Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(7), 415–426
https://doi.org/10.12913/22998624/204321
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.04.14
Accepted: 2025.05.14
Published: 2025.05.25

Anomaly detection and security threats in Internet of Things: A study of data integrity in the measurement process

Adam Paluch[1], Monika Paluch-Ferszt[2], Michał Kalisz[3], Michalina Gryniewicz-Jaworska[4*]

[1] Under Ant sp. z o. o. sp. k., Łączyny 4, 02-820 Warsaw, Poland
[2] Heavy Ion Laboratory, University of Warsaw, ul. Pasteura 5A, 02-093, Warszawa, Poland
[3] Department of the Foundations of Computer Science, Faculty of Philosophy, The John Paul II Catholic University of Lublin, Aleje Racławickie 14, 20-950 Lublin, Poland
[4] WSEI University, Projektowa 4, 20-209 Lublin, Poland
* Corresponding author's e-mail address: michalina.gryniewicz-jaworska@wsei.pl

ABSTRACT

The publication presents potential places of conscious or unconscious interference or degradation of data collected using a distributed system of sensors and/or other Internet of Things devices. Places of possible interference in the measurement data are indicated from the very beginning of the measurement chain – from a single measurement element, through data conversion, transmission, processing, storage, analysis to interpretation. Threats are indicated not only of technical but also economic nature. Examples are presented that are intended to show how the flow of data can affect decision-making and how the lack of knowledge about the measurement context can affect their interpretation. Possible anomaly detection mechanisms are also indicated, considering new, developing techniques.

Keywords: mesh measurement, big data, machine learning, Internet of Things, influence point, telemetry.

INTRODUCTION

Since man gained consciousness, he has always used data to make decisions. As awareness grew, the amount of data analyzed continued to expand. At some point, their analysis became more complicated – a person began to consider making decisions that affect not only his life (sometimes he did not even consider that his decisions may affect something/someone else) but also the individuals in his environment. He then realized that these decisions impacted an even wider area of his life. With the current development of civilization, we have come to consider the impact of our activities on other planets, and perhaps even galaxies [1].

This type of analysis requires, on the one hand, a large amount of data from many sources, as well as capturing patterns, dependencies or repeatability. For this purpose, man has developed a number of tools, from the simplest comparison of single data, through statistical tools that can capture trends, to tools based on neural networks, machine learning and artificial intelligence (AI) that can help capture (previously noticed and developed) patterns and trends in input data. All these tools also enable forecasting (predictions) of events in the future.

On the surface, the subject of experimental data is quite simple and defined, but it is based on idealized and hermetic assumptions. In practice, the number of data sources increases, science has not determined (and will not determine for a long time) all possible relationships and the impact of one data on another (unless in a very narrow field or scope). In almost all considerations, we take it for granted that we have data, there is enough of it and it is of appropriate quality.

Modern advancements in data acquisition and processing, particularly in the realm of Internet

of Things (IoT) systems, have introduced unprecedented opportunities and challenges. This development necessitates rigorous scrutiny of data integrity, as errors, whether arising from technical limitations or human oversight, can lead to significant consequences, particularly in critical domains such as environmental monitoring or public safety. Acquisition, transmission, processing and analysis of big data is a relatively new issue and therefore many mechanisms and standards have yet to be developed. In many aspects, such as in the case of data transmission, already developed technologies can be used, adapting them to the specificity of the issue. In others (storage and processing of large amounts of data), technologies and solutions are just being developed and many solutions exist in parallel.

This paper aims to systematically examine potential vulnerabilities in data collection and processing pipelines, offering both theoretical insights and practical recommendations for mitigating risks.

Considering that IoT devices and systems of such devices are a relatively new field, sets of good practices and recommendations are still being created. The following studies deserve attention: [2–3]. These studies present the practical knowledge collected by the authors. Real threats are presented and a set of good practices on how to minimize threats. The authors devoted a lot of space to explaining the mechanisms of potential attacks and weak points in hardware and software implementations. Equipment that can be used for attacks is also presented. Penetration testing is also widely discussed to help find weak points during your own IoT implementations.

The Message Queuing Telemetry Transport (MQTT) protocol, widely used in various industries for Machine-to-Machine (M2M) communication, plays a central role in the IoT architecture. It is also worth mentioning an interesting article [4] that presents an understanding of the vulnerabilities in the MQTT protocol, highlights the critical importance of strengthening security, and serves as a reference point for new researchers in this field.

In the rapidly expanding field of the IoT, ensuring the reliability and integrity of sensor data has become a pressing concern. Modern IoT systems operate across distributed environments and are often deployed in critical applications, from healthcare monitoring to climate observation. As such, the reliability of the data they generate directly impacts decision-making processes that may have wide-ranging consequences. Existing literature addresses various components of data systems individually, but comprehensive frameworks that consider the entire measurement chain from sensor to interpretation remain limited. The purpose of this article is to fill this gap by analyzing potential threats and data degradation across all stages of the IoT data life cycle.

