Big Data and Its Utility

The term “Big Data” is pervasive, and still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. The Big Data dimensions like the volume, variety, velocity, veracity, volatility and variability are the major bottlenecks. However, it can produce operational and business value at an unprecedented scale and specificity.

There are some technical factors that have given an uprising to this Big Data adoption irrespective of the challenges that it poses. Storage cost has dramatically decreased in the last few years. With the Internet of Things scenario, the data deluge will become rather overwhelmed. Big Data tools such as the Hadoop ecosystem provide the technology to increase the processing speed of complex queries and analytics. The main idea behind all this is to extract knowledge from the data and use it for the betterment of the Earth.

‘Big-data’ is similar to ‘Small-data’, but bigger and with added dimensions. It therefore requires different approaches, techniques, tools & architectures to analyze. Often, because of vast amount of data, modeling techniques can get simpler as long as we deal with the scale. In this article, the challenges and opportunities in Big Data analytics have been discussed. It also shows the analytical role in various fields of science. Various Big Data platforms like Cloud Computing, High Performance Computing (HPC) and Graphics Processing Unit (GPU) have been helping the organization to gain insight, by turning data into high quality information, providing a better understanding of the business situations. In this context, the Hadoop environment is discussed along with the MapReduce design and other components of Hadoop. Finally, some case studies have been shown that depict how the Big Data analytics have started shaping our world.

Keywords: Big Data, Hadoop, Map Reduce, Importance of Big Data.

1. What is Big Data?

The concept of “Big Data” (BD) was first introduced to the computing world by Roger Magoulas from O’Reilly media [1]. It defines a huge volume of data that traditional data management techniques cannot manage and process due to its complexity and size. The story of how data became big started many years before the current buzz around big data. Almost seventy years ago we had encountered the first attempts to quantify the growth rate in the volume of data or what had popularly been known as the “information explosion” (a term first used in 1941, according to the Oxford English Dictionary). However,
today the concept is treated from different perspectives rather than only its size. The data becomes big data when its volume, velocity, or variety exceeds the abilities of an information system to ingest, store, analyze, and process it [2]. It has therefore been covering various applications in many fields [3]. In 2014, it was estimated that the worldwide data generation would increase by a staggering pace of 7 Zeta Bytes by the end of 2015. As a mere example, it is expected that by 2018 each smartphone will generate 2GB of data every month. At the same time, it is understood that the amount of data will grow at a rate of 40 percent each decade over the present volume. Thus, it is evident that the growth of data is exponential. With such a pace in growth of data, the need for evolution of analytics tools is also necessary to process it.

Big data analytics is the process of examining the big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. The term big data analytics, when used by software vendors, refers to the technology (which includes tools and processes) that an organization requires to handle data at extreme scales. Such analytics has edges over the conventional analytics and business intelligence solutions in extracting useful information from huge volume of data. This has helped data scientists and others in unwinding the complex hidden causal relationship among the attributes. For the case of handling one billion rows of data with millions of data combinations in multiple data stores and in abundant formats, big data analytics provide distributed, scalable and generic framework to process or mine them [4]. Not only does this make Big Data management and storage vastly different from normal (or structured) data that most people are accustomed to handling, but it also means that organizations now require powerful, integrated solutions for making this information usable and applicable for business, analytics practices.

The main aim of this article is to raise interest for research on “big data” and to inform the audience about the challenges in BD research, rather than to discuss all topics related to big data in detail.

2. HACE Theorem: Characteristics of Big Data

The HACE theorem models the BD characteristics [5]. It states that BD starts with large volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from data. This can be understood by a small example. Suppose there are six blind men who had no idea of what an elephant is. They went to feel it anyway and so everyone touched it. One of them said that the elephant is like a pillar. The other five also recognized the elephant as rope, hand fan, huge wall, solid pipe and a thick branch of a tree depending upon which part of the elephant they touched. Now, because each person’s view is limited to a local region, it is not surprising that each blind man will conclude independently about the elephant. The problem gets more complicated if the elephant is gradually getting increased in size (that is what happens in real time perspective of big data).

Thus HACE theorem suggests that big data can be characterized by the huge volume of the data with miscellaneous representation due to the distributed, autonomous and decentralized nature of the source in generation of the data. For example, the data generated by Facebook, Twitter, MySpace, WhatsApp and every other social media on a particular event can have different representations as each of them is autonomous and has its independent method of recording the information. Thus it is not only lots of data or massive data as the name might suggest but also encompasses other dimensions [6] as discussed below.

