Big Data Analytics

V Rajaraman

The volume and variety of data being generated using computers is doubling every two years. It is estimated that in 2015, 8 Zettabytes (Zetta=10^{21}) were generated which consisted mostly of unstructured data such as emails, blogs, Twitter, Facebook posts, images, and videos. This is called big data. It is possible to analyse such huge data collections with clusters of thousands of inexpensive computers to discover patterns in the data that have many applications. But analysing massive amounts of data available in the Internet has the potential of impinging on our privacy. Inappropriate analysis of big data can lead to misleading conclusions. In this article, we explain what is big data, how it is analysed, and give some case studies illustrating the potentials and pitfalls of big data analytics.

V Rajaraman is at the Indian Institute of Science, Bengaluru. Several generations of scientists and engineers in India have learnt computer science using his lucidly written textbooks on programming and computer fundamentals. His current research interests are parallel computing and history of computing.

Introduction

I got an email from my brother-in-law, who was to visit me, giving his travel schedule with flight details. He was staying with me for a few days. On the day of his departure, I got a message from Google on my smart phone that he should be leaving my home at 8.30 pm as his plane was delayed by 55 minutes and it will take 41 minutes for him to reach the airport from my home as no traffic snarls were reported. I was shocked. How did Google know that my brother-in-law was staying with me, which flight he was taking, where my home was, traffic conditions, and flight information? I then realized that Google was performing big data analysis on the data it had gathered from the emails we exchanged about the flights my brother-in-law was taking, informing that he will be staying with me, along with GPS (Global Positioning System) data and Google map co-ordinates to find the location of my smart phone which it guessed was my home, and real-time data from the airline's database on the flight's delay. On the one hand, it was nice to be told when my brother-in-law should leave

Keywords
Big data, data science, fourth paradigm, MapReduce, Hadoop.

our house and not waste time at the airport. On the other hand, I was concerned that I had lost my privacy. Google was reading all my emails and knew where I lived. This is the power of big data analytics. It is a double-edged sword.

There are other situations that you would have encountered if you shop using the websites of e-commerce companies such as Flipkart or Amazon. The web portals remember what items you bought during the past few months, and even suggest that you buy a knife sharpener (with picture and prices) when you place an order for a knife. If you order a book it will give you a list of other books you would probably like to buy as some other buyers with taste similar to yours have bought.

These are two typical examples of the power of big data analytics. By now you will be wondering what big data analytics is all about and how does it do what seems like magic. The purpose of this article is to explain the technology behind big data analytics and the pros and cons of using this tool.

What is Big Data?

There are several definitions of big data which look at it from different points of view. We give some of them here:

Mills *et al* [1] define big data as:

Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

National Institute of Standards and Technology, USA [2] defines big data as:

Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant scaling (more nodes) for efficient processing.

Ward and Barker [3] after surveying twelve definitions by Gartner, Microsoft, Oracle, Intel etc., distil them and define big data as:

Big data is a term describing the storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce, and machine learning.

(see *Box* 1 for an explanation of NoSQL. We discuss MapReduce later in this article. A *Glossary* on p.715 of the article provides quick reference to new terminology used in this article)

An analysis of these definitions shows that all of them emphasize the massiveness of the data which the term implies, its variety and rate of change and that the data need to be analysed to gain insight into what it implies. In addition, the analysis needs new tools and massive computing resources.

All definitions of big data emphasize the massiveness of the data which the term implies, its variety and rate of change and that the data need to be analysed to gain insight into what it implies. In addition, the analysis needs new tools and massive computing resources.

Data Science, of which big data analytics is an important component, has ushered in the fourth paradigm[1] [4] in science. Two earlier paradigms were theoretical science and experimental science. This was followed in the last two decades by a third paradigm, simulation. The advent of high performance computer simulations reduced the number of experiments conducted to authenticate a hypothesis and led to the field of computational science. The latest, namely, the fourth paradigm has been enabled by Data Science. The fourth paradigm is data-driven

[1] Paradigm refers to a philosophical/theoretical framework.

Box 1. What is NoSQL?

Traditionally data bases of computers were used primarily to store data appropriate for business data processing. From the late 70's data was organized as tables with each row containing a key and attributes of a data item e.g., item code as key, with item name, quantity in stock, price per item etc. This was called Relational Data Base Management System (RDBMS). A query language called SQL (Structured Query Language) was standardised and widely used to retrieve relevant data from relational data bases. With the advent of new variety of data such as documents, images, audio etc., in which data grows and shrinks dynamically, the organization and retrieval of data required a different approach. This led to the development of NoSQL (for Not SQL) which are different database storage structures (non-relational) with a more flexible querying to retrieve data. Some examples of NoSQL database organizations are Mongo DB and Cassandra. (See SQL Vs NoSQL – What you need to know – www.dataconomy.com/sql-vs-nosql-need-know/)

Box 2. Data Science and Data Scientist

Data Science is a field of study concerned with the collection, cleaning, and anonymising (if required) large quantities of data of diverse variety relevant for solving real-life problems and analysing them to initiate meaningful actions. A data scientist normally works as a member of an inter-disciplinary group and builds models and develops algorithms to extract meaningful relations or inferences which would normally be used to build a data product. A data scientist should know how to use the tools of statistics, machine learning, software engineering, visualisation, and should have an education in a discipline such as life sciences, physical sciences, chemical sciences, social sciences, or engineering. A data scientist requires skills to communicate in clear language and use visualization tools, when necessary, to convey to the team the implications derived from data to enable data-driven decision making*.

