Technical Report: Big Data – Concepts, Infrastructure, Analytics, Challenges and Solutions

Kaya Kuru
School of Engineering and Computing
University of Central Lancashire
Preston, UK
https://orcid.org/0000-0002-4279-4166

Abstract

Industry 5.0 is emerging while swarms of Cyber-Physical Systems (CPSs) are being integrated with other swarms of CPSs and humans to work co-actively in collaborative and sustainable environments, which leads to the development of cyber-physical-social-systems (CPSSs) by leveraging the insights filtered from the passed experiences, i.e. Big Data (BD). With Industry 5.0, the borders between humans and intelligent machines cannot be readily distinguished in the new decentralised world – what is created by AI? What is produced by humans or what is built by both? The recent advances in the CPSs and domains, cloud and edge platforms along with the advanced communication technologies are playing a crucial role in connecting the globe more than ever, which is creating large volumes of data at astonishing rates and a tsunami of computation within hyper-connectivity. Data analytic tools (e.g. ChatGPT, Gemini) are evolving rapidly to harvest these explosive increasing data volumes. In this direction, this technical report analyses the concept of BD, infrastructure of BD, BD analytics, and challenges in the processing of BD and discusses practical solutions for these challenges toward Industry 5.0.

Index Terms

Big Data, Big Data Analytics, Cyber-Physical Systems (CPS), Cyber-Physical-Social-Systems (CPSSs), Industry 5.0, Internet of Everything (IoE), Cloud Platform, Edge/fog computing, Federated Learning (FL), Collaborative Learning (CL).

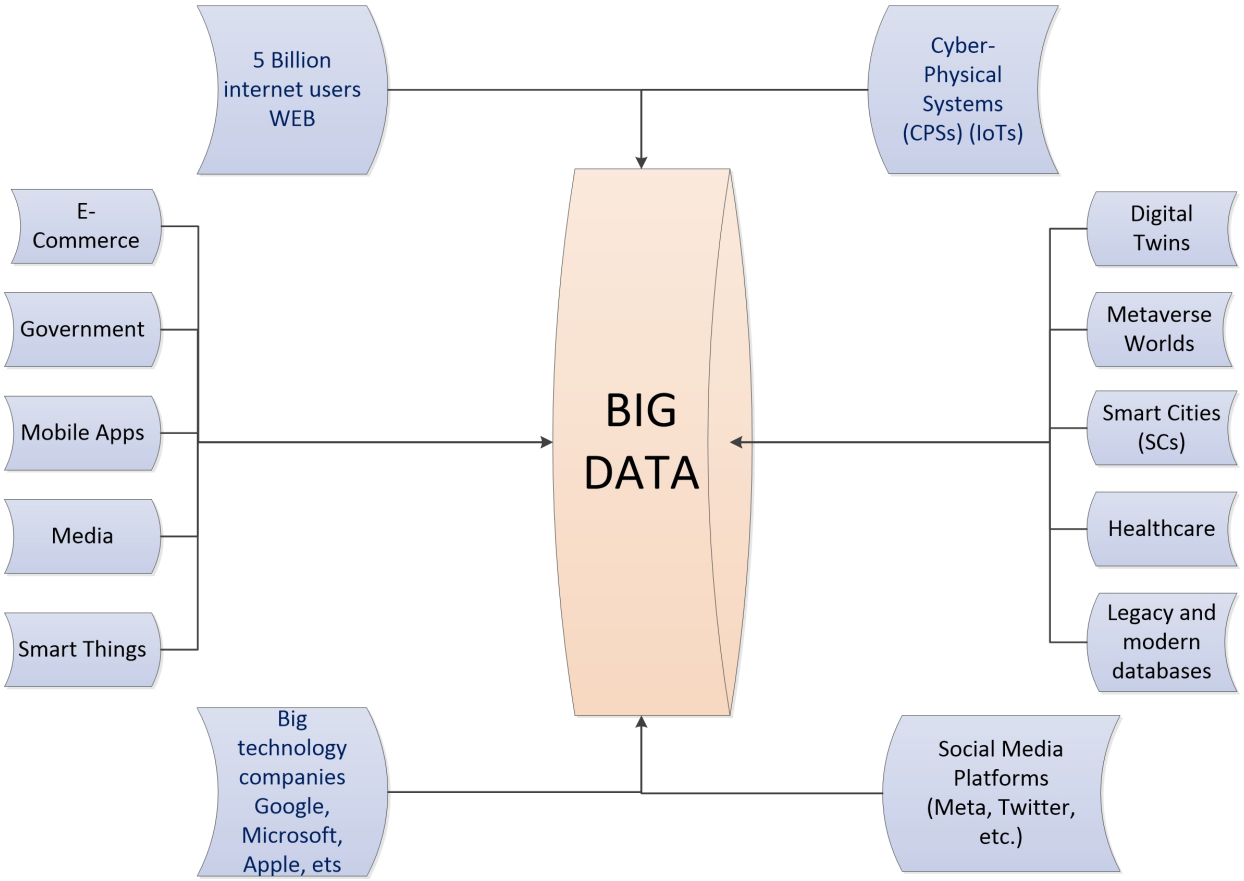

Fig. 1: Major sources of Big Data.

I. Concepts of Big Data (BD)

Humans and robots are bound to work co-actively in collaborative environments in Industry 5.0 to produce value-added outcomes using intelligence obtained from the passed experiences, i.e. Big Data (BD). Not only does Industry 5.0 aim to automate intelligent robots to work as a teamwork, but also it aims to automate human and machines in a highly synergistic collaborative environment, enabling to increased Quality of Products (QoP), Quality of Experiences (QoE), and Quality of Life (QoL).

Fig 1 demonstrates the major sources of BD. Fig 2 shows a glimpse at the data being created each minute on the internet and it's a breathtaking view of how the volume and variety of data keep accelerating, showing no signs of slowing down [1] on distributed platforms. The world's internet population is growing significantly year by year; as of April 2022, the internet reaches 63% of the world's population representing 5 billion people - an 11% increase from January 2022 – over 93% were social media users [1]. With increased digital consumption the world is creating massive amounts of data on a daily basis [1]. IDC (a technology research firm) estimates that data has been constantly growing at a 50%

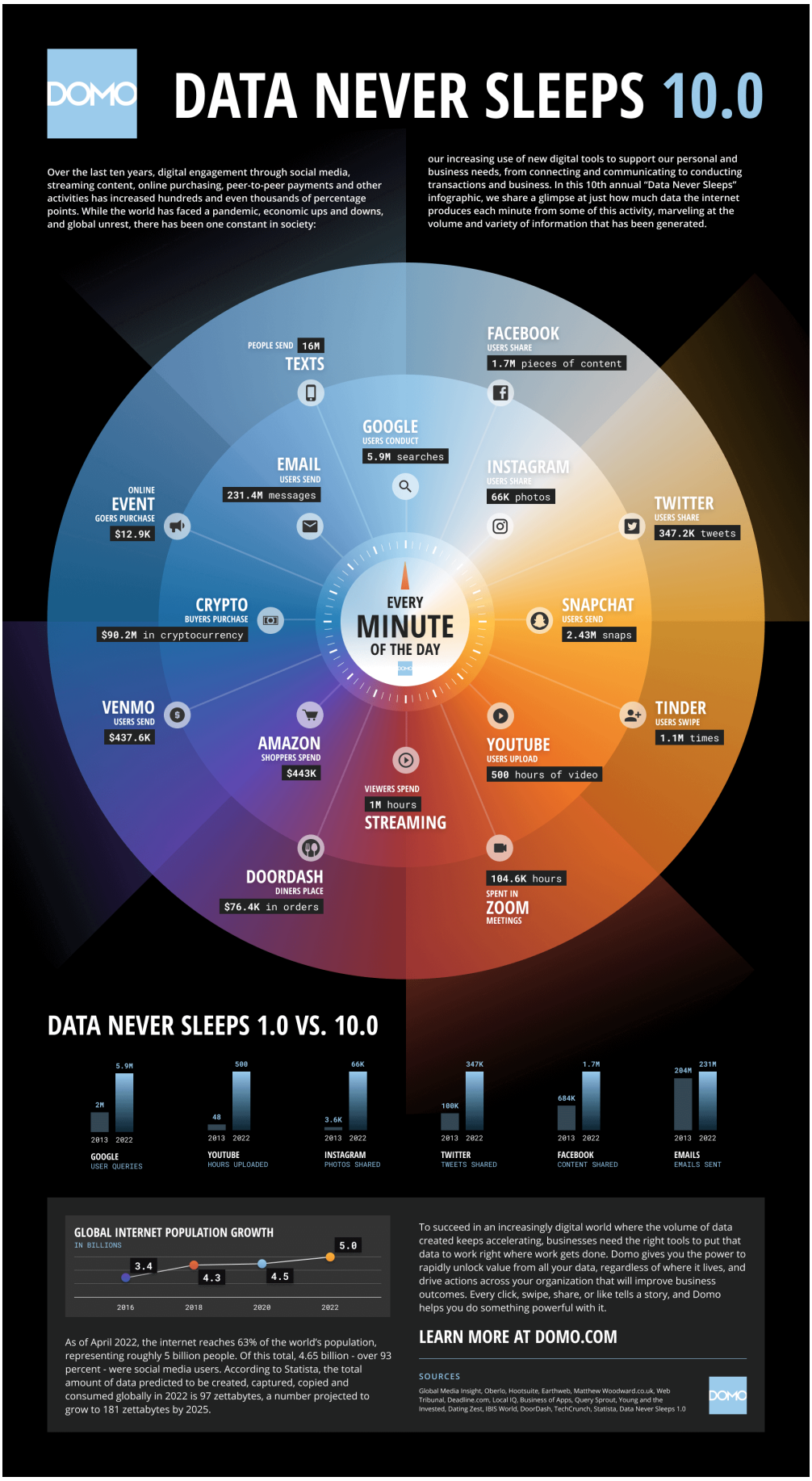

Fig. 2: Data Never Sleeps 10.0: data being created each minute on the internet in 2013 and 2022 [1].

increase each year, more than doubling every two years [2]. According to Statista, the total amount of data predicted to be created, captured, copied and consumed globally in 2022 is 97 zettabytes, a number projected to grow to 181 zettabytes by 2025 [1]. Digital ownership of high-fidelity virtual assets on Web 3.0 is increasing exponentially, which is an indication that the economic value of the digital virtual worlds using blockchain technologies and the metaverse will increase significantly in the years to come, promising much further economic growth [3], leading to the generation of very large volumes of BD. According to DOMO's Data Never Sleeps 6.0 report, more than half the world's web traffic comes from smartphones, and it's predicted that more than 6.1 billion people have access to a smartphone in 2020 [4]. We can safely conclude that more than 90% of the existing BD in the world has been generated in the last several years. For some companies, like Facebook, Twitter, and LinkedIn, nearly the entire value of the company lies in the analytic and predictive value of the voluminous data from their social networks; Meta was worth more than double General Motors and Ford combined even though they manufacture no products and sell no services in the traditional sense to their users [5]. Amazon and Pandora, use social network data as important components of predictive engines aimed at selling products and services [5] to targeted customers. The most recent BD created every minute by leading companies and everybody across several industries, including tech, media, retail, financial services, travel, and social media is presented in [1]. With such a large scale and variety of data, social network analysis becomes increasingly crucial for classifying end users, predicting buying interests, foretelling event occurrences [105]. As in smart domains, the data sharing between these companies is highly limited regarding privacy and security issues, insufficient rules and regulations, which extremely reduce the chance of unfolding and gaining numerous insights, and consequently, a tremendous amount of global economic wealth is unfortunately lost.

