Chapter 2

Big Data and its Tools

This chapter introduces evolution of Big Data and its various tools, Big Data analytics in cloud computing, requirement for Big Data analytics, data mining & its process, data mining methods and roles of Big Data in various sectors.

2.1 Evolution of Big Data

In Big Data from digital data to health data are included. It has evolved from various stages i.e. from primitive and structured data to complex relational data. Now a days very complex and unstructured data are also included. Figure 2.1 shows evolution of Big Data over the years. The concept of Big Data came into the light when the growth rate in volume of data was known as information explosion (about 70 years ago). In 1944 Fremont Rider, a librarian estimated that size of American Universities libraries is getting doubled every sixteen years. He estimated that by this growth rate there would be 200,000,000 volumes by 2040. In decade of 90 IBM introduced the relational database concept in which data can be stored in tables and can be analysed easily by using different analysis techniques. By the end of 2003 there was 5 exabytes of data that was created, in just two days this amount of information is created. This data is generated from different sources and includes online transactions, sensor data, social networking, health records, census and science data and live streaming data. This digital data is of 2.72 zettabytes and is predicted to be doubled every two years. Figure 2.1 shows that how significantly data has grown from 2005 onwards [29][30][31].
2.2 Tools Used with Big Data

For the development of data, machine learning is used. The main aim to develop algorithms that permits computers to form a pattern based on observed data. Generally supervised and unsupervised learning algorithms are used with Big Data. To clarify the data the most fundamental algorithm is used and it is known as support vector machine (SVM). However in present time recently parallel support vector machine overcomes the drawback of SVM such as scalability, time, and memory problems. Neural Networks and Artificial Neural Networks ANN are mature techniques that are used in adaptive control, image analysis, and pattern recognition. ANNs are nowadays used with artificial intelligence where control theory, and statistical estimation needs to be used. Social Network Analysis uses Big Data to study the communication, social psychology, and economics. Figure 2.2 shows the Big Data techniques used.
To handle Big Data for getting the better insights and judgments we will always be in need of the efficient tools to manage and process the giant. When we talk about Big Data technology, the first thing that comes in the mind is “Hadoop”.

Big Data produces Big challenge to manage massive amount of structured and unstructured data to handle. Cloud Computing offers scalable solutions to manage such a large amount of data in cloud environment to take advantage of
both technology. To effectively incorporate and manage Big Data in cloud
environment it is important to understand tools and services offered by them.
Some vendors like Amazon Web Service (AWS), Google, Microsoft and IBM
offers Cloud based Hadoop and NoSQL database platforms that are supporting
Big Data applications. In addition to, many cloud providers offer their own
Big Data services as: AWS’s Elastic MapReduce(), Google’s BigQuery etc.
Most of the cloud service provider offers Hadoop framework that scale
automatically on-demand of customers for data processing.

Following are the major technologies that are being used to process the Big Data:

2.2.1 Apache Hadoop and MapReduce()

2.2.1.1 Hadoop

Hadoop provides an open-source software framework for distributed storage
and processing applications on very large datasets, written in java. Hadoop
platform includes higher level declarative languages for writing queries and
data analysis pipelines. Hadoop is used by approximately 63% of organizations
to manage and analyze huge number of unstructured logs and events (Sys. Con
Media, 2011). Hadoop is composed of many components but in Big Data two
mostly components Hadoop Distributed File System (HDFS) and MapReduce() are used. The other components provide complementary services and higher-
level of abstraction.

Data-intensive distributed applications use a sophisticated software platform
called Apache Hadoop. It uses Map/Reduce as its computational paradigm.
The mapper is responsible for mapping the word with a value. This is then
given to the reducer to sum up the values. Figure 2.3 and Figure 2.4 shows
various tools used in Hadoop and a simple architecture of Apache Hadoop.
Figure 2.3: Tools of Apache Hadoop
Source: “Hadoop Ecosystem” [33]

Figure 2.4: Hadoop architecture
2.2.1.2 MapReduce()

MapReduce() system is the main part in Hadoop framework that is used for processing and generating large datasets on a cluster with distributed or parallel algorithm. It is a programming paradigm used to process large volume of data by dividing the work into various independent nodes. A MapReduce() program corresponds to two jobs, A Map() method which include obtaining, filtering & sorting datasets and A Reduce() method which include finding out summaries and generate final result. MapReduce() system arranges distributed servers, manage all communications, parallel data transfers, also provide redundancy and fault tolerance. Figure 2.5 and 2.6 illustrates the working of Map/Reduce process.
2.2.1.3 Hadoop Distributed File System (HDFS)

HDFS is used to store large data files that are too much to store on a single machine typically in gigabyte to terabyte. HDFS is a distributed, scalable and portable file system written in java for Hadoop framework [36]. It maintains reliability by replicating data across multiple hosts to facilitate parallel processing, for that it split a file into blocks that will stored across multiple machines. The cluster of HDFS has master-slave relationship with single name-node and multiple data-node.