RESEARCH METHODOLOGY

The research methodology involved the design and deployment of a modular IoT system equipped with sensors for temperature, humidity, and pressure. In our research, we started with a hypothesis that all selected, commercially available sensors would work the same way under identical external conditions. After testing the sensors, we performed a thought experiment to identify potential threats that may arise when transmitting data from sensors and when analyzing large amounts of measurement data.

Sensor modules were evaluated based on accuracy, documentation, communication interface (I2C), and implementation feasibility. Calibration procedures were conducted in a climatic chamber set at 25 °C and 50% humidity to provide reference data for offset calculations. Data collection was carried out both under controlled environmental conditions and in natural outdoor settings to compare measurement deviations. The system architecture included microcontroller-based data acquisition, GSM-based transmission, and cloud-based storage using standard telemetry protocols like MQTT.

To conduct experiments with data, a device was developed that allows environmental measurements to be made with temperature, humidity and pressure sensors and to transmit the results via the GSM network. The device includes support for solar panels and a buffer battery, which allows the device to be installed in virtually any external place. The device has additional pins (UART, I2C) for any future expansion and testing. The device is divided into three functional modules: a motherboard containing a microcontroller, a power supply system and expansion slots, a GSM module containing a GSM modem and GPS, and a module with measurement sensors. The most important issue for creating the device was the selection of measurement sensors.

Search criteria: completeness of documentation including measurement errors, availability, ease of implementation, available materials on implementation and experience in use, I2C interface, price. The following sensors were selected, presented in Table 1.

Due to the nature of the measurements, i.e. temperature, pressure and humidity measurements, it was necessary to use appropriate covers to prevent the sensors from being exposed to direct wind, sun, rain, snow, etc. A standard, widely used formula was chosen. A set of all sensors in covers (3 pcs) was mounted at a similar height on a 40 cm high support. The measurement system is shown in Figure 1.

Then, several series of measurements were carried out in the climatic chamber. After testing and analyzing the measurement data, the best and comparable sensors were selected.

MEASUREMENTS AND RESULTS

Data quality

Data quality, encompassing dimensions such as accuracy, completeness, consistency, and timeliness, is foundational to the reliability of any dataset. Measurement uncertainty, degradation of sensors, and environmental variables all contribute to potential inconsistencies. For instance, periodic calibration against recognized standards is critical to mitigate the inevitable wear and tear of sensors over time. Furthermore, contextual data, such as measurement conditions or external disturbances, must be integrated to ensure a robust interpretation of results. Addressing these dimensions systematically can transform raw data into actionable insights while minimizing errors that could skew analysis.

The term data quality can be used to describe the correctness of data, but also its usefulness for creating information. Data quality is a measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability and whether it's up to date.

Data quality can also be related to the measurement uncertainty, defined (or not) measurement conditions, the influence of the measurement on the tested system or the repeatability of the measuring sensor. To achieve repeatability of the measuring sensor, typical physical (physicochemical) characteristics should be taken into account. Depending on the measured value, it must be assumed that the sensor/probe degrades over time. Countermeasure to this is periodic control and calibration of readings with a recognized standard.

Data can also be descriptive values that aggregate, interpret or convert one quantity to another. An example can be colors, the basis is the parametric reflected wavelength, the name of the color is an interpretation in a certain measurement resolution. It all depends on their purpose of use.

Table 1. List of sensors used for the experiment

No.	Producer	Symbol	Measured quantities			Measurement range	Comments
			Temperature	Pressure	Humidity		
1	Texas instruments	TMP117MAIDRVT	+	-	-	-55–150 °C	Accuracy within ranges
2	Sensirion	SHTC3	+	-	+	-40–125 °C 0–100%	Measurement accuracy table
3	TE Connectivity	MS8607-02BA01	+	+	+	-40–85 °C 10–2000 mbar 0–100%	Accuracy at reference temperature
4	Bosch	BME280	+	+	+	-40–85 °C 300–1100 hpa 0–100%	
5	Infineon	DPS368XTSA1	+	+	-	-40–85 °C 300–1200 hpa	The accuracy of temperature measurement is low, but the sensor is dedicated to measuring pressure

Note: Temperature measurement is possible with all the sensors mentioned above, but not all of them are dedicated to this. Temperature is needed for internal calibration and internal correction of readings. The following sensors are dedicated to temperature measurement: TMP117MAIDRVT, SHTC3 and MS8607-02BA01. For comparison purposes, the measurement was performed from all sensors.

