NATIONAL RESEARCH UNIVERSITY

HIGHER SCHOOL OF ECONOMICS

Faculty of Business and Management

School of Business Informatics

Sara EL Elimy

METHODS TO EFFICIENTLY HANDLING BIG DATA IN 5G NETWORKS

MASTER THESIS

Respective Field of Study: 38.04.05 Business Informatics

Master's Program «Big Data Systems»

Supervisor
National Research University HSE

Prof. Evgeny Kucheryavy

Adviser
UAS Technikum Wien

FH-Prof. Dipl.-Ing. Helmut Gollner

Submission date: May 2018

Moscow 2018

ABSTRACT

Telecommunications has a huge influence on our social life nowadays and leaves big impact on the innovation, growth and success of other sectors. Thanks to the new trend and dynamic environment of this sector, most of Mobile network operators MNOs start to face a lot of challenges in the digital era, especially with high demands from customers, and that is why they started to move from the 4th generation LTE toward 5G (5th Generation). As a result, we started to highlight the importance of effectively applying big data analytics for current mobile cellular network development to achieve 5G requirements and for telecoms companies in developing their business and operation aspects. We provide methods and frameworks for handling Big data effectively to adapt fast and performance with high efficient for new business changes and challenges from new trends like Digitalization and IoT by using different analysis techniques. Due to the need for low latency for 5G networks, we will focus on caching as a solution and application based on big data analytics. WINTERsim simulator is used to figure out results from different caching techniques.

Key words: Big Data, 5G, WINTERsim , Caching, Telecom, MNOs

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my academic supervisor Prof.Evegeny Kucheryavy for his continuous support for my master study and research. I also would like to thank him for his patience, motivation, enthusiasm, and immense knowledge. His guidance has helped me in all the time of research and writing of this thesis.

My sincere thanks also goes to my second adviser FH-Prof. Dipl.-Ing. Helmut Gollner for his continuous support during my study in UAS Technikum Wien.

I also thank Prof. Petr Panfilov and Dr. Petr Baranov for their guidance, insightful comments during research seminars

My sincere thanks also goes to Dr. Michael M. Komarov for his guidance to me as an international student during my study in National Research University, Higher School of Economics and in UAS Technikum Wien.

I am grateful to Prof. Vasily Kornilov for providing me an internship opportunity, which helped me to work on diverse exciting projects.

I am so grateful to the Russian Federation Quota scholarship scheme and the Faculty of Business and Management at National Research University, Higher School of Economics, and School of Business informatics for making it possible for me to study here. I give deep thanks to the professors and lecturers at Big Data Systems program.

Last but not the least, I would like to thank my family who encouraged and supported me throughout the time of my research.

Table of Contents

List of Figures

INTRODUCTION

Telecommunications is a big and rich industry; many diverse companies work in variant aspects of this business. Many specialist companies produce hardware and software to service providers. Telecommunication sector, moreover, plays a critical role in all other industries and in people's lives; it has changed how people communicate and how companies and organizations can do their business. This sector requires an environment that is dynamic and rapidly changeable. In the past, all ways of communication depended on wired technologies and equipment, while wireless and mobility communications started to be dominant and produced large amount of data that are booming and become vastly required.

Nowadays most of Mobile network operators (MNOs) have started to face many challenges with digital era, especially with extraordinary demands from customers for OTT application (e.g. WhatsApp, Facebook, etc.). As a result, they started to look for innovative solutions to face these challenges and offer perfect customer experience by managing their complex networks by efficiently handling network resources.

One of these challenges, is to utilize vast number of cell phones and generate vast amounts of traffic on a daily basis with data characteristics Volume, Velocity and Variety generated from end user. Mobile networks have been booming exponentially (Musolesi 2014) and this, in turn, has had significant effect on society. This data traffic created with diverse data types from various domains (e.g. D2D, M2M, connected cars…etc.) (Gupta et al. 2015) (Agiwal, et al. 2016).

Concurrently, MNOs have started to move from 4th Generation LTE to 5G (5th Generation) wireless network, which is the promising and upcoming required solution to meet specifications of wireless broadband of 2020 owing to its new technology components.

Therefore, Big Data has already become very important in our life and will be firmly established by the next Generation (5G) very soon (Boccardi 2014). Knowing about special characteristics of Big Data for MNOs becomes an crucial part, which is critical for 5G requirements and KPIs. (Zheng et al. 2015).

Due to certain needs of 5G network and advanced Big Data Analytics, this research is indispensable for both scientific and industrial stakeholders in the telecommunications sector. Researchers and telecom companies have been recently investigating diverse and innovative methods for handling Big Data in an effective manner to discover hidden patterns and information from the data which is collected from different sources and help companies to provide new smart services with high revenues and to achieve reduction of capitals and expenditure.

The main goal of this thesis is to investigate effectiveness application of big data for current mobile cellular network development and to fulfill the 5G requirements of telecoms companies in developing their business and operational aspects.

The main goal can be achieved by achieving the following sub-goals, which we address in this thesis are to:

- Provide methods and framework for handling big data effectively inside MNOs.
- Know what kind of big data sources are inside MNOs and its features and what analytic tools to be used
- Provide different Big Data applications and different scenarios, from current mobile cellular network to upcoming 5G
- Recognize how big data and machine learning are vital for 5G network
- Investigate if caching at the edge as one of key solution to achieve requirement for low latency in 5G and its importance in future wireless big data applications for MNOs and its integration with big data analytical tools.

This thesis could be interesting for:

- The telecommunications companies sector, which are planning or getting ready for applying big data analytics.
- Academics in the fields of big data and telecommunications.
- Wide-ranging range of readers who are involved with ICT.

THESIS STRUCTURE AND PURPOSE

The chapters are organized as follows:

Chapter 1 describes the development of Mobile networks generations, why 5G is become a need and the main challenges to achieve 5G goals. Connection between 5G and Big Data, by the effective handling of Big Data in mobile cellular networks, can expand different capabilities for 5G network.

Chapter 2, Big Data Driven Methods and Frameworks, aims to effectively handle Big Data in MNOs and can also be applied for 5G network as the key technique for decreasing deployment costs and improving network efficiency.

Chapter 3 includes beneficial Data sources and features, which can be collected in mobile network operators and Big Data analytics tools, to be used for analyzing those data types.

Chapter 4 discusses different case studies and applications for handling Big Data effectively in mobile network operators and applies methods on these applications to demonstrate possible new solutions for improving the performances of mobile networks towards 5G.

Chapter 5 is mainly concerned with caching in mobile network as application of big data Analytics and as a solution to latency requirement for 5G network and challenges facing mobile network operators . It also includes caching at the edge trend and popularity-based caching will be introduced and different case studies results will be shown after applying this technique.

Chapter 6 consists of the method and techniques employed in thesis, which is divided into two parts. The first part refers to the call details records (CDRs) datasets that will be used for applying different analysis techniques and draws attention to the importance of end results for the industrial sector in business development and enhancing network performance.

It also includes the description and features of Datasets used from “Telecom Italia Big Data challenge“.

The second part relates to caching based on popularity of end users requests content using WINTERsim simulation tool to figure out the results of applying popularity based caching and check different parameters affecting network performance.

CHAPTER 1

1.1. Generations of Mobile Network

From the evolution of cellular network generations, it is evident that there have not just been liberal additions in mobile network capacity, there have additionally been changes in operating systems and introducing various services.

First Generation (1G) of cellular network was based on the Advanced Mobile Phone System (AMPS) and analogue communication techniques that were less difficult to produce but facing problem of power consumption and interference

The move to the second generation (2G) used two network technologies were digital, much secure than 1G, with less power consumption which are , GSM and CDMA. They made more efficient use of spectrum which is consider an important and rare resource for cellular network. However, their design was based on that usage would be mainly for phone calls without keeping in consideration internet accessing demands. For internet access in 2G, it provided slow mobile data speed typically less than 10 kilobits per second.

Vast demands on internet services and high speed of mobile data, 3G was presented to support mobile operators to cover this demand and provide new services. From speeds with hundreds of kilobits per second , 4G with it new principles in all Internet Protocol (IP) provided tens of megabits per second. (Philip Branch, 2018).

1.2. Why *5G?*

The fifth generation (5G) is still in the developing phase, the International Telecommunications Union ITU has formed a number of focus groups to examine how it will be implemented. One of its advantages will be an increase in the number of customers that can be serviced with the same amount of spectrum, but with improved speed (ITU-R 2015).

The appearance of the Internet of things is one of the reason behind the emerging of new generation. With interaction of a lot of things and trends of smart houses and smart cities during our life or usage, and monitors and controls by internet these all make a unique advantages of 5G (Parvez et al.2017).

"Cisco believes that there will be 50 billion devices connecting to internet by 2020 compared to the current 15 billion devices". 5G will provide up to 10 times faster broadband services as compared to 4G. Until 2019, we will be able to see the 5G smartphones (Orange 2018).

In several ways, the recent 3G and 4G networks are not suitable for a lot of new trends and applications. The future use and breakthrough in the media transmission framework is very encouraging and now the wide range use of radio wave in industries, restorative, data framework , instruction than the range a bit mess. In the near future by 2020 it is estimated that over 50 billion gadgets will be web linked, so irrespective of the use of wide range for gadgets to access the web have to be considered else framework will be unable to meet high request (Kadir et al.2015). There are several Billions of sensors that have been built and linked into security frameworks, machines such as cars, live screens, clothing and door locks. from critical observation, An intensive study from Firm Gartner predicts the number of organized gadgets accessing the web has rapidly increased from approximately five billions in 2015 to a raging 25 billion by 2020 (B. Commision Report 2014).

5G mainly aims to achieve improvement of capacity efficiency in energy and increase connection density and low latency than other generation of wireless network as illustrated in Fig.1.1 (Parvez et al .2017).

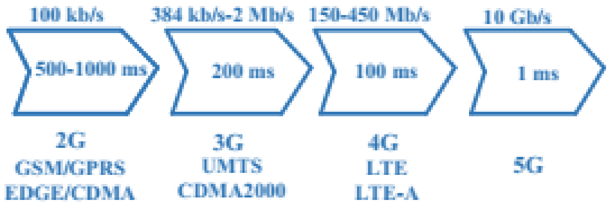


Fig.1.1: latency for mobile network generations (Parvez et al .2017)

One of 5G’s benefits is that it provides real time communication for machines and devices to build IoT with personal communication, empowered by different application at the edge of network as in Fig1.2 (Parvez et al .2017)

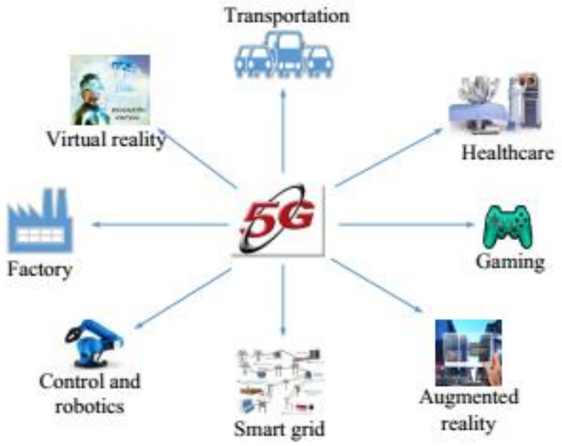


Fig1.2: Application with low latency requirement at the edge 5G network (Parvez et al .2017).

1.3. Connection between 5G and Big Data

5G networks and big data are correlated in two direction as 5G is mandatory for big data transmission and inversely generation of big data depends on new recent efficient mobile system (Chimeh 2015). He studied the characteristic and specification of big data and 5G, with special full description, to the rule of 5G component as a reliable and accurate base for handling the big data.

These competent are cloud computing , network function virtualization NFV , and massive 3d MIMO , results of this work reviled that , 5G and big data are correlated in two direction as 5G is mandatory for big data transmission and inversely generation of big data depends on new recent efficient mobile system

In addition , he concluded that big data needs improvement in wireless network eg. Data rate , accelerating the mobility , user speed And decrease the delay this advantages has been included in 5G and its new technology .(Chimeh 2015)

5G has interrelated diverse things as far as drivers that will structure up its framework's uniqueness from the past ages. IoT will be made accessible with 5G to interface billions of gadgets and improve the basic things in the human everyday life. 5G also focuses on varying the presence of big data from human-to-human interface towards machine-to-machine platform.

KPIs recognized, for 5G force the use of extensive, complex and efficient energy algorithms and different solutions for the core access elements in the network. It is generally comprehended that the entire network's performance will be improved by a dependable and prompt access to exact information and right data in a wide sense.

Concerning the future wireless network, gaining a lot of information to gather relationships and likelihood are conceived to empower proactive decisions and to enhance system execution and proficiency in this manner. Hence, utilization of ML based application might be vital. Despite the fact that the information structures can be organized from numerous points of view, data correlation with geo-location is a promising idea giving both efficiency and visualization. Relocation, on the other hand, can empower enhanced short and long management plan of resources in the network in a proactive manner. (Chakraborty, 2017)

CHAPTER 2

2.1. In The Era of Big Data and Machine learning

MNOs are considered a source and carrier of huge amount of data starting 2014 as the mobile users worldwide penetration have achieved 97% of mobile data around the world, creating staggeringly 10.7 ExaBytes (10.7 × 1018). This increase on mobile data traffic is due to the prominence of cell phones and other smart devices that demand mobile data broadband applications, such as online music, video and gaming as appeared in Fig. 2.1 (Bi et al., 2015).

With yearly growth rate of more than 40%, it is normal for the portable information activity to increase 5 times from 2015 to 2020 due to the high demand of data and that expansion in advanced wireless technologies and mobile applications (Liu et al., 2015).

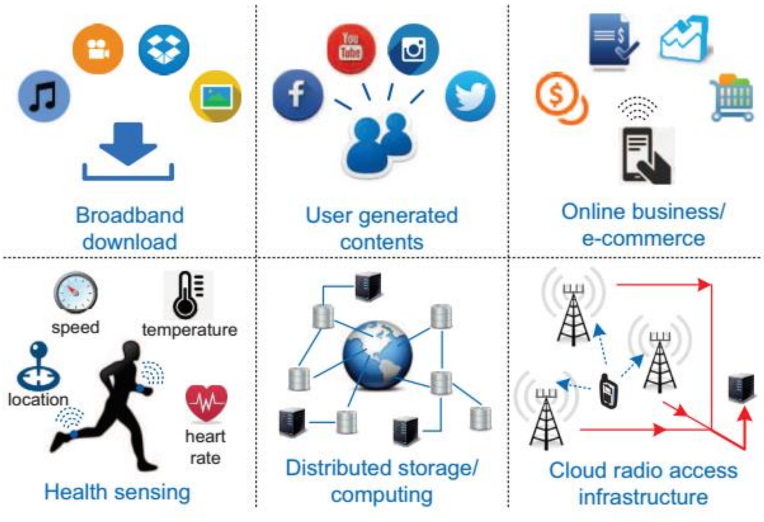

Broadband download
User generated contents
Online business/ e-commerce
speed
temperature
location
heart rate
Health sensing
Distributed storage/ computing
Cloud radio access infrastructure

Fig.2.1 : Different sources of Mobile Data Traffic

Before the shift to big data analytics, specialists used traditional data analytics. This was one of the setbacks that led to paying no attention to the massive data, which was accumulating on mobile cellular networks database. The structure of data was the solely factor affecting on usage of traditional data analytics was solely based on the structure of data ,this was a problem because a large amount of the accumulated app-based data was unstructured (He et al., 2016). It also paid less attention to operational data and did not focus on transactional data.

A good example is how the complete data that is related to a client or subscriber are usually fragmented in several different business departments in case of using the traditional way of analyzing these data. However, using the big data analytics, the capability of collecting this scattered data so as to form a pattern of customer behaviors and their preferences from several perspectives like places they frequently visit , what they access on the Internet and range of different activities to form an integrated picture elevated by a highly significant can be found in home location Register HLR (He et al., 2016).

However, big data analytics is beneficial in numerous ways in comparison to the traditional ways of analyzing the accumulated data. Since the bandwidth of the mobile cellular networks is scarce, big data analytics can define which data are useful and compressor transmit data (He et al., 2016).Real time decision-making in large part of application is another advantage of Big Data analytics, by monitoring infrastructure and development of network performance. MNOs will easily and effectively provide and support numerous smart and new services by analyzing different types and sources of data.

Moreover, with unification of 5G, the quality of services and customer experience is vastly expected to increase by merging mobile network operators MNOs with big data analytics.

However, the nature of Big Data presents vast challenges in relation to data mining, mobile sensing and knowledge discovery (Lomotey & Deters 2014). New technologies are required to handle Big Data in a highly scalable, cost-effective, and fault-tolerant fashion (Zeng.et al., 2015) (Lei et.al, 2013). Not only challenges related to huge volume of data can MNOs faced. Also, non-homogenous structure is considered another challenge due to incomplete and ambiguous information.

Furthermore, MNOs are expected to play a key role in the standardization of 5G networks. However, a critical challenge is to understand the requirements of utilizing Big Data analytics to provide user services with personalized Quality of Experience (QoE), and to enable highly efficient resource utilization in 5G networks.

That is why it has become essential to have good knowledge of the unique characteristics of Big Data in mobile networks. Such knowledge is crucial for the optimization of 5G mobile networks (Zheng et al. 2016).

With current improvement and development in data analytics, mobile networks based on Big Data have become an attractive research area for many researchers all over the world (Samulevicius et al., 2015) (Ramaprasath et.al., 2015).

In addition, researchers in the industrial sector have recently investigated and built frameworks for handling Big Data in an effective way in mobile networks.

2.2 Frameworks based on Big Data inside Mobile Network

Different frameworks are introduced to support the big data analytics in mobile cellular networks as discussed below.