- **Volume:** The quantity (volume) of data that is generated is very important in this context. It is the size of the data which determines the value and potential of it and whether it can actually be considered as BD or not. 90% of all data ever created was generated in the past 2 years. If we take all the data generated in the world till 2008, the same amount of data will soon be generated every minute. Scale is certainly a part of what makes the data ‘Big’. We therefore need to use tools that can store and analyze data across databases that are dotted around anywhere in the world.
however, is the sheer size of video data. However, BD technologies turn this challenge into an opportunity [12]. The primary application of video analytics in recent years has been in automated security and surveillance systems [10, 11]. Video analytics can efficiently and effectively perform surveillance functions such as detecting suspected objects, identifying objects removed or left unattended, detecting loitering in a specific area, recognizing suspicious activities, detecting camera tampering, etc.

Automatic video indexing and retrieval constitutes another domain of video analytics applications. The widespread emergence of online and offline videos has highlighted the need to index multimedia content for easy search and retrieval. The indexing of a video can be performed based on different levels of information available in a video including the metadata the soundtrack, the transcripts, and the visual content of the video [13].

5.2 Role in Social Media Analysis

Social media is a broad term encompassing a variety of online platforms that allow users to create and exchange contents from social media channels. Social media can be categorized into the following types: social networks (e.g., Facebook, LinkedIn), messaging apps (e.g., Viber, WhatsApp, Telegram etc.), scientific networks (e.g., ResearchGate, Academia.edu, Mendeley, ResearchID), blogs (e.g., Blogger and WordPress), microblogs (e.g., Twitter and Tumblr), social news (e.g., Digg and Reddit), social bookmarking (e.g., Delicious and Stumble Upon), media sharing (e.g., Instagram and YouTube), wikis (e.g., Wikipedia and Wikihow), question-and-answer sites (e.g., Yahoo! Answers and Ask.com) and review sites (e.g., Yelp, Trip Advisor) [14]. The key characteristic of the modern social media analytics is its data-centric nature. Research on social media analytics spans across several disciplines, including psychology, sociology, anthropology, computer science, mathematics, physics, and economics.

Social network analysis is a branching discipline of social science, which focuses on individual molecules, such as sequence of nucleotide acids and amino acids, systems biology focuses on systems that are composed of molecular components and their interactions. Analysis of gene regulatory networks
Involves: finding an optimal pathway, and effect of gene expression/regulation on other pathways. For a particular tissue, every gene is not expressed, only a subset of it is expressed and it exhibits some pattern over time. The objective is to find this temporal pattern from such a huge volume of data.

5.4 Role in Health Care

Radiological sciences in the last two decades have witnessed a tremendous progress in medical imaging and computerized processing of medical images [16]. Healthcare organizations can leverage proven innovations to put big data and analytics to work for them. Medical imaging is an important cornerstone of modern healthcare and will continue to play a role of ever increasing importance at all levels of the healthcare system. The data come in various forms like waveforms, 3D reconstructions, videos, genomic data, ECG waveforms, ultrasound images, MRI, lab results etc. Medical images are now archived longer and are increasing by 20-40% annually. Reasons for such a growth are obvious. Many people are aged but still the population is affluent. Also technologies are improved and the medicines are quite defensive. Genomics has been the cutting edge of the big data revolution in life sciences that holds considerable promise for enabling personalized medicine.

Big data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. With the world’s population increasing and everyone living longer, models of treatment delivery are rapidly changing, and many of the decisions behind those changes are being driven by data.

To succeed, big data analytics in healthcare needs to be packaged so as to make it menu driven, user-friendly and transparent. Real-time big data analytics is a key requirement in healthcare. Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions.

5.5 Role in Scientific Data Analysis

Scientific data collection involves more care because scientists build on their own work and the work of others based on measurements. It is therefore important that they are systematic and consistent in their data collection methods and make detailed records so that others can see and use the data they collect.

But collecting data is only one step in a scientific investigation, and scientific knowledge is much more than a simple compilation of data points. The world is full of observations that can be made, but not every observation constitutes a useful piece of data. For example, a meteorologist could record the outside air temperature every second of the day, that will not make the forecast more accurate than recording it once an hour. All scientists make choices about which data are most relevant to their research and what to do with those data: how to turn a collection of measurements into a useful dataset through processing and analysis, and how to interpret those analyzed data in the context of what they already know.

The thoughtful and systematic collection, analysis, and interpretation of data allow them to be developed into evidence that supports scientific ideas, arguments, and hypotheses.

Nearly everything about high-energy physics research is big - thousands of scientists are working on a single experiment, the miles-long accelerators, the stories-high detectors. Yet “big” does not begin to describe the torrent of data that gushes out of high-energy physics experiments every day. Out of sheer necessity, particle physicists are virtuosos at plying vast troves of data into meaningful physics.