Figure 1. Mounting system with 3 set of sensorsin covers

In the case of preparing a measurement of any quantity, we must choose the appropriate measurement method (e.g. direct or indirect measurement) and the appropriate measuring sensor for this purpose. It should also be considered whether in the case of a given measurement it is not necessary to additionally compensate for some undesirable phenomena or to protect the moment of measurement against an incorrect state. It is also important to take into account the possible wear/aging of the measuring sensor and/or to establish its calibration intervals. At this point, we can talk about qualitative measurements. It is also possible to use quantitative measurements where statistical methods can be used to eliminate or at least reduce the number of measurements made in inappropriate conditions.

When considering the quality of data obtained from the calibration of a test device in a climatic chamber, several issues should be considered. The temperature of 25 °C and the humidity of 50% were assumed as calibration conditions. This is how the program of the climatic chamber was set. Inside the chamber, measurement sensors were placed on a stand, and data processing and acquisition modules were placed outside the chamber, as in Figure 1. The climatic chamber makes it possible to carry out long-term measurements in given environmental conditions. However, it should be remembered that the conditions in the chamber are controlled by automation, which also has its own control algorithms and mechanisms for maintaining these conditions. Through the measurements, it was possible to observe humidity fluctuations in a limited, but still visible, range (Figure 2). In the case of temperature measurements, the inertia in achieving the set ambient temperature and the temperature stabilization of the measuring module itself were also visible (Figure 3). Pressure measurements showed how sensitive electronic barometers are (Figure 4). After a series of measurements in the chamber, due to the above conditions, it will be necessary to take control measurements in natural conditions and check whether the calculated offsets are correct.

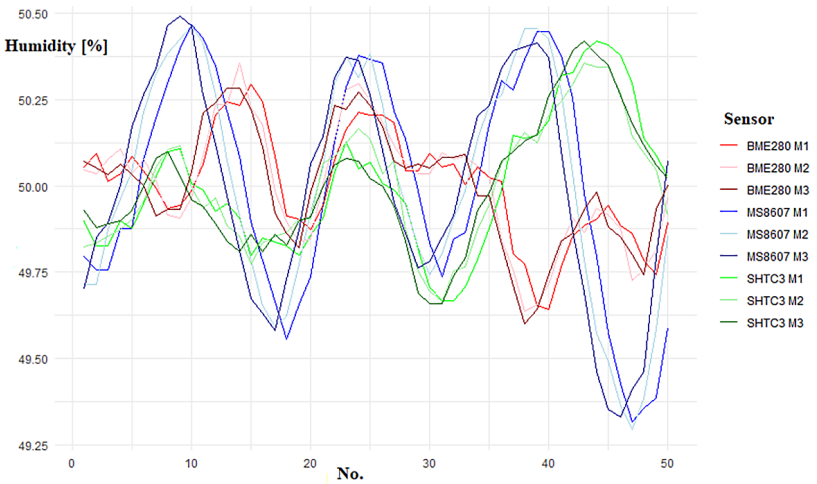

Figure 2. Graph of humidity readings, after applying the calculated offsets, for all 3 measurement modules with sensors

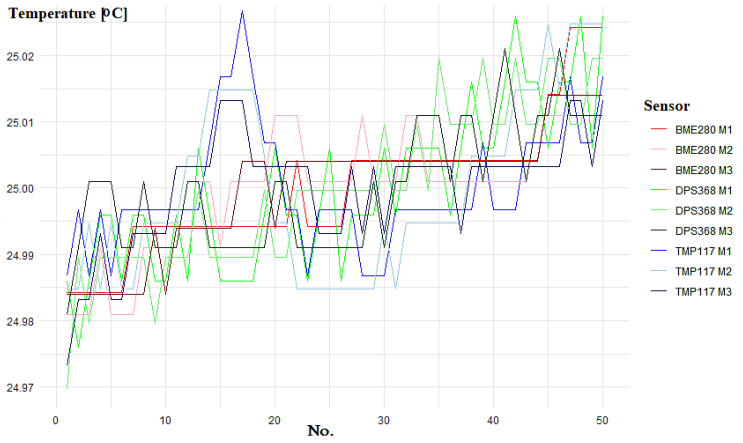

Figure 3. Graph of temperature readings, after applying the calculated offsets, for all 3 measurement modules with sensors

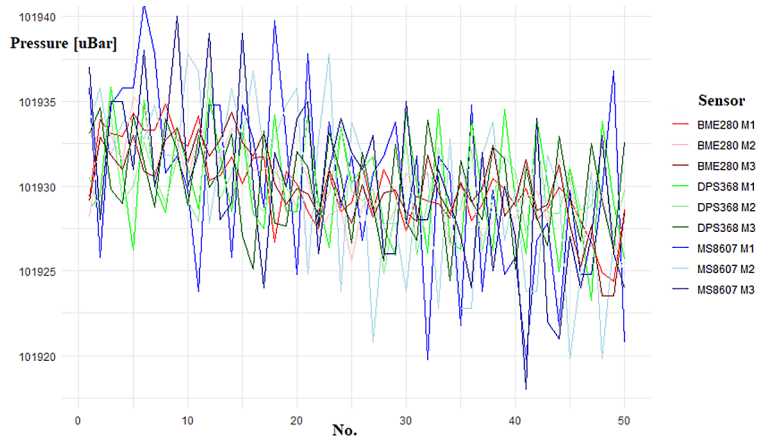

Figure 4. Graph of pressure readings, after applying the calculated offsets, for all 3 measurement modules with sensors

Single data and big data

Consider a single data, i.e. a value describing one of the characteristics of an object or phenomenon.

