International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Big Data Analytics: An Indian Perspective

Ashish Kumar Jha, Sudhir Kumar Gupta, Ajay Kumar, Mahesh Kumar Chaubey, Jitendra Singh

Abstract: ***To gain competitive advantage and to improve efficiency, reliance on data has grown manifold. This work explored the emerging sources of big data generation in India, particularly post implementation of 'Digital India' scheme. Agencies that have implemented the big data are accentuated along with the potential area. Further, a model is proposed to integrate key services utilized within a country with the 'Aadhar' to gain insight in real time. Seamless integration of key services with 'Aadhar' will be increasingly helpful in curbing the crime from Indian landscape. Further, challenges in implementation of the model are perceived, and their solutions have been proposed.***

Keywords: big data analytics, big data in India, data mining, big data and crime, big data enabler, Hadoop, big data and cloud

I. INTRODUCTION

Growth of Information technology in public services, community and personal use has led to the large data generation. In developed countries, usage of information technology (IT) is widespread, however, in developing countries particularly in emerging markets, only recently it gained the momentum [1]. In India, usage of IT in routine usage is limited. In several ministries, shift towards IT is either partial or lacking altogether. This leads to inefficiency, corruption and excessive delay in rendering the public services. Demand of prompt services has further worsened due to huge population growth in recent past [2]. Census also reflects the huge diversity in terms of literacy, economic status, infrastructure etc. across India [2].

Usage of data is not homogeneous across the world; instead developed countries are the major source of data generation. In developing countries, reliance on data is gaining momentum. In developing countries, the usage of IT for data services is primarily restricted to the services that include banking, railway reservation, ticket booking, income tax return, retail chains to name a few [3, 4].

Manuscript published on 30 September 2019
* Correspondence Author
Ashish Kumar Jha*, Assistant Professor, Dept. of Computer Science, College of Vocational Studies, University of Delhi , New Delhi, India, Email: ashishkumarjha.cs@cvs.du.ac.in
Sudhir Kumar Gupta, Assistant Professor Dept. of computer science, Keshav Mahavidhalaya, University of Delhi, New Delhi, India, email: cs.sudhirg@gmail.com
Ajay Kumar, Assistant Professor, Bharati Vidyapeeth (Deemed to be) university Institute of Management & Research New Delhi, India, Email: kumar.ajai@yahoo.co.in
Mahesh Kumar Chaubey, Assistant Professor, Bharati Vidyapeeth (Deemed to be) university Institute of Management & Research, New Delhi, India, Email: mkchaubey@gmail.com
Jitendra Singh, Dept. of computer science, Dyal Singh Evening College, University of Delhi, New Delhi, India-110003, Email: jitendra.singh0705@gmail.com

According to several studies, it is revealed that data generation is growing rapidly in government sectors including railway, Income tax, Bank etc. however, gaining insight from the data is limited due to variety of reasons including lack of awareness, limited capability of existing resources to name a few [5].

Big data has envisioned recently due to the increasing reliance on IT. Big data can be defined with the help of four 'V's. Each 'V' denotes as Volume, Velocity, Variety, and Veracity [6]. In order to qualify the definition of the big data, it should be generated at a higher speed and volume [7]. At the same time, data should be generated from one or variety of sources that include computers, smartphone, devices etc. Significant enough, data generated should be reliable that is represented with the help of veracity [8].

A. *Leveraging Analytics*

Significance of data analytics is growing rapidly and same is indicated in prominent predictions by some of the leading research and analysis firms that include Gartner, IDC, Forrester, Aberdeen etc [9]. The entire aforementioned firm highlighted that the analytics will dominate in the upcoming years [10]. Key reason behind the growing trend of analytics usage is that the organizations have the data but they are unable to exploit it fully to derive the business benefit. Owing to the huge storage of data, operating cost increases, at the same time profitability decreases [11]. In India, a person on an average receives several calls from the marketing company to purchase the specific plan. Majority of the phone call receivers are reluctant to answer such calls thereby not a prospective customer. This results in waste of effort, time, and money.

With the IT growth, users are generating the huge amount of data from the variety of sources that include social media writing, tweets, blog, sensors, cameras etc. These data is generated with tremendous pace, for instance sensors are capturing huge amount of data at every second. Twitter generates around 5787 tweets every second [12]. Every minute nearly 510 comments are posted, around 2, 93,000 statutes are updated and 1, 36,000 photos are uploaded on twitter [12]. Data posted on social media site is ranging from structured data, unstructured data that includes images, videos, etc. Information available in unstructured form such as comments, audio, video may contain valuable information [4]. Usage of analysis will assist in reaching out to the prospective client with a higher accuracy [13].

Big data is transforming the one's life in range of areas that includes home devices, security at home and office, number of social media post [14]. A great future is predicted for cloud computing despite of several issues and challenges underlying big data [15, 16]. Several verticals including wind energy, biomedical, business analytics are leveraging the big data to gain the competitive advantage [17, 18].

B. *Big Data and Cloud Computing*

Cloud computing is consistently growing from last several years [19]. It can act as a catalyst for the big data since, Big data computation needs massive resources that may not be feasible for many organizations. Even for the big giant, resources may not support the magnitude of data that need to be store [20]. Consequently, they may need the scalability of resources. Cloud computing is well suited in scenario where the massive and immediate resources are needed for short duration [21]. Technologies such as Hadoop, cloud computing, analytics are set to widely used for the data that is large in size, in order to gain insight from the data [22, 23]. Several researches are already active in order to generate the efficient algorithm that should deeply analyzed the data in a timely manner [24]. Same can be determined with the help of big data analytics and deep learning methods undertaken in this area [25].