As illustrated in Figure 2.7, Hadoop architecture is comprised of following components:

a) Name Node
b) Job Tracker
c) Data Node
d) Task Tracker
e) HDFS

![Hadoop Architecture](source: “Hadoop Demystified” [37])
2.2.1.4 Cassandra and HBase

Both are open-source, non-relational, distributed DBMS written in java, supports structured data storage for large tables and runs on top of HDFS. It is columnar data model with features like compression, in-memory operations and provides fault tolerance way of storing large quantities of sparse data.

2.2.1.5 Hive

It is a warehouse infrastructure by facebook providing data summarization, adhoc querying and analysis. It provides SQL like language (HiveQL) to make powerful queries and get results in real time.

The data after cleaning was supposed to be stored in some order and had to be in some shape(schema). The hive was used for the same. However, the alternative was kept to tackle the hive failure situation like if the data was coming from the multiple sources and complex transformations were to be done e.g. there were two datasets one from sensors and other one was the global survey data.

2.2.1.6 Pig

It is a high-level data flow language (Pig Latin) and execution framework for parallel computation.

Pig has two major components [38]
1. Pig Latin is a high-level data processing language.
2. A compiler that compiles Pig Latin script in a choice of evaluation mechanisms is known as Hadoop. Pig also supports a local mode for development purposes.
2.2.1.7 Zookeeper

It is a high performance coordination service for distributed application that can store configuration information and have master-slave node.

2.2.2 Spark

Apache Spark open sourced tool that is originally developed in AMP Lab at UC Berkley. It provides in-memory analytics which is faster than Hadoop (up to 100 times). It is designed for running iterative algorithms and interactive analytics. It is highly compatible with Hadoop’s Storage APIs. It can run on existing Hadoop Cluster Setup. Developers can write driver programs using multiple programming languages on Spark. The use of Spark is necessary because of machine learning algorithms that are iterative and each iteration can improve the results [39]. Figure 2.8 and 2.9 illustrates the working of Hadoop MapReduce() Vs Spark and Spark process.

![Hadoop MapReduce Vs Spark](image)

**Figure 2.8: Hadoop MapReduce() Vs Spark**

Source: “Hadoop MapReduce() Vs Spark”[40]
2.2.3 Dryad

Dryad [42] is also used to perform parallel and distributed computing from a very small cluster to a large cluster. To use Dryad the programmer uses the computer clusters to run the programs in a distributed manner. Dryad uses directed graphs consisting of communicational channels and vertices for running an application.

2.2.4 Apache Mahout

Apache Mahout [43] provides commercial and scalable machine learning techniques for intelligent and large-scale data analysis application. It alleviates the Big Data problems. Apache Mahout runs on top of Hadoop with the Map/Reduce for identifying patterns, filtering and so on.

2.2.5 Jaspersoft BI Suite

Jaspersoft [44] BI suite is an open source software that are used in many business intelligent systems. It can provide fast data visualization from the Big Data stored on various databases such as MongoDB, CouchDB, Riak, and so on. It can also build effective HTML5 reports and dashboards from the Big Data.
2.2.6 Karmasphere Studio and Analyst

Karmasphere [45] provides an efficient approach in a collaborative way for analytics and self-service access to Big Data. It can used with Hadoop and also provides a user friendly platform for analysing Big Data.

2.2.7 Pentaho Business Analytics

Pentaho [46] can produce results from both structured and unstructured data. It can be used to deliver services for business with visualization of data. It enables business users to get results with positive effect on the output of the organization. It also provides several security and scalability features.

2.2.8 Skytree Server

Skytree server [47] can generate accurate results from large volume data sets. It uses machine learning in a sophisticated way. It is generally used for performing real time analytics. It can also produce results from both structured and unstructured data in an efficient way.

2.2.9 SAP-HANA

SAP-HANA is a new tool for “In-memory computing”. It processes on block of the data by using advanced parallel architecture and algorithms for faster speed [39].