* C O'Neil and R Schutt, *Doing Data Science*, O'Reilly Media, 2013.

discovery of new hypotheses which may lead to the formulation of novel theories. See *Box* 2 for elaboration of Data Science and Data Scientist.

Returning to our example, the recommendation by Amazon that I would be interested in certain types of books based on my past purchases and also the buying pattern of persons with "similar taste" was enabled by analysing massive amount of data which Amazon's computers had gathered from millions of customers. This is a typical recommendation system. The recommendation that my brother-in-law should leave for the airport at a specified time was also based on the analysis of a variety of data, namely emails, GPS data, Google maps, and airline's database collected over an extended period. Discovering patterns in massive data of diverse varieties to formulate new theories is the essence of the fourth paradigm. One has to be cautious in relying entirely on big data analysis to formulate theories, as we will see later in this article.

The fourth paradigm is data-driven discovery of new hypotheses which may lead to the formulation of novel theories.

5 'V's of Big Data

The term big data emphasises volume or size. Size is a relative term. In the 1960s, 20 Megabytes was considered large. Now data is not considered big unless it is several hundred Petabytes (Peta = 10^{15}). Size is not the only property used to describe big data. In

addition to volume which we expand on, there are other important properties that we will discuss in what follows:

Size is not the only property used to describe big data. In addition, other properties are Volume, Variety, Velocity, Veracity and Value – called 5 'V's of big data.

***Volume*:** Amount of global digital data created, replicated, and consumed in 2013 was estimated by the International Data Corporation (a company which publishes research reports) as 4.4 Zettabytes (ZB) (Zetta = 10^{21}). It is doubling every 2 years. By 2015, digital data grew to 8 ZB and is expected to grow to 12 ZB in 2016. To give an idea of a ZB, it is the storage required to store 200 billion high definition movies which will take a person 24 million years to watch! In *Box* 3 we tabulate some of the big data sources and the volume of data generated.

***Variety*:** In the 1960s, the predominant data types were numbers and text. Today, in addition to numbers and text, there are image, audio, and video data. Large Hadron Collider (LHC), earth and polar observations generate mainly numeric data. Word processors, emails, tweets, blogs, and other social media generate primarily unstructured textual data. Medical images and billions of photographs which people take using their mobile phones are image data. Surveillance cameras and movies produce video data. Music sites store audio data. Most data in the 80s were structured and organized as tables with keys. Today they are unstructured and multimedia data often used together.

***Velocity*:** Data in conventional databases used to change slowly. Now most data are real time. For example, phone conversations, data acquired from experiments, data sent by sensors, data exchanged using the Internet, and stock price data are all real time. Large amounts of data are transient and need to be analysed as and when they are generated. They become irrelevant fast.

***Veracity*:** A lot of data generated are noisy, e.g., data from sensors. Data are often incorrect. For example, many websites you access may not have the correct information. It is difficult to be absolutely certain about the veracity of big data.

***Value*:** Data by itself is of no value unless it is processed to obtain information using which one may initiate actions. The large

Box 3. Big Data Sources and Volume of Data Generated per Minute*

Sources of numeric data	**Amount of data generated**	
Large Hadron Collider	1.5TB/minute	
Sensor data in all flights	5 PB/minute	
Sources of unstructured data		
Emails sent/received	300 GB/minute	(205 billion mails/day)
Websites created	600/minute	(1 billion - Total)
Google searches	4 million/minute	(4 PB data processed/day)
Tweets	35000/minute	(465 million accessed)
App downloads(Apple)	30000/minute	
Facebook	4×10^6 posts/minute	
Sources of image data		
Photos loaded/shared	1.25 million/minute	
Medical images	131 GB/minute	
Sources of video data		
Skype	110000 calls/minute	
YouTube	300 hours of material uploaded/minute	
Netflix movies seen	7700 hours of video/minute	
Sources of transactions data		
Walmart	16700 transactions/minute (2.5PB in database)	
Amazon (USA)	4300 visitors/minute	

Estimate : 8 ZB data generated and consumed in 2015. Internet user population estimated as 3.2 billion in 2015.

* Marcia Conner, *Big Data Statistics*, www.marciaconner.com/blog/data-on-big-data/

A huge number of inexpensive processors working in parallel has made it feasible to extract useful information to detect patterns from big data.

volume of data makes processing difficult. Fortunately, computing power and storage capacity have also increased enormously. A huge number of inexpensive processors working in parallel has made it feasible to extract useful information to detect patterns from big data. Distributed file systems such as Hadoop Distributed File System (HDFS) coupled with parallel processing programs such as MapReduce are associated with big data as software tools to derive value from big data. (We discuss Hadoop and MapReduce later in this article.)