6. Big Data Platforms

Big data platforms must operate and process data at a scale that leaves very small place for errors. Clusters to handle big data must be built for speed, scale and efficiency [12]. These platforms help the organization to gain insight, by turning data into high quality information, providing a better understanding of the business situations. This basically means turning of raw data into powerful statistics. Here are some of the data management platforms.

6.1 Cloud Computing

Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on
our own hard disk or updating applications for our needs, we use a service over the Internet, at another location, to store our information or use its applications [17]. Cloud services allow individuals to use software and hardware that are managed by third parties at remote site. Online file storage, social networking sites, webmail and online business applications are certain common examples for this. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power and specialized applications. The following definition of cloud computing has been developed by the U.S. National Institute of Standards and Technology (NIST) [17]:

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models [18].

Cloud computing and big data are conjoined. Cloud computing includes virtualized resources, parallel processing, security, and data service integration with scalable data storage. This can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access. Big data utilizes distributed storage technology based on cloud computing rather than local storage attached to a computer or electronic device. Big data evaluation is driven by fast-growing cloud-based applications developed using virtualized technologies. Therefore, cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model.

6.2 Internet of Things

The Internet of Things (IoT) is a scenario in which objects, animals or people are provided with unique identity and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction. A thing, in the IoT, can be a person with a heart monitor implant, a farm animal with a biochip transponder, an automobile that has built-in sensors to alert the driver when tyre pressure is low - or any other natural or man-made object that can be assigned an IP address and provided with the ability to transfer data over a network. So far, the IoT has been most closely associated with machine-to-machine (M2M) communication in manufacturing and power, oil and gas utilities. Products built with M2M communication capabilities are often referred to as being smart.

Fig. 1: IoT connects all things

A significant increase in connected devices is due to happen at the hands of the IoT. This will, in turn, lead to an exponential increase in the data that an enterprise is required to manage. Here is where IoT intersects wonderfully with Big Data - and where it becomes evident that the two trends fit one another like a glove. The internet-accessible cloud now provides an easily accessible, low-cost platform for storing and processing data collected, while big data tools support the analysis of vast amounts of data. The promise of future social and economic benefits that these developments could offer is stoking current interest in the IoT.

The IoT is expected to make impacts in government, education, finance and transportation. On the consumer side there are nearly endless combinations of applications.

6.3 High Performance Computing (HPC) Clusters

HPC clusters [19], also called blades or supercomputers, are machines with thousands of cores. They can have a different variety of disk organizations, cache, communication mechanism etc. depending upon the user
requirement. These systems use well-built powerful hardware which is optimized for speed and throughput. Because of the top quality high end hardware, fault tolerance in such systems is not problematic since hardware failures are extremely rare. The initial cost of deploying such a system can be very high because of the use of the high-end hardware. They are not as scalable as clusters of commodity hardware but they are still capable of processing terabytes of data. The cost of scaling up such a system is much higher. The communication scheme used for such platforms is typically MPI (Message Passing Interface) [20]. Since fault tolerance is not an important issue in this case, MPI's lack of fault tolerance mechanism does not come as a significant drawback here. HPC is well positioned to enable big data use cases through all three phases of typical workflows, including: data capture and filtering, analytics and results visualization. In addition to the three phases, the speed of computation matters just as much as the scale. In order to unlock the full potential of big data, we have to pair it with “big compute”, or HPC.

6.4 Graphics Processing Unit (GPU)

Graphics Processing Unit (GPUs) [21] is a specialized hardware designed to accelerate the creation of images in a frame buffer intended for display output. Until the past few years, GPUs were primarily used for graphical operations such as video and image editing, accelerating graphics-related processing etc. However, due to their massively parallel architecture, recent developments in GPU hardware and related programming frameworks have given rise to GPGPU (general-purpose computing on graphics processing units) [21]. A GPU has a large number of processing cores (typically around 2500 to date) as compared to a multi-core CPU. In addition to the processing cores, a GPU has its own high throughput DDR5 memory which is many times faster than a typical DDR3 memory. If the process is optimized for parallel computing, a GPU could perform calculations 100 times faster than a CPU. GPU performance has increased significantly in the past few years compared to that of CPU. Recently, Nvidia has launched Tesla series of GPUs which are specifically designed for high performance computing. Recently, the machine learning community has started embracing GPU-based systems for some of their most challenging problems. In particular, the training of deep learning algorithms has had some great success stories. Training these algorithms can traditionally take weeks or even months on CPU-based systems. On GPU-based systems, however, they are now able to train in days or even hours. Hadoop leveraging GPU technologies such as CUDA and OpenCL can boost big data performance by a significant factor.