The main features of BD — 10 V's are 1) velocity (i.e., the speed at which tremendous amounts of data are being generated, collected and analysed), 2) volume (i.e., tremendous amounts of data are being generated each minute), 3) value (i.e., worth of the data), 4) variety (i.e., different types of data in different formats), 5) veracity (i.e., quality and trustworthiness of the data), 6) variability (i.e., a multitude of data dimensions regarding multiple data types and sources), 7) validity (i.e., accuracy and correctness regarding its intended use), 8) vulnerability (i.e., breaches and security concerns), 9) volatility (i.e., required intervals for retrieval of information) and 10) visualisation (i.e., presentation in reasonable response time).

Today, the world is connected more than ever and the "Business Intelligence (BI)" landscape enabling improved and optimised decisions and performance by leveraging analytics software to transform data into intelligence [6] is dominated by intelligent inferences acquired from BD using Business Analytics that is the practice and art of bringing quantitative data to bear on decision-making [5]. The concept of BD

has fundamentally changed the way organisations manage, analyse and leverage data in any industry [7]. Bigger data means more insights and better decision-making regarding its increasing power in representing real-world cases better. For instance, big healthcare data has considerable potential to improve patient outcomes, predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the cost of healthcare delivery and improve the QoL in general [7]. The next level of business analytics, now termed BI, refers to data visualisation and reporting for understanding "what happened and what is happening" [5] and most importantly "what is going to happen". Online retailers are rapidly adopting BD analytics solutions, particularly predictive analytics aiming to foresee the market and customer needs [8]. The era of BD has attracted much attention and accelerated the use of BD analytics and new advanced analytical tools and techniques (e.g. ChatGPT, Gemini) are on the rise to support innovation, promote productivity, improve efficiency, and manage intelligent autonomous traffic on the global network, cloud platforms, and smart domains from the perspective of transforming data into wiser formats (i.e., insights) — information, knowledge and wisdom. Data is the uninterpreted raw quantities, characters, or symbols collected, stored, and transmitted; "information" is a collection of interpreted, structured, or organised, meaningful and useful for certain applications; "knowledge" is acquaintance or familiarity about facts, truths, or principles gained through study or investigation; "wisdom" is sagacity, discernment, or insight to know what's true or right for making correct judgments, decisions, and actions [9]. Wisdom and knowledge extraction and information harvesting on distributed platforms are the main developing and promising subjects of doing business intelligently in today's business environment. However, there is chaos in the effective and efficient management of the globally distributed knowledge and wisdom regarding various cloud platforms, numerous smart domains, privacy and security concerns, and insufficient rules and regulations. As a result, it is unlikely to create a synergistic environment in which the globally created voluminous data and intelligence can be exploited effectively and efficiently benefiting the whole globe. Building up a global synergistic intelligent infrastructure requires the orchestration of resources in a new concept that can mitigate the chaos by generating, disseminating and acquiring desired insights from globally generated voluminous data within Automation of Everything (AoE) and Internet of Everything (IoE) using Advanced Insight Analytics (AIA). In this direction, this technical report analyses the concept of "Big Data" (BD), infrastructure of BD, BD analytics, and challenges in the processing of BD and it discusses practical solutions for these challenges.

II. Infrastructure of Big Data: Cloud and Edge/Fog/MEC Platforms

The cloud platform with vertically expandable data storage and processing capabilities has the advantages for massive storage, heavy-duty computation, global coordination, and wide-area connectivity,

while edge/fog, particularly Mobile-Edge Computing (MEC) [10] is useful for real-time processing, rapid innovation, user-centric service, edge resource pooling [11] and decentralisation of datacenter computation [12]. The main cloud service providers are IBM, Amazon EC2, Microsoft Azure, Fiware and Google providing an open public network connecting businesses, individuals, organisations, and governments all around the world under an umbrella with Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) using the pay-per-use model. Each cloud platform is being expanded with main distributed advanced data centres around the globe where each one contains tens of thousands, in some cases, hundreds of thousands of servers called supercomputers using high-performance computing systems and tens of thousands of high-bandwidth endpoints. Amazon EC2, Windows Azure and Google Cloud Engine currently provide services from 16, 36 and 11 geo-distributed regions, respectively [13]. The cloud is approaching the edge as the massive network in a wider infrastructure [8] along with the deployment of an increasing number of datacentres and multiple Virtual Machines (VMs) is being constructed by the leading providers [14], in particular, using the smaller versions of geo-distributed VM-based cloud platforms — cloudlets that are much closer to end-user applications, enabling reduced latency. Experiments show that the use of cloudlets decreases response time by 51% and reduces energy consumption by up to 42% in a mobile device compared to cloud [15]. By leveraging low-latency offload, cloudlets enable a new class of real-time cognitive assistive applications on mobile cloud convergence [16] by massive virtualisation with VMs managed by using hypervisors servers [15]. The start-up costs of enterprises, particularly emerging ones, on the cloud are quite small for starting a rapid worldwide business from scratch, which makes this platform very appealing along with the advanced abilities of 1) automatic detection of compromised accounts, malware, data breaches, and malicious insiders [17], and 2) reducing the risk of data loss using redundancy [17], archiving, VM backups at multiple locations.

Cloud networking uses 1) Software-Defined Network (SDN) [18] enabling the configuration of the datacenter at a high level with much less human intervention and allocation of VMs, virtual networking, and virtual storage to a new tenant with specified service level guarantees, and 2) specialised cloud operating systems (e.g., OpenFlow, OpenStack) to manage server, storage, and networking resources to support multiple applications from third parties. Edge (or fog) or most recent popular platform, so-called MEC is an emergent architecture for computing, storage, control, and networking that distributes these services closer to end-users [11] to enable a more independent processing and organisation, particularly for applications requiring real-time decision making, low-latency, ultra-low-latency, high privacy, and security. The segmentation of what tasks go to the fog/edge and what tasks go to the back-end cloud is application-specific and can change dynamically based upon the state of the network, including processor loads, link bandwidths, storage capacities, fault events, and security threats [19]. One of the primary

problems in cloud computing today is how to manage sensitive workloads running on the cloud [20]. Service providers are trying to assure their customers about this issue using various approaches and policies involving their users with Shared Responsibility Model (SRM) generally promoted by Amazon Web Services (AWS) and Elastic Compute Cloud (EC2) on the cloud and edge/fog platforms.

The number of connected things will exceed 7 trillion by 2025 which makes 1000 devices per person and an estimated value of 36 trillion of dollars [21]. The emergence of Intenet of Things (IoT) has led to increasing data volumes [22]. IoT with resource constraint characteristics is composed of physical objects embedded with electronics, software, and sensors, which allows objects to be sensed and controlled remotely across the existing network infrastructure, facilitates direct integration between the physical world and computer communication networks, and significantly contributes to enhanced efficiency, accuracy, and economic benefits [23]. With IoT, physical objects are seamlessly integrated globally so that the physical objects can interact with each other and to cyber-agents in order to achieve mission-critical objectives [24]. IoT envisions a future in which digital and physical entities can be linked through appropriate information and communication technologies.

The emerging smart domain structure, rising over the above-mentioned infrastructure, enables the classification of data regarding their subject fields and eases the management, processing and exploitation of BD in these particular domains. Some of the main smart domains are smart city, smart home, smart building, smart transportation, smart health, smart industry, smart factory, smart shopping and manufacturing, smart logistics and retail, smart energy and smart grid, and smart agriculture. The connected IoT devices or mechatronics devices in these smart domains not only talk to each other within their smart domains, but also they can talk to other devices in the other smart domains as well, e.g., security, fire or gas alarm using intelligent sensors in the smart home domain may trigger an action for police or fire department in the smart city domain [8]. There are no strict boundaries between these smart domains; some of the outputs of the smart devices may input for other smart devices within both their domains and other domains in which the smart city is in the centre to indicate people-focused cyber-physical understanding, since more than 60% of the population will be living in an urban environment by 2030 [25] and the global population is expected to double by 2050 [26]. Cities with heavy populations escalate burden on energy, water, buildings, public places, transportation and many other things [25]. Interested readers are referred to the study [17] for more detailed information about the transformation of the cities into smart cities and for more information about the smart domains with many examples. Recent advances in Cyber-Physical Domains (CPD) and smart vision-based platforms [27], [28], [29], [30], [31], [32], [33] which physical objects around us become an irrevocable part of the global information system and BD have created the chance to develop an open architecture with effective sharing and intelligent services [2].

The proliferation of devices with communicating-actuating capabilities is bringing closer the vision of an IoT, where the sensing and actuation functions seamlessly blend into the background and new capabilities are made possible through the access of rich new information sources [34] within smart domains. A voluminous data is created in these domains. Still, there are serious restrictions in sharing of data between these domains regarding privacy and security concerns where data can be reached from anywhere anytime with worldwide distributed computing environments, and sharing of data between these domains within an effective and efficient infrastructure addressing cyberattacks is required, which would make our life functional and easier in many aspects. In this regard, one of the recent prominent trends is to integrate all smart domains in a combined architecture of the cloud platform [35] and a revolutionary networking model called Information-Centric Networking has recently attracted the attention of the research community working on data dissemination in various smart domains [36]. Similar data-sharing attempts between the smart domains will increase in the following years where privacy and security concerns are addressed well. For instance, blockchain technologies are promising to provide users with decentralised security and privacy-preserving capabilities while using and sharing their data with ML applications [37].