The object of measurement can be static or dynamic phenomena. The value of interest may be related to an antecedent phenomenon (e.g. phenomena following a collision of particles). Other times, the measurement of interest is the time between the occurrence of other quantities or a different pattern. As mentioned, it all depends on additional information about the measurement context itself and whether we can measure a given value directly or indirectly.

When we consider large amounts of data, their transmission and collection becomes an important issue. Each of these operations carries further potential sources of error. For example, uncorrected transmission errors, interruptions in transmission, unauthorized interference in transmissions, intentional data substitution are possible. Unauthorized interference, loss or substitution of data is also possible on the data collector side. Another category of threats is data leakage.

When considering single data points, it is essential to highlight their correlation with contextual variables. For instance, a single temperature measurement may have minimal interpretative value without accounting for factors such as environmental conditions, time of measurement, or location. A multilayered approach is necessary to link measurement data with metadata, enabling more precise analysis. Furthermore, in processing large datasets, anomaly detection methods play an increasingly significant role by eliminating erroneous or irrelevant data, thereby enhancing the reliability of analyses.

Under big data, we have (generally speaking) data in context, so not a single measurement, but a measurement with additional contextual data. Let's use an example of how the value (context) of a given measurement can be extended and how its meaning can change by providing additional information (data). Suppose a unit temperature measurement is 46.3 °C.

What does this measurement really tell us? Not much. We need to add context and additional data (constants or variables). As additional constant data, we can know what the measurement is about – let's assume that in this case we want to know about the ambient temperature; geographic location of measurement – let our measurement sensor be located on one of the buildings on the market square in Wroclaw, Poland.

With these two pieces of information above, we can determine some variability of ambient temperature over time over time, but even for statistical purposes, further important elements are missing. So let's add the date and time of the measurement as variable data. This procedure, it allows a wider view of the measurement. So let's assume that it is 6:00 p.m. on May 15, 2022. This is quite a lot of information, so we can already wonder if the previously mentioned value is within certain averages, or whether it is low or high compared to historical data.

Let's add another variable data: atmospheric pressure and humidity.

At this point it is possible to convert the temperature to that which occurs according to the definition of the International Standard Atmosphere (as defined in ISO 2533:1975). Then an absolute reference value must be obtained. Given the above measurement, it should be taken for granted that the value given is the correct, true value. Can we be certain of this, considering the points at which data degradation can occur, as described above?

Information on the calibration and/or inspection of the measuring device, information on data integrity, possibly on the occurrence of attacks at a given time, etc. should be added here. The above information is variable, often independent and may occur long or short term and relate to aspects of data transmission. However, physical "disturbances" can occur, so let's add two more pieces of information to the example. The sensor is installed on the south wall and that there was a fire in the building at the time mentioned. The above two pieces of information eliminate the entire measurement value of the sample measurement at this point. In the example above, we used a minimum of 12 measurement values to obtain one complete piece of information, which, additionally, at the very end cannot be used directly for a complete analysis of the previously mentioned measurement context, i.e. ambient temperature.

The above example also shows that for a measurement to have a specific value, it is not just one number, but in this case nine. So let's calculate how much data can be collected per day. Let's assume that we measure every 5 minutes from 20 locations. During the day (measurement every 5 minutes) it gives 288 entries per day, i.e. 4608 kB. With 20 locations, it already gives 92160 kB, which is approximately 92 MB of data per day. On a monthly basis, it is already 2.8 GB and in a year about 1TB of data. The above calculations do not consider any optimization and possible compression mechanisms. Looking at such amounts of data, the economic issue cannot be overlooked, because any data space is associated with costs. A 1TB HDD is about 90 EUR, an SSD is about 180 EUR (2025). Space in the cloud the costs are dependent on selected option. Whether such costs are acceptable or not depends directly on the economic analysis of the project and its short and long-term goals. It is possible that after such an analysis it is necessary to revise the measurement assumptions. It should also be remembered that the cost of data storage is only one of the costs. Processing systems and implementation of data processing and administration algorithms are still needed.

Having a set of data (dataset) often from a long period of time, it is possible to arrange certain patterns. Such patterns are defined in some domain, or it is possible to analyze in many domains. An example may be the daily temperature distribution in a given place and then comparing this distribution between different locations. However, determining certain regularities/repeatability always requires observation, and often the use of additional statistical, visualization and analytical tools. Comparative data is also sometimes used for analysis.