Massive parallelism in GPU makes it a more appealing option for parallel computing applications. However, GPU has its own drawbacks. The primary drawback is the limited memory that it contains. Once the data size is more than the size of the GPU memory, the performance decreases significantly as the disk access becomes the primary bottleneck. Another drawback is the limited amount of software and algorithms that are available for GPUs.

7. Big Data: Technological Perspective

Big Data technologies are important in providing more accurate analysis which may lead to more concrete decision making resulting in greater operational efficiencies, cost reduction and reduced risks for business. BD handling needs an infrastructure that can manage and process huge volumes of data (both structured and unstructured) in real time with privacy and security. The BD landscape is dominated by two classes of technologies: systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored; and systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data. These classes of technologies are complementary and frequently deployed together.

One of the most famous and powerful batch-process based BD analysis tools is Apache Hadoop. Hadoop is an open-source software framework for storing and processing BD in a distributed fashion on large network of commodity hardware. It provides infrastructures and platforms for other specific BD applications. Hadoop fills a gap in the market by effectively storing and providing
computational capabilities over substantial amount of data. It's a distributed system made up of a distributed file system and it offers a way to parallelize and execute programs on a cluster of machines. Hadoop has been adopted by technology giants like Yahoo, Facebook, and Twitter to address their BD needs, and it is making roads across all industrial sectors.

Hadoop is actually the name that creator Doug Cutting's son gave to his toy elephant. Hadoop is a top-level Apache project in the Apache Software Foundation written in Java. For all intents, Hadoop can be thought of as a computing environment built relying on the distributed file system designed specifically for very large-scale data operations. It was inspired by Google's work on its Google File System (GFS) and MapReduce programming paradigm. Unlike transactional systems, Hadoop is designed to scan through large data sets to produce results through a highly scalable, distributed batch processing system.

8. Working of Hadoop

The objective of Hadoop is to support running of applications on BD. It is an open source system and is distributed under Apache license. This guarantees that no particular company is controlling the direction of Hadoop and it is maintained by Apache. Hadoop addresses BD problems at three levels: volume, velocity and variety.

Hadoop breaks the data into pieces and therefore it is able to deal with the Big Data [Figure: 2].

It breaks the computation as well down into smaller pieces and it sends each piece of computation to each piece of data. So, the data is broken down into equal pieces such that the computation is fast and finished in equal amount of time. Once the computation is finished, the combined result is given to the application as the job done.

The main working of Hadoop depends on distributed data storage and distributed computation. At a very high level it has a simple architecture. Hadoop has two main components: MapReduce and a distributed file system (HDFS).

The data processing framework is the tool used to work with the data itself. By default, MapReduce is a Java-based system. Hadoop also has another Apache application - Hive, that helps convert query language into MapReduce jobs. The jobs are independently handled by some commodity servers. Each commodity has two parts: Task Tracker and Data Node.

The job of the task tracker component is to process the smaller pieces of task that has been given to a particular node. The job of the data node is to manage the piece of

![Fig. 2: Modern framework of Big Data analysis](image-url)

![Fig. 3: Hadoop distributed model](image-url)

![Fig. 4: The Job Tracker is responsible for managing the worker nodes i.e. Task Trackers and the Task Tracker has simple responsibilities - launch tasks on orders from the Job Tracker and provide task-status information to the Job Tracker periodically](image-url)
data that has been given to this particular node. All these computers are called slaves. There is a master computer too. The master computer will have two additional components than the slaves - Job Tracker and Name Node.

8.1 Job Tracker

The Job Tracker is the service within Hadoop that frames out MapReduce tasks to specific nodes in the cluster, ideally the nodes which have the data, or at least are in the same rack. The job tracker is responsible for assigning the job to a task tracker near the location of the data and carrying out the housekeeping activities related to the job. It communicates with the client program regarding the status of the assigned job. The Job Tracker is a point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted.

8.2 Task Tracker

A Task Tracker is a node that accepts tasks (Map, Reduce and Shuffle operations) from a Job Tracker. Every Task Tracker is configured with a set of slots; these indicate the number of tasks that it can accept. When the Job Tracker tries to find somewhere to schedule a task within the MapReduce operations, it first looks for an empty slot on the same server that hosts the Data Node containing the data, and if not, it looks for an empty slot on a machine in the same rack. The Task Tracker spawns a separate Java Virtual Machine (JVM) processes to do the actual work; this is to ensure that process failure does not take down the task tracker. The Task Tracker monitors these spawned processes, capturing the output and exit codes.

