

A Study on Big Data Privacy Protection Models using Data Masking Methods

Archana R. A.¹, Ravindra S. Hegadi², Manjunath T. N.³

¹R& D Centre, Bharathiar University, Coimbatore, Tamil Nadu, India

²School of Computational Sciences, Solapur University, Maharashtra, India

³Dept of ISE, BMS Institute of Technology, Bangalore, Karnataka, India

Article Info

Article history:

Received Jan 26, 2018

Revised Apr 20, 2018

Accepted Jul 2, 2018

Keyword:

Big data privacy

Business domains

Data masking

Dynamic data masking

ABSTRACT

In today's predictive analytics world, data engineering play a vital role, data acquisition is carried out from various source systems and process as per the business applications and domain. Big Data integrates, governs, and secures big data with repeatable, reliable, and maintainable processes. Through volume, speed, and assortment of information characteristics try to reveal business esteem from enormous information. However, with information that is frequently deficient, conflicting, ungoverned, and unprotected, which is hazardous and enormous information being a risk instead of an advantage. What's more, with conventional methodologies that are manual and unpredictable, huge information ventures take too long to acknowledge business esteem. Reasonably and over and over again conveying business esteem from enormous information requires another technique. In this connection, raw data has to be moved between onsite and offshore environment during this course of action, data privacy is a major concern and challenge. A Big Data Privacy platform can make it easier to detect, investigate, assess, and remediate threats from intruders. We tried to do complete study of Big Data Privacy using data masking methods on various data loads and different types. This work will help data quality analyst and big data developers while building the big data applications.

*Copyright © 2018 Institute of Advanced Engineering and Science.
All rights reserved.*

Corresponding Author:

Archana R. A.,

R&D Centre,

Bharathiar University,

Coimbatore, Tamil Nadu, India.

Email: archana.tnm@gmail.com

1. INTRODUCTION

Big Data is growing from systems around us at faster rate, every second enormous amount of data is getting generated, and these data has characteristics of volume, variety and velocity. There is a need of big data management with respect to big data integration, Big data governance and quality of Big Data Privacy. In this connection data development and expansion, associations have poor perceivability into the area and utilization of their sensitive information. However security laws and directions require an exact comprehension of information hazard in view of different application domains and use crosswise over different frameworks [1], [2]. Enormous Data Privacy finds and arranges information to drive an exhaustive 360-degree perspective of the data for different purposes so you can group sensitive information with 360-degree perceivability De-distinguishes information so it can be securely utilized as a part of improvement and creation conditions. This ensures consistence with corporate approaches and industry directions data at the undertaking level, as a common administration [3].

The big data privacy framework provides a common infrastructure for development, testing and support, enabling scalability and repeatability across applications. We achieve reuse and scalability via the

use of process-oriented metadata that defines the way masking is to be carried out for each item of data in terms of subsetting, encryption, manipulation and so on. Used in conjunction with Extract Transform Load (ETL) tools and operational scripts, the metadata results in a completely standard masking process whenever anyone in the organization needs to mask a given piece of data for a given purpose. Tools we typically use are Informatica Power Center with Power Exchange for Extract Transform Load (ETL) with masking capability. Figure 1 shows the big data privacy protection model using data masking methods.

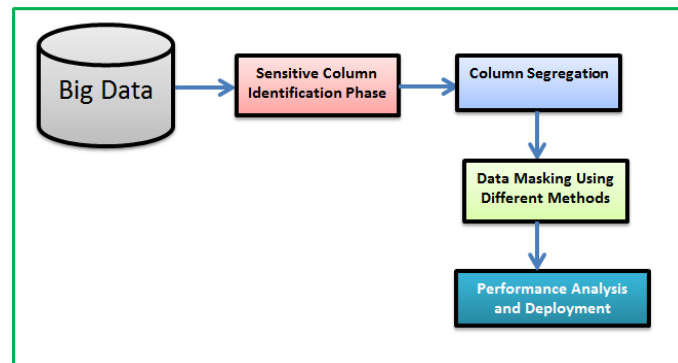


Figure 1. Big data privacy protection model using data masking methods

2. RESEARCH METHOD

As per the Ponemon Institute's 2015 State of Data Security Intelligence Report, IT's greatest stress not knowing where sensitive information dwells is just developing. Deliver fit-for-purpose big data, with a scalable, role-based data quality environment Big data transforms the way businesses innovate and improve their operational processes. However, as organizations begin to bring big data into their environments they struggle to make these projects pay off. One of the key difficulties is that information quality issues debase the uprightness and trust in enormous information resources. Any inquiry of information quality is a genuine, if not acquire mountable obstruction to an associations capacity to settle on shrewd choices, decrease costs, create development which advances development [3], [4].

