Big data Quality Assessment in the IoT era

Chapter · November 2024

DOI: 10.1016/B978-0-443-21640-4.00003-X

Ikbal Taleb
Concordia University

Nadia Dahmani
Zayed University

Sujith Samuel Mathew
Zayed University

Big data Quality Assessment in the IoT era

I. Taleb {ikbal.taleb@zu.ac.ae}, **N. Dahmani** {Nadia.dahmani@zu.ac.ae},
College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates
S.S. Mathew{Sujith.Mathew@zu.ac.ae}, **K. Hayawi** {abdul.hayawi@zu.ac.ae}
College of Interdisciplinary Studies, Zayed University, Abu Dhabi, United Arab Emirates

Abstract— The Internet of Things (IoT) has shown unprecedented data generation and connectivity through disruptive applications across various domains. This data, referred to as big data, contains valuable information about various aspects of our lives and the physical world. As the IoT ecosystem continues to expand, ensuring the quality of the vast volumes of data it produces has become crucial. Therefore, big data quality assessment is important in the IoT era to ensure that the data collected and analyzed is accurate, trustworthy, and usable. By performing big data quality assessments, organizations can make better decisions, improve their operations, and gain a competitive advantage. This paper aims to enhance data quality assessment in IoT by providing an overview of its state-of-the-art. Big data and IoT data properties and their new lifecycles are discussed. Moreover, selected works from the literature on IoT data quality are exhaustively reviewed. Additionally, a holistic architecture of IoT quality management model is proposed capturing key characteristics of IoT applications. Finally, open challenges and possible future research directions are discussed.

Keywords— IoT, Big Data, Big Data Quality, Data Quality Assessment, Edge Computing.

I. Introduction

In the modern data-driven world, everything is connected. Since IoT refers to a network of connected devices and sensors that collect and exchange high amounts of data, this data is referred to as big data (Extremely large data sets with various data types and formats) that contains valuable information about various aspects from which it has been collected. Therefore, the quality of such data is subject to various factors, such as data collection and transmission errors, the heterogeneity of devices and sensors, and the dynamic nature of the IoT environment (Ann and Wagh, 2019; Mansouri et al., 2023; Singh et al., 2023). Big data quality plays a paramount role in the IoT era, especially to guarantee that the data collected and analyzed is accurate, reliable, and usable. Data quality assessment is particularly important for IoT systems because they often involve complex interactions between multiple devices and networks. A failure or vulnerability in one device can have a cascading effect on the entire system, potentially leading to security breaches or operational disruptions. In addition, IoT devices are often deployed in critical applications such as healthcare, transportation, and energy management, where reliability and safety are crucial.

IoT Quality assessments are complex processes and challenging tasks that involve multiple layers of hardware, software, and network infrastructure, as well as the interactions between them due to the large number, and heterogeneity of devices and the complex interactions between them (Byabazaire et al., 2022, 2023a; Martín et al., 2023a). Furthermore, the quality processes should be performed at various stages of the big data analytics lifecycle, including data acquisition/collection, transmission, storage, processing, and analysis. This complexity requires the use of a framework, model, or architecture to guide the quality assessment process. Such a framework or model can provide a systematic way to identify and evaluate the key quality attributes of IoT systems, from reliability, security, performance, scalability, and usability, and must include quality attributes, and metrics relevant to IoT systems, Assessment methods (simulation, monitoring, and analysis), and tools and techniques (simulators, emulators, monitoring tools, and analysis tools).

Some examples of quality frameworks and architectures such as IoT-A Architecture Reference Model (Fortino et al., 2022), and Open Platform for IoT (OPI) Architecture (Hansch et al., 2019; Pang et al., 2013; Shin et al., 2022) are available. IoT-A provides a comprehensive framework for the interoperability and guidance across different IoT systems. It aims to create a shared understanding of IoT concepts to ensure compatibility and scalability, addressing the heterogeneity challenge. OPI Architecture, on the other hand, focuses on open platforms that enable the integration of diverse IoT devices and applications, promoting an ecosystem where data can be reliably and securely shared across platforms, addressing security and reliability challenges in dynamic IoT environments. These frameworks tackle the core issues of IoT quality assessment by providing structured approaches to ensure that IoT systems are interoperable, scalable, secure, and reliable. These frameworks, models, or architectures can help identify and evaluate the key quality attributes of IoT systems and guide the decision-making process in IoT system development and deployment.

While quality assessment can be tackled at the back-end or at the edge, it is essential to prioritize addressing it at the edge. By processing data locally, closer to the source of the data, organizations can improve efficiency, reduce latency, and identify potential issues in real-time, leading to more reliable, secure, and efficient IoT deployments. This is known as edge computing which provides a different model than the big data value chain lifecycle stages in data analytics(Casado-Vara et al., 2018; Song and Zhang, 2020). This completely changed the way quality assessment is performed as it will happen close to the IoT devices, which highlights the importance of IoT-specific quality characteristics.

Since the IoT devices or edge gateways will perform some processing on the data before it is transmitted to a central repository, this can have significant benefits like the reduction of the amount of data sent over the network, which can help minimize data loss or corruption during transmission and improve accuracy and relevance. Moreover, edge computing can also improve the scalability and reliability of the IoT system, as it allows distributed processing and storage across multiple devices. (Casado-Vara et al., 2018; Truong and Karan, 2018; Yu et al., 2018)

In this chapter, we tackle data quality assessment in the IoT era from a different angle and consider both quality assessment at the back-end (e.g., cloud) and at the Edge with a focus on the processes of Quality assessment. The remainder of this chapter is organized as follows: Section II introduces the key concepts of quality assessment, processes, and dimensions. Section III presents a comprehensive classification of quality assessment processes in the IoT domain. Section IV proposes a quality assessment model that captures key characteristics of IoT applications. Section V identifies the main challenges in assessing quality in the IoT and draws some research directions in this research area.

II. Background

As IoT becomes more pervasive, it is likely that big data capabilities will undergo a revolution in the coming years, with enhanced domain sensing capabilities. Nevertheless, many IoT -related projects are hampered by real-time connectivity issues, insufficient computing power for the level of information processing, and data transport capabilities, necessitating the execution of complex data analysis on heterogeneous computing platforms. Due to these constraints, the outcomes of data processing, such as aggregations and approximations, are susceptible to information loss, which ultimately impacts data accuracy (Klein and Lehner, 2009). Additionally, environmental factors, reduced sensor precision, communication latencies, short battery life, a limited number of available sensor/actuator sets, data breaches, etc. lead to reduced data accuracy.

Data quality (DQ) issues must be carefully addressed at the source and effectively managed before they can affect the decision-making process in order to preserve the integrity of the generated data and support the highest standards of decision-making.

Gartner (M. Chien and A. Jain, 2021) estimates, 60% of organizations will utilize machine-learning-enabled DQ technology by 2022 to reduce human operations for data quality improvement, and 50% will use data quality solutions by 2024 to promote digital business. DQ is critical for tracking data value and relevance, and we believe its use in quantifying data will give us a handle on what data is available, what its value might be for business decision-making, and it should be assessed primarily during the data transformations at the pre-processing and processing stages of the data. The accuracy of a classification model is heavily dependent on the DQ, so measuring DQ (Bello et al., 2021) is critical for estimating task complexity earlier. DQ attributes should be verified, improved, and regulated throughout their life cycle, as they have a direct impact on the conclusions drawn from data analysis. To capture the quality requirements, characteristics, dimensions, scores, and applications of quality rules, data profiling (Loshin, 2010; Taleb et al., 2021)has become a popular approach. DQ assurance and this approach have become so intertwined that they are often referred to as the same. It is a collection of techniques used to facilitate a variety of data management tasks, such as data quality evaluation and metadata management.

A. Big Data

An exponential rise in global inter-network activity and data storage has sparked the beginning of the Big Data (BD) Era. Popular applications like Facebook, Amazon, Twitter, YouTube, IoT sensors, and mobile devices are prominent players and data generators in this market. This ecosystem as a whole produces an astounding 2.5 quintillion bytes of data per day, (equivalent to 2.5 Exabytes, where 1 EB equals 10^18 Bytes).

IBM characterizes BD as an information asset distinguished by its high volume, high velocity, and high variety, necessitating cost-effective and innovative approaches to information processing to glean enhanced insights and inform decision-making. It serves as an umbrella term encompassing both structured and unstructured data, rendering traditional database and software tools ill-suited for efficient processing. Furthermore, BD encompasses the requisite technologies and storage infrastructure required by organizations to effectively manage large data repositories. (Taleb et al., 2021)

Many significant architectural frameworks for BD systems have been presented in (Ramaswamy et al., 2013). Within the realm of BD, the data ecosystem is enriched by diverse sources, including (1) heterogeneous data origins such as e-Government datasets like Census data, social networking platforms like Facebook, and web-derived data like Google page rank statistics. Additionally, (2) this data further exhibits a myriad of formats, spanning from video to text, and (3) varying manifestations, encompassing unstructured data like raw text without predefined schemas, as well as semi-structured data incorporating metadata and text-based graph structures.

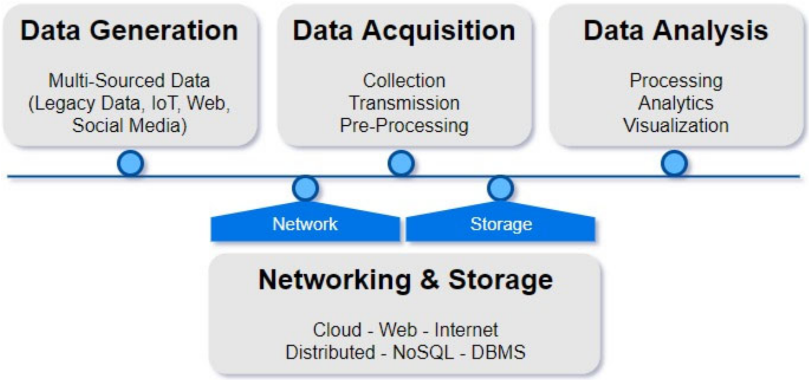

Fig. 1. Big Data Lifecycle

The BD lifecycle value chain includes four main phases that can be described as follows ((see Fig. 1):

Data Generation: This initial phase revolves around the creation of data, with diverse sources contributing to its generation. These sources encompass a wide spectrum, ranging from electrophysiology signals and climate-monitoring sensors to surveillance devices, social media posts, videos, images, transaction records, stock market indices, GPS location data, and more.