When the process finishes, successfully or not, the tracker notifies the Job Tracker. The Task Trackers also send out heartbeat messages to the Job Tracker, usually every few minutes, to reassure the Job Tracker that it is still alive. These messages also inform the Job Tracker of the number of available slots, so the Job Tracker can stay up-to-date with where are the tasks delegated.

9. Hadoop Distributed File System (HDFS)

There is a distinction between an HDFS file and a native file on the host computer. A computer in an HDFS installation which is (typically) allocated to one Name Node or one Data Node. Each computer has its own file system and information about an HDFS file (the metadata) is managed by the Name Node and persistent information is stored in the Name Node’s host file system. The information contained in an HDFS file is managed by a Data Node and stored on the Data Node’s host computer file system.

HDFS exposes a file system namespace and allows user data to be stored in HDFS files. An HDFS file consists of a number of blocks. Each block is typically 64 MB or 128 MB. Each block is replicated some specified number of times. The replicas of the blocks are stored on different Data Nodes chosen to reflect loading on a Data Node as well as to provide both speed in transfer and resiliency in case of failure of a rack.

A standard directory structure is used in HDFS. That is, HDFS files exist in directories that may in turn be subdirectories of other directories, and so on. There is no concept of a current directory within HDFS. HDFS files are referred to by their fully qualified name which is a parameter of many of the elements of the interaction between the client and the other elements of the HDFS architecture.

9.1 Name Nodes and Data Nodes

The Name Node executes HDFS filesystem namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to Data Nodes. The list of HDFS files belonging to each block, the current location of the block replicas on the Data Nodes, the state of the file, and the access control information is the metadata for the cluster and is managed by the Name Node.

The Data Nodes are responsible for serving read and write requests from the HDFS file system’s clients. The Data Nodes also perform block replica creation, deletion, and replication upon instruction from the Name Node. The Data Nodes are the arbiter of the state of the replicas and they report this to the Name Node.

The existence of a single Name Node in a cluster greatly simplifies the architecture of the system. The Name Node
Dr. Ashish Ghosh

is the arbitrator and repository for all HDFS metadata. The client sends data directly to and reads directly from Data Nodes so that client data never flows through the Name Node.

HDFS is a file system designed for storing very large files with streaming data access patterns, running on clusters on commodity hardware. Hadoop has an abstract notion of file system, of which HDFS is just one implementation. In the context of HDFS, we will discuss some terminologies in the following sections.

Hadoop MapReduce applications use storage in a manner that is different from general purpose computing. To read an HDFS file, client applications simply use a standard Java file input stream, as if the file was in the native file system. Behind the scenes, however, this stream is manipulated to retrieve data from HDFS instead. First, the name node is contacted to request access permission. If granted, the name node will translate the HDFS filename into a list of the HDFS block IDs comprising the file and a list of data nodes that store each block, and return the lists to the client. Next, the client opens a connection to the "closest" data node (based on Hadoop rack-awareness, but optimally the same node) and requests a specific block ID. This HDFS block is returned over the same connection, and the data delivered to the application. To write this data to HDFS, client applications HDFS file is seen as a standard output stream. Internally, however, stream data is first fragmented into HDFS-sized blocks (64MB) and then smaller packets (64KB) by the client thread. Each packet is enqueued into a FIFO (First In First Out) that can hold upto 5MB of data. This may lead to the decoupling of the application thread from storage system latency during normal operation. A second thread is responsible for de-queuing packets from the FIFO, coordinating with the name node to assign HDFS block IDs and destinations, and transmitting blocks to the data nodes (either local or remote) for storage. A third thread manages acknowledgements from the data nodes that data has been written to disk.

10. MapReduce

MapReduce is a parallel programming data processing model introduced by Google. In this model, a user specifies the computation by two functions, Map and Reduce. In the mapping phase, MapReduce takes the input data and feeds each data element to the mapper. In the reducing phase, the reducer processes all the outputs from the mapper and arrives at a final result. In simple terms, the mapper is meant to filter and transform the input into something that the reducer can aggregate over. At a first glance, one would think MapReduce is fairly inefficient because it must break up the problem, distribute the problem (which may be sub-divided yet again), and then assemble all the results from the worker node to create the final answer. This seems like a great deal of work just to set up the problem and execute it. For small problems, it is faster to execute the application on a single node than to use MapReduce.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a network is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

MapReduce has two main stages of operation

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the HDFS. The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data. In other words, it maps the input key/value pairs to a set of intermediate key/value pairs. This basically serves the task of mapping an abstract structure to unstructured formats of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
10.1 Other Components of Hadoop

There are also other supporting components associated with Apache Hadoop framework.

