Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(REVIEW ARTICLE)

Check for updates

Big data and data analytics in 5G mobile networks

Panagiotis Leliopoulos and Athanasios Drigas *

Net Media Lab IIT NCSR Demokritos, Athens, Greece.

Global Journal of Engineering and Technology Advances, 2023, 15(03), 165–190

Publication history: Received on 12 May 2023; revised on 27 June 2023; accepted on 29 June 2023

Article DOI: https://doi.org/10.30574/gjeta.2023.15.3.0114

Abstract

In this paper, we study the features of big data and data analytics. We see how Big Data contributes to mobile networks. We give a term in which big data generally refers to a large amount of digital data. Also, we estimate that the amount processed by "Big Data" systems will double every two years. Hence, Big Data on mobile networks need to be analyzed in-depth according to retrieve exciting and useful information. Big Data provides unprecedented opportunities for internet service providers to understand the behavior and requirements of their users, which in turn enables real-time decision making across a wide range of applications. After that, we mention the dimensions often describe the 4Vs of Big Data. We continue with the study about the use of big data analytics in mobile networks. As we see, new technologies for managing big data in a highly scalable, cost-effective, and damage-resistant manner are required. So, beyond 2020 the system capacity and data rates in mobile networks must support thousands of times more traffic than 2010 levels. Furthermore, we mention the end-to-end latency, the massive number of connections, the cost, the Quality of Experience, the Issues, and finally, the big data management. We continue with the study about the big data analytics in 5G. The 5G networks standardizing and the 5G mobile optimization are crucial areas. There are new research areas were exploring new analytics techniques in big data according to discover new patterns and extract knowledge from the data are collected. Big data analytics can provide organizations with the ability to profile and segment customers based on distinct socioeconomic characteristics and increase customer satisfaction and retention levels. Also, Big Data analytics techniques can provide telecom providers with in-depth knowledge of networks before making informed decisions. Also, as we see, these analytics techniques can help Telecommunication providers to monitor and analyze various types of data as well as event messages on networks. Important information, like business intelligence, can be extracted from momentary and stored data. Hence, the mass adoption of smartphones, mobile broadband modems, tablets, and mobile data applications has been overwhelmingly wireless. Operators bend under the pressure and cost of continuously adding capacity and improving coverage while maximizing the use of the existing components of their range. Advanced radio access technologies, and all Internet Protocols, open internet network architectures must evolve smoothly from 4G systems. So those needs are leading us to make a study about the heterogeneous network or else HetNet for 5G networks. We are continuing with the challenges, and we mention about the curse of modularity, dimensions procedure, feature engineering, non-linearity, Bonferonni's principle, category report, variance and bias, data locality, data heterogeneity, noisy data, data availability, real-time processing, and streaming, data provenance, and data security.

Keywords: Big Data; Machine Learning; Mobiles; AI

* Corresponding author: Athanasios Drigas.

1 Introduction

In recent years, "Big Data" has been a rapidly evolving application [1]. The evolution of the field of "Big Data Analytics" (An Algorithmic Massive Data), which offers innovative methodologies for reusing this considerable amount of data produced daily to extract valuable information from "information chaos" [2].

So, the term "Big Data" generally refers to the enormous amount of digital information produced daily by humans and their environment and stored by various companies and governments for further processing and exploitation. Besides, estimating that the amount processed by "Big Data" systems will double every two years [3].

Also, recent statistical surveys testify to the rapid production of data in various information systems. So, Google processes over 40,000 search queries every second and send more than 200 million emails every day. YouTube users transfer 720K hours of new videos. Facebook users share more than 510,000 comments, 293,000 statuses are updated, and 136,000 photos are uploaded every minute. Finally, Twitter users generate 200 billion tweets per year [4] [5] [6] [7].

Besides, the New York Stock Exchange generates 1 Terabyte of financial data per day. The web site "Ancestry.com" has about 2.5 Petabytes volume stored. The Internet storage increases by 20 Terabytes per month, and the Large Hadron Collider in Genoa, Switzerland, generates 15 Petabytes of data annually [8] [9] [10].

From the above, we can see that big data has invaded all areas of modern life and is expected to affect it significantly. However, the results of Big Data and Big Data Analytics applications are most acceptable to everyone, such as when they support predicting climate change or the spread of an epidemic or even the side effects of a drug. It is expected, at the same time, for Big Data and Big Data Analytics to raise serious concerns about the protection of both confidentiality and personal data [11]. So as "Big Data" and "Big Data Analytics" come to stay, it cannot be accepted that this will be at the price of privacy [12].

In recent years, there has been a massive increase in mobile telephony, which has contributed to billions of mobile devices and vast internet traffic [13]. Worldwide cellular network traffic from mobile devices is exponentially increasing every year, 2022 expected to exceed 77 Exabytes per Month (Figure 1.1) [14]. Such a massive volume of mobile telephony constitutes a large-scale mobile and extensive archive that records human activities in the natural world, cyberbullying behaviors, and interactions with urban social ecology [13].

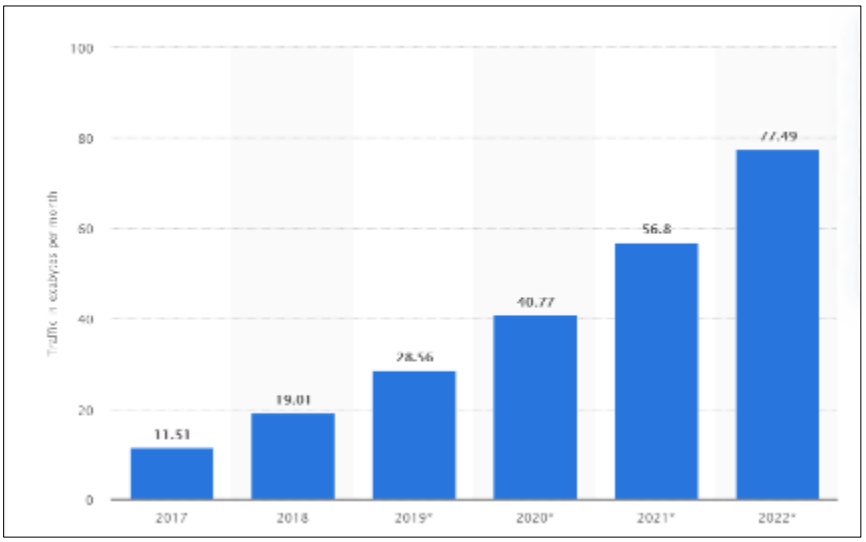

Figure 1 Global mobile data traffic from 2017 to 2022 (in exabytes per month) [14]

Hence, Big Data on mobile networks need to be analyzed in-depth according to retrieve interesting and useful information. Big Data provides unprecedented opportunities for I.S.P.s to understand the behavior and requirements of mobile users, which in turn enables real-time decision making across a wide range of applications. By analyzing these data, mobile networks can provide and support different smart services. However, the nature of Big Data presents enormous challenges in terms of data mining, mobile scanning, and knowledge discovery [15].

Thus, avoiding the limited capabilities of relational data models used to date, the ultimate goal is to identify new correlations or capture new and unexpected uses of that data. Indeed, these areas offer a whole new era of opportunities in various areas of modern human life, from commerce and online trading to science, health, research, and security, significantly changing everyday life [16].

2 The 4Vs Of Big Data

At this point, we must mention the 4Vs of Big Data [17], which are often described by its dimensions. The older definitions of Big Data focused on three Vs, (volume, velocity, and variety). However, the value is defined as the desired result of Big Data processing and not a defining feature of Big Data itself. This value provides an opportunity to relate the challenges directly to the defining characteristics of Big Data, making the origin and cause of each explicit. Figure 2.1 shows the dimensions of Big Data, along with the relevant challenges [18].

As we can see, the volume is the evaluation of Big Data by the size of the dataset. Velocity is the speed of data in and out. Variety is the range of data types. Moreover, veracity is the source and quality of the data. Consequently, regardless of these highlights, Big wireless data is usually considered the dataset that cannot be transmitted, accessed, processed, and served over a continuous period by existing communication and network systems [17]. Below we can see a short description of the 4Vs.

2.1 Volume

The first and most discussed feature of Big Data is volume. It is the quantity, size, and scale of data. In the context of machine learning (ML), size can be defined either vertically by the number of records or samples in a set of data or horizontally by the number of features or characteristics it contains. Hence, the volume is related to the data type; a smaller number of very complex data points can be considered equivalent to a more massive amount of simple data [18].

2.2 Velocity

Big Data's speed dimension refers not only to the speed at which data is generated but also to the rate at which it needs to be analyzed. This need drives to the ubiquity of smartphones and sensors in real-time. Again, this urgent need must quickly interact with the environment by developing technologies such as smart homes. The speed of Big Data has become an essential factor to consider [18].

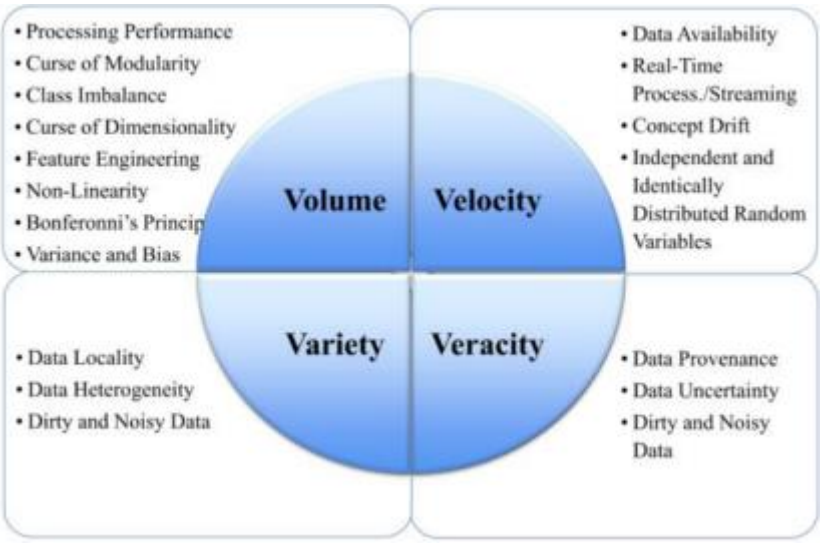

Figure 2 Big Data characteristics with associated challenges [18]

2.3 Variety

Big Data variety describes the structural variation of a set of data and the types of data it contains. However, the variety in what it represents, its semantic interpretation, and its sources. Although not as many as for other V dimensions, the challenges associated with this dimension have a significant impact [18].

2.4 Veracity

Big Data veracity lies not only in the reliability of the data that forms a dataset but also in the inherent reliability of the data sources. The origin and quality of Big Data together determine the element of truth, but they also pose some challenges [18].

3 The Use of Big Data Analytics in Mobile Networks and Challenges

As we can see from the above, new technologies for managing big data in a highly scalable, cost-effective, and damage-resistant manner are required [15]. So, beyond 2020 the system capacity and data rates in mobile networks must support thousands of times more traffic than 2010 levels. Also, it is necessary a 15-fold increase in transmission rates even in high mobility and crowded areas, if they continue the current trends [19].

The invoicing systems can manage and possibly reduce the rise in data consumption, as already demonstrated by market operators. However, as customers are willing to pay for the service instead of the volume of data, pricing models may not effectively suppress future traffic [19].

Today's consensus is that addressing these challenges in Radio Access Network (R.A.N.) requires a combination of more spectrum, higher spectrum performance, network condensation, and unloading (mmW), unauthorized range, and fragmented spectrum aggregation using carrier aggregation techniques. The dual connectivity of terminals to multiple base stations can take advantage of the general use of the spectrum deployed at different base stations [19] [20].

