Large-Scale Data Streaming, Processing, and Blockchain Security

Hemraj Saini
Jaypee University of Information Technology, India

Geetanjali Rathee
Jaypee University of Information Technology, India

Dinesh Kumar Saini
Sohar University, Oman

A volume in the Advances in Information Security, Privacy, and Ethics (AISPE) Book Series

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Saini, Hemraj, 1977- editor. | Rathee, Geetanjali, 1990- editor. | Saini, Dinesh Kumar, 1974- editor.
Title: Large-scale data streaming, processing, and blockchain security / Hemraj Saini, Geetanjali Rathee, and Dinesh Kumar Saini, editors.
Description: Hershey, PA : Information Science Reference, an imprint of IGI Global, [2020] | Includes bibliographical references and index. | Summary: "This book explores the latest methodologies, modeling, and simulations for coping with the generation and management of large-scale data in both scientific and individual applications"-- Provided by publisher.
Identifiers: LCCN 2019052981 (print) | LCCN 2019052982 (ebook) | ISBN 9781799834441 (hardcover) | ISBN 9781799834458 (paperback) | ISBN 9781799834465 (ebook)
Subjects: LCSH: Data mining. | Streaming technology (Telecommunications) | Blockchains (Databases)
Classification: LCC QA76.9.D343 L368 2020 (print) | LCC QA76.9.D343 (ebook) | DDC 006.3/12--dc23
LC record available at https://lccn.loc.gov/2019052981
LC ebook record available at https://lccn.loc.gov/2019052982

This book is published in the IGI Global book series Advances in Information Security, Privacy, and Ethics (AISPE) (ISSN: 1948-9730; eISSN: 1948-9749)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Advances in Information Security, Privacy, and Ethics (AISPE) Book Series

ISSN:1948-9730
EISSN:1948-9749

Editor-in-Chief: *Manish Gupta* State University of New York, USA

MISSION

As digital technologies become more pervasive in everyday life and the Internet is utilized in ever increasing ways by both private and public entities, concern over digital threats becomes more prevalent.

The **Advances in Information Security, Privacy, & Ethics (AISPE) Book Series** provides cutting-edge research on the protection and misuse of information and technology across various industries and settings. Comprised of scholarly research on topics such as identity management, cryptography, system security, authentication, and data protection, this book series is ideal for reference by IT professionals, academicians, and upper-level students.

COVERAGE

- Information Security Standards
- Internet Governance
- IT Risk
- Cyberethics
- Technoethics
- Risk Management
- Electronic Mail Security
- CIA Triad of Information Security
- Privacy Issues of Social Networking
- Data Storage of Minors

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: http://www.igi-global.com/publish/.

The Advances in Information Security, Privacy, and Ethics (AISPE) Book Series (ISSN 1948-9730) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit http://www.igi-global.com/book-series/advances-information-security-privacy-ethics/37157. Postmaster: Send all address changes to above address.

Titles in this Series

For a list of additional titles in this series, please visit:
http://www.igi-global.com/book-series/advances-information-security-privacy-ethics/37157

Applied Approach to Privacy and Security for the Internet of Things
Parag Chatterjee (National Technological University, Argentina & University of the Republic, Uruguay) Emmanuel Benoist (Bern University of Applied Sciences, Switzerland) and Asoke Nath (St. Xavier's College, Kolkata, India)
Information Science Reference • © 2020 • 295pp • H/C (ISBN: 9781799824442) • US $235.00

Advanced Localization Algorithms for Wireless Sensor Networks
M. Vasim Babu (Institute of Technology and Sciences, India)
Information Science Reference • © 2020 • 300pp • H/C (ISBN: 9781799837336) • US $195.00

Social, Legal, and Ethical Implications of IoT, Cloud, and Edge Computing Technologies
Gianluca Cornetta (Universidad CEU San Pablo, Spain) Abdellah Touhafi (Vrije Universiteit Brussel, Belgium) and Gabriel-Miro Muntean (Dublin City University, Ireland)
Information Science Reference • © 2020 • 333pp • H/C (ISBN: 9781799838173) • US $215.00

Multidisciplinary Approaches to Ethics in the Digital Era
Meliha Nurdan Taskiran (Istanbul Medipol University, Turkey) and Fatih Pinarbaşi (Istanbul Medipol University, Turkey)
Information Science Reference • © 2020 • 300pp • H/C (ISBN: 9781799841173) • US $195.00

Sensor Network Methodologies for Smart Applications
Salahddine Krit (Ibn Zohr University, Morocco) Valentina Emilia Bălaş (Aurel Vlaicu University of Arad, Romania) Mohamed Elhoseny (Mansoura University, Egypt) Rachid Benlamri (Lakehead University, Canada) and Marius M. Bălaş (Aurel Vlaicu University of Arad, Romania)
Information Science Reference • © 2020 • 279pp • H/C (ISBN: 9781799843818) • US $195.00

For an entire list of titles in this series, please visit:
http://www.igi-global.com/book-series/advances-information-security-privacy-ethics/37157

701 East Chocolate Avenue, Hershey, PA 17033, USA
Tel: 717-533-8845 x100 • Fax: 717-533-8661
E-Mail: cust@igi-global.com • www.igi-global.com

Table of Contents

Detailed Table of Contents

Trust is a firm belief over a person or a thing in distributed environment based on its feedback on review based on its performance by others. Similarly, in cloud, trust models play an important role in solving various open challenges in cloud environment. This chapter showcases all such issues that can be solved by trust management techniques. This work discourses various trust management models and its categorization. The work discourses existing work using trust models from the field of grid computing, cloud computing, and web services because all these domains are sub child of each other. The work provides an abstract view over all trust models and find the suitable one for cloud and its future prospects.

Section 2

The future applications of blockchain are expected to serve millions of users. To provide variety of services to the users, using underlying technology has to consider large-scale storage and assessment behind the scene. Most of the current applications of blockchain are working either on simulators or via small blockchain network. However, the storage issue in the real world is unpredictable. To address the issue of large-scale data storage, the authors have introduced the data storage scheme in blockchain (DSSB). The storage model executes behind the blockchain ledger to store large-scale data. In DSSB, they have used hybrid storage model using IPFS and MongoDB(NoSQL) in order to provide efficient storage for large-scale data in blockchain. In this storage model, they have maintained the content-addressed hash of the transactions on blockchain network to ensure provenance. In DSSB, they are storing the original data (large-scale data) into MongoDB and IPFS. The DSSB model not only provides efficient storage of large-scale data but also provides storage size reduction of blockchain ledger.