A Big Data Driven (BDD) framework for mobile network optimization is proposed by (Zheng et al., 2016) in Fig.2.2

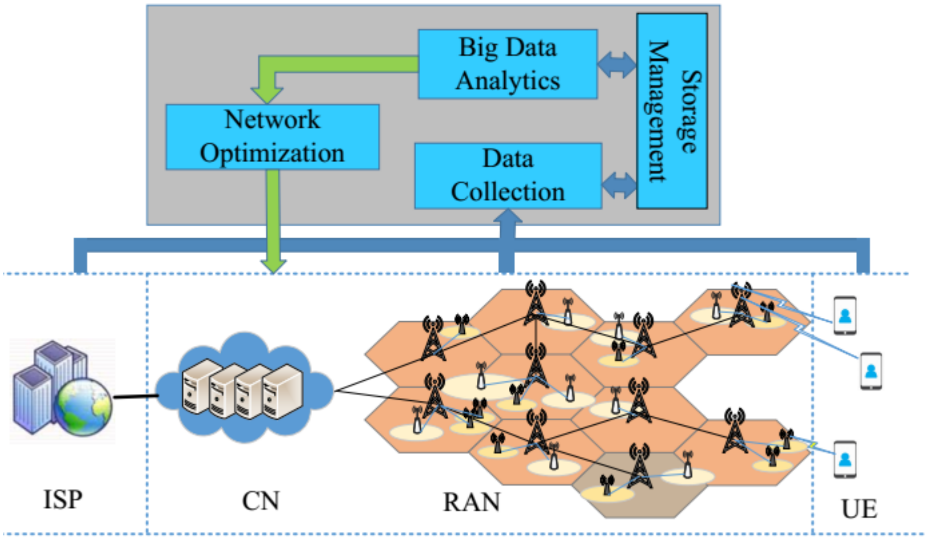


Fig.2.2 Big Data Driven Framework (Zheng et al., 2016)

Fig.2.2 shows Big Data collection, where data are collected from different sources in and out of the network: User Equipment (UEs), the Radio Access Network (RAN), the Core Network (CN) and the Internet Service Providers (ISPs). (Zheng et al., 2016).

The next process is storage management and preprocessing, in both frameworks (Zheng et al 2016) and (He et al. 2016) highlighted the importance of this step, whereas Big Data storage infrastructure needs to have scalable capacity as well as scalable performance. Thus, storage management needs to be simple and efficient so that storing and sorting of Big Data can be easily achieved. In addition, the preprocessing of data before data analysis is crucial so as to avoid the consumption of unnecessary storage space and thus ensure the efficiency of the power of processing (He et al. 2016).

After data collection process and storage, serious problem for MNOs is the processing of such huge volume of data. The collected data are multi-source, heterogeneous, real-time and voluminous. Here comes the main and important role of data analytics and knowledge extraction techniques to process the data and convert them into actionable knowledge. Subsequently, this knowledge could be employed to build an sutiable schemes for network optimization (Zheng et al. 2016).

In Zheng et al. (2016) framework, the BDD network optimization functions will :

- analyzing Big Data to recognize problems, and to decide what level of optimization and how can be processed .
- Able to predict traffic variations either in a local area or over the network coverage.

Based on the optimization results, there will be improvement measurements to be implemented by the control functions in the RAN and through prediction of traffic variations results that will help to improve network and user performance.

He et al. (2016) presented an architectural framework that supports the big data analytics in mobile cellular networks.

The first process is the collection of data, Big data in mobile cellular networks are collected from either internal or external sources and they will be defined in the next chapter. Data are collected through:

1- Data sources
2- Auxiliary tools

Also, smartphones are considered some of the tools for data collection and their functions include collection of audio information through the microphones, collection of geological locations through GPS, Wi-Fi, or Bluetooth among other applications, and collection of pictures, videos, and other multimedia information through cameras. (He et al. 2016)

The next step is big data analysis and preprocessing. The collected data from network logs, CDRs, weblogs, etc. (He et al. 2016), will not be transmitted to storage directly as they should be preprocessed for correlation and normalization. There are three common preprocessing techniques: integration, cleaning and redundancy elimination (Chen et al 2014).

Meanwhile, these massive amounts of collected data with different types cannot be fit in relational database. These datasets collected on a large scale are in several different locations under various formats, and, therefore, NoSQL databases are becoming the new technology implemented for big data.

The following step is the big data analytics platforms and tools; Hadoop comes to mind as for Big Data. It is an open-source software framework meant for the distribution of storage and processing power of large datasets. To cater to the various models of the business applications, Hadoop has to turn into a multi-purpose big data operating platforms where manipulations of data and their analytical operations can be plugged (Meng et al. 2015). These features make Hadoop mainly adjusted to the processing of data in mobile cellular networks including CDRs, GPS data, web clickstream, and network logs to say the least (He et al. 2016).

The last step is big data analytics applications. The application of big data analytics can be divided into two aspects:

- internal business supporting applications
- external innovative business model development

The internal business supporting applications include operational efficiency that is offered to subscribers of given mobile cellular networks in terms of experience improvement, as well as customized marketing (He et al. 2016).

From the different frameworks introduced and mentioned in this chapter, there are common processes for the big data driven framework that can be applied for any application in MNOs network:

1. Big Data collection from different sources.

2. Storage management and preprocessing of Data.
3. Big Data analytics.
4. Big Data analytics application and network optimization

CHAPTER 3

3.1. Data Sources with its features And Big Data Analytic Tools

3.1.1. Data Sources

Huge Data which collected from mobile network operators can be classified at different levels. Zheng et al (2016) divided into two types subscriber’s data and network’s data from general prospective and different features are represented in fig.3.2 .

3.1.1.1. Subscriber's data

Subscriber's data are usually generated from UE and contained a lot of information regarding their location, with whom, when, and how they communicate and their behaviors patterns. In addition, different types of installed applications on subscribers’ devices have also become a source of big data.

3.1.1.2. Operator's data

The main sources of these data are CN and RAN which are mainly large amounts of data including information about network performance, signaling, cell information and faults. Many network performance inductors will be discussed later in details.

In addition, He et al. (2016) classified the data into Internal sources from the operational, business, and other supporting systems. External sources are derived from the local statistics.

From general classification, they can be deeply classified according to Mobile network operators ‘KPIs and different types of stored logs and databases used in the network as introduced by Imran et al.(2014) shown in Fig.3.1

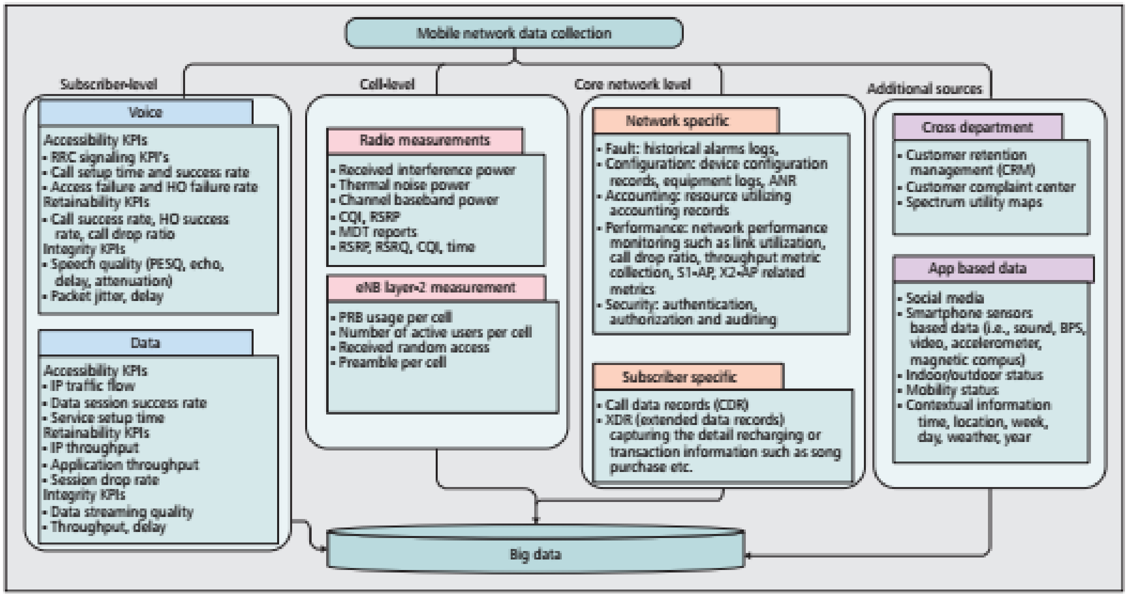

Mobile network data collection
Subscriber-level
Cell-level
Core network level
Additional sources
Voice
Accessibility KPIs
• RRC signaling KPI's
• Call setup time and success rate
• Access failure and HO failure rate
Retainability KPIs
• Call success rate, HO success rate, call drop ratio
Integrity KPIs
• Speech quality (PESQ, echo, delay, attenuation)
• Packet jitter, delay
Data
Accessibility KPIs
• IP traffic flow
• Data session success rate
• Service setup time
Retainability KPIs
• IP throughput
• Application throughput
• Session drop rate
Integrity KPIs
• Data streaming quality
• Throughput, delay
Radio measurements
• Received interference power
• Thermal noise power
• Channel baseband power
• CQI, RSRP
• MDT reports
• RSRP, RSRQ, CQI, time
eNB layer-2 measurement
• PRB usage per cell
• Number of active users per cell
• Received random access
• Preamble per cell
Network specific
• Fault: historical alarms logs,
• Configuration: device configuration records, equipment logs, ANR
• Accounting: resource utilizing accounting records
• Performance: network performance monitoring such as link utilization, call drop ratio, throughput metric collection, S1-AP, X2-AP related metrics
• Security: authentication, authorization and auditing
Subscriber specific
• Call data records (CDR)
• XDR (extended data records) capturing the detail recharging or transaction information such as song purchase etc.
Cross department
• Customer retention management (CRM)
• Customer complaint center
• Spectrum utility maps
App based data
• Social media
• Smartphone sensors based data (i.e., sound, BPS, video, accelerometer, magnetic compus)
• Indoor/outdoor status
• Mobility status
• Contextual information time, location, week, day, weather, year
Big data

Fig.3.2 : Data classification of mobile network for big data 5G -based SON(Imran et al. 2014)

From fig.3.2 it is clear that the data is classified into :

3.1.1.3. Subscriber's data

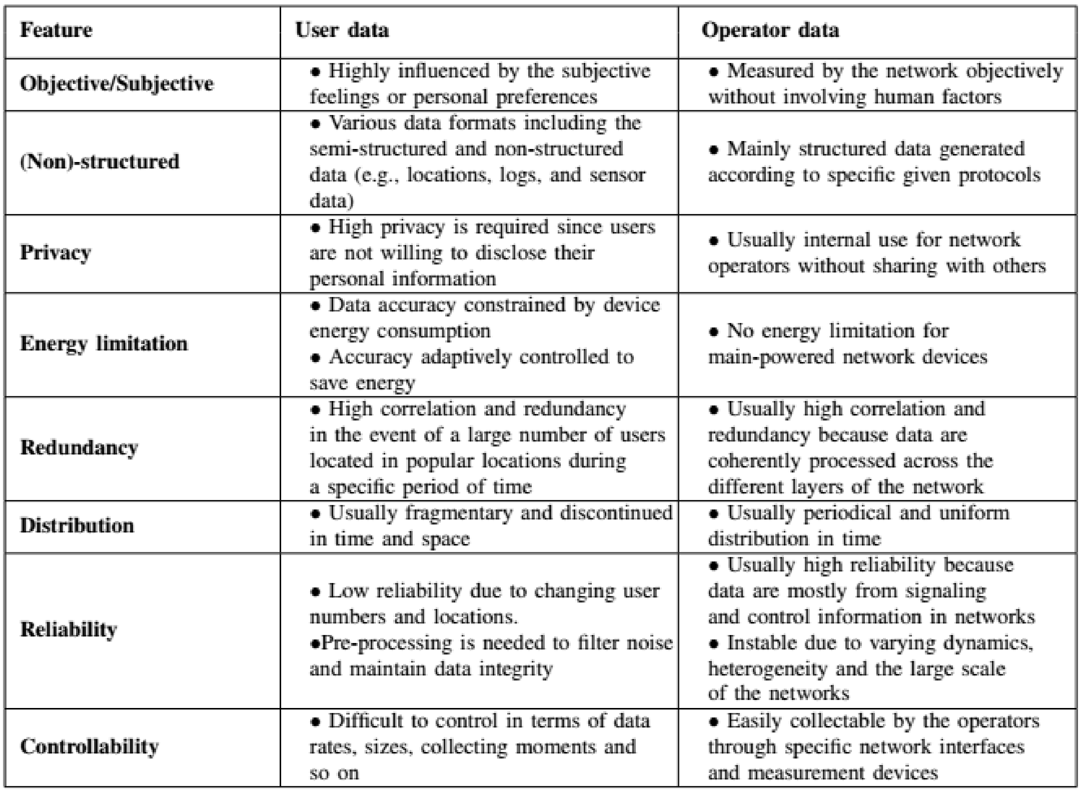
this contains controlled data and important data which not as it were can be overused in order to enhance and improve and arrange network-centric operations, but this is similarly necessary to support key trade forms such as client's implication and management improvement.

3.1.1.4. Cell-Level Data

The utilities of cell-level data can complement the subscriber's level data. We can group these data streams within a big data as a cell-level data For instance Minimization of drive test (MDT) measurements, which comprise the Reference Signal Received Power (RSRP) and Reference Signal Received Quality (RSRQ) degree of the functioning and close cells are specifically of use for independent coverage appraisal and optimization. These same measurements can be use to develop automated error discovering and utilization of solutions for identifying coverage holes, sleeping cells, or cells in suspension. Similarly, on a sum level, layer 2 measurements such as average cell load in context with subscriber level data and contextual data such as time, day, and weather data can serve as input to the SON engine for executing load balancing, traffic control and prediction function.

3.1.1.5. Core-Network-Level Data

The recognition of data made available by the central network can be classed as a core network data. Recently such data is only done autonomously and hence refuses to control the total network operation. Error recognition and diagnoses of network-level problems can be resolved using Data from streaming (Imran et al. 2014).



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

Fig.3.2: shows comparison between user data and operator data features (Zheng et al 2016).

3.1.2. Analytics Tools

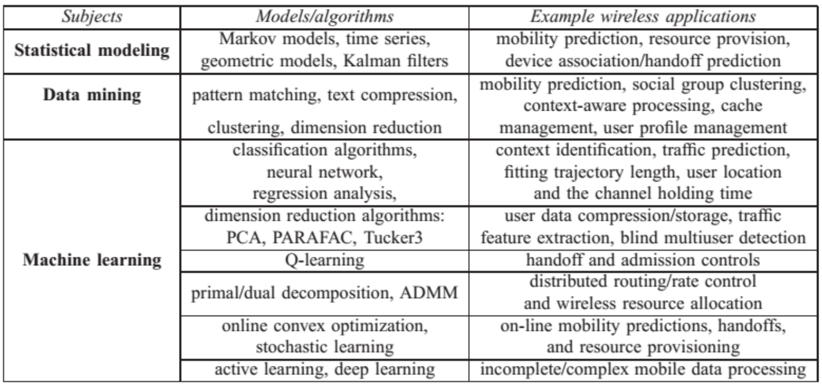
Collection of raw data is the first step of huge Data processing in mobile network. For example, MNOs can receive information from mobile users who downloads information connected to their movement. Localization errors and eco-interferences are often huge problems to the usage of Big Data. Moreover, the power source of the user device may be exhausted, and hence the required data cannot be received in sure occasion.

Approaching this end, data mining, selection and removal processes are developed with the aim of removing interference or non-usable data that belongs to the presumed error-prone grouping schemes. However, it is still a significant difficulty to possess useful data from unfinished, reduced and unpredictable Big Data. One assuring solution for data mining over such data is multi-source dynamic data mining, since data are basically received from broad sources (Wu et al. 2014).

Researchers are nowadays massively studying new and effective big data techniques and tools To explore and find out the previously undefined patterns and knowledge from the collected data , one of these research efforts resulted in various big data applications with proscriptive and predictive analytics using strong machine learning techniques such as support vector machine (SVM) and deep learning which reflects the advance of multi varieties statics , pattern identification and data mining (Zheng et al. 2016).

Another is sequence classification the author stated that SVM is proven to be highly effected in classifying feature sequences.

In the same prospective, Bi et al. 2014 summarized the most common used analytics tools for wireless traffic analysis and their main application to wireless communication , their base are the embedded data – analytics algorithms and they classified it into three categories: statistical modeling , Data Mining and Machine learning modeling methods as per fig.3.3.



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

Fig.3.3 Popular Big Data Analytics tool mobile network data and application (Zheng et al (2016)

CHAPTER 4

4.1. Applications of Big data and Machine learning For Mobile Cellular Network

4.1.1. Big data in accordance to Signaling

Superiority of network services enhancement and real–time network problems identification (Fei et al.2016) are depending on signaling monitoring which is also a vital part in resources allocation for mobile network(Fei et al.2016). Traditional systems for signaling monitor unable to handle efficiently the massive amount of data, which produced inside network as a result of continuous and fast improvement for mobile network.

Fig. 4.1 illustrates Big Data analytics for signaling monitoring Architecture. As mentioned before, there are main processes in framework for handling Big Data in mobile cellular network which are data collection, data analysis and application, by applying them in the below architecture (He et al. 2016):

- Data collection: different signaling protocols collected from various network components without affecting any other network operations. Through protocol processor, the preprocessing step is done before storage and analysis of data.
- Data analysis: different algorithms are used for analyzing, like correlation analysis and decomposition.
- Data application: results from the analysis process used in different applications.

This is one of case studies and applications proposed by Celebi *et al* (2013), for enhancement network converges using BSSAP message data from Hadoop platform interface and analyzed data to recognize handovers from 3G to 2G. Simulation results demonstrate that recognized 3G coverage gaps are reliable with the drive test.

