Studies in Big Data 11

Hrushikesha Mohanty
Prachet Bhuyan
Deepak Chenthati *Editors*

Big Data

A Primer

Springer

Studies in Big Data

Volume 11

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence incl. neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/11970

Hrushikesha Mohanty · Prachet Bhuyan
Deepak Chenthati
Editors

Big Data

A Primer

Springer

Editors
Hrushikesha Mohanty
School of Computer and Information Sciences
University of Hyderabad
Hyderabad
India

Prachet Bhuyan
School of Computer Engineering
KIIT University
Bhubaneshwar, Odisha
India

Deepak Chenthati
Teradata India Private Limited
Hyderabad
India

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-81-322-2493-8 ISBN 978-81-322-2494-5 (eBook)
DOI 10.1007/978-81-322-2494-5

Library of Congress Control Number: 2015941117

Springer New Delhi Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer (India) Pvt. Ltd. is part of Springer Science+Business Media
(www.springer.com)

Preface

Rapid developments in communication and computing technologies have been the driving factors in the spread of the internet technology. This technology is able to scale up and reach out to more and more people. People at opposite sides of the globe are able to remain connected to each other because of the connectivity that the internet is able to provide now. Getting people together through the internet has become more realistic than getting them together physically at one place. This has led to the emergence of cyber society, a form of human society that we are heading for with great speed. As is expected, this has also affected different activities from education to entertainment, culture to commerce, goodness (ethics, spiritual) to governance. The internet has become a platform of all types of human interactions. Services of different domains, designed for different walks of people, are being provided via the internet. Success of these services decisively depends on understanding people and their behaviour over the internet. For example, people may like a particular kind of service due to many desired features the service has. Features could be quality of service like response time, average availability, trust and similar factors. So service providers would like to know of consumer preferences and requirements for designing a service, so as to get maximum returns on investment. On the other side, customers would require enough information to select the best service provider for their needs. Thus, decision-making is key to cyber society. And, informed decisions can only be made on the basis of good information, i.e. information that is both qualitatively and quantitatively sufficient for decision-making.

Fortunately for cyber society, through our presence on the internet, we generate enough data to garner a lot of meaningful information and patterns. This information is in the form of metaphorical due to footsteps or breadcrumbs that we leave on the internet through our various activities. For example, social networking services, e-businesses and search engines generate huge data sets every second of the day. And these data sets are not only voluminous but also in various forms such as picture, text and audio. This great quantum of data sets is collectively christened *big data* and is identified by its three special features *velocity*, *variety* and *volume*.

Collection and processing of big data are topics that have drawn considerable attention of concerned variety of people ranging from researchers to business makers. Developments in infrastructure such as grid and cloud technology have given a great impetus to big data services. Research in this area is focusing on big data as a service and infrastructure as a service. The former looks at developing algorithms for fast data access, processing as well as inferring pieces of information that remain hidden. To make all this happen, internet-based infrastructure must provide the backbone structures. It also needs an adaptable architecture that can be dynamically configured so that fast processing is possible by making use of optimal computing as well as storage resources. Thus, investigations on big data encompass many areas of research, including parallel and distributed computing, database management, software engineering, optimization and artificial intelligence. The rapid spread of the internet, several governments' decisions in making of smart cities and entrepreneurs' eagerness have invigorated the investigation on big data with intensity and speed. The efforts made in this book are directed towards the same purpose.

Goals of the Book

The goal of this book is to highlight the issues related to research and development in big data. For this purpose, the chapter authors are drawn from academia as well as industry. Some of the authors are actively engaged in the development of products and customized big data applications. A comprehensive view on six key issues is presented in this book. These issues are big data management, algorithms for distributed processing and mining patterns, management of security and privacy of big data, SLA for big data service and, finally, big data analytics encompassing several useful domains of applications. However, the issues included here are not completely exhaustive, but the coverage is enough to unfold the research as well as development promises the area holds for the future. Again for the purpose, the Introduction provides a survey with several important references. Interested readers are encouraged to take the lead following these references.

Intended Audience

This book promises to provide insights to readers having varied interest in big data. It covers an appreciable spread of the issues related to big data and every chapter intends to motivate readers to find the specialities and the challenges lie within. Of course, this is not a claim that each chapter deals an issue exhaustively. But, we sincerely hope that both conversant and novice readers will find this book equally interesting.