The meaningful data extraction from the biological big data or omics data is a remaining challenge in bioinformatics. The deep learning methods, which can be used for the prediction of hidden information from the biological data, are widely used in the industry and academia. The authors have discussed the similarity and differences in the widely utilized models in deep learning studies. They first discussed the basic structure of various models followed by their applications in biological perspective. They have also discussed the suggestions and limitations of deep learning. They expect that this chapter can serve as significant perspective for continuous development of its theory, algorithm, and application in the established bioinformatics domain.

We can name many industries that are still based on the same working practices and business models that they have had for a long time – maybe since they started. Despite the wealth of modern technology now available, public infrastructure, a critical component for the well-being of the society, is still an industry based on the paperwork, letters, emails, manual approvals, and a large amount of guess work. It involves a lot of manual effort and is also error prone. It is really very hard for the stakeholders and end users to get an update on the progress of the project, which impacts them directly or indirectly. The authors intend to develop a groundbreaking blockchain platform that can meet the needs of all the different stakeholders involved in creating and providing a better infrastructure. They plan to automate the entire process by using smart contracts to minimize paperwork for the government officials. This will not only eliminate the errors that can happen during manual execution but will also provide a real-time update to all the stakeholders in making the process more transparent.

The usage of information is essential for data-driven capabilities in artificial intelligence. The data-driven AI techniques lead to several security and privacy concerns. Among various digital techniques, digital rights management is required as one of collaboration scheme that ensures the security and privacy of intellectual rights. Though a number of researchers have proposed various security techniques, none of them have proposed an efficient and effective privacy procedure for digital rights. Recently, blockchain technique is considered as one of the major security methods to ensure a transparent communication among individuals. It can be used by various applications such as industries, marketing, transportation systems, etc. The aim of this chapter is to propose an ensured resource allocation algorithm that validates the scheme by comparing various security measures against previous approaches. Further, the proposed phenomenon ensures the transparency on security and privacy due to its integration.

In this chapter, the authors have detailed the need of blockchain technology along with its case studies in different domains. The literature survey is described that describes how blockchain technology is rising. Further, a number of domains where blockchain technology can be applied along with its case studies have been discussed. In addition, the authors have considered the various use cases with their recent issues and how these issues can be resolved using the blockchain technology by proposing some new ideas. A proposed security framework in certain applications using blockchain technology is presented. Finally, the chapter is concluded with future directions.

Internet of things(IoT) is the conception of interfacing the devices to the internet to make life more efficient. It comprises the large amount of data in its network where it fails to assure complete security in the network. Blockchain is a distributed ledger where it mainly focuses on the data security. Every block in the blockchain network is connected to its next block, which prevents threats like large data loss. In the area of agri-food supply chain, where IoT plays a very important role, there occurs data integrity issues or data tampering. This can lead to improper supply chain management, timely shortage of goods, food spoilage, etc. So the traceability of agri-food supply chain is necessary to ensure food safety and to increase the trust between all stakeholders and consumers. Many illegal activities can be prevented, and cold chain monitoring can be achieved by bringing in transparency and traceability.

India is the largest democracy in the world, and in spite of that, it faces various challenges on a daily basis that hinder its growth like corruption and human rights violations. One of the ugliest phases of corruption and political mayhem is visible during the election process where no stone is kept unturned in order to gain power. However, it is the common citizen who suffers most in terms of clarity as well as security when it comes to his/her vote. Blockchain can play a very important role in ensuring that the voters registering their votes are legit and the counting of votes is not manipulated in any way. It is also needed in today's times where the world is available to people in their smart phones to also give them the opportunity to register their votes hassle free via their smart phones without having to worry about the system getting hacked. Therefore, in this chapter, the proposed layout will be based on a smart contract, using Ethereum software to create an e-voting app. In this chapter, the authors have proposed a secure e-voting framework through blockchain mechanism.

Foreword

In recent years, there has been an enormous diffusion of large-scale data technologies, usually oriented to data processing, omitting an equally important aspect related to the transformation of data to be ready for this process. In fact, it is increasingly urgent to address the issue of heterogeneity, diversity, and complexity of data, and how to normalize, integrate, and transform the data from many sources into the format required to run large-scale analysis. This edited book addresses the research about large-scale data management with semantic technologies as a unified data access layer and a consistent approach to analytic execution. Semantic technologies have been used to create domain models describing mutually relevant datasets and the relationships between them.

In addition, security of the information is also a primary challenge for the large-scale data which can be ensured by blockchain. This book is intended as an exploration of the broader concepts, features, and functionality of Bitcoin and blockchain technology, and their future possibilities and implications; it does not support, advocate, or offer any advice or prediction as to the industry's viability. The blockchain industry is in an emergent and immature phase and very much still in development with many risks. Right now is the time to learn about the underlying technologies; their potential uses, dangers, and risks; and perhaps more importantly, the concepts and their extensibility. The objective here is to provide a comprehensive overview of the nature, scope, and type of activity that is occurring in the cryptocurrency industry and envision its wide-ranging potential application. The account is necessarily incomplete, prone to technical errors, and could likely soon be out-of-date as different projects described here fail or succeed. Or, the entire Bitcoin and blockchain technology industry as currently conceived could become outmoded or superseded by other models.

The challenges in Large-Scale Data Streaming, Processing, and Blockchain Security are both difficult and interesting. People are working on them with enthusiasm, tenacity, and dedication to develop new methods of analysis and provide new solutions to keep up with the ever-changing threats. In this new age of global interconnectivity and interdependence, it is necessary to provide security

practitioners, both professionals and students, with state-of-the art knowledge on the frontiers in information assurance. This book is a good step in that direction.

