Lecture Notes on Data Engineering and Communications Technologies 54

Ashish Khanna · Deepak Gupta · Zdzisław Pólkowski · Siddhartha Bhattacharyya · Oscar Castillo *Editors*

Data Analytics and Management

Proceedings of ICDAM

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 54

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/15362

Ashish Khanna · Deepak Gupta ·
Zdzisław Pólkowski · Siddhartha Bhattacharyya ·
Oscar Castillo
Editors

Data Analytics and Management

Proceedings of ICDAM

Springer

Editors
Ashish Khanna
Maharaja Agrasen Institute of Technology
New Delhi, India

Deepak Gupta
Maharaja Agrasen Institute of Technology
New Delhi, India

Zdzisław Pólkowski
Jan Wyzykowski University
Polkowice, Poland

Siddhartha Bhattacharyya
CHRIST (Deemed to be University)
Bengaluru, India

Oscar Castillo
Tijuana Institute of Technology
Tijuana, Mexico

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-15-8334-6 ISBN 978-981-15-8335-3 (eBook)
https://doi.org/10.1007/978-981-15-8335-3

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dr. Ashish Khanna would like to dedicate this book to his mentors Dr. A. K. Singh and Dr. Abhishek Swaroop for their constant encouragement and guidance and his family members including his mother, wife and kids. He would also like to dedicate this work to his (Late) father Sh. R. C. Khanna with folded hands for his constant blessings.

Dr. Deepak Gupta would like to dedicate this book to his father Sh. R. K. Gupta, his mother Smt. Geeta Gupta for their constant encouragement, his family members including his wife, brothers, sisters, kids and to his students who are close to his heart.

Dr. Zdzisław Pólkowski would like to dedicate this book to his wife, daughter and parents.

Prof. (Dr.) Siddhartha Bhattacharyya would like to dedicate this book to his father Late Ajit Kumar Bhattacharyya, his mother Late Hashi Bhattacharyya, his beloved wife Rashni and his colleagues Jayanta Biswas and Debabrata Samanta.

ICDAM-2020 Committees

Steering Committee Members

Patrons

Dr. Tadeusz Kierzyk, Professor of UJW, Rector of Jan Wyzykowski University, Polkowice, Poland
Lukasz Puzniecki, Mayor of Polkowice, Poland
Mr. Rajeev Jain, Chairman, B. M. Institute of Engineering and Technology, India
Mr. Rakesh Kuchhal, Manager BOG, B. M. Institute of Engineering and Technology, India

General Chairs

Prof. Dr. Janusz Kacprzyk, Polish Academy of Sciences, Systems Research Institute, Poland
Prof. Dr. Cesare Alippi, Polytechnic University of Milan, Italy
Prof. Dr. Siddhartha Bhattacharyya, CHRIST (Deemed to be University), Bengaluru, India

Honorary Chairs

Prof. Dr. Aboul Ella Hassanien, Cairo University, Egypt
Prof. Dr. Vaclav Snasel, Rector, VSB-Technical University of Ostrava, Czech Republic

Conference Chairs

Dr. Zdzisław Pólkowski, Professor of UJW, Jan Wyzykowski University, Polkowice, Poland
Prof. Dr. Harish Mittal, Principal, B. M. Institute of Engineering and Technology, India
Prof. Dr. Joel J. P. C. Rodrigues, Universidade Estadual do Piau Teresina, Brazil

Prof. Dr. Abhishek Swaroop, Bhagwan Parshuram Institute of Technology, Delhi, India
Prof. Dr. Anil K. Ahlawat, KIET Group of Institutes, Ghaziabad, India

Technical Programme Chairs

Dr. Stanislaw Piesiak, Professor of UJW, Jan Wyzykowski University, Polkowice, Poland
Dr. Jan Walczak, Jan Wyzykowski University, Polkowice, Poland
Dr. Anna Wojciechowicz, Jan Wyzykowski University, Lubin, Poland
Dr. Abhinav Juneja, Associate Director, B. M. Institute of Engineering and Technology, India
Dr. Vishal Jain, Dean Academics, B. M. Institute of Engineering and Technology, India

Technical Programme Co-chairs

Prof. Dr. Victor Hugo C. de Albuquerque, Universidade de Fortaleza, Brazil
Dr. Gulshan Shrivastava, National Institute of Technology Patna, India

Conveners

Dr. Ashish Khanna, Maharaja Agrasen Institute of Technology (GGSIPU), New Delhi, India
Dr. Deepak Gupta, Maharaja Agrasen Institute of Technology (GGSIPU), New Delhi, India
Dr. Pradeep Kumar Mallick, KIIT Deemed to be University Bhubaneswar, Odisha, India
Dr. Akash Kumar Bhoi, Sikkim Manipal University, India

Publication Chair

Dr. Jerzy Widerski, Professor of UJW, V-Rector of Jan Wyzykowski University, Polkowice, Poland