In addition to introducing the concepts involved, the authors have made attempts to provide a lead to realization of these concepts. With this aim, they have presented algorithms, frameworks and illustrations that provide enough hints towards system realization. For emphasizing growing trends on big data application, the book includes a chapter which discusses such systems available on the public domain. Thus, we hope this book is useful for undergraduate students and professionals looking for an introduction to big data. For graduate students intending to take up research in this upcoming area, the chapters with advanced information will also be useful.

Organization of the Book

This book has seven chapters. Chapter "Big Data: An Introduction" provides a broad review of the issues related to big data. Readers new to this area are encouraged to read this chapter first before reading other chapters. However, each chapter is independent and self-complete with respect to the theme it addresses.

Chapter "Big Data Architecture" lays out a universal data architecture for reasoning with all forms of data. Fundamental to big data analysis is big data management. The ability to collect, store and make available for analysis the data in their native forms is a key enabler for the science of analysing data. This chapter discusses an iterative strategy for data acquisition, analysis and visualization.

Big data processing is a major challenge to deal with voluminous data and demanding processing time. It also requires dealing with distributed storage as data could be spread across different locations. Chapter "Big Data Processing Algorithms" takes up these challenges. After surveying solutions to these problems, the chapter introduces some algorithms comprising random walks, distributed hash tables, streaming, bulk synchronous processing and MapReduce paradigms. These algorithms emphasize the usages of techniques, such as bringing application to data location, peer-to-peer communications and synchronization, for increased performance of big data applications. Particularly, the chapter illustrates the power of the Map Reduce paradigm for big data computation.

Chapter "Big Data Search and Mining" talks of mining the information that big data implicitly carries within. Often, big data appear with patterns exhibiting the intrinsic relations they hold. Unearthed patterns could be of use for improving enterprise performances and strategic customer relationships and marketing. Towards this end, the chapter introduces techniques for big data search and mining. It also presents algorithms for social network clustering using the topology discovery technique. Further, some problems such as sentiment detection on processing text streams (like tweets) are also discussed.

Security is always of prime concern. Security lapses in big data could be higher due to its high availability. As these data are collected from different sources, the vulnerability for security attacks increases. Chapter "Security and Privacy of Big Data" discusses the challenges, possible technologies, initiatives by stakeholders and emerging trends with respect to security and privacy of big data.

The world today, being instrumented by several appliances and aided by several internet-based services, generates very high volume of data. These data are useful for decision-making and furthering quality of services for customers. For this, data service is provided by big data infrastructure to receive requests from users and to accordingly provide data services. These services are guided by Service Level Agreement (SLA). Chapter "Big Data Service Agreement" addresses issues on SLA specification and processing. It also introduces needs for negotiation to avail data services. This chapter proposes a framework for SLA processing.

Chapter "Applications of Big Data" introduces applications of big data in different domains including banking and financial services. It sketches scenarios for the digital marketing space.

Acknowledgments

The genesis of this book goes to 11th International Conference on Distributed Computing and internet Technology (ICDCIT) held in February 2015. Big data was a theme for industry symposium held as a prelude to the main conference. The authors of three chapters in this book presented their ideas at the symposium. Editors took the feedback from participants and conveyed the same to the chapter authors for refining their contents.

In preparation of this book, we received help from different quarters. Hrushikesha Mohanty expresses his sincere thanks to the School of Computer and Information Sciences, University of Hyderabad, for providing excellent environment for carrying out this work. I also extend my sincere thanks to Dr. Achyuta Samanta, Founder KIIT University, for his inspiration and graceful support for hosting the ICDCIT series of conferences. Shri. D.N. Dwivedy of KIIT University deserves special thanks for making it happen. The help from ICDCIT organizing committee members of KIIT University is thankfully acknowledged. Deepak Chenthati and Prachet Bhuyan extend their thanks to their respective organizations Teradata India Pvt. Ltd. and KIIT University. Thanks to Shri Abhayakumar, graduate student of SCIS, University of Hyderabad, for his help in carrying out some pressing editing work.

Our special thanks to chapter authors who, despite their busy schedules, contributed chapters for this book. We are also thankful to Springer for publishing this book. In Particular, for their support and consideration for the issues we have been facing while preparing the manuscript.