Rakesh Belwal
Faculty of Business, Sohar University, Oman & Business School, University of Queensland, Australia

Preface

At the highest-level description, this book is about large-scale data mining. However, it focuses on data streaming, processing, and security of very large amounts of data, that is, data is so large and does not fit in traditional category. Further, the identified topics for call for book chapters provide it emphasis over streaming, processing and blockchain security of large-scale data. The principle topics of this book cover Large scale data, Large scale Data streaming, Large scale data streaming models, Large scale data processing models, Large scale data and machine leaning, blockchain Security concerns in large scale data, blockchain Security models for large scale data, Large scale data in cloud or fog, Scheduling of Large scale data processing on clouds or fog, and blockchain Security and privacy in big data clouds or fog. The identified contents of this book will be helpful to a set of companies or organizations those are flooding an enormous amount of data and need frequent mining of required contents. In addition, presently, IoT structures are frequently used in a number of applications and generating large scale data which needs affective processing and streaming with sufficient security and our book will help in this aspect.

The book will provide proper understanding, methodologies, modeling, and simulation to cope up the current requirement of the technological world generating large scale data not only in scientific applications but also in applications affecting individual's day to day life.

This book will aim to provide relevant theoretical frameworks and the latest empirical research findings in the area. It will be written for professionals who want to improve their understanding of the strategic role of Large-Scale Data Streaming, Processing, and Blockchain Security at different levels in the related applications.

The target audience of this book will be composed of professionals and researchers working in the field of Large-Scale Data Streaming, Processing, and Blockchain Security in various domains, e.g. social networking, banking, agriculture, chemistry, data mining, cloud computing, finance, marketing, stocks, BDA, health care etc. Moreover, the book will provide insights and support executives concerned with the management of expertise, knowledge, information, innovative technologies and

organizational development in different types of work communities and environments. A short review about the commitments for this book is as underneath-

Chapter 1: Data processing has become an important field in today's big data dominated world. The data has been generating at a tremendous pace from different sources. There has been a change in the nature of data from batch-data to streaming-data, and consequently, data processing methodologies have also changed. Traditional SQL is no more capable of dealing this big data. This book chapter describes the nature of data and various tools, techniques, and technologies to handle this big data. The chapter also describes the need of shifting big data on to cloud and the challenges in big data processing in the cloud, the migration from data processing to data analytics, tools used in data analytics, and the issues and challenges in data processing and analytics. Then the chapter touches an important application area of streaming data: sentiment analysis, and tries to explore it through some test case demonstrations and results.

Chapter 2: Expansion of data in the dimensions of volume, variety or velocity is leading to big data. Learning from this big data is challenging and beyond capacity of conventional machine learning methods and techniques. Generally, big data getting generated from real time scenarios is imbalance in nature with uneven distribution of classes. This imparts additional complexity in learning from big data since the class that is under-represented is more influential and its correct classification becomes critical than that of over-represented class. This chapter addresses the imbalance problem and its solutions in context of big data along with a detailed survey of work done in this area. Subsequently, it also presents an experimental view for solving imbalance classification problem and a comparative analysis between different methodologies afterwards.

Chapter 3: The emerging trends in fog computing has increased the interests and focus in both industry and academia. Fog computing extends cloud computing facilities like the storage, networking, and computation towards the edge of networks wherein it offloads the cloud data centres and reducing the latency of providing services to the users. This paradigm is like cloud in terms of data, storage, application and computation services, except with a fundamental difference - it is decentralized. Furthermore, these Fog systems can process huge amount of data locally and can be installed on hardware of different types. These characteristics make Fog to be suitable for time and location-based applications like Internet of Things (IoT) devices which can process large amount of data. In this chapter we present fog data streaming, its architecture and various applications.

Chapter 4: Trust is a firm belief over a person or a thing in distributed environment based on its feedback on review based on its performance by others. Similarly, in cloud trust models plays an important role to solve various open challenges in cloud environment. This chapter showcases all such issues that can be solved by trust

management techniques. This work discourses various trust management models and its categorization. The work discourses existing work using trust models from the field of grid computing cloud computing and web services because all these domains are sub child of each other. The work main focus it provide abstract view over all trust models and find the suitable one for cloud and its future prospects.

Chapter 5: The future applications of blockchain are expected to serve millions of users. To provide variety of services to the users using underlying technology has to consider large-scale storage and assessment behind the scene. Most of the current applications of blockchain are working either on simulators or via small blockchain network. However, the storage issue in the real world is unpredictable. To address the issue of large-scale data storage, we have introduces the data storage scheme in blockchain (DSSB). Our storage model executes behind the blockchain ledger to store large-scale data. In DSSB, we have used hybrid storage model using IPFS and MongoDB (NoSQL) in order to provide efficient storage for large-scale data in blockchain. In this storage model, we have maintained the content-addressed hash of the transactions on blockchain network to ensure provenance. In DSSB, we are storing the original data (Large-Scale data) into MongoDB and IPFS. The DSSB model not only provides efficient storage of large-scale data but also provide storage size reduction of blockchain ledger.

Chapter 6: The meaningful data extraction from the biological big data or omics data is a remaining challenge in bioinformatics. The deep learning methods, which can be used for the prediction of hidden information from the biological data, are widely used in the industry and academia. We have discussed the similarity and differences in the widely utilized models in deep learning studies. We first discussed the basic structure of various models followed by their applications in biological perspective. We have also discussed the suggestions and limitations of deep learning. We expect that this chapter can serve as significant perspective for continuous development of its theory, algorithm, and application in the established bioinformatics domain.

Chapter 7: We can name many industries that are still based on the same working practices and business models that they had since a long time – maybe the time they started. Despite the wealth of modern technology now available, public infrastructure - a critical component for the well-being of the society, is still an industry based on the paperwork, letters, emails, manual approvals, and a large amount of guess work. It involves a lot of manual effort and is also error prone. It is really very hard for the stakeholders and end users to get an update on the progress of the project which impacts them directly or indirectly. We intend to develop a ground-breaking blockchain platform that can meet the needs of all the different stakeholders involved in creating and providing a better infrastructure. We plan to automate the entire process by using smart contracts to minimize paperwork for the government officials. This will not only eliminate the errors which can happen

during manual execution but will also provide real time update to all the stakeholders in making the process more transparent.

