Smart Data Analysis and Data Governance

Yunzhong Feng
Hebei Normal University, Shijiazhuang, P.R. China
fyz02817@sina.com

Xiaohua Feng*
University of Bedfordshire, Bedfordshire, United Kingdom
xiaohua.feng@beds.ac.uk

ABSTRACT

Data governance is still an outstanding issue currently, especially in big data cases. There is a planned big event in 2022 at Chongli. A Smart Chongli has up and running. In order to run the event properly, information security protection should be taken into account. This research is to consider generalize this issue to the future, including the Smart Chongli winter event, but not only limited there. Because of many Smart Cities services are running by online devices, to protect their data, cyber security is in demand, for the Smart Cities' quantity. As a case study, Smart Chongli is discussed, in terms of participants' data. Analysis of these data needs to pay attention to the "Measures for the Administration of Data Security", "GDPR" and so on. Cyber security should be taken care of, from the start to its final ending. Nevertheless, for Smart Chongli, overall development and management need to consider data security impact. A solution is suggested to fill this gap, according to MKSmart trial project Phase 2 experience. Which could also be applied to other smart cities similar project for the time being.

CCS CONCEPTS

• **Security and privacy**; • **Human and societal aspects of security and privacy**; • **Social aspects of security and privacy**;

KEYWORDS

Information security, Smart Chongli, Measures for the data governance, Big data privacy, General data protection regulation, Data security law

ACM Reference Format:
Yunzhong Feng and Xiaohua Feng*. 2021. Smart Data Analysis and Data Governance. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487112

1 INTRODUCTION

Smart Chongli has been set to provide various winter competitive sports services offer to the Winter Olympics society. Some of Smart Chongli data are public data; the electronic data collected from different sources. The data which has been stored into data hub, can be used to make decisions aiming at efficiently managing Smart Chongli assets and resources. Confidentiality, Integrity, Availability and Audit security principle concepts could be applied on these data, to fulfill information security.

2 LITERATURE REVIEW

2.1 Recent Computer Law Development

In May 2019, one of the departments of Chinese government, the Cyberspace Administration of China office has published "Measures for the Administration of Data Security Act 2019". This Internet law is made corresponding to the Chinese government "Network Security Act" 2016. The purpose of this act in terms of Internet data protection, is to safeguard national security, social and public interests, protect citizens and legal persons and to ensure the security of personal information and important data protection take place. It will work along with GDPR (General Data Protection Regulation) [1].

In this "Measures for the Administration of Data Security Act", Article 9 stated: any rule of collection and use data is included in the privacy policy shall be relatively concentrated and clearly indicated for easy reading. In addition, a network operator may not collect personal information until the user has been aware of its rules of collection and use, expressly given their consent. And in Article 10 described to ISP (Internet Service Provider), a network operator shall strictly abide by its rules of collection and use, and the functional design of a website or application program to collect or use personal information shall be consistent with the privacy policy and be adjusted concurrently. Cyber Security technology could be applied to protect data, systems and networks from malicious intentions; including to safeguard Smart Chongli data. Alexakos et al [2] suggested an example, to use digital forensics technique on smart cities AV (Autonomous Vehicle) as part of smart city transport sensor data governance. There were numbers of experiments to support their proposal. The idea required more real case testing to prove feasibility.

2.2 Data Breach Incidence Examples in UK

As the audience requested, some examples of data lost incidents in United Kingdom were listed here. In 2017, at Heathrow Airport, a 2.5GB unencrypted USB containing 76 folders of sensitive information for the airport was lost, which included badges, maps, CCTV camera locations, and so on security data. Furthermore, in 2012 an incident happened with Greater Manchester Police. An unencrypted USB stick containing details of witnesses with links to serious criminal investigations that was being kept in a police officer's house, which was stolen in a burglary. The Police force had been fined £120,000 by the state. The constant data incidences made government realize the importance of data governance, triggered DPA updates for public and private data [3]. Alexakos et al [2] suggested to use forensics technique on smart cities autonomous

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487112

vehicle sensor and IoT (Internet of Things) as Smart city data governance example, still needed real trial feasibility proof.