The need to develop Digital Twins (DTs) is increasing the amount of data significantly [38], [39], [40], [41], [42], [43], [44], [45]. DTs was first verbalised by Michael Grieves in 2002 during a conference at Michigan University. The industrial concept of DTs attributed to NASA (2012) was constructed to integrate ultra-high fidelity simulation with the vehicle's on-board integrated vehicle health management system, maintenance history and all available historical and fleet data to mirror the life of its flying twin and enable unprecedented levels of safety and reliability [46]. DTs, with the main components i) a physical entity, ii) a virtual representation entity of that entity, and iii) bidirectional near-real-time data streaming between these two entities, have changed the way of procedures in each discipline. The once-in-a-lifetime pandemic has forced many to rely on digital technology as the only way to communicate, collaborate, learn and sustain our lives [47]. In the same manner, with decreasing contact with the real physical world and people, the recent pandemic has confirmed the importance of location and time-independent DTs (i.e., digital replicas) of cities and their enabled remote services that can provide everybody with equity and accessibility, in particular, senior citizens and disabled residents by democratising all types of services leading to increased QoL with societal flourishing. The pressure is growing on city governments to leverage every opportunity to improve QoL for inhabitants [48] due to the ever-growing demands of citizens, economic concerns, and imminent environmental risks. In this manner, to alleviate the problems of rapid urbanisation and improve the liveability of citizens, there are many concerns to be taken into account in urban development and efficient urban management with effective public services. In a broader

inclusive definition, SC can be defined as an opportunistic concept that enhances harmony between the lives and the environment around those lives perpetually in a city by harnessing smart technology enabling a comfortable and convenient living ecosystem paving the way towards smarter countries and a smarter planet [17]. SCs are being implemented to combine governors, organisations, institutions, citizens, environment, and emerging technologies in a highly synergistic synchronised ecosystem to increase QoL and enable a more sustainable future for urban life with increasing natural resource constraints [17].

Considering another emerging infrastructure, the metaverse [17], as a potential to produce voluminous BD, can be defined as "democratised, decentralised, user-driven virtual and augmented immersive 3D spaces where two worlds — virtual and physical existence — can be more tangibly connected and people who are not in the same physical space can come together with their avatars to feel many different types of experiences". The metaverse — a blended harmonised virtual and physical existence — aiming at developing high-fidelity virtual worlds as rich as the real world, is still in its infancy, requiring a great deal of evolution. First and foremost, open-source metaverse development platforms are in high demand and they are expected to expedite the development of many undiscovered metaverse experiences. Truly persistent and immersive computing, at scale and accessible by billions of humans in real time, will require a 1000 times increase in computational efficiency from todays state of the art [47] where a standard kind of Moore's law curve is only going to get us to about 10 times growth over the next five years based on hardware improvement and high computation requirement can be alleviated by energy-efficient computing, better algorithms, better architectures as stated by Intel to balance 1000 times improvement [49]; an example of which is Ethereum's "transition to a mechanism known as proof of stake" to reduce their energy use around 99% with energy-efficient computing and processing. Limited by resources, computing power with edge intelligence, and sensory devices, the metaverse is still far from realising its vision of thorough immersion, materialisation, and interoperability [50].

III. BD Analytics

Data analytic tools (e.g. ChatGPT, Gemini) are evolving rapidly to harvest BD volumes. Deriving meaningful insights from voluminous geo-distributed data of all kinds as a strategic asset is fuelling innovation, facilitating e-commerce and revolutionising the industry and businesses in the transition from digital to the intelligent way of doing business. The era of BD has attracted much attention and accelerated the use of BD analytics and new advanced analytical tools and techniques are on the rise to support innovation, promote productivity, improve efficiency, and manage the intelligent autonomous traffic on the global network, cloud platforms, and smart domains. Analytics is the science of using data to build models that lead to better decisions that in turn add value to individuals, companies, and institutions [51].

Cloud analytics enables businesses to carry out analytics through an integration of hosted data warehouses, BI, and other analytics [6]. Most recent data analytic tools designed to work on the cloud platforms are analysed in [52]. The orchestration of resources and network traffic across geo-distributed nodes are provided using specialised interfaces such as OpenStack and OpenFlow enabling SDN controller to manage distributed nodes effectively and efficiently. The evolution of SDN allows for a logically centralised but physically distributed control plane by eliminating vendor dependency and compatibility issues between different networking devices [53]. More and more sensors and devices [54], [55], [56] are being interconnected via IoT techniques, and these sensors and devices generate massive data and demand further processing, providing intelligence to both service providers and users [57].

The leading cloud service providers employ various BD analytics infrastructures along with their software to process large volumes of data such as Big Query as Database-as-a-Service (DaaS) by Google, No-SQL, BigInsights by IBM, Apache Hadoop, Apache Spark enabling interactive SQL on Hadoop, Hive built based on the Hadoop Distributed File System (HDFS) with a master-slave (i.e, JobTracker-TaskTrackers) architecture enabling ad hoc query processing, and Elastic MapReduce by Amazon. Regarding the BD programming language tools, Pig is used by Yahoo; Hive with SQL-like language called HiveQL is used by Facebook; Jagl with the ability of processing structured and non-traditional data is used by IBM. The components of the most commonly used BD analytics infrastructure, Hadoop in a broader perspective are 1) high-level languages (i.e., Pig (execution framework), Cascading, Hive (data warehouse)), 2) execution engine (i.e., MapReduce) where the map process distributes the processes independently from each other in parallelisation and the reduce process combines the results obtained from the independent map tasks, 3) distributed light-weight database (i.e., HBase), 4) distributed file system (i.e., HDFS) and 5) centralised tool for coordination (i.e., Zookeeper). Hadoop has no support for hyperscale geo-distributed data processing [58] even though it has a Java-based software framework for distributed processing of large datasets across large clusters of computers/racks with limitless concurrent tasks. Hadoop has no support for hyperscale geo-distributed data processing [58] even though it has a Java-based software framework for distributed processing of large datasets across large clusters of computers/racks with limitless concurrent tasks. Therefore, BD is collected from geo-distributed data centres and locally stored in HDFS to be processed by MapReduce software computing, which makes it highly difficult to gain timely insights, more importantly it gets more difficult to store BD locally as the data volume grows exponentially regarding the restricted computing abilities, hardware limitations such as processing power, network bandwidth, storage limitations, and high-latency. The bandwidth availability between different data centres significantly varies over time which usually is the bottleneck of such an evaluation [59], particularly for low-latency requirements (e.g., intelligent transportation systems). Raw

BD stored in no-SQL database are messily scattered and can't be used directly for two reasons [26]: first of all, for most of them, data cannot be interpreted by the model itself, and additional features have to be handled in the application logic [26]. Secondly, the overwhelming majority of BD are few of value while only a drop in the bucket is valuable [60]. Unlike Hadoop using a new file system HDFS, Dryad is based on the daily-used and mature New Technology File System, which is much easier to use for developers [61]. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provide performance up to 100 times faster for certain applications with interactive data exploration [58]. Some of the other BD analytics projects established to solve the different problems are Ambari (cluster management), Avro (data serialisation), Cassandra (multi-master database), Chukwa (data collection), Hbase (distributed database), MaHout (Machine Learning (ML) and Data Mining (DM)), Tez (data-flow programming framework intended to replace MapReduce), Cloudera Hortonworks, Microsoft (HDInsight), MapR, Map Altiscale, Factor in Apache storm (stream processing) and Kafta.

Regarding the approaches mentioned above, the dominant approach is to collect all the data across the world to a central datacenter location and to process using data-parallel applications or data analytics, which is neither efficient nor practical as the volume of data grows exponentially [62], [63]. Several particular approaches in various studies have recently been presented in order to process exponentially increasing very big geo-distributed data efficiently and effectively by pinpointing different aspects of the problem space, in particular, privacy and resource availability — e.g., storage, computation and networking. Jimenez-Peris *et al.* [64] proposed a new transactional highly distributed cloud platform, CumuloNimbo in which each component can be scaled independently. Xiang *et al.* [65] designed a framework titled Unicorn for multi-domain, geo-distributed data analytics to manage a large set of distributively owned heterogeneous resources, with the fundamental objective of efficient resource utilisation, following the autonomy and privacy of resource owners by providing accurate resource availability information. Zhou *et al.* [13] introduced the special privacy requirements in geo-distributed data centres and formulate the geo-distributed process mapping problem as an optimisation problem with multiple constraints. Hu *et al.* [62] proposed an approach titled Flutter in which a new task scheduling algorithm aims to reduce both the completion times and the network costs of BD processing jobs across geographically distributed data centres by focusing on how tasks may be scheduled closer to the data across geo-distributed data centres. Zhao *et al.* [63] studied how to optimise geo-distributed data analytics with coordinated task scheduling and routing. Luo *et al.* [66] proposed a deep learning-based framework for scaling of the geo-distributed virtual network function chains, exploring the inherent pattern of traffic variation and good deployment strategies over time.

Sending the entire datasets across the extreme ends is unrealistic and the approaches that collect data and perform computational processing near the data source present a more practical and realistic alternative [67] due to large volumes of data leading to bandwidth limitations. Analysing data at the early stages of infrastructure pipeline presents additional benefits of data and communication security in the overall system, owing to the fact that raw data is now processed closer to the data source, and only processed data is sent further [67]. More explicitly, data analytics can be used on fog/edge platforms for processing large volumes of multi-modal and heterogeneous data from various sensor devices and other Internet of Things (IoT) devices to achieve real-time and fast processing for decision making [68] where the processing power with increasing storage units is getting bigger locally. In other words, local advanced Hybrid Cloud-Edge Analytics (HCEA) should enable performing local analytics, identifying usable information from raw data, extracting insights in abstract forms and finally transmitting the result to the cloud platform for the use of any global entities and/or for the further exploitation of other AIA on the cloud platform. In this direction, Federated Learning (FL), introduced by Google, has gained prominence as an effective solution for addressing data silos, enabling collaboration among multiple parties without sharing their data [69]. In FL, each entity trains its own data locally, and only the locally generated model itself is sent to the central server to aggregate all the models to form the final model for each entity to utilise. Collaborative Learning (CL) and FL have been used interchangeability in the literature to train global models using swarm AI. The concepts and applications of FL is analysed in [70]. With the increasing need for collaborative work, as well as the increasing concern in data privacy, the Collaborative Deep Learning (CDL) has become much more common [71] regarding its successful application with Deep Neural Network (DNN) models established on Big Data (BD) with some of these instances of success being reached on imperfect conditions. CDL models enable parties to locally train their deep learning structures and only share a subset of the parameters in the attempt to keep their respective training sets private [72]. The CDL framework allows local devices to cooperate on training models without sharing private data, which resolves .the contradiction of the availability and privacy of data [73]. From a technical standpoint, DL can be performed in a collaborative manner, where a parameter server is required to maintain the latest parameters available to all parties [74]. Data, particularly BD, is distributed among multiple entities due to changing distributed architecture (e.g. cloud, metaverse), its strategic value, data privacy and security, which necessitates CL – with distributed multiple entities. In CL, a learning model is constructed using multiple distributed data points, possibly by exploiting whole data, either belonging to a single user, multiple users, a single platform, or multiple platforms to extract common features or patterns by preserving data privacy. Although local data is not directly shared with FL, models trained on this data may also be spied on by malicious adversaries, semi-honest

parties, or honest but curious parties, when local models are aggregated into a centre. Moreover, under the circumstance of knowing the local model, spies may adopt some attacks to restore the original data, which indirectly leads to information leakage [75].