Hyderabad
March 2015

Hrushikesha Mohanty
Prachet Bhuyan
Deepak Chenthati

Contents

Editors and Contributors

About the Editors

Hrushikesha Mohanty is currently a professor at School of Computer and Information Sciences, University of Hyderabad. He received his Ph.D. from IIT Kharagpur. His research interests include distributed computing, software engineering and computational social science. Before joining University of Hyderabad, he worked at Electronics Corporation of India Limited for developing strategic real-time systems. Other than computer science research publications, he has penned three anthologies of Odia poems and several Odia short stories.

Prachet Bhuyan is presently an associate professor at KIIT University. He completed his bachelor and master degrees in computer science and engineering from Utkal University and VTU, Belgaum, respectively. His research interests include service-oriented architecture, software testing, soft computing and grid computing. Before coming to KIIT, he has served in various capacities at Vemana Institute of Technology, Bangalore, and abroad in Muscat, Oman. He has several publications in indexed journals as well as conferences. He has been generously awarded by several organisations including IBM for his professional competence.

Deepak Chenthati is currently a senior software engineer at Teradata India Private Limited. His Industry experience includes working on Teradata massively parallel processing systems, Teradata server management, Teradata JDBC drivers and administration of Teradata internal tools and confluence tool stack. His research interests include Web services, Teradata and database management. He is currently pursuing his doctorate from JNTU Hyderabad. He received his master and bachelor degrees in computer science from University of Hyderabad and Sri Venkateswaraya University, respectively.

Contributors

Hareesh Boinepelli Teradata India Pvt. Ltd., Hyderabad, India

Raoul P. Jetley ABB Corporate Research, Bangalore, India

VenkataSwamy Martha @WalmartLabs, Sunnyvale, CA, USA

Hrushikesha Mohanty School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India

P. Radha Krishna Infosys Labs, Infosys Limited, Hyderabad, India

Srini Ramaswamy US ABB, Cleveland, USA

Bhashyam Ramesh Teradata Corporation, Dayton, USA

Sithu D. Sudarsan ABB Corporate Research, Bangalore, India

Supriya Vaddi School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India

Acronyms

AAA	Authentication, authorization and access control
ACID	Atomicity, consistency, isolation and durability
BI	Business intelligence
BSP	Bulk synchronous parallel
CIA	Confidentiality, integrity and availability
CII	Critical information infrastructure
COBIT	Control objectives for information and related technology
CPS	Cyber-physical system
DHT	Distributed hash tables
DLP	Data loss prevention
DN	Data node
EDVAC	Electronic discrete variable automatic computer
EDW	Enterprise data warehouse
ER	Entity relation
ETL	Extract-transform-load
HDFS	Hadoop distributed file system
IaaS	Infrastructure as a service
iMapReduce	Iterative MapReduce
IoT	Internet of things
kNN	k nearest neighbour
MOA	Massive online analysis
MPI	Message passing interface
MR	MapReduce
NN	Name node
NSA	National security agency
PaaS	Platform as a service
PAIN	Privacy, authentication, integrity and non-repudiation
PII	Personally identifiable information
POS	Parts of speech
RWR	Random walk with restart
SaaS	Software as a service
SLA	Service-level agreement

SOA	Service-oriented architecture
SRG	Service relation graph
WEKA	Waikato environment for knowledge analysis
YARN	Yet another resource negotiator

Big Data: An Introduction

Hrushikesha Mohanty

Abstract The term big data is now well understood for its well-defined characteristics. More the usage of big data is now looking promising. This chapter being an introduction draws a comprehensive picture on the progress of big data. First, it defines the big data characteristics and then presents on usage of big data in different domains. The challenges as well as guidelines in processing big data are outlined. A discussion on the state of art of hardware and software technologies required for big data processing is presented. The chapter has a brief discussion on the tools currently available for big data processing. Finally, research issues in big data are identified. The references surveyed for this chapter introducing different facets of this emergent area in data science provide a lead to intending readers for pursuing their interests in this subject.

Keywords Big data applications · Analytics · Big data processing architecture · Big data technology and tools · Big data research trends

1 Big Data

"Big data" the term remains ill-defined if we talk of data volume only. It gives an impression before data size was always small. Then, we run into problem of defining something small and big. How much data can be called big—the question remains unanswered or even not understood properly. With relational database technology, one can really handle huge volume of data. This makes the term "big data" a misnomer.