2.3 The Legal Background of Data Governance

2.3.1 Data Protection Act (DPA) 2018. The Data Protection Act (DPA) 2018 is the UK's implementation of the General Data Protection Regulation (GDPR). Everyone responsible for using personal data has to follow the controls on how personal information could be used and your rights to ask for your information data. The nowadays Data Protection Act 2018 replaced the Data Protection Act 1984 and the Access to Personal Files Act 1987, and implemented the EU Data Protection Directive 1995. The Privacy and Electronic Communications (EC Directive) Regulations 2003 altered the consent requirement for most electronic marketing to "positive consent" such as an opt-in box. Exemptions remain for the marketing of "similar products and services" to existing customers and enquirers, which can be given permission on an opt out basis.

Now, everyone in United Kingdom being responsible for using personal data has to follow strict rules called 'data protection principles'. They have to make sure the information is: used fairly, lawfully and transparently; used for specified, explicit purposes; used in a way that is adequate, relevant and limited to only what is necessary; accurate and, where necessary, kept up to date; kept for no longer than is necessary; handled in a way that ensures appropriate security, including protection against unlawful or unauthorized processing, access, loss, destruction or damage. For instance, here is stronger legal protection for more sensitive information, such as: race, ethnic background, political opinions, religious beliefs, trade union membership, genetics, biometrics and so on [3].

2.3.2 GDPR (General Data Protection Regulation). The GDPR (General Data Protection Regulation) according to the running of DPA, reorganized to the 6 data protection principles:

1. Lawfulness, fairness and transparency: the first principle is relatively self-evident: organizations need to make sure their data collection practices do not break the law and that they aren't hiding anything from data subjects.
2. Purpose limitation: organizations should only collect personal data for a specific purpose, clearly state what that purpose is, and only collect data for as long as necessary to complete that purpose.
3. Data minimization: organizations must only process the personal data that they need to achieve its processing purposes.
4. Accuracy: the accuracy of personal data is integral to data protection. The GDPR states that "every reasonable step must be taken" to erase or rectify data that is inaccurate or incomplete.
5. Storage limitation: similarly, organizations need to delete personal data when it's no longer necessary.
6. Integrity and confidentiality: this is the only principle that deals explicitly with security. The GDPR states that personal data must be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures". The GDPR is deliberately vague about what measures organizations should take, because technological and organizational best practices are constantly changing. Currently, organizations should encrypt and pseudo-science personal data wherever possible, but they should also consider whatever other options are suitable. GDPR defined the responsibility of data privacy governance, especially in EU countries. As the impact is on data protection, which includes organization based in European, or has branches in European or even provides services to the European residents [4].

2.3.3 Data Security Law. The Data Security Law of the People's Republic of China (adopted June 10, 2021, effective Sept. 1, 2021) is very useful for international matters, such as the Smart Chongli event and so on. NPCSC (the National People's Congress Standing Committee) states that Data Security Law has offered data protection by classification Data Protection, to protect legal right for citizen and organizations Security and development interests, defend state sovereignty. Data Security Law has 7 chaptered 51 Articles which promotes the development and utilization of data. Big data cases will be based on facts regarding to this law to process. Furthermore, as GDPR has to be enforced for European Union (EU) services etc. and if any case for EU residence, Data Security Law and GDPR would all be applied on the case. For big data, the management needs to consider their users and customers' requirements as one of the top priorities. Since new techniques like Big data, Smart cities, AI, IoT, edge intelligence the Cloud and so on growth very fast, while most computer laws only developed or updated in the recent decade, there are still many gaps need to be taken care of [5].

3 CASE STUDY

We use Smart Chongli as a data governance case study of information security protection future development. Usually, smart cities design put user's requirements first. For the Chongli Winter Olympics, the Smart Chongli developers have put the sportsman, athletics and the audience in the highest priority, to ensure that their security, privacy and safety are satisfied. A research on the data analysis for personal privacy and safety purpose have been published in the 2010th century [6] [7]. Furthermore, GDPR needs to be enforced for European Union (EU) athletics and the audience; the management needs also to consider their users and customers' requirements as one of the top priorities. Since new techniques like Big data, Smart cities, AI, IoT, edge intelligence, the Cloud and so on developed rapidly, while most computer laws only developed or updated in the recent decade, there are still many gaps need to be taken care of [1]. Especially during the pandemic, Ali and Dyo et al.'s work includes public key infrastructure (PKI) has attracted many interests [8].

3.1 Technical Discussion

Similarly, Smart Chongli data protection could be achieved through PKI infrastructure to implement and execute data governance in a hypothetical way.