In order to store the big data one need huge and sophisticated data center. Maintaining the data center may not be feasible for all the enterprise, at the same time, procuring and deploying resources is a time taking activity, whereas, subscribing to the data center is immediate [3]. In cloud subscription, one need to register only at the cloud provider and the data center aimed for usage. With the growing compliance at data center, security apprehension is done away with to a great extent.

C. *Big data analytics in world landscape*

Big data analytics is widely used in range of services offered by the government and private agencies [26]. For instance, in Chicago America, big data analytics is used to analyze the data in real time in order to curb the crime. It is also used by the municipal corporation to understand the citizen's problems and to improve civic services. Such services need huge resources and same can be provisioned by subscribing the cloud computing [27].

Rest of the manuscript is organized as: Section 2 elucidates the related work carried out using data mining technique or big data analytics. Section 3 highlights the major source of data generation. Section 4 describes the major implementation of big data in India and compared it with the world. Finally, this work has proposed a framework to curb the crime by integrating the services with Digital based individual's identity.

II. Related Work

Big data is a new emerging domain with huge potential. To harness it benefit, variety of organization and researchers are active in this area and contributing significantly to evolve this domain further. Prominent work related to us has been presented herein.

To improve the data processing speed, [28] introduced improvement in Mapreduce with the help of data placement in heterogeneous Hadoop clusters. Introduction of Hadoop in Facebook was presented by [29]. Transition to big data implementation is not smooth instead carries several challenges that may creep during the introduction of big data considering the cultural, technologies and scholarly phenomenon that rests on the interplay of technology [30]. Method to smoothly transit from the legacy based method to big data was proposed by [30], authors also proposed modification in Hadoop to improve its efficiency.

Role of Business intelligence and analytics (BI& A) was elucidated by [31] along with 03 models of BI&A that includes BI & A 1.0, BI & A 2.0 and BI & A 3.0. Benchmark in big data analytics is lacking, thereby [32] have stressed need for benchmark in big data analytics and suggested the benchmark with the name 'BigBench' as testbed. Recent achievement is in introduction of new technologies such as Hadoop and Spark particularly in Social network was highlighted by [33]. [8] highlighted the utility of Apache Spark for big data processing. A novel approach to detect the breast cancer was introduced by [34]. Music library association with mailing list using text mining was explored by [35].

III. Problem Statement

Social media, YouTube, are great sources of information generation. On Twitter and Facebook millions of transactions are carried out by the users in the form of posting comments, photos, posts etc. The available data at the social network sites are consumed by the analytical firms to analysis of the trend and sentiment of users [36]. According to the user's mood and liking marketing strategies are articulated to target the wider user's spectrum [37, 38]. According to several reports, rate of data generation is high particularly in last few years. According to the estimates, data generated in last few five years is higher than the one generated in last twenty years.

In India, information is generated with the digital India programme, mobile phones, social network such as Facebook, twitter, Instagram etc. All these sources are generating huge amount of data and gaining insight from this data in a timely manner is challenging. However, real-time analytics of data is extremely desired in order to determine the need of the citizens, maintaining law and order and to discover the gaps in the services offered to the citizen's and the one sought by them. In such cases, data is aimed to be maintained at distributed location in order to improve the reliability and to improve performance in cost effective manner [11].

Organizations are storing huge amount of data generated. In order to improve efficiency and to understand the past trend, gaining insight from the data is needed. In addition, maintaining the archive data require huge expenditure. Gaining insight with the data will assist in reduction of the maintenance cost.

With the free flow of information, unsocial elements are also leveraging this tool to spread hatred and misguide the general public [39, 40, 41, 42]. Social media and you tube has emerged as a cost effective at the same time with global accessibility tool to spread terrorism and spread their ideology [42]. Accordingly, it is imperative to analyze the video and other material posted on internet.

IV. Sources of Big Data Generation in India

In 2015, government of India has launched the digital India programme [43]. With this scheme, government is aiming to shift the majority of its activities online in order to improve the transparency and for the prominence governance [43]. Another programme namely 'Smart city' is also underway, that is also perceived as working only on data. This work highlights the prominent services that may be generating the huge amount of data.

A. *'Aadhar' Data*

'Aadhar' is the unique identification issued to resident of India bearing the photograph and biometric details for identification. Sole purpose of 'Aadhar' is to eliminate the duplicate and multiple identities [44]. This method is robust relative to other methods, since it carries multiple biometric details of residents. UIDAI(Unique Identification Authority of India) is the statutory body responsible for issuing, maintaining, the 'Aadhar' in a cost effective manner. Currently, 120 crore (1.2 billions) resident in all range of age groups have been issued 'Aadhar'. Total number of digits used to identify is 12 [45]. In addition, it carries four hidden digits that make the total digit to '16'. To avoid typing digits, 'Aadhar' also bears 'quad code'. To enhance security and minimize instances of leaks, UIDAI have also introduced the 16 digit virtual ID.

B. *'DigiLocker'*

'DigiLocker' service is commissioned by the government of India to store the citizen's record in digital form. Limited amount of space is provided free of cost. Certificates such as driving license, 'Aadhar' number can be stored at 'DigiLocker'. To safeguard the data, best in class security has been introduced in order to secure the citizen's data. Since, this service is gaining growth, consequently, it is transforming as a big data. It is widely popular among the youth and registered the robust growth in the year 2017-18. Same has been illustrated with the help of figure 1 and figure 2.