Days of yesteryears were not as machine-driven as we see it today. Changes were also not as frequent as we find now. Once, data repository defined, repository was

H. Mohanty (✉)
School of Computer and Information Sciences, University of Hyderabad,
Gachhibowli 500046, Hyderabad, India
e-mail: hmcs_hcu@yahoo.com

H. Mohanty et al. (eds.), *Big Data*, Studies in Big Data 11,
DOI 10.1007/978-81-322-2494-5_1

used for years by users. Relational database technology thus was at the top for organisational and corporate usages. But, now emergent data no longer follow a defined structure. Variety of data comes in variety of structures. All accommodating in a defined structure is neither possible nor prudent to do so for different usages.

Our world is now literally swamped with several digital gadgets ranging from wide variety of sensors to cell phones, as simple as a cab has several sensors to throw data on its performance. As soon as a radio cab is hired, it starts sending messages on travel. GPS fitted with cars and other vehicles produce a large amount of data at every tick of time. Scenario on roads, i.e. traffic details, is generated in regular intervals to keep an eye on traffic management. Such scenarios constitute data of traffic commands, vehicles, people movement, road condition and much more related information. All these information could be in various forms ranging from visual, audio to textual. Leave aside very big cities, in medium-sized city with few crores of population, the emerging data could be unexpectedly large to handle for making a decision and portraying regular traffic conditions to regular commuters.

Internet of things (IoT) is the new emerging world today. Smart home is where gadgets exchange information among themselves for getting house in order like sensors in a refrigerator on scanning available amount of different commodities may make and forward a purchase list to a near by super market of choice. Smart cities can be made intelligent by processing the data of interest collected at different city points. For example, regulating city traffic in pick time such that pollution levels at city squares do not cross a marked threshold. Such applications need processing of a huge data that emerge at instant of time.

Conducting business today unlike before needs intelligent decision makings. More to it, decision-making now demands instant actions as business scenario unfolds itself at quick succession. This is so for digital connectivity that makes business houses, enterprises, and their stakeholders across the globe so closely connected that a change at far end instantly gets transmitted to another end. So, the business scenario changes in no time. For example, a glut in crude oil supply at a distributor invites changes in status of oil transport, availability at countries sourcing the crude oil; further, this impacts economy of these countries as the productions of its industries are badly affected. It shows an event in a business domain can quickly generate a cascade of events in other business domains. A smart decision-making for a situation like this needs quick collection as well as processing of business data that evolve around.

Internet connectivity has led to a virtual society where a person at far end of the globe can be a person like your next-door neighbour. And number of people in one's friend list can out number to the real number of neighbours one actually has. Social media such as Twitter, Facebook, Instagram and many such platforms provide connectivity for each of its members for interaction and social exchanges. They exchange messages, pictures, audio files, etc. They talk on various issues ranging from politics, education, research to entertainment. Of course, unfortunately such media are being used for subversive activities. Every moment millions of people on social media exchanges enormous amount of information. At times for different usages ranging from business promotions to security enhancement,

monitoring and understanding data exchanged on social media become essential. The scale and the speed at which such data are being generated are mind bugging.

Advancement in health science and technology has been so encouraging in today's world that healthcare can be customised to personal needs. This requires monitoring of personal health parameters and based on such data prescription is made. Wearable biosensors constantly feed real-time data to healthcare system and the system prompts to concerned physicians and healthcare professionals to make a decision. These data can be in many formats such as X-ray images, heartbeat sounds and temperature readings. This gives an idea for a population of a district or a city, the size of data, a system needs to process, and physicians are required to handle.

Research in biosciences has taken up a big problem for understanding biological phenomena and finding solution to disorders that at times set in. The research in system biology is poised to process huge data being generated from coding information on genes of their structure and behaviour. Researchers across the globe need access to each others data as soon as such data are available. As in other cases these data are available in many forms. And for applications like study on new virus and its spread require fast processing of such data. Further, visualisation of folds that happen to proteins is of importance to biologists as they understand nature has preserved gold mine of information on life at such folds.

Likewise many applications now need to store and process data in time. In year 2000, volume of data stored in the world is of size 800,000 petabytes. It is expected to reach 35 zettabytes by the year 2020. These figures on data are taken from book [1]. However, the forecast will change with growing use of digital devices. We are storing data of several domains ranging from agriculture, environment, house holdings, governance, health, security, finance, meteorological and many more like. Just storing such data is of no use unless data are processed and decisions are made on the basis of such data. But in reality making use of such large data is a challenge for its typical characteristics [2]. More, the issues are with data capture, data storage, data analysis and data visualisation.

