

**People's Democratic Republic of Algeria**  
**Ministry of Higher Education and Scientific Research**  
**University M'Hamed BOUGARA – Boumerdes**



**Institute of Electrical and Electronic Engineering**

**Department of Electronics**

Final Year Project Report Presented in Partial Fulfilment of  
the Requirements for the Degree of the

**MASTER**

**In Telecommunications Engineering**

**Option: Telecommunications Engineering**

Title:

**Quality of Service Key Performance  
Indicators Analysis and Optimization in  
4G/LTE Core Network**

Presented by:

- **LAIMECHE Mohammed Salim**
- **AMROUCHE Yacine**

Supervisor:

**Dr. CHERIFI Dalila**

Co-Supervisor:

**Mr. SAKILANI Mohammed (DJEZZY)**

Registration Number:...../2023

# Abstract

In order to face competition and meet the requirements of their customers as well as national regulatory authorities, mobile telephone operators must constantly monitor the quality of their services. This important task in the overall process of running a telecommunications network requires in-depth knowledge of the operation and constitution of the network.

This project aims to address this need by developing a comprehensive mechanism for analyzing and optimizing the key performance indicators (KPIs) related to the quality of service (QoS) in the 4G/LTE core network for DJEZZY mobile operator. Through the collection and processing of relevant KPI data, we evaluate technical success rates across various processes. Subsequently, we employ advanced analytics techniques to detect anomalies or malfunctions in the network, allowing for timely troubleshooting and optimization. To enhance accessibility and usability, we have implemented this mechanism within a user-friendly web application, this empowers service workers within the organization to easily access and utilize valuable insights gained from the analysis. The intuitive interface facilitates efficient decision-making and prompt actions to rectify identified issues, ultimately improving the network's performance. By focusing on QoS analysis and optimization, this project contributes to enhancing the overall user experience and ensuring the operator's compliance with regulatory standards.

# Dedication

*I dedicate my dissertation work to my family and many friends. To my mother and my father Thanks to their tender encouragement and their great sacrifices, they were able to create an affectionate climate conducive to the pursuit of my studies. Dedication could not express my respect, my consideration and my deep feelings towards them. I pray Allah to bless them, watch over them, hope they will always be proud of me.*

*As a testimony to the attachment, love and affection that I bear for you. I dedicate this work to you with all my wishes for happiness, health and success.*

*Salim*

*To my incredible family, especially my beloved parents, supportive siblings, and wonderful sisters. Your constant love, encouragement, and belief in me have been the cornerstone of my achievements. Thank you for being by my side and for just being a part of my life.*

*Special dedication to all my friends especially Abdelmajid, Yanis and Youcef.*

*Yacine*

# Acknowledgements

Praise be to Allah for guiding us on our educational path and helping us to complete this project

First and foremost, we would like to extend our heartfelt appreciation to our supervisor, Dr.CHERIFI Dalila, for their invaluable guidance, expertise, and unwavering support throughout this journey. Her deep knowledge and insightful feedback have been instrumental in shaping the direction and quality of this work.

We owe special thanks to our co-supervisor Mr SAKILANI Mohammed (IT & CORE (PS/CS) quality performance engineer DJEZZY) for his help, patience, understanding and encouragement. Also, special words of thanks with gratitude are devoted to all DJEZZY employees for their reception and treatment during the period of the Internship.

In addition, we are grateful to our friends and family for their unwavering encouragement, understanding, and support throughout this endeavor. Their belief in us has provided the motivation and strength to overcome challenges and strive for excellence.

To all those who have played a part, directly or indirectly, in making this project possible, we extend our sincere thanks and gratitude. Your support has been invaluable, and we are truly grateful for your contributions.



# Contents

<b>Abstract</b>	<b>I</b>
<b>List of Figures</b>	<b>VI</b>
<b>List of abbreviations</b>	<b>IX</b>
<b>General Introduction</b>	<b>XI</b>
<b>1 Mobile Networks Overview</b>	<b>1</b>
1.1 Introduction . . . . .	2
1.2 Cellular communication concept . . . . .	2
1.3 Key data transmission concepts . . . . .	2
1.4 Cellular Telecommunication generations . . . . .	2
1.5 GSM network . . . . .	3
1.5.1 GSM Network Architecture . . . . .	3
1.5.2 Mobile Station MS . . . . .	4
1.5.3 Network Switching Subsystem (NSS) . . . . .	4
1.5.4 Base Station Subsystem (BSS) . . . . .	5
1.5.5 Operation and Support Subsystem (OSS) . . . . .	6
1.5.6 EGPRS:GPRS/EDGE . . . . .	6
1.6 Third Generation network (3G)/UMTS . . . . .	6
1.7 Fourth Generation (4G)/LTE . . . . .	7
1.8 LTE Network Architecture . . . . .	9
1.9 Summary . . . . .	11
<b>2 Quality of Service and KPIs</b>	<b>13</b>
2.1 Introduction . . . . .	14
2.2 Quality of service . . . . .	14
2.3 QOS in mobile networks . . . . .	14
2.4 Counters . . . . .	15
2.5 Key Performance indicators . . . . .	16
2.6 The attach procedure . . . . .	18
2.7 Tracking area update procedure . . . . .	19
2.8 Manual monitoring . . . . .	20
2.9 New monitoring application . . . . .	20
2.9.1 Functional needs . . . . .	22
2.9.2 Non-functional needs . . . . .	22
2.10 Summary . . . . .	22

<b>3</b>	<b>Implementation and Results</b>	<b>23</b>
3.1	Introduction	24
3.2	Development Tools	24
3.2.1	Python	24
3.2.2	FileZilla FTP	24
3.2.3	Pentaho Data Integration	25
3.2.4	KNIME Analytics Platform	26
3.2.5	Microsoft SQL Server	26
3.2.6	Microsoft Power BI	27
3.3	Environment Preparation	27
3.3.1	Creation of FTP server	27
3.3.2	Creation of SQL database	28
3.4	Data Description	30
3.5	Mechanism Implementation Steps	32
3.5.1	Data collection	32
3.5.2	Data Integration	35
3.5.2.1	Data Integration using Pentaho	36
3.5.2.2	Data integration using knime:	39
3.5.3	Data Staging	41
3.5.4	Data visualization	45
3.6	Data Analysis for Anomaly Detection	49
3.6.1	Non coded anomaly detection:	49
3.6.1.1	Clustering method:	49
3.6.1.2	Outlier detection method:	50
3.6.1.3	Predictive analysis:	52
3.6.2	Coded Anomaly detection	58
3.6.2.1	Data preprocessing:	58
3.6.2.2	Getting normal range from main KPIs:	58
3.6.2.3	Detecting anomalies using normal range:	59
3.6.2.4	Anomaly report creation:	60
3.7	Mechanism Deployment	61
3.7.1	Development tools:	61
3.7.1.1	Flask library:	61
3.7.1.2	HTML:	61
3.7.2	Web application presentation	61
3.8	Summary	64
	<b>General Conclusion</b>	<b>65</b>
	<b>A Tracking Area Optimization</b>	<b>66</b>
	<b>B Machine Learning Models</b>	<b>68</b>
B.1	K-means	68
B.2	GBM	68
B.3	Random Forest	69
	<b>Bibliography</b>	<b>71</b>

# List of Figures

1.1 Evolution of mobile networks [6]	3
1.2 GSM system architecture [2]	4
1.3 GSM system cell representation [8]	5
1.4 GPRS/EDGE network [10]	6
1.5 UMTS network architecture [11]	7
1.6 3GPP Standard evolution [14]	8
1.7 LTE network description [6]	8
1.8 LTE network architecture [16]	9
1.9 eNodeB main functions [16]	10
1.10 EPC architecture [17]	10
1.11 MME connections to other nodes [16]	11
2.1 4G/LTE Key performances indicators [18]	16
2.2 Key performances indicators types	17
2.3 Counters,KPIs and QOS reports	17
2.4 Attach procedure call flow [3]	18
2.5 Tracking area procedure callflow [14]	20
2.6 Monitoring System Workflow	21
3.1 FileZilla client interface	25
3.2 Pentaho User interface	25
3.3 KNIME User interface	26
3.4 Power BI User interface	27
3.5 Creation of SQL database	29
3.6 Login-New tab	30
3.7 Naming scheme of csv files	31
3.8 Organization of the data inside the csv file	31
3.9 Get files from emails flowchart	32
3.10 Importing table structure into SQL database	33
3.11 Skipping data rows	34
3.12 Column names check flowchart	35
3.13 Pentaho transformation	36
3.14 Pentaho text file input node	37
3.15 Pentaho split fields node	37
3.16 Database connection tab	38
3.17 Table output configuration	39
3.18 Knime workflow	40
3.19 A zoom at the workflow	40
3.20 Query for managing the primary key columns.	42
3.21 Attach and TAU table creation	42
3.22 Power BI front page	45

3.23 Pie chart representing attach request and success times	45
3.24 Pie chart representing attach failure times	46
3.25 Line chart representing attach failure times	47
3.26 Line chart representing Intra TAU success rate	47
3.27 Pie chart representing Intra TAU failure times due to USN causes	48
3.28 Pie chart representing Intra TAU request times	48
3.29 K-means workflow	49
3.30 K-means centers	50
3.31 K-means line plot	50
3.32 Numeric outliers workflow	51
3.33 Numeric outliers output table	51
3.34 Numeric outliers output line plot	52
3.35 AutoML learner node	53
3.36 Best model	54
3.37 Leaderboard table	54
3.38 Gradient boosters workflow	55
3.39 Gradient boosters classifier	56
3.40 Gradient boosters regressor	57
3.41 Random forest workflow	57
3.42 Random forest prediction table	58
3.43 Anomaly visualization for USN_Blida on 2023-04-07.	60
3.44 Anomaly report on 2023-04-07.	61
3.45 Login page for web app	62
3.46 Navigation dashboard for web app	62
3.47 Visualization dashboard for web app	63
3.48 Anomaly report dashboard for web app	63
A.1 Tracking Area scheme	66
B.1 Gradient boosted trees for regression	69



---

# List of Abbreviations

3GPP	Third Generation Partnership Project
AuC	Authentication Centre
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CN	Core Network
CS	cellular systems' Circuit-Switched
CSFB	Circuit Switched Fall Back
CSV	Comma Separated Values
DBMS	database management system
E-RAB	Evolved Radio Access Bearer
E-UTRAN	Evolved-UTRAN
EDGE	Enhanced Data Rates for GSM Evolution
EGPRS	Enhanced GPRS
EIR	Equipment Identity Register
EMM	EPS Mobility Management
EPC	Evolved Packet Core
EPS	Evolved Packet System
ESM	EPS Session Management
ETL	extraction, transformation, and loading
FDMA	Frequency Division Multiple Access
FTP	File Transfer Protocol
GBM	Gradient Boosting Machine
GSN	Gate GPRS Serving Node
GMSC	Gateway Mobile Switching Centre
GPRS	General Packet Radio Service
GUTI	Globally Unique Temporary Identity
HLR	Home Location Register
HSS	Home Subscriber Server
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
IMEI	International Mobile Equipment Identity
IMS	IP Multimedia Sub-System
IMSI	International Mobile Subscriber Identification
IP	Internet Protocol
JDBC	Java Data Base Connection
KPIs	key performance indicators
LTE	Long Term Evolution
ME	mobile equipment
MME	Mobility Management Entity
MS	Mobile Station
MSC	Mobile Services Switching Center

---

MSE	Mean Square Error
NAS	Non Access Stratum
NSS	Network Switching Subsystem
OFDMA	Orthogonal Frequency Division Multiple Access
OMC	Operation and Maintenance Center
OSS	Operation and Support Subsystem
P-GW	Packet Gateway
PCEF	Policy Control Enforcement Function
PCRF	Policy Control and Charging Rules Function
PDN	Packet Data Network
PS	Packet Switched
QoS	quality of service
RAN	Radio Access Network
RNC	Radio Network Controller
RRC	Radio Resource Control
S-GW	serving gateway
SGSN	Serving GPRS Support Node
SGW	Serving Gateway
SIM	Subscriber Identity Module
SMS	Short Message Service
SSMS	SQL Server Management Studio
TA	Tracking Area
TAI	Tracking Area Identity
TAU	Tracking Area Update
TDMA	Time Division Multiple Access
UE	User Equipment
URL	Uniform Resource Locator
VLR	Visitor Location Register
VoIP	Voice over IP
NCASr	Non Combined Attach Success rate
AST	Attach Success Times
NUSNCE	non USN Causes Excluded
ART	Attach Request Times
CASr	Combined Attach Success rate
CAST	Combined Attach Success Times
CART	Combined Attach Request Times
CITAUSr	Combined Intra Tracking Area Update Success rate
ITAUST	Intra Tracking Area Update Success Times
ICTAUST	Intra Combined Tracking Area Update Success Times
PTAUST	Periodic Tracking Area Update Success Times
ITAURT	Intra Tracking Area Update Request Times
PTAURT	Periodic Tracking Area Update Request Times

# General Introduction

It is important to note that, during the past century, telecommunications have been the area of modern society that has seen the most innovation. As the competition among mobile operators intensifies, satisfying customer demands and ensuring regulatory compliance become paramount. One crucial aspect influencing customer satisfaction and differentiation is the quality of service (QoS) provided by mobile networks. Mobile operators must continuously monitor and improve their services to maintain optimal performance and gain a competitive edge.

This project addresses the pressing need for comprehensive analysis and optimization of key performance indicators (KPIs) related to QoS in the 4G/LTE core network, in case of DJEZZY network. By developing a robust mechanism for analyzing these KPIs, valuable insights can be gained into the network's technical success rates across various processes.

The primary objective of this project is to promptly detect anomalies or malfunctions within the network. Leveraging advanced analytics techniques, deviations from expected performance levels can be identified, enabling timely troubleshooting and optimization. Proactive measures to prevent service disruptions, reduce downtime, and enhance the overall user experience are essential.

We will specifically focus on QoS analysis and optimization within the 4G/LTE core network, acknowledging its crucial role in determining the user experience. Through comprehensive monitoring and analysis of KPIs, areas requiring attention and improvement can be identified. Furthermore, this project ensures compliance with regulatory standards, as maintaining QoS benchmarks is a key requirement imposed by regulatory authorities.

The subsequent chapters of this project will explore the mobile network landscape, its evolution, and underlying architectures. Chapter 2 will delve into the principles of QoS, emphasizing its significance in delivering satisfactory user experiences, and highlight the vital role of KPIs in monitoring and evaluating network performance. In the third chapter we are going to show the implementation and design of our application, followed by the results in the form of various screenshots as well as the requested interpretations.



# **Chapter 1**

## **Mobile Networks Overview**

---

## 1.1 Introduction

Mobile networks are critical in today's interconnected society, allowing billions of people globally to communicate wirelessly. These networks provide seamless connectivity and enable people to stay connected, access information, and interact while on the move. They employ a variety of technologies, including 2G, 3G, 4G, and the forthcoming 5G, each of which provides faster data speeds, lower latency, and more capacity. Mobile networks rely on a network of base stations and towers to send and receive signals, allowing voice conversations, text messaging, and internet access [1]. The aim of this chapter is to provide a thorough understanding of cellular networks, including its characteristics and concepts. We will look at the core notion of cellular networks and the stages of evolution of the technologies applied, from the first to the fourth generation.

## 1.2 Cellular communication concept

The cellular concept is a fundamental principle underlying the design and operation of cellular telecommunication systems. It involves dividing the coverage area into smaller cells served by base stations or towers. It allows for efficient frequency reuse, maximizing system capacity and supporting a large number of mobile devices. As users move between cells, seamless handovers ensure uninterrupted communication. The cellular concept has been the basis for generations of mobile networks, enabling efficient wireless connectivity and continuous advancements in speed and capacity [2]. It continues to evolve to meet the growing demands of a connected world.

## 1.3 Key data transmission concepts

Packet switching and circuit switching are two different methods of data transmission used in telecommunications. Circuit switching involves the creation of a dedicated physical communication path, or circuit, between two devices for the duration of a communication session. Once the circuit is established, data is transmitted along the circuit until the communication session is complete. Circuit switching is often used for real-time voice communication, such as traditional phone calls, where a dedicated and consistent connection is required [3]. Packet switching, on the other hand, involves the breaking up of data into packets that are transmitted independently and can take different paths to their destination, depending on network traffic and other factors. Each packet includes information about its source, destination, and sequencing, and the packets are reassembled into their original form at the destination. Packet switching is more efficient than circuit switching, as network resources can be shared among multiple users and devices, and can handle different types of traffic, including voice, video, and data. In summary, circuit switching provides a dedicated, consistent connection for real-time voice communication, while packet switching allows for more efficient and flexible transmission of different types of data. The choice of which method to use depends on the requirements of the specific communication task [4].

## 1.4 Cellular Telecommunication generations

Mobile networks have evolved through different generations to meet the increasing demands for wireless communication. The first generation (1G) networks emerged in the 1980s, utilizing analog technology for voice calls but with limited coverage and call quality [5]. The second

generation (2G) networks, introduced in the 1990s, marked a significant improvement with digital technology, better call quality, increased capacity, and the advent of text messaging. The third generation (3G) networks arrived in the early 2000s, offering high-speed internet access, video calling, and mobile TV. In the 2010s, the fourth generation (4G) networks revolutionized mobile communication with faster data speeds, lower latency, and the rise of app-based services [5]. The latest and most advanced standard is the fifth generation (5G), employing advanced technologies like higher-frequency radio waves and massive MIMO to provide unprecedented data speeds of up to 20 Gbps. 5G enables transformative applications such as the Internet of Things, autonomous vehicles, and virtual and augmented reality [6].

Technology ⇔	1G	2G	3G	4G	5G
Feature ↴					
Start/Deployment	1970 – 1980	1990 – 2004	2004-2010	Now	Soon (probably 2020)
Data Bandwidth	2kbps	64kbps	2Mbps	1 Gbps	Higher than 1Gbps
Technology	Analog Cellular Technology	Digital Cellular Technology	CDMA 2000 (1xRTT, EVDO) UMTS, EDGE	Wi-Max LTE Wi-Fi	WWWW(coming soon)
Service	Mobile Telephony (Voice )	Digital voice, SMS, Higher capacity packetized data	Integrated high quality audio, video and data	Dynamic Information access, Wearable devices	Dynamic Information access, Wearable devices with AI Capabilities
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	CDMA
Switching	Circuit	Circuit, Packet	Packet	All Packet	All Packet
Core Network	PSTN	PSTN	Packet N/W	Internet	Internet

Figure 1.1: Evolution of mobile networks [6]

## 1.5 GSM network

GSM (Global System for Mobile Communications) is a widely adopted digital cellular network technology that revolutionized mobile communication. It provides seamless voice and data services, operating on specific frequency bands for compatibility. Utilizing TDMA and FDMA techniques, GSM efficiently allocates network resources. With standardized protocols, it ensures reliable and secure communication, supporting features like call forwarding, texting, and internet connectivity. GSM's impact on wireless communication is significant, serving as a foundation for subsequent mobile network generations [7].

### 1.5.1 GSM Network Architecture

The Mobile Station (MS), Base Station Subsystem (BSS), Network Switching Subsystem (NSS), and Operation and Support Subsystem (OSS) are the four primary parts of the GSM network architecture. Each one of these elements is essential to the GSM network's operation. The basic diagram of the 2G GSM mobile communications system's overall system architecture shown in [1.2] highlights the four key components:

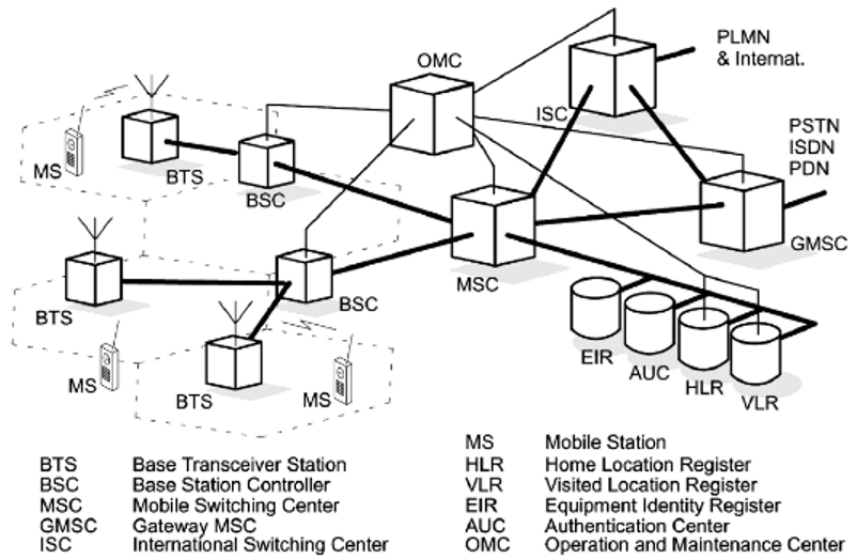


Figure 1.2: GSM system architecture [2]

Within this diagram the different network nodes can be seen - they are grouped into the four areas that provide different functionality, but all operate to enable reliable mobile communications to be achieved. The whole network design proved to be very effective and was further expanded to permit data transmission in the 2G evolution, and later with additional developments to enable the establishment of 3G [3].

### 1.5.2 Mobile Station MS

The mobile station (MS) consists of two main components: the mobile equipment (ME) and the SIM card. The ME comprises the hardware, including the antenna, radio transceiver, display, keypad, and device-specific features. The SIM card holds the subscriber's identity information, such as phone number and authentication keys. When the mobile device is turned on, it searches for available GSM networks and registers with the nearest Base Transceiver Station (BTS) by sending a registration request message. The MS maintains its location through periodic location update messages, enabling the network to route calls and messages accurately. The SIM card allows the MS to authenticate with the network and access subscribed services, while also storing contact lists and user-specific data. The SIM card's International Mobile Subscriber Identification (IMSI) provides network access and allows users to switch phones easily. The MS can utilize data services like SMS and MMS in addition to voice calls and messages. The ME also contains the International Mobile Equipment Identity (IMEI), a unique identifier that remains unchanged and is verified by the network during registration to prevent unauthorized equipment usage [8].

### 1.5.3 Network Switching Subsystem (NSS)

The nodes and functionalities required for call switching, subscriber management, and mobility management are contained in the Network Subsystem (NSS), sometimes known as the "core network". The major elements within the core network are:

- **Mobile Services Switching Centre (MSC):** The Mobile Switching Center is the primary component of the core network area of the overall GSM network design (MSC). The MSC

---

performs the same functions as a standard switching node within a PSTN or ISDN while also offering extra capability to satisfy mobile user needs.

- **Home Location Register (HLR):** The Home Location Register (HLR) serves as a fundamental database within the GSM network, containing crucial administrative information about each subscriber and their last known location. It acts as a central repository for subscriber details, including telephone numbers, service subscriptions, permissions, and authentication data.
- **Visitor Location Register (VLR):** VLR is responsible for a group of location areas and stores the data of all users that are currently located in these areas. The VLR can be implemented as a separate entity, but it is commonly realized as an integral part of the MSC, rather than a separate entity. In this way access is made faster and more convenient [3].
- **Equipment Identity Register (EIR):** The EIR is the entity that decides whether a given mobile equipment may be allowed onto the network.
- **Authentication Centre (AuC):** The AuC is a protected database that contains the secret key also contained in the user's SIM card. It is used for authentication and for ciphering on the radio channel.
- **Gateway Mobile Switching Centre (GMSC):** The GMSC serves as the interface between the GSM network and external networks, facilitating the routing of incoming calls to the GSM network and outgoing calls to external networks.

#### 1.5.4 Base Station Subsystem (BSS)

The Base Station Subsystem (BSS) is a crucial component of a GSM network, responsible for managing communication between mobile devices and the core network. It consists of two main elements: the Base Transceiver Station (BTS) and the Base Station Controller (BSC). The BTS facilitates wireless communication with mobile devices within its coverage area, handling data transmission and reception. The BSC optimizes network performance by managing radio resources, controlling handovers, and overseeing call setup and release. By strategically locating base stations, the BSS ensures comprehensive coverage, enabling seamless connectivity between mobile devices and the core network [8].

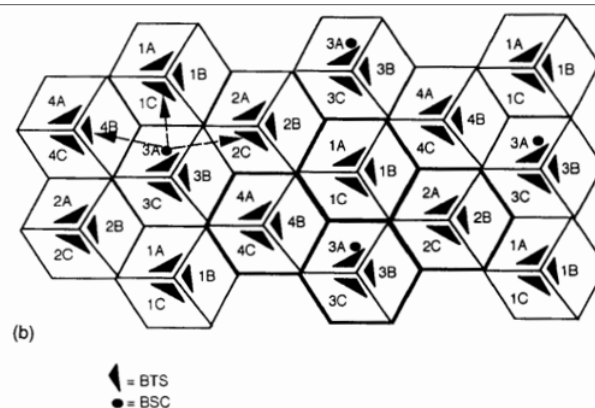


Figure 1.3: GSM system cell representation [8]

### 1.5.5 Operation and Support Subsystem (OSS)

The OSS, also known as the "operation support subsystem," is a part of the overall GSM mobile communications network architecture and is linked to elements of the NSS and the BSS. It also helps manage the BSS's traffic load and is used to monitor and operate the entire GSM network. It must be noted that as the number of BS increases with the scaling of the subscriber population some of the maintenance tasks are transferred to the BTS, allowing savings in the cost of ownership of the system. In general, OSS is a supervising system for the overall network.

### 1.5.6 EGPRS:GPRS/EDGE

The main application of the 2G network is to provide voice calls between two individuals in the network. In order to support the concept of packet service, the GPRS system was deployed as an overlay of GSM with two new network nodes, SGSN and GGSN, new interfaces and new functionalities in Base-station controller, BSC. EDGE stands for Enhanced Data rate for GSM Evolution. It is a further evolution of GPRS, giving the option for an increased system data rate using extended modulation schemes at the air interface with no impact to other parts and nodes of the system. The combination of GPRS and EDGE is usually referred to as EGPRS [9]. Gateway GPRS Support Node is abbreviated as GGSN. Incoming packets are routed by the GGSN to the mobile's current location. As a result, it must interface with the HLR in order to obtain the necessary location information for mobile terminating packet transfers. The Serving GPRS Support Node is the second node (SGSN). The SGSN establishes a mobility management context for a connected MS. The SGSN also ciphers packet-service traffic. This is distinct from the encrypted circuit switch traffic between the MS and BSC [10].

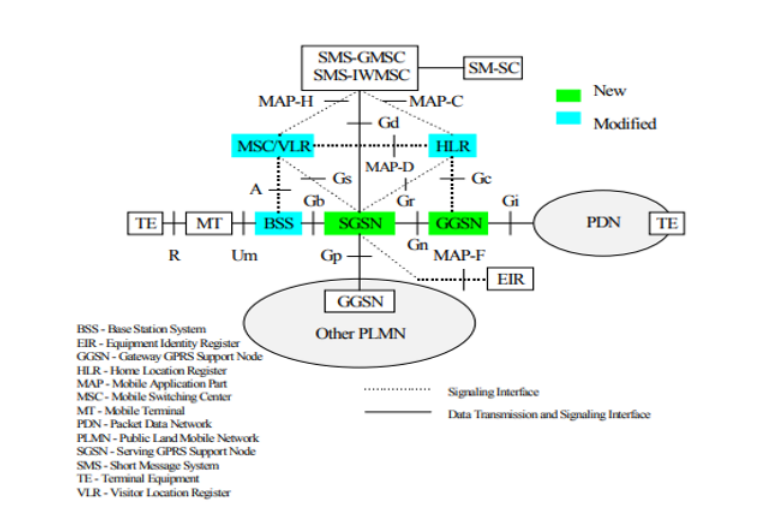


Figure 1.4: GPRS/EDGE network [10]

## 1.6 Third Generation network (3G)/UMTS

UMTS, introduced in 1999 as the 3rd generation (3G) mobile network, followed GSM/EDGE. It adhered to the European approach to 3G standardization and maintained backward compatibility with GSM through the 3GPP specification. UMTS shares a network structure similar to GSM.