Emrouznejad [10] has point out, only the authorized member can have the private key to access the protected team training data and the analysis result. NIST (National Institute of Standards and Technology) has published some standard, guidelines, articles and so on [11]. Some of them are quite appropriate to smart cities scenarios, which could be used for these team information data security [6]

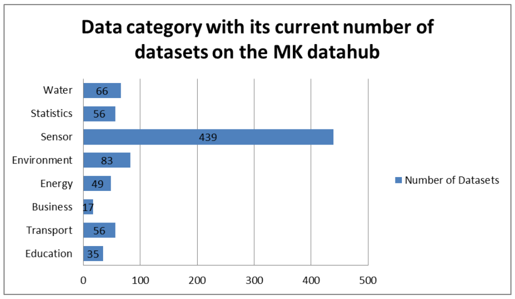

Figure 1: Data Profile of a Smart City Trial [9].

DEA (Data Envelopment Analysis) [10] and AI (Artificial Intelligence) technology [12] could be applied to the Smart Chongli data process. Based on the data statistics, carry out a significant analysis, then working out a pattern prediction and future possible trends. Utilizing these data of forecast, in order to produce a schematic planning for the similar events in the future. Take the data process into account, to consider preparing contest like Smart Chongli, to make future strategical development more and reasonably.

3.2 The Data in the Hub

During the Phase 1 of MKSmart project, activities supported by Milton Keynes City Council [13]; the University of Bedfordshire as one of the organization of participations had many experiments. Okai and et al. have already reported them at the International Workshop ACE-2018 and 2019 [14] [9]. Most of smart cities data are big data. The smart cities data process had something in common with normal data process. Smart cities data process also include: data acquisition, validation, storage, protection and analysis. Smart cities data process are required to ensure data accessibility, reliability, timeliness etc. Data acquisition should be objective. That is an organized foundation to guarantee an efficient analysis duration.

The percentage of the data profile of data-set in MKSmart project data hub is shown in Figure 1. For some peers' interests, Figure 1 shows the profile of data-set distribution in the MKSmart extracted from the smart city hub. The data-sets from the data hub had been processed, which would be one of the first trial of smart cities project in United Kingdom. This data profile of the data-set in the data hub shows a kind of distribution of the MKSmart project, which might represent a typical smart city data percentage of these kind. From Figure 1, in smart cities, Sensor data collected are still the largest portion, 54.8%. Environment monitor data are the second largest data-set, 10.4%. While Business data, because they are hard to monitoring, since the willingness, the Business data-set is the smallest, 2.1%.As for the other collection of the smart city hub data-set, water, energy resource, city transport and education and so on are more or less at the similar level of the monitored data amount, around about 7% to 8%. Therefore, we could through make use of these data by DEA method [9] to analyses and using AI technical to predict, such as this kind of smart cities' development trends and design accordingly, also on IoT collected data. Emrouznejad et al. [10] had already many successful examples, I wont show here, because of the paper's limit.

A live online data acquisition obtains real time data. And an online data analysis could process user's request in real time and permit user to make alterations or changes upon constraint and restriction. Smart cities data analysis include by certain AI algorithm, carries out statistics, display result in figures. Then, according to the output result, classify these data, to mine certain patterns, in order to predict trends. Therefore, strategic planning could be made on integrity, access control (AC), resource configuration, difference than expected value, privacy and security, reasonability, if any risk within data governance acceptability and so on [15]. Take the experience of MKSmart data governance into account, a solution is proposed [9] at later section. Following the legacy of the MKSmart data hub process of the trial project, those have been analyzed. From the data analysis of the trial project, smart cities experience had been gained [9]. It is significant to pass on to the future Smart Chongli event. In fact, both authentication access control, proactive intrusion detection system (IDS), 24/7 close monitoring and effective disaster recovery plan and so on should be all in place in smart cities strategic plan.

3.3 The Solution Suggested

Legal Consideration, on Cyber Law aspect, from legal theory of cyber science to a practical design application [16], GDPR, Network Security Act, Measures for the Administration of Data Security Act, E-Commerce Act. Data Security Law and so on need to be embedded to the data process governance project. Oki [9] has introduced Hadoop to Phase 2 of MKSmart project as smart cities data analysis tool; with Knox and Ranger, which is a REST (Representational State Transfer) API (Application programming interfaces) base perimeter security gateway that performs authentication, support monitoring, auditing, authorization management, and policy enforcement on Hadoop clusters. Hadoop has advantages as a data analysis tool in: high reliability, scalability, efficiency and high fault tolerance. Hadoop is a free framework and provides considerable storage for all sorts of data. With Hadoop robust analytical power and ability to performing multiple tasks, Hadoop may keep people reassuring about hardware availability. There are three areas to implement security in Hadoop: via cryptography, authentication, authorization [17]. The encryption ensures the information being trustworthy. Authentication ensures only genuine user, service accesses cluster. Authorization ensures accessibility service administration correct. In order to avoid strategy input error, Hadoop advantage could generate security shell. Other advantage is to monitor, recognize sensitive data and check any reveals and so on. The process in MKSmart Trial Project introduced [13] could be a kind of experience for future contest to take the references into account. Big data analysis of the MKSmart represented in Phase 1 and Phase 2 could be drawn support to the Smart Chongli event to speed up success and avoid detours.