PPML or more specifically, Privacy-Preserving Deep Learning (PPDL) schemes have been developed and employed to further preserve sensible data and privacy while performing FL/CL. Multiple distributed encrypted data points can be uploaded by their owners to a central server, collected by the platform, or processed data models using specific agreed-upon transparent DL training models, which are then later aggregated to establish the global model without sharing the data itself. The Homomorphic Encryption (HE) scheme allows data to be processed without needing to decrypt it. HE enables multiple entities to perform complex queries and computations on encrypted data without compromising the privacy of data and its encryption. The processed result still may remain in encrypted form for the owner of the data to decrypt it using the private key for visualisation. Concretely speaking, sensitive data can be shared and computed without the need to decrypt, but with a large computational overhead. The ciphertext operation's computational complexity is much higher than that of the plaintext operation's, both in terms of memory consumption and processing time [76]. There are three types of HE, namely: partially HE, somewhat HE, and fully HE. Fully HE produces the largest computational overhead compared to the other two HEs, while having infinite addition and multiplication operations on ciphertexts. Fully HE is being employed by many giant companies such as Microsoft to compute sensitive data in the public domain despite its computational overhead and complexity. Most importantly, it allows to training of homomorphic-based encryption structures to build larger learning models, namely CDL models, using SAI. HEs goal is to prevent recovery of the original data in order to protect the data from unauthorised access and users privacy. ML-as-a-Service (MLaaS) techniques using the processing of encrypted data with HE-like approaches will be focused on in the future, particularly, for applications which need a high level of privacy-preserving requirements on data that is stored in public domains and need to be computed by multiple entities. Another privacy preservation technique, which has captured a wide range of attention, is differential privacy which was developed in [77], by which noise is added to the data to secure the data from attacks. However, the more noise added to the data to provide further security and privacy, the less the model accuracy is obtained. SEALion, CryptoNet [78], and CryptoDL are the early implementation examples (trained networks) of the PPDL scheme via encrypted outputs using HE. A PPDL system in which many learning participants perform NN-based DL over a combined dataset of all, without revealing the participants' local data to a central server is presented in [79] using asynchronous stochastic gradient descent, in combination with HE. An FL-enabled network data analytics function architecture with partial HE to secure ML model sharing with privacy-preserving mechanisms is proposed

in [80]. A full HE scheme to the standard DNN, ResNet-20, is applied in [81] to implement PPML. A universal multi-modal vertical FL framework is proposed in [69] to effectively acquire cross-domain semantic features on homomorphic-encrypted data. FL mechanism is introduced into the deep learning of medical models in Internet of Things (IoT)-based healthcare system in [75] in which cryptographic primitives, including masks and HE, are applied for further protecting local models, so as to prevent the adversary from inferring private medical data by various attacks such as model reconstruction attack or model inversion attack or model inference attacks.

IV. Challenges and Practical Solutions

The emerging BD pose new challenges, some of which cannot be adequately addressed by existing infrastructure, data analytics, state-of-the-art cyber-security solutions, cloud and host computing models alone, which exacerbates the current situation with respect to exponentially increasing voluminous BD. The major challenges are summarised as follows.

A. Sanitisation of BD and Cybersecurity Risks

Sanitisation, e.g., anonymisation/de-identification, and cybersecurity measures to prevent breaches of sensitive information allow the sharing of data for secondary purposes, such as research, the establishment of decision-making tools, extraction of other meaningful information that can be an input to other systems. Sharing of data should protect individual privacy via sanitisation, but still ensures that the data is of sufficient quality that the analytics are useful and meaningful [82]. Large amounts of data stored on the cloud are very sensitive, and so data privacy remains one of the top concerns for many reasons; mainly those relating to legal or competition issues [83]. In a Gallup poll, 27% of respondents said they or someone within their household had credit card information stolen [84]. 1 billion user accounts in Yahoo were compromised in 2013 [85]; hackers attacked on Apple's iCloud platform that resulted in the release of the personal photographs of many high profile figures in 2014 [86]. LinkedIn, Sony, Oracle, T-Mobile, Dropbox and many others were also attacked similarly by hackers. In 2020, more than 70% of enterprises continuously monitors for sensitive data incidents [87]. In this sense, the BD analytics technologies are instrumental for organisations to improve their capabilities in discovering potential threats, detecting actual threats, gathering intelligence about attacks, and deploying a comprehensive response to minimise the business impacts of cyberattacks [88] such as theft of credit card data and trade secrets. In this perspective, privacy-preserving data analysis is an emerging discipline within data science, which posts several challenges currently being simultaneously tackled from several areas such as hardware, systems security, cryptography, statistics, and ML [89]. Several privacy-enhancing techniques evolved in the last

decade have different trade-offs, maturity levels, and privacy guarantees, and in some cases solve slightly different problems [89]. In 2020, large global enterprise use of data masking or similar pseudonymisation techniques increases to 40%, from 10% in 2016 [87].

There have been various cases in which the personal identities of the owners of the sensitive data have been unveiled on the sanitised datasets placed in the public domain for several reasons such as research [90]. For instance, when the state of Massachusetts released sanitised medical records summarising every state employee's hospital record in the mid-1990s, the Governor gave a public assurance that it had been anonymised by removing all identifying information such as name, address, and social security number and he was surprised to receive his health records (which included diagnoses and prescriptions) in the mail [90]. 50% of the Americans can be identified from city, birth date, and sex; 85% can be identified if you include the zip code as well [90]. You will probably be left with nothing useful if you do remove all possible identification information from a database [90]. Sanitisation of the data is not an easy process and it requires data scientists expertised particularly in sanitisation to address the privacy and security concerns to mitigate the possible risks. Security is "confidentiality, integrity and availability" of data whereas privacy is the appropriate use of user's information [7]. A study by Skyhigh Networks, a cybersecurity firm, found that 18.1% of all documents uploaded to cloud-linked systems contain sensitive data [91]. Zhang *et al.* demonstrated that 20% of the big image data was found sensitive and maintained on the edge platform whereas 80% was found non-sensitive and encrypted, then, subsampled and stored on the cloud platform [92]. Another research estimates that 90% of the data generated by the endpoints will be stored and processed locally rather than processed on the cloud [93] where sending all the data to the cloud requires prohibitively high network bandwidth [11]. New studies that aim to sanitise data at its source before sending to cloud platforms are emerging such as [94] in which a privacy-preserving smart home system connects a single home controller with data-hiding capabilities through community networking and integrates the data to a hierarchical architecture on a cloud platform for a data analytical access control mechanism. Sanitisation starts on the edge platform. However, quite an important amount of raw sensitive data is placed on the cloud platforms even though most of the sensitive data can be processed, maintained or deleted on edge platforms. In this respect, sensitive data, in particular, private data on the cloud platforms such as sensitive personal data, medical data, credit card information and transactions should be managed carefully regarding the privacy and security aspects. What happens if a smartphone operating as an edge platform is hacked by a cyber attacker; cameras that are meant for surveillance may turn into cameras that are violating our privacy [8]. Anonymisation is more than simply masking or generalising certain fields – anonymised datasets need to be carefully analysed to determine whether they are vulnerable to attacks [7].

Cryptographic schemes and practical systems analysed in a recent study by Moghadam *et al.* [83] which enable the execution of queries over encrypted data (e.g., homomorphic encryption, property-preserving encryption) without decryption using analytics are both non-trivial and costly in terms of analytic processing difficulties. Hence, new sanitisation approaches are on the rise to protect the privacy and security in addition to conventional most commonly used sanitisation techniques such as k-anonymity [95], privacy [96] and l-diversity [97]. Various new sanitisation approaches specific to BD on the cloud and edge platforms have recently been extensively introduced in order to both mitigate the shortcomings of existing ones and process specific types of BD effectively and efficiently such as 1) LinkMirage to address the link privacy in the social media data [98], 2) HCMPSO in an IoT Environment [99], 3) data-sanitisation for preventing sensitive information in social networks [100], 4) automatic unsupervised general-purpose sanitisation of textual documents by detecting and hiding sensitive textual information while preserving its meaning [101], 5) collaborative search log sanitisation toward differential privacy and boosted utility [102], 6) ant colony system sanitisation approach to hiding sensitive item sets — ACS2DT [103], in order to hide sensitive and critical information by decreasing sanitisation side effects [103], and enhancing the performance of the sanitisation process [103], and 7) individual trajectory data sanitisation — Lclean, using a plausible replacement method [104]. In addition to sanitisation, laws, and regulations should be amended to ensure the privacy and security wherever breaches emerge, which is not within the scope of this study and not explored in this research in detail. Cryptographic schemes and practical analytics which enable the execution of queries over encrypted data without decryption will be immensely focused both in order to mitigate the security and privacy concerns and in order to reduce the computation overhead caused by the encryption in the following years.

To summarise, the need for robust privacy-preserving data analysis technologies has been recognised by both regulators and industry [89]. New effective privacy-protective techniques and mechanisms are needed to retain privacy when analysing users' data and these mechanisms should be active in AIA. The moral is that if you do remove all possible identification information from a database, you will probably be left with nothing useful [90]. It takes a long time to sanitise BD and BD may lose its meaning with the side effects of the sanitisation process. A fully-fledged approach to privacy-preserving data analysis would still require significant interdisciplinary effort, some of which have to do with issues such as effective personal data management and consent [89]. Processing of BD at the data centre and transportation of insights free-from private information helps provide reliability, security and privacy.