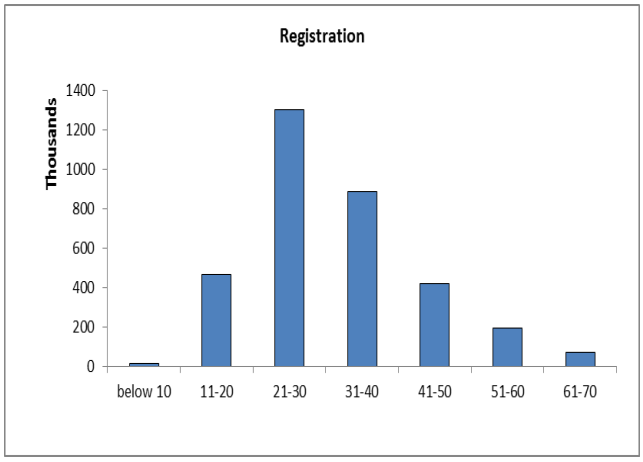

Figure 1: 'DigiLocker' Registration Age wise

Statistics of the figure 1 clearly highlights the inclination of the youth towards the digiLocker. Among youth, users in the age group of 21-30 hold the majority of the share. That is followed by the users in the age group of 31-40.Whereas, it is least popular among the age group of 61-70 as well as lower than 10.

'Aadhar', driving license, PAN card, Degrees are the popular category that are widely stored and accessed by the users.

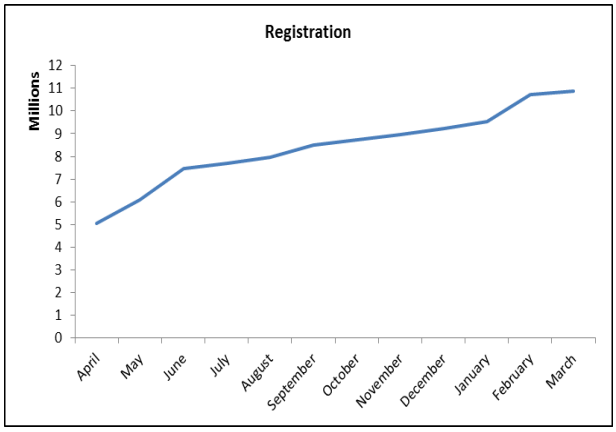

Figure 2: 'DigiLocker' total Registration

C. *Big data in Income tax*

Big data is also set to boost the income tax collection in the country. According to the sources, government is aiming to use the big data analytics to improve the tax collection. Presently only 6.84 crore people are paying taxes in the country with over 1.3 billion of population [46]. Government is also aiming to analyze the spending pattern, variety of income source, sources of income and expenditure. This will greatly help in understanding the underlying pattern and tax evasion undergoing. By collecting and correlating the expenses incurred in foreign visit, jewelry purchased payment of hotel bills will be helpful to understand the one's lifestyle this could be further combined with the income tax return filed by an individual. If any mismatch appears then broader enquiry can be conducted in order to prevent the tax evasion and bring larger population under the tax ambit.

D. *Data generation in Banking*

Large amount of data is being generated by banking sector. A number of facilities are offered to the customer in order to ease their life. For instance, using internet banking one can transfer the fund from one account to another account. Similarly, funds can be withdrawn using ATM machine. A large number of transactions are happening in the ATM's across the country. According to the data released by Reserve bank of India, around 158 million transactions are happening on a monthly basis [47]. This can immensely help the banker's to understand the withdrawal habit of a customer, period of a year when the extraordinary withdrawal happens [36]. Understanding the spending pattern will enable the banks to prepare well in advance for the fund management. Eventually, measure placed will result in strong customer relationship.

V. BIG DATA IMPLEMENTATION IN INDIA AND OPPORTUNITY

In India big data analytics has already been implemented and its benefits are being harnessed.

A. *Big Data Implementation*

Being a leader in software, India is not untouched with the implementation of big data. Instead, it is already put in place at several sectors to improve their services. Areas big data is already implemented have been presented herein:

1. Water management in Bangluru

Bangluru civic body has implemented the big data implementation in its water distribution system. The IT city of India also known as Silicon Valley of India has around 10 Million population and water need to be served to each household [48]. For the effective delivery of water system, it has adopted the big data. In order to effectively implement the plan it has tied up with IBM that will create the dashboard based on IBM's intelligent operational center. This implementation will enable the implementer in monitoring, administering and managing the water supply across the Bangluru. Here the key role will be to analyze the water distributed and the one reaches to the consumer [48]. Bangluru Municipal Corporation analyzes the gap in distribution to curb the water wastage, theft, and pipeline leakage. This system has achieved the huge success in improving the water supply and curbing the water loss.

2. IRCTC

IRCTC is the acronym of 'Indian Railway Catering and tourist Corporation' and established primarily to facilitate the railway reservation and catering need of passengers during train travel. A large number of consumers are booking the ticket daily. In the year 2001, only 10 persons have utilized its website to get their reservation done. From 29 people booked a ticket in 2002 to 1.3 million passengers booking in a single day on 1st April 2015. Indian Railway reservation system has undergone radical changes since 2001.