Relevant, timely, and trustworthy data is essential for success. Informatica Big Data Quality empowers any organization to take a holistic approach to manage data quality by leveraging the power of Hadoop. This makes a genuine information privacy driven condition that backings better business conditions basic leadership and investigation paying little respect to your information's size, configuration, or stage. It conveys definitive, put stock in information privacy to all partners, tasks, and business applications on Hadoop, local or in the cloud. With Informatica Big Data Quality, you can enrich and standardize more data at scale, enable business and IT collaboration in the governance of data and prepare and share fit-for-use data into trusted insights Powerful Data Discovery and Profiling tools such as Informatica Big Data Quality which as set of unified, role-based data discovery and profiling for quickly identifying critical data problems hidden across the enterprise. Powerful and versatility of these tools allow business and IT to collaborate and quickly identify data quality issues, easily design and apply business rules and policies, as well as proactively monitor the data quality process [3], [4].

Informatica Analyst is a simple to-utilize, browser based instrument that engages the business to effortlessly take part in enhancing the nature of information, without the requirement for IT intercession. Rich Set of Data Quality Transformations and Universal Connectivity Informatica Big Data Quality should ensure confidence to all the stakeholders and should reconcile and sync any information [5]. It features standardization, matching, worldwide address cleansing, and versatile data quality management for all project types. The product additionally empowers you to send pre-constructed information quality guidelines to enhance quality over the venture. As indicated by Experian's 2015 Data Quality Benchmark Report, associations speculate 26% of their information to be incorrect and inconsistent [5].

3. BIG DATA MASKING

Focus on each field with an information security you select from twelve diverse insurance category based on your business guidelines. For instance say encryption and tokenization for credit card esteems, pseudonymization for names, randomization for a very long time, redaction for equations, and character covering on national ID esteems programming can find, arrange, and ensure sensitive information and

encourages security law consistence with the broadest cluster of static information concealing or dynamic information veiling capacities accessible for databases and records. On the off chance that you have PII in Excel spreadsheets, see the buddy item, IRI CellShield. Find and group sensitive information in different sources Encrypt with our consistent libraries, De-distinguish by means of covering characters or jumbling controls, Pseudonymize, encode, hash, randomize, tokenize, Filter or redact fields or records in view of conditions [6], [7].

FieldShield produces XML review logs you can secure and question to report and check your assurances and consistence with information protection laws. FieldShield can likewise cover information subsets for testing [8]. Not with standing, consider IRI RowGen for creating safe, referentially rectify test information starting with no outside help rather, particularly on the off chance that you can't get to generation information or need better information. Pick the assurance work you requirement for each field. Take after your own particular business rules with respect to: approval (RBAC), security quality, reversibility, and appearance. Secure like segments (and protect referential respectability) crosswise over tables with capacities attached to information class or administer libraries. Target existing or new tables, records, applications, and even custom reports. Set controls at the field and employment level for various beneficiaries (one target, differential access). Randomization is another approach to anonymize or de-distinguish actually identifiable data [8], [9]. Figure 2 shows the big data sets migrations – security need.

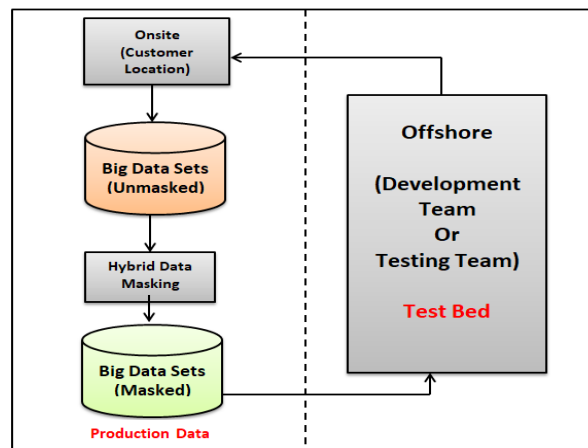


Figure 2. Big Data Sets Migrations – Security Need

FieldShield software in the IRI Data Protector suite provides simple access to non-reversible functions for: Random data generation - non-deterministic: replaces original field value with randomly generated data Random data selection - rarely deterministic: specific identifiers are pulled at random from the source, losing their association with other values in the original row. This dialog in the IRI Workbench GUI for FieldShield users can address any source file field or database column, along with other data masking functions [10], [11]. Data shuffling is also possible through custom random functions or match logic you can define in the IRI Workbench expression builder. Hashing is a hard to-turn around information covering system that changes over a variable length "message" (e.g., somebody's secret word) into a muddled, settled length, alphanumeric string.