2.3 Big Data Tools for Stream Processing

Hadoop’s performance goes down slightly when it comes for real time data. While processing real-time data stream processing is necessary. Tools such as Strom [47], SQLstream [47], and StreamCloud [48] are used in the Big Data platform to process real-time data.
2.3.1 Storm

Storm is an open source system used to process high streaming data. It has high fault tolerance and is distributed in manner. It is generally used to process real-time applications and contrasts with Apache Hadoop which is used to batch systems. Hadoop works on Map/Reduce whereas Strom makes use of different topologies for different tasks [47].

2.3.2 SQLstream s-Server

SQLstream can process real-time data in an intelligent manner. It can determine patterns from very large unstructured data in a real-time fashion. The in-memory processing used in SQLstream is generally called as ‘NoDatabase’ technology. It processes the data at a high speed using multiple cores, and parallel data analysis [47].

2.3.3 Apache Kafka

LinkedIn developed Kafka [49] to get high-throughput messaging system. It is used for real-time decision making, and also publishes and subscribes the messages in a distributed system. The four main categories of Kafka are high-throughput, distributed support, disk structures, and parallel data load.

2.4 Data Analytics in Cloud Computing

Cloud Computing is a successful paradigm of service oriented computing, capable in providing on-demand, Pay-as-you-use, secure storage management, easy and agile development, and ubiquitous access of computing resources to its shared remote users for data storage and processing [50]. The world of Cloud Computing is totally virtual to its users that require minimum effort from user to manage with features like: on-demand, scalability, reliability, maintenance, cost-effective, and flexibility. The services are delivered to users of it through the use of Internet and sharing of resources can be done using
network of remote servers to store, manage and process data with distributed data processing system. Its service-oriented architecture supports "everything as a service", offers their "services" according to different models with infrastructure-, platform- and software-as-a-service.

As illustrated in Figure 2.10 Big Data and cloud computing technologies are continually evolving and are complementary technologies. The benefits of using them in an integrated way are: scalability, agility and elastic on-demand availability of data [51]. Environment of Big Data need cluster of servers to maintain the tools capable to process large data volumes with high velocity and varied formats. This type of service is offered by cloud computing in a cost-effective way with deployment of cluster of servers, storage and networking resources that can scale us or down as needed. With the use of Cloud Computing a single server can serve multiple customers to retrieve and update their data without paying for different applications.

![Figure 2.10: Big Data and Cloud Computing integration](source: “Applications of Cloud Computing” [52])

Cloud Computing technology deliver parallel-processing as a solution for handling Big Data technologies with advanced analytical application. Several
benefits can be achieved using Infrastructure-as-a-Service (IaaS) in cloud such as: reduce data centre costs, improve utilization, secure and scalable data solution.

2.5 Resources in Cloud for Big Data

An ideal computing environment can be created using cloud and offering many different products for Big Data users. IaaS requires more investment of IT resources in implementing Big Data analytics with installation of software like: Hadoop framework, NoSQL database as Cassandra, MongoDB etc. Some examples of such type of cloud providers with IaaS for Big Data include: Amazon.com, AT&T and IBM. Some examples of using Cloud with Big Data:

2.5.1 IaaS in Public Cloud

Using IaaS provider can be capable to create on-demand virtual machines with unlimited storage and large processing power. The infrastructure of public cloud provider would be used for Big Data services as anybody doesn’t want to use their infrastructure. An example: Amazon Elastic Compute Cloud (Amazon EC2) service to run real-time predictive model, requires parallel-processing of massively distributed data in a scalable manner [53].

2.5.2 PaaS in a Private Cloud

PaaS provides tools and libraries to its developers in cloud to fast develop, run and deploy applications in a private or public cloud without worry about maintaining complexities of Hadoop like implementation environment. PaaS integrated with Big Data is a fully packaged infrastructure that includes Big Data software, infrastructure, tools and managed services. Using PaaS enterprises can rapidly develop secure tools and techniques to Big Data analytics applications. PaaS developers are moving to enhance capabilities of Hadoop and MapReduce() like Big Data analytics applications [51]. An example: Google Cloud Engine offers cloud based capabilities for virtual machine computing with secure and flexible environment.
2.5.3 SaaS in Hybrid Cloud

SaaS can be provided as a standalone application or a solution to its developers by cloud. SaaS provides specific cloud-based application to its customers as required or analyzed by them. Applications that are required by business users are merged into SaaS and these applications are provisioned to its users in a Pay-as-you-go fashion. An example: Amazon Elastic MapReduce() that provides a Hadoop framework with easy, fast, and cost-effective processing of vast amount of data across dynamically scalable Amazon Cloud Computing instances [36].