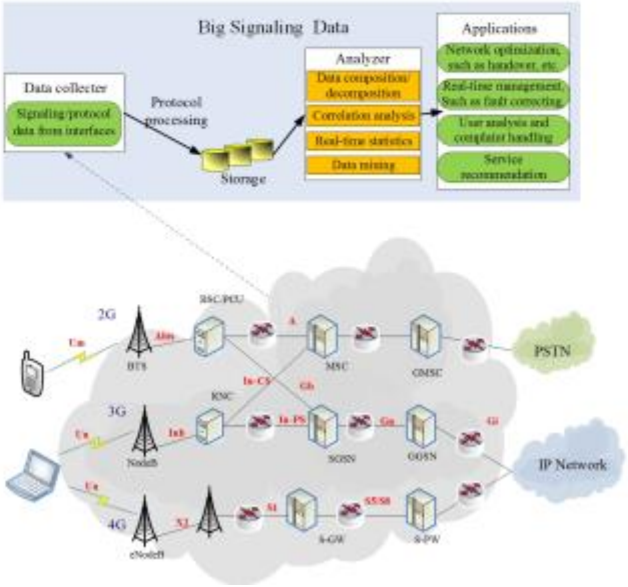


Fig. 4.1: illustrates signaling monitor based on Dig Data driven Architecture (He et al. 2016)

4.1.2. Big data related to Location

Costless data about the communication behavior class of people is given by analysis of location data as human mobility is related to location.

In Fig. 4.2 , the data collection based on location can be collected from different sources . As mentioned before, there are main processes in framework for handling Big Data in the mobile cellular network, data collection, data analysis and application, by applying them in the following architecture (He et al. 2016):

- Data collection
- Data analysis
- Data application

Kolar *et al*.(2014) and Deng et al.(2013) emphasized gain from these collected data, especially CDRs call detailed records, from business perspective .

framework for mobile advertising and marketing was introduced by Deng et al. (2013) based on big data analytics, as Kolar *et al*.(2014) used CDRs datasets to develope end to end based on Hadoop with diverse algorithms .

It was not only researchers who highlighted the importance of CDRs datasets analysis, but vendors and Mobile operators like"Orange Telecom and Telecom

Italia” proposed many Big Data challenges for gaining big data analytics benefit in revenue and development of network .

The practical of this work will include more details about the importance of CDRs, different analyses and gaining that can be achieved from these datasets.

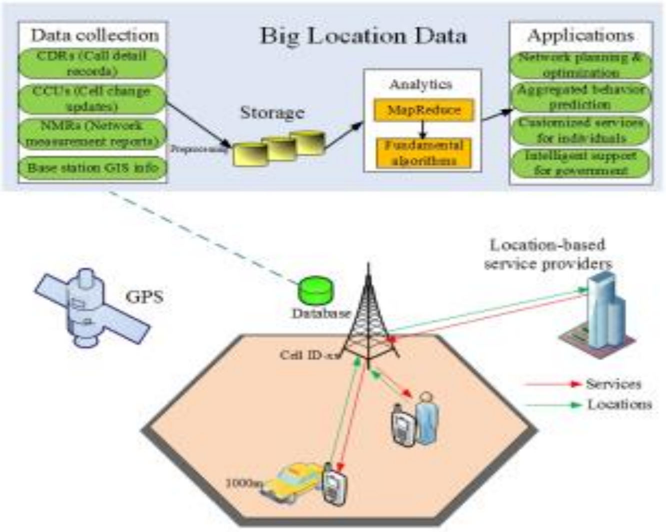

Data collection
CDRs (Call detail records)
CCUs (Cell change updates)
NMRs (Network measurement reports)
Base station GIS info
Big Location Data
Preprocessing
Storage
Analytics
MapReduce
Fundamental algorithms
Applications
Network planning & optimization
Aggregated behavior prediction
Customized services for individuals
Intelligent support for government
GPS
Database
Location-based service providers
Cell ID
Services
Locations
1000m

Fig. 4.2: illustrates BDD Architecture Based on Location Information (He et al. 2016)

4.1.3. Big Data Driven network optimization based on QoE

Network optimization function in combination to Data analytics are able to recognize the exact reasons of network problems and faults, and determine the proper action . As a result ,MNOs will achieve high QoE and QoS and objectives of resource optimization and reducing cost. (Zheng et al. 2016).

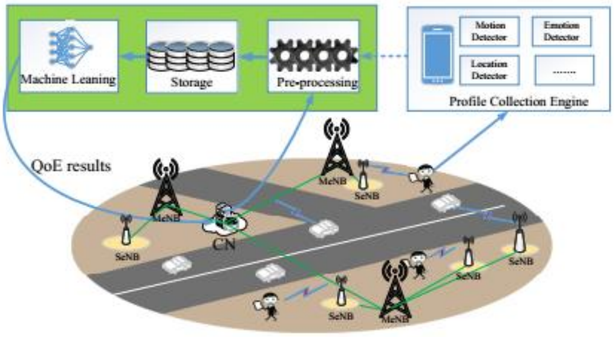


Fig.4.3: Big Data Driven network optimization based on QoE (Zheng et al 2016)

4.1.4. Big Data related to Traffic

Owing to the massive and worldwide usage of mobile internet, the amount of traffic data increased in an unexpected rate. analyzing this data plays a key role in network performance enhancement detection of failures as shown in Fig.4.4 and also in traffic monitoring and load distributed by using prediction modeling and techniques rather than Classical methods and approaches, which are insufficient in handling this amount of data.

In addition, the relation between big data and soft defined network SDN discussed by (Kreutz et al. 2015;Yan, 2015; Cai et al.).

Liu et al .(2015) suggested a new Hadoop platform based system to monitor and analysis and it is implemented in commercial cellular with massive daily input of traffic data and his results proved that system is able to process large scale data and define traffic patterns and user behaviors .

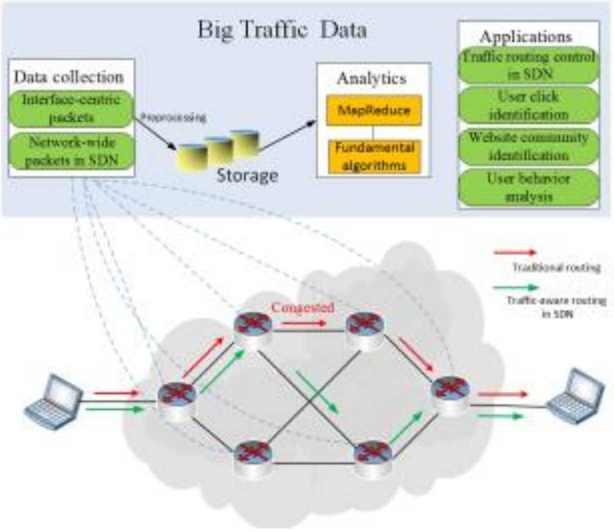

Big Traffic Data
Data collection
Interface-centric packets
Network-wide packets in SDN
Preprocessing
Storage
Analytics
MapReduce
Fundamental algorithms
Applications
Traffic routing control in SDN
User click identification
Website community identification
User behavior analysis
Traditional routing
Traffic-aware routing in SDN
Congested

Fig.4.4: Big Data Driven based on Traffic(He et al. 2016)

4.1.5. Network Planning and Resources Allocation

4.1.5.1. Resource allocation:

Resources requirements are changed from one place to another in certain specific time, and they are predictable when using data analytics. In centralized baseband units, resource allocation prediction provides assistance to define the exact location at the proper time, by getting knowledge of the traffic peak hours (Zheng et al. 2016).

4.1.5.2. Network planning

When planning for future mobile networks, one must consider multiple challenges due to the high complexity of the network to be managed. 4G and 5G networks are depicted by high densification of nodes and heterogeneity of layers, Radio Access Technologies (RAT) and applications (Moysen et al. 2017).

Moysen et al. (2017) presented such a smart network-planning tool that uses Machine Learning (ML) techniques. The applied approach should be able to predict the (QoS) the users experience based on the measurement history of the network. A selection of Physical Resource Block (PRB) per Megabit (Mb) used as primary QoS indicator to optimize should occur. This metric allows the same service to be offered to users by consuming fewer resources.

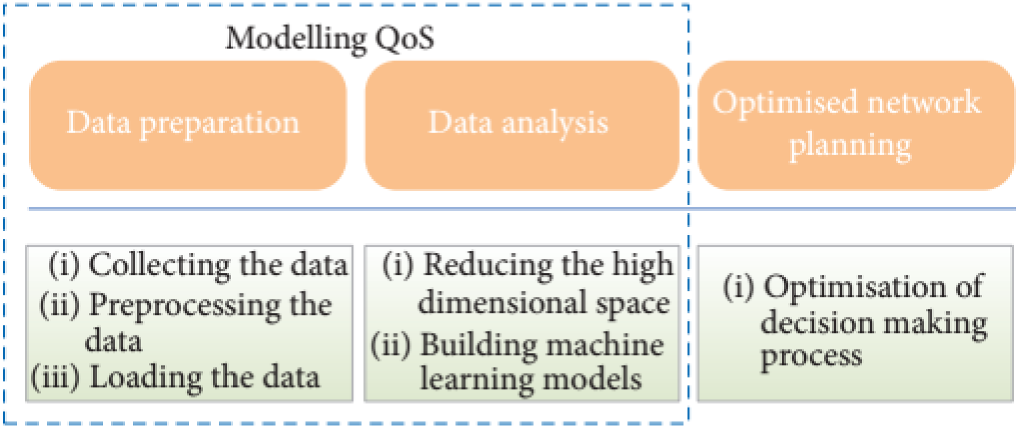

Modelling QoS
Data preparation
Data analysis
Optimised network planning
(i) Collecting the data
(ii) Preprocessing the data
(iii) Loading the data
(i) Reducing the high dimensional space
(ii) Building machine learning models
(i) Optimisation of decision making process

Fig.4.5 Data Driven framework based on QoS Data (Moysen et al. 2017).

For application and testing performance of their work, they introduced two application based on this framework:

1- Locating the small cell in an indoor deployment by using prediction of QoS.
2- Readjusting (antenna tilt) to recover outage due to loss of service and faulty cell problem in the network.

4.1.6. Cache-assisted wireless resource allocation

Future expectation indicates that BS-level caching will have a major role in the future wireless big data processing, because it is simple cost-effective and can reasonably be combined with big data analytics tools. Research efforts towards wireless cache-assisted wireless are still limited and massive studies need to be conducted as there is a shortage for cache-assisted wireless . In addition, effective and optimized integration of various identified big data characteristics in cache assisted network design is an interesting problem that awaits future investigations (Bi et al. 2015). Therefore, the next chapter will focus on caching significant in relation to 5G and Big Data analytics.

CHAPTER 5

5.1. Caching at the edge

As a result of The vast amount of data in 5G network, new different methods and approaches are required to handle these demands of data as the currently applied methods are an ineffective regards to cost, scalability and flexibility. so mobile operators start to move from traditional methods e.g : increase number of base station or nodes , or new spectrum acquiring and go towards applying big data analytics , edge computing with 5G context aware networks (Zeydan et al. 2016).

Dramatic changes in network topology need to be involved according to the vision of IMT-2020 for development 5G , which will achieve throughput enhancement, more than 100 billion connections and almost zero latency (Zhang et al. 2014).

In the term to achieve zero latency, several solutions are introduced to solve such problem and achieve requirements for 5G networks as fig.13 illustrates , there are three main solution based on the main sources of latency these solution is introduced and discussed in details in (Parvez et al. 2014).

The latency is a continuation of core network, radio access network RAN, backhaul and internet as per below equation, for one way transmission time is shown in fig .14 and Equation **Error! Reference source not found.** end to end time will be time T multiplied by 2 (Garcia-Perez et al. 2016).

$$T = T\text{Radio} + T\text{Backhaul} + T\text{Core} + T\text{Transport} \qquad (1)$$

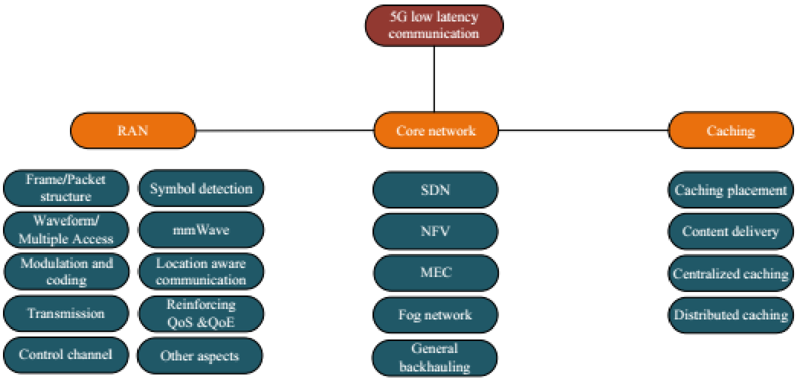


Fig.5.1 Different solutions for achieve Low Latency in 5G (Parvez et al. 2014)

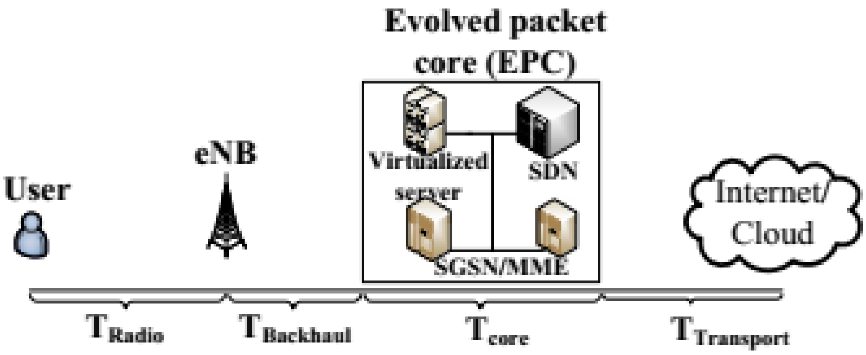


Fig.5.2 Total time and latency end to end contributors (Garcia-Perez et al. 2016)

Our focus here is on Caching solutions, caching and content text awareness. Data acentric networks are considered one of promising solutions to provide a big shift in latency reduction in 5G networks (Parvez et al. 2014). Low latency is important for the quality of services and that of the experience in 5G, and one source of latency can be attributed to the long delay in peak traffic hours due to the high requests of many subscribers during this time (Zhang et al.2015 ; Ioannou& Weber 2016).

Clarity that mobile contents’ nearness to the edge is necessary when the user’s connectivity time is out while engaging in data streaming and/or download. In order to reduce this, data are to be given accessibility to be nearer to users by minimizing the space of content to users and promoting the proper applications and content at the edge which produces greater user experience, by delay minimization (Ji et al. 2016 ; Poularakis et al. 2016).

5.2. Caching Techniques

Caching at the edge of mobile cellular network (BS level and UE) has become an interesting area for investigation in (Paschos et al. 2016), (Ji et al. 2016) and (Poularakis et al. 2016). These caching schemes reduce the latency by fulfilling user content requests without involving core network by using backhaul links.

Fig.5.3 shows different caching schemes: “In fact, each user starts from the nearest source to look for its desired content and proceed until finding it in any of the proposed sources “:

1. Local caching: if the user's equipment starts to access content, the first step is to check in itself when it is confirmed in storage local caching so access for required contents is done without any delay(Boccardi et al.2014).
2. Device to device caching: in case the required content is not found in local storage, the user's request will be checked within its device to device communication range. Once it is available in near devices, it will be delivered to the user(Chen et al. 2016).
3. SBS caching: the user's equipment requested content would be delivered by local SBS(Gregori et al. 2016).
4. MBS caching: once the content is not available from the previous three cases, the content will be delivered by MBS caching(Chen et al. 2016).

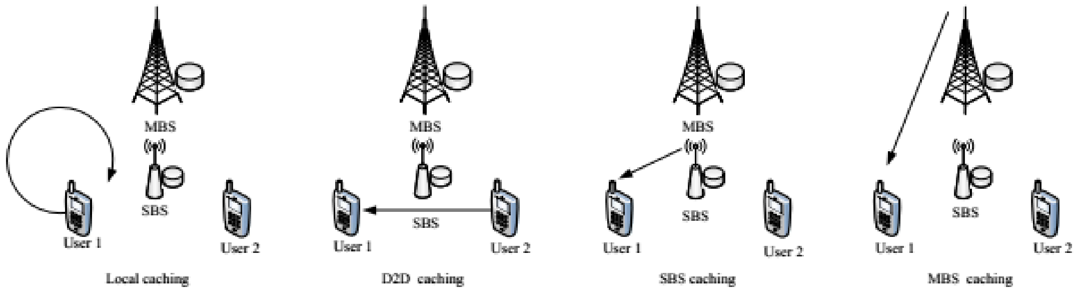


Fig.5.3 Different Caching Scenarios (Parvez et al. 2014).

5.3. Content popularity Based Caching

Popularity often does not depend on content only but also on the users as shown in Fig.16; Therefore substances of great popularity are often likely to be placed on the cache servers in order to enhance the cache access ratio. Nevertheless, user mobility may also make the content in the cache to transform often, giving rise to inefficiency in content caching. So as shown in Fig, the data analytics process needs to evaluate the data in relation to both content and users in order to forecast content popularity or determine correctly (Zeng et al.2016).

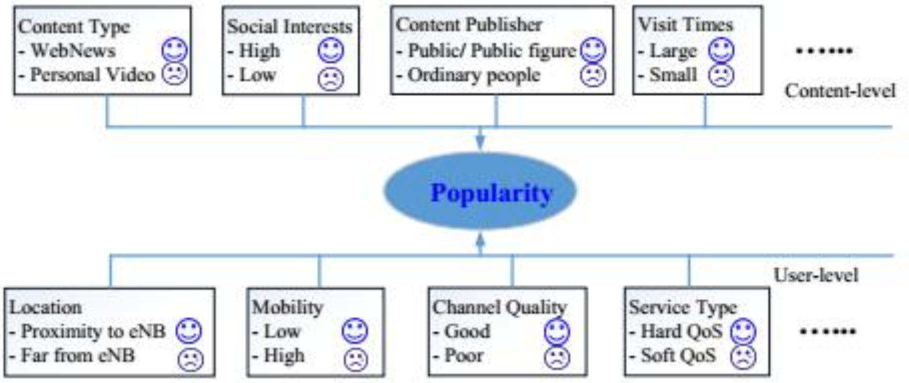


Fig.5.4 Popularity factors from user and content level (Zeng et al.2016).