The message process, or "hash esteem," can be a list search up for the message. Some of the time there is more than one message for each record (a "crash"). Since hashing isn't as solid as encryption, or as dependably reversible, it is at times reasonable for covering alone. All the more usually, be that as it may, hashing is utilized with encryption [12], [13]. Hash capacities are likewise used to create checksums or Message Authentication Codes (MAC). These are made and sent alongside messages like messages, EFTs, or passwords. At the point when the message is gotten, its substance is go through a similar hash capacity to make another MAC. In the event that the first and new MACs coordinate, the message is legitimate; on the off chance that they don't, the message is probably going to have been modified, and in this way traded off. Utilize the field-level hashing capacities in both FieldShield in the IRI Data Protector Suite, and CoSort in the IRI Data Manager suite, to help cover PII [14], [15].

Make a MAC for at least one section esteems in each line. Incorporate it as an extra field or give it in a different document. Utilize it to check that the information in the record was undisturbed. Huge

information veiling support on Hadoop use the qualities of both the Hadoop and Camouflage stages, enabling clients to cover at the size of Hadoop while exploiting the consistency and refinement of Camouflage concealing. Local incorporation with Hadoop gives coordinate guide/lessen execution and perceivability of veiling employments from straightforwardly inside the Hadoop condition. Put away information is covered reliably with every single other datum sources so that the conceal Hadoop condition remains in a state of harmony with other veiled information sources. The most recent arrival of Enterprise additionally presents an elite alternative that drastically speeds up social information veiling ventures, accomplishing execution picks up 15-80 times speedier than past cycles [16], [17].

Besides, execution improvements are naturally and shrewdly connected, boosting throughput without the requirement for arrangement changes, bringing about predominant usability and essentially diminished concealing invigorate windows. Information concealing is the most ideal approach to agree to information security laws, invalidate the impacts of an information rupture, and bolster the hazard and controls system of your enterprise. IRI FieldShield rapidly fulfills the information recognizable proof, assurance and confirmation prerequisites of your data stewardship, administrative consistence, and information misfortune anticipation programs. You can run FieldShield capacities independently using ETL [18].

3.1. Dynamic data masking

Dynamic information covering utilizes information insurance controls continuously to keep favored faculty, for example, DBAs, production staff members, and business clients from getting to delicate and actually identifiable data that isn't required for them to play out their employments. The estimation of dynamic information covering lies in its capacity to apply distinctive veils to various sorts of information found underlying databases, applications, and detailing and improvement instruments. Since covering is connected powerfully in view of client parts and benefit levels, just people with a need to see the completely uncovered information could do as such; all others see conceal information. In an open area association, this would imply that a DBA or unapproved client would not have the capacity to see genuine Social Security numbers, singular understudy evaluations, or citizens' altered balanced gross pay figures in light of the fact that these qualities and other by and by identifiable data would be specifically mixed, hashed, covered, or blocked.

Information covering can be utilized to stretch out insurance to unstructured and semistructured information. Tenacious information covering can likewise be utilized as a part of conjunction with encryption to make encoded information more sensible searching for advancement and testing purposes. In either case, concealing information rather than scrambling it applies almost no execution punishment. Informatica Dynamic Data Masking has been exhibited and demonstrated to secure delicate data without affecting database execution. This exceptional favorable position has engaged a worldwide versatile correspondences supplier to take a noteworthy jump toward keeping unapproved clients from getting to individual information.