Analysing Big Data

Data analytics is concerned with extraction of actionable knowledge and insights from big data. This is done by hypothesis formulation that is often based on conjectures gathered from experience and discovering (sometimes serendipitously) correlations among variables. There are four types of data analytics. They are:

Data analytics is concerned with extraction of actionable knowledge and insights from big data.

***Descriptive Analytics*:** This essentially tells what happened in the past and presents it in an easily understandable form. Data gathered is organized as bar charts, graphs, pie charts, maps, scatter diagrams, etc., for easy visualization which gives insight into what the data implies. This form of data presentation is often called a dash board, mimicking the dashboard of a car which gives information on speed, engine status, petrol left in the tank, distance travelled etc. A typical example of descriptive analytics is presentation of population census data which classifies population across a country by sex, age groups, education, income, population density and similar parameters.

***Predictive Analytics*:** It extrapolates from available data and tells what is expected to happen in the near future. The tools used for extrapolation are time series analysis using statistical methods, neural networks, and machine learning algorithms. One major use of predictive analytics is in marketing by comprehending customers' needs and preferences. An example is the advertisement on socks that appears when you buy shoes from an e-shop. Another common use is in managing election campaigns by collecting variety of data such as composition of electorates in various locations, perception of their requirements such as infrastructure and local issues.

Predictive Analytics extrapolates from available data and tells what is expected to happen in the near future.

***Exploratory or Discovery Analytics*:** This finds unexpected relationships among parameters in collections of big data. Collection of data from a variety of sources and analyzing them provides additional opportunities for insights and serendipitous discovery. One of the major applications is discovering patterns in

Discovery Analytics finds unexpected relationships among parameters in collections of big data.

Prescriptive analytics identifies opportunities to optimize solutions to existing problems, based on data gathered.

customers' behaviour by companies using their feedback, tweets, blogs. Facebook data, emails, sales trends etc. Based on the customers' behaviour it may be possible for companies to foretell their actions such as renewing magazine subscription, changing mobile phone service provider, cancelling a hotel reservation. A company may then come up with an attractive offer to try and change the customer's anticipated action.

***Prescriptive analytics*:** This identifies, based on data gathered, opportunities to optimize solutions to existing problems. In other words, the analysis tells us what to do to achieve a goal. One of the common uses is in airlines' pricing of seats based on historical data of travel patterns, popular origins and destinations, major events, holidays, etc., to maximize profit.

Examples of Big Data Analytics

In this section, nine examples of use of big data analytics in a variety of applications are outlined:

1. *Predicting whether a wait-listed train ticket will be confirmed*: India has one of the largest railway networks in the world. It runs around 8500 trains each day and there are around 7000 railway stations of which about 300 are major stations. 20 million passengers travel on a given day. The total number of reserved seats/berths issued every day is around 250,000 and reservation can be made 60 days in advance. The allocation of reserved seats is very complicated. There are quotas for VIPs, ladies, emergency travel, handicapped persons etc. There are also quotas for starting and intermediate stations. Besides these, there are 'Tatkal' berths available at a higher cost for last minute reservations. A person reserving berths/seats well ahead of time would normally get confirmed berths/seats. Else, he or she would get a RAC (Reservation against Cancellation) or a wait-listed ticket. A passenger with a wait-listed ticket would like to know the probability of getting the ticket confirmed.

The problem of predicting this is very complicated as it depends

on several factors such as weekends, festivals, night train, starting or intermediate station, etc. A company in the travel business gathered data on 10 million reservations on various trains over a period of 2 years. The data gathered was on wait-listed tickets and those which got confirmed. Using the data and complex analytics, which took into account all the constraints noted above, the company was able to predict with 90 to 95% accuracy the probability that a wait-listed ticket with a given waiting list number in a given class of travel on a specified train from a given starting station on a given date would lead to a confirmed reservation or not. The algorithm required machine learning based on past big data collection [5].

A company in the travel business gathered data on 10 million reservations on various trains over a period of 2 years on wait-listed and confirmed tickets. Using the data and complex analytics, the company was able to predict with 90 to 95% accuracy the probability that a wait-listed ticket would lead to a confirmed reservation or not.

***2. Success in matching men and women in a matrimonial site*:** With the advent of the Internet, the method of finding potential marriage partners has undergone a big change. There are over a dozen matrimonial sites in India such as Shaadi.com and Matrimony.com. The competition among them is intense. It is predicted that by 2017, the turnover of matrimony sites will exceed Rs.150 billion. Matrimony.com is one of the largest sites. Around 8000 applicants register on this website every day. Nearly 1500 successful matches resulting in marriages are reported every day to the site. Prospective applicants are from various countries in the world and Non-Resident Indians from 150 countries register on this site. The major goal of this marriage site is to assist applicants to properly enter their profiles. The company employs over 2000 enquiry assistants who attend to 300,000 telephone calls, SMSs, and emails every day from the applicants. The customer support team visits applicants to verify their credentials. Data is gathered on pairs of applicants who corresponded.

The major goal of this marriage site is to assist applicants to properly enter their profiles to maximise matchmaking success by modifying the profile form of applicants with mandatory and optional fields.