Publicity Chairs

Dr. Pawel Gren, Professor of UJW, V-Rector of Jan Wyzykowski University, Polkowice, Poland
Dr. Aditya Khamparia, Lovely Professional University, Punjab, India

Co-convener

Mr. Moolchand Sharma, Maharaja Agrasen Institute of Technology, India

Local Committee

Leszek Cybulski, Director of Jan Wyzykowski University, Polkowice, Poland

Advisory Committee

Dr. Tadeusz Kierzyk, Professor of UJW, Rector of Jan Wyzykowski University, Polkowice, Poland
Prof. Vincenzo Piuri, University of Milan, Italy
Prof. Aboul Ella Hassanien, Cario University, Egypt
Prof. Marcin Paprzycki, Polish Academy of Science, Poland
Prof. Valentina Emilia Balas, Aurel Vlaicu University of Arad, Romania
Prof. Marius Balas, Aurel Vlaicu University of Arad, Romania
Prof. Mohamed Salim Bouhlel, Sfax University, Tunisia
Prof. Cenap Ozel, King Abdulaziz University, Saudi Arabia
Prof. Ashiq Anjum, University of Derby, Bristol, UK
Prof. Mischa Dohler, King's College London, UK
Prof. David Camacho, Universidad Autonoma de Madrid, Spain
Prof. Parmanand, Dean, Galgotias University, Uttar Pradesh, India
Prof. Maryna Yena, Medical University of Kiev, Ukraine
Prof. Giorgos Karagiannidis, Aristotle University of Thessaloniki, Greece
Prof. Tanuja Srivastava, Department of Mathematics, IIT Roorkee
Dr. D. Jude Hemanth, Karunya University, Coimbatore
Prof. Tiziana Catarci, Sapienza University, Rome, Italy
Prof. Salvatore Gaglio, University Degli Studi di Palermo, Italy
Prof. Bozidar Klicek, University of Zagreb, Croatia
Prof. A. K. Singh, NIT Kurukshetra, India
Prof. Anil Kumar, KIET Group of Institutes, India
Prof. Chang-Shing Lee, National University of Tainan, Taiwan
Dr. Paolo Bellavista, Alma Mater Studiorum–Università di Bologna
Prof. Sanjay Misra, Covenant University, Nigeria
Prof. Benatiallah Ali, Adrar University, Algeria
Prof. Suresh Chandra Satapathy, PVPSIT, Vijayawada, India
Prof. Marylene Saldon-Eder, Mindanao University of Science and Technology
Prof. Özlem Onay, Anadolu University, Eskisehir, Turkey
Prof. Kei Eguchi, Department of Information Electronics, Fukuoka Institute of Technology
Prof. Zoltan Horvath, Kasetsart University
Dr. A. K. M. Matiul Alam Vancouver British Columbia, Canada
Prof. Joong Hoon Jay Kim, Korea University
Prof. Sheng-Lung Peng, National Dong Swa Uinversity, Taiwan
Dr. Dusanka Boskovic, University of Sarajevo, Sarajevo
Dr. Periklis Chat Zimisios, Alexander TEI of Thessaloniki, Greece
Dr. Nhu Gia Nguyen, Duy Tan University, Vietnam
Dr. Ahmed Faheem Zobaa, Brunel University, London
Prof. Ladjel Bellatreche, Poitiers University, France
Prof. Victor C. M. Leung, The University of British Columbia, Canada
Prof. Huseyin Irmak, Cankiri Karatekin University, Turkey