2.6 Design Principles for Big Data Systems

The seven principles in designing Big Data systems are as follows:

1. Good and proper architectures and frameworks are required for proper functioning of Big Data systems. The use of Lambda architecture can putrefy the Big Data problem into three layers such as serving layer, batch layer, and speed layer.

2. Big Data applications must support a variety of analytical methods, making the application perform complex tasks.

3. Having tools that can perform different task for any given data sets.

4. Getting the right data to analyse

5. The Big Data processing must be distributed across multiple clusters.

6. The data sets must be distributable for in-memory storage.

7. A proper coordination needs to happen between the processing and the data sets used for processing.
2.7 Data Mining

2.7.1 Introduction to Data Mining Systems

Recent inventions and advances in various fields of IT industry and information communication have given a sudden blow to the rapidly increasing data used by different organizations or even by an individual. With central and remote data storage facility, the one like service provided by the Cloud Computing technology, the amount of data to be managed and secured is increasing day by day at a rapid speed. The type of data which is being stored at distributive locations over the internet or at a single repository is dynamic. Dynamic data is continuously changing and updating data and hence requires the dynamic management and processing over the data information. This real time data challenge problem requires dynamic knowledge discovery and pattern identification whenever any new information or data is added.

To extract and determine the useful information, knowledge and certain data patterns from this large unstructured, wide and distributed data organizations use various data mining tools. Large amount of data stored in databases can be discovered by the process of data mining or KDD (knowledge discovery process). These data can also be extracted from data warehouses, or other information repositories. Various methods are used for data mining process such as the intersection of artificial intelligence, machine learning, statistics and database systems. By this process not only the analysis task can be done, but also data management aspects, data pre-processing, inference rules, interestingness measures, complexity considerations, visualization and online updating be done. Typical data mining process involve data selection from target data, data pre-processing, and transformation of data to the appropriate type for mining process, mining of useful patterns, and finally followed by interpretation and evaluation.
For the successful application of data mining for pattern discovery and knowledge in this frequently changing uncertain data, behavioural data, organizational and distributed data, several data mining models are used. These models are implemented in different techniques used by different Data Mining Tools.

For data mining task different tools are used such as Orange, WEKA (Waikato Environment for Knowledge Analysis), Rapid Miner, KNIME and R etc. By these tools user friendly interface, visualization tools and modelling algorithms for the analysis of data and to aid with the ease of performing the task can be performed. For machine learning, pattern recognition, information retrieval, clustering, and data analysis process, all these tools can be applicable. These tools also includes characterization and discrimination, mining frequent patterns, associations and correlations, classification of data and prediction over this data to identify the recurring patterns. This particularly helps in business analysis by identifying the customer buying behaviour. Other important functionalities of these tools include cluster analysis, outlier analysis to detect noises within the useful data and evolution analysis.

Some challenges with the data mining are as follows:

1. In every process dynamic updating is not possible.
2. Real time pattern evolution is not achieved.

The solution of above challenges can be sort out by adopting certain ways such as:

1. Selection of the most suitable tool for particular data cluster can easily be extracted by comparative study.
2. Enhancement over some functionality provided by the tools for more exact knowledge mining.
2.7.2 Data Mining

Several different types of data are being stored in the data repositories and each of them may store data in the single data base structure or the data as parts belonging to a particular field, in the form of data marts. No matter what is the way of the storage of the data the important thing is to determine the value of that data and its future use.

Many of the organizations or even an individual store the data which is required to return some sort of information of value to the owner. Thus data are stored in a fashion which may related to each and every other data and provide some hidden fact or information necessary for evolution to further data. This hidden file of facts and some vital information is not just visible to the user but is present in the vast data in some form. The larger the data size the larger may be the size of the hidden information and greater is the possibility of deriving patterns and rules from it. The data mining is as has been mentioned in several articles and papers and books, is the technique to extract this fact or information from the stored data repository. It is the process to work out the data using some methods of extraction.

The structure of stored data is according to the requirement of the user, may be needed to organize in different manner. These tasks include characterization, classification, clustering of data which is then followed by, if necessary, and the task of determining the association with different data of different fields. The then obtained data can be said as knowledge or information gain.

There are two approaches used for mining the much needed data:

2.7.2.1 Descriptive Approach

This approach for mining data is based on the definition of the attributes defining the instances in the data. Several data are related to each other on the bases of the characteristics of the data tuples which includes attributes for that
data. One data is related to the other data by the degree of relevance of the attribute tuples of those data, and visualization of the data in the data sets and the most important by the association of the data.