The major aim of collecting all the data and analysing them is to maximise matchmaking success by modifying the profile form of applicants with mandatory and optional fields. Analysing the data it has gathered, the company has found that women are more interested in hobbies of their prospective partners rather than their aspirations. Another surprising fact that has emerged is that

The main advantage of almost real-time monitoring of consumer reactions is the ability to take a quick decision leading to better sales and profit.

the women prefer non-smokers, smokers, and social smokers in that order rather than non-smokers, social smokers, and smokers. By using big data analytics, the company expects to increase their matchmaking success by ten to fifteen percent in the near future.

3. *Introducing a new coffee type*: A coffee shop with a number of branches located all over the country was introducing a new variety of coffee and wanted to know the reaction of its consumers. As soon as they introduced the new coffee variety, they started monitoring the social media: twitter, blogs, and coffee drinkers' groups for the reaction of customers. By noon they discovered that most people who tasted the new variety liked it but were unhappy about the high price. The shops were asked to lower the price and within hours they found that the negative perception (monitored on Twitter etc) almost disappeared. The main advantage of almost real-time monitoring of consumer reactions is the ability to take a quick decision leading to better sales and profit.

4. *Use of big data and neural networks in playing GO, a board game* [6]: GO[2], a board game, is 2500 years old (probably invented in China) and is popular in East Asia. GO is played by two players who compete by moving white and black polished stones on a 19 × 19 grid. The player who controls the maximum territory on the board at the end of the game is declared the winner. There are 250 possible moves in a given board position. A new position has another 250 possible moves. It has been pointed out that there are more possible moves in a GO game than atoms in the Universe. Beating a human at GO was one of the most challenging problems in Artificial Intelligence (AI) as it is not possible to analyse each possible move in a reasonable time even by the fastest supercomputer. The strategy used by a computer to play chess is largely brute force as in chess at any given board position there are only 35 possible moves and the consequence of executing all these moves can be analysed in a reasonable time by a fast computer such as IBM's Deep Blue supercomputer[3]. But a similar brute force will not work in playing GO. A clever approach to learn the strategy of moving polished

[2] See www.kiseido.com to find out how to play GO.

[3] K S R Anjaneyulu, *Resonance,* Vol.2, No.7, pp.89–90, 1997.

Beating a human at GO was one of the most challenging problems in Artificial Intelligence (AI) as it is not possible to analyse each possible move in a reasonable time even by the fastest supercomputer.

stone at any given board position to capture maximum territory in the long run is needed. Many AI researchers felt that it was impossible for a computer to beat a human champion.

In March 2016, a computer program called AlphaGO defeated the GO world champion Lee Sedol by 4 games to 1. AlphaGO was developed by Deep Mind, a British company specialising in AI software that was acquired by Google. AlphaGO used a technique in AI called 'deep learning' which relies on neural networks. A neural network consists of a network of simulated web of neurons as in a human brain [7]. Neural networks analyze huge amounts of data and attempt to learn a particular task. For example, if enough photographs of automobiles of different shapes and sizes are shown to a neural network, it adjusts itself to learn to recognize automobiles. With increasing speed of computers, learning using neural network has now become a standard tool of AI. Around 30 million moves were used to train AlphaGO's neural network to play GO. This was the initial step. To hone the skill of AlphaGO the developers of the software made AlphaGO play with itself. This led to a vast collection of moves which could be used to train AlphaGO further. The most significant aspect of AlphaGO program is that unlike earlier game playing programs it uses big data analytics. The data from millions of GO games are used to assist machine learning techniques to beat the world champion.

In March 2016, a computer program called AlphaGO defeated the GO world champion Lee Sedol by 4 games to 1 using a technique in AI called 'deep learning' which relies on neural networks.

5. ***Transportation and logistics*****:** Public road transport companies are using RFID (Radiofrequency Identification) and GPS to track buses and arrive at realistic schedules. Data on number of passengers using the buses in different routes are used to optimize bus routes and the frequency of trips. A real-time system has been implemented to facilitate a passenger to get information on when to expect the next bus which will take him/ her to the desired destination.

6. ***Elections*****:** Political parties in India mine data on voter sentiments in all constituencies. This is done by combining social network communications, personal interviews, and voter

Political parties in India mine data on voter sentiments in all constituencies by combining social network communications, personal interviews, and voter compositions to detect local issues in addition to national issues.

compositions to detect local issues in addition to national issues. These affect the decisions taken by voters to vote for a candidate belonging to a political party. This data is used to campaign in diverse constituencies besides raising funds. Information Technology professionals are being employed by political parties in both state and national elections.

7. *Public utilities*: Utilities such as water supply organizations are placing sensors in the pipelines to monitor flow of water in the complex water supply networks. It is reported in the Press that Bangalore Water Supply and Sewage Board is implementing a real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city, reducing the need for valve operators.

8. *Computer/mobile phone based game developers*: Reliance Communications in India employs over 600 game developers and creates games for over 2000 different types of devices. The company collects data on active users of each game every day, repeat users, those who drop out, and user comments about games. Over 30 million records are collected each day. Based on this data, they improve popular games and withdraw games that are less played.