When analyzing data sets, it is necessary to pay attention to a certain repeatability (in context) from which certain patterns and repetitions can be arranged. There may also be anomalies, they are more difficult to find, the more that they may relate to different categories. Searching for

anomalies and their categorization can be used as datasets for machine learning.

Referring to the mentioned example, such an anomaly may be a temperature reading correlated with placing the sensor on a wall with direct sunlight, or a building fire.

Assuming two degenerating conditions occurring at the same time, it is necessary to consider whether such a set of data can be accepted as a pattern defining an anomaly in these circumstances.

In the case of cyclical data, it seems necessary to use anomaly detection mechanisms to ensure the appropriate quality of data and increase the reliability of these data. In many cases, this role is performed by a qualified and experienced data operator who knows the specifics of a given measurement field. Due to the increasing amount of data and the speed of measurements, the current trend is the use of automation. In order for automation to be implemented, it is necessary to build datasets corresponding to specific anomalies and valid data. Here we encounter another threat, which is categorization. It is possible to make an error by categorizing events incorrectly or categorizing them in the wrong context.

For applications in remote sensing (remote temperature measurement) as well as radiative transport, and modeling of environmental energy balances accurate knowledge of surface emissivity is essential.

When there is considerable background radiation at the same wavelength as the emitted radiation direct measurements of surface emissivity are difficult. This occurs, for example, when objects at temperatures near room temperature are measured in a terrestrial environment by use of the infrared 8–14 μm band.

This problem is usually treated by assumption of a perfectly diffuse surface or of diffuse background radiation. However, real surfaces and actual background radiation are not diffuse; therefore, there will be a systematic measurement error. In article [5] presented analysis of and several cases in which the surface properties and the background radiation are not perfectly diffuse. The resulting errors in the measured emissivity can be large: the magnitude and distribution of error depend on the individual case, and in some cases errors as large as 0.10–0.20 have been found

An important aspect is also matching the measuring equipment to operate at a certain height. In the dissertation [6] voltage dependence of the neutron-induced soft error rate (SER) was investigated for a static memory. In conclusion of dissertation, neutron-induced SER increases approximately linearly with decreasing technology feature size. SER also increases with height, i.e. the number of neurons. For example, a 32 Mbit static memory implemented in a 0.1 μm process will fail on the average each 5.7 years at sea level, or each 20 days at airplane flight altitudes. The same trend should apply also to other logic circuits, e.g. flip-flops, latches, registers, but also combinational circuits.

An interesting example describing the previously mentioned places of possible data degradation is the example of radiation measurements performed by grassroots social movements after the nuclear power plant accident in Fukushima. Various radiation meters were used and measurements were made in an unsystematic manner. The challenge was to analyze and process such a large amount of data and to present it in a useful way [7].

The above examples show the identified mechanisms, places and types of possible degradation or interference with data. This is not a closed list and with the development, popularity or the need to use a large amount of data from various sources, it will also be necessary to use new anomaly detection mechanisms. In addition to the previously mentioned manual and statistical methods, the natural stage is the use of ML and later AI mechanisms. The development of samples of correct and incorrect data will still rest on man and his intelligence and natural skills of perceiving patterns. Developing insights and converting them into usable mechanisms for machine use in the case of various data can be time-consuming and describe many variants. For measurements of environmental quantities, the Polish Institute of Meteorology and Water Management has published an extensive study based on data from a section of the research network [8] (in Polish). It can be assumed that the use of ML algorithms will enable much faster analysis of large amounts of data and data in a broad context by AI.

With respect to the research conducted, we can summarize where we can expect possible data degradation. These are:

- measurement sensor system – starting from the design and the sensor used and software and method of processing measurement data;
- data transmission devices – data transmission via wire or wireless including data conversion and possible attacks on the protocol or device or server;

- data aggregators – databases, information brokers, conversion and recording functions, possible attack on the database;
- memory systems – damage and/or interference with memory media;
- data presentation mechanisms – another stage of processing where possible conversion errors, rounding errors, etc. may occur.

In practice, educating IoT system operators and administrators about data management and potential threats is paramount. Many incidents can be avoided by implementing straightforward procedures, such as conducting regular security audits and employing basic data protection principles, including encryption of transmitted information. Moreover, automating these processes through the deployment of machine learning algorithms for real-time anomaly detection is becoming increasingly effective and widespread. For instance, analyzing telemetry data streams can trigger alerts upon detecting deviations from expected norms, enabling rapid corrective action.

PURPOSE OF USING THE DATA

When consider the purpose of using the data, we can consider the potential threat of the conscious or unconscious use of too poor-quality data to analyze a phenomenon that requires precise data. Determining the quality of the data needed is another issue that may pose a potential threat. The decisions made may not only concern the individual, but also the ever-wider reality. At this point, the question of the possible deliberate use of data to exert a certain influence should be raised. Quantitative data can be used, for example, to interfere with behavior, pressure to decide, or generally speaking as a source for manipulation techniques. This is a relatively new field included in the definition of cybersecurity.