Online reservation system has surpassed the ticket booking happening at the counter. Comparing ticket booked from the ticket window with the one booked from IRCTC, online (IRCTC) has touched the whopping share of 60 percent in its recent report [49]. In order to serve the consumer effectively, IRCTC has deployed the powerful server. Capabilities of IRCTC servers are enhanced frequently owing to the growth in traffic. Ease of access, wider connectivity, emergence to tech savvy youth, and reach of digital payment are cited as a few reasons driving the IRCTC's growth.

With the introduction of next generation ticketing system, number of users those can book the ticket concurrently have reached to the tune of 7200 person in a second. This result in massive data generation, since, IRCTC needs several field that includes name, age, address, food habit, mobile number etc. to store the data. Government officials are analyzing the data to understand it on several dimensions that include, age factor, concession, route highly demanded, peak season, idle season etc.

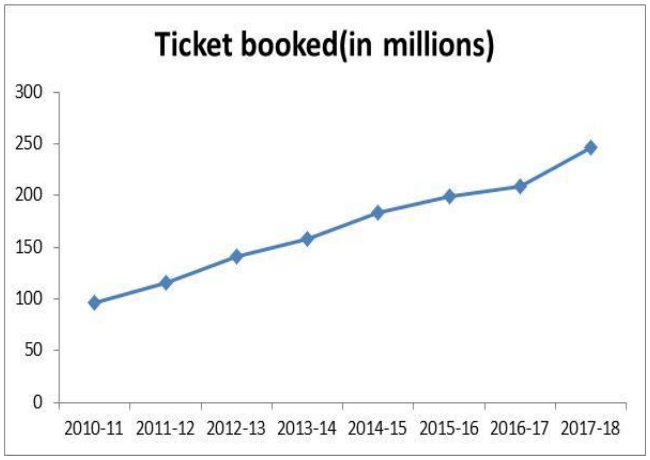

Figure 3: Ticket Booked IRCTC [49]

Realizing the significance of huge data, government is also aiming to generate additional revenue from non-operational services, data is one among them. Currently, revenue generation is of non-operational data is around 5 percent that is far below with that of global standard that lies around 10 percent. Government is also aiming to enhance the revenue generation to the level of 10 percent.

3. Twitter analysis

Twitter is the leading social media platform where an individual submit his post and can enjoy high visibility. According to the followers count and individual's visibility, followers improve over time. Primarily, celebrity, leaders, authors are leveraging the twitter to gain higher visibility and dissemination of policies as well as changes that need to be incorporated. In addition, other citizens across the world are also connected to the twitter and post their liking and disliking on twitter. Consequently, it acts as a great source to analyze the sentiments of citizen's. Although new figures are not available, however according to its last released hundreds of millions tweets are posted in a day [50], which is fairly a large number.

Twitter analysis is widely used in the Parliament election held in 2014 to determine the citizen's mood [51]. Accordingly, the manifesto was prepared to cover the citizen's mood mentioned in their post [51]. Presently twitter analysis is also used in some of the government organization dealing with the public services to understand the problem and analyze them [52]. For instance, The Indian Railway is using the Twitter to diagnose the passenger's problem both in real time and offline mode. If need for any intervention is experienced by the administration, it is exercised immediately to address the individuals and group problems [52]. In offline mode, problems require to be undertaken of long term in nature and are considered to help the government in framing the policies.

4. Digitization of Health Services

Government is laying increasing focus to strengthen the health system in order to offer the ease of medical treatment [26]. Indeed, digital measures are required due to the huge population in hand and the limited number of doctors and hospital. According to the AIIMS detail, around 1.5 million patients are treated every year [53]. Online registration for OPD (Outpatient Department) registration has been launched in July, 2015 as a part of digital India initiative [53]. Accordingly huge amount of data is being generated due to the large number of patient's arrival. Once registered, patients can easily access their medical history without any pain. However, storing the detailed information of a patient particularly x-rays and other clinical report would need a lot of storage resources [9].

Big data analytics can immensely assist the doctors and researchers to understand the diseases and treat the customer [54]. Analytics method will prove to be highly beneficial for the patient with higher paying capacity seeking immediate relief due to their higher risk diseases [55].

B. *Prospective area*

In addition to the present implementation of digitization, following are the prospective area that will increasingly rely on data for their functioning and survival. In the event of failure of information technology the entire system will collapse.

1. 'Make in India' and Industry 4.0

Government of India has introduced 'Make in India' programme that aims to boost the manufacturing to cater the domestic requirement and focus on export. Eventually, country should establish itself as prominent export countries. Industrial revolution 4.0 (Industry 4.0) lays emphasis on integration of machines with technological aspect in order to boost production, minimum human intervention, reduction in production cost etc. Integration of machines with its surrounding device will lead to generation of tremendous amount of data that can be analyzed with the legacy system, thereby will need further enhancement of technology and adoption of big data technologies [56].

2. Analysis of video in real time

Total population of India has already touched the mark of 1.3 billion and continues to grow as well. To manage such a large population and tracing their activities need deployment of CCTV at strategic locations such as bus stands, railway station, airports, hotels and business centers. Although, CCTV deployment in metropolitan city is gaining ground yet the wide gap lies with tier-2 and tier-3 cities that are lacking this facility.

In metropolitan city, CCTV installed is monitored by the security personnel remotely and probability of human errors remains high. Accordingly, there is urgent need to deploy the computerized analytical method in order to gain insight and trace the targeted people needed by agencies such as police, crime branch, enforcement directorate etc. Deployment of real time analysis will enable to alert the nearby area for the presence of 'wanted' individual(s).

