combined during data integration
3) Data Selection – relevant data to the analyze are retrieved from the database.
4) Data Transformation –Here, data is transformed into forms which are appropriate or valid for data mining by performing various summary operations.
5) Data Mining –Many methods/tasks such as class description, association, classification, prediction, clustering is applied to extract patterns
6) Pattern Evaluation − patterns obtained from previous steps are evaluated.
7) Knowledge Presentation – knowledge is represented in required form in this step [4,9].



The above diagram indicates how the Knowledge is stored in database.

Text Analytics

It is a method for extracting useful data or experiences from content based information for example, messages, archives, ads, gatherings, web journals, news content, interpersonal organization content, site content, call focus logs, client remarks and surveys, tweets and so on. It includes Statistical Analysis, Computational Linguistics and Machine Learning.

IE (Information Extraction) converts unstructured content into structured data. There are two tasks in IE: Entity Recognition: Finds useful information in text and classifies. Relation Extraction: Finds the semantic relationship between entities.

Text Summarization is another technique in analytics that produces a summary of data from multiple data sources. There are two approaches: Extractive: In this type of approach, original text from multiple sources are extracted and summarized as subset of data. Abstractive: Uses natural language processing(NLP) to parse the original text and produce summary.

Sentiment Analysis is also one of text analytics which analyses the opinion of people. It is very useful for deciding the marketing strategy of a certain product. It can be done at document, sentence and aspect level. Document level can be effective to obtain the positive or negative sentiment has been portrayed in that document. Sentence level can be used to determine the entity. Aspect based technique is used to determine all sentiments specific to a particular entity in the document[1].

b) Social Media Analytics

It is the most important and booming type of analytics in the real world. In this type of analytics there are lot of privacy concerns involved. There can also be misuse of public data as in the case of Cambridge Analytica obtaining Facebook data to mould the shape of politics in nations such as US, UK and Mexico.

Social Media analytics can be categorized into two parts: Content-based analytics and Structure-based analytics. Content based analytics is performed on the content posted by user on social media. We can analyse personal traits and interests from it. The data here is highly noisy, unstructured and dynamic in nature.

Social Influence Analysis analyses the influence of connections in a social network. It is based on the assumption that the behaviour of an entity is influenced by others. Such data gives an insight into the actor's influence, strength of connections and the patterns of influence in the network. It can be helpful in deciding marketing strategy of product [1].

The **tools** for unstructured data are

a) Apache Hadoop
b) HBase
c) Apache Splunk
d) Apache Flume
e) Apache Pig

f) MongoDB
g) Hadoop MapReduce

a) Apache Hadoop

Apache Hadoop software library is one of the best new approach to unstructured data analytics. Hadoop is an open-source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms and an application layer that manages distributed processing, parallel computation, workflow and configuration management. In addition to offering high availability, Hadoop is cost efficient for handling large unstructured data sets at great speed. Hadoop is implemented on MapReduce model for analysis of datasets and uses HDFS as storage file system. Its provides portability across different platforms-hardware or software. Here computation is divided into map function and reduce function. The map function obtains key-value pair and generates an intermediate key-value pair. This intermediate key value pair is passed to the reduce function and merges all values corresponding to a single key. It is a very popular choice when we need to filter, sort or pre-process large amounts of data. Facebook stores 100PB of data- structured and unstructured using Hadoop.

b) HBase

HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File system), providing BigTable-like capabilities for Hadoop. That is, it provides a fault tolerant way of storing large quantities of sparse data.

It features linear and modular scalability. It provides automatic and configurable shading of tables. It supports automatic failover between Region servers. It has convenient base classes for backing Hadoop MapReduce Jobs with Apache HBase tables. It consists of Java APIs for client access that are easy to use[5].

c) Apache Splunk

It is a real-time and intelligent platform developed for analysis of text-series data, mainly machine data generated from business industries. The Splunk engine is optimized for quickly indexing and persisting unstructured data loaded into the system. The main objective of Splunk is to provided metrics for all possible types of applications. It combines both big data and up-to-the-moment cloud technologies. Since it is mainly used for providing metrics, the output generated is usually in the form of graphs and reports. The Splunk Add-on for Apache Web Server allows a Splunk software administrator to collect and analyze data from Apache Web Server using file monitoring. After the Splunk platform indexes the events, you can analyze the data using the prebuilt panels included with the add-on[10, 18, 21].

