FUSION FROM BIG DATA TO SMART DATA TO ENHANCE QUALITY OF INFORMATION SYSTEMS

Enhancing the Quality of Information Systems using smart data

Frank, Febiri*

Faculty of Economics and Administration, University of Pardubice, frank.febiri@upce.cz

Meseret Yihun, Amare

Faculty of Economics and Administration, University of Pardubice, Meseret.Amare@upce.cz

Miloslav, Hub

Faculty of Economics and Administration, University of Pardubice, Miloslav.Hub@upce.cz

The use of the term “smartness” in the context of data indicates relevancy based on the intended purpose of data. The Internet of Things (IoT) and advancements in technology have resulted in an ever-increasing pool of data available to all institutions to derived meaning and make sound decisions from them. The research presented in this paper explored the role smart data play in information systems quality through a qualitative study of how using the large pool of data (big data) and fusing it to smart data organizations can make sound and smart decisions using the available techniques. We use an existing architecture adopted for a public institution to analyse how data ingestion can be achieved with minimum challenges. The findings suggest that even though there exist large pool of data for most organization, it is becoming more difficult using this data to make organizational sense due to the challenges pose by such data. The realisation of smart data and its benefits in information systems helps to improved quality of information system which reduce cost and promote the smartness agenda of today’s organization.

Additional Keywords and Phrases: Information Systems, Smart Data, Big Data, Quality Measure

---

* Place the footnote text for the author (if applicable) here.

1 INTRODUCTION

In recent years, gathering data about a system from various sources, from different granularity levels, and of multiple aspects has been a common practice in the field of science and technology to enhance the quality of information and improve government and planning (Khan et al., 2021). With the help of the Internet of Things (IoT) and other data collection and data mining tools, a plethora of data sources are available to both public and private sectors.

Data fusion is the process of data integration from various sources and from multiple aspects to produce comprehensive, unified, and specific data for an entity. The objective of data fusion is to solve the problematic data, extract the data that contains a higher level of information and increase the completeness of data (Sun et al., 2018). Data is generally divided into two data types, big data, and small data.

Big data is a huge volume of data that is sometimes structured and organized and sometimes random and unstructured. It refers to a large volume of data and that is defined by the key factors of data volume, data variety, data velocity, and data value, and data veracity. The aspects of volume, velocity, and variety refer to the size of data and its diversity, the aspects of value and veracity refer to the quality of data and to what extent this data is useful. These aspects in recent years have been more crucial where the usefulness of data does not depend on its huge size, rather on its quality. While smart data is the type of data in which patterns and signals are extracted with the smart algorithm and using appropriate architecture. Smart data filters some certain features of the data such as noise and produces worthy data that enhances the quality of the information system. The fusion of big data to smart data is acquired with the help of machine learning, data mining and IoT that enables reliable, efficient, and accurate management and decision making of ubiquitous atmosphere and ultimately improves the quality of information system (Triguero et al., 2019).

The fusion of big data into smart data allows integrating multiple data that creates useful, accurate, and consistent results than data from an individual source. The study of data from multiple sources and multiple aspects enhances the quality of the information system. Smart data addresses the issue of noise to a great extent and only provides the data that is useful and structured. With an enormous growth of the urban population, the demand for energy resources and advancements in the field of information systems has also increased to create a sustainable living environment for the urban population. The fusion of big data to smart data serves this purpose best and provides ways to enhance the quality of information systems adding competency, comprehensiveness, and reliability of the data.

The available superfluity of data poses difficulties in addressing the issue of data noise filtration with traditional means. In this regard, new architecture and algorithms should be considered for the fusion of big data to smart data and to address the issue of noise and other problems to enhance the quality of information systems. The quality of information has a direct relation to the data from which the information is drawn. Therefore, scientists and researchers emphasize improving the quality of available data and other errors that the available data contains. Another major problem with this form is data is its complexity in managing. Moving from a legacy data management system and integrating a new solution comes as a challenge in itself. Data coming from multiple sources, and IT teams creating their own data while managing data, systems can become complex quickly and may cause data confusion which can distort the effective administration of information systems.

The importance of data security cannot go unnoticed with the chunk of data available to information systems. To enhance maximum security of personal and other corporate data that come through these systems, data needs to be stored properly, which starts with encryption and constant backups of the required data. This paper aims to explore the ways and techniques of big data fusion to smart data and its applications to enhance the quality of information systems using an adopted architecture of a public university.