Below we can see some significant challenges for the 5G Mobile Networks (Figure 3.1):

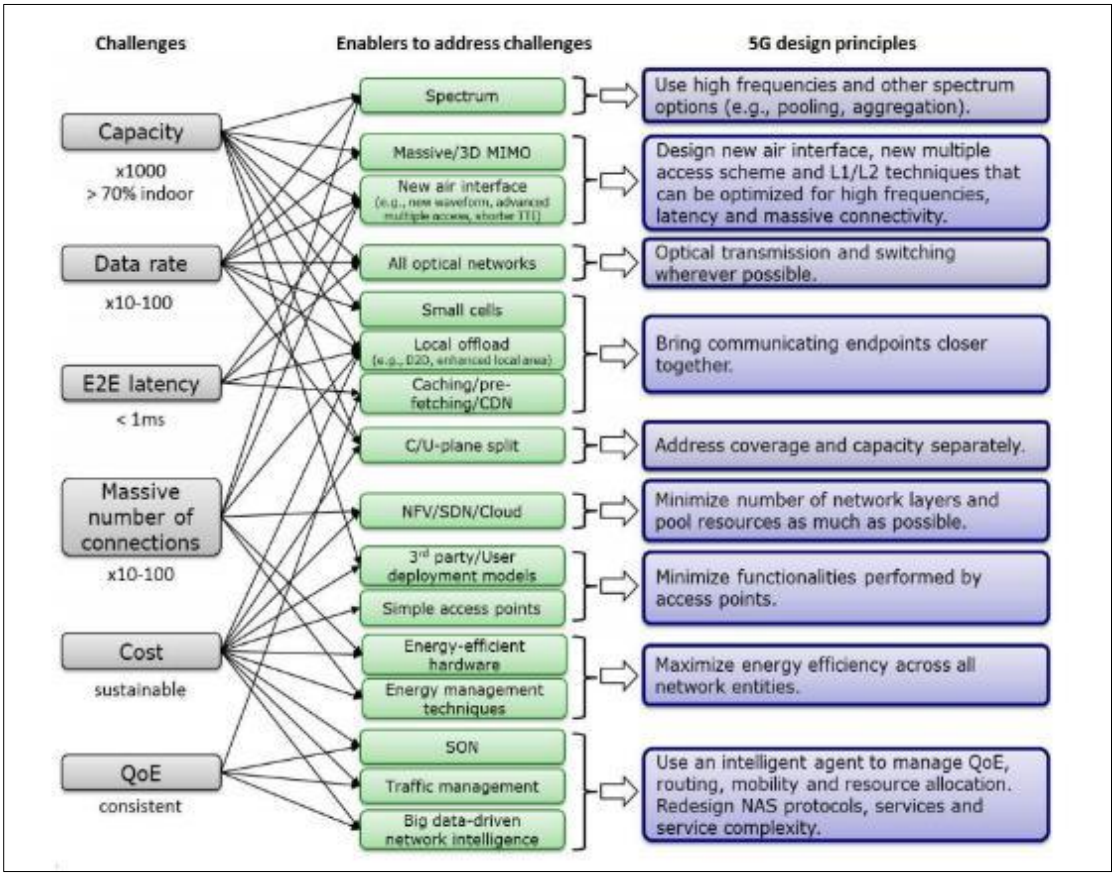

Figure 3 5G challenges, potential enablers, and design principles [19]

3.1 End-to-End Latency

End-to-end delay is vital for triggering new applications in real-time. For example, remote-controlled robots for medical, first responders, and industrial applications require fast feedback-control cycles to work well. Critical applications for cars and humans, based on vehicle-to-vehicle (V2V) communication, also require speedy application response and

feedback-control cycles with high availability and reliability. Enhanced applications and virtual reality applications (e.g., dynamic visuals and environments) require breakneck application-response cycles to mitigate cyberspace. For these applications to implement, networks must support a millisecond end-to-end latency target with high reliability. A new air interface with new counts, such as the shorter transmission time (T.T.I.), can reduce the over-the-air delay by a few hundred microseconds. The smaller T.T.I. It requires a high bandwidth available, but this can be supported using higher frequency bands [19].

3.2 A Massive Number of Connections

The number of connected devices expected to increment between ten and a hundred times after 2020. These will range from devices with limited resources that require only intermittent connectivity for reporting (e.g., sensors) to tools that always need connectivity to monitor and surveillance (e.g., security cameras, transport fleet). Apart from the significant number of connected devices, the challenge is to support the diversity of methods and service requirements in a scalable and efficient way. Combining advances in air interface design, signaling optimization, and smart clustering and relaying techniques can support hyperlinking. For example, using one device as a gateway or relay to garner traffic from multiple devices may reduce network signaling load. Also, offline access processes can effectively support Machine-type-communication (M.T.C.) applications that only require intermittent connectivity to transmit small packets. Not all devices can equip with high-precision tools to handle, e.g., with strict synchronization to maintain signal rectangle in a multi-access environment when new numerology is introduced to reduce latency (Figure 3.2) [19] [21].

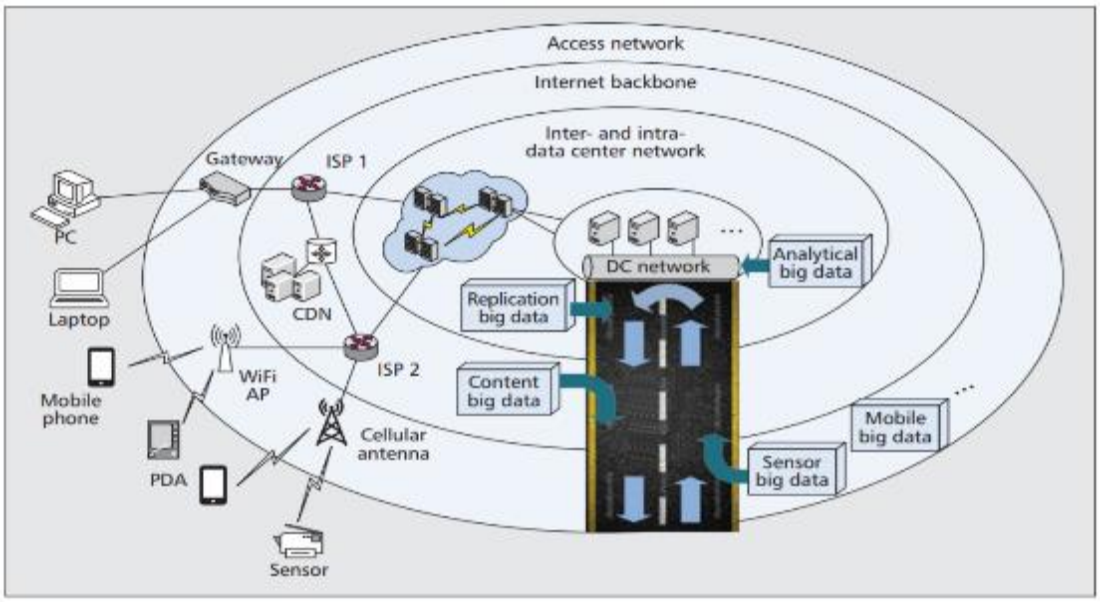

Figure 4 Three-layered system engineering from the point of big data applications [22]

3.3 Cost

Large data centers have been developed around the world, providing services to hundreds of thousands of users. According to that, a data center can consist of many servers and consume megawatts of power. Millions of dollars in electricity costs have caused a massive burden on operating costs for data center providers. As a result, reducing electricity costs has received considerable attention from academia and industry [23].

Therefore, it is crucial to lowering the cost of infrastructure as well as the costs associated with their development, maintenance, management, and operation to make connectivity a universally available, accessible, and sustainable utility. 5G should be a network covering all new requirements at a cost that will make the service viable. One way to reduce equipment costs is to minimize the number of base station operations. Reducing the number of functions leads to more unaffected base stations deployed by users and remotely or autonomously reduced installation and operating costs. Therefore, intelligent energy management techniques, especially in R.A.N., could provide a viable means of reducing the overall cost of running the network. Energy-efficient equipment design, low power backhaul, and smart power management techniques, especially on ultra-dense systems, to put base stations to rest when not being used can help reduce the cost of operating a 5G network (Figure 3.3) [19] [24].

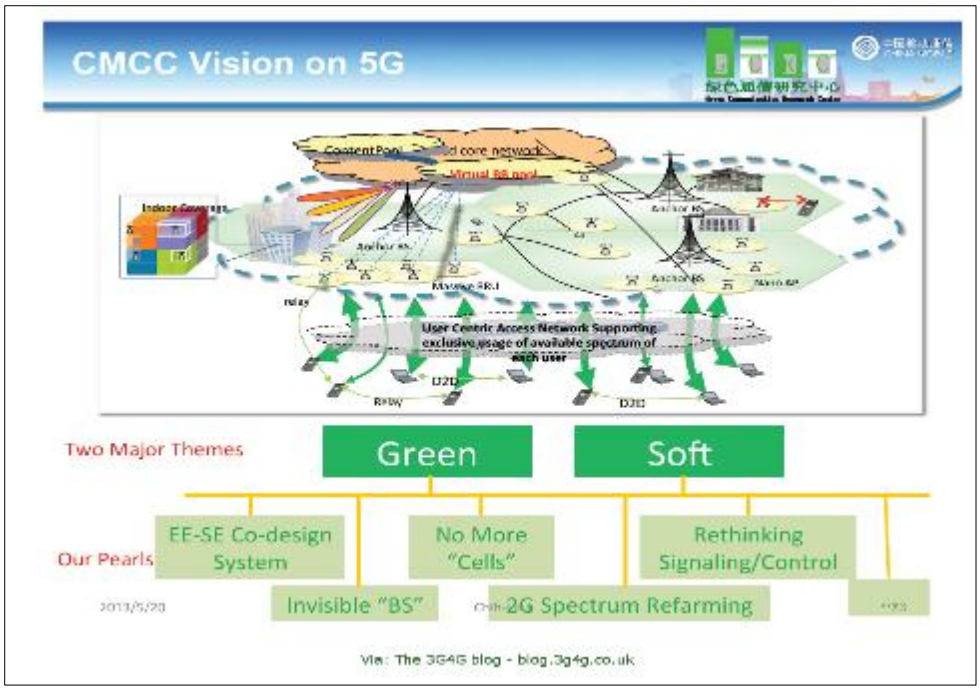

Figure 5 The prevailing perceptions and ideas for 5G mobile networks undermine the importance of green technology [25].

3.4 QoE

Quality of Experience (QoE) describes the user's subjective perception of how well an application or service performs. For example, the video QoE applications depend on the quality of the encoded and delivered video within the screen on which the video is displayed. Providing a form with very low QoE leads to user dissatisfaction, while excessively high QoE unnecessarily removes resources both from the user's battery and from the operator's side. Therefore, a 5G challenge is to support applications and services with an optimal and consistent level of QoE anywhere and anytime [19].

3.5 Issues

Also, A few problems will appear as concerning as the 5G network architecture. One point is how legacy networks will be interconnected and working with the new network architecture. So there is a need for a new network to ensure interoperability. Another problem is to determine the optimal physical implementation of the cloud to achieve performance and cost goals. While pooling resources could lead to savings from concentration, it could also lead to performance bottlenecks, higher delays, and a single point of failure. Besides, the merger could lead to the need for higher capacity and capacity for the central entity to process and transfer the overall traffic, reducing the cost savings achieved by the merger. On the other hand, resource allocation could lead to performance improvements and reduced latency but could be costly due to reduced pooling profits and an increase in the number of data center locations at correspondingly higher operating costs [26] [19].

As shown in Figure 3.4, different deployment options have other implications for the network. Also, support for local unblocking makes network traffic invisible, which affects QoE smart provisioning. Furthermore, the issues featured over the seamless provision of mobility between different types of a locally developed and wide range of technologies with other functionalities need to be addressed to improve the overall QoE quality for end-users. Mechanisms will also be required to support concurrent sessions and smooth session mobility across various access networks to support consistent QoE for end-users. Different network systems will need to be integrated into the 5G network architecture [27] [19].

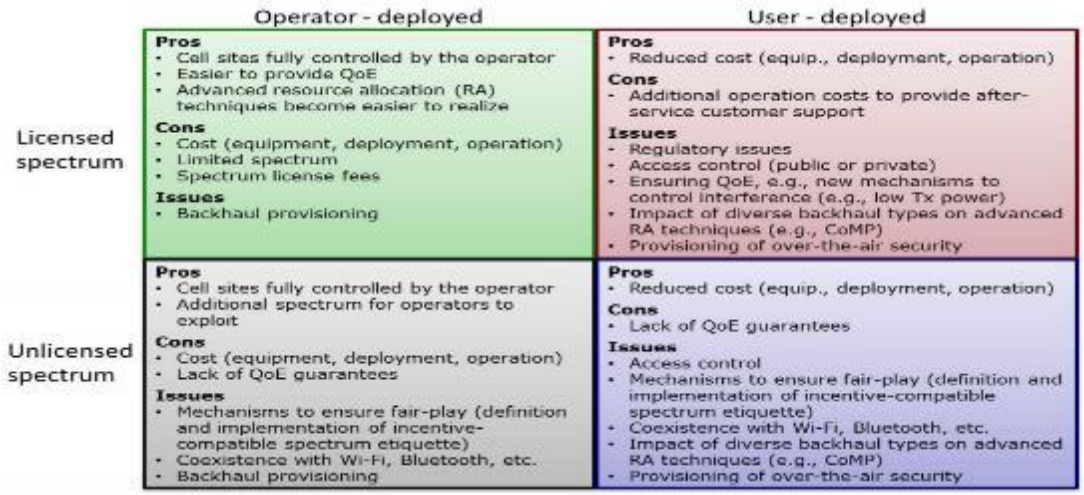

Figure 6 Small cell deployment options and issues [19]

3.6 Big Data Management

The key to managing big data is reliable and efficient data placement. The use of flexibility in data placement policy is to increase energy efficiency in data centers and to use an algorithm that takes into account energy efficiency in addition to the properties of justice and local data. Also, the way resources are allocated to tasks has also attracted a great deal of attention. In this case, it is proposed new design philosophies, techniques, and experiences that provide a further magnetic, agile, and in-depth data analysis. For example, one of the largest ad networks in the world on the Fox Audience Network, using parallel database systems. Therefore, it is recommended to use an innovative data-driven algorithm to reduce energy costs and ensure the thermal reliability of servers [23].