d) Apache Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It is mainly used to aggregate and transfer large amounts of data. Apache Flume agents consists of source, channel and sink. Sources listen and consume events. They usually are files, system logs. Sink refers to HDFS and HBase mostly. Sink can remove events from a channel and transmit them to next agent in flow[22].

e) Apache Pig

It is a framework, initially developed by Yahoo Research that uses high level scripting language and consists of a run-time platform that can be used to execute MapReduce on Hadoop. It has its own data type map that represents semi-structured data such as data in XML and JSON format. It provides scripting language for describing operations such as reading, writing, filtering, joining data. The languages for Apache Pig is known as Pig Latin. The key features are ease of programming, optimization opportunities and extensibility [22].

f) MongoDB

It is one of the open-source tool used for analysis of unstructured data. It is NoSQL database that uses master-slave mechanism. It is written in C++. Master performs read and write operations on the database whereas slave copies data from the master to read and backup the data. Slave is not involved during write operations. MongoDB has a binary format similar to JSON and uses dynamic schemas instead of relational databases. MongoDB has drivers installed for supporting all types of programming languages. The main features of MongoDB are indexing, replication and load balancing. MongoDB supports capped collections ie., collections that have fixed size. It behaves as a circular queue once the fixed size has been reached[21].

g) Hadoop MapReduce

It performs 2 operations:

1) Map- reading the data from the database and putting it into an appropriate format for analysis.
2) Reduce- perform certain operations to obtain the reduced data set

MapReduce is a distributed data processing algorithm introduced by Google. It is used to process huge amount of data in parallel, reliable and efficient way in cluster environments.
MapReduce refers to two separate tasks: 1) Map program or the mapper takes set of data and coverts into another set of data where they are broken down into key-value pairs. 2) Reduce program inputs the output from Map program and combines that data into smaller set of key-value pairs. Therefore, reduce job is always performed after Map program has completed its execution [18].
The tools and technologies used in big data analysis are the above ones [13].

IV. BIG DATA ANALYSIS REQUIREMENTS

There are three significant requirements for data analysis:

1. Minimize data movement
2. Use existing skills
3. Provide security to data

It saves lot of time and space when the processing of data occurs in the place where it is stored rather moving the data. Movement of large volumes data for processing reduces efficiency as well as degrades the performance of the process.

It is of less use if new skills have to be acquired. It increases the expenses of organization for training, hiring and explaining new tools that are used to investigate the issue. Most of the organizations have associate who can easily work on SQL than on MapReduce, it is important for it to be able to work on both the programming languages.

Data warehouse not only has dimensions of arrays for storing data but also set of policies for providing security to data. It is the most essential component of corporate applications.

V. CHALLENGES IN UNSTRUCTURED DATA MINING

- Usability: organizations should think of an approach to find, locate, extract, organize, and store the Data for using unstructured data.
- Volume: It is one of the major issue that affects business. The volume of unstructured data is growing at a rate of approximately 60% per year which shows that it is problematic for organization to storage and use data efficiently.
- Relevance: One way by which relevance becomes possibly the most important factor is absence of understanding into the past story of specific bits of information.
- Heterogeneity: The challenges of analysis are caused due to its extensive scale and the presence of mixed data based on different patterns in the collected and stored data. Data therefore has several patterns and rules that vary greatly.
- Incompleteness: It is the major issue during analysis and must be managed.
- Privacy: The current technologies are static data-oriented that are used in maintaining data security but big data analytics deals on data that is dynamic and keep changing regularly. Protecting individual personal information is a challenging task.
- Requirements for real-time data analysis has been predominant like for weather predictions, ex-stock trading, time series, etc. [4].