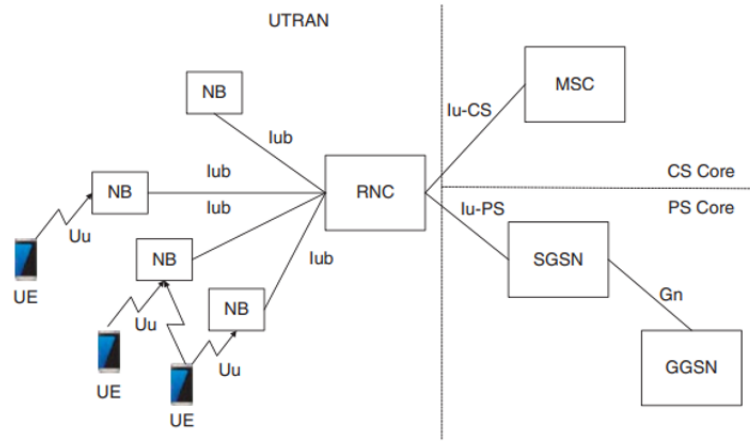


Figure 1.5: UMTS network architecture [11]

The mobile terminal for the 3G system is referred to as the User Equipment (UE). Most terminals are physically designed as dual-mode multiband devices, capable of supporting both 2G and 3G communications. The collective term for the radio access component of the network is the Radio Access Network (RAN), which encompasses both 2G and 3G technologies. Within the realm of WCDMA UMTS, the specific radio access is known as Universal Terrestrial Radio Access (UTRA) or UTRAN. In UMTS, the base-station controller is denoted as the Radio Network Controller (RNC). While the switching system can be shared between GSM and UMTS, there are notable distinctions. In UMTS, specifically within the UTRAN and the Core Network (CN), a new set of protocols is introduced, necessitating different hardware, software, and interfaces than those used in GSM [12]. The Core Network (CN) structure, derived from GSM, comprises two distinct domains that depend on user traffic:

- The CS domain handles circuit-switched traffic.
- The PS domain manages packet-switched traffic. Both GSM and UMTS networks rely on common network entities such as the Home Location Register (HLR), Authentication Centre (AuC), and Equipment Identity Register (EIR) for subscriber management, roaming, and service handling. The HLR contains subscriber information for GSM, GPRS, and UMTS [11]. The packet switched elements of the 3G UMTS core network architecture are mainly the same as the GPRS which are the SGSN and the GGSN. This architecture ensures efficient management of different types of traffic and facilitates seamless integration between GSM and UMTS networks.

## 1.7 Fourth Generation (4G)/LTE

LTE, or Long-Term Evolution, is the fourth-generation (4G) wireless communication standard that emerged after UMTS. Developed by the 3rd Generation Partnership Project (3GPP), LTE revolutionized mobile communication by delivering higher data rates, improved spectral efficiency, and enhanced network capacity. With its all-IP infrastructure, low latency, and simplified architecture, LTE provides seamless connectivity and exceptional throughput for time-sensitive data traffic. It introduced innovative radio access technologies like OFDMA and SC-FDMA, enabling faster data transmission and better spectral efficiency. LTE's compatibility



with existing GSM and UMTS networks allowed for a smooth transition and efficient utilization of infrastructure. Overall, LTE represents a significant advancement in telecommunication networks, setting the stage for the future of mobile communications [13].

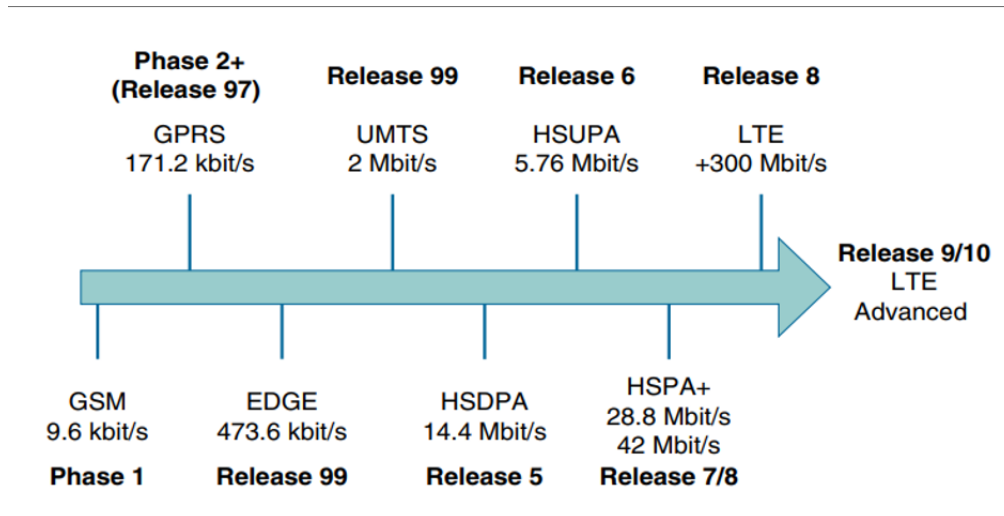


Figure 1.6: 3GPP Standard evolution [14]

LTE network was designed to offer solely Packet Switched (PS) services, as opposed to prior cellular systems' Circuit-Switched (CS) approach. It seeks to enable continuous Internet Protocol (IP) communication between User Equipment (UE) and the Packet Data Network (PDN) during mobility, with no disruption to end users' applications. While the term "LTE" refers to the evolution of radio access via the Evolved-UTRAN (E-UTRAN), it is complemented by an evolution of non-radio features known as "System Architecture Evolution" (SAE), which includes the Evolved Packet Core (EPC) network [14].

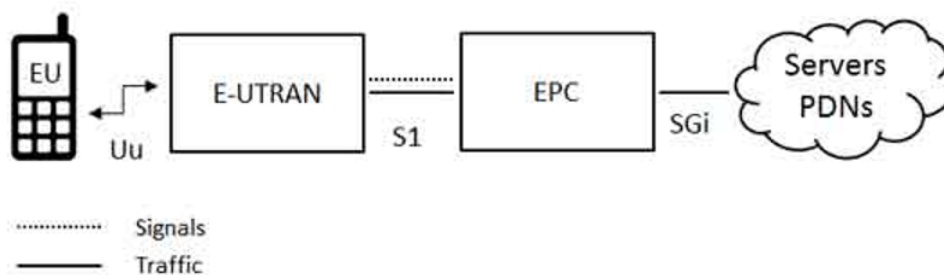


Figure 1.7: LTE network description [6]

In LTE, the transport of IP traffic from a gateway in the Packet Data Network (PDN) to the User Equipment (UE) is facilitated through the use of EPS bearers. An EPS bearer refers to an IP packet flow that is assigned a specific Quality of Service (QoS) level. This QoS level ensures that the bearer receives the required priority and resources for delivering the IP traffic effectively. EPS bearers play a crucial role in maintaining the desired level of service and ensuring efficient and reliable transport of IP traffic within the EPS network [15].



## 1.8 LTE Network Architecture

The LTE network architecture is a sophisticated system that combines essential components to ensure rapid and dependable IP connectivity. Figure 1.8 demonstrates the architecture's division into four primary high-level domains: User Equipment (UE), Evolved UTRAN (E-UTRAN), Evolved Packet Core Network (EPC), and the Services domain. These domains collectively play a vital role in facilitating seamless and efficient communication within the LTE network.

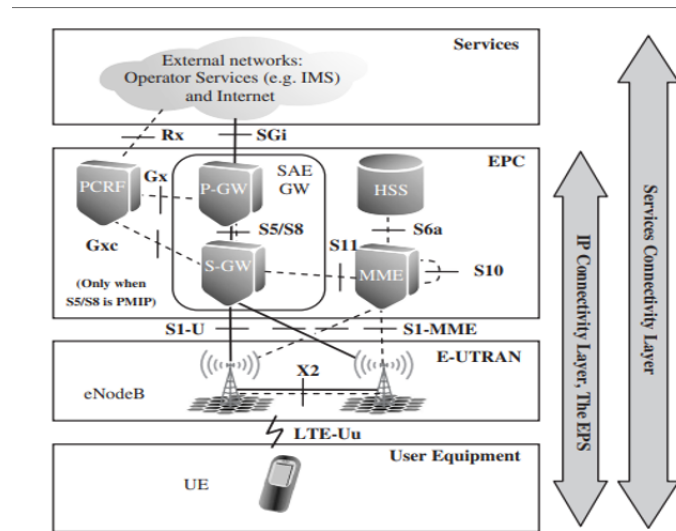


Figure 1.8: LTE network architecture [16]

UE, E-UTRAN and EPC together represent the Internet Protocol (IP) Connectivity Layer. This part of the system is also called the Evolved Packet System (EPS). The main function of this layer is to provide IP based connectivity, and it is highly optimized for that purpose only [UMTS EVOLUTION TO LTE]. The main elements of the EPS are introduced below:

### 1. User Equipment (UE):

The User Equipment (UE) serves as the communication device used by end users, typically a smartphone or data card. It incorporates the Universal Subscriber Identity Module (USIM), which is a separate module that identifies and authenticates the user, as well as provides security keys for protecting radio transmission.

### 2. E-UTRAN Node B (eNodeB):

The E-UTRAN consists of the eNodeB, serving as the radio base station responsible for radio operations within the system. Positioned strategically throughout the coverage area, multiple eNodeBs facilitate data transmission between User Equipment (UE) and the Evolved Packet Core (EPC). Acting as the termination point for radio protocols, the eNodeB ensures seamless connectivity between the radio connection and IP-based connectivity to the EPC. It establishes connections with neighboring eNodeBs for smooth handover operations, and Figure 1.9 illustrates the connections and primary functions performed across these interfaces [15].

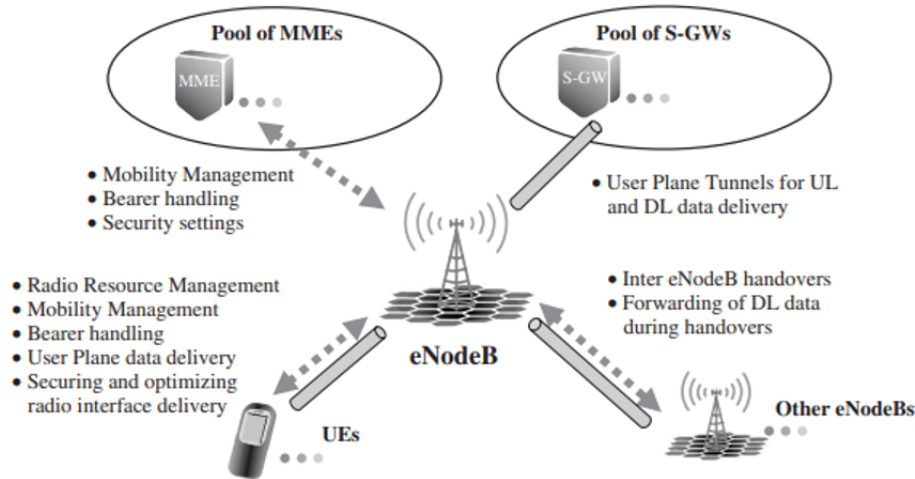


Figure 1.9: eNodeB main functions [16]

### 3. The Evolved Packet Core (EPC):

The Evolved Packet Core (EPC) is responsible for managing the flow of data between mobile devices and the internet or other networks, and it provides a variety of services, including mobility management, quality of service (QoS) control, and security. The EPC is made up of several different network elements as shown in the figure below:

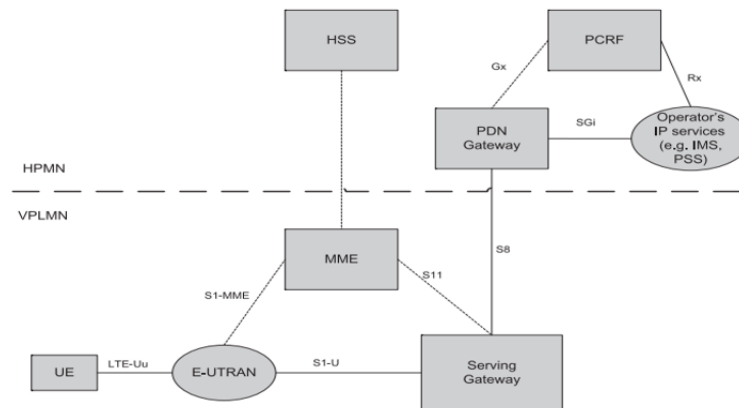


Figure 1.10: EPC architecture [17]

Below is a brief description of each of the components shown in the above architecture:

- The Home Subscriber Server (HSS) is a data base which stores subscription data for users, such as the EPS-subscribed QoS profile and any roaming access limitations.
- The Packet Data Network (PDN) Gateway (P-GW) serves as the communication bridge between the LTE network and external packet data networks (PDNs). Its primary functions include allocating dynamic IP addresses to users, routing user plane packets, and facilitating connectivity with the outside world.

- The serving gateway (S-GW) acts as a router, it is responsible for managing user data tunnels between the eNode-Bs in the radio network and the Packet Data Network Gateway (PDN-GW), which is the gateway router to the Internet [16].
- The Policy Control and Charging Rules Function (PCRF) is in charge of policy control decision-making as well as controlling the flow-based charging functionalities in the Policy Control Enforcement Function (PCEF), which is located in the P-GW.
- The Mobility Management Entity (MME) is a crucial network node responsible for signaling exchanges between base stations, the core network, and users. It facilitates tasks such as authentication, establishment of bearers, mobility management, handover support, and SMS/voice services. The MME handles authentication information exchange during network attachment, coordinates with other network components to establish IP tunnels, manages UE location and resource allocation, facilitates handovers between different access networks, and collaborates with IMS and CSFB functions for voice services [3] .

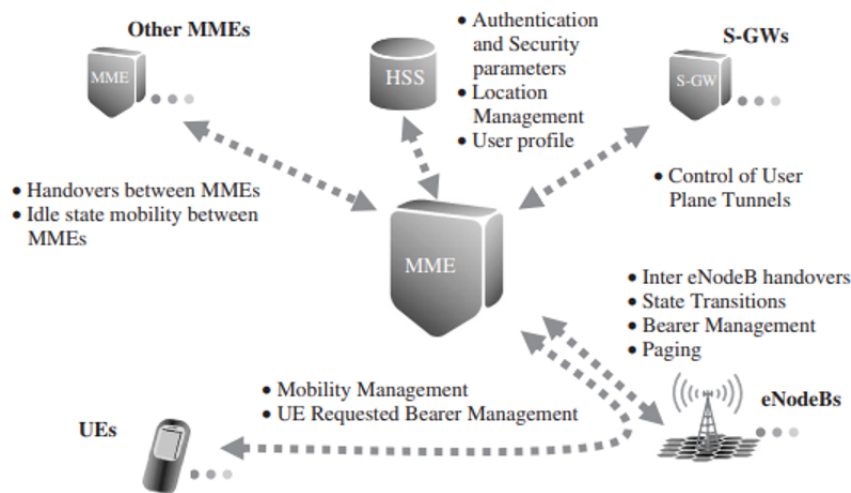


Figure 1.11: MME connections to other nodes [16]

The Services domain encompasses multiple sub-systems that offer a wide range of services. One notable example is the IP Multimedia Sub-System (IMS), which operates within the Services Connectivity Layer and enables the provision of services over the underlying IP connectivity. For instance, IMS can facilitate Voice over IP (VoIP) services and establish connections with traditional circuit-switched networks like PSTN and ISDN through its controlled Media Gateways. This allows for seamless integration between IP-based services and legacy communication systems [16].

## 1.9 Summary

Throughout this chapter, we have been able to see the principles of the cellular concept and also to see the state of the art in the evolution of mobile networks. This first chapter allowed us to distinguish between the many technologies employed in each type of network and thus

---

see the unique characteristics of each mobile network. The International Telecommunications Union has enabled the categorization of various mobile telephony technologies by the standardization of these various technologies. We can now see and distinguish between different types of networks, as well as understand their differences. Because each type of network has unique characteristics, it is now easy to navigate and refer to the various mobile networks. And, of course, each technology has its own performance and, as a result, each has its own key performance indicators.

## **Chapter 2**

### **Quality of Service and KPIs**

---

## 2.1 Introduction

The 4G/LTE optimization process is very complex task for the Mobile Network Operators (MNOs). This process includes the effects of multiple factors, which should be considered separately. Today, the MNOs are facing many challenges such as dynamically changing service requirements, technologies, competition, etc. In some ways, this has changed the MNOs' business model and new tools are required not only to manage the network, but also the subscribers. Imagine a situation where you are hardly able to hear what your friend is talking over the phone or the phone gets cut when you are talking something important. These things are highly undesirable and you do not want to get low quality service for paying high monthly bills. Communication plays a major role in today's world and to support it QoS has to be given maximum priority. It is important to differentiate the traffic based on priority level. Some traffic classes should be given higher priority over other classes, Example: voice should be given a higher priority compared to data traffic as voice is still considered as the most important service. It should be noted that more preference has to be given to customers who pay more to get better service, without affecting the remaining customers who pay normal amount. To realize all these things effective QoS schemes are needed.

## 2.2 Quality of service

Quality of service (QoS) refer to the measurement of the overall performance of a service experienced by the users of the network. To quantitatively measure the QoS packet loss, bit rate, throughput, transmission delay, availability, etc. In the absence of QoS, network data becomes disorganized and congests the network. This often leads to severe network performance degradation or even a complete network shutdown. Moreover, when QoS is low, security and data integrity can be jeopardized. People depend on the communication services to work, and poor QoS leads to poor work quality. Businesses face the need to provide reliable, consistent services for both staff and customers. Since QoS shapes the user experience, reputation can be negatively impacted when services are unstable. Ultimately, QoS mechanisms give network administrators the power to prioritize applications as determined by the needs of the business. This makes it easy to assign higher importance to particularly data delivery types over others.

## 2.3 QOS in mobile networks

The evolution of the mobile network has several phases, initially it is the deployment of the network which is followed by its commercial opening to finally be in operation. Once the network is in operation, the operator monitors its quality of service, the aim is multiple: it allows traffic to be optimized and areas lacking in QoS to be visualized in order to make the necessary modifications. As explained above, QoS is a performance control mechanism, this service can be applied to all circuit-switched services. In the context of mobile telephony, the difficulty will be in the mobility of the customer for several reasons:

- A call or other session may be interrupted while roaming, if the new base station is overloaded. Unpredictable handovers make it impossible to absolutely guarantee quality of service during a session initiation phase.
- A crucial part of QoS in mobile communications is quality of service, involving outage probability (the probability that the mobile station is outside the service coverage area,

---

or affected by co-channel interference, crosstalk for example) blocking probability (the probability that the required level of QoS cannot be offered) and scheduling starvation.

The expectations of the mobile network are therefore as follows:

- Network availability (probability of obtaining a new call).
- Maintaining communications (the probability of a communication being cut off).
- Auditory quality of the communication (signal strength, interference).

Mobile QoS measurements are then made using the following methods:

1. OMC measurements: these are obtained by collecting radio statistics from the OMC among the operators, which make it possible to assess the behavior of an operator's mobile network.
2. Drive tests: radio measurements of one or more mobile networks, operators take the role of users and measure quality of service, QoS problems are discovered by employees.
3. END to END measurements: for the evaluation of the quality of service as perceived by the end user.
4. Customer complaints: this is an important source of network service quality that cannot be ignored.
5. Protocol analyzers: they are connected to the BTS, BSC and MSC over a given period to check for problems on the mobile network. When a problem is discovered, it is escalated to employees for analysis.

In our work we are going to focus on the OMC measurements. the measurements are based on the collection of counters calculated by the network equipment.

## 2.4 Counters

Counters play a crucial role in monitoring the performance of mobile networks. They are software variables that are stepped according to specific criteria, and they are divided into three main types:

- Event Counters: Event counters are used to count events such as seizure, congestion, handover, and so on. Event counters are only incremented and have values between 0 and a unit-specific maximum. Event counters are used to monitor and analyze network behavior and help network operators to identify and diagnose problems that may occur in the network.
- Level Counters: Level counters show the status of units in a certain condition, such as the number of blocked devices or available traffic channels, etc. Level counters can be both incremented and decremented, and they provide a real-time snapshot of network status. They are used to monitor and manage network resources and help network operators to optimize network performance.

- **Accumulation Counters:** Accumulation counters are used for summing the value of level counters during a specific time period. They are used in conjunction with an event counter, which records the number of summations performed. Accumulation counters are used to monitor network usage and help network operators to identify trends in network behavior.

The counters give an absolute number whose exploitation requires relativizing it to the cardinality of its starting set. In other words, instead of judging, for example, the level of congestion on a cell by the number of attempts which have congested, one should, from the total number of attempts, determine the rate of congestion, which this time expresses the situation better. It is in this context that performance indicators fit. They are deduced from the raw meters read from the network and give an overall state in a relativized space.

## 2.5 Key Performance indicators

The key performance indicators or KPIs of an LTE access network help monitor and optimize the performance of the radio network, in order to provide better service to the subscribers or to obtain better use of the installed network resources. KPIs are rates calculated from the counters discussed previously using formulas and compiling different data. The KPIs essentially evaluate the maintenance of the call, the volume of traffic, the quality of service on the whole network. The KPIs thus make it possible to detect faulty cells, peak hours, etc. A limit threshold is determined for each KPI, if it is exceeded an alarm is sent to the supervision to indicate the presence of a problem on the function that the KPI measures.

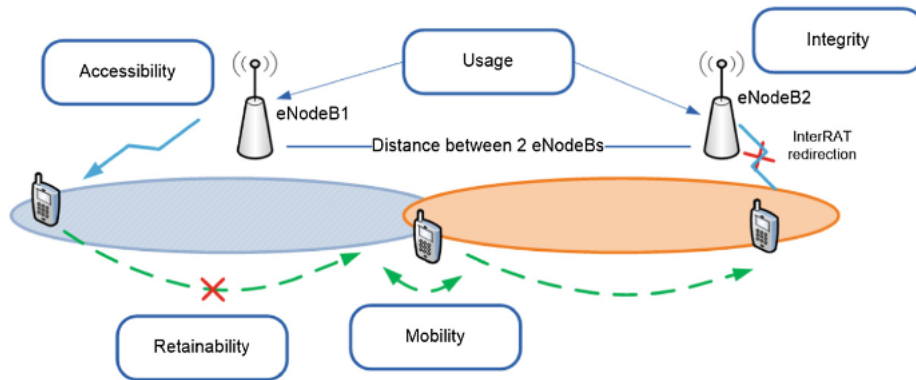


Figure 2.1: 4G/LTE Key performances indicators [18]

The 3GPP standardized 4G/LTE networking KPIs are presented in Fig. 1 and are categorized as follows: accessibility, retainability, mobility, integrity, availability and usage.

1. Integrity KPIs' are used to measure the character or honesty of network to its user, such as what is the throughput, latency which users were served.
2. Accessibility, which has KPIs indicating the possibility to access to a service: RRC connection establishment, paging records, and discards. It is a combined metric including RRC, S1, and E-RAB establishment success rate. In the case of poor accessibility, each success rate must be analyzed individually.



3. Retainability, which is defined as the ability of a user to retain the E-RAB once connected for the desired duration, which has KPIs indicating the ability to hold/sustain the call( call drop, call completion, E-RAB drop, E-RAB normal release, RRC connection re-establishment IP incoming traffic error rate).
4. Mobility, which has KPIs indicating performance of handovers (handover preparation, handover success rate, and handover failure rate).  
Reasons for poor mobility include but are not limited to missing neighbor relations, poor radio conditions, or badly tuned handover parameters.
5. Usage, which has KPIs indicating how LTE network is loaded in terms of data volume, throughput, number of users (active and connected), PRB usage, and cell availability.
6. Availability KPI's are used to measure the availability of network, suitable or ready for users to use services.



Figure 2.2: Key performances indicators types

Based on the estimation of these indicators, it must be generated a set of periodic reports of QoS according to the desired temporal scale: daily, weekly, monthly, etc., which will allow the service provider to get an overall assessment of the network status and provided services. The relationship between event counters, KPIs and QoS reports is shown in [2.3](#).



Figure 2.3: Counters,KPIs and QOS reports

The service provider could, among other things, accomplish the following goals by keeping an eye on key performance indicators and evaluating QoS reports:

- Identifying occasional faults in the hardware equipment of base stations.
- Identifying some interference issues and service degradations may encourage the implementation of corrective measures like new frequency allocations, antenna adjustments, or radio parameter changes.
- Detecting some congestion or oversizing issues with the capacity would require doing a more proper sizing in accordance with user demand.
- Monitoring network performance and noting fluctuations and trends enables the service provider to prepare some preventative measures [19].

In our work we are going to focus mostly on the KPIs of two basic LTE procedures : attach and tracking area update.

## 2.6 The attach procedure

The LTE attach call flow describes the process that a user device undergoes to establish a connection with the LTE network. The attach procedure is the first step in the process of accessing the network, and it involves the exchange of various messages between the user device and the LTE network.

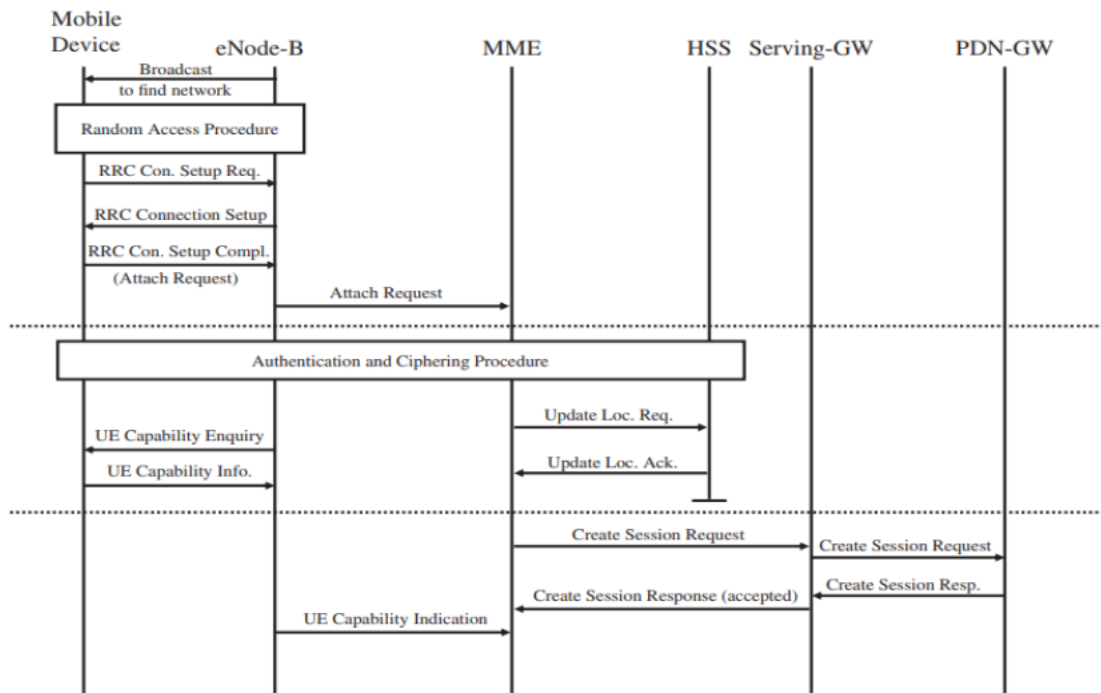


Figure 2.4: Attach procedure call flow [3]

The following is a brief description of the LTE attach call flow:

1. **Cell Search:** The user device scans available LTE frequencies to find a suitable cell and acquires system information about the cell.
2. **Random access preamble:** After acquiring the system information, the user device initiates a random access procedure by sending a random access preamble with a unique code to the e-NodeB.

- 
3. **Random access response:** The e-NodeB responds to the random access preamble by sending a random access response message, including a timing advance value.
  4. **Radio Resource Control Connection request:** The user device broadcasts an RRC connection request message to create an RRC channel with the e-NodeB and the core network.
  5. **RRC Response:** If access is granted, the network responds with an RRC connection setup message, including assignment parameters for a specialized radio signaling bearer (SRB- 1).
  6. **RRC connection setup complete:** The user device sends an RRC connection setup complete message, including information about the previously connected MME (Mobility Management Entity) and an Attach Request message.
  7. **Authentication request:** Mutual authentication between the network and the mobile device takes place, ensuring secure connectivity. A Security Mode Command message enables integrity checking and encryption of messages.
  8. **Update Location Request:** The MME notifies the HSS of successful authentication by sending an update location request message.
  9. **Session Creation:** The MME initiates the core network session establishment procedure, creating a tunnel for user IP packets via a create session request message.
  10. **Establishing a tunnel in the radio network:** The e-NodeB and the serving-GW establish a user data tunnel through an Initial Context Setup Request message, configuring signaling radio bearers and data radio bearers [20]. Upon completion of these steps, the user device can send and receive IP packets over the LTE network. The MME handles overall session management, while the e-NodeB and serving-GW facilitate data transfer. Additional RRC Reconfiguration messages may be exchanged to set up neighbor cell measurements and reporting for potential cell switches.

## 2.7 Tracking area update procedure

During a Tracking Area Update (TAU) procedure, the user equipment (UE) initiates the process by sending a Tracking Area Update Request message to the eNodeB. The message includes the UE's current Globally Unique Temporary Identity (GUTI) or International Mobile Subscriber Identity (IMSI), old Tracking Area Identity (TAI), and EPS bearer status information. The eNodeB forwards this message to a Mobility Management Entity (MME). If necessary, a new MME is selected and authentication takes place, with communication between the old and new MMEs occurring through GTP-C context request messages [17]. Upon successful authentication, the new MME examines whether a serving gateway (GW) change is required and notifies the old MME of its readiness to assume control of the UE through a context acknowledge message. The old MME starts a timer and waits for the subscriber record to be canceled. Meanwhile, the new MME sends a GTP-C create bearer request message to the chosen serving GW, establishing new S1 tunnels and updating the PDN GW. The traffic path from the PDN GW to the new serving GW and eNodeB is modified accordingly. Concurrently, the MME updates the Home Subscriber Server (HSS) and the HSS cancels the subscriber record in the old MME. Finally, the UE receives a Tracking Area Update Accept NAS message containing a new GUTI and new TA(L). The UE acknowledges with a Tracking Area Update Complete

message, indicating the completion of the TAU procedure.