5.4 Proactive Caching

Bastug et al. (2015) investigated that proactive caching has remarkable and vital part in 5G wireless network by applying various proactive caching seniors in two case studies to explore the important role in the spatial and social structure of the network. They proposed solutions to backhaul congestion problem using proactive caching for files in off-peak demands based on file popularity and correlation between file patterns and users. Secondly, they researched influence of social networks and device-to-device (D2D) communications and proposed a method to derive benefit from the social structure of the network through predicting effective users' proactive cache vital contents and distributing them to their social ties through D2D communications. Their research results reveled that applying of such mechanisms for proactive caching is beneficial for the following reasons:

- Backhaul saving up to 22%
- Higher ratio of satisfied users up to 26%

- Increase of storage capacity at the network edge, which achieves higher benefits from applying these mechanisms.

Bastug et al. (2015) continued their research area keeping in mind that Telecoms started to face complicity management for mobile cellar networks by using classic and current techniques, which are more costly and ineffective.

Therefore, their work aimed to develop new approaches inside framework of Big Data using up to date advantages of storage, edge/cloud computing, and context-awareness.

They identified the optimization of 5G networks through the concept of proactive caching at BTS (base stations). They investigated benefit of proactive caching regarding backhaul off loadings and request satisfaction.

They analyzed mobile traffic user data collected from telecom operator BTSs to estimate context popularity on Big Data platform, and investigated the benefit of proactive caching on BTS through numerical simulation.

Their research results showed that proactive caching achieved under 10% of content rating and storage size of 15.4 GB:

- 100% request satisfactions
- 98 % of the backhaul.

Su et al. (2015) represented a model based on Small Cell Networks (SCNs) and indicated that QoE matrix presented an algorism called “PropCaching” based on proactive caching popularity to overcome the problem of capacity limitation.

Comparison between performance of Proactive caching and Random caching results: Propcahing is 85%, higher than the Random one. Also, the usage of backhaul is decreased by 10%,under the assumption of SCN (Small Cell Network) with high storage and limited capacity for backhaul links.

CHAPTER 6

6.1. Methodology and Techniques adopted for practical

An empirical scientific approach “Fourth Tradition “ which is different than tradition ones qualitative quantitative and mixed methods (Fan, Han, and Liu, 2014), this approach is based on knowledge discovery fundamental by applying predictive, visualization and data mining tools (Daniel, 2017).

This chapter is divided to two parts:

6.1.1 First part

The first part focuses on applying common big data driven framework, which consists of four processes data collection, data preprocessing ,data analysis and application , more details about several frameworks were discussed in chapter 2. with special attention to Call Details Records (CDRs) as an output of Data collection process due to the importance of these datasets for both scientific and industrial aspects.

6.1.2 Contribution of CDRs Datasets in industrial and scientific aspects

Call Data Records was usually required from finical aspects for mobile operators, but in the Big Data era, it started to receive huge attention in both scientific and industrial aspects, because CDRs datasets are rich with information regarding communication between millions of users and where, with whom, when and how they communicate.

CDRs datasets analysis has become an interesting area for research (Blondel et al. 2015) as it provides numerous uses of these datasets for different research purposes resulting in the development of different methods of handling these kinds of datasets and in the creation of different analytic techniques and multiple types of analysis from different perspectives using big data analysis tools.

As for the telecommunication industry, especially the mobile network operators, CDRs contributes massively in the improvement of business and operation performance for MNOs.

Orange is one of the biggest Telecom operators; they launched the first challenge in 2013 “D4D challenge”. Through this challenge, they invited different candidates from around the world and provided access to massive CDRs datasets for developing purposes of their infrastructures and customer satisfaction as a source to gain more revenues.

The scientific work resulted in successful outcomes which encourage the company to launch a second challenge in April 2015 during the NET Mobile conference (Blondel et al. 2015).

Telecom Italia challenge was another example when they launched the first edition of the Big Data Challenge in 2014 (Telecom Italia challenge, 2014) and this will be discussed in details in the “Data Source” part.

6.1.3. Data Source

6.1.3.1. Telecom Italia data-set

Italian Telecom launched the first edition of the Big Data challenge at the start of the year 2014. The Big Data field is introduced in order to compete a new design for stimulating the formation and expansion of innovative ideas based on new technologies.

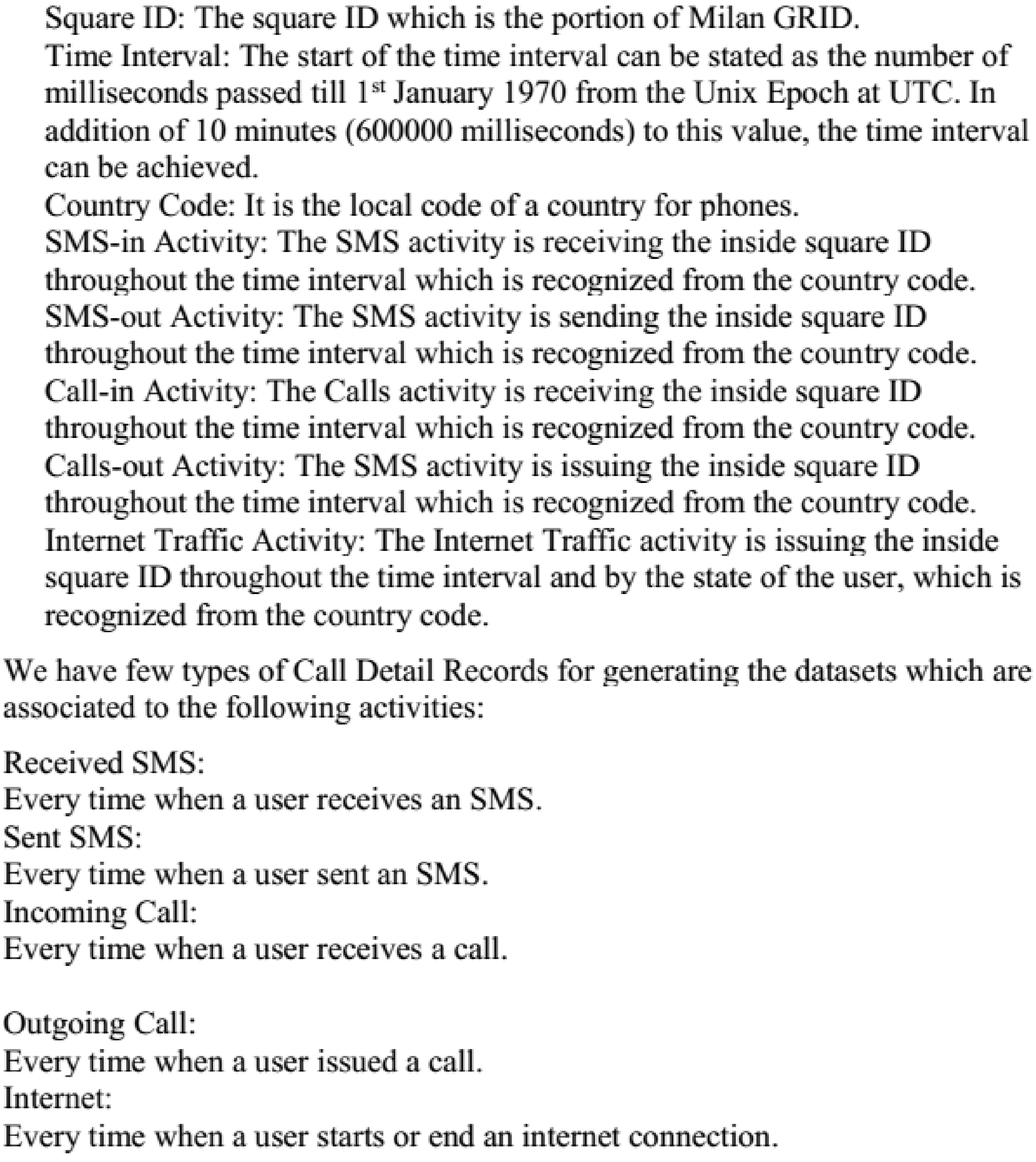
The attention of the partakers was taken to account before releasing the first Dataset. But the demand is being increasing rapidly for the datasets at the end of the competition which became an initiative towards “Open Big Data”. According to (Data Telecom, 2014), the Datasets were published freely to improve the uses of the Datasets in the entire society.

These Datasets were part of the “Telecom Italia Big Data Challange” in 2014, which was an ironic and having combination of more telecommunications including social networking, news, weather forecasting and data of the electricity from the Milan city and one of the province of Trentino (Italy). Telecom Italia has formed an original Dataset with connotation of some Labs which includes the following institutes:

- MIT Media Lab
- Polytechnic University of Milan
- Trento and Trento RISE University
- EIT ICT Labs
- Fondazione Bruno Kessler
- SpazioDati
- Northeastern University

Between November-December 2013, the Dataset were holding millions of records. These records were taken out from telecommunication which includes social network events, energy, weather forecast and transportation of public-private people. Some more details by (Barlacchi et al.2015) about the construction and collection the data.

The information about the activity of telecommunication of Milano has been provided in the above dataset. This dataset is the outcome of calculation upon the CDRs produced in Milano city by the "Telecom Italia" cellular network. The user activities are recorded by the Call Detail Records (CDRs) for the management of network and billing. The dataset consists of the information below(Telecom Italia challenge, 2014):



Square ID: The square ID which is the portion of Milan GRID.
Time Interval: The start of the time interval can be stated as the number of milliseconds passed till 1st January 1970 from the Unix Epoch at UTC. In addition of 10 minutes (600000 milliseconds) to this value, the time interval can be achieved.
Country Code: It is the local code of a country for phones.
SMS-in Activity: The SMS activity is receiving the inside square ID throughout the time interval which is recognized from the country code.
SMS-out Activity: The SMS activity is sending the inside square ID throughout the time interval which is recognized from the country code.
Call-in Activity: The Calls activity is receiving the inside square ID throughout the time interval which is recognized from the country code.
Calls-out Activity: The SMS activity is issuing the inside square ID throughout the time interval which is recognized from the country code.
Internet Traffic Activity: The Internet Traffic activity is issuing the inside square ID throughout the time interval and by the state of the user, which is recognized from the country code.

We have few types of Call Detail Records for generating the datasets which are associated to the following activities:

Received SMS:
Every time when a user receives an SMS.
Sent SMS:
Every time when a user sent an SMS.
Incoming Call:
Every time when a user receives a call.

Outgoing Call:
Every time when a user issued a call.
Internet:
Every time when a user starts or end an internet connection.

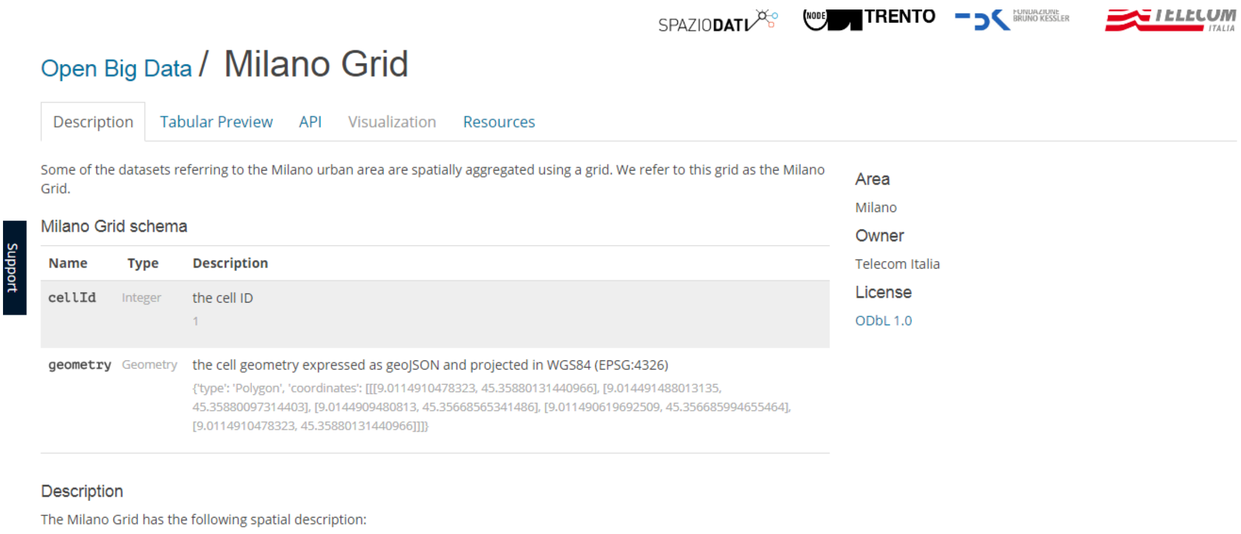
Throughout the similar internet connection one of the below restrictions is reached

- 15 Minutes after producing the final CDR
- 5 MB after producing the final CDR

This Dataset was formed by accumulating the above stated records, to deliver Internet Traffic, SMSs and Calls activities. The level of collaboration between users and mobile network is calculated through this. For instance, more SMS sending by a user results more activity of the SMSs which sent by the user. The SMSs and Calls activities are having the similar scale of sizes “Therefor they are analogous to each other”.

According to (Data Telecom , 2014) Datasets are combined in a four-sided cells grid as shown below:

> -“The area of a Milan is composed of a grid overlay, which is 1,000 squares having the size of 235*235 meters.”
>
> -“The grid is probable with WGS84 (EPSG:4326) standard.”



Some of the datasets referring to the Milano urban area are spatially aggregated using a grid. We refer to this grid as the Milano Grid.

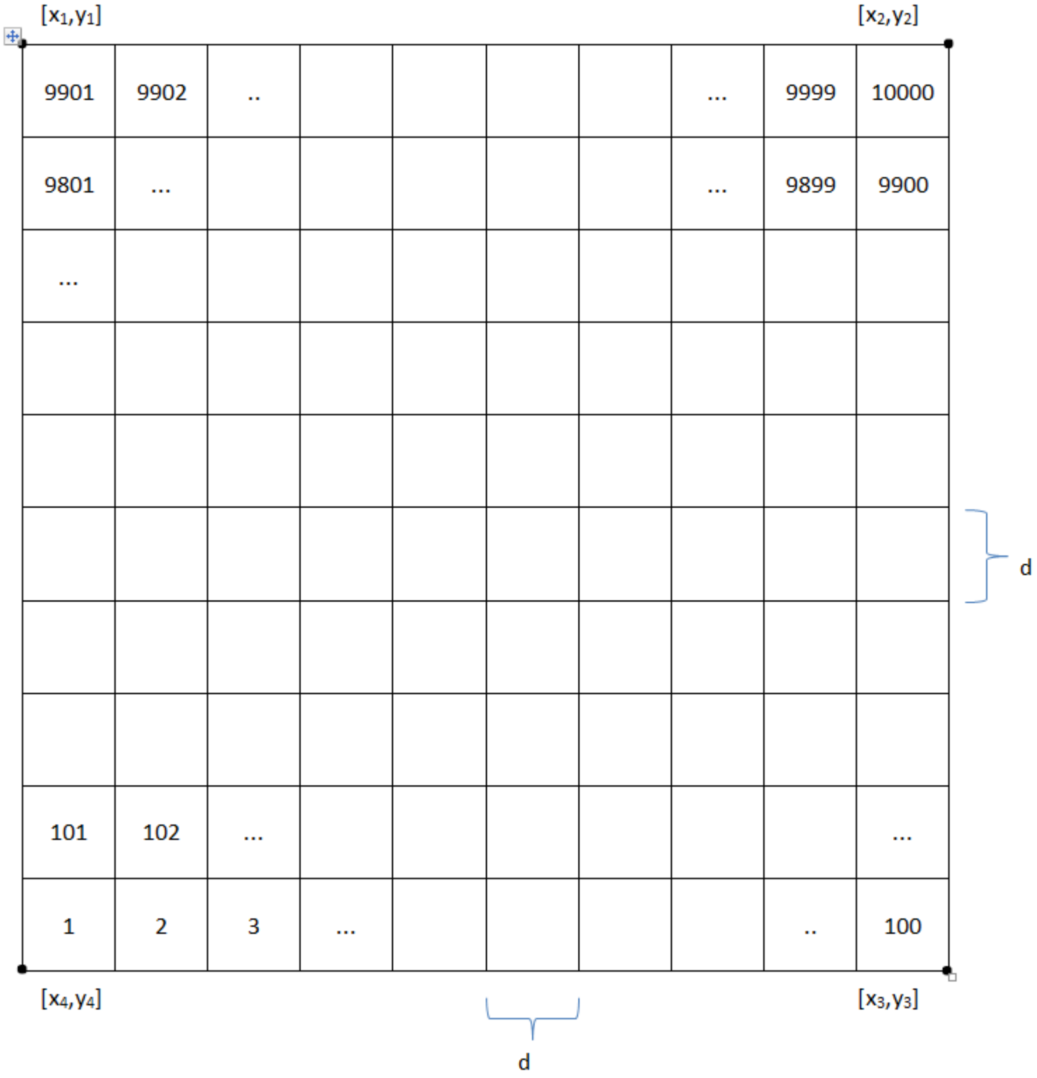
Milano Grid schema

Name	Type	Description
cellId	Integer	the cell ID 1
geometry	Geometry	the cell geometry expressed as geoJSON and projected in WGS84 (EPSG:4326) {'type': 'Polygon', 'coordinates': [[[9.0114910478323, 45.35880131440966], [9.014491488013135, 45.35880097314403], [9.0144909480813, 45.35668565341486], [9.011490619692509, 45.356685994655464], [9.0114910478323, 45.35880131440966]]]}

Area
Milano
Owner
Telecom Italia
License
ODbL 1.0

Description

The Milano Grid has the following spatial description:



9901	9902	..					...	9999	10000
9801	...						...	9899	9900
...									
101	102	...							...
1	2	3	...					..	100

Fig.6.1: Milan GRID

As previously shown, squares are numbered with ids. The square id numbering starts from the bottom left corner of the grid and grows till its right top corner.