Dr. Alex Norta, Tallinn University of Technology, Estonia
Prof. Amit Prakash Singh, GGSIPU, Delhi, India
Prof. Abhishek Swaroop, Bhagwan Parshuram Institute of Technology, Delhi
Prof. Christos Douligeris, University of Piraeus, Greece
Dr. Brett Edward Trusko, President and CEO (IAOIP) and Assistant Professor, Texas A&M University, Texas
Prof. Joel J. P. C. Rodrigues, National Institute of Telecommunications (Inatel), Brazil; Instituto de Telecomunicações, Portugal
Prof. Victor Hugo C. de Albuquerque, University of Fortaleza (UNIFOR), Brazil
Dr. Atta ur Rehman Khan, King Saud University, Riyadh
Dr. João Manuel R. S. Tavares, FEUP-DEMec
Prof. Ku Ruhana Ku Mahamud, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia
Prof. Ghasem D. Najafpour, Babol Noshirvani University of Technology, Iran
Prof. Sanjeevikumar Padmanaban, Aalborg University, Denmark
Prof. Frede Blaabjerg, President (IEEE Power Electronics Society) Aalborg University, Denmark
Prof. Jens Bo Holm Nielson, Aalborg University, Denmark
Dr. Abu Yousuf, University Malaysia Pahang Gambang, Malaysia
Dr. Ahmed A. Elngar, Faculty of Computers and Information, Beni-Suef University, Egypt
Prof. Dijana Oreski, Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia
Prof. Prasad K. Bhaskaran, Ocean Engineering and Naval Architecture, IIT Kharagpur
Dr. Yousaf Bin Zikria, Yeungnam University, South Korea
Dr. Sanjay Sood, C-DAC, Mohali
Prof. Ajay Rana, Senior Vice President and Advisor—Amity Education Group, Amity University, Noida, India
Dr. Florin Popentiu Vladicescu, University Politehnica of Bucharest, Romania
Dr. Pawel Gren, Professor of UJW, Jan Wyzykowski University, Polkowice, Poland
Prof. Joanna Jozefowska, Pro-Rector for Research (etc.) of Poznan University of Technology
Prof. Gerhard-Wilhelm Weber, Poznan University of Technology, Poland
Prof. Dr. Sung-Bae Cho, Yonsei University, South Korea
Prof. Carlos A. Coello Coello, CINVESTA, Mexico
Dr. L. C. Jain, Founder KES International and Adjunct Professor University of Canberra, Australia
Dr. Debahuti Mishra, ITER, SOA University, Odisha, India
Dr. Ebrahim Aghajari, Islamic Azad University of Ahvaz, (IAUA), Iran
Dr. Hongyan Yu, Department of Computer Science, Shanghai Maritime University, Shanghai, China
Dr. Benson Edwin, Raj Higher College of Technology Fujairah Women's College, United Arab Emirates

Dr. Mohd. Hussain, Faculty of Computer Science and Information System, Islamic University, Medina, Saudi Arabia
Dr. Vahid Esmaeelzadeh, Department of Computer Engineering, Iran University of Science and Technology, Narmak, Tehran, Iran
Dr. Avinash Konkani, Clinical Engineer, University of Virginia Health System Charlottesville, VA, USA
Prof. Yu-Min Wang, National Chi Nan University, Taiwan
Dr. Ganesh R. Naik, Centre for Health Technologies (CHT), Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia
Dr. Yiguang Liu, Institutes of Image and Graphics College of Computer Science and Engineering, Yihuan Road, Chengdu Sichuan, China
Dr. Karol Morawski, The Karkonosze University of Applied Sciences, Jelenia Gora, Poland

Preface

We hereby are delighted to announce that Jan Wyzykowski University, Polkowice, Poland, and B.M. Institute of Engineering and Technology, Haryana, India, have hosted the eagerly awaited and much coveted International Conference on Data Analytics and Management (ICDAM-2020). The first version of the conference was able to attract a diverse range of engineering practitioners, academicians, scholars and industry delegates, with the reception of abstracts including more than 1540 authors from different parts of the world. The committee of professionals dedicated towards the conference is striving to achieve a high-quality technical programme with tracks on data analytics, data management, big data, computational intelligence and communication networks. All the tracks chosen in the conference are interrelated and are very famous amongst present-day research community. Therefore, a lot of research is happening in the above-mentioned tracks and their related sub-areas. More than 380 full-length papers have been received, among which the contributions are focused on theoretical, computer simulation-based research and laboratory-scale experiments. Amongst these manuscripts, 70 papers have been included in the Springer proceedings after a thorough two-stage review and editing process. All the manuscripts submitted to the ICDAM-2020 were peer reviewed by at least two independent reviewers, who were provided with a detailed review proforma. The comments from the reviewers were communicated to the authors, who incorporated the suggestions in their revised manuscripts. The recommendations from two reviewers were taken into consideration while selecting a manuscript for inclusion in the proceedings. The exhaustiveness of the review process is evident, given the large number of articles received addressing a wide range of research areas. The stringent review process ensured that each published manuscript met the rigorous academic and scientific standards. It is an exalting experience to finally see these elite contributions materialize into a book volume as ICDAM proceedings by Springer entitled "Data Analytics and Management: Proceedings of ICDAM".

ICDAM-2020 invited three keynote speakers, who are eminent researchers in the field of computer science and engineering, from different parts of the world. In addition to the plenary sessions on each day of the conference, six concurrent

technical sessions are held every day to assure the oral presentation of around 70 accepted papers. Keynote speakers and session chair(s) for each of the concurrent sessions have been leading researchers from the thematic area of the session. A technical exhibition is held during the conference, which has put on display the latest technologies, expositions, ideas and presentations. The delegates were provided with a book of extended abstracts to quickly browse through the contents, participate in the presentations and provide access to a broad audience of the audience. The research part of the conference was organized in a total of six special sessions. These special sessions provided the opportunity for researchers conducting research in specific areas to present their results in a more focused environment.