VI. PROPOSED SOLUTION AND FURTHER CHALLENGES

Data is being generated at tremendous pace. Part of the data is structured whereas the part is unstructured data. Importance of data varies with the business carried out by the stakeholders. For instance details such as income, asset, deposit and withdrawal are significant to the banks in order to approve the bank loans or issue of credit card. Whereas unstructured data may be important to marketing company in order to understand the feedback and users experience. Similarly, unstructured data can be equally important to political parties to sense the voter's mood and to generate wave in favour of specific political parties.

With the inception of digital India campaign, data is being generated at considerably higher rate relative to earlier. Government is aiming to integrate the key services with 'Aadhar' that also bears the biometric details. The move will immensely help the agencies to establish the unique identification. Need of an hour is to integrate the 'Aadhar' with other services such as Bank, Income tax, hotels, travel agencies, mobile in order to learn deeper insight related to an individual. Further, the move will promote the higher transparency and curb the crime rate in financial and non-financial sectors.

A. *Big Data to curb Black Money and crime*

Big data can greatly help in curbing the black money. In order to accomplish this task, an 'Aadhar' is to be linked with the bank. Same can be verified at the time of opening bank account. This will ensure that the genuine and authentic person is opening the account. At the same time number of account open by a person with manipulated name can be well connected to each other with the help of 'Aadhar'.

Aadhar linking will also help the Income tax department to pool the bank statement from different banks with the help of 'Aadhar'. By Integrating the services, hiding one's detail will prove to be cumbersome. Upon tracing of anomaly, account holder's account can be audited by Income tax department or in coordination with other investigation agencies. This measure will greatly help in tax evasion and curbing the crimes. Government of India has already initiated a program to connect the 'Aadhar' with bank account.

B. *Challenges*

However, integration of variety of domains also carries several challenges since people and other non-government group protests this move. On mobile linking with 'Aadhar', several entities have already approached the court, citing infringement of privacy.

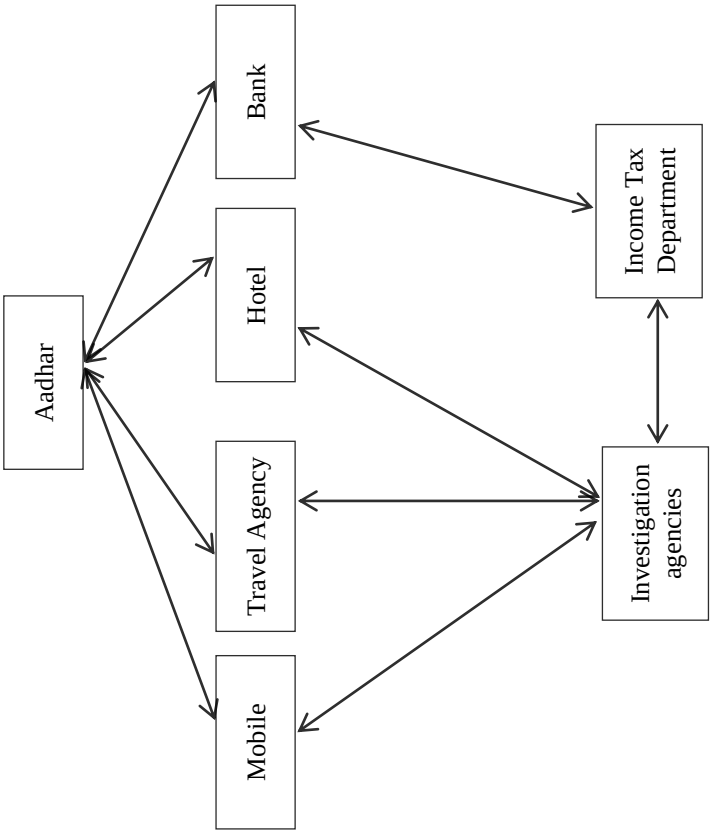

Figure 4: Proposed model to curb crimes

Presently, privacy and security is covered with the help of IT act 2010 and people are apprehensive that it does not adequately address their concern. Since then, technology has witnessed tremendous transformations. Paradigm such as cloud computing has emerged, whereas the act is not suitably modified to cover the cloud paradigm.

C. *Proposed Solution*

Undoubtedly integration of data among variety of entities carries huge risk as far as data privacy is concerned. Therefore, appropriate law and regulation need to be placed, in order to preserve the privacy and security. Security need to be enhanced at both the levels that include a) Hardware and Software level b) Human access. Strong security in place at software and hardware level will prevent the leakage of data at their respective level.

At the same time, regulated environment will discourage human to indulge in data sharing that may be deliberately or inadvertently. Therefore, designation should be identified that can access the data stored and managed by other agencies and the degree of freedom at which data access can be permitted.

VII. CONCLUSION

Big data is the norm of the day and truly applicable in Indian landscape that holds huge population. Big data analytics is the best fit to identify the target and reach. In India, it has already been implemented in several domains including water supply, railway booking, twitter analysis etc. Although it's implemented is limited to few city or a singly organization. Once compared with the global implementation, it is far behind and need to be expanded further in other areas. Seamless integration of proposed 'Aadhar' based system with several other agencies and services that include Bank, hotels, and airport will enable to gain deep insight about individuals to a great extent. Any potential threat by criminal can be traced and linked thereby assist in proactively initiate an action that will help in curbing the crime.