In the following picture, the grid is overlaid to the city of Milan city.

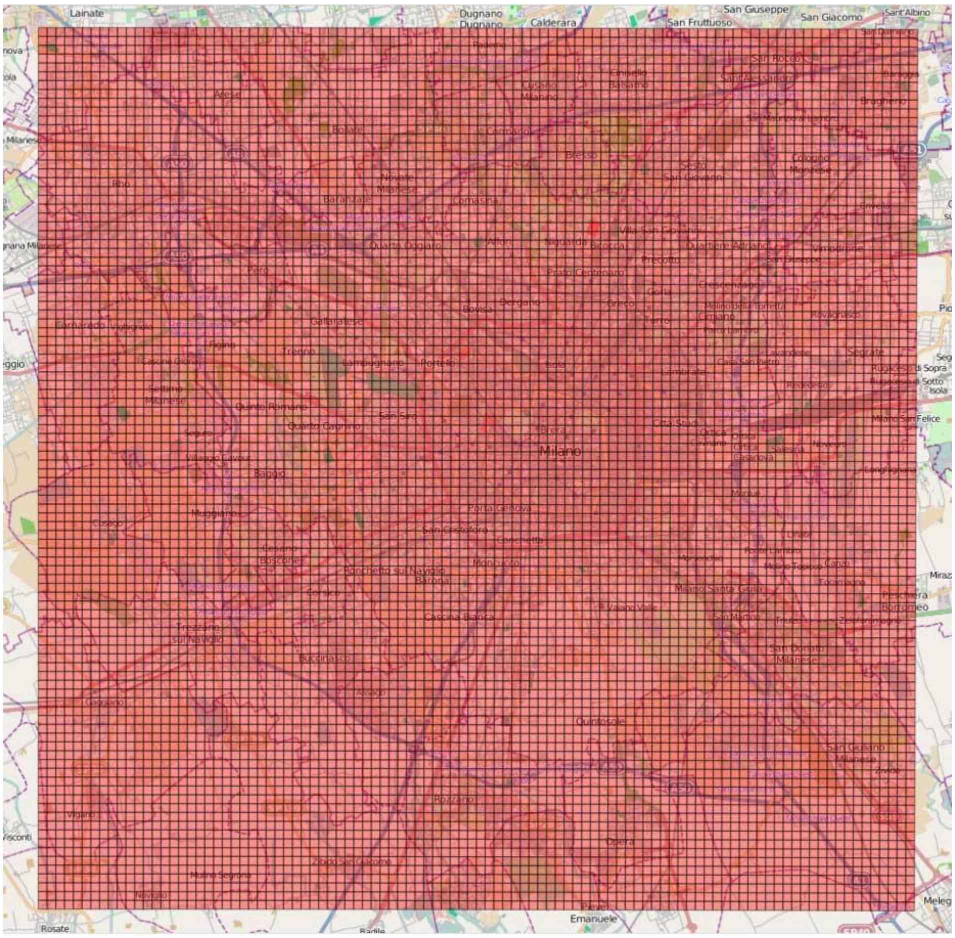


Fig.6.2: Milan GRID on Milan city Map

6.1.4. Methods and techniques:

Different methods and techniques are used for analysis of datasets in order to do clustering, prediction and data visualization. Data tests were also performed to check the best suitable model for application on data. The techniques are detailed below.

6.1.4.1 Clustering :

In the field of data mining, Clustering procedures constitute crucial tools (Xu et al. 2005) Because of their notable high abilities to deduce connections between data objects, they have been to a great extent utilized by scientists for the investigation of datasets for mobile tracing.

The techniques of clustering are either accept a separated approach or hierarchical approach(Xu et al. 2005). Hierarchical techniques arrange items into a hierarchical structure, which can visually be represented diagrammatically. Hierarchical algorithms can follow an organized method or separated one. However, partitioned clustering algorithms eg. ISODATA and K-means, directly group objects into numbers of categories K. A relevant comment is that hierarchical algorithms can be used also in categorizing objects into a definite number of categories, which can be finished by ending the algorithm at the required point/level. In all instance, there is no stipulated rule to determine the definite number of categories, the decision still remains either ascertained definitely relying on the accordance to certain clustering quality measures or knowledge about the data (Naboulsi,2015).

In the mobile phone datasets analysis, k-means is the most frequently used partitioned clustering algorithm is k-means algorithm, which aims at grouping objects in X into a set of k clusters f, with the aim of reducing the inner-cluster distances. The algorithm functions as follows Equation (2) begins by previous initialization of objects k-partition based selections randomly.

(2)

$$\mathbf{m}_k = \frac{1}{|\mathbb{C}_k|}\sum_{i\in\mathbb{C}_k}\mathbf{x}_i$$

For the analysis of mobile phone datasets, k-means is most applied and have good results based on previous work of other researchers in the same area (Lin et al.2007; Willkomm, et al. 2008;Soto et al.2011 , Shafiq et al.2012; Zhang et al. 2012;Mucelli et al.2015; Csaji et al.2013 ;Becker et al.2011).

6.1.4.2. Data visualization

Using right type of visualization brings insight to data analysis process. Although we can use many ways to explore data but one of the best ways is to visualize it and very large quantity of papers and research on how to make acute observation and deduction on data visualizations has been made available.

Explanatory Data Analysis (EDA) and Execute in a proper order to study and expound the presented dataset. Often times, the dataset is sometimes clear and simple for forward steps.

The purpose of detailed data analysis are (Wickham et al. 2014);

- To explain outcome of data analysis.
- To examine data patterns.
- To authenticate data.
- Apprehending the functions and restrictions of data.
- To inspect data need for cleaning.

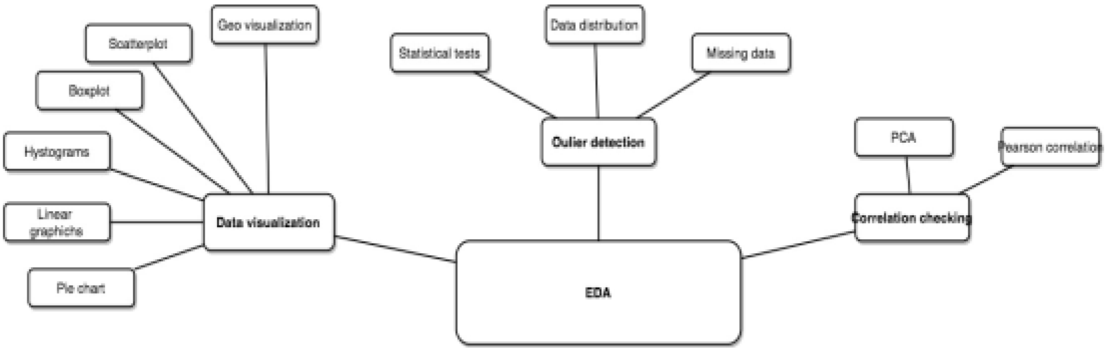


Fig.6.3: Explanatory Data Analysis EDA

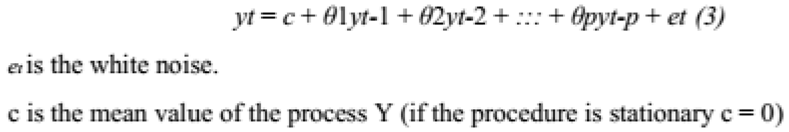
6.1.4.3. Prediction

Prediction is consider an important part for mobile operators as it is essential to take decision related to network optimization and one of main parts of ML. One of the most famous prediction algorithms is described in (Zhang, 2003) and it is called Autoregressive Integrated Moving Average (ARIMA) model.

6.1.4.3.1. The autoregression model :

In this model, using past values combination on linear base for estimating a specific variable (Hyndman et al. 2013). The term auto regression is a regression of this variable against itself. An autoregressive model is typically confined to stationary data, thus, data should have constant mean value, variance and autocorrelation constant(Trioa ,2106).

In this model the yield variable directly on its previous values and on a stochastic term. The notation AR(p) shows an autoregressive model of order p, and it is characterized as Equation (3):



$$yt = c + \theta 1yt-1 + \theta 2yt-2 + ::: + \theta pyt-p + et \quad (3)$$

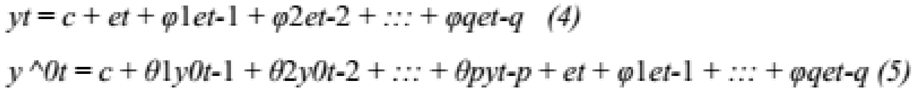
e_t is the white noise.

c is the mean value of the process Y (if the procedure is stationary c = 0)

ARIMA is one of important prediction models for time series data practically and statically.

ARIMA (p; d; q) model can be obtained using below equation (4) by combing difference in moving average model and an autoregresstion model.

As moving average model (5) using pervious forecast errors instead of past forecast values in autoregreesion model



$$yt = c + et + \varphi 1et-1 + \varphi 2et-2 + ::: + \varphi qet-q \quad (4)$$

$$y\ \hat{}0t = c + \theta 1y0t-1 + \theta 2y0t-2 + ::: + \theta pyt-p + et + \varphi 1et-1 + ::: + \varphi qet-q \quad (5)$$

6.1.4.3.2. LSTM

Recurrent Neural Network (RNN) model with layers based on Long Short Term memory blocks is usually called LSTM , it is consist of memory blocks and by Backpropagation can be trained , problem of gradient is decreased over time exponentially in this model (Sundermeyer et al.2012) .

Neural networks are build by basic components working in parallel and are made out of slender inputs that are for the most part not known. We can train a to get a specific function by altering the values of the associations (weights) between components. As in nature, the system work is determined to a great extent by the associations between component (Trioa ,2106).

Three main thing are important need to be consided in design introduced by (Troia 2016):

Learning rule: The learning principle determines the manner by which the neural system's weights alter over time;

Activity rule: it is weights depended depend on the weights (the parameters) in the system. Most neural system models have local guidelines and characterize how the activity of the neurons change in response each other;

Architecture: The design determines which factors are related to the network and their topological relationships.

RNN and ARIMA both performed better than other models for prediction of time series data since they provide lower predection errors and higher correct reversal detection percentage, but in short term prediction they give more accurate results , and it was found that RNN sometimes performs better with optimal factor of weighting (Ho et al. 2002).

6.1.5. Data Analysis

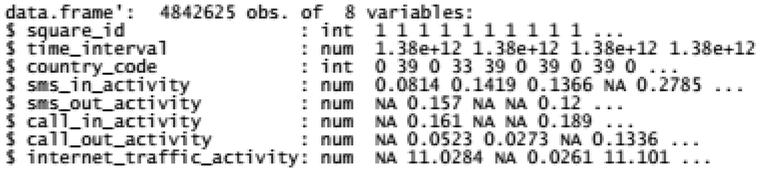
6.1.5.1. First Analysis:

The objective of this analysis is to figure out: the total activity, define peak hours and top country code dialed by subscribers on daily activity sub-set data in November of 600 square grids, from main datasets which are already described in the “Data source” section. As part of first analysis, preprocessing and clustering were done.

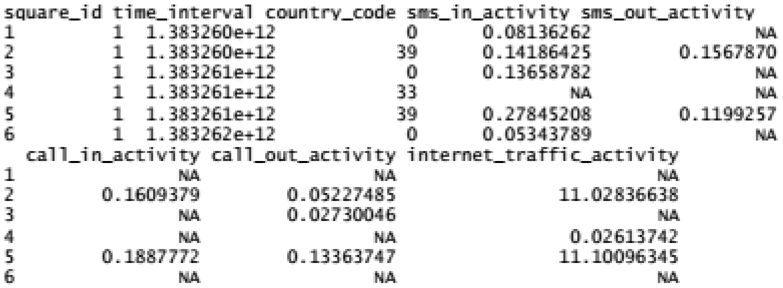
In order to achieve this analysis, the practical process includes preprocessing of data, data visualization using EDA and clustering as shown in the below sample of data.

6.1.5.1.1. Sample of Data :

As shown the following we have 8 variables ,



```
data.frame':  4842625 obs. of  8 variables:
 $ square_id               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ time_interval           : num  1.38e+12 1.38e+12 1.38e+12 1.38e+12
 $ country_code            : int  0 39 0 33 39 0 39 0 39 0 ...
 $ sms_in_activity         : num  0.0814 0.1419 0.1366 NA 0.2785 ...
 $ sms_out_activity        : num  NA 0.157 NA NA 0.12 ...
 $ call_in_activity        : num  NA 0.161 NA NA 0.189 ...
 $ call_out_activity       : num  NA 0.0523 0.0273 NA 0.1336 ...
 $ internet_traffic_activity: num  NA 11.0284 NA 0.0261 11.101 ...
```



```
square_id time_interval country_code sms_in_activity sms_out_activity
1         1  1.383260e+12            0      0.08136262               NA
2         1  1.383260e+12           39      0.14186425        0.1567870
3         1  1.383261e+12            0      0.13658782               NA
4         1  1.383261e+12           33              NA               NA
5         1  1.383261e+12           39      0.27845208        0.1199257
6         1  1.383262e+12            0      0.05343789               NA
  call_in_activity call_out_activity internet_traffic_activity
1               NA                NA                        NA
2        0.1609379        0.05227485               11.02836638
3               NA        0.02730046                        NA
4               NA                NA                0.02613742
5        0.1887772        0.13363747               11.10096345
6               NA                NA                        NA
```

6.1.5.1.2. k-means clustering

K-means was applied on total activity and activity time per hour to recognize the clusters with three categories : first category clusters with high usage for 24 hours , clusters with low usage for 24 hours and clusters with high usage during specific hours and identify these specific activity hours . fig6.6 illustrates clustering with activity time

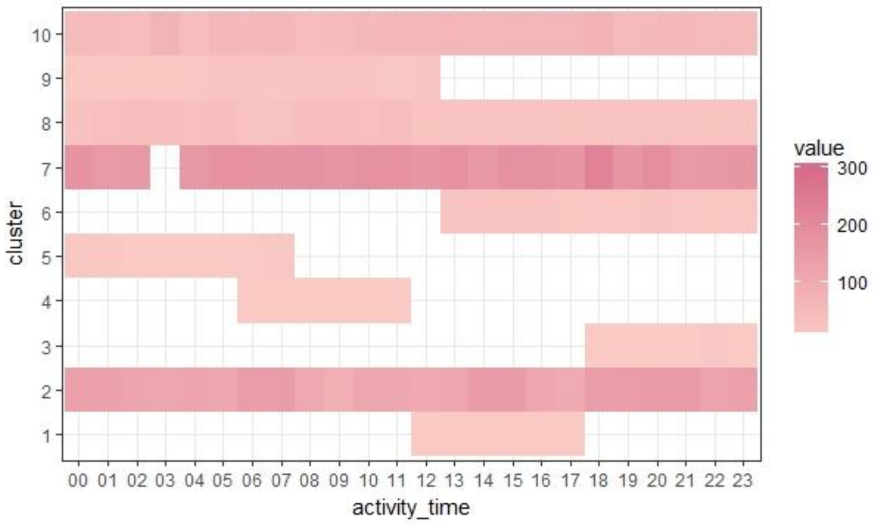


Fig.6.6: Clustering with activity hours

6.1.5.1.3. Data preprocessing

Data preprocessing is an essential steps before starting the analysis and is focusing total activity. Therefore, data processing was achieved through:

- Adding new variables (activity date and total activity).
- Data cleaning and detection of missing value.
- Determine date time from time interval

```
  activity_date  total_activity
1    2013-11-01      0.08136262
2    2013-11-01     11.54023043
3    2013-11-01      0.16388829
4    2013-11-01      0.02613742
5    2013-11-01     11.82175589
6    2013-11-01      0.05343789
```

In this analysis, we applied “Exploratory Data Analysis” EDA (Ehrenberg, 2012) as we mentioned in methods section .

a. Top square girds per total activity was carried out to identify which highest square gird id reveled the highest daily total activity during this day as shown in the following fig.6.4.

b. Total activity vs activity time to identify peak hours with 24 hours as shown in fig. 6.5.