Chapter 8: The usage of information is essential for data driven capabilities in Artificial intelligence. The data driven AI techniques leads to several security and privacy concerns. Among various digital techniques, digital rights management is required as one of collaboration scheme that ensures the security and privacy of intellectual rights. Through number of researchers have proposed various security techniques, however, none of them have proposed an efficient and effective privacy procedure for digital rights. Recently, Blockchain technique is considered as one of the major security method to ensure a transparent communication among individuals. It can be used by various applications such as industries, marketing, transportation systems etc. The aim of this chapter is to propose an ensured resource allocation algorithm, that validates the scheme by comparing various security measures against previous approaches. Further, the proposed phenomenon ensures the transparency on security and privacy due to its integration.

Chapter 9: In this chapter, we have detailed the need of Blockchain technology along with its case studies in different domains. The literature survey is described that intricate how Blockchain technology is rising now-a-days. Further, number of domains where Blockchain technology can be applied along with its case studies has been discussed. In addition, we have considered the various use case with their recent issues and how these issues can be resolved using the blockchain technology by proposing some new ideas. A proposed security framework in certain applications using blockchain technology is presented. Finally, the chapter is concluded with future directions.

Chapter 10: Over the last few years, development of technology took a major role in day to day life. Internet of Things (IoT) is the conception of interfacing the devices to the internet to make life more efficient. It comprises the large amount of data in its network were it fails to assure complete security in the network. Blockchain is a distributed ledger and it mainly focuses on the data security. Every block in the blockchain network is connected to its next block which prevents threats like large data loss. In the area of Agri-Food Supply Chain, where IoT plays a very important role, there occur data integrity issues or data tampering. This can lead to improper Supply Chain Management, timely shortage of goods, food spoilage, etc. So the traceability of Agri-Food Supply Chain is necessary to ensure food safety and to increase the trust between all stakeholders and consumers. Many illegal activities can be prevented and cold chain monitoring can be achieved by bringing in transparency and traceability.

Chapter 11: India is the largest democracy in the world and in spite of that faces various challenges on a daily basis which hinder its growth like corruption and human rights violations. One of the ugliest phases of corruption and political mayhem is visible during the election process where no stone is kept unturned in order to gain power. However, it is the common citizen who suffers most in terms of clarity as well as security when it comes to his/her vote. Blockchain can play a very important role in ensuring that the voters registering their votes are legit and the counting of votes is not manipulated in any way. It is also needed in today's times where the world is available to people in their smart phones to also give them the opportunity to register their votes hassle free via their smart phones without having to worry about the system getting hacked. Therefore, in this paper, the proposed layout will be based on a smart contract, using Ethereum software to create an e-voting app. In this paper, we have proposed a secure e-voting framework through blockchain mechanism.

We hope that the quality chapters published in this book will be able to serve the concerned humanity, science and technology at the best.

ACKNOWLEDGMENT

The editors are thankful to the authors and reviewers who contributed to this book with their scientific work and useful comments, respectively.

Hemraj Saini
Jaypee University of Information Technology, India

Geetanjali Rathee
Jaypee University of Information Technology, India

Dinesh Kumar Saini
Sohar University, Oman

Acknowledgment

We take this opportunity to express our gratitude to our Vice Chancellor, Professor Vinod Kumar and Director & Academic Head, Professor Samir Dev Gupta for their continuous motivation and for the support given in allowing using Institutional resources for the topic. We thank our colleagues for their support, and also the students who helped us in organizing the materials. We are also very thankful to the reviewers for providing their valuable input for the chapters.

We would also like to thank the team of IGI Global for the enthusiasm and support extended to us during various stages of the project. Finally, we would like to thank all the valuable contributors for the book.

Hemraj Saini, Geetanjali Rathee, & Dinesh Saini

Introduction

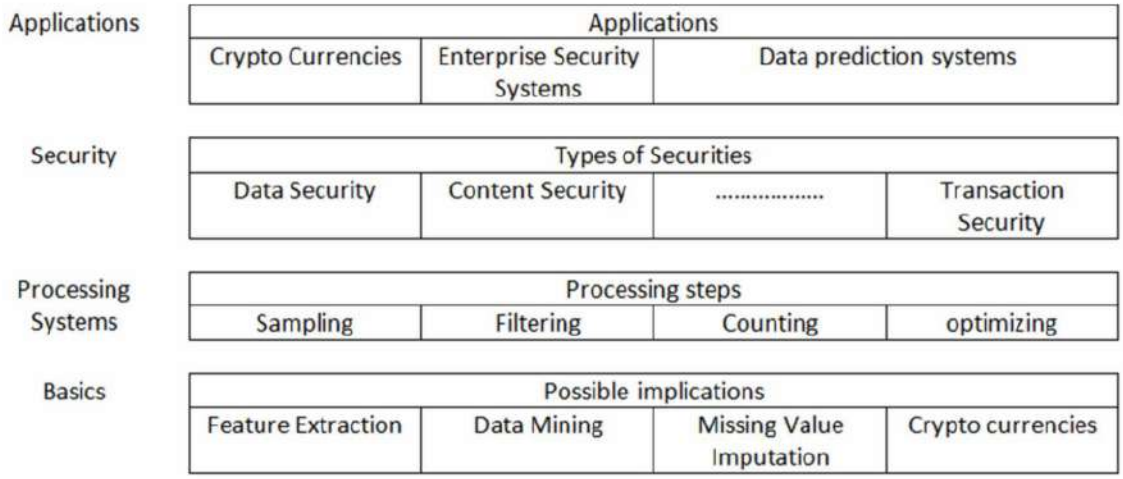
It feels like every distinct day we are hesitant across added use-cases for the blockchain technology. Large number of industries is verdict out that either the blockchain technology is going to take them to the next level, or may end up fetching their major threat. One of the various fields that have exposed a symbiotic association with blockchain is large data (big data) streaming and processing. In this book, we are going to explore the relationship among large data and blockchain technique. Before we go further, let's appreciate what blockchain and large data mean.