The idea of caching on the edge of wireless networks has also been studied in various projects, including preventive caching for 5G wireless networks. A cache architecture is introduced based on cache users and small base stations. Aspects of storing base stations can save via stochastic geometry tools, where the probability output is a function of the signal to interference plus noise ratio (S.I.N.R.), the base station density, and the storage size. Hence, to optimize the cache allocation, an approximation framework must be based on known system installation, according to succeed an advantage of multicast transmission, along with caching at the base station (Figure 3.5) [28].

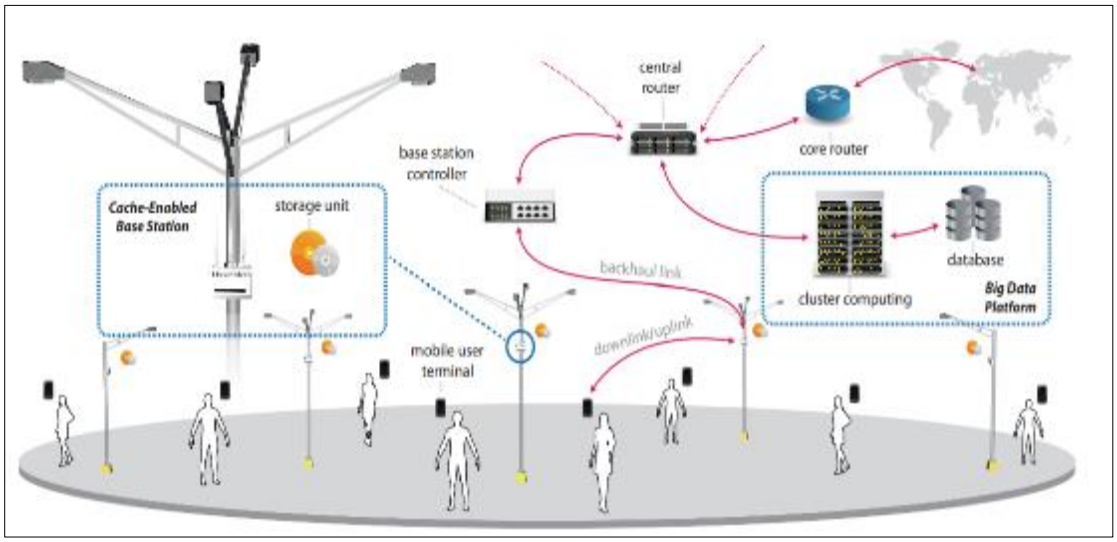

Figure 7 An illustration of the network model A large data platform is responsible for monitoring/forecasting user demand, while cache-enabled base stations store the strategic content provided by the massive data platform [28].

4 Big Data Analytics In 5G

Researchers are currently exploring new analytics techniques in Big Data according to discover new patterns and extract knowledge from the data are collected [15]. For example, big data analytics can provide organizations with the ability to profile and segment customers based on distinct socioeconomic characteristics and increase customer satisfaction and retention levels. Also, Big Data analytics techniques can provide telecom providers with in-depth knowledge of networks before making informed decisions. For example, these analytics techniques can help Telecommunication providers to monitor and analyze various types of data as well as event messages on networks. Important information, like business intelligence, can be extracted from momentary and historical data [15].

Those methods can allow them to make more informed marketing decisions and buy different departments based on their preferences and recognize sales and marketing opportunities. By analyzing emotions in this data, companies can be notified in advance when customers turn against them or move to different products and, as a result, to take action. Also, those technics can track customers' feelings about brands and identify individuals' influences, which can help organizations react to trends and engage in direct marketing [29].

Even organizations that have been using segmentation for many years have begun to develop more sophisticated Big Data methods, such as real-time micro-segmentation of customers, to target promotions and advertising. As a result, Big Data analytics can benefit organizations by allowing better targeted social media marketing, determining and anticipating market sentiments, and analyzing and understanding customer turmoil and other behaviors [29]. Useful information, such as the correlation between user behavior and network traffic, can help providers not only make decisions based on long-term strategies but also to optimize resource allocation to minimize installation and operating costs [15]. All those technics can be very useful in telecommunication companies and provide more personalized and accurate products and services.

4.1 5G Networks Standardizing

Providers are expecting to play a crucial role in standardizing 5G networks [15]. In recent years, many suppliers, operators, universities, research centers, and standardized and regulators have announced the launch of their work to develop a new generation of 5G mobile communications [30]. Standardization must play an essential role in the built-in 5G front-wheel-drive / reversible architecture for many suppliers [31].

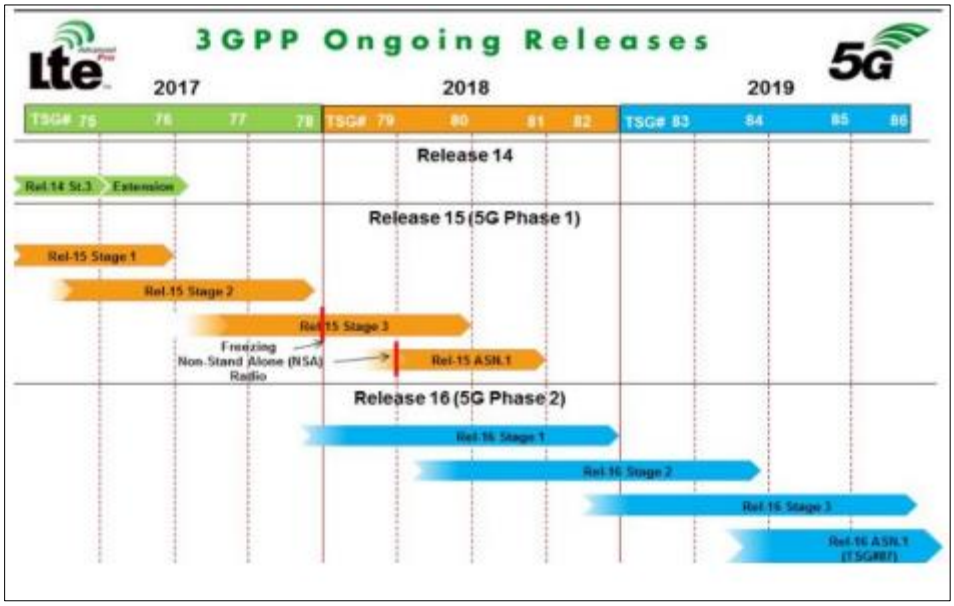

Figure 8 Timeline of 3GPP Releases and 5G Phases [32]

The convergence of different technologies is a crucial aspect of the 5G networks, paves the way for new services and business models. However, these new opportunities come at the cost of increased complexity and stricter or new requirements. The ongoing standardization of the 5G Phase 1 feature (Figure 4.1) introduces some new services, such as mass communication Machine Type (mMTC), improved Mobile Broadband (eMBB), and extraordinarily reliable and Low Latency Communications (uRLLC), which can only partially cover all these new needs. 3GPP introduced 5G Phase

2 to host such advanced services as the eMBB and uRLLC combination in a new category, called Ultra-High-Speed Low Latency Communications (uHSLLC) (Figure 4.2) [32].

Communication networks are currently undergoing a significant evolutionary change to be able to flexibly serve the needs and requirements of a large number of connected users and devices and to enable the operation of all new applications and services in a flexible and programmable manner. Key terms in this context are Internet-of-things, virtualization, softewaralization, and cloud-native [31].

To maintain and run these 5G sliced networks, such as Network Function Virtualization (N.F.V.) and Network-Defined Networking (S.D.N.). Several organizations (ETSI NFV, O.N.F., ETSI MEC, N.G.M.N., 3GPP, IEEE, B.B.F., M.E.F., and others) are working to standardize the architectural frameworks and interfaces required to combine the multitude of elements into a functional system that can be applied [31]. However, a critical challenge is to understand the requirements of using Big Data analytics is to provide personalized Quality Experience services to users and allow extremely efficient use of resources on 5G networks [15].

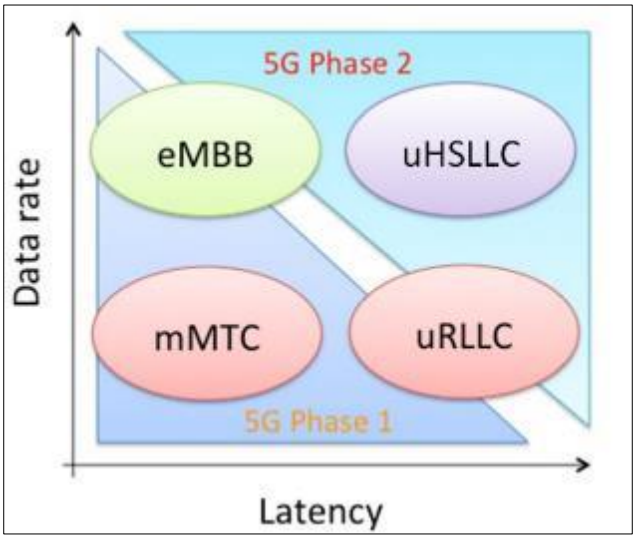

Figure 9 5G Phase 1 and Phase 2 primary services and key system KPIs [32]

4.2 5G Mobile Optimization

In addition to mobile phone users, providers can also take advantage of big data. They can quickly obtain vast volumes of data, which are generated by mobile devices belonging to their customers, as well as by various network elements in their networks. All data can be explored together to improve the efficiency of network operation [15]. Real-time reactions make better decisions, not only for network optimization but also for the quality of the user experience. Big data analysis has many impressive advantages. However, at this point, its analysis tools are generally complicated and require planning, not as friendly as traditional ones [33].

Datasets from mobile networks have distinctive features, including a wide variety of recording scales, time resolution, and various types of data. For example, after recording and retrieving data from mobile networks, we can extract location, mobility, proximity, and application usage information. However, unlike traditional big data on computer networks, wireless extensive mobile data has their typical statistical characteristics [34].

4.2.1 Spatial-temporal distribution

Data on a mobile network can be divided into different details (for example, 5 minutes, 15 minutes, one hour, or even one day). Similarly, in rural areas, mobile data can be studied from an entire city, a common area, a base station (B.S.), or even a mobile phone. The movement of mobile phones in different locations, such as the stadium, the campus, and the densely populated areas, present different patterns at other times. The full use of such features has already been imposed on mobile networks [34].

4.2.2 Data collection property

Typically, mobile feature data concentration characteristics are more important for network performance optimization. Concentration capabilities will be further used in predictive modeling to improve network performance [34].

4.2.3 Social correlations

In the surrounding areas of wireless cellular networks, due to the social nature and habits of users, it is observed that they tend to show similar habits, behaviors, and mobility rules. For example, social correlations of users lead to a larger scale of correlation of movement in both temporal and spatial areas, such as self-correlation and interrelationship. Thus, social correlations are a global phenomenon in mobile data. So, a provider needs to take full advantage of the mobile network [34].

The challenges are not only due to the massive volume of data but also to the non-homogeneous structure that is often associated with incomplete and ambiguous information. Therefore, it is imperative to have a good knowledge of the unique features of Big Data on mobile networks, which is vital for optimizing 5G mobile networks. With the recent developments in data analysis, the optimization of Big Data mobile networks has attracted intense efforts from researchers worldwide [15].

4.3 Big Data Driven Mobile Network Optimization and Vehicular Communications Networks

Researchers are trying to find a general framework to support a variety of Big Data-Driven (BDD) mobile network. Some frameworks allow engineers to use data from both the network and users when optimizing networks, instead of user-only data. BDD programs can improve QoE performance and utilize the entire network. The characteristics of the data collected by users and operators are also investigated, followed by the study of specific analytical data systems. Hence, it is intended to provide a case study of optimizing the BDD mobile network for further validation [15].

An excellent example of BDD applications is the Vehicular communications networks (V.A.N.E.T.s). A smart solution to support Big Data V.A.N.E.T.s is the widespread cellular network. As the 4G LTE network struggles to keep the ever-increasing volume of data and emerging mobile telephony services with different QoS requirements, 5G networks create a way to solve problems [35]. According to optimize the performance, private short-range communication (D.S.R.C.) can provide actionable real-time information exchange between vehicles without the need for penetrating road communication infrastructure. However, both D.S.R.C. and mobile networks have their limitations when used in-vehicle environments. On the other hand, although mobile networks can provide broad geographical coverage, they cannot adequately support the exchange of real-time information for local areas [36].