9. *Aviation industry*: An Airbus A350 plane uses 6000 sensors across the plane. There are sensors in the wings, landing gear, engine, air conditioning system and other critical areas. Each day around 300 GB data is collected and analysed primarily to optimize engine performance to minimize fuel consumption. Another important application is predictive analytics is, as to when a critical component may fail (based on past performance data) and replacing it. Another important analytics is to decide how many critical components are to be stocked. It is reported that there is a lead time of over 3 months to obtain some critical spare parts. The need for big data analytics in planes is bound to increase. The next generation Airbus A380 will carry 1000 passengers. It will have 10, 000 sensors in each wing giving real

Each day, around 300 GB of data is collected by Airbus to optimize engine performance and to minimize fuel consumption.

time data. In addition, there will be sensors in engines and other subsystems. Smarter aircrafts will be analysing big data in real time.

Hardware and Software Systems for Analysing Big Data

We saw in the last section that an enormous amount of data from various sources are needed to solve many interesting problems. Analysing big data is essential to recognize patterns in data collections and based on that initiate meaningful actions. The challenge in analysing big data is the sheer size of data, variety, and the need quite often to analyse the data in real time. Fortunately, the processing power of computers is doubling every two years and the storage capacity too is doubling every 15 months, both without any increase in cost. Parallel processing of data using thousands of inexpensive commodity processors (e.g., those used in desktop PCs) is eminently suitable for analysing big data. Chunks of big data can be processed independently and simultaneously by a large number of computers and the results combined to solve many problems. We will illustrate this later in this section with an example.

Parallel processing of data using thousands of inexpensive commodity processors is eminently suitable for analysing big data. Chunks of big data can be processed independently and simultaneously by a large number of computers and the results combined to solve many problems.

Hadoop System

Hadoop is the name (see *Box* 4) of a software system developed to process big data. It is open source software written in a high level programming language (Java) and the source code is freely available for download from the Internet. The development of Hadoop was funded by Apache Software Foundation.

Box 4. Origin of the Name Hadoop

A system for big data analysis was developed by Doug Cutting and Mike Cafarella and released in 2006. When the time came to name it, Doug Cutting remembered that his young son had named his cuddly toy elephant Hadoop. The toy was flexible and large and Doug Cutting jocularly suggested that the software system be named Hadoop as it is easy to pronounce and could be used as a unique search term in Google. Mike Cafarella readily agreed.

Figure A. Picture of soft toy elephants named Hadoop.

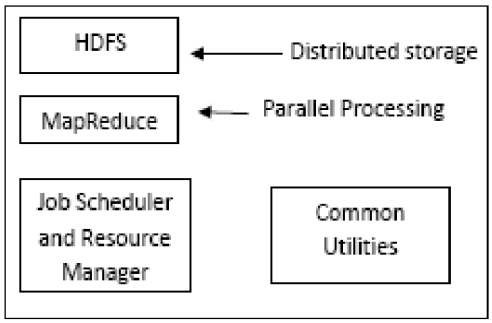

Figure 1. A simplified block diagram of Hadoop.

Hadoop system [8] is designed to be used with large clusters (of the order of thousands) of commodity processors each with both local main memory and large secondary memory (normally disks). It consists of two major parts: one part for storing files to be processed and the other to process in parallel the data stored in these files. They are called Hadoop Distributed File System (HDFS) and MapReduce respectively. HDFS is modelled after Google file system and MapReduce was also originally developed by Google as a proprietary system. The idea came from Map and Reduce functions used in a very old programming language called LISP[4].

Two facts which were important considerations during the design were:

i) When thousands of low-cost commodity processors and disks are used, the probability of one or more processors/disks failing is high. Thus the design has to be fault tolerant.

ii) If the system has to be widely used, the complexity of how faults are handled by the system should be hidden from users. In other words, a user should assume that the system is fault free. In *Figure* 1 we give a simple block diagram of the Hadoop System.

Yahoo uses over 42,000 computers each with a local disk in its Hadoop installation. This gives an idea of the size of Hadoop systems.

***Hadoop Distributed File System* (*HDFS*):** HDFS is designed to store large files in which data is written once and read many times. Files are split and stored as 64 bytes or 128 bytes blocks. Three sets of blocks of the same file are stored in three different computers (which may be in different racks), to tolerate faults. The HDFS system works in a master–slave mode. The master process is executed in a node (i.e., a computer) called Name node. The Name node maintains information about files such as file names, where they are stored, their attributes, etc., so that it has a global view of the file system. The Name node is vital as the whole

[4] V Rajaraman, John McCarthy – Father of Artificial Intelligence *Resonance*, Vol.19, No.3, pp.198– 207, 2014.
Harish Karnick, LISP, *Resonance*, Vol.19, No.3, pp. 208–221, 2014.

system will stop functioning if it fails. The slave nodes, called Data nodes, obey commands sent by the Name node on data blocks stored locally. HDFS is used to store both input and output of MapReduce jobs, intermediate data if any (e.g., output from Map tasks) that are stored in local disks of slave nodes. In *Figure* 2 we give a picture of a rack with a number of computers. Each box in the figure contains one or more computers with local disks. These boxes are assembled in a rack as seen in the figure. A large data centre such as the ones maintained by Amazon or Google will have hundreds of such racks containing over 100,000 computers.