4 EVALUATION

As shown in Figure 1 and Figure 2, a Smart Chongli data analysis needs safety and security protection solutions to be implemented. Furthermore, PII (personal recognizable information) data privacy need to be take into consideration [17] [18]. These should not be in the normal data hubs, unless there is a consent in place. Therefore,

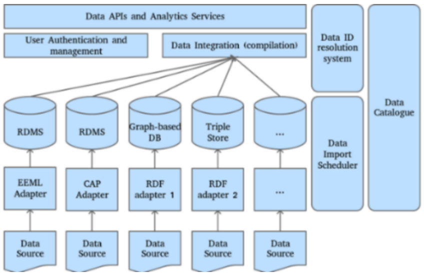

Figure 2: Architecture of an MKSmart Trial Project on Layered System Management of the Data Hub [9].

the data storage in the cloud, authority access and process will be an issue. The data monitoring and analysis could be a daily task, also needs to pay attention to. Perimeter access control and extra host based access control [12] should be applied to these data storage. Robust cipher using modern PKI cryptography technique could be applied. Figure 2 had been approved working efficiently in MKSmart Phase 1. Figure 3 had been worked well by some projects' experience, to be one of the better recent data process implementation tools to date. Hadoop is Suitable for MKSmart Phase 2, but there is still a pitfall that not 10% distributed when process metadata. Hopefully the next version could improve on this. A schematic table of Hadoop architecture is shown in Figure 3. Bean [19] had further discussion on DPA 2018 and so on. Nevertheless, the contribution of this paper is: through MKSmart trial project Phase 1, the authors recognized the existing issue, which was in common in smart cities globally. Through this paper, the authors would like to suggest Smart Chongli (as well as the future smart cities strategy planners) to take these outstanding issues and our data governance suggestions including PII into account, for consider designing and executing comfortable and well-served smart cities for the participating citizens in the world [12].

5 CONCLUSIONS

In this paper, we have discussed the recent cyber security threats and possible solutions for Smart Chongli event. We suggested that the event organizers not only follow the Measures for the Administration of Data Security 2019 and Network Security Act 2016 (Cybersecurity Law of the People's Republic of China 2016), meantime, but also trying to meet the sports team's requirements and audients' expectations. Cryptography being applied on Hadoop to execute security protection for Smart Chongli's event. In addition, by means of proactive IDS system plus AC methodology and so on, to realize a possible maximum protection against potential cyber security attack on PII and privacy and so on. The research work reported is interim of a trial Smart Chongli project international consideration. The analysis of big data result will be represented in the forthcoming paper. The contribution of this paper include: through MKSmart trial project Phase 1, the authors recognized the existing outstanding issue. The authors would like to by publish this paper, to suggest Smart Chongli organizers to take the issue and our suggestions into account, to design and implement comfortable and good serviced smart cities in the world.

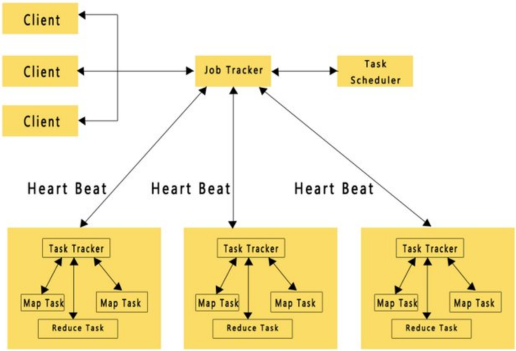

Figure 3: The Hadoop Architecture Scheme [19].

A suggested future development is for a more appropriate approach to finalize the method, which related with potential issues, including GDPR and PII [6]. The process of private data storage, access and analysis will also be applied to the development of Administration of Data Security, with security audit for the time being [7]. A good compensation will be an ongoing process in the near future and far future. Furthermore, developing AI tools maybe helpful in the kind of circumstances. Our problem solving could also be suggested to other smart cities similar events to sort out their gap for GDPR or DPA (Data Protection Act) 2018 or Data Security Law smart cities big data governance issues [20] [21].