Hadoop has become a brand name which contains the following components:

- **HBASE** - Developed using the Java programming language, HBASE is a layer on top of HDFS and is non-relational, scalable, and fault tolerant. In HBASE, every single row is identified using a key. The number of columns is also not fixed. These columns are grouped into column families.

- **Zookeeper** - This is a centralized service which maintains configuration information, naming information and synchronization information. On top of these, Zookeeper also provides group services and is used by HBASE. It is also used by the MapReduce programs.

- **Solr/Lucene** - This is used as the search engine. This library is developed by Apache and it took over 10 years to have this robust search engine.

- **Programming Languages** - The following two languages are identified as original Hadoop programming languages: Hive and Pig.

11. Importance of Hadoop Platform

Since its inception, Hadoop has become one of the most talked about technologies. One of the top reasons is its ability to handle huge amounts of data, any kind of data, and speed. With volumes and varieties of data growing each day, especially from social media and automated sensors, speed is the key consideration for most organizations. Other reasons include:

- **Low cost**: The open-source framework is free and uses commodity hardware to store large quantities of data.

- **Computing power**: Its distributed computing model can quickly process very large volumes of data. The more computing nodes you use the more processing power you have.

- **Scalability**: You can easily grow your system simply by adding more nodes. Little administration is required.

- **Storage flexibility**: Unlike traditional relational databases, we do not have to pre-process data before storing it. This includes unstructured data like text, images and videos. You can store as much data as you want and decide on how to use it later.

- **Inherent data protection and self-healing capabilities**: Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure that the total system does not fail. Also it automatically stores multiple copies of all data.

12. Solving Big Problems with Big Data

Delivering actionable data that can create value requires significant analytical capabilities [22]. Big data analytics is not only limited to the efficient storage of massive volumes of data, but also the processing, predicting and decision making by having proper insights of the data. Significant challenges still exist in the pursuit of novel analytics and novel uses of data. Some domains where big data analytics can be applied are:

- **Data mining**: Data mining techniques can be used to determine why a specific event happened. This causal-effect relationship can bring out several important rules for the occurrence of event. BD mining is the capability of extracting valuable information from huge datasets, which was earlier not an easy task because of the volume, variety and variability of the data.

- **Prediction**: Sophisticated analysis tools may predict the most likely situation that arises based on some current and historical facts. BD approaches may be used to predict a natural disaster like earthquake before it had actually happened and also to provide accurate information regarding the consequences.

- **Risk management**: Risks increase with complexity of an industry. Risk management tool
may identify the potential risks in advance. They may be used to analyze and hence take precautions to reduce the effect of the risk [23]. An enterprise may improve its predictive power of risk modeling and can thus improve its system response, generate significant cost savings and provide more risk coverage.

- **Social media analysis:** The analysis of media content has been central in social sciences, due to the key role that media plays in shaping public opinion. This kind of analysis typically relies on the preliminary coding of the text being examined, a step that involves reading and annotating it, and that limits the sizes of the corpora that can be analyzed. The use of modern data analysis technologies allows researchers to automate the process of applying different codes in the same text and also enable the automation of data collection, preparation, management and visualization. This provides opportunities for performing massive scale investigations, real time monitoring, and system-level modeling of the global media system.

- **Scientific data analysis:** From genome sequencing machines capable of reading a human’s chromosomal DNA (about 1.5 gigabytes of data) in half an hour to particle accelerators like the Large Hadron Collider at CERN (which generates close to 100 terabytes of data a day), researchers are flooded with data. One of science’s primary goals is to understand the relationship between cause and effect. Scientists devote substantial effort to explain why things happen as a consequence of a few fundamental principles. First, we can predict an effect, just by knowing about the cause. Second, we use models to understand why something is a consequence of its causes. Prediction and understanding have been intimately tied together in science, but the influence of big data in science is now breaking them apart. If we measure enough variables, it doesn’t matter whether we understand the relationship between cause and effect; all it needs is a relationship between one variable and another. With enough data, these data analysis models can find such trends, among enormous numbers of variables in complex data that the unaided human mind would never spot.

- **Market analytics:** Marketing analytics is the practice of measuring, managing and analyzing marketing performance to maximize its effectiveness and optimize return on investment (ROI). Marketing analytics, Internet (or Web) marketing analytics in particular, allow the enterprises to monitor campaigns and their respective outcomes, enabling them to spend each penny as effectively as possible.