Figure 2. A rack in which many computers are assembled.
Courtesy:
https://commons.wikimedia.org/wiki/File:Datacenter Rack.jpg

***MapReduce*:** MapReduce processes data stored in HDFS in parallel. We illustrate how MapReduce works with a simple example. The problem we consider is to analyse six emails. The task is to find the number of occurrences of the words Ram, Delhi, Anu, and Chennai in these emails. The first task is to split the six emails into three sets of two emails each (see *Figure* 3). After splitting, a mapping function counts the occurrences of the four words Ram, Delhi, Anu, and Chennai (called keywords) in the three split sets of records. Three Map functions are used in this example, one corresponding to each split. The Map function counts the number of occurrences of the keywords in each split file. The Map function feeds shuffle-sort (*Figure* 3) which gathers key words Ram, Delhi, Chennai and Anu to facilitate 'Reduce function' to count the number of occurrences of the specified keywords. In our example, we show keywords in sorted order in the shuffle box to enable Reduce to just count the number of occurrences of the specified keywords.

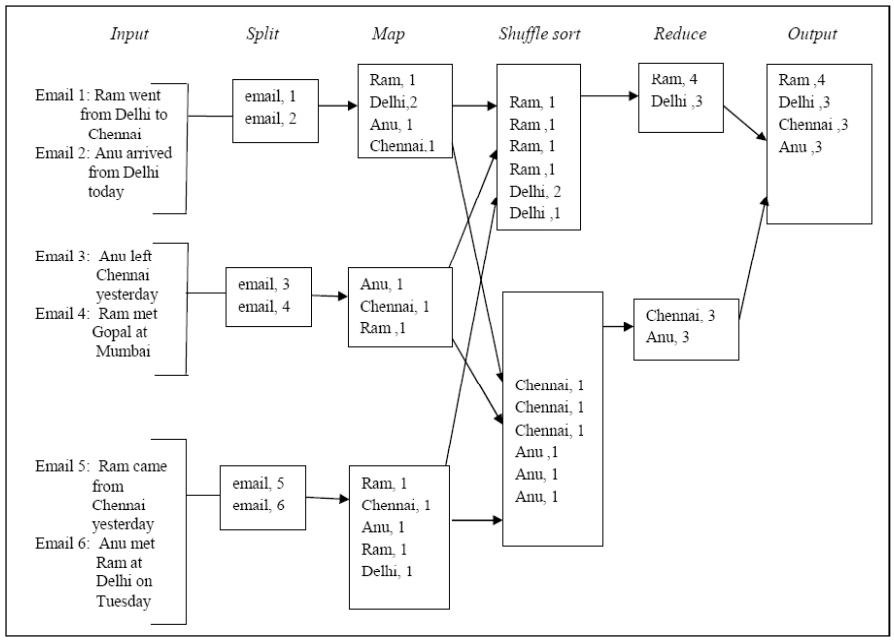

Figure 3. MapReduce to count specified words in emails.

We have used three Map tasks and two Reduce tasks in the example. Observe that all three Map tasks can be run in parallel. The two Reduce tasks can also be run in parallel. Reduce has to follow Map. The choice of number of Map tasks and Reduce tasks is done automatically by the Hadoop system to balance the work load of processors in the cluster.

In the example, we took just six emails and four keywords to search which enabled us to illustrate the basic idea used in MapReduce with a small figure. In actual practice, millions of emails are searched by using MapReduce to select hundreds of keywords.

The choice of number of Map tasks and Reduce tasks is done automatically by the Hadoop system to balance the work load of processors in the cluster.

Implementation of MapReduce in Hadoop [9]

There are two types of processes which control users' programs in Hadoop. One is called a Job tracker that runs on the master node and the other is called a Task tracker that runs on slave nodes. (The term node is used for a computer in a cluster). The Job tracker splits a job submitted to the system into Map tasks and Reduce tasks. These are scheduled to Task trackers. The scheduler tries to assign Map tasks close to the node where the

requisite data is located. The Job tracker monitors the tasks that run on slave nodes. When a Task tracker has no work to do, the Job tracker picks a new task from the queue and assigns it to that Task tracker. When a Reduce task is to be performed it is assigned to a Task tracker which has completed the appropriate Map task. The Job tracker waits for a response from a Task tracker (called a heart-beat signal) before any more tasks are assigned to it. Due to the fact that inexpensive nodes are used as Task trackers in the cluster, there is high probability that one of them would fail. Thus, there must be a mechanism to handle a failed Task tracker. The system does it by transferring a task assigned to a failed Task tracker to a healthy node. The fact that a node has failed is detected by the Job tracker if it does not receive a task completion signal from the Task tracker to which it had assigned the task. If a Task tracker, for some reason, performs an assigned task slowly, the run time system of Hadoop sends the same task to another node (called a Backup task). The Job tracker takes the results from the Task tracker which finishes first and aborts the other. The complexity of fault handling is hidden from the user.