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the support received on the trial data hub from the MKSmart consortium.

REFERENCES

[1] X. Feng, Y. Feng, *et al.* (2019a). "*Computer Laws Consideration on Smart City Data Planning of Chongli 2022*" IEEE Xplore 2019 Smart City Congress, Proceeding of International Workshop ACE-2019, Leicester, UK

[2] C. Alexakos, C. Katsini, K. Votis, A. Lalas, D. Tzovaras and D. Serpanos (2020) "*Enabling Digital Forensics Readiness for Internet of Vehicles*" 23rd EURO Working Group on Transportation Meeting, EWGT 2020, Paphos, Cyprus, ScienceDirect. .sciencedirect.com. Transportation Research Procedia 52. 339–346.

[3] L. Woehler (2014). "The first cloud computing platform to conform to ISO/IEC 27018, the only international set of privacy controls in the cloud", Microsoft.

[4] N. Hawthorn, *et al* (2015), "*White paper: How European Union data protection affects your data in the cloud*", the new EU data protection regulations, Skyhigh and DMH Stallard LLP, Euro Cloud Expo 2015.

[5] J. Xu (2021) "*Legal Protection of Personal Data in China*", IEEE International Workshop ACE-2021, Canada.

[6] ICO (2016) "*Overview of the General Data Protection Regulation (GDPR)*". https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr.

[7] X. Feng and Y. Feng (2019) "*Smart Cities Design Consideration*", represented at the 3rd ACM International Conference on CS & Application Eng. CSAE 2019, China.

[8] J. Ali and V. Dyo (2021) "*Cross Hashing: Anonymizing Encounters in Decentralised Contact Tracing Protocols*", The 35th International Conference on Information Networking (ICOIN 2021), Jeju, Korea. 13-16 January 2021.

[9] E. Okai, X. Feng, *et al.* (2019) "*Smart City Challenges*", Proceeding of IEEE International Workshop ACE-2019.

[10] Ali Emrouznejad (2016). "*Big Data Optimization: Recent Developments and Challenges.* In the series of "Studies in Big Data", Springer-Verlag, ISBN: 978-3-319-30263-8.

[11] National Institute of Standards and Technology (2018) "*Law Enforcement*". NIST. USA. [Accessed 19/04/2019]

[12] X. Feng, *et al.* (2021) "*NHS Big Data Intelligence on Blockchain Applications*". Big Data Intelligence for Smart Applications. Editors B. Youssef, Yassine M, *et*

al. Series: Studies in Computational Intelligence, Vol 994, Publisher Springer International Switzerland. eBook ISBN 978-3-030-87954-9.

[13] MKSmart, (2017) "*About MKSmart*". Milton Keynes City Council. http://www.mksmart.org/about/ [Accessed 20/May/2019].

[14] E. Okai, *et al.* (2018): "*Smart Cities Survey*". The IEEE 16th International Conference on Smart City; 4th International Conference on Data Science and Systems (SmartCity/DSS). Exeter, UK. https://ieeexplore.ieee.org/abstract/document/8623018. [Accessed: 10/10/2019].

[15] X. Feng, *et al.* (2020) "*Artificial Intelligence and Cyber Security*", IEEE International Workshop ACE2020, Calgary, Canada.

[16] J. Holt *et al.* (2011) "*A Managers Guide to IT Law*", British Computer Society, 2nd Ed. ISBN 10: 1906124752, ISBN 13: 9781906124755.

[17] A. Adegoroy, *et al.* (2020) "*A new perspective on the issue of privacy: Covid-19 pandemic vs. privacy*". The 19th International Conference of Internet.

[18] R. Feng, *et al.* (2021) "*Robot, Edge Intelligence and Data Survey*". The 5th International Workshop ACE-2021, Athabasca University.

[19] J. Bean (2016) "*Introduction to Hadoop Security: A Hands-on Approach to Securing Big Data Clusters*" O'Reilly Media, ISBN: 1771375051.

[20] E. Washington (2017) "*Why Data Governance is Crucial for Big Data Environments*", Tdwi. https://tdwi.org/articles/2017/09/15/diq-all-data-governance-in-big-data-world.aspx

[21] The Standing Committee of the National People's Congress (2021). "*Data Security Law*", Data Security Law, states data classified protection, the legislation session passed by the Standing Committee of the National People's Congress September 2021.