VI. REAL WORLD BIG DATA APPLICATIONS AND RESEARCH

1) Social Media Analytics
2) Government and transportation
3) Fraud detection
4) Business (Marketing, Stock Market)
5) Defence

Social Media Analysis:

Due to increasing use of social media such as Facebook, Twitter, large amount of data has been produced every day. There are two types of social media analytics: Content based analytics, Structure-based analytics. Content-based analytics is performed mainly on the content that is posted by the user on the social media sites. Such content is unstructured and is of large size. To perform analytics on such high volume of data we use big data technologies. Structure based analytics can be used to understand the relationship between the entities. Social Media analytics can be used to analyze the mentality of group of people or reaction of a group to a certain social, political or economic issue. Also based on the issue, the solution can be obtained. Social media is very powerful as there are large number of people using it and they can be helpful in manipulating the mindset of many people on certain product, subject or even about a party in politics.

Government and transportation:

Big data plays a major in government sector. It was very helpful in Barack Obama's reelection campaign in 2012. BJP in India also uses Big data analytics to understand the response of the Indian citizens to certain schemes, policies and development changes. Big data analytics performed on social media can help politicians understand the effect of the government policies in an easy manner.

Transportation is also one of the major area where Big data plays an important role. It can be helpful to perform real time analytics, such as: Estimating the time to reach a destination, Estimating the changes in traffic on the route selected, finding the shortest route possible, changing the route based on landslides, quick suggestions and also providing landmarks for better understanding.

Fraud Detection

It is one of the critical fields in which big data analytics play a role. Fraud detection in bank transaction can be very crucial. It mainly detects anomalies from the transactions. Usually fraud detection is performed after the damage has been done. It is useful to understand the damage done, what was the path used for the damage to be done, identifying the network holes and using the obtained information to fix and minimize the harm caused. It can also be used to prevent such frauds from occurring again. Supervised, unsupervised and social network learning can be used for fraud detection.

Business

Business environment changes rapidly these days and it must be in such a way that, it reacts quickly and be dynamic to adopt changes. Data mining and big data analytics can be used to predict the changes that may occur so that the company is prepared to face and overcome the issues that may occur. It also helps in taking important decisions that will affect the company.

It can be useful in customer relationship management also. The cost of managing and retaining the customers is lesser than replacing the customer. We can use big data analytics to identify the customer who is about to leave the services provided, and therefore we can reduce the risk of losing a customer. We can use big data analytics to identify the recurring problem. An organization with large amount of data can use big data analytics to optimize the performance.
It can be useful for designing promotions for the organization. It can also use demographic data to predict the reaction of individuals, group of individuals. It can be useful to customize the services accordingly.

In stock market, it is important to analyze data in real time of both buyers and sellers. They can be used to predict the trends in prices, help in predicting the prices for buying and selling shares, detecting illegal activities.

Defense

Data analytics can be helpful in winning wars. It can be used to understand, predict the enemies' strength and react accordingly. It can provide almost accurate information and assist in taking right decision. In war the data generated is of high volume and is very dynamic in nature. The information can be collected to analyze the region where maximum damage has occurred to revive that region. It can be used to analyze which equipment's need to be replaced, upgraded or repaired.

Therefore these are some of the fields in which big data plays major role and shapes our future[19, 20, 21].

BIGDATA AND IoT IN HEALTHCARE

IoT analytics in health care is nothing but collection of facts, that is, information from the sensors, storing them in data warehouse and processing them to find results. The obtained results can be useful for diagnosis of a patient health [15].

The most recent proposed concept is that, in labs there are many devices that interact and send data through Wi-Fi. Instead the data between devices within small range can send by using heat waves that are generated by the device. Heat waves can be used as mode of communication between the devices. For two systems to engage in such data transfer, initially there should be a handshake between those two systems, this can be termed as "thermal pings".

Thermal pings work on the principal of increasing the existing temperature readings value by +1C for binary value 1 else it's considered to be Binary value 0.
Further on implementation stage, once all the air-gapped systems are communicating with eachother, there are huge volumes of binary bits transferred between them.
If there are any issues caused or any unread bits (which may lead to wrong data) to do any kind of analysis and backtrack study on the missing bits BIG DATA would help us do the basic analysis and have a fixed way of storing all the data[2].