Big data looks for techniques not only for storage but also to extract information hidden within. This becomes difficult for the very characteristics of big data. The typical characteristics that hold it different than traditional database systems include *volume, variety, velocity and value*. The term *volume* is misnomer for its vagueness in quantifying the size that is fit to label as big data. Data that is not only huge but expanding and holding patterns to show the order exist in data, is generally qualifying volume of big data. *Variety* of big data is due to its sources of data generation that include sensors, smartphones or social networks. The types of data emanate from these sources include video, image, text, audio, and data logs, in either structured or unstructured format [3]. Historical database dealing with data of past has been studied earlier, but big data now considers data emerging ahead along the timeline and the emergence is rapid so *Velocity* of data generation is of prime concern. For example, in every second large amount of data are being generated by social networks over internet. So in addition to volume, velocity is also a dimension for such data [4]. *Value* of big data refers to the process of extracting hidden

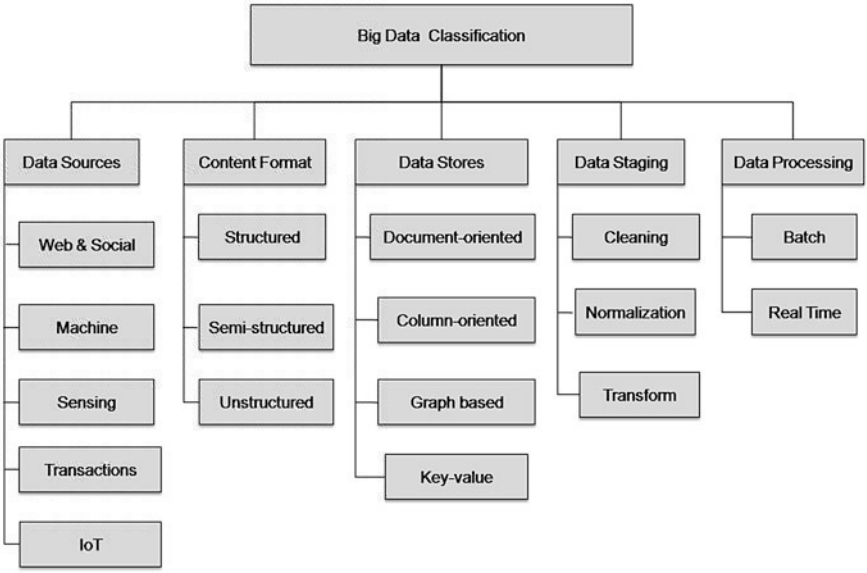

Fig. 1 Big data classification

information from emerging data. A survey on generation of big data from mobile applications is presented in [5].

Classification of big data from different perspectives as presented in [6] is presented in Fig. 1. The perspectives considered are data sources, content format, data stores, data staging, and data processing. The sources generating data could be web and social media on it, different sensors reading values of parameters that changes as time passes on, internet of things, various machinery that throw data on changing subfloor situations and transactions that are carried out in various domains such as enterprises and organisations for governance and commercial purposes. Data staging is about preprocessing of data that is required for processing for information extraction. From data store perspective, here the concern is about the way data stored for fast access. Data processing presents a systemic approach required to process big data. We will again touch upon these two issues later in Sect. 3.

Having an introduction on big data, next we will go for usages of these data in different domains. That gives an idea why the study on big data is important for both business as well as academic communities.

2 Big Data as a Service

In modern days, business has been empowered by data management. In 1970s, RDBMS (Relational Database Management System) has been successful in handling large volume of data for query and repository management. The next level of

data usage has been since 1990s, by making use of statistical as well as data mining techniques. This has given rise to the first generation of Business Intelligence and Analytics (BI&A). Major IT vendors including Microsoft, IBM, Oracle, and SAP have developed BI platforms incorporating most of these data processing and analytical technologies.

On advent of Web technology, organisations are putting businesses online by making use of e-commerce platforms such as Flipkart, Amazon, eBay and are searched for by websearch engines like Google. The technologies have enabled direct interactions among customers and business houses. User(IP)-specific information and interaction details being collected by web technologies (through cookies and service logs) are being used in understanding customer's needs and new business opportunities. Web intelligence and web analytics make Web 2.0-based social and crowd-sourcing systems.