B. Compromise of BD

The data stored on the cloud platform might be compromised. The studies conducted by ABCNews and Boston Globe show that it is achievable to infer the sexual orientation of a user through mining a Facebook subnetwork involving the user's friendship relations, gender, and other attributes [105]. There are many studies related to data security and privacy issues on the cloud platform to alleviate these similar concerns. Encryption was found to be the most widely applied technique [106], [107] for protecting highly sensitive data such as passwords, physical locations, sexual orientation, names, ID numbers, images, personal files, bank transactions from unauthorised access by all entities, including service providers, which in turn makes the third parties not able to reach and analyse most of the data on the cloud platforms. In the metaverse worlds, owners of most of the generated data will be their users who have the data with data sovereignty, and no central authority is required for transactions. Blockchain technologies, enabling individual data ownership, are already being used by many applications to store, share and process the information that is under the control of their individual owners using Web 3.0 and Web3. New effective approaches need to be established to open this BD to everybody in order to unveil its potential and economic value without compromising privacy and security concerns. One way is to incentivise owners of data (e.g. financially) to share their experiences without compromising their privacy and security.

C. Unstructured formats of BD

Most of the BD is unstructured (e.g., text, speech, video) around 85-90%, which makes the analysis and interpretation so difficult in gaining insights even though there are promising attempts to develop new tools (e.g., text mining, web mining, image mining, Social Network Analysis (SNA)) that can analyse unstructured BD. Large volumes of BD being generated exponentially in different formats are in the geo-distributed cloud platforms and likely input for all other smart systems and enterprises as insights, which will contribute to the smooth working of these systems and enterprises substantially. As the amount of BD being processed on data centres in multiple cloud platforms increases, the network resource consumption also increases and BD management across multiple data centres in multiple cloud platforms turns out to be an important and challenging task [108]. Streaming of this exponentially increasing voluminous data at once may choke the underlying current network infrastructure [108]. Recognising the growing demand for ways to handle geo-distributed data cost-effectively, researchers have begun to focus on how to efficiently analyse geo-distributed datasets [58]. However, many solutions address only how to store data across data centres and few efforts have investigated how to effectively compute with it [58]. In cloud platforms, 1) most of the time BD can not be reached because of privacy and security issues and effective tools are being deployed to detect and respond faster to cyber threats, attacks, and breaches of data, 2)

BD should be sanitised before published in the public domains to be explored and exploited for further analysis — not to mention that sanitised data still carry high risks of leaking sensitive information, and 3) processing a substantial amount of data within a very small time interval is a great challenge for low-latency cloud applications [109] where analysis of BD through the geo-distributed data centres incur huge communication cost, particularly where BD is needed to be collected from geo-distributed data centres and stored locally to be processed using the current processing techniques such as Apache Hadoop. Therefore, new approaches needed first to reduce privacy and security risks to a minimum, and second to make the most out of BD regarding extracting thorough and up-to-date insights in a timely manner. Several approaches have been proposed to find solutions to "no quality public domain to establish a quality decision-making platform for a specific field" while "we are drowning in data, but starving for knowledge".

D. Constraints with Network Resources and Analytics

1) Resource-constrained devices: From the point of mobile computing and IoT, the devices' limited computational capacity and limited battery life span are major challenges [110] in the establishment of smart applications. The highly geo-distributed IoT devices in smart city and smart transportation applications are examples for these types of limitations. Traditional cloud computing architectures were simply not designed with an IoT, characterised by extreme geographic distribution, heterogeneity and dynamism in mind and a novel approach is required to meet the requirements of IoT including transversal requirements (scalability, interoperability, flexibility, reliability, efficiency, availability, and security) as well as Cloud-to-Thing (C2T)-specific computation, storage and communication needs [111]. Therefore, the economical use of the scarce resources is vital and by offloading resource-intensive compute tasks to more powerful nodes — such as servers in a datacenter or compute resources at the network edge — the range of possible applications can be widened significantly [112]. In this sense, most of the resource-hungry computing should be carried out in cloud centres to mitigate these limitations.

2) Insufficient tools for harvesting insights: Robust AIA tools that can acquire insights are needed along with efficient resource orchestration, insight and content distribution enabling audits to ensure compliance. These tools should embrace the text mining, web mining, speech mining, image/video mining abilities and should be able to tackle the unstructured BD very well since most of the BD is in unstructured format, enabling discovering intrinsic relationships [117], text clustering and categorisation, speech categorisation, concept/entity extraction, production of granular taxonomies, document summarisation, sentiment analysis, and entity relation modelling.

3) Lack of hybrid cloud-edge analytics at the edge/fog/MEC platform: Data is subject to attacks and security breaches when stored on the cloud platform. Effective AIA at the edges producing insights to be submitted to the cloud platform along with sanitised and filtered data rather than unprocessed raw data is extremely crucial to reduce the workload burden on the cloud and network and to improve the overall efficacy of the network architecture.

4) Networking, bandwidth and load balancing challenges: the orchestration and reallocation of computing and networking resources between edge and cloud platforms can be managed effectively with the rapid migration of insights rather than transforming BD without causing any overhead, which will both reduce workload and facilitate distributed computing across edge/fog/MEC nodes with limited resource-constrained environments.

5) Latency challenge: Processing time of several ready-to-use up-to-date insights is far more less than processing BD and the application latency can be reduced significantly, particularly for low-latency applications.

E. Diverse data centres

The management of cloud data centres scattered around the globe supported by cloudlets and large network webs is extremely expensive. The main cloud service providers invested an amount of $383 billion into their cloud infrastructures in 2020 [134]. In this regard, it is envisaged that leading competitive cloud service providers such as Google, Amazon, Microsoft and IBM will need to merge their powers under bigger joint ventures not only to reduce the current immense management costs and future investment costs substantially, but also to increase the efficacy of their services in a fruitful and prosperous way of exploiting many more data centres and much larger global networks, particularly, by maximising the resource utilisation while mitigating the performance degradation. The aforementioned joint ventures can expedite data sharing abilities resulting in both increased productivity through the booms of insights gained from gigantic data, and effective way of doing business by intertwining smart platforms and businesses, particularly conducted by leading companies, which will definitely make our life cheaper, easier, and more intelligent.

F. Lack of Regulatory Framework

1) Concrete agreement difficulties among cloud service providers: Agreement between parties on many issues is not easy and this process will effect how regulation and legislation will evolve.

2) Integration of smart domains: Main actors as decision-makers in smart domains avoid sharing their data due to serious security and privacy concerns, which reduces the further integration of the smart

domains even though the number of successful examples and attempts of combining these domains in open data sharing platforms [1] is increasing.

3) Reluctance in data sharing: Data sharing between prominent leading companies that generate BD is highly limited regarding privacy and security concerns, insufficient rules and regulations, and competitive and commercial risks. Additionally, data owners do not volunteer to share their data particularly due to the breaches of their privacy and security. Effective incentive mechanisms along with effective privacy and security preserving tools should be developed and employed to encourage data sharing.

4) National/international rules and regulations: The legal challenges, in particular, about the privacy of people using CPD have yet to be solved [113]. The responsibilities of cloud service providers against the customers are not well defined within the national and international laws of electronic commerce. Moving data from a country with more restricted data privacy regulations to a less restricted country can violate the data privacy requirements and is hence prohibited [13]. Due to the multi-level and asymmetric data privacy requirements, there can be many constraints on data movement between different data centres, which complicates the geo-distributed process mapping problem [13].

5) Management of communication regarding emerging future networks: The studies on the management of large volumes of BD traffic regarding emerging communication technologies (e.g., 5GB) are not adequate to deploy these technologies effectively. A recent paper [114] discusses how to manage the E2E traffic with recently emerging communication technologies effectively. More experimental studies and more real-world examples are needed in this field to be able to deploy these communication technologies in an efficient and effective fashion regarding very BD.

6) Energy consumption and optimisation concerns: Data centres cost hundreds of millions per year and consume more than 61 billion kilowatt-hours, one-fifth percent of the country's entire energy consumption [26]. Amazon commits to 100% renewable energy on its cloud platforms and Apple, Facebook and Google, three of Amazon's peers and rivals, all have laid out road maps that explain how they intend to achieve their goals of procuring 100% renewable energy on their platforms [115]. Advanced optimisation and data analytics tools are needed to be built to increase the efficacy of these platforms for environmental benefits and cost reduction; an attempt for which iSpot is developed to be able to guarantee the performance of BD analytics running on cloud transient servers while reducing the job budget by up to 83.8% in comparison to the state-of-the-art server provisioning strategies, yet with acceptable runtime overhead [116]. Another attempt to run the BD analytics in geo-distributed data centres effectively can be found in the study [59]. Furthermore, software-defined-network-based optimisation of BD management

[1]Interested readers are referred to the study [17] for real-world implementations of open data hubs (Table 1).

across multi-cloud data centres was analysed in [108].

7) Lack of standards: Agreed-upon standards and protocols for effective policies and mechanisms are necessary. The General Data Protection Regulation (GDPR) and the ongoing e-privacy regulation effort are significant steps in regulating the protection of sensitive information by placing obligations on data controllers and data processors, as well as specifying user's rights. However, no specific algorithms are mentioned, and hence we are far from effective standardisation guidelines [89]. The complexity of secure data analysis requires several kinds of standards, related not only to the different aspects of privacy-preserving analytics, but also related issues like personal data management and consent [89].

V. Discussion and Conclusion

Cheap ubiquitous computing enables the collection of massive amounts of personal data in a wide variety of domains [118]. The extreme explosion of BD imposes a heavy burden on computation, storage, and communication resources in today's infrastructure [61]. Cloud, with sufficient resources in large-scale data centres, is widely regarded as an ideal platform for BD processing and how to explore these resources has become the first concern in BD [61]. The management and processing of voluminous BD is still in its infancy. What makes the processing of BD more difficult on the cloud platforms is that most of the time BD can not be reached because of the privacy and security concerns and effective tools are being deployed to detect and respond faster to cyberthreats and breaches of data. One of the recent popular trends in order to make BD more valuable is to integrate all smart domains in a combined architecture of the cloud platform [119] to create bigger synergies, which is not a trivial task and involves many challenges.