Preceding actualizing Informatica Dynamic Data Masking, this supplier had been consistently ending a normal of three individuals for each month for getting to the classified data of its clients. This procedure was trading off the organization's operational productivity and harming its notoriety. While trying to address the issue, different methodologies were investigated yet none of them addressed the supplier's issues for security and execution. Encryption, for instance, was precluded because of execution debasement in the generation condition. It would have required various changes and nonstop updates to applications—a restrictive errand given that a significant number of the applications the association was utilizing were bundled with shut information models. The organization required a more hearty, superior approach. As depicted before in this paper, Informatica Dynamic Data Masking utilizes a straightforward visual execution system. This empowered the correspondences supplier to rapidly secure an abundance of individual distinguishing proof information in a few of the most perplexing and requesting business applications, including charging, Siebel, Clarify, and cloned applications. Informatica Dynamic Data Masking enabled individual data to be secured from the organization's business clients, recently enrolled and existing workers, contracted staff, and outsourced and IT staff enabling them all to get to that data while conforming to "have to-know" information get to arrangements [18], [19].

Notwithstanding significantly diminishing the danger of an information rupture, the product provided the correspondences supplier with the adaptability to rapidly alter information covering abilities for various administrative or business prerequisites. Manage spread outfitted fast assurance crosswise over basic generation, preparing, and nonproduction conditions. Also, consistence with security directions as accomplished cost-adequately and with no effect to database execution. Besides, the organization could sidestep costly and tedious changes to applications that would have brought about long advancement and testing forms. Remarking on the effect of Informatica Dynamic Data Masking programming, the

organization's main data security officer stated, "In only half a month, the Informatica Platform straightforwardly veiled individual data on our charging, CRM, and custom application screens and bundled reports underway and nonproduction situations. The Informatica programming is currently a foundation of our hazard administration and consistence methodology." because of these difficulties, associations are in more noteworthy need of powerful information concealing programming to anticipate breaks and uphold information security [18], [19].

Such an answer ought to enable IT associations to: Mask the sensitive information uncovered underway conditions, Shield creation applications and databases without changes to source code, Respond rapidly to decrease the dangers of information ruptures and the subsequent costs, Customize database security for various administrative or business necessities. Informatica Dynamic Data Masking encourages associations to achieve these overwhelming assignments, proactively tending to information security challenges continuously. As the main genuine dynamic information covering item available, Informatica Dynamic Data Masking de-recognizes information and controls unapproved access to generation situations. It has many focal points, among the key preferences of information covering, both constant and dynamic, is its flexibility from expecting changes to databases or application source code [19], [20].

This implies veiling can be connected rapidly and unpretentiously to secure private information over an association, paying little mind to estimate. Information concealing is likewise granular, in that it empowers associations to specifically veil information down to the line, segment, or cell level. Besides, information concealing innovation can incorporate with existing verification arrangements, including ActiveDirectory, LDAP, and Identity Access Management programming. Also, it supplements other information insurance advances, for example, encryption, database movement checking (DAM), and security data and occasion administration (SIEM), all things considered giving complete information protection assurance.

Algorithm-1 Big Data Privacy Model Using Data Masking Methods

Input: Raw Big Data Sets

Output: Secured Big Data Sets

1. Consider a big data set consists of R records $BD = \{r_1, r_2, \dots, r_m\}$
 2. Each record in R consists of set of columns $R = \{c_1, c_2, \dots, c_m\}$
 3. Record Validation process
 4. Identify appropriate data masking methods to be applied for every column
 5. $DM = \{S, KR, M, R, Shuf\}$
 6. store all masked data dynamically
 7. Repeat step 1 to 5 for all the files in the big data sets under test
-

3.2. Mathematical model

The Anonymity of the data is measured and analysed using the following mathematical equations, in a big data sets key definitions are defined and measure the distance between the keys to maintain the integrity and calculate Variational distance using Equation (1).

$$VD[X, Y] = \sum_{i=1}^n \frac{1}{2} |x_i - y_i| \quad (1)$$

We are masking the datato preserve the privacy and we analyzed the divergence factor to check the data validity and protection form using Kullback-Leibler (KL) distance formula using Equation (2).

$$KLD[X, Y] = \sum_{i=1}^n x_i \log \frac{x_i}{y_i} = H(X) - H(X, Y) \quad (2)$$

Other measurement is done to know the order and relationship between two attributes which measures the similarity and dissimilarity of the two attributes under privcy protection, order distance between two numeric attributes are measured using Equation (3).

$$OD[X, Y] = \frac{1}{n-1} \sum_{i=1}^{i=n} \sum_{j=1}^{j=i} r_j \quad (3)$$

For Categorical attributes, equal distance is measures as whose order does not always matter, we can either view the ground distance between 2 categorical attributes as always being 1 (equal Distance). As the distance between any two values is 1, for each point that $x_i - y_i > 0$, one just needs to move the extra to some other points.