An international conference of such magnitude and release of the ICDAM-2020 proceedings by Springer has been the remarkable outcome of the untiring efforts of the entire organizing team. The success of an event undoubtedly involves the painstaking efforts of several contributors at different stages, dictated by their devotion and sincerity. Fortunately, since the beginning of its journey, ICDAM-2020 has received support and contributions from every corner. We thank them all who have wished the best for ICDAM-2020 and contributed by any means towards its success. The edited proceedings volumes by Springer would not have been possible without the perseverance of all the steering, advisory and technical programme committee members.

All the contributing authors owe thanks from the organizers of ICDAM-2020 for their interest and exceptional articles. We would also like to thank the authors of the papers for adhering to the time schedule and for incorporating the review comments. We wish to extend our heartfelt acknowledgment to the authors, peer reviewers, committee members and production staff whose diligent work put shape to the ICDAM-2020 proceedings. We especially want to thank our dedicated team of peer reviewers who volunteered for the arduous and tedious step of quality checking and critique on the submitted manuscripts. We wish to thank our faculty colleague Mr. Moolchand Sharma for extending their enormous assistance during the conference. The time spent by them and the midnight oil burnt is greatly appreciated, for which we will ever remain indebted. The management, faculties, administrative and support staff of the college have always been extending their services whenever needed, for which we remain thankful to them.

Lastly, we would like to thank Springer for accepting our proposal for publishing the ICDAM-2020 conference proceedings. Help received from Mr. Aninda Bose, the acquisition senior editor, in the process has been very useful.

New Delhi, India

Ashish Khanna
Deepak Gupta
Organizers, ICDAM-2020

About This Book

International Conference on Data Analytics and Management (ICDAM-2020) was held on 18 June via virtual mode and jointly organized by Jan Wyzykowski University, Polkowice, Poland, and B. M. Institute of Engineering and Technology, Haryana, India. This conference was able to attract a diverse range of engineering practitioners, academicians, scholars and industry delegates, with the reception of papers including more than 1540 authors from different parts of the world. Only 70 papers have been accepted and registered with an acceptance ratio of 18% to be published in one volumes of prestigious Springer Lecture Notes on Data Engineering and Communications Technologies series.

Contents

About the Editors

Dr. Ashish Khanna has expertise in Teaching, Entrepreneurship, and Research & Development of 16 years. He received his Ph.D. degree from National Institute of Technology, Kurukshetra, in March 2017. He has completed his M.Tech. and B.Tech. from GGSIPU, Delhi. He has completed his PDF from the Internet of Things Lab at Inatel, Brazil. He has around 100 research papers along with book chapters including more than 40 papers in SCI Indexed Journals with cumulative impact factor of above 100 to his credit. Additionally, he has authored and edited 19 books. Furthermore, he has served the research field as a Keynote Speaker/Session Chair/Reviewer/TPC member/ Guest Editor, and many more positions in various conferences and journals. His research interest includes image processing, distributed systems and its variants, and machine learning. He is currently working at the CSE, Maharaja Agrasen Institute of Technology, Delhi. He is Convener and Organizer of ICICC Springer conference series.

Dr. Deepak Gupta is an eminent academician and plays versatile roles and responsibilities juggling between lectures, research, publications, consultancy, community service, Ph.D. and postdoctorate supervision, etc. With 12 years of rich expertise in teaching and two years in industry, he focuses on rational and practical learning. He has contributed massive literature in the fields of–computer interaction, intelligent data analysis, nature-inspired computing, machine learning, and soft computing. He is working as an Assistant Professor at Maharaja Agrasen Institute of Technology (GGSIPU), Delhi, India. He has served as Editor-in-Chief, Guest Editor, and Associate Editor in SCI and various other reputed journals. He has authored/edited 33 books with national/international level publisher (Elsevier, Springer, Wiley, Katson). He has published 96 scientific research publications in reputed international journals and conferences including 46 SCI Indexed Journals of IEEE, Elsevier, Springer, Wiley, and many more.

Dr. Zdzisław Pólkowski is an Adjunct Professor at Faculty of Technical Sciences at the Jan Wyzykowski University, Poland. He is also the Rector's Representative for International Cooperation and Erasmus Programme and Former Dean of the

Technical Sciences Faculty during the period of 2009–2012. His area of research includes management information systems, business informatics, IT in business and administration, IT security, small medium enterprises, CC, IoT, big data, business intelligence, and block chain. He has published around 60 research articles. He has served the research community in the capacity of Author, Professor, Reviewer, Keynote Speaker, and Co-Editor. He has attended several international conferences in the various parts of the world. He is also playing the role Principal Investigator.