### 13. Growing Population Fed by Big Data

There are expectations to have more than 9 billion people on the planet by 2050, and you can expect that type of population growth to strain the world’s resources like energy, water and food. But it turns out that the BD tools will be particularly adept at helping organizations track and attempt to solve severe shortages in these resources. To feed the world’s rapidly expanding population in the coming decades, agriculture must produce more. BD holds one of the keys for farmers.

#### 13.1 Big Data in Farming

During the technical boom of recent decades, the agricultural world has quietly been introduced to data aggregation technology. Data systems are being built into a farmer’s machinery. They started enabling Wi-Fi in barns. Larger farms started using software to manage their operations. Adoption has been slow, and systems are often unwanted because they create lock-in to the software, incompatible with the variety of other tools and brands used on the farm.

#### 13.2 Global Water Management using Big Data

Lack of water is a critical constraint to increasing food production, particularly as droughts and other consequences of climate change are making water scarcer. To help solve this enormous challenge, the agriculture and water communities are harnessing BD to ramp up food production with less pressure on water resources.
Satellites gather vast amounts of data that are used at global and local scales. For example, satellites can track atmospheric patterns, precipitation and ocean currents. Combined with weather data, researchers are using satellite data to develop better forecasting and risk-management tools to help farmers make better decisions, and to help governments better plan for droughts and floods. Satellite data also can be used on local areas to precisely map landscapes, analyze soils or assess crop yields, among many other uses.

Globally, agriculture consumes 70% of all freshwater withdrawals, primarily for irrigation, and groundwater is a key source of irrigation water. Data regarding aquifer conditions, groundwater withdrawals and other metrics are critical for water managers to prevent catastrophic aquifer depletions. It has been successfully used in areas of Nebraska for more than 40 years to help maintain groundwater levels, despite having the most irrigated acres in the nation. BD analysis and modeling can reach even subsistence farmers in remote areas. Many farmers in poor and rural areas long isolated by insufficient phone service and roads, have cell phones now. They now have access to weather forecasting and market information to make better decisions, manage money and develop a wider support network, thereby improving livelihoods as well as local water and food security.

13.3 Data Science of Global Warming

Global warming is a really challenging problem. Many people and many species are being affected by the climate change. Scientists have made it clear that human activities are the major causes of this change and warn that damages caused so far (rise of sea levels, acidification of the oceans, melting of glaciers) will remain for centuries if governments don't take drastic measures to counteract global warming.

BD is trying to enable new approaches to achieve sustainable growth and energy solutions through innovations in computation, as well as computational technologies themselves, including data analytics, modeling and simulation, optimization, high performance computing, and distributed computing platforms.

13.4 Sustainable Development and Big Data

Big data technologies can translate information from smart meter and smart grid projects into meaningful operational insights and identify energy consumption trends and needs. By turning one-way transmission and distribution networks into intelligent, two-way information networks through big data analytics, companies can not only take advantage of automation, but also manage resources more efficiently.

We should be collecting BD that can be used to model and test an array of different scenarios for sustainable transforming the production and consumption of energy, improving food and water security, and eradicating poverty. Managing these issues will also help to re-balance important bio-geochemical cycles (especially the carbon, nitrogen and phosphorus cycles), mitigate climate change, reverse ocean acidification and reduce the loss of biodiversity. BD will help to illuminate the origins, nature and scale of these challenges, and how they relate to one another.

National databases and research centers can be linked to create huge databases. Initiatives similar to those of the Inter-governmental Panel on Climate Change and the Global Ocean Observing System could fill the gaps in scientific, technical and socio-economic data. New initiatives such as Global Pulse (www.unglobalpulse.org) could help in mining and mobilizing big data, which are available in real time as a result of the explosive growth in new media.

14. Case Studies

Every government faces numerous challenges, the biggest perhaps being making sense of the massive amount of information they receive every day and making decisions based on the same, which in turn, may affect the entire country or even multiple nations. Not only is it tough to scrutinize all the information, but it is even more difficult to verify it (problem of veracity). Flawed information can have devastating consequences.

With the help of BD, governments can derive crucial insights to aid decision making in real-time from the heaps
of ever-mounting data received from a myriad of sources, including the Web, biological and industrial sensors, video, email, and social communications. Government can utilize BD to serve their citizens better and overcome countless challenges like increasing health care costs, unemployment, natural calamities, poverty, illiteracy, terrorism, international trade relations, and so on. BD in government can be the touchstone of a nation’s global standing.