2 RELATED RESEARCH

The term “Smart Data” contrasts with the term “Big Data”, which usually refers to a combination of structured and unstructured data that may be measured in petabytes or exabytes. Smart Data, in contrast, consists of usable datasets derived from Big Data repositories. Volume, Velocity and Variety are the three Vs in the original definition of the key characteristics of Big Data according to the research report published by META Group (D. Laney, 2001). Volume refers to the size of the data, Velocity refers to the speed of data generation and Variety refers to different types/sources of data. Since then, other factors have also been considered, such as Veracity (trustworthiness of the data obtained) and Value (usefulness of data) (A. Splendiani). As a huge amount of research has been done in Big Data processing, usually focused on Volume, Velocity and Variety, we focus on how we can reduce the noise and identify the most reliable data that are useful for information systems purposes (Veracity and Value).

There are ample sources and literature already written to address the issue and to provide insights for data fusion from big data to smart data. Sagiroglu and Sinanc (2013), in this regard note that “Big data is a term for massive data sets having a large, more varied and complex structure with the difficulties of storing, analysing and visualising for further processes or results”. The huge random and unstructured data contains noise that produces distortion in information system and affects its quality negatively. As García-Gil et al., state in a paper, “Among all the problems that may appear in the data, the presence of noise in the dataset is one of the most frequent. Noise can be defined as the partial or complete alteration of the information gathered for a data item, caused by an exogenous factor not related to the distribution that generates the data” (García-Gil et al., 2019). To improve the quality of information systems in both private and public sectors, data from multiple sources and of different aspects is required that is executed with the help of data fusion of big data. In this regard, With the availability of parallel data sources in various smart city domains, data fusion techniques that combine multiple data sources, lie at the heart of smart governance platform integration.

The conventional data fusion approach, which is regarded as a part of data integration, is a process of integration of multiple data representing the same real-world object into a consistent, accurate, and useful representation (J. Bleiholder & F. Naumann, 2008). However, in the big data era, there are multiple datasets generated in different domains, which are implicitly connected by a latent object. For instance, traffic conditions, POIs and demography of a region describe the region's latent function collectively, while they are from three different domains. Literally,

records from the three datasets describe different objects, i.e., a road segment, a POI, and a neighborhood, respectively. Thus, we cannot merge them straightforwardly to obtained the needed knowledge for decision making, Instead, we need to extract knowledge from each dataset by different methods, fusing the knowledge from them organically to understand a region's function collectively. This is more about knowledge fusion rather than the quantum of data available.

As Zeng argues, "Smart Data aims to filter out the noise and produce valuable data. Although an unprecedentedly large amount of sensory data can be collected with the advancement of the Cyber Physical-Social systems, the key is to explore how Big Data can become Smart Data and offer intelligence" (Zeng, 2017).

Studies suggest that the fusion of big data to smart data has proven to be enhancing the quality of the living environment in urban cities and contributing to improved provision of energy and power facilities. As Marjani et al., describe in their paper "IoT has witnessed its recent adoption in smart cities with interest in developing intelligent systems, such as smart office, smart retail, smart agriculture, smart water, smart transportation, smart healthcare, and smart energy" (Marjani et al., 2017). United Nations estimates that by 2050, 68% of the global population will be living in urban populations (UN, 2018).

The concepts of smart cities and smart governance in recent years have emerged as "A city combining ICT and Web 2.0 technology with other organizational, design and planning efforts to de-materialize and speed up bureaucratic processes and help to identify new, innovative solutions to city management complexity, to improve sustainability and livability" (Lau, 2019). The quality of information system, for better government, planning and decision-making process, becomes very important in such cities.

Random, partial, and single-sourced data are not sufficient in this regard. Therefore, the fusion of big data into smart data emerges as an important concept that assists in improving transportation, human mobility, the quality of information system, and better decisions for the provision of quality facilities in such cities. With such advancements and emerging urbanization trends, smart data is essential to improve information systems, decision making, government, and planning in facets of the economy. Although some of the desired results can be produced by mining and analyzing big data, the fusion techniques explore the ways to integrate both kinds of data that offer more intelligent and efficient solutions to information system management. This papa thus explores the advancements in using smart data to improve quality of information systems by using a public educational institution's architecture and a mean of collecting big data to smart data.