Data Acquisition: This stage comprises the comprehensive process of data collection, transmission, and preprocessing (AlNuaimi et al., 2019; Hu et al., 2014). In the backdrop of an exponential surge in data production from heterogeneous sources, a staggering volume of structured, semi-structured, and unstructured data becomes available. As a result, traditional data preparation processes like data integration, enhancement, transformation, reduction, discretization, and purification are included in the field of BD preprocessing.

Data Analysis (Processing, Analytics, and Visualization): This critical phase is characterized by the application of data mining and machine learning algorithms to process data and unearth valuable insights that inform superior decision-making. Data scientists emerge as pivotal users of this phase, harnessing its capabilities to derive actionable intelligence from the vast sea of data.

Data Storage & Transmission: Within this phase, data finds its abode in the intricate infrastructure of data centers through high bandwidth networks, disseminated across multiple clusters and geographically dispersed data facilities worldwide. The software storage component is underpinned by the robust Hadoop ecosystem, ensuring a requisite level of fault tolerance, storage reliability, and efficiency through data replication. This stage shoulders the responsibility of managing all input and output data coursing through the entire data lifecycle.

B. IoT - Internet of Things

The IoT represents a vast network of millions of interconnected objects distributed worldwide. These objects continuously communicate and exchange data through their sensors, generating an immense volume of data every second. IoT stands as a transformative evolution of the Internet, as outlined by (Karkouch et al., 2016), and its definition varies depending on the perspective adopted. One perspective focuses on how IoT redefines roles in the data ecosystem. In this view, interconnected smart devices play a major role as the primary data producers and consumers, eclipsing human involvement. This paradigm shift entails a flow of data from the physical world into the digital realm, enabling computers to gain a heightened awareness of their surroundings and the ability to act on behalf of humans through ubiquitous services.

IoT 's impact is far-reaching and touches upon numerous facets of daily life, both personal and business-related. It permeates various domains, including cities, homes, healthcare, and more. Moreover, it has assumed a significant role in society, emerging as a symbol of power and influence (Karkouch et al., 2018). (Gubbi et al., 2013) present a taxonomy of IoT applications based on factors such as network availability, coverage, scale, heterogeneity, repeatability, user involvement, and impact. This taxonomy identifies four distinct application domains: Home and Personal, Enterprise, Utilities, and Mobile. Within the IoT vision, applications that bridge the physical and cyber worlds have already been realized and continue to proliferate, encompassing areas like healthcare, home energy monitoring, smart cities, intelligent products, and more, as observed by (Aggarwal, 2016) and (Kiritsis, 2011).

C. IoT Data at the Edge

Data holds a significant role within the IoT paradigm, serving both as a wellspring for extracting insights and as a medium for communication. It bridges the gap between the cyber and physical worlds. Its significance becomes evident with the emergence of the IoT semantic-oriented vision (Atzori et al., 2010). This vision derives its relevance from the necessity to establish methods for representing and handling the vast volumes of raw data projected to be generated by "things" in the IoT ecosystem. Notably, this data is autonomously and continuously collected by these "things," such as RFID readers and sensor nodes, often surpassing the capacity of manually input data.

Before cloud technology, the data lifecycle followed a straightforward and circular pattern. During this period, data was more structured, had lower diversity, and moved through a limited number of channels towards a handful of designated destinations. Each phase within the traditional data lifecycle was conceptually uncomplicated consisting of planning, acquiring, processing, analyzing, integrating, and storing. With cloud technology, now possibilities for connecting the “thing” rose. The IoT data lifecycle is divided into three components: the edge, where IoT collects data; the cloud which is the front end where the processes and communications happen, and the back end, where data is stored subject to mining and analysis (see Fig. 2). In this new architecture, rather than sending data generated by sensors and devices to a distant data center, edge computing processes it efficiently and in close proximity to its point of origin. This approach involves the local collection and processing of substantial data volumes by edge devices, which retain critical information. Edge computing is strategically positioned closer to end-users, ensuring the delivery of Quality of Service (QoS) enhancements to them. These edge computing nodes are often referred to as edge or cloudlet servers. They offer several advantages, including cost reduction, real-time analysis capabilities, alleviation of network congestion, and overall performance improvements for various applications (Ann and Wagh, 2019).

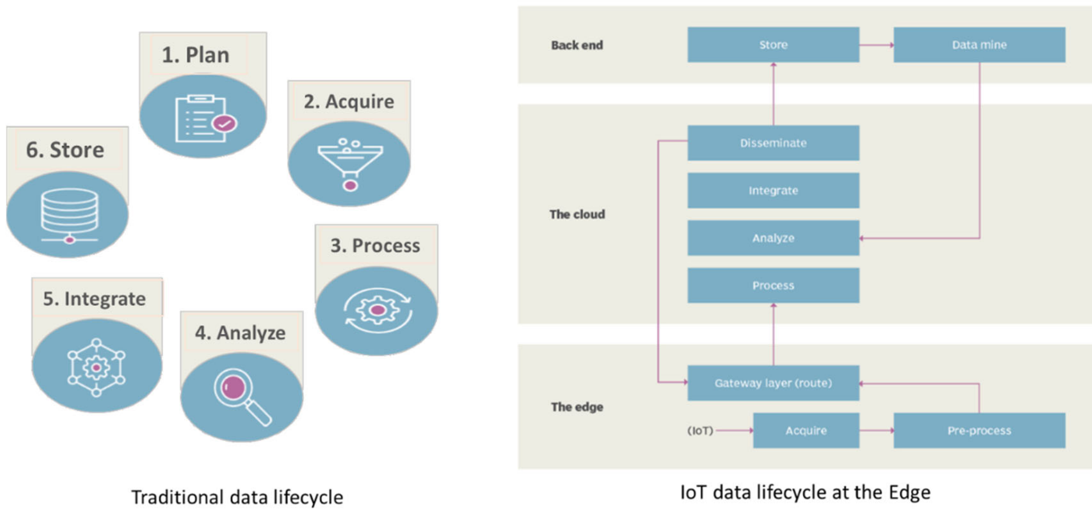

Fig. 2. Data lifecycle on the Internet and IoT (Scott, 2020)

IoT sensors typically observe a specific variable in the physical world, such as temperature or sleep patterns. Additionally, the environments where data is collected are marked by rapid and unpredictable changes. Consequently, IoT data often exhibits a range of characteristics (Karkouch et al., 2016). Some of these traits can be considered universally present, including uncertainty, errors, noise, distribution, and large volumes. On the other hand, certain characteristics are context-dependent and closely tied to the observed phenomena, such as smooth variations, continuity, correlations, periodic patterns, and Markovian behavior.

D. Data Quality: Dimensions and Metrics

Ensuring data quality emerges as a crucial requirement, especially for data consumers in IoT, such as pervasive services and their users. Numerous studies (Berti-Équille, 2007) (Hand, 2007) (Hipp et al., 2007) emphasize the important role of data quality (DQ) in data mining processes. They highlight how low DQ can severely compromise the validity of the outcomes and interpretations derived from these processes. Consequently, the consensus is that ensuring DQ, and accuracy is imperative.

An entirely new set of metrics incorporating "data weights" has been proposed by the authors in (Vaziri et al., 2019). Choosing a set of dimensions to work with is an important part of the approach, and to measure each dimension, a metric must be selected. To accomplish continuous improvement, TDQM (Total data quality methodology) (Wang, 1998) is one of the few approaches that operate in a cyclical or revolutionary fashion. When it comes to metrics, the only thing it uses is "basic percentages," such as the percent of missing data for the "completeness" dimension. DQA combines subjective and objective evaluations. The root causes of discrepancies can give great insight into the DQ problems and guide data quality improvement efforts. According to (Heinrich et al., 2018b) five requirements of data quality metrics are: (1) the existence of minimum and maximum metric values, (2) interval scaling of the metric values, (3) quality of the configuration parameters, (4) the determination of metric values, (5) aggregated metric values, and (6) economic efficiency of the metric values.

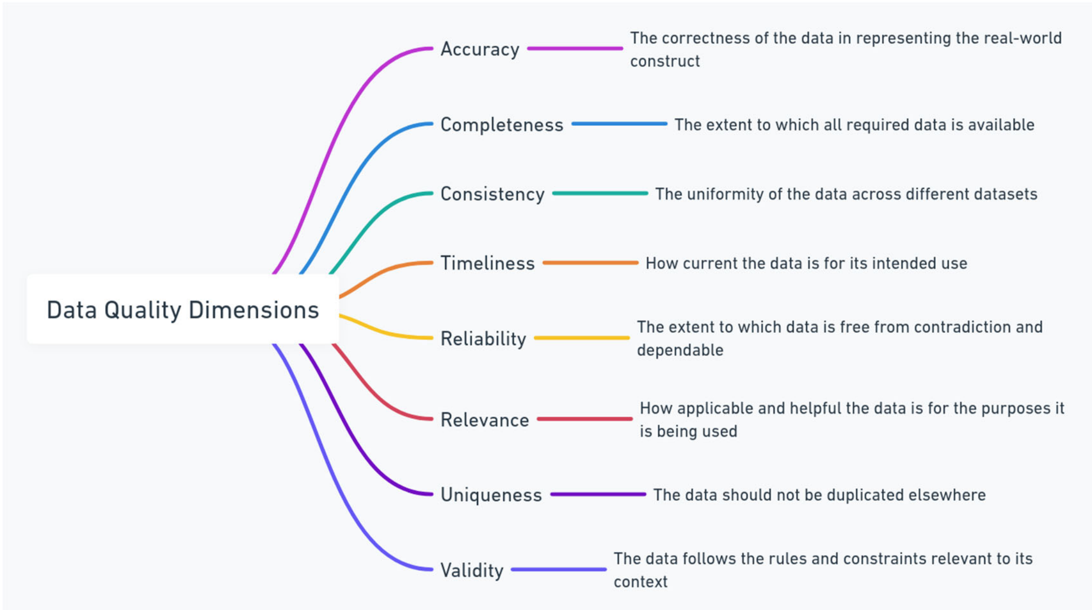

Fig 3. Data Quality Dimensions

DQ dimensions

The term dimension is used to describe elements of data that can be measured, as well as the ways in which data quality can be evaluated and quantified. Data accuracy, completeness, uniqueness, timeliness, and validity are the six primary criteria that determine the quality of data (Batini and Scannapieco, 2016a). The following are the accepted definitions of each of these metrics (Fig. 3.)