The reason why blockchain and large data can have a very abundant relationship is that the blockchain technique can easily and efficiently cover the defects of large data. There are three major motives why this corporation can be productive:

- **Security and Privacy**: Blockchain's biggest benefit is the security that it conveys to the information stored inside it. All the information that is inside the blockchain is non-altered.
- **Transparent**: The transparent networking structure of the blockchain can help further to trace information back to its origin point.
- **Flexibility**: The blockchain can record all types and kinds of information.
- **Decentralization:** All the information that is recorded inside a blockchain is not possessed by one single individual. Therefore, there is no chance of information stolen and alteration even if that individual gets concessioner in any way.

By considering all these features, the conclusion that we can sketch is that whatever information comes out of the blockchain network is worthy. The information is fraud-proof and already been cleaned through it. Now, this brings us to the next question what exactly are the characteristics of blockchain that enables this relationship? Now let us understand how the blockchain mechanism can further provide a better relationship in large data streaming and processing along with transparency and worthy security.

Figure 1. Global presentation of the most important fields, as discussed in the book

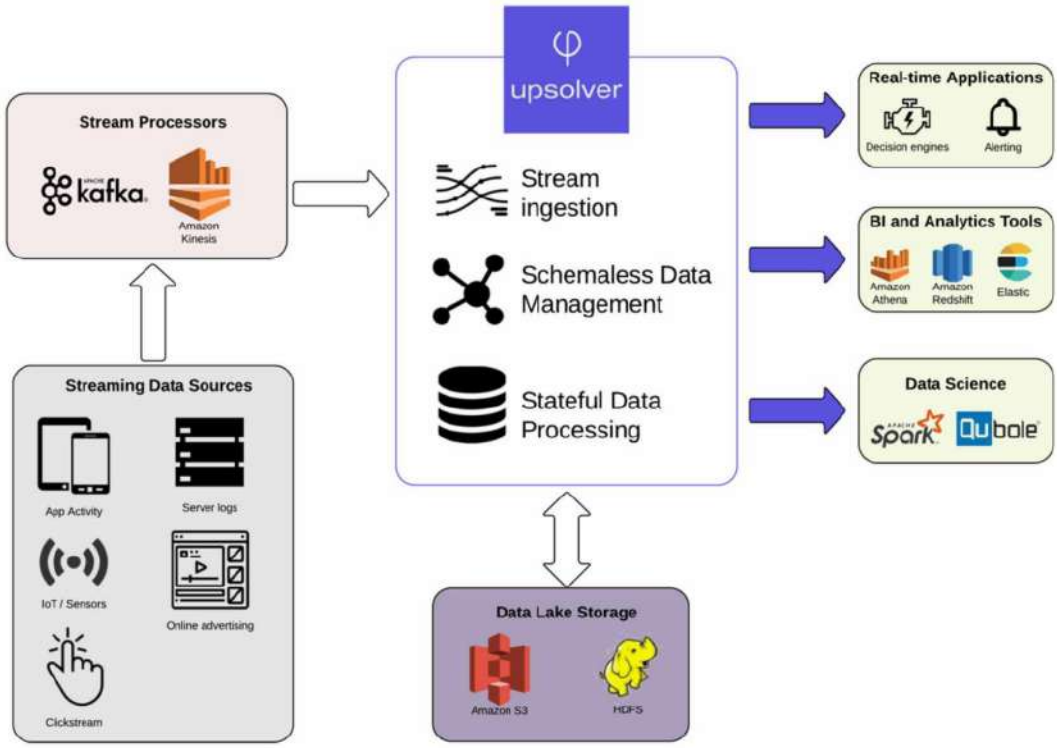


1.3.1 Large-Scale Data Streaming

Numbers of organizations are adopting modern data streaming deployment full stack approaches, rather than relying on focusing together open-source technologies. The recent information platform is erect on organization-centric value chains instead of IT-centric coding schemes, wherein the complexity of conventional architecture is preoccupied into a solitary self-service platform that revolves event flow into analytics-ready information.

The idea behind this new technique is to act as the centralized information platform that regulates the manual parts of working with streaming information such as streaming and batch ETL, message ingestion, preparing data for analytics and storage management. A sample overview of the modern streaming architecture is depicted in Figure 2.

Figure 2. Modern streaming architecture



Benefits of using modern large streaming architecture:

- Can eliminate the necessitate for large information engineering projects
- Built in high availability, Performance and fault tolerance
- Can be deployed easily as newer platforms are cloud-based with no upfront investment
- Supportive and flexible for multiple use cases

1.3.2 Large-scale data processing

A. Data collection

The first step in data processing is information gathering where information is pulled from available resources, including data warehouses and data lakes. Further, it is significant that the data resources available are legitimate and well-built so the information gathered (and later used as meaningful information) is of the peak possible quality.

B. Data Preparation

Once the data is gathered, it then comes into data preparation stage. Data preparation is also referred as "pre-processing". It is the phase at which raw data is organized and cleaned up for the following phases of data processing. During the data preparation, raw information is initially diligently checked for any errors and mistakes. The purpose of this step is to delete or remove bad information such as incomplete, redundant, and incorrect data and then begin to generate high-quality information for the best organization intelligence.

C. Data Input

The clean information is then entered into its final destination (perhaps a CRM like data warehouse like Redshift and Salesforce), and translated into a language that it can be easily understand. Data input is the first phase in which raw data starts to take the form of serviceable information.

D. Processing

During this stage, the information inputted to the computer in the previous phase is actually practiced for interpretation. Processing is done using machine/deep learning techniques and algorithms, though the data processing itself may vary vaguely depending on the source of information being processed such as social networks, data lakes, and connected devices etc. and its anticipated use such as medical diagnosis from connected devices, examining advertising patterns, and determining customer needs.

E. Data Output/Interpretation

The interpretation/ output stage is the phase at which information is finally serviceable to non-data scientists. It is readable, translated and often in the form of videos, graphs, plain text, and images, etc. Members of the institution or organization can now start to self-serve the information for their own analytics projects.

F. Data Storage

The final stage of information processing is storage. After all of the information is processed, it is then stored for the future purpose. While some data may be put to use at once, much of it can be served as a purpose later on. In addition, properly stored information is a necessity for observance with protection legislation like GDPR.

Whenever the information is properly stored, it can be easily and quickly accessed by members of the businesses when needed.