The fact that a node has failed is detected by the Job tracker if it does not receive a task completion signal from the Task tracker to which it had assigned the task.

Citizens and Big Data Analytics

At the beginning of this article we mentioned that big data analytics is like a double-edged sword. While it does provide useful information to citizens it also leads to loss of privacy. One can, of course, use privacy settings or encrypt all emails but many users do not know how to do it. Vast amounts of medical data of individuals may be analysed and may lead to better diagnosis and treatment if properly used but has the potential of irresponsible use. There is no legislation in India which protects privacy of individuals who use social media extensively. Email, Twitter messages, and Facebook data would allow unscrupulous organizations to 'profile' individuals. There has been extensive debate on this point. Many countries have privacy laws which require anonymizing before citizens' data is put in the public domain.

Vast amounts of medical data of individuals may be analysed and may lead to better diagnosis and treatment if properly used but has the potential of irresponsible use.

A simple anonymizing technique used in medical databases is to

There are also large quantities of data on citizens as well as other entities collected by governments. It is argued that if free access to such data is made available (after suitable anonymization even though it may not be foolproof) it could help social scientists and other researchers to suggest improved governance.

remove the name of a patient and substitute it with a number. More complex anonymization methods have been used. However, it is reported in the literature [10] that no anonymization tool is foolproof. A determined hacker who has access to publicly available data can, by combining many such data bases and using educated guesses, construct the details of a particular person's private confidential data. Does it mean that medical data of populations, for example, should not be put in a publicly available database? This would be counterproductive as analysis of large quantities of medical data could potentially lead to better health care to citizens. There are also large quantities of other data on citizens as well as other entities collected by governments. It is argued that if free access to such data is made available (after suitable anonymization even though it may not be foolproof) it could help social scientists and other researchers to suggest improved governance. There are also movements all over the world on the rights of citizens to have access to data collected using their tax money. Transparency leads to better governance. Open data initiatives have appeared all over the world. India's Open Governance Initiative may be found from the URL https://data.gov.in/about-us. There are also committees promoting easy access to data of relevance to citizens (http://datameet.org). With the availability of large amounts of data and ease of processing, big data analytics will have a very beneficial impact on our daily lives provided analysis is carried out with care.

It is dangerous to implicitly trust the results obtained from big data analytics. Often it is assumed that big data collected include *all* data points relevant to the problem. This is far from true as data collected can seldom be *all* data.

Big Data – Need to be Cautious

It is dangerous to implicitly trust the results obtained from big data analytics. Often it is assumed that big data collected include *all* data points relevant to the problem. Many believe that with *all* data the "numbers speak for themselves" and that correlations among data tell all we need to know about systems. This is far from true as data collected can seldom be *all* data. We use a few examples to illustrate this.

In Boston, the Mayor of the city wanted to use big data analytics

to find the potholes in the roads which need to be filled. An application was developed which used data from the accelerometer in the smart phone of a road user to detect bumps in a road and automatically sent a message to a command centre giving the location coordinates where the car bumped due to a pothole. Messages were sent to all mobile phone owners in Boston area and they were requested to download the free app on their smart phones to assist the city engineers in identifying potholes to enable them to be promptly filled. The city's Mayor announced that they found an inexpensive way using data analytics to fill potholes. As it turned out only the young affluent citizens with smart phones downloaded the app and the potholes in affluent areas of Boston were detected and filled. From poorer areas (which probably had more potholes) only few messages came as most of the population in these areas did not have smart phones and even among those who had them, many did not download the app with the result that most potholes remained.

Another example is the programs used to predict the results of elections. The forecasts are often way off the mark due to poor sampling. It is impossible to collect data from *all* voters after they vote to get an accurate prediction. It is essential to use good statistical analysis methods developed over many years to pick diverse samples to get near accurate prediction.There is also a tendency to correlate often unrelated data from collections of big data to predict outcomes of games. A recent example is the publication of tables in many newspapers predicting the outcome of World Cup T20 cricket matches by correlating performance of players, scores, toss winning, etc. of two teams which were often laughable.

Programs used to predict the results of elections are often way off the mark due to poor sampling.

Google Flu Trends Program

Many of us use Google search engine to obtain information useful in our daily lives. For example, when we fall ill we use Google to find information about symptoms, medicines, side effect etc. We often use Google for available hotels when we plan a trip. Google surmised that their users' searches reflect trends in illness, travel,

The methodology used by Google was a spectacular demonstration of the power of big data analytics in prediction until it miserably failed in 2013.

fashions, product sales, and may even be able to anticipate the results of elections. In 2008 Google's analysts explored whether detecting frequently used search terms could forecast the spread of flu in USA. Google's assumption was that when persons fall ill with symptoms of flu they would use Google's search engine and type terms relevant to flu symptoms. This would give Google an immediate signal on the prevalence of flu in a specified geographical area of USA. When Google started its investigation, the Centre for Disease Control and Prevention (CDC), a department of the US Government, was tracking the occurrence of flu in different areas by collecting data from doctors, hospitals, and pharmacies. Data collection is time consuming and thus there was a delay in detecting the spread of flu.