- **Accuracy:** The degree to which data correctly portrays the thing or event under investigation in the "actual world." Validity is a quality dimension linked to accuracy.
- **Completeness:** The ratio of data saved to the potential for "completeness." Validity and accuracy are two quality dimensions related to completeness.
- **Uniqueness:** No object, regardless of how it is identified, will be recorded more than once. Consistency is the quality dimension connected to uniqueness.
- **Timeliness**: The degree to which data correctly represents reality at a given point in time. The quality component associated with timeliness is accuracy.
- **Validity:** Data is accepted if it adheres to the syntax of the specification (format, type, and range). The validity-related quality dimensions include accuracy, completeness, consistency, and uniqueness.

Data quality features appear in the data collection process as well as the preprocessing stage - this includes both the upstream and downstream stages of the data processing process (Kirchen et al., 2017). The upstream influencing factors are determined by the data collection system. The loss of data quality is expressed by missing values in the event of data storage failures or the inability to measure the requested physical values. The Completeness indicator considers any missing values. Accessibility, Mobility, and Recovery all fall under this umbrella term. The data analyst cannot use the signal data if it cannot be accessed or if it cannot be transferred to a database or data mining software. In the event of a failure or loss of data, a lack of recoverability results in a lack of information. When it comes to

Traceability, the impact isn't caused by missing values in the time series, but rather by a lack of details about the dataset itself. The various sub-dimensions of Completeness also cover this influence. When compared to the factors that have an impact on data quality upstream, the factors that have an impact on downstream DQ that appear during data preprocessing are Accuracy, Credibility, Consistency, and Relevance. It is also worth noting that the conversion of signal data to SI units, for example, does not have a negative impact on the quality of the data. Compliance is a problem for data quality if there is a lack of information that prevents conversion to a particular standard.

DQ Metrics

Different definitions for data quality metrics have been presented by many research studies. Below we list a few metrics (Batini and Scannapieco, 2016b; Heinrich et al., 2018a; Kirchen et al., 2017).

- **Timeliness Metric:** The ratio between currency and volatility determines the timeliness and they must be measured using the same units of time. Time tags provide information about the date the data item was acquired. The data quality metric for timeliness is defined as:

  $$\textbf{Timeliness} = \{1 - \text{currency}/\text{volatility}, 0\}^{s}$$

  The metric's sensitivity to the ratio age of the data value depends on the exponent $s > 0$, which must be determined based on expert estimations.

- **Completeness Metric:** Complete data has been described as data with all values recorded. Missing data can typically be indicated by null or another indicator in most applications. The metric for completeness is:

  $$\textbf{Completeness} = 1 - (T/N)$$

  where **T** is the proportion of tuples in relation with null values to the total number of tuples and **N** is the total number of tuples.

- **Correctness Metric:** Hinrichs defines correctness, as a metric to evaluate the accuracy of a stored data value, defined by:

  $$\textbf{Correctness} = 1 / (d(V, V_x) + 1)$$

  where **V** is the stored data value, V_x is the corresponding real-world value and **d** is a domain-specific distance measure.

E. IoT Data Quality: Dimensions and influencing Factors

IoTDQ Dimensions

To effectively address IoT Data Quality (IoTDQ) challenges and ensure the integrity and usability of data within IoT ecosystems, it's essential to focus on key elements that contribute to data quality. These elements are fundamental to assessing, improving, and maintaining the quality of data collected from diverse IoT devices and sensors. Here is a comprehensive outline of IoT Data Quality elements, some of them already listed previously (D) like Accuracy, Completeness, Timeliness, Validity, Uniqueness:

- **Consistency**: The uniformity of data across different sources and over time. Consistency is vital in IoT ecosystems that integrate data from multiple types of devices and protocols.
- **Reliability**: The dependability of data, including its stability and availability over time. Reliable data is essential for IoT systems to perform consistently and deliver expected outcomes.:
- **Integrity**: The accuracy and consistency of data over its entire lifecycle. Data integrity in IoT involves maintaining the correctness of data as it is collected, transmitted, stored, and processed.
- **Security**: Protecting data from unauthorized access and ensuring that data privacy is maintained. In IoT, security is paramount to protect sensitive information transmitted between devices and over networks.

These elements form the foundation of a robust IoT Data Quality management framework, guiding the implementation of practices and technologies designed to ensure that data collected and used in IoT ecosystems is of high quality. Addressing these elements comprehensively can help in overcoming common data quality challenges, enhancing the reliability of IoT applications, and facilitating better decision-making based on IoT data.

IoTDQ Influencing Factors

In the following are the factors that impact and influence the IoT data quality such as technical factors and human factors. This list is illustrated in Fig.5.

Technical Factors:

- **Hardware Quality**: Relates to the reliability and accuracy of the IoT devices and sensors themselves. Poor-quality hardware can produce inaccurate or unreliable data.
- **Software and Firmware**: Encompasses the algorithms and operating systems running on IoT devices. Bugs or limitations in software can lead to data quality issues, such as incorrect data processing or loss of data.
- **Network Connectivity and Bandwidth**: Affects the timely and successful transmission of data. Network failures or limited bandwidth can result in data loss or delays, impacting data freshness and relevance.

Environmental Factors:

- **Physical Conditions**: Environmental conditions (e.g., temperature, humidity) can affect sensor performance and, consequently, the quality of the data collected.
- **Interference**: Electromagnetic interference or physical obstructions can distort or block the data being transmitted from sensors to data collection points.

Human Factors:

- **Configuration Errors**: Improper setup or configuration of devices can lead to inaccurate data collection.
- **Maintenance Practices**: Inadequate maintenance of sensors and devices can degrade their performance over time, affecting data accuracy.

Data Management Factors:

- **Data Handling and Storage**: Involves the processes of data collection, transfer, and storage. Data can be corrupted or lost due to mishandling or technical failures in storage systems.
- **Data Processing and Analysis**: Incorrect or inadequate processing algorithms can lead to misleading analysis results, affecting decision-making based on this data.

Security and Privacy Factors:

- **Data Tampering**: Unauthorized access to IoT devices can lead to intentional data manipulation, compromising data integrity.
- **Privacy Leaks**: Inadequate data protection measures can lead to privacy breaches, affecting the trustworthiness and ethical use of data.

Fig. 5. Factors Influencing IoT Data Quality

Impacts and Influence on IoT Data Quality

- **Reliability**: Technical and environmental factors directly impact the reliability of data. Unreliable data can lead to erroneous conclusions and decisions.
- **Accuracy**: Influenced by technical, environmental, and human factors, accuracy is critical for making informed decisions based on IoT data.
- **Timeliness**: Network connectivity and bandwidth affect the timeliness of data, which is essential for real-time monitoring and response systems.
- **Completeness**: Data management practices influence the completeness of the data collected. Incomplete data sets can skew analysis and insights.
- **Consistency**: Software and firmware quality, along with data management practices, ensure consistency across data sets, facilitating comparative analysis.
- **Security and Privacy**: Security and privacy factors affect the integrity and ethical use of data, influencing user trust and regulatory compliance.

By understanding these categories and their impacts, IoT system designers and operators can implement targeted strategies to improve data quality across the IoT ecosystem. This structured approach aids in identifying specific challenges and applying best practices for data quality management, ensuring that the data collected, processed, and analyzed is reliable, accurate, and meaningful.

F. Some IoT Use Case Examples

In the following some case studies that highlight how IoT technology addresses several everyday life challenges. These examples underscore the transformative impact of IoT across various sectors, addressing connectivity, computing power, and data quality challenges through innovative applications and solutions.

Smart City Traffic Management: In several world big cities, IoT sensors are deployed throughout to monitor and manage traffic flow. These sensors collect data on vehicle movements, enabling real-time traffic management and reducing congestion. By analyzing this data, the city adjusts traffic signals and reroutes traffic to improve urban mobility. This demonstrates how IoT connectivity and computing power can address urban traffic challenges, leading to significant improvements in traffic flow and pollution reduction.

Healthcare Patient Monitoring: The use of IoT in healthcare is exemplified by the remote patient monitoring system implemented by the Cleveland Clinic. Patients with chronic conditions are provided with wearable IoT devices that monitor vital signs in real-time, transmitting data back to healthcare providers. This system allows for continuous monitoring without the need for hospital stays, improving patient outcomes and reducing healthcare costs. IoT in healthcare is enabling remote care delivery.

Industrial IoT in Manufacturing: IIoT (Industrial Internet of Things) application collects and analyzes data from industrial equipment to predict maintenance needs and optimize operations. For instance, data from sensors on jet engines enable predictive maintenance, reducing downtime and saving costs. IIoT illustrates how IoT can lead to more efficient and proactive management of resources.

III. IoT Data Quality: Related Works

The advent of the IoT has heralded a new era in data generation, with billions of devices continuously producing data (Manyika et al., 2015). As these data integrate with Big Data ecosystems, ensuring their quality becomes paramount (Zikopoulos and Eaton, 2011). This literature review delves into key studies, methodologies, and findings pertinent to data quality assessment for IoT in the Big Data realm.

Data quality plays a pivotal role in ensuring the credibility of analyses drawn from the IoT Big Data ecosystem (Vermesan and Friess, 2014). For businesses and decision-makers relying on these analyses, it's crucial that the data is accurate, reliable, and timely. Poor quality can result in costly mistakes, from misguided business strategies to malfunctioning automation systems. Furthermore, for end-users or stakeholders, trust in the system's output is paramount. If they perceive that the data is of low quality, they might be reluctant to adopt or invest in IoT solutions.

Many challenges must be resolved when managing data quality in IOT as the IOT application domains are expanding rapidly like Healthcare, Agriculture, Smart Cities, ...etc. The IOT data used is going through many stages that are essential to assure its quality and provenance but not mandatory in all sectors. The data is collected, integrated, stored, shared, or directly accessed in the Edge for local processing including quality assessment and verification of sources.