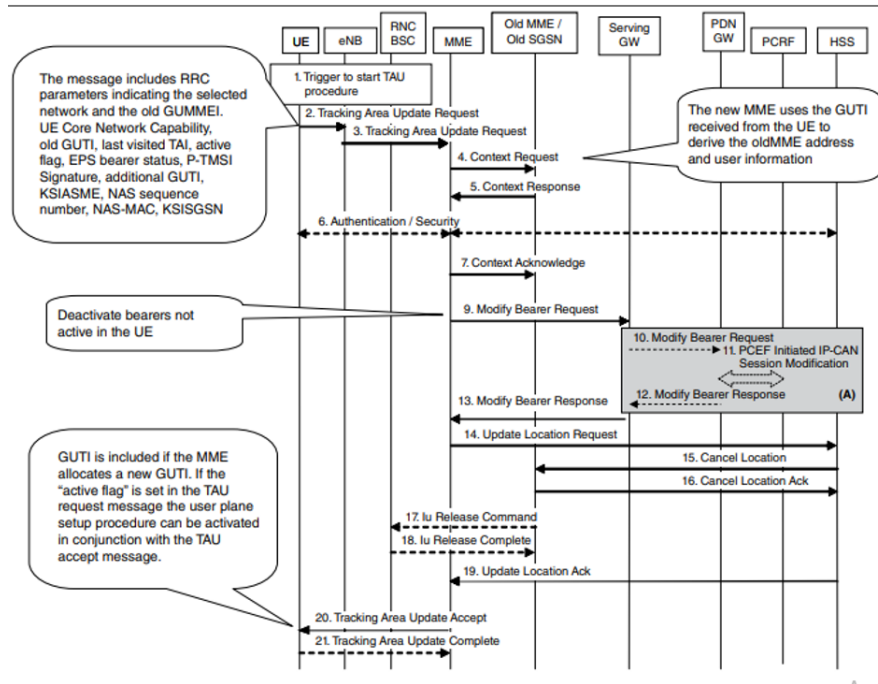


Figure 2.5: Tracking area procedure callflow [14]

## 2.8 Manual monitoring

Creating comprehensive reports for monitoring key performance indicators (KPIs) in a packet switching LTE network is a meticulous and time-consuming task. The engineer receives diverse data in different formats and time intervals every morning, requiring careful validation and cleansing to ensure accuracy. After uploading the data, the engineer performs in-depth analysis, including calculations, statistical analyses, and data aggregation, to derive meaningful KPI measurements. The analyzed data is then transformed into visually appealing reports with charts, graphs, and tables to summarize the network's performance. The engineer interprets the findings, compares KPIs with historical data or thresholds, and provides recommendations for addressing performance issues and optimizing KPIs. This process demands expertise, attention to detail, and a methodical approach. Documentation of reports and findings is crucial for future analysis and effective communication with stakeholders. Automation and streamlining of processes are essential for efficiency and accuracy in this complex task.

## 2.9 New monitoring application

Our proposed application aims to automate the process of data collection, integration, staging, analysis and visualization. It is a question of presenting the different KPIs according to several forms of display (curves, histograms, tables), and this to allow an intelligent analysis of customer experiences and a diagnosis of possible quality of service problems. The workflow is shown in figure 2.6:

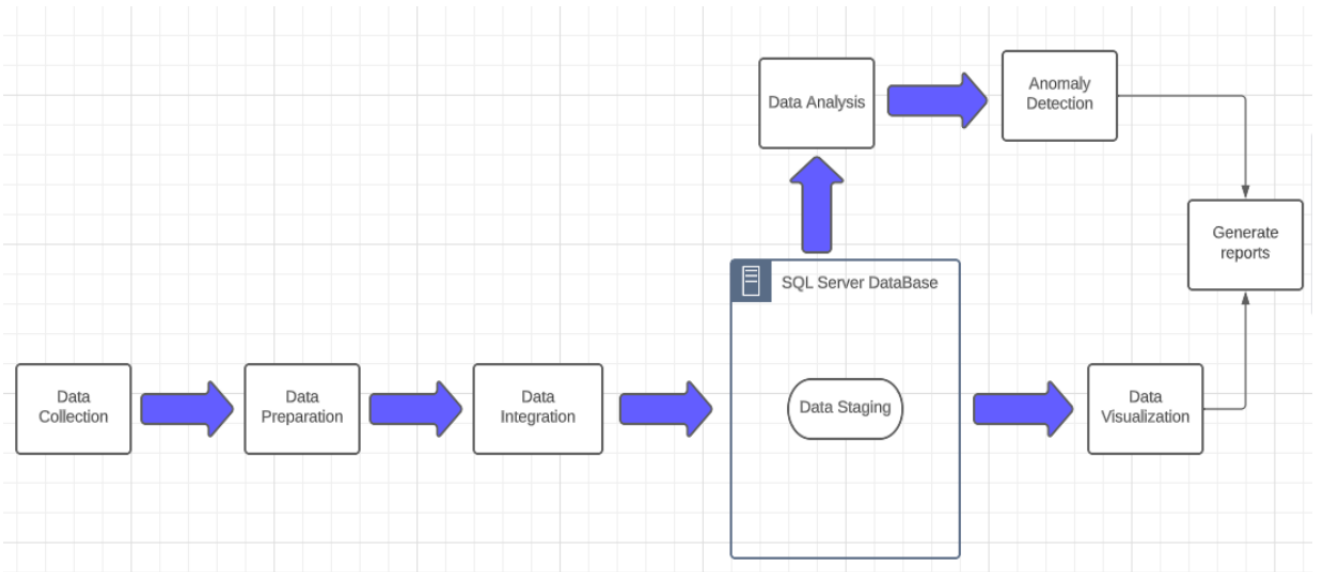


Figure 2.6: Monitoring System Workflow

Here's a general outline of the steps we would need to follow to create reports:

1. **Data Collection:** Ensure that the data for each hour of the previous day is available on the engineer's laptop. This could involve receiving data files, accessing a database, or any other means of obtaining the required data.
2. **Data Preparation:** Review the data received and ensure its integrity and accuracy. Cleanse and preprocess the data if necessary, addressing any inconsistencies or missing values.
3. **Data Integration:** After preparing the data, we need to upload it to the desired table in our data base so it would be stored there in historical tables.
4. **Data Staging:** Add or create the needed columns for the analysis purpose.
5. **Data Visualization:** Enhance the reports with visual elements to make them more understandable and insightful. Use appropriate visualizations such as line charts, bar graphs, or heat-maps to represent the KPI trends and comparisons over time.
6. **Data Analysis:** Perform the necessary calculations and analysis on the collected data to derive the desired KPI measurements. This could involve aggregating data for each hour, calculating average values, identifying outliers, or computing other relevant metrics based on the specific KPIs of interest.
7. **Report Generation:** Based on the analyzed data, create reports that summarize the KPI measurements for the previous day. This could be in the form of tabular reports, charts, graphs, or any other suitable format that effectively presents the information.

---

### 2.9.1 Functional needs

Our application must fulfill the following conditions:

- Collection of our data.
- Save the data in a local folder.
- Send the saved folder into an FTP server.
- Creation and backup of the database.
- Being able to access the database as soon as a connection is established.
- Display of query execution result.
- The application must be updated each time the actualization button is pressed.
- The application allows the operator engineers to visualize the different KPIs and diagnose network problems.

### 2.9.2 Non-functional needs

These are the needs that characterize the system. These are performance requirements, hardware type or design type. These needs may relate to constraints related to the implementation (programming language, operating system, etc.) or to general interoperability (not eating up all the resources of the machine). These requirements may be set up by the users (optional functions), or by the developer (implementation constraints). Among the non-functional needs of our application we mention:

- The application must be efficient and must ensure continuity of operation.
- The application must be efficient in terms of time.

## 2.10 Summary

In this chapter we talked about the quality of service and its importance in the different businesses. We then talked about the QOS in mobile networks and how it is evaluated. We proceeded to talk about the role of counters in monitoring the performance of mobile networks, and the key performance indicators. Our work focuses mainly on the KPIs of two basic LTE procedures: attach and tracking area update. These two procedures have been thoroughly explained in the chapter. After that, we discussed the manual monitoring and the problem that our application aims to solve. The chapter ends with a brief description of the role of our mechanism and the different implementation steps.

## **Chapter 3**

# **Implementation and Results**

---

## 3.1 Introduction

Data integration is the process of bringing data from disparate sources together to provide users with a unified view. The premise of data integration is to make data more freely available and easier to consume and process by systems and users. Data integration contributes to QOS performance by providing high-quality information for analysis. This is extremely valuable if you are dealing with big data, but no matter the size of your company or customer base, you need to know what is happening. When you use a data integration tool, all of your information is neatly gathered for you and presented in the source of your liking. If you choose a tool with more features, it will also offer you detailed reports on what is going on with your data, and some even notice certain trends and patterns. This is crucial for improving QOS and detecting in a timely manner what needs to be fixed and improved.

## 3.2 Development Tools

### 3.2.1 Python

Python is a high-level, interpreted, and general-purpose programming language that emphasizes code readability and simplicity. It was created by Guido van Rossum and first released in 1991. Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming. One of Python's distinguishing advantages is its clean and simple syntax, which encourages efficient and expressive coding. It defines code blocks with indentation and whitespace, which improves code readability. Python's broad standard library includes modules and functions for a variety of activities such as file I/O, networking, web development, data manipulation, and more [21]. Python's popularity has skyrocketed due to its adaptability and breadth of uses. It is frequently used in web development, data analysis, scientific computing, artificial intelligence, machine learning, automation, and scripting, among other things. Python's vast and active community has helped to produce a slew of third-party libraries and frameworks, significantly enhancing its capabilities.

### 3.2.2 FileZilla FTP

FileZilla is a popular free and open-source FTP (File Transfer Protocol) software. FileZilla allows users to transfer files between a local computer and a remote server. It supports various file transfer protocols such as FTP, FTPS, and SFTP. FileZilla has an easy-to-use interface with a dual-pane architecture that allows users to navigate and manage files on both local and remote systems. Drag-and-drop file transfers, directory comparison, remote file editing, and the option to resume interrupted transfers are all available. FileZilla is compatible with Windows, Mac OS X, and Linux. Web developers, system administrators, and anyone who need to upload or download files to and from servers all use it.



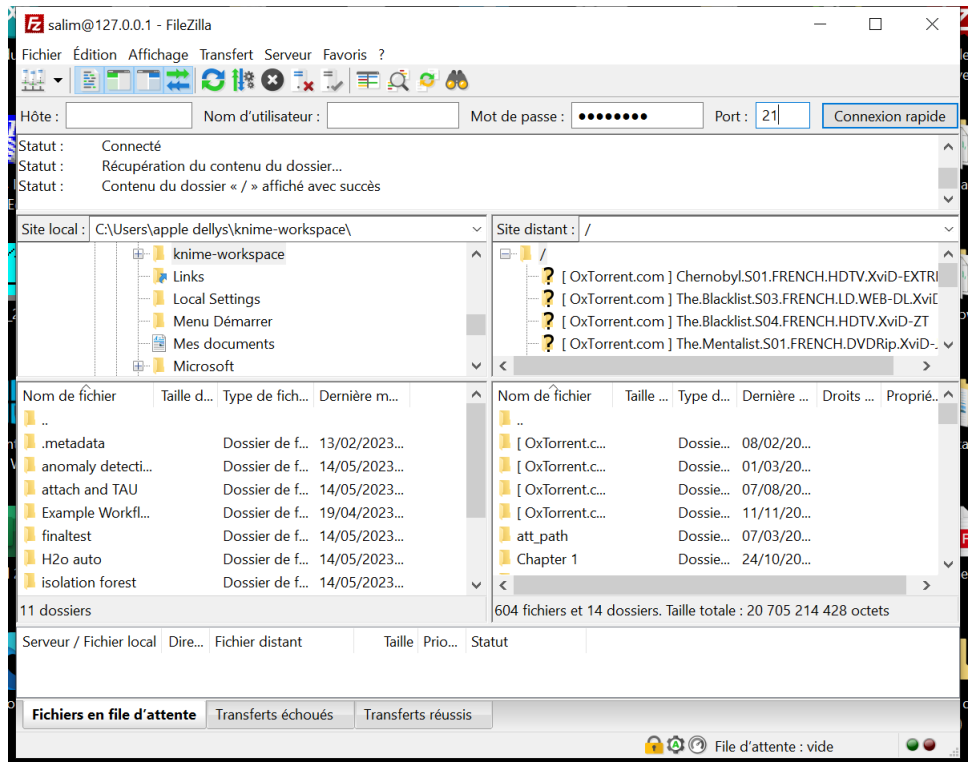


Figure 3.1: FileZilla client interface

### 3.2.3 Pentaho Data Integration

Pentaho is an open-source BI and data integration platform that offers tools for data extraction, transformation, and loading (ETL), reporting, analytics, and dash boarding. It enables users to connect and analyze data from several sources, generate interactive reports, and create visually appealing dashboards. Pentaho is adaptive, scalable, and flexible enough to suit a wide range of business requirements. It is offered in community and enterprise editions, giving users alternatives based on their needs. Overall, Pentaho enables businesses to better use their data, make data-driven decisions, and improve business performance [22].

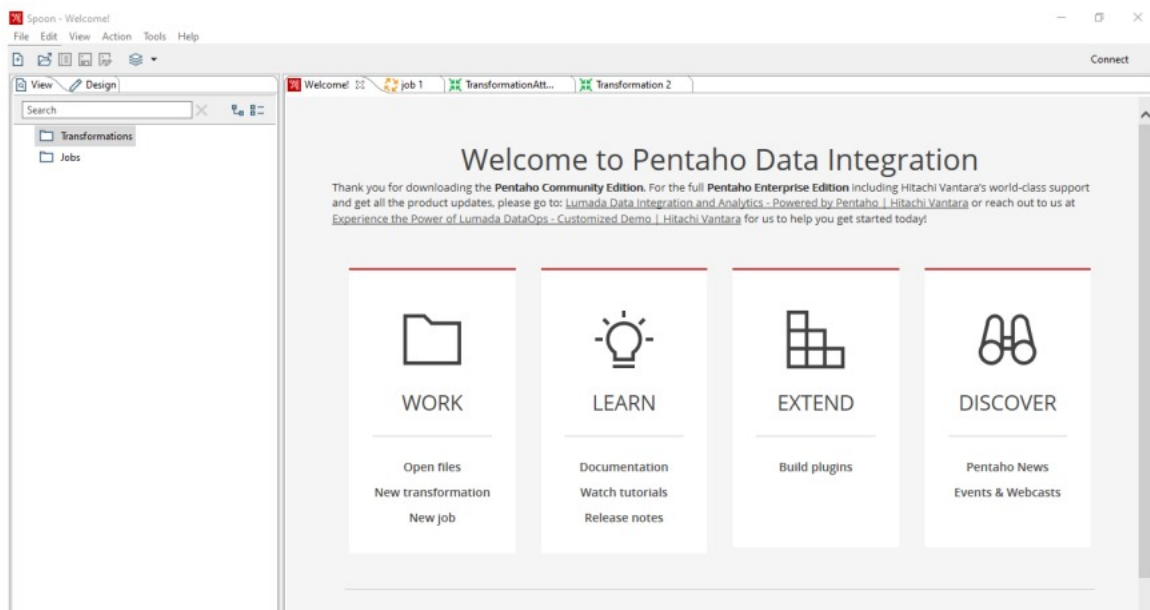


Figure 3.2: Pentaho User interface

### 3.2.4 KNIME Analytics Platform

KNIME Analytics Platform is an open-source data analytics and visual workflow solution that allows users to integrate, analyze, and visualize data in real time. Users can easily build and execute complex data workflows using its user-friendly graphical interface. The platform has significant data integration features, allowing users to connect and manage data from a variety of sources, including databases, files, and online services. KNIME has a robust set of analytics and machine learning algorithms that allow users to execute a variety of tasks such as classification, regression, clustering, text mining, and more [23]. KNIME also enables interactive data visualization, allowing users to generate dynamic and visually appealing charts, graphs, and maps in order to effectively communicate data insights. The platform encourages cooperation by allowing workflows to be shared and allowing for remote execution. It is also scalable and can easily interface with big data technologies such as Apache Hadoop and Apache Spark, allowing users to efficiently handle massive datasets. The KNIME community is extremely active and supportive, with rich documentation, tutorials, and example workflows available to help users at every stage. In conclusion, the KNIME Analytics Platform is a strong and versatile platform that enables users to successfully integrate, analyze, and visualize data, making it a significant asset for data-driven enterprises.

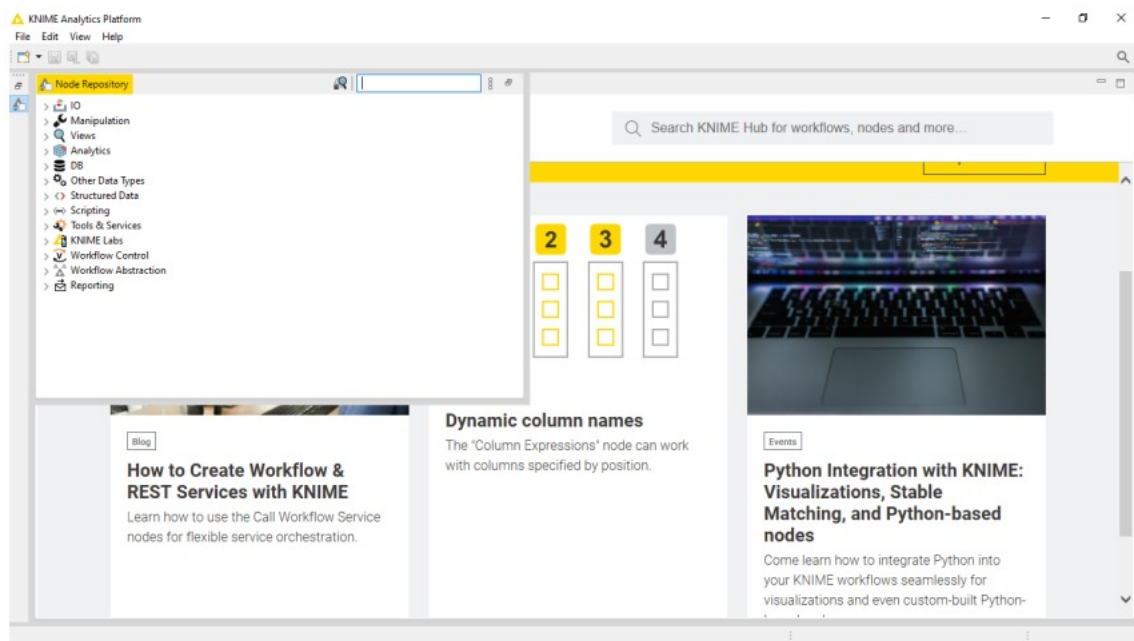


Figure 3.3: KNIME User interface

### 3.2.5 Microsoft SQL Server

Microsoft SQL Server, usually known as MS SQL, is a comprehensive relational database management system (RDBMS) developed by Microsoft. It provides a strong foundation for storing, managing, and retrieving structured data. MS SQL Server includes a wide range of features for data management, querying, and administration. The SQL query language, which stands for Structured Query Language can be used to execute actions such as filtering, sorting, aggregating, and joining data across different tables. MS SQL Server also offers stored procedures, triggers, and user-defined functions, which allow users to add business logic and custom actions into their databases. It also supports programming languages such as C, Java, and Python, allowing developers to create apps that interface with the database using familiar programming paradigms. MS SQL Server's versatility, combined with its scalability, security,

and integration capabilities, makes it a popular choice for managing and leveraging data in a wide range of applications and industries [24].

### 3.2.6 Microsoft Power BI

Power BI is a complete business intelligence platform built by Microsoft that allows users to connect to many data sources, transform data into meaningful insights, and generate interactive visualizations and reports. Power BI Desktop's intuitive drag-and-drop interface allows users to quickly create sophisticated dashboards and reports. The platform interfaces with diverse data sources easily, provides real-time data streaming, and allows automatic data refreshing. Power BI's sharing tools encourage collaboration by allowing users to securely share reports with colleagues and stakeholders. Power BI also has mobile apps for accessing reports while on the go. Power BI Report Server allows enterprises that want on-premises reporting to host and share Power BI reports within their own environment. With Power BI Report Server, organizations can maintain data security and have greater control over their reporting infrastructure. Overall, Power BI simplifies data analysis and reporting, empowering organizations to make informed decisions based on visually appealing and up-to-date information, both in the cloud with Power BI service and on-premises with Power BI Report Server.

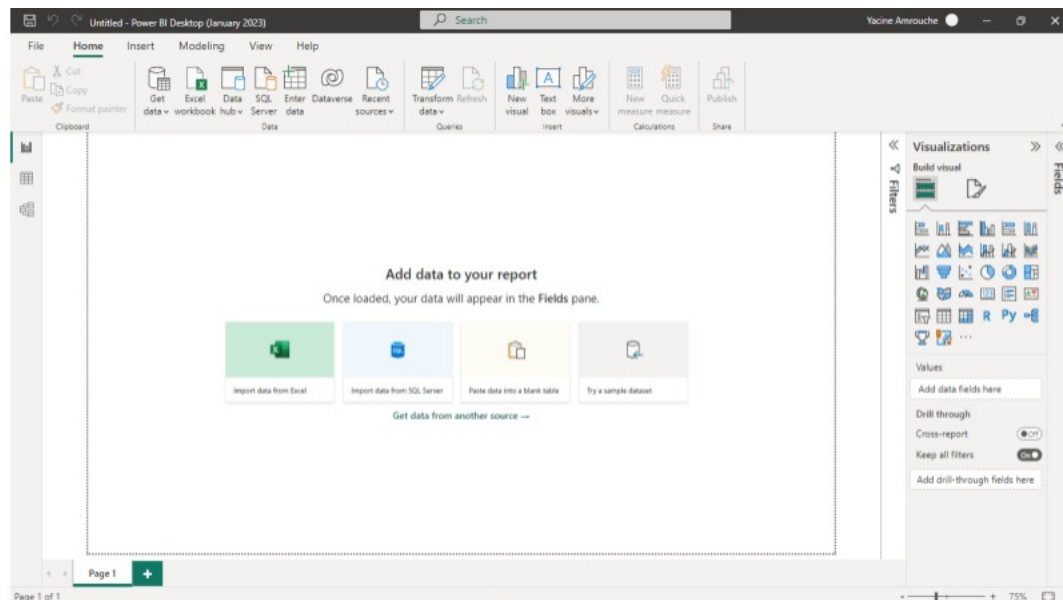


Figure 3.4: Power BI User interface

## 3.3 Environment Preparation

It is critical to prepare the environment before creating the system to ensure a smooth and efficient implementation. Setting up the appropriate infrastructure components is part of the environment preparation.

### 3.3.1 Creation of FTP server

FTP (File Transfer Protocol) is a standard network protocol used for the transfer of files from one host to another over a TCP-based network. FTP servers are the software solutions used for transferring files across the internet. They are primarily used for two essential functions,

---

“Put” and “Get.” It allows uploading files to the server from the client device and downloading files from the server on the client device.

First, we need to set up an FTP server for storing the data files using FileZilla, we begin by launching FileZilla Server and configuring the general settings by providing the server’s IP address and port. Next, we create user accounts with the necessary credentials and assign their home directories to our desired storage location. Careful attention is given to setting the appropriate permissions, enabling file uploads and downloads, and customizing folder access and user quotas if needed. We then designate shared folders specifically for storing the data files, configuring the permissions to allow read and write access. Encryption is enabled and passive mode settings are fine-tuned to ensure secure and efficient file transfers. After saving the changes, the FTP server is launched, allowing users to connect using FileZilla Client or any other FTP client program by entering the server’s IP address, port, username, and password. This setup simplifies the uploading and downloading of files, facilitating efficient data management and exchange throughout our project. The configuration of the FileZilla FTP server is a crucial step in ensuring secure storage and seamless access to our project-related data files.

### **3.3.2 Creation of SQL database**

A database is an organized collection of structured information, or data, typically stored electronically in a computer system. A database management system (DBMS) typically has control over a database[25]. The term “database system,” which is frequently abbreviated to “database,” refers to the combination of the data, the DBMS, and the applications that are connected to it. To create our database we need to use Microsoft’s SQL Server Management Studio (SSMS). First, After Opening SSMS, We connect to the SQL Server instance and navigate to the “Databases” folder in the Object Explorer. We then select “New Database” to open the New Database dialog box. We give the database a unique name, grant administrative capabilities to the owner, and tweak parameters like collation, recovery model, and compatibility level. We configure file and file group settings, describing the location, size, and arrangement of data files. Optional extras include default schema, confinement type, and auto growth parameters. If required, we add advanced capabilities such as file stream storage, database snapshots, encryption, and change tracking. Clicking “OK” generates the necessary T-SQL script and creates the database.

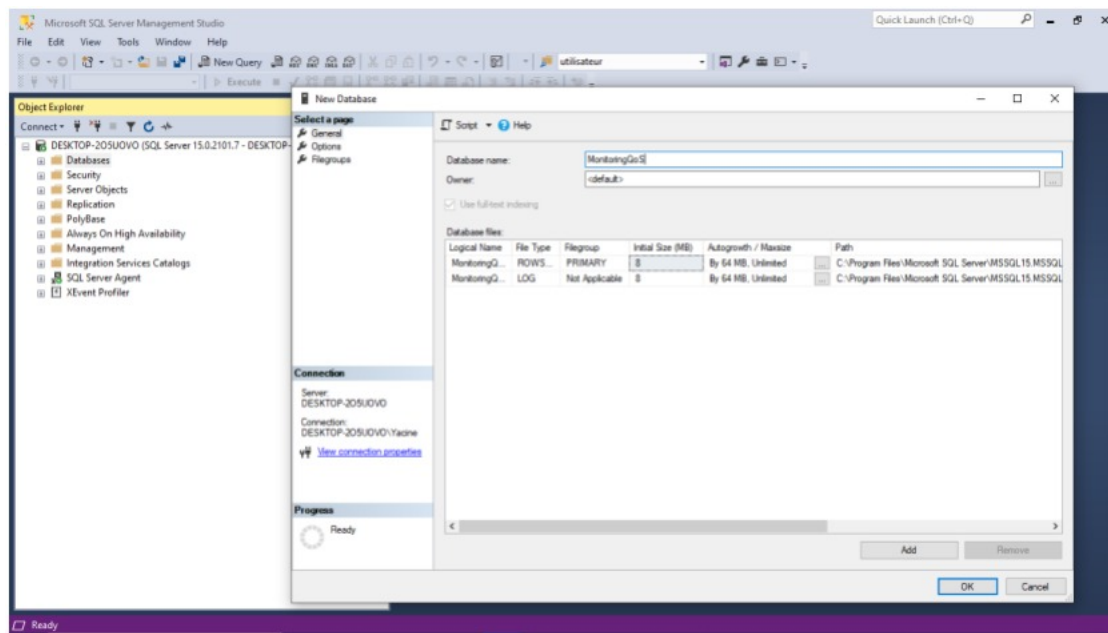


Figure 3.5: Creation of SQL database

After creating the database, the next step is to create a user account and grant the necessary privileges for database management. To do this, navigate to the security folder in the Object Explorer menu and select "New Login" by right-clicking on it. The login tab shown below will be displayed in order to fill the new user details. We fill in the login name and password for the user's credentials. Then we need to specify the default database and assign the user to relevant database roles such as "db\_owner", "db\_datareader" or "db\_datawriter" from the server roles page. Once configured, the user will have the required access and privileges to connect to the SQL Server instance and work with the designated database.