Dr. Siddhartha Bhattacharyya [FIEI, FIETE, LFOSI, SMIEEE, SMACM, SMIETI, LMCSI, LMISTE, MIET (UK), MIAENG, MIRSS, MIAASSE, MCSTA, MIDES, MISSIP, MSDIWC] is currently the Principal of RCC Institute of Information Technology, Kolkata, India. In addition, he is also serving as the Professor of Computer Application and Dean (Research and Development and Academic Affairs) of the institute. He is the recipient of several coveted awards like Adarsh Vidya Saraswati Rashtriya Puraskar, Distinguished HoD Award, Distinguished Professor Award, Bhartiya Shiksha Ratan Award, Best Faculty for Research, and Rashtriya Shiksha Gaurav Puraskar. He received the Honorary Doctorate Award (D. Litt.) from The University of South America and the SEARCC International Digital Award ICT Educator of the Year in 2017. He has been appointed as the ACM Distinguished Speaker for the tenure 2018–2020. He is a co-author of 4 books and the co-editor of 16 books and has more than 200 research publications in international journals and conference proceedings to his credit.

Dr. Oscar Castillo holds the Doctor in Science degree (Doctor Habilitatus) in Computer Science from the Polish Academy of Sciences (with the Dissertation "Soft Computing and Fractal Theory for Intelligent Manufacturing"). He is a Professor of Computer Science in the Graduate Division, Tijuana Institute of Technology, Tijuana, Mexico. Currently, he is President of HAFSA (Hispanic American Fuzzy Systems Association) and Past President of IFSA (International Fuzzy Systems Association). Prof. Castillo is also Chair of the Mexican Chapter of the Computational Intelligence Society (IEEE). His research interests are in Type-2 Fuzzy Logic, Fuzzy Control, Neuro-Fuzzy, and Genetic-Fuzzy hybrid approaches. He has published over 300 journal papers, 10 authored books, 40 edited books, 200 papers in conference proceedings, and more than 300 chapters in edited books, in total 865 publications according to Scopus (H index=60), and more than 1000 publications according to Research Gate (H index=72 in Google Scholar).

A Forecasting-Based DLP Approach for Data Security

Kishu Gupta and Ashwani Kush

Abstract Sensitive data leakage is the major growing problem being faced by enterprises in this technical era. Data leakage causes severe threats for organization of data safety which badly affects the reputation of organizations. Data leakage is the flow of sensitive data/information from any data holder to an unauthorized destination. Data leak prevention (DLP) is set of techniques that try to alleviate the threats which may hinder data security. DLP unveils guilty user responsible for data leakage and ensures that user without appropriate permission cannot access sensitive data and also provides protection to sensitive data if sensitive data is shared accidentally. In this paper, data leakage prevention (DLP) model is used to restrict/grant data access permission to user, based on the forecast of their access to data. This study provides a DLP solution using data statistical analysis to forecast the data access possibilities of any user in future based on the access to data in the past. The proposed approach makes use of renowned simple piecewise linear function for learning/training to model. The results show that the proposed DLP approach with high level of precision can correctly classify between users even in cases of extreme data access.

Keywords Data leakage · Data leakage prevention · Forecast · Guilty agent · Statistical analysis

1 Introduction

The NIST explains computer security as "protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability, and confidentiality of information system resources (includes hardware,

K. Gupta (✉)
Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India
e-mail: kishugupta2@gmail.com

A. Kush
Institute of Integrated and Hons. Studies, Kurukshetra University, Kurukshetra, India
e-mail: akush20@gmail.com

A. Khanna et al. (eds.), *Data Analytics and Management*, Lecture Notes on Data Engineering and Communications Technologies 54,
https://doi.org/10.1007/978-981-15-8335-3_1

software, firmware, information/data, and telecommunications)" [1, 2]. Advancement in technology allows easy and speedy transfer of data. Data is the key to conduct business activities nowadays, and hence, a need arises to share data among various stakeholders/third parties like human resources working from outside the site (e.g., on laptops), business colleague, and clients [3]. For example, service provider requires access to the company intellectual property and other confidential information to carry out their services [4, 5].

Data loss/leakage has emerged as the biggest threat that organizations are facing today. In the present scenario, almost all business activities depend on extensive sharing of sensitive/confidential data, within or outside the organization [6, 7]. Data leakage is an event that may occur either accidentally or maliciously that permits data access to unauthorized user. Sensitive data loss/leakage rigorously hampers reputation of organization, confidence/faith of customers in company which may ultimately lead to shut down company or even may lead to severe political crisis like WikiLeak leaks [8]. Leakage is thus a subset of data loss with a spotlight on the data security goal.

To minimize the risk of data loss, organizations usually make use of DLP solutions as a protection/defense mechanism. Prior to DLP security, mechanisms/technologies like firewalls and IDS were in use [9]. DLPSs are used to protect all kind of data, that is, data in use, data at rest and data in transit. DLPSs use the statistical/ analytical approach, data fingerprinting, regular expressions on context and content of data to identify and avoid unauthorized access to data [10]. DLP approach performs deep content analysis and observes the data access by users to discover improper usage [11–13]. DLP systems employ a model using either knowledge of an expert or may train/learn from available past records (Fig. 1).