Guaranteeing IOT data quality requires managing it at different stages of Big Data values chain including the specification of IOT data with edge computing. Several works addressed many aspects of IOT DQ from the possible challenges in IoT Data Quality, management framework, evaluation processes, enhancement procedures, ...etc.

In the following a brief classification of works addressing IoT data Quality from several facets.

A. Specific Challenges in IoT Data Quality

The data generated within the IoT Big Data Ecosystem stands out due to its inherent characteristics. It's often real-time or near-real-time, demanding swift actions. It's heterogeneous, given the diverse range of devices. It can be structured (like database entries) or unstructured (like video feeds). And its generation can be constant (like a temperature sensor reporting every second) or event-driven (like a smoke detector triggering only when there's smoke). Ensuring data quality in the expansive and intricate IoT Big Data Ecosystem demands multifaceted approaches that address specific challenges. These are related to the diverse, dynamic, and decentralized nature of the IoT Big Data ecosystem that brings about unique challenges in maintaining data quality. While some of these challenges have traditional parallels, others are distinctly characteristic of the IoT paradigm.

a. **Volume and Velocity**: As billions of devices continuously feed data into the ecosystem, managing and monitoring this massive influx becomes daunting (Chen et al., 2014). The sheer volume makes manual checks implausible and stresses automated systems. Moreover, real-time or near-real-time data processing demands are required by many IoT applications (Bello-Orgaz et al., 2016), which leaves minimal room for comprehensive data quality checks before decisions or actions are based on the data. (Y. Li et al., 2020; Zhang et al., 2021)
b. **Variety**: the heterogeneous aspect of data and its diverse array of devices in the IoT landscape means it comes in all shapes and sizes—from numerical values of sensors to video feeds of security cameras (Gubbi et al., 2013). Each data type requires specific quality assessment metrics and tools. Add to that, different Sources where data originate like numerous manufacturers, regions, and technologies. This diversity can lead to inconsistencies in data formats, units of measurement, or even data granularity.(Hasan and Curry, 2015; Kollolu, 2020)
c. **Veracity**: With numerous touchpoints—from device to storage to analysis—the potential for data corruption or alteration exists at every step which may impact data Integrity. Therefore, not all devices are created equal. Variations in device quality, calibration, or even wear and tear can impact the accuracy and precision of the data generated.(Assiri, 2020; Liu et al., 2018; Moke et al., 2021)
d. **System Vulnerabilities**: IoT devices have been targeted in numerous cyberattacks. Such breaches can tamper with data, making it unreliable or even maliciously misleading. Also, Physical Vulnerabilities where Devices exposed to harsh environments or physical tampering can produce compromised data (Stankovic, 2014).
e. **Lack of Standardization**: The IoT world lacks universal standards. This can lead to challenges in data integration, especially when systems from different vendors are expected to communicate and collaborate. Without standard guidelines, what one entity considers high-quality data might be deemed subpar by another. (Al-Qaseemi et al., 2016)
f. **Legacy Systems & Integration Challenges**: Older, non-IoT systems might need to integrate with newer IoT devices. These legacy systems, with their own data quality issues, can introduce challenges when harmonizing with the dynamic nature of IoT data.(Liu et al., 2023) (Jha, 2021) (Tedeschi et al., 2018)
g. **Resource Constraints**: Transmitting vast amounts of data requires substantial bandwidth. These constraints can lead to data loss or forced data reduction techniques, impacting data quality. Add to that, the limited energy in many IoT devices that are battery-powered. Energy constraints can lead to reduced data transmission frequencies or even device downtimes, causing gaps in data.(Li et al., 2023; Mukherjee et al., 2022)

In navigating the evolving maze of the IoT Big Data Ecosystem, these challenges form the intricate barriers that entities—be they businesses, governments, or developers—need to surmount. While daunting, recognizing these challenges is the first step toward formulating strategies and solutions to ensure impeccable data quality.

B. Data Quality Assessment

Many approaches for detecting noisy data have been proposed in the literature for IoT. (Teh et al., 2020) presents a systematic review of the literature to identify various dimensions of data quality and their associated challenges in the context of IoT sensory data.

Data profiling can help address DQ issues such as noise and outliers, inconsistent data, duplicate data, and missing values (Azeroual et al., 2018) (Taleb et al., 2019). Class imbalance issues develop during classification because of the data's minimal class representation, and as a result, the dominant class is favorably viewed by the classifiers. Like this, there are several occasions where ML systems' general classification accuracy incorrectly labels minority classes. Increased misclassification expenses, time, and risk assessment are the results of this. Since the training sessions may be too brief in the context of Edge computing, the dependability of the taught data values for each IoT device stored data (training set) depends on the IoT device (Fantacci and Picano, 2020).

It might be challenging to filter through this deluge of information in today's digital age where massive amounts of data are being generated at a rapid rate to identify the bits of worth. Because the majority of the data and resources needed to successfully train machine learning models are under the hands of huge tech companies, there is a centralization problem. The topic of relevant data selection is explored in traditional machine learning settings under the assumption that all relevant data is available for computation in a single location(AlNuaimi et al., 2019; Bijarbooneh et al., 2015). The traditional methods for selecting relevant data are not applicable in this case since the data in FL settings are scattered among several clients and the server is unable to review them due to privacy restrictions (Nagalapatti et al., 2022).

Studies that have already been done on data quality have been tackled from various angles. The majority of the publications' authors concur that the phases or processes of a data's lifetime are related to the quality of that data (Chen and Zhang, 2014; Kim et al., 2022; Mouha, 2021). Particularly, the stages of data generation and/or the source of the data are strongly correlated with data quality. The procedures used to evaluate data quality are based on conventional data management techniques, and Big Data should be accommodated. Additionally, how the quality evaluation metrics are developed and used depends on the application domain and type of information (Content-based, Context-based, or Rating-based). While context-based metrics employ meta-data as quality indicators, content-based metrics use the information itself as a quality indicator.

According to (Glowalla et al., 2014; Sidi et al., 2012), there are two basic approaches to improving data quality: data-driven and process-driven. The first technique uses pre-processing procedures including cleansing, filtering, and normalizing to address data quality during the pre-processing stage. These steps are crucial and take place prior to the data processing stage, ideally as soon as possible. The process-driven quality method, however, is utilized at every link in the Big Data value chain. Early on in the literature, data quality assessment was covered (Hu et al., 2014). Subjective and objective categories serve as its main divisions. Additionally, a strategy that combines these two groups was presented to give organizations practical data quality measures to assess their data. However, the suggested approach, was not created to handle Big Data characteristics. In conclusion, data quality needs to be addressed early in the data lifecycle, during pre-processing. The aforementioned Big Data quality issues have not been thoroughly explored in the IoT data quality literature. There are still a lot of unresolved problems, particularly in the pre-processing step, which must be addressed at the IoT device in the Edge which is a particular characteristics of Big Data IoT only.

C. IoT Data Quality Frameworks

Like data quality frameworks that have been established to manage quality in any environment and system that deals with data in its different forms. Many frameworks have been proposed like for Big Data (Taleb et al., 2021) that tackles data quality from the inception of the data till its analytics with a continuous feedback to assess the quality requirements established in their data quality profile.

(Vermesan and Friess, 2013) establish a broad overview of IoT's converging technologies, with discussions on the role of data quality and the necessity for proper frameworks specialized of IoT is necessary since its specific characteristics that cannot be dealt as Big Data only which create a novel direction of quality assessment IoT aware only.

(Byabazaire et al., 2023b) propose a framework solution that integrates different fusion methods (Techniques used to combine data from multiple sources or sensors to produce more consistent, accurate, and useful information than that provided by any individual data source or sensor) for different IoT applications' that require specific and unique data quality needs. The paper explores the effects of several fusion methods on data quality scores and their pertinency in maintaining diverse applications. It also evaluates the computational efficiency of the fusion methods to optimize service location. Therefore, still some limitations related to the representativeness of the used scenarios with a reduced fusion methods probing remains to be more explored including resource needs and scalability concerns.

In IoT and Smart City networks, heterogeneous sensor device networks with numerous maintainers, data gathered from social media, as well as crowdsourcing, frequently contain aspects of uncertainty. There is frequently no available ground truth that can be utilized to verify the veracity and consistency of the new information. (Kuemper et al., 2018) propose "Valid-IoT" Framework as a connectable IoT framework element that can be associated with various platforms to produce a Quality of IoT vectors and interpolated sensory data with estimates of plausibility and quality. By fusing crowdsourced and device-generated sensor data, the system uses enhanced infrastructure knowledge and infrastructure-aware interpolation methods to validate and assess the data.

In another IoT federated learning context, (Navaz et al., 2023) proposed a Federated Data Quality profiling that tackles various quality issues at multiple edges while keeping data localized, thereby enabling data privacy and reducing data resource consumption and transportation costs. They use a federated profiling to lessen the costs of enhancing data quality and ML classification algorithms at multiple IoT edges.

Last but not least, AI frameworks are the newcomers in Big Data quality assessment and enhancement, particularly for IoT. (Martín et al., 2023b) suggest an AI-based framework architecture to improve and evaluate data quality in IoT data. The paper analyzes and defines the most important DQ dimensions in IoT data streams as well as the evaluation methods for each one. It also presents data curation (The process of organizing, integrating, cleaning, and enhancing data to ensure it is of high quality and suitable for specific analytical purposes) techniques that make use of AI algorithms to improve the quality aspects of IoT data streams. They implemented AI-based data curation mechanisms, evaluated the results of their experiments on artificially generated data, and showed appropriate behavior in terms of removing IoT observations that would reduce DQ by utilizing linked-data principles and integrating AI-based IoT data curation mechanisms within a Data Enrichment Toolchain.

D. IoT Communication Networks

To convey data, communication is a crucial aspect of IoT, and ways to reduce communication, such as local updating and model compression, must be understood through a thorough examination of the trade-off between accuracy and communication for each strategy. Many researchers have been interested in the impact of the communication network on IoT data quality, particularly in the context of big data. The dependability, bandwidth, latency, and security of the network have a direct impact on the quality and integrity of data transported from IoT devices to central systems or cloud platforms. It's also interesting to look at the effects of this more realistic device-centric communication architecture, in which each device may choose when to connect with the central server in an event-triggered way (C et al., 2022; T. Li et al., 2020) .