A solution about that can be provided from heterogeneous networks (we can see them more below). Those networks not only can give extensive area coverage for all vehicles on large-scale networks but also support the delivery of real-time security messages in local areas to reduce traffic accidents. Therefore, heterogeneous network vehicles (HetVNETs), which incorporate D.S.R.C. with cellular networks, may support the communication requirements of the Intelligent Transport System (ITS) [36].

5 5G Heterogeneous Network (HetNet)

The mass adoption of smartphones, mobile broadband modems, tablets, and mobile data applications has been overwhelmingly wireless. Operators bend under the pressure and cost of continuously adding capacity and improving coverage while maximizing the use of the existing components of their range [37]. Advanced radio access technologies, and all Internet Protocols (I.P.s), open internet network architectures must evolve smoothly from 4G systems [38].

Those critical discoveries in the basic zone and radio frequency (R.F.) are required to enable computational intensity and adaptation to new air interfaces in 5G systems. Meanwhile, findings are needed in the integrated access node and heterogeneous convergence to allow too dense radio nodes to operate efficiently. Besides, software-defined wireless interface technologies should be seamlessly integrated into the architecture of the 5G R.A.N. [38].

Hence, for this reason, the considerable growth of small cells is necessary to support the required capacity and unload traffic on existing unlicensed access technologies. So, creating the concept of Heterogeneous Networks or HetNet, where macrophages, small cells, and Wi-Fi access points work together (Figure 5.1)[37]. As we can see from the bibliography, HetNet is a widely accepted solution to improve the overall network performance. The Heterogeneous network aims to combine all wireless transmission protocols, spectrum bands, and standards under a single network control plane.

HetNets introduce a high number of small cells, such as pico-cells and femtocells, as well as Wi-Fi hot spots to increase bandwidth per one section and provide higher throughputs for the end-users [39]. So, HetNets usually consists of two layers, i.e., the macrocells and small cell layers, where the former provides mobility, while the latter enhances coverage and ability [15]. All these are necessary for better management of the capacity and user experience, so as we can see, is critical the role of small cells [37].

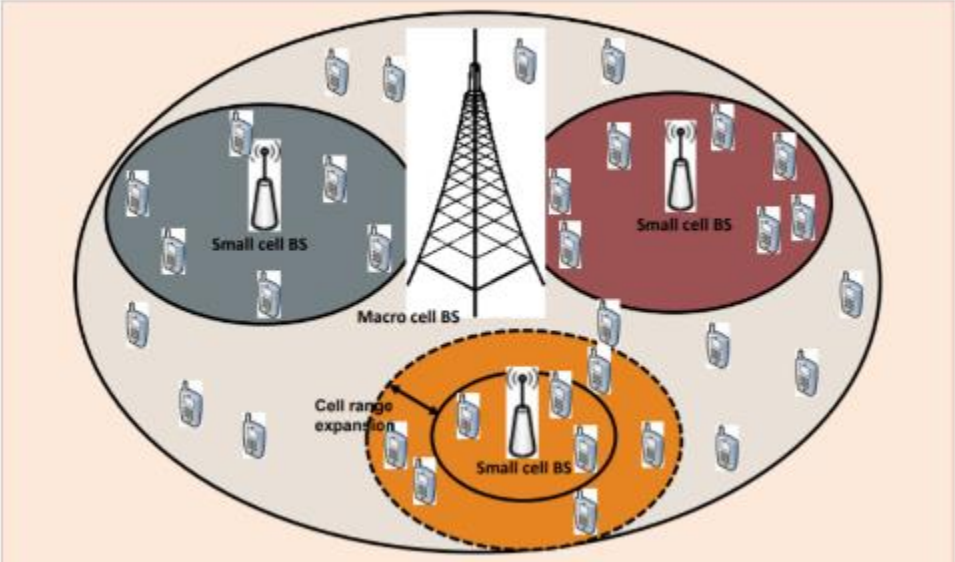

Figure 10 Small cell range expansion technology for the heterogeneous network [39]

Wireless access infrastructures based on the computational cloud will provide necessary resource processing, delayed storage, and high network capacity where needed. Some other advanced technologies, such as the cloud radio access network (C-RAN) access network and the heterogeneous network, have been identified as possible 5G solutions. In C-RANs, a considerable number of low-cost remote radio heads (R.R.H.) are randomly developing and connecting to the base-band unit (B.B.U.) via the fronthaul connections. Taking advantage of the collection of resources and the gain of statistical multiplexing, C-RAN is much more efficient both in terms of energy and cost, because it is unnecessary to dimension the computational resource of each traditional B.S. depending on the individual peak load. However, fronthaul limitations have a significant impact on the deterioration of C-RAN performance, and the scale size of R.R.H.s accessing the same B.B.U. group is limited. It could not be too large due to the complexity of the application [38]. Hence, HetNet is very suitable for providers to make the use of spectra more efficient in densely populated areas as they move in the direction of 5G development [15].

To increase the capacity of cellular networks in densely populated areas with high traffic requirements, the low-power node (L.P.N.) serving the high-capacity "data only" service is recognized as one of the critical elements of HetNets. L.P.N.s have only control levels, while wireless control channels and especially L.P.N. Reference signals can be fully converted to macro base station (M.B.S.) stations. Therefore, it is necessary to suppress interference through advanced signal processing techniques to be able to earn the potential full profits of HetNets. Such as adopting advanced multi-point coordinated transmission technology (CoMP) to suppress both in-house and in-house and transit interventions [38].

A practical HetNet setting that includes two key features:

- Users are concentrated in certain areas in the cell, thereby forming hotspots.
- Small cells are deployed in the vicinity of some hotspots, as is typically the case in current capacity-driven deployments [40].

Assuming that the size of the hotspot is much smaller than the macrocell radius, the first feature facilitates the separation of the vacuum, under which a vast M.I.M.O. A long cell can focus its energy in the direction of the hotspots it serves while allowing the simultaneous transmission of small cells. The second feature allows low complexity cell coordination strategies, simplifying the cell selection process [40].

5.1 HetNets Issues

A multi-layered network architecture like HeNet can allow for high coverage and high capacity, thus providing users with enhanced QoE. However, while using small cells can improve the ability of the entire network, it is not able to support the adaptation of various network resources according to the traffic characteristics that change over time [15]. Despite all the advantages provided by HetNets, some issues limit the network [39].

- First, the problem of interference is a significant challenge due to the shorter distances between cellular cells.
- Second, the complexity of resource allocation limits the number of network levels.
- Third, many small cells lead to high-cost reconstruction because it is difficult to provide reliable reversal in a dense urban environment [39].

Besides, it is necessary to consider the density of users in each area, because the heterogeneous network can provide a significant advantage only in areas with high user density. Because user density is not uniform and changes frequently, it is challenging to design a heterogeneous network topology [39].

Network densification allows to reduce path loss between base stations (B.S.) transmitter and user equipment (U.E.) receiver and increase bandwidth per single U.E. that allows providing higher throughputs and quality of service. The denser deployment also allows using of low power and cheaper B.S.s (access points), reducing capital and operational expenditures comparing to intense and expensive macro B.S.s. However, such a dense network deployment brings additional challenges to deployment and management [39].

The difference in transmission power leads to the handover problem. Sometimes the user equipment can be close enough to pico B.S. and macro B.S. that will decrease power-based handover [39]. Moreover, in this case, some small cells may be underutilized to reduce the advantages of HetNet. Although small cells usually are optimally placed, they also may be underutilized or overloaded due to the various traffic demands. In this case, cell range expansion techniques may be applied to decrease or increase the coverage area of small cells [39].

According to improve the operational efficiency of network infrastructure in different environments, providers are encouraged to adapt to network traffic requirements and improve resource allocation efficiency by using big data and database analytics [15].

As shown in Figure 5.2, the proposed BDD network optimization framework includes:

- Large data collection
- Storage management
- Data analysis
- Network optimization [15].

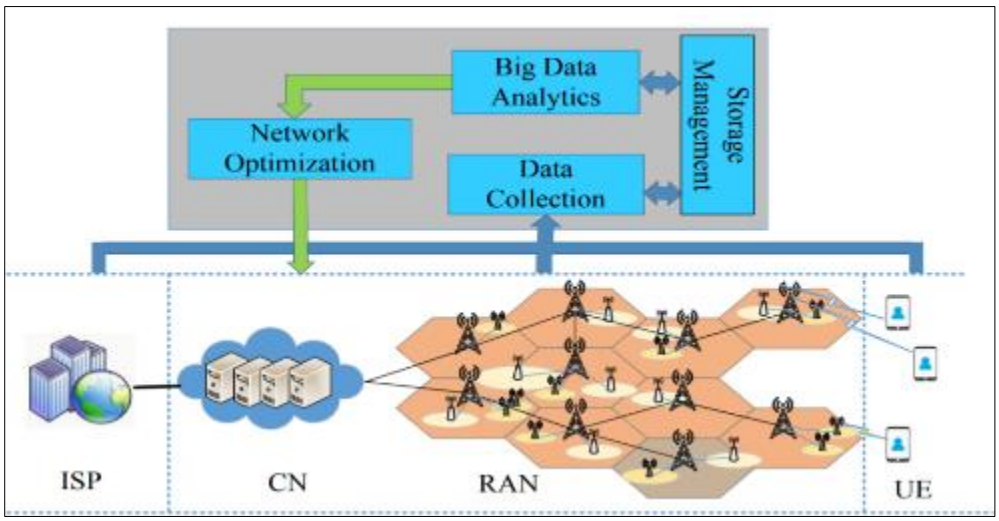

Figure 11 Illustration of the proposed BDD network optimization framework [15]

5.2 HetNets With Mobile Cloud Computing

A promising solution to the increasing data and computational demands from mobile users is both the HetNets and mobile cloud computing (MCC) [41]. Many methods have been proposed to meet the needs of mobile device resources through the integration of MCC and HetNet technologies. Researchers introduce the R.A.N. network as a service, which utilizes the cloud system to implement a flexible, functional separation of R.A.N. functions for optimized use of resources on dense networks. Hence for portable cloud computing, dynamic resource allocation, and parallel execution in the cloud system [41].

We must mention the importance of resource management in 5G HetNet. In a cloud-based approach to network resource management can be applied by coordinating small cell transmissions. It is also a method that combines encryption and compression techniques to create a stable computer environment in which users can store their data safely and efficiently after downloading it to the cloud. Additionally, a two-level authentication scheme is also can be used to ensure that only legitimate users can access the data in the cloud [41].

As we can see from studies, cloud computing download techniques are related to technical issues that need to be considered when designing efficient transmission of extensive data in cloud storage. It is serious about investigating data storage and the same processing in the cloud for mobile. Unlike previous methods that focus more on data security and data delays in cloud storage, some approaches focus on reducing the data rate while transmitting large data files to/from the cloud storage center [41].

The system model of the network infrastructure that integrates HetNet and MCC technologies along with a cloudlet development. In Figure 5.3, the network consists of two B.S. macros, three Femto BS with overlapping coverage areas. Each BS has access to the public cloud system via a backhaul connection, and there is a cloudlet associated with each BS [41].

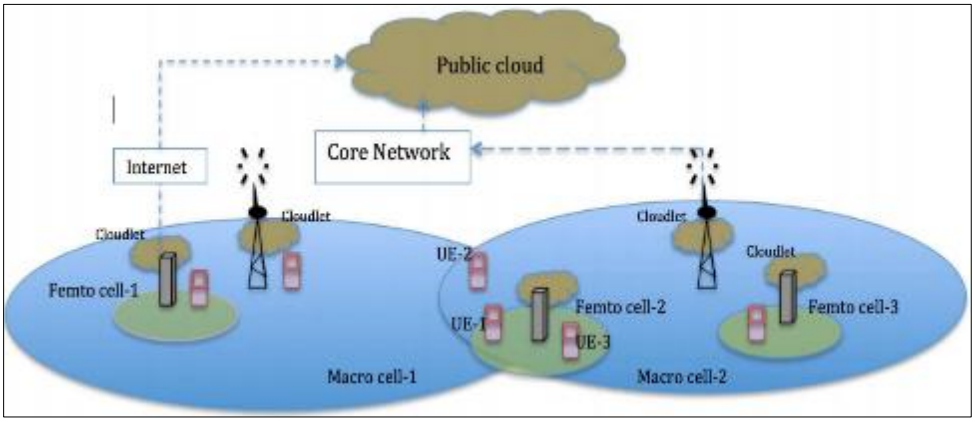

Figure 12 Typical LTE HetNet architecture with macro and femto base stations [41]

5.3 Cloud-Based Architecture

In Figure 5.4 illustrates the architecture of cloud-based wireless networks. The structure consists of a mobile cloud, a cloud-based radio access network (Cloud-RAN), a remodeling network, and data centers [42].