REFERENCES

1. *Apache spark: a unified engine for big data processing.* **Zaharia, Matei, et al., et al.** s.l. : ACM, 2016, Communications of the ACM, Vol. 59, pp. 56-65.
2. *Improving mapreduce performance through data placement in heterogeneous hadoop clusters.* **Xie, Jiong, et al., et al.** 2010. Parallel \& Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on. pp. 1-9.
3. *Data mining with big data.* **Wu, Xindong, et al., et al.** s.l. : IEEE, 2014, IEEE transactions on knowledge and data engineering, Vol. 26, pp. 97-107.
4. *Going dark: Terrorism on the dark web.* **Weimann, Gabriel.** s.l. : Taylor & Francis, 2016, Studies in Conflict \& Terrorism, Vol. 39, pp. 195-206.
5. *Finding the missing link for big biomedical data.* **Weber, Griffin M., Mandl, Kenneth D. and Kohane, Isaac S.** s.l. : American Medical Association, 2014, Jama, Vol. 311, pp. 2479-2480.
6. *Big data analytics in logistics and supply chain management: Certain investigations for research and applications.* **Wang, Gang, et al., et al.** s.l. : Elsevier, 2016, International Journal of Production Economics, Vol. 176, pp. 98-110.
7. *Apache hadoop yarn: Yet another resource negotiator.* **Vavilapalli, Vinod Kumar, et al., et al.** 2013. Proceedings of the 4th annual Symposium on Cloud Computing. p. 5.
8. *Big data: New tricks for econometrics.* **Varian, Hal R.** 2014, Journal of Economic Perspectives, Vol. 28, pp. 3-28.
9. **Suneja, Kirtika.** 16-digit Aadhaar to have four hidden numbers. *Business Standard.* [Online] Jan 20, 2013. https://www.business-standard.com/article/economy-policy/16-digit-aadhaar-to-have-four-hidden-numbers-110042800076_1.html.
10. *Dendroid: A text mining approach to analyzing and classifying code structures in android malware families.* **Suarez-Tangil, Guillermo, et al., et al.** s.l. : Elsevier, 2014, Expert Systems with Applications, Vol. 41, pp. 1104-1117.
11. **stricker, Gabriel.** Twitter Blog. [Online] 2014. https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html.
12. *Big data technologies and Management: What conceptual modeling can do.* **Storey, Veda C. and Song, Il-Yeol.** s.l. : Elsevier, 2017, Data \& Knowledge Engineering, Vol. 108, pp. 50-67.
13. **Srinivas, V.** Digital AIIMS: Leading a Revolution in Indian Healthcare. *eHealth.* [Online] Dec 2015. https://ehealth.eletsonline.com/2017/03/digital-aiims-leading-a-revolution-in-indian-healthcare/.
14. **Smith, Eileen and Shirer, Michael.** Worldwide Public Cloud Services Spending Forecast to Reach $160 Billion This Year, According to IDC. *IDC.* [Online] Jan 18, 2018. https://www.idc.com/getdoc.jsp?containerId=prUS43511618.
15. *Digital India, cloud computing.* **Singh, Jitendra and Chaubey, Mahesh kumar.** NewDelhi : s.n., 2015. BVIMR.
16. *Big data analytics to improve cardiovascular care: promise and challenges.* **Rumsfeld, John S., Joynt, Karen E. and Maddox, Thomas M.** s.l. : Nature Publishing Group, 2016, Nature Reviews Cardiology, Vol. 13, p. 350.
17. **Poushter, Jacob.** Emerging, developing countries gain ground in the tech revolution. *Pew Research Center.* [Online] Feb 22, 2016. http://www.pewresearch.org/fact-tank/2016/02/22/key-takeaways-global-tech/.
18. *Predicting the future_big data, machine learning, and clinical medicine.* **Obermeyer, Ziad and Emanuel, Ezekiel J.** s.l. : NIH Public Access, 2016, The New England journal of medicine, Vol. 375, p. 1216.
19. **Nirmala.** Launch of NASSCOM Report- Cloud: Next Wave of Growth in India-2019. *NassCom.* [Online] https://community.nasscom.in/communities/digital-transformation/launch-of-nasscom-report--cloud-next-wave-of-growth-in-india-2019.html.
20. *Deep learning applications and challenges in big data analytics.* **Najafabadi, Maryam M., et al., et al.** s.l. : Springer, 2015, Journal of Big Data, Vol. 2, p. 1.
21. *Nosql database: New era of databases for big data analytics-classification, characteristics and comparison.* **Moniruzzaman, A. B. M. and Hossain, Syed Akhter.** 2013, arXiv preprint arXiv:1307.0191.
22. **Magdalena, KAMINSKA and SMIHILY, Maria.** Cloud computing statistics on the use by enterprised: statistics explained. [Online] 2018. https://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_-_statistics_on_the_use_by_enterprises.
23. *The impact of social media influencers on purchase intention and the mediation effect of customer attitude.* **Lim, Xin Jean, Cheah, Jun-Hwa and Wong, Mun Wai.** 2017, Asian Journal of Business Research, Vol. 7, pp. 19-36.
24. *On the energy (in) efficiency of hadoop clusters.* **Leverich, Jacob and Kozyrakis, Christos.** s.l. : ACM, 2010, ACM SIGOPS Operating Systems Review, Vol. 44, pp. 61-65.
25. *Service innovation and smart analytics for industry 4.0 and big data environment.* **Lee, Jay, Kao, Hung-An and Yang, Shanhu.** s.l. : Elsevier, 2014, Procedia Cirp, Vol. 16, pp. 3-8.
26. *Big-data applications in the government sector.* **Kim, Gang-Hoon, Trimi, Silvana and Chung, Ji-Hyong.** s.l. : ACM, 2014, Communications of the ACM, Vol. 57, pp. 78-85.
27. *Trends in big data analytics.* **Kambatla, Karthik, et al., et al.** s.l. : Elsevier, 2014, Journal of Parallel and Distributed Computing, Vol. 74, pp. 2561-2573.
28. *Big data: Issues and challenges moving forward.* **Kaisler, Stephen, et al., et al.** 2013. System sciences (HICSS), 2013 46th Hawaii international conference on. pp. 995-1004.
29. **John Walker, Saint.** Big data: A revolution that will transform how we live, work, and think. s.l. : Taylor & Francis, 2014.
30. 30. *Significance and challenges of big data research.* **Jin, Xiaolong, et al., et al.** s.l. : Elsevier, 2015, Big Data Research, Vol. 2, pp. 59-64.
31. *This is not your mother_s terrorism: Social media, online radicalization and the practice of political jamming.* **Huey, Laura.** s.l. : Centre for the Study of Terrorism and Political Violence, University of St _, 2015, Journal of Terrorism Research.
32. *Exploring the Music Library Association Mailing List: A Text Mining Approach.* **Hu, Xiao, et al., et al.** 2017. The 18th International Conference on Music Information Retrieval (ISMIR).
33. *The rise of _big data_ on cloud computing: Review and open research issues.* **Hashem, Ibrahim Abaker Targio, et al., et al.** s.l. : Elsevier, 2015, Information Systems, Vol. 47, pp. 98-115.
34. *Recent advances and emerging applications in text and data mining for biomedical discovery.* **Gonzalez, Graciela H., et al., et al.** s.l. : Oxford University Press, 2015, Briefings in bioinformatics, Vol. 17, pp. 33-42.
35. *Key player identification in terrorism-related social media networks using centrality measures.* **Gialampoukidis, Ilias, et al., et al.** 2016. 2016 European Intelligence and Security Informatics Conference (EISIC). pp. 112-115.
36. *BigBench: Towards an Industry Standard Benchmark for Big Data Analytics.* **Ghazal, Ahmad, et al., et al.** New York, NY, USA : ACM, 2013. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 1197-1208. ISBN: 978-1-4503-2037-5.