4. RESULTS AND DISCUSSION

The proposed methods provide flexibility around how the data will be masked and ensure that business rules of the enterprise application will not be impacted. After data segregation, the masking type will be decided based on the data such as substitution, replacement, multiplier, randomizer and shuffling, the same is illustrated in Figure 3 below with example.


Masking Type	Example
Substitution	Ravi becomes Manju
Replacement	Z123456 becomes A999123
Multiplier	22/11/1978 becomes 15/02/1986
Randomizer	City will become xxxxaaa
Shuffling	

Figure 3. Example of hybrid data masking method for data security

Proposed method is a general approach that deals with the needs of privacy problems faced by various organizations when onsite-offshore business delivery models are used. The proposed framework ensures two principles while operations are carried out (i) Masking is not reversible. There is no way to reverse engineer the original data from the masked data and (ii) Masked data is usable. For example, when testing valid addresses the masked data must include valid zip codes not random numbers which fit the data type. After preserving the data calculated the performance factors between original data and preserved data and Comparison study of Statistical performance is done. So as to figure the factual properties, for example, mean, change and standard deviation for unique information and adjusted data. Microaggregation strategy returns just the mean esteem is same as the first. In any case, other measurable property, for example, change and standard deviation does not deliver similar outcomes. We have connected diverse size of informational collections for check. Figure 4 shows the statistical performance of the original data and modified data.

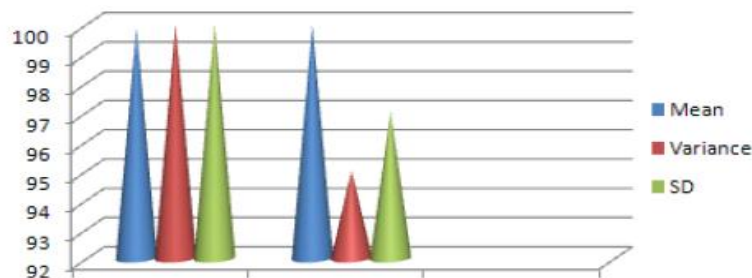


Figure 4. Statistical performance of the original data and modified data

5. CONCLUSION

We have attempted to propose the big data privacy models using data masking methods, this work will help the big data engineers and big data scientist and predominantly analyst as a general to provide the customized solutions for the customers based on the need with accepted level of security in a big data environment.

ACKNOWLEDGEMENTS

We would like to thank Dr.Ravikumar G. K, Technical Architect, Big Data Projects, Wipro Technologies, USA. Mr.Govardhan Meti, Architect, HP, Bengaluru for their valuable inputs in validating the proposed work.