3 DATA CHALLENGES AND FUSION TECHNIQUE

3.1 Big Data collection

Data ingestion is the first step in data information systems management. The data ingestion layer is responsible for the collection of data and transformation for analysis. As the strategic and modern approach to designing the data pipeline ultimately drives organizations value, data analysts, managers, and decision-makers need to understand data ingestion and its associated technologies. Data ingestion helps to bring various types of data sources from its source into a system where it is easy to be analyzed and stored. Data collection is critical and should be emphasized for any big data project, as the volume of data is becoming very large (Singh & Kumar, 2019). Data handling is always a challenge and critical activity if the amount of data becomes huge and consists of various formats. As big data information systems are designed to process unstructured or semi-structured data, it becomes complex to capture data from different sources. Data ingestion is generally becoming complex in data systems as data sources and processing now includes batch and stream formats which, increase the complexity and management of information systems. Besides, the ever-increasing IoT devices are resulting in a large volume of various data sources. Hence, extracting data using traditional data ingestion approaches becomes a challenge (Shekhar, 2019). The following section explores the challenges of data collection in public administration organizations information systems.

3.2 Big Data Challenges

The absence of a complete big data architecture framework tailored for higher organizations and institutions that serve as a guideline for an overarching process is one of the existing challenges of implementing big data analytics at public institutions (Matsebula and Mnkandla, 2017). Existing architecture does not give a detailed learning analytics process of big data in public institutions. There are no designated specific tools or methodologies for gathering, cleansing and using captured data (Hadwer et al, 2019).

The traditional standards of data architecture are changing at an ever-increasing rate. Preceding enterprise architectures are undergoing significant technological changes in the face of new trends, including big data, non-relational data stores, IoT, machine learning, and artificial intelligence, data lakes, and many others. The following are the key challenges that can impact data ingestion and pipeline performances (Shekhar, 2019):

- Data quality: Data quality as a challenge crops up when working with diverse data sources. Inconsistent data formats, data repetition, and missing values would make analysis unreliable. Thus, bringing the data together should be done after the data is appropriately analyzed and prepared.
- Slow processes: Writing codes to ingest data and manually creating mappings for extracting, cleaning, and loading data can be cumbersome as data today has grown in volume and has become highly diversified. Therefore, there is a move towards data ingestion automation. The old methods of ingesting data are not quick enough to persevere with the large volume and variety of data sources.

- High complexity: The ever-increasing of new data sources and IoTs, public institutions are confronting challenges to make data integration in order to insights value from their data. The main challenge is the ability to attach to that data source, recognizing and error elimination, and inconsistent data structures.
- Cost: The infrastructure that enables the data acquisition process from the sources and the cost of associated ingestion tools makes the data ingestion task very costly.
- Scaling: The overall performance may decrease if the issue is not addressed correctly during the planning phases of building the architecture.
- Unreliability: Incorrectly ingesting data could result in unpredictable connectivity. This further can disrupt communication and cause loss of data in the information system stream.
- Security: Data security is the biggest challenge that could occur when the data moved from one source to the storage system since data are staged in numerous phases throughout the ingestion process. Consequently, it is challenging to accomplish security standards during the data ingestion process. Data ingestion can compromise compliance and data security regulations, creating further complexity and cost. It requires to introduce advanced security techniques of data encryption and anonymous access to sensitive information.

4 DATA FUSION TECHNIQUE

The main goal of big data collection is to read data provided in various communication channels, frequencies, sizes, and formats. This layer takes care of categorizing the data for the smooth flow of data into the further layers of the architecture. In order to load data, REST APIs, streaming tools, and web scraping techniques are mostly applied.

By analyzing a selected available, it was possible to explore data collection challenges, identify data sources, and develop data fusion technique. This proposal of data collection architecture was based on the guidelines and best practices of general big data architectures from works of literature. In order to cope up with the challenges that the data ingestion process is facing, the following best practices can be incorporated in the process of developing the big data ingestion architecture:

- Design for performance: In order to benefit from big data at higher institutions, it’s necessary to design an efficient mechanism to ingest available data from different sources and prepare for learning analytics. As recommended by (Amare & Simonova, 2021), the proposed architecture was designed to process big data by dividing large data sets to allow it to run independently in parallel to satisfy high-performance computing and work efficiently without having to worry about intra-cluster complexities, monitoring of tasks, node failure management.
- Data collection automation: As the data in the institution continues to grow both in volume and complexity, it’s wouldn’t be possible to curate a huge amount of data using manual techniques. Automating the ingestion process would shorten the time takes for ingestion, increases productivity, and reduces

manual efforts. The learning analytics needs several new data sources to be ingested on-demand with minimal user intervention.