### 2.7.2.2 Predictive Approach

Predictive approach is as the title suggests is used to predict the values of the data attributes. This approach is what is used with the classifier models to classify the given data set and predict its value for the future data to be used. The given data sets are preprocessed and executed over with the classifier model. Among several classification algorithms available, one is used to build the model over with training data and is processed over the test data which is the data for which the attribute values are to be determined. The training data is the supplied data used to make learn the classifier about the pattern of the attribute values of the data in the training set.

### 2.8 Data Mining Methods

#### 2.8.1 Classification

Classification is the task of finding groups and structures which are similar to one another in some way without knowing structures in data. The most common algorithms are AQ Algorithm and decision tree algorithms.

Classification is the method of classifying the presented data to the particular class label. The classification is done in order to determine the potential of the supplied data if where it will be best suited in classification categories. In this method of data mining the data to be classified is presented as test data and classified using a classifier built for the purpose.

The classifier is built using the classification algorithms and over a data set where it can be made to learn about the classification pattern for the data. That supplied data set is called training data which is used to train the classifier.
The classifier then learns the rules for classification and applies to the test data. The classified label is identified using the attributes for any data tuple and is then labelled to the specific class. The number of classes for categories remains same as for what is there defined in training data set.

The learning pattern used by the classifier is not changed over the test data so it is required for training and test data to be of compatible types.

2.8.2 Regression

Regression is the type of data mining method all same to that of classification but different in terms of the type of data which is used for the classification task. In regression the data type used in the data sets is of the ordinal or numeric type. The classifier algorithms which are used for the classification task are can be easily used for the similar task of data mining with regression. The modification required for the algorithms to perform regression is supported by the tools used for mining. Pre-processing the data can also help with the regression as the attributes and data type can be changed from discrete to continuous and from continuous to discrete. This process is same to the method of data mining through classification task.

2.8.3 Clustering

Clustering is another way of data mining. This is somewhat different from the methods of classification and regression. In this method the class labels are not known prior to the execution and labelling and this is the difference between the classification and clustering and hence this approach is called as unsupervised learning method of data mining. In clustering the data instances are labeled with some category or can be said by class labels. Then for all data instances, the instances with similar properties and attributes are places under one class label. This process involves discovery of some new class labels not known previously during the process. The technique used by this mining
method is of measuring the distance if the data tuples over a graph or feature space for each tuple. The close ones and similar are classified under the same label and those at some distance are kept in different labels. The measure of distance to be considered which labelling the tuples or objects, depends upon the density of the data over the feature space.

The tuples with same labels are then classified under one cluster and the others are in different. This process is called clustering. The distance measure can be of any one used either Euclidean distance or other.

The advantage of clustering over classification is that clustering is adaptable to changes and easily distinguishes different clusters, also it requires less collection cost for tuples and pattern as compared to classification [54]. The clustering method is appropriate for discovering the internal relationship among datasets for preliminary assessment of sample structures. With automatic clustering one can easily performs it on one, two or three dimensional data but with higher dimensions it is difficult to construe data from high dimensional space. Clustering methods are difficult to categorize in a crisp form as there is an overlapping of these categories which may have features of different categories.

These clustering algorithms can be broadly classified into many techniques e.g. hierarchical, partition and density-based clustering. The hierarchical clustering method can be further divided into agglomerative and divisive [55].

### 2.8.3.1 Partition Clustering

In partition based clustering techniques the dataset is spitted into various clusters. By partitioning each data element is given the cluster index. In this approach user has to predefine the number of clusters with some criteria or parameters on the basis of which the solution will be evaluated. For example one of the parameter could be distance between the data points on the basis of which each cluster has certain number of data elements. The most popular
algorithms of partitioning cluster are K-Means, PAM (Partitioning Around Mediods) and CLARA (Clustering LARge Applications). The well known distance method for this category is Euclidean distance method for K-Means.

2.8.3.2 Hierarchical Clustering

In hierarchical clustering, the hierarchy of clusters is build that is a tree of clusters which is also called dendrogram. It represents the result of cluster analysis. In this technique the prior number of clusters is not to be mentioned. There are two types of Hierarchical clustering (i) agglomerative and (ii) divisive approach. The agglomerative is bottom-up as it starts from bottom with one cluster and merges two or more clusters. The divisive is top-down approach in which each cluster is split into different number of clusters. The most popular algorithm of this category is CLUE.