This scenario provides ample space to produce requirement of a mechanism that can identify leakage with more precision for greater data security. The proposed DLP model tries to provide data security by observing users' trend to access the date, uses learning-based approach to highlight the user whoever is performing different

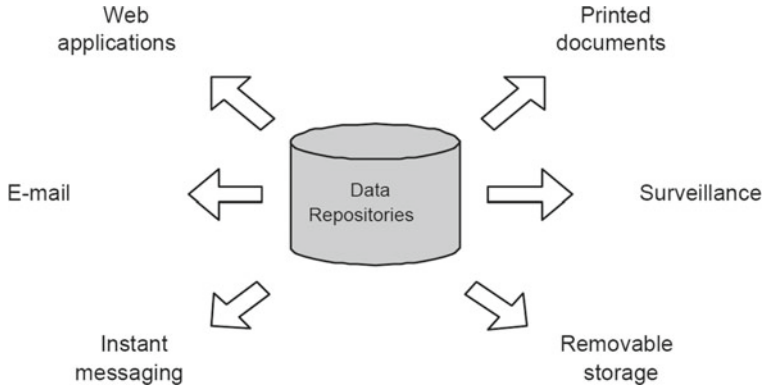


Fig. 1 Possible leak channels [14]

to trend observed. This enables organization to take suitable action like imposing access restriction on sensitive data for data security.

The paper layout is as follows: in Sect. 2, an overview of DLP solutions related to work has been presented. Furthermore, an overview of the proposed data fitting model framework has been discussed in Sect. 3. Section 4 deals with the experimental results, and Sect. 5 represents conclusion of paper, respectively.

2 Overview of DLP Approach

This section specifies major benefits of DLP solution. DLP model generally distinguishes suspicious activity from normal activity and performs either detection, i.e., raise alert if doubtful activity happens, or prevention, i.e., block nasty activity. DLP model can be characterized by various dimensions like model construction, filtering approach, network-based, host-based, etc. Model building approach to describe how the model is constructed is most relevant and best suited for the proposed work. Specification and learning-based are two approaches for model construction. Specification-based approach uses expert's knowledge and hence more precise/accurate, while learning-based model automatically learns using statistical techniques, i.e., machine learning.

The proposed framework reflects numerous benefits over existing solutions for DLP. First, learning-based framework tailors itself to the user's behavior and hence makes feasible to detect unknown and insider attacks. Additionally proposed DLP approach provides better control on data from being misused along with providing flexibility to access the data simultaneously. Moreover, this approach integrates data protection with user identity, thus making organization capable to implement data protection policy based on user identity and their role. The proposed approach tries to forecast guilty user based on available records of user access to the organization data. Finally, this forecasting-based framework proves to be more practical and efficient (Fig. 2).

3 Data Training Model

The proposed DLP model fulfills the objective of data security by employing machine learning-based approach and provides forecast for further action. The proposed model considers multiple agents (m) which have accessibility to organization data anytime and for any number of times. Each and every access to organization data by any user is entered in the form of a user's accessibility dataset containing many important details like date to access data, duration for which organization data was in use by particular user, i.e., y. The user's accessibility dataset continues to grow with time and is fitted and trained by machine learning techniques to obtain trend. Trend in this study determines the pattern of data accessibility by various users along

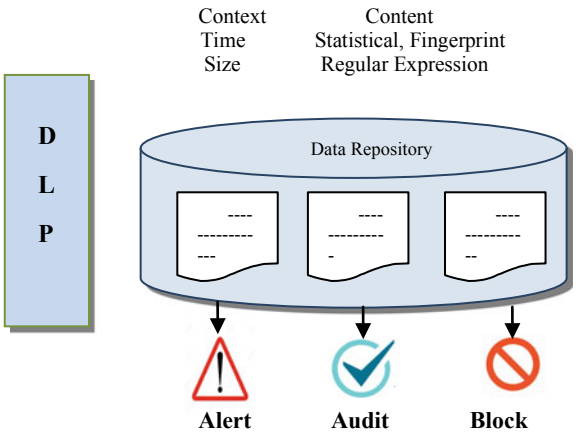


Fig. 2 DLP overview

a period of time. Thus, the proposed model provides future insight by studying the past events. This section uses dataset of 2014 to 2018 to train the model and then predicts the trend of particular user after 2018. If the trend exceeds the defined upper limit, then it raises alert to prevent or restrict the user access to organization data. The discussed approach is explained through Fig. 3.

3.1 Model Equations

The model uses a simple piecewise linear model-based function as shown in Eq. (1) named as model fitting equation.