14.1 Colorado Data Sharing Network (CDSN)

The Colorado Data Sharing Network (CDSN) is the primary project of the Colorado Water Quality Monitoring Council (CWQMC). CDSN is a revolutionary way for organizations to share data with the public and with each other. After data is collected by organizations around the state, it can be easily uploaded and made viewable on the web. Each time data is shared, it not only increases the knowledge of water quality around the state, but hopefully saves time and resources for others who are interested in monitoring in a given basin. It is helping to enhance Colorado’s watersheds and water quality by providing CDSN’s data partners and Colorado’s citizens with an accessible and affordable tool kit for data management, data analysis, and data sharing.

14.2 South Korea Government Improves Citizen Engagement through Open Data and Big Data

The South Korean Ministry of Government Legislation (MOLEG) has significantly improved citizen engagement by enabling easy access to and search of accurate and timely legal information. It gathers all kinds of information related to legislation, current laws and their histories, constitution, laws passed in the national assembly, treaties, presidential decrees, decrees produced by each ministry, and other rules including local governments’ ordinances and regulations. MOLEG has also created a mobile app so that citizens can access the Centre on the go. Besides making legal information open and easily searchable by citizens, MOLEG wants to involve the public in the lawmaking process.

14.3 Indian Government Investing in Big Data

Already, the Prime Minister’s Office is implementing an attendance system for India’s Central Government employees through attendance.gov.in. Similarly, the state Government of Telangana is employing Big Data Analytics for the data collected from nearly 3.5 crore people across the state. Gujarat will emerge as the fourth centre for Analytics after Mumbai, Bangalore and Delhi. The number of analytics companies based in Gujarat has been on the rise due to which there will be an increased demand in that region for trained data scientists. Governments will embark on several initiatives leveraging Big Data—from trying to improve national security to solving infrastructure problems. Government is planning to use Analytics to increase its revenue base. India’s tax-payer base is just about 3 crore and the number has been inching its way slowly for the last 5-10 years, which the government would like to see growing at a faster pace. According to government data, the total tax payers in the country stood at about 3.24 crore during the 2011-12 fiscal. India’s total direct taxes are only 9 per cent of the GDP, whereas it should be about 18 per cent, and it cannot be raised by taxing people who you have already taxed. So, analytics will be put to use. As a nation, all the data can be put together. If someone travels abroad, buys expensive jewellery, government can check his digital footprints on online shopping and piece together his lifestyle and through that can create a taxable database.

14.4 Indian Government using Big Data Analytics to Revolutionize Democracy

The Prime Minister’s Office is using BD techniques to process ideas thrown up by citizens on its crowd-sourcing platform mygov.in, place them in context of the popular mood as reflected in trends on social media, and generate actionable reports for ministries and departments to consider and implement. Ministries are being asked to revert with an action taken report on these ideas and policy suggestions currently being generated on 19 different policy challenges such as expenditure reforms, job creation, energy conservation, skill development and government initiatives such as Clean India, Digital India and Clean Ganga. The Prime Minister is inviting Indian
Big Data and Its Utility

... in other countries to join the online platform, which he has termed a ‘mass movement towards Surajya’. The traffic handling capacity of mygov.in is therefore being scaled up consistently. The biggest issue for governments today is how to be relevant. In the government’s view, if all citizens are treated with dignity and invited to collaborate, it can be easier for administrations to have a direct finger on the pulse of the nation rather than lose it in transmission through multiple layers of bureaucracy.

15. Conclusion and Discussions

Big Data power lies in the analytical realm. It can certainly help us to solve many fundamental problems. Major innovations in BD are still to take place; but, it is believed that emergence of such novel analytics is to come in near future. Data is evolving much more rapidly than humans.

It is not very wrong to say that such a Big Data surge will transform our lives in possibly the next five to ten years. Each student would be taught in a particular method by which he/she can learn easily rather than in a way as if one size fits all. Doctors will use our DNA to keep us well. Smart cities will be developed where problems of the citizens could be anticipated and responded to before they actually develop. For instance, the joint polar satellite mission launching in 2018 will use sensor technology and data to forecast the path of hurricanes and storms. A recent hype has also been created regarding smart service platform for users based on big data perspective.

To some extent rise of Big Data may seem to mean degrading privacy, but it will surely be used in future to serve the mankind in a smarter way. Individuals could be identified by only some information like location, face, fingerprint, behavior, banking information etc. Big Data analytics can harness various real time potential of data that is being generated each moment. However, before that, we need to resolve some legal issues like intellectual property rights, data privacy, cyber security, exploitation liability and Big Data code of conduct. Only if these threats related to privacy breach and personal freedom are mitigated, Big Data can be exploited in a transparent manner to bring out immense possibilities.

To conclude, the benefits seem to outweigh the risks. In the near future, such smart tools will be seen as our helpers to make better decisions and live a better life.

References