- Real-time collection technique: Some institutional data is time-sensitive and needs to be processed as soon as they are collected. Data is extracted, processed, and stored as soon as it is generated for real-time decision-making.
- Batch collection technique: There are requirements where data in organization needs to move at regularly scheduled intervals. This type of data fusion is suitable for repeatable processes.

The proposed architecture combines the batch and real-time ingestions to utilize batch processing to offer broad views of batch data and real-time processing to increase the quality of data used in the information system .

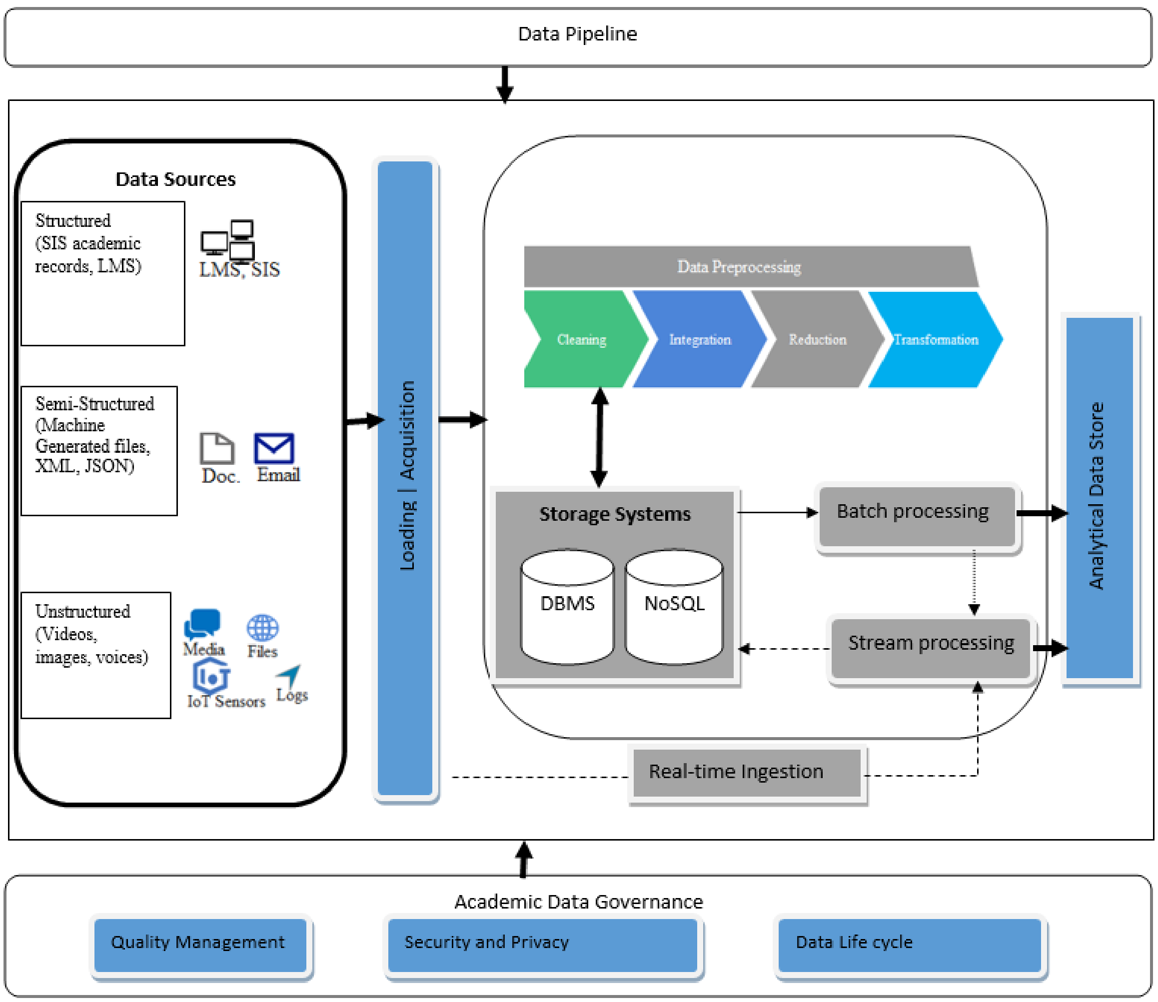

Figure 1:Big data collection architecture (Amare & Simonova)

To make the system processes easier, data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation.