Google published a paper in the journal *Nature* demonstrating that by examining search terms used by people in Google searches and tuning them with the flu tracking information provided by CDC it could provide an accurate estimate of the prevalence of flu in different areas of USA two weeks ahead of CDC. Google's flu trends program was used all over the world by Google giving not only flu trends but also predicting the spread of dengue fever. The methodology used by Google was a spectacular demonstration of the power of big data analytics in prediction until it miserably failed in 2013. In 2013, Google's Flu Trends over predicted the prevalence of flu by 140%. The possible reason may be the appearance of several scary stories in the media about a flu epidemic which provoked many healthy persons to search for flu symptoms using Google which misled Google's algorithm. If Google had correlated flu symptoms with search for sale of medicines in pharmacies and prescriptions given by doctors, the incorrect prediction may not have happened. In 2015, Google stopped giving flu trends information. An analysis on why Google's Flu Trends failed may be found in an article published in the journal *Science* [11].

The availability of huge amount of data and the ease of analysing them with software tools and large cluster of computers has numerous uses and is bound to change the way we solve many problems. However, big data analytics can lead to misuse unless citizens are vigilant.

Conclusions

The availability of huge amount of data and the ease of analysing

them with software tools and large cluster of computers has numerous uses and is bound to change the way we solve many problems. This has been amply demonstrated with some examples in this article. However, big data analytics can lead to misuse unless citizens are vigilant. Blind reliance on big data analytics without clear thinking on the relevance of the data, veracity of data, and meaningful correlations of variables may lead to incorrect and misleading results. Once should not ignore years of research done by statisticians on picking unbiased samples, error estimates, and testing hypotheses with data.

GLOSSARY

Big Data: It is a very large collection of structured, semi-structured, and unstructured data that may change rapidly. It is analysed using large clusters of computers to reveal patterns, trends, and correlations which may be used for making informed decisions. (See text for three formal definitions).

5 Vs of Big Data: Five important properties of big data. They are huge Volume, large Variety, high Velocity, acceptable Veracity, i.e., trustworthy, and good Value for decision making when appropriately processed.

Data Science: See *Box* 2 on page 698.

Fourth Paradigm: Data driven discovery of new hypotheses which may lead to the formulation of novel theories

Hadoop: It is an Open-Source, Java-based programming framework sponsored by Apache software foundation for storing large data sets in a distributed computing environment. It runs applications on clusters of thousands of commodity hardware and has the ability to handle virtually limitless concurrent tasks.

HDFS: Hadoop Distributed File System is a highly fault-tolerant distributed file system designed for very large data sets to be deployed on huge clusters of low-cost computers. It provides high throughput access to application data to programs running on clusters.

Machine Learning: It is a branch of Artificial Intelligence (AI). By providing large data sets to computers we endow them with the ability to learn without explicitly programming them. The learning is adaptive and adapts when new data is provided. In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed".

MapReduce: It is a parallel programming model in which data to be processed is split into a number of sets which are distributed to individual computers in a large cluster of computers. Each computer processes the data assigned to it simultaneously and independently using identical programs. The results are then gathered by a master processor.

NoSQL: See *Box* 1 on page 697 for definition.

Acknowledgements

I thank Dr S Ramani, Dr Radha Balakrishnan and the anonymous referee of this article for their critical comments and suggestions which improved it.

Suggested Reading

[1] H J Watson, Tutorial: Big data analytics: Concepts, technologies, and applications, *Communication of the Association for Information Systems,* Vol.34, Article 65, pp.124–168, 2014.

[2] NIST definition of big data and data science, www.101.datascience.community/2015/nist-defines-big-data-and-data-science.

[3] J S Ward and A Baker, *A Survey of Big Data Definitions*, arxiv.org/abs/1309.5821 VI.

[4] T Hey, S Tansley and K Tolle (Editors), *The Fourth Paradigm – Data-Intensive Scientific Discovery,* Microsoft Research, Richmond, WA., USA, 2009.

[5] R P Srikant, 8 Innovative examples of big data usage in India, *Dataquest*, August 20, 2016.

[6] C Metz, In a huge breakthrough, Google's AI beats a top player at the game of GO, www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-GO-champion/

[7] M H Hassoun, *Fundamentals of Artificial Neural Networks,* Prentice-Hall of India, 1998.

[8] Big Data Hadoop Tutorial, www.tutorial point.com/hadoop/hadooptutorial.pdf

[9] V Rajaraman and C Siva Ram Murthy, *Parallel Computers: Architecture and Programming,* 2nd Edition, Chapter 6, PHI Learning, New Delhi, 2016.

[10] S Smith, Data privacy: Now you see me; New model of data sharing: Modern governance and statisticians, *Significance*, Vol.11, No.4, pp.10–17, Oct 2014. Available at URL:http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00762.x/full

[11] D Lazer, *et al*, The parable of Google Flu: Traps in big data analysis, *Science*, Vol.343, April 2014.

Address for Correspondence
V Rajaraman
Supercomputer Education and Research Centre
Indian Institute of Science
Bengaluru 560 012, India.
Email: rajaram@serc.iisc.in