In our device, we send data about temperature, humidity, and pressure from reading devices to the server, so an important issue to consider is the vulnerability of protocols to attacks and ways of interfering with the data. We have presented below the most important issues related to these problems for the most popular network protocols.

In the context of IoT, the role of data extends beyond its basic utility, opening avenues for its use as a tool for manipulation. For example, while analyzing weather data trends can help predict future events, it could also be misused to fabricate misleading narratives about climate change. This underscores the necessity of audit mechanisms that verify data accuracy and trace its origin. The integration of blockchain technology into data processing systems can provide an additional layer of security and transparency, particularly in critical applications.

MQTT protocol vulnerabilities

The MQTT protocol, widely adopted in IoT systems, exhibits vulnerabilities that could compromise data integrity and confidentiality. Key weaknesses include the lack of mandatory encryption and insufficient mechanisms for authentication and authorization. Such gaps render systems susceptible to attacks, including data interception (Man-In-The-Middle), unauthorized data access (Topic Snooping), and data manipulation. This paper explores these vulnerabilities and proposes countermeasures, emphasizing the necessity of adopting version 5.0 or later, which incorporates enhanced security features. Additionally, custom security layers tailored to specific applications may provide effective protection against sophisticated threats.

Researchers conducted a study of the popular MQTT protocol and discovered 33 security vulnerabilities, highlighting the need for stronger cybersecurity measures in IoT devices using this messaging protocol. The article describes 33 possible vulnerabilities of the MQTT protocol [9].

The most significant types of vulnerabilities and attacks include:

- Sending data without authentication and authorization – the MQTT protocol allows you to send data to the broker without the need for data authentication, this is a specific type of configuration, but it is possible.
- Unenabled data encryption – MQTT allows data to be transferred openly without the need for encryption – this allows data to be obtained by third parties.
- Man In The Middle attacks – the lack of encryption and authorization enabled greatly facilitates this type of attacks. Additionally, the lack of control mechanisms in the protocol may significantly enable data interception and modification.
- Lack of protection of the broker/server against DoS attacks [10] – if additional security measures are not applied on the server side, the

broker itself is susceptible to overload attacks, as are the devices themselves (often due to limited hardware resources, including the basic implementation of the TCP/IP stack) are not resistant to attacks of this type. This may lead to temporary loss of data from devices.
- Privacy violations – due to the lack of mechanisms for controlling access to topics, it may lead to unauthorized persons subscribing to data to which they should not have access. This type of attack is called Topic Snooping.
- Information injection attack – after the attacker knows the topics and structure of the message, if no security mechanisms are implemented – this type of attack is very simple and allows for significant interference with data.
- Replay attack – in the absence of security, it is possible to intercept and send the same message to the broker, which is quite a significant interference with data. Due to the popularity of IoT devices and the use of the MQTT protocol in them, the number of vulnerabilities and attacks is systematically increasing.

Due to this, the protocol itself is also being developed mainly in terms of security. It is important to remember that when it comes to IoT devices, there are 2 sides that are vulnerable to attacks. The device itself, which in most cases is a device based on a microcontroller and a module GSM/Wi-Fi/LoraWAN/SigFox etc. where limited hardware resources do not allow the use of more complex filtering and protection methods against network attacks. The server side also requires appropriate security measures and correct implementation of mechanisms in the protocol itself. It is important to highlight the development of the protocol itself. The first practically used version 3.1 (released in 2010) included a basic implementation. In 2014, corrections were introduced and the version number was raised to 3.1.1. It should be noted that this version (despite security shortcomings) is still widely used in many IoT systems (including home automation) due to the popularity of implementations in many OpenSource projects. Only since version 5.0 (2019) have stronger mechanisms for authorization, authentication, session handling, etc. been implemented.

CoAP protocol vulnerabilities

The CoAP protocol may be vulnerable to threats similar to MQTT and additional ones threats resulting from similarities to the HTTP protocol. These are:
- XSS attacks, i.e. the possibility of executing a script on the server side; – reflective injections – sending code that is processed by server and displayed to the user;
- DOM attacks – injecting code directly into the page object.

The origins of the CoAP protocol date back to 2009. The first version (until 2014) did not have any security mechanisms implemented. CoAP v2 with encryption and authentication mechanisms has become a much more secure protocol.