- Data cleaning: Data cleaning refers to techniques to clean data by removing outliers, replacing missing values, smoothing noisy data, and correcting inconsistent data.
- Data integration: As the institutions data is collected from multiple sources, data integration is a vital part of the process. The following are the most common activities to integrate data: (1). physically bringing the data all to one data store. This usually involves Data Warehousing. (2). Copying data from one location to another using application. It could be synchronous or asynchronous and is event-driven. (3). Virtualization of data using an interface to provide a real-time and unified view of data from multiple sources. The data can be viewed from a single point of access.
- Data reduction: The purpose of data reduction is to have a condensed representation of the data set which is smaller in volume, while maintaining the integrity of original data.
- Data transformation: Transforming the data into form appropriate for Data-modeling and information systems reprocessing is the final step of data pre-processing. Some of the strategies that enable data transformation include: Smoothing, attribute construction, aggregation, normalization, discretization, and concept hierarchy generation for nominal data.
- The transformation engine is capable of moving, cleaning, splitting, translating, merging, sorting, and validating data. For example, structured data such as that typically resided in staff record might be extracted from staff information systems (SIS) and subsequently converted into a specific standard data format, sorted by the specified and then the record validated against data quality rules.

5 SMART DATA IMPACT ON INFORMATION SYSTEMS

Previous study indicates that public organisations most commonly adopt innovative ways of improving their information systems. At thew core of these approaches is the data used in these systems and how they are processed. (Cordella & Bonina, 2012; J. R. Gil- Garcia et al., 2016). Quality of information systems is important to the management processes of both public and private services, as organisations always strive to provide measures of quality to demonstrate the responsible use of funds. In this study, a key examples of quality factor in information systems have been discussed. We linked quality information system to quality data processing and to increased productivity through a decrease in the time required to complete tasks, and distortions caused by noisy and bulky information.

As the population across the globe is moving towards urbanization, many countries are building smart cities, smart industries, smart infrastructure, smart economics, and smart government. This situation makes big data less efficient in managing smart cities and providing an efficient environment. With Today's advanced technologies such as IoT, cloud computing, FOG, and others, a plethora of data is available (Huh et al., 2017). This data, however, is random, unstructured, and of a single perspective that does not contribute to the quality of the information system. This big data also makes the quality of information system poor and hence puts obstacles towards making better decisions for the management of institutional resources, government and planning.

Some Techniques involved in data fusion such as association techniques, input and output data techniques, data relation fusion techniques and the JDL technique are available to help change bulky data to smart and more meaning data for information systems use. The steps of the data ingestion techniques help to improve the data by providing information from different sources or combining the information of a new meaningful source with the existing information. All these techniques involve a set of operation and practices to fuse the data from big data to smart data with the single goal to improve the quality of the information system. A decision is typically taken based on the knowledge of the perceived situation, which is provided by many sources in the data fusion domain. These techniques aim to make a high-level inference about the events and activities that are produced from the detected targets. These techniques often use symbolic information, and the fusion process requires reasoning while accounting for the uncertainties and constraints. These methods fall under level 2 (situation assessment) and level 4 (impact assessment) of the JDL data fusion model. The improved information system is then used to make certain decisions development and planning and resource management.

To enhance the quality of information system data fusion from big data to smart data is important. The result of the aforementioned details suggests that the fusion of data sources provides various perspectives and multiple data sources and produces better results for the decision-making process. Furthermore, the literature review suggests that the fusion of data improves the quality of information system for better planning and management of data resources. Data fusion identifies different perspective of data that improves the entire information system and leads to taking broad decisions. The data fusion process improves the reliability of data, extracts higher level and smart information from data sources, enhances the completeness of data, and fixes the data that has errors. Additionally, a fusion of data addresses the noise in data as well that disturbs the information and hence improves the information system.