With large, heterogeneous IoT devices and networks, a communication-efficient strategy must be characterized by a minimum of communication rounds, or a low number of gradients in any communication cycle. Another issue is that the majority of IoT analytics machine learning models have many redundant parameters that do not need to be communicated to the server, wasting communication resources. Furthermore, utilizing data from edge devices is not always practicable due to the devices' inability to relay data quickly enough or due to poor network conditions. (Merino et al., 2020; Xiong et al., 2006).

It is important to note that communication aspects have a specific impact on IoT and its data quality. Most of these aspects can be limited to the data itself like (1) Data Loss and Incompleteness that happens due to network instabilities or congestion, with probabilities that data packets will not reach their destination (Al-Fuqaha et al., 2015), (2) Network induced-Latency and Timeliness that can degrade the value of real-time data (Bello and Zeadally, 2016), (3) Data Integrity and Consistency that might be tampered with during data transit in insecure networks (Sicari et al., 2015), (4) Bandwidth Limitations impacting and reducing data granularity that might require data compression (Pratap et al., 2019), (5) Significant energy consumption in IoT devices Network communication which impact on the quality of the data in the long term transmission (Adelantado et al., 2017), (6) the Security and Privacy Concerns that can lead to data breaches or tampering (Roman et al., 2013), (7) Scalability Challenges with a highly estimated increase in the future of networks need to handle increased data traffic as IoT deployments is growing everyday (Gubbi et al., 2013), wireless Interference and Noise might introduce errors or necessitate data retransmissions leading to an increase of energy consumption and delays (Wei et al., 2020).

E. IoT DQ Management Innovative Solutions

The proliferation of IoT devices has guided in a new era of data generation, requiring rigorous quality assessment methodologies to ensure the reliability and accuracy of insights derived from this data. The landscape of IoT data quality is troubled with challenges, including but not limited to, data heterogeneity, volume, and velocity, all of which compound the difficulty of maintaining high data quality standards. In

The following we distills the essence of the literature, pointing in on pivotal challenges and groundbreaking solutions that mark significant advancements in the field.

Addressing Data Heterogeneity: One of the quintessential challenges in IoT data quality is the inherent heterogeneity of data that needs to be integrated from diverse sources while optimizing for quality and computational efficiency. There is a need for more efficient and flexible frameworks capable of accommodating the varied nature of IoT data.

Enhancing Data Veracity with Federated Learning: Tackling data quality issues at the edge of the network with an efficient data privacy preservation and mitigating resource overhead, represents a shift towards decentralized data quality enhancement, reflecting the distributed nature of IoT networks.

AI-driven Data Quality Improvement: The integration of AI for data quality management highlights the potential of artificial intelligence algorithms to automate and enhance the data curation process. Through AI-based frameworks, IoT ecosystems can dynamically improve data quality, addressing issues such as noise, inconsistency, and incompleteness in real-time.

Finally, it is important to note that there is still much work to be done on the development and standardization of IoT Big Data quality. Research projects and solutions in this area are underway. This current research field is diverse, complex, and multivariant, and novel assessment approaches, processing and analytics algorithms, storage and processing technologies, and platforms will be crucial to the growth and maturity of this field. We believe that academic scholars will contribute to the creation of new methodologies that will go beyond the conventional DQ approaches presently in use. These methodologies will be primarily based on AI, federated learning processes, and optimization techniques. Additionally, industries will lead development initiatives of new platforms, solutions, and technologies optimized to support end-to-end quality management within the Big Data lifecycle. Still, some of the challenges related to IoT data that is distributed and from several sources which present its own set of issues, as opposed to a centralized dataset that is simply dispersed around IoT nodes.

IV. Holistic Architecture of IoT Data Quality Management Model

Data management in Big Data is agnostic to quality management. As a result, we must identify quality concerns and requirements at each stage of the Big Data lifecycle. A quality improvement process is unavoidable in order to ensure the high quality of the Big Data value chain, and it should be embedded into each phase of the Big Data lifecycle. The goal is to carry out quality management tasks without introducing unnecessary communication, processing, and cost overhead on the various Big Data levels.

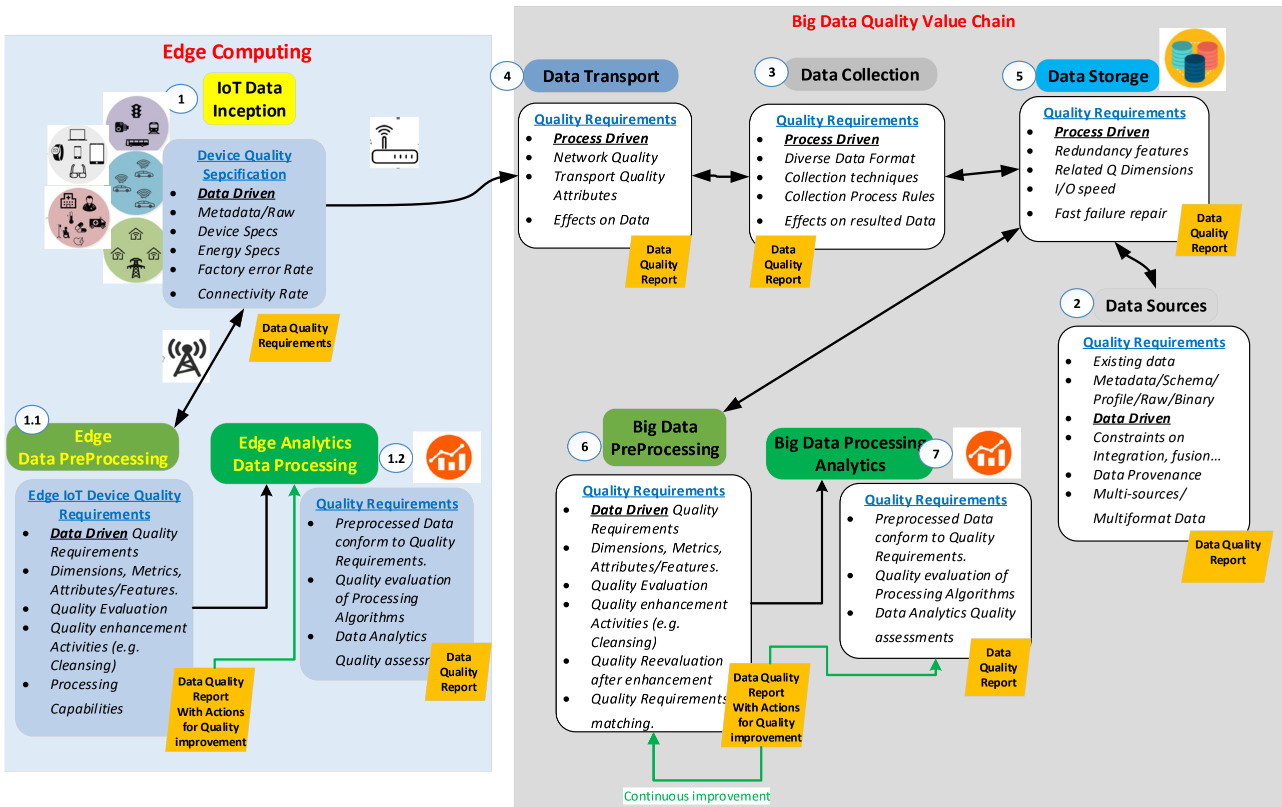

Fig 4. A Holistic Architecture of IoT Quality Management Model for Big Data Value Chain

To get started on a quality Big Data project, a set of characteristics and concepts must be developed in order to identify the processes involved and characterize the data and workflow type used as important inputs for its administration. Data-driven methods and process-driven strategies are utilized in data quality (Taleb et al., 2021). The data-driven method works on the data itself, examining its quality and aiming to improve it. The process-driven method is a predictive strategy that focuses on the process used to generate or transform data.

To guarantee Data Quality in the expansive and intricate IoT Big Data Ecosystem necessitates multifaceted approaches that address the specific challenges we've identified. In figure 4, We propose here a prominent holistic approach to IoT Data Quality management that captures important quality aspects and explores how to deal with Big Data quality throughout its lifecycle.

To ensure effective data management, we detailed the processes that must handle and address data quality issues, as well as provide quality evaluation methodologies. The most important stages are, in this order: (1) data inception, (2) data source, (3) data collecting, (4) data transfer, (5) data storage, (6) data pretreatment, and (7) data processing and analytics. Furthermore, to develop an end-to-end IoT Big Data quality management driven lifecycle, Big Data lifecycle stages, as well as quality information and associated processes, must be addressed.

In the following we emphasize the significance of quality requirements considerations, quality implementation, and enforcement across various processes in the Big Data value chain. To help quality management, we also characterize quality propagation across multiple processes in the value chain, as well as the level of cooperation between these activities. A continuous quality improvement through loopbacks and inter-process involvement is paramount to keep the quality requirements level accomplished through the lifecycle.

a. Quality enabled data acquisition

Data acquisition is divided into two stages: data inception and data collection. However, data sources (2 in Fig. 4) are linked to previously stored data in a variety of formats. Before addressing Big Data quality at its start, consider existing data that has been acquired in the past and create a knowledge foundation from this data to better redesign its production. This is aided by the introduction of Unified Data Models, which define conventions for standardization and interoperability. Adopting or creating these models means that data coming from multiple sources can be seamlessly combined. Without excluding the actual developed or existing open standards utilized in the IoT ecosystem, which can aid in mitigating interoperability difficulties and encouraging consistent data quality measurements.

The adoption of Unified Data Models is important. These models are similar to universal translators, helping different data "languages" from various sources understand each other. They set common standards and rules, ensuring that data, regardless of its origin, can be seamlessly merged and communicated. This is crucial for standardization and interoperability.

Unified Data Models serve as a toolkit to overcome the Tower of Babel scenario in IoT data, where each device or system might use its own unique way of describing and formatting data. By adopting or creating these models, we ensure that data from a temperature sensor, a smartwatch, or a manufacturing robot can all be integrated, enhancing the quality and usability of the data collected. This integration is vital for making informed decisions, as it allows for a comprehensive view of data collected across different devices and platforms.