Computational resources that are typically concentrated in cloud data centres are now approaching users and edge resources based on user device requests using cloudlets and VMs equipped with advanced communication technologies and AIA. This paves the way for both federating geographically distributed cloud and fog/edge resources and the effective management of geo-distributed insights leading to decreased latency, which enables the effective use of low-latency applications in addition to supporting delay-tolerant applications. The increasing interdependencies between people, advanced intelligent devices, and the infrastructure requires the orchestration, harmonisation and optimised use of these resources. The exponential increase in the volume of data being created and analysed has triggered interest in a new interdisciplinary form of scientific inquiry referred to as "data science" and "data analytics" [120] which has given rise to a new profession: the data scientist with analytics skills [5] in order to make the tremendous potential explicit, i.e., secrecy of life, in BD and exploit it as a profitable business. Data science is a mix of skills in the areas of statistics, ML, math, programming, business,

and IT [5]. AI/ML applications have proven successfully in many different fields with successful applications [121], [122], [123], [124], [125], [126], [127]. BD created on the cloud platform as Data-as-a-Service (DaaS) can turn into Information-as-a-Service (InaaS); InaaS can turn into Knowledge-as-a-Service (KaaS) using BD analytics, and finally, KaaS transforms into Wisdom-as-a-Service (WaaS) [128]. In other words, BD is turned into insights to be input to other systems or merged with other insights to gain further insights. Insights are created both in the edge and cloud platforms. The WaaS standards and service platforms are expected to be fine-tuned continuously as a core infrastructure for intelligence industry and smart city to support the development of various intelligent IT applications and it is anticipated that this will bring a huge economic value for intelligence IT industry by realising the pay-as-you-go concept [9].

More than 87% of organisations are classified as having low BI and analytics maturity, according to a survey [129]. Another survey conducted by MIT found that organisations that "strongly agreed that the use of business information and analytics differentiates them within their industry" were more likely to be top performers and such organisations use analytics in a wide spectrum of decision making, both to guide future strategies and for day-to-day operations [6]. The digital business future confronts individuals and companies with almost unlimited possibilities to create business value through data and analytics by deploying and running BD analytics on the cloud [87], over 60% of global enterprises adopts public clouds for BD analytics [116]. It can be safely concluded that the way of doing business based on insights that can be gained from the globally generated BD from multiple channels using BD analytics triggers a new industrial revolution where bigger data increases the chance of representing the real world better and consequently, extracting better insights for better decision making. The BD analytics with its decision support capability, provide critical information such as historical reports, statistical analyses, time-series comparison, forecasting business opportunities and executive summaries to managers and executives to facilitate better decision making [130]. Furthermore, the BD analytics produce and deliver insights to be used as an input into other systems or merged with other insights created locally for better working IoT and AMS to make autonomous and dynamic real-time decisions [8], [17]. To become and remain competitive, enterprises must seek to adopt advanced analytics, and adapt their business models, establish specialist data science teams and rethink their overall strategies to keep pace with the competition [87]. The transformation of smart domains and platforms into smarter domains will foster the development of the industry to make our life better and simpler.

In future mobile networks, such as 5G/6G, emerging smart services are expected to support billions of smart devices with unique characteristics and traffic patterns [131] to facilitate the application of wireless BD and to achieve a flexible and efficient communication, consequently, an excellent synergy using the smart platforms and wiser domains. High-level topics concerning today's production of goods and

services include sustainability, flexibility, efficiency, and competitiveness [132]. In this manner, today's rapid changing technological and business environment urges the companies to be agile in order to adapt to upheaval market fluctuations, cope with unprecedented threats and most importantly thrive in a competitive business environment, even during recessions by foreseeing and exploiting the emerging business opportunities. The exponentially increasing volumes of data needs to turn into insights using advanced data analytics, which, in turn, will lead to an aggregation of wisdom to optimise all processes to ensure higher quality services and goods are manufactured at a lower cost with high quality. More explicitly, transforming the business and products into more intelligent, more autonomous services and products with increased customisable functionalities will be possible to maintain a competitive edge by meeting the market dynamics. Moreover, insights obtained from the high quality BD will bring huge economic value for the wiser IT industry based on the instant pay-as-you-go services. These insights will provide a wiser environment and particularly, this will be the core architecture of Industry 5.0 in the coming age of CPSSs. To prevent chaos in the hyper-connected world, businesses need to make every effort to reduce the complexity of connected systems, enhance the security and standardisation of applications, and guarantee the safety and privacy of users anytime, anywhere, on any device [133]. By focusing on user interaction and configurability, lifetime optimisation, intelligent analysis of BD, location independent monitoring and control, data security and reduced system complexity using effective management of gained insights, the way of doing business unsustainably and more intelligently will be realised with Industry 5.0.

ACKNOWLEDGMENT

I would like to thank DOMO Inc. (https://www.domo.com/), a cloud software company specialised in business intelligence tools and data visualisation, for permitting me to use Fig. 2.

REFERENCES

[1] "Data never sleeps 10.0," 2023. [Online]. Available: https://www.domo.com/data-never-sleeps

[2] J. Chen, J. Ma, N. Zhong, Y. Yao, J. Liu, R. Huang, W. Li, Z. Huang, Y. Gao, and J. Cao, "Waas: Wisdom as a service," *IEEE Intelligent Systems*, vol. 29, no. 6, pp. 40–47, Nov 2014.

[3] K. Kuru, "Metaomnicity: Toward immersive urban metaverse cyberspaces using smart city digital twins," *IEEE Access*, vol. 11, pp. 43 844–43 868, 2023.

[4] DOMO, "Data never sleeps 6.0," 2018. [Online]. Available: https://www.domo.com/learn/data-never-sleeps-6

[5] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and J. Kenneth C. Lichtendahl, *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, 2nd ed. NJ, USA: Wiley, 2018.

[6] A. C. Victor and S. Rao, "Analytics on the cloud," *IEEE Potentials*, vol. 37, no. 4, pp. 24–27, July 2018.

[7] K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: preserving security and privacy," *Journal of Big Data*, vol. 5, no. 1, p. 1, Jan 2018. [Online]. Available: https://doi.org/10.1186/s40537-017-0110-7

[8] K. Kuru and H. Yetgin, "Transformation to advanced mechatronics systems within new industrial revolution: A novel framework in automation of everything (aoe)," *IEEE Access*, vol. 7, pp. 41 395–41 415, 2019.

[9] J. Chen, J. Ma, N. Zhong, Y. Yao, J. Liu, R. Huang, W. Li, Z. Huang, and Y. Gao, "Waaswisdom as a service," in *Wisdom Web of Things*, N. Zhong, J. Ma, J. Liu, R. Huang, and X. Tao, Eds. Springer International Publishing Switzerland: Springer Nature, 2016, ch. 2, pp. 27–46.

[10] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, October 2016.

[11] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, Dec 2016.

[12] R. Buyya and S. N. Srirama, *Wiley Series on Parallel and Distributed Computing*, A. Y. Zomaya, Ed. John Wiley & Sons, Ltd, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119525080.scard

[13] A. C. Zhou, Y. Xiao, Y. Gong, B. He, J. Zhai, and R. Mao, "Privacy regulation aware process mapping in geo-distributed cloud data centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 8, pp. 1872–1888, Aug 2019.

[14] R. S. Montero, E. Rojas, A. A. Carrillo, and I. M. Llorente, "Extending the cloud to the network edge," *Computer*, vol. 50, no. 4, pp. 91–95, April 2017.

[15] Y. Ai, M. Peng, and K. Zhang, "Edge computing technologies for internet of things: a primer," *Digital Communications and Networks*, vol. 4, no. 2, pp. 77 – 86, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352864817301335

[16] M. Satyanarayanan, R. Schuster, M. Ebling, G. Fettweis, H. Flinck, K. Joshi, and K. Sabnani, "An open ecosystem for mobile-cloud convergence," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 63–70, March 2015.

[17] K. Kuru and D. Ansell, "Tcitysmartf: A comprehensive systematic framework for transforming cities into smart cities," *IEEE Access*, vol. 8, pp. 18 615–18 644, 2020.

[18] M. Gharbaoui, B. Martini, D. Adami, S. Giordano, and P. Castoldi, "Cloud and network orchestration in sdn data centers: Design principles and performance evaluation," *Computer Networks*, vol. 108, pp. 279 – 295, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128616302821

[19] O. Consortium, "Openfog architecture overview," 2016. [Online]. Available: https://www.alibabacloud.com/blog/bringing-iot-to-the-cloud-fog-computing-and-cloudlets_593824

[20] K. A. Beaty, J. M. Chow, R. L. F. Cunha, K. K. Das, M. F. Hulber, A. Kundu, V. Michelini, and E. R. Palmer, "Managing sensitive applications in the public cloud," *IBM Journal of Research and Development*, vol. 60, no. 2-3, pp. 4:1–4:13, March 2016.

[21] E. Borgia, "The internet of things vision: Key features, applications and open issues," *Computer Communications*, vol. 54, pp. 1 – 31, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366414003168

[22] Q. Zhang, Z. Yu, W. Shi, and H. Zhong, "Demo abstract: Evaps: Edge video analysis for public safety," in *2016 IEEE/ACM Symposium on Edge Computing (SEC)*, Oct 2016, pp. 121–122.

[23] Z. Sheng, S. Yang, Y. Yu, A. V. Vasilakos, J. A. Mccann, and K. K. Leung, "A survey on the ietf protocol suite for the internet of things: standards, challenges, and opportunities," *IEEE Wireless Communications*, vol. 20, no. 6, pp. 91–98, December 2013.

[24] D. Miorandi, S. Sicari, F. D. Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497 – 1516, 2012. [Online]. Available:

http://www.sciencedirect.com/science/article/pii/S1570870512000674

[25] M. J. Kaur and P. Maheshwari, "Building smart cities applications using iot and cloud-based architectures," in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, March 2016, pp. 1–5.

[26] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.

[27] K. Kuru, S. Clough, D. Ansell, J. McCarthy, and S. McGovern, "Wildetect: An intelligent platform to perform airborne wildlife census automatically in the marine ecosystem using an ensemble of learning techniques and computer vision," *Expert Systems with Applications*, vol. 231, p. 120574, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742301076X

[28] ——, "Intelligent airborne monitoring of irregularly shaped man-made marine objects using statistical machine learning techniques," *Ecological Informatics*, vol. 78, p. 102285, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495412300314X

[29] K. Kuru, D. Ansell, and D. Jones, "Airborne vision-based remote sensing imagery datasets from large farms using autonomous drones for monitoring livestock," 2023.

[30] K. Kuru, D. Ansell, B. Jon Watkinson, D. Jones, A. Sujit, J. M. Pinder, and C. L. Tinker-Mill, "Intelligent automated, rapid and safe landmine and unexploded ordnance (uxo) detection using multiple sensor modalities mounted on autonomous drones," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[31] K. Kuru and D. Ansell, "Vision-based remote sensing imagery datasets from benkovac landmine test site using an autonomous drone for detecting landmine locations," *IEEE Data Port*, pp. 1–10, 2023.