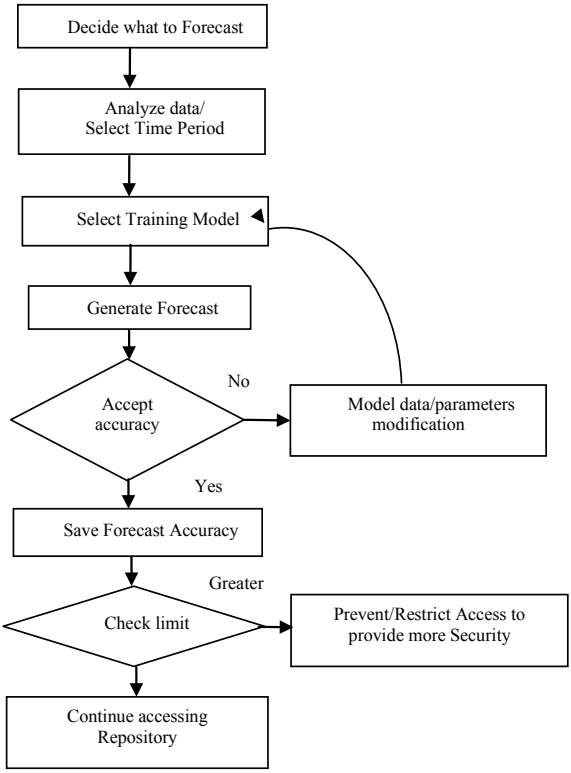
$$g(t) = (k + \alpha(t)'\delta)t + (m + \alpha(t)'\gamma)\ . \tag{1}$$

The above model fitting equation is used to fit and train the user's accessibility dataset to evaluate the predicted time value to be spent by user $\hat{y}$ along with many other parameters required for study. The model executes all entries in user accessibility database, i.e., $i = 1$ to n.

3.2 Model Accuracy

After obtaining predicted time value, the model calculates the error existing between actual and predicted value to determine the accuracy of approach. Here, ϵ_i shown in Eq. (2A) is error existing between actual and predicted values of time spend by user to access data. £$_i$ is percentage error represented by Eq. (2B).

Fig. 3 Data training model



$$\epsilon_i = \left((y_i + \alpha) - (\hat{y}_i - \alpha)\right) \tag{2A}$$

$$\epsilon_i = 100 * \frac{\epsilon_i}{(y_i + \alpha)}. \tag{2B}$$

3.3 Calculating Limit

To define the upper bound and lower bound for a user to access the data, the study computes $Æ_i$, L_i and υ_i in Eqs. (3A, 3B and 3C), respectively.

$$Æi = |\varepsilon i|. \tag{3A}$$

$$\mathcal{L}\mathrm{i} = \left(\hat{\mathrm{y}}\mathrm{i} - (\mu * \varsigma * \sigma)\right). \tag{3B}$$

$$\mathrm{Ui} = \left(\hat{\mathrm{y}}\mathrm{i} + (\mu * \varsigma * \sigma)\right). \tag{3C}$$

$Æ_i$ is absolute error, i.e., nonnegative value of error as calculated in Eq. (2A). L_i is lower bound and υ_i is upper bound for users to access data. Here, μ and σ are mean and standard deviation of Æ calculated as shown below.

$$\mu = \frac{\sum_{i=1}^{n} Æi}{n} \text{ and } \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Æi - \overline{Æ})^2}$$

4 Experimental Result

The proposed model is implemented using Python on Jupyter Notebook platform in Anaconda environment to conduct the experiments. Figure 4a shows the overall half yearly forecast, and Fig. 4b shows overall annual forecast; it is observed that from 2014 to 2018, data was accessed for less than 80 min, given a year or half yearly. These graphs show access time forecast of overall system, i.e., all users together. Separate graphs for particular users can also be generated. Forecasting results depict that some user is going to access the data for longer duration in 2019 which is unusual as compared to the previous years. Based on trend, it can be a scenario/possibility of data leakage; hence, more restriction can be imposed to the user by checking on the access time limit for particular user. This is how this model will help to prevent data leakage for database being in use by multiple users and multiple repositories.

5 Conclusion

The biggest challenge in the present era is to shield sensitive data from leakage, which imposes a big threat for organization's growth/security/health. The paper highlights DLP approach with better control on data to protect data from being misused and also provide flexibility to access the data simultaneously. Moreover, this approach integrates data protection with user identity, hence enabling organization to enforce data protection policy based on user identity and their role. The proposed approach tries to forecast seems to be guilty user based on available records of user access to the organization data. On basis of model prediction output; access rights of any particular user can be restricted or blocked completely, hence proposed model provides enhanced data security for organization data. Thus, the proposed model provides future insight by studying the past events. Conclusion drawn from this study is that the system based on forecasting approach to identify guilty user is more practical and efficient.