In this context, data design must be an iterative process that compromises data inspection and data quality review (typically using a data-driven method when using current data) to improve procedures that provide high-quality data suitable to be used by Big Data lifecycle processes.

In the data creation phase, a process-driven technique is used, with quality restrictions specified to eliminate bad data before it is created. Furthermore, calibration and maintenance for IoT devices are addressed in this phase. The continual calibration of Internet of Things devices, particularly sensors, which may deviate from their calibrated states over time. Calibration of devices on a regular basis ensures that they continue to produce reliable data. In addition, predictive maintenance is implemented by forecasting when a device is likely to break or degrade, allowing for proactive maintenance and ensuring continuous, high-quality data flow.

b. Quality-aware data transport and storage

It is critical to enable QoS throughout the data transport phase because of the requirements that must be met during service provisioning. Ensuring IoT data quality is more specifically related to the underlying networks that transfer data, as well as the security mechanisms that ensure data transmission across many IoT devices and node points without data loss or corruption. Data network quality is determined by device QoS requirements, provider QoS offerings, and real-time QoS measurements. QoS's major purpose is to offer priority, which includes dedicated bandwidth, regulated jitter, and delay (e.g., IoT real-time stream processing), and enhanced loss characteristics.

Furthermore, improving IoT security by implementing robust security processes at the device level, such as regular firmware updates and encrypted communication, can defend against unauthorized data modification. Secure data transmission methods and intrusion detection systems will assure data integrity during transmission.

Data quality in Big Data and IoT is regulated for data storage via data replication and delivery. For example, the Hadoop ecosystem's storage is based on several nodes that duplicate data to prevent catastrophic data loss and to ensure continuity in the case of a breakdown. Furthermore, the I/O performance of storage medium influences the quality of data storage for various types of Big Data. Reading and writing ratios must be followed by the I/O requirements for each type of data, such as video HD data, streaming, and stream processing.

c. Quality driven data preprocessing

Preprocessing is the last line of defense in Big Data quality management. It is the process of cleaning, integrating, normalizing, converting, and integrating data before it is ingested by the processing stage. It takes a data-driven approach that mainly focuses on data values. It is represented as Automated Data Validation and Cleaning processes in IoT. Either in real-time, where data is validated as it is generated and sent using tools and algorithms, or in batch mode, where data is validated as it is generated and transferred. These technologies can detect anomalies or disparities in real-time, alerting users to the need for additional research or immediate correction. The cleansing process works its way through the acquired data, detecting and fixing errors, filling gaps, and removing noise or irrelevant information.

d. Quality enabled data processing and analytics

Processing and analytics benefit from the quality assessments completed in prior quality assessment activities. Quality management in the processing step consists of checking that preprocessed data meets processing quality requirements. Furthermore, IoT Edge Computing, which is defined by local data processing at the source (i.e., the device itself or a nearby node), can handle bandwidth and data quality concerns. The only relevant, processed data is then delivered to central systems. In terms of Big Data preparation, IoT Data processing frameworks, like those used for Big Data, have been developed to run on the edge(Namiot, 2015).

e. Quality propagation and continuous quality improvement

The quality management model depicted in Fig. 4 shows how quality management activities spread across the various operations of the IoT Big Data value chain. A quality need is considered at each stage, quality measurements are recorded, and a quality report is prepared. This quality report is delivered to the next level of the value chain to track, validate, enrich, and adapt the data and process quality. Continuous quality improvement is enabled through process loopback to monitor, revise, and adapt quality as needed. This is an essential component that ensures accuracy and improves quality evaluation throughout the IoT Big Data value chain. Continuous real-time monitoring using dashboards that provide data quality indicators and aid teams in promptly finding and addressing issues. Systems that can adapt and improve over time by continuously learning from earlier data quality concerns, adjusting algorithms, or alerting human overseers are examples of feedback mechanisms.

The IoT Big Data Ecosystem faces complicated data quality concerns that call for a coordinated fusion of different techniques. While technological solutions are essential, it's as important to promote a culture of data quality, place a premium on training and constant learning, and adjust to the ecosystem's ever-changing conditions.

Use case Example: Smart Farming System with Edge Computing

Considering the comprehensive Holistic Architecture of the IoT Data Quality Management Model provided, a real use case that encapsulates all the described processes, including data inception at the edge, along with related data preprocessing, processing, analytics, and quality management at the edge, can be depicted through an advanced smart farming system.

A large-scale agricultural enterprise aims to leverage IoT and edge computing technologies to revolutionize its farming operations. The goal is to maximize crop yields, optimize resource usage, and ensure environmental sustainability through precise, data-driven decision-making.

1. Quality-Enabled Data Acquisition

- **Data Inception and Collection at the Edge**: The system deploys a network of soil moisture, temperature, pH sensors, and drones equipped with multispectral cameras across various farm locations. These edge devices collect real-time data on soil conditions, crop health, and environmental factors. Unified Data Models are employed to standardize the diverse data formats and ensure seamless integration and interoperability across devices.
- **Consideration of Existing Data**: Historical crop yield data, weather patterns, and soil analysis reports are integrated into the system. This existing data forms a knowledge foundation, aiding in the refinement of data collection strategies and predictive analytics models.

2. Quality-Aware Data Transport and Storage

- **Secure Data Transport**: Data transmitted from edge devices to the central processing units and cloud storage undergoes encryption, maintaining data integrity and confidentiality. Quality of Service (QoS) protocols ensure priority data delivery with minimal delay and data loss.
- **Robust Data Storage Solutions**: Utilizing the Hadoop ecosystem for distributed storage, data is replicated across multiple nodes. This ensures data redundancy and high availability for continuous, quality-driven farming operations.

3. Quality-Driven Data Preprocessing

- **Automated Data Validation and Cleaning at the Edge**: Data collected in real-time is preprocessed directly on edge devices. Automated algorithms validate data integrity, correct anomalies, fill missing values, and remove irrelevant information, ensuring only high-quality data is forwarded for further processing.

4. Quality-Enabled Data Processing and Analytics

- **Edge Processing and Analytics**: Local edge computing nodes perform initial data analysis, running machine learning algorithms to detect immediate insights like pest detection, nutrient deficiencies, or irrigation requirements. This reduces latency and conserves bandwidth by transmitting only relevant, processed data to centralized systems for further analysis.

5. Quality Propagation and Continuous Quality Improvement

- **Continuous Monitoring and Feedback Loops**: The system continuously monitors data quality and system performance through real-time dashboards. Feedback from processed data and analytics results informs adjustments in data collection and preprocessing algorithms at the edge, fostering a cycle of continuous quality improvement.

6. Implementation and Impact

The smart farming system dynamically adjusts irrigation schedules, fertilizer application, and pest control measures based on real-time data and predictive analytics. Edge computing plays a pivotal role in enabling immediate, localized responses to detected conditions, significantly reducing the time between data collection and actionable insights. Continuous quality improvement mechanisms ensure that the system evolves, adapting to changing environmental conditions and emerging challenges, ultimately leading to higher crop yields, reduced resource waste, and enhanced sustainability of farming operations.

This use case exemplifies the holistic application of the IoT Data Quality Management Model in a real-world scenario, showcasing how edge computing can be leveraged to ensure high-quality data management throughout the IoT Big Data lifecycle, from inception to analytics.

V. Challenges and Future Research Directions

As IoT continues to expand and evolve, Big Data quality assessment faces numerous challenges that require new approaches and methodologies. These challenges derive from the unique characteristics of IoT data and the sheer scale and complexity of IoT ecosystems. In the following, we describe some of the key challenges and trends that will likely shape the future.

The heterogeneity of data sources, spanning sensors, devices, and social media, introduces variations in data formats and structures, while the high volume and velocity of data overwhelms traditional assessment methods. The diverse types of data, including time-series data, images, and text, demand specialized quality assessment approaches. Imbalanced data distributions, privacy concerns, resource constraints on IoT devices, and the need for resource-efficient algorithms further complicate the landscape.

In order to meet these problems, continual research and innovation are needed in a number of promising areas, all of which are meant to ensure the validity and usefulness of IoT-generated Big Data.

- Machine learning (ML) and artificial intelligence (AI) integration: ML and AI algorithms will play a key role in automating the processes of data validation, anomaly detection, and data cleansing. These algorithms will continuously improve based on historical data in order to handle brand-new problems with data quality.
- Emphasis on Edge and Fog Computing: More localized data processing will be required as IoT grows, requiring edge and fog computing. These approaches will lessen the risk of transmission-related errors by improving data quality at the source in addition to addressing latency and bandwidth constraints.
- Quantum Computing for Data Validation: With its better computational capabilities, quantum computing holds the potential for managing massive IoT datasets. Data validation procedures

may undergo a revolution as quantum technologies develop, especially in complex networks with millions of connected devices.

- Blockchain for Data Integrity and Provenance: The foundation of data quality is ensuring data integrity and establishing data lineage. Particularly in decentralized IoT systems, blockchain's immutable ledgers can provide a transparent and tamper-proof means to validate and trace data.
- Growth of Data Quality as a Service (DQaaS): The market for data quality as a service (DQaaS) is expanding as more companies use the IoT without having the internal resources to monitor data quality. Service providers with specialized knowledge will give complete data quality solutions designed for certain markets or applications.
- Emphasis on Data Ethics and Privacy: As IoT devices grow in private and sensitive areas, maintaining data quality also entails moral data gathering, processing, and storage. Regulations are likely to change, placing more emphasis on ethical issues and data quality.
- Improved User Interfaces for Data Quality Management: As IoT systems get more sophisticated, it will be essential to have user-friendly interfaces that provide data quality indicators, anomalies, and system health. Virtual reality (VR) and augmented reality (AR) may contribute to providing immersive user interfaces for large-scale.
- Sustainable and Green IoT: Large IoT networks' environmental impact will be investigated. Ensuring data quality may also entail utilizing sustainable energy sources, developing low-impact technology, and optimizing data flows to save energy.