REFERENCES

- [1] C. S. Sindhu, *et al.*, “A Novel Integrated Framework to Ensure Better Data Quality in Big Data Analytics over Cloud Environment”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, October 2017, pp. 2798-2805.
- [2] Mario Diván, *et al.*, Towards a Consistent Measurement Stream Processing from Heterogeneous Data Sources”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 6, December 2017, pp. 3164-3175.
- [3] Thu Yein Win, Huaglorry Tianfield, Quentin Mair, “Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing”, *IEEE Transactions on Big Data*, vol. PP1-1, no. 99, June 15, 2017.
- [4] Ryoichiro Obukata, Miralda Cuka, Donald Elmazi, Shinji Sakamoto, Tetsuya Oda, Leonard Barolli, “Performance Evaluation of an Aml Testbed for Improving QoL: Evaluation Using Clustering Approach Considering Distributed Concurrent Processing”, *Advanced Information Networking and Applications Workshops (WAINA) 2017 31st International Conference on*, pp. 271-275, 2017.
- [5] Youssef Gahi, Mouhcine Guennoun, Hussein T. Mouftah, “Big Data Analytics: Security and privacy challenges”, *Computers and Communication (ISCC) 2016 IEEE Symposium on*, pp. 952-957, 2016.
- [6] Md Tanzim Khorshed, Neeraj Anand Sharma, Aaron Vinek Dutt, A B M Shawkat Ali, Yang Xiang, “Real Time Cyber Attack Analysis on Hadoop Ecosystem using Machine Learning Algorithms”, *Computer Science and Engineering (APWC on CSE) 2015 2nd Asia-Pacific World Congress on*, pp. 1-7, 2015.
- [7] Clement Almeida, Harshitha K, Manjunath T.N, “A Study on Column Segregation for Data Security”, *IJRCSIT*, vol. 2, no. 2, February 2014.
- [8] Manjunath T.N, Ravindra S Hegadi, “Data Quality Assessment Model for Data Migration Business Enterprise”, *International Journal of Engineering and Technology (IJET)*, vol. 5, no. 1, Feb-Mar 2013.
- [9] Manjunath T.N, Ravindra S Hegadi, “Statistical Data Quality Model for data Migration business Enterprise”, *International Journal of Soft Computing*, vol. 8, no. 5, pp. 340-351, 2013.
- [10] Ravikumar.G.K, *et al.*, “A Survey on Recent Trends, Process and Development in Data Masking for Testing”, *IJCSI International Journal of Computer Science Issues*, vol. 8, no. 2, March 2011.
- [11] Ravikumar G.K, *et al.*, “Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing”, *IJEST11-03-06-217*, vol. 3, no. 6, pp. 5150-5159, June 2011.
- [12] Understanding and Selecting Data Masking Solutions - Creating Secure and Useful Data-Securosis, L.L.C. Data Masking: What You Need to Know What You Really Need To Know Before You Begin A Net 2000 Ltd. White Paper.
- [13] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run, Dan Wolfson, “Enterprise Master Data Management: An SOA Approach to Managing Core Information”, Dorling Kindersley (India) Pvt. Ltd. 2008.
- [14] Ralph Kimball and Joe Caserta, “The Data Warehouse ETL Toolkit”, Wiley Publishing, Inc. Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications - Batini, Scannapieco – 2006.
- [15] Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park, “A Data Quality Management Maturity Model”, *ETRI Journal*, vol. 28, no. 2, April 2006.
- [16] Manjunath T.N *et al.*, “Analysis of Data Quality Aspects in Data Warehouse Systems”, (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, vol. 2, no. 1, pp. 477-485, 2011.
- [17] Manjunath T.N., Ravindra S. Hegadi, Ravi Kumar G.K., “Design and Analysis of DWH and BI in Education Domain”, *IJCSI International Journal of Computer Science Issues*, vol. 8, no. 2, pp. 545-551, March 2011.
- [18] Manjunath T.N., Ravindra S. Hegadi and Mohan H.S., “Automated Data Validation for Data Migration Security”, *International Journal of Computer Applications*, vol. 30, no. 6, pp. 41-46, September 2011.
- [19] Muralidhar, K., D. Batra, and P. Kirs, “Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach”, *Management Science*, vol. 41, no. 9, pp. 1549-1564, 1995.
- [20] Muralidhar, K. and R. Sarathy, “A Theoretical Comparison of Data Masking Techniques for Numerical Microdata”, *3rd IAB Workshop on Confidentiality and Disclosure - SDC for Microdata, Nuremberg, Germany, November 20-21, 2008*.

BIOGRAPHIES OF AUTHORS



Archana R. A., Received her Bachelor’s degree in computer Science and engineering from VTU, Belgaum, Karnataka, India during the year 2007 and Master of Technology in year 2010 in computer science and engineering from VTU, Belgaum, Karnataka, India. She is currently pursuing Ph.D degree in Bharathiar University, Coimbatore, Tamilnadu. She is having 7 years of experience. Her area of interests is Image Mining, Databases and Business Intelligence. She has published and presented papers in journals, international and national level conferences.



Dr. Ravindra S. Hegadi, Received his Master of Computer Applications (MCA) & M.Phil and Doctorate of Philosophy (Ph.D) in year 2007 in Computer Science from Gulbarga University, Karnataka. He is having 22 years of experience. He has visited overseas to various universities as Subject Matter Expert (SME). His area of interests is Image Mining, Image Processing and Databases and Business Intelligence. He has published and presented papers in journals, international and national level conferences.



Manjunath T. N., Received his Bachelor's Degree in Computer Science and Engineering from Bangalore University, Bangalore, during the year 2001 and M. Tech in Computer Science and Engineering from VTU, Belgaum, during the year 2004. Ph.D degree from Bharathiar University, Coimbatore, Tamilnadu during 2015. He is having total 16 years of Industry and Teaching experience. His areas of interests are Data Warehouse & Business Intelligence, Multimedia and Databases. He has published and presented papers in journals, international and national level conferences.