- Mobile Cloud: Actually, it consists mainly of two parts, the local and remote cloud. The primary function is to relocate inner computing work from a portable device in the cloud, to overcome the limitation of computing mobile device power, storage capacity, and battery life [42].
- Cloud RAN: The main functions include achieving shared resource utilization, dynamic resource allocation by virtualizing network access devices, using collaboration functionality, improving spectral efficiency, while reducing energy consumption, achieving low-cost services [42].
- Configurable network: It can configure the network through network virtualization and open network control. The goal is to bring together the management and control of network resources and create a flexible network architecture to enhance the user experience [42].
- Big data center: Not only can it provide mobile phone terminals and storage capacity, improving the user experience, but it also analyzes network needs in different network traffic scenarios. Optimizes resource allocation and improves network utilization. Along with the continued progress in cloud computing, many

researchers apply cloud computing to mobile networks. Depending on the location of the data center in the cloud can be divided into three types, such as remote cloud, local cloud, and hybrid cloud [42].

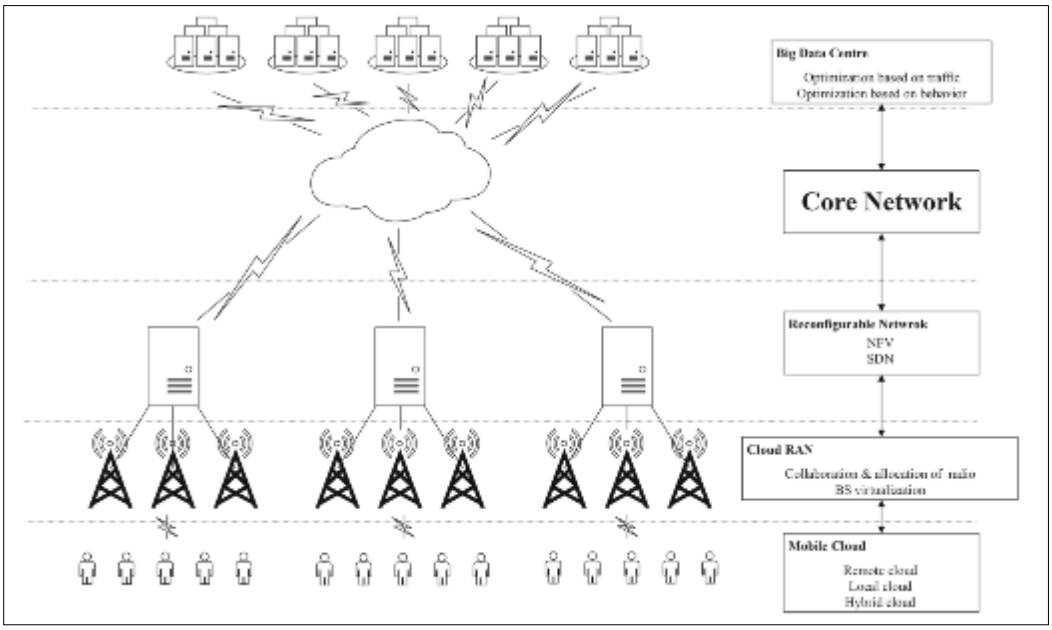

Figure 13 Cloud-based wireless network architecture [42]

5.4 Cloud RAN

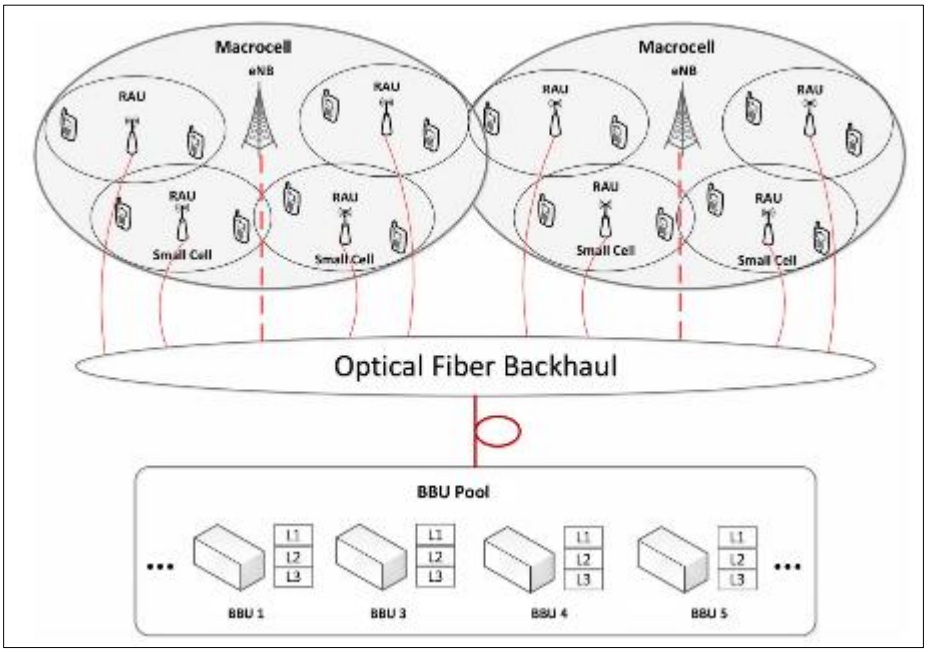

Figure 14 Cloud RAN architecture [42]

As shown in Figure 5.5, the baseband unit (BBU) is separated from the radio access unit (RAU). It forms a pool BBU to achieve centralized digital processing and management. Cloud RAN disconnects the inner connection between RAU and BBU. Each sending or receiving of a signal from RAU is completed on a virtual basis, and BBU assigns the processing capacity of the virtual base station in real-time, which can achieve optimal global use of natural resources. Therefore, cloud research RAN focuses mainly on the following aspects:

- Dynamic resource allocation and collaborative wireless processing
- Base station visualization [42].

5.5 HetNets With Big Data Collection and Storage

Data collecting requires an extensive effort and underlying investment, which often means that the collected datasets are only available in a limited way. The verification of the findings can be difficult if access to the data is minimal. Protecting users' privacy behind the information is the main reason for Big Data's access and use restrictions [43].

Big Data collection can be achieved from the user equipment, Radio Access Networks, Core Network (C.N.), and Internet Service Providers (I.S.P.s). Events that occur in U.E.s are collected either through user applications or through a control signal. R.A.N.'s evolving NodeB (eNB) collects data on the hive, which including wireless signaling and instantaneous measurement reports. Meanwhile, providers have vast amounts of data about users and services in C.N. When the cell size becomes smaller in HetNets, the number of eNBs increases. As this trend continues, network data may explode and place a heavy burden on data collection. Also, Big Data storage infrastructure must have a scalable capacity as well as scalable performance. Thus, storage management must be efficient and straightforward, so that large data storage and sorting can be easily achieved [15]. Below we can see a Collecting Data Architecture and some Data Storage Protocols.

5.6 Reduced Variable Neighborhood Search (R.V.N.S.) Queue Collecting Data Architecture

As shown in Figure 5.6, interest-based reduced queue architecture (I.R.Q.A.) can be divided into three levels: data collection layer (D.G.L.), data integration layer (D.I.L.), and data analysis layer (D.A.L.) [44].

D.G.L. undertakes the collection and storage of relevant data collected. It also checks the status of mobile nodes and reports interoperable nodes with a fault tolerance mechanism. Robust data enter DIL, which is used to control the effectiveness of the data. Also, the data is divided into different levels in the DIL. Finally, D.A.L. places data in an R.V.N.S. queue and reports valuable data. Data Gathering Layer (D.G.L.) in the proposed architecture, the data collected from developed mobile phone nodes will be converted to a local server and wait for processing. Usually, on large mobile data networks, nodes are portable and carried by individuals, transport, or movement. The continuous movement of these nodes serves as a basis for matching interests. Each node is assigned a special index during preparation as recognition in the system. After collecting data, the nodes pack the data into packets using node indexes to identify parcels [44].

On the other hand, a local server can retrieve the source node of a packet by searching the node index in that packet. Based on this timing, a server can sort the data collected in a time sequence—initially, packet transfer nodes according to the original routing algorithm. Meanwhile, due to node interruptions, a server cannot access complete data [44].

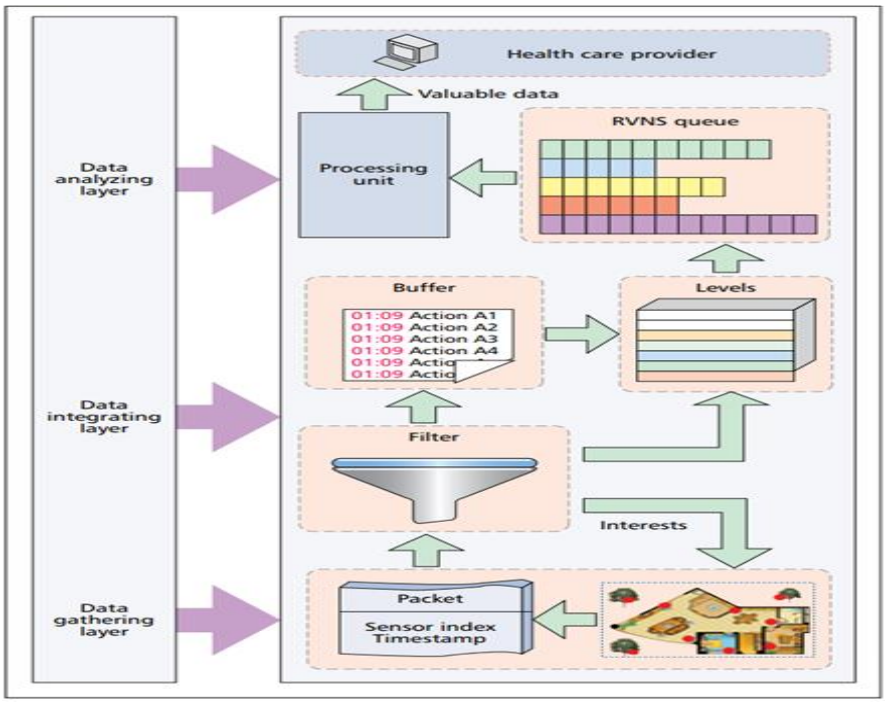

Figure 15 Structure of I.R.Q.A. [44].

Researchers propose a fault-tolerant mechanism designed using the thinking of matching interests, as shown in Figure 5.7. Lost packages can be found concerning node indexes when sorting packets based on time marking. An index is then generated and transmitted for each lost packet according to the timekeeping, and node index found-nodes related to search packets between their temporary memories as soon as the interests are taken. Instead, the node sends again the corresponding packets that match interests. The reason the interest matching mechanism works are the existence of mobile nodes. The routing algorithm sets a routing path for packet transmission, but the server does not receive packets due to routing interruption. During this time, the routing changes because the nodes are continually moving [44].

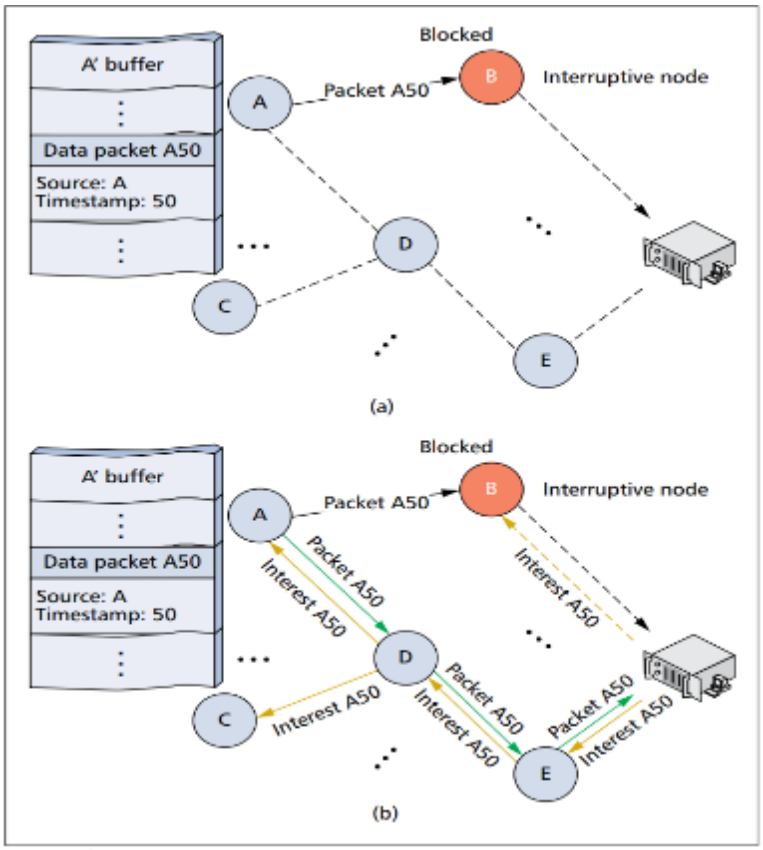

Figure 16 Fault-tolerant mechanism: a) original transmission process, b) fault-tolerant transmission process [44].