6 CONCLUSION

The fusion of data from big data to smart data has a wide range of potential implications. As the world is moving towards new trends and smart cities, it is very important to have diverse and useful data to make decisions regarding planning and resource management Sagiroglu and Sinanc (2013). Hence it becomes of paramount importance to evaluate the data based on multiple-perspective and from multi-data sources that involve more than a single data source and domains. The improved quality of information system that is obtained from the fusion of big data sources to smart ones ultimately improves the smart industry, transportation, urban management, communication, resource management, education, governance, planning, and others. In the energy sector, its potential applications are improved smart grid systems, improved communication, and improved supply of energy resources. Smart data is changing the future of big data. It lessens the risk of data loss, adds value, increases efficiency and minimizes expected loss. Quality data is more important than the volume of data. Privacy protection is one example of the benefits of using smart data, as Koeck suggests "there is a big tendency to leverage unstructured data gathered about the customer via social media and other sources in order to increase a number of variables in the algorithms for the credit decision for example. By removing this big data from the equation and only using the valuable smart data, the less an institution will leave itself vulnerable.

Arguably the most important function of this data is to enhance organizational performance, and transforms organizational operations (Nick Ismail, 2016).

Additionally, this large size of data includes a number of data sources such as cyber data sources and hybrid data sources (Huh et al., 2017) which difficult to analyze, mine or extract and create good from them. Although, big data sources contain a huge volume of data, yet this data is not complete and useful to get smart and comprehensive results. Therefore, in order to make data useful for the purposes of information systems, to produce accurate and precise outcomes it is important to improve the quality of information that the data contain. This is solely possible by mining and analyzing several data sources at the same time with the help of machine learning tools and data mining techniques.

Conclusively, this study has two main contributions for the field of information systems developments and improvements. First, the study shows how the use of smart data in information systems provides both internal and external benefits and affects several dimensions of the smartness framework. This can be extended to demonstrate that – in both the public and private sectors – smartness is created through the combination of information systems, people and organisations. Second, we show how the dimensions of quality, effectiveness, efficiency and cost reduction can be improved using smart data in place of big data. Furthermore, by taking into account contextual challenges and limitations of big data, we make conclusions that smart data is the way forwards in improving the quality of information systems. Finally, this study was subject to several limitations that must be acknowledged. Our focus on information systems was generalized to include all systems both in public and private administrations, However, new and emerging technologies adopted in some information systems are able to eliminate some of the above-mentioned challenges of big data discussed above. We identified and adopted an architecture which we think is of great related to avoiding problems with data ingestion, this may have its own limitations when used for other private and public settings as the case may be. Again, certain externalities may affect information system quality which may not be caused by big data challenges, these were not discussed. Future research is needed to reflect other factors that affect the quality of information systems. Smart data alike big data comes with its own challenges, future studies can be conducted to look into more challenges that comes with using smart data as a means of measuring the quality of information systems.

REFERENCES

[1] A. Splendiani, M. Donato, and S. Drăghici, "Ontologies for bioinformatics," in Springer Handbook of Bio- /Neuroinformatics, 2014, pp. 441–461. Available: https://doi.org/10.1007/978-3-642-30574-0_27.

[2] Creswell, J.W. and Creswell, J.D., 2017. Research design: Qualitative, quantitative, and mixed methods approach. Sage publications.

[3] D. Toppeta, "The smart city vision: how innovation and ict can build smart,livable, sustainable cities," The Innovation Knowledge Foundation, vol. 5, pp. 1–9, 2010.

[4] García-Gil, D., Luengo, J., García, S. and Herrera, F., 2019. Enabling smart data: noise filtering in big data classification. *Information Sciences*, *479*, pp.135-152.

[5] Himeur, Y., Alsalemi, A., Al-Kababji, A., Bensaali, F. and Amira, A., 2020. Data fusion strategies for energy efficiency in buildings: Overview, challenges and novel orientations. *Information Fusion*, *64*, pp.99-120.

[6] Huh, S., Cho, S. and Kim, S., 2017, February. Managing IoT devices using blockchain platform. In *2017 19th international conference on advanced communication technology (ICACT)* (pp. 464-467). IEEE.

[7] J. Bleiholder and F. Naumann. Data fusion. ACM Computing Surveys, vol. 41, No. 1, pp. 1–41, 2008

[8] Johnston, M.P., 2017. Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries*, *3*(3), pp.619-626.