c. Country code

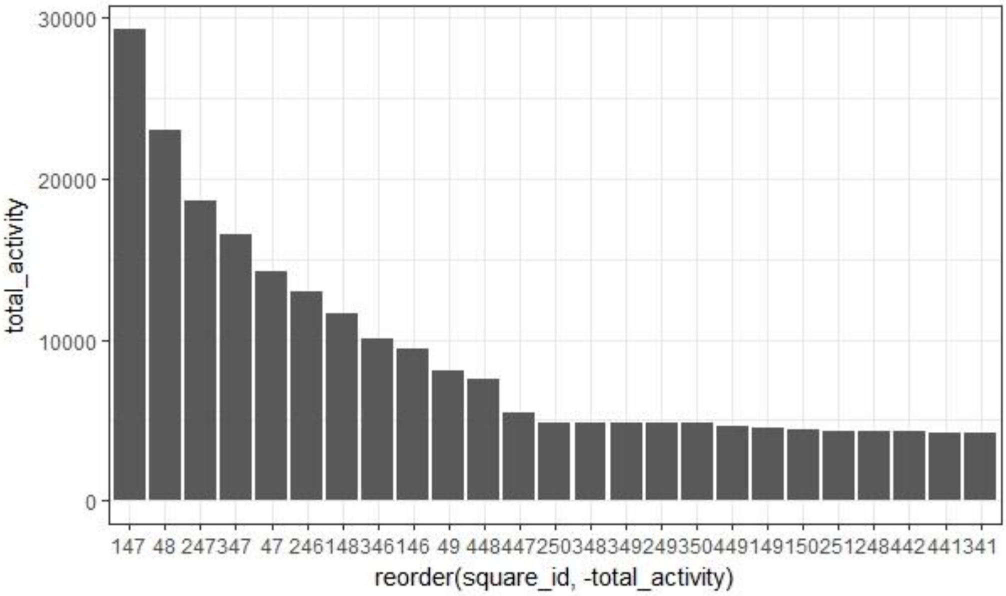


Fig.6.4: illustrating the highest Square gird regards to total activity

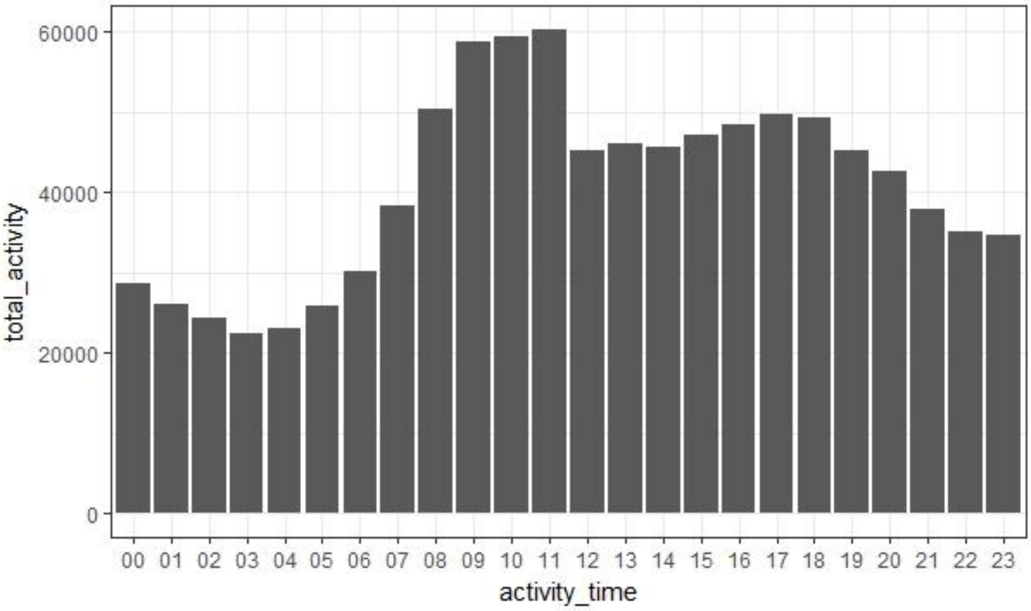


Fig.6.5: shows total activity peak hours during the day

country_code	count
39	72000
0	55776
46	13520
33	7183
41	5121
49	5010

Table 6.1: Country code with highest activity

6.1.5.1.4 Results

Results of the first analysis showed in respect to activity time and total activity ,peak hours during this day are 09:00 AM, 10:00 AM and 11:00 Am , and 03:00 AM is less activity hour during this day as observed in fig.6.6 , and squire gird ID number 149 is the highest total activity during this day as per fig. 6.5. From heat map cluster fig. 6.6 it is clear that clusters 7, 2 ,and 10, high usage over 24 hours and cluster 3,1 and 4 have some activity at different times. And highest activity country code during this day showed in table 1 and from highest to lowest are (39,0,46,33,41,49).

6.1.5.2. Second Analysis

This analysis is aimed to illustrate and compare weekly internet usage in November for three selected cells ID representing several areas for different categories in Milan city .

- Duomo (corresponding to downtown)
- Bocconi (corresponding to university)
- Navigli (corresponding to night life)

.

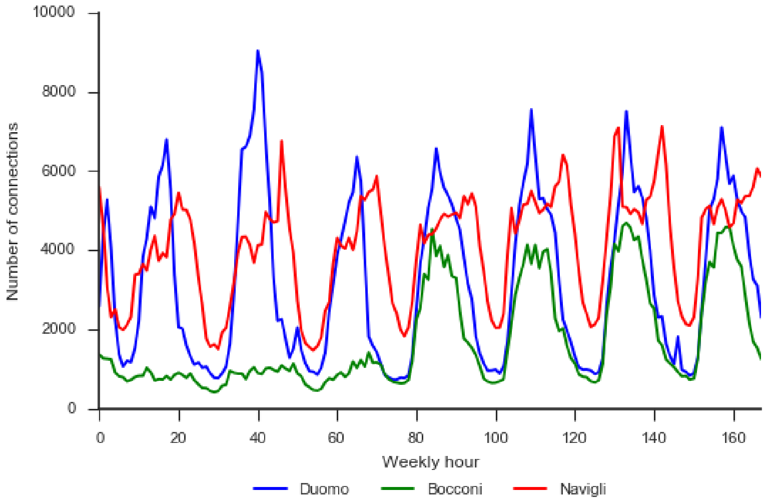


Fig.6.7 weekly internet traffic for different three regions

6.1.5.2.1 Results

The results of second analysis inducated , The peak of the Duomo (downtown) is earlier than the Navigli (night live) Bocconi have less phone calls on the weekends The volume of calls on the weekends decreased as illustrated in fig.6.7.

6.1.5.3. Third Analysis

Since forecasting is an important, factor in efficient network planning and support in optimization decision and resource allocation. Therefore, the objective of this analysis is focusing on internet traffic prediction.

In this analysis, Three different methods were employed for internet usage modeling and prediction .

first model is ARIMA autoregressive model , second is neural network model LSTM long short term modeling and last model is based on previous model introduced in Kaggle Competition and we validated this model on different weekly based data to investigate if modeling for one week is considered effective and enough to have the same results and can be applied on different dataset collected in several intervals of time .

Results of each model are shown in the following :

6.1.5.3.1. ARIMA :

ARIMA (2,1,0) is applied for one week data sets , we will focus first on specific three cell IDs for main three regions and below are obtained results :

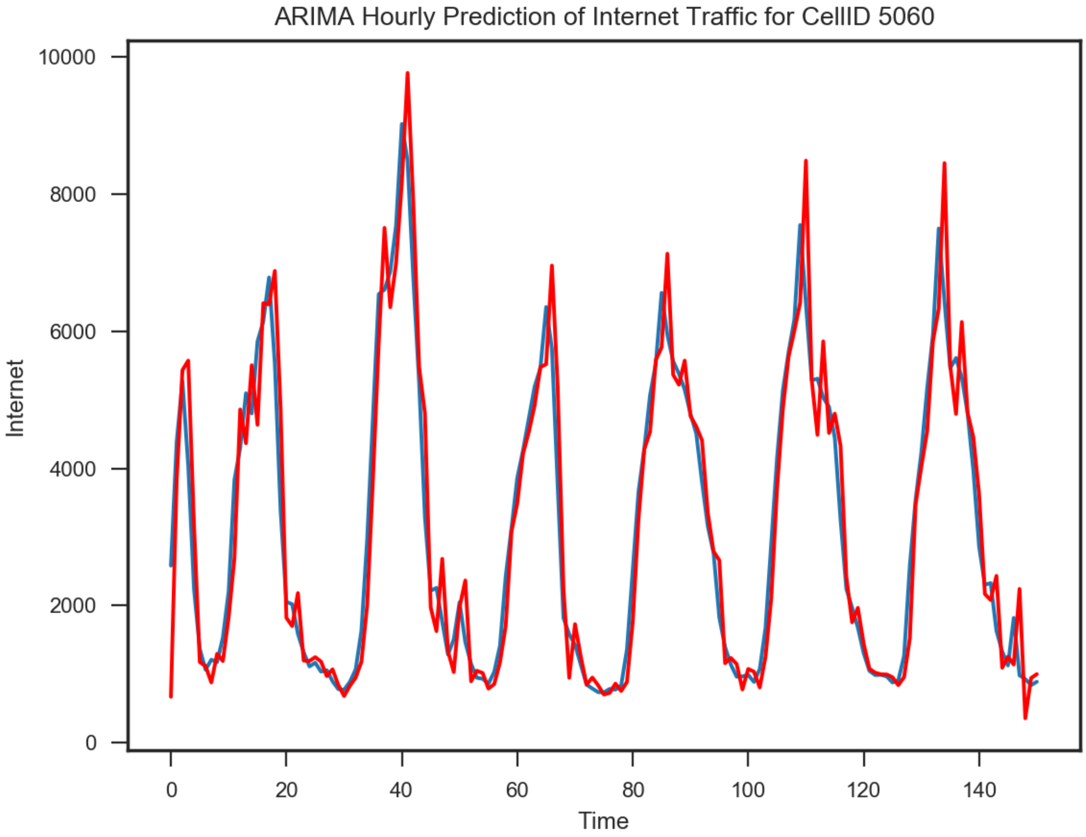


Fig .6.8 ARIMA Hourly Prediction of Internet Traffic for Cell ID 5060

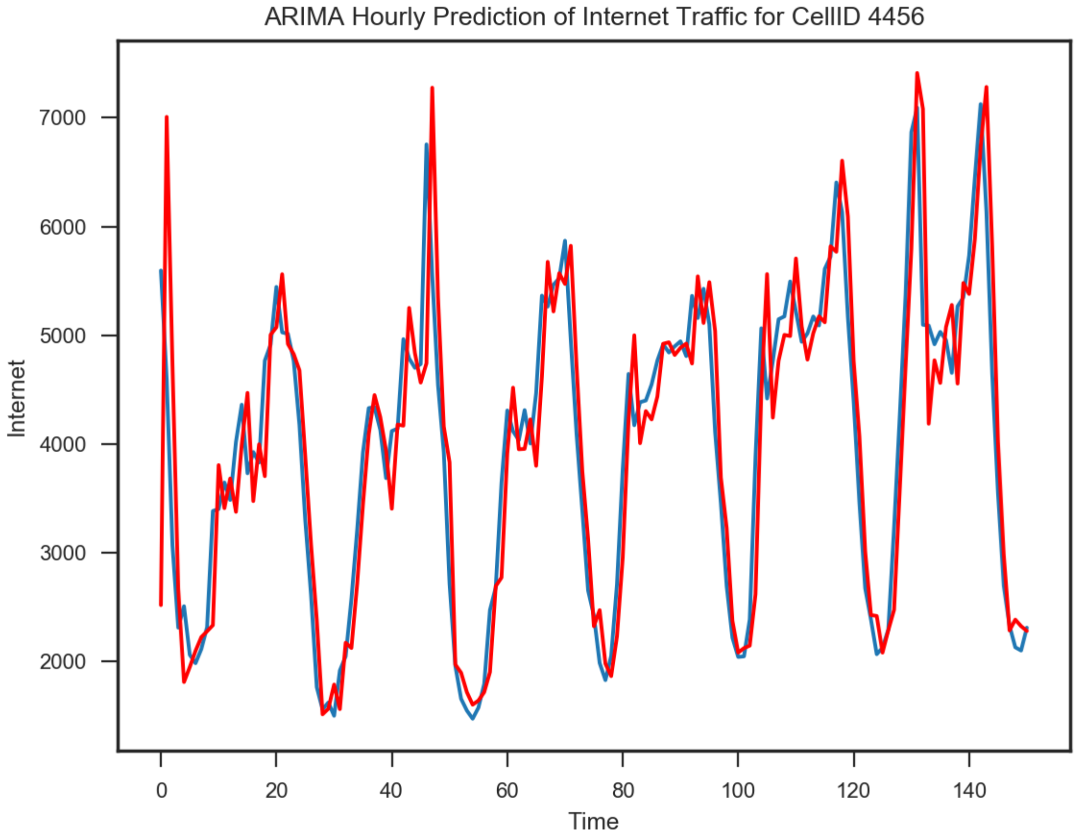


Fig .6.9 ARIMA Hourly Prediction of Internet Traffic for Cell ID 4456

After that we applied it on all cells (9999 cells) and we got below results :

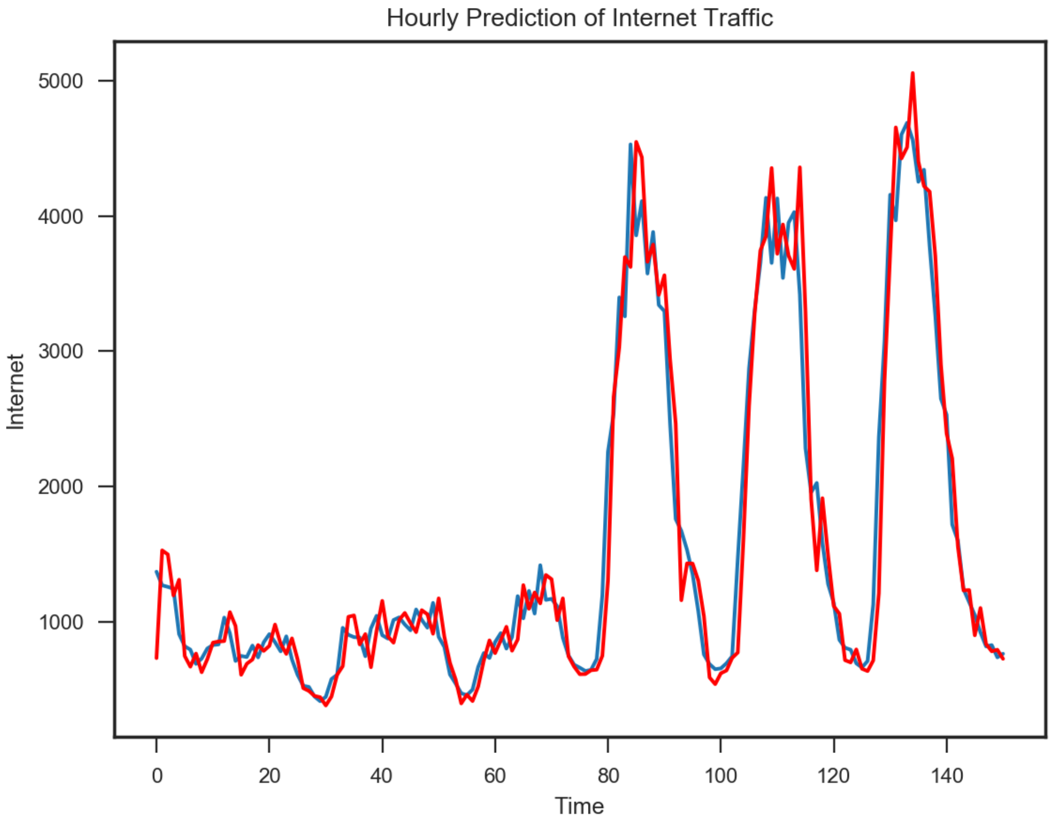
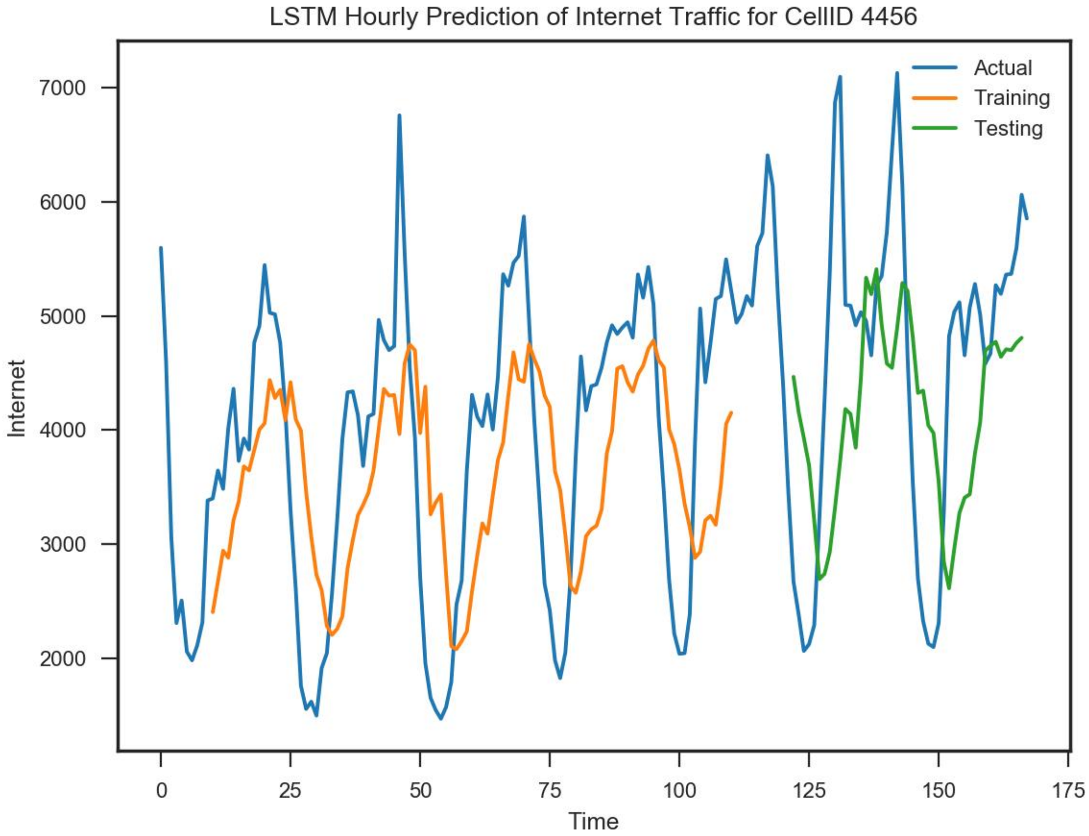


Fig .6.10 ARIMA Hourly Prediction of Internet Traffic

6.1.5.3.2. LSTM

In this model consists of one input for visible layer and 4 blocks in hidden layer, and single value prediction in output layer. We applied LSTM on the same week data which used in ARIMA models and we got below results for each cell ID internet usage prediction:



Fig,6.11 LSTM results for Internet Traffic Prediction per cell

6.1.5.3.3. Third prediction model (validated model):

This model is based on the pervious analysis which shows data sets is periodically data per cell every 24 hours and internet traffic is SIN behavior. In addition, this model is introduced in Kaggel competition where it applied on different period in contrast to our data.