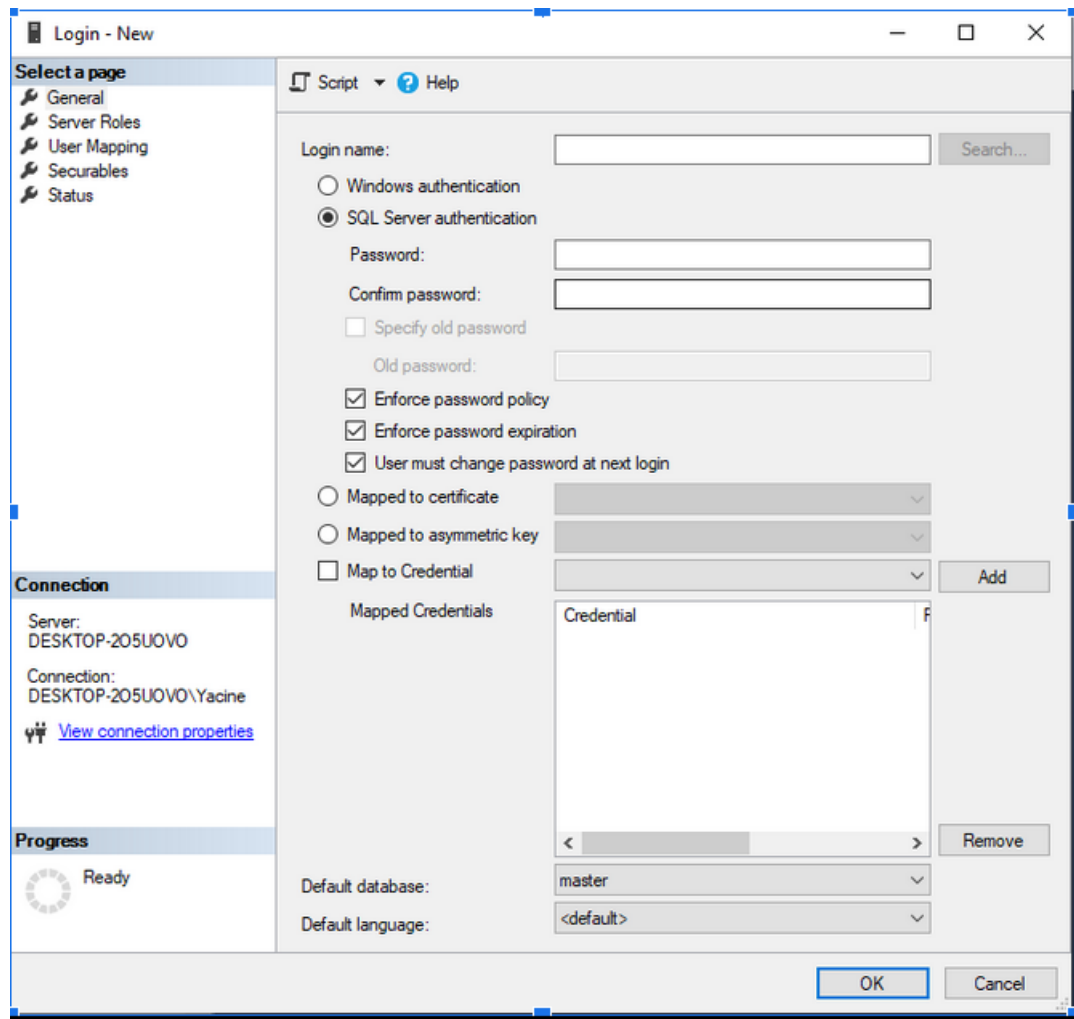


Figure 3.6: Login-New tab

## 3.4 Data Description

Before beginning the system implementation, it is important to have a thorough understanding of the data and its properties. First of all, what is data? If you look for the word data in the dictionary you will find that it is "facts and statistics collected together for reference or analysis." Or in computing terms, data is information that has been translated into a form that is efficient for movement or processing. In our project, the data consists of KPI counter values that are recorded at regular intervals of time for a specific network component. These KPI counter values are saved in a sequence of CSV files, each corresponding to a different type of KPI counter. The CSV files are named in a specified way to aid identification and interpretation. This naming strategy provides useful information about the content of each CSV file, allowing us to get insights before opening them. The figure [3.7](#) illustrates the naming strategy that was deployed:



HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_GGSN_2	4/4/2023 2:14 AM	Microsoft Excel C...	11 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_GGSN_4	4/4/2023 2:14 AM	Microsoft Excel C...	28 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_01_S1 mode MM	4/4/2023 2:14 AM	Microsoft Excel C...	12 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_02_S1 mode MM-S1	4/4/2023 2:14 AM	Microsoft Excel C...	12 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_03_S1 mode MM-SM	4/4/2023 2:14 AM	Microsoft Excel C...	12 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_04_S1 mode MM-SM	4/4/2023 2:14 AM	Microsoft Excel C...	12 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_05_S1 mode SM	4/4/2023 2:14 AM	Microsoft Excel C...	11 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_06_Other	4/4/2023 2:14 AM	Microsoft Excel C...	7 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_07_S1 mode MM(3)	4/4/2023 2:14 AM	Microsoft Excel C...	3 KB
HOST11_pmresult_60_202304040100_202304040200_SK_MG_4G_MME_07_S1 mode MM(4)	4/4/2023 2:14 AM	Microsoft Excel C...	9 KB

Figure 3.7: Naming scheme of csv files

In our naming scheme for the CSV files, the first word represents the name of the server from which we retrieve the records. In this case, it is “Host11”. The next element is the Granularity Period, which indicates that we have records at a 60-minute interval, making them hourly. Following that is the Time Interval, which specifies the timeframe during which the records were taken. For example, it shows that the records span from 01:00 AM to 02:00 AM on April 4, 2023. The “SK MG” identification in the name refers to the engineer who generated the CSV files, which provides assistance to people who deal with them. The “4G” code identifies the network type, which in this case is the LTE network. Following that, the name refers to the specific network component to which the counters are linked. It is the Mobility Management Entity (MME) in our scenario. The number “05” in the name represents the code for the table inside the file, providing information about the content of the table and the included counters. In our work, we will deal only with files with table codes from 01 to 08 because only these tables includes the needed counters for our work. Lastly, the last part of the name indicates the type of counters present in the file. In this case, “S1 mode SM” signifies that the counters are related to the Session Management procedure and are recorded from the S1 interface, which is the interface between the MME and the eNodeB.

The CSV files contain structured data organized in a tabular format, with each row representing an individual record and each column denoting a specific attribute. Figure 3.8 outlines the content and organization of the data within the CSV files:

Result Time	Granularity	Object Name	Reliability	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac	CS Fallbac
minutes			times	times	times	times	times	times	times	times	times	times	times	times	times	times	times	times
4/4/2023 1:00	60	USN9810_Dely/Whole System:USN9810_Dely	Reliable	3051	5	1691	1355	227940	102	146491	81347	0	0	0	0	0	0	0
4/4/2023 1:00	60	USN9810_Blida/Whole System:USN9810_Blida	Reliable	2629	2	1508	1119	214246	122	138270	75854	0	0	0	0	0	0	0
4/4/2023 1:00	60	USN9810_Annaba/Whole System:USN9810_Annaba	Reliable	2880	1	1649	1230	227482	92	146950	80440	0	0	0	0	0	0	0

Figure 3.8: Organization of the data inside the csv file

Within the CSV files, there are two primary key columns: “Result Time” and “Object Name”. The “Result Time” column contains the precise date and time at which the values were recorded, while the “Object Name” column includes the name of the Mobility Management Entity (MME) to which the records are associated. Additionally, the “Granularity” column indicates the time interval for which the results were captured, in this case, 60 minutes or hourly records. The “Reliability” column denotes the reliability status of the recorded data. The remaining columns are the file-specific KPI counters. The unit of measurement for each column is

normally represented in the first row of the CSV file, providing context for the recorded values. Following rows provide KPI counter records for a specific date, time, and network component.

## 3.5 Mechanism Implementation Steps

### 3.5.1 Data collection

The initial phase of our work encompasses the data collection process, which precedes the integration of the data into our database. Data collection is the process of collecting the information needed for our analysis. This crucial step consists of several sub-steps, which will be detailed below:

- **Downloading files from mails:** To streamline the retrieval and handling of CSV files, we have developed a Python script that automates the process. Each morning, the script logs into the designated email account and retrieves the corresponding zip file containing the CSV files as an attachment. The zip file is then extracted, and the individual CSV files are transferred to a specific folder for organized storage. As a security measure, the script promptly deletes the zip file and the received email. The flowchart below describes the workflow of the script:

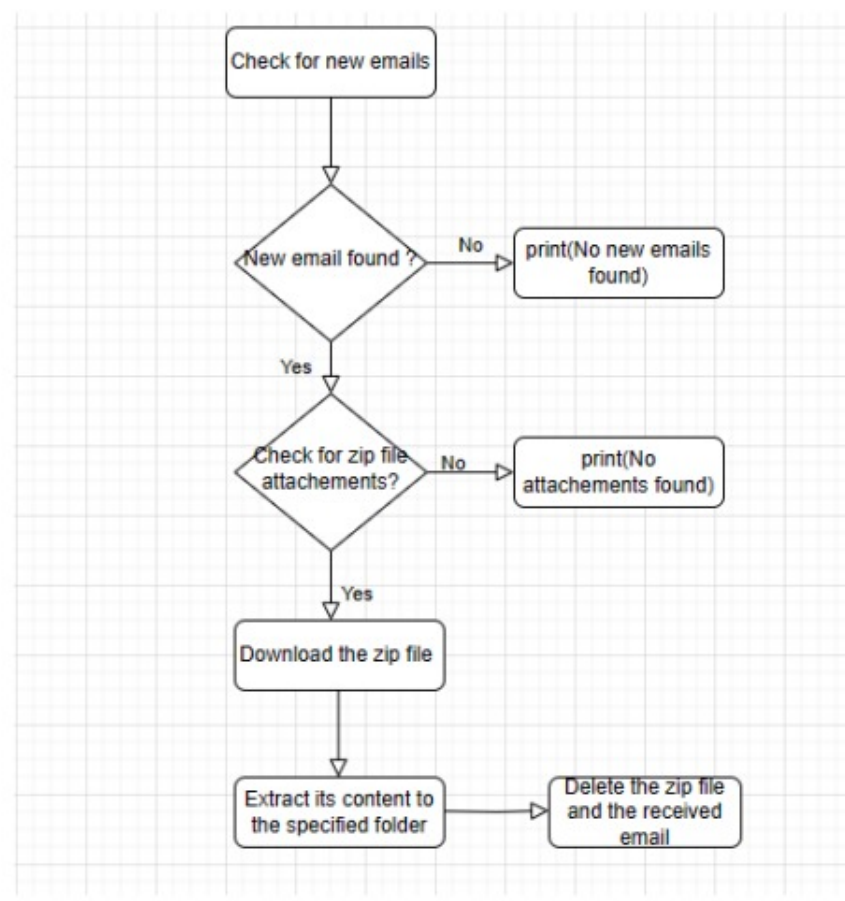


Figure 3.9: Get files from emails flowchart

- **Creation of SQL tables:** To store the data from the CSV files in our SQL database, we need to create individual tables for each table code found in the CSV file names. This allows us to organize the data effectively. It is crucial that each table follows a consistent structure, with identical column names and column order. This ensures a smooth and



error-free data upload process, preventing any issues when transferring the data to their respective tables in the database. By maintaining uniformity in the column structure across all tables, we can seamlessly integrate the CSV data into our database. In order to create the needed tables we should follow these steps:

- First, we need to launch MS SQL Management Studio and connect to the desired SQL server instance.
- Next, in the Object Explorer, we right-click on the target database where we want to create the table. From the context menu, we select "Tasks" and then "Import Data" from the submenu. This opens the SQL Server Import and Export Wizard.
- In the SQL Server Import and Export Wizard, we select the "Data Source" as the CSV file from which we want to create the table. We can specify the file location, format, and delimiter of the CSV file.

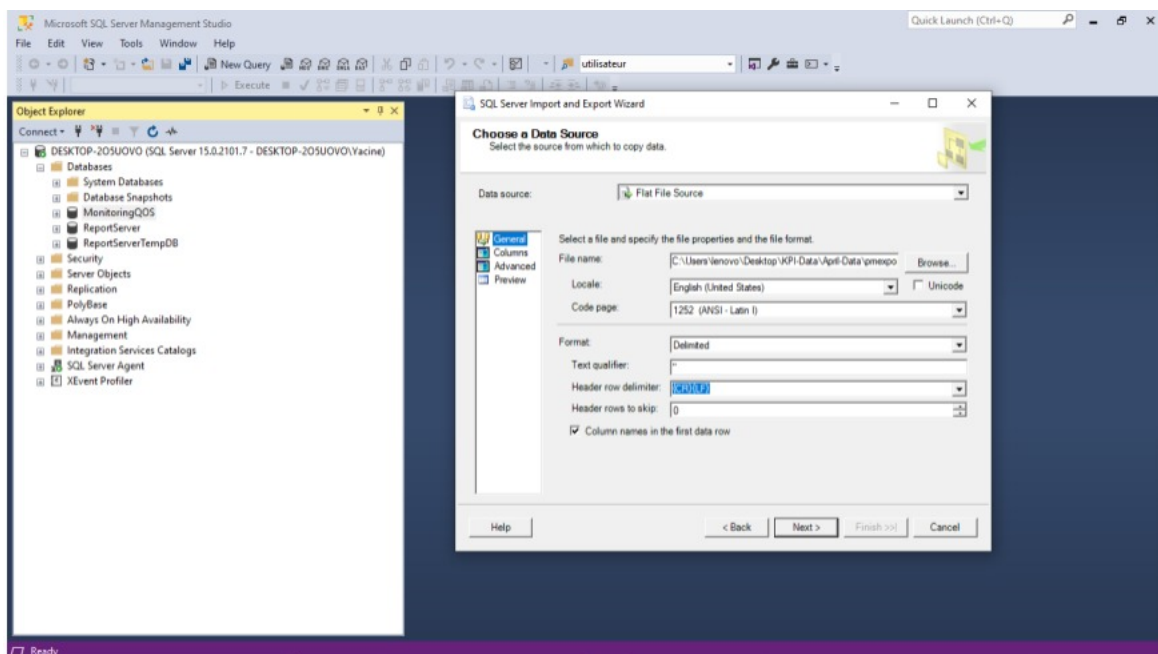


Figure 3.10: Importing table structure into SQL database

- After specifying the CSV file, we proceed to skip all the rows except for the header of the file.

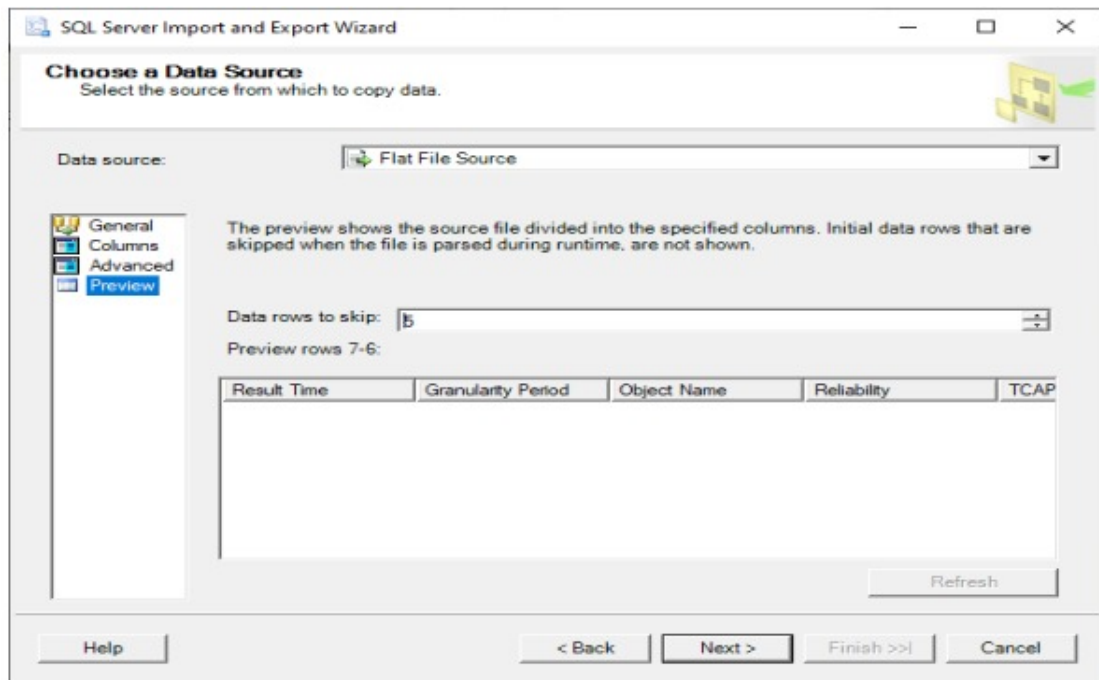


Figure 3.11: Skipping data rows

- We specify the "Destination" as our SQL Server database and choose the appropriate authentication method for the connection, such as Windows Authentication or SQL Server Authentication.  
-Before completing the process, we review the summary of the import settings and click "Finish" to initiate the import process. The wizard then creates the new table in the SQL Server database based on the structure of the CSV file.  
The creation of tables should be done only in the first time when we initialize the system.
- **Checking file columns:** In order to maintain consistency between the received CSV files and the database tables, we have developed a Python script that verifies the files before uploading them into our database. The purpose of this script is to ensure that the table structure of the CSV file matches the corresponding table in the database. The script performs several checks based on the table code. Firstly, it compares the column names of the CSV file with the predefined column names of the corresponding table in the database. If the columns exist but are in a different order, the script reorders them to match the expected structure. Additionally, the script detects any additional columns present in the CSV file that are not found in the database table. In such cases, it raises an error to signal the discrepancy and halts the uploading process. This rigorous verification step guarantees that only valid and consistent data is uploaded to the database. By implementing this Python script, we ensure that the integrity of our database is maintained, preventing any inconsistencies or mismatches between the CSV files and the corresponding tables. The flowchart bellow explain the workflow of the program:

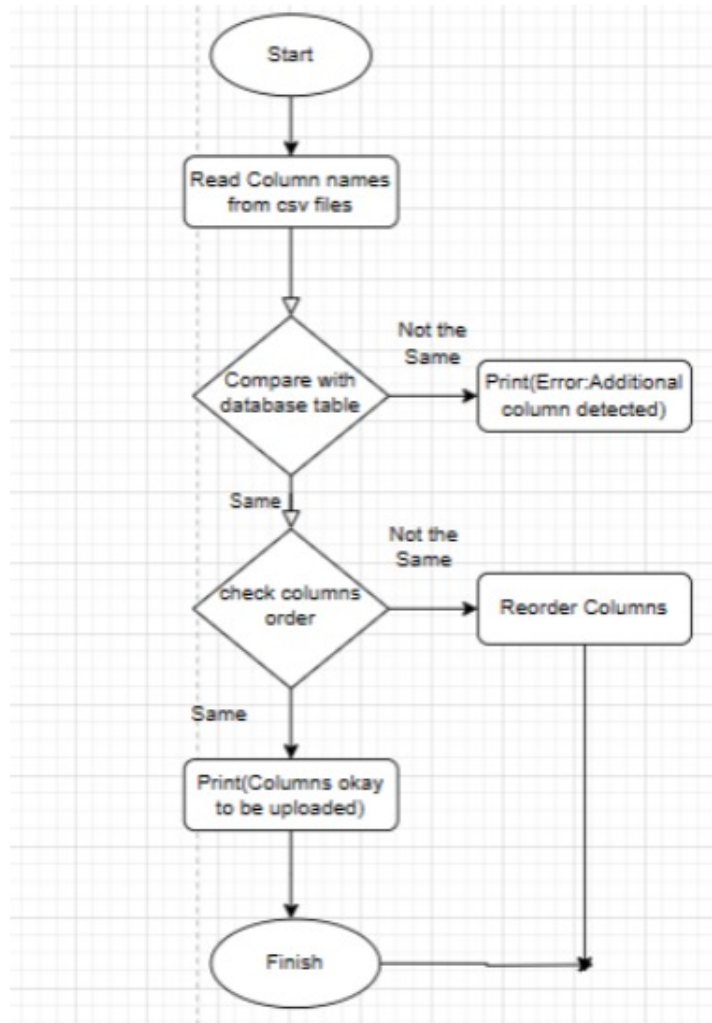


Figure 3.12: Column names check flowchart

- **Uploading files into FTP server:**

Upon finalizing the preparation of our CSV files, the next crucial step is to transfer them to our designated FTP server location on the desktop. This allows seamless accessibility for other users who rely on this data. The FTP server plays a pivotal role in our project by serving as a centralized repository for storing and sharing the CSV files. It ensures efficient and secure data transfer between different stakeholders involved in our project. By utilizing the FTP server, we establish a structured and organized approach to file management, enabling authorized individuals to access the data easily. This centralized location enhances collaboration, promotes data integrity, and simplifies the overall data sharing process. The FTP server serves as a reliable platform that facilitates the timely availability of the CSV files to the relevant project members, supporting smooth workflow and effective data utilization. Using a simple python script we can move the content of the download folder directly to ftp server location.

### 3.5.2 Data Integration

To streamline the content uploading process of our data files into designated tables, we have implemented two distinct approaches utilizing various data integration tools. These approaches employ different integration software solutions.

### 3.5.2.1 Data Integration using Pentaho

Pentaho is a powerful approach to extract, transform and load the data into the desired database. We can use pentaho graphical interface to design the data integration process. This involve defining the data sources, transformation, and the destination. We can utilize a wide range of built-in transformation steps and connectors to manipulate and move the data as required. The workflow deployed in the transformation for uploading the data files is shown in fig 3.13:

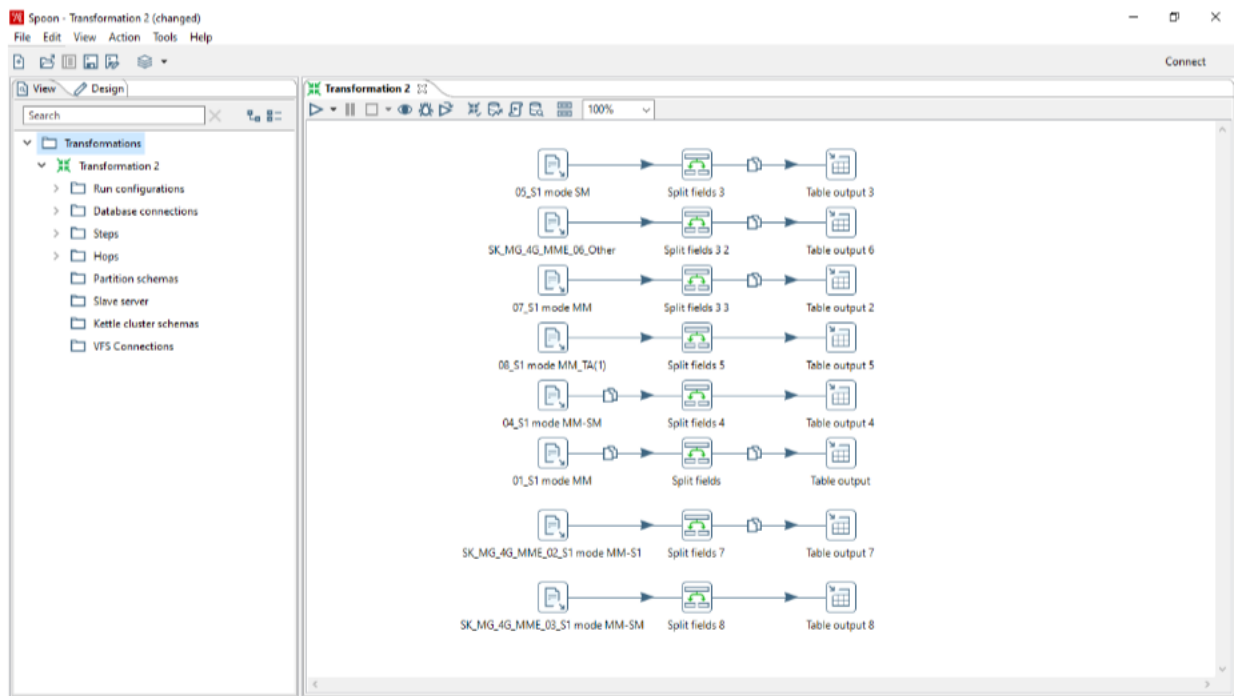


Figure 3.13: Pentaho transformation

We can divide the above transformation into three main parts:

1. **Data extraction:** The Text File input feature in Pentaho allows us to apply regular expressions to filter data files, extract desired information, and retrieve specific data fields. The regular expression engine looks for patterns in input text that match a regular expression. One or more character literals, operators, or structures make up a pattern.



the corresponding columns in the output table. To begin, we connect to our SQL database by navigating to the "Database Connection" tab. In this tab, we will need to provide all the necessary information about our database, as outlined in [3.16](#):

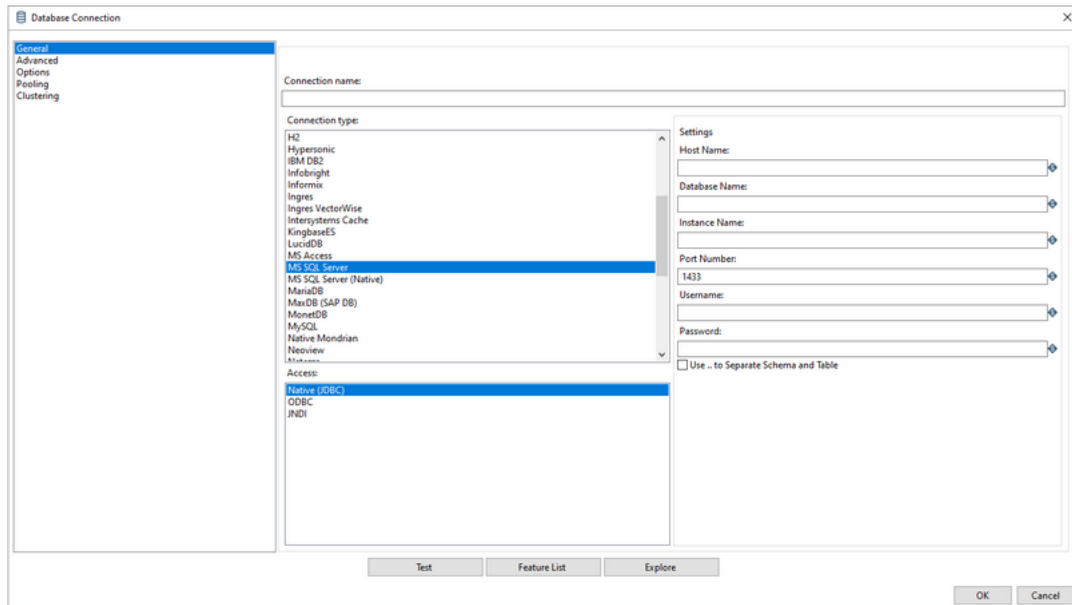


Figure 3.16: Database connection tab

We need to provide a name for this database connection by filling in the "Connection Name" input field. Since we are using an MS SQL Server database, we should select it as the connection type. On the other hand, we need to define the database parameters as follows:

- Host name: We should enter the server or host name where the database is located. This can be an IP address or a domain name.
- Database name: We need to specify the name of the database we want to connect to.
- Instance name: The name of the server hosting the database.
- Port number: For an MS SQL database, the default port number is usually 1433, but it may vary depending on the specific configuration
- Username and Password: We should enter the user credentials required to access the database. These credentials must be valid and have appropriate permissions to perform the required operations.

The JDBC driver allow us to interact with a database using the JDBC API. In order to successfully connect to the database we need to include the appropriate JDBC driver library in the project's class path. After successfully establishing the connection, we should proceed by selecting the desired table in the "Target table" input field. To import the input fields from the previous step, we can use the "Get Fields" button. This action will map the obtained fields to their corresponding counterparts in the database table. Consequently, the content of these fields will be uploaded to the relevant columns in our table. The results obtained after completing the configuration are illustrated in [3.17](#):

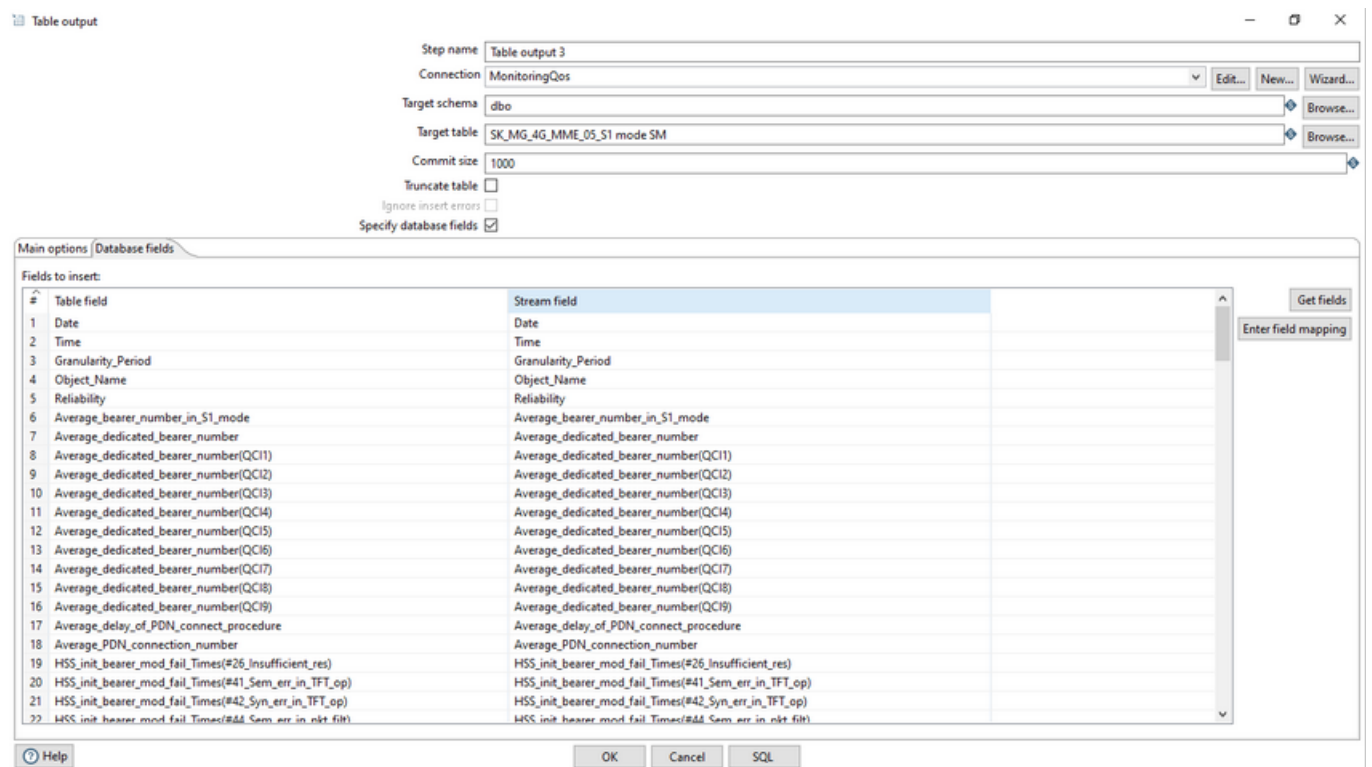


Figure 3.17: Table output configuration

For each row in our transformation, we need to configure this setup to guarantee the accurate mapping of fields into the correct table. By doing so, we ensure that the desired data is uploaded correctly every time, consistently aligning the fields with their corresponding table columns.