In summary, a convergence of technology developments, cooperative efforts, legislative evolutions, and a raised emphasis on ethics and sustainability will shape the future of data quality in the IoT Big Data Ecosystem. Organizations and stakeholders who foresee these developments and adjust to them will be in the forefront, maximizing IoT's full potential while preserving the quality and integrity of the data.

VI. Conclusion

IoT holds immense promise as it interconnects everyday objects on a global scale, offering intelligent and ubiquitous services for the benefit of humanity. This interconnected network generates an unprecedented volume of data, serving as a foundational resource for deriving insights about individuals, entities, and phenomena, all aimed at enhancing IoT services. However, the quality of this data is crucial. In fact, user engagement and the adoption of the IoT paradigm depend critically on the reliability of the data. In this paper, we investigated data quality assessment in the IoT context. We discussed big data and IoT data properties and their new lifecycles. Moreover, we reviewed the literature on IoT data quality where we dressed a comprehensive classification of quality assessment processes in the IoT. Additionally, we proposed a holistic architecture for IoT quality assessment management model that captures key characteristics of IoT applications. Finally, we discussed IoT-related challenges when assessing the quality of data and drew some research directions in this research area.

In this ever-evolving tapestry of the IoT Big Data Ecosystem, one dictum stands the test of time: Quality isn't just a metric; it's a commitment. A commitment to accuracy, reliability, and above all, to the countless stakeholders who hinge their decisions, big and small, on the data that flows through the veins of our interconnected world.

VII. References

Adelantado, F., Vilajosana, X., Tuset-Peiro, P., Martinez, B., Melia-Segui, J., Watteyne, T., 2017. Understanding the Limits of LoRaWAN. IEEE Commun. Mag. 55, 34–40. https://doi.org/10.1109/MCOM.2017.1600613

Aggarwal, A., 2016. Identification of quality parameters associated with 3V's of Big Data, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Presented at the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1135–1140.

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M., 2015. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Commun. Surv. Tutor. 17, 2347–2376. https://doi.org/10.1109/COMST.2015.2444095

AlNuaimi, N., Masud, M.M., Serhani, M.A., Zaki, N., 2019. Streaming feature selection algorithms for big data: A survey. Appl. Comput. Inform. https://doi.org/10.1016/j.aci.2019.01.001

Al-Qaseemi, S.A., Almulhim, H.A., Almulhim, M.F., Chaudhry, S.R., 2016. IoT architecture challenges and issues: Lack of standardization, in: 2016 Future Technologies Conference (FTC). Presented at the 2016 Future Technologies Conference (FTC), pp. 731–738. https://doi.org/10.1109/FTC.2016.7821686

Ann, F.N., Wagh, R., 2019. Quality assurance in big data analytics: An IoT perspective. Telfor J. 11, 114–118. https://doi.org/10/ggzhg3

Assiri, F., 2020. Methods for Assessing, predicting, and improving data veracity: A survey. ADCAIJ Adv. Distrib. Comput. Artif. Intell. J. 9, 5. https://doi.org/10.14201/ADCAIJ202094530

Atzori, L., Iera, A., Morabito, G., 2010. The Internet of Things: A survey. Comput. Netw. 54, 2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010

Azeroual, O., Saake, G., Schallehn, E., 2018. Analyzing data quality issues in research information systems via data profiling. Int. J. Inf. Manag. 41, 50–56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007

Batini, C., Scannapieco, M., 2016a. Data Quality Dimensions, in: Data and Information Quality, Data-Centric Systems and Applications. Springer, Cham, pp. 21–51. https://doi.org/10.1007/978-3-319-24106-7_2

Batini, C., Scannapieco, M., 2016b. Data and Information Quality, Data-Centric Systems and Applications. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-24106-7

Bello, M., Nápoles, G., Vanhoof, K., Bello, R., 2021. Data quality measures based on granular computing for multi-label classification. Inf. Sci. 560, 51–67. https://doi.org/10.1016/j.ins.2021.01.027

Bello, O., Zeadally, S., 2016. Intelligent Device-to-Device Communication in the Internet of Things. IEEE Syst. J. 10, 1172–1182. https://doi.org/10.1109/JSYST.2014.2298837

Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: Recent achievements and new challenges. Inf. Fusion 28, 45–59. https://doi.org/10.1016/j.inffus.2015.08.005

Berti-Équille, L., 2007. Measuring and Modelling Data Quality for Quality-Awareness in Data Mining, in: Guillet, F.J., Hamilton, H.J. (Eds.), Quality Measures in Data Mining, Studies in Computational Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 101–126. https://doi.org/10.1007/978-3-540-44918-8_5

Bijarbooneh, F.H., Du, W., Ngai, E.C.-H., Fu, X., Liu, J., 2015. Cloud-assisted data fusion and sensor selection for internet of things. IEEE Internet Things J. 3, 257–268. https://doi.org/10/f3rr7v

Byabazaire, J., O'Hare, G.M., Collier, R., Delaney, D., 2023a. IoT Data Quality Assessment Framework Using Adaptive Weighted Estimation Fusion. Sensors 23, 5993. https://doi.org/10.3390/s23135993

Byabazaire, J., O'Hare, G.M.P., Collier, R., Delaney, D., 2023b. IoT Data Quality Assessment Framework Using Adaptive Weighted Estimation Fusion. Sensors 23, 5993. https://doi.org/10.3390/s23135993

Byabazaire, J., O'Hare, G.M.P., Delaney, D.T., 2022. End-to-End Data Quality Assessment Using Trust for Data Shared IoT Deployments. IEEE Sens. J. 22, 19995–20009. https://doi.org/10.1109/JSEN.2022.3203853

C, N., PhamQuoc-Viet, N, P., DingMing, SeneviratneAruna, LinZihuai, DobreOctavia, HwangWon-Joo, 2022. Federated Learning for Smart Healthcare: A Survey. ACM Comput. Surv. CSUR. https://doi.org/10.1145/3501296

Casado-Vara, R., de la Prieta, F., Prieto, J., Corchado, J.M., 2018. Blockchain framework for IoT data quality via edge computing, in: Proceedings of the 1st Workshop on Blockchain-Enabled Networked Sensor Systems - BlockSys'18. Presented at the the 1st Workshop, ACM Press, Shenzhen, China, pp. 19–24. https://doi.org/10.1145/3282278.3282282

Chen, C.P., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. 275, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

Chen, M., Mao, S., Liu, Y., 2014. Big Data: A Survey. Mob. Netw. Appl. 19, 171–209. https://doi.org/10.1007/s11036-013-0489-0

Fantacci, R., Picano, B., 2020. Federated learning framework for mobile edge computing networks. CAAI Trans. Intell. Technol. 5, 15–21. https://doi.org/10.1049/trit.2019.0049

Fortino, G., Guerrieri, A., Savaglio, C., Spezzano, G., 2022. A Review of Internet of Things Platforms Through the IoT-A Reference Architecture, in: Camacho, D., Rosaci, D., Sarné, G.M.L., Versaci, M. (Eds.), Intelligent Distributed Computing XIV, Studies in Computational Intelligence. Springer International Publishing, Cham, pp. 25–34. https://doi.org/10.1007/978-3-030-96627-0_3

Glowalla, P., Balazy, P., Basten, D., Sunyaev, A., 2014. Process-Driven Data Quality Management – An Application of the Combined Conceptual Life Cycle Model, in: 2014 47th Hawaii International Conference on System Sciences (HICSS). Presented at the 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 4700–4709. https://doi.org/10.1109/HICSS.2014.575

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. Future Gener. Comput. Syst. 29, 1645–1660. https://doi.org/10.1016/j.future.2013.01.010

Hand, D.J., 2007. Principles of data mining. Drug Saf. 30, 621–622. https://doi.org/10.2165/00002018-200730070-00010

Hansch, G., Schneider, P., Fischer, K., Böttinger, K., 2019. A unified architecture for industrial IoT security requirements in open platform communications, in: 2019 24th Ieee International Conference on Emerging Technologies and Factory Automation (Etfa). IEEE, pp. 325–332. https://doi.org/10.1109/ETFA.2019.8869524

Hasan, S., Curry, E., 2015. Thingsonomy: Tackling variety in internet of things events. IEEE Internet Comput. 19, 10–18. https://doi.org/10.1109/MIC.2015.26

Heinrich, B., Hristova, D., Klier, M., Schiller, A., Szubartowicz, M., 2018a. Requirements for Data Quality Metrics. J. Data Inf. Qual. 9, 1–32. https://doi.org/10.1145/3148238

Heinrich, B., Klier, M., Schiller, A., Wagner, G., 2018b. Assessing data quality – A probability-based metric for semantic consistency. Decis. Support Syst. https://doi.org/10.1016/j.dss.2018.03.011

Hipp, J., Müller, M., Hohendorff, J., Naumann, F., 2007. Rule-Based Measurement Of Data Quality In Nominal Data., in: ICIQ. pp. 364–378.

Hu, H., Wen, Y., Chua, T.-S., Li, X., 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access 2, 652–687. https://doi.org/10.1109/ACCESS.2014.2332453

Jha, S., 2021. A big data architecture for integration of legacy systems and data (thesis). CQUniversity. https://doi.org/10.25946/16735342.v1

Karkouch, A., Mousannif, H., Al Moatassime, H., Noel, T., 2018. A model-driven framework for data quality management in the Internet of Things. J. Ambient Intell. Humaniz. Comput. 9, 977–998. https://doi.org/10/gd4hmg

Karkouch, A., Mousannif, H., Al Moatassime, H., Noel, T., 2016. Data quality in internet of things: A state-of-the-art survey. J. Netw. Comput. Appl. 73, 57–81. https://doi.org/10.1016/j.jnca.2016.08.002

Kim, S., Pérez-Castillo, R., Caballero, I., Lee, D., 2022. Organizational process maturity model for IoT data quality management. J. Ind. Inf. Integr. 26, 100256. https://doi.org/10.1016/j.jii.2021.100256

Kirchen, I., Schütz, D., Folmer, J., Vogel-Heuser, B., 2017. Metrics for the evaluation of data quality of signal data in industrial processes, in: 2017 IEEE 15th International Conference on Industrial Informatics (INDIN). Presented at the 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), pp. 819–826. https://doi.org/10.1109/INDIN.2017.8104878

Kiritsis, D., 2011. Closed-loop PLM for intelligent products in the era of the Internet of things. Comput.-Aided Des. 43, 479–501. https://doi.org/10.1016/j.cad.2010.03.002

Klein, A., Lehner, W., 2009. Representing Data Quality in Sensor Data Streaming Environments. J Data Inf. Qual. 1, 10:1-10:28. https://doi.org/10.1145/1577840.1577845

Kollolu, R., 2020. A Review on wide variety and heterogeneity of iot platforms. Int. J. Anal. Exp. Modal Anal. Anal. 12, 3753–3760.