1.3.3 Blockchain Security

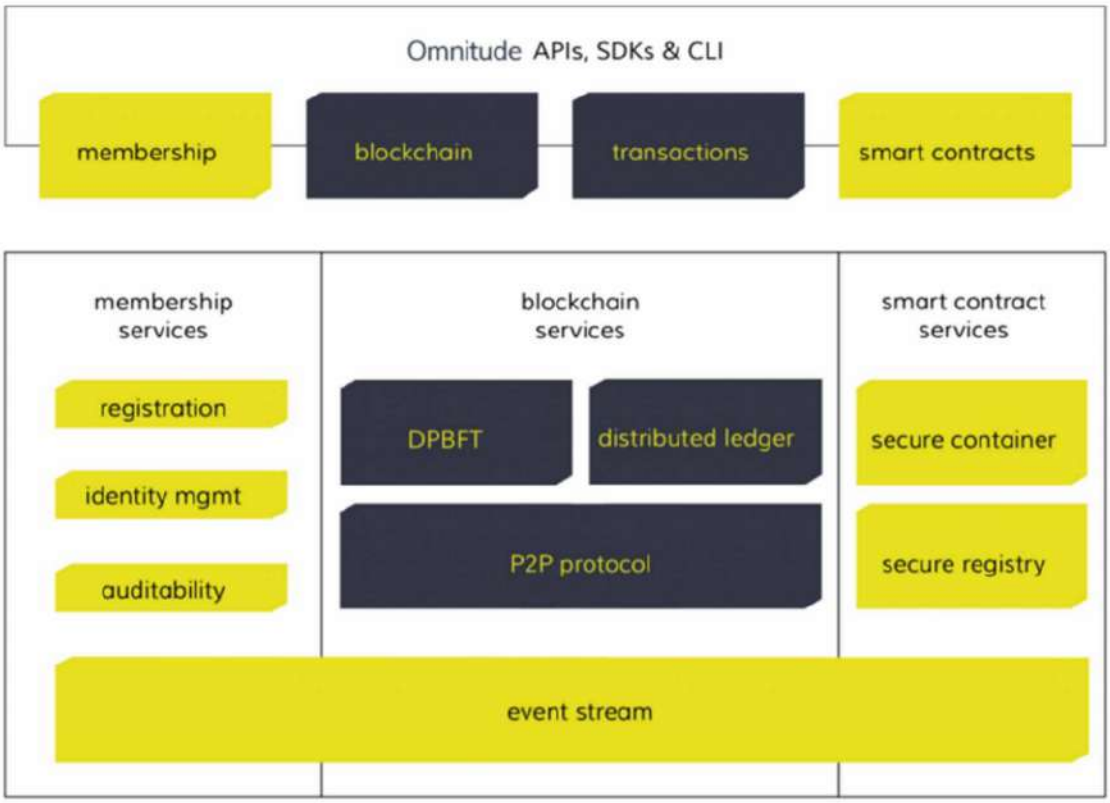
A blockchain, as the name entails, is a chain of digital "blocks" that hold records of transactions. Each block is gathered to all the blocks after and before it. This makes it hard to interfere with a single data because an intruder would require changing the block having that record as well as those associated to it to delete detection. This alone might not seem like much of deterrence, however, blockchain has some other inherent features that provide further means of security.

The data or information stored on a blockchain is secured through cryptography. Network applicants have their own private keys that are allocated to the transactions they make and act as an individual digital signature. If stored information is altered, the signature will turn out to be invalid and the peer network will know right away that incredible has happened. Early announcement is critical to preventing additional damage.

Unfortunately for those ambitious intruders, blockchains are distributed and decentralized across peer-to-peer networks that are repeatedly kept and updated in sync. Because they aren't enclosed in a central location, therefore, blockchains don't have a central point of failure and cannot be altered from a single device. It would necessitate massive records of computing power to entrance every instance (or at least a 51 percent majority) of a definite blockchain and change them all at the same instance. There has been some argue about whether this defines smaller blockchain networks that could be susceptible to threat, however a verdict hasn't been accomplished. In any case, the bigger your network is, the more tamper-resistant your blockchain will be. At a glance, blockchains have some desirable features that would help to secure your transaction information. Though, there are other requirements and conditions to believe when you desire to use a blockchain for commerce.

1.3.4 Blockchain in Streaming and Processing

Figure 3. Reference architecture for blockchain and middleware

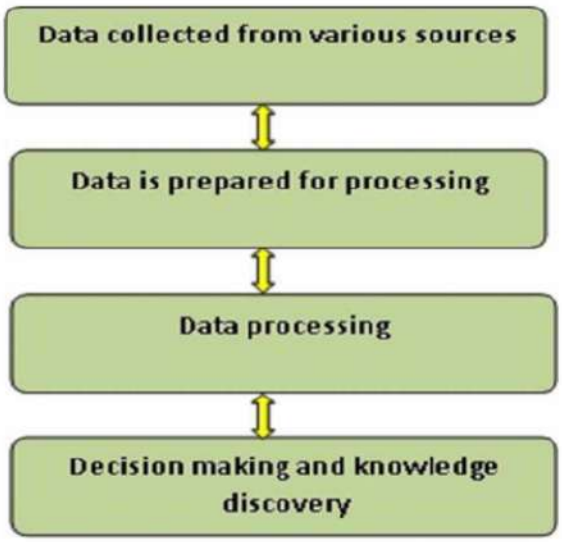


Blockchain is the next big thing for middleware! There is no such question around this. You need to intersect other applications, cloud offerings and micro-services with a blockchain infrastructure to acquire real value out of it. Further, machine learning and visual analytics have to be leveraged to get patterns and insights in non-blockchain and blockchain data. Finally, streaming analytics is worn to apply these patterns and insights to new events in a blockchain communications. There is a variety of applications like compliance issues, fraud detection, supply chain processes, optimization of manufacturing or any kind of circumstances with the Internet of Things (IoT). Reference architecture for blockchain and middleware is depicted in Figure 3.