[9] Kakati, T., Bhattacharyya, D.K. and Kalita, J.K., 2020. X-Module: A novel fusion measure to associate co-expressed gene modules from condition-specific expression profiles. *Journal of biosciences*, *45*(1), pp.1-13.

[10] Khan, S., Nazir, S., García-Magarino, I. and Hussain, A., 2021. Deep learning-based urban big data fusion in smart cities: towards traffic monitoring and flow-preserving fusion. *Computers & Electrical Engineering*, *89*, p.106906.

[11] Lau, B.P.L., Marakkalage, S.H., Zhou, Y., Hassan, N.U., Yuen, C., Zhang, M. and Tan, U.X., 2019. A survey of data fusion in smart city applications. *Information Fusion*, *52*, pp.357-374.

[12] Lloret, J., Canovas, A., Sendra, S. and Parra, L., 2015. A smart communication architecture for ambient assisted living. *IEEE Communications Magazine*, *53*(1), pp.26-33.

[13] Lund, H., Østergaard, P.A., Connolly, D. and Mathiesen, B.V., 2017. Smart energy and smart energy systems. *Energy*, *137*, pp.556-565.

[14] Mallia, J.G., 2014. Inductive and Deductive Approaches to Teaching English Grammar. *Arab World English Journal*, *5*(2).

[15] Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I.A.T., Siddiqa, A. and Yaqoob, I., 2017. Big IoT data analytics: architecture, opportunities, and open research challenges. *ieee access*, *5*, pp.5247-5261.

[16] Momoh, J.A., 2012. *Smart grid: fundamentals of design and analysis* (Vol. 63). John Wiley & Sons.

[17] Pacifici, K., Reich, B.J., Miller, D.A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. and Collazo, J.A., 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, *98*(3), pp.840-850.

[18] Sagiroglu, S. and Sinanc, D., 2013, May. Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

[19] Singh, G. and Kumar, E., 2017. Input data scale impacts on modeling output results: A review. *Journal of Spatial Hydrology*, *13*(1).

[20] Sun, G., Chang, V., Guan, S., Ramachandran, M., Li, J. and Liao, D., 2018. Big Data and Internet of Things—Fusion for different services and its impacts. *Future Generation Computer Systems*, *86*, pp.1368-1370.

[21] Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S. and Herrera, F., 2019. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(2), p.e1289.

[22] UN, 68% of the world population projected to live in urban areas by 2050, says UN. UN, 2018. [Online]. Available: https://www.un.org/development/desa/en/news/population/ 2018-revision-of-world-urbanization-prospects.html

[23] Yan, P.Y., Ding, Y., Chen, Z., Liu, S., Michalopoulos, A., Chen, R., Gulzar, Z.G., Yang, B., Cieply, K.M., Luvison, A. and Ren, B.G., 2014. Novel fusion transcripts associate with progressive prostate cancer. *The American journal of pathology*, *184*(10), pp.2840-2849.

[24] Žitnik, M. and Zupan, B., 2014. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, *37*(1), pp.41-53.

[25] Singh, C., Kumar, M. (2019). Mastering Hadoop 3: Big data processing at scale to unlock unique business insights. Packt Publishing Ltd.

[26] Shekhar, C.(2019, August 20). Big Data Ingestion: Parameters, Challenges, and Best Practices. Retrieved from:https://www.datapine.com/

[27] Hadwer, A., Gillis, D., Rezania, D. (2019). Big Data Analytics for Higher Education in The Cloud Era, *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*(pp. 203-207), Suzhou: IEEE.

[28] Matsebula, F., Mnkandla, E. (2017). A big data architecture for learning analytics in higher education, *2017 IEEE AFRICON*Conference,(951-956),Cape Town:IEEE.

[29] Amare, M.Y. & Simonova, S. (2019). Overview of Big Data Challenges, Opportunities and its Applications in the Context of Public Administration Organizations.In Proceedings of the 34rd International Business Information Management Association Conference 2019 (IBIMA) (pp. 12200 – 12209), Madrid: IBIMA Publishing.

[30] Amare, M. Y., & Simonova, S. (2021). Learning analytics for higher education: proposal of big data ingestion architecture. In SHS Web of Conferences (Vol. 92). EDP Sciences.