### 3.5.2.2 Data integration using knime:

Although there are many other types of data analysis software, Knime is one of the most popular options. Knime is an open-source program that is totally free to use, which makes it easily available to many users who are looking for a better data analysis tool without spending a lot of money. This is one of the key reasons for its popularity. Knime uses a drag and drop interface, therefore using it is also rather simple. Its machine learning techniques also improve it as a tool for data analysis. Knime also includes various algorithms such as logistic regressions, decision trees, random forests, linear regressions, and polynomial regressions. The figure [3.18](#) illustrates the workflow deployed for uploading the data:



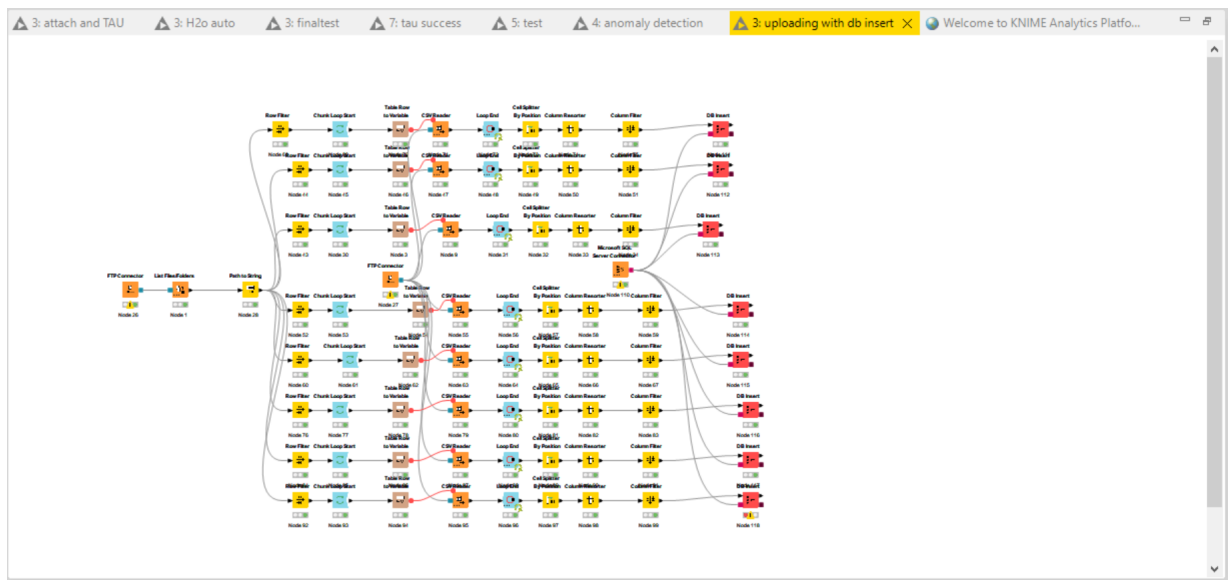


Figure 3.18: Knime workflow

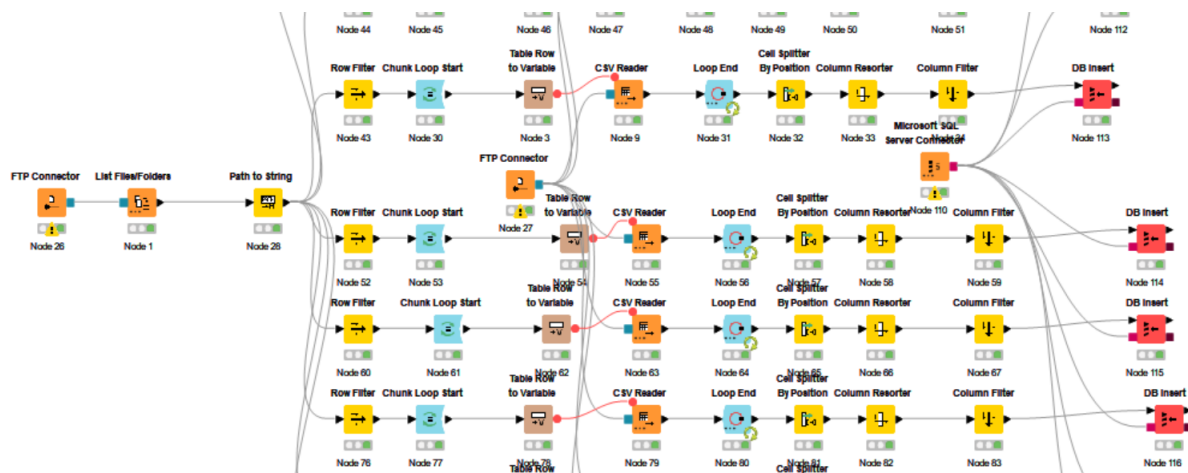


Figure 3.19: A zoom at the workflow

The workflow does the following job:

1. **File collection and handling:** The workflow starts with the FTP connector node, which as the name suggests connects us to our FTP server. The resulting output port allows downstream nodes to access the files of the remote server, e.g. to read or write, or to perform other file system operations (browse/list files, copy, move, ...) [23]. The special thing about this node is that it is a "community node". KNIME Community Extensions offer a wide range of KNIME nodes from different application areas, such as chemo- and bioinformatics, image processing, or information retrieval. In contrast to the extensions available via the standard KNIME Update Site they are provided and maintained by various community developers.

It is then followed by list files/folders node and path to string node. The path to string node is needed to transform the path to the files into a string that is read by the succeeding node, the row filter. The row filter allows for row filtering according to certain criteria. It can include or exclude: certain ranges (by row number), rows with a certain row ID, and rows with a certain value in a selectable column (attribute). We include the rows that match our regular expression only.



- 
2. **File reading:** We first start with a chunk loop start node, where each iteration processes another (consecutive) chunk of rows. We then use a table row to variable node to turn the path into a variable for the CSV reader. The CSV reader is the node responsible for reading the CSV files, it needs to be connected to the table row to variable and also the FTP connector nodes as it needs to know the location of the files. This node is then succeeded by a loop end node.
  3. **Date and Time:** The files that we receive have a datetime column. For better analysis we want to split it into date and time. The cell splitter node splits the content of a selected column into several separate new columns. The data is always split at the same position(s) specified in the node's dialog. This is then followed by column resorter and column filter nodes respectively to sort out the order of the columns and filter the datetime column.
  4. **Uploading into the database tables:** After reading and handling the files, we need to upload them into our database. For that we will need to first establish a connection using Microsoft SQL server connector node. This node creates a connection to a Microsoft SQL Server, an Azure SQL Database or Azure Synapse SQL pool via its JDBC driver. You need to provide the server's hostname (or IP address), the port, and a database name. Login credentials can either be provided directly in the configuration, via credential variables or via the dynamic input port. Then we need to use a DB insert node, which must be connected to both the MS SQL connector and the node containing the files, which in this case is the column filter node. This node data rows into the selected database table based on the values from the selected KNIME input table columns. All selected column names need to exactly match the column names within the database table. The column order of the KNIME table and the database table do not need to match.

### 3.5.3 Data Staging

Data staging is the process of preparing and organizing data before it is used for analysis, processing, or storage. It entails gathering, purifying, transforming data into a uniform and structured manner. Data staging is an important stage in the data pipeline because it guarantees that the data is correct, complete, and ready for further processes. In our case the data staging process is divided into a set of sub processes.

1. **Removing duplicates:** to begin the data staging process, it is crucial to verify that our table does not include any duplicate data rows. Declaring primary key columns becomes essential to prevent the presence of duplicate data within the tables. These key columns serve as unique identifiers for each record, ensuring data integrity by guaranteeing that no two records share the same key value. They provide a means to uniquely identify and differentiate each row in the table. In our specific case, we designate the date and time column, along with the column identifying the network component, as the primary keys of the table. This combination of three factors uniquely identifies each data row. By running a straightforward query for each table, we can declare the primary key columns. The figure [3.20](#) provides a visual representation of the query's functionality.

```
SQLQuery1.sql - D...5UOVO\Yacine (68)) * -> X
ALTER TABLE [dbo].[SK_MG_4G_MME_01_S1 mode MM]
ADD CONSTRAINT [SK_MG_4G_MME_01_S1 mode MM] PRIMARY KEY ([Date], [Time], [Object_Name])
```

Figure 3.20: Query for managing the primary key columns.

- We use the ALTER TABLE statement to modify the table structure.
  - We add the primary key constraint using the ADD CONSTRAINT clause.
  - We need to specify the primary key columns within parentheses after the PRIMARY KEY keyword.
2. **Uploading data to the new tables:** After successfully uploading the data into our tables, we should proceed with deploying an additional transformation to upload the data into the attach table and the tracking area tables. This workflow [3.21](#) specifically selects the required columns from each table and fills our new tables accordingly.

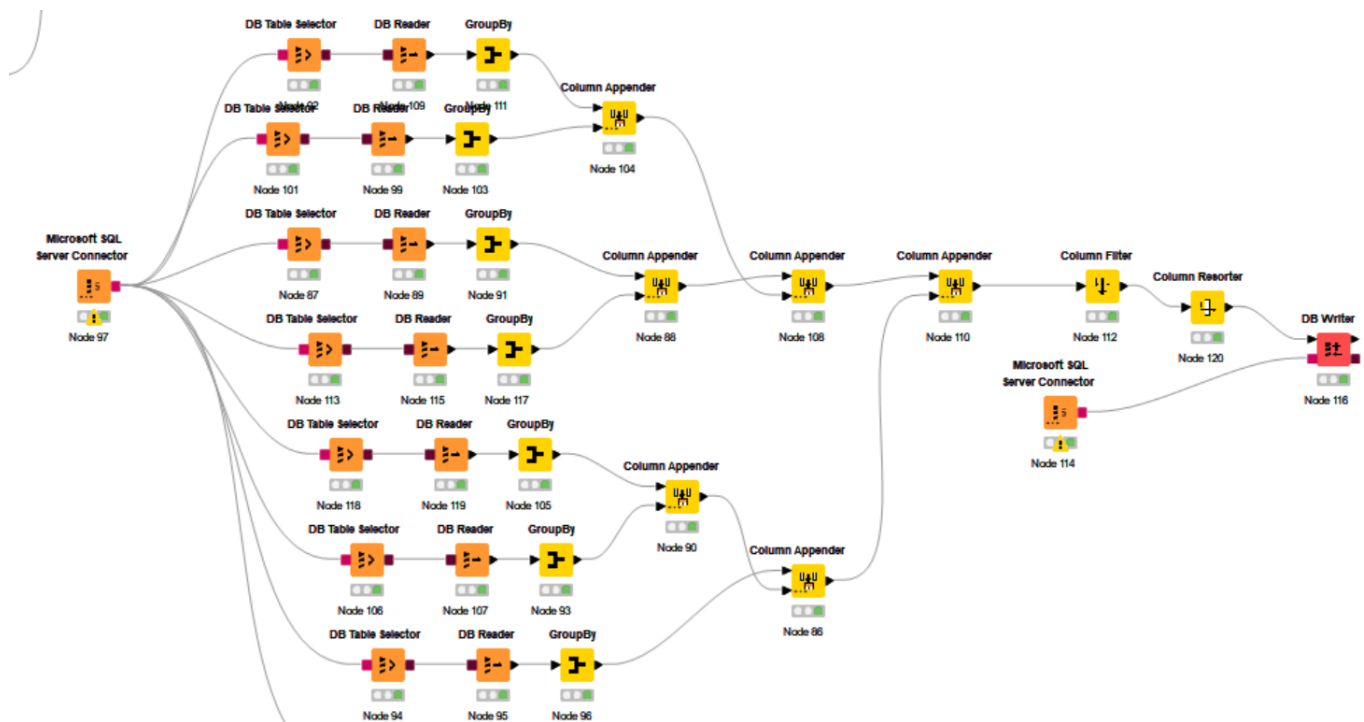


Figure 3.21: Attach and TAU table creation

The workflow starts by connecting to the database, then selecting our tables using the DB table selector node. The contents of each table are then read by the DB reader node. After that, GroupBY node Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns. In our case we group columns based on the procedure. Then we use a column appender node to join the columns of each table into one table containing all the filtered columns.

The result is then connected to a DB writer node, this node writes into an already selected table in the database or creates a new one if it does not exist.

### 3. Creation of technical success rate columns:

The data within the generated tables provides valuable information regarding the attach and tracking area update procedures. The success rate of these procedures is of utmost significance as it denotes the percentage or proportion of processes that have been accomplished successfully. It indicates the effectiveness and reliability of the procedure in successfully establishing the desired connections or associations between entities or components. The success rate is calculated by dividing the number of successful procedures by the total number of attempted procedures and then multiplying the result by 100 to express it as a percentage. Failures have a significant impact on the success rate, as they decrease its value. Each failure reduces the numerator (number of successful attach procedures) and potentially increases the denominator (total number of attempted attach procedures), resulting in a lower success rate. As per the specifications of **3GPP TS 24.301**, failures in these procedures can be attributed to either user causes or network causes. To determine the technical success rate of our network, it is necessary to recalculate the success rate by excluding failures caused by users and considering only failures caused by the network. The following formulas have been deployed to calculate the new success rates:

- **Non combine attach success rate:** This metric indicates the success rate of the attach procedure to the CS network. It is calculated using the following formula:

$$NCASr = \frac{AST + NUSNCE}{ART} \times 100\% \quad (3.1)$$

- **Combined attach success rate:** This metric represents the success rate of the attach procedure to both the LTE network and the CS network. It is calculated using the following formula:

$$CASr = \frac{CASTEpsservices + CASTEpsservices + NUSNCE}{CART} \times 100\% \quad (3.2)$$

- **Combined intra TAU success rate:** This metric reflects the rate at which TAU operations, involving both SGW change and SGW not change scenarios, are successfully completed when the user is attached to both the CS and PS networks. It is calculated using the following formula:

$$CITAU_{sr} = \frac{ITAUST + ICTAUST + PTAUST + NUSNCE}{ITAURT + CTAURT + PTAURT} \times 100\% \quad (3.3)$$

- **Non combined intra TAU success rate:** This metric represents the rate of successful TAU over the TAU request times when the user is attached to the CS network only. It is calculated using the following formula:

$$NCTAU_{sr} = \frac{ITAUST + PTAUST + NUSNCE}{PTAURT + ITAURT} \times 100\% \quad (3.4)$$

- **Combined Inter TAU success rate:** This metric indicates the rate at which TAU operations, involving both SGW change and SGW not change scenarios, are successfully completed when the user is attached to both the CS and PS networks, but the user is connected to a different MME. It is calculated using the following formula:

$$CiTAU_{sr} = \frac{iTAUST + iCTAUST + NUSNCE}{iTAURT + iCTAURT} \times 100\% \quad (3.5)$$

- **Non combine Inter TAU success rate:** This metric refers to the rate of successful TAU over the TAU request times when the user is attached to the CS network only, but the user will be connected to a different MME when changing location. It is calculated using the following formula:

$$NCiTAU_{sr} = \frac{iTAUST + NUSNCE}{iTAURT} \times 100\% \quad (3.6)$$

Determining the USN causes and Non-USN causes to be excluded proved to be quite challenging. However, after careful examination of more than forty different causes and discussion, we were able to find the KPIs that must be included in the formulas. Some of these KPIs description from 3GPP [26] will be listed below:

#### 1. Non-USN causes:

- #3 Illegal UE: This EMM cause is sent to the UE when the network refuses service to the UE either because an identity of the UE is not acceptable to the network or because the UE does not pass the authentication check, i.e. the RES received from the UE is different from that generated by the network.
- #8 EPS services and non-EPS services not allowed: This EMM cause is sent to the UE when it is not allowed to operate either EPS or non-EPS services.
- #12 Tracking area not allowed: This EMM cause is sent to the UE if it requests tracking area updating in a tracking area where the HPLMN determines that the UE, by subscription, is not allowed to operate.
- #15 No suitable cells in tracking area: This EMM cause is sent to the UE if it requests tracking area updating in a tracking area where the UE, by subscription, is not allowed to operate, but when it should find another allowed tracking area in the same PLMN.

#### 2. USN-causes

- #17 Network failure: This EMM cause is sent to the UE if the MME cannot service an UE generated request because of PLMN failures.
- #19 ESM failure : This EMM cause is sent to the UE when there is a failure in the ESM message contained in the EMM message.
- #111 Protocol error, unspecified: This ESM cause is used to report a protocol error event only when no other ESM cause in the protocol error class applies.

### 3.5.4 Data visualization

To make data easier for the human brain to grasp and draw conclusions from, data visualization is the practice of putting information into a visual context, like a map or graph. Data visualization's major objective is to make it simpler to spot patterns, trends, and outliers in big data sets. For our work we decided to use Microsoft's Power BI:

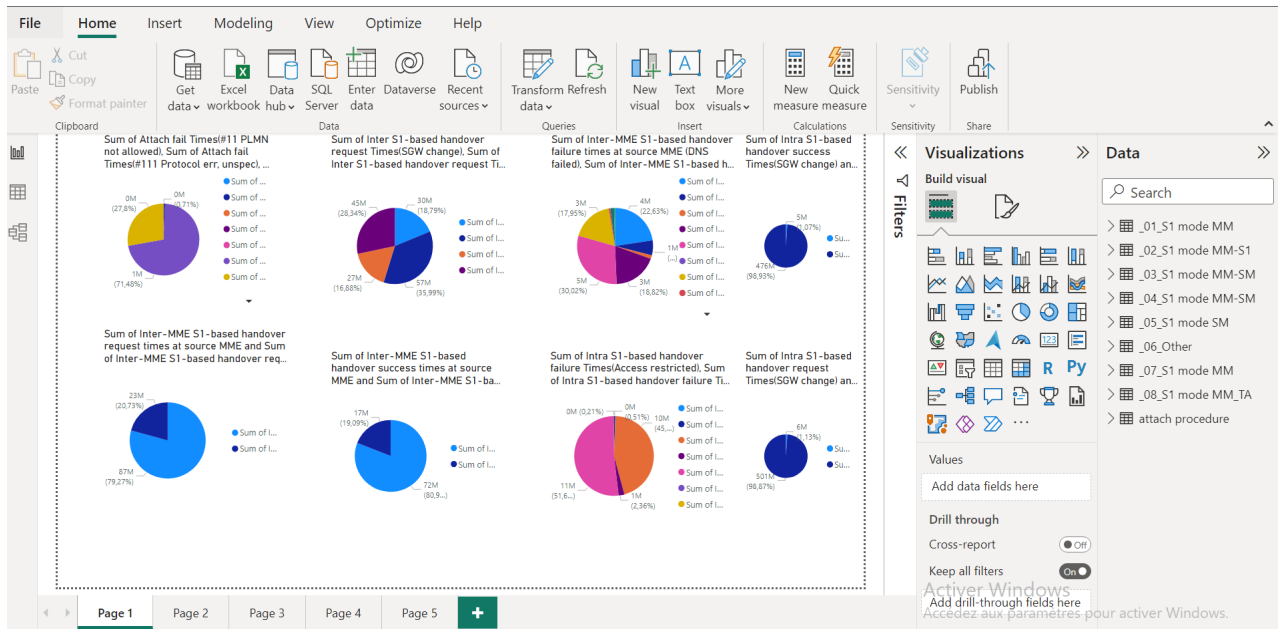


Figure 3.22: Power BI front page

Power BI is able to connect to and get data from many different databases. In our case, we click on the SQL server and enter the server and database names. As soon as a connection is established we get the tables displayed on the right. We can display our data using different diagrams and curves. The figure 3.22 shows some data sums displayed as a pie chart, this type of display allows for a better understanding of the data in terms of percentage. In this part, we will focus on the Non-combine attach success rate and Intra TAU combined and non combined success rate of one MME which we have chosen to be Dely Brahim's MME for a clearer view and better understanding. First let's start with the attach procedure visualization:

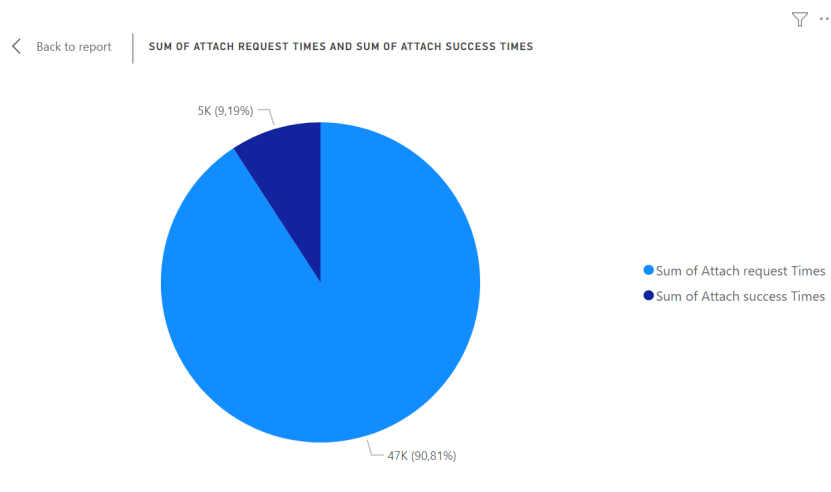


Figure 3.23: Pie chart representing attach request and success times

We used a segment to choose USN-DELY and set the date to 7th April. We can see that the 'pure' success times is only 11%, however the technical success rate that is calculated using the formula shown in the previous section is much higher than that. That suggests that most of the failures are due to non-USN causes. The figure 3.24 displays the different failure causes as a pie chart to see which failure happens the most:

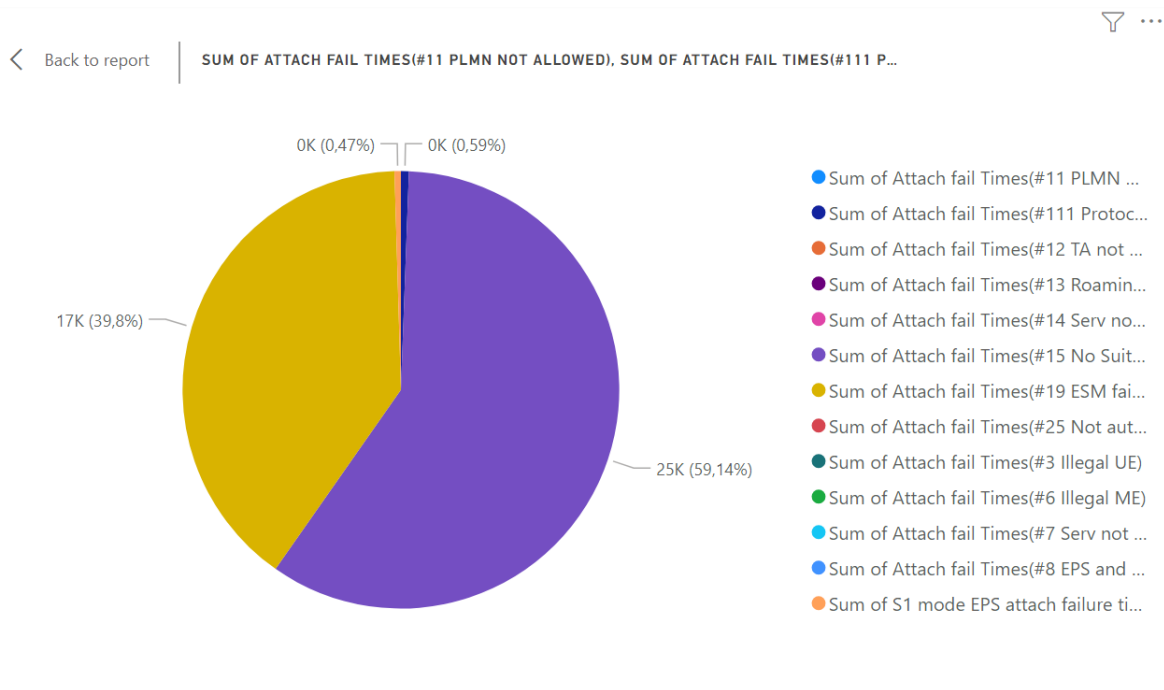


Figure 3.24: Pie chart representing attach failure times

We can see that 59.4% of the failures are due to [15 no suitable cells in tracking area] which is a non-USN cause. The other failures are due to ESM failure, network failure and unspecified protocol error which are USN causes. Next, we will visualize these three different causes using a line chart for a better understanding :

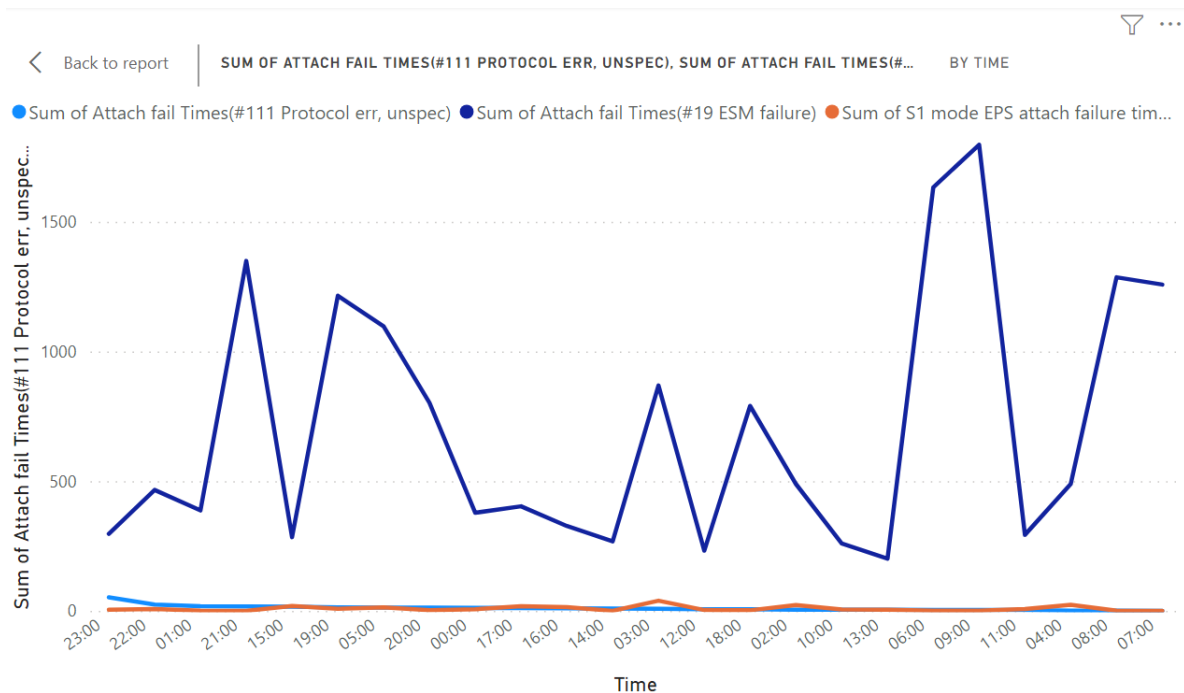


Figure 3.25: Line chart representing attach failure times

As shown by the pie chart previously, the failures due to ESM failure causes are much higher than the other two causes. ESM failure has a range from 300 to 1800, while unspecified protocol errors and network failure have ranges of 0 to 60 and 0 to 40 respectively. Now let us look at the TAU success rate on 10-03-2023:

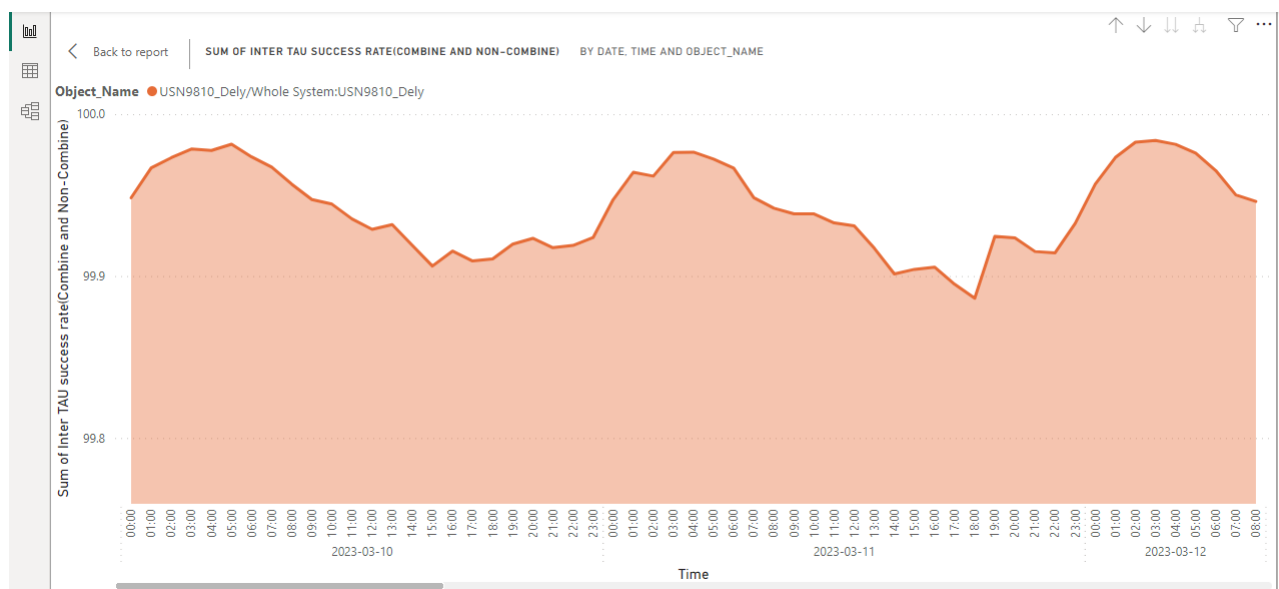


Figure 3.26: Line chart representing Intra TAU success rate

We can notice that unlike the attach procedure, the success here is very high and close to 100%. To understand why we will see the number of failures due to USN causes compared to the number of requests:

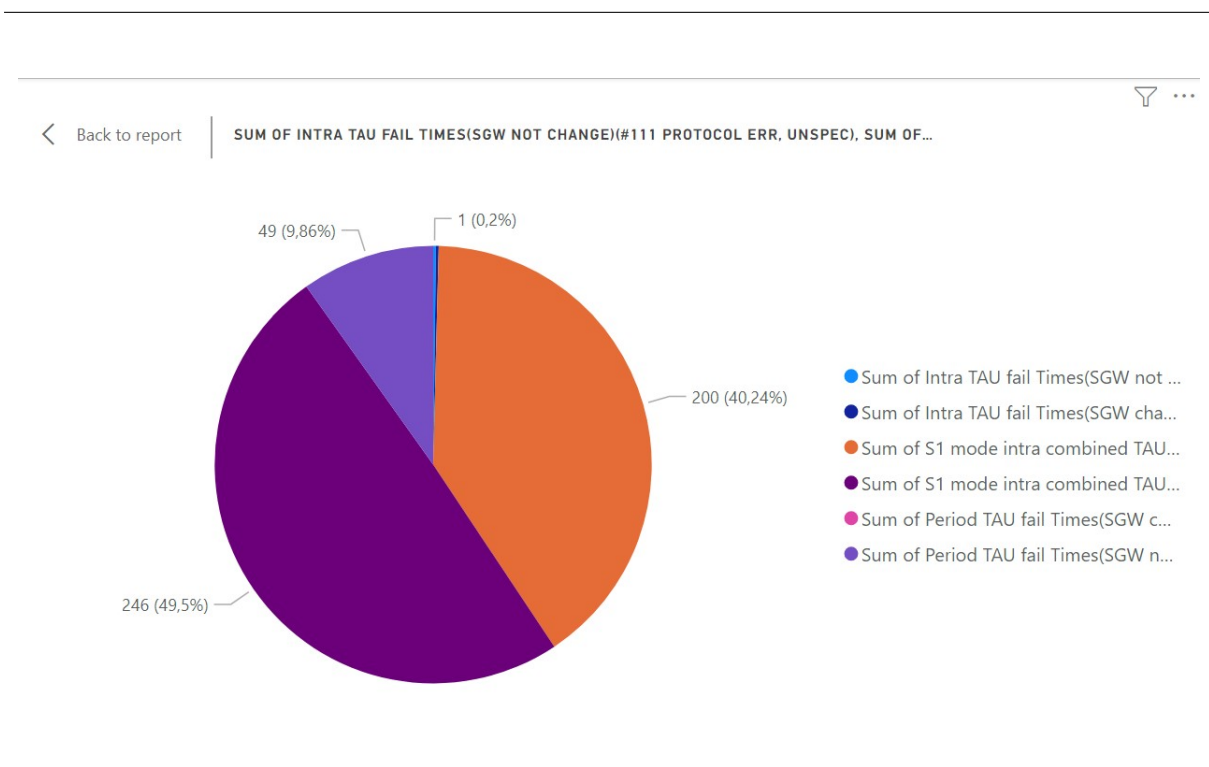


Figure 3.27: Pie chart representing Intra TAU failure times due to USN causes

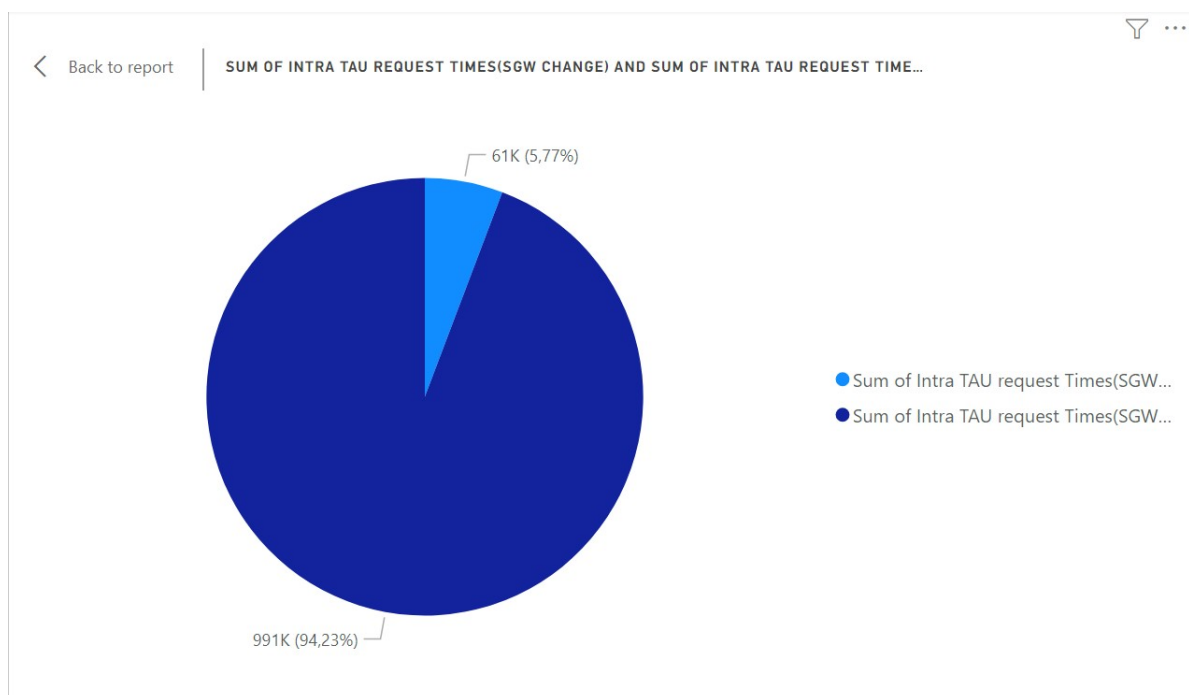


Figure 3.28: Pie chart representing Intra TAU request times

The number of failures due to USN causes is very low compared to the number of request times. That explains the very high success rate. From the previous attach success rate line chart, we might see some unusual behaviour and irregular patterns. However, we need to analyze it first to decide if it is an anomaly or not. The following part will be devoted to the analysis of these unusual data values.



## 3.6 Data Analysis for Anomaly Detection

Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. It can be used to identifying unusual or abnormal patterns, events, or observations within a dataset by detecting deviation from expected behavior, which may indicate potential fraud, errors, security breaches, or other anomalies. It involves various techniques and methods to extract meaningful insights from data, which can be in various formats such as numerical, textual, or visual. The process of anomaly detection is divided into the following sub processes.

### 3.6.1 Non coded anomaly detection:

In this part we will try to detect anomalies without any Python code, using only Knime analytics platform.

#### 3.6.1.1 Clustering method:

The first method that we will try is the clustering method. Clustering is one of the more popular data mining techniques. The process of clustering involves grouping the entire dataset into distinct clusters based on the instances' shared properties. Grouping or clustering the instances of a big database based on similarities between the data instances, regardless of the size of the database, is regarded as an important portion of data mining. [27]. There are plentiful clustering techniques but we will use the K-means algorithm. The figure 3.29 below shows the workflow used for this method:

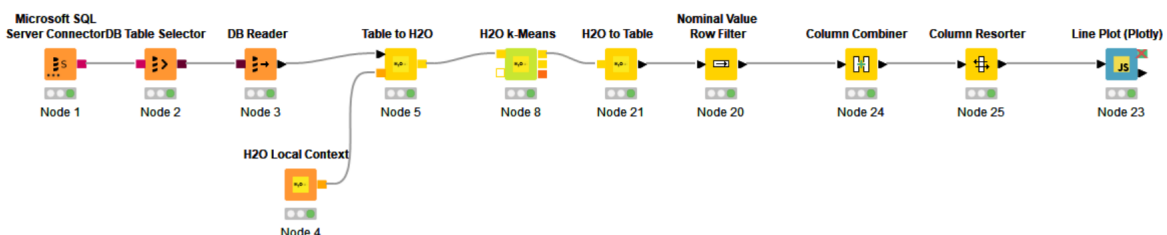


Figure 3.29: K-means workflow

The H2O nodes are part of Knime's H2O extension. H2O is a fully open source, distributed in-memory machine learning platform with linear scalability. H2O supports the most widely used statistical and machine learning algorithms including gradient boosted machines, generalized linear models, deep learning and more [28]. The H2O k-means node applies a k-means model using H2O, where you can select the number of clusters and iterations number, or you can just let the algorithm estimate. It is also possible to give user-specified initial cluster centers, but we preferred to avoid that. You can also choose what columns to apply the clustering to, we have chosen are newly calculated technical non combine attach success rate. The clustering centers obtained are shown in figure 3.30:

H2O Frame - Rows: 3 Flow Variables					
Row ID	S Date	S Time	S Object Name	D technic...	S Cluster
Row0	2023-03-15	15:00	USN9810_Blida/Whole System:USN9810_Blida	83.067	cluster_0
Row1	2023-04-07	19:00	USN9810_Blida/Whole System:USN9810_Blida	54.787	cluster_1
Row2	2023-03-18	03:00	USN9810_Dely/Whole System:USN9810_Dely	74.587	cluster_2

Figure 3.30: K-means centers

Since we are trying to detect anomalies, we need to focus about the low values around cluster 1. So the succeeding node is a nominal value row filter which keeps only the rows that belong to cluster 1. The obtained results are then plotted by the Line plot node which is based on Plotly.js library. The results are shown in the figure 3.31:

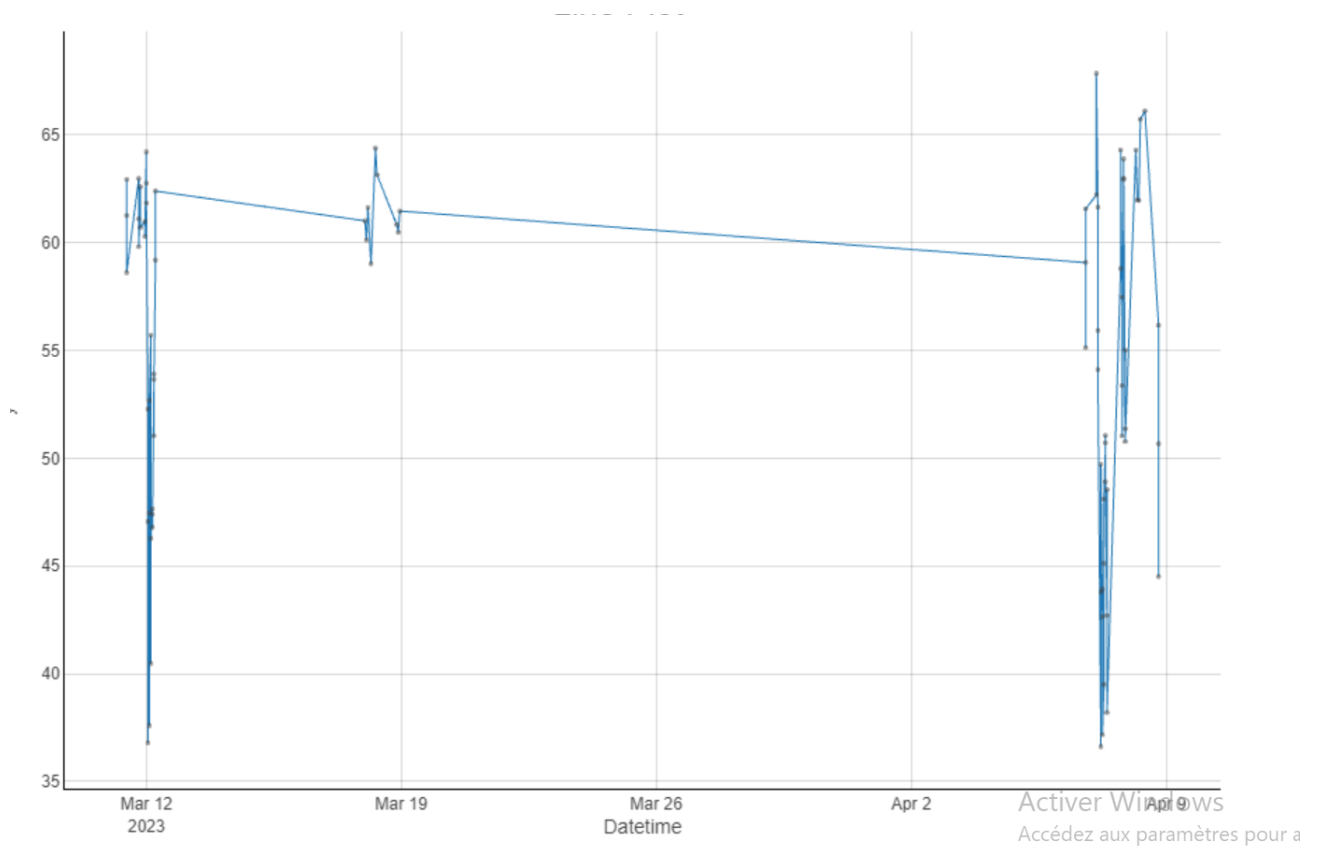


Figure 3.31: K-means line plot

We can see that while most of the anomalies detected in the previous part are detected here, there are some additional "normal" values and that is because k-means algorithm uses the nearest mean as a prototype of the cluster. In hope of a more accurate detection of anomalies, we are going to try another method.

### 3.6.1.2 Outlier detection method:

This method is based on filtering out the values lying outside of the upper and lower bounds of our interval, that are calculated using the interquartile range IQR. To calculate the IQR the 1st and 3rd quartile are calculated, the first quartile Q1 is the value under which 25% of data points are found when they are arranged in increasing order. is the value under which 75% of

data points are found when arranged in increasing order[29]. The inter quartile range is then calculated as  $IQR = Q3 - Q1$ . An observation point is said to be an outlier if it lies outside the range  $R = [Q1 - (k \times IQR), Q3 + (k \times IQR)]$ . In Knime we use the numeric outliers node combined with numeric outliers apply as shown in figure 3.32:

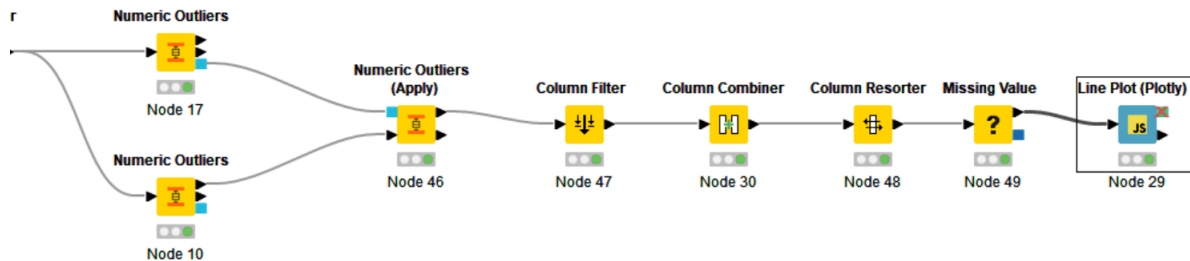


Figure 3.32: Numeric outliers workflow

The reason for using two numeric outliers nodes separately is that If an observation is flagged an outlier, one can either replace it by some other value or remove/retain the corresponding row. In the first one we select the USN-cause attach failure columns and choose to replace outliers above the upper bound by a missing value. In the second node we choose to remove non outlier rows and keep only outliers below the lower bound as we are focusing on the technical non combine attach success rate. The output of these two nodes are then joined in one table to try and find the attach failure causes anomalies in the technical success rate anomalies.

Table "default" - Rows: 42 Spec - Columns: 7 Properties Flow Variables									
Row ID	S Date	S Time	S Object Name	I Attach ...	I Attach ...	I S1 mod...	D technic...		
Row 147	2023-03-12	01:00	USN9810_Annaba/Whole System:USN9810_Annaba	3	?	40	52.281		
Row 148	2023-03-12	01:00	USN9810_Blida/Whole System:USN9810_Blida	4	?	68	36.801		
Row 149	2023-03-12	01:00	USN9810_Dely/Whole System:USN9810_Dely	15	?	73	47.051		
Row 150	2023-03-12	02:00	USN9810_Annaba/Whole System:USN9810_Annaba	4	?	38	47.448		
Row 151	2023-03-12	02:00	USN9810_Blida/Whole System:USN9810_Blida	7	?	91	37.594		
Row 152	2023-03-12	02:00	USN9810_Dely/Whole System:USN9810_Dely	12	?	84	52.683		
Row 153	2023-03-12	03:00	USN9810_Annaba/Whole System:USN9810_Annaba	?	?	96	55.692		
Row 154	2023-03-12	03:00	USN9810_Blida/Whole System:USN9810_Blida	12	?	92	40.469		
Row 155	2023-03-12	03:00	USN9810_Dely/Whole System:USN9810_Dely	8	?	65	46.267		
Row 156	2023-03-12	04:00	USN9810_Annaba/Whole System:USN9810_Annaba	15	?	107	47.655		
Row 157	2023-03-12	04:00	USN9810_Blida/Whole System:USN9810_Blida	7	?	92	47.407		
Row 158	2023-03-12	04:00	USN9810_Dely/Whole System:USN9810_Dely	6	?	54	46.786		
Row 159	2023-03-12	05:00	USN9810_Annaba/Whole System:USN9810_Annaba	11	?	97	53.908		
Row 160	2023-03-12	05:00	USN9810_Blida/Whole System:USN9810_Blida	13	?	119	51.033		
Row 161	2023-03-12	05:00	USN9810_Dely/Whole System:USN9810_Dely	8	?	56	53.659		
Row 850	2023-04-06	19:00	USN9810_Blida/Whole System:USN9810_Blida	6	?	0	55.121		
Row 874	2023-04-07	03:00	USN9810_Blida/Whole System:USN9810_Blida	0	?	717	23	55.931	
Row 875	2023-04-07	03:00	USN9810_Dely/Whole System:USN9810_Dely	8	?	39	54.114		
Row 879	2023-04-07	05:00	USN9810_Annaba/Whole System:USN9810_Annaba	3	?	19	36.621		
Row 880	2023-04-07	05:00	USN9810_Blida/Whole System:USN9810_Blida	3	?	16	43.787		
Row 881	2023-04-07	05:00	USN9810_Dely/Whole System:USN9810_Dely	12	?	13	49.704		
Row 882	2023-04-07	06:00	USN9810_Annaba/Whole System:USN9810_Annaba	0	?	4	42.636		
Row 883	2023-04-07	06:00	USN9810_Blida/Whole System:USN9810_Blida	0	?	3	43.923		
Row 884	2023-04-07	06:00	USN9810_Dely/Whole System:USN9810_Dely	3	?	1	37.164		
Row 885	2023-04-07	07:00	USN9810_Annaba/Whole System:USN9810_Annaba	1	?	0	39.497		
Row 886	2023-04-07	07:00	USN9810_Blida/Whole System:USN9810_Blida	1	?	0	45.126		
Row 887	2023-04-07	07:00	USN9810_Dely/Whole System:USN9810_Dely	0	?	0	48.102		
Row 888	2023-04-07	08:00	USN9810_Annaba/Whole System:USN9810_Annaba	1	?	6	50.713		
Row 889	2023-04-07	08:00	USN9810_Blida/Whole System:USN9810_Blida	2	?	0	51.056		
Row 890	2023-04-07	08:00	USN9810_Dely/Whole System:USN9810_Dely	1	?	1	48.908		
Row 891	2023-04-07	09:00	USN9810_Annaba/Whole System:USN9810_Annaba	0	?	1	42.698		
Row 892	2023-04-07	09:00	USN9810_Blida/Whole System:USN9810_Blida	5	?	3	48.537		
Row 893	2023-04-07	09:00	USN9810_Dely/Whole System:USN9810_Dely	3	?	1	38.192		
Row 921	2023-04-07	19:00	USN9810_Annaba/Whole System:USN9810_Annaba	16	?	5	57.461		
Row 922	2023-04-07	19:00	USN9810_Blida/Whole System:USN9810_Blida	10	?	3	53.363		
Row 923	2023-04-07	19:00	USN9810_Dely/Whole System:USN9810_Dely	13	?	8	51.028		
Row 927	2023-04-07	21:00	USN9810_Annaba/Whole System:USN9810_Annaba	14	?	0	51.362		

Figure 3.33: Numeric outliers output table

We can see that most of the values of the ESM failure column are missing, so an assumption can be made that the technical success rate anomalies happened due to a spike in attach failures due to ESM failure. To get a better understanding through a visualization we used a missing value handling node and replaced missing values by the value 1000. The obtained plot is shown in 3.34:

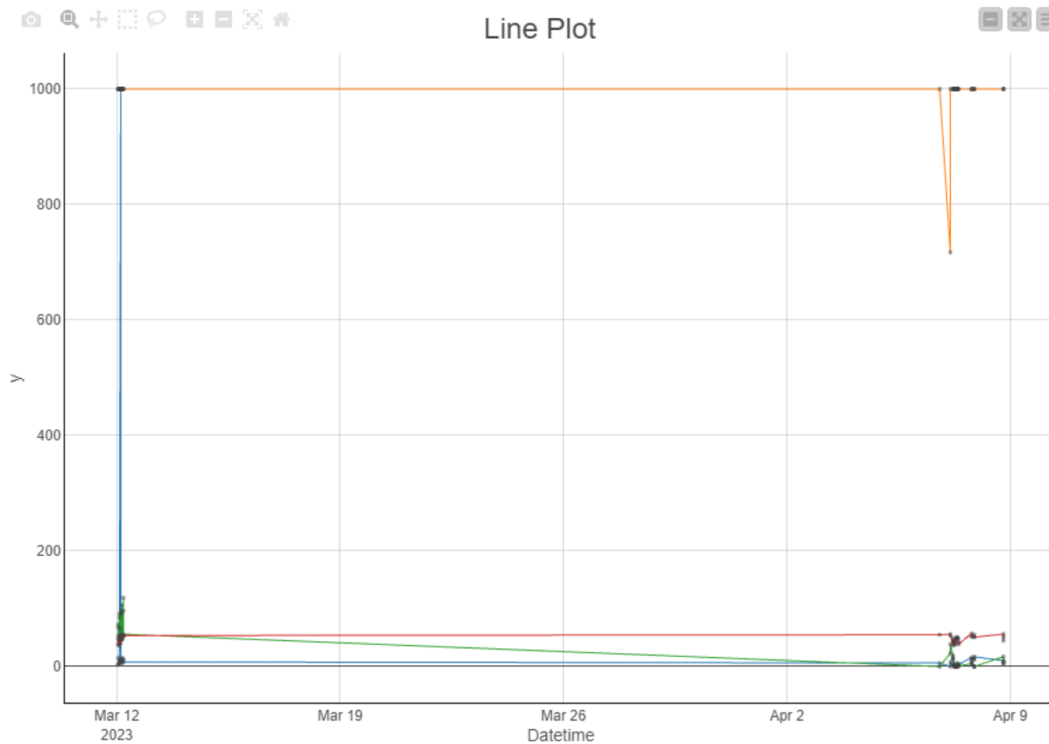


Figure 3.34: Numeric outliers output line plot

Outliers detection method does a better job at detecting anomalies in our data than the clustering method. This is due to the type of data that we have. Next we will try to see if we can predict the values and anomalies in our data.

### 3.6.1.3 Predictive analysis:

The use of statistics and modeling approaches to forecast future results and performance is known as predictive analytics. With predictive analytics, data patterns in the past and present are examined to see if they are likely to recur. Before choosing what model to use, we will first execute an AutoML learner node to decide which model will yield the better results.

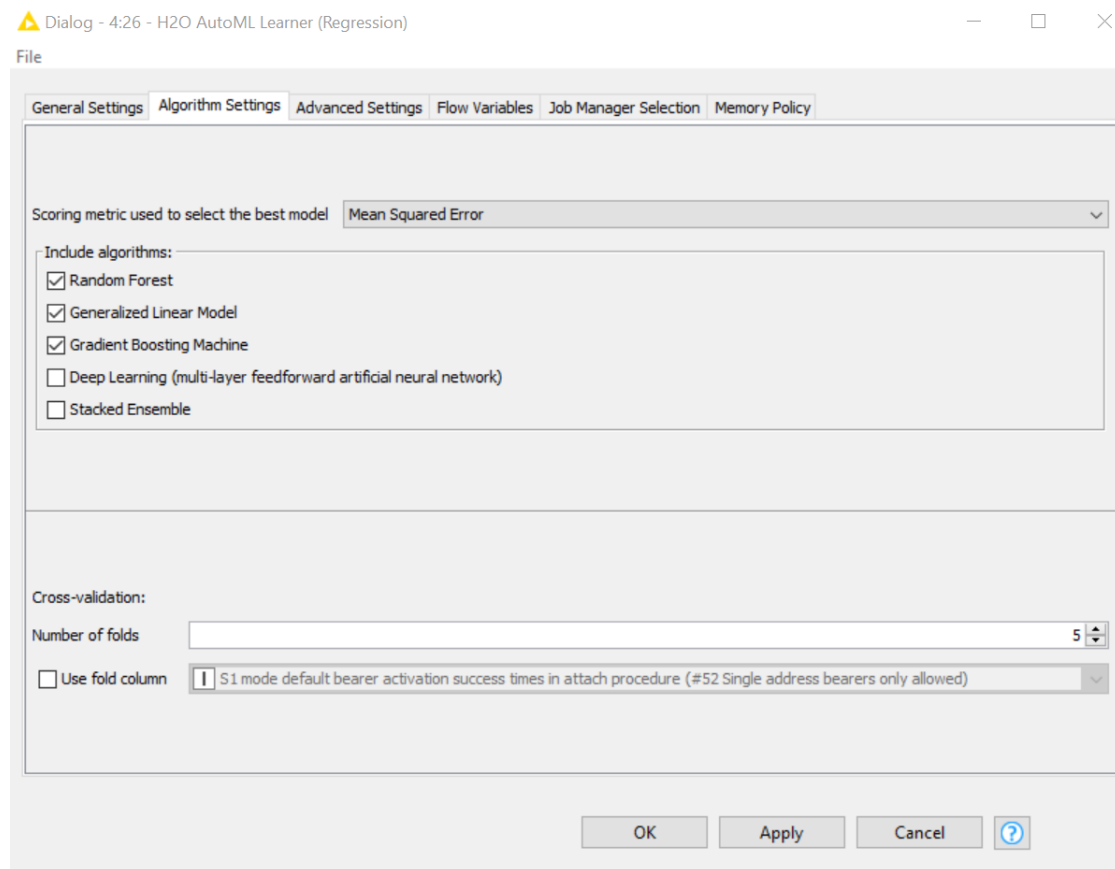


Figure 3.35: AutoML learner node

AutoML learner node Learns the specified types of models using H2O AutoML and returns the leading model amongst these. As part of the learning process, hyperparameters are automatically optimized by H2O using a random grid search. The scoring meter selected to determine the best model was chosen to be the mean squared error.

$$MSE = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (3.7)$$

Where :

N = number of data points.

xi = observed values.

yi = predicted values.