5.7 Data Storage Protocols

Researchers propose two network models. The first considers a tree topology rooted in the sink and a subset of them selected as storage nodes responsible for storing data collected by their offspring on the tree. In the second model, tree topology is constructed after the planned development of the storage nodes. The positions of which are obtained by a linear programming optimization. However, node failures are not considered in both models, leading to the loss of all storage nodes [45].

Researchers also suggest a light random participation service for ad hoc networks called RaWMS. The protocol provides each node with a partially uniformly selected view of the network nodes. In RaWMS, each data generator node starts a maximum reverse RW, whose message carries the identifier and the node data. The last node in RW appears as if it was randomly selected from all network nodes and will be responsible for storing the data transferred from RW. The authors prove that when RW is completed, each node will have a uniform random view with data from the nodes in the network. In other words, RaWMS evenly distributes network data to the sensor nodes. However, the results of RaWMS for uniform distribution of the monitored data across the network are quite encouraging in terms of data collection efficiency [45].

Finally, another data storage protocol proposed by researchers is ProFlex. It is a distributed data storage protocol for heterogeneous wireless sensor networks with portable sinks. At the end of the three phases, ProFlex guarantees that a node in the total node storage will store a quantity of data proportional to the significant factor. For example, if all nodes in the node storage set have the same element of importance, then ProFlex guarantees uniform distribution of network

data between the nodes in the total node storage [45]. Once the data is collected and stored, another major challenge for providers is how to process such vast volumes of data [15].

6 Big Data Analysis Techniques and Knowledge Extraction

The collected data are multi-source, heterogeneous, real-time, and bulky. For this reason, data analysis techniques and knowledge extraction techniques are required to process the data and turn it into practical knowledge. As a result, this knowledge can design adaptive schemes for network optimization [15].

6.1 Data Analytical Tools

Data analysis allows providers to manage networks and provide services to customers in a systematic way. Not only network measurements but also the application/service status for each area can be monitored and analyzed over time. BDD network optimization functions can analyze Big Data to identify problems and decide what/how to optimize the appropriate level, e.g., user, cell, or service [15]. The ability to acquire, analyze, and exploit mobile phone traffic characteristics can be achieved with specially designed learning units (LUs) installed in both Base Station (BS) and Central Unit (CU). Their main activating factors are built-in analytical data algorithms. Hence, some algorithms are commonly used for wireless traffic analysis. Their main applications in wireless communications are classified as follows and are summarized (Table 6.1), we can see a brief description of them [46].

Table 1 Summary of standard wireless Big Data analytic tools and example applications [46]

Subjects	*Models/algorithms*	*Example wireless applications*
Statistical modeling	Markov models, time series, geometric models, Kalman filters	mobility prediction, resource provision, device association/handoff prediction
Data mining	pattern matching, text compression, clustering, dimension reduction	mobility prediction, social group clustering, context-aware processing, cache management, user profile management
Machine learning	classification algorithms, neural network, regression analysis,	context identification, traffic prediction, fitting trajectory length, user location and the channel holding time
	dimension reduction algorithms: PCA, PARAFAC, Tucker3	user data compression/storage, traffic feature extraction, blind multiuser detection
	Q-learning	handoff and admission controls
	primal/dual decomposition, ADMM	distributed routing/rate control and wireless resource allocation
	online convex optimization, stochastic learning	on-line mobility predictions, handoffs, and resource provisioning
	active learning, deep learning	incomplete/complex mobile data processing

6.2 Stochastic Modeling

Stochastic modeling methods use probabilistic models to capture the exact characteristics and dynamics of data circulation. Stochastic models commonly used include the Markov-K series model, the Markov hidden model, the geometric model, the time series, the linear or non-linear random dynamical systems, etc. For example, Markov models and Kalman filters are widely used to predict mobility and service requirements. The data collected by users is often used to estimate the parameters of meditative models [46].

6.3 Data Mining

Data mining focuses on the exploitation of indirect structures in mobile datasets. Also, considering the mobility prediction problem. For example, the mobility pattern of an individual user could be extracted and discovered by finding the most common orbital sections in the mobility record file. Clustering is another useful technique for identifying different patterns in datasets. It is used in a mobile-based environment, where the user's content and the behavior of a mobile user. For example, sleep and work are recognized by wireless detection data services related to the environment [46].

6.4 Large-Scale Data Analysis

Wireless Big Data pose many challenges to the aforementioned conventional data analysis methods. Such as due to their large volume, large size, data inequalities, and complex characteristics. To improve signal processing efficiency, one can

combine the following techniques to reduce complexity with conventional data analysis tools for large-scale data processing [46]. Below we can see some large scale data analysis methods.

- Distributed optimization algorithms, such as primary or double decomposition and alternative propagation method (ADMM), are beneficial for disconnecting large-scale statistical learning problems in small sub-problems for parallel calculations to be relieved. The bandwidth presses on the fronthaul or backhaul links [46].
- Dimension reduction methods are useful for reducing the volume of data to be processed while recording the critical features of Big Data. Also, tensor decomposition methods are popular in mobile data processing, which seeks to represent several high-order multiple lines (tensor) as a linear combination of low-order external tensor products. In this way, the demand for hardware and the cost of mobile phone storing data arrays could be reduced [46].
- Other advanced learning methods could be used to handle incomplete or complex datasets. Interesting examples include active learning, which deals with partially marked datasets. Also, deep learning is useful to model problematic behaviors contained in a set of data [46].

The control functions in R.A.N then apply the optimization measures based on the optimization results. Also, it can be optimized at the user level. For users who are close to the same cell, optimization can be customized for each user, again depending on their service category. Besides, BDD network optimization functions can predict traffic variations either in a local area or through network coverage and ultimately improve network and user performance [15].

6.5 Big Data on Mobile Networks

The data can be divided into two types: the users and the data from the network providers. A comprehensive analysis of two kinds of data can provide valuable information, which can be used by providers for network optimization. Telecommunication providers can analyze data to plan the network, to allocate the spectrum, to manage the resources, and so on [15].

6.6 Mobile User Data

The data collected by UE is primarily related to the user's profile and behavior, which provides much information about users, such as their location, mobility, and personal behavior or communication pattern [15]. A Big data-driven framework, such as Device to Device (D2D) (Figure 6.1), depends on exploring multidimensional capabilities derived from the problematic behaviors of mobile users. For example, users' geographical differences, social relationships, interests' similarity, and mobility models, complex content properties, and others. [47].

In particular, sharing activities between mobile users can form a connected graph so that a Big Data framework can apply advanced theories and algorithms based on the complex idea of the network. However, integrating a Big Data framework has not been implemented and analyzed for significant volumes of real data collected from practical D2D sharing applications on mobile networks. Although online virtual community social networking sites, people with a common interest, such as a particular hobby or activity, can interact and socialize with each other. Also, have been thoroughly explored, content-based interactions can be further socialized and linked to a promising future with the help of Big Data techniques [47].

At this point, we must mention, with the rapid expansion of mobile networks and the massive increase in developed smart mobile devices, excessive amounts of data are being created by applications installed on users' mobile devices. Application-level data has become one of the primary sources of big mobile data [15]. However, applications (i.e., instant messaging (IM), web browsing, and video) in cellular networks are affected by different factors and significantly different from those in wired networks. Therefore, instead of directly applying the results generated by the wired network traffic, it is necessary to reconsider the relevant traffic characteristics in mobile networks and check the accuracy of the predictability of traffic at the application level [48].

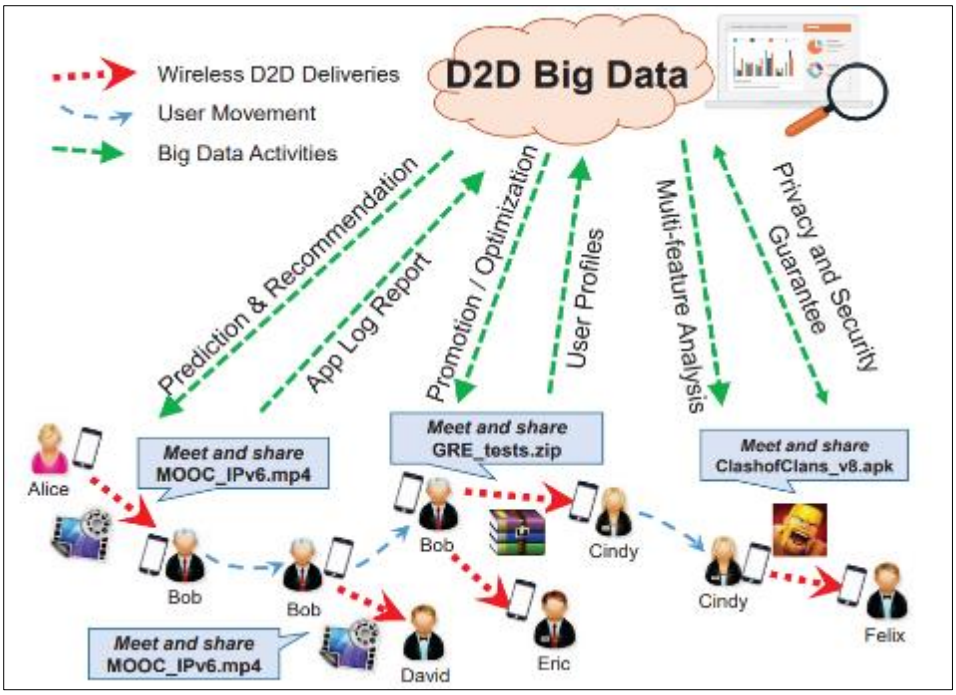

Figure 17 Illustration of wireless D2D deliveries integrated and supported with D2D Big Data [47]

6.7 Mobile Network Operator Data

The data collected by the operators comes mainly from CN and RAN. CN has plenty of data carrier or service data regarding, e.g., network performance information, successful calls, and usage index per application [15]. A mobile network operator (MNO) records many incidents on its primary network interfaces by installing network detectors. Each incident recorded on the primary network has the following characteristics [49].

- Timing: specifies the exact time of the event
- ID: identifies the subscriber
- Cell ID: identifies the cell to which the subscriber is connected
- LAC: specifies the location area to which the subscriber has been assigned
- Event: specifies the type of recorded event

Some of these events are closely related to the mobility of subscribers. The most critical events in this regard are the creation and termination of calls, anonymous handover (HO), and location area updates (LAU). These events may be related to a specific subscriber ID and the cell where the event occurred. In the case of an HO event, this is the cell to which the subscriber is delivered (target cell). In the case of an LAU event, this is the target cell in the recently assigned location area. A case of a sequence of events related to a specific subscriber recorded on the central network is shown in Table 62. The first event describes a call installation, followed by three HO events and a call termination event. Using the position of the base station involved in conjunction with the configuration of their antenna (i.e., the angles of electrical vision), an approximate part of the subscriber at the time of the event can be determined [49].

Table 2 Events captured in the core network of the mobile network operator A1 [49].

Timestamp	ID	Cell ID	LAC	Event
1327678833	4	51520	5501	Call establishment
1327678854	4	31510	5501	Handover update
1327678935	4	31210	5502	Handover update
1327678949	4	40510	5502	Handover update
1327679027	4	40510	5502	Call termination

On the other hand, there is a large amount of data in RAN, including cellular information. For example, eNB configuration information, resource status information, interference information, transfer reports, mobility information, error status,

use of links, and call drop ratio. Also, RRC messages for the installation and transmission of connection and radio signal measurements. For example, we received signal strength power, quality of received reference signal, and others. The key features of these two types of data are summarized below in Table 6.3 [15].

Table 3 Key features of the user and operator data [15].