VII. CONCLUSION

In the current world, handling unstructured data is of utmost important. Analysis of unstructured data can shape our future. One small analysis and its effects can make a huge difference. It affects business, politics, environmental concerns, societal rules. In this paper, we have discussed in detail about unstructured data, approaches to handle unstructured data and challenges faced while handling unstructured data.

VIII. REFERENCES

1) Unravelling Unstructured Data: A wealth of Information in Big Data, Mona Tanwar, Reena Duggal, Sunil Kumar Khatri, 2015 IEEE
2) Application of deep learning technique for acutomatic data exchange with air-gapped systems and its security concerns, V. Dhanush, A. R. Mahendra, M V Kumudavalli, DebabrataSamanta, 2017, pg 324-328, IEEE
3) Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis, Rolf Sint1, Sebastian Schaffert, Stephanie Stroka and Roland Ferstl
4) Unstructured Data Analysis- A Survey, K. V. Kanimozhi, Dr. M. Venkatesan, Vol. 4, Issue 3, March 2015, International Journal of Advanced Research in Computer and Communication Engineering
5) Unstructured Data Mining and its applications, JagrutiJangalWagh, JidnyasaDharmikGondane, Ashvini Tulshiram Dukare, INTERNATIONAL JOURNAL OF CURRENT ENGINEERING AND SCIENTIFIC RESEARCH (IJCESR)
6) BIG Data Analytics: A Framework for Unstructured Data Analysis, T.K.Das, P.Mohan Kumar, Vol 5 No 1 Feb-Mar 2013, International Journal of Engineering and Technology (IJET)\
7) https://en.wikipedia.org/wiki/Semi-structured_data
8) J. McKendrick. Survey on unstructured data, produced by Unisphere Research 2011; Available from: http://www.ciosummits.com/media/pdf/solution_spotlight/Marklogic_ 20 11- survey.pdf, 2011.
9) Erik P. Blasch, Stephen Russell, Guna Seetharaman.Joint DataManagement for MOVINTData-to-Decision Making, ISIF 11.14thInternational Conference on Information Fusion; 2011 July 5-8;Chicago, Illinois: USA; 2011. 978-0-9824438-3-5 ©2011 ISIF
10) Han &Kamber& Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann.
11) https://docs.splunk.com/Documentation/AddOns/released/ApacheWebServer/About
12) Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities, Stephanie B. Baker, Wei Xiang, and Ian Atkinson, Volume 5, 2017, IEEE.
13) Big Data Analytics, Jasmine Zakir, Tom Seymour, Kristi Berg, Issues in Information Systems, Volume 16, Issue II, pp. 81-90, 2015
14) Extracting Structured Data from Web Pages, Arvind Arasu, Stanford University.
15) A study on Data Analytics: Internet of Things & Health-Care, N. Nalini, P. Suvithavani, IJCSMC, Volume 6, Issue 3, March 2017, pg 20-27
16) Medical Internet of Things and Big data in HealthCare, Dimiter V Dimitrov, Healthcare informatics Research, July 2016, Volume 22, no. 3
17) Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics, Amir M. Rahmani, BaharFarahani, Future Generation Computer Systems, September 2017
18) A survey on Big Data Analytics: challenges, Open Research Issues and Tools; Debi Prasanna Acharjya, KauserAhmned, International Journal of Advanced Computer Science and Applications, Volume 7, no. 2, Feb 2016
19) The Application of Big Data Analytics in Business World, O. Liu, W. K. Chong, K. L. Man, and C. O. Chan, IMECS 2016, March 2016, Vol II
20) Applications of Real-Time Big Data Analytics, Akinul Islam Jony, International Journal of Computer Applications, Volume 144- no. 5, June 2016
21) Big Data: Tools and Applications, SofiyaMujawar, Soha Kulkarni, International Journal of Computer Applications, Volume 115- no. 23, April 2015
22) Big Data: Survey, Technologies, Opportunities, and Challenges, Nawsher Khan, Ibar Yaqoob, Ibrahim AbakerTargio Hashem, Zakir Inayat, Waleed Kamaledin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz and Abdullah Gani, The Scientific World Journal, Volume 2014, July 2014