Scenario of an attack on an MQTT broker (mosquito):
- Port scanning by NMAP to determine on which ports the broker is running.
- In the case of the default configuration or using the Man in the Middle technique to take over the username and password, we can easily display what messages are being sent (using the mosquitto_sub and mosquitto_pub commands).
- After connecting to the broker (e.g. using MQTT Explorer), we can listen to the data.
- Interaction with devices – if the devices can execute commands, manipulation becomes simple.
- Additionally, you can use the “Fuzing” technique, i.e. sending any data to the broker and observing what effects it will have.
- DoS (Denial of service) attack scenario on the CoAP server:
- Reconnaissance – scanning a range of ports to see which of them the CoAP server is listening on.
- Use the script to generate query packets to the CoAP server; to increase the attack potential, it should be run from multiple machines.
- Generating an additional stream (PING flood) of requests from the server.
- Possible use of other tools to generate CoAP packages.

Economic and physical aspects of performing measurements and related risks to data

The economic approach to measurements is manifested in estimating the costs of the entire measurement system in relation to a single measurement and its significance.

There are a number of sensors available on the market, from very poor quality, the use of which may be limited only to detection (is/is not), to ultra-accurate, suitable for high-precision laboratory measurements. In many cases, the choice of a solution must be economically and technically justified. Economic balancing in this place is the second of the risks – in terms of measurement, usually the cheaper sensor, the less accurate it is.

Standards are used to maintain the quality of a given measurement category. The standards indicate under what conditions a given measurement should be carried out. In addition to standards that are not developed for all possible measurements, especially when they occur in specific industries or conditions or are performed for the purposes of, for example, a scientific experiment, a set of good practices is often used, which includes recommendations or in specific conditions the method of measurement is selected in experimental form. The experimental form has the feature that it is sometimes modified and adapted during the research.

When consider a single measurement of any physical quantity in a natural environment with regard to the information obtained from the measurement. The very nature of a single measurement returns us a value at one specific point in time. Due to the fact that measurements are made in the natural environment, it becomes possible that the measured value of a given object/part of an object may be in a different state at a given moment than we assumed in the conditions of measurement. Therefore, we may get a result that we do not expect. In a given state the result may be completely correct or it could be completely incorrect due to the wrong measurement method in relation to the given state of the object. At this point, it becomes important to first link the measurement data with additional information, which is the determination of the measurement conditions and the determination of the range of expected results.

At this point, we already encounter the first degradation of the data, but this degradation does not necessarily mean an error or inaccuracy, which in turn will eliminate the measurement value. On the contrary, at this stage we can adjust the measurement quality to the needs. At this point, we also encounter the first threat, i.e. possible, conscious or not, non-adjustment of the measurement to the expected precision of the data for the analysis of a specific phenomenon.

During the operation of the measuring device, we can encounter another threat: damage to the measuring sensors. The damage can be unambiguous (indications differ diametrically or are always the same) or latent (e.g. constant under- or over-indication). In some cases, the detection of such a failure may be significantly impeded and its discovery may take place only during the planned maintenance and calibration works. It is then necessary to decide what to do with invalid data.

A single data item does not carry much value. Additional data is required to describe the conditions, nature or location of the measurements. Without them, data interpretation may be incorrect, and this is another threat.

When we consider measurements other than point measurements, we see further threats. In the case of distributed measurements, carried out in a certain area, sometimes and all over the globe, we should take into account factors directly related to the phenomenon of scale and a new area, i.e. the issues of data transmission.

In the case of the scale effect, we have to take into account possible measurement deviations between the sensors, a certain dispersion of performance parameters, random, statistically higher failure rate. If it is necessary to mount the sensors in a special way – repeatability of the assembly.

The indicated problems are partly related to economics – the expenditure of financial and personal resources for the needs of one measuring station.

In the case of popular term: the Internet of Things , or generally speaking: things connected to the Internet, equipped with dedicated or additional sensors, we are dealing with the overlapping of threats mainly due to economics. For mass or large-scale measurements, you need the right amount of equipment and the provision of appropriate infrastructure and management of the measurement project. Ultimately, it is possible to calculate the cost of a single measurement. The higher the cost related to the measuring device itself and its operation, the higher the cost of a single measurement. Depending on the type of project and its scale of importance, the result may be degraded by the reluctance to incur costs or the desire to save.

In the case of distributed measurements, the threat of data degradation may occur due to the selection of sensors (their quality in relation to costs), the effect of scale and the summation of inaccuracies, threats from the data transmission side.

A separate category of threats to data integrity is data transmission. As data transmission, we understand any method of data transmission, be it analog (e.g. as voltage states) or in digital form, i.e. properly encoded. Among the threats in data transmission, we can distinguish: interruptions in transmission, jamming, eavesdropping, data substitution. The listed threats apply to every available transmission medium and every technology (cable, fiber optic, radio). In addition to physical interference, there may be threats in the higher layer, i.e. at the level of the network protocol used.

Popular telemetry data transmission protocols are currently characterized by very simple and therefore insufficient security (MQTT, CoAP, DDS). An example of an attack on the MQTT protocol is described in detail in [11]. The mechanisms described are not too different from any other type of attack on critical systems. In many cases, measurement data is also critical data and requires security at an equally high level, and this requires the development of protocols that ensure speed, reliability and data security.