[32] K. Kuru, "Iotfauav: Intelligent remote monitoring of livestock in large farms using autonomous uninhabited aerial vehicles," *Computers and Electronics in Agriculture*, 2023.

[33] K. Kuru, D. Ansell, and D. Jones, "Intelligent airborne monitoring of livestock using autonomous uninhabited aerial vehicles," 2023.

[34] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645 – 1660, 2013, including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services and Cloud Computing and Scientific Applications Big Data, Scalable Analytics, and Beyond. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X13000241

[35] S. Tayeb, S. Latifi, and Y. Kim, "A survey on iot communication and computation frameworks: An industrial perspective," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan 2017, pp. 1–6.

[36] M. Amadeo, C. Campolo, A. Iera, and A. Molinaro, "Information centric networking in iot scenarios: The case of a smart home," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 648–653.

[37] K. Kuru and K. Kuru, "Blockchain-based decentralised privacy-preserving machine learning authentication and verification with immersive devices in the urban metaverse ecosystem," *Preprints*, February 2024. [Online]. Available: https://doi.org/10.20944/preprints202402.0317.v1

[38] K. Kuru, S. Worthington, D. Ansell, J. M. Pinder, A. Sujit, B. Jon Watkinson, K. Vinning, L. Moore, C. Gilbert, D. Jones *et al.*, "Aitl-wing-hitl: Telemanipulation of autonomous drones using digital twins of aerial traffic interfaced with wing," *IEEE Access*, vol. 11, 2023.

[39] K. Kuru, "Planning the future of smart cities with swarms of fully autonomous unmanned aerial vehicles using a novel framework," *IEEE Access*, vol. 9, pp. 6571–6595, 2021.

[40] K. Kuru and W. Khan, "A framework for the synergistic integration of fully autonomous ground vehicles with smart city," *IEEE Access*, vol. 9, pp. 923–948, 2021.

[41] K. Kuru, "Conceptualisation of human-on-the-loop haptic teleoperation with fully autonomous self-driving vehicles in the urban environment," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 2, pp. 448–469, 2021.

[42] K. Kuru, J. M. Pinder, B. J. Watkinson, D. Ansell, K. Vinning, L. Moore, C. Gilbert, A. Sujit, and D. Jones, "Toward mid-air collision-free trajectory for autonomous and pilot-controlled unmanned aerial vehicles," *IEEE Access*, vol. 11, pp. 100 323–100 342, 2023.

[43] K. Kuru, "Trustfsdv: Framework for building and maintaining trust in self-driving vehicles," *IEEE Access*, vol. 10, pp. 82 814–82 833, 2022.

[44] ——, "Technical report: Essential development components of the urban metaverse ecosystem," *CLoK*, 2024.

[45] K. Kuru, D. Ansell, W. Khan, and H. Yetgin, "Analysis and optimization of unmanned aerial vehicle swarms in logistics: An intelligent delivery platform," *IEEE Access*, vol. 7, pp. 15 804–15 831, 2019.

[46] E. Glaessgen and D. Stargel, *The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles*. ARC, 2012. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2012-1818

[47] R. Koduri, "Powering the metaverse," 2021. [Online]. Available: https://www.intel.com/content/www/us/en/newsroom/opinion/powering-metaverse.html#gs.nkd8ql

[48] K. Benouaret, R. Valliyur-Ramalingam, and F. Charoy, "Crowdsc: Building smart cities with large-scale citizen participation," *IEEE Internet Computing*, vol. 17, no. 6, pp. 57–63, Nov 2013.

[49] S. Nover, "Intel wants to take you inside the metaverse," 2021. [Online]. Available: https://qz.com/2101581/intel-is-ready-to-talk-about-the-metaverse

[50] L. Chang, Z. Zhang, P. Li, S. Xi, W. Guo, Y. Shen, Z. Xiong, J. Kang, D. Niyato, and X. Q. Y. Wu, "6g-enabled edge ai for metaverse:challenges, methods,and future research directions," *Journal of Communications and Information Networks*, vol. 7, no. 2, p. 107, 2022.

[51] J. Han, M. Kam, and J. Pei, *Data Mining Concepts and Techniques,*, 3rd ed. San Francisco, USA: Elsevier, 2012.

[52] G. T. Lakshmanan and R. Khalaf, "Leveraging process-mining techniques," *IT Professional*, vol. 15, no. 5, pp. 22–30, Sep. 2013.

[53] A. C. Baktir, C. Sonmez, CemErsoy, A. Ozgovde, and B. Varghese, "Addressing the challenges in federating edge resources," in *Fog and Edge Computing*, A. Y. Zomaya, Ed. John Wiley & Sons, Ltd, 2019, pp. 25–47. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119525080.scard

[54] K. Kuru, "Sensors and sensor fusion for decision making in autonomous driving and vehicles," 2023.

[55] K. Kuru, O. Erogul, and C. Xavier, "Autonomous low power monitoring sensors," *Sensors*, vol. 21, 2021.

[56] K. Kuru, "Definition of multi-objective deep reinforcement learning reward functions for self-driving vehicles in the urban environment," *IEEE Transactions on Vehicular Technology*, 2023.

[57] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE Access*, vol. 6, pp. 6900–6919, 2018.

[58] P. Eugster, C. Jayalath, K. Kogan, and J. Stephen, "Big data analytics beyond the single datacenter," *Computer*, vol. 50, no. 6, pp. 60–68, 2017.

[59] Q. Xia, W. Liang, and Z. Xu, "Data locality-aware big data query evaluation in distributed clouds," *The Computer Journal*, vol. 60, no. 6, pp. 791–809, June 2017.

[60] Y. Sun, H. Yan, J. Zhang, Y. Xia, S. Wang, R. Bie, and Y. Tian, "Organizing and querying the big sensing data with

event-linked network in the internet of things," *International Journal of Distributed Sensor Networks*, vol. 10, no. 8, p. 218521, 2014. [Online]. Available: https://doi.org/10.1155/2014/218521

[61] D. Zeng, L. Gu, and S. Guo, *Cloud Networking for Big Data*, 1st ed. Switzerland: Springer, 2015.

[62] Z. Hu, B. Li, and J. Luo, "Time- and cost- efficient task scheduling across geo-distributed data centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 3, pp. 705–718, March 2018.

[63] L. Zhao, Y. Yang, A. Munir, A. X. Liu, Y. Li, and W. Qu, "Optimizing geo-distributed data analytics with coordinated task scheduling and routing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 2, pp. 279–293, 2020.

[64] R. Jiménez-Peris, M. Patiño-Martínez, B. Kemme, I. Brondino, J. Pereira, R. Vilaça, F. Cruz, R. Oliveira, and M. Y. Ahmad, "Cumulonimbo: A cloud scalable multi-tier SQL database," *IEEE Data Eng. Bull.*, vol. 38, no. 1, pp. 73–83, 2015. [Online]. Available: http://sites.computer.org/debull/A15mar/p73.pdf

[65] Q. Xiang, S. Chen, K. Gao, H. Newman, I. Taylor, J. Zhang, and Y. R. Yang, "Unicorn: Unified resource orchestration for multi-domain, geo-distributed data analytics," in *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug 2017, pp. 1–6.

[66] Z. Luo, C. Wu, Z. Li, and W. Zhou, "Scaling geo-distributed network function chains: A prediction and learning framework," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1838–1850, 2019.

[67] M. Taneja, N. Jalodia, and A. Davy, "Distributed decomposed data analytics in fog enabled iot deployments," *IEEE Access*, vol. 7, pp. 40 969–40 981, 2019.

[68] S. K. Sharma and X. Wang, "Live data analytics with collaborative edge and cloud processing in wireless iot networks," *IEEE Access*, vol. 5, pp. 4621–4635, 2017.

[69] M. Gong, Y. Zhang, Y. Gao, A. K. Qin, Y. Wu, S. Wang, and Y. Zhang, "A multi-modal vertical federated learning framework based on homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1826–1839, 2024.

[70] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, jan 2019. [Online]. Available: https://doi.org/10.1145/3298981

[71] P. Li, Z. Zhang, A. S. Al-Sumaiti, N. Werghi, and C. Y. Yeun, "A robust adversary detection-deactivation method for metaverse-oriented collaborative deep learning," *IEEE Sensors Journal*, pp. 1–1, 2023.

[72] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 603618. [Online]. Available: https://doi.org/10.1145/3133956.3134012

[73] Z. Chen, J. Wu, A. Fu, M. Su, and R. H. Deng, "Mp-clf: An effective model-preserving collaborative deep learning framework for mitigating data leakage under the gan," *Knowledge-Based Systems*, vol. 270, p. 110527, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123002770

[74] L. Lyu, Y. Li, K. Nandakumar, J. Yu, and X. Ma, "How to democratise and protect ai: Fair and differentially private decentralised deep learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 2, pp. 1003–1017, 2022.

[75] L. Zhang, J. Xu, P. Vijayakumar, P. K. Sharma, and U. Ghosh, "Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 5, pp. 2864–2880, 2023.

[76] R. Podschwadt, D. Takabi, P. Hu, M. H. Rafiei, and Z. Cai, "A survey of deep learning architectures for privacy-preserving machine learning with fully homomorphic encryption," *IEEE Access*, vol. 10, pp. 117 477–117 500, 2022.

[77] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[78] J. W. Bos, K. Lauter, J. Loftus, and M. Naehrig, "Improved security for a ring-based fully homomorphic encryption scheme," in *Cryptography and Coding*, M. Stam, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 45–64.

[79] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1333–1345, 2018.

[80] C. Zhou and N. Ansari, "Securing federated learning enabled nwdaf architecture with partial homomorphic encryption," *IEEE Networking Letters*, vol. 5, no. 4, pp. 299–303, 2023.

[81] J.-W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, E. Lee, J. Lee, D. Yoo, Y.-S. Kim, and J.-S. No, "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network," *IEEE Access*, vol. 10, pp. 30 039–30 054, 2022.

[82] L. Arbuckle and K. E. Emam, *Anonymizing Health data*, 1st ed. Surrey, UK: O'Reilly Media, Inc., 2013.

[83] S. Sobati Moghadam and A. Fayoumi, "Toward securing cloud-based data analytics: A discussion on current solutions and open issues," *IEEE Access*, vol. 7, pp. 45 632–45 650, 2019.