DATA PROCESSING

Since last decade, rapid development of Internet enabled services such as social media, Internet of Things, and cloud based services have led to tremendous growth of data termed as big data. This data has become very difficult to be handled and managed for further processing (Jin et al., 2015). It has been estimated that around 2.5 quintillion bytes of new data is generated per day and expected to be more in near future as the number of internet users are growing unprecedentedly. This exponential growth of data has posed many challenges in front of researchers, academia and Industry across the globe. Moreover, the big data is unstructured: it varies in volume, velocity, veracity and variety makes (4Vs) it more challenging to manage and process (Mishra, R. K., & Mishra, R. K., 2018). This sudden explosion of data in terabytes, petabytes and exabytes could not be handled by the traditional database such as SQL led to the emergence of new tools and techniques to process the big data (Storey, V. C., & Song, I. Y., 2017).

Figure 1. Big data chain



Big data processing and analysis have become very crucial for better decision making, knowledge discovery, business intelligence and actionable insights. The Fig-1 represents the big data chain i.e. from data collection to decision making (Janssen, M., van der Voort, H., & Wahyudi, A., 2017). Big data is collected in raw form from various sources of interest which need to be prepared for processing. Next the quality data sets are prepared for further processing using data cleansing and standardization. After that, data processing takes place which includes transformation,

aggregation and pattern generation. Once the data processing is completed, various reports are generated and analyzed for better decision making, knowledge discovery and insight or trends. Analysis of data could be classified as descriptive, diagnostic, predictive and prescriptive (Perwej, Y., 2017).

This book chapter proposes to show various tools, techniques, and technologies of data processing and analytics. Later, the use streaming data for sentiment analysis through executable test cases is presented. Sentiment analysis is performed on run-time tweets with Python using twitter API "tweepy" and obtained results are presented through plots.

A survey on various sentiment analysis methods used by researchers is also presented. This would also help in identifying the best one and possibly may be in predicting a newer one.

FAILURE OF TRADITIONALSQL IN HANDLING BIG DATA

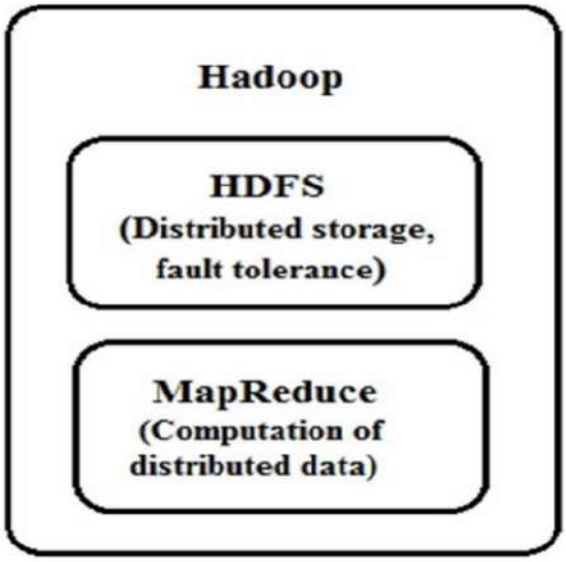
The volume of data is expected to grow 50% per year, and data production by 2020 will be 50 times larger than what it was in 2009. This rapid increase in volume requires powerful tools and techniques to process big data (Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V., 2016). The conventional tools such as SQL are unable to process it due to high volume, velocity and veracity of data. With such a diversification of data, ACID properties (Atomicity, Consistency, Integrity, and Durability) of databases are very difficult to meet using conventional tools; also desired outcome is difficult to produce within a reasonable frame of time period.

Secondly, most of the data are being generated in semi-structured or unstructured format in the form of images, text, audio, video and mails. Traditional tools are mainly designed to deal with structured data only. Therefore, new and advanced technologies have been devised to cope up the processing of big data in batches. In the next section, Hadoop based technologies to handle this increasing amount data has been discussed.

Database Technologies for Big Data Based on Hadoop

Apache Hadoop is one of widely used open source batch processing software for big data. Hadoop serves the basis for software that aim to work on parallel processing on large volume of data (Mishra, A.D., & Singh, Y.B., 2017). Hadoop works in two main phase i.e. storage and computation. Hadoop is assisted by two main components as shown in Figure 2, the first component is Hadoop distributed file system (HDFS) and the second component is MapReduce.

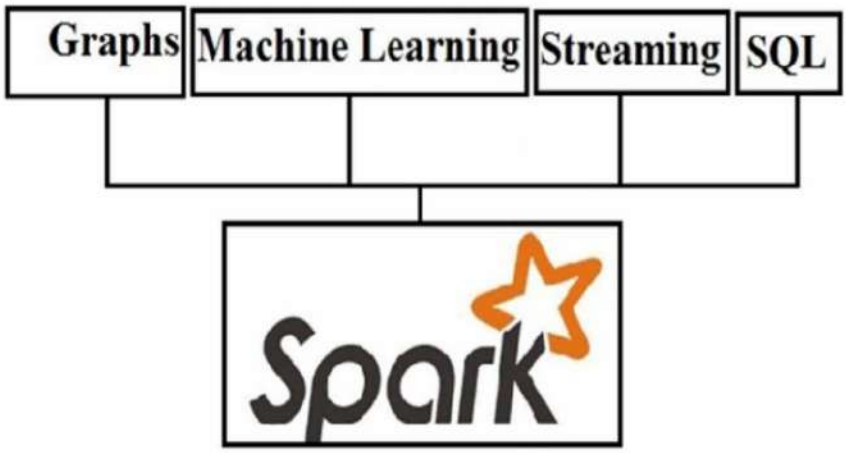
Figure 2. Hadoop component



HDFS allows a network of computers to form a cluster for data storage and processing. The MapReduce performs computation on stored data (Huang, W., Wang, H., Zhang, Y., & Zhang, S., 2017). The HDFS follows a master slave model to process data. The main issue with MapReduce is that it is unable to process iterative algorithms up to an optimum level. This section discusses some advanced technologies (Hadoop eco system) which have contributed in improving performance in batch processing of big data.