Validation of this model was achieved to be adapt on our datasets in different periods and figure out if it can works for all period, or just suitable for datasets on specific period .validated model was applied using first cell ID for downtown are traffic :

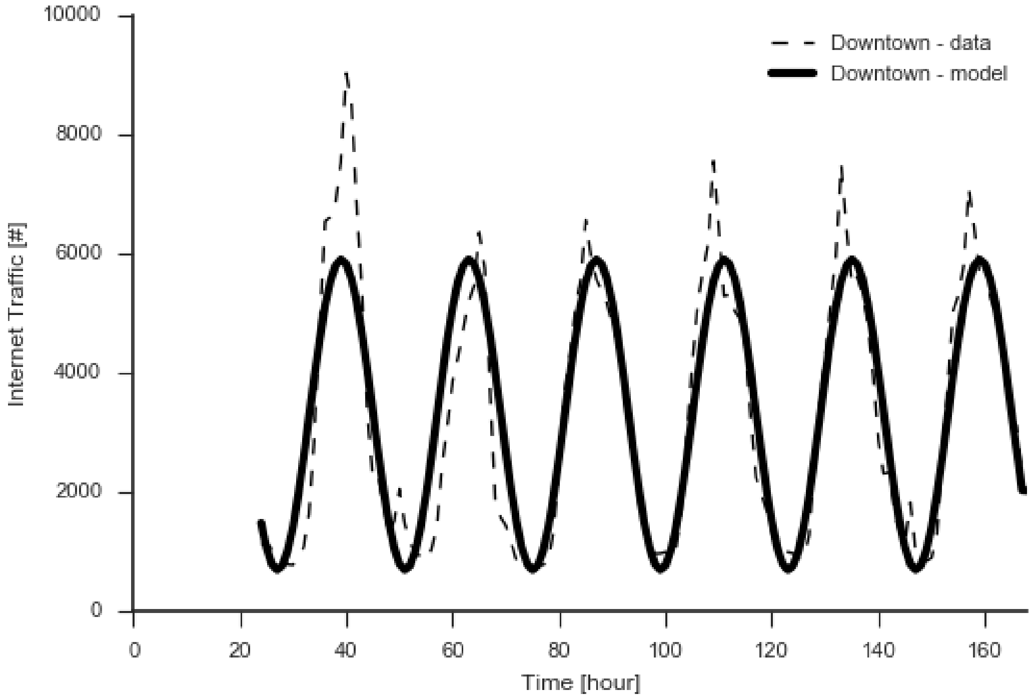


Fig.6.12 internet traffic real data and model results downtown area

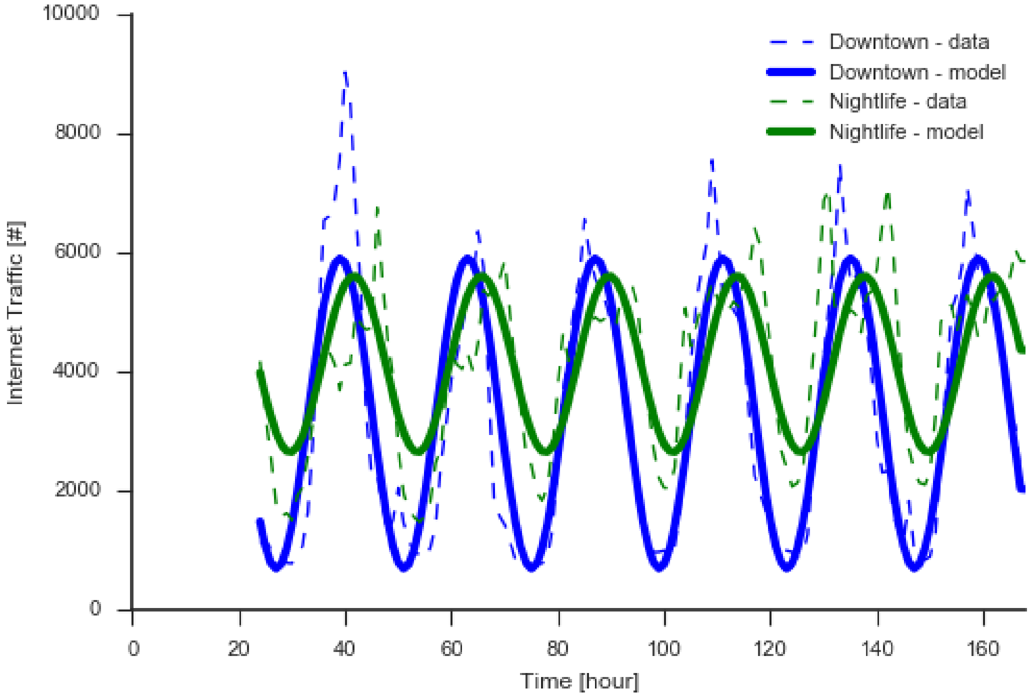


Fig.6.13Internet traffic real data and model results downtown and nightlife areas

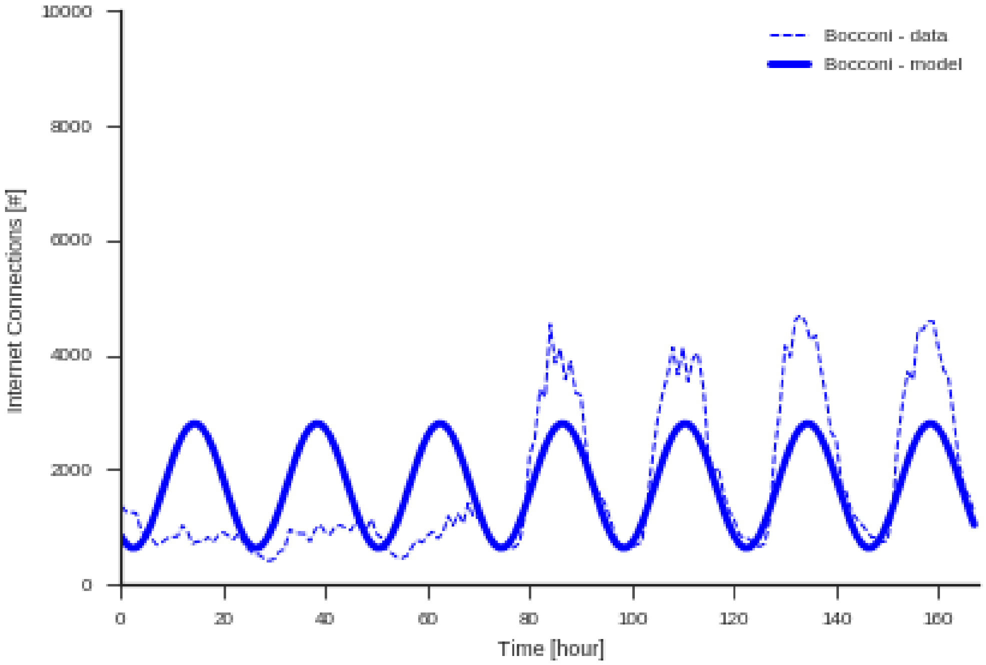


Fig.6.14 Internet traffic real data and model results university area

6.1.5.3.4. Results

Results of the third analysis based which we applied three models for prediction Internet traffic based on one week data and hourly, results explained that ARIMA prediction model is accurate prediction for chosen three cells and with training data set 70% and 30% test set as shown in fig.6.8 ,6.9 ,6.10. while the results of LSTM prediction 69% training set and 21 % test sets was not sufficient in data/cell ID as shown in fig. 6.11. the validated model the obtained results indicated that this model is suitable for all datasets (9998 cells), except for Boccioni area data still have some issue as shown in fig.6.12,6.13,6.14. Therefore, it was proven that this model fits for most of the datasets. At the same time, we investigated peak hours of internet traffic using the model and we can see that downtown peak hour accrued at 2 pm and night life at 6pm it expected since usually downtown traffic at evening start to decrease and night life areas start to have more activity in the evening.

6.1.6. Second Part

The second part is related to caching using WINTERsim simulation tool to figure out the results from different caching seniors and check the different parameters affecting network performance.

6.1.6.1 Performance evaluation of caching

In what follows we provide performance evaluation for different caching strategies and compare the results obtained using WINTERsim simulation tool.

To reflect the actual parameters of 3GPP in this simulation framwork we used the following parameters:

- LTE Release 12, 20MHz channel
- Base Station scheduler type: round-robin
- 3 Pico BSs, 30 UE - BS and UEs are uniformly distributed across the scenario of the size 100x100 m.
- Both BSs and UEs have isotropic antennas, BS tx power = 30 dbm (power control switched off), UE tx power = 23 dbm
- Traffic type: CBR, 100-byte packets with 15Mbps per UE
- None-cached traffic latency distribution: Poisson-based, 14ms expectation value (expectation is measured for youtube servers)

To numerically evaluate the importance of different caching schemas listed before on Fig 5.3, we have performed three different simulation setups. We did not simulate the first local caching since its performance depends on the hardware and software of the local device, we analyzed D2D, SBS and MBS scenarios.

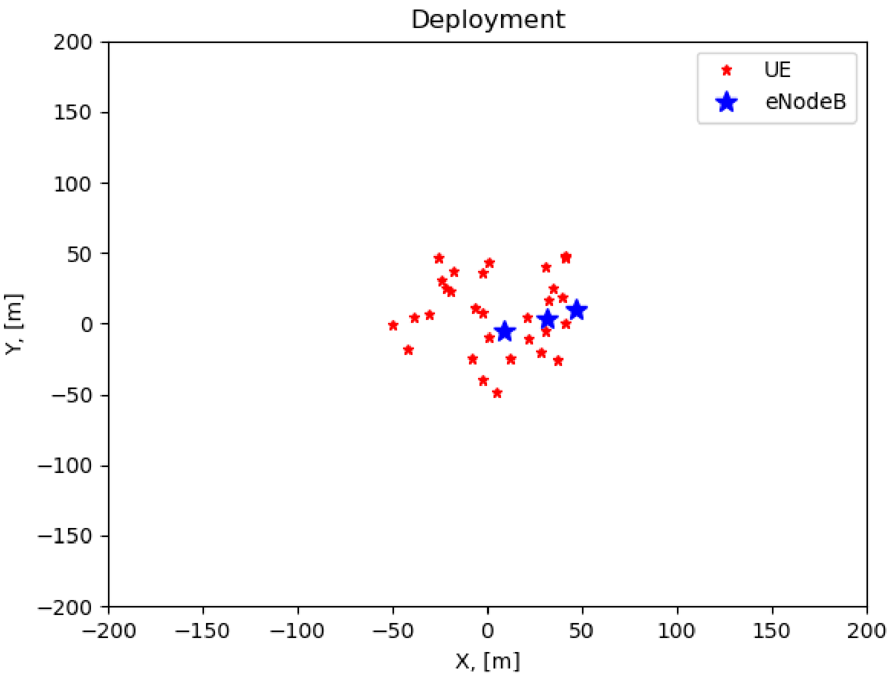


Fig. 6.15 Reference deployment scenario for the caching simulation framework

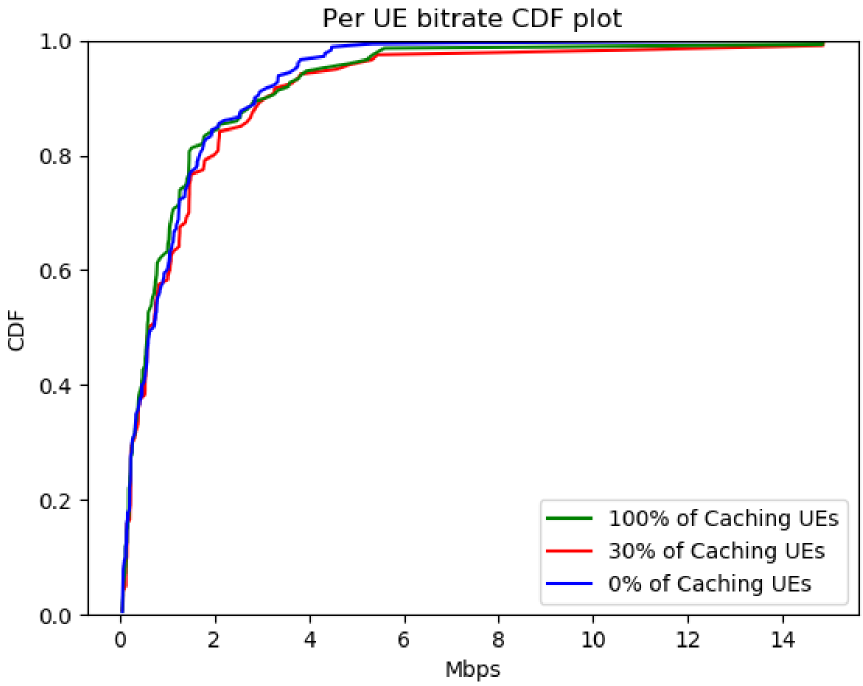


Fig. 6.16 CDF for per UE bitrate in Mbps for different caching scenarios blue: MBS, red: SBS, green: D2D

Let us discuss the obtained results. Fig 6.15 provides us with the visual representation of the reference deployment scenario for the caching simulation framework. Once can clearly see three base stations and a number of UEs that have all the opportunities to collaborate in D2D manner since some of them are in the vicinity of each other. Then Fig. 6.16 represents CDF (cumulative distribution function) for per UE bitrate in Mbps for three different caching scenarios. It is clear that the achievable bitrate for the given scenarios differs for the high rates increase of the bitrate. In practice it means that in SBS and D2D caching scenarios assure higher bitrates, than MBS. There is no much difference between bitrates for D2D and MBS since target bitrates in this simulation are not that high. The next Fig. 6.17 represents CDF for per UE SINR (Signal to Interference plus Noise Ratio) in dB for different caching scenarios. Then the last Fig. 6.18 represents CDF for per UE delay in ms for different caching scenarios and shows significant value for D2D caching compared to other MBS and SBS scenarios, i.e. average delay for the D2D caching is lower than any other caching scenarios.

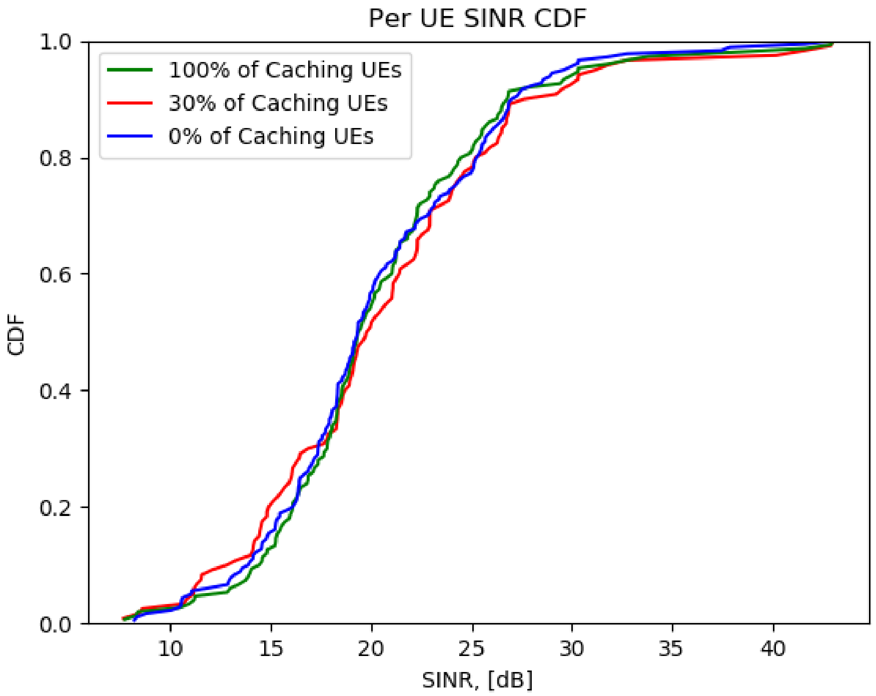


Fig. 6.17 CDF for per UE SINR in dB for different caching scenarios blue: MBS, red: SBS, green: D2D

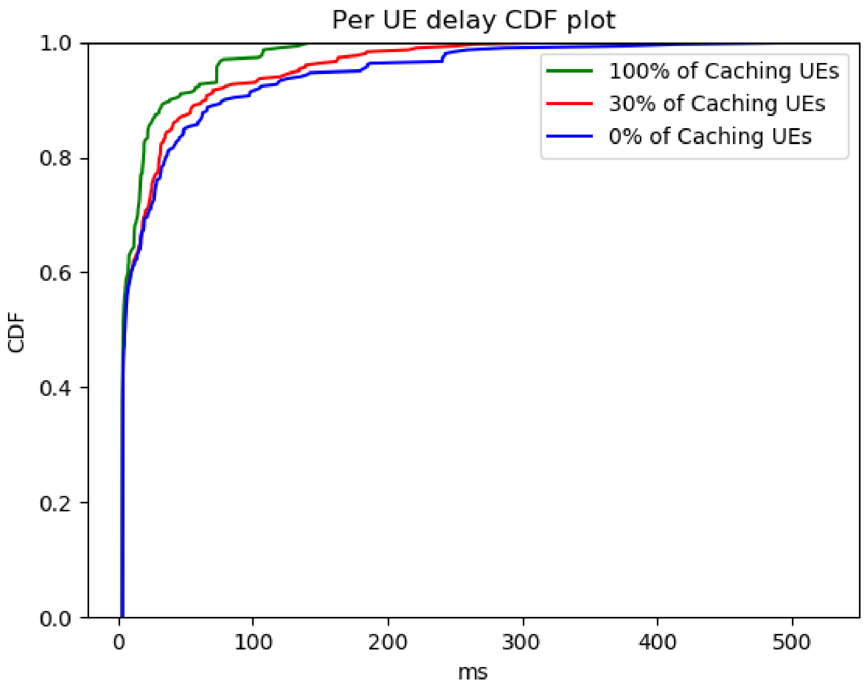


Fig. 6.18 CDF for per UE delay in ms for different caching scenarios blue: MBS, red: SBS, green: D2D

Discussion

Massive and continuous 5G researches have been introduced for standardization and establishment of 5G network, on the other hand ,Big Data enormous of structured and unstructured data and are generated from different sources constitutes a serious obstacle and real challenge for 5G network .