2.8.3.3 Density-based Clustering

The purpose of density-based clustering is to determine arbitrary y-shaped clusters. In this technique the low density regions are separated from the high density region. This algorithm is also used for determining connected graph as it checks the distance between core point with other data points and checks whether distance is not less than the defined radius. The density of each data point is calculated by detecting the number of data points in its neighbourhood. The cluster is only said to be dense if it would has more than minimum points; the minimum points are the number of data points which should be present in the cluster. The most common algorithm of this category is DBSCAN which deals with noisy data as well.

2.8.4 Association Rule Learning

In data mining, purpose of association rule learning is to discover or extract interesting links between data items from large databases. The common
algorithm is Apriori algorithm. Let us take example of supermarket where data is gathered about the purchasing habits of customer.

Association rule as the name suggests is the process of determining the rules within the observed data result for any classification. These rules are derived on the bases of the predictions and classification values. They describe the relation between the several attributes and their complementing nature. The rules are used to make the target more specific for achieving more desired results with enhanced efficiency and improved output.

The rule is defined with two parts one antecedent and other consequent. The consequent is dependent on the antecedent of the rule and is formed on its basis. These rules are then can be used to derive more patterns and facts and some more rules from the new data.

2.8.5 Stream Data Mining

The stream data mining is the concept in which the dynamic data is to be analyzed and is used for business and marketing purposes. The data collected from the internet, clicking data etc is being generated day by day thus to analyze this data is one of the difficult task as the rate of flowing of data is different. There are very few techniques to analyze online streaming data.

2.9 Applications of Big Data

Before jotting down a brief description of the applications of Big Data, let us examine the following comparison table which presents some of the implications of Big Data in business intelligence – moving from traditional decision making to data driven decision making [56].
Table 2.1: Traditional (vs) Big Data environments [56]

<table>
<thead>
<tr>
<th>Traditional Data Making environs</th>
<th>Big Data Extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Determine and analyse current business situation.</td>
<td>• Provides complete answers, predict future business situation and investigates new opportunities.</td>
</tr>
<tr>
<td>• Integrated data sources.</td>
<td>• Virtualised and blended data sources.</td>
</tr>
<tr>
<td>• Supports only structured Data.</td>
<td>• Supports multi-structured data.</td>
</tr>
<tr>
<td>• Have aggregated and detailed data within limits.</td>
<td>• Have got large volumes of data (detailed) without any limits.</td>
</tr>
<tr>
<td>• One size fits all data management.</td>
<td>• Flexible and optimized data management.</td>
</tr>
</tbody>
</table>

The Big Data covers wide area for its application and implementation. In this Political sector, Education sector, Healthcare sector , Pervasive Computing industry and Weather Forecasting field are included.

2.9.1 Role of Big Data in Political Campaigning

Big Data analysis is playing the important role in political campaigning because of using it BJP and its allies has won a highly successful Indian General Elections of 2014.

Our lives and environment can be transformed by the use of Big Data. The Data Scientist are well aware of the thing that the use and combination of the different data sources can create competitive advantage when running public posts. In the book of Sacha Issengber titled as “The Victory Lab”, insists on the use of Big Data technique to win political campaigns [56].

2.9.2 Role of Big Data analytics in Education

Education systems at all levels will be benefited from Big Data analytics. By adopting predictive models and analytics, our education system can be modified and value added education system can be developed.
This value added education further can be achieved by increasing collaboration between students, teachers, administrators and parents. In this way satisfaction and better performance can be achieved [56].

2.9.3 Role of Big Data Analysis in Healthcare

By the use of Big Data analytics the personal care of patient can be highly served to advanced and improvement in the health of patient can be acquired at minimum cost. EHR’s (Electronic Health Records) coupled with new analytical tools will significantly open doors to mining information for most effective outcomes across large population [57] surveys by which the diseases can be easily curbed for the analysis. The data scientists can predict the outbreak of the epidemic and it can be controlled easily with the help of Big Data analysis.

2.9.4 Role of Big Data in Weather Forecasting

The Big Data analysis is also useful in weather forecasting where the unstructured data can be analysed. In September 2012; in Palau, Japan, Korea, China and Russia “Typhoon Sanba” (a kind of typhoon), damaged $378.8 million populations (2012 USD) [57].

After that such natural calamity South Korea upgraded its national weather information system that can predict the location and ferocity of weather events [58]. By the Big Data analytics weather data is stored in repositories and the analysis of data and its implementation can be easily achieved.