Measurement results, especially in large quantities, must be stored or processed as data streams, and after processing, stored. This creates further points of possible data disintegration. The first element may be the problem of data types and their subsequent conversions [12]. This type of obstruction can occur many times when we adopt additional mechanisms for updating and inserting/rewriting data in streams. For example, it can be MQTT Broker, Hive, Apache Hadoop etc.

Problems can also be more prosaic, such as data corruption due to storage failure, data loss due to incorrect operation and/or configuration of a given service. The second category may be conscious and intentional attacks on data centers in order to steal, replace or delete them (these types of attacks are practically everyday in banking systems). While data loss is easily identifiable and often relatively easy to restore (backups, etc.), data manipulation is more difficult to detect and requires at least one source of verification (trusted) of the stored data, which also involves the need to maintain the entire system integrity checks. In the case of the need to store more and more data, this causes performance problems in terms of processing speed and data transmission (network and internal for mass storage).

A separate category of threat in data storage systems is their conscious or unconscious manipulation. Unconscious manipulation can take place during any operations related to optimization, regeneration and data recovery. A possible error in the code/script may cause an unintended effect and the result may not be detected. Deliberate tampering with data can take place after an attack and unauthorized entry into the system, or because of internal sabotage. It should be mentioned that detecting sabotage from the inside is often much more difficult to detect, and the consequences can be more dangerous and long-term.

CONCLUSIONS

This research highlights a comprehensive spectrum of risks associated with data acquisition, transmission, and processing in distributed IoT systems. By identifying at least 12 critical points of potential data degradation, it underscores the necessity of adopting robust methodologies and security protocols to safeguard data integrity.

The general conclusions from this work are presented below:

- Data integrity in IoT systems is vulnerable at multiple stages—from sensing to transmission to storage.
- Selection of sensors must consider environmental conditions, degradation over time, and economic constraints.
- Protocol-level vulnerabilities (e.g., MQTT, CoAP) necessitate additional custom security layers in IoT architectures.
- Systematic calibration, anomaly detection algorithms, and redundancy checks are essential in ensuring data quality.

REFERENCES

1. Sagan C., Salzman Sagan L., Drake F. A message from Earth. Science. 1972; 175(4024): 881–4. https://doi.org/10.1126/science.175.4024.881
2. Ren J., Dubois D. J., Choffnes D., Mandalari A. M., Kolcun R., Haddadi H. Information Exposure for Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach. In: Proceedings of the Internet Measurement Conference; 2019. https://doi.org/10.1145/3355369.335557
3. Bella G., Biondi P., Bognanni S., Esposito S. PET-IoT: Penetration testing the internet of things. Internet of Things 2023; 22: 100707. https://doi.org/10.1016/j.iot.2023.100707
4. Kombate Y., Hounge P., Ouya S. Securing MQTT: unveiling vulnerabilities and innovating cyber range

solutions. Procedia Comput. Sci. 2024; 241: 69–76. https://doi.org/10.1016/j.procs.2024.08.012

5. Kribus A., Vishnevetsky I, Rotenberg E., Yakir D. Systematic errors in the measurement of emissivity caused by directional effects. Appl Opt. 2003; 42(10): 1839–46. https://doi.org/10.1364/AO.42.001839
6. Hazucha P. Background radiation and soft errors in CMOS circuits [dissertation]. Linköping University; 2000.
7. Hultquist C., Oravecz Z., Cervone G. A Bayesian approach to estimate the spatial distribution of crowdsourced radiation measurements around Fukushima. ISPRS Int J Geo-Inf. 2021; 10(12): 822. https://doi.org/10.3390/ijgi10120822
8. Instytut Meteorologii i Gospodarki Wodnej. Monitoring wiarygodności pomiarów temperatury [in Polish]. IMGW; 2019 [cited: 2025 Apr 7]. https://www.imgw.pl/sites/default/files/2019-12/monitoring-wiarygodnosci-pomiarow-temperatury
9. Mitchell R. 33 Critical Vulnerabilities Found in Popular IoT Protocol MQTT. Electropages. 2022 Feb [cited: 2025 Apr 7]. https://www.electropages.com/blog/2022/02/researchers-find-mqtt-have-33-vulnerabilities
10. Paolone M., Tamburri D.A. Serverless computing: a systematic mapping study. ACM Comput Surv. 2021; 54(8): 1–32. https://doi.org/10.1145/3465481.3470049
11. Syaiful A., a, Rahardjo B., Hanindhito B. Attack Scenarios and Security Analysis of MQTT Communication Protocol in IoT System. In: Conference: 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI); 2017. https://doi.org/10.1109/EECSI.2017.8239179
12. Zhang G, Mariano B, Shen X, Dillig I. Automated translation of functional big data queries to SQL. Proc ACM Program Lang. 2023; 7(OOPSLA1): 580–608. https://doi.org/10.1145/3622825