The best model selected is shown in figure 3.36 :

File	
Gradient Boosting Machine	Flow Variables
Model Summary	
Number of Trees	98
Number of Internal Trees	98
Model Size in Bytes	75185
Min. Depth	0
Max. Depth	10
Mean Depth	9.020409
Min. Leaves	1
Max. Leaves	63
Mean Leaves	56.489796

Figure 3.36: Best model

Gradient booster machine was selected as the best model. The table shown in 3.37 is a leaderboard of models trained in the AutoML process. The models are ranked by the selected scoring metric:

File Edit Hilite Navigation View						
Table "default" - Rows: 1968 Spec - Columns: 6 Properties Flow Variables						
Row ID	S model_id	D mean_r...	D mse	D rmse	D mae	D rmsle
0	GBM_grid_1_AutoML_1_20230609_122708_model_1895	9.811	9.811	3.132	2.222	0.045
1	GBM_grid_1_AutoML_1_20230609_122708_model_682	9.982	9.982	3.159	2.243	0.045
2	GBM_grid_1_AutoML_1_20230609_122708_model_50	10.009	10.009	3.164	2.245	0.046
3	GBM_grid_1_AutoML_1_20230609_122708_model_1826	10.029	10.029	3.167	2.253	0.046
4	GBM_grid_1_AutoML_1_20230609_122708_model_417	10.037	10.037	3.168	2.233	0.046
5	GBM_grid_1_AutoML_1_20230609_122708_model_1028	10.04	10.04	3.169	2.242	0.046
6	GBM_grid_1_AutoML_1_20230609_122708_model_228	10.08	10.08	3.175	2.27	0.046
7	GBM_grid_1_AutoML_1_20230609_122708_model_1438	10.091	10.091	3.177	2.246	0.046
8	GBM_grid_1_AutoML_1_20230609_122708_model_1144	10.118	10.118	3.181	2.278	0.046
9	GBM_grid_1_AutoML_1_20230609_122708_model_1424	10.12	10.12	3.181	2.254	0.046
10	GBM_grid_1_AutoML_1_20230609_122708_model_985	10.128	10.128	3.182	2.267	0.046
11	GBM_grid_1_AutoML_1_20230609_122708_model_1827	10.143	10.143	3.185	2.272	0.046
12	GBM_grid_1_AutoML_1_20230609_122708_model_1809	10.146	10.146	3.185	2.302	0.046
13	GBM_grid_1_AutoML_1_20230609_122708_model_1917	10.153	10.153	3.186	2.231	0.046
14	GBM_grid_1_AutoML_1_20230609_122708_model_539	10.153	10.153	3.186	2.231	0.046
15	GBM_grid_1_AutoML_1_20230609_122708_model_407	10.153	10.153	3.186	2.231	0.046
16	GBM_grid_1_AutoML_1_20230609_122708_model_722	10.162	10.162	3.188	2.306	0.046
17	GBM_grid_1_AutoML_1_20230609_122708_model_1295	10.162	10.162	3.188	2.267	0.046
18	GBM_grid_1_AutoML_1_20230609_122708_model_1937	10.163	10.163	3.188	2.238	0.046
19	GBM_grid_1_AutoML_1_20230609_122708_model_255	10.167	10.167	3.189	2.262	0.046
20	GBM_grid_1_AutoML_1_20230609_122708_model_784	10.171	10.171	3.189	2.252	0.046
21	GBM_grid_1_AutoML_1_20230609_122708_model_716	10.192	10.192	3.192	2.26	0.046
22	GBM_grid_1_AutoML_1_20230609_122708_model_895	10.2	10.2	3.194	2.245	0.046
23	GBM_grid_1_AutoML_1_20230609_122708_model_1752	10.212	10.212	3.196	2.228	0.046
24	GBM_grid_1_AutoML_1_20230609_122708_model_822	10.216	10.216	3.196	2.241	0.046
25	GBM_grid_1_AutoML_1_20230609_122708_model_233	10.216	10.216	3.196	2.284	0.046
26	GBM_grid_1_AutoML_1_20230609_122708_model_938	10.226	10.226	3.198	2.238	0.046
27	GBM_grid_1_AutoML_1_20230609_122708_model_1810	10.227	10.227	3.198	2.269	0.046
28	GBM_grid_1_AutoML_1_20230609_122708_model_192	10.233	10.233	3.199	2.263	0.046
29	GBM_grid_1_AutoML_1_20230609_122708_model_1224	10.242	10.242	3.2	2.236	0.047
30	GBM_grid_1_AutoML_1_20230609_122708_model_1798	10.255	10.255	3.202	2.243	0.046
31	GBM_grid_1_AutoML_1_20230609_122708_model_470	10.262	10.262	3.203	2.269	0.046
32	GBM_grid_1_AutoML_1_20230609_122708_model_733	10.267	10.267	3.204	2.252	0.046
33	GBM_grid_1_AutoML_1_20230609_122708_model_1538	10.269	10.269	3.205	2.279	0.046
34	GBM_grid_1_AutoML_1_20230609_122708_model_1215	10.273	10.273	3.205	2.225	0.046
35	GBM_grid_1_AutoML_1_20230609_122708_model_1952	10.289	10.289	3.208	2.267	0.046
36	GBM_grid_1_AutoML_1_20230609_122708_model_1692	10.29	10.29	3.208	2.251	0.046

Figure 3.37: Leaderboard table

The first row corresponds to the GBM.

1. **Gradient boosting machine:** Boosting models were originally developed for classification problems and were later extended to the regression setting[30]. Both continuous and categorical target variables can be predicted using the gradient boosting approach (as a regressor or classifier). The cost function is Mean Square Error (MSE) when it is used as a regressor, while it is Log loss when it is used as a classifier. The workflow for these models is shown in 3.38:

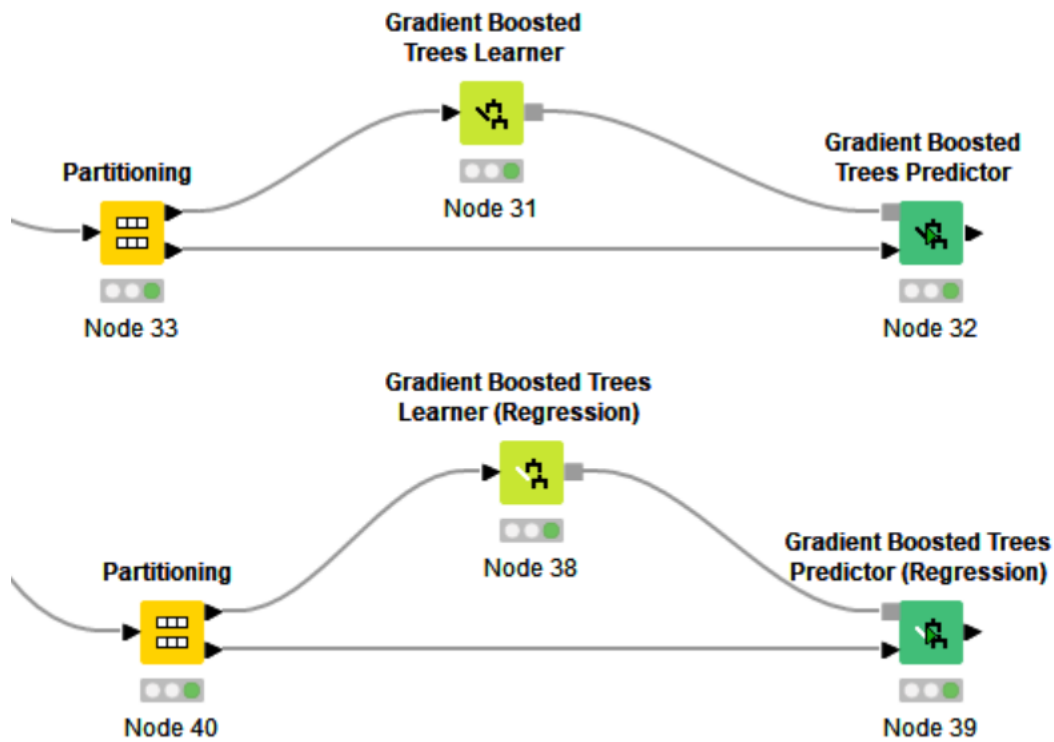


Figure 3.38: Gradient boosters workflow

Our dataset was partitioned using a partitioning node. 70% was dedicated to the learning node, while the predicting node tried to predict the results of the remaining 30%. The learning nodes Learn Gradient Boosted Trees with the objective of classification. The algorithm uses very shallow regression trees and a special form of boosting to build an ensemble of trees. The implementation follows the algorithm in section 4.6 of the paper "Greedy Function Approximation: A Gradient Boosting Machine" by Jerome H. Friedman (1999). While the predicting nodes classify data using the learned model. The prediction results of the classifier are shown in 3.39:



Table "default" - Rows: 324 Spec - Columns: 6 Properties Flow Variables						
Row ID	[S] Date	[S] Time	[S] Object Name	[S] Cluster	[S] Predict...	[D] Predict...
Row0	2023-03-10	00:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	0.999
Row1	2023-03-10	00:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	0.999
Row3	2023-03-10	01:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	1
Row7	2023-03-10	02:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_2	cluster_2	0.999
Row8	2023-03-10	02:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	1
Row9	2023-03-10	03:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_0	0.754
Row10	2023-03-10	03:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_2	cluster_2	1
Row13	2023-03-10	04:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row16	2023-03-10	05:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row24	2023-03-10	08:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	1
Row29	2023-03-10	09:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	0.998
Row32	2023-03-10	10:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	0.999
Row34	2023-03-10	11:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row35	2023-03-10	11:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_0	cluster_0	1
Row37	2023-03-10	12:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row43	2023-03-10	14:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row44	2023-03-10	14:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_0	cluster_0	1
Row45	2023-03-10	15:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_0	cluster_0	1
Row47	2023-03-10	15:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_0	cluster_0	1
Row48	2023-03-10	16:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_0	cluster_0	1
Row51	2023-03-10	17:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_0	cluster_0	1
Row52	2023-03-10	17:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row56	2023-03-10	18:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	1
Row59	2023-03-10	19:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	0.997
Row60	2023-03-10	20:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	0.997
Row61	2023-03-10	20:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	0.999
Row63	2023-03-10	21:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_0	cluster_0	1
Row64	2023-03-10	21:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row65	2023-03-10	21:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	0.999
Row67	2023-03-10	22:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row73	2023-03-11	00:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_0	cluster_0	1
Row74	2023-03-11	00:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_2	cluster_2	1
Row76	2023-03-11	01:00	USN9810_Blida/Whole System:USN9810_Blida	cluster_2	cluster_2	0.999
Row90	2023-03-11	06:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	0.988
Row92	2023-03-11	06:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_0	cluster_0	0.998
Row98	2023-03-11	08:00	USN9810_Dely/Whole System:USN9810_Dely	cluster_0	cluster_0	1
Row99	2023-03-11	09:00	USN9810_Annaba/Whole System:USN9810_Annaba	cluster_2	cluster_2	1

Figure 3.39: Gradient boosters classifier

This node predicted the clusters of each row, with an additional column about the confidence of the decision. We can see that although the dataset used for learning was small, the results were satisfying regardless of the very few mistakes made.

Next we will see the results of the regressor prediction:



Row ID	S Date	S Time	S Object Name	D technic...	D Predicti...
Row63	2023-03-10	21:00	USN9810_Annaba/Whole System:USN9810...	80.24	81.721
Row68	2023-03-10	22:00	USN9810_Dely/Whole System:USN9810_Dely	80.12	82.957
Row73	2023-03-11	00:00	USN9810_Blida/Whole System:USN9810_Blida	82.611	76.871
Row75	2023-03-11	01:00	USN9810_Annaba/Whole System:USN9810...	78.079	77.019
Row76	2023-03-11	01:00	USN9810_Blida/Whole System:USN9810_Blida	71.765	79.107
Row79	2023-03-11	02:00	USN9810_Blida/Whole System:USN9810_Blida	77.606	76.059
Row83	2023-03-11	03:00	USN9810_Dely/Whole System:USN9810_Dely	74.591	70.384
Row86	2023-03-11	04:00	USN9810_Dely/Whole System:USN9810_Dely	79.215	75.535
Row90	2023-03-11	06:00	USN9810_Annaba/Whole System:USN9810...	78.606	75.223
Row92	2023-03-11	06:00	USN9810_Dely/Whole System:USN9810_Dely	91.985	73.295
Row93	2023-03-11	07:00	USN9810_Annaba/Whole System:USN9810...	81.42	77.673
Row95	2023-03-11	07:00	USN9810_Dely/Whole System:USN9810_Dely	91.915	78.436
Row97	2023-03-11	08:00	USN9810_Blida/Whole System:USN9810_Blida	81.779	79.24
Row98	2023-03-11	08:00	USN9810_Dely/Whole System:USN9810_Dely	90.097	76.094
Row106	2023-03-11	11:00	USN9810_Blida/Whole System:USN9810_Blida	61.257	77.08
Row107	2023-03-11	11:00	USN9810_Dely/Whole System:USN9810_Dely	58.605	75.925
Row108	2023-03-11	12:00	USN9810_Annaba/Whole System:USN9810...	81.256	79.557
Row112	2023-03-11	13:00	USN9810_Blida/Whole System:USN9810_Blida	87.868	83.758
Row115	2023-03-11	14:00	USN9810_Blida/Whole System:USN9810_Blida	75.406	78.224
Row116	2023-03-11	14:00	USN9810_Dely/Whole System:USN9810_Dely	74.819	75.771
Row121	2023-03-11	16:00	USN9810_Blida/Whole System:USN9810_Blida	86.339	81.411
Row127	2023-03-11	18:00	USN9810_Blida/Whole System:USN9810_Blida	72.677	74.171
Row131	2023-03-11	19:00	USN9810_Dely/Whole System:USN9810_Dely	59.808	60.675
Row141	2023-03-11	23:00	USN9810_Annaba/Whole System:USN9810...	65.855	72.756
Row142	2023-03-11	23:00	USN9810_Blida/Whole System:USN9810_Blida	60.973	74.249
Row143	2023-03-11	23:00	USN9810_Dely/Whole System:USN9810_Dely	60.281	73.302
Row148	2023-03-12	01:00	USN9810_Blida/Whole System:USN9810_Blida	36.801	55.513
Row150	2023-03-12	02:00	USN9810_Annaba/Whole System:USN9810...	47.448	55.084
Row154	2023-03-12	03:00	USN9810_Blida/Whole System:USN9810_Blida	40.469	56.823
Row163	2023-03-12	06:00	USN9810_Blida/Whole System:USN9810_Blida	59.187	73.852
Row164	2023-03-12	06:00	USN9810_Dely/Whole System:USN9810_Dely	62.388	67.873
Row171	2023-03-12	09:00	USN9810_Annaba/Whole System:USN9810...	81.476	77.223
Row173	2023-03-12	09:00	USN9810_Dely/Whole System:USN9810_Dely	77.472	76.596
Row174	2023-03-12	10:00	USN9810_Annaba/Whole System:USN9810...	81.591	79.65
Row175	2023-03-12	10:00	USN9810_Blida/Whole System:USN9810_Blida	79.334	81.722
Row179	2023-03-12	11:00	USN9810_Dely/Whole System:USN9810_Dely	83.239	80.571
Row181	2023-03-12	12:00	USN9810_Blida/Whole System:USN9810_Blida	81.101	81.361

Figure 3.40: Gradient boosters regressor

The results in this case are less accurate than the classifier, that is expected since predicting classes is more tolerant than predicting values. Trying a larger dataset may improve the results of this node. Now we will try another model and see if the GBM was truly the best model as determined by the AutoML learner.

2. **Random forest algorithm:** Leo Breiman and Adele Cutler are the creators of the widely used machine learning technique known as random forest, which mixes the output of various decision trees to produce a single outcome. Its widespread use is motivated by its adaptability and usability because it can solve classification and regression issues. We will use only the classification model :

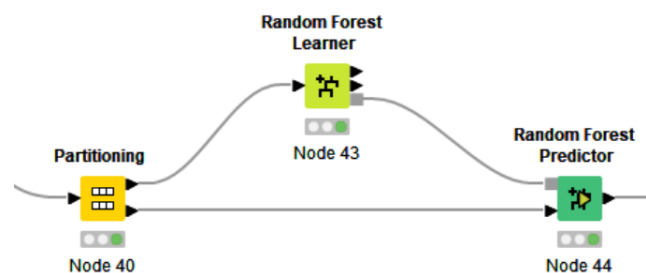


Figure 3.41: Random forest workflow

We used the same dataset and partitioning used in the previous model for comparison. The predictions are shown in the following table [3.42](#):

Row ID	S Date	S Time	S Object Name	D technic...	S Cluster	S Prediction (Cluster)	D Prediction (Cluster) (Confidence)
Row56	2023-03-10	18:00	USN9810_Dely/Whole System:USN9810_Dely	82.373	cluster_2	cluster_2	0.54
Row57	2023-03-10	19:00	USN9810_Annaba/Whole System:USN9810...	82.308	cluster_0	cluster_0	0.78
Row58	2023-03-10	19:00	USN9810_Blida/Whole System:USN9810_Blida	82.866	cluster_0	cluster_0	0.96
Row59	2023-03-10	19:00	USN9810_Dely/Whole System:USN9810_Dely	82.77	cluster_2	cluster_2	0.55
Row63	2023-03-10	21:00	USN9810_Annaba/Whole System:USN9810...	80.24	cluster_0	cluster_2	0.52
Row68	2023-03-10	22:00	USN9810_Dely/Whole System:USN9810_Dely	80.12	cluster_2	cluster_2	0.66
Row73	2023-03-11	00:00	USN9810_Blida/Whole System:USN9810_Blida	82.611	cluster_0	cluster_0	0.79
Row75	2023-03-11	01:00	USN9810_Annaba/Whole System:USN9810...	78.079	cluster_2	cluster_2	0.55
Row76	2023-03-11	01:00	USN9810_Blida/Whole System:USN9810_Blida	71.765	cluster_2	cluster_2	0.75
Row79	2023-03-11	02:00	USN9810_Blida/Whole System:USN9810_Blida	77.606	cluster_0	cluster_0	0.88
Row83	2023-03-11	03:00	USN9810_Dely/Whole System:USN9810_Dely	74.591	cluster_2	cluster_2	0.8
Row86	2023-03-11	04:00	USN9810_Dely/Whole System:USN9810_Dely	79.215	cluster_2	cluster_2	0.62
Row90	2023-03-11	06:00	USN9810_Annaba/Whole System:USN9810...	78.606	cluster_2	cluster_2	0.62
Row92	2023-03-11	06:00	USN9810_Dely/Whole System:USN9810_Dely	91.985	cluster_0	cluster_0	0.47
Row93	2023-03-11	07:00	USN9810_Annaba/Whole System:USN9810...	81.42	cluster_0	cluster_0	0.58
Row95	2023-03-11	07:00	USN9810_Dely/Whole System:USN9810_Dely	91.915	cluster_0	cluster_0	0.58
Row97	2023-03-11	08:00	USN9810_Blida/Whole System:USN9810_Blida	81.779	cluster_0	cluster_0	0.96
Row98	2023-03-11	08:00	USN9810_Dely/Whole System:USN9810_Dely	90.097	cluster_0	cluster_0	0.6
Row106	2023-03-11	11:00	USN9810_Blida/Whole System:USN9810_Blida	61.257	cluster_1	cluster_1	0.55
Row107	2023-03-11	11:00	USN9810_Dely/Whole System:USN9810_Dely	58.605	cluster_1	cluster_1	0.68
Row108	2023-03-11	12:00	USN9810_Annaba/Whole System:USN9810...	81.256	cluster_0	cluster_0	0.8
Row112	2023-03-11	13:00	USN9810_Blida/Whole System:USN9810_Blida	87.868	cluster_0	cluster_0	0.95
Row115	2023-03-11	14:00	USN9810_Blida/Whole System:USN9810_Blida	75.406	cluster_0	cluster_0	0.8
Row116	2023-03-11	14:00	USN9810_Dely/Whole System:USN9810_Dely	74.819	cluster_2	cluster_2	0.92
Row121	2023-03-11	16:00	USN9810_Blida/Whole System:USN9810_Blida	86.339	cluster_0	cluster_0	0.93
Row127	2023-03-11	18:00	USN9810_Blida/Whole System:USN9810_Blida	72.677	cluster_2	cluster_2	0.59
Row131	2023-03-11	19:00	USN9810_Dely/Whole System:USN9810_Dely	59.808	cluster_1	cluster_1	0.57
Row141	2023-03-11	23:00	USN9810_Annaba/Whole System:USN9810...	65.855	cluster_2	cluster_2	0.82
Row142	2023-03-11	23:00	USN9810_Blida/Whole System:USN9810_Blida	60.973	cluster_1	cluster_1	0.43
Row143	2023-03-11	23:00	USN9810_Dely/Whole System:USN9810_Dely	60.281	cluster_1	cluster_2	0.66
Row148	2023-03-12	01:00	USN9810_Blida/Whole System:USN9810_Blida	36.801	cluster_1	cluster_1	0.82
Row150	2023-03-12	02:00	USN9810_Annaba/Whole System:USN9810...	47.448	cluster_1	cluster_1	0.84
Row154	2023-03-12	03:00	USN9810_Blida/Whole System:USN9810_Blida	40.469	cluster_1	cluster_1	0.85
Row163	2023-03-12	06:00	USN9810_Blida/Whole System:USN9810_Blida	59.187	cluster_1	cluster_1	0.65
Row164	2023-03-12	06:00	USN9810_Dely/Whole System:USN9810_Dely	62.388	cluster_1	cluster_2	0.5
Row171	2023-03-12	09:00	USN9810_Annaba/Whole System:USN9810...	81.476	cluster_0	cluster_0	0.86
Row173	2023-03-12	09:00	USN9810_Dely/Whole System:USN9810_Dely	77.472	cluster_2	cluster_2	0.77

Figure 3.42: Random forest prediction table

We can see that the confidence of the prediction is less than that of the GBM, also there are more errors in the classification. This shows that GBM was indeed the best model for our predictive analysis. Overall, the results of our predictive analysis were not 100% accurate. The few prediction mistakes are due to the insufficient amount of data that we have for better learning. However, the results of the GBM classifier were satisfying.

## 3.6.2 Coded Anomaly detection

### 3.6.2.1 Data preprocessing:

Data preprocessing is a crucial initial phase in data analysis, focused on preparing and refining raw data to enhance its suitability for further analysis. Its primary objectives include resolving data quality concerns, addressing missing values, normalizing or scaling the data, and ensuring it conforms to the necessary format for analysis. When dealing with data stored in database tables, the initial step involves retrieving the relevant data from the tables and storing each required table in a Python data frame to facilitate analysis. The subsequent step entails managing any missing values using the mean imputation technique, while also ensuring the absence of duplicate rows within the data frame.

### 3.6.2.2 Getting normal range from main KPIs:

After cleaning and storing the data in a data frame, we can move forward with the analysis by determining the normal range for our key performance indicators (KPIs). It is crucial to select the relevant columns for the analysis of each procedure, as we do not need to calculate the

normal range for every column. In our scenario, the main KPI counters consist of the success rates columns and the columns describing network-related failures. To calculate the normal range, we begin by filtering the data using a specified date range. This allows us to focus on the relevant time period for our analysis. Then, for each column of interest, we calculate the mean and standard deviation for that column within the filtered data using the following formulas:

$$\mu = \frac{\sum_{i=1}^N (x_i)}{N} \quad (3.8)$$

Where:

- $\mu$  is the mean,
- $x_i$  represents each individual value in the dataset,
- $N$  is the total number of values in the dataset.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (3.9)$$

Where:

- $\sigma$  is the standard deviation,
- $x_i$  represents each individual value in the dataset,
- $\mu$  is the mean,
- $N$  is the total number of values in the dataset.

The normal range is determined by establishing a range centered around the mean value. For the success rate columns, the normal range is defined as three standard deviations below and above the mean:  $[\mu - (3 \times \sigma), \mu + (3 \times \sigma)]$ . On the other hand, for the failure columns, the normal range is defined as a minimum value of 0 and the mean plus two times the standard deviation:  $[0, \mu + (2 \times \sigma)]$ . This range is considered "normal" and is used to determine if a value falls within expected bounds. The calculated normal ranges are stored in dictionaries, where each column is associated with a range for each unique object in the dataset. This allows for easy reference and comparison of values against their respective normal ranges during further analysis or reporting.

### 3.6.2.3 Detecting anomalies using normal range:

The process of detecting anomalies using the calculated normal ranges involves a systematic comparison of column values against their respective normal ranges. We iterate over the success rate columns, examining whether values fall outside the established normal range for each unique MME in the table. Any identified anomaly is marked as True in the "Anomaly" column of the corresponding DataFrame. Additionally, critical anomalies are identified by assessing the anomaly's distance from the lower bound of the normal range relative to the overall range. If an anomaly exceeds 30% of the overall range, it is labeled as critical. To provide comprehensive insights, we generate detailed reports for each critical anomaly, including crucial information such as the object name, timestamp, column name, anomaly value, and the minimum value allowed within the normal range. Furthermore, we conduct an in-depth examination of related columns to understand the underlying causes of anomalies. By retrieving the normal ranges for these failure columns, we analyze whether any values exceed the expected range. If anomalies are found in the related columns, they are promptly included in the anomaly report, offering visibility into the specific column and the anticipated maximum value within the normal range.

Overall, our code adopts a comprehensive approach to anomaly detection, leveraging calculated normal ranges to effectively identify and report anomalies within the success rate columns. The inclusion of related columns enhances anomaly understanding and facilitates robust anomaly analysis and reporting, empowering network operators to gain valuable insights into network performance deviations and enabling proactive measures for resolution and optimization.

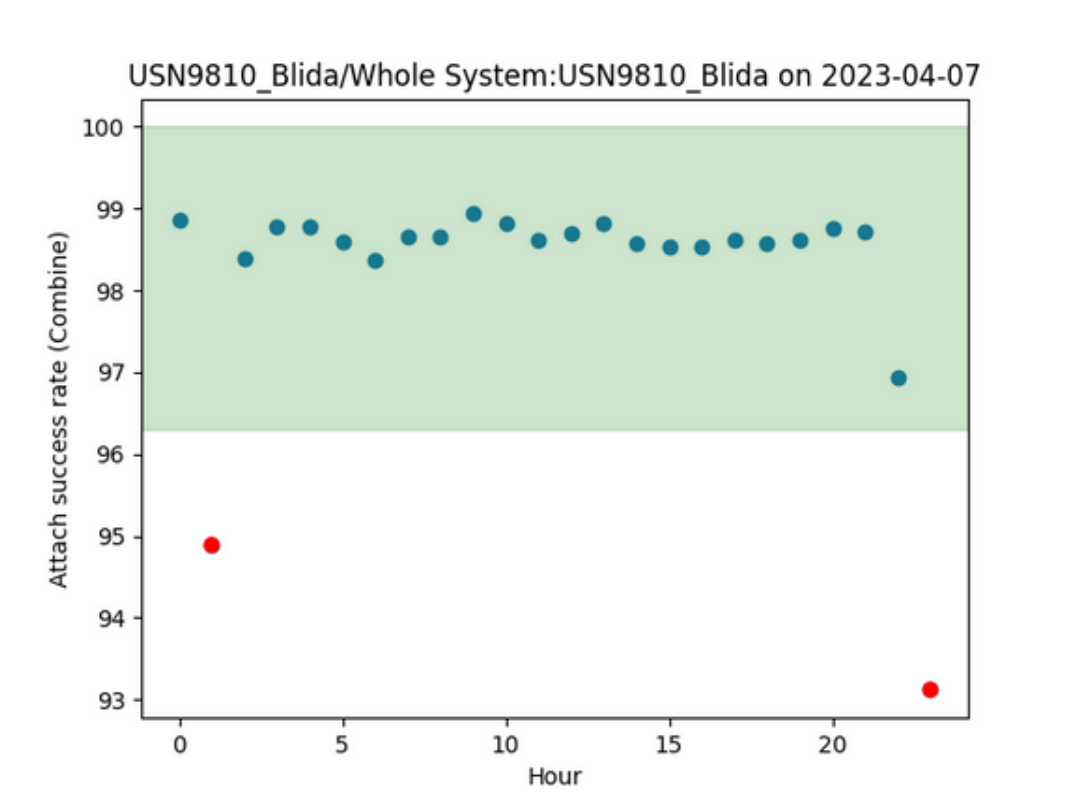


Figure 3.43: Anomaly visualization for USN\_Blida on 2023-04-07.