Therefore, in this work we investigated how effectively handling Big Data inside mobile network operator to achieve network development and future 5G requirements and KPIs , the main idea is to applied Big Data Driven frame works inside their network , and using different effective Big Data analytics methods and techniques , which will give the ability for mobile operators to develop their operational and business aspects based on future traffic demands .

Applying of a specific process and define framework on big data is considered a vital and a highly effective pre requisite process to achieve aims of Big data analytics inside the mobile network. In chapter 2, different proposed big data driven frameworks were discussed and we can conclude that four main processes should be performed: data collection, storage management and preprocessing, as data analysis and Define data application or optimization based on pervious steps.

In addition, the understanding of which kind of useful information and data to be collected and some characteristics of this data and its sources , is important from prospective of achieving targeted application , and which analysis tools need to be applied most of this data sources and tools are introduced in chapter 3.

Moreover, several case studies and application of Big Data analytics inside mobile network were described in chapter 4, which already applied and proved significant result from this application in both business and operational value.

In our practical part, we applied data-intensive approach, by analysis and applying different machine learning techniques on CDRs data sets since it is have an important contribution value in both scientific and business aspects.

We carried out data preprocessing by adding new variables to our data sets, data cleaning and detection of missing values. We employed different types of methods and is included: EDA, K-means Clustering, ARIMA and LSTM. These types and techniques were selected based on our data type is time series and we applied some tests on our data to identify which models can be suitable for our work. Moreover the findings of pervious work which they applied some of these techniques on similar data characteristics, are proven satisfactory performance similar to our work.

Results of the first analysis showed in respect to activity time and total activity ,peak hours during this day are 09:00 AM, 10:00 AM and 11:00 Am , and 03:00 AM is less activity hour during this day, and squire gird ID number 149 is the highest total activity during this day. From heat map cluster, it is clear that clusters 7, 2 ,and 10, high usage over 24 hours and cluster 3,1 and 4 have some activity at different times. And highest activity country code during this day and from highest to lowest are (39,0,46,33,41,49).

The application of this analysis and their results are beneficial in network development and business aspects, since it Will help in identifying which area can be developed or needs more resources, in identifying which square gird or country code produces more traffic as result of this application companies can gain more revenues by targeting high customer based on their geo- location and decrease its expenses by resources management.

In second analysis based on weekly data and identified regions with regards to specific Cell IDs , we can find that peak hours of Downtown areas is earlier than night life areas . For university area it is shown that in weekends have less call traffic and also internet may be because student can travel or another regions and mainly call traffic decreases at weekends.

These observation will help in resource allocation and optimization by defining which area is loaded and when, and can define in this area different temporary solution for peaks hour like Pico cell deployment.

Results of the third analysis based which we applied three models for prediction Internet traffic based on one-week data and hourly, results explained that ARIMA prediction model is accurate prediction for chosen three cells and with training data set 70% and 30% test set. While the results of LSTM prediction 69% training set and 21 % test sets was not sufficient in data/cell ID. the validated model the obtained results indicated that this model is suitable for all datasets (9998 cells), except for university area data still have some issue. Therefore, it was proven that this model fits for most of the datasets. At the same time, we investigated peak hours of internet traffic using the model and we can see that downtown peak hour accrued at 2 pm and night life at 6pm it expected since usually downtown traffic at evening start to decrease and night life areas start to have more activity in the evening.

Traffic prediction is an essential and play a vital role for mobile operators from prediction by hour, which I will be useful in traffic routing, or prediction yearly to support network optimization and investment planning.

The second practical part focused on caching at the edge using simulation tool to analysis and numerically evaluate the importance of caching at the edge using different caching seniors. The analysis of the results shows that for all caching scenarios SINR stays pretty much on the same level, this can be easily explained taking into account that UEs are communicating to BS similarly in case of MBS and SBS, while D2D is considered to be implemented using WiFi, that doesn’t affect SINR. It is clear that the achievable bitrate for the given scenarios differs for the high rates increase of the bitrate. In practice, it means that in SBS and D2D caching scenarios assure higher bitrates, than MBS.

CONCLUSION

This work is dedicated to application and handling of big data on an effective way inside mobile networks operator in telecommunication sector and specially to achieve requirements and KPIs of next generation 5G.

From my point of view, it is clear that effective implementation of Big Data inside networks appears to be serious issue. So we concluded that handling and applying Big Data in an effective manner by defining a framework is a mandatory to achieve a significant results from Big data Application and this framework was summarized in four main processes which we provided in our work.

Understanding the available data, which type of useful information and data to be collected, which analytics tools are suitable and should be applied are important consideration for any service provider to harvest best results from its data. Pervious successful case studies and application of Big Data for mobile operators was discussed and proved that significant result obtained from this application in both business and operational value.

The practical aspects of this work has proven how benefits in operation and business aspect of telecom companies can be achieved by efficient application of Big data techniques rather than traditional methods. We applied predictive models like ARIMA and LSTM for traffic prediction and explained that reveled results are beneficial in short and strategic plans for operator. Our selection of CDRs database to perform our practical part is based on the importance of this data set for mobile operator; as our results shows that analyzing CDRs has a prospective significance currently and beyond in areas such as resources allocation , network optimization , investment plans based on traffic prediction, fault detection and self-organized and optimization network .

In last part using a WINTERsim simulation tool, we evaluated numerically the importance of caching at the edge to achieve some of KPIs required by 5G and performance of different caching seniors.

Further investigation are required for future work on analysis another datasets inside mobile networks and area of proactive caching based on big data and machine learning techniques for 5G networks .

Acronyms list

MNOs	Mobile network operators
5G	Fifth Generation
IoT	Internet Of Things
OTT	On The Top
D2D	Device to Device
ICT	Information and Communication Technology
M2M	Machine to Machine
CDRs	Call Details Records
2G	Second Generation
GSM	Global System for Mobile communications
ITU	International Telecommunication Union
GPS	Global Positioning System
MIMO	Multi input Multi output
UE	User Equipment
CN	Core Network
MDT	Measurement Driver Test
SVM	support vector networks
SDN	Self Defined Network
ML	Machine Learning
BS	Base Station
D4D	Data for development
BDD	Big Data Driven
ISPs	Internet service provider
RAN	Radio Access Network
QoE	Quality of Experience
Qos	Quality of Services
ARIMA	Auto Regressive Integrated Moving Average
LSTM	Long Short Time Memory
LTE	Long Term Evolution

REFERENCES:

5G, the mobile connectivity of the future. (2018, February 07). https://www.orange.com/en/Human-Inside/Thematic-feature/5G-the-mobile-connectivity-of-the-future.

Agiwal, M., Roy, A., & Saxena, N. (2016). Next generation 5G wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *18*(3), 1617-1655.

Bastug, E., Bennis, M., & Debbah, M. (2014). Living on the edge: The role of proactive caching in 5G wireless networks. *IEEE Communications Magazine*, *52*(8), 82-89.

Baştuğ, E., Bennis, M., Zeydan, E., Kader, M. A., Karatepe, I. A., Er, A. S., & Debbah, M. (2015). Big data meets telcos: A proactive caching perspective. *Journal of Communications and Networks*, *17*(6), 549-557.

Bastug, E, Jean-Louis Gu´en´ego, M´erouane Debbah. (2013). Proactive Small Cell Networks. ICT ,2013, May 2013, Casablanca, Morocco. pp.1-5,

Barlacchi G, De Nadai M, Larcher R, Casella A, Chitic C, Torrisi G et al. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci. Data*. 2015; 2: 150055 pmid:26528394

Bi, S., Zhang, R., Ding, Z., & Cui, S. (2015). Wireless communications in the era of big data. *IEEE communications magazine*, *53*(10), 190-199.

Blondel, Vincent D., Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis." *EPJ Data Science 4*, no. 1 (2015): 1.

Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L., & Popovski, P. (2014). Five disruptive technology directions for 5G. *IEEE Communications Magazine*, *52*(2), 74-80.

B. Commision, "The State of Broadband 2014 : Broadband for All," *International Telecommunication Union*, vol. Geneva, September 2014

Cai, Y., Yu, F. R., Liang, C., Sun, B., & Yan, Q. (2016). Software-defined device-to-device (D2D) communications in virtual wireless networks with imperfect network state information (NSI). *IEEE Transactions on Vehicular Technology*, *65*(9), 7349-7360.

Carlton, A., & IDG Contributor Network. (2017, May 25). Big Data will enable better network and application intelligence in 5G.

https://www.computerworld.com/article/3198454/internet-of-things/big-data-will-enable-better-network-and-application-intelligence-in-5g.html

Celebi, O. F., Zeydan, E., Kurt, O. F., Dedeoglu, O., Iieri, O., AykutSungur, B., & Ergut, S. (2013, May). On use of big data for enhancing network coverage analysis. In *Telecommunications (ICT), 2013 20th International Conference on* (pp. 1-5). IEEE.

Chakraborty, A. (2017). *Data-Driven Performance Optimization in Wireless Networks* (Doctoral dissertation, State University of New York at Stony Brook).

Chen, M., Hao, Y., Qiu, M., Song, J., Wu, D., & Humar, I. (2016). Mobility-aware caching and computation offloading in 5G ultra-dense cellular networks. *Sensors*, *16*(7), 974.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.

Chimeh, J. (2015). 5G mobile communications: a mandatory wireless infrastructure for big data. In *Proc. International Conference on Advances in Computing, Electronics and Electrical Technology (CEET)* (Vol. 2015).

Cui, L., Yu, F. R., & Yan, Q. (2016). When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE network*, *30*(1), 58-65.

Daniel, B. K. (2017). Contestable professional academic identity of those who teach research methodology. *International Journal of Research & Method in Education*, 1-14.

Data available at: Telecom Italia Big Data Challenge 2014, https://dandelion.eu/datamine/open-big-data/.

Deng, L., Gao, J., & Vuppalapati, C. (2015, March). Building a big data analytics service framework for mobile advertising and marketing. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on* (pp. 256-266). IEEE.

Ehrenberg, A. (2012). Andrew SC Ehrenberg. *Business Theory: High-impact Strategies-What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*, 110.

Fei, Y., Wong, V. W., & Leung, V. (2006). Efficient QoS provisioning for adaptive multimedia in mobile communication networks by reinforcement learning. *Mobile Networks and Applications*, *11*(1), 101-110.

Garcia-Perez, C. A., & Merino, P. (2016, September). Enabling low latency services on LTE networks. In *Foundations and Applications of Self* Systems, IEEE International Workshops on* (pp. 248-255). IEEE.

Gregori, M., Gómez-Vilardebó, J., Matamoros, J., & Gündüz, D. (2016). Wireless content caching for small cell and D2D networks. *IEEE Journal on Selected Areas in Communications*, *34*(5), 1222-1234.

Gupta, A., & Jha, R. K. (2015). A survey of 5G network: Architecture and emerging technologies. *IEEE access*, *3*, 1206-1232.

He, Y., Yu, F. R., Zhao, N., Yin, H., Yao, H., & Qiu, R. C. (2016). Big data analytics in mobile cellular networks. *IEEE access*, *4*, 1985-1996.

Ho, S. L., Xie, M., & Goh, T. N. (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. Computers & Industrial Engineering, 42(2-4), 371-375.

Hyndman ,R.J and Melbourne Australia. http://otexts.org/fpp/. Athanasopoulos, G. 2013.

Imran, A., Zoha, A., & Abu-Dayya, A. (2014). Challenges in 5G: how to empower SON with big data for enabling 5G. *IEEE network*, *28*(6), 27-33.

ITU-R, "Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Feb. 2015

Ji, M., Caire, G., & Molisch, A. F. (2016). Wireless device-to-device caching networks: Basic principles and system performance. *IEEE Journal on Selected Areas in Communications*, *34*(1), 176-189.

Jun, L., Tingting, L., Gang, C., Hua, Y., & Zhenming, L. (2013). Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis. *China Communications*, *10*(12), 25-36.

Kadir, E. A., Shamsuddin, S. M., Rahman, T. A., & Ismail, A. S. (2015). Big data network architecture and monitoring use wireless 5G technology. *Int. J. Advanced Soft Compu. Appl*, *7*(1), 1-14.

Kreutz, D., Ramos, F. M., Verissimo, P. E., Rothenberg, C. E., Azodolmolky, S., & Uhlig, S. (2015). Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, *103*(1), 14-76.

Kolar, V., Ranu, S., Subramainan, A. P., Shrinivasan, Y., Telang, A., Kokku, R., & Raghavan, S. (2014, January). People in motion: Spatio-temporal analytics on call detail records. In *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on* (pp. 1-4). IEEE.

Lei, L., Zhong, Z., Zheng, K., Chen, J., & Meng, H. (2013). Challenges on wireless heterogeneous networks for mobile cloud computing. *IEEE Wireless Communications*, *20*(3), 34-44.

Liu, J., Chang, N., Zhang, S., & Lei, Z. (2015). Recognizing and characterizing dynamics of cellular devices in cellular data network through massive data analysis. *International Journal of Communication Systems*, *28*(12), 1884-1897.

Lomotey, R. K., & Deters, R. (2014, June). Terms mining in document-based nosql: Response to unstructured data. In *Big Data (BigData Congress), 2014 IEEE International Congress on* (pp. 661-668). IEEE.

Moysen, J., Giupponi, L., & Mangues-Bafalluy, J. (2017). A mobile network planning tool based on data analytics. *Mobile Information Systems*, *2017*.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., & Xin, D. (2016). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, *17*(1), 1235-1241.

Musolesi, M. (2014). Big mobile data mining: Good or evil?. *IEEE Internet Computing*, *18*(1), 78-81.

Naboulsi, D. (2015). *Analysis and exploitation of mobile traffic datasets* (Doctoral dissertation, Lyon, INSA).

Parvez, I., Rahmati, A., Guvenc, I., Sarwat, A. I., & Dai, H. (2017). A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions. *arXiv preprint arXiv:1708.02562*.

Paschos, G., Bastug, E., Land, I., Caire, G., & Debbah, M. (2016). Wireless caching: Technical misconceptions and business barriers. *IEEE Communications Magazine*, *54*(8), 16-22.

Philip Branch. (2018, April 09). The 'G' in 5G: How mobile generations have evolved. Retrieved April 16, 2018, from http://theconversation.com/the-g-in-5g-how-mobile-generations-have-evolved-53102.

Poularakis, K., Iosifidis, G., Argyriou, A., Koutsopoulos, I., & Tassiulas, L. (2016, April). Caching and operator cooperation policies for layered video content delivery. In *INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE* (pp. 1-9). IEEE.

Ramaprasath, A., Srinivasan, A., & Lung, C. H. (2015, May). Performance optimization of big data in mobile networks. In *Electrical and Computer*

Engineering (CCECE), 2015 IEEE 28th Canadian Conference on (pp. 1364-1368). IEEE.

Samulevicius, S., Pedersen, T. B., & Sorensen, T. B. (2015, May). MOST: Mobile broadband network optimization using planned spatio-temporal events. In *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st* (pp. 1-5). IEEE.

Shafiq, M. Z., Ji, L., Liu, A. X., Pang, J., & Wang, J. (2012, March). Characterizing geospatial dynamics of application usage in a 3G cellular data network. In *INFOCOM, 2012 Proceedings IEEE* (pp. 1341-1349). IEEE.

Soto, V., & Frías-Martínez, E. (2011, June). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch* (pp. 17-22). ACM.A. Ioannou and S. Weber, “A Survey of Caching Policies and Forwarding Mechanisms in Information-Centric Networking,” *IEEE Commun.Surv. Tutor.*, vol. 18, no. 4, pp. 2847–2886, Fourthquarter 2016.

Su, Z., Xu, Q., Zhu, H., & Wang, Y. (2015). A novel design for content delivery over software defined mobile social networks. *IEEE Network*, *29*(4), 62-67. -- Content deliver.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth Annual Conference of the International Speech Communication Association.

Tolle, K.M.; Tansley, D.S.W.; Hey, A.J. The Fourth Paradigm: Data-intensive Scientific Discovery; IEEE: Piscataway, NJ, USA, 2011; pp. 1334–1337.

TROIA, S. (2016). Machine learning-based traffic prediction and pattern extraction for dynamic optical routing in SDN mobile metro networks.

Wickham H. et al. Tidy data //Journal of Statistical Software. – 2014. – Т. 59. – №. 10. – С. 1-23.

WINTERsim simulation tool http://winter-group.net/download/.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

Yan, M. (2005). *Methods of determining the number of clusters in a data set and a new clustering criterion* (Doctoral dissertation, Virginia Tech).

Yan, Q., & Yu, F. R. (2015). Distributed denial of service attacks in software-defined networking with cloud computing. *IEEE Communications Magazine*, *53*(4), 52-59.

Yang, J., Zhang, S., Zhang, X., Liu, J., & Cheng, G. (2013, June). Characterizing smartphone traffic with MapReduce. In *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on* (pp. 1-5). IEEE.

Zeng, D., Gu, L., & Guo, S. (2015). Cost minimization for big data processing in geo-distributed data centers. In *Cloud networking for big data* (pp. 59-78). Springer, Cham.

Zeydan, E., Bastug, E., Bennis, M., Kader, M. A., Karatepe, I. A., Er, A. S., & Debbah, M. (2016). Big data caching for networking: Moving from cloud to edge. *IEEE Communications Magazine*, *54*(9), 36-42.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.

Zhang, M., Luo, H., & Zhang, H. (2015). A survey of caching mechanisms in information-centric networking. *IEEE Communications Surveys & Tutorials*, *17*(3), 1473-1499.

Zhang, S., Xu, X., Wu, Y., & Lu, L. (2014, November). 5G: Towards energy-efficient, low-latency and high-reliable communications networks. In *Communication Systems (ICCS), 2014 IEEE International Conference on* (pp. 197-201). IEEE.

Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., & Xiang, W. (2016). Big data-driven optimization for mobile networks toward 5G. *IEEE network*, *30*(1), 44-51.