Kuemper, D., Iggena, T., Toenjes, R., Pulvermueller, E., 2018. Valid.IoT: a framework for sensor data quality analysis and interpolation, in: Proceedings of the 9th ACM Multimedia Systems Conference. Presented at the MMSys ’18: 9th ACM Multimedia Systems Conference, ACM, Amsterdam Netherlands, pp. 294–303. https://doi.org/10.1145/3204949.3204972

Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Process. Mag. 37, 50–60. https://doi.org/10.1109/MSP.2020.2975749

Li, Y., Ma, S., Yang, G., Wong, K.-K., 2020. Secure localization and velocity estimation in mobile IoT networks with malicious attacks. IEEE Internet Things J. 8, 6878–6892. https://doi.org/10.1109/JIOT.2020.3036849

Li, Y., Wang, W., Yuan, W., Wei, Y., Zhu, J., 2023. Resource scheduling for smart IOT system quality management method research, in: Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023). Presented at the Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), SPIE, pp. 383–390. https://doi.org/10.1117/12.2689881

Liu, X., Tamminen, S., Su, X., Siirtola, P., Röning, J., Riekki, J., Kiljander, J., Soininen, J.-P., 2018. Enhancing veracity of IoT generated big data in decision making, in: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, pp. 149–154. https://doi.org/10.1109/PERCOMW.2018.8480371

Liu, Y., Yu, W., Rahayu, W., Dillon, T., 2023. An Evaluative Study on IoT Ecosystem for Smart Predictive Maintenance (IoT-SPM) in Manufacturing: Multiview Requirements and Data Quality. IEEE Internet Things J. 10, 11160–11184. https://doi.org/10.1109/JIOT.2023.3246100

Loshin, D., 2010. Data Quality Fundamentals 47.

Mansouri, T., Sadeghi Moghadam, M.R., Monshizadeh, F., Zareravasan, A., 2023. IoT Data Quality Issues and Potential Solutions: A Literature Review. Comput. J. 66, 615–625. https://doi.org/10.1093/comjnl/bxab183

Martín, L., Sánchez, L., Lanza, J., Sotres, P., 2023a. Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation. Internet Things 22, 100779. https://doi.org/10.1016/j.iot.2023.100779

Martín, L., Sánchez, L., Lanza, J., Sotres, P., 2023b. Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation. Internet Things 22, 100779. https://doi.org/10.1016/j.iot.2023.100779

Merino, J., Xie, X., Parlikad, A.K., Lewis, I., McFarlane, D., 2020. Impact of data quality in real-time big data systems.

Moke, K.C., Low, T.J., Khan, D., 2021. IoT blockchain data veracity with data loss tolerance. Appl. Sci. 11, 9978. https://doi.org/10.3390/app11219978

Mouha, R.A., 2021. Internet of Things (IoT). J. Data Anal. Inf. Process. 9, 77–101. https://doi.org/10.4236/jdaip.2021.92006
Mukherjee, A., Gazi, F., Pathak, N., Misra, S., 2022. AquaStream: Multihop Multimedia Streaming Over Acoustic Channel in Severely Resource-Constrained IoT Networks. IEEE Internet Things J. 9, 12085–12092. https://doi.org/10.1109/JIOT.2021.3133341
Nagalapatti, L., Mittal, R.S., Narayanam, R., 2022. Is your data relevant?: Dynamic selection of relevant data for federated learning, in: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7859–7867.
Namiot, D., 2015. On big data stream processing. Int. J. Open Inf. Technol. 3.
Navaz, A.N., Serhani, M.A., El Kassabi, H.T., Taleb, I., 2023. Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling. Sensors 23, 6443. https://doi.org/10.3390/s23146443
Pang, Z., Chen, Q., Tian, J., Zheng, L., Dubrova, E., 2013. Ecosystem analysis in the design of open platform-based in-home healthcare terminals towards the internet-of-things, in: 2013 15th International Conference on Advanced Communications Technology (ICACT). IEEE, pp. 529–534.
Pratap, A., Gupta, R., Siddhardh Nadendla, V.S., Das, S.K., 2019. On Maximizing Task Throughput in IoT-Enabled 5G Networks Under Latency and Bandwidth Constraints, in: 2019 IEEE International Conference on Smart Computing (SMARTCOMP). Presented at the 2019 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 217–224. https://doi.org/10.1109/SMARTCOMP.2019.00056
Ramaswamy, L., Lawson, V., Gogineni, S.V., 2013. Towards a quality-centric big data architecture for federated sensor services, in: Big Data (BigData Congress), 2013 IEEE International Congress On. IEEE, pp. 86–93.
Roman, R., Zhou, J., Lopez, J., 2013. On the features and challenges of security and privacy in distributed internet of things. Comput. Netw. 57, 2266–2279. https://doi.org/10.1016/j.comnet.2012.12.018
Scott, R., 2020. IoT data lifecycle adapts to AI at the edge | TechTarget [WWW Document]. IoT Agenda. URL https://www.techtarget.com/iotagenda/feature/IoT-data-lifecycle-adapts-to-AI-at-the-edge (accessed 9.29.23).
Shin, D.-H., Kim, G.-Y., Euom, I.-C., 2022. Vulnerabilities of the Open Platform Communication Unified Architecture Protocol in Industrial Internet of Things Operation. Sensors 22, 6575. https://doi.org/10.3390/s22176575
Sicari, S., Rizzardi, A., Grieco, L.A., Coen-Porisini, A., 2015. Security, privacy and trust in Internet of Things: The road ahead. Comput. Netw. 76, 146–164. https://doi.org/10.1016/j.comnet.2014.11.008
Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A., 2012. Data quality: A survey of data quality dimensions, in: 2012 International Conference on Information Retrieval Knowledge Management (CAMP). Presented at the 2012 International Conference on Information Retrieval Knowledge Management (CAMP), pp. 300–304. https://doi.org/10.1109/InfRKM.2012.6204995
Singh, P.K., Wierzchoń, S.T., Pawłowski, W., Kar, A.K., Kumar, Y. (Eds.), 2023. IoT, Big Data and AI for Improving Quality of Everyday Life: Present and Future Challenges: IOT, Data Science and Artificial Intelligence Technologies, Studies in Computational Intelligence. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-35783-1
Song, S., Zhang, A., 2020. IoT Data Quality, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3517–3518. https://doi.org/10.1145/3340531.3412173

Stankovic, J.A., 2014. Research directions for the internet of things. IEEE Internet Things J. 1, 3–9. https://doi.org/10.1109/JIOT.2014.2312291

Taleb, I., Serhani, M.A., Bouhaddioui, C., Dssouli, R., 2021. Big data quality framework: a holistic approach to continuous quality management. J. Big Data 8, 76. https://doi.org/10.1186/s40537-021-00468-0

Taleb, I., Serhani, M.A., Dssouli, R., 2019. Big Data Quality: A Data Quality Profiling Model, in: Xia, Y., Zhang, L.-J. (Eds.), Services – SERVICES 2019, Lecture Notes in Computer Science. Springer International Publishing, pp. 61–77.

Tedeschi, S., Rodrigues, D., Emmanouilidis, C., Erkoyuncu, J., Roy, R., Starr, A., 2018. A cost estimation approach for IoT modular architectures implementation in legacy systems. Procedia Manuf., Proceedings of the 6th International Conference in Through-life Engineering Services, University of Bremen, 7th and 8th November 2017 19, 103–110. https://doi.org/10.1016/j.promfg.2018.01.015

Teh, H.Y., Kempa-Liehr, A.W., Wang, K.I.-K., 2020. Sensor data quality: a systematic review. J. Big Data 7, 11. https://doi.org/10.1186/s40537-020-0285-1

Truong, H.-L., Karan, M., 2018. Analytics of performance and data quality for mobile edge cloud applications, in: 2018 IEEE 11th International Conference on Cloud Computing (CLOUD). IEEE, pp. 660–667. https://doi.org/10.1109/CLOUD.2018.00091

Vaziri, R., Mohsenzadeh, M., Habibi, J., 2019. Measuring data quality with weighted metrics. Total Qual. Manag. Bus. Excell. 30, 708–720. https://doi.org/10.1080/14783363.2017.1332954

Vermesan, O., Friess, P., 2014. Internet of things applications-from research and innovation to market deployment. Taylor & Francis.

Vermesan, O., Friess, P., 2013. Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems. River Publishers.

Wang, R.Y., 1998. A product perspective on total data quality management. Commun. ACM 41, 58–65. https://doi.org/10.1145/269012.269022

Wei, Z., Masouros, C., Liu, F., Chatzinotas, S., Ottersten, B., 2020. Energy- and Cost-Efficient Physical Layer Security in the Era of IoT: The Role of Interference. IEEE Commun. Mag. 58, 81–87. https://doi.org/10.1109/MCOM.001.1900716

Xiong, H., Pandey, G., Steinbach, M., Kumar, V., 2006. Enhancing data analysis with noise removal. IEEE Trans. Knowl. Data Eng. 18, 304–319. https://doi.org/10.1109/TKDE.2006.46

Yu, W., Liang, F., He, X., Hatcher, W.G., Lu, C., Lin, J., Yang, X., 2018. A Survey on the Edge Computing for the Internet of Things. IEEE Access 6, 6900–6919. https://doi.org/10.1109/ACCESS.2017.2778504

Zhang, Q., Li, L., Wang, T., Jiang, Y., Tian, Y., Jin, T., Yue, T., Lee, C., 2021. Self-sustainable flow-velocity detection via electromagnetic/triboelectric hybrid generator aiming at IoT-based environment monitoring. Nano Energy 90, 106501. https://doi.org/10.1016/j.nanoen.2021.106501

Zikopoulos, P., Eaton, C., 2011. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data.