Feature	User data	Operator data
Objective/Subjective	• Highly influenced by the subjective feelings or personal preferences	• Measured by the network objectively without involving human factors
(Non)-structured	• Various data formats including the semi-structured and non-structured data (e.g., locations, logs, and sensor data)	• Mainly structured data generated according to specific given protocols
Privacy	• High privacy is required since users are not willing to disclose their personal information	• Usually internal use for network operators without sharing with others
Energy limitation	• Data accuracy constrained by device energy consumption • Accuracy adaptively controlled to save energy	• No energy limitation for main-powered network devices
Redundancy	• High correlation and redundancy in the event of a large number of users located in popular locations during a specific period of time	• Usually high correlation and redundancy because data are coherently processed across the different layers of the network
Distribution	• Usually fragmentary and discontinued in time and space	• Usually periodical and uniform distribution in time
Reliability	• Low reliability due to changing user numbers and locations. •Pre-processing is needed to filter noise and maintain data integrity	• Usually high reliability because data are mostly from signaling and control information in networks • Instable due to varying dynamics, heterogeneity and the large scale of the networks
Controllability	• Difficult to control in terms of data rates, sizes, collecting moments and so on	• Easily collectable by the operators through specific network interfaces and measurement devices

6.8 Mobile Network Collection Data

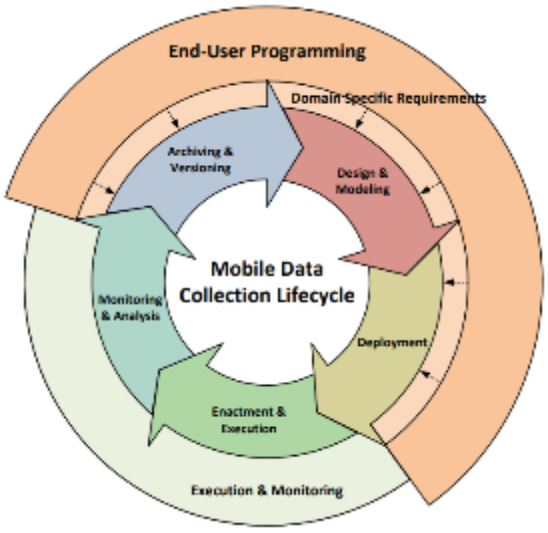

Figure 18 Mobile Data Collection Lifecycle [50]

Figure 6.2 shows the life cycle of the five-phase mobile data collection. These provide the basis for empowering industry experts to carry out advanced mobile data collection applications themselves. In the design and modeling phase, mobile data collection tools include end-users create complex navigation logic. The development phase allows the robust development of the applications developed in smart portable devices. In the activate and run stage, multiple displays of the realized data collection media on smart mobile devices can be created and executed. During the monitoring and

analysis phase, the collected data is analyzed in real-time on the smart mobile device and the backend system. Finally, the archiving and versioning phase provides advanced techniques for managing the circulation cycles of mobile application services and their issuance [50].

Once data is collected from different sources, the next challenge is to use those data in the best way. Data generated from all sources must be processed and converted into knowledge that can be used after that can be used to design adaptive algorithms and optimization strategies to improve network performance. Advanced data collection and analysis techniques are essential for network optimization [15].

6.9 Big Data Analytical Systems on Mobile Networks

In mobile networks, unprocessed data collection is the first step in a Big Data analysis. For example, ISPs can collect data from mobile phone users who share or download information related to their mobility [15]. As we see from the above, a good example is the D2D Big Data framework (Figure 6.3). In this framework, there are two main conceptual elements, the introduction of large-scale data sets, including user profiles, statistical content, activity log files, motion records, and others. While users move and share content files, the system performs flow and batch-based processing. Prediction and suggestions are made with learning-based tools that include machine learning, unsupervised learning, and collaborative group proposal algorithms for content promotion and friendship suggestions. Meanwhile, detailed data and findings can further contribute to marketing and mobile application development strategies [47].

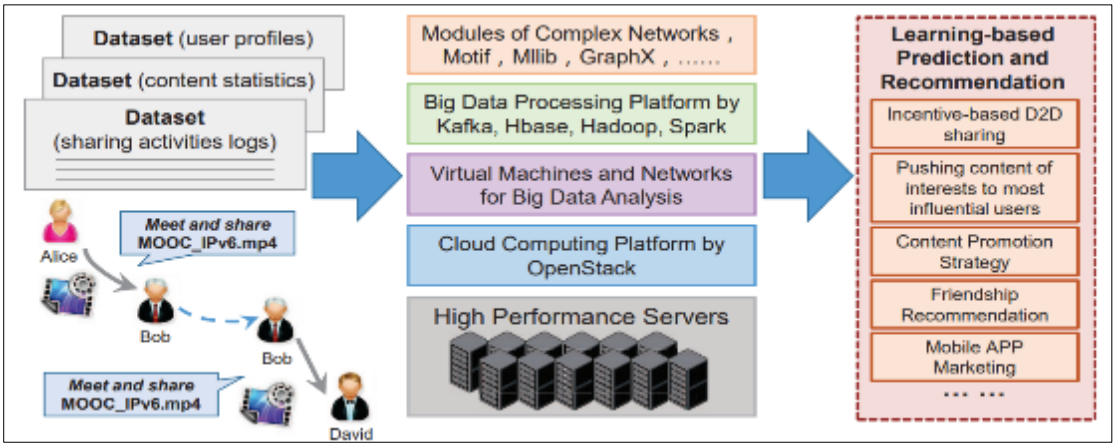

Figure 19 The framework of the D2D Big Data system and the related processing workflow [47]

6.10 Mobile Collection Data Issues

However, user position information may not be possible if mobile phone users disable location on their mobile devices. Alternatively, location information obtained directly from eNBs may not be accurate due to the inaccuracy of tracking techniques. Location errors and environmental interference are often significant barriers to Big Data usability [15]. The lack of availability of such data has limited many research and application activities. The different challenges related to the availability of sufficient data can be divided into several categories, as summarized below [51].

.1 Spatial and temporal analysis: Many geolocation processes are very spatial and temporal variables, but most existing data collection techniques are not able to capture this variant adequately.
.2 Cost: Traditional data collection tools are expensive, limiting the amount of data collected within the limits of available resources.
.3 Accessibility: Many sites where data is required are difficult to access physically, or the services needed to manage data are not available.
.4 Availability: In many cases, real-time information is needed. Therefore, sometimes data collection and transmission are not able to access data when needed.
.5 Uncertainty: There can be significant uncertainty about the quality of the data provided.
.6 Dimension: Collecting the different types of data required for application areas that need a higher degree of social interaction can be a challenge [51].

Also, the battery of the user's device can be depleted, and therefore the desired information cannot be collected at specific times. To this end, data mining, filtering, and export techniques are being developed to eliminate interference

or useless data, which belongs to the so-called ranking schemes with prone errors. However, it is still a big challenge to get useful information from incomplete, unnecessary, and uncertain Big Data [15].

Researchers propose a scenario in which two parties holding confidential databases want to run a data mining algorithm in joining their databases without revealing unnecessary information. Their proposal is based on the need to protect both the privileged information and its ability to be used for research or other purposes. However, data mining algorithms are usually complex, and, besides, the input usually consists of massive datasets [52].

7 Data Mining Techniques in Big Data

Data mining is a recent field that connects the three worlds. Databases, Artificial Intelligence, and Statistics. The information age has allowed many organizations to collect large volumes of data. However, the functionality of this data is negligible if "significant information" or "knowledge" cannot be extracted. Data mining, or else known as knowledge discovery, attempts to address this need. Unlike standard statistical methods, data mining techniques look for exciting information without requiring prior hypotheses. It has also applied in well-known algorithms of machine learning, such as learning inductive rule (e.g., with decision trees) in the environment where extensive databases are involved. Data mining methods are used in business and research and are becoming increasingly popular over time [52].

In the Mobile Telecommunication case, a system for creating a Big Data flow-based analytics framework for the quick response and real-time decision making will be fair. Key challenges and research issues including, designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing and creating prediction models from Big Data flows [52]. A promising solution for data mining in such data is the dynamic extraction of multiple source data, as the information is usually collected from various sources [15]. This can be a knowledge indexing framework to ensure real-time monitoring and classification of Big Data applications [52].

7.1 Big Data Technical Tools

With an unprecedented increase in data collected, both users and network operators require useful data analysis and forecasting tools for real-time fast response and classification [15]. Researchers are proposing a robust data processing platform based on server groups and the Spark system. An excellent method to solve the problem of storing large data sets and processed results are to apply the Hadoop Distributed File System (HDFS) according to ensure the scalability and robustness of the system [47].

The Spark application memory platform is being developed to improve the framework's high data processing capability. The use of Python scripts along with the Anaconda scientific package and the Graphics Toolbar, Hadoop, Spark, MLlib, GraphX, Spark Streaming, and data processing tools based on batches, such as Kafka, is perfect for supporting analytics. However, a suitable method is not too developed Big Data systems directly on the servers, but first to be applied to a Cloud Computing infrastructure through the servers. Then Big Data processing tools and software packages can be configured on the virtual machines [47].

One right solution for a cloud computing infrastructure is OpenStack. OpenStack is an open-source cloud computing software with cloud computing functions that control the cloud in large groups of computer resources, storage, and networking in a data center at the IaaS level. Virtual machines with the widely used large data processing software, including Spark, Hadoop, and Hive, can be created and used through OpenStack. In this way, the platform can enjoy the flexibility and optimal capacity of the integrated cloud resource continuously. For testing and evaluation, a Big Data framework can capture and import a large-scale set of data from services such as Xender. According to an experiment about data measurement for 13 weeks, the total size was 3.56 TBytes [47].

All of these datasets include truly multidimensional features. Where users' online behaviors, location relationships, meeting dynamics, content properties, social characteristics, popularity trends, and so on. Are extracted by Cloud Analysis to take advantage of social graph properties. User preferences and content properties are then analyzed by Cloud of Interests and Cloud of Promoting, respectively [47].

8 Conclusion

Mobiles have an important impact in all the aspects of everyday life both in social and in personal level, in business and in education [53-64], in this paper, we study the features of big data and data analytics. We see how Big Data contributes

to mobile networks. We give a term in which Big Data generally refers to a large amount of digital data, and we estimate that the amount processed by "Big Data" systems will double every two years. Hence, Big Data on mobile networks need to be analyzed in-depth according to retrieve exciting and useful information. Big Data provides unprecedented opportunities for I.S.P.s to understand the behavior and requirements of mobile users, which in turn enables real-time decision making across a wide range of applications.

After that, we mention the 4Vs of big data. As we see, the 4Vs of big data are often described by their dimensions. However, the value is defined as the desired result of big data processing and not a defining feature of big data itself. This value provides an opportunity to relate the challenges directly to the defining characteristics of Big Data, making the origin and cause explicit. Consequently, regardless of all of these highlights, big wireless data is usually considered as the dataset that cannot be transmitted, accessed, processed, and served over a continuous period by existing communication and network systems.

Continuing, we study the use of big data analytics in mobile networks. As we see, today's consensus is that addressing these challenges in R.A.N. requires a combination of more spectrum, higher spectrum performance, network condensation, and unloading, unauthorized range, and fragmented spectrum aggregation using carrier aggregation techniques. The dual connectivity of terminals to multiple base stations can take advantage of the general use of the spectrum deployed at different base stations. Furthermore, we mention the end-to-end latency, the massive number of connections, the cost, the Quality of Experience, the Issues, and finally, the big data management.

We continue our study with the crucial aspects of the big data analytics in 5G, the 5G networks standardizing, and the 5G mobile optimization. We see that big data analytics can benefit organizations by allowing better targeted social media marketing, determining and anticipating market sentiments, and analyzing and understanding customer turmoil and other behaviors. Also, providers are expecting to play a crucial role in standardizing 5G networks. In recent years, many suppliers, operators, universities, research centers, and standardized and regulators have announced the launch of their work to develop a new generation of 5G mobile communications. Hence, standardization must play an essential role in the built-in 5G front-wheel-drive / reversible architecture for many suppliers. In the end, 5G mobile optimization can help providers to provide useful statistical characteristics such as spatial-temporal distribution, data collection property, and users' social correlations.

5G heterogeneous network or else HetNet is appropriate for providers to make the use of spectra more efficient in densely populated areas as they move in the direction of 5G development. According to an increase in the capacity of cellular networks in densely populated areas with high traffic requirements, the low-power node serving the high-capacity "data only" service is recognized as one of the critical elements of HetNets. A HetNet setting includes two key features. Users are concentrated in certain areas in the cell, thereby forming hotspots. Moreover, microcells are deployed in the vicinity of some hotspots, as is typically the case in current capacity-driven deployments. Because HetNets using small cells can improve the ability of the entire network, but it is not able to support the adaptation of various network resources according to the traffic characteristics that change over time. Also, many small cells lead to high-cost reconstruction because it is difficult to provide reliable reversal in a dense urban environment.

After that, mention about the big data on mobile networks. The data can be divided into two types: the users and the data from the network providers. A comprehensive analysis of two kinds of data can provide valuable information, which can be used by providers for network optimization. Users' applications mobile devices created this optimization, including the mobile user data, excessive amounts of data. Those users' data are events were are closely related to the mobility of subscribers. These events may be related to a specific subscriber ID and the cell where the event occurred. In the case of an HO event, this is the cell to which the subscriber is delivered (target cell). The first event describes a call installation, followed by three HO events and a call termination event. Of course, it is vital to be a method of mobile network collection data. Hence, the next step after collection is the process and the analysis of those data, with suitable technical tools for big data.

Compliance with ethical standards

Acknowledgments

The Authors would like to thank Net Media Lab Mind-Brain R&D Team for their support.

Disclosure of conflict of interest

The Authors proclaim no conflict of interest.

References

[1] J. Williams, 'Managing the Big Data Flood Smartly', PromptCloud, Oct. 14, 2015. https://www.promptcloud.com/blog/Managing-the-Big-Data-Flood-Smartly/ (accessed Dec. 20, 2019).