[84] GALLUP, "Americans: Credit card information still getting hacked," 2016. [Online]. Available: https://news.gallup.com/poll/196802/americans-credit-card-information-getting-hacked.aspx

[85] L. H. Newman, "Hack brief: Hackers breach a billion yahoo accounts. a billion," 2016. [Online]. Available: https://www.wired.com/2016/12/yahoo-hack-billion-users/

[86] ICloudPE, "How cloud storage became a target for hackers and what can be done about it," 2016. [Online]. Available: https://icloud.pe/blog/how-cloud-storage-became-a-target-for-hackers-and-what-can-be-done-about-it/

[87] D. Laney and A. Jain, "100 data and analytics predictions through 2021," 2018. [Online]. Available: https://www.gartner.com/en/doc/3746424-100-data-and-analytics-predictions-through-2021

[88] P. O. Obitade, "Big data analytics: a link between knowledge management capabilities and superior cyber protection," *Journal of Big Data*, vol. 6, no. 1, p. 71, Aug 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0229-9

[89] J. Crowcroft and A. Gascn, "Analytics without tears or is there a way for data to be anonymized and yet still useful?" *IEEE Internet Computing*, vol. 22, no. 3, pp. 58–64, May 2018.

[90] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Illinois, USA: Morgan Kaufmann, 2016.

[91] T. Barrabi, "Why hackers love the cloud," 2016. [Online]. Available: https://www.foxbusiness.com/features/why-hackers-love-the-cloud

[92] Y. Zhang, H. Huang, Y. Xiang, L. Y. Zhang, and X. He, "Harnessing the hybrid cloud for secure big image data service," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1380–1388, Oct 2017.

[93] L. Mearian, "Internet of things data to top 1.6 zettabytes by 2022." 2016. [Online]. Available: ttps://campustechnology.com/articles/2015/04/15/internet-of-thingsdata-to-top-1-6-zettabytes-by-2020.aspx

[94] Y. Lee, W. Hsiao, Y. Lin, and S. T. Chou, "Privacy-preserving data analytics in cloud-based smart home with community hierarchy," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 2, pp. 200–207, May 2017.

[95] L. SWEENEY, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002. [Online]. Available: https://doi.org/10.1142/S0218488502001648

[96] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[97] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*, April 2006, pp. 24–24.

[98] C. Liu and P. Mittal, "Linkmirage: Enabling privacy-preserving analytics on social relationships," in *NDSS*, 2016.

[99] J. C. Lin, J. M. Wu, P. Fournier-Viger, Y. Djenouri, C. Chen, and Y. Zhang, "A sanitization approach to secure shared data in an iot environment," *IEEE Access*, vol. 7, pp. 25 359–25 368, 2019.

[100] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, July 2018.

[101] D. Snchez, M. Batet, and A. Viejo, "Automatic general-purpose sanitization of textual documents," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 853–862, June 2013.

[102] Y. Hong, J. Vaidya, H. Lu, P. Karras, and S. Goel, "Collaborative search log sanitization: Toward differential privacy and boosted utility," *IEEE Transactions on Dependable and Secure Computing*, vol. 12, no. 5, pp. 504–518, Sep. 2015.

[103] J. M. Wu, J. Zhan, and J. C. Lin, "Ant colony system sanitization approach to hiding sensitive itemsets," *IEEE Access*, vol. 5, pp. 10 024–10 039, 2017.

[104] Q. Han, D. Lu, K. Zhang, X. Du, and M. Guizani, "Lclean: A plausible approach to individual trajectory data sanitization," *IEEE Access*, vol. 6, pp. 30 110–30 116, 2018.

[105] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, Jan 2018.

[106] S.-S. Yau, H. An, and A. Buduru, "An approach to data confidentiality protection in cloud environments," *International Journal of Web Services Research*, vol. 9, no. 3, pp. 67–83, 7 2012.

[107] A. De Salve, R. D. Pietro, P. Mori, and L. Ricci, "A logical key hierarchy based approach to preserve content privacy in decentralized online social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 1, pp. 2–21, 2020.

[108] R. Chaudhary, G. S. Aujla, N. Kumar, and J. J. P. C. Rodrigues, "Optimized big data management across multi-cloud data centers: Software-defined-network-based analysis," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 118–126, Feb 2018.

[109] S. Ji and B. Li, "Wide area analytics for geographically distributed datacenters," *Tsinghua Science and Technology*, vol. 21, no. 2, pp. 125–135, April 2016.

[110] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the internet of things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, Aug 2016.

[111] Z. A. Mann, "Optimization problems in fog and edge computing," in *Fog and Edge Computing*, A. Y. Zomaya, Ed. John Wiley & Sons, Ltd, 2019, pp. 103–122. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119525080.scard

[112] K. Kumar and Y. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 43, no. 4, pp. 51–56, April 2010.

[113] R. H. Weber, "Internet of things new security and privacy challenges," *Computer Law and Security Review*, vol. 26, no. 1, pp. 23 – 30, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364909001939

[114] E. Pateromichelakis, F. Moggio, C. Mannweiler, P. Arnold, M. Shariat, M. Einhaus, Q. Wei, . Bulakci, and A. De Domenico, "End-to-end data analytics framework for 5g architecture," *IEEE Access*, vol. 7, pp. 40 295–40 312, 2019.

[115] M. G. Richard, "Amazon.com caves to pressure, commits to 100% renewable energy on its cloud platform,"

2014. [Online]. Available: https://www.treehugger.com/corporate-responsibility/amazon-caves-pressure-greens-commits-100-renewable-energy-cloud-platform.html

[116] F. Xu, H. Zheng, H. Jiang, W. Shao, H. Liu, and Z. Zhou, "Cost-effective cloud server provisioning for predictable performance of big data analytics," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 5, pp. 1036–1051, May 2019.

[117] K. Kuru and W. Khan, "Novel hybrid object-based non-parametric clustering approach for grouping similar objects in specific visual domains," *Applied Soft Computing*, vol. 62, pp. 667 – 701, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494617306701

[118] B. Li, Y. Vorobeychik, M. Li, and B. Malin, "Scalable iterative classification for sanitizing large-scale datasets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 3, pp. 698–711, March 2017.

[119] A. Sharma, T. Goyal, E. S. Pilli, A. P. Mazumdar, M. C. Govil, and R. C. Joshi, "A secure hybrid cloud enabled architecture for internet of things," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, Dec 2015, pp. 274–279.

[120] J. K. Winn and B. Wright, *The Law of Electronic Commerce*, 4th ed. NY, USA: Wolters Kluwer, 2019.

[121] K. Kuru, M. Niranjan, Y. Tunca, E. Osvank, and T. Azim, "Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics," *Artificial Intelligence in Medicine*, vol. 62, no. 2, pp. 105–118, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365714000980

[122] K. Kuru, S. Girgin, K. Arda, and U. Bozlar, "A novel report generation approach for medical applications: The sisds methodology and its applications," *International Journal of Medical Informatics*, vol. 82, no. 5, pp. 435–447, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138650561200113X

[123] K. Kuru, D. Ansell, M. Jones, C. De Goede, and P. Leather, "Feasibility study of intelligent autonomous determination of the bladder voiding need to treat bedwetting using ultrasound and smartphone ml techniques," *Medical & Biological Engineering & Computing*, Dec 2018. [Online]. Available: https://doi.org/10.1007/s11517-018-1942-9

[124] K. Kuru, D. Ansell, M. Jones, B. J. Watkinson, N. Caswell, P. Leather, A. Lancaster, P. Sugden, E. Briggs, C. Davies, T. C. Oh, K. Bennett, and C. De Goede, "Intelligent autonomous treatment of bedwetting using non-invasive wearable advanced mechatronics systems and mems sensors: Intelligent autonomous bladder monitoring to treat ne," *Medical & biological engineering & computing*, vol. 58, no. 4, pp. 1 – 123, February 2020. [Online]. Available: https://doi.org/10.1007/s11517-019-02091-x

[125] N. Caswell, K. Kuru, D. Ansell, M. J. Jones, B. J. Watkinson, P. Leather, A. Lancaster, P. Sugden, E. Briggs, C. Davies, C. Oh, K. Bennett, and C. DeGoede, "Patient engagement in medical device design: Refining the essential attributes of a wearable, pre-void, ultrasound alarm for nocturnal enuresis," *Pharmaceutical Medicine*, vol. 34, no. 1, p. 3948, Jan. 2020. [Online]. Available: http://dx.doi.org/10.1007/s40290-019-00324-w

[126] K. Kuru, D. Ansell, D. Hughes, B. J. Watkinson, F. Gaudenzi, M. Jones, D. Lunardi, N. Caswell, A. R. Montiel, P. Leather, D. Irving, K. Bennett, C. McKenzie, P. Sugden, C. Davies, and C. Degoede, "Treatment of nocturnal enuresis using miniaturised smart mechatronics with artificial intelligence," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 12, pp. 204–214, 2024.

[127] K. Kuru, B. J. Watkinson, D. Ansell, D. Hughes, M. Jones, N. Caswell, P. Leather, K. Bennett, P. Sugden, C. Davies, and C. DeGoede, "Smart wearable device for nocturnal enuresis," in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, 2023, pp. 95–96.

[128] K. Kuru, "Management of geo-distributed intelligence: Deep insight as a service (dinsaas) on forged cloud platforms (fcp)," *Journal of Parallel and Distributed Computing*, vol. 149, pp. 103–118, 2021.

[129] Gartner, "Gartner data shows 87 percent of organizations have low bi and analytics maturity," 2019.

[Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2018-12-06-gartner-data-shows-87-percent-of-organizations-have-low-bi-and-analytics-maturity

[130] B. Marr, *Big data: using SMART big data. analytics and metrics to make better decisions and improve performance*, 5th ed. Chichester, UK: Wiley, 2015.

[131] B. Nguyen, N. Choi, M. Thottan, and J. V. der Merwe, "Simeca: Sdn-based iot mobile edge cloud architecture," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, May 2017, pp. 503–509.

[132] J. Delsing, "Local cloud internet of things automation: Technology and business model features of distributed internet of things automation solutions," *IEEE Industrial Electronics Magazine*, vol. 11, no. 4, pp. 8–21, Dec 2017.

[133] I. Lee and K. Lee, "The internet of things (iot): Applications, investments, and challenges for enterprises," *Business Horizons*, vol. 58, no. 4, pp. 431 – 440, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0007681315000373

[134] Gartner, "Gartner says worldwide public cloud services market to grow 18 percent in 2017," 2017. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2017-02-22-gartner-says-worldwide-public-cloud-services-market-to-grow-18-percent-in-2017