37. *Beyond the hype: Big data concepts, methods, and analytics.* **Gandomi, Amir and Haider, Murtaza.** s.l. : Elsevier, 2015, International Journal of Information Management, Vol. 35, pp. 137-144.
38. *Performance and energy efficiency of big data applications in cloud environments: A Hadoop case study.* **Feller, Eugen, Ramakrishnan, Lavanya and Morin, Christine.** s.l. : Elsevier, 2015, Journal of Parallel and Distributed Computing, Vol. 79, pp. 80-89.
39. *Mining big data: current status, and forecast to the future.* **Fan, Wei and Bifet, Albert.** s.l. : ACM, 2013, ACM sIGKDD Explorations Newsletter, Vol. 14, pp. 1-5.
40. *Trend Topic Analysis for Wind Energy Researches: A Data Mining Approach Using Text Mining.* **Erouglu, Yunus and Seckiner, Serap U.** 2016, Journal of Technology Innovations in Renewable Energy, Vol. 5, pp. 44-58.
41. . *Big Data consumer analytics and the transformation of marketing.* **Erevelles, Sunil, Fukawa, Nobuyuki and Swayne, Linda.** s.l. : Elsevier, 2016, Journal of Business Research, Vol. 69, pp. 897-904.
42. **Elayidom, M Sudheep.** *Data Mining and business intelligence.* Delhi : Cengage learning India, 2015.
43. **Cooper, Paige.** 28 Twitter Statistics All Marketers Need to Know in 2019. *Hootsuite.* [Online] Jan 2019. https://blog.hootsuite.com/twitter-statistics/.
44. *Business intelligence and analytics: from big data to big impact.* **Chen, Hsinchun, Chiang, Roger H. L. and Storey, Veda C.** s.l. : JSTOR, 2012, MIS quarterly, pp. 1165-1188.
45. *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* **Chen, C. L. Philip and Zhang, Chun-Yang.** s.l. : Elsevier, 2014, Information Sciences, Vol. 275, pp. 314-347.
46. *A novel approach for breast cancer detection using data mining techniques.* **Chaurasia, Vikas and Pal, Saurabh.** 2017.
47. *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.* **Boyd, Danah and Crawford, Kate.** s.l. : Taylor & Francis, 2012, Information, communication \& society, Vol. 15, pp. 662-679.
48. *Apache Hadoop goes realtime at Facebook.* **Borthakur, Dhruba, et al., et al.** 2011. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. pp. 1071-1080.
49. *Social big data: Recent achievements and new challenges.* **Bello-Orgaz, Gema, Jung, Jason J. and Camacho, David.** s.l. : Elsevier, 2016, Information Fusion, Vol. 28, pp. 45-59.
50. *Big data in health care: using analytics to identify and manage high-risk and high-cost patients.* **Bates, David W., et al., et al.** 2014, Health Affairs, Vol. 33, pp. 1123-1131.
51. **Bala, Myneni Madhu and Ravilla, Venkata Krishnaiah and Prasad, Kamakshi and Dandamudi, Akhil.** Dynamic Behavior Analysis of Railway Passengers. *Innovative Applications of Big Data in the Railway Industry.* s.l. : IGI Global, 2017, pp. 157-182.
52. **Babu, Shivnath and Herodotou, Herodotos.** Cost-based optimization of configuration parameters and cluster sizing for hadoop. s.l. : Google Patents, 6 2016. US Patent 9,367,601.
53. *Big Data computing and clouds: Trends and future directions.* **Assuno, Marcos D., et al., et al.** s.l. : Elsevier, 2015, Journal of Parallel and Distributed Computing, Vol. 79, pp. 3-15.
54. *The 2014 Indian Elections on Twitter: A comparison of campaign strategies of political parties.* **Ahmed, Saifuddin, Jaidka, Kokil and Cho, Jaeho.** 4, 2016, Telematics and Informatics, Vol. 33.
55. **ET Bureau.** With 99.49 lakh new tax filers, income tax returns surge 26% in 2017-18 . *The Economics Times.* [Online] Apr 2018. https://economictimes.indiatimes.com/news/economy/finance/with-99-49-lakh-new-tax-filers-income-tax-returns-surge-26-in-2017-18/articleshow/63586783.cms.
56. **CSA.** The Treacherous 12 - Top Threats to Cloud Computing. [Online] 2017b. https://downloads.cloudsecurityalliance.org/assets/research/top-threats/treacherous-12-top-threats.pdf.
57. **NPCI.** Statistics. *NPCI.* [Online] 2019. https://www.npci.org.in/statistics.
58. **Census India.** Office of the Registrar General & Census Commissioner, India. *http://censusindia.gov.in/.* [Online] 2011. http://censusindia.gov.in/pca/Searchdata.aspx.
59. **Microsoft.** Microsoft Cloud Services Vision Becomes Reality With Launch of Windows Azure Platform. *Microsoft Blog.* [Online] Nov 17, 2009. https://news.microsoft.com/2009/11/17/microsoft-cloud-services-vision-becomes-reality-with-launch-of-windows-azure-platform/.
60. **Gartner.** Gartner IT Glossary. *Gatner.* [Online] https://www.gartner.com/it-glossary/big-data/.
61. Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.3 Percent in 2019. *https://www.gartner.com.* [Online] Sept 12, 2018. https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019.
62. **IDG.** Data & Analytics Survey. *CDN.* [Online] 2016. https://cdn2.hubspot.net/hubfs/1624046/IDGE_Data_Analysis_2016_final.pdf.
63. **CSA.** CSA Security Guidance for Critical Areas of Focus in Cloud Computing v4.0. *Cloud Security Alliance.* [Online] 2017. https://cloudsecurityalliance.org/guidance/#_overview.
64. **ETCIO.** Bengaluru uses big data analytics to check unaccounted water supply. *ETCIO.com.* [Online] Feb 20, 2014. https://cio.economictimes.indiatimes.com/news/case-studies/bengaluru-uses-big-data-analytics-to-check-unaccounted-water-supply/30732789.
65. **irctc.** Annual Report. *IRCTC.* [Online] 2017. http://www.irctc.com/annualReport_En.jsp.
66. **UIDAI.** About UIDAI. *Unique Identification Authority of India.* [Online] Dec 25, 2018. https://uidai.gov.in/about-uidai/unique-identification-authority-of-india/about.html.