[2] P. Zikopoulos and C. Eaton, Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[3] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, 'Critical analysis of Big Data challenges and analytical methods', J. Bus. Res., vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

[4] 'More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute', Tubefilter, May 07, 2019. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/ (accessed Jan. 01, 2020).

[5] 'Google Search Statistics - Internet Live Stats'. https://www.internetlivestats.com/google-search-statistics/ (accessed Jan. 01, 2020).

[6] 'The Number of tweets per day in 2019', David Sayce, Dec. 03, 2019. https://www.dsayce.com/social-media/tweets-day/ (accessed Jan. 01, 2020).

[7] 'Top 20 Facebook Statistics - Updated November 2019', Zephoria Inc., Nov. 12, 2019. https://zephoria.com/top-15-valuable-facebook-statistics/ (accessed Jan. 01, 2020).

[8] T. I. Nath, 'How Big Data Has Changed Finance', Investopedia. https://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp (accessed Jan. 01, 2020).

[9] 'EDN - What the Large Hadron Collider scientists teach us about avoiding collisions - EDN'. https://www.edn.com/what-the-large-hadron-collider-scientists-teach-us-about-avoiding-collisions/ (accessed Jan. 01, 2020).

[10] W. John, Encyclopedia of Business Analytics and Optimization. IGI Global, 2014.

[11] K. Vassakis, E. Petrakis, and I. Kopanakis, 'Big Data Analytics: Applications, Prospects and Challenges', 2018, pp. 3–20. doi: 10.1007/978-3-319-67925-9_1.

[12] S. Mohanty, M. Jagadeesh, and H. Srivatsa, Big Data Imperatives: Enterprise 'Big Data' Warehouse, 'BI' Implementations and Analytics. Apress, 2013.

[13] F. Xu, Y. Li, M. Chen, and S. Chen, 'Mobile cellular big data: linking cyberspace and the physical world with social ecology', IEEE Netw., vol. 30, no. 3, pp. 6–12, 2016.

[14] 'Global mobile data traffic 2022', Statista. https://www.statista.com/statistics/271405/global-mobile-data-traffic-forecast/ (accessed May 31, 2020).

[15] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, 'Big Data Driven Optimization for Mobile Networks towards 5G', Nov. 2015. doi: 10.13140/RG.2.1.2389.1923.

[16] ParamitaGhosh, 'Data Modeling Trends in 2019', DATAVERSITY, Feb. 12, 2019. https://www.dataversity.net/data-modeling-trends-in-2019/ (accessed Dec. 09, 2019).

[17] L. Qian, J. Zhu, and S. Zhang, 'Survey of wireless big data', J. Commun. Inf. Netw., vol. 2, no. 1, pp. 1–18, Mar. 2017, doi: 10.1007/s41650-017-0001-2.

[18] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, 'Machine learning with big data: Challenges and approaches', IEEE Access, vol. 5, pp. 7776–7797, 2017.

[19] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, 'Design considerations for a 5G network architecture.', IEEE Commun. Mag., vol. 52, no. 11, pp. 65–75, 2014.

[20] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, 'Network slicing and softwarization: A survey on principles, enabling technologies, and solutions', IEEE Commun. Surv. Tutor., vol. 20, no. 3, pp. 2429–2453, 2018.

[21] K.-R. Jung, A. Park, and S. Lee, 'Machine-type-communication (MTC) device grouping algorithm for congestion avoidance of MTC oriented LTE network', in International Conference on Security-Enriched Urban Computing and Smart Grid, 2010, pp. 167–178.

[22] X. Yi, F. Liu, J. Liu, and H. Jin, 'Building a network highway for big data: architecture and challenges', IEEE Netw., vol. 28, no. 4, pp. 5–13, 2014.

[23] L. Gu, D. Zeng, P. Li, and S. Guo, 'Cost minimization for big data processing in geo-distributed data centers', IEEE Trans. Emerg. Top. Comput., vol. 2, no. 3, pp. 314–323, 2014.

[24] M. Morshed, 'Wireless Backhaul in Future Cellular Communication', 2018.

[25] Muljati Muli, '5 g peek from cmcc 20may2013', 05:39:40 UTC. Accessed: Dec. 09, 2019. [Online]. Available: https://www.slideshare.net/mulimuljati/5-g-peek-from-cmcc-20may2013

[26] M. Condoluci and T. Mahmoodi, 'Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges', Comput. Netw., vol. 146, pp. 65–84, 2018.

[27] E. Liotou et al., 'Shaping QoE in the 5G ecosystem', in 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, pp. 1–6.

[28] E. Baştuğ et al., 'Big data meets telcos: A proactive caching perspective', J. Commun. Netw., vol. 17, no. 6, pp. 549–557, 2015.

[29] N. Elgendy and A. Elragal, 'Big Data Analytics: A Literature Review Paper', in Advances in Data Mining. Applications and Theoretical Aspects, vol. 8557, P. Perner, Ed. Cham: Springer International Publishing, 2014, pp. 214–227. doi: 10.1007/978-3-319-08976-8_16.

[30] G. Bochechka and V. Tikhvinskiy, 'Spectrum occupation and perspectives millimeter band utilization for 5G networks', in Proceedings of the 2014 ITU kaleidoscope academic conference: Living in a converged world-Impossible without standards?, 2014, pp. 69–72.

[31] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, 'NFV and SDN—Key technology enablers for 5G networks', IEEE J. Sel. Areas Commun., vol. 35, no. 11, pp. 2468–2478, 2017.

[32] V. Frascolla et al., '5G-MiEdge: Design, standardization and deployment of 5G phase II technologies: MEC and mmWaves joint development for Tokyo 2020 Olympic games', in 2017 IEEE Conference on Standards for Communications and Networking (CSCN), 2017, pp. 54–59.

[33] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, 'Big Data Analytics in Mobile Cellular Networks', IEEE Access, vol. 4, pp. 1985–1996, 2016, doi: 10.1109/ACCESS.2016.2540520.

[34] X. Zhang et al., 'Social computing for mobile big data', Computer, vol. 49, no. 9, pp. 86–90, 2016.

[35] N. Cheng et al., 'Big data driven vehicular networks', IEEE Netw., vol. 32, no. 6, pp. 160–167, 2018.

[36] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, 'Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions', IEEE Commun. Surv. Tutor., vol. 17, no. 4, pp. 2377–2396, 2015.

[37] J. Hoadley and P. Maveddat, 'Enabling small cell deployment with HetNet', IEEE Wirel. Commun., vol. 19, no. 2, pp. 4–5, 2012.

[38] M. Peng, Y. Li, Z. Zhao, and C. Wang, 'System architecture and key technologies for 5G heterogeneous cloud radio access networks', IEEE Netw., vol. 29, no. 2, pp. 6–14, 2015.

[39] T. Maksymyuk, M. Brych, and V. Pelishok, 'Stochastic geometry models for 5G heterogeneous mobile networks', SmartCR, vol. 5, no. 2, pp. 89–101, 2015.

[40] A. Adhikary, H. S. Dhillon, and G. Caire, 'Massive-MIMO meets HetNet: Interference coordination through spatial blanking', IEEE J. Sel. Areas Commun., vol. 33, no. 6, pp. 1171–1186, 2015.

[41] R. Siddavaatam, I. Woungang, G. Carvalho, and A. Anpalagan, 'Efficient ubiquitous big data storage strategy for mobile cloud computing over HetNet', in 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–6.

[42] M. Chen, Y. Zhang, L. Hu, T. Taleb, and Z. Sheng, 'Cloud-based wireless network: Virtualized, reconfigurable, smart wireless network to enable 5G technologies', Mob. Netw. Appl., vol. 20, no. 6, pp. 704–712, 2015.

[43] J. K. Laurila et al., 'The mobile data challenge: Big data for mobile computing research', 2012.

[44] K. Wang, Y. Shao, L. Shu, C. Zhu, and Y. Zhang, 'Mobile big data fault-tolerant processing for ehealth networks', IEEE Netw., vol. 30, no. 1, pp. 36–42, 2016.

[45] G. Maia, D. L. Guidoni, A. C. Viana, A. L. Aquino, R. A. Mini, and A. A. Loureiro, 'A distributed data storage protocol for heterogeneous wireless sensor networks with mobile sinks', Ad Hoc Netw., vol. 11, no. 5, pp. 1588–1602, 2013.

[46] S. Bi, R. Zhang, Z. Ding, and S. Cui, 'Wireless communications in the era of big data', IEEE Commun. Mag., vol. 53, no. 10, pp. 190–199, 2015.

[47] X. Wang, Y. Zhang, V. C. Leung, N. Guizani, and T. Jiang, 'D2D big data: Content deliveries over wireless device-to-device sharing in large-scale mobile networks', IEEE Wirel. Commun., vol. 25, no. 1, pp. 32–38, 2018.

[48] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, 'The learning and prediction of application-level traffic data in cellular networks', IEEE Trans. Wirel. Commun., vol. 16, no. 6, pp. 3899–3912, 2017.

[49] G. Ostermayer, C. Kieslich, and M. Lindorfer, 'Trajectory estimation based on mobile network operator data for cellular network simulations', EURASIP J. Wirel. Commun. Netw., vol. 2016, no. 1, p. 242, Oct. 2016, doi: 10.1186/s13638-016-0718-x.

[50] J. Schobel, R. Pryss, M. Schickler, M. Ruf-Leuschner, T. Elbert, and M. Reichert, 'End-user programming of mobile services: empowering domain experts to implement mobile data collection applications', in 2016 IEEE International Conference on Mobile Services (MS), 2016, pp. 1–8.

[51] F. Zheng et al., 'Crowdsourcing methods for data collection in geophysics: State of the art, issues, and future directions', Rev. Geophys., vol. 56, no. 4, pp. 698–740, 2018.

[52] R. Elankavi, R. Kalaiprasath, and R. Udayakumar, 'DATA MINING WITH BIG DATA REVOLUTION HYBRID.', Int. J. Smart Sens. Intell. Syst., vol. 10, 2017.

[53] Stathopoulou, et all 2018, Mobile assessment procedures for mental health and literacy skills in education. International Journal of Interactive Mobile Technologies, 12(3), 21-37, https://doi.org/10.3991/ijim.v12i3.8038

[54] Kokkalia G, AS Drigas, A Economou 2016 Mobile learning for preschool education. International Journal of Interactive Mobile Technologies 10 (4), 57-64 https://doi. org/10.3991/ijim.v10i4.6021

[55] Stathopoulou A, Karabatzaki Z, Tsiros D, Katsantoni S, Drigas A, 2019 Mobile apps the educational solution for autistic students in secondary education Journal of Interactive Mobile Technologies 13 (2), 89-101https://doi.org/10.3991/ijim.v13i02.9896

[56] Drigas A, DE Dede, S Dedes 2020 Mobile and other applications for mental imagery to improve learning disabilities and mental health International Journal of Computer Science Issues (IJCSI) 17 (4), 18-23, DOI:10.5281/zenodo.3987533

[57] Alexopoulou A, Batsou A, Drigas A, 2020 Mobiles and cognition: The associations between mobile technology and cognitive flexibility iJIM 14(3) 146-15, https://doi.org/10.3991/ijim.v14i03.11233

[58] Papoutsi, C. and Drigas, A. (2017) Empathy and Mobile Applications. International Journal of Interactive Mobile Technologies 11(3). 57. https://doi.org/10.3991/ijim.v11i3.6385

[59] Drigas A, Mitsea E 2022 Conscious Breathing: a Powerful Tool for Physical & Neuropsychological Regulation. The role of Mobile Apps Technium Social Sciences Journal 28, 135-158. https://doi.org/10.47577/tssj.v28i1.5922

[60] Stathopoulou A., Loukeris D., Karabatzaki Z., Politi E., Salapata Y., and Drigas, A. S., 2020 "Evaluation of Mobile Apps Effectiveness in Children with Autism Social Training via Digital Social Stories," Int. J. Interact. Mob. Technol. (iJIM); Vol 14, No 03,

[61] Drigas, A. S., and Angelidakis P., 'Mobile Applications within Education: An Overview of Application Paradigms in Specific Categories', International Journal of Interactive Mobile Technologies(iJIM), vol. 11, no. 4, p. 17, May 2017. https://doi.org/10.3991/ijim.v11i4. 6589

[62] Drigas AS, Pappas MA, 2015 A review of mobile learning applications for mathematics. International Journal of Interactive Mobile Technologies 9 (3)

[63] Drigas, A.S., Ioannidou, R.E., Kokkalia, G. and Lytras, M. (2014), "ICTs, mobile learning and social media to enhance learning for attention difficulties", Journal of Universal Computer Science, Vol. 20 No. 10, pp. 1499-1510.

[64] Drigas, A., Kokkalia, G. & Lytras, M. D. (2015). Mobile and Multimedia Learning in Preschool Education. J. Mobile Multimedia, 11(1-2), 119–133.