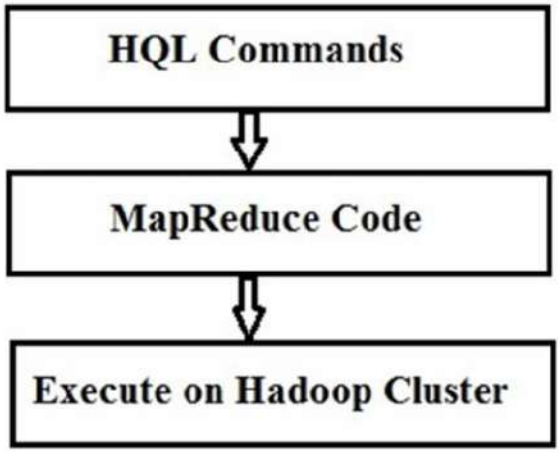
1. **Apache Spark:** Apache Spark is also a general purpose, distributed open source project that extends the capabilities of MapReduce by supporting processing of multiple data types such as SQL-like queries, streaming, machine learning, graph and data flow processing (Mavridis, I., & Karatza, H., 2017). Spark is considered to be very good for iterative as well as batch processing algorithms which processes data in memory. It reduces usages of disk by keeping data in memory during map and reduce phases. Spark has many higher level specialized library items to process specific kind of data as shown in Figure 3. Many programming tools such as Java, Python, R and Scala can be used for implementation of algorithms.

Figure 3. Spark with specialized library to process the data



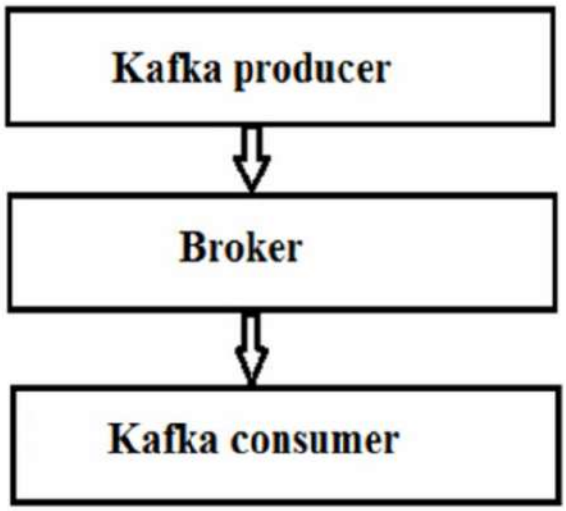
2. **Apache PIG:** Apache Pig is a scripting platform to process and analyze the large volume of data set present in Hadoop cluster (Kaur, R., Chauhan, V., & Mittal, U., 2018). The language used for scripting is known as PIG Latin. PIG runs the programs, convert it into the map reduce tasks, and finally executes the tasks. PIG is best suitable for the programming of the data in semi structured form.
3. **Apache Hive:** Hive is a Haddop eco-system tool which acts as an interface for the data warehouse for MapReduce programming. Hive has its own SQL, known as Hive query language (HQL). HQL is used to query data from the HDFS, generate MapReduce code and finally execute on Hadoop cluster as shown in Figure 4. Hive is not compatible with only HDFS, but also with Spark and other big data frameworks. Hive is fast, extensible and scalable, mainly developed for the OLAP (Mahmood, Z., 2016).
4. **HBase:** HBase is an open source, column-oriented, distributed, and non-relational database management system that runs on top of HDFS. HBase belong to the family of NoSQL database with the capability to handle massive amounts of data from terabytes to petabytes. Tables in Hbase are stored logically in the form of rows and columns. The benefit of such table storage is that they can process a million of rows and columns (Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S., 2018). It provides many features at low latency such as, natural language processing, real-time queries, linear and modular scalability, and consistent access to Big Data from various sources. However, HBase has the limitation of not supporting a structured query language like SQL.

Figure 4. Hive code execution



5. **Apache Kafka:** Apache Kafka is an open-source stream-processing software platform written in Scala and Java with an aim to provide low-latency platform for handling real-time data feeds with high throughput. The main components of Kafka architecture are Producer, Consumer, Broker and Topic (Vohra, D., 2016). In Kafka, a message is termed as the smallest unit of data that can flow from a producer to a consumer through a Kafka server (Broker) as shown in Figure 5. The message can persist on the server to be processed at a later time and feeds in topics. A topic, in Kafka, is a stream of messages of a similar category. In comparison to other messaging systems, Kafka has better built-in partitioning, replication, inherent fault-tolerance and throughput which make it one of best suitable platform for large-scale message processing applications.

Figure 5. Message flow in Apache Kafka



Whereas, worker nodes process assigned tasks for analysis (Achariya, D., & Kauser, A., 2016).

3. **Apache Spark streaming:** Spark is distributed platform stream processing is written in Java and Sacala. Spark has special libraries called Spark Streaming to support the stream processing with short latency. An Apache Storm topology consumes streams of data; repartition the streams between each stage of the computation for real time processing. It is primarily based on micro-batch processing mode where events are processed together based on specified time intervals. Spark has three main components; the driver program responsible for the proper scheduling the task and creating spark context; cluster mangers are responsible for the resource allocation between applications; task managers responsible for computation and storage. The processing rate of Spark is lower as compared to Strom and Flink due to formation of micro-batch before processing (Hesse, G., & Lorenz, M., 2016).

CLOUD COMPUTING IN DATA PROCESSING

As discussed, 4Vs has posed many challenges in efficient processing of big data. Now, need of the hour is transformation of 4Vs in to 5Vs. Value is big issue for the processing capacity (Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y., 2017). Cloud computing has become an amazing computation utility to address issues associated with big data with on demand service, ubiquitous network access, location independent resource pooling, rapid expansion and metered services (Verma, D.C., Mohapatra, A.K., & Usmani, K., 2012).

The rapid development in virtualization has made computation more economical sharable and accessible. Cloud computing eliminate the need of expensive resources such as processor, storage, operating system and memory for the large scale processing and complex computation. Large amount of data from the web and cloud are kept in a fault-tolerant distributed database and processed by a programming model for large volume of dataset with the help of parallel distributed algorithm in a cluster (Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S., 2015).

After processing of a large dataset, data visualization is used to present results in different graphs for decision making. The Figure 7 depicts the use of cloud computing for big data processing and analysis. Data sources in Figure 7 represent main contributor of data such as web, IOT, sensors and cloud. The main components of cloud data processing are: fault tolerant databases to store captured data, programming data model to process the clustered data through parallel computing and the query engine to execute queries.