AUTHORS PROFILE

Ashish Kumar Jha is working as an assistant professor in the Department of Computer Science, College of Vocational Studies, University of Delhi. With over a decade experience in teaching, he has keen interest in the research areas that include cloud computing, artificial intelligence. He is working in cloud computing, AI and big data.

Sudhir Kumar Gupta is working as an assistant professor in the Keshav Mahavidyalaya, University of Delhi. Besides teaching, he has worked in Industry for long and developed some of the leading application used today. He has keen interest in the research areas namely cloud computing, artificial intelligence, machine learning. Currently he is active in cloud computing, AI and big data.

Ajay Kumar working as assistant professor in Bharati Vidyapeeth (Deemed to be university) Institute of management & research New Delhi. He has 10 years' experience in teaching and his area of interest is Network Security, machine learning and big data. He is Pursuing Ph.D from Mewar University.

Mahesh Kumar Chaubey is working as Asst. Professor with Bharati Vidyapeeth Deemed to be University. He is an oracle certified trainer with 12+ years of experience in teaching, research and administration. In addition, he has been key resource person to deliver lecture on various research writing tool and on quality research writing. His two articles have received the best paper award in their respective categories. He is deeply involved in machine learning, big data and CC.

Jitendra Singh pursued PhD in the area of cloud computing and same is completed in the 2013. He has qualified the prestigious UGC-NET examination conducted by the UGC of India in the year 2006. With over 16 years of experience in teaching, research and administration, currently he is working with a college of University of Delhi. In addition, contributed as faculty member with the Stratford University, USA, India Campus, as a part time faculty for over five and half years. Beyond, he has contributed more than two dozen of research articles in the area of cloud computing, security, machine learning. Several of them have been published in reputed journals indexed in Scopus, Inspec, DBLP etc. Besides, he has authored three books namely 'Cloud computing for beginner to researcher', "Data structure simplified: Implementation with c++",
"Python: Principles and practice