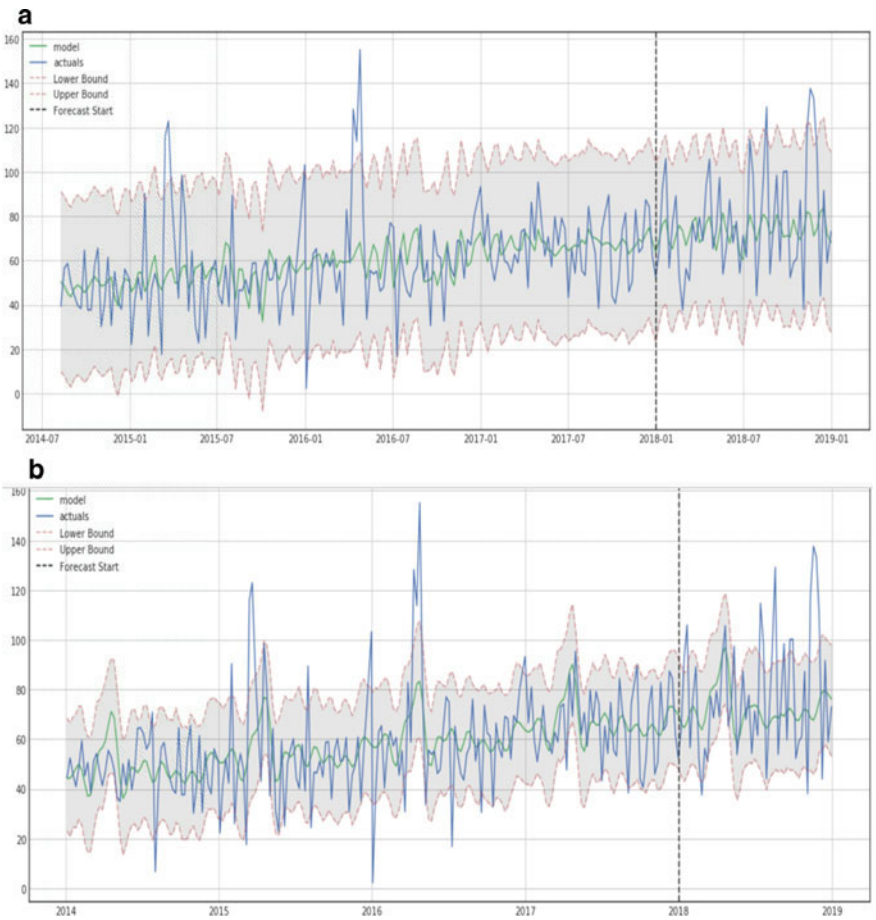


Fig. 4 **a** Overall half yearly forecast, **b** overall annual forecast

References

1. Gupta K, Kush A (2017) A review on data leakage detection for secure. Int J Eng Adv Technol (IJEAT) 7(1):153–159
2. Gupta K, Kush A (2018) Performance evaluation on data leakage detection for secure communication. In: 5th international conference on computing for sustainable global development: INDIACom, New Delhi, India
3. Alneyadi S, Sithirasenan E, Muthukkumarasamy V (2016) A survey on data leakage prevention systems. Elsevier J Netw Comput Appl 62(1):137–152
4. Papadimitriou P, Molina H (2011) Data leakage detection. IEEE Trans Knowl Data Eng 23(1):51–63
5. Sodagudi S, Kurra RR (2016) An approach to identify data leakage in secure communication. In: 2nd international conference on intelligent computing and applications, Singapore
6. Shabtai A, Elovici Y, Rokach L (2012) A survey of data leakage detection and prevention solutions. Springer, pp 1–87

7. Dattana V, Gupta K, Kush A (2019) A probability based model for big data security in smart city. In: International conference on big data and smart city (ICBDSC), Muscat
8. Liu F, Shu X, Yao D, Butt A (2015) Privacy-preserving scanning of big content for sensitive data exposure with MapReduce. In: 5th ACM conference data application security, privacy (CODASPY), Texas, USA
9. Shu X, Yao D (2015) Privacy-preserving detection of sensitive data exposure. IEEE Trans Inform Forensics Sec 10(5):1092–1103
10. Shu X, Zhang J, Yao D, Feng WC (2016) Fast detection of transformed data leaks. IEEE Trans Inform Forensics Sec 11(3):528–542
11. Fin Y, Lina W, Rongwei Y, Xiaoyan M (2013) A distribution model for data leakage prevention. In: International conference on mechatronic sciences, electric engineering and computer (MEC), IEEE, Shenyang, China
12. Peneti S, Rani B (2016) Data leakage prevention system with time stamp. In: International conference on information communication and embedded system (ICICES)
13. Kadu RS, Gadicha VB (2017) Review on seuring data by using data leakage prevention and detection. Int J Recent Innov Trends Comput Commun 5(5):731–735
14. Raman P, Kayacık HG, Somayaji A (2011) Understanding data leak prevention. In: 6th annual symposium on information assurance (ASIA'11), Albany, New York, USA