Now social media analytics provide unique opportunity for business development. Interactions among people on social media can be traced and business intelligence model be built for two-way business transactions directly instead of traditional one-way transaction between business-to-customer [7]. We are need of scalable techniques in information mining (e.g. information extraction, topic identification, opinion mining, question-answering, event detection), web mining, social network analysis, and spatial-temporal analysis, and these need to gel well with existing DBMS-based techniques to come up with BI&A 2.0 systems. These systems use a variety of data emanating from different sources in different varieties and at different intervals. Such a collection of data is known as big data. Data in abundance being accompanied with analytics can leverage opportunities and make high impacts in many domain-specific applications [8]. Some such selective domains include e-governance, e-commerce, healthcare, education, security and many such applications that require boons of data science.

Data collected from interactions on social media can be analysed to understand social dynamics that can help in delivering governance services to people at right time and at right way resulting to good governance. Technology-assisted governance aims to use data services by deploying data analytics for social data analysis, visualisation, finding events in communities, extracting as well as forecasting emerging social changes and increase understanding of human and social processes to promote economic growth and improved health and quality of life.

E-commerce has been greatly benefited in making use of data collected from social media analytics for customer opinions, text analysis and sentiment analysis techniques. Personalised recommender systems are now a possibility following long-tail marketing by making use of data on social relations and choices they make [9]. Various data processing analytics based on association rule mining, database segmentation and clustering, anomaly detection, and graph mining techniques are being used and developed to promote data as a service in e-commerce applications.

In healthcare domain, big data is poised to make a big impact resulting to personalisation of healthcare [10]. For this objective, healthcare systems are planning to make use of different data the domain churns out every day in huge quantity. Two main sources that generate a lot of data include genomic-driven study, probe-driven

treatment and health management. Genomic-driven big data includes genotyping, gene expression and sequencing data, whereas probe-driven health care includes health-probing images, health-parameter readings and prescriptions. Health-management data include electronic health records and insurance records. The health big data can be used for hypothesis testing, knowledge discovery as well as innovation. The healthcare management can have a positive impact due to healthcare big data. A recent article [11] discusses on big data impact on host trait prediction using meta-genomic data for gastrointestinal diseases.

Security has been a prime concern and it grows more, the more our society opens up. Security threats emanating across boundary and even from within boundary are required to be analysed and understood [12]. And the volume of such information flowing from different agencies such as intelligence, security and public safety agencies is enormous. A significant challenge in security IT research is the information stovepipe and overload resulting from diverse data sources, multiple data formats and large data volumes. Study on big data is expected to contribute to success in mitigating security threats. Big data technology including such as criminal association rule mining and clustering, criminal network analysis, spatial-temporal analysis and visualisation, multilingual text analytics, sentiment and affect analysis, and cyber attacks analysis and attribution should be considered for security informatics research.

Scientific study has been increasingly collaborative. Particularly, sharing of scientific data for research and engineering data for manufacturing has been a modern trend, thanks to internet providing a pervading infrastructure for doing so [13]. Big data aims to advance the core scientific and technological research by analysing, visualising, and extracting useful information from large, diverse, distributed and heterogeneous data sets. The research community believes this will accelerate the progress of scientific discovery and innovation leading to new fields of enquiry that would not otherwise be possible. Particularly, currently we see this happening in fields of research in biology, physics, earth science, environmental science and many more areas needing collaborative research of interdisciplinary nature.

The power of big data, i.e. its impact in different domains, is drawn from analytics that extracts information from collected data and provide services to intended users. In order to emphasise on vast scope of impending data services, let us discover some analytics of importance. *Data Analytics* are designed to explore and leverage unique data characteristics, from sequential/temporal mining and spatial mining, to data mining for high-speed data streams and sensor data. Analytics are formulated based on strong mathematical techniques including statistical machine learning, Bayesian networks, hidden Markov models, support vector machine, reinforcement learning and ensemble models. Data analytics are also looking into process mining from series of data collected in sequence of time. Privacy security concerned data analytics ensure anonymity as well as confidentiality of a data service. *Text Analytics* aims at event detection, trend following, sentiment analysis, topic modelling, question-answering and opinion mining. Other than traditional soft computing and statistical techniques, text analytics take the help of several well-researched natural language processing techniques in parsing