#### 3.6.2.4 Anomaly report creation:

Once anomalies have been detected, we need to generate detailed anomaly reports to provide comprehensive insights. The reporting process is further enhanced through visualization techniques. After identifying critical anomalies in the success rate columns based on the comparison with normal ranges, the code generates detailed reports for each critical anomaly. These reports include essential information such as the object name, timestamp, column name, anomaly value, and the minimum value allowed within the normal range. Additionally, the code incorporates visualizations by plotting the data points and highlighting the anomalies in scatter plots. The plots utilize a green shaded region to represent the normal range. These detailed reports and visualizations are created for each date within the specified date range and are stored in separate output files. In cases where no anomalies are detected, the code generates a report indicating "No Anomaly Detected Today." This combined approach of structured reports and visualizations enhances the overall analysis and enables network operators and analysts to easily review and investigate anomalies, providing valuable insights into potential network performance issues. The anomaly report creation process facilitates effective anomaly tracking, resolution, and communication among stakeholders, contributing to the overall efficiency and optimization of network operations.

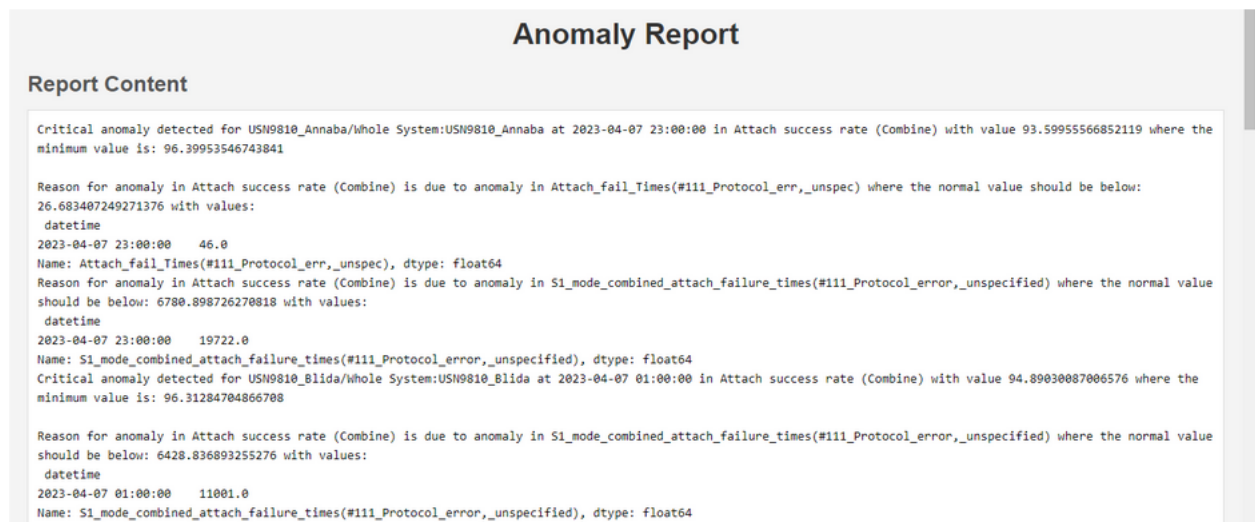


Figure 3.44: Anomaly report on 2023-04-07.

## 3.7 Mechanism Deployment

Once the mechanism has been implemented, one way to utilize it is by integrating it into a local web application. This application should be installed on the server, granting access to the entire working team simply by entering the application's URL.

### 3.7.1 Development tools:

#### 3.7.1.1 Flask library:

Flask is a lightweight web framework for building web applications in Python. It follows a minimalistic approach, providing essential tools for web development without imposing a specific project structure. With Flask, you can easily define routes using decorators and handle different HTTP methods. It supports Jinja2 templating for creating dynamic HTML pages. Flask also offers request and response handling, along with a wide range of extensions for integrating additional functionalities like database integration and authentication. Its simplicity, flexibility, and active community make it a popular choice for developing small to medium-sized web applications.

#### 3.7.1.2 HTML:

HTML (Hypertext Markup Language) is the standard markup language used for creating web pages and applications. It provides the structure and content of a webpage, defining the elements and their relationships. HTML uses tags to markup different parts of a document, such as headings, paragraphs, images, links, forms, and more. It allows for the inclusion of multimedia elements like audio and video, as well as the organization and presentation of data. HTML documents can be viewed in web browsers, which interpret the markup and render the content accordingly. With HTML, developers can create interactive and visually appealing web pages that are accessible across different devices and platforms.

### 3.7.2 Web application presentation

The web app built using Flask serves as a user interface for various functionalities. Users access the app through a login page where they can enter their credentials. The app verifies the

provided username and password against a predefined dictionary of valid users. If the user does not have login credentials, he can fill out a form with his name, job, and email address. Upon submission, the app validates the form data and checks for an active internet connection. If the conditions are met, the app composes an email using the MIMEText module, including the submitted information. It then establishes a connection with the SMTP server, logs in with the appropriate credentials, and sends the email to a specified recipient. Upon successful delivery, the app provides a confirmation message to the user.

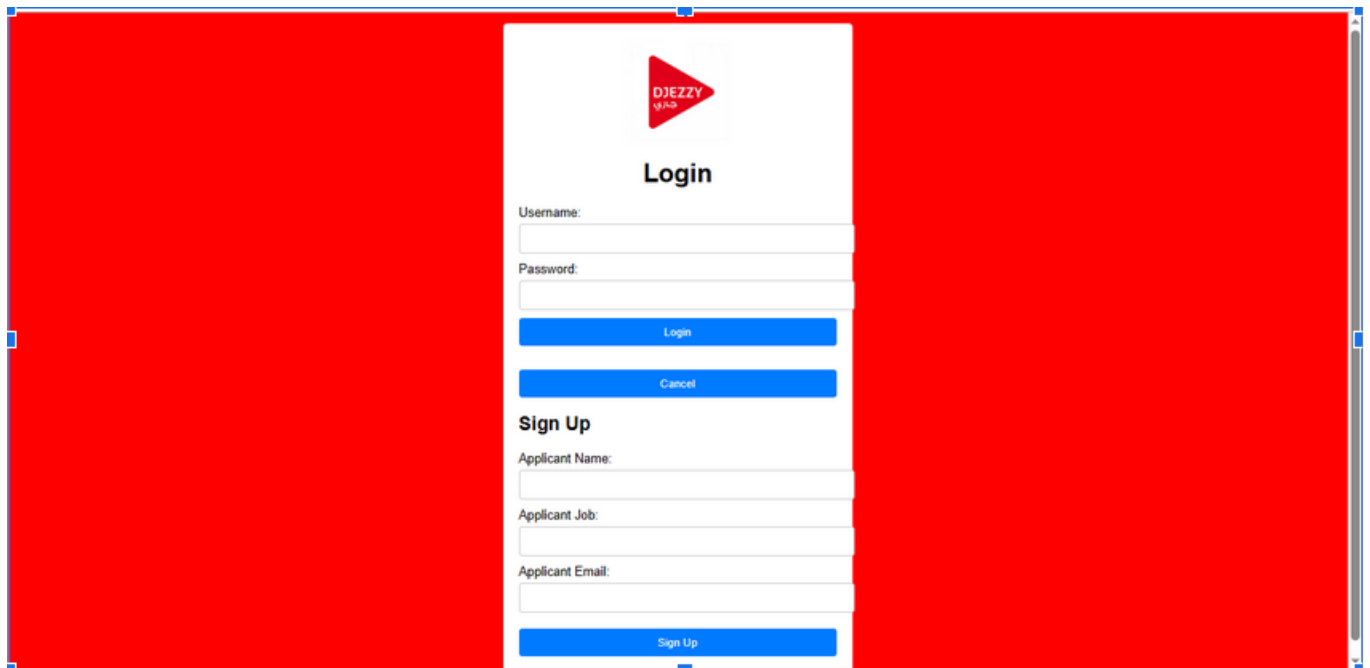


Figure 3.45: Login page for web app

If the authentication is successful, the user is redirected to a dashboard page, which serves as the main navigation hub which is shown in figure [3.46](#):

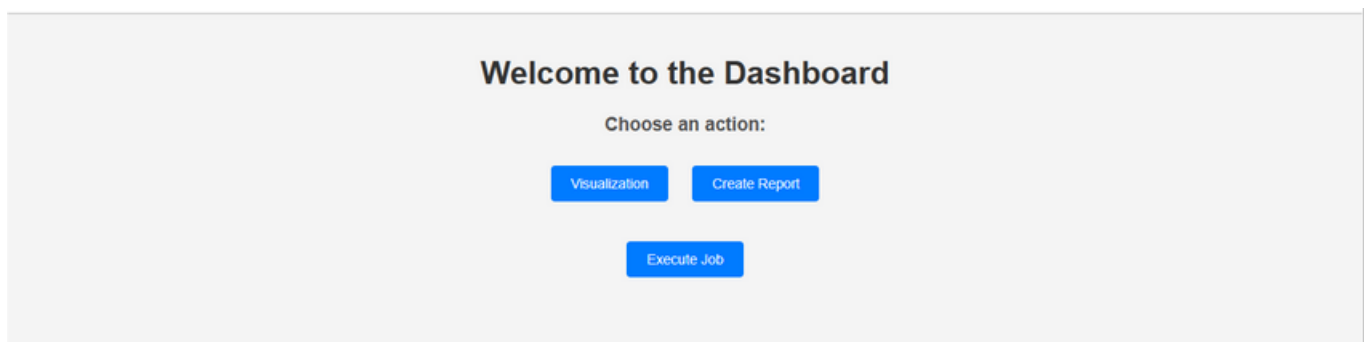


Figure 3.46: Navigation dashboard for web app

The web app also includes a visualization page that dynamically displays HTML content. This functionality utilizes Flask's integration with Jinja2 templating. The app accesses a Power BI report server URL and obtains an embed URL. The embed URL is then passed to the visualization page, where it is rendered within an HTML iframe. This allows users to view and interact with the Power BI report directly on the web app. To keep the visualization up to date, the web app provides a mechanism to refresh the report. When triggered, the app sends a POST



request to the Power BI report server URL, simulating the action of clicking the "Actualize" button. This ensures that the report data is refreshed, providing users with the most recent information and insights.

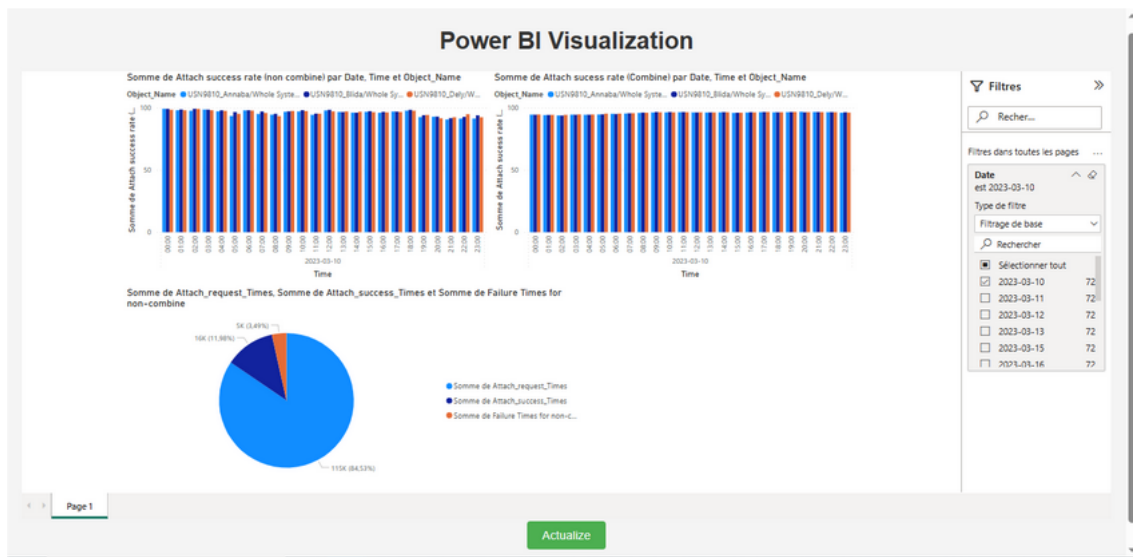


Figure 3.47: Visualization dashboard for web app

Furthermore, the web app facilitates the creation of anomaly reports. Users can input a specific date, and the app checks if a corresponding report file exists. If found, the app reads the content of the report file and generates plots based on the provided date. This functionality relies on pandas and matplotlib libraries for data manipulation and visualization. The generated plots are then encoded as base64 strings and passed to the anomaly report page, where they are rendered as images. This allows users to visualize and analyze anomalies in the data associated with the specified date.

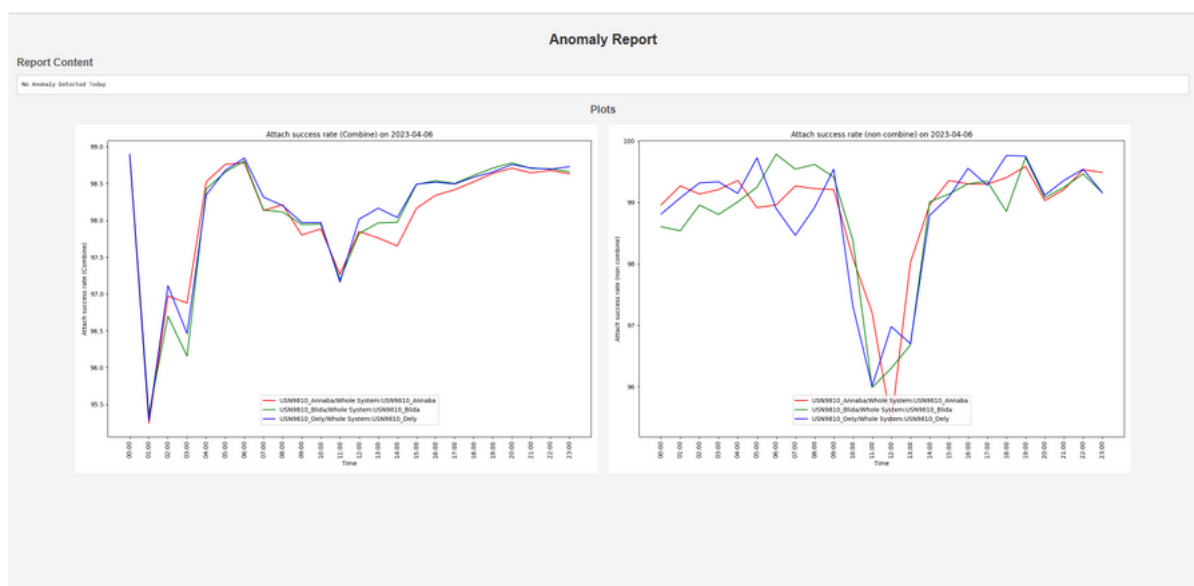


Figure 3.48: Anomaly report dashboard for web app

Additionally, the web app offers the capability to execute a job. Before executing the job, it checks for an active internet connection. Once validated, the app logs in to an Outlook account

---

using the provided credentials and searches for unread emails. If any emails contain zip file attachments, the app downloads and extracts them to a designated folder. It then proceeds to execute an additional task, which involves running a Pentaho job. The results of the job execution are displayed on a dedicated job result page. The page also includes buttons that show or hide the execution buttons based on the internet connection status, providing users with appropriate feedback.

In summary, the web app combines user authentication, email sending, dynamic visualization, anomaly report generation, and job execution functionalities. It leverages Flask's routing, request handling, and templating features to create an intuitive and interactive user interface, enabling users to perform various tasks efficiently within a web-based environment.

## 3.8 Summary

In this chapter, we outlined the implementation process for our mechanism, starting with the description of development tools and the introduction of the data used in the analysis. We then proceeded to collect and preprocess the data, followed by the integration of the data into specific database tables. To facilitate data analysis, we performed data staging and calculated the necessary technical success rates. Various methods were employed for data analysis, and the final step in the mechanism involved data visualization. Finally, we embedded our mechanism into a local web application to ensure accessibility for all service workers.



# General Conclusion

In conclusion, this project has addressed the critical need for comprehensive analysis and optimization of key performance indicators (KPIs) related to the quality of service (QoS) in the 4G/LTE core network for DJEZZY mobile operator. By developing a robust mechanism for analyzing these KPIs, valuable insights into the network's technical success rates across various processes have been obtained. The main contribution of this project was to promptly detect anomalies or malfunctions within the network. Through the utilization of advanced analytics techniques, deviations from expected performance levels have been identified, enabling timely troubleshooting and optimization. These proactive measures have proven instrumental in preventing service disruptions, reducing downtime, and enhancing the overall user experience. The implementation of a user-friendly web application has significantly enhanced accessibility and usability for service workers within the organization. This intuitive interface has empowered them to easily access and utilize the valuable insights gained from the analysis, facilitating efficient decision-making and prompt actions to rectify identified issues. As a result, the network's performance has been improved, leading to higher customer satisfaction and regulatory compliance.

The platform developed in this project meets the specified requirements and provides valuable insights for QoS optimization. However, it is important to acknowledge the limitations of the work. Firstly, the scope of the project was focused on the 4G/LTE core network of DJEZZY on specific procedures, and expanding the platform to incorporate other network components could provide a more comprehensive solution. Additionally, the project's findings and recommendations are based on the data available during its implementation, and ongoing monitoring and adjustments may be necessary to adapt to evolving network conditions.

Despite these limitations, this project has made a significant contribution to improving the quality of service provided by DJEZZY. It has provided a foundation for further enhancements and refinements in the future, enabling the platform to reach its full potential.

Overall, this project has successfully fulfilled its objectives by creating a platform that empowers DJEZZY mobile operator to monitor and optimize their quality of service. The project has not only enhanced technical skills but also facilitated the integration into the professional sphere, providing a valuable foundation for future endeavors in the telecommunications industry.

# Appendix A

## Tracking Area Optimization

In an LTE network, the coverage area of the network is divided into smaller regions called Tracking Areas. Each Tracking Area consists of a group of cells, and each cell covers a specific geographic area. An eNB can contain cells from different tracking areas, and a single cell can belong to more than one TA. A single TA may span many MME areas or be contained within a single MME region. When a UE registers in the network, it is responsible for registering inside a certain TA, and the core network records information about the tracking area when the registration is completed. When the MME pages a UE, a paging message is sent to all eNBs in the TAI list that includes 1 to 16 TA to notify the user about incoming data connections. The TA to which a cell belongs is sent via System Information Block 1 (SIB1) once every 80 ms [17]. In order to get around some of the drawbacks of the traditional TA, the LTE standard introduces the tracking area list (TAL) idea. With this method, one cell may contain a list of TAs rather than having one TA assigned to each cell. Instead of planning one TA, these TAs can be combined into a TAL. So when a UE repeatedly switches between two or more nearby cells in various TAs, no TAU is triggered if they are inside one TAL. The user receives the TA list from a cell, and will not change its list, until it moves to a cell, which is not included in the list. TA list can cover one or more TAs, TA list benefits over adding eNBs to single TA is that TA list can be created based on subscriber needs according to history of UE movement.

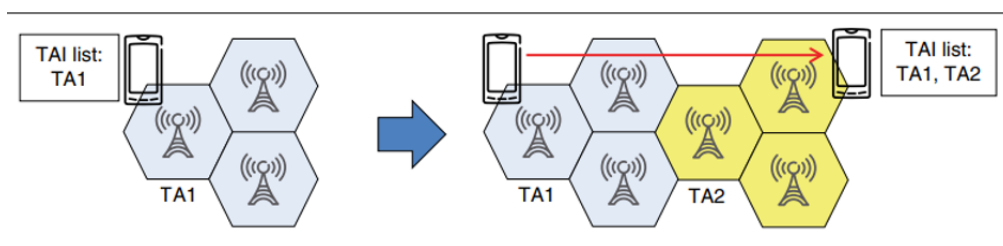


Figure A.1: Tracking Area scheme

To refresh its location, the UE must send frequent location updates to MME. TAUs (TA updates) are often classified into two types: periodic TAUs and mobility TAUs. With periodic TAU, the UE wakes up and attaches to the network at fixed intervals to refresh the EPC of its existence. This allows the MME to keep track of UEs that have gone out of coverage. If a tracking area update isn't received by MME within the expected timeframe, it will delete the context of the UE from its database. For mobility TAU, the UE must notify the MME of the change when it determines that it is traveling between different tracking zones, and the MME will then update its location database.

The UE initiates a new registration either when it enters a new TA/TA list where the registration is no longer valid (periodic registration) or after a predetermined amount of time. In order to

---

minimize the tracking area update signaling, an MME can allocate the UE multiple tracking areas. The TAI is the TA history of a UE that has been relocated to; if the UE has moved to another TA, the new TA is added to the TA list. One MME pool area must contain all of the TAs from the TA list given to the UE.

# Appendix B

## Machine Learning Models

### B.1 K-means

K-means clustering algorithm computes the centroids and iterates until it finds the optimal centroids. It presumes that there are already known quantities of clusters. It is also called flat clustering algorithm. The 'K' in K-means stands for the number of clusters that the algorithm was able to identify from the data. In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. Less diversity within the clusters will result in more identical data points inside the same cluster.

### B.2 GBM

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

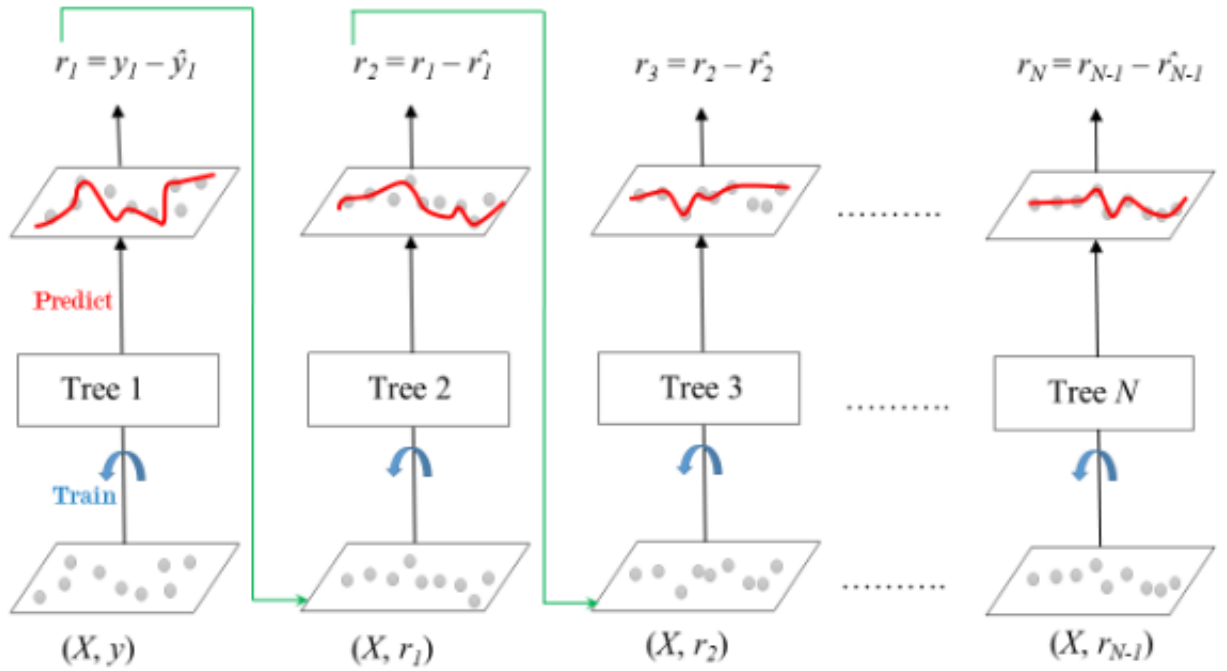


Figure B.1: Gradient boosted trees for regression

The ensemble consists of  $M$  trees. Tree1 is trained using the feature matrix  $X$  and the labels  $y$ . The predictions labeled  $\hat{y}_1$  are used to determine the training set residual errors  $r_1$ . Tree2 is then trained using the feature matrix  $X$  and the residual errors  $r_1$  of Tree1 as labels. The predicted results  $\hat{r}_1$  are then used to determine the residual  $r_2$ . The process is repeated until all the  $M$  trees forming the ensemble are trained. There is an important parameter used in this technique known as Shrinkage. Shrinkage refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate ( $\eta$ ) which ranges between 0 to 1. There is a trade-off between  $\eta$  and the number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Since all trees are trained now, predictions can be made. Each tree predicts a label and the final prediction is given by the formula:

$$y(pred) = y_1 + (\eta * r_1) + (\eta * r_2) + \dots + (\eta * r_N) \quad (B.1)$$

### B.3 Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. When performing Random Forests based on classification data, you should know that you are often

---

using the Gini index, or the formula used to decide how nodes on a decision tree branch.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2 \quad (B.2)$$

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here,  $p_i$  represents the relative frequency of the class you are observing in the dataset and  $c$  represents the number of classes. You can also use entropy to determine how nodes branch in a decision tree.

$$Entropy = \sum_{i=1}^C (-p_i \times \log_2(p_i)) \quad (B.3)$$

Entropy uses the probability of a certain outcome in order to make a decision on how the node should branch. It is more mathematically complex than the Gini index since a logarithmic function is utilized to calculate it.

# Bibliography

- [1] Shakil Akhtar. “2G-5G networks: Evolution of technologies, standards, and deployment”. In: *Encyclopedia of Multimedia Technology and Networking* (2009).
- [2] A Kukushkin. “Introduction to Mobile Network Engineering GM, 3G-WCDMA”. In: *LTE and the Road to 5G* 5 (2018).
- [3] Martin Sauter. *From GSM to LTE-advanced: an introduction to mobile networks and mobile broadband*. John Wiley & Sons, 2014.
- [4] Siretta Limited - Enabling Industrial IoT, <https://www.siretta.com>.
- [5] Pankaj Sharma. “Evolution of mobile wireless communication networks-1G to 5G as well as future prospective of next generation communication network”. In: *International Journal of Computer Science and Mobile Computing* 2.8 (2013), pp. 47–53.
- [6] Erik Dahlman, Stefan Parkvall, and Johan Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.
- [7] EP Ivanova et al. “Evolution of mobile networks and seamless transition to 5G”. In: *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 2021.
- [8] Siemund M Redl, Mathias K Weber, and Malcolm W Opliphant. *GSM and Personal Communication Handbook*. Artech House, Inc., 1998.
- [9] Asha Mehrotra. *GSM system engineering*. Artech House, Inc., 1997.
- [10] Jorg Eberspacher et al. *GSM-architecture, protocols and services*. John Wiley & Sons, 2008.
- [11] Ashraful Arefin. “UNIVERSAL MOBILE TELECOMMUNICATION SYSTEM”. In: (2013).
- [12] Amitabh Mishra. “Performance and architecture of SGSN and GGSN of general packet radio service (GPRS)”. In: *GLOBECOM’01. IEEE Global Telecommunications Conference (Cat. No. 01CH37270)*. Vol. 6. IEEE, 2001.
- [13] Volkan Sevindik et al. “Performance evaluation of a real long term evolution (LTE) network”. In: *37th Annual IEEE Conference on Local Computer Networks-Workshops*. Ieee, 2012, pp. 679–685.
- [14] Aderemi A Atayero et al. “3GPP long term evolution: Architecture, protocols and interfaces”. In: *International Journal of Information and Communication Technology Research* 1 (2011).
- [15] Stefania Sesia, Issam Toufik, and Matthew Baker. *LTE-the UMTS long term evolution: from theory to practice*. John Wiley & Sons, 2011.
- [16] Harri Holma and Antti Toskala. *LTE for UMTS: Evolution to LTE-advanced*. John Wiley & Sons, 2011.
- [17] Xincheng Zhang. *LTE optimization engineering handbook*. John Wiley & Sons, 2018.

- 
- [18] Fidel Krasniqi, Liljana Gavrilovska, and Arianit Maraj. “The analysis of key performance indicators (KPI) in 4G/LTE networks”. In: *Future Access Enablers for Ubiquitous and Intelligent Infrastructures: 4th EAI International Conference, FABULOUS 2019, Sofia, Bulgaria, March 28-29, 2019, Proceedings* 283. Springer. 2019, pp. 285–296.
- [19] Jose Dario Luis Delgado and Jesus Maximo Ramirez Santiago. “Key performance indicators for QOS assessment in TETRA networks”. In: *arXiv preprint arXiv:1401.1918* (2014).
- [20] 3gpp “EventHelix”. <https://www.EventHelix.com/EventStudio>.
- [21] about python. <https://www.python.org/about/>.
- [22] Maria Carina Roldan. *Pentaho Data Integration Beginner’s Guide*. Packt Publishing Ltd, 2013.
- [23] Michael R. Berthold et al. “KNIME: The Konstanz Information Miner”. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007. ISBN: 978-3-540-78239-1.
- [24] D.Petkovic. *Microsoft SQL Server 2019: A Beginner’s Guide, Eighth Edition*. McGraw-Hill Education, 2020.
- [25] Oracle definition of a database. <https://www.oracle.com/database/what-is-database/>.
- [26] “Universal Mobile Telecommunications System (UMTS); LTE; 5G; Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3”. In: *3GPP TS 24.301 version 15.4.0 Release 15* (2018).
- [27] Swati Patel. *K-means Clustering Algorithm: Implementation and Critical Analysis*. Scholars’ Press, 2019.
- [28] H2O definition. <https://h2o.ai/platform/ai-cloud/make/h2o/>.
- [29] Statistics Canada. “Statistics: Power from Data